Tim Beens

Dealing with competing requirements surrounding machine learning performance and explainability

Rijksdienst voor Identiteitsgegevens
Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

Dealing with competing requirements surrounding machine learning performance and explainability

Designing a decision support framework for choosing specific machine learning models while dealing with competing model requirements; A case study for the RvIG

Master thesis submitted to Delft University Of Technology in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

IN **Engineering and Policy Analysis**

FACULTY OF TECHNOLOGY, POLICY AND MANAGEMENT

by
Tim Beens
Student number: 4598024

To be defendend in public on August the 29th 2022

Thesis committee:	Dr. H. G. van der Voort	TU Delft, chair & supervisor
	Dr. N. Metoui	TU Delft, supervisor
	Drs. N. Mulder	RvIG, external supervisor
	Ir. C. Mostert	RvIG, external supervisor

Keywords: Machine learning, Competing model requirements, Decision support framework, Case study, Multivariate time-series analysis

August 9, 2022

Preface

This thesis marks the end of my student career and the start of a new chapter in my life. I still remember starting as a first-year TPM bachelor student, choosing the I&C track and always wanting more programming and modelling. No wonder I felt right at home within the EPA MSc program. Although most of my masters was spent online it did not prevent me from making some amazing new friends. Not only did they keep me sane during that first year working from home, they were also always ready to discuss whatever thesis related issue I would have at the time later on. Thank you for that Alma, Auriane, Anna, Sherman and Yasin. I would also like to give a special thanks to Fieke and Ragnhild for being my day-to-day study buddies without whom I would have never had the motivation to come to campus every day. Also a big shout out to both of their employee cards for providing me with much needed coffee and saving me a lot of money. Although I might have broken even by the amount of Coffeestar coffees I ordered in compensation...

I can of course not forgot the help of my supervisors, both from the RvIG and the university. I would therefore like to thank both Nita and Chris for their continuous support during my thesis and their helpful attitude with which they always reassured me that I could come to them with questions whenever I would have any. I would also like to thank my second supervisor, Nadia. Even though I did not approach you as often as I should have, I still very much appreciate your sharp feedback and critical looks at my thesis. Last but certainly not least, I would also like to thank my first supervisor Haiko for always being available for some quick feedback, be it through mail or in person, which really helped me out in the moments I would feel stuck or did not know what to do next. I also very much appreciated the casual conversations we would have and your belief in me even when I would doubt myself. It was a pleasure working together. A final thanks to my family for always supporting me, my friends from Vleeskeuring, and of course my rowing team Impils.

I still can not believe I forgot to request a passport in time for my summer holiday after working with identification documents for over half a year.

I hope you enjoy reading my thesis.

Tim Beens

Executive Summary

Research context
In the Netherlands, the supply of identification documents is based on a future demand prediction performed by the Rijksdienst voor Identiteitsgegevens (RvIG). They aim to predict the demand for at least five years ahead using monthly intervals. To more accurately predict the demand for identification documents there is a need for a new prototype prediction model. With a better prediction model, the supply and demand of identification documents can be better adjusted to one another and civilians will be more likely to receive their requested document on time. A better prediction of the future demand also enhances the accuracy of the financial budget made by the RvIG since it is partly based on the estimated demand for identification documents.

For the new prototype prediction model the use of Machine learning (ML) algorithms for multivariate time-series analysis is explored. Machine learning algorithms have a lot of potential since they have shown effective in many different fields dealing with similar problems. The main downside to machine learning algorithms is that more complex algorithms are often needed to reach better results which become increasingly more difficult to explain. This creates an issue in a context in which both model performance and model explainability are important, as is the case when exploring the possibilities for a new prototype prediction model for the RvIG. From the perspective of the RvIG the performance of the model is very important because it means more accurate predictions. However, since they are a government organisation and the use of (machine learning) algorithms lies under a magnifying glass ever since the child benefit scandal, they also need to make sure the algorithm(s) they use are explainable. This creates competing requirements for the new prototype prediction model which can be hard to fulfil since a more explainable model is usually less accurate and vice versa. Additionally, determining what algorithm to use can be very challenging since there exist a lot of different machine learning algorithms that each have their own unique level of performance and explainability. In order to help decision-makers this research, therefore, aims to develop a solution that can support decision-makers in their choice of an algorithm when dealing with competing model requirements.

Research methodology
To find a solution to the proposed problem the following main research question is formulated: "*How to deal with competing requirements regarding model explainability and performance while choosing what machine learning model to use for multivariate time-series analysis?*". To answer the main research question five distinct subquestions are formulated, these are:

- How can the term explainability be operationalised in order to assess different machine learning models?
- What machine learning algorithms are suitable for long-term multivariate time-series analysis?
- How do the different machine learning models score with regards to explainability?
- What predictive performance levels can be reached using the identified machine learning algorithms?
- How do competing requirements regarding explainability and performance influence the choice of multivariate time-series analysis models?

To answer these questions Hevner's design science research methodology approach is adopted which consists of three cycles that are continuously iterated upon while working on developing new designs. By combining the results of all five subquestions a final design is developed through which the main research question is answered.

The main structure of the research is as followed: First, the concept of explainability is operationalised and a list of potentially suitable machine learning algorithms for multivariate time-series analysis is identified. Afterwards the performance and explainability per model are analysed. To

conclude, the influence of competing requirements on the choice of model is explored and linked to specific model recommendations using different requirement scenarios. To answer the main research question the previous results are combined to develop a decision support framework.

Results

Based on state of the art scientific literature a novel model explainability assessment table is designed which can be seen in table 1. The assessment table consists of four distinct categories on which machine learning models can be assessed. By standardising the explainability assessment of the models the results between different models can be compared. Since assessing explainability is a relatively subjective task the categories use an ordinal rating system consisting of three different levels. If a category is scored High it is easily explainable while a low score indicates the category is hard to explain.

Table 1: Explainability assessment table

Criteria	Score (Low/Med/High)	Argument
Explainability of the feature importance		
Explainability of the general algorithmic structure		
Explainability of the performance criteria		
Explainability of the error minimalisation function		

Further literature research, focused on multivariate time-series analysis predictions using machine learning algorithms, identified six potentially suitable machine learning algorithms for the future demand prediction of identification documents. These are multiple linear regression (MLR), decision tree regression (DTR), random forest (RF), extreme gradient boosting (XGB), support vector regression (SVR) and multilayer perceptron (MLP) Regression. Assessing the algorithms using the explainability assessment table showed that each of the algorithms scores the same with regards to the explainability of the performance criteria since they all use the R-squared value. Further results show that Multiple Linear Regression and Decision Tree Regression are by far the most explainable. Following up are the Random Forest Regression and Extreme Gradient Boost algorithms. The least explainable models are Support Vector Regression and Multilayer Perceptron Regression. Although literature usually defines these last two algorithms as black boxes, the explainability assessment performed in this research shows that this is not necessarily valid. Although feature importance is indeed difficult to explain the other explainability criteria can still be explained.

The models are trained and tested using historic demand and feature data. Before applying the models hyperparameter optimisation is performed to increase their performance. The R-squared values for MLR and DTR were both negative which indicates that they are unfit to predict the future demand. The RF algorithm reached below-average results with R-squared values around 0.2. Both the XGB and SVR algorithm score roughly the same with R-squared values around 0.6-0.7. The MLP algorithm surprisingly reached underwhelming results with most predictions reaching only a negative R-squared value.

To answer the final subquestion three decision scenarios are sketched in which varying model implementation contexts and model requirements are discussed which resulted in three decision profiles. The scenarios and corresponding decision profiles show how the choice of algorithm changes depending on the varying contexts in which the algorithm is to be applied and the requirements that correspond with that. The decision profiles can be further used as a tool to help decision-makers determine their own requirements.

By combining the results of the five subquestions a final design is developed which functions as an answer to the main research question. Figure 1 shows the high-level decision support framework that is developed to support decision-makers with their choice for a machine learning model. The framework gives a high-level overview of the processes that need to be performed in order to determine which algorithm to use. The processes identified in each phase are further elaborated upon in the form of more detailed sub processes in the second part of the framework.

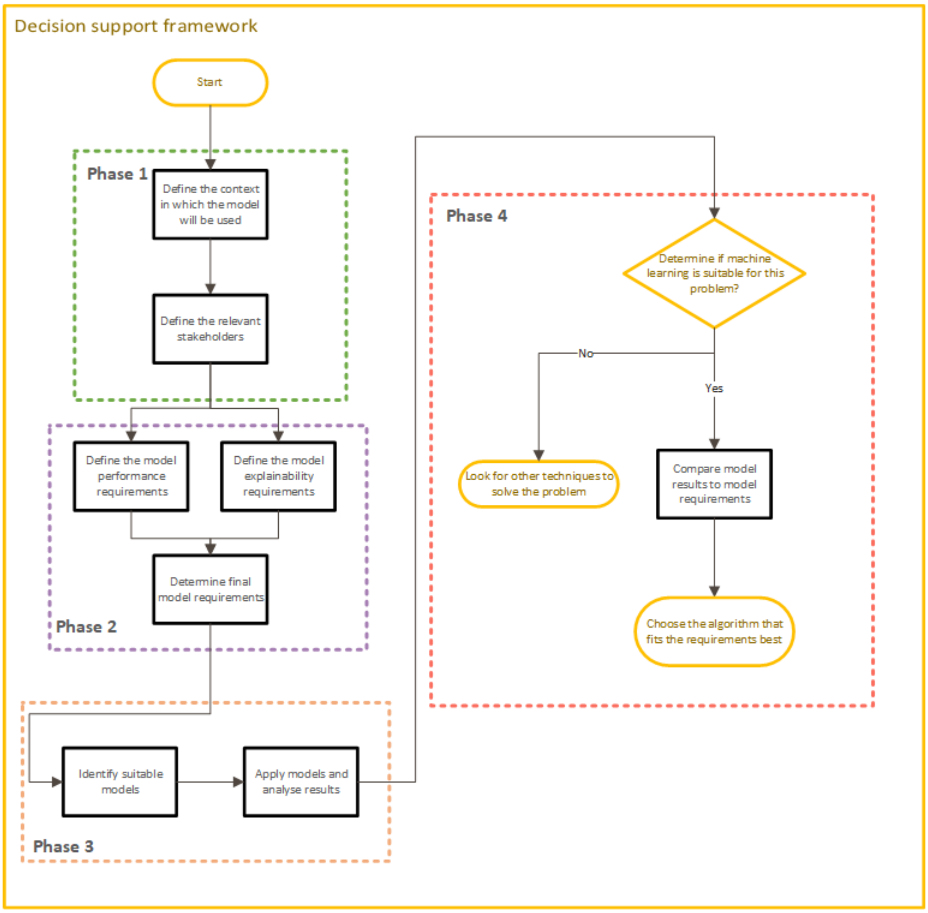

Decision support framework
Start
Phase 1
Define the context in which the model will be used
Define the relevant stakeholders
Define the model performance requirements
Define the model explainability requirements
Determine final model requirements
Phase 2
Identify suitable models
Apply models and analyse results
Phase 3
Phase 4
Determine if machine learning is suitable for this problem?
No
Yes
Look for other techniques to solve the problem
Compare model results to model requirements
Choose the algorithm that fits the requirements best

Figure 1: High-level decision support framework

Conclusion

Based on the model implementation context and the model requirements that have been identified for the RvIG they could be linked to the pragmatic performance profile. They want a combination of high performance of the model so that accurate predictions can be made and medium to high explainability so that trust in the results is created. These requirements fit best with the RF and XGB algorithms. However, since the model results showed that the RF algorithm performed quite poor the XGB algorithm is recommended instead even though it is slightly less explainable.

Recommendations

Future researchers willing to extend upon this research are suggested to work on one of the following three research suggestions:

- Researchers are recommend to further analyse the future demand prediction of identification documents using different data, algorithms, and hyperparameter optimisation methods
- Researchers are recommend to further develop the decision support framework to make it more accessible to people with a less technical background
- Researchers are recommend to analyse what the limitations to a standardised definition and application of explainability are and how these can best be tackled

Table of contents

List of Abbreviations **7**

List of Figures **8**

List of Tables **9**

1 Introduction **10**

1.1 Research context and problem statement . . . 10
1.2 Relevance of the research . . . 12
1.3 Structure of the report . . . 12

2 The future demand prediction of identification documents **14**

2.1 Introduction . . . 14
2.2 Case description . . . 14
2.3 Stakeholder requirements . . . 15

3 Machine learning, trust and explainability **17**

3.1 Introduction . . . 17
3.2 Machine learning versus traditional statistics . . . 17
3.3 The need for trustworthy machine learning . . . 17
3.4 Creating trust through explainability . . . 18
3.5 Defining explainability . . . 19
3.6 Formalizing explainabilty . . . 21
3.7 Assessment of machine learning algorithms . . . 22
3.8 Machine learning algorithms for times series analysis . . . 23

4 Model design process, implementation and prerequisites **25**

4.1 Introduction . . . 25
4.2 Model design strategy . . . 25
4.3 Programming language and packages . . . 26
4.4 Data sources . . . 26
4.5 Data cleaning methodology . . . 27
4.6 Data cleaning results . . . 28
4.7 Model implementation methodology . . . 30

5 Model implementation **32**

5.1 Introduction . . . 32
5.2 Multiple linear regression analysis . . . 32
5.3 Decision Tree, Random Forest, Gradient Boosting Machine . . . 33
5.4 Support Vector Regression . . . 36
5.5 Artificial Neural Networks . . . 38
5.6 Model hyper parameter optimisation . . . 39

6 Results **43**

6.1 Introduction . . . 43
6.2 Prediction results . . . 43
6.3 Future demand prediction using direct multi-step forecasting . . . 56
6.4 Conclusion . . . 57

7 Competing model requirements and the choice of algorithm **59**

7.1 Introduction . . . 59
7.2 Summary of the explainability and prediction results . . . 59
7.3 The influence of varying model requirements on the choice of algorithm . . . 60
7.4 Notes on model recommendations . . . 63
7.5 Model requirements by the RvIG . . . 64
7.6 Decision support framework . . . 64

7.7 Validating the framework . . . 67
7.8 Conclusion . . . 68

8 Conclusion, Discussion and Recommendations 69
8.1 Introduction . . . 69
8.2 Conclusion . . . 69
8.3 Discussion . . . 70
8.4 Recommendations . . . 73

References 75

A Survey results 80
A.1 Survey response Chris Mostert . . . 82
A.2 Survey response Nita Mulder . . . 82
A.3 Survey response Joshua Bannor . . . 82

B Prediction results 84
B.1 Multiple linear regression results . . . 84
B.2 Decision tree regression results . . . 85
B.3 Random forest regression results . . . 87
B.4 Extreme gradient boost regression results . . . 89
B.5 Support vector regression results . . . 91
B.6 Artificial neural network regression results . . . 93
B.7 Direct forecasting results . . . 95

List of Abbreviations

RvIG Rijksdienst voor Identiteitsgegevens 2

ML Machine learning 2

ID Identification card 14

MLR multiple linear regression 3

RF random forest 3

SVM support vector machine 36

SVR support vector regression 3

XGB extreme gradient boosting 3

MLP multilayer perceptron 3

DTR decision tree regression 3

RBF radial basis functions 24

CNN convolutional neural networks 24

AI artificial intelligence 18

XAI explainable artificial intelligence 21

MSE mean squared error 32

List of Figures

1 High-level decision support framework . . . 4
2 Overview of the different research elements . . . 13
3 Current step in the research process . . . 14
4 Current step in the research process . . . 17
5 Different target audiences for explainability (Barredo Arrieta et al., 2020) . . . 20
6 Overview of the different research elements . . . 25
7 The different design cycles as defined by Hevner (Hevner, 2007) . . . 26
8 Issued documents per type and validity period . . . 29
9 General implementation of a machine learning algorithm . . . 30
10 Current step in the research process . . . 32
11 Example of a decision tree (Charbuty & Abdulazeez, 2021) . . . 33
12 Example of a hyperplane dividing data in three dimensions (AYLIEN, 2022) . . . 36
13 Example of a hyperplane and its ϵ values for support vector regression (Sethi, 2020) 37
14 Example of a neural network (Lahiri & Ghanta, 2009) . . . 38
15 Current step in the research process . . . 43
16 Predicted 5 year valid identification cards using MLR . . . 45
17 Predicted 10 year valid identification cards using SVR . . . 45
18 Predicted 5-year valid identification cards using regression tree . . . 47
19 Predicted 10-year valid identification cards using regression tree . . . 47
20 Predicted 5-year valid identification cards using random forest . . . 49
21 Predicted 10-year valid identification cards using random forest . . . 49
22 Predicted 5 year valid identification cards using XGB . . . 51
23 Predicted 10 year valid identification cards using XGB . . . 51
24 Predicted 5 year valid identification cards using SVR . . . 53
25 Predicted 10 year valid identification cards using SVR . . . 53
26 Predicted 5 year valid identification cards using MLP . . . 55
27 Predicted 10 year valid identification cards using MLP . . . 55
28 Current step in the research process . . . 59
29 Different target audiences for explainability (Barredo Arrieta et al., 2020) . . . 60
30 Model decision profiles . . . 62
31 Decision support framework . . . 66
32 Decision support framework continued . . . 67
33 Current step in the research process . . . 69
34 Predicted number of 5 year valid passports using MLR . . . 84
35 Predicted number of 10 year valid passports using MLR . . . 85
36 Predicted number of 5 year valid passports using DT . . . 86
37 Predicted number of 10 year valid passports using DT . . . 87
38 Predicted number of 5 year valid passports using RF . . . 88
39 Predicted number of 10 year valid passports using RF . . . 89
40 Predicted number of 5 year valid passports using XGV . . . 90
41 Predicted number of 10 year valid passports using XGB . . . 91
42 Predicted number of 5 year valid passports using SVR . . . 92
43 Predicted number of 10 year valid passports using SVR . . . 93
44 Predicted number of 5 year valid passports using ANN . . . 94
45 Predicted number of 10 year valid passports using ANN . . . 94

List of Tables

1 Explainability assessment table . . . 3
2 Types of explainability . . . 20
3 Explainability assessment table . . . 23
4 Overview of identified datasets . . . 27
5 Combined datasets . . . 28
6 Explainability assessment table of Multiple Linear Regression . . . 33
7 Explainability assessment table of a Decision Tree . . . 34
8 Explainability assessment table of Random Forest . . . 35
9 Explainability assessment table of Gradient Boosting . . . 36
10 Explainability assessment table of SVR . . . 38
11 Explainability assessment table of an MLP . . . 39
12 Hyper parameter space for Decision Tree . . . 40
13 Hyper parameter space for Random Forest . . . 41
14 Hyper parameter space for XGBOOST . . . 41
15 Hyper parameter space for Support Vector Regression . . . 41
16 Hyper parameter space for Artificial Neural Network . . . 42
17 Dependent variable names . . . 44
18 Multiple linear regression prediction results . . . 44
19 Decision Tree Regression prediction results . . . 46
20 Random Forest Regression prediction results . . . 48
21 Extreme Gradient Boosting prediction results . . . 50
22 Support Vector Regression prediction results . . . 52
23 Multilayer Perceptron Regression prediction results . . . 54
24 Direct multi-step forecasting results for Support Vector Regression . . . 56
25 Summary of the explainability and performance assessment per model . . . 60
26 Multiple linear regression prediction results . . . 84
27 Decision Tree Regression prediction results . . . 85
28 Optimal hyper parameters used for Decision Tree Regression prediction . . . 86
29 Random Forest Regression prediction results . . . 87
30 Optimal hyper parameters used for Random Forest Regression prediction . . . 88
31 Extreme Gradient Boosting prediction results . . . 89
32 Optimal hyper parameters used for XGB Regression prediction . . . 90
33 Support Vector Regression prediction results . . . 91
34 Optimal hyper parameters used for SVR Regression prediction . . . 92
35 Multilayer Perceptron Regression prediction results . . . 93
36 Direct forecasting results for MLR . . . 95
37 Direct forecasting results for DTR . . . 95
38 Direct forecasting results for RF . . . 96
39 Direct forecasting results for XGB . . . 96
40 Direct forecasting results for SVR . . . 96
41 Direct forecasting results for MLP . . . 97

1 Introduction

1.1 Research context and problem statement

Identification documents serve as a "hidden" backbone of modern-day developed countries. In the Netherlands alone at least 62% of the population has a passport and 49% of the population has a so-called NIK (Dutch identification card) (Brummelkamp, Hoevenagel, & Witkamp, 2020). Although double possession of documents is included in these numbers, they still show that more than half the Dutch adult population has at least one of either two documents. The widespread use and possession of identification documents is also acknowledged by the Dutch government who aims to continually develop their identification system (Brummelkamp et al., 2020).

On average Dutch citizens use their identification documents 8.2 times a year (Brummelkamp et al., 2020). The specific use of identification documents can vary widely. It ranges from voting at the national elections to accepting a package at a local pick-up point. It also includes opening a new bank account, renting a car, setting up a new telephone plan, registering at a (new) hospital/doctor, travelling abroad or when switching jobs just to name a few (Brummelkamp et al., 2020). The use of identification documents is thus widely spread throughout Dutch society.

The widespread use of identification documents within Dutch society also means that a lack of availability can have a significant impact on people's lives. A lack of supply of identification documents can mean that someone won't be able to travel, set up a new bank account, change jobs, or do any of the other previously mentioned examples. Furthermore, Dutch law demands that citizens can always identify themselves and risk a fine if they are not able to (Rijksoverheid, 2022). Last but not least a lack of identification documents might also negatively impact the image of the Dutch Ministry of Internal affairs since they are responsible for the supply of identification documents (Mulder, 2022). At the same time, an oversupply of identification documents is also not desired since they might end up not being used. Under normal circumstances this does not pose a significant issue, however, the moment a change is made to the physical appearance of any of the identification documents all the previously made, non personalised documents, become invalid and have to be thrown away. This means a significant loss of tax money that could have been spent elsewhere (Mulder, 2022).

In the Netherlands, the supply of identification documents is based on a demand prediction performed by the Rijksdienst voor Identiteitsgegevens (RvIG) which is part of the Ministry of Internal affairs. They aim to predict the demand for identification documents for at least five years ahead using monthly intervals. To more accurately predict the demand for identification documents there is a need for a new prediction model. With a better prediction model the supply and demand of identification documents can be better adjusted to one another and civilians will be more likely to receive their requested document on time. Additionally, a better prediction of the future demand also enhances the accuracy of the financial budget made by the RvIG since it is partly based on the estimated demand for identification documents (Mulder, 2022).

Within the scientific literature little to no prior scientific research regarding the future demand prediction of identification documents exists. However, a lot of research has been done in which machine learning (ML) models are effectively used to predict or classify variables. Some examples of this are in the banking sector (Hu et al., 2021), medical sector (Bi, Goodman, Kaminsky, & Lessler, 2019) and even the media sector (Portugal, Alencar, & Cowan, 2018) just to name a few. Due to the historic success of machine learning models for prediction problems the use of machine learning models for the future demand prediction of identification documents is explored during this research.

However, although machine learning as a method has proven to be successful, the lack of previous (empirical) research means many challenges still exist that can hinder the successful implementation of machine learning algorithms for this specific prediction problem. For example, to capture the complexity present within while predicting the future demand for identification documents multiple time dependent predictor variables are needed. This means the prediction problem is a multivariate time-series analysis prediction problem instead of a simple univariate time-series analysis problem. Further complexity arises due to the dependency between the future demand and policy changes made by the parliament and Ministry of Internal Affairs. Historic demand data shows that policy changes can have a huge influence on demand. This suggests that future policy changes will also affect the demand, however, what the exact influence on the demand will be is very hard if not impossible to predict.

Apart from the general model implementation challenges, machine learning algorithms also bring forth challenges concerning competing requirements regarding model performance and explainability. Depending on the specific machine learning algorithm used, the explainability of the model can vary from relatively simple to extremely complex. The more complex a model is the harder it becomes to explain. However, the results of more complex models are usually more precise than those of less complex models. Here competing requirements occur as to what is preferable in a model. For the RvIG a precise future demand prediction is preferred so that they get a good overview of the estimated future demand. To do this a complex, harder to explain model is likely needed. On the other hand, the Dutch government, especially after the child benefits scandal, requires the use and development of (machine learning) algorithms to be transparent and explainable to prevent future scandals (Mulder, 2022). This means the used model also has to be explainable and should thus be less complex.

For the RvIG to reach a successful implementation of a new prototype prediction model they need to successfully manage the competing requirements. If the predictive performance of the model is too low the model provides no added value. If the used model is too complex and can not be properly explained anymore the use of the model might get cancelled by the government in fear of future scandals. Determining what algorithm to use can however be very challenging. Many machine learning algorithms exist and each of them has its own level of explainability and performance. Additionally, clear guidelines on when to use what machine learning algorithm are lacking. To streamline the decision-making process regarding the use of machine learning algorithms this research therefore aims to develop a solution that can support decision-makers in their choice of an algorithm when dealing with competing interests. An added difficulty concerning this arises in and around the fact that the knowledge levels of the people that are intended to use the model, manage the model and the people that are responsible for the end results of the model are likely all different. Therefore, what is seen as explainable depends per person. This has to be taken into account when working with explainability.

This research is done in the context of a case study for the Dutch Rijksdients voor Identiteitsgegevens, specifically the financial department who are interested in a new prediction model for identification documents. By developing a new prototype prediction model an overview of suitable multivariate time-series analysis models is generated which can be further assessed on their explainability and performance. The results of this analysis not only add to the development of a new prototype prediction model but also add valuable information for developing the solution to assist future decision makers dealing with similar competing requirements. Additionally, the model design process performed in this research can serve as a handhold on how to develop and apply similar models in the future. The aim of this research is formalised through the use of the following research question:

"*How to deal with competing requirements regarding model explainability and performance while choosing what machine learning model to use for multivariate time-series analysis?*"

To answer the main research question formulated above the following subquestions are formulated as well:

- How can the term explainability be operationalised in order to assess different machine learning models?

- What machine learning algorithms are suitable for long-term multi-variate time-series analysis?
- How do the different machine learning models score with regards to explainability?
- What predictive performance levels can be reached using the identified machine learning algorithms?
- How do competing requirements regarding explainability and performance influence the choice of multivariate time-series analysis models?

By answering the first four subquestions a clear overview is created of the explainability and performance levels of different machine learning models. By combining these results with the new insights in how competing requirements influence the choice of multivariate time-series analysis models a framework can be developed that can support decision-makers in their choice of a model when dealing with competing requirements. Through this framework, the main research question can be answered.

1.2 Relevance of the research

The relevance of this research comes forth in multiple ways as is shown by the following bullet points.

- The research has societal relevance in the sense that a better prediction model can prevent the under and oversupply of identification documents which prevents many negative consequences for both Dutch citizens and the Ministry of Internal affairs.
- The research has specific relevance for the RvIG because a better future demand prediction means a more precise financial budget can be developed.
- The research has scientific relevance since no previous machine learning implementation exists for this specific subject and it deals with a multivariate time-series analysis problem in which the aim is to perform multi-step forecasting on a monthly timescale which is a relatively small research area.
- By developing the framework as part of a case study empirical knowledge is included in the design which adds added value compared to a framework which is only based on theory. The developed framework should therefore be more relevant and usable in practice.

1.3 Structure of the report

This report is structured as follows: chapter two consists of a case study description that dives deeper into the intricacies of developing a new prototype prediction model. Chapter three consists of a literature review in which the relevant literature with regards to machine learning models and explainability is discussed. Chapter four explains the model design process and prerequisites such as data collection and cleaning. Chapter five goes over the identified machine learning algorithms and their hyperparameter optimisation while also assessing them based on their explainability. Chapter six applies the different algorithms to the prediction problem and describes their results. Chapter seven reflects on the model results concerning the competing requirements present for the decision maker and provides a novel solution on how to deal with this. Chapter 8 concludes the report with a conclusion, discussion and recommendations. Figure 2 gives a visual overview of the different elements that will be discussed in this research.

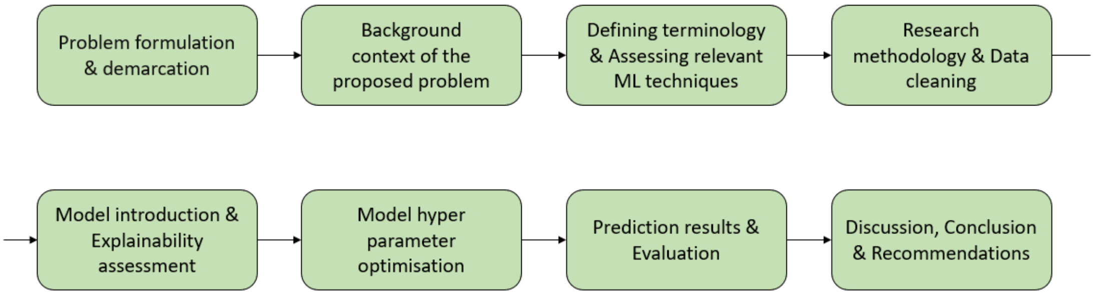

Problem formulation & demarcation
Background context of the proposed problem
Defining terminology & Assessing relevant ML techniques
Research methodology & Data cleaning
Model introduction & Explainability assessment
Model hyper parameter optimisation
Prediction results & Evaluation
Discussion, Conclusion & Recommendations

Figure 2: Overview of the different research elements

2 The future demand prediction of identification documents

2.1 Introduction

This section further elaborates on the case study proposed in the introduction. The case study performed for this research looks at the development and implementation of a new prototype prediction model for the future demand prediction of identification documents. Since the RvIG deals with competing requirements surrounding model performance and explainability when working with algorithms this case also proves to be very suitable for developing a decision support framework dealing with this issue. Given that findings from the case can be used as input for the framework. Additionally, where most organisations are relatively secretive when discussing their use of algorithms the RvIG is quite open which is not seen very often. This provides a lot of opportunities to learn about the considerations that are made when working with algorithms in general and machine learning models specifically which makes this a very interesting and unique case. The section starts with a general introduction to the case and its intricacies and ends with a concise discussion of the stakeholder requirements proposed by employees of the RvIG. Figure 3 shows the research element this section discusses.

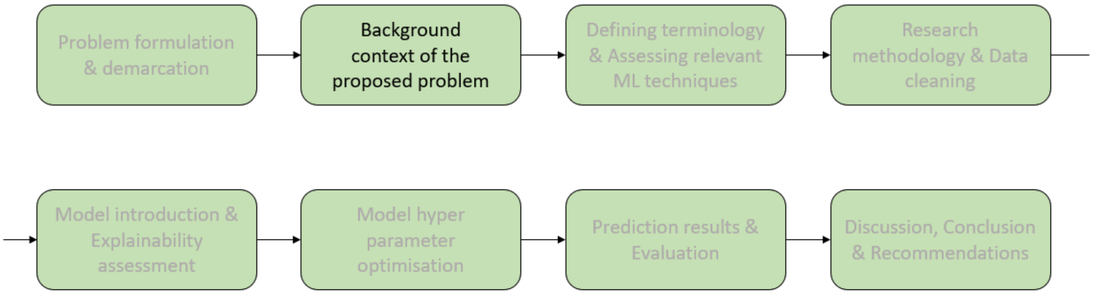

Problem formulation & demarcation
Background context of the proposed problem
Defining terminology & Assessing relevant ML techniques
Research methodology & Data cleaning
Model introduction & Explainability assessment
Model hyper parameter optimisation
Prediction results & Evaluation
Discussion, Conclusion & Recommendations

Figure 3: Current step in the research process

2.2 Case description

In the Netherlands, the Rijksdienst voor Identiteitsgegevens (RvIG), part of the Dutch Ministry of Internal affairs, is tasked with predicting the demand for identification documents. Predicting the demand for identification documents means predicting the demand from the entire Dutch population which poses many challenges given the plethora of uncertain factors involved. Many hypotheses regarding the predictor variables are present, however little empirical evidence exists. Factors such as the Dutch economic development are expected to play an important role since a thriving economic environment means citizens have more money to spend and are thus able to go on international holidays more often which requires a valid identification document. However, this relatively simple factor already poses multiple challenges on its own. Predicting economic development is uncertain and even if the direction of the economic development is known its direct influence on the demand for identification documents is unclear. The latter is made even more difficult due to the fact the demand prediction for identification documents consists of two distinct categories, namely the demand for passports and the demand for Identification card (ID). Ideally, the demand for both documents is predicted individually since both require a different document type.

The influence of political decision-making, or rather the influence of changing policy with regards to identification documents is also a challenging factor. Policy changes are pushed by the Dutch first and second chambers together with the Ministry of Internal affairs and have occurred multiple times over the past decades. Policy changes usually affect the use and validity period of identification documents. Some examples are the new legislation starting in 2014 which made it mandatory for youths aged 14 and above to be able to identify themselves, abolishing the possibility of adding children to the passport of their parent(s), and changing the validity period of both passports and

ID cards from 5 to 10 years (Mulder, 2022). The last change is only applicable to adults which means that youths still have documents valid for only five years. Each of these changes has had a significant impact and long-lasting influence on the demand for identification documents as is shown by historic demand data (Mulder, 2022).

A policy change usually leads to a peak in demand just before or after the policy change is implemented depending on the type of change. This peak then reoccurs every 5 or 10 years depending on the validity period of the document requested. Similar to economic developments political changes are also challenging to predict. There is a small window of opportunity to estimate the influence of a policy change when the new policy is announced but not yet implemented. However, new policies are not always known on time, especially given the time horizon desired for the demand prediction which is 5 to 10 years. Furthermore, even if a new policy is known to be implemented its exact impact on the demand remains uncertain. This lack of insight into the effects of policy changes forces certain decisions when developing a prediction model.

Challenges also occur due to the responsibility that the Ministry of Internal affairs and thus the RvIG bears towards Dutch citizens. Due to them being part of the Dutch government extra care is needed to prevent the occurrence of privacy concerns, misuse of data, or irresponsible design in general. Given the development of a new prototype prediction model, this means extra emphasise is laid on the need for trustworthy algorithms. To ensure this the used algorithms and the model design process need to be transparent, explainable, interpretable, and comprehensible where possible. This adds an extra dimension to the model design process since it means a trade-off between the performance of the model and its explainability. Extra emphasis on explainability also brings forth questions with regard to whom the model must be explainable, interpretable, etc... since explaining a machine learning model to a machine learning expert is easier than explaining it to civilians without any machine learning background.

Given the governmental nature of the RvIG and their need for a new prediction model this case fully encompasses the trade-off between explainability and performance while working with machine learning algorithms. On the one hand, the model needs to be able to make accurate predictions since otherwise it does not add any added value. On the other hand, the RvIG needs to be able to explain the new prediction model to a certain degree in order to create trust in the prediction model and to ensure that it's reliable. Thus explainability is needed. Internally the model design process adds further scientific relevance since it requires special design choices due to the many uncertain factors involved.

2.3 Stakeholder requirements

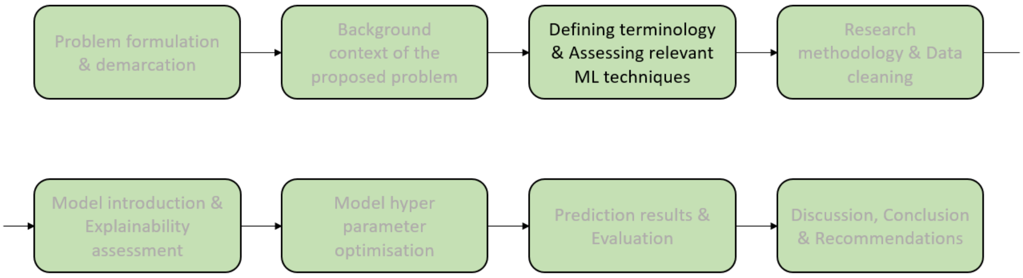
To ensure the developed model has practical value a small survey is performed. Four employees who either work with the old prediction model or the new prototype model are sent a list of questions consisting of three parts. The first part of the questions goes over some general aspects with regards to their job function and their relation to the current model. The second part of questions asks questions about the current model and its implementation and the third part goes over potential requirements for a new prototype prediction model. A full discussion of the answers can be found in appendix A. The following gives a brief overview of the common requirements proposed by the respondents.

