
# Predicting Freight Mode Choice with Machine Learning

## A case study of the NEAC model

Author: Saskia Veldkamp
Date: September 15, 2025

TU Delft

Panteia

# Predicting Freight Mode Choice with Machine Learning

## A case study of the NEAC model

by

Saskia Veldkamp

In partial fulfilment of the requirements for the degree of:

**Master of Science**
in Transport, Infrastructure, and Logistics

at the Delft University of Technology,
to be defended publicly on Monday, September 15, 2025 at 11:00 AM.


| | | |
|---|---|---|
| Student number: | 6060846 | |
| Thesis committee: | Dr. Bilge Atasoy, | TU Delft, chair |
| | Dr. Sander van Cranenburgh, | TU Delft, supervisor |
| | Jan Kiel, | Panteia, supervisor |

# Preface

I am grateful for all the support I received during this project as well as during my master's program. I would like to thank my thesis supervisors Bilge Atasoy and Sander van Cranenburgh for their feedback and direction throughout this thesis. Thank you as well to Jan Kiel, my company supervisor, for his guidance and for giving me the opportunity to research this topic. I am thankful to Yuko Kawabata who generously offered me her time and expertise numerous times to answer my questions. I am grateful to my classmates and friends in TIL, the interns and other colleagues at Panteia, and members of the CityAI lab for their help in completing this project. Finally, thank you to my family for their constant support in all my endeavours.

Though difficult at times, I am happy that I was able to research this topic and contribute to the emerging field that is the application of machine learning in transportation analysis and planning. I hope this report proves to be of interest to its readers and of use to Panteia in the ongoing work to enhance the NEAC model.

*Saskia Veldkamp*
*Delft, September 2024*

# Summary

Freight transportation makes up 5% of the Gross Domestic Product (GDP) and 25% of total greenhouse gas emissions in the European Union (EU). In order for the EU to meet its emission reduction goals, inland freight transportation, 77% of which is currently by road, must shift more to inland waterway and rail (Eurostat, 2020) (Eurostat, 2022). Mode choice models are necessary to evaluate how well transport policies can affect this desired change. Although these are typically estimated using a Multinomial Logit (MNL) or other Logit-based model, machine learning (ML) models have gained more popularity recently due to their often higher predictive accuracy.

The identified research gap is that previous studies using machine learning models to predict freight mode choice employed disaggregate data at the individual shipment level. This data includes details about the shipment including its weight, monetary value, and commodity type, as well as information about the shipper's industry, all of which are known important factors in freight mode choice (Xu, et al., 2024). It is therefore unclear whether ML models trained with aggregated data of total goods transported between regions could achieve comparable predictive performance. In the EU, aggregate freight transport data is more easily accessible than shipment-level data; exploring the performance of models trained with aggregate data could thus precede the creation of an EU-wide machine learning-based freight mode choice model.

Secondly, earlier research has focused on evaluating and improving the predictive performance of ML models for freight mode choice and has not considered whether or how these models may be used for policy analysis. Further investigation is needed into the criteria such models must meet to be used in practice. This would support the development of a model whose results could be used by analysts and policymakers in transport policy evaluations and other real-world applications.

The main research question was:

*What role can machine learning-based approaches have in freight mode choice modeling for policy analysis?*

This research compared three machine learning models: logistic regression (LR), chosen as a baseline model, Random Forest (RF), and XGBoost (XGB). In response to the research gaps, the models were trained using aggregate data at a Nomenclature of territorial units for statistics (NUTS) 2 level, consisting of basic regions in the EU. To answer the second research gap, a case study of the NEAC MNL mode choice model, developed and maintained by Panteia, was undertaken. The NEAC MNL model is currently used by Panteia to estimate and forecast transport demand in response to cost and infrastructure changes. To assess whether ML models are suitable for this type of application, the ML models trained in this study were compared to the NEAC MNL model based on seven criteria: predictive performance, interpretability, practicality, computation time, generalizability, robustness, and data efficiency.

The dataset was constructed using road, rail, and inland waterway freight flows datasets from 2015 accessed through Eurostat. The explanatory variables included from Eurostat and other sources were generalized costs, commodity type, inland waterway availability, distance from terminals, East/West Europe dummy variables, and rail and inland waterway service quality. Due to some missing or

unavailable data from Eurostat, namely for road, some adjustments were made to the data resulting in a dataset that differs from real-world mode shares and patterns: inland waterway is overrepresented by 154.2%, rail by 10.8%, and road by 7.9%. Additionally, compared with Panteia's cost estimates from 2017, the calculated generalized costs were on average lower for road and inland waterway and higher for rail. In some OD pairs, the relative cost ranking between modes also shifted, revealing possible data quality issues due to inaccuracies in the feature values.

Several highly correlated variables were not included in the models, as high correlations cause feature importance measures to be unreliable and less interpretable. Many variables had high Variance Inflation Factor (VIF) values, indicating widespread multicollinearity. These variables were kept in the models despite the multicollinearity as removing all of them would overly simplify the models and reduce interpretability.

After feature selection and engineering, the data was split into a training set and test set stratified by OD pair and mode. The best hyperparameters for each algorithm were found using RandomizedSearchCV. The hyperparameter ranges were adjusted based on the gap between the training set log loss and test set log loss in the 5 cv-folds, a larger gap being an indication of model overfitting (Hawkins, 2004). The models were trained and evaluated using the tonnage amounts in each row as sample weights, which penalizes the model for incorrect predictions on higher tonnage rows. To evaluate model interpretability, Shapley Additive exPlanations (SHAP) values were calculated for RF and XGB, and the estimated Logit coefficients for LR were discussed.

The models were assessed and compared to each other and the NEAC MNL model using the previously mentioned criteria. The XGBoost model had the best overall predictive performance, followed by Random Forest, and finally, logistic regression. The predictive performance of the models was evaluated using the accuracy, precision, recall, log-loss, overfitting gap (i.e., difference between training and test set log-loss), and differences between actual and predicted mode shares.

The MNL model was considered the most interpretable and practical. Likely due to the high multicollinearity of some variables, some of the Logit coefficients appeared counterintuitive and poorly identified, which reduced the LR model's interpretability. Using beeswarm plots to observe the mean average SHAP values for each variable per class, some of the RF and XGB SHAP values appeared counterintuitive as well. In all three models, a higher rail service (i.e., better quality service) negatively affected the probability of choosing rail. XGB had more behaviorally realistic SHAP values than RF due to each mode's respective costs being more important in predicting that mode in the former model.

The model with the shortest computation time was the logistic regression/MNL model. Logistic regression was the most data efficient, as it had the smallest change in performance with greater amounts of training data. XGB had the greatest improvement in performance with more data, making it the least data efficient.

To assess generalizability, one country was omitted from the training set and used as the test set. This was repeated for each country in the dataset for each algorithm. None of the models were particularly generalizable to new, unseen countries, as they had an average 25.3-26.3 percentage point drop in accuracy across all countries and an average 0.485-0.583 increase in log loss compared to when they were trained and tested on the same countries. Logistic regression was the most generalizable model as it performed similarly or only slightly worse than RF and XGB on almost all countries and substantially better on three countries.

Robustness was tested by introducing varying levels of noise with a Gaussian distribution into the training set cost and terminal distance variables. As noise increased, accuracy and log loss worsened for all models, but LR performance dropped the least, making it the most robust to noise.

In earlier research comparing machine learning models for mode choice predictions, the differences in accuracy between linear models such as logistic regression and more complex models such as RF or XGB were much greater than the 2 to 3 percentage points exhibited in this study. This could be because the relationships in the dataset were largely linear, thus the slight improvements in accuracy with RF and XGB were in the few areas where there were nonlinearities. Recall and precision for the minority classes of rail and inland waterway were lower than for road, similar to other previous studies. This could be partly due to some data incompleteness and feature inaccuracies, with the expectation that if these data quality issues are addressed, the performance of the models would improve.

The ML and MNL models were determined to be complements rather than substitutes. To achieve the best performance, the ML models should be used for short- or medium-term forecasting, applied to scenario analyses that involve only small variations in feature values, and retrained whenever new scenario features are introduced. The XGBoost model was recommended to Panteia for further exploration of its implementation into the NEAC framework. Data quality issues should be addressed prior to the model's use, and after these are addressed, the model should be retrained using the existing workflow in this report.

Future studies could assess the performance of other algorithms as well as ensembled learning techniques for this use case. Secondly, adding more explanatory variables, including physical locations of shippers and buyers and mode-specific characteristics, could help the models to differentiate inland waterway and rail from road, potentially improving minority class precision and recall. Lastly, hybrid models offer an opportunity to combine the advantages of both Logit and ML models into a single model. This would provide a more seamless solution compared to this study's recommended course of action where either the MNL or ML model is employed depending on the project.

# Contents

# 1 Introduction

## 1.1 Background

Freight transportation represents 5% of the Gross Domestic Product (GDP) of the EU, as well as 25% of the EU's total greenhouse gas emissions, making it a highly impactful industry for European businesses and citizens (Eurostat, 2022). The sector has experienced significant changes, including the disruption caused by COVID-19, which interrupted the steady growth in the volume of goods transported since 1995. Other changes include policies that aim to reduce emissions by encouraging a shift to renewable energy and more sustainable transportation modes, such as inland waterways and rail, which have lower environmental footprints than road transport (European Commission, 2024). Currently, approximately 77% of European freight transport occurs by road, followed by 17% by rail, and 5% by inland waterways (Eurostat, 2020).

Freight mode choice models are used to estimate, analyze, and forecast transport demand and assess how various factors influence mode choice decisions. These models help to evaluate the impact of environmental policies, infrastructure projects, and regulatory changes on freight transport patterns. By providing insights into modal distribution, mode choice models support strategic planning, sustainability efforts, and logistics optimization.

## 1.2 Problem Definition

Traditionally, mode choice models have been estimated using the Multinomial Logit (MNL) model, introduced by McFadden in the 1970s for transportation applications. This model remains a highly popular tool for modeling mode choice due to its ease of development, transparency, and interpretability (Benjdiya, Rouky, Benmoussa, & Fri, 2023). Other Logit-based models, namely Mixed Logit and Nested Logit, were created to overcome some of the limitations of the MNL model and are also commonly used to predict mode choice.

One of the disadvantages of Logit-based models is that they typically perform poorly in predictive accuracy. For this reason, the mode share output of these models often does not closely reflect real-world behavior, leading these models to either under- or overestimate the effects of policies and other changes in transportation patterns (Hillel, Bierlaire, Elshafie, & Jin, 2021). This means that these policies cannot more precisely target certain behaviors in order to produce the desired outcome.

Machine learning models have more recently been explored as an alternative to Logit-based models. These have been shown to produce higher accuracy. However, the application of machine learning for freight policy analysis as well as these models' performance with aggregated data has not been explored. Even if machine learning models produce better accuracy, there are other important model characteristics that these models must meet in order to be suitable for freight policy analysis.

## 1.3 Research Objective and Questions

The aim of this study is to explore the contribution that machine learning approaches can make to freight mode choice modeling. Specifically, the focus of this study is on the performance of machine learning mode choice models using aggregated data and on the application of such models in freight policy analysis. This research will use a case study of the NEAC European freight transport model.

The main research question is:

*What role can machine learning-based approaches have in freight mode choice modeling for policy analysis?*

This question is answered by the following sub-questions:

1. What are the criteria a machine learning mode choice model should meet to be suitable for freight transport policy analysis?
2. Which machine learning methods are most suitable for modeling freight mode choice?
3. What additional explanatory variables and external datasets can enhance model performance?
4. How does the performance of machine learning models compare to that of an MNL model?
5. Based on the results, should a machine learning mode choice model be incorporated into the NEAC framework, and if so, under what conditions?

## 1.4 Case Study Description: NEAC Freight Transport Model

NEAC is a European-wide freight transport model developed and maintained by Panteia. The model is used for forecasting and assessing the effects of transport policy and infrastructure changes on transportation patterns (Leest, Duijnisveld, & Hilferink, 2006). It is used for both predictions now and forecasting based on future possible conditions. NEAC contains several submodels, including a trade model, mode choice model, and assignment model, as well as a mode chain builder which was used to create the 2010 base year database. The mode chain builder takes national-level trade data and converts it to origin-destination (OD) matrices for each mode. This is then calibrated using known transport data. The 2010 base year database contains estimated tonnes per commodity type transported between regions in Europe. There are 10 commodity types, based on the Standard goods classification for transport statistics (NST). The regions are based on the EU Nomenclature of territorial units for statistics (NUTS) level 3 regions of which there are 1,585 in Europe.

The NEAC mode choice model uses a Multinomial Logit (MNL) model to predict changes in road, rail, and inland waterway mode shares between time periods. These predicted mode share changes are based on defined scenarios, such as changes in the cost structure or in the networks (Newton, Kawabata, & Smith, 2015). The probability that a given mode is chosen is calculated using the formula:

$$P_{m|cij} = \frac{e^{V_{m|cij}}}{\sum_{i \in M} e^{V_{l|cij}}}$$

$$with: V_{m|cij} = \beta_{m0} + \sum_{k} \beta_{mk} x_{cijmk}$$

Where:

*M:* set of available modes
$P_{m|cij}$: choice probability of mode *m* given commodity group *c* and OD relation *ij*
$V_{m|cij}$: systematic utility of mode *m* given commodity group *c* and OD relation *ij*
$x_{cijmk}$: level of service *k* for mode *m* given commodity group *c* and OD relation *ij*
$\beta_{mk}$: logit parameter for mode *m* and level of service *k*

Level of service refers to the explanatory variables which are the generalized costs and regional and border resistance dummy variables. The estimated shift per mode is calculated as the difference between the mode probabilities in the base year and the scenario year.

The NEAC mode choice model was formulated in 2005, with a more recent update on the base year data in 2015 (Newton, Kawabata, & Smith, 2015). At the time the mode choice model was updated in 2005, the MNL formulation was chosen over machine learning because firstly, machine learning was deemed to not be transparent enough for NEAC use, and secondly, user-friendly machine learning software tools were not readily available at the time (Leest, Duijnisveld, & Hilferink, 2006). Lastly, it was determined that Artificial Neural Networks (ANN) are better for spatial rather than temporal forecasting.

Since 2005, a wide number of advancements have been made in the field of machine learning. User-friendly software is now available, making it easier for researchers without a data science background to train their own models. Many other methods besides ANN have been developed and applied to predict mode choice. Other tools such as Shapley Additive exPlanations (SHAP) are used to help make machine learning models more transparent. Given the large number of changes that have occurred since 2005, it now appears more appropriate to explore the potential integration of machine learning into the NEAC mode choice model. Therefore, in this study, NEAC's current MNL model will be compared to machine learning approaches.

## 1.5 Report Structure

Chapter 2 presents the literature review which consists of recent research on using machine learning for mode choice predictions, criteria used to evaluate the suitability of machine learning models for policy analysis, and freight transport demand predictors. Chapter 3 describes the data collection, preparation, and analysis steps conducted prior to training the models. The methodology of the machine learning models trained in this research are discussed in Chapter 4. In Chapter 5, the results of the different machine learning models are discussed and compared. Chapter 6 includes a discussion of the research's limitations as well as the type of policy analysis projects where these models could be applied. Finally, Chapter 7 summarizes the report, answering the research questions, and Chapter 8 provides recommendations for future research.

# 2 Literature Review

The literature review in this chapter presents existing studies on machine learning mode choice models for both freight and passenger transportation as well as hybrid models (models that incorporate Logit and machine learning characteristics). Section 2.4 provides an explanation of evaluation criteria useful for determining the suitability of machine learning models for freight transport policy analysis. Section 2.5 synthesizes factors influencing freight transport demand and mode choice, followed by a conclusion and discussion in Section 2.6.

## 2.1 Machine Learning for Freight Mode Choice Models

Several studies have compared the performance of various machine learning methods to more traditional models such as the MNL model in predicting freight mode choice. Uddin, Anowar, & Eluru (2021) developed eight machine learning models and an MNL model and evaluated their accuracy and precision in predicting mode choice with US commodity flow data. Accuracy is the percentage of correctly classified observations across all modes, while precision refers to the number of correct predictions for a specific mode. All machine learning models with the exception of Support Vector Machine (SVM) had a higher mean accuracy than the MNL model, with Random Forest (RF) having the best results (75%) and ANN (51%) having a slightly higher accuracy than the MNL (42%). SVM had an accuracy of 36.7%.

Lui et al. (2024) sought to improve the predictive accuracy of machine learning models with ensemble learning techniques where various models are combined for the best result. They used the same US commodity flow data as Uddin, Anowar, & Eluru (2021) and incorporated additional spatial information in the form of derived distances for all modes. They also created local models for each commodity and industry type and achieved 92% accuracy with ensemble learning and derived distances. Tree-based methods, particularly RF and Bagging Decision Tree, had the highest accuracy, aligned with earlier findings from Uddin, Anowar, & Eluru (2021). Similarly, Xu et al. (2024) compared the performance of RF, XGBoost, and CatBoost models with an MNL model using the same US commodity flow data and also found that the machine learning models performed better than the MNL model in terms of accuracy and precision.