Based on the survey responses some initial requirements for the proposed model can already be identified. The consensus among stakeholders is that data should, where possible, not contain any personal data and if it does it has to be aggregated in order not to be able to be traced back to an individual person. The specific decision-making process and features used by an algorithm should be explainable where possible. If this is not possible clear explanations with regards to the limits and workings of the used algorithm should be given. Where possible insights should be given into the relationship between input data and output data. This is important for the end user to explain to stakeholders how the model has come to certain results. Furthermore, transparency and explainability of the model design process are seen as important to answer potential questions posed by the house of representatives and WOB requests. If possible a graph showing the difference between the prediction and the actual demand is seen as a bonus. Lastly, the model should be kept as simple as possible to ensure it is used since not everyone is familiar with the Python programming language.

3 Machine learning, trust and explainability

3.1 Introduction

The following chapter starts by discussing the difference between traditional statistics and machine learning. This is followed by an extensive literature discussion regarding the need for trust in machine learning algorithms and how this can be achieved by explainable machine learning models. Afterwards, a new definition of explainability is proposed and an explainability assessment table is formalised. The section ends by discussing the state-of-the-art literature regarding machine learning models for future demand prediction. Figure 4 shows the research element this section discusses.



Problem formulation & demarcation
Background context of the proposed problem
Defining terminology & Assessing relevant ML techniques
Research methodology & Data cleaning
Model introduction & Explainability assessment
Model hyper parameter optimisation
Prediction results & Evaluation
Discussion, Conclusion & Recommendations

Figure 4: Current step in the research process

3.2 Machine learning versus traditional statistics

Policymakers increasingly rely on scientists when dealing with prediction problems. Models are used to predict natural phenomena such as hurricanes and earthquakes, but also when predicting the effect of human-induced problems such as the effect of climate change. According to Sarewitz and Pielke (1999) prediction serves two roles in modern-day societies. Prediction is used to test a scientific hypothesis by comparing what is expected to occur with what really occurs and it is used as a tool to support (political) decision-making. In the latter, prediction models are used because decision-makers lack the knowledge to make accurate predictions themselves such as is the case when predicting the demand for Identification documents.

When dealing with prediction problems the use of machine learning techniques is becoming more and more popular. Machine learning algorithms are for example used for are water demand prediction (Anele, Hamam, Abu-Mahfouz, & Todini, 2017), electricity demand prediction (Ghalehkhondabi, Ardjmand, Weckman, & Young, 2017), demand forecasting in transport (Tsekeris & Tsekeris, 2015) and sales prediction in fashion or electronics (Beheshti-Kashi, Karimi, Thoben, Lütjen, & Teucke, 2015). The increasing use of machine learning techniques should be of no surprise given the growing popularity of machine learning over the last 20 to 30 years (Hu et al., 2021). Machine learning algorithms are algorithms that "learn" from data by modifying their actions to attain more accuracy. This makes them very suitable for tackling complex real-world problems (Alzubi, Nayyar, & Kumar, 2018). Compared to other techniques such as traditional statistics, machine learning distinguishes itself due to its focus on prediction whereas traditional statistics focus on the inference between variables (Bzdok, Altman, & Krzywinski, 2018). Given the purpose of this research, a machine learning approach is thus most suitable.

3.3 The need for trustworthy machine learning

Although machine learning algorithms are very good at recognising patterns in data and making predictions, they are not a miracle solution to all of societies problems. Machine learning algorithms suffer from issues regarding transparency, bias, and explainability which influence the trust people have in the predictions made by them (Das & Rad, 2020).

Trust is needed for people to be able to accept the predictions made by machine learning algorithms. Within the context of the case study performed for this research, this specifically means that the person intended to use the new model needs to have enough trust in the results to base the new financial budget upon them. Without trust, it does not matter how well an algorithm can predict since it will likely mean that the machine learning algorithm will not be used for decision-making anyway.
Simultaneously, too much trust in machine learning algorithms can lead to issues as well. For example in cases where the machine learning algorithm is trained on biased data. In such cases, the corresponding predictions are likely to be biased themselves and should not be used without a proper review (Lee, 2021) (Schmidt & Biessmann, 2019).

An adequate amount of trust in machine algorithms is thus an important aspect for the successful implementation and future use of machine learning algorithms. Too little trust in the algorithm means the prediction results won´t be used, while too much trust in the algorithm can mean potential harmful results if the algorithm is based on biased data.

3.4 Creating trust through explainability

Most researchers agree that to create trust an algorithm needs to be interpretable, explainable and comprehensible. However, no clear consensus consists about which specific criteria are needed to create trust and especially not which definitions to use in order to interpret these criteria (Dosilovic, Brcic, & Hlupic, 2018). de Bruijn, Warnier, and Janssen (2021) for example assume that explainability creates transparency and trust in artificial intelligence (AI), whereas Dosilovic et al. (2018) instead argue that the interpretability of a model is important to create trust. Dosilovic et al. (2018) further mention that the terms interpretability and explainability are often used interchangeably in literature while they do not necessarily mean the same thing. According to them, interpretability refers to the ability to explain or present in understandable terms what a machine learning algorithm does to humans. Explanation on the other hand refers to the collection of features of the interpretable domain that have led to a specific decision of the machine learning algorithm. To add to the confusion they continue by stating that both the terms comprehensibility and transparency are used as a synonym for interpretability in literature. Doran, Schulz, and Besold (2017) on the other hand define interpretability as 'interpretable systems'. A system where a user cannot only see but also study and understand how inputs are mathematically mapped to outputs. This implies model transparency and requires a level of understanding of the technical details of the mapping. This definition is in line with the explainable AI definition from Dosililovic but is now used for interpretability instead of explainability.

As has become apparent, a lack of consensus exists about which definitions to use when working with these terms. Due to the lack of industry standards, it is unclear as to what concretely adds to trustworthy AI which makes the terms hard to apply in practise. Krishnan (2019) even uses the lack of industry standards and definitions as an argument to reason that these criteria might not even be useful at all. According to her the definitions of these criteria often use other undefined terms which "move the bump under the rug". In other words, the inherent vagueness of these definitions means that they substitute one undefined term with another which means that they remain hard to apply to machine learning algorithms. She furthermore questions if understanding the decision-making progress adds any added value. She compares the decision-making progress of neural network algorithms to that of human decision-making in such a way that she explains it by neurons shooting electrical signals at each other. On a very specific level, this is indeed how humans make decisions, but it does not tell us anything useful about how a decision has been made. Thus understanding the decision-making process of neural networks might also tell us nothing. One can argue about whether or not this specific comparison is valid, but it does pose an interesting perspective. She furthermore states that these criteria are often a means to an end and not an end in themselves. Therefore she claims that it is important to look further than just these criteria and check if it is possible to reach goals such as non-discrimination and fairness in other ways as well. In the end, she concludes by acknowledging that these criteria can be useful when trying to find causality between inputs and outputs and also to create trust.

Since trust is an essential part for the use and implementation of machine learning algorithms and no other approaches to create trust currently exist, the lack of industry standards regarding the different definitions is accepted. However, given the aim of this research to assess the explainability of ML models and their performance only the term explainability is extended upon. Explainability is chosen because it is seen as one of the most important characteristics of machine learning algorithms when talking about trust. According to Burkart and Huber (2021) explainability serves as a prerequisite for creating trust because it enables people to understand how an algorithm has come to a decision which in turn gives people more confidence in the results. Siau and Wang (2018) underline this by stating that a lack of explainability influences the amount of trust in the application. Explainability has further value in automatic decision-making systems where humans are not in the decision-making loop anymore. In such cases, explanations of how model decisions are made are essential to create trust in the decisions made by these types of algorithms (Barredo Arrieta et al., 2020). Lastly, apart from trust, explainability also adds to increased accountability, it enables adjustments to the used algorithms and it helps to create fair and ethical decision-making (Burkart & Huber, 2021).

3.5 Defining explainability

Similar to the terms transparency, interpretability, comprehensibility, and understandability the definition of the term explainability also struggles with a lack of industry standards. Within the literature two main trends with regards to the definition of explainability exist. The first interpretation of explainability sees explainability as something the model itself is capable of. It refers to the extent to which a model can generate (useful) information that helps in understanding the algorithm and its results. A simple example of this is the ability of a decision tree algorithm to show which features were the most important to come to a certain prediction (Burkart & Huber, 2021). In this interpretation, explainability is thus not tied to how well a model can be explained, but to how well a model can generate extra information that helps in understanding it afterwards. This type of explainability will further be referred to as indirect explainability. The second interpretation of explainability sees explainability as the ability to explain how a model works, and especially to explain how a model has come to a decision in an understandable way. This form of explainability is seen as essential in cases where a model can predict most cases correctly but fails on a few data instances. This interpretation of explainability is further used when trying to validate a model before its actual implementation and deployment since it allows understanding of the underlying decision-making process and thus creates trust in its results (Burkart & Huber, 2021). This interpretation of explainability will further be referred to as direct explainability.

When working with indirect explainability, the term interpretability is usually used to assess the general understanding of an algorithm instead of (direct) explainability. Interpretability is then seen as the extent to which a model is inherently understandable to humans Barredo Arrieta et al. (2020). However, when working with direct explainability, interpretability is seen as a part of explainability. Gilpin et al. (2019) states that a good explanation needs to be both interpretable and complete. Here interpretability refers to the extent to which a human can understand an explanation while completeness refers to the accurate description of the system to be explained. However, since a more complete explanation usually becomes more complex it reduces the interpretability which means a trade-off exists between the interpretability of an explanation and its completeness. Take for example a deep neural network. It can be explained as a black box that uses an input and creates an output. This explanation of a deep neural network is easily interpretable, however, it is far from complete as it lacks a lot of details. To make this explanation more complete at least part of the underlying mechanisms used by the deep neural network to come to a prediction needs to be elaborated upon which inherently makes the explanation more complex and less interpretable. The different forms of explainability and the trade-off between concise and complete explanations are recapped in table 2. Note that none of the combinations of explainability and completeness is inherently better than the others. Each of them has its pros and cons and their use depends on the situation and the goal of the researcher.

Table 2: Types of explainability

	Direct explainability	Indirect explainability
Concise explanation	Easily understandable, but lacks depth	Effective as long as generated information is useful and understandable
Complete explanation	Most suitable when explaining to a target audience with background knowledge	Effective when trying to give detailed insights into the mechanisms of an algorithm

Since this research aims to assess different machine learning algorithms on their explainability and performance the concept of indirect explainability is deemed too abstract compared to direct explainability in which a clear trade-off between interpretability and completeness is mentioned. Together with the use of direct explainability, the focus of this research lies at giving complete explanations because those are expected to create the most trust.

Lastly, when dealing with the explainability of machine learning algorithms, specific emphasis must be given to the question of explainability for whom? Within machine learning, and most advanced fields in general, the level of knowledge someone has about a subject can differ greatly from person to person. What is understandable to one person is not always necessarily understandable to another. Think for example of the knowledge difference between a machine learning researcher and a beginning student. A clear overview of potentially different target groups and their motivation for wanting to understand what a ML algorithm is doing is given by Barredo Arrieta et al. (2020) and is shown in figure 5.

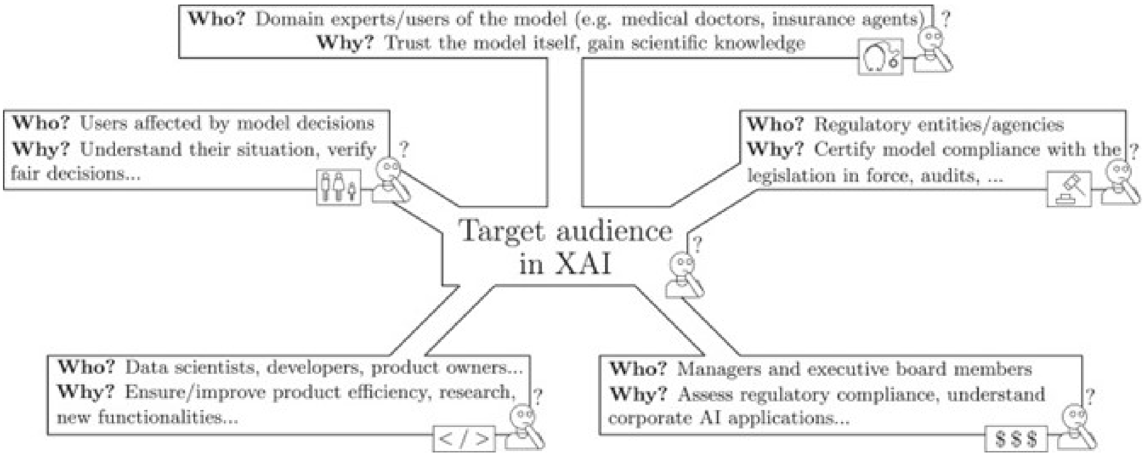

Who? Domain experts/users of the model (e.g. medical doctors, insurance agents)
Why? Trust the model itself, gain scientific knowledge
Who? Users affected by model decisions
Why? Understand their situation, verify fair decisions...
Who? Regulatory entities/agencies
Why? Certify model compliance with the legislation in force, audits, ...
Target audience in XAI
Who? Data scientists, developers, product owners...
Why? Ensure/improve product efficiency, research, new functionalities...
Who? Managers and executive board members
Why? Assess regulatory compliance, understand corporate AI applications...
</>
$$$

Figure 5: Different target audiences for explainability (Barredo Arrieta et al., 2020)

The figure shows that depending on the target audience not only does the level of knowledge differ, but also the information that has to be explained changes. This clearly shows the impact the ‘who?’ Question has on the result of explainability since different persons have different reasons for wanting to understand a model.

Based on the different target audiences shown in the figure this research focuses the level of explainability to that of the domain experts and users of the model as well as managers and executive board members. The first group is chosen because it is assumed that in order to effectively use a model it is helpful to understand how it works, at least to a certain degree. The second group is chosen because they are the stakeholder that is accountable if required. To give adequate answers about the model they, therefore, need to understand how it works. Since the model is developed by the data scientist it is assumed that they already understand how the model(s) work.

Given the choice for complete as possible direct explanations and the previously defined target audience, the following definition is formulated based on Gilpin et al's research which will be used in the remainder of the research.

"Explainability is the extent to which the underlying mechanism, feature importance, and results of a machine learning algorithm can be explained in an interpretable manner to a pre-defined target audience without having to make significant concessions with regards to the completeness of the explanation"

Thus, algorithms that can be explained in an interpretable manner while the explanations remain relatively complete are seen as explainable while algorithms that require explanations that need to sacrifice a lot of their completeness in order to remain interpretable are seen as less explainable. However, since the aforementioned definition is subjective in the sense that something can be interpretable by one person and not by another specifying the target audience is very important. When a specific target audience is specified a level of background knowledge can be assumed to which the explanation can be fitted. However, further research and general scientific developments are needed in the field of explainable artificial intelligence (XAI) to operationalise explainability further. Since this is still an ongoing research area the following subsection dives deeper into the general literature regarding criteria for proper explanations and what is meant by explaining something in an interpretable way to make explainability more quantitative.

3.6 Formalizing explainabilty

The underlying reason for formalising explainability lies in the fact that the term explainability remains relatively abstract even with the definition formulated in the previous subsection. Assessing different machine learning algorithms on their explainability using the given definition therefore also remains abstract. Thus to be able to properly assess the explainability of different algorithms more concrete criteria are needed. This also enables the possibility of comparing the different algorithms on their explainability and performance which is central for this research.

As a start, Carnap's criteria of adequacy can be used as inspiration when looking for criteria to operationalise explainability. Carnap's criteria define four aspects which an explanation should adhere to (Dutilh Novaes & Reck, 2015). The first criteria 'similarity to the explicandum' states that the explanation of a concept should be similar to the original concept, preferably in such a way that the explanation can substitute the original concept in most cases where it has been previously used. The second criteria 'exactness' refers to how accurately the explanation describes the subject to be explained. Exactness can be compared to the completeness of an explanation which has been mentioned previously. The third criteria 'fruitfulness' can be seen as the extent to which the explanation is interpretable and thus useful. Here the same trade-off as for completeness and interpretability occurs. A very exact explanation will likely display the concept to be described very well, but if the subject being explained for example consists of very complex maths it is likely not fruitful. As is the case when an explanation is complete but not interpretable. One can give the exact mathematical function that explains a concept, but without any added explanation a layperson will not be able to understand what it represents. Lastly, there exist the 'simplicity' criteria which states that once the previous criteria are met the simplest explanation should be used (Sovrano, Sapienza, Palmirani, & Vitali, 2022).

Looking at the ideas of Carnap with regards to criteria for an explanation especially the trade-off between exactness and fruitfulness stands out. It is almost similar to the trade-off between completeness and interpretability which underlines the importance of an adequate balance between the two. The idea of simplicity is also roughly mentioned in (Burkart & Huber, 2021) where it is mentioned that concise explanations are more interpretable than long and extensive explanations. Furthermore, Miller (2019) states that explanations containing probabilities are less important than explanations containing information about causality since most humans have a hard time dealing with uncertainty. He also states that humans prefer contrasting explanations. People do not just like to know why result A occurred, but also why result B did not occur. Lastly, Miller mentions that explanations are ‘social’. Meaning that an explanation works better if it fits within the beliefs of the person the concept is being explained to. The latter is inherently subjective, but combined with the selected target audience certain beliefs can be assumed.

Zhou, Gandomi, Chen, and Holzinger (2021) categorize explanations in rationale explanations, data explanations and safety and performance explanations. Here rationale explanations mostly deal with the why questions behind machine learning techniques I.E. how does a model come to a decision? Data explanations specifically deal with the feature importance question which is used to clarify which features are most important for a model when making predictions. The last category deals with the safety and design decisions that are made during the entire model development process. Although important, this criteria is not directly applicable to specific ML model techniques and will therefore not be included. Zhou et al. (2021) also mentions more quantitative metrics that can be used to assess a model’s explainability. By using model complexity as an indicator for explainability, where more complex models are harder to explain than less complex models, he mentions that model size, runtime operation counts, main effect complexity (the influence that each feature has on the output) and the interaction strength between features are important model characteristics because they influence the complexity. Lastly, Sovrano et al. (2022) mention whether or not a model (or its elements) can only be explained through text or also through other types of media such as image, whether or not a models output is inherently understandable or needs extra clarification and whether or not the model requires any prior experience/does it assume a certain level of knowledge/ does it need extra clarification to be able to understand what is happening?

3.7 Assessment of machine learning algorithms

Using the findings from the literature discussed in the previous subsection two elements regarding the explainability of machine learning algorithms can be directly derived from (Zhou et al., 2021). Their statements regarding rationale and data explanations add a lot of value with regards to understanding machine learning algorithms and will therefore be included in the assessment table as criteria regarding the explainability of the feature importance and the explainability of the general algorithmic structure. Based on Sovrano et al. (2022) statements regarding the importance of understanding the output of a model another explainability assessment criteria can be derived namely the explainability of the performance criteria. Understanding a prediction or classification results is usually not too complex, however, understanding the performance metric associated with such predictions or classifications can sometimes be quite difficult to understand and can potentially hinder the implementation of these types of algorithms. To keep track of this it is added to the assessment table. The final criteria, the explainability of the error minimisation function, is loosely based on Miller (2019) statements regarding the importance of contrasting explanations. By understanding the error minimisation function it becomes clearer why sometimes result A is reached while other times result B is reached. Understanding where these differences come from should enhance the understanding of the machine learning model as a whole which is why it is seen as an important aspect of understanding ML models in general and added to the assessment table.

Not all literary findings, specifically those from Dutilh Novaes and Reck (2015), Burkart and Huber (2021) and Sovrano et al. (2022), can be directly translated to explainability assessment criteria. However, they still provide a lot of useful information regarding the quality and inherent understandability of explanations. So although they are not directly used in the explainability assessment table, their findings will be kept in mind while assessing the different algorithms on their explainability and determining their explainability score.

With the assessment criteria finalised an explainability assessment table is developed as can be seen in 3. The assessment table consists of three columns. The first column contains the explainability criteria on which the algorithm is to be assessed, the second column contains the score which is based on an ordinal scale consisting of three values: Low, Medium, and High. Here low refers to a criteria that is hard to explain and high means the criteria can be easily explained. The last column contains the argumentation corresponding to the criteria and its score. By operationalising the term explainability into four distinct criteria the algorithms to be used during this research can each be assessed in a way that allows them to be compared to each other. The results of the explainability assessment analysis can be combined with the models and their predictive performances to draw a conclusion regarding their practical use for this specific prediction problem.

Table 3: Explainability assessment table

Criteria	Score (Low/Med/High)	Argument
Explainability of the feature importance		
Explainability of the general algorithmic structure		
Explainability of the performance criteria		
Explainability of the error minimalisation function		

Additionally, scoring the algorithms on these criteria further adds scientific value in general because it creates a more nuanced perspective regarding the complexity of machine learning algorithms. Currently, most literature very quickly labels more complex algorithms as black-box algorithms while most of them usually aren't fully black-box at all. This creates a wrong image of machine learning algorithms which can hinder their implementation, especially in the context of real-life applications where understanding part of the algorithm might already be enough to create trust in the results but is not even considered because the algorithm is labelled as a black box. This research, therefore, argues that complex algorithms should not always be immediately labelled as black boxes, especially when their underlying mechanisms can still be explained. Instead one has to be honest and clear about the limits of different algorithms and let the end user decide whether it can be useful or not. It might indeed not always be possible to determine which specific feature was used for a certain prediction, but if this limitation is clear and there is trust in the underlying decision-making logic this does not always have to be an issue, especially if the corresponding prediction is a lot better than that of a more explainable algorithm.

3.8 Machine learning algorithms for times series analysis

Time series analysis is a prediction method that uses historical data that is gathered periodically in order to predict future values. Depending on whether or not the prediction problem contains linear or non-linear relationships different algorithms are needed. When dealing with non-linear relationships more complex algorithms such as machine learning algorithms are needed. Machine learning techniques can be used for time series analysis by transforming the time series data in such a way that it is compatible with supervised machine learning algorithms (Brownlee, 2019). This usually entails shifting the data set so that data from a previous observation is used to predict the next observation.

There exist many different machine learning algorithms for times series analysis, however, they can generally be categorised in short-term, medium term and long-term forecasting algorithms (Selvin, Vinayakumar, Gopalakrishnan, Menon, & Soman, 2017). Current literature with regards to demand prediction focuses mainly on short to medium-term predictions. These predictions usually range from minutes to weeks but can sometimes also include monthly predictions. Short to medium-term predictions are usually more suitable for even more complex algorithms such as recurrent neural networks, deep neural networks, and long short-term memory neural networks since these algorithms become less accurate the further away in time the prediction tends to be. This means different machine learning algorithms are needed for long-term predictions.

Marcelino, de Lurdes Antunes, Fortunato, and Gomes (2019) use random forest techniques for the prediction of pavement health in the next five to ten years. However, Abedi et al. (2021) show that long-term prediction (up to 5 years) is also possible using different machine learning algorithms such as logistic regression, Extreme Gradient Boosting, Gradient Boosting Machine, Support Vector Machine, and Decision Trees for recurrent stroke prediction. Elkamel, Schleider, Pasiliao, Diabat, and Zheng (2020) use multiple regression analysis and convolutional neural networks (CNN) for energy demand prediction in Florida for the upcoming 24 months. The use of CNN's is special because they do not take into account historic patterns. This is useful when short-term trends are not representative of the real demand curve. Because of this they sometimes work better than traditional neural network algorithms that do take historic data into account.

Research from Pinho, Costa, Silva, and Furtado (2018) argue that neural networks such as multilayer perceptrons (MLP) and radial basis functions (RBF) work well when data is non-linear. They applied both to a case of telecom demand behavior where they used 3 to 6 months of data to predict the demand one month ahead. Their results showed that the MLP was able to most accurately predict the demand. Although the prediction period in the example is smaller than what is applicable in this research, the results show potential for future medium-long term demand prediction.

Since there are no clear guidelines regarding which algorithm to use for long-term predictions this research will explore the effectiveness of several different techniques. The machine learning techniques in question are chosen based on the results discussed in the previous paragraphs. The chosen algorithms are Decision Tree Regression, Support Vector Regression, Random Forest, Gradient Boosting Machine, Multilayer Perceptron, and Multiple Regression Analysis. All of these algorithms are suitable options for non-linear prediction problems, however, their effectiveness for long-term predictions remains to be seen due to a lack of empirical research. Research comparing the effectiveness of different machine learning algorithms for standard prediction problems, however, does often come to the conclusions that more complex algorithms such as SVR and ensemble methods such as RF and XGB tend to perform well in most cases as is shown by Osisanwo et al. (2017), Caruana and Niculescu-Mizil (2006), Parreco et al. (2019), Deist et al. (2018) and Uddin, Khan, Hossain, and Moni (2019). These algorithms are therefore expected to perform relatively well when used for the future demand prediction of identification documents. Caruana and Niculescu-Mizil (2006) however also state that even though some of the models they analysed performed better than others this is no guarantee that they always perform better. Similarly, it can be possible for the algorithms that score worse on average to still perform very well given the right prediction problem.

4 Model design process, implementation and prerequisites

4.1 Introduction

The following section starts by discussing the model design strategy. Afterwards, the used programming language and important packages are briefly discussed. The section continues by describing the identified data sources and the steps taken to clean the data. The section ends by showing some preliminary results based on the data cleaning and by describing the general implementation steps that are needed to implement most machine learning algorithms. Figure 6 shows the research element this section discusses.

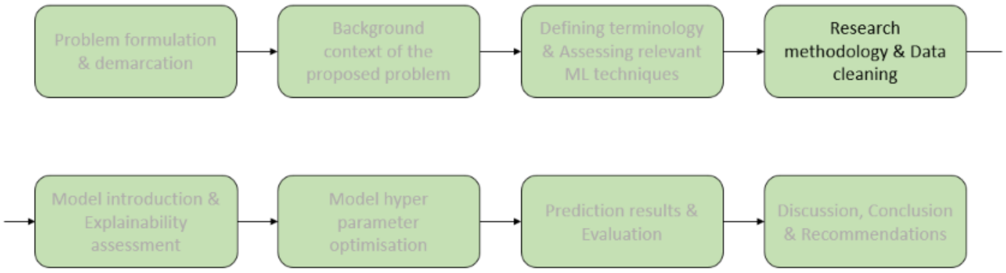

Problem formulation & demarcation
Background context of the proposed problem
Defining terminology & Assessing relevant ML techniques
Research methodology & Data cleaning
Model introduction & Explainability assessment
Model hyper parameter optimisation
Prediction results & Evaluation
Discussion, Conclusion & Recommendations

Figure 6: Overview of the different research elements

4.2 Model design strategy

This research will adopt the design research approach as defined by Hevner (2007). A design science research approach is adopted because the aim is to develop a new decision support framework from scratch. This inherently means something new is designed which requires a design science approach. Additionally, to develop a successful design one has to learn by doing which requires a lot of trial and error I.E. design iterations. Since Hevner's design science approach is based on different iteration cycles it fits well with the research at hand.

Hevner's approach consists of three main cycles that each fulfil an important aspect of design science. The first cycle, called the relevance cycle, initiates the design science research by identifying relevant opportunities and problems within an actual application domain. It then defines what new artefact should be developed and continues by defining the metrics through which the new artefact can be evaluated. It ends with a reflection of the added benefit of the new artefact by performing field testing. If the field test shows a mismatch between the required artefact and the developed artefact or the results of the evaluation metrics are not satisfactory the relevance cycle starts a new iteration with the purpose of finding a solution to the identified disparities by incorporating feedback from the field test and adjusting research requirements were needed.

The second cycle is called the rigor cycle. This cycle uses a combination of scientific literature, expertise from domain experts and knowledge from existing artefacts to ensure the innovation of the developed artefact. This is done through the relevant selection and application of theories and methods for constructing and developing the artefact. A key element within the rigor cycle is that the research results add back to the existing knowledge base by for example extending upon existing theories or methods.

The final cycle, the Design cycle, combines the results from the two previous cycles to develop a new and relevant artefact. A key element within this cycle is the vast amount of iterations that are made based on continuous evaluation of the artefact until a satisfactory design is achieved. Hevner stresses the importance of a proper balance between constructing and evaluation of the artefact. He states that developed artefacts without proper evaluation lead to insufficient results with regards to design research. During the design process, it is therefore important to continuously iterate over the artefact by constructing and evaluating.

In short, the relevance cycle identifies the relevance and requirements of the to build artefact based on the problems and opportunities in the actual application domain. The rigor cycle builds upon this by using expert knowledge, previous artefacts and literature to define a new and innovative artefact. The design cycle combines everything by developing a new artefact based on the requirements defined in the relevance cycle and the innovative theories and methods from the rigor cycle. Figure 7 shows an overview of the different cycles.

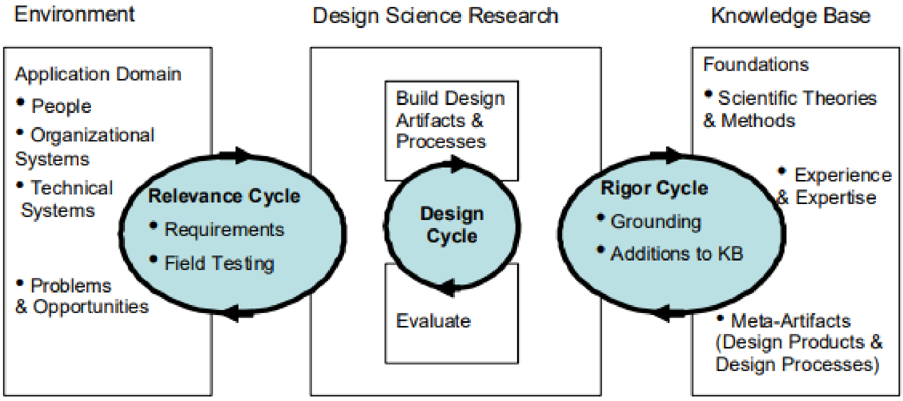


Figure 7: The different design cycles as defined by Hevner (Hevner, 2007)

4.3 Programming language and packages

To apply the machine learning algorithms identified in the previous chapter and perform data cleaning the Python programming language is used. Python is chosen because it is one of the easiest languages to work with, has a lot of online support and has access to multiple powerful packages for both data cleaning and applying machine learning algorithms. For data cleaning the Pandas package is used. This package is ideal for reading in CSV files and working with them. For the application of the different machine learning algorithms the SKlearn package is used. The SKlearn package provides access to algorithms such as decision tree regression, multiple linear regression, random forest regression, support vector regression and multi-layer perceptron regression. Apart from providing access to the different algorithms the SKlearn package also provides access to multiple useful functions needed for the implementation of these algorithms. Some examples are the scalar functions used to scale the input data, the train test split function to divide the feature and dependent data into training and test sets and the cross-validation functions which can be used for hyperparameter optimisation.

4.4 Data sources

In order to use machine learning algorithms data with predictive capabilities is needed. For this research data is collected based on discussions with domain experts within the RvIG who used their expertise in the field of identification documents to suggest different data sources (Mulder, 2022). An overview of these sources together with a small description and their source is shown in table 4.

Table 4: Overview of identified datasets

Dataset name	type of data	source
Age distribution Netherlands	Number of people per age from 0 till 105 for the years 2010 till 2034	(CBS, 2022a)
Adjusted Gross Disposable Income	Adjusted gross disposable income data per year from 2012 till 2019	(Eurostat, 2022)
Approaches of domestic product GDP National accounts	GDP data per year for the Netherlands from 2000 till 2020	(CBS, 2022b)
People with a driver's license per age category	Number of people with a driver's license divided over age categories for 2014 till 2022	(CBS, 2022c)
Dutch Holidays	Number of national and international holidays by Dutch citizens per year from 2017 till 2020	(CBS, 2022d)
Issued documents data	Demand per document type from 2011-04 till 2022-03	Mulder (2022), RvIG
Expired documents data	expired documents per document type from 2016-04 till 2032-03	Mulder (2022), RvIG
Lost documents data	lost documents per document type from 2016-04 till 2022-03	Mulder (2022), RvIG
Possessed documents data	possession data per document type from 2016-05 till 2022-03	Mulder (2022), RvIG

Although the datasets regarding the identification documents (issued, expired, lost and possessed documents) are provided by the RvIG they are not 100% correct. Due to some rounding issues when retrieving the data it can be possible that a row value contains a document too many or too little. Since it is impossible to retrieve these errors and their impact is very small, monthly document values are in the 10's of thousands if not more than 100 thousand, this error is accepted while working with the data. However, it is mentioned in order to remain transparent about the data that is used.

The issued documents data is the dependent data that will be used in this research. This data consists of four variables representing IDs with a five and ten year valid period and passports with a five and ten year valid period. Since the future demand for each of these four variables is important a separate ML model is trained for each of them so that they can each be predicted. Regardless of the specific document type that is predicted the same feature data is used.

4.5 Data cleaning methodology

Since machine learning algorithms learn to make decisions based on data the quality of the used data is very important. The higher the quality of data the likelier it is that the algorithm predicts better, but vice versa, the worse the data quality is the worse the results will likely be (Vogelsang & Borg, 2019).

To ensure that the data is of sufficient quality data cleaning is performed. To do this, all data sets are checked for missing values and potential outliers. Additionally, all datasets containing yearly data values are interpolated to generate monthly values. To do this a linear relationship between data values is assumed. In practice, this means that the monthly values increase or decrease with the same value every month. Although this method is likely not the most representative of real life, a more detailed relationship can not be modelled due to a lack of detailed data and knowledge. Lastly, the data provided by the RvIG contains a minor error which has to be cleaned. Identification documents with a validity period of ten years should only exist from march 2014 and onwards since the new policy regarding these types of documents went live in march 2014, however, they sometimes already occur before that date in the data provided by the RvIG. These occurrences regard a very small number of documents and should be removed since these types of documents did not exist yet.

Apart from having to clean the datasets internally, the datasets also need to be combined. As can be seen in table 4 most datasets contain data from different periods that only partly overlap with each other. Ideally one wants to have one big data set that contains all data together, however since most machine learning algorithms are not capable of handling missing values, different datasets have to be created. To make the most out of the data at hand three different datasets are created. Each of the three datasets aimed to cover as much time as possible while also using different feature variables. The resulting datasets can be seen in table 5.

Table 5: Combined datasets

Dataset	variables	data availability
Dataset 1	Date, GDP, PPS, Population, Drivers license possession	2014-2018
Dataset 2	Date, GDP, Population, Drivers license possession, Number of total, summer, winter and (inter)national holidays, Document possession, Expired documents and lost documents	2017-2019
Dataset 3	Date, Population, Possession of drivers licences, Document possession, Expired documents and lost documents	2016-2022

On top of that, the date variable is transformed from a string value to an ordinal value so that it can be used as input for the algorithm and lagged data is added to incorporate historic values. The lagged data consists of historic values from t-1, t-2 and t-3 time steps before. Lastly, the dependent data is normalised to prevent variables with large values from overshadowing variables with small values.