Another recent study used machine learning methods to predict freight vehicle type choice for logistics firms, comparing RF to multinomial and mixed logit models using commercial travel surveys from Toronto, Canada (Ahmed & Roorda, 2022). They found that RF provided more accurate predictions (49.5% accuracy) compared to the MNL (41.7%) and mixed logit (39.9%) models. The low accuracies in general were attributed to the small sample size, and the mixed logit model accuracy was lower than the MNL model likely due to its lower number of explanatory variables. Finally, other earlier studies investigated the application of neural networks for freight mode choice, though the number of papers on this topic is quite limited. This is in part because machine learning methods are more effective when applied to disaggregate data, and this type of freight data is difficult and expensive to obtain (Samimi, Kawamura, & Mohammadian, 2011) (Benjdiya, Rouky, Benmoussa, & Fri, 2023).

One of the identified research gaps is that existing studies on machine learning for freight mode choice focus mainly on predictive accuracy, with less emphasis on other important model characteristics such

as interpretability. Although some of the studies mentioned above also use Shapley Additive exPlanations (SHAP) to explain the output of the machine learning models used, a more extensive exploration of interpreting machine learning models for freight mode choice is lacking. Secondly, most of these studies use North American disaggregated shipment-level data, and all of them apply models at either a national or even local (city-wide) level. Thus, the use of machine learning to predict freight mode choice between two or more countries using aggregated data has not been studied. Finally, additional factors such as shipping reliability and quality of service could be included in future models to further improve predictive accuracy, as these also influence freight mode choice (Xu, et al., 2024).

## 2.2 Machine Learning for Passenger Mode Choice Models

Although machine learning models have been shown to have significantly higher predictive accuracy for freight mode choice than MNL and mixed logit models, the results for passenger mode choice are more mixed. Martin-Baos, Lopez-Gomez, Rodriguez-Benitez, Hillel, & Garcia-Rodenas (2023) compared the MNL model to five machine learning models: SVM, RF, XGBoost, ANN, and Deep Neural Network (DNN). They found that while all of the machine learning models had higher accuracy than the MNL model, this accuracy was at most only 6.16% higher. This aligns with previous research showing that discrete choice models (DCMs) had only 3-4% lower accuracy than the top machine learning classifiers (Wang, Wang, & Zhao, Deep neural networks for choice analysis: Extracting complete economic information for interpretation, 2020), as cited in (Martin-Baos, Lopez-Gomez, Rodriguez-Benitez, Hillel, & Garcia-Rodenas, 2023).

Zhao, Yan, Yu, & van Hentenryck (2020) compared two logit models (MNL and mixed logit) with seven machine learning methods using stated preference survey data. They found that while tree-based methods such as RF and BAG were significantly more accurate than the logit models, Naïve-Bayes (NB), CART, and ANN performed either as well as or slightly worse than the logit models. Using ensemble techniques, Zhang, Zhang, Liu, & Zhang (2023) achieved 83% accuracy on passenger travel mode choice in Jinan, China, compared to 66% predictive accuracy for the MNL model, which was also the same accuracy as the AdaBoost model they tested. Literature on passenger mode choice indicates that predictive accuracy can vary greatly depending on the data used and the selected machine learning method.

Additionally, Hillel, Bierlaire, Elshafie, & Jin (2021) highlight a number of limitations in previous studies that may have led to overestimations in the predictive improvements gained from machine learning methods. These limitations include using input features which are dependent on output choice, for instance using trip duration for prediction even though this is dependent on the mode taken. Other limitations include using inappropriate validation schemes, using incorrect sampling methods, and optimizing hyperparameters on test data, all of which may result in inaccurate estimates of model performance (Hillel, Bierlaire, Elshafie, & Jin, 2021). Of the 70 studies included in their research, only one did not have any of the identified limitations.

Despite the higher predictive ability of machine learning models, many passenger mode choice studies still prefer DCMs due to their advantages in understanding causality, greater interpretability, and better generalization (Benjdiya, Rouky, Benmoussa, & Fri, 2023). Similarly to freight mode choice studies, much research on passenger mode choice has focused on comparing the predictive accuracy of various machine learning methods with logit models. The findings range from significant accuracy

improvements with some tree-based methods such as RF and BAG to either the same or lower accuracy in ANN and NB compared to logit models.

## 2.3 Hybrid Models

Other studies have investigated how machine learning methods can be combined with DCMs to potentially gain improvements in predictive accuracy without losing interpretability. Aboutaleb, Danaf, Xie, & Ben-Akiva (2021) argue that machine learning models cannot replace DCMs for policy analysis because they can only capture correlations, not causations. Additionally, if the results are counterintuitive, machine learning models do not provide a clear way of understanding what went wrong and how it should be fixed (Aboutaleb, Danaf, Xie, & Ben-Akiva, 2021). For these reasons, they suggest using optimization techniques, regularization, and out-of-sample validation to determine the specification of the random component of the utility equations in logit models and discuss how this can be applied to nested logit and mixed logit models.

Similarly, Sifringer, Lurkin, & Alahi (2020) propose a hybrid approach that integrates neural networks into discrete choice modeling, referring to it as the Learning-Multinomial Logit Model (L-MNL). This approach allows for the utility function to be partially specified using neural networks, enabling the capturing of non-linear relationships. In their framework, the utility function is divided into one interpretable, knowledge-driven part, written as a convoluted neural network, and a data-driven part, obtained using a dense neural network. The L-MNL model was applied to Swissmetro revealed preference data and resulted in higher accuracy (76%) compared to the 66% for the standard MNL model, likely because the traditional MNL model was not able to capture non-linear dependencies among variables. The authors argue that interpretability is also maintained, as key variables such as time, cost, and trip distance are kept in the MNL part of the model, while sociodemographic variables such as age and education are added to the data-driven part (Sifringer, Lurkin, & Alahi, 2020). An extension of the L-MNL model, the TasteNet-MNL model, uses a neural network to learn individual-specific taste parameters, which are then inputted into a utility function (Han, Camara Pereira, Ben-Akiva, & Zegras, 2022). This results in an interpretable model with improved model fit, indicated by a lower negative log-likelihood, compared to models with manually specified parameters.

Another framework combining Logit models and machine learning is the architecture with alternative-specific utility functions (ASU-DNN) in which a deep neural network (DNN) is designed to use only the data for each specific alternative to calculate that alternative's utility (Wang, Mo, & Zhao, Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions, 2020). In a fully connected DNN (F-DNN), the utility of each alternative is computed using the attributes of all alternatives, which violates the independence of irrelevant alternative (IIA) constraint in Logit models. By designing the DNN to adhere to IIA, the model provides more interpretable behavioral insights than an F-DNN does and also gains an 8% increase in accuracy compared to a baseline MNL model.

Areas for future research related to hybrid models include developing a framework for the inclusion of a data-driven part to mixed logit, latent class, or other more advanced DCMs. As existing studies on hybrid models focus on passenger data, the question of interpretability and the application of these models for freight mode choice is also another research gap.

## 2.4 Machine Learning Model Evaluation Criteria for Freight Policy Analysis

A study on passenger travel mode choice defined the following criteria for selecting machine learning models used to inform policy decisions: predictive performance, behavioral interpretability and explainability, computational complexity, and data efficiency (Martin-Baos, Lopez-Gomez, Rodriguez-Benitez, Hillel, & Garcia-Rodenas, 2023). In a separate paper, the AP-GRIP framework was introduced to evaluate train delay prediction models. Using this framework, the models are assessed based on their accuracy, precision, generalizability, robustness, interpretability, and practicality (Yong, Ma, & Palmqvist, 2025). The criteria mentioned in these two studies are described below. No other studies were found that define evaluation criteria for machine learning models applied to freight transport policy analyses or other policy areas; existing work emphasizes predictive performance as the primary basis for assessing and comparing machine learning models.

*Predictive performance:* in machine learning studies, predictive performance is most commonly assessed using accuracy, which measures how often the model correctly predicts the outcome. It is also the most widely used metric for model comparisons (Martin-Baos, Lopez-Gomez, Rodriguez-Benitez, Hillel, & Garcia-Rodenas, 2023). Another commonly used metric is precision which measures the variability or uncertainty in prediction errors (Yong, Ma, & Palmqvist, 2025).

*Interpretability:* for policy applications, interpretability includes both behavioral and technical aspects. Behavioral interpretability involves deriving economic indicators such as willingness to pay or elasticities, which are particularly relevant in passenger transportation studies (Martin-Baos, Lopez-Gomez, Rodriguez-Benitez, Hillel, & Garcia-Rodenas, 2023). More generally, interpretability can be defined in two ways: first, as providing explanations for users to assess the impact of the inputs on the outputs, such as through feature importance analysis. The second, deeper definition of interpretability allows users to understand not only the importance of the features involved in the prediction but also how the model itself learns from the input data (Yong, Ma, & Palmqvist, 2025). For this deeper interpretability, post-hoc techniques are often applied to black-box machine learning models.

*Practicality:* lastly, practicality considers whether the model's outputs are usable and meaningful for end-users. For example, in train delay predictions, a model whose outputs fluctuate excessively may lead to user distrust, making this model less practical in real-world applications (Yong, Ma, & Palmqvist, 2025).

*Computational complexity (computation time):* different models vary in computational complexity, which depends on a number of factors including the algorithm itself, the number of input features, and the hyperparameters. Models that are less computationally complex may be preferred depending on the resource availability and the required level of predictive accuracy (Martin-Baos, Lopez-Gomez, Rodriguez-Benitez, Hillel, & Garcia-Rodenas, 2023).

*Generalizability:* this criterion relates to a model's performance on new, "unseen" data that differs from the data that it was trained on. Generalizability is assessed through external validation, which involves training on one dataset and testing on another, either from a different time period or geographic region (Yong, Ma, & Palmqvist, 2025).

*Robustness:* robustness in machine learning can refer to adversarial robustness, robustness to natural distribution shifts (similar to the previously described generalizability), shortcut learning, or other

specific robustness notions (Freiesleben, T & Grote, 2023). In this study, robustness refers to a model's ability to produce acceptable predictions under poor data quality or exceptional conditions. This can be evaluated by testing model performance under various realistic perturbations in the data (Yong, Ma, & Palmqvist, 2025).

*Data efficiency:* data efficiency refers to a model's ability to learn effectively from limited data. In practical applications where large, high-quality datasets may not be available, models that can perform well with less data are preferable (Martin-Baos, Lopez-Gomez, Rodriguez-Benitez, Hillel, & Garcia-Rodenas, 2023).

The relative importance of these evaluation criteria may differ depending on the individual priorities of the model user or domain expert. In this study, the ordering of the criteria is based on the perspective of a potential end-user of the model. Predictive performance is considered the most important, as improving predictions is one of the main motivators in exploring an alternative to the current MNL model. Interpretability and practicality are also highly valued; models that are too complex become less useful if this shifts the focus of the client away from the model's results to the more technical aspects of the model itself. Next, computation time is important due to the need to complete analyses within a limited project timelines. In freight transport policy analysis, robustness and generalizability are important as they help ensure the models produce credible results under noisy or imperfect data, which is common in freight datasets, and when applied to shifting conditions, which such models are made to assess. However, these criteria are considered secondary to predictive performance, interpretability, and computation time. Lastly, data efficiency may be important in contexts where large amounts of high-quality data are unavailable, but in this case it is considered least important, as the potential end-user prioritizes other factors over minimizing data requirements.

## 2.5 Predictors of Freight Transport Demand and Mode Choice

One of the main drivers of freight transport demand is Gross Domestic Product (GDP), which has been shown to account for 81% to 92% of the variation in total tonne-kilometres of transported goods across countries (van de Riet, de Jong, & Walker, 2012). GDP acts as an indirect driver of freight demand, primarily by influencing consumer demand, which more directly drives freight activity. Other key factors influencing freight demand include the economic structure and logistics systems within countries, as well as mode characteristics (van de Riet, de Jong, & Walker, 2012). The relationships between these factors are shown in Figure 2.1 below.

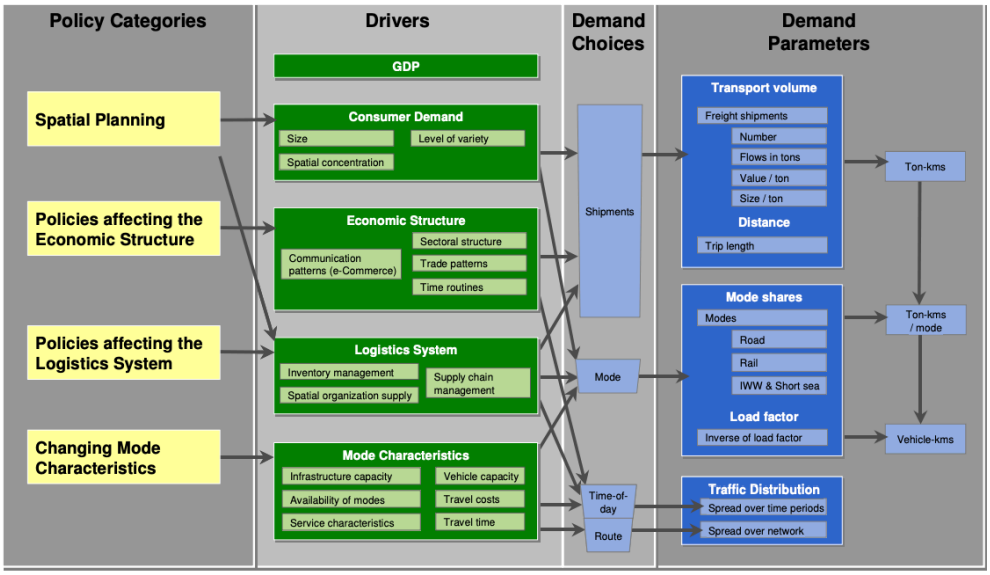


*Figure 2.1. Drivers of Freight Transport Demand (van de Riet, de Jong, & Walker, 2012)*

Factors affecting freight mode choice can be divided into four categories: industry of the shipper, commodity characteristics, including commodity type, shipment size, and shipment value, mode characteristics, and infrastructure characteristics, which include network density and the presence of intermodal facilities (Xu, et al., 2024). Mode characteristics which affect the choice of mode include infrastructure and vehicle capacity, availability of modes, service characteristics, travel costs, and travel time (van de Riet, de Jong, & Walker, 2012). Another influencing factor is the location of the sellers, buyers, and shipping carriers (Xu, et al., 2024).

Regarding mode availability, road transportation tends to be more widely available than rail or inland waterways due to the former's denser networks. Service characteristics include cost, travel time, reliability, flexibility, tracing of freight, use of infrastructure, scale/volume, service of terminals, legislation, safety, and security (van de Riet, de Jong, & Walker, 2012). Road transportation is commonly preferred, especially for short and medium distance trips, due to its greater flexibility and availability. For large volumes of lower-value bulk goods, rail and inland waterways are preferred, whereas air and truck transport are preferred for goods with higher value-to-weight ratios (van de Riet, de Jong, & Walker, 2012).

## 2.6 Conclusion and Discussion

Previous studies on both freight and passenger mode choice demonstrate that machine learning methods often offer greater predictive accuracy compared to Logit models, with tree-based methods such as RF providing the highest accuracy. In freight mode choice models, accuracy can also be improved through ensembled learning techniques, the inclusion of additional spatial information such as derived distances for all modes, and the specification of local models per commodity type and industry. Machine learning methods work best with large datasets, but they can still offer higher accuracy than MNL models on smaller datasets (Ahmed & Roorda, 2022). Some studies use SHAP to explain the output of machine learning models, but most research focuses on the predictive performance of these models and does not explore the interpretability or application of machine learning for freight policy analysis. Hybrid models such as the Learning-MNL model offer a way to maintain

interpretability while improving predictive accuracy, though these frameworks have only been developed for passenger mode choice predictions.

This research aims to address two key gaps in the existing literature. First, previous studies on machine learning for predicting freight mode choice have used disaggregate shipper survey data, with detailed information about shipment value, shipment weight, commodity type, trip distance, and shipper industry type, to train machine learning models. This research explores the performance of machine learning models trained with aggregated freight data and how this compares with Logit models as well as what additional factors can be incorporated to enhance the predictive accuracy of machine learning models for cross-border freight transport. The second research gap lies in the suitability of machine learning-based freight mode choice models for policy analysis. Previous studies have focused mainly on the predictive accuracy of machine learning models and have not considered other model characteristics such as interpretability and the role this may play in the adoption of machine learning for freight policy studies. By comparing different machine learning models and considering their advantages and disadvantages compared to an MNL model, this research will explore whether and how machine learning can enhance mode choice models used in freight policy studies.

# 3 Data Preprocessing

In the first section of this chapter, the freight flows data sources and how this data was assembled are described. Section 3.2 defines the explanatory variables used in this research. Section 3.3 provides a description of the data analysis conducted on the combined dataset, including the correlations between explanatory variables. Lastly, Section 3.4 summarizes the main conclusions from this chapter.