4.6 Data cleaning results

After cleaning the data some interesting patterns can already be spotted in graphs of the dependent variables. Figure 8 shows the number of issued documents per type and validity period. From 2014 and onwards the validity period of passports and identification cards was changed from five to ten years (Plasterk, 2014). The effect of this policy change can be seen in the figure. Before 2014 everyone, both minors and adults received the same documents. This explains why the peaks in the number of issued documents are a lot higher before 2014 than after 2014. From 2014 onwards only minors receive passports and IDs with a five-year validity period and since there are a lot less minors than adults the demand is also much lower than before. The number of five-year valid issued documents remains relatively constant after 2014 since the number of minors does not change much, however, the number of issued documents does show a periodic peak roughly once a year. Since these peaks occur roughly around June and July every year they can likely be explained by the summer holiday period. When people travel abroad they need a valid identification document which explains the increased demand during summer.

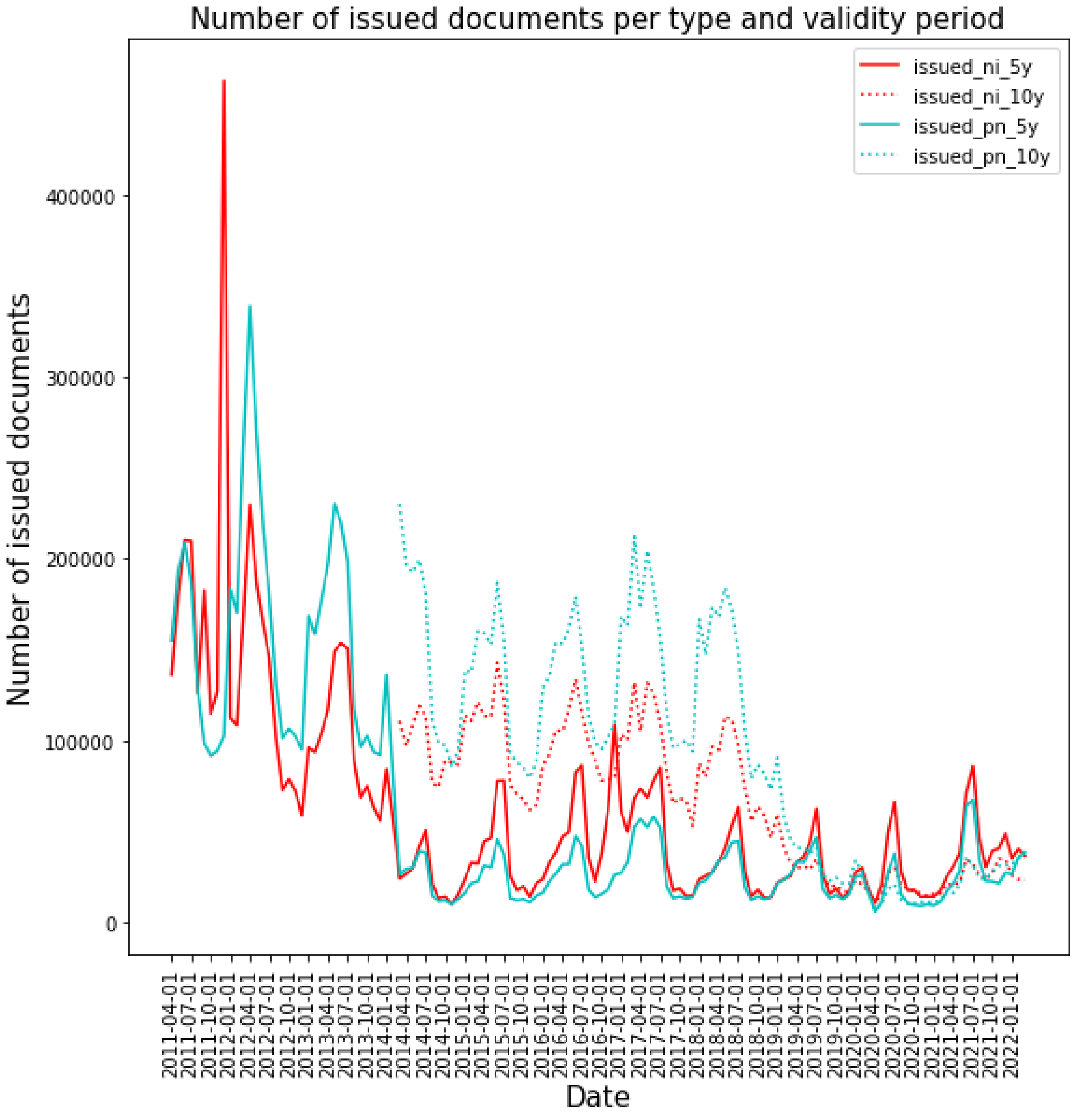


Figure 8: Issued documents per type and validity period

The number of issued documents with a validity period of ten years is shown by the red and blue dotted lines in the figure. Both these lines start from 2014 since that is when the new policy went into effect. Similar to the number of issued documents with a five-year validity period a yearly periodic peak occurs during summer. Here the same line of reasoning also goes. People mostly go abroad during summer for which they need a valid identification document and thus the number of issued documents increases just before/during summer. Whereas the number of issued documents with a five-year validity period remains relatively constant throughout the years the number of issued documents with a validity period of ten years drops quite drastically after 2019. Although this might look quite strange, the explanation is relatively simple. Although the new policy went live in March 2014 a lot of people still had five-year valid identification documents from before that time. Since these all expire between 2014 and 2019 there is a relatively constant demand for new documents during this period as well. After 2019 however, every adult citizen should have a ten-year valid document. Since the earliest date these have been distributed is March 2014, the earliest date adult citizens will need a new ten-year valid document is March 2024. Thus there is a drop in demand between 2019 and 2024. The only demand left during this period is that of minors that have become 18. This also means that from 2024 to 2029 a similar number of issued documents is expected to those that have occurred between 2014 and 2019. Lastly, It is of course possible for adults with a ten-year valid identification document to lose their document which would require them to get a new one in the period between 2019-2024, however, these numbers are relatively small and thus do not make up a large part of the demand.

4.7 Model implementation methodology

The implementation of different supervised machine learning algorithms each follows a similar pattern. First, the data has to be split into a training and a test dataset. This goes for both the dependent and independent data. The model is then fitted to the training data where it learns to transform the input data to a correct output value. The fitted model is then applied to the test data to check how well it can predict the dependent variable using new input data. This is done for each of the different models for each of the different datasets. Initially, this is done using the default parameter settings of the algorithms as defined by SKlearn. Afterwards, cross-validation is applied to tune the hyperparameters and find the best scoring values. For the hyperparameter optimisation through cross-validation a grid search technique is applied. Here the hyperparameters are varied using pre-defined values of which the corresponding model result is kept track of. After applying cross-validation the parameter values with the best model results are adapted. Since each algorithm has its own unique hyperparameters the specific grid search values are specified when performing the hyperparameter optimisation in the next chapter. Lastly, the results are checked for potential over and underfitting. Once finished the results are plotted to show the prediction results and the actual data. Depending on the algorithm some visualisation with regards to the decision structure are also plotted. Given the prediction results of the different algorithms, the assessment of the performance and explainability trade-off is continued and a small reflection is given. The results of this analysis are then shown to the RvIG for a final reflection and a conclusion from their point of view. An overview of the different processes is shown in figure 9.

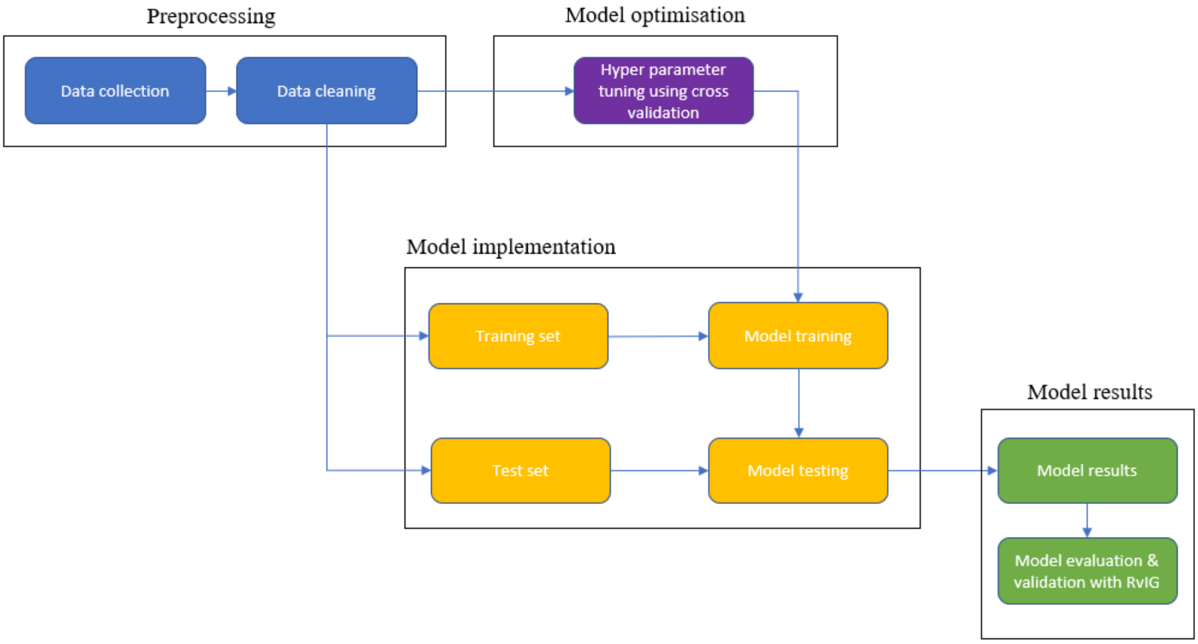


Figure 9: General implementation of a machine learning algorithm

The initial training and testing of the models is done using historic data. Here only the theoretical performance of the models is analysed. For actual future demand predictions future feature data is required which does not exist. To tackle this problem two options exist. Firstly, each of the feature values can be extrapolated assuming a univariate relationship between the date variable and the feature value. The extrapolated values can then function as future values for the future demand prediction of identification documents. The second option is to use direct multi-step forecasting. In normal forecasting feature data from month x is used as training data to predict the dependent variable in that same month. In direct multi-step forecasting each model is trained using 'historic'/'current' feature data to predict the dependent variable in the next time step. By training the model to predict the dependent variable in t+n time steps ahead using historic data the feature data from the latest T-n months can be used to predict the future. Direct multi-step forecasting is considered more reliable in comparison to extrapolating the feature values since the feature values themselves are often quite complex as well and can likely not be properly predicted using only the date value. Therefore direct multi-step forecasting is applied instead.

5 Model implementation

5.1 Introduction

In this section the workings of the different machine learning techniques are elaborated upon. First, a general description of their mechanisms is discussed after which they are assessed based on their explainability. To conclude, model hyperparameter optimisation is performed to find the best hyperparameters per machine learning technique. The results of this chapter are used for the final demand prediction of identification documents which is done in the next chapter. Figure 10 shows the research element this section discusses.

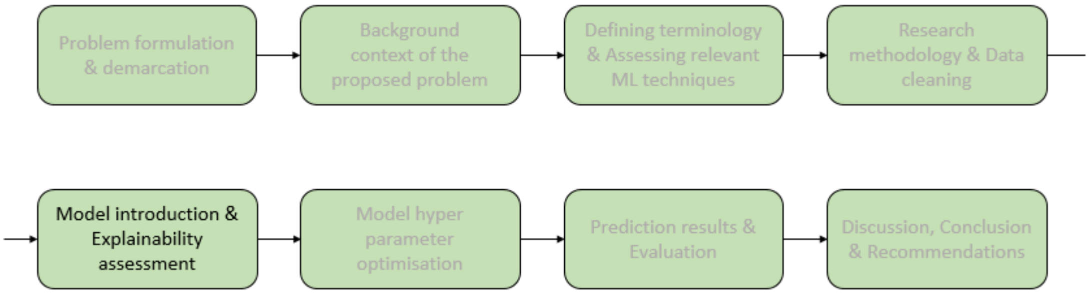

Problem formulation & demarcation
Background context of the proposed problem
Defining terminology & Assessing relevant ML techniques
Research methodology & Data cleaning
Model introduction & Explainability assessment
Model hyper parameter optimisation
Prediction results & Evaluation
Discussion, Conclusion & Recommendations

Figure 10: Current step in the research process

5.2 Multiple linear regression analysis

Multiple linear regression is a type of regression analysis in which multiple independent variables are used to predict a single dependent variable. This is different from normal- and polynomial linear regression where only a single independent variable is used (Ghani & Ahmad, 2010). The formula for multiple linear regression follows the same structure of the linear regression formula except that there are x1, x2, x3 ... xn predictor variables that each have their own slope value (B1, B2, B3 ... Bn) (Eberly, 2007). A simple example formula looks as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \; ... + \beta_n x_n + \epsilon$$

Similar to normal linear regression β_0 refers to the y-intercept and ϵ refers to the potential error with the actual y value. To optimise the slopes for the different predictor variables an optimisation problem is solved. Since it is a regression problem this is usually done by minimising the mean squared error (MSE). This is the average error between the real y value and the predicted value overall predictions. By changing the slopes of the prediction variables the mean squared error changes. This creates an optimisation problem which when solved results in a formula for which the MSE is lowest and the prediction is thus optimised.

Based on the previous description of the multiple linear regression algorithm the explainability assessment is performed. The results are shown in table 6. Across all criteria the Multiple Linear Regression scores well with regards to explainability.

Table 6: Explainability assessment table of Multiple Linear Regression

Criteria	Score (Low/Med/High)	Argument
Explainability of the feature importance	High	The coefficients corresponding to each variable indicate which variables were most important for the prediction
Explainability of the general algorithmic structure	High	A function is fit on the input data that finds a combination of features and coefficients that can best predict the dependent variable
Explainability of the performance criteria	Medium	R-squared variable is used which needs some background knowledge to understand
Explainability of the error minimisation function	High	The least squares method is used which means varying the coefficients so that the difference between the predicted and real value is minimised

5.3 Decision Tree, Random Forest, Gradient Boosting Machine

Decision trees are a type of machine learning algorithm in which feature values are compared to a threshold value in a series of tests. The series starts from a so-called root node I.E. start value and goes down through a series of decision nodes until a so-called leaf node is reached. Based on the result of a decision node (threshold test) two directions are possible. Either a follow-up test is performed based on the result of the previous test or the result is seen as the end and a leaf node is created. A leaf node indicates the end of a 'decision branch' which means no more tests are performed. The Decision Tree Regressor algorithm uses the MSE metric to determine which feature to use per decision node. A feature that reduces the MSE a lot is seen as a good predictor which should be used in a decision node used by the tree. Features that do not reduce the MSE a lot are seen as bad predictors which means they will not be used in the decision tree. An example of a decision tree is shown in figure 11 (Charbuty & Abdulazeez, 2021).

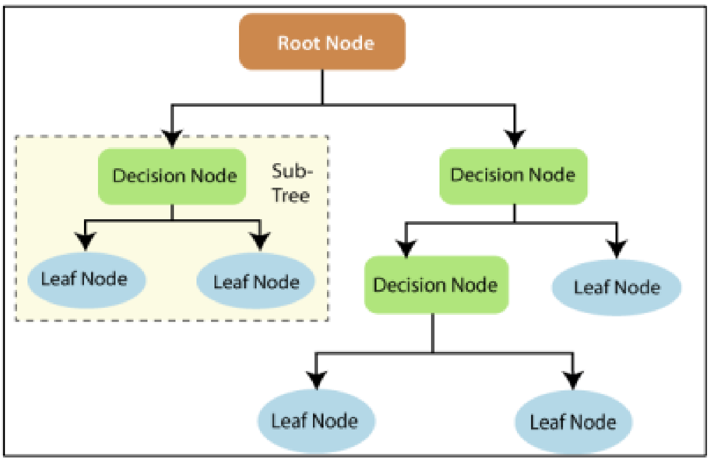


Figure 11: Example of a decision tree (Charbuty & Abdulazeez, 2021)

The explainability assessment of the Decision Tree algorithm is shown in table 7. The Decision Tree algorithm scored high on most of the explainability criteria due to the algorithm's relative simplicity. Decision Trees are especially comprehensible because of the logical rules they use for decision-making. These are a lot easier to interpret than for example the weights used between nodes in a neural network (Kotsiantis, 2011).

Table 7: Explainability assessment table of a Decision Tree

Criteria	Score (Low/Med/High)	Argument
Explainability of the feature importance	High	The decision tree algorithm enables the extraction of the important features
Explainability of the general algorithmic structure	High	Based on the data a sequence of decisions is generated to predict the dependent variable
Explainability of the performance criteria	Medium	R-squared variable is used which needs some background knowledge to understand
Explainability of the error minimisation function	High	The least squares method is used which means varying the decisions so that the difference between the predicted and real value is minimised

More advanced versions of the decision tree algorithm exist. Two of these, the Random Forest algorithm and the Gradient Boosting Algorithm, are chosen due to their proven effectiveness in previous long-term prediction research as has been found in section 3.

The random forest algorithm uses similar logic to that of the traditional decision tree in the sense that it makes use of the same tree-like structure and feature selection criteria. It is unique, however, because it creates many different trees based on random samples of the input (Naghibi, Pourghasemi, & Dixon, 2016). In the end, it combines the results from the different decision trees through so-called majority voting when used for classification or by averaging the results when used for regression. By creating and averaging the results of many different trees better predictions are reached compared to when only a single decision tree is used (Sruthi, 2021).

The increase in the predictive performance of the Random Forest algorithm does come at a cost. Whereas a single decision tree can immediately show the user which features were most important to come to a prediction this is not possible anymore when using the Random Forest algorithm. Because the Random Forest algorithm combines the results of many different decision trees the specific feature importance is lost. Instead, only a rough estimate can be given based on the Gini importance which indicates how well a future was able to reduce the error criterion or by the permutation feature importance which randomly removes one variable from the data to analyse its effect on the error criteria. The reduction in feature importance explainability is reflected in the explainability assessment of the Random Forest algorithm which is shown in table 8. Overall the model still scores relatively well for most categories even though it is starting to become more complex than the previous two algorithms.

Table 8: Explainability assessment table of Random Forest

Criteria	Score (Low/Med/High)	Argument
Explainability of the feature importance	Medium	Can be roughly assessed by the Gini importance or Permutation importance analysis.
Explainability of the general algorithmic structure	Medium	Similar to a normal decision tree except multiple ones of which the result is based on an average score over all the trees
Explainability of the performance criteria	Medium	R-squared variable is used which needs some background knowledge to understand
Explainability of the error minimisation function	High	The least squares method is used which means varying the decisions so that the difference between the predicted and real value is minimised

Similar to the random forest technique, Gradient boosting techniques also use the traditional decision tree structure as a basis and use multiple trees to reach a new and better-performing model. However, whereas the random forest technique creates many different decision trees parallel to each other and combines the results at the end, gradient boosting works more sequentially. In gradient boosting a decision tree is made and evaluated immediately. This is done by analysing the corresponding loss function. For a regression problem this could mean analysing the MSE. Based on the result of this analysis observations with a high error get a high weight assigned while observations with a low error get a low weight assigned. After tuning the weights of the observations a new tree is added that, by adding itself to the first tree, aims to perform better than the first tree based on the newly specified weights for the observations. This process repeats itself until a pre-specified stopping criterion is met (Singh, 2018). By adding trees that optimise themselves for the cases that are hard to predict the overall performance of the model increases.

Retrieving the feature importance is mainly possible using permutation importance or approaches similar to that of the Gini importance but with different loss functions. The corresponding explainability assessment can be seen in table 9. The model also scores slightly worse with regard to the explainability of the general algorithmic structure since gradient boosting is less intuitive than random forest.

Table 9: Explainability assessment table of Gradient Boosting

Criteria	Score (Low/Med/High)	Argument
Explainability of the feature importance	Medium	Permutation importance and Gini importance can be applied to retrieve an indication of future importance
Explainability of the general algorithmic structure	low/medium	Trees are added sequentially to improve upon the errors of the previous tree(s) to reach better results
Explainability of the performance criteria	Medium	R-squared variable is used which needs some background knowledge to understand
Explainability of the error minimisation function	High	The least squares method is used which means varying the decisions so that the difference between the predicted and real value is minimised

5.4 Support Vector Regression

Support vector machine algorithms can be applied for both classification and regression predictions. When used for regression instead of classification the term Support Vector Regression (SVR) is used instead of support vector machine (SVM). Very simply said the goal of SVM's is to find a hyperplane that best divides a dataset into two classes. Once a hyperplane is found new data points are classified based on their distance to the hyperplane. Points further away from the hyperplane are more likely to belong to a specific class while points closer to the hyperplane are less likely to belong to a specific class. The shape and position of the hyperplane are based on the data points closest to it on any of its sides. These data points are called the support vectors. If any of these points are removed the hyperplane changes its position. The support vectors are therefore seen as critical data points. When data becomes non-linear simple 2d hyperplanes become less effective. To solve this issue more dimensions in which to separate the data are added.

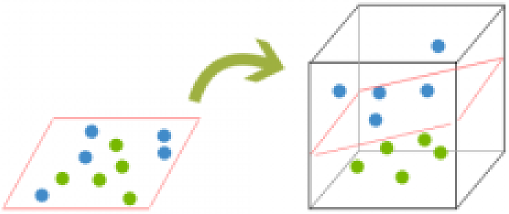


Figure 12: Example of a hyperplane dividing data in three dimensions (AYLIEN, 2022)

This is also known as kernelling and will be done until a number of dimensions is found in which a hyperplane can be fitted that properly separates the data. A three-dimensional example of this is shown in figure 12. This is also the reason why SVM's are more suitable for small datasets since bigger datasets require more dimensions to be added before a proper hyperplane is found. This can become quite computationally expensive (AYLIEN, 2022).

Support vector regression extends upon the support vector machine algorithm by adding a ϵ-insensitive region around the original function which is called the ϵ-tube. The ϵ value can be seen as a maximum allowed error value during prediction. This means that any predicted values further than ϵ away from the real value are penalised harder than predicted values equal or lower than ϵ from the real value. A simple example of this is shown in figure 13. The so-called decision boundaries are the positive and negative epsilon values between which the data should ideally stay (Sethi, 2020). The aim is to minimise the ϵ-insensitive loss function (I.e. the sum total error between the predicted and real values) to find the flattest ϵ-tube (the two red lines in the example figure) that still contains most of the training data (Awad & Khanna, 2015).

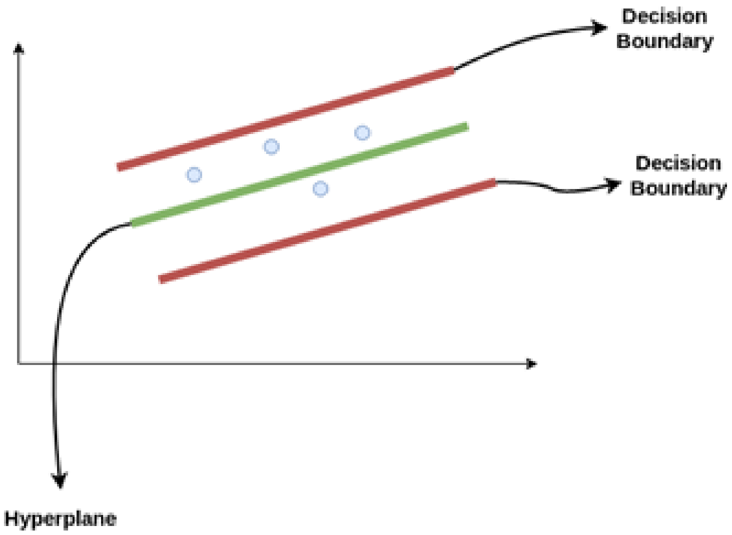


Figure 13: Example of a hyperplane and its ϵ values for support vector regression (Sethi, 2020)

SVM's have proven to be able to accurately forecast time series data even when the prediction problem contains non-linear elements. Additionally, they have shown to be able to outperform other non-linear techniques such as multilayer perceptron models (Sapankevych & Sankar, 2009). Given the non-linear elements in the prediction problem the SVM algorithm is therefore anticipated to perform well. For a more detailed mathematical explanation of support vector machines and support vector regression the reader is referred to Smola and Schölkopf (2004).

The complexity of the SVM algorithm influences its explainability assessment which can be seen in table 10. Compared to previous algorithms the feature importance can only be explained when a linear kernel is used which will not be the case since the problem contains non-linear elements. Therefore retrieving the feature importance can not be done using the algorithm itself. Permutation importance might still be possible, however. Furthermore, the general algorithmic structure, especially when higher order kernels come into play, as well as the error minimisation function has become more complex compared to earlier algorithms discussed.

Table 10: Explainability assessment table of SVR

Criteria	Score (Low/Med/High)	Argument
Explainability of the feature importance	Low	Impossible for kernels other than linear, however, permutation importance might be possible.
Explainability of the general algorithmic structure	Low/Medium	Simple support vector structure is understandable. Higher order support vectors become more difficult
Explainability of the performance criteria	medium	R-squared variable is used which needs some background knowledge to understand
Explainability of the error minimisation function	Low	SVR creates a function that maximises the % between epsilon. Requires understanding of the algorithmic structure of SVM's which is complex

5.5 Artificial Neural Networks

Neural networks are a type of machine learning that is loosely based on the human brain. This is shown through the use of so-called "neurons" that are connected to each other. The simplest neural network consists of three layers; an input layer, a hidden layer and an output layer (Ozoegwu, 2019). Each layer usually consists of multiple neurons, also called nodes, that either represent a value or transform a value. In the input layer each node represents an input feature which is used to predict the dependent variable. This layer is connected to a so-called hidden layer which in turn is connected to the output layer. Neurons between two layers always have at least one connection to a neuron in the upcoming layer. If all neurons from the previous layer are connected to all the neurons in the upcoming layer it is called a fully connected neural network. Each connection between two neurons from different layers converts the output from the previous layer by multiplying it with a weight value (Sadek et al., 2019). Neurons in the hidden layer sum up the input from the nodes connected to them from the previous layer and transform it through the use of an activation function. Here a bias value from a bias node is also added to ensure an output even when the input space is zero (Patra, Meenakshisundaram, Hung, & Simmons, 2017). An example figure showing this structure is shown in figure 14.

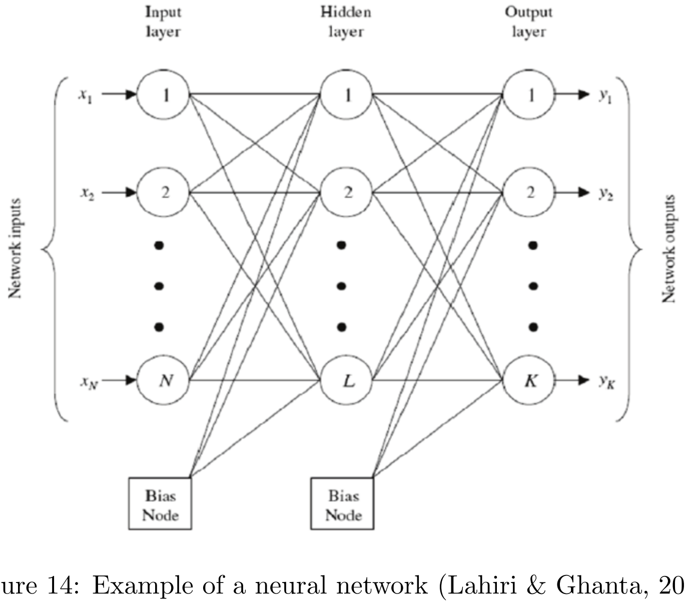


Figure 14: Example of a neural network (Lahiri & Ghanta, 2009)

The output is then forwarded to the next layer which in this case is the output layer. Before arriving at the output layer the value is again multiplied with a weight value. Whereas neurons in the hidden layer apply an activation function to the sum of the incoming values and bias, neurons in the output layer apply their own function to the sum of the incoming values and bias. For regression problems this is usually a linear function. Through this function, a prediction can be made of which the error with the actual value can be calculated. Through the use of a cost function, usually mean squared error or root mean squared error in case of a regression problem, weights are sought that minimise the error. This is done in a backwards manner (from the output layer to input layer) and is therefore called back-propagation (Ahmed & Khalid, 2017). The algorithm stops optimising once the error can't be lowered further or a different stopping criterion is met. Different algorithms for the optimisation problem exist, but the most well-known example is the so-called gradient descent algorithm. Gradient descent uses the derivative (gradient) of the error to determine if a weight has to be increased or decreased. Gradient descent aims to find local/global minima I.E. points where the gradient is zero and the error is the lowest. A deep neural network is an artificial neural network that contains more than one hidden layer (Cranenburgh S, 2021).

The results of the explainability assessment are shown in table 11. Similar to the explainability assessment of the SVM algorithm the increased complexity of the model has influenced its explainability scores. Feature importance can only be approached using permutation importance which makes the actual feature importance hard to explain. The basic algorithmic structure might still be reasonably explainable, however, if one wants to understand all the details it becomes a lot more complex and hard to explain. The error minimisation function minimises the squared error, however, the specific algorithm used to do this is hard to explain.

Table 11: Explainability assessment table of an MLP

Criteria	Score (Low/Med/High)	Argument
Explainability of the feature importance	Low	Permutation importance analysis can be applied
Explainability of the general algorithmic structure	Low/Medium	General structure can be explained relatively easily. Details require extensive background knowledge
Explainability of the performance criteria	Medium	R-squared variable is used which needs some background knowledge to understand
Explainability of the error minimisation function	Low	The specific implementation of this model uses LBFGS to minimise the squared error

5.6 Model hyper parameter optimisation

To get the most optimal results hyperparameter optimisation is performed for the different machine learning techniques. Since there are four different dependent variables and three different datasets each machine learning technique ends up with 12 different sets of optimal hyperparameter values.

Hyperparameter optimisation is performed using a grid search approach. This means a predefined set of hyperparameter values is analysed. To get reliable results the different combinations of hyperparameters are tested multiple times on different parts of the data using cross-validation. Since the data is a time series ordinary cross-validation can not be applied since that would mean future data would be used to train a model to predict historic data. Instead, a special kind of cross-validation is used in which the model is trained on different sets of historic data while always testing on future data. This is done using the TimeSeriesSplit function from SKlearn (2022c).

Since each algorithm has its own set of hyperparameters the hyperparameter optimisation is tailored to each algorithm individually. The specific hyperparameter values to analyse per algorithm are shown in table 12 till 16. Because the number of hyperparameter combinations grows exponentially the number of combinations is kept relatively small to ensure reasonable computation times even though this might mean a local minimum is found instead of a global minimum. Multiple linear regression is excluded from this analysis since it does not have any hyperparameters to tune.

When applying the decision tree regressor scikit-learn's optimised version of the Classification and Regression Trees algorithm is used (SKlearn, 2022a). When tuning hyperparameters for the CART algorithm research from Mantovani et al. (2018) has shown that the minimum samples per split and the minimum samples per leaf hyperparameters are the most important for the final performance of the model. When defining the grid search values to analyse for the hyperparameter optimisation these hyperparameters will therefore be given a larger spread of values compared to the other hyperparameters. As for all hyperparameter optimisation performed in this research, the optimisation criteria is the R-squared value. The grid search values of the Decision Tree hyperparameter optimisation are shown in table 12.

Table 12: Hyper parameter space for Decision Tree

Hyper parameter	Value range	Explanation of parameter
Maximum tree depth	[1, 2, 3, 5, 10, 20, 30, 50]	Maximum allowed depth of the three
Minimum samples per split	[2 - 15]	Minimum number of data samples that are required after a split
Minimum samples per leaf	[1 - 10]	Minimum number of data samples that need to be present in a leaf node
Maximum number of features	auto, sqrt, log2	Number of features to take into account when making a decision
Maximum number of leaf nodes	[5, 10, 20, 30]	Maximum number of leave nodes allowed in the tree

For the Random Forest algorithm, the corresponding Scikit-Learn documentation states that the main hyperparameters to adjust are the number of trees generated and the maximum number of features to consider when splitting a node (SKlearn, 2022b). Research from Probst, Wright, and Boulesteix (2019) further show that the sample size I.E. the number of observations drawn for each tree and the minimum number of samples per leaf node are also important albeit less influential on the results than the number of trees and features to consider when splitting a node. However, since the available data is small and the cross-validation already splices the data into different sets the number of samples per tree is kept at a maximum and is not further analysed during hyperparameter optimisation. The grid search values for the Random Forest hyperparameter optimisation are shown in table 13.

Table 13: Hyper parameter space for Random Forest

Hyper parameter	Value range	Explanation of parameter
n_estimators	[2, 5, 10, 20, 50, 100]	Number of trees to generate
max_features	auto, sqrt, log2	Number of features to consider when splitting a node
max_depth	1, 2, 3, 5, 10, 20, 50	Maximum number of layers in the tree
min_samples_split	2, 5, 10	Minimum number of samples needed in a node to allow it to split
min_samples_leaf	1, 5, 10	Minimum number of samples per leaf node

The Gradient Boosted Trees algorithm is applied using XGBoost. The XGBoost algorithm uses a few more hyperparameters compared to single Decision Trees and Random Forests since gradient boosted trees, among others, also include a learning rate and regularisation parameters. For XGBoost hyperparameter optimisation Luellen (2021) argues that the most important hyperparameters are the number of estimators, the maximum tree depth, the learning rate, gamma and the regularisation functions alpha and lambda. For XGBoost hyperparameter optimisation these values will therefore be analysed. The corresponding grid search values per parameter are shown in table 14.

Table 14: Hyper parameter space for XGBOOST

Hyper parameter	Value range	Explanation of parameter
n_estimators	[2, 5, 10, 20, 50, 100]	Number of trees to create
max_depth	[2, 5, 10, 20, 50]	Maximum allowed depth per tree
reg_alpha	[0,1,10,20,30,40,100]	L1 regularisation term on weights
reg_lambda	[0.0001, 0.001, 0.01, 0.1]	regularisation term on weights
gamma	[0, 0.1, 0.2, 0.3, 0.4, 0.5]	Minimum loss reduction required

For Support Vector Regression the most important hyperparameters are the epsilon value, C value and gamma value (Laref, Losson, Sava, & Siadat, 2019). The epsilon value refers to the epsilon insensitive area around the prediction, whereas C determines the strength of the regularisation and gamma determines the similarity radius used by the kernel function (Yildrim, 2020). Apart from these variables, the use of different kernel functions is also analysed. The corresponding grid-search values are shown in table 15.

Table 15: Hyper parameter space for Support Vector Regression

Hyper parameter	Value range	Explanation of parameter
kernel	RBF, Poly, Sigmoid	Specifies which kernel type to use
Degree	[3,5,7]	Number of degrees used in the polynomial kernal
epsilon	[1, 10, 500, 1000, 5000, 7500, 10000, 15000]	Defines the epsilon insensitive area
C	[0.01,0.1,1,10,100,1000,10000]	Define the strength of the regularisation
gamma	[scale,auto]	Determines the similarity radius used by the kernel function

The hyperparameter optimisation of Neural Networks is a relatively opaque process wherein the optimal hyperparameter values are often dependent on the tuning experience of the researcher and whether or not the combination of hyperparameter values is workable rather than optimal. This impacts the credibility since there is a lack of logical reason behind the decision for a certain set of hyper parameters(Yu & Zhu, 2020). However, since there are no set rules to determine which hyperparameter values to use trial and error is still needed to find adequate hyperparameter values. However, the hyperparameter search space can be narrowed down by using the hyperparameters identified by Yu and Zhu (2020) to be important. These are the Learning rate, Number of hidden layers, number of nodes per layer, Regularisation lambda and Activation functions. However, tuning the learning rate is only relevant when either the stochastic gradient descent (SGD) solver or the ADAM solver is used which is not the case for this analysis since SKlearns documentation advises the lbfgs solver when working with small data-sets (SKlearn, 2022d). Initial analysis with the SGD and ADAM confirms this since both solvers showed negligible results compared to the lbfgs solver. To prevent unnecessary computations they will therefore not be analysed. Instead, the maximum number of iterations before early stopping of the models optimisation is analysed instead. The corresponding grid search values are shown in table 16

Table 16: Hyper parameter space for Artificial Neural Network

Hyper parameter	Value range	Explanation of parameter
max_iter	200, 300, 500	The solver iterates until convergence or until this number of iterations is reached
Number of hidden layers & nodes per layer	(10), (20,20), (50), (100), (150,100,50), (40,20,10)	Number of hidden layers and nodes per hidden layer in the model. A value of (20,20) means two hidden layers with 20 nodes per layer
Alpha	0.0001, 0.001, 0.01, 0.1	Strength of the L2 regularisation term
Activation function	ReLu	Activation function of the hidden layer
Solver	lbfgs	Algorithm used to solve the weights for the prediction problem

6 Results

6.1 Introduction

In this chapter the prediction results of the different machine learning models are discussed. The section starts by discussing the results based on historic test data in order to show the theoretical performance possible when using machine learning models for the future demand prediction of identification documents. The section continues by discussing the results of the multi-step forecasting analysis and the section ends with a conclusion discussing and reflecting on the different results per model. Figure 15 shows the research element this section discusses.