## 3.1 Freight Transport Data Sources

The data for actual freight flows per mode is from Eurostat. As described earlier in Section 1.4, the NEAC mode choice model uses a base year database to estimate modal shifts in a scenario year. In order for the machine learning models to learn the patterns in real-world behavior rather than assumptions made in the base year database, actual freight flows data was used when possible. The NEAC base year estimated data was used to fill in some of the gaps where real data was not available.

For inland waterway, the dataset used contains information about freight transported from region of loading to region of unloading per 22 cargo types (Eurostat, 2024). These regions are at a NUTS 2 level, which are basic regions suitable for regional policy use (Eurostat, n.d.). There are 20 countries in the EU for which there is inland waterway transportation data available, and these figures are reported by eight different reporting countries. The data is reported on an annual basis from 2014 to 2023; for this research, the 2015 annual data was utilized.

The rail transportation dataset used from Eurostat also provides freight flows by NUTS 2 region of loading/unloading (Eurostat, 2024). However, this is only provided in terms of total cargo, not by cargo type. Other Eurostat datasets break down total rail flows by cargo type, but these are only available by country, not NUTS 2 or 3 zones. There are 20 EU countries with available rail transportation data, and each of these report transported freight by tonnes. In this dataset, annual data is provided for 2005, 2010, 2015, and 2020.

Finally, two datasets for road transportation are used: one on road freight transport by region of loading, and the second of road freight transport by region of unloading (Eurostat, 2024) (Eurostat, 2024). There is no single dataset for road flows from region of loading to region of unloading as there is for inland waterway and rail. Instead, these two datasets provide region of loading/unloading by reporting country. Each reporting country reports amount transported by vehicles registered in that country inside and outside of the country. These amounts are reported in terms of total annual tonnes from 1999 to 2023 for NUTS 3 level zones for 20 countries in the EU.

In the rail and inland waterway datasets, there are multiple tonnes amounts for the same OD pair reported by different countries. When downloading the data, only one of these values was selected, with a preference for the amount reported by the country where the goods were loaded. If this was unavailable, then the value reported by the country of unloading was used. For inland waterways, because not all countries of loading/unloading were also reporting countries, if these had no reported values for a given OD pair, then the value reported by Germany was used as Germany is well-connected to many of the other countries in the EU by inland waterways and thus provides a lot of data for the region.

The road datasets were restructured into a format matching the inland waterway and rail datasets. First, in the “region by loading” dataset, an assumption was made that the reporting country was the same as the destination country. In reality, the reporting country is the country where the vehicle is registered, not necessarily the vehicle’s destination. However, this was assumed in order to be able to adjust the dataset into a usable format; the results are further discussed in Section 3.3. Data Analysis. Then, total unloaded values within a given destination country were allocated to NUTS 3 zones in that country based on the distribution of unloaded goods in those NUTS 3 zones from the same origin country in the “region by unloading” dataset. This resulted in a matrix with road OD flows at NUTS 3 level, based on the totals from the region of loading dataset. The data was aggregated to NUTS 2 to match with the other mode datasets, and all origin-destination data for all modes was combined into a single dataset to be used in model training.

## 3.2 Explanatory Variables

For each origin-destination pair, additional information about the generalized costs, commodity types, inland waterway availability, travel time, travel distance, distance from rail and inland waterway terminals, and regional characteristics was acquired from other data sources. These variables are those which have been identified in previous studies as being influential factors in freight mode choice and for which data was available. An explanation of each variable and how it was integrated with the Eurostat datasets is provided in the following sections.

### 3.2.1 Generalized Transportation Costs

Transportation costs are a key factor in freight mode choice, with an increase in the costs for one mode typically leading to a shift to other transportation modes (van de Riet, de Jong, & Walker, 2012). In standard Logit models, the influence of transportation costs on mode choice is usually assumed to be linear, although several studies have shown that nonlinear transformations of costs can improve model fit (Jensen, et al., 2019) (Xu, et al., 2024). In this research, these costs are included as linear inputs.

The NEAC generalized cost formula contains five basic elements: track or infrastructure, traction or haulage, equipment (wagons, containers, etc.), terminals or transshipment/loading points, and service (Newton, Kawabata, & Smith, 2015). These are divided into distance-based variable costs and time-based fixed costs. For each mode and OD pair in the dataset, the generalized costs as well as several individual cost components which are incurred across all modes are calculated. Rail and road costs are calculated based on country, and inland waterway costs are determined based on CEMT size, or the size of the vessel.

Total kilometres and minutes for each mode and OD pair were taken from a NEAC cost model output from 2017. This output contains the time and distance as well as generalized costs per mode for each OD pair at the NUTS 3 level. To use these values for NUTS 2 level freight flow data, the time and distance for all NUTS 3 OD pairs within each NUTS 2 OD pair were averaged and used as the NUTS 2 time and distance. The OD pairs for which this data was not available in this dataset were dropped. To obtain country-specific distances for calculating rail and road costs, the shortest path between each NUTS 2 zone’s centroid along each mode’s network was calculated in QGIS. The total kilometres and minutes for each OD pair and mode from the 2017 data were then distributed to the countries along this path using the QGIS amounts as weights. The QGIS calculations were only used where necessary to fill in the gaps from the 2017 file in order to reduce data validation time. For inland waterway, the

assumed CEMT size for each OD pair is the one that produced the generalized cost amount closest to the 2017 cost model amounts.

The final cost calculations are the total generalized cost, seven fixed time-based costs (depreciation, insurance, drivers' wages, other fixed traction costs, fixed equipment costs, fixed service costs, fixed maintenance costs), five variable distance-based costs (variable maintenance costs, track costs, fuel costs, other variable traction costs, and variable equipment costs), and one fixed terminal cost for each mode. For each row in the dataset, the costs of the chosen mode as well as the other possible alternatives are calculated.

Several of these individual costs are zero for certain modes. These are fixed equipment costs, fixed service costs, fixed maintenance costs, variable maintenance costs, other variable traction costs, and variable equipment costs. These are eliminated from the dataset as they would allow the model to learn to predict mode choice with just these variables (e.g., learning that any time fixed service costs = 0, the correct mode is inland waterway) instead of learning from the entirety of the data.

The generalized costs per OD pair and mode differ from the 2017 costs for a number of reasons. Firstly, the 2017 costs are based on NUTS 3 OD pairs, whereas these calculations are adjusted for NUTS 2 zones. The assumed CEMT size for inland waterway and the countries in the shortest path calculated in QGIS for road and rail may also differ from the 2017 values. For inland waterway, the newest costs are on average €8.33 lower than the 2017 costs. For road, the newer costs are also lower, with an average difference of €15.46. For rail, the newer costs are on average €10.18 greater than the earlier costs. For some OD pairs, the relative difference in costs between modes has shifted; for instance, where rail ought to be less expensive than road, in this new calculation, the road cost is lower than rail.

### 3.2.2 Commodity Types

Commodity type is a known factor in freight mode choice: for instance, low-value, bulk items such as coal, grains, and chemicals are more often transported by rail, whereas higher-value items such as prepared foods, electronics, and textiles are more commonly transported by road (van de Riet, de Jong, & Walker, 2012).

For inland waterway, the commodity types from Eurostat were used. Since this was not available for the other two modes at a NUTS 2 level, the commodity types from the NEAC 2010 base year estimated data were used for rail and road. To add commodity type for road and rail, first the estimated tonnes data was aggregated from NUTS 3 to NUTS 2 zones. In the Eurostat freight flows data, each tonnage amount per mode and OD pair is represented by one row in the dataset. This was split into multiple rows, one per commodity type, based on the commodity types transported along this OD pair in the NEAC base year estimated data. The Eurostat tonnes for each OD pair were then allocated to the commodity-specific data rows based on the estimated share of tonnes of each commodity type transported between that OD pair.

The inland waterway Eurostat data contains 22 cargo types, whereas the NEAC base year database is based on the Standard Goods Classification for Transport Statistics (NST) of 10 commodity types. In order to match these, the Eurostat cargo types were converted to the NST commodity type they most closely align with. NST 10 was added to encompass all unknown commodity types. This mapping is shown below in Table 3.1.