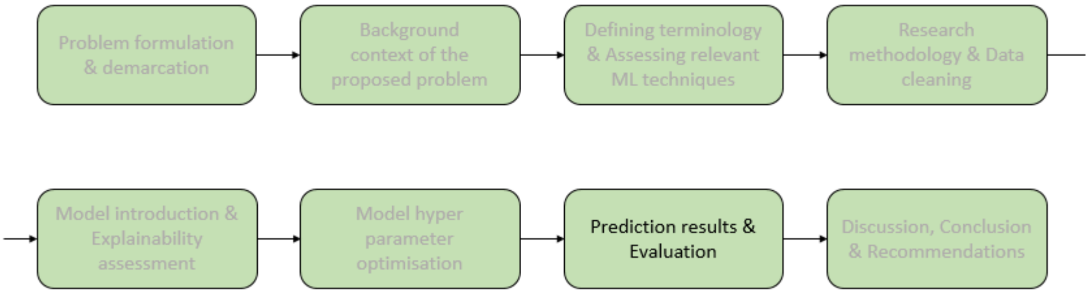

Problem formulation & demarcation
Background context of the proposed problem
Defining terminology & Assessing relevant ML techniques
Research methodology & Data cleaning
Model introduction & Explainability assessment
Model hyper parameter optimisation
Prediction results & Evaluation
Discussion, Conclusion & Recommendations

Figure 15: Current step in the research process

6.2 Prediction results

The prediction results of the different algorithms are based on two different factors. The dataset that is used for prediction and the ´optimal´ hyperparameter values corresponding to the used algorithm for a specific dataset and dependent variable. When presenting the results these factors will therefore also be shown.

While running the algorithms to predict the demand per dataset and dependent variable the training and test data is kept static by using a pre-defined seed value. This ensures that the different results can be compared since it means that the same training and test data is used for each algorithm. When fitting the algorithm, however, no seed was set. This means that each algorithm theoretically reaches a slightly different result each time the algorithm is run since most of the algorithms train themselves based on random decisions. In a robust model a different seed value might influence the performance metric with a few percentages. The results found when predicting the demand, however, showed a much higher variance in performance when different seeds were used. Given that the training and test data are always kept the same, these fluctuations solely occurred due to slight differences in the training of the models. Since a robust model should only show slight variations in performance when using different seeds the high variance in model performance found when predicting the demand using different seeds indicates that the models are not always able to properly predict the demand. This means the results are less reliable than desired and should be looked at with a relatively high degree of mistrust.

Because of the variance in the results each of the algorithms is run 25 times and the mean R-squared score is calculated to provide a more reliable result. Additionally, the variance in the R-squared score over the 25 runs is also calculated. This way, when discussing the results, the reliability of the results in the form of their variance can also be shown. When discussing the results of the different algorithms the optimal combination of algorithm and dataset is determined by the maximum mean R-squared score. The results are discussed in order of model complexity. This means the results are discussed in the following order: Multiple linear regression, regression tree, Random Forest, Extreme Gradient Boosting, SVR and MLP.

When discussing the results per model a table is shown containing the type of model applied, what dependent variable is predicted, which dataset is used, what the maximum average R-squared value is and what the variance in R-squared is. The column dependent variable contains the dependent variable names as supplied by the RvIG and say 'uitgifte aantal'. This is the Dutch translation of the number of issued documents. The different dependent variable names can be seen in table 17. The ni_5y, pn_5y, ni_10y and pn_10y behind 'uitgifte aantal' refer to the type of document issued and its validity period. 'Ni' refers to ID cards and 'pn' refers to passports. '5y' refers to a validity period of five years and '10y' refers to a validity period of ten years.

Table 17: Dependent variable names

dependent_variable
uitgifte_aantal_ni_5y
uitgifte_aantal_pn_5y
uitgifte_aantal_ni_10y
uitgifte_aantal_pn_10y

6.2.1 Multiple Linear Regression

Multiple linear regression is arguably the least complex model of the five. Although the model can be easily explained and excels at extracting the important feature variables, the algorithm is unable to capture the complexity of the prediction problem and this also shows itself in the results. Table 18 shows the prediction results with the highest averaged R-squared score over 25 runs. R-squared scores range between 0 and 1 and refer to the percentage of variance that can be explained by the model. A negative R-squared score means that the model performs worse than a horizontal line representing the mean of data. The negative R-squared scores in table 18 thus mean that the model is incapable of predicting the dependent variable and, given the size of the negative number, also is not even close to predicting the dependent variable.

Table 18: Multiple linear regression prediction results

model	dependent_variable	dataset	max_average_r2	variance in R2
LinearRegression	uitgifte_aantal_ni_5y	lagged_feature_data_2016_2022_t3.csv	-1026.967063	0.0
LinearRegression	uitgifte_aantal_pn_5y	lagged_feature_data_2016_2022_t3.csv	-2014.046870	0.0
LinearRegression	uitgifte_aantal_ni_10y	lagged_feature_data_2016_2022_t3.csv	-4409.456620	0.0
LinearRegression	uitgifte_aantal_pn_10y	lagged_feature_data_2016_2022_t3.csv	-10188.174676	0.0

The lack of predictive power can be visualised by plotting the prediction results. Example figures for the prediction of 5 and 10 year valid identification cards are shown in figures 16 and 17. The blue line in the figure represents the training data while the dotted orange line represents the test data I.E. the real value that should have been predicted. The green line represents the predicted value given the test data. The prediction is many times higher than the actual demand which shows the inability of the Linear Regression model to predict the future demand. Figures for the prediction of 5 and 10-year valid passports are not shown here to remain concise, but they show similar results to the figures shown here.

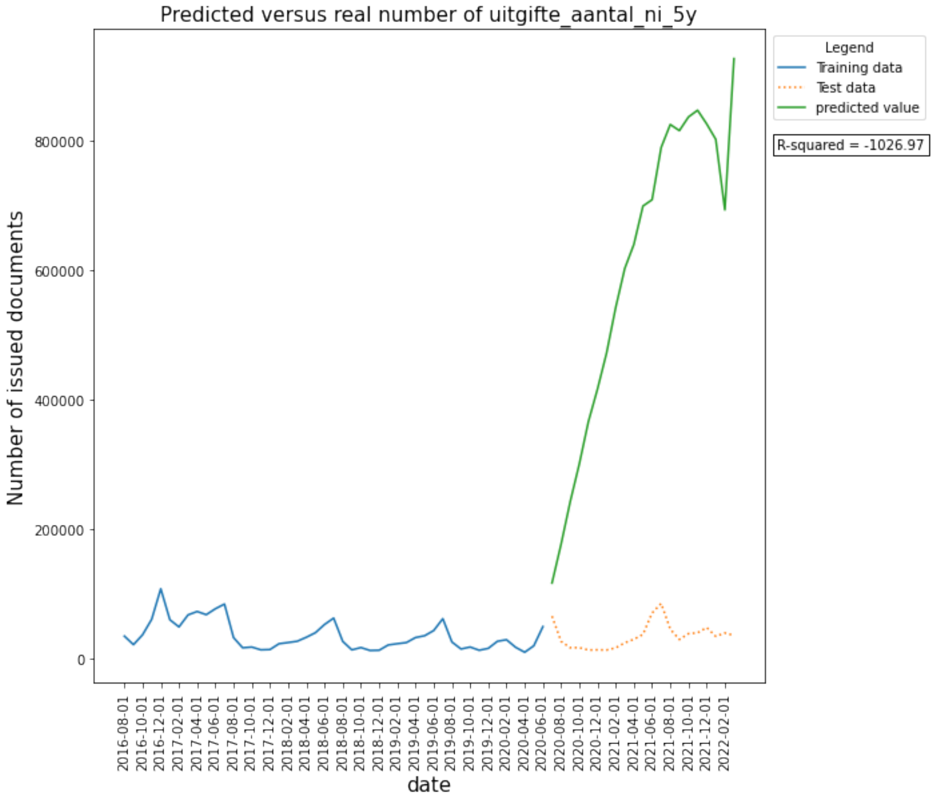


Figure 16: Predicted 5 year valid identification cards using MLR

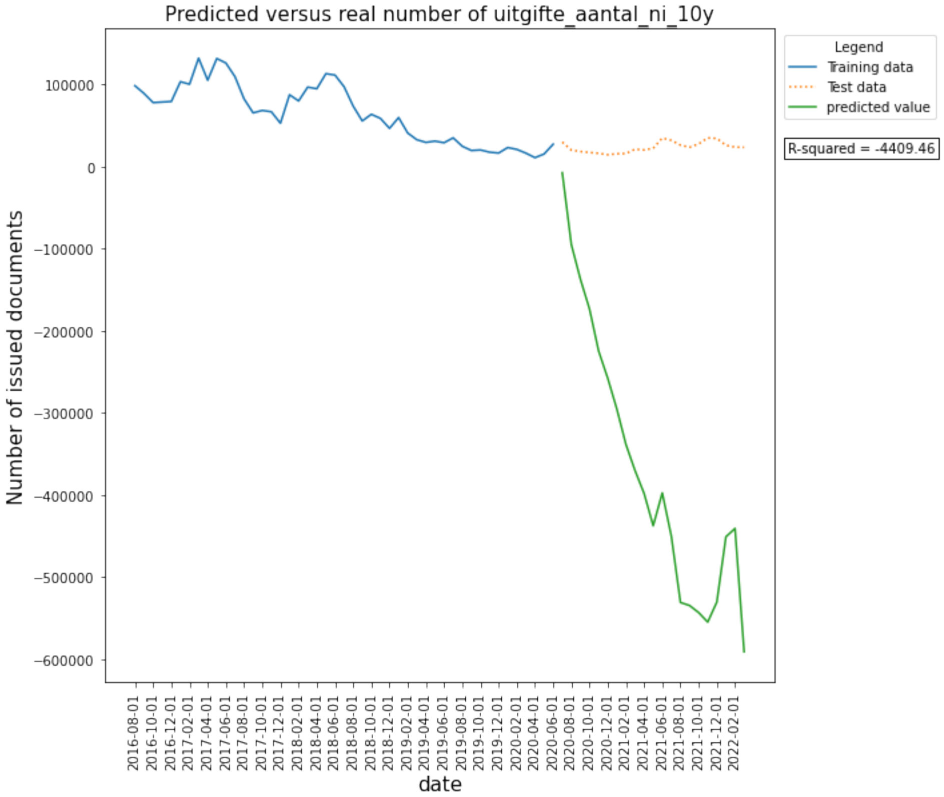


Figure 17: Predicted 10 year valid identification cards using SVR

Combining the prediction results with the explainability score given in the previous chapter leads to the conclusion that the Linear Regression algorithm is very explainable due to its relative simplicity, but that same simplicity leads to a lack of predictive power with regard to the prediction problem. This ultimately leads to the algorithm not being useful even though its explainability would be very suitable for real-life applications.

6.2.2 Regression Tree

The results of the regression tree analysis are shown in table 19. Similar to the results of the multiple linear regression analysis the model is unable to predict the dependent variable. The results might be relatively better than those of the multiple linear regression analysis, they are still all negative which means a line representing the average of the data would be a better predictor than the model. On top of that, the variance in the results is also relatively high. Especially the variance for the ID cards and passports with a 10-year validity time. Given that the R-squared value has a range between 0 and 1, a variance of 1.3 and 0.74 respectively is quite high. This further undermines the already unreliable results of the regression tree model.

Table 19: Decision Tree Regression prediction results

model	dependent_variable	dataset	max_average_r2	Variance in r2
DecisionTreeRegressor	uitgifte_aantal_ni_5y	lagged_feature_data_2016_2022_t-1.csv	-0.102824	0.267882
DecisionTreeRegressor	uitgifte_aantal_pn_5y	feature_data_2016_2022.csv	-0.020852	0.112970
DecisionTreeRegressor	uitgifte_aantal_ni_10y	feature_data_march_2014_dec_2018.csv	-1.991366	1.302158
DecisionTreeRegressor	uitgifte_aantal_pn_10y	standardized_feature_data_2014_2018.csv	-0.452015	0.749400

Figure 18 and 19 show the results of the regression tree algorithm using a pre-defined seed value of 1234. As can be seen in figure 18 the R-Squared value of the prediction is 0.1 and the line representing the predicted values sort of follows the actual demand values. However, an R-squared value of 0.1 means only 10% of the variance can of the dependent variable can be explained by the model leaving 90% of the variance left to other factors. Combine this with the variance present in the results and the model becomes even less reliable. Figure 19 shows what happens when the R-squared value is negative, namely a horizontal line representing the average of the data becomes the prediction.

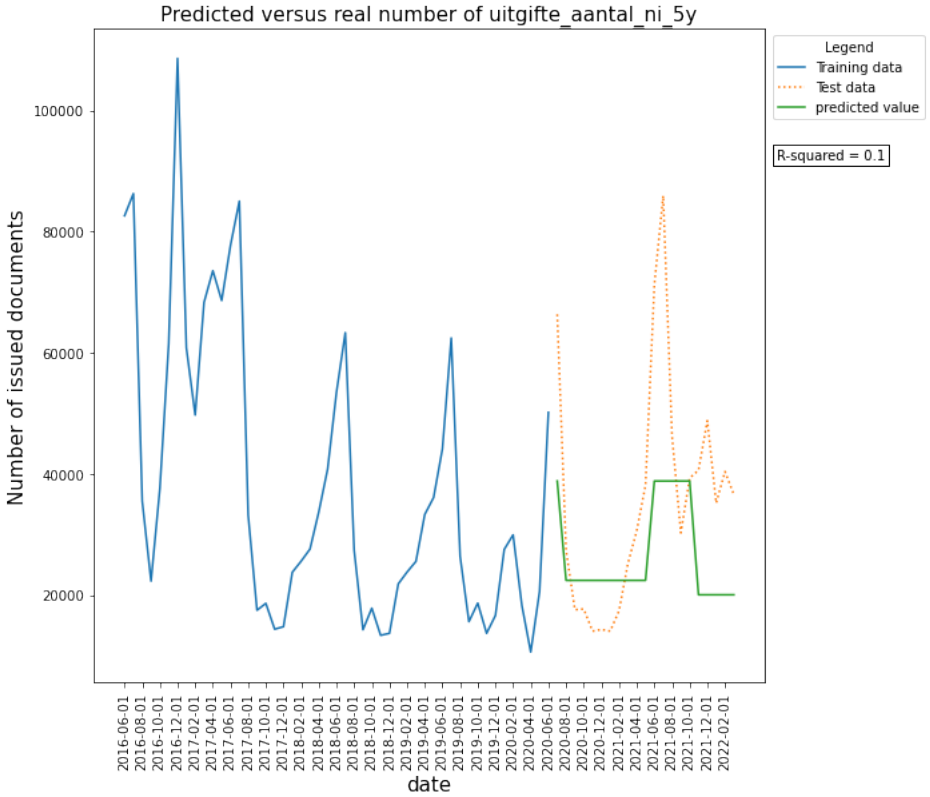


Figure 18: Predicted 5-year valid identification cards using regression tree

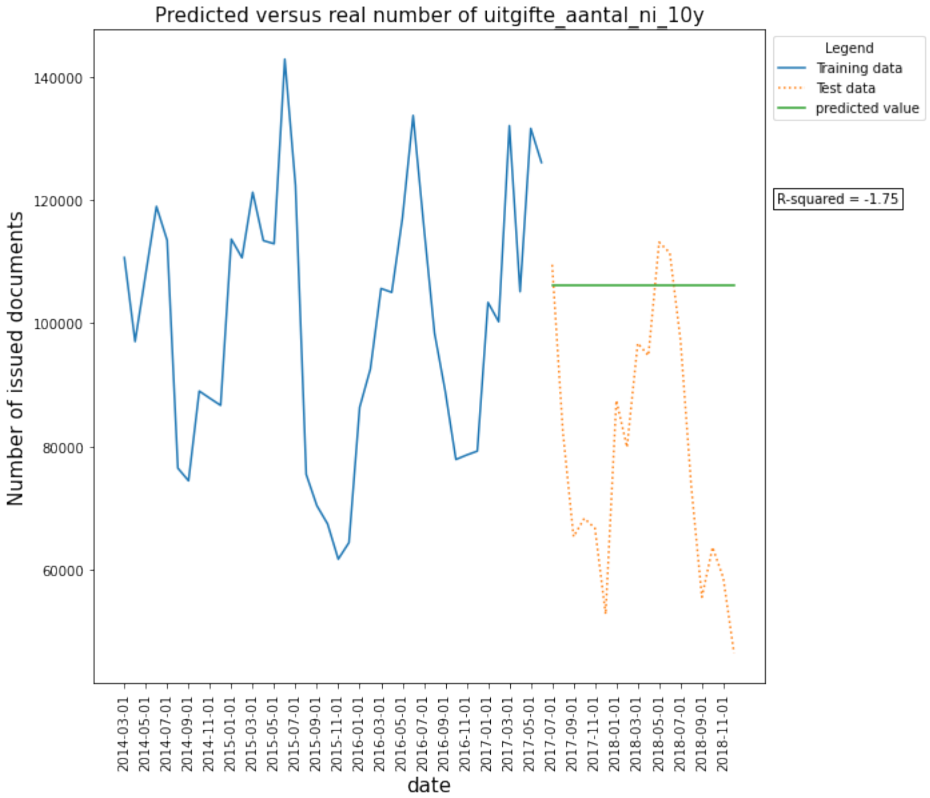


Figure 19: Predicted 10-year valid identification cards using regression tree

Coupling back the prediction results to the explainability score given to the regression tree model in the previous chapter it must be concluded that the model is very easily explainable, but lacks the complexity to properly predict future demand. Thus the regression tree algorithm is seen as unfit for the prediction problem.

6.2.3 Random forest

The results for the random forest predictions are shown in table 20. The random forest algorithm is the first algorithm that can capture part of the complexity of the prediction problem. Note that the results still are not noteworthy, the highest R-squared value is still only 0.26. But the results are a step up compared to the previous two models. What might be even more important is that the variance in R-squared is relatively low for the dependent variables with a positive R-squared value. Interestingly though the model seems to have a hard time predicting the demand for ID cards with a 10-year validity period even though the prediction for passports with the same validity period does have a positive R-squared value.

Table 20: Random Forest Regression prediction results

model	dependent_variable	dataset	max_average_r2	Variance in r2
RandomForest Regressor	uitgifte_aantal_ni_5y	lagged_feature_data _2016_2022_t3.csv	0.227153	0.013558
RandomForest Regressor	uitgifte_aantal_pn_5y	lagged_feature_data _2016_2022_t2.csv	0.098252	0.062277
RandomForest Regressor	uitgifte_aantal_ni_10y	lagged_feature_data _2016_2022_t-1.csv	-0.749718	0.892041
RandomForest Regressor	uitgifte_aantal_pn_10y	lagged_feature_data _2016_2022_t2.csv	0.259962	0.118852

Figure 20 and 21 show the results of the random forest algorithm using a pre-defined seed value of 1234. As was to be expected due to the low variance in R-squared, the R-squared value in figure 20 is close to the maximum average R-squared value found. Here the results of a higher R-squared value become clear. The green line, representing the predicted demand, starts to follow the actual demand. Simultaneously it also becomes clear why an R-squared value of 0.25 is not high enough since the predicted demand is multiple times lower and sometimes higher than the actual demand. Figure 21 once again shows the result of a negative R-squared value which is almost a purely horizontal line.

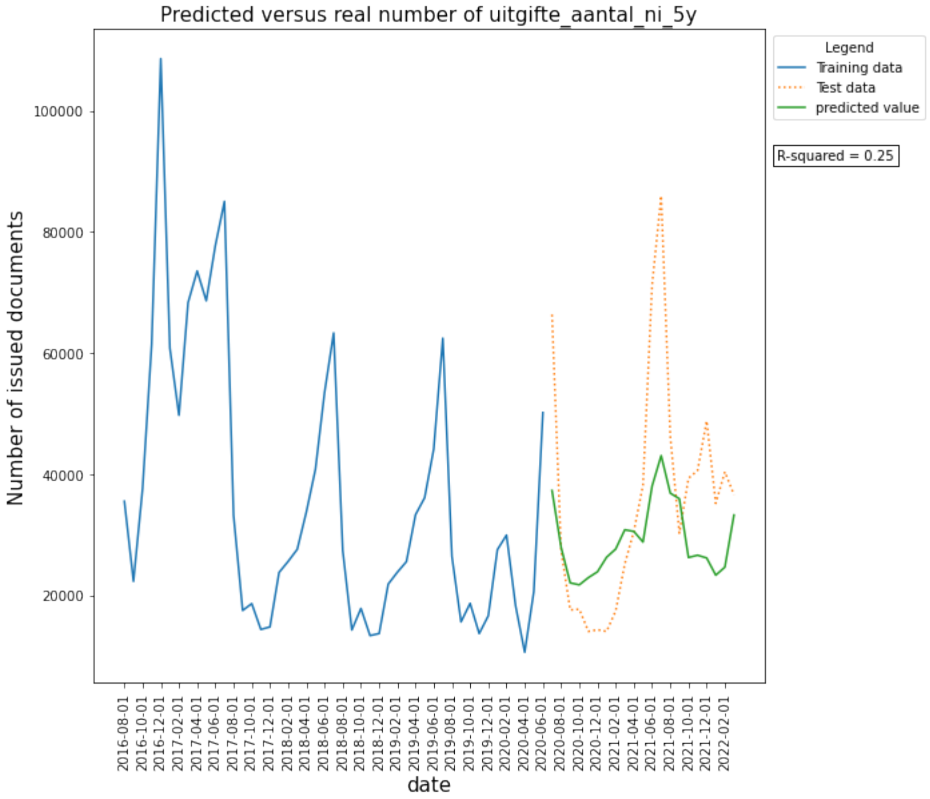


Figure 20: Predicted 5-year valid identification cards using random forest

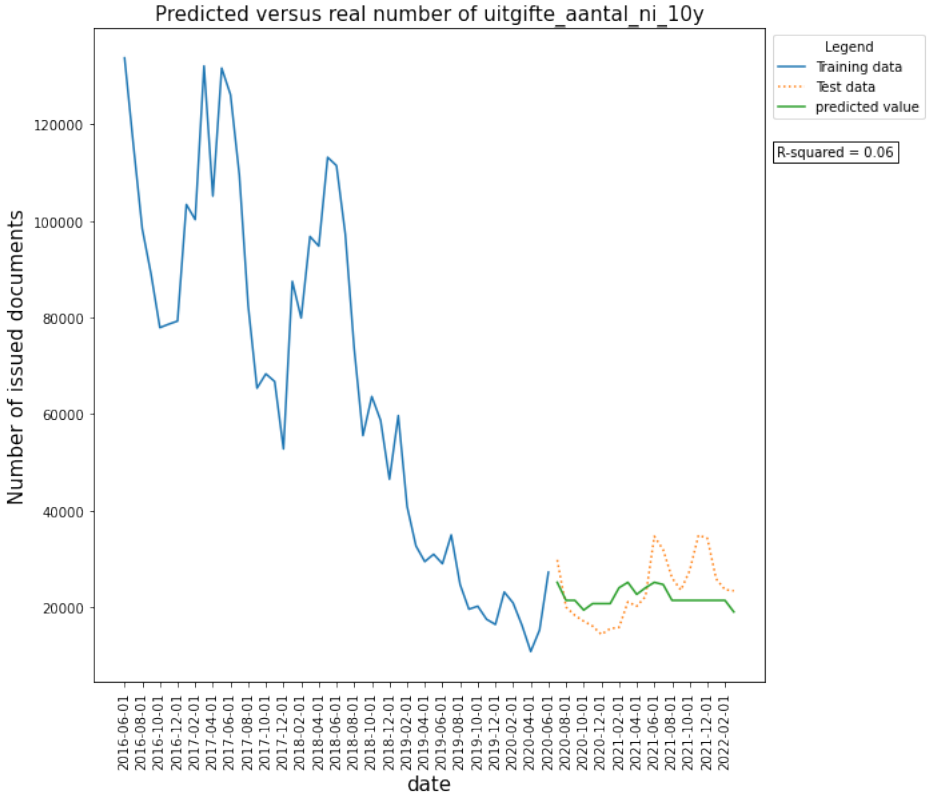


Figure 21: Predicted 10-year valid identification cards using random forest

Combing the results of the random forest algorithm with its explainability score the conclusion is that the random forest algorithm should be analysed further. The RF algorithm remains one of the most explainable algorithms inside the group of more complex algorithms and although the current results are not yet up to par, it might be possible to get better predictions with more fine-tuning of the hyperparameters and extra data.

6.2.4 Extreme Gradient Boosting

Table 21 contains the results for the Extreme Gradient Boosting algorithm. Although the subsampling and colsample hyperparameters are all set to 0.75 no randomness seems to occur in the results even though that is expected given the hyperparameter settings. A possible reason for the lack of randomness could be the fact that the training and test data is static since the same seed is used to create train and test data every time the model is applied. Other reasons could be a bug in the XGBRegressor code or an implementation error caused by the researcher. Nevertheless, the XGBregressor algorithm is able to reach a relatively high performance. Both the R-squared value for five-year valid IDs and passports are almost 0.5 and the R-squared of ten-year valid passports even reaches 0.74. Interestingly the R-squared value for ten-year valid IDs is negative. No clear reason for this exists. One possibility might be that there exist less clear patterns in the demand for ID cards with a ten-year validity period because people age 18 and above usually have a driver's license which often functions as a substitute for an ID card. Instead, people aged 18 and above will likely buy a passport. This, however, remains purely speculative and should be investigated further by the domain experts.

Table 21: Extreme Gradient Boosting prediction results

model	dependent_variable	dataset	max_average_r2	Variance in r2
XGBRegressor	uitgifte_aantal_ni_5y	lagged_feature_data _2016_2022_t2.csv	0.477636	0.0
XGBRegressor	uitgifte_aantal_pn_5y	lagged_feature_data _2016_2022_t3.csv	0.494608	0.0
XGBRegressor	uitgifte_aantal_ni_10y	lagged_feature_data _2016_2022_t-1.csv	-0.614031	0.0
XGBRegressor	uitgifte_aantal_pn_10y	lagged_feature_data _2016_2022_t-1.csv	0.739382	0.0

The predictions are once again visualised and can be seen in figure 22 and 23. Figure 22 shows the result of one of the best predictions up until now. Although the prediction does not follow the different peaks in demand perfectly in terms of height it can follow their form relatively well. Figure 23 interestingly enough shows a plot that still follows the actual demand relatively well even though the corresponding R-squared value is negative. Where a horizontal line representing the average of the test data was expected a rough but relatively well-fitting plot is present instead.

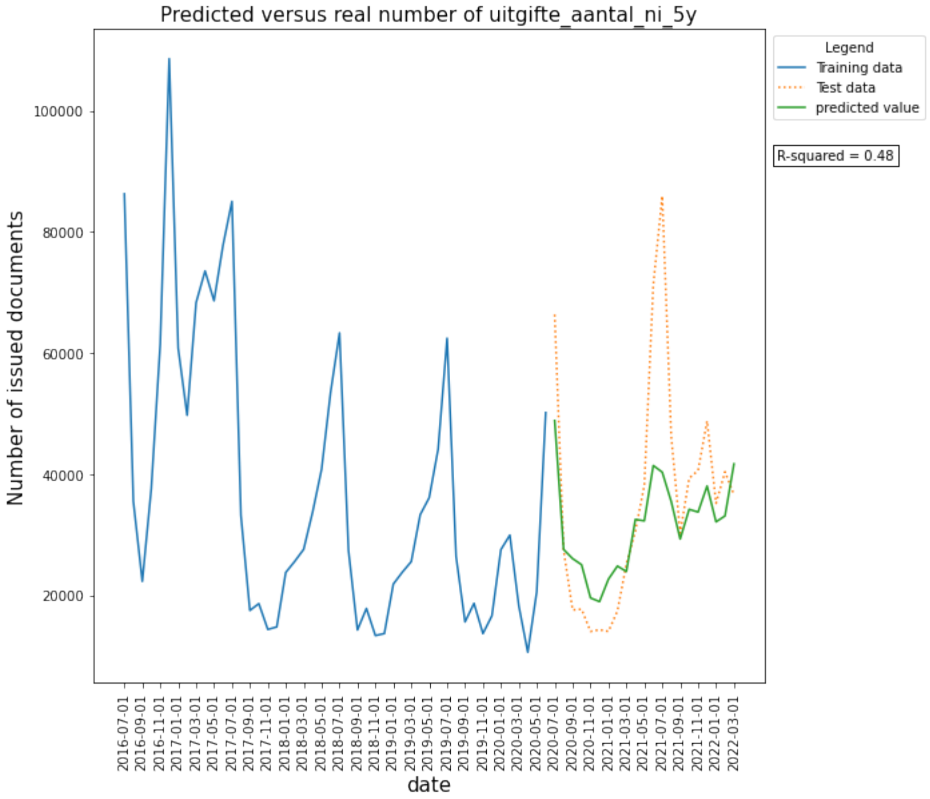


Figure 22: Predicted 5 year valid identification cards using XGB

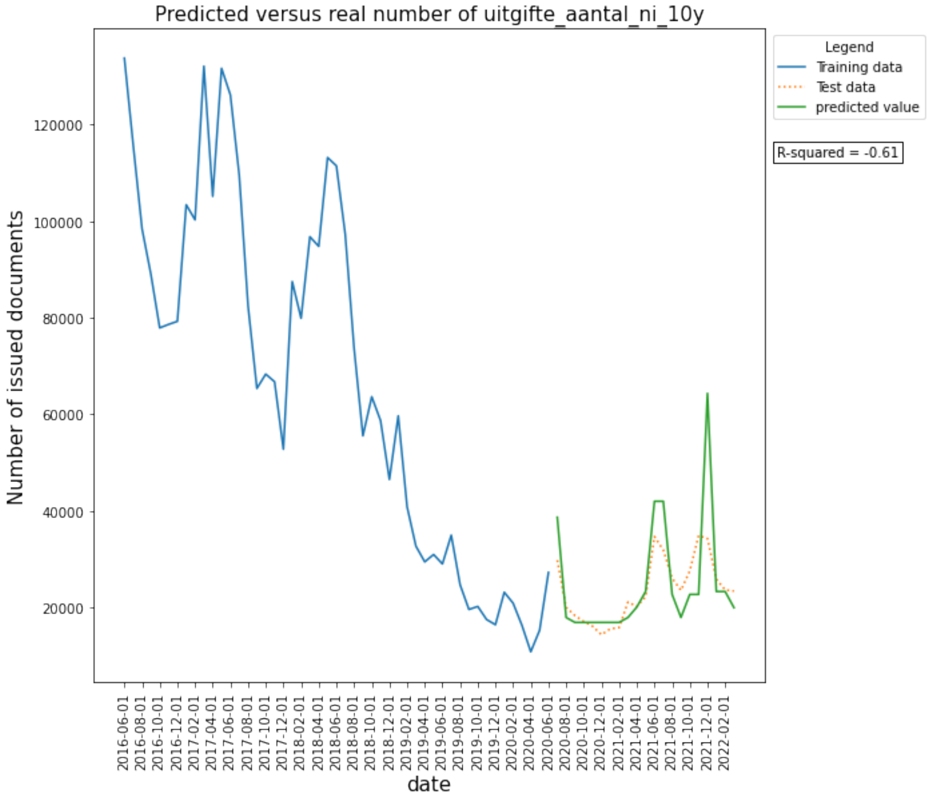


Figure 23: Predicted 10 year valid identification cards using XGB

Combining the performance results of the XGB algorithm with its explainability score leads to the conclusion that this algorithm might be one of the most suitable ones for this specific prediction problem. The XGB algorithm is relatively explainable and can reach relatively high-performance scores. This makes it a very interesting algorithm to use for the RvIG.

6.2.5 Support vector regression

Table 22 contains the prediction results for the SVR algorithm. Similar to the multiple linear regression algorithm no variance occurs in the prediction results when fitting and applying the model multiple times. This makes sense since the SVR documentation by SKlearn does not enable the user to set a random_state which suggests the model implementation by sklearn does not contain any stochastic elements (SKlearn, 2022e). Which, if the case, means the results are very reliable. The SVR results are also generally higher than any of the previous results received by the other models. The lowest R-squared value is retrieved for the demand prediction for passports with a five-year validity period and is equal to 0.51. The other R-squared results for the other three predictions fluctuate around 0.7 with the prediction for five-year valid IDs reaching an R-squared value of 0.79 while the prediction for ten-year valid passports reached an R-squared value of 0.67. The prediction for ten-year valid IDs lies between both of them and has an R-squared value of 0.74.

Table 22: Support Vector Regression prediction results

model	dependent_variable	dataset	max_average_r2	Variance in r2
SVR	uitgifte_aantal_ni_5y	MaxAbs_feature_data _2016_2022.csv	0.790205	0.0
SVR	uitgifte_aantal_pn_5y	MaxAbs_feature_data _2016_2022.csv	0.510128	0.0
SVR	uitgifte_aantal_ni_10y	lagged_feature_data _2016_2022_t3.csv	0.740977	0.0
SVR	uitgifte_aantal_pn_10y	feature_data _2016_2022.csv	0.673658	0.0

Since the R-squared metric represents the variance the model can explain, higher R-squared values mean more of the variance can be explained and thus a better predicting model. This immediately shows in the plots corresponding to the predictions discussed above. Figure 24 shows the demand prediction for five-year valid ID's. As can be seen, the predicted values follow the actual demand very well apart from a few areas in which the prediction is slightly higher. Similar results can be seen in figure 25 in which the demand prediction for ten-year valid IDs is plotted. Here the predicted demand is often slightly different than the actual demand, but overall the model can predict the demand quite well. Although not plotted here, the demand prediction for five and ten-year valid passports show similar results.

The results from the SVR algorithm show that future demand prediction of both passports and ID cards is possible using machine learning and more importantly, with the feature variables that have been identified. However, the explainability score of the SVR algorithm is low which might make it hard for the RvIG to implement. It must further be noted that to make actual predictions for the future, future feature data is needed which does not exist (yet). The current results, therefore, show the theoretical possibilities for future demand prediction, but more research needs to be done to make actual future predictions. Different strategies to tackle this issue exist, one of which will be explored in the next subsection.