*Table 3.1. NST Commodity Type and Inland Waterway Cargo Type Mapping*

| NST Code | NST Description | Cargo Type Description |
|---|---|---|
| 0 | Agricultural products and live animals | Dry bulk – agricultural products |
| 1 | Other food products and animal feed | |
| 2 | Solid mineral fuels | Dry bulk – coal |
| 3 | Petroleum and petroleum products | Liquid bulk – refined oil products<br>Liquid bulk – other<br>Liquid bulk – unspecified |
| 4 | Ores, metal waste, roasted iron oxide | Dry bulk – ores |
| 5 | Iron, steel, and non-ferrous metals (incl. semi-finished products) | Other cargo – iron and steel products |
| 6 | Crude minerals and manufactured products; building materials | Dry bulk – construction materials<br>Dry bulk – other<br>Dry bulk – unspecified |
| 7 | Fertilizers | |
| 8 | Chemical products | Liquid bulk – chemicals |
| 9 | Vehicles, machinery and other goods (including general cargo) | Other general cargo<br>Other cargo – forestry products |
| 10 | Other/unknown | Other cargo - unknown |

### 3.2.3 Inland Waterway Availability

Mode availability can refer to both the mobile (e.g., trucks, wagons) and fixed (e.g., roads, railways) infrastructure required to ship goods, with road being more often the preferred mode due to its greater availability in terms of its network density (van de Riet, de Jong, & Walker, 2012). Including mode availability as an explanatory variable gives the model additional information about the nature of the transportation networks.

For each row of data, a column is added with a binary value indicating whether inland waterway transportation is available for this OD pair. Availability is determined based on whether a valid route is found between the two zones with Dijkstra's shortest path algorithm. The road and rail networks are much denser compared to inland waterway, thus it is assumed that road and rail are always available as mode alternatives for each OD pair. There are 8,509 unique OD pairs in this dataset. The total number of OD pairs for which inland waterway is available are 6,731 (79.10% of the total).

### 3.2.4 Travel Time and Distance

Travel time and distance are two other mode-specific characteristics that influence mode choice; a previous study on machine learning for freight mode choice found that travel time, distance, and cost were the explanatory variables with the greatest impact on mode predictions (Xu, et al., 2024). Different

modes are also known to be preferred for different trip lengths, with for example, the amount of goods transported by road typically dropping as travel distance increases (van de Riet, de Jong, & Walker, 2012). As with transport costs, travel time and distance are often assumed to have a linear relationship to mode choice, although nonlinear specifications of these variables in discrete choice models have been shown to result in an improved model fit than linear ones (Koppelman, 1981).

As described in Section 3.2.1, the generalized costs for each OD pair and mode are calculated based on the trip's length and travel time. Because of this, the variables time, distance, and cost are highly interconnected. For this research, travel time and distance are inputted as the total minutes and kilometres for each OD trip, based on the NEAC 2017 cost model output.

### 3.2.5 Distance from Terminals

The distance of an origin or destination zone to rail and inland waterway terminals is related to infrastructure characteristics, which as mentioned earlier, have been identified as influencing mode choice (Xu, et al., 2024). Although this distance would be included in the total distance of the trip and in the generalized cost, adding this as a separate variable may allow the model to identify other non-monetary costs associated with the distance to terminals that were not captured in the cost calculations.

The distances are calculated in QGIS as the Euclidean distance in kilometres between the NUTS 2 zone centroids and the nearest rail and inland waterway terminals. For each OD pair in the dataset, four distances are included: the distance from the origin zone to the closest rail terminal, origin zone to nearest inland waterway terminal, destination zone to closest rail terminal, and destination zone to closest inland waterway terminal.

### 3.2.6 Regional Characteristics

The quality of service and reliability of rail and inland waterway transportation are also factors in mode choice decisions (van de Riet, de Jong, & Walker, 2012). These can vary greatly across different countries. In the current NEAC MNL model, a distinction is made between Eastern and Western Europe, with separate parameters estimated per mode and region.

In this research, two additional columns are added to the dataset identifying whether the origin and destination are both in Eastern Europe, or one of the two zones is in Western Europe and the other in Eastern Europe, or vice versa. In the first column, a value of 1 indicates that both zones are in Eastern Europe. In the second of these columns, 1 indicates that the origin zone is in Western Europe and the destination zone is in Eastern Europe, or vice versa. The countries considered part of Western Europe are Austria, Belgium, Switzerland, Germany, Denmark, Spain, Finland, France, Ireland, Italy, Netherlands, Sweden, and Luxembourg (Statistics Netherlands (CBS), n.d.). Eastern European countries are Bulgaria, Croatia, Czech Republic, Hungary, Poland, Romania, and Slovakia.

In addition to distinguishing between Eastern and Western Europe, several other columns are added to represents differences in level of service in different countries. For rail, OD pairs are categorized into three levels of service: high quality, medium quality, and low quality. These are based on the 2017 European Railway Performance Index which grouped European countries into three tiers based on the intensity of use, quality of service, and safety of their passenger and freight railway systems (Duranton, Audier, Hazan, Langhorn, & Gauche, 2017). The ranking of countries is shown in Table 3.2; Croatia was not included in the 2017 performance index, so it was placed in the same category as its

neighboring countries Slovakia and Hungary. The OD pair is assigned to one of the categories based on the worst of the origin and destination countries' service levels.

*Table 3.2. Railway Systems Level of Service Categories by Country*

| Service Level | Countries |
|---|---|
| High | Switzerland, Denmark, Finland, Germany, Austria, Sweden, France |
| Medium | The Netherlands, Luxembourg, Spain, Czech Republic, Belgium, Italy |
| Low | Ireland, Hungary, Slovakia, Poland, Romania, Bulgaria, Croatia |

Three service level categories were also defined for inland waterway systems. The categories and the countries in each one are shown below in Table 3.3. There is no service performance index for inland waterway as there is for rail. For this reason, each country was ranked based on the average speed across all trips that originate in that country in the NEAC 2017 cost model output file. Although service level is typically measured by frequency and length of delays (van de Riet, de Jong, & Walker, 2012) (Duranton, Audier, Hazan, Langhorn, & Gauche, 2017), in this case average speed is used as a proxy as it can also be indicative of infrastructure bottlenecks, terminal delays, and other reliability issues. As with rail, the OD pairs in the dataset used in this research are assigned to one of the categories below based on the worst of their origin and destination countries' service levels.

*Table 3.3. Inland Waterway Systems Level of Service Categories by Country*

| Service Level | Countries |
|---|---|
| High | Belgium, Switzerland, Germany, Netherlands |
| Medium | Bulgaria, France, Italy, Romania |
| Low | Austria, Czech Republic, Hungary, Poland, Slovakia |

## 3.3 Data Analysis

The resulting dataset has 67,210 total data points, with 8,716 of these of goods transported by inland waterway, 25,639 of rail, and 32,855 of road. The mode shares by the number of data points is inland waterway (12.97%), rail (38.15%), and road (48.89%). The modal shares by the amount of tonnes transported per mode is inland waterway (9.32%), rail (11.5%), and road (79.17%). The actual mode split of inland freight transport in the EU in 2015 based on an aggregated dataset of modal shares in tonne-kilometres provided by Eurostat is inland waterway (6.9%), rail (18.8%), and road (74.2%) (Eurostat, 2025).

Compared with Eurostat's aggregate mode-specific data, the tonnage per mode in this dataset is under-estimated. 499,574 thousand tonnes are transported by inland waterway across all countries in this dataset, compared to 541,599 thousand tonnes in the aggregate data (Eurostat, 2025). Similarly, 616,377 thousand tonnes are transported by rail here, compared to 1,530,899 thousand tonnes in the aggregate rail transportation data (Eurostat, 2025). Road tonnage is the most underestimated, with 4,241,777 thousand tonnes in this dataset compared to the actual 12,665,191 thousand tonnes (Eurostat, 2024). These underestimations reflect excluded flows due to both missing or confidential values in the disaggregated Eurostat data as well as trips lacking a generalized cost estimation. For road, the underestimation is largely due to the way the two Eurostat datasets were combined.

The mode shares based on the Eurostat datasets described above are 3.67% inland waterway, 10.39% rail, and 85.95% road. Based on these percentages, the dataset used in this research overrepresents inland waterway by 154.2% and rail by 10.8% and underrepresents road by 7.9%.

Most total transported tonnes are domestic flows within Germany (2,725,777 thousand tonnes), Czech Republic (444,294), Austria (335,216), Switzerland (281,613), and Belgium (262,614). The most non-domestic flows are between Germany and The Netherlands (281,068 thousand tonnes), Belgium and The Netherlands (111,605), Belgium and Germany (51,449), Austria and Germany (45,652), and Germany and Poland (42,244). There are also seven country pairs with less than 2,000 tonnes transported between them: these are Denmark-Hungary, Switzerland-Denmark, Denmark-Romania, Belgium-Romania, Czech Republic-Norway, Spain-Poland, and Spain-Slovakia.