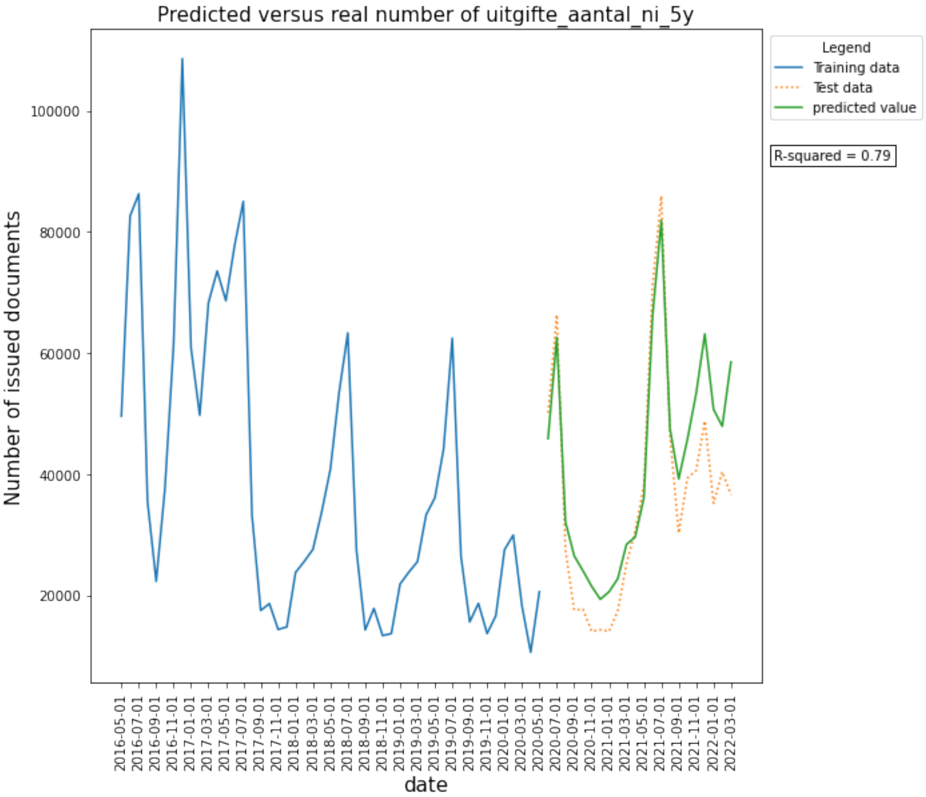


Figure 24: Predicted 5 year valid identification cards using SVR

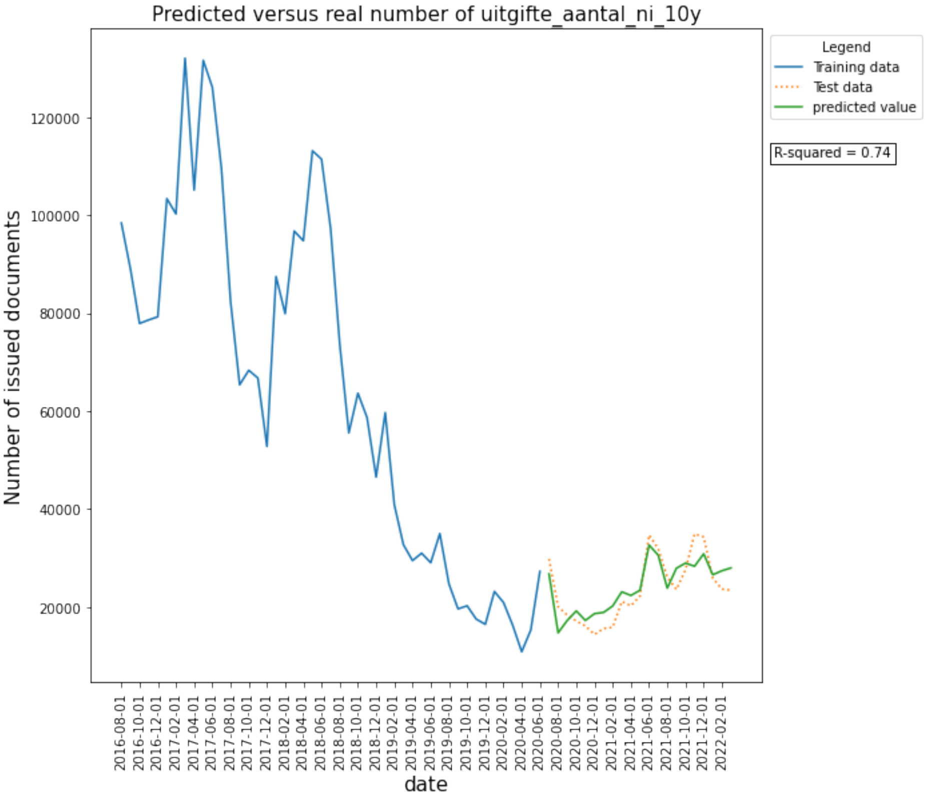


Figure 25: Predicted 10 year valid identification cards using SVR

6.2.6 Multilayer Perceptron

The results from the Multilayer Perceptron model can be seen in table 23 and are surprisingly quite underwhelming. Only the demand prediction for five-year valid passports shows a positive R-squared value of 0.41. However, some variance exists in this result which makes it less reliable. The other predictions all show a negative R-squared value which means the model is not capable of predicting the demand. These results are surprising since the complexity of Multilayer Perceptron models suggests that they should be able to capture the relationship between the feature variables and dependent variables, especially since previous models have also been able to.

Table 23: Multilayer Perceptron Regression prediction results

model	dependent_variable	dataset	max_average_r2	Variance in r2
MLPRegressor	uitgifte_aantal_ni_5y	standardized_feature _data_2016_2022.csv	-0.234234	2.938335
MLPRegressor	uitgifte_aantal_pn_5y	lagged_feature_data _2016_2022_t2.csv	0.408624	0.139951
MLPRegressor	uitgifte_aantal_ni_10y	MaxAbs_feature_data _2016_2022.csv	-0.109329	0.556050
MLPRegressor	uitgifte_aantal_pn_10y	MaxAbs_feature_data _2017_2019.csv	-0.905083	4.163745

Figure 26 and figure 27 show a plot of the demand prediction for five and ten year valid identification cards respectively. Keep in mind that even though the R-squared values present in the figure are relatively high, 0.56 and 0.43, these can not be relied upon due to the high variance in the prediction results. The R-squared values shown are therefore more a matter of "luck" than anything else. The lack in performance might be explainable by the use of sub-optimal hyperparameters. Similar to the explanation for the random forest results, the hyperparameters found using grid search might have led to a local optimum instead of a global optimum if a local optimum was even found at all. A more extensive hyperparameter optimisation to ensure better hyperparameters might lead to better results using the MLP model.

The performance results of the ANN algorithm combined with its explainability score lead to the conclusion that the ANN algorithm is unfit for this prediction problem. Normally a more complex model reaches better results which acts compensates for the lack of explainability. However, in this specific case, the model does not perform well and it is hard to explain. The current implementation of the ANN algorithm is therefore not recommended for this prediction problem. It might however be possible that a different implementation of the algorithm can reach better results. The ANN algorithm should therefore not be fully discarded.

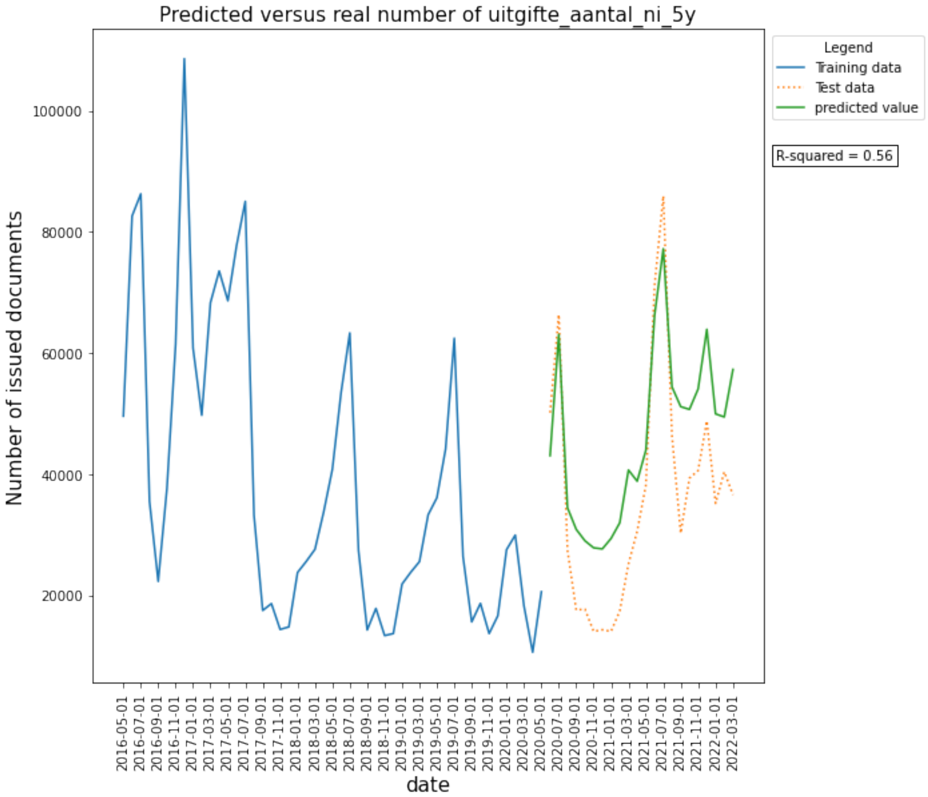


Figure 26: Predicted 5 year valid identification cards using MLP

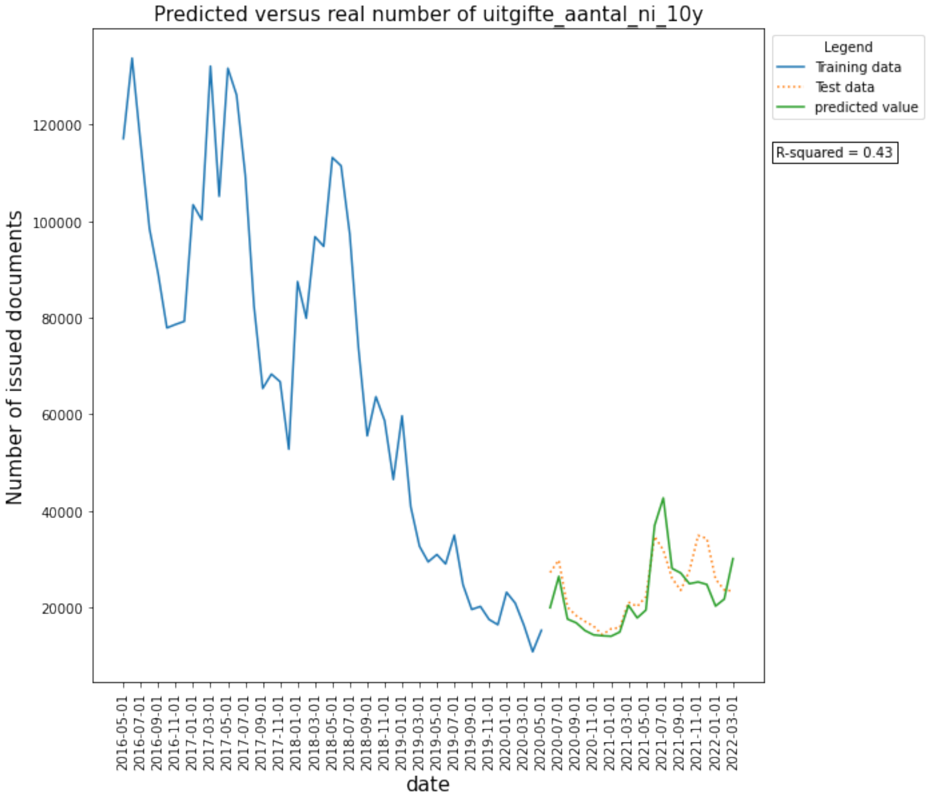


Figure 27: Predicted 10 year valid identification cards using MLP

6.3 Future demand prediction using direct multi-step forecasting

The results in the previous sub-section have shown the theoretical capabilities of different machine learning algorithms for predicting the future demand of identification documents. However, no actual future predictions were made. Here the results of the direct multi-step forecasting analysis are discussed in which future predictions are made for t+1, t+2 and t+3 time steps ahead.

Direct multi-step forecasting is performed using the 2016-2022 data set. Based on this dataset different combinations of lagged feature data and future dependent variables are created which are used to train and test the models. Lagged feature data is created for t-1 till t-5 steps back while the dependent data is shifted forward t+1 till t+3 steps in the future. When creating lagged data for t-3 steps back, for example, all lagged steps before are also added to the dataset. Thus a dataset with lagged steps equal to three contains lagged data for t-1, t-2 and t-3. The same does not apply to the dependent variable since the models are only trained to predict one step ahead at a time. While training, testing and predicting the hyperparameter values used per model are based on the hyperparameter values found in the previous section using the 2016-2022 feature dataset with t-1 lagged variables. Similar to the results in the previous section variance existed in the predictions made by the models. Therefore the models are once again run 25 times and the mean results are taken. Since many datasets are tested per model only the results of the dataset that managed to reach the highest average R-squared value per dependent variable are looked at. The corresponding results for the SVR algorithm are shown in table 24. A full overview of the results for all the different models is shown in appendix B.

Table 24: Direct multi-step forecasting results for Support Vector Regression

Model	Dependent variable	Lagged steps	Max average R^2	Variance in R^2
SVR	uitgifte_aantal_ni_5y_t+1	3	-1.119743	0.0
SVR	uitgifte_aantal_pn_5y_t+1	1	0.065676	0.0
SVR	uitgifte_aantal_ni_10y_t+1	5	-1.306219	0.0
SVR	uitgifte_aantal_pn_10y_t+1	5	-10.604207	0.0
SVR	uitgifte_aantal_ni_5y_t+2	2	-1.210848	0.0
SVR	uitgifte_aantal_pn_5y_t+2	4	-0.045891	0.0
SVR	uitgifte_aantal_ni_10y_t+2	5	-1.513715	0.0
SVR	uitgifte_aantal_pn_10y_t+2	5	-5.62213	0.0
SVR	uitgifte_aantal_ni_5y_t+3	4	-1.13186	0.0
SVR	uitgifte_aantal_pn_5y_t+3	3	-0.060252	0.0
SVR	uitgifte_aantal_ni_10y_t+3	5	-1.086283	0.0
SVR	uitgifte_aantal_pn_10y_t+3	3	-6.37712	0.0

The R-squared values correspond to the maximum average R-squared values the SVR model could reach. The number of lagged steps corresponds to the number of lagged steps used in the dataset for that specific prediction. This is thus the number of lagged steps with which the model reached the highest R-squared value. However, as can be seen by the R-squared results for the SVR model in the table it is mostly incapable of predicting future demand. Almost all of the applied SVR models have a negative R-squared value except for a few which are still only just able to reach a positive value. For the models reaching a negative R-squared value, a line representing the average demand would be a better predictor than the current predictions. The positive SVR results are only very slightly better which makes them far from good enough to use as input for decision making. Similar, if not worse, results occur for all of the other models as well.

Multiple possible explanations exist for the lack of predictive performance achieved by the models. First of all the number of lagged steps used in the datasets might be too small. Because only lagged data up to 5 months back in time was used this might have prevented the models from properly learning the periodic demand fluctuations which are present throughout the year. Secondly, the models all used hyperparameters which were optimised for a different purpose than direct multi-step future demand prediction. Although the results are not expected to change significantly with different hyperparameters the lack of hyperparameter optimisation per model, lagged feature data and dependent data is still seen as one of the possible reasons for the lack of performance. Lastly, the models might be underperforming simply because of a lack of data due to which the models are not able to properly capture the relationship between the lagged feature data and future demand.

6.4 Conclusion

The results have shown that the future demand prediction of identification documents is theoretically possible with (part of) the identified machine learning models in combination with the used feature data. Especially the SVR and Gradient Boosting models have shown good performance reaching R-squared values up to 0.7 for part of the dependent variables. Both models also do not have any variance in their prediction which makes them more reliable. For the XGB algorithm this might however be caused by a bug or a wrong implementation of the algorithm. These results further strengthen the findings discussed in the last paragraph of subsection 3.8 which showed that the SVR and XGB algorithms often perform well.

The Multiple Linear Regression and Single Regression Tree algorithms were not able to capture the complexity between the feature data and the dependent data. Both algorithms only reached negative R-squared values for all of the dependent variables. On top of that, the Regression Tree also showed relatively high variance in its results which makes them less reliable. The relative simplicity of these algorithms withholds them from capturing the complexity that is present in the prediction problem. When trying to predict the future demand for identification documents these models should therefore not be used.

The Random Forest algorithm surprisingly did not perform very well. In most literature, the RF algorithm usually performs quite well, but here the algorithm was only able to reach an R-squared value of 0.22 and 0.25 for NI 5y and PN 10y demand predictions respectively. The R-squared for PN 5y demand prediction was only 0.09 and the R-squared for the NI 10y demand prediction was negative. Given the results, it is not recommended to use the current implementation of the RF algorithm for the future demand prediction of identification documents.

Lastly, the r-squared results of the multilayer Perceptron algorithm were also low. Only one of the dependent variables was predicted with a positive R-squared value. However, this prediction has a variance of 0.14 which means it is still not very reliable. The other predictions all had a negative R-squared value. The MLP algorithm was expected to overfit due to the small amount of data but the opposite occurred. No clear reason for the lack in performance exists apart from a potential sub-optimal combination of hyperparameters and a lack of data. Both will be discussed more extensively later on in this section. Ultimately, the RvIG should not use this algorithm.

The previous paragraphs all talked about the results in which historic data was used to train and test the models. Here none of the models predicted any actual future values. The direct multi-step forecasting approach tried to fill this gap by training a new model using historic feature data to predict the dependent value on time step t+n. Unfortunately, the R-squared results for all of the models trained on the different combinations of data were either negative or very close to zero. None of the models was thus able to accurately predict the future demand using this approach.

In its current form, the models are unable to accurately predict the future demand for identification documents. The R-squared values reached by the models are too low to be able to put any trust in the results. In more practical terms the low R-squared value means that the predictions that are made using these models are likely way too high or way too low compared to the actual demand values. However, even though the current implementation of the machine learning algorithms is not able to reach satisfactory results it does not mean machine learning algorithms are inherently incapable of solving this prediction problem.

The demand for identification documents shows a strong seasonal pattern which, due to assumptions made in the data preparation, is currently underrepresented in the feature data. The biggest factor in this is that while creating monthly data out of yearly values a linear growth between years is assumed. This means that the feature data either always increases or decreases throughout the year while the demand for identification documents fluctuates both up and down throughout the year. For a model that aims to find relationships between feature data and dependent data this is very confusing which is likely why the models have a hard time predicting the demand. Before discarding the use of machine learning algorithms as a whole it is therefore strongly recommended to further investigate if the seasonality in the demand can be better represented using higher quality feature data. Additionally, since the data that is currently available sits in between the transition period of changing from five-year valid documents to ten-year valid documents for adults the dependent data itself is relatively inconsistent as well. This can be a reason for the lack of performance of the models since it makes it harder for the models to identify relationships between the feature data and dependent data. It might therefore be worthwhile to wait a few years until this transition has eased out a bit more so that the dependent data is more consistent and easier to predict.

Another potential reason for the results is that the hyperparameters that are currently used for the models are likely sub-optimal due to the relatively small hyperparameter optimisation analysis that is performed. Better results can therefore likely be reached with a more extensive hyperparameter optimisation analysis. On top of that the current number of lagged data steps that are added only go up to five steps back. Since the demand shows yearly seasonality this might be too few steps to effectively represent the yearly patterns, thus limiting the performance of the models. Adding lagged data from a year ago to try and incorporate more seasonality might therefore enhance the results. Furthermore, even though the feature data is determined using the input of domain experts it is still possible that effective predictor variables are not included in the current feature data. This potentially limits the accuracy of the results as well. Looking for additional feature variables that potentially have (more) predictive power can therefore also increase the performance of the models. Lastly, the current amount of data is quite small, especially for a machine learning problem. Because of this the model likely does not have enough examples to learn from which makes it hard for the model to find effective patterns/relationships between the feature and dependent data. This limits the prediction accuracy of the models. Where possible the RvIG should therefore try to increase the available amount of data. Doing this, however, remains a challenge since the RvIG can only access dependent data up until a certain amount of years back in time and feature data corresponding to the available periods does not always exist.

In the case that none of the aforementioned suggestions were able to increase the model performance substantially the use of machine learning algorithms can be assumed to be unfit for this specific prediction. In that case, the RvIG should extend its scope and explore the use of different prediction techniques.

7 Competing model requirements and the choice of algorithm

7.1 Introduction

This section analyses how competing requirements regarding the performance and explainability of models influence the choice of model used by the decision maker. First, the models and their explainability and performance scores are briefly repeated. Afterwards, three decision-making profiles are formalised to which the different models are linked. The section ends with a novel framework that can assist future decision-makers in their choice of a model when dealing with competing requirements. Figure 28 shows the research element this section discusses.

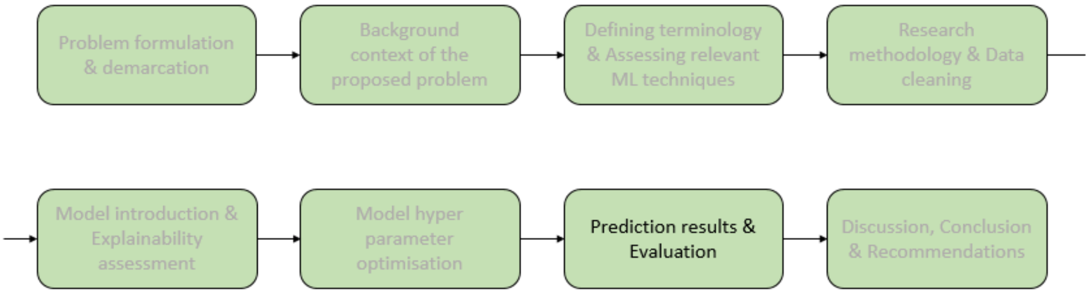

Problem formulation & demarcation
Background context of the proposed problem
Defining terminology & Assessing relevant ML techniques
Research methodology & Data cleaning
Model introduction & Explainability assessment
Model hyper parameter optimisation
Prediction results & Evaluation
Discussion, Conclusion & Recommendations

Figure 28: Current step in the research process

7.2 Summary of the explainability and prediction results

Each of the different machine learning models has been assessed on their explainability in section 5 and have been assessed on their performance in chapter 6. Table 25 shows a brief overview of both the explainability score and performance reached per machine learning model. The MLR and DTR algorithms both scored the same in the explainability assessment analysis. They each scored high with regards to the explainability of the feature importance, explainability of the general algorithmic structure and explainability of the error minimisation function. They both scored a medium for the explainability of the performance criteria. Both algorithms reached negative R-squared results for all of the dependent variables. The random forest algorithm scored medium for its explainability of the algorithmic structure, feature importance and performance criteria and scored high with regards to its explainability of the error minimisation criteria. It reached R-squared scores of 0.22, 0.09, -0.75 and 0.26 for NI 5y, PN 5y, NI 10y and PN 10y respectively. The XGB algorithm scored a medium for its explainability of the feature importance and explainability of the performance criteria, a low to medium for its explainability of the general algorithmic structure and a high for its explainability of the error minimisation function. It reached R-squared scores of 0.48, 0.49, -0.61 and 0.74 for NI 5y, PN 5y, NI 10y and PN 10y respectively. Both the SVR and MLP algorithms scored similarly on the explainability assessment analysis. They scored a low for the explainability of their feature importance, a low/medium for the explainability of their general algorithmic structure, a medium for the explainability of the performance criteria and a low for the explainability of the error minimisation function. The SVR algorithm reached R-squared scores of 0.79, 0.51, 0.74 and 0.67 for NI 5y, PN 5y, NI 10y and PN 10y respectively while the MLP reached R-squared scores of -0.23, 0.4, -0.1 and -0.9 for NI 5y, PN 5y, NI 10y and PN 10y respectively. The results show that the level of explainability, but also the performance of models varies greatly depending on the model that is looked at. Because of this, depending on the specific requirements regarding model explainability and performance different algorithms can be suitable in different scenarios.

A rough summary of the results is shown in table 25. Keep in mind that the discussed R-squared results represent the model results on historic data. The negative R-squared results for the Multiple

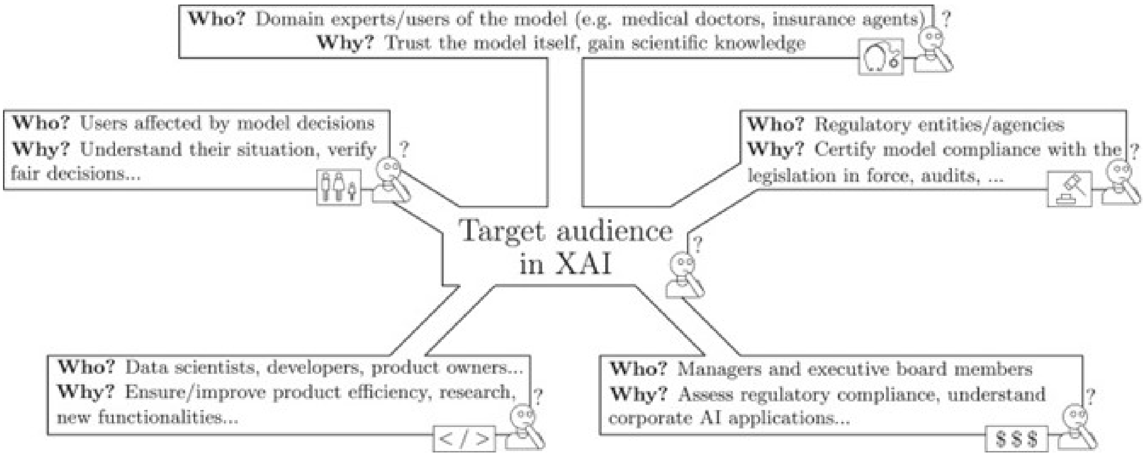
Linear Regression, Decision Tree Regressor and the Artificial Neural Network models indicate that these models are not suitable for the future demand prediction of identification documents. Since the results are negative the predictions made with these models are worse than a horizontal line representing the average historic demand. They can therefore not be used for the future demand prediction of identification documents. However, although these models currently do not perform well for this specific problem, they might still be effective for future machine learning problems tackled by the ministry. Therefore they are still included in the upcoming analysis with the side note that their results for this specific problem are inadequate. When further discussing these models the focus will therefore be on their explainability scores and their theoretical performance.

Table 25: Summary of the explainability and performance assessment per model

model	Average Explainability score	Average R-squared results
Multiple Linear Regression	High	Negative
Decision Tree Regressor	High	Negative
Random Forest	Medium	0.15 - 0.2 excluding NI 10y
Gradient Boosting	Medium	0.55 excluding NI 10y
Support vector regression	Low	0.7
Artificial Neural Network	Low	Negative excluding PN 5y

7.3 The influence of varying model requirements on the choice of algorithm

The specific requirements regarding the level of explainability of a model and the performance of a model are very dependent on the context in which the model is used. For example, when using a model that deals with sensitive personal data the need for an explainable model is much higher than when a model is used to predict the weather where predictive performance is more important. Additionally, the requirements can vary a lot depending on the target audience whose input is asked as was seen in the paper by (Barredo Arrieta et al., 2020) which was briefly discussed during the literature review. The upcoming section will use the findings from (Barredo Arrieta et al., 2020) and especially their figure as seen in 29, as a basis for further analysis by using it as inspiration to develop four hypothetical scenarios. Each scenario describes how the requirements for a model change depending on the target audience that is kept in mind and the context in which the model is to be implemented.



Who? Domain experts/users of the model (e.g. medical doctors, insurance agents)
Why? Trust the model itself, gain scientific knowledge

Who? Users affected by model decisions
Why? Understand their situation, verify fair decisions...

Who? Regulatory entities/agencies
Why? Certify model compliance with the legislation in force, audits, ...

Who? Data scientists, developers, product owners...
Why? Ensure/improve product efficiency, research, new functionalities...

Who? Managers and executive board members
Why? Assess regulatory compliance, understand corporate AI applications...

Figure 29: Different target audiences for explainability (Barredo Arrieta et al., 2020)

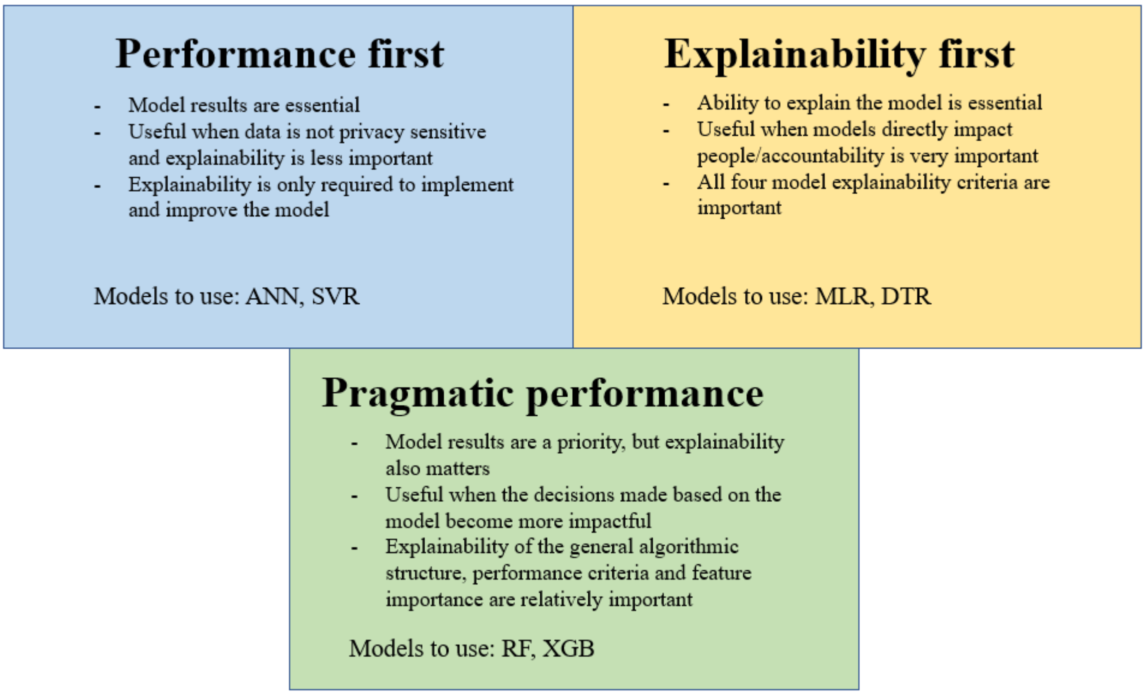
For the first scenario the preferred requirements from the perspective of a Data scientist, shown at the bottom left of the figure, are described. Imagine a data scientist working on a prediction model in which no sensitive data is used and model accuracy is essential. In such a scenario model performance is the main requirement while model explainability is less relevant. All that matters is that the data scientist him- or herself understands how the model works so that it can be optimised to reach the highest results. This requires a personal understanding of the model but does not require the model to be easily explainable to other people. In a scenario such as this model performance is the main priority while model explainability is less relevant. This scenario is labelled as performance first.

The second scenario discusses the preferred requirements for a machine learning model from the perspective of the people affected by the model. Since models are increasingly used in everyday life for example to determine whether or not someone is eligible for a loan, it makes sense to think about possible requirements posed by the people affected by a model's decision(s). Especially when a model negatively influences someone's life it is only fair that the decision can be explained. For the people affected by machine learning models, decision explainability is therefore of utmost importance. Not only to understand why a certain decision has been made, but also to validate whether the decision was valid at all. In a scenario such as this, the requirements are opposite of those discussed in the previous scenario. Although model performance is of course still important, the main requirement for the users affected by a model's decision is explainability instead of performance. This scenario is labelled as explainability first.

The two aforementioned examples gave two relatively extreme examples of different possible requirements regarding model explainability and performance. A more lenient example, however, also exists. The third and final scenario sketches the important requirements from the perspective of the domain experts and users of the model shown at the top of the figure. Imagine a scenario in which a machine learning algorithm is used to diagnose non-terminal patients. In such a scenario the users of the model want to understand the model in order to trust the results, but also require accurate results since the model would have no added value otherwise. Therefore they need a model that can perform well while at the same time can be explained to an extent that trust is created in the results. Thus the main requirements consist of a combination of both performance and explainability. This scenario is labelled as pragmatic performance.

A fourth scenario discussing the combined requirements of the regulatory entities, the managers and board members can also be designed. However, from a practical point of view, such a scenario would not differentiate much from the pragmatic performance scenario. Again a combination of performance and explainability would namely be required. Explainability so that the stakeholders understand what the models are about while and performance so that the models add value for the stakeholders. Since the scenarios would be almost similar it is not further developed and the pragmatic explainability scenario is used for both groups of stakeholders.

The previous scenarios have shown that the model requirements are very sensitive to the context in which the model is used. Additionally, they have shown that when a decision maker has to incorporate the requirements of multiple different stakeholders in his decision for a specific model the occurrence of competing requirements is almost unavoidable. In the case of competing requirements, it can be difficult to decide which algorithm fits the requirements the best due to the many different algorithms that exist. As a first step in streamlining this decision-making process, the scenarios described are used to develop three profiles representing different situations with regard to model requirements. The profiles are shown in figure 30 and give a brief overview of the context in which they are valid, which explainability criteria are likely to be important and which models are expected to be useful. A more extensive description of the profiles is given in the upcoming subsections.



Performance first

- Model results are essential
- Useful when data is not privacy sensitive and explainability is less important
- Explainability is only required to implement and improve the model

Models to use: ANN, SVR

Explainability first

- Ability to explain the model is essential
- Useful when models directly impact people/accountability is very important
- All four model explainability criteria are important

Models to use: MLR, DTR

Pragmatic performance

- Model results are a priority, but explainability also matters
- Useful when the decisions made based on the model become more impactful
- Explainability of the general algorithmic structure, performance criteria and feature importance are relatively important

Models to use: RF, XGB

Figure 30: Model decision profiles

7.3.1 Performance first

As the name of the profile indicates, the performance first profile focuses on model performance. This profile represents the scenario in which model performance is the main requirement while model explainability is almost if not completely irrelevant. As seen in the profile the statement 'explainability is only required to be able to implement and improve upon the model' is added to the profile. This statement is based on findings from Burkart and Huber (2021) and is used to show that even if the focus is purely on performance explainability is still required if only so that person implementing the model can reach optimal results. However, since model explainability does not play a big role anymore the reduced explainability present in the more complex algorithms is not a limitation any more. This makes the more complex algorithms attractive for the performance first profile since more complex models such as support vector regression usually tend to perform better than less complex algorithms as was found in section 3.8. The biggest downside to the use of more complex models is that they require more expertise by the developer to be able to implement and optimise them. Additionally extracting the feature importance is almost impossible as was seen during their explanation in section 5. This also showed in the explainability assessment of the SVR and MLP algorithms where they both scored low for the explainability of the feature importance.

A decision maker should look at this profile when the main model requirement is model performance. This might be the case when model predictions have little to no real-life impact, the model does not use any sensitive/personal data or the accuracy of the result is very important and the prediction result can be validated by an expert afterwards.

7.3.2 Explainability first

The explainability first profile distinguishes itself by its main focus on explainability. In this regard, it is the exact opposite of the performance first profile. The algorithms relating to this profile, multiple linear regression and decision tree regression, are both relatively simple which makes them easy to explain no matter the explainability assessment criteria looked at. This profile is therefore very interesting when trust in the algorithm, which is one the main reasons for explainability as was shown by both Burkart and Huber (2021) and Siau and Wang (2018), is important.