The most transported commodity type is NST6 (1,448,207 thousand tonnes), following by NST9 (1,192,152), and NST4 (666,703). The least transported commodity types are NST10 (27,158 thousand tonnes), NST7 (92,028), and NST3 (167,031). Figure 3.1 shows the percentage of tonnes of each commodity type transported by each mode. As shown in the figure below, two commodity types, NST1 (other food products and animal feed) and NST7 (fertilizers) are not transported by inland waterway. NST3 (petroleum and petroleum products) and NST2 (solid mineral fuels) have the highest percentage of tonnes transported by inland waterway compared to other commodity types. The commodities with the highest percentage of tonnes transported by rail is NST2, NST7 (fertilizers), and NST10 (other/unknown). Eight of the eleven commodity types are most commonly transported by road.

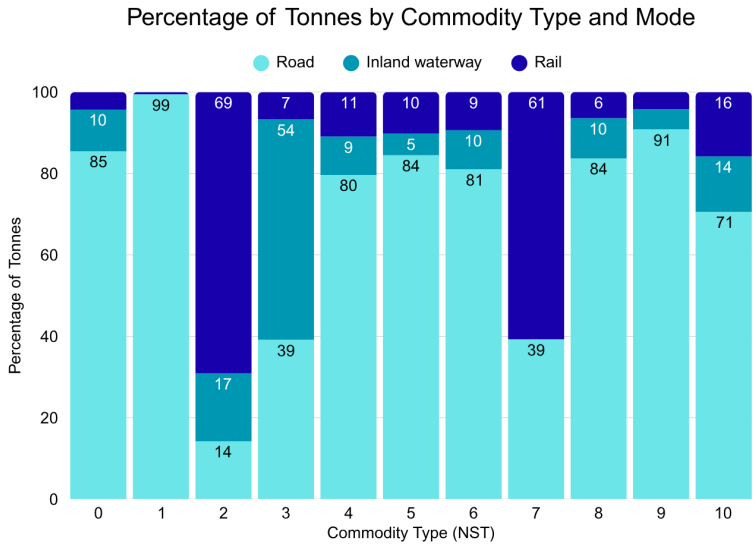


*Figure 3.1. Percentage of Tonnes Transported per Commodity Type and Mode*

For road flows, the Germany-Netherlands corridor is the largest both in this dataset and in regional reports (European Commission, 2017). However, the magnitude of some of the other trade corridors has shifted in this dataset, with Belgium-Netherlands and Germany-Poland transporting the next most amount of goods by road, whereas in official data, Belgium-France and Netherlands-Belgium are identified as the second and third biggest trade corridors in the EU (European Commission, 2017). Additionally, most domestic road flows are missing, but for the six countries that have domestic flows,

these amounts are largely consistent with aggregate reports (Eurostat, 2024). While some major corridors and relative rankings are preserved, differences arise due to missing entries and data filtering.

### 3.3.1 Correlations and Multicollinearity

High correlations between predictor variables cause feature importance measures to be unreliable, leading the model to become less interpretable (Kashifi, Jamal, Kashefi, Almoshaogeh, & Rahman, 2022). To check for correlations, the Phik correlation coefficients for all variable pairs are calculated. Phik can calculate correlations between categorical, ordinal, and interval variables, compared to Pearson correlation which is primarily intended for numerical variables (Baak, Koopman, Snoek, & Klous, 2020). Figure 3.2 displays a heatmap with the Phik coefficients for all variables except the individual cost components.

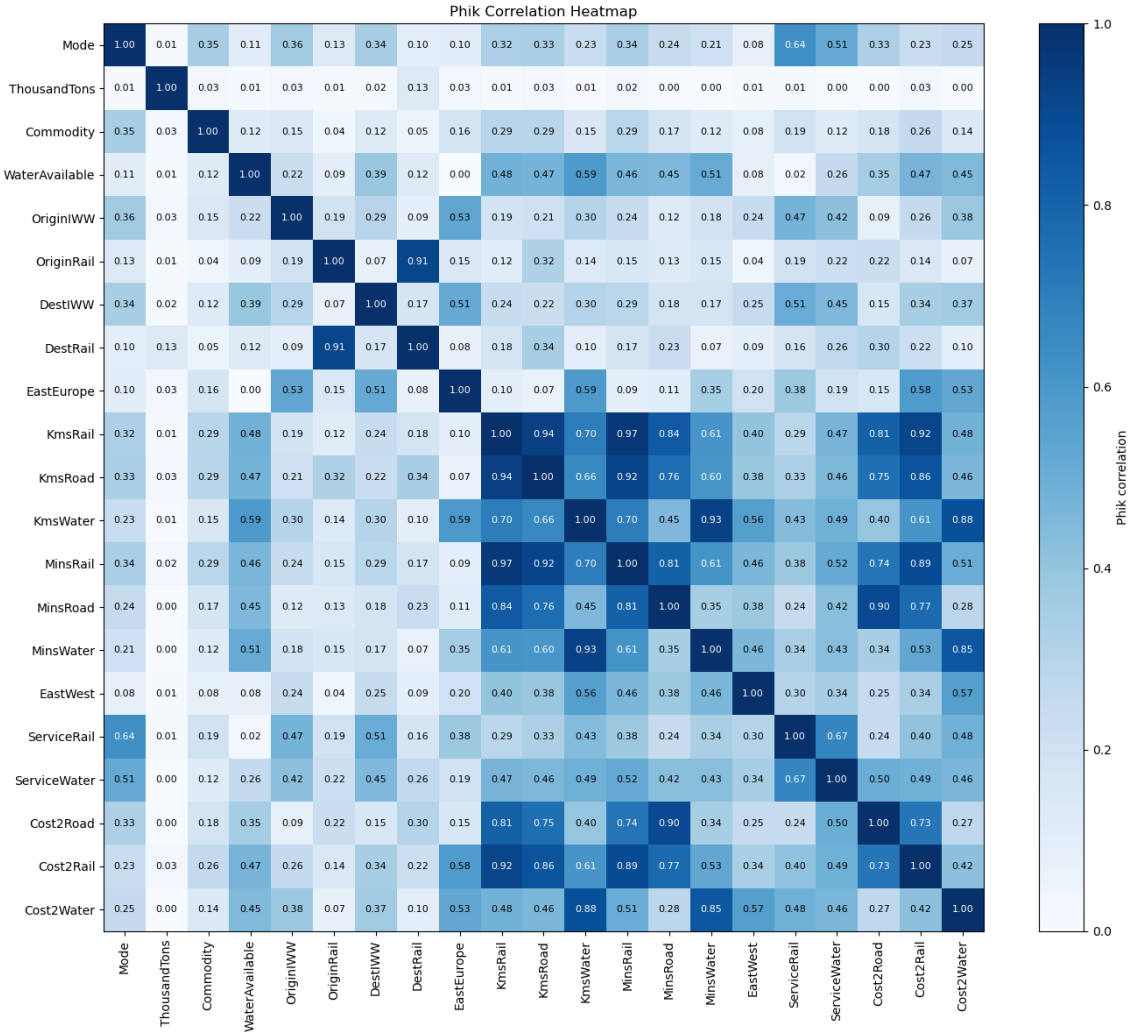


*Figure 3.2. Phik Correlations Between Explanatory Variables*

A correlation coefficient of 0.8 or 0.9 is typically used as a cut-off to indicate high correlation (Yi-Le Chan, et al., 2022). As seen in the heatmap, OriginRail (the distance from the origin zone to nearest rail terminal) and DestRail (distance from destination zone to nearest rail terminal) are highly correlated, with a Phik coefficient of 0.91. Total distance and time variables are highly correlated with each other

and with the generalized cost variables. Since the generalized cost variables are calculated from the individual cost component variables, these are also highly correlated. The set of variables with Phik correlations less than 0.80 are shown in Table 3.4 below.

*Table 3.4. Explanatory Variables without High Phik Correlations*

| Variable | Name in Dataset | Type of Variable |
|---|---|---|
| Generalized cost | Cost2Road<br>Cost2Rail<br>Cost2Water | Continuous |
| Commodity type | Commodity | Nominal |
| Inland waterway availability | WaterAvailable | Binary |
| Distance from terminals | OriginIWW<br>OriginRail<br>DestIWW | Continuous |
| East/West Europe | EastWest<br>EastEurope | Binary |
| Service levels | ServiceRail<br>ServiceWater | Ordinal |

High correlations between variables may indicate multicollinearity, when a linear or near-linear relationship exists between two or more predictive variables (Yi-Le Chan, et al., 2022). This increases the standard errors and makes it harder to measure the impact of an individual variable on the prediction. One way to measure multicollinearity is with the Variation Inflation Factor (VIF), defined as

$$VIF_j = \frac{1}{(1 - R_j^2)}$$

where $R_j^2$ is the R-squared from regressing $x_j$ on every other predictor. A VIF of 10 or more is considered the cut-off for high multicollinearity (Yi-Le Chan, et al., 2022).