When dealing with complex prediction problems the explainability first profile is, however, likely not the first profile to be used. Although the profile represents the model requirements for people affected by a model very well, the level of explainability required in such cases is usually so high that it can only be accomplished using MLR or DTR which often struggle when the problem gets to complex. The main weakness of fully focusing on the explainability of a model is therefore that it potentially limits the predictive performance to such an extent that the predictions themselves become unreliable. This would for example be the case when the MLR or DTR algorithms would be used to predict the future demand for identification documents even though their R-squared results showed that the models are incapable of accurately predicting the demand. If, which is highly unlikely, these models would still be used the corresponding results would be very inaccurate and would therefore negatively impact the people affected by them more than the added explainability would add as value. Only in cases where the prediction problem is simple enough that these models can reach adequate results should they be used. Thus when choosing the explainability first profile it is very important to first analyse the predictive performances before continuing. If no adequate results can be reached the explainability standards should either be lowered to enable the use of less explainable algorithms or a different, more explainable, technique should be looked for all together. At the same time, if the simple model can predict well, the added explainability can create a lot more trust and positive attitudes with regard to the model's results and the future use of machine learning algorithms.

This profile fits well when the decision maker has to decide upon a model to use when the explainability of the model is very important and the prediction problem is simple enough that models such as MLR and DTR can reach satisfying results. Additionally, this profile fits well when the aim is to introduce people to machine learning algorithms.

7.3.3 Pragmatic performance

The pragmatic performance profile is applicable in situations where model performance is the main requirement but model explainability is important to create trust in the model results. To create a general sense of trust in the algorithms it is assumed that the algorithmic structure, the feature importance and the performance criteria of the algorithm need to be at least partly explainable. The explainability of the error minimisation function is left out since it is deemed too complex given the aim of creating a general sense of trust. This profile is therefore most suitable for algorithms such as Random Forest or Extreme Gradient Boosting due to their generally good performance while still being relatively explainable as has been shown by their results in the explainability assessment. Depending on whether or not there are specific elements of the model that need to be explainable the choice between RF and XGB can be influenced since XGB's general algorithmic structure is a bit more complex than RF's. Additionally, since this profile focuses on pragmatic performance more complex algorithms such as SVR or MLP may also be used if the explainability requirements set by the decision makers are met by these algorithms. However, as a general guideline RF and XGB are recommended. The potential downside to these algorithms is that they might not perform well in which case they neither succeed at performance and explainability as is currently the case for the random forest algorithm when predicting the future demand for identification documents. In such cases, different models might be more suitable.

This profile fits well with decision makers that need to decide on a model to use which will have an impact in real life for which trust is required. If this profile is chosen the decision maker needs to know what the specific requirements are with regards to explainability since it might allow the use of more complex algorithms than those which are initially suggested.

7.4 Notes on model recommendations

The profiles discussed above each recommend a set of machine learning algorithms to use given the specific context of the profile. The models suggested are recommendations that are taught to generally fit the model requirements proposed within the corresponding scenario quite well. However, the performance of machine learning algorithms depends on a lot of different factors. Is the prediction linear or non-linear? Is the data clean and does it contain predictive features? Are the hyperparameters of the algorithms properly optimised etc... Because of this the performance of the algorithms can vary greatly between prediction problems. Just because the decision tree algorithm did not reach satisfying results for the future demand prediction of identification documents does not mean it will not work well for other prediction problems. When working with ML algorithms it is therefore always important to test how well they can perform. The recommended algorithms should therefore not be seen as set in stone. Instead, they should be seen as a general handhold for decision makers that need to make an initial decision regarding what algorithm to use given their specific model implementation context.

7.5 Model requirements by the RvIG

Within the context of the case study certain requirements with regards to model performance and explainability also exist as have been identified in section 2. The following discussion goes over these requirements and provides a recommendation with regard to the most suitable profile. For the RvIG the aim of the model is to accurately predict the future demand for identification documents. Since no personal or privacy-sensitive data is needed for this prediction and the model results do not directly impact people, the level of explainability of the model does not have to be the main priority. However, since the RvIG is a government organisation they can be held accountable when something goes wrong and therefore should be able to explain to stakeholders that are likely not very experienced with machine learning how the model works. This requires that at least part of the model is explainable and thus that the more complex models are likely not suitable. Furthermore, although the decisions made based on the model results might not directly impact people, they are used to determine the financial budget for the upcoming years. This means the prediction accuracy of the model is very important, but also that trust in the model results is required since a lot of money is involved. Ultimately this leads to the Pragmatic performance profile which represents a situation in which the model results are the priority while maintaining a certain degree of explainability. Suggested models in this profile are the Random Forest algorithm and the Extreme Gradient Boosting algorithm. Given the results of the models for the future demand prediction of identification documents, the XGB algorithm is recommended since it performed better than the Random Forest algorithm.

7.6 Decision support framework

The discussion of the model implementation scenarios and corresponding profiles has shown that the decision for the type of machine learning algorithm to use depends on a lot of different factors that are not always known beforehand. Given the findings by Barredo Arrieta et al. (2020) and the model implementation scenarios discussed previously, it is clear that the requirements regarding model explainability and model performance can vary greatly depending on the decision maker that is asked. When deciding on the type of model to use it is therefore very important to first identify the context in which the model is to be applied and to identify the stakeholders that are relevant. By first identifying the context of the model an initial assessment regarding potentially sensitive elements such as the use of personal data and the target audience of the model is done. The context of the model can then be discussed with the relevant stakeholders to identify model requirements. To facilitate this process the model decision profiles can be used to quickly help the stakeholders determine what kind of model requirements they think are important. Afterwards, specific requirements can be identified by asking the stakeholder about the four model explainability criteria. The decision maker can then decide what requirements to include and which to exclude.

The next important step is to, based on the context of the situation, identify which algorithms can potentially be used and what performance each of these algorithms can reach. This is important because the performance of machine learning algorithms can vary greatly between prediction problems (Caruana & Niculescu-Mizil, 2006). Looking at the results for the future demand prediction of identification documents, for example, the prediction results have shown that the only viable algorithms are the XGB and SVR algorithms. If the performance requirements are lowered, which may or may not be possible within the boundaries of the performance requirements, and the random forest algorithm can be further optimised the RF algorithm might also be a viable choice. Knowing this, the decision for what algorithm to use can already be specified to just three choices. To analyse the performance of the different algorithms someone familiar with machine learning algorithms is required. This means that either the decision maker needs to have this knowledge or he/she has colleagues available that can perform the tasks that need to be performed. Assuming this criterion is met, the model implementation steps performed during this research should be applied. First data has to be collected which likely requires some literature study and or domain knowledge. Then the data has to be cleaned and split into training and test data. Afterwards, each of the different machine learning algorithms has to be set up so that predictions can be made and their parameters have to be optimised. Then predictions can be made with the different algorithms using the training and test data and their relative performances can be analysed. If none of the models perform well and they can not be optimised further it must be concluded that machine learning algorithms are not a suitable solution for the prediction problem.

Lastly, depending on the specific requirements identified for model explainability the explainability assessments of the different algorithms may or may not have to be fine-tuned to include the specific requirements better. If a new machine learning algorithm is applied its explainability assessment should be performed during phase three as shown in the framework. By comparing the model results regarding performance and explainability with the model requirements a decision can be made regarding the type of algorithm to use for the prediction problem at hand. The steps described in this subsection have been formalised into a decision support framework consisting of two parts that aim to support decision-makers with their choice of a model when dealing with competing requirements. The first part of the framework is shown in figure 31 and gives a high-level overview of the different processes that are required to find a suitable algorithm. This part of the framework follows a traditional flow chart structure in which the different high-level steps required to determine what model to use are shown. The blocks with round edges represent the start and end blocks. The black rectangles represent processes the decision maker should perform and the yellow diamond represents a decision the user should make. The processes are categorised into four phases which are indicated by the coloured rectangles made out of dashed lines surrounding certain processes. Each phase contains a few processes which are elaborated upon in more detailed sub processes in figure 32. The second part of the framework consists of four squares representing each of the phases mentioned in the high-level framework. Each square contains more detailed processes that are suggested to fulfil the steps mentioned in the corresponding phase in the high-level framework. The arrows between the different squares represent the output of the previous square that is used as input for the following square. Underneath each square the chapters from this research corresponding to the phase are mentioned in case a more hands-on explanation is required.

Whereas the model decision profiles are meant to help stakeholders identify what model requirements they think are important by giving a high-level overview of different scenarios, the decision support framework goes deeper than that by including more of the nuances that occur when working with machine learning algorithms and competing requirements. Since it goes into more detail, using the framework requires more work but will also lead to a more specific result regarding what algorithm to use given the model requirements identified.

Since the framework focuses on identifying and incorporating different requirements regarding model performance and explainability so that a suitable model can be found, the framework is most suitable for decision makers that work in organisations where stakeholders have different requirements regarding model performance and model explainability. As is the case for the RvIG and most other governmental organisations. When competing model requirements occur it becomes harder to determine which algorithm to use which is when the framework adds the most value. The framework can also be used by decision-makers were competing requirements are less apparent. However, the decision support framework would then mostly add value as a handhold regarding the different steps required to implement and analyse machine learning models. The framework is not intended to support decisions on algorithms other than machine learning since it requires a comparative analysis based on explainability criteria that are specific to machine learning to determine which algorithm fits the requirements best.

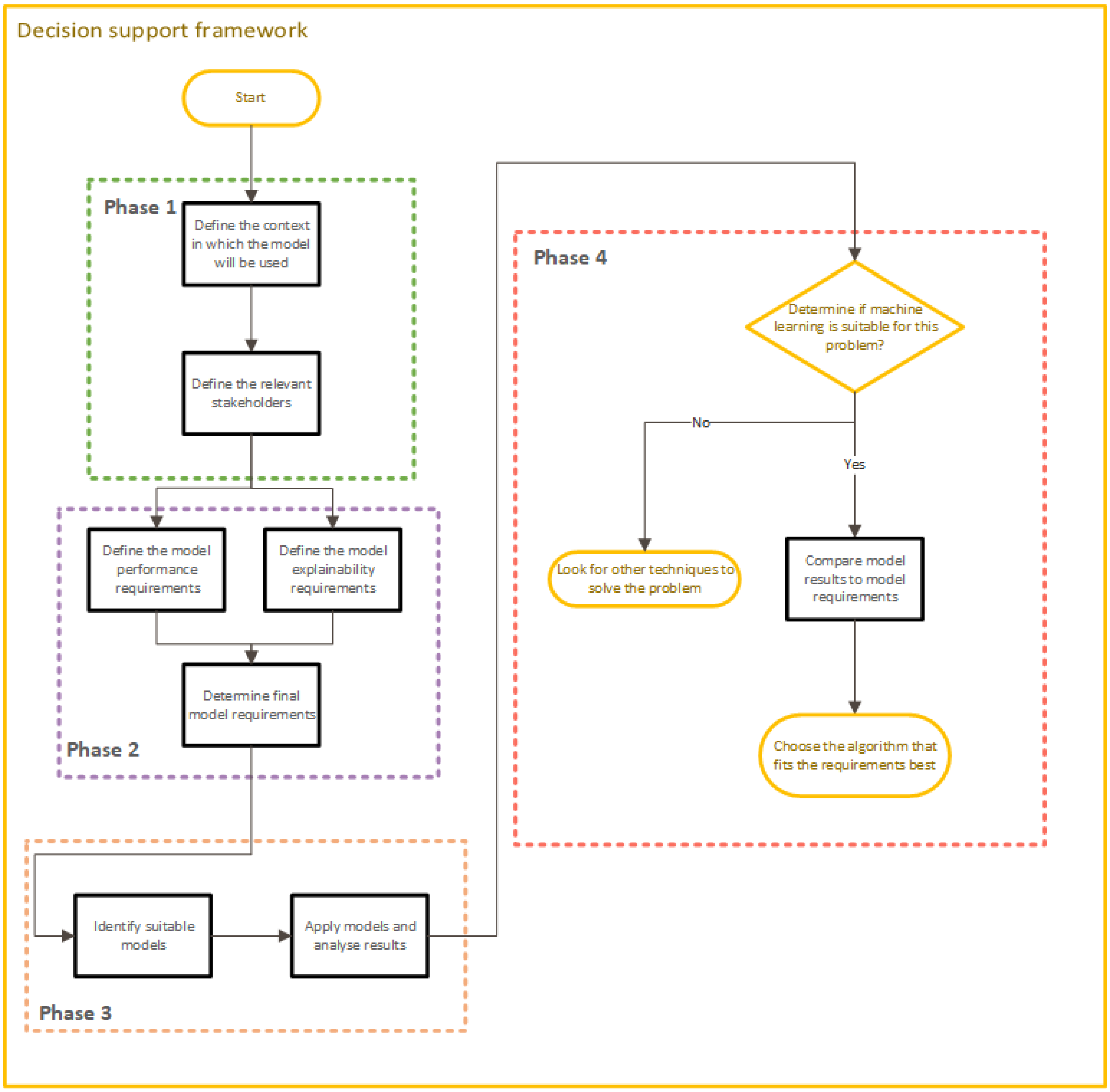

Decision support framework
Start
Phase 1
Define the context in which the model will be used
Define the relevant stakeholders
Define the model performance requirements
Define the model explainability requirements
Determine final model requirements
Phase 2
Identify suitable models
Apply models and analyse results
Phase 3
Phase 4
Determine if machine learning is suitable for this problem?
No
Yes
Look for other techniques to solve the problem
Compare model results to model requirements
Choose the algorithm that fits the requirements best

Figure 31: Decision support framework

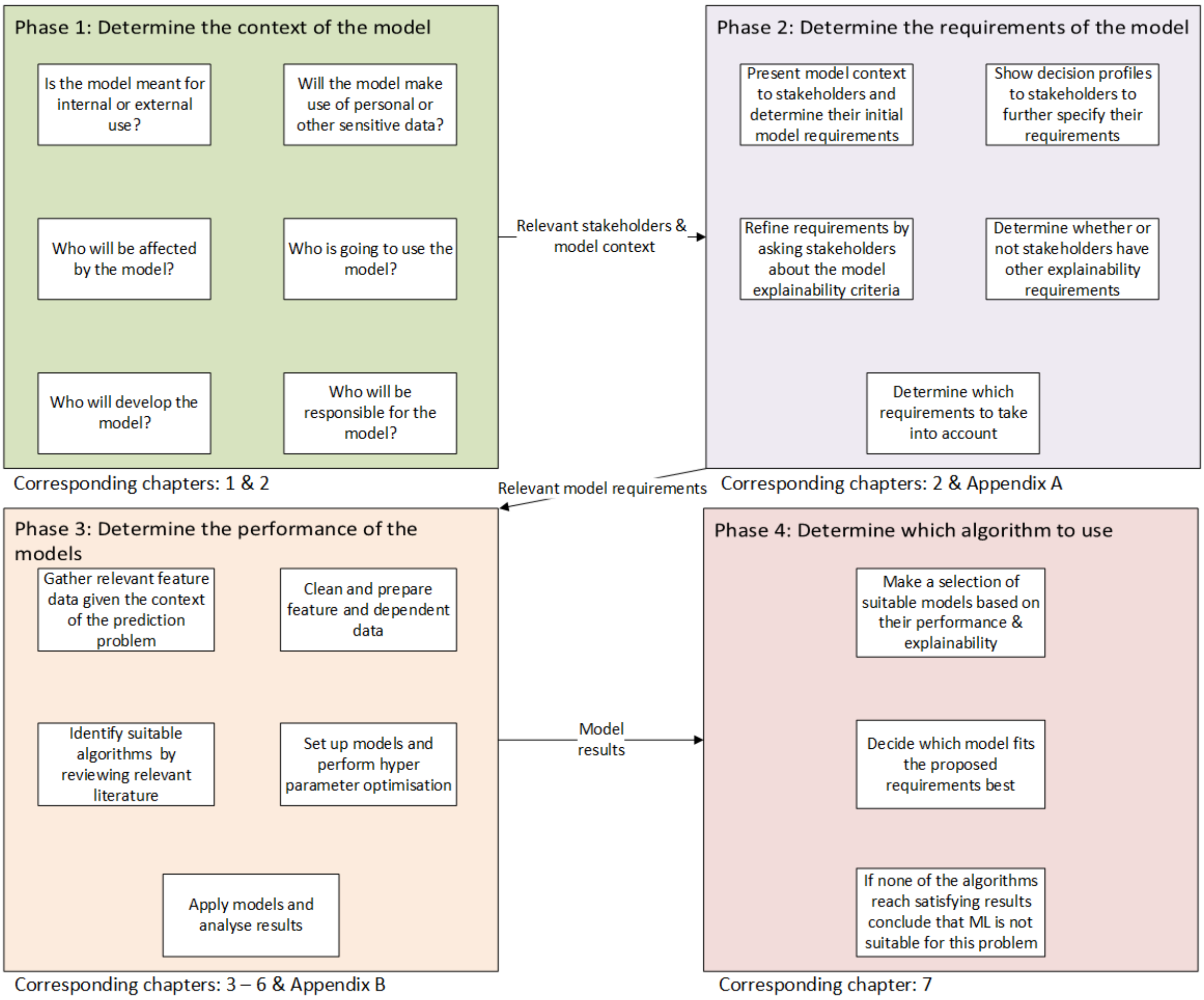

Phase 1: Determine the context of the model
Is the model meant for internal or external use?
Will the model make use of personal or other sensitive data?
Who will be affected by the model?
Who is going to use the model?
Who will develop the model?
Who will be responsible for the model?
Corresponding chapters: 1 & 2
Relevant stakeholders & model context
Phase 2: Determine the requirements of the model
Present model context to stakeholders and determine their initial model requirements
Show decision profiles to stakeholders to further specify their requirements
Refine requirements by asking stakeholders about the model explainability criteria
Determine whether or not stakeholders have other explainability requirements
Determine which requirements to take into account
Corresponding chapters: 2 & Appendix A
Relevant model requirements
Phase 3: Determine the performance of the models
Gather relevant feature data given the context of the prediction problem
Clean and prepare feature and dependent data
Identify suitable algorithms by reviewing relevant literature
Set up models and perform hyper parameter optimisation
Apply models and analyse results
Corresponding chapters: 3 – 6 & Appendix B
Model results
Phase 4: Determine which algorithm to use
Make a selection of suitable models based on their performance & explainability
Decide which model fits the proposed requirements best
If none of the algorithms reach satisfying results conclude that ML is not suitable for this problem
Corresponding chapter: 7

Figure 32: Decision support framework continued

7.7 Validating the framework

The developed framework is a novel design developed based on the findings of this research. To determine the value of the design it needs to be validated by the target audience it is intended for. A simple way of validating the framework would be to present it to either the RvIG or another organisation that aims to work with machine learning and deals with competing model requirements that make it difficult to determine which algorithm to use and simply ask them whether or not they see added value in the framework. Validating the framework in such a way is easy to perform and quickly gives an idea about the value of the framework. However, validating the framework in such a way is very subjective since each organisation likely has its own specific goals it aims to achieve using such a framework. Additionally, since the validation of the framework is subjective it also makes it hard to compare the framework's usefulness to that of other frameworks.

To better validate the framework more quantitative criteria should ideally be used by which to determine the value of the framework. This is also a key element of the relevance cycle proposed by Hevner (2007) which is discussed in section 4 as part of the methodological approach of this research. However, since the framework does not deal with quantitative elements itself it is hard to develop quantitative criteria for evaluation. Therefore inspiration is instead taken from (Stetler et al., 2006) who discuss the use of formative evaluation during the implementation phase of research findings as a way of evaluation. By keeping close contact with the people implementing the research findings, in this case the decision support framework, they evaluate if the finding is implemented and used correctly, evaluate what goes right and evaluate what goes wrong. By doing this they can make adjustments so that the research finding better fits within the context of the people intended to implement and use it. Based on this a set of qualitative criteria in the form of questions is developed which can be used to evaluate whether or not the framework reaches its intended goals. Since the use of the framework is seen as a success when it leads to a choice of model that fits the identified requirements well, either in terms of performance, explainability or a combination of both while being easy to use the following questions are proposed:

- Is the framework easy to interpret?
  - Is the structure of the framework easy to follow?
  - Are the processes clear and detailed enough?
  - Are all the detailed processes relevant?
- Does the use of the framework lead to a decision regarding what algorithm to use where it otherwise would not have or would have taken longer?
- Does the use of the framework lead to a model that fits the identified requirements better than previous models did?
  - Does the new model reach better results?
  - Is the new model more explainable?
  - Does it do both?

Due to time constraints the validation of the framework by the RvIG using the aforementioned question could not be performed. Future researchers are therefore recommended to apply and evaluate the framework to determine its empirical value. Given sufficient time and resources two experiment groups could be created in which both groups have to find a suitable algorithm given a specific context. One group could then be given the framework while the other group does not get any added materials. Afterwards the recommended algorithms can be compared to the proposed requirements and it can be determined which group recommended the best algorithm. If the group with the framework reaches a better recommendation enough times its added value can be concluded.

7.8 Conclusion

Given the varying and sometimes competing requirements regarding model explainability and model performance three different scenarios are discussed to indicate how different requirements lead to different models. Each scenario led to a profile which gives a general sense of the situation in which it is applicable which should help decision-makers to get a better idea of which model to use in situations with different model requirements. Using the findings from the research and the aforementioned profiles a novel decision support framework is conceptualised that shows what steps are needed to determine the context in which the model will be used, determine the relevant stakeholders, identify the model requirements and determine which models are suitable. By performing these steps the decision maker can then determine which algorithm fits the identified requirements the best. In case the machine learning algorithms are unable to accurately predict or the model implementation context turns out to be too sensitive, it is also possible for the decision maker to decide not to use machine learning models at all. In the case of the future demand prediction of identification documents, however, the requirements fitted best with the Pragmatic performance profile and the recommended algorithm to use is the the XGB algorithm due to its relatively high performance and while remaining relatively well explainable.

8 Conclusion, Discussion and Recommendations

8.1 Introduction

The following section consists of three elements. First, the conclusion of the research is given. Secondly, the research choices are reflected upon in the discussion. To conclude, the section ends with a recommendation for future research. Figure 33 shows the research element this section discusses.

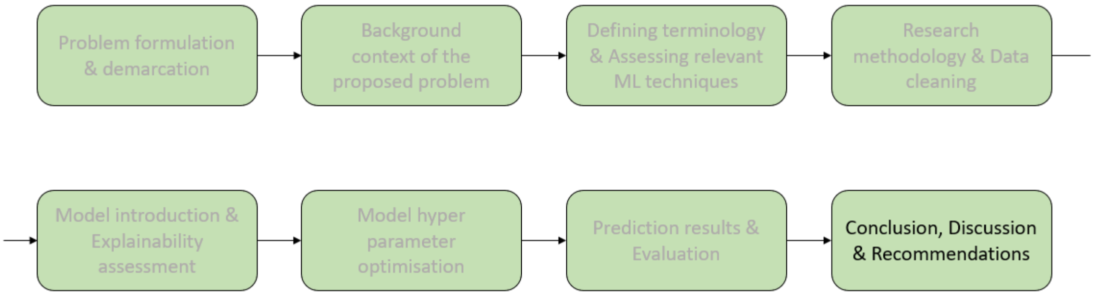

Problem formulation & demarcation
Background context of the proposed problem
Defining terminology & Assessing relevant ML techniques
Research methodology & Data cleaning
Model introduction & Explainability assessment
Model hyper parameter optimisation
Prediction results & Evaluation
Conclusion, Discussion & Recommendations

Figure 33: Current step in the research process

8.2 Conclusion

This research aimed to answer the following research question:

How to deal with competing requirements regarding model explainability and performance while choosing what machine learning model to use for multivariate time-series analysis?

To answer this research question five supporting sub-questions were formulated. Each of the sub-questions focused on its own specific research element that when combined could be used to answer the main research question. The five sub-questions were:

- How can the term explainability be operationalised in order to assess different machine learning models?
- What machine learning algorithms are suitable for long-term multi-variate time-series analysis?
- How do the different machine learning models score with regards to explainability?
- What predictive performance levels can be reached using the identified machine learning algorithms?
- How do competing requirements regarding explainability and performance influence the choice of multivariate time-series analysis models?

To answer the main research question a design science approach was adopted based on Hevner (2007). Using the relevance and rigor cycles a decision support framework is developed that can support decision-makers in their choice of machine learning model when dealing with competing requirements. The basis of the framework is based on the first three subquestions. Here the rigor and design cycles are applied to their full potential. Using scientific literature the different applications of explainability were categorised to develop a new explainability assessment table. Further literature search identified six machine learning algorithms that showed potential for multivariate time-series analysis, namely: MLR, DTR, RF, XGB, SVR and MLP. By combining the results of the first two subquestions the identified algorithms were assessed on their explainability. The assessment showed that MLR and DTR were the most explainable and that complex algorithms such as SVR and MLP, which are usually labelled as black box algorithms, can still be explained when looking at specific explainability criteria. Labelling them as black-box algorithms is therefore not valid. The fourth subquestion identified the predictive performance per algorithm. For the future demand prediction of identification documents using historic data the XGB and SVR algorithms turned out to perform best. These results correspond with literature discussed in section 3.8 which argued that XGB and SVR often perform better than most other models. The lacking performance of the MLP further shows that more complex models do not always result in better results. Instead, model performance also depends a lot on the complexity of the prediction problem and the available data. None of the models were able to reach satisfying results when performing direct multi-step forecasting.

The fifth subquestion analysed how competing requirements regarding model explainability and performance, as identified previously, influence the choice of machine learning model. Here, three profiles were conceptualised that represented three different model implementation scenarios. The profiles were called Performance first, Explainability first and Pragmatic performance. Each profile represents a specific scenario in which specific model requirements are required. Linked to these profiles are recommended models that likely fit the model requirements based on their explainability and expected performance score. The results of subquestion five showed that the choice of model is very dependent on the context in which the model is to be used and the requirements the model should adhere to. However, due to the many different machine learning models that exist, even when knowing the context and requirements it can still be hard to determine which machine learning algorithm to use. Therefore, as a final design and answer to the main research question a novel decision support framework is conceptualised. This framework aims to support decision-makers in choosing a fitting machine learning model while dealing with competing requirements.

The first part of the framework goes over the processes required to determine which algorithm to use on a high level of abstraction. It starts by defining the context of the model and the relevant stakeholders. Then the model requirements are identified and suitable models are found. The high-level framework ends by implementing the models, analysing their results and determining which model fits the requirements best. Within the high-level framework four phases are identified which are further elaborated upon in the second part of the decision support framework. In the second part of the framework each of the four phases identified in the initial high level framework is elaborated upon in more detail by showing what specific steps should be performed. Additionally, each of the phases refers to chapters in the research that correspond to the suggested steps for when more detail is needed. Based on the requirements identified for the RvIG the suggested model to use resulted in the XGBoost algorithm. Given the main need for performance in combination with explainability to create trust, the RvIG could be linked to the pragmatic performance profile. This profile initially suggests two algorithms, namely random forest and XGBoost, however, after implementing the algorithms the results showed that the XGB algorithm can reach much higher accuracy levels than the RF algorithm. So although the RF is slightly more explainable the XGB algorithm is still recommended due to its higher performance. The SVR algorithm reached even better results but is also more less explainable which is why it is not recommended instead.

8.3 Discussion

During the development of this research multiple decisions, both in terms of model design and general conceptualisation of ideas, have been made which have influenced the results of the research. The following two paragraphs start with a brief discussion relating to the model-specific decisions. Afterwards, the design approach and the developed designs are reflected upon.

To start, multiple assumptions were made with regard to the used data. While transforming the yearly data to monthly values a linear growth between months was assumed. This assumption was made due to a lack of knowledge regarding the real monthly development of the feature variables and made it possible to still retrieve monthly values. However, due to the assumed linear growth the feature values always increase or decrease linearly in value which is not representative of real life. Take for example the feature regarding the number of holidays taken per month. The number of holidays generally shows a periodic peak during December/January and July/August which correspond to the school holiday periods. This periodicity is currently not taken into account since the number of holidays grows linearly each month. This likely influenced the prediction performance of the models since the four dependent variables all show a yearly periodic peak during the holiday seasons as well. Instead of assuming a linear growth more realistic assumptions should be used to create monthly values. This should increase the quality of the feature data and therefore also the performance of the models. A similar assumption was made when transforming the date value from a string value to an integer value. Since dates represent a period in time, earlier dates received a lower ordinal value and more recent dates received a higher ordinal value. However, since the demand goes up and down throughout the year and shows a yearly peak during summer, a constantly increasing date value does not represent this properly. Instead, better results might be reached by disregarding the year value and only using the month values. The months can then be coded in a binary fashion so that it is always clear which month each row of data represents.

When performing hyperparameter optimisation a grid search optimisation algorithm was used. Because this algorithm checks every combination of hyperparameters that is supplied the computational costs increase exponentially without a guarantee for optimal hyperparameters. Because of this and the lack of computational power that was available only a limited grid search could be performed. It is therefore likely that part of the machine learning algorithms used sub-optimal hyperparameters when predicting. Instead of using a grid search algorithm for hyperparameter optimisation a random search or Bayesian/Gradient-based optimisation algorithm could have been used. Since the logic behind these optimisation algorithms is more advanced they should be more efficient and have a higher chance of finding better-performing hyperparameters. When adding lagged data to the datasets, the lagged data only went back half a year. Since the demand shows yearly patterns better results might have been reached if lagged data of up to a year was added.

Apart from reflecting on the model-specific decisions that have been made, the general design approach and decisions can also be reflected upon. For this research's design approach Hevner's design cycles were adopted. Generally, this approach worked quite well. The relevance cycle was used to determine the knowledge gap, the need for a new prediction model and validate the final design. The rigor cycle worked well when dealing with the formalisation of different concepts such as Machine Learning and explainability. The use of the rigor cycle could have been extended upon more by also including the analysis of existing designs a bit more. This was not included during the research which meant the entire design cycle had to be performed from scratch. Although the design cycle usually implies something previously non-existent is designed, a more extensive rigor cycle could have potentially identified design standards that have been empirically proven to work.

While formalising the explainability criteria the focus lay on finding four distinct elements within machine learning algorithms to explain. It is however uncertain whether or not these four categories are indeed the main categories stakeholders care about when talking about the explainability of machine learning models. Depending on the stakeholder the important explainability categories could potentially be different. Nonetheless, the identified categories fit well when the aim is to get a general understanding of an algorithm and it provides more nuance compared to simply dividing algorithms into black box and non black box algorithms. In case a stakeholder indeed desires to assess the models on different criteria, the current categories can still function as good inspiration.

A further point of discussion is that the scoring of the algorithms using these categories remains inherently subjective. To reduce subjectivity an ordinal scoring system was developed using Low, Medium and High scores and the importance of defining the target audience was stressed. However, the eventual assessment of the algorithms was still performed by the researcher. It, therefore, remains relatively unsure whether or not the results of the explainability assessment correspond to what the stakeholder thinks. These results could have been made more reliable through a more extensive relevance cycle in which the stakeholder was included more during the explainability assessment of the models.

The three decision profiles that were created give an easy-to-interpret idea of different model implementation contexts together with recommendations of models that will likely fit well. Due to the way the profiles are represented they can only include a limited amount of information before getting too crowded. because of this the amount of information shown in the profiles is limited on purpose. This made the profiles themselves clear but also less in-depth. If the decision maker cares about a specific explainability category this is now only briefly discussed in the profile. To compensate for this the textual descriptions of the profiles go into a bit more depth, however, this is less easy to interpret than when just a figure is used. To increase the detail of the profiles maybe a different design can be used in which a separate figure is made for each of the three profiles individually so that they can contain more information compared to the current situation in which one figure contains all three profiles. Lastly, instead of extending upon the profiles by adding more detailed information to them, the use of more profiles might also help. If more profiles are added each of them can focus on a more specific model implementation scenario which can be based on even more specific model requirements I.E. different explainability criteria. This way the individual profiles should remain easy to interpret while also being more detailed due to their specific focus.

The final design of this research resulted in a decision support framework aimed to support decision-makers with their choice of what model to use while having to account for competing requirements. The framework consists of two parts. The first part uses a flow chart structure to show which processes should be performed to decide which algorithm to use. It does this on a relatively high level of detail so that the main processes remain clear and easy to interpret. To compensate for the lack of detail in the first part of the framework the processes are categorised into four distinct phases which are further elaborated upon in the second part of the framework. Here more detailed processes are suggested that should help the decision maker perform the processes mentioned in the corresponding phase. The detailed processes suggested assume that the level of detail in these processes is enough for the decision maker to know what to do. It might however be possible that a decision-maker would need more detailed instructions before being able to perform the suggested processes. If this is the case it would likely mean that each suggested process needs to be elaborated upon using an even more detailed step by step description. If this is the case the use of a framework is not recommended anymore because the amount of detail needed for such a framework would be too high. Instead the use of a specific manual per step might then be a good solution.

The framework also contains an important decision regarding the use of machine learning algorithms stating that if none of the algorithms shows suitable results machine learning should not be used. Although not directly included in the framework because it is slightly out of scope, this statement can be extended by stating that even if the machine learning algorithms can reach acceptable results there might still be other, none machine learning, techniques that can reach better results. Additionally, this framework assumes that the user is already aware of the potential that machine learning might have in his/her field of work, however situations in which the use of machine learning is not responsible of course also exist. For example, when a lot of personal data is handled, explainability of all parts of the algorithm is very important or when the model implementation context is under high interest. In such cases, the use of machine learning might not be optimal due to its potential lack of explainability. Other, more explainable, algorithms might then be a better solution given that it is a problem than can be fixed with algorithms at all. Ultimately the decision-maker should be aware of the fact that machine learning is not a one size fits all solution. Machine learning can be very useful in situations where enough (clean) data is available, enough expertise is at hand and the model implementation context is not very sensitive. If any of these criteria are not the value of machine learning drops significantly and other solutions are likely better.

Lastly, in order to know the practical value of the framework it needs to be validated. However, due to time constraints this was not possible. Instead, only a suggestion is made on how to validate the frame work which remains quite qualitative. if possible, a quantitative validation strategy should be used instead to get a more robust and trustworthy validation.

8.4 Recommendations

Based on the performed research a small research agenda of three different potential research areas is proposed. The three suggestions focus on different areas that build upon the results found in this research.

The first research suggestion focuses on an extended analysis of the future demand prediction of identification documents. Due to the limited time and computational power available, the implementation of the algorithms during this research was not always optimal. The main example is the limited hyperparameter optimisation that was performed due to a lack of computational power and the choice for a grid search optimisation algorithm. Future research could dive deeper into what data has predictive value by analysing more data sources and using feature importance analysis to determine which features were most important for the prediction. With new data the research could further analyse the predictive performance of different algorithms and/or the same algorithms but with better-optimised hyperparameter values. If the latter is chosen it can be done by running different optimisation algorithms or by running grid search optimisation with more combinations of hyperparameters. Although the current research focused specifically on the use of machine learning algorithms to predict future demand, future research could take a step and look for different prediction techniques altogether. This research suggestion can be formalised through the following main research question: "How can the future demand of identification documents be predicted?"

The second research suggestion focuses on a more detailed version of the decision support framework. Currently, the framework assumes that the person deciding what algorithm to use under competing requirements has enough knowledge regarding machine learning algorithms to assess all the different processes mentioned in the framework on his own. Instead, the current framework could be used as a baseline to develop a more detailed version, either for the entire framework or for parts of the framework that require less individual knowledge. Since developing an actual model would not be necessary for this research the researchers could spend more time doing empirical research. For example in the form of interviews with stakeholders that have either already implemented machine learning models or are deciding on whether or not to use machine learning models. The empirical knowledge gathered through these interviews can potentially lead to new insights with regard to factors that influence the decision on what model to use which can provide added value to a new decision support framework. Instead of focusing purely on the factors influencing the decision for a specific algorithm the research could also use the interviews to gather knowledge on what is desired from a decision support framework. Insights from such an analysis might increase the effectiveness of the current decision support framework without having to add new elements or extend upon old ones. This suggestion can be formalised using the following research question: "How can the decision support framework used to determine what model to use under competing requirements be improved upon?"

The third and final research suggestion focuses on creating standard definitions and applications of explainability further. The results from the literature review have shown that there exists a lack of standards with regards to the application and definition used for describing explainability. Due to the lack of standards researchers need to either evaluate tons of different definitions to find a definition that fits their research or use the different definitions as inspiration to define their own definition. Not only does this cost a lot of valuable time that could have been spent elsewhere, but it also adds to the overall chaos surrounding the definition of the term. Since the definition of explainability varies per research, designing standard metrics to measure the explainability of different machine learning algorithms is also challenging. Similar to a standard definition of explainability, standard metrics would save a lot of time and would make it easier to compare results between different research. This research tried to categorise the different definitions of explainability to create an overview that researchers can use when using the term explainability for their research. Similarly, a novel explainability assessment table was developed to provide general guidelines on which to assess machine learning algorithms. Both designs are a start to further standardisation of the term and use of explainability, but can certainly be improved upon. However, before entirely new frameworks, standards and definitions are developed it might be more useful to first analyse what the requirements are for the successful implementation of a standard definition and use of the term explainability. This research goal can be formalised using the following research question: “What are the limiting factors for the implementation of a standardised definition and use of the term explainability and how can these be solved?”

References

Abedi, V., Avula, V., Chaudhary, D., Shahjouei, S., Khan, A., Griessenauer, C. J., . . . Zand, R. (2021, 3). Prediction of Long-Term Stroke Recurrence Using Machine Learning Models. *Journal of Clinical Medicine 2021, Vol. 10, Page 1286*, *10*(6), 1286. Retrieved from https://www.mdpi.com/2077-0383/10/6/1286/htmhttps://www.mdpi.com/2077-0383/10/6/1286 doi: 10.3390/JCM10061286

Ahmed, A., & Khalid, M. (2017). Multi-step Ahead Wind Forecasting Using Nonlinear Autoregressive Neural Networks. *Energy Procedia*, *134*, 192–204. Retrieved from www.sciencedirect.comwww.sciencedirect.comwww.elsevier.com/locate/procedia1876-6102 doi: 10.1016/J.EGYPRO.2017.09.609

Alzubi, J., Nayyar, A., & Kumar, A. (2018, 11). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, *1142*(1), 012012. Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012/meta doi: 10.1088/1742-6596/1142/1/012012

Anele, A. O., Hamam, Y., Abu-Mahfouz, A. M., & Todini, E. (2017, 11). Overview, Comparative Assessment and Recommendations of Forecasting Models for Short-Term Water Demand Prediction. *Water 2017, Vol. 9, Page 887*, *9*(11), 887. Retrieved from https://www.mdpi.com/2073-4441/9/11/887/htmhttps://www.mdpi.com/2073-4441/9/11/887 doi: 10.3390/W9110887

Awad, M., & Khanna, R. (2015). Support Vector Regression. *Efficient Learning Machines*, 67–80. Retrieved from https://link.springer.com/chapter/10.1007/978-1-4302-5990-9_4 doi: 10.1007/978-1-4302-5990-9{_}4

AYLIEN. (2022). *Support Vector Machines for dummies; A Simple Explanation - AYLIEN News API.* Retrieved from https://aylien.com/blog/support-vector-machines-for-dummies-a-simple-explanation

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., . . . Herrera, F. (2020, 6). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. doi: 10.1016/J.INFFUS.2019.12.012

Beheshti-Kashi, S., Karimi, H. R., Thoben, K. D., Lütjen, M., & Teucke, M. (2015, 1). A survey on retail sales forecasting and prediction in fashion markets. *http://mc.manuscriptcentral.com/tssc*, *3*(1), 154–161. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/21642583.2014.999389 doi: 10.1080/21642583.2014.999389

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019, 12). What is Machine Learning? A Primer for the Epidemiologist. *American Journal of Epidemiology*, *188*(12), 2222–2239. Retrieved from https://academic-oup-com.tudelft.idm.oclc.org/aje/article/188/12/2222/5567515 doi: 10.1093/AJE/KWZ189

Brownlee, J. (2019, 8). *How to Convert a Time Series to a Supervised Learning Problem in Python.* Retrieved from https://machinelearningmastery.com/convert-time-series-supervised-learning-problem-python/

Brummelkamp, G., Hoevenagel, R., & Witkamp, A. (2020). *Monitor Identiteit 2019, Bezit, Gebruik en Misbruik van Identiteitsmiddelen.* Rijksoverheid. Retrieved from https://www.rijksoverheid.nl/documenten/rapporten/2020/02/29/monitor-identiteit-2019

Burkart, N., & Huber, M. F. (2021, 1). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, *70*, 245–317. Retrieved from https://www.jair.org/index.php/jair/article/view/12228 doi: 10.1613/JAIR.1.12228

Bzdok, D., Altman, N., & Krzywinski, M. (2018). THIS MONTH Statistics versus machine learning. *Nature Publishing Group*, *15*. doi: 10.1038/nmeth.4642

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *ACM International Conference Proceeding Series*, *148*, 161–168. doi: 10.1145/1143844.1143865

CBS. (2022a). *Bevolkingspiramide.* Retrieved from https://www.cbs.nl/nl-nl/visualisaties/dashboard-bevolking/bevolkingspiramide

CBS. (2022b). *StatLine - Approaches of domestic product (GDP); National Accounts.* Retrieved from https://opendata.cbs.nl/statline/#/CBS/en/dataset/84087ENG/table?ts=1647521285705

CBS. (2022c). *StatLine - Personen met een rijbewijs; rijbewijscategorie, leeftijd, regio, 1 januari.* Retrieved from https://opendata.cbs.nl/#/CBS/nl/dataset/83488NED/table?defaultview&dl=135C2

CBS. (2022d). *StatLine - Vakanties van Nederlanders; kerncijfers.* Retrieved from https://opendata.cbs.nl/#/CBS/nl/dataset/84363NED/table

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(1), 20–28. Retrieved from https://www.jastt.org/index.php/jasttpath/article/download/65/24

Cranenburgh S. (2021). *Lecture: Introduction to Artificial Neural Networks (ANNs).*

Das, A., & Rad, P. (2020, 6). *Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey* (Tech. Rep.). Retrieved from https://arxiv.org/abs/2006.11371v2 doi: 10.48550/arxiv.2006.11371

de Bruijn, H., Warnier, M., & Janssen, M. (2021, 12). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 101666. doi: 10.1016/J.GIQ.2021.101666

Deist, T. M., Dankers, F. J., Valdes, G., Wijsman, R., Hsu, I. C., Oberije, C., . . . Lambin, P. (2018, 7). Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. *Medical Physics*, *45*(7), 3449–3459. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/mp.12967https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.12967https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.12967 doi: 10.1002/MP.12967

Doran, D., Schulz, S., & Besold, T. R. (2017). *What Does Explainable AI Really Mean? A New Conceptualization of Perspectives .* Retrieved from https://www.researchgate.net/publication/320180555_What_Does_Explainable_AI_Really_Mean_A_New_Conceptualization_of_Perspectives

Dosilovic, F. K., Brcic, M., & Hlupic, N. (2018, 6). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings*, 210–215. doi: 10.23919/MIPRO.2018.8400040

Dutilh Novaes, C., & Reck, E. (2015, 7). Carnapian explication, formalisms as cognitive tools, and the paradox of adequate formalization. *Synthese 2015 194:1*, *194*(1), 195–215. Retrieved from https://link.springer.com/article/10.1007/s11229-015-0816-z doi: 10.1007/S11229-015-0816-Z

Eberly, L. E. (2007). Multiple Linear Regression. *Methods in molecular biology (Clifton, N.J.)*, *404*, 165–187. Retrieved from https://link.springer.com/protocol/10.1007/978-1-59745-530-5_9 doi: 10.1007/978-1-59745-530-5{_}9

Elkamel, M., Schleider, L., Pasiliao, E. L., Diabat, A., & Zheng, Q. P. (2020, 8). Long-Term Electricity Demand Prediction via Socioeconomic Factors—A Machine Learning Approach with Florida as a Case Study. *Energies 2020, Vol. 13, Page 3996*, *13*(15), 3996. Retrieved from https://www.mdpi.com/1996-1073/13/15/3996/htmhttps://www.mdpi.com/1996-1073/13/15/3996 doi: 10.3390/EN13153996

Eurostat. (2022, 5). *Statistics — Eurostat.* Retrieved from https://ec.europa.eu/eurostat/databrowser/view/TEC00113/default/table

Ghalehkhondabi, I., Ardjmand, E., Weckman, G. R., & Young, W. A. (2017, 5). An overview of energy demand forecasting methods published in 2005–2015. *Energy Systems*, *8*(2), 411–447. Retrieved from https://link.springer.com/article/10.1007/s12667-016-0203-y doi: 10.1007/S12667-016-0203-Y/TABLES/8

Ghani, I. M. M., & Ahmad, S. (2010, 1). Stepwise Multiple Regression Method to Forecast Fish Landing. *Procedia - Social and Behavioral Sciences*, *8*, 549–554. doi: 10.1016/J.SBSPRO.2010.12.076

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019, 1). Explaining explanations: An overview of interpretability of machine learning. *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, 80–89.

doi: 10.1109/DSAA.2018.00018

Hevner, A. R. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, *19*(2). Retrieved from https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1017&context=sjis

Hu, L., Chen, J., Vaughan, J., Aramideh, S., Yang, H., Wang, K., … Nair, V. N. (2021, 12). Supervised Machine Learning Techniques: An Overview with Applications to Banking. *International Statistical Review*, *89*(3), 573–604. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/insr.12448https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12448https://onlinelibrary.wiley.com/doi/10.1111/insr.12448 doi: 10.1111/INSR.12448

Kotsiantis, S. B. (2011, 6). Decision trees: a recent overview. *Artificial Intelligence Review 2011 39:4*, *39*(4), 261–283. Retrieved from https://link.springer.com/article/10.1007/s10462-011-9272-4 doi: 10.1007/S10462-011-9272-4

Krishnan, M. (2019, 8). Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy & Technology 2019 33:3*, *33*(3), 487–502. Retrieved from https://link.springer.com/article/10.1007/s13347-019-00372-9 doi: 10.1007/S13347-019-00372-9

Lahiri, S. K., & Ghanta, K. C. (2009). Artificial neural network model with the parameter tuning assisted by a differential evolution technique: The study of the hold up of the slurry flow in a pipeline. *Chemical Industry and Chemical Engineering Quarterly*, *15*(2), 103–117. doi: 10.2298/CICEQ0902103L

Laref, R., Losson, E., Sava, A., & Siadat, M. (2019, 1). On the optimization of the support vector machine regression hyperparameters setting for gas sensors array applications. *Chemometrics and Intelligent Laboratory Systems*, *184*, 22–27. doi: 10.1016/J.CHEMOLAB.2018.11.011

Lee, E. (2021). *How do we build trust in machine learning models?* Retrieved from https://ssrn.com/abstract=3822437 doi: 10.2139/ssrn.3822437

Luellen, E. (2021, 7). *Mastering XGBoost. Hyper-parameter Tuning & Optimization — by Eric Luellen — Towards Data Science.* Retrieved from https://towardsdatascience.com/mastering-xgboost-2eb6bce6bc76

Mantovani, R. G., Horváth, T., Cerri, R., Junior, S. B., Vanschoren, J., & de Carvalho, A. C. P. d. L. F. (2018, 12). *An empirical study on hyperparameter tuning of decision trees* (Tech. Rep.). Retrieved from https://arxiv.org/abs/1812.02207v2 doi: 10.48550/arxiv.1812.02207

Marcelino, P., de Lurdes Antunes, M., Fortunato, E., & Gomes, M. C. (2019). Machine learning approach for pavement performance prediction. *https://doi.org/10.1080/10298436.2019.1609673*, *22*(3), 341–354. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/10298436.2019.1609673 doi: 10.1080/10298436.2019.1609673

Miller, T. (2019, 2). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. doi: 10.1016/J.ARTINT.2018.07.007

Mulder, N. (2022, 2). *Interview.*

Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016, 1). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental Monitoring and Assessment*, *188*(1), 1–27. Retrieved from https://link.springer.com/article/10.1007/s10661-015-5049-6 doi: 10.1007/S10661-015-5049-6/FIGURES/19

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., & Akinjobi, J. (2017, 6). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, *48*(3). Retrieved from https://www.researchgate.net/profile/J-E-T-Akinsola/publication/318338750_Supervised_Machine_Learning_Algorithms_Classification_and_Comparison/links/596481dd0f7e9b819497e265/Supervised-Machine-Learning-Algorithms-Classification-and-Comparison.pdf doi: 10.14445/22312803/IJCTT-V48P126

Ozoegwu, C. G. (2019, 4). Artificial neural network forecast of monthly mean daily global solar radiation of selected locations based on time series and month number. *Journal of Cleaner Production*, *216*, 1–13. doi: 10.1016/J.JCLEPRO.2019.01.096

Parreco, J., Soe-Lin, H., Parks, J. J., Byerly, S., Chatoor, M., Buicko, J. L., … Rattan, R. (2019). Comparing machine learning algorithms for predicting acute kidney injury. *American Surgeon*, *85*(7), 725–729. doi: 10.1177/000313481908500731

Patra, T. K., Meenakshisundaram, V., Hung, J. H., & Simmons, D. S. (2017, 2). Neural-Network-Biased Genetic Algorithms for Materials Design: Evolutionary Algorithms That Learn. *ACS Combinatorial Science*, *19*(2), 96–107. Retrieved from `https://pubs.acs.org/doi/full/10.1021/acscombsci.6b00136` doi: 10.1021/ACSCOMBSCI.6B00136/SUPPL{_}FILE/CO6B00136{_}SI{_}001.PDF

Pinho, A., Costa, R., Silva, H., & Furtado, P. (2018, 9). Comparing Time Series Prediction Approaches for Telecom Analysis. In *Theory and applications of time series analysis* (pp. 331–345). Springer, Cham. Retrieved from `https://link.springer.com/chapter/10.1007/978-3-030-26036-1_23` doi: 10.1007/978-3-030-26036-1{_}23

Plasterk, R. (2014, 1). *Staatsblad van het Koninkrijk der Nederlanden.* Retrieved from `https://zoek.officielebekendmakingen.nl/stb-2014-10.html#`

Portugal, I., Alencar, P., & Cowan, D. (2018, 5). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, *97*, 205–227. doi: 10.1016/J.ESWA.2017.12.020

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019, 5). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1301. Retrieved from `https://onlinelibrary.wiley.com/doi/full/10.1002/widm.1301` doi: 10.1002/WIDM.1301

Rijksoverheid. (2022, 2). *Wat is de identificatieplicht?* Rijksoverheid. Retrieved from `https://www.rijksoverheid.nl/onderwerpen/paspoort-en-identiteitskaart/vraag-en-antwoord/wat-is-de-identificatieplicht#:~:text=De%20identificatieplicht0houdt%20in%20dat,vragen%20om%20een%20geldig%20identiteitsbewijs.`

Sadek, R. M., Mohammed, S. A., Rahman, A., Abunbehan, K., Karim, A., Ghattas, H. A., … Abu-Naser, S. S. (2019). Parkinson's Disease Prediction Using Artificial Neural Network. *International Journal of Academic Health and Medical Research*, *3*(1), 1–8. Retrieved from `www.ijeais.org/ijahmr`

Sapankevych, N., & Sankar, R. (2009, 5). Time series prediction using support vector machines: A survey. *IEEE Computational Intelligence Magazine*, *4*(2), 24–38. Retrieved from `https://ieeexplore.ieee.org/abstract/document/4840324` doi: 10.1109/MCI.2009.932254

Sarewitz, D., & Pielke, R. (1999, 4). Prediction in science and policy. *Technology in Society*, *21*(2), 121–133. doi: 10.1016/S0160-791X(99)00002-0

Schmidt, P., & Biessmann, F. (2019, 1). *Quantifying Interpretability and Trust in Machine Learning Systems.* Retrieved from `https://arxiv.org/abs/1901.08558v1` doi: 10.48550/arxiv.1901.08558

Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2017, 11). Stock price prediction using LSTM, RNN and CNN-sliding window model. *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, *2017-January*, 1643–1647. doi: 10.1109/ICACCI.2017.8126078

Sethi, A. (2020, 3). *Support Vector Regression In Machine Learning.* Retrieved from `https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/`

Siau, K., & Wang, W. (2018). Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter Business Technology jJournal*, *31*(2). Retrieved from `https://www.researchgate.net/profile/Keng-Siau-2/publication/324006061_Building_Trust_in_Artificial_Intelligence_Machine_Learning_and_Robotics/links/5ab8744baca2722b97cf9d33/Building-Trust-in-Artificial-Intelligence-Machine-Learning-and-Robotics.pdf`

Singh, H. (2018, 11). *Understanding Gradient Boosting Machines — by Harshdeep Singh — Towards Data Science.* Retrieved from `https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab`

SKlearn. (2022a). *1.10. Decision Trees — scikit-learn 1.1.1 documentation.* Retrieved from `https://scikit-learn.org/stable/modules/tree.html#tree-algorithms-id3-c4-5`

-c5-0-and-cart

SKlearn. (2022b). *1.11. Ensemble methods — scikit-learn 1.1.1 documentation.* Retrieved from https://scikit-learn.org/stable/modules/ensemble.html#forest

SKlearn. (2022c). *sklearn.model_selection.TimeSeriesSplit — scikit-learn 1.1.1 documentation.* Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html

SKlearn. (2022d). *sklearn.neural_network.MLPRegressor — scikit-learn 1.1.1 documentation.* Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html#sklearn.neural_network.MLPRegressor

SKlearn. (2022e). *sklearn.svm.SVR — scikit-learn 1.1.1 documentation.* Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

Smola, A. J., & Schölkopf, B. (2004, 8). A tutorial on support vector regression. *Statistics and Computing 2004 14:3*, *14*(3), 199–222. Retrieved from https://link.springer.com/article/10.1023/B:STCO.0000035301.49549.88 doi: 10.1023/B:STCO.0000035301.49549.88

Sovrano, F., Sapienza, S., Palmirani, M., & Vitali, F. (2022, 2). Metrics, Explainability and the European AI Act Proposal. *J Multidisciplinary Scientific Journal*, *5*(1), 126–138. Retrieved from https://www.mdpi.com/2571-8800/5/1/10/htm doi: 10.3390/J5010010

Sruthi, E. (2021, 6). *Random Forest — Introduction to Random Forest Algorithm.* Retrieved from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Stetler, C. B., Legro, M. W., Wallace, C. M., Bowman, C., Guihan, M., Hagedorn, H., ... Smith, J. L. (2006, 2). The Role of Formative Evaluation in Implementation Research and the QUERI Experience. *Journal of General Internal Medicine*, *21*(Suppl 2), 8. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2557128/ doi: 10.1111/J.1525-1497.2006.00355.X

Tsekeris, T., & Tsekeris, C. (2015). Demand Forecasting in Transport: Overview and Modeling Advances. *http://www.tandfonline.com/action/authorSubmission?journalCode=rero20&page=instructions*, *24*(1), 82–94. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/1331677X.2011.11517446 doi: 10.1080/1331677X.2011.11517446

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019, 12). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 1–16. Retrieved from https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8 doi: 10.1186/S12911-019-1004-8/FIGURES/12

Vogelsang, A., & Borg, M. (2019, 9). Requirements engineering for machine learning: Perspectives from data scientists. *Proceedings - 2019 IEEE 27th International Requirements Engineering Conference Workshops, REW 2019*, 245–251. doi: 10.1109/REW.2019.00050

Yildrim, S. (2020, 4). *Hyperparameter Tuning for Support Vector Machines — C and Gamma Parameters — by Soner Yıldırım — Towards Data Science.* Retrieved from https://towardsdatascience.com/hyperparameter-tuning-for-support-vector-machines-c-and-gamma-parameters-6a5097416167

Yu, T., & Zhu, H. (2020, 3). *Hyper-Parameter Optimization: A Review of Algorithms and Applications* (Tech. Rep.). Retrieved from https://arxiv.org/abs/2003.05689v1 doi: 10.48550/arxiv.2003.05689

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021, 3). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics 2021, Vol. 10, Page 593*, *10*(5), 593. Retrieved from https://www.mdpi.com/2079-9292/10/5/593/htmhttps://www.mdpi.com/2079-9292/10/5/593 doi: 10.3390/ELECTRONICS10050593

A Survey results

In this appendix the responses to the survey questions are discussed. The survey consisted of three parts that tackled a different subject each.

The first group of questions consisted of introductory questions such as what do your daily tasks look like and how are you connected to the prediction model? The second group of questions is meant for the people actually using the model as part of their job. Here questions were asked with regards to the respondents knowledge about the underlying mechanisms of the current prediction model questions, how easy the current model is to work with, what parts could be better and what kind of extra information is seen as useful. This set of question can be skipped by the respondents who don't actively work with the model itself, but are connected to it through other ways. The last group of questions regard the requirements a new prediction model should fulfil. Questions such as whether or not needing to be able to program in Python are asked, what are other requirements apart from a better prediction, how transparent/explainable should the model be and what requirements should there be with regards to the data that is used. An overview of the interview questions is shown on the following page. Afterwards the answers per respondent are briefly discussed.

Introductory questions:

- What is your function within the RvIG?
  - What are your general daily tasks and how do these relate to the prediction model for identification documents?
- How are you involved with the current and/or new prediction model?
  - Do you use the current prediction model?
  - If so, how often and with which purpose?
  - If you do not use the model, but are involved with the model. What aspects are you involved with?

The following questions are relevant if you use the current prediction model. If you do not use the current prediction you can ignore these questions.

- Do you understand the underlying logic of the current prediction model?
  - Is understanding the underlying logic important?
- Is the current prediction model easy to operate?
  - Why?
- Can the current model be easily adjusted?
- Are there other model elements, apart from a better prediction, that can be improved upon?
- The current model does not predict the future demand perfectly. How is the model currently used to make predictions?

The following questions regard the desired requirements for a new prototype prediction model. You are asked to answer these questions even if you do not directly use the current or future prediction model. You might still indirectly use the results of the model and are helped with a better prediction model.

- How familiar are you with the Python programming language?
  - Would needing to be able to program in Python be a reason not to use the new prototype prediction model?
  - If so, would a simple interface for which no programming knowledge is required suffice in order to use the model?
- Apart from a better prediction, what other requirement should the model adhere to in order to be used in practise?
  - Are insights into the feature importance for example also important?
  - Are there specific graphs that are desired?
  - Does the model need to be easily adjustable?
- What are the requirements with regards to model transparency?
  - Is it important that the model design process is transparent?
  - Is it important that the used algorithm is explainable?
- What requirements are there with regards to the data that will be used?

If you have any other comments or suggestions please let me know.

A.1 Survey response Chris Mostert

Data analist for the department of Research and Analysis. Works with (data) analysis questions, developing existing products and researching the possibility of new data driven techniques such as machine learning. Chris has not worked on the existing prediction model, but plays a supporting role in the development of the prototype model. He does not use the prediction model himself.

According to Chris using the prototype model should be as simple as possible to ensure that it is used by the end users and to reduce the possibility for mistakes. If the end users do not have a programming background he thinks a simple model interface with proper manual on how to use the model is needed.

Further requirements are reproduceability to a certain extent in order to create trust in the results, being able to verify the results and proper documentation.

Insights into the relationship of input data and their predictive power with regards to the future demand of identification documents is seen as an important requirement to be able to "sell" the new model to the end-users. Transparency is deemed important and should be maximised wherever possible. However, the trade-off between model performance and transparency/explainability is acknowledged thus transparency should be present where possible. For example about model design choices, data choices, assumptions that are made, internal mechanisms from the models etc... In the case of more complex models the need for transparency and explanations of the model design process are even more important. To round up, the data used should be "right" since wrong data leads to the wrong results and data should remain on an aggregated level as to not lead to any potential privacy issues.

A.2 Survey response Nita Mulder

Senior data annalist at the department of research and analysis. Among others supervisor, contact person and adviser for the development of the prototype prediction model. Personally says she would use the prototype model even if it requires some programming knowledge. Says that understanding the relationship between input data and the prediction results is important. Would like to see a graph that shows the trend of the prediction because it would increase insightfullness. Thinks being open about the model design process is important and also that the used machine learning algorithms should be insightful. Lastly data used should not contain data that can be used to trace back to a single person or other potentially privacy intruding data.

A.3 Survey response Joshua Bannor

The following discuses the response from Joshua Bannor who was supported by Django ter Bals. Joshua is financial advisor at the department of business operations who, among others, works on travel documents. Joshua is responsible for calculating the budget for travel documents for which he uses the current prediction model to determine the revenue and costs. Apart from the latter he uses the model to calculate new costs when working with new innovative additions.

Understanding the logic behind the prediction model is important in order to be able to explain to stakeholders which conditions were included while creating the budget. Furthermore he states that everyone is aware that the model is a prediction and therefore not fully reliable. To ensure the predictions are kept in line predictions made for t+3 at time step t are/can be evaluated at t+1 if need be. Apart from the insights in demand for identification documents, double possession of documents, population growth, possession of drivers licenses and age is also seen as important.

Joshua states that he is not familiar with the Python programming language. Thus to ensure the model is used a simple interface/python script is likely useful. Additionally he states that insights into the relationship between feature and dependent data are important in order to be able to explain which assumptions have been made. Transparency in the design process is also important in case questions arise from the house of representatives (second chamber) or from WOB requests. Furthermore he states that explainability is important and that the data used should keep in mind privacy and should remain up to date. Lastly he states that a benchmark function to show the difference between the actual demand and the predicted demand would be useful.

B Prediction results

In this appendix the results not shown in the main text are briefly discussed as well as the hyper parameters that were used for the predictions. The results are discussed per algorithm following the same order as the one used in the main text

B.1 Multiple linear regression results

Figure 34 and 35 show the predicted demand for five and ten year valid passports respectively. Table 26 recaps the models results and the variance present in the prediction. Both predictions were made using historic data and therefore only show the theoretical performance of the model. As the figures clearly show the multiple linear regression is unable to accurately predict the demand.

Table 26: Multiple linear regression prediction results

model	dependent_variable	dataset	max_average_r2	variance in R2
LinearRegression	uitgifte_aantal_ni_5y	lagged_feature_ data_2016_2022_t3.csv	-1026.967063	0.0
LinearRegression	uitgifte_aantal_pn_5y	lagged_feature_ data_2016_2022_t3.csv	-2014.046870	0.0
LinearRegression	uitgifte_aantal_ni_10y	lagged_feature_ data_2016_2022_t3.csv	-4409.456620	0.0
LinearRegression	uitgifte_aantal_pn_10y	lagged_feature_ data_2016_2022_t3.csv	-10188.174676	0.0

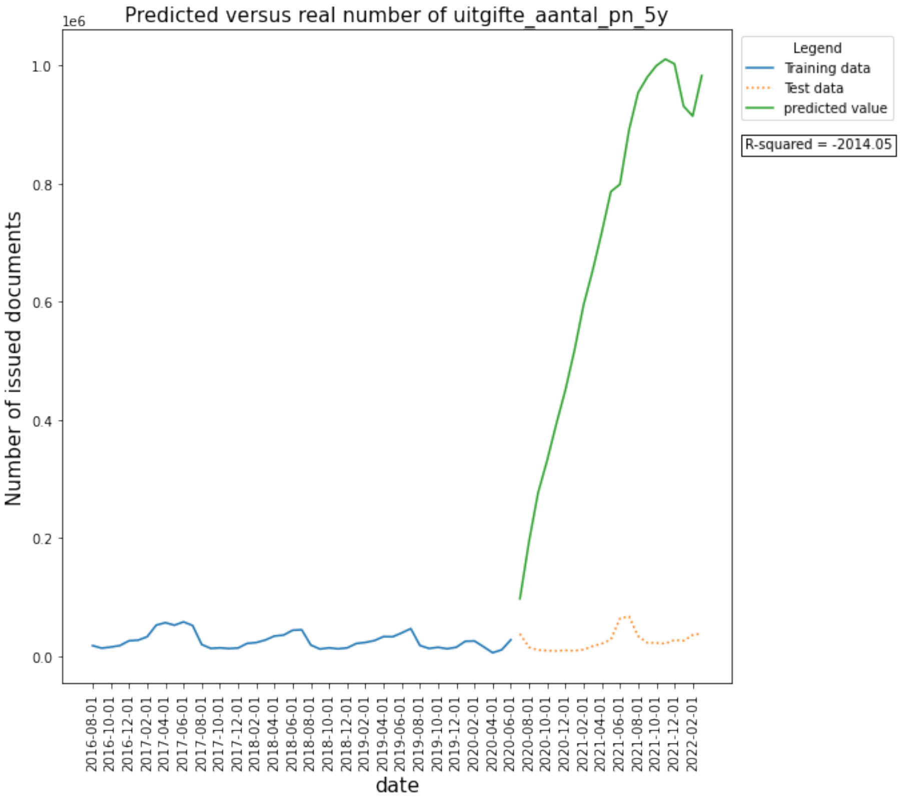


Figure 34: Predicted number of 5 year valid passports using MLR

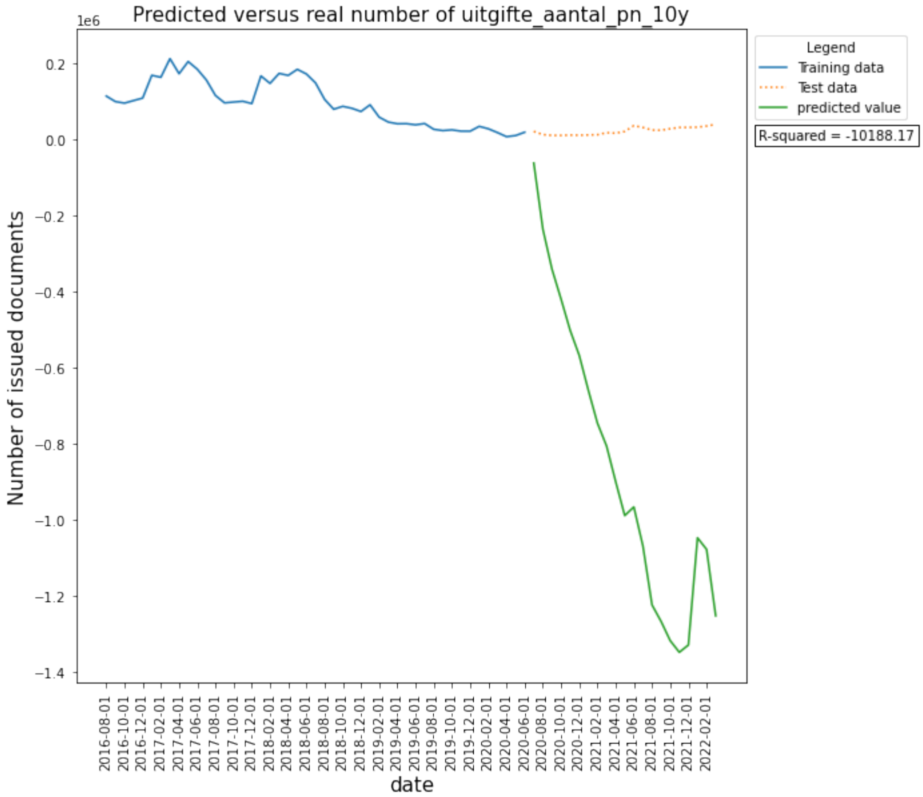


Figure 35: Predicted number of 10 year valid passports using MLR

B.2 Decision tree regression results

Figure 36 and 37 show the predicted demand for five and ten year valid passports respectively. Table 27 recaps the models results and the variance present in the prediction. Table 28 shows the optimal hyper parameters used for the prediction. Both predictions were made using historic data and therefore only show the theoretical performance of the model. When compared to the multiple linear regression results mentioned previously, the results look better. However, the negative R-squared values present indicate that the decision tree model is also unfit for the future demand prediction of passports.

Table 27: Decision Tree Regression prediction results

model	dependent_variable	dataset	max_average_r2	Variance in r2
DecisionTreeRegressor	uitgifte_aantal_ni_5y	lagged_feature_data_2016_2022_t-1.csv	-0.102824	0.267882
DecisionTreeRegressor	uitgifte_aantal_pn_5y	feature_data_2016_2022.csv	-0.020852	0.112970
DecisionTreeRegressor	uitgifte_aantal_ni_10y	feature_data_march_2014_dec_2018.csv	-1.991366	1.302158
DecisionTreeRegressor	uitgifte_aantal_pn_10y	standardized_feature_data_2014_2018.csv	-0.452015	0.749400

Table 28: Optimal hyper parameters used for Decision Tree Regression prediction

dependent variable	max_depth	max_features	max_leaf_nodes	min_samp_leaf	min_samp_split
uitgifte_aantal_ni_5y	3.0	sqrt	30.0	4.0	4.0
uitgifte_aantal_pn_5y	50.0	log2	10.0	8.0	14.0
uitgifte_aantal_ni_10y	5.0	log2	30.0	10.0	3.0
uitgifte_aantal_pn_10y	1.0	sqrt	5.0	6.0	4.0

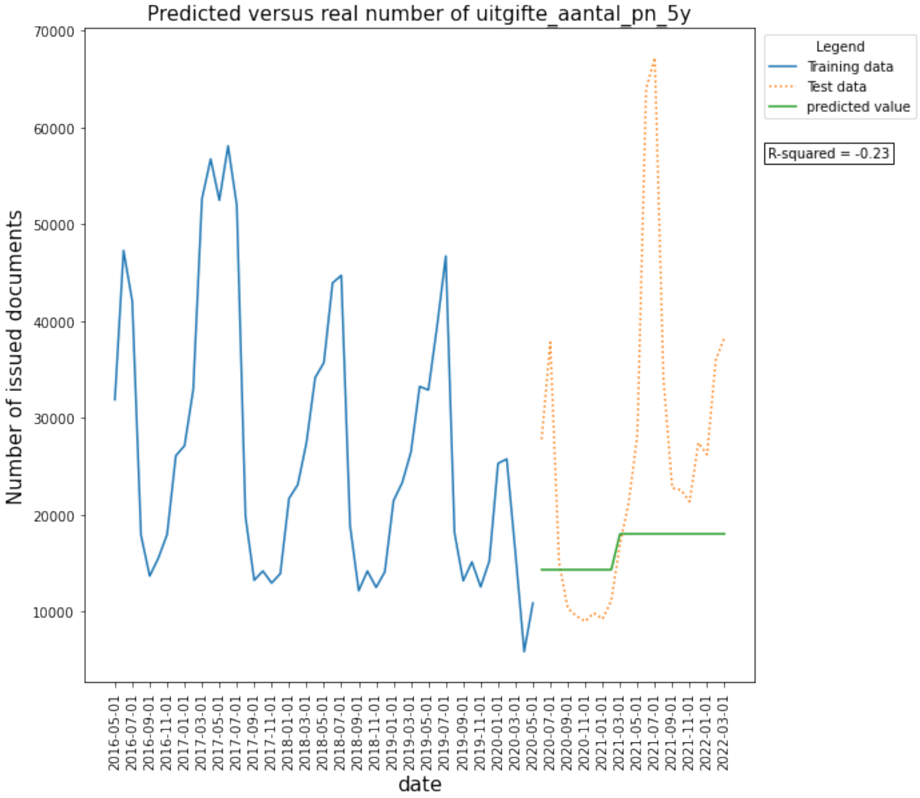


Figure 36: Predicted number of 5 year valid passports using DT

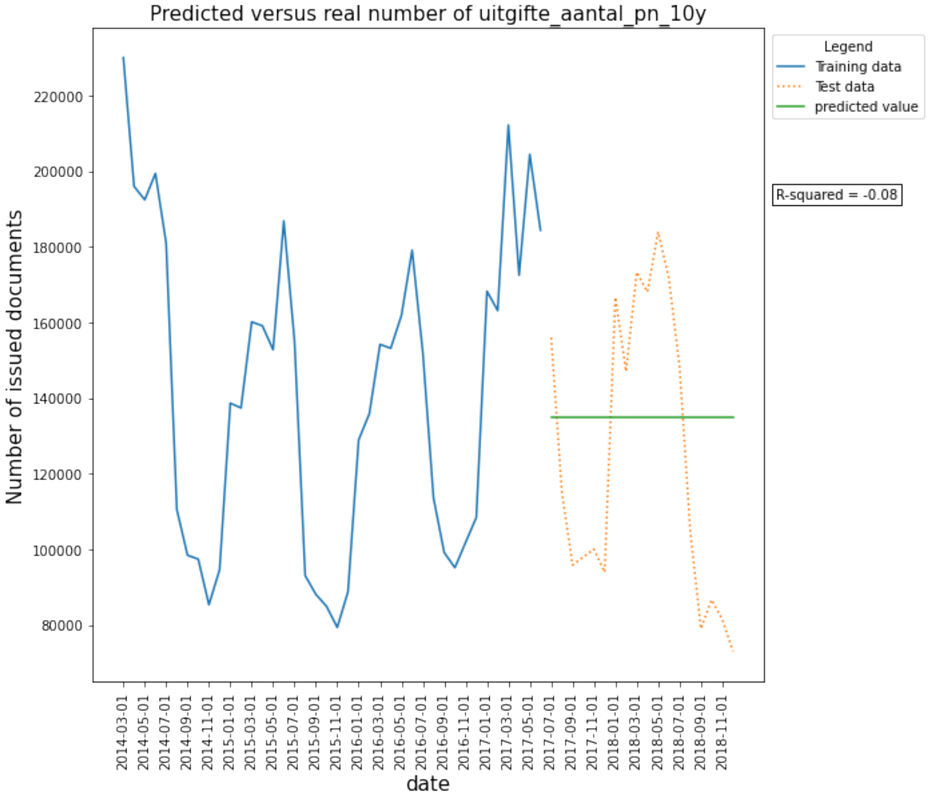


Figure 37: Predicted number of 10 year valid passports using DT

B.3 Random forest regression results

Figure 38 and 39 show the predicted demand for five and ten year valid passports respectively. Table 29 recaps the models results and the variance present in the prediction. Table 30 shows the optimal hyper parameters used for the prediction. Both predictions were made using historic data and therefore only show the theoretical performance of the model. The random forest results are the first results with a positive R-squared value. The increased prediction accuracy can be seen in both figures although it also clear that the predictions are not fully accurate yet. The predictions for the five year valid passports is unable to capture the big peak at the end which means a lot of demand is missed.