Table 3.5 reveals the VIF values for many of the variables are very high, indicating high multicollinearity. These values fall below 10 only after dropping the variables road cost, rail service, origin distance rail, inland waterway service, and inland waterway availability. If road cost is not removed, the VIF value for rail cost remains high at 24.0 which shows there is high multicollinearity between road and rail cost. However, filtering out important variables solely due to their high multicollinearity may oversimply the models and reduce interpretability.

*Table 3.5. VIF Values*

| Variable | VIF |
|---|---|
| Cost2Rail | 39.73 |
| ServiceRail | 15.01 |
| WaterAvailable | 13.95 |
| ServiceWater | 12.72 |
| Cost2Road | 10.37 |
| Cost2Water | 3.93 |
| Commodity | 3.73 |
| OriginRail | 2.73 |
| EastWest | 2.44 |
| DestIWW | 2.43 |
| OriginIWW | 2.23 |
| EastEurope | 2.16 |

Another strategy to reduce multicollinearity is to merge correlated variables. Road and rail cost could be merged into a single variable representing the difference between these two costs. The estimated coefficient of the cost difference variable signifies the change in utility as the road cost increases relative to rail cost, or vice versa. Although the coefficient estimate will be more reliable due to reduced multicollinearity, the merging of the two variables means it is not possible to estimate the change in utility as only one cost increases or as both costs increase. When the separate road and rail cost variables are replaced by the cost difference variable, the variables inland waterway availability, rail service, and inland waterway service still have VIF values over 10. Due to the large number of multicollinear variables, eliminating or merging these would restrict model complexity and potentially result in less accurate predictions. Thus, all variables as they are listed in Table 3.4 are included as explanatory variables, with the understanding that the high VIF values may affect coefficient estimate reliability.

## 3.4 Data Preprocessing Conclusions

Although real data was used as much as possible when constructing the dataset used in this research, various estimations in and adjustments to the data as well as some confidential or otherwise unavailable data led to some deviations from known real-world values. The key areas where this dataset differs from reality are in the relative mode shares in the EU, absolute values of tonnage transported by each mode, and the road tonnage amounts transported between certain OD pairs, namely for domestic trips in numerous countries in the EU. Despite these limitations, this dataset may still be useful to compare different machine learning algorithms' performance and demonstrate how machine learning models could be used for mode choice modeling for freight policy studies. However, it is important to acknowledge that the best machine learning algorithm found for this dataset may not be the best algorithm for a less biased one. Lastly, the data limitations in the dataset should be addressed prior to its use in real-world policy applications.

Several explanatory variables were chosen and added to the dataset, based on a literature review and on data availability. The Phik correlations calculated for each variable pair showed high correlations

between many of the time, distance, and cost variables. VIF values were also calculated, and these indicated high multicollinearity among many variables. In order to estimate feature importance, which is a valuable element of model interpretation, the variables with very high Phik correlations were excluded from the set of explanatory variables to be used during model training. The remaining variables were kept as predictors, with the acknowledgement that the high VIF values for some of these may result in unreliable coefficient estimates and may make it more difficult to derive policy insights from the model results.

# 4 Methodology

In this chapter, Section 4.1 describes the machine learning algorithms selected for this research. Section 4.2 details the model training process for each algorithm, including the selection of hyperparameters and the use of explainability tools. Finally, Section 4.3 provides conclusions on the methods.

## 4.1 Method Selection and Description

Three machine learning algorithms were selected: logistic regression (LR), Random Forest (RF), and XGBoost (XGB). Logistic regression was chosen as a baseline of comparison to the two more advanced machine learning methods. Random Forest and XGBoost were chosen due to their high accuracy demonstrated in previous freight mode choice studies (Xu, et al., 2024) (Liu, et al., 2024). While other studies have found that neural networks and support vector machines can estimate behavioral outputs such as willingness to pay more accurately than Random Forest and XGBoost (Martin-Baos, Lopez-Gomez, Rodriguez-Benitez, Hillel, & Garcia-Rodenas, 2023), these models were not included in this research because deriving and comparing behavioral outputs is outside the scope of this research and, due to time constraints, only a limited number of algorithms could be trained. The following sections provide a description of the three selected algorithms.

### 4.1.1 Logistic Regression

Logistic Regression is a supervised machine learning method that uses the same mathematical formulation as a Random Utility Model (RUM) in Logit discrete choice models. The model estimates the coefficients that minimize the multinomial log-loss, or the negative log-likelihood of the observed choices. In multiclass classification problems, LR uses the SoftMax logistic function to compute the probabilities of the input data belonging to each class (Hillel, Bierlaire, Elshafie, & Jin, 2021). The difference between LR and the RUM approach is that in LR, regularization is applied automatically through either L1 (or lasso) regularization which adds a penalty based on the absolute value of the model's weights or L2 (or ridge) regularization, where the penalty is based on the square values of the model's weights. L2 regularization is often applied to address multicollinearity in logistic regression problems (Oluwadare, 2020). The amount of regularization is controlled by a hyperparameter called C, where a smaller C allows for a greater penalty on large weights, and a larger C means less penalties, or weaker regularization.

In 38 studies on travel mode choice models between 2020 and 2023, only two utilized LRs, compared to 10 using Artificial Neural Networks (ANNs) and 11 employing Random Forest/Decision Trees (Benjdiya, Rouky, Benmoussa, & Fri, 2023). This is in part because LR tends to offer lower accuracy than other more complex machine learning models, as it fails to capture nonlinear relationships between variables (Kashifi, Jamal, Kashefi, Almoshaogeh, & Rahman, 2022).

### 4.1.2 Random Forest

Random Forest is a combination machine learning algorithm used for classification and regression problems. It combines multiple decision trees, each trained on a random subset of the data, to make a

more robust and accurate prediction (Roßbach, 2018). Each tree makes a prediction based on the input data, and the final output of the model is the average of the predictions of all the decision trees (Benjdiya, Rouky, Benmoussa, & Fri, 2023). This approach helps to reduce overfitting, a common issue with decision trees. An illustration of a random forest is shown below in Figure 4.1, where each box contains a decision tree. The circles in the image represent decision nodes, starting with the root node at the top which is then split until the final leaf node is reached.

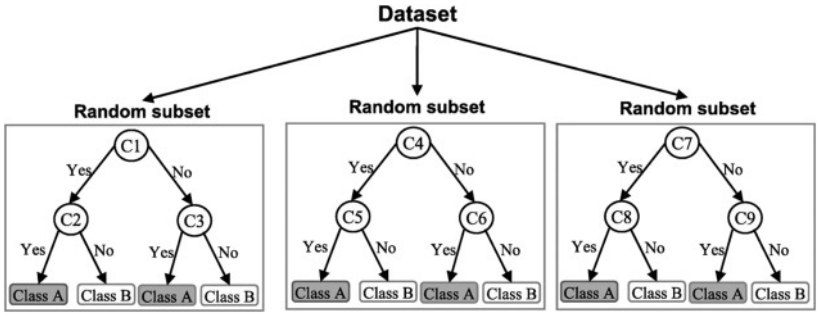


*Figure 4.1. Random Forest Structure (Dutta, Paul, & Kumar, 2021)*

RF has become one of the most popular machine learning techniques applied to mode choice problems in recent years due to its high accuracy, computational efficiency, and ability to handle large, complex datasets with minimal tuning. In a previous study comparing various machine learning methods for mode choice predictions, RF was determined to be both the most computationally efficient and accurate (Xu, et al., 2024). RFs are able to handle different types of input data, including both categorical (e.g. commodity type) and numerical (e.g. distance) variables as well as missing data.

RFs have several hyperparameters that must be set by the user; different combinations of hyperparameter values can be explored through tuning strategies to find the appropriate values for the given dataset. Some of the hyperparameters that are typically tuned in RF for classification problems and their default values in the scikit-learn library are described in Table 4.1.

*Table 4.1. Random Forest Default Hyperparameter Values*

| Hyperparameter | Description | Default value |
|---|---|---|
| n_estimators | Number of trees in the forest | 100 |
| max_depth | Maximum depth of each tree | None |
| min_samples_split | Minimum samples to split a node | 2 |
| min_samples_leaf | Minimum samples in each leaf | 1 |
| max_features | Number of features considered at each split | sqrt |
| bootstrap | Whether to bootstrap samples for each tree | True |

A greater number of trees (n_estimators) typically lead to better performance