Table 29: Random Forest Regression prediction results

model	dependent_variable	dataset	max_average_r2	Variance in r2
RandomForest Regressor	uitgifte_aantal_ni_5y	lagged_feature_data _2016_2022_t3.csv	0.227153	0.013558
RandomForest Regressor	uitgifte_aantal_pn_5y	lagged_feature_data _2016_2022_t2.csv	0.098252	0.062277
RandomForest Regressor	uitgifte_aantal_ni_10y	lagged_feature_data _2016_2022_t-1.csv	-0.749718	0.892041
RandomForest Regressor	uitgifte_aantal_pn_10y	lagged_feature_data _2016_2022_t2.csv	0.259962	0.118852

Table 30: Optimal hyper parameters used for Random Forest Regression prediction

dependent variable	max_depth	max_features	min_samp_leaf	min_samp_split	n_estimators
uitgifte_aantal_ni_5y	10.0	auto	1.0	5.0	20.0
uitgifte_aantal_pn_5y	30.0	auto	1.0	5.0	5.0
uitgifte_aantal_ni_10y	20.0	log2	1.0	2.0	5.0
uitgifte_aantal_pn_10y	5.0	auto	1.0	2.0	10.0

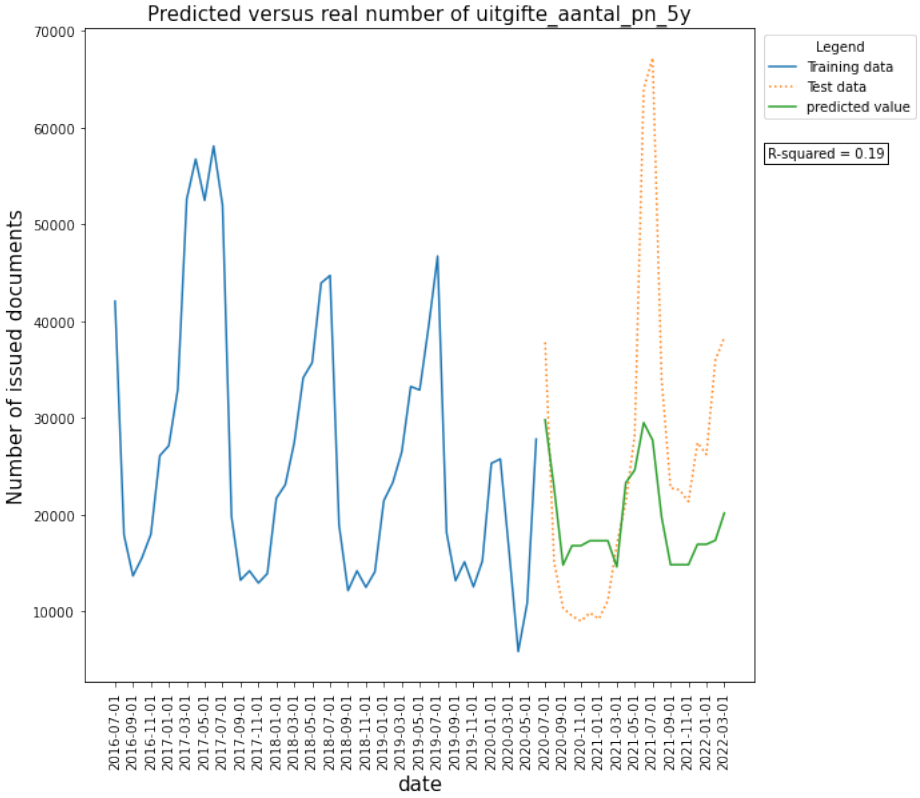


Figure 38: Predicted number of 5 year valid passports using RF

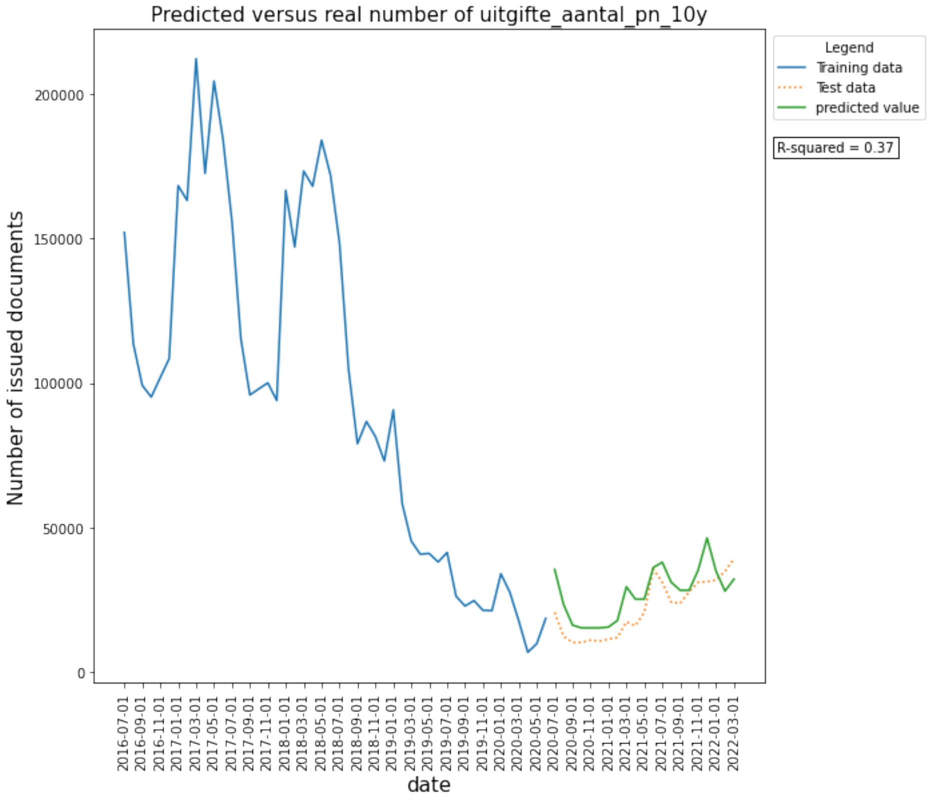


Figure 39: Predicted number of 10 year valid passports using RF

B.4 Extreme gradient boost regression results

Figure 40 and 41 show the predicted demand for five and ten year valid passports respectively. Table 31 recaps the models results and the variance present in the prediction. Table 32 shows the optimal hyper parameters used for the prediction. Both predictions were made using historic data and therefore only show the theoretical performance of the model. The results for the xgb algorithm show increased performance compared to the rf results. The prediction for five year valid passports still does not capture the big peak at the end, but it is able to capture most of the remaining demand. The ten year valid prediction shows an R-squared value of 0.74 and is able to capture most of the future demand.

Table 31: Extreme Gradient Boosting prediction results

model	dependent_variable	dataset	max_average_r2	Variance in r2
XGBRegressor	uitgifte_aantal_ni_5y	lagged_feature_data_2016_2022_t2.csv	0.477636	0.0
XGBRegressor	uitgifte_aantal_pn_5y	lagged_feature_data_2016_2022_t3.csv	0.494608	0.0
XGBRegressor	uitgifte_aantal_ni_10y	lagged_feature_data_2016_2022_t-1.csv	-0.614031	0.0
XGBRegressor	uitgifte_aantal_pn_10y	lagged_feature_data_2016_2022_t-1.csv	0.739382	0.0

Table 32: Optimal hyper parameters used for XGB Regression prediction

dependent variable	gamma	max_depth	n_estimators	reg_alpha	reg_lambda
uitgifte_aantal_ni_5y	0.0	2.0	100.0	30.0	0.0001
uitgifte_aantal_pn_5y	0.0	2.0	20.0	100.0	0.1000
uitgifte_aantal_ni_10y	0.0	5.0	5.0	0.0	0.1000
uitgifte_aantal_pn_10y	0.0	10.0	10.0	0.0	0.0100

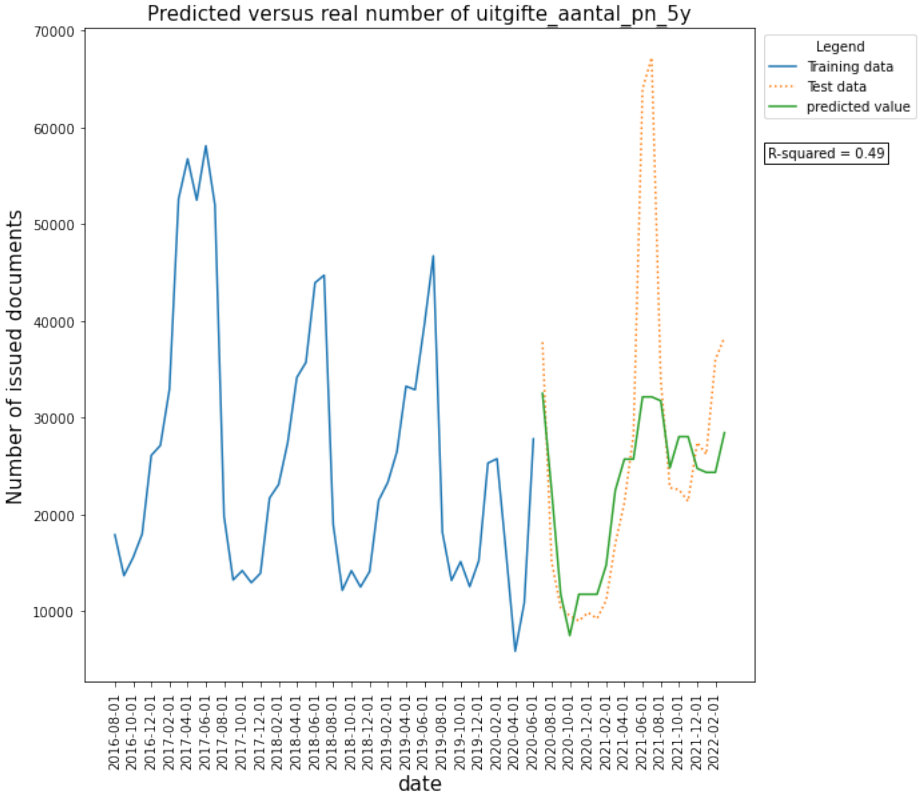


Figure 40: Predicted number of 5 year valid passports using XGV

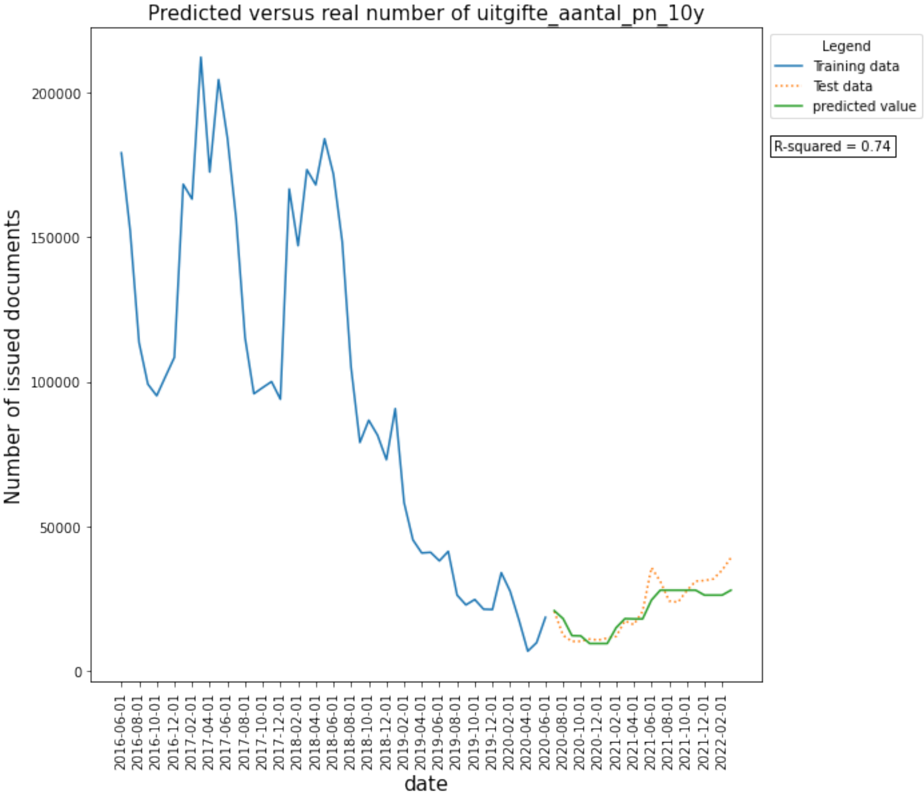


Figure 41: Predicted number of 10 year valid passports using XGB

B.5 Support vector regression results

Figure 42 and 43 show the predicted demand for five and ten year valid passports respectively. Table 33 recaps the models results and the variance present in the prediction.Table 34 recaps the models results and the variance present in the prediction. Both predictions were made using historic data and therefore only show the theoretical performance of the model. The SVR results are similar to the results reached using the XGB algorithm. However, the five year valid demand prediction for passports is able to slightly better predict the peak in demand near the end.

Table 33: Support Vector Regression prediction results

model	dependent_variable	dataset	max_average_r2	Variance in r2
SVR	uitgifte_aantal_ni_5y	MaxAbs_feature_data_2016_2022.csv	0.790205	0.0
SVR	uitgifte_aantal_pn_5y	MaxAbs_feature_data_2016_2022.csv	0.510128	0.0
SVR	uitgifte_aantal_ni_10y	lagged_feature_data_2016_2022_t3.csv	0.740977	0.0
SVR	uitgifte_aantal_pn_10y	feature_data_2016_2022.csv	0.673658	0.0

Table 34: Optimal hyper parameters used for SVR Regression prediction

dependent variable	C	degree	epsilon	gamma	kernel
uitgifte_aantal_ni_5y	10.0	7.0	10000.0	scale	poly
uitgifte_aantal_pn_5y	1.0	5.0	10000.0	scale	poly
uitgifte_aantal_ni_10y	10000.0	5.0	5000.0	scale	poly
uitgifte_aantal_pn_10y	10000.0	5.0	7500.0	scale	poly

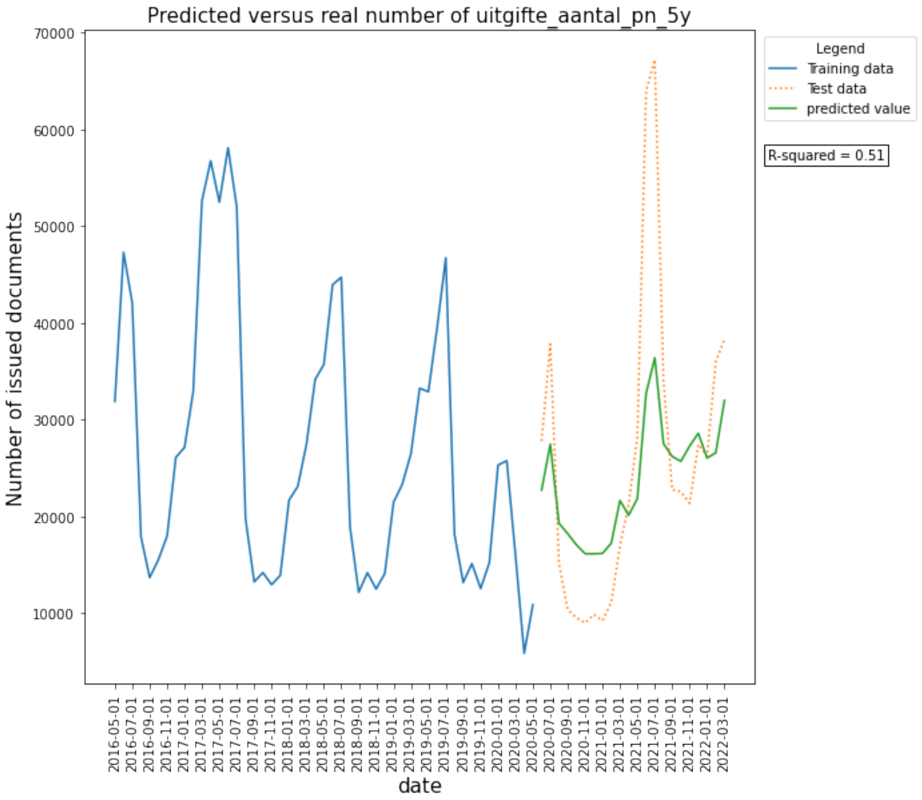


Figure 42: Predicted number of 5 year valid passports using SVR

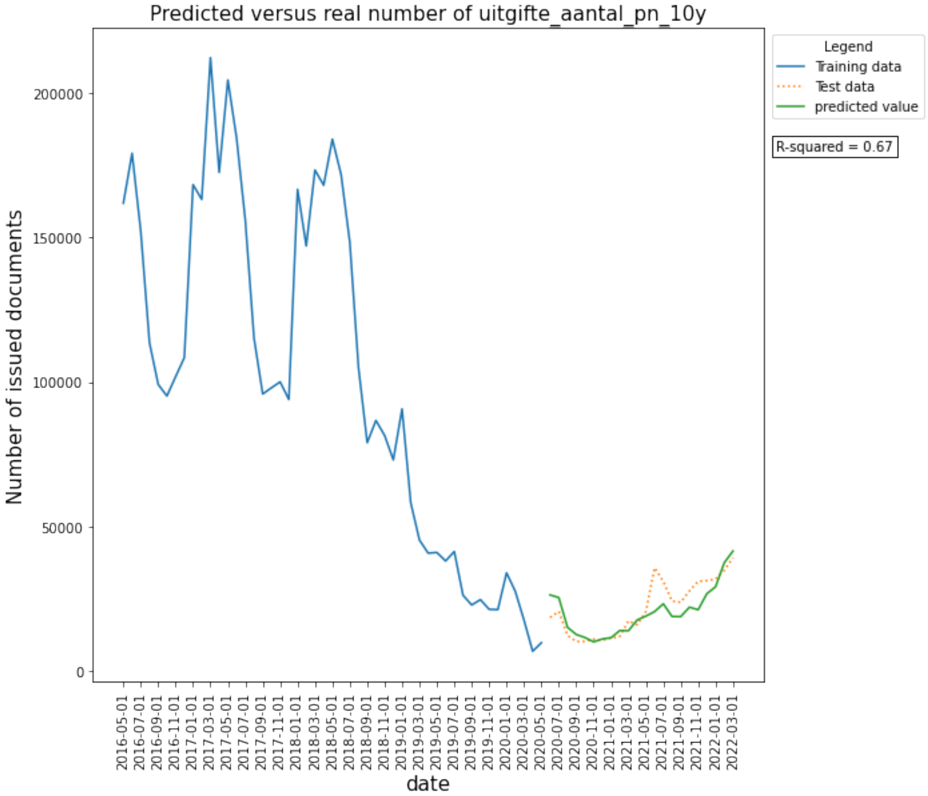


Figure 43: Predicted number of 10 year valid passports using SVR

B.6 Artificial neural network regression results

Figure 42 and 43 show the predicted demand for five and ten year valid passports respectively. Table 35 recaps the models results and the variance present in the prediction. Both predictions were made using historic data and therefore only show the theoretical performance of the model. The ANN results show a huge polarity in performance. The five year valid passports prediction almost perfectly follows the real demand with a R-squared value of 0.88 while the ten year valid prediction reaches a R-squared value of -0.82. However, both results are unreliable due to the variance present in the results.

Table 35: Multilayer Perceptron Regression prediction results

model	dependent_variable	dataset	max_average_r2	Variance in r2
MLPRegressor	uitgifte_aantal_ni_5y	standardized_feature _data_2016_2022.csv	-0.234234	2.938335
MLPRegressor	uitgifte_aantal_pn_5y	lagged_feature_data _2016_2022_t2.csv	0.408624	0.139951
MLPRegressor	uitgifte_aantal_ni_10y	MaxAbs_feature_data _2016_2022.csv	-0.109329	0.556050
MLPRegressor	uitgifte_aantal_pn_10y	MaxAbs_feature_data _2017_2019.csv	-0.905083	4.163745

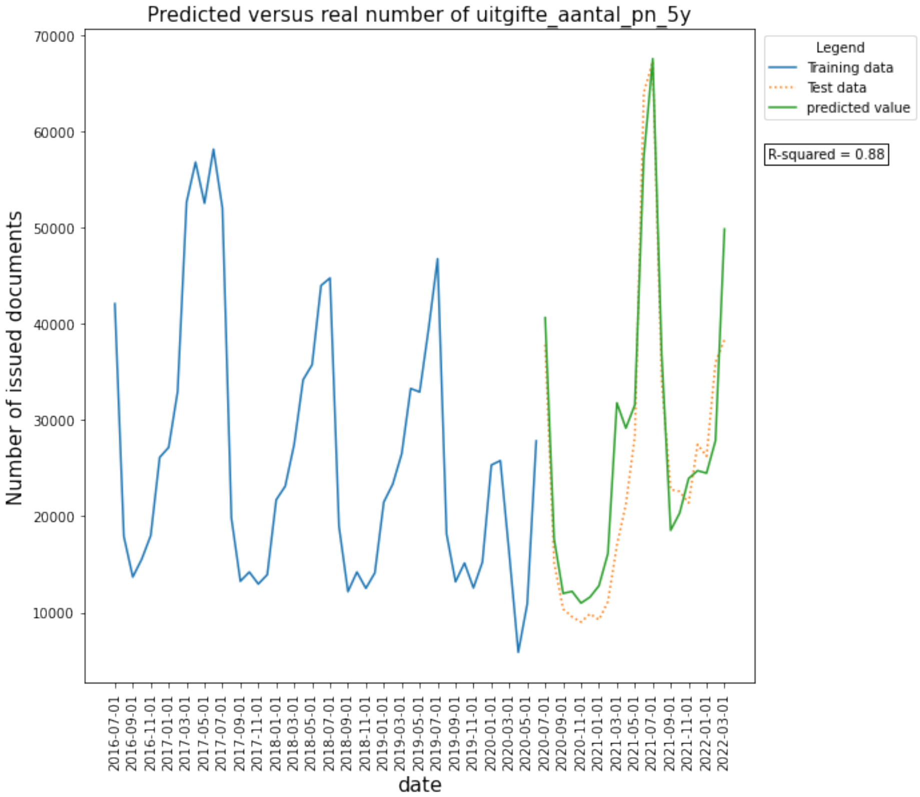


Figure 44: Predicted number of 5 year valid passports using ANN

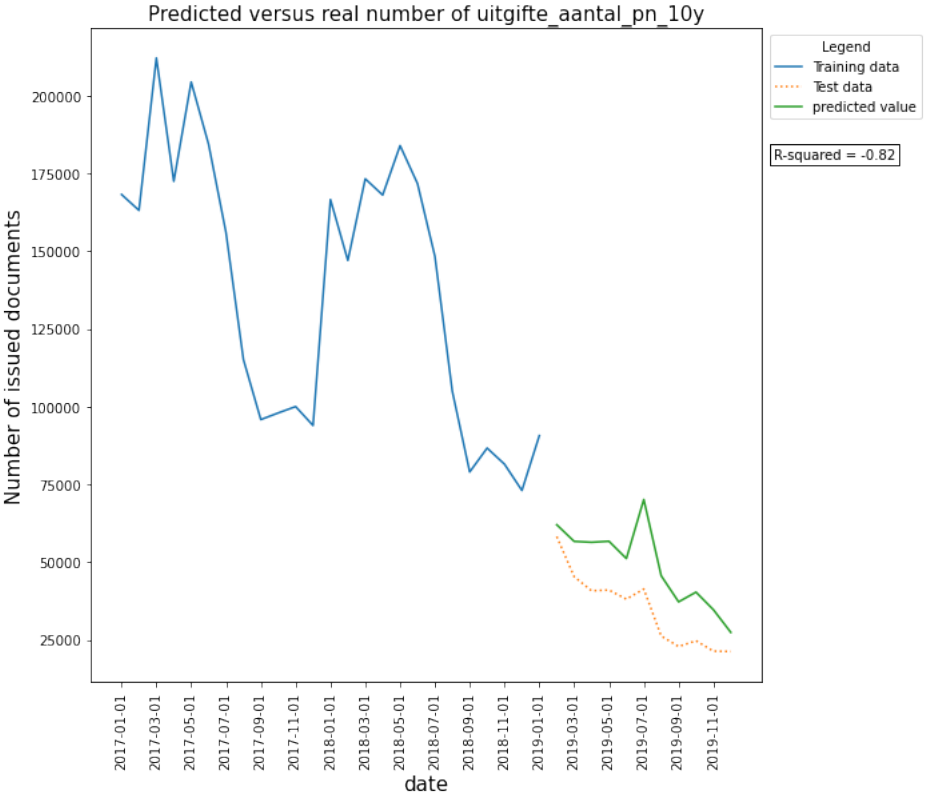


Figure 45: Predicted number of 10 year valid passports using ANN

B.7 Direct forecasting results

In this subsection the direct forecasting results for each of the different machine learning algorithms are discussed. As can be seen by the average R-squared results none of the algorithms is able to reach acceptable prediction results. The reader is referred to the results section for the full reflection on the results. Table 36 till table 41 show the direct forecasting results for MLR, DTR, RF, XGB, SVR and MLP respecively. The predictions are made based on the 2016-2022 dataset with varying number of lagged datasteps. The number of lagged steps that reached the highest R-squared values are used. The number of optimal lagged steps is shown in the lagged steps column.

Table 36: Direct forecasting results for MLR

Model	Dependent variable	Lagged steps	Max average R^2	Variance in R^2
LinearRegression	uitgifte_aantal_ni_5y_t+1	4	-2.876749	0.0
LinearRegression	uitgifte_aantal_pn_5y_t+1	5	-15.296275	0.0
LinearRegression	uitgifte_aantal_ni_10y_t+1	5	-67.415979	0.0
LinearRegression	uitgifte_aantal_pn_10y_t+1	5	-39.022178	0.0
LinearRegression	uitgifte_aantal_ni_5y_t+2	4	-21.616654	0.0
LinearRegression	uitgifte_aantal_pn_5y_t+2	5	-50.723159	0.0
LinearRegression	uitgifte_aantal_ni_10y_t+2	5	-29.025732	0.0
LinearRegression	uitgifte_aantal_pn_10y_t+2	5	-38.782297	0.0
LinearRegression	uitgifte_aantal_ni_5y_t+3	5	-90.396844	0.0
LinearRegression	uitgifte_aantal_pn_5y_t+3	5	-27.159358	0.0
LinearRegression	uitgifte_aantal_ni_10y_t+3	5	-89.431392	0.0
LinearRegression	uitgifte_aantal_pn_10y_t+3	5	-24.316891	0.0

Table 37: Direct forecasting results for DTR

Model	Dependent variable	Lagged steps	Max average R^2	Variance in R^2
DecisionTreeRegressor	uitgifte_aantal_ni_5y_t+1	3	-0.440544	0.16869
DecisionTreeRegressor	uitgifte_aantal_pn_5y_t+1	4	-0.179663	0.090365
DecisionTreeRegressor	uitgifte_aantal_ni_10y_t+1	2	-1.304954	3.970898
DecisionTreeRegressor	uitgifte_aantal_pn_10y_t+1	2	-1.865384	1.806417
DecisionTreeRegressor	uitgifte_aantal_ni_5y_t+2	2	-0.453153	0.213317
DecisionTreeRegressor	uitgifte_aantal_pn_5y_t+2	5	-0.109999	0.073817
DecisionTreeRegressor	uitgifte_aantal_ni_10y_t+2	2	-1.566379	15.80618
DecisionTreeRegressor	uitgifte_aantal_pn_10y_t+2	4	-1.798653	6.49692
DecisionTreeRegressor	uitgifte_aantal_ni_5y_t+3	5	-0.442126	0.101103
DecisionTreeRegressor	uitgifte_aantal_pn_5y_t+3	5	-0.109812	0.012558
DecisionTreeRegressor	uitgifte_aantal_ni_10y_t+3	3	-0.461509	0.288102
DecisionTreeRegressor	uitgifte_aantal_pn_10y_t+3	3	-5.785681	295.210296

Table 38: Direct forecasting results for RF

Model	Dependent variable	Lagged steps	Max average Rˆ2	Variance in Rˆ2
RandomForestRegressor	uitgifte_aantal_ni_5y_t+1	3	-0.560048	0.223165
RandomForestRegressor	uitgifte_aantal_pn_5y_t+1	4	-0.109841	0.012806
RandomForestRegressor	uitgifte_aantal_ni_10y_t+1	2	-0.638828	0.703085
RandomForestRegressor	uitgifte_aantal_pn_10y_t+1	2	-1.552104	2.486191
RandomForestRegressor	uitgifte_aantal_ni_5y_t+2	2	-0.639683	0.12415
RandomForestRegressor	uitgifte_aantal_pn_5y_t+2	4	-0.069709	0.016346
RandomForestRegressor	uitgifte_aantal_ni_10y_t+2	2	-1.359107	2.735407
RandomForestRegressor	uitgifte_aantal_pn_10y_t+2	1	-6.193946	136.411948
RandomForestRegressor	uitgifte_aantal_ni_5y_t+3	5	-0.61377	0.358113
RandomForestRegressor	uitgifte_aantal_pn_5y_t+3	2	-0.095271	0.008643
RandomForestRegressor	uitgifte_aantal_ni_10y_t+3	4	-1.379274	7.4778
RandomForestRegressor	uitgifte_aantal_pn_10y_t+3	5	-10.200804	295.356361

Table 39: Direct forecasting results for XGB

Model	Dependent variable	Lagged steps	Max average Rˆ2	Variance in Rˆ2
XGBRegressor	uitgifte_aantal_ni_5y_t+1	3	-0.084731	0.0
XGBRegressor	uitgifte_aantal_pn_5y_t+1	5	-0.262256	0.0
XGBRegressor	uitgifte_aantal_ni_10y_t+1	5	-1.90828	0.0
XGBRegressor	uitgifte_aantal_pn_10y_t+1	1	-1.083394	0.0
XGBRegressor	uitgifte_aantal_ni_5y_t+2	2	-0.090316	0.0
XGBRegressor	uitgifte_aantal_pn_5y_t+2	5	-0.271625	0.0
XGBRegressor	uitgifte_aantal_ni_10y_t+2	5	-0.145749	0.0
XGBRegressor	uitgifte_aantal_pn_10y_t+2	1	-1.148105	0.0
XGBRegressor	uitgifte_aantal_ni_5y_t+3	1	-0.090316	0.0
XGBRegressor	uitgifte_aantal_pn_5y_t+3	5	-0.037391	0.0
XGBRegressor	uitgifte_aantal_ni_10y_t+3	3	0.006668	0.0
XGBRegressor	uitgifte_aantal_pn_10y_t+3	5	0.040809	0.0

Table 40: Direct forecasting results for SVR

Model	Dependent variable	Lagged steps	Max average Rˆ2	Variance in Rˆ2
SVR	uitgifte_aantal_ni_5y_t+1	3	-1.119743	0.0
SVR	uitgifte_aantal_pn_5y_t+1	1	0.065676	0.0
SVR	uitgifte_aantal_ni_10y_t+1	5	-1.306219	0.0
SVR	uitgifte_aantal_pn_10y_t+1	5	-10.604207	0.0
SVR	uitgifte_aantal_ni_5y_t+2	2	-1.210848	0.0
SVR	uitgifte_aantal_pn_5y_t+2	4	-0.045891	0.0
SVR	uitgifte_aantal_ni_10y_t+2	5	-1.513715	0.0
SVR	uitgifte_aantal_pn_10y_t+2	5	-5.62213	0.0
SVR	uitgifte_aantal_ni_5y_t+3	4	-1.13186	0.0
SVR	uitgifte_aantal_pn_5y_t+3	3	-0.060252	0.0
SVR	uitgifte_aantal_ni_10y_t+3	5	-1.086283	0.0
SVR	uitgifte_aantal_pn_10y_t+3	3	-6.37712	0.0

Table 41: Direct forecasting results for MLP

Model	Dependent variable	Lagged steps	Max average R^2	Variance in R^2
MLPRegressor	uitgifte_aantal_ni_5y_t+1	4	-2.471432	3.483755
MLPRegressor	uitgifte_aantal_pn_5y_t+1	4	-1.254545	1.628535
MLPRegressor	uitgifte_aantal_ni_10y_t+1	3	-3.883033	11.042927
MLPRegressor	uitgifte_aantal_pn_10y_t+1	4	-13.526044	400.066575
MLPRegressor	uitgifte_aantal_ni_5y_t+2	5	-2.559482	12.84914
MLPRegressor	uitgifte_aantal_pn_5y_t+2	3	-1.471729	2.272324
MLPRegressor	uitgifte_aantal_ni_10y_t+2	3	-3.504947	18.216416
MLPRegressor	uitgifte_aantal_pn_10y_t+2	3	-15.837415	409.080381
MLPRegressor	uitgifte_aantal_ni_5y_t+3	5	-1.849992	1.686406
MLPRegressor	uitgifte_aantal_pn_5y_t+3	4	-0.930784	0.160632
MLPRegressor	uitgifte_aantal_ni_10y_t+3	5	-4.283729	21.524817
MLPRegressor	uitgifte_aantal_pn_10y_t+3	5	-11.56435	156.127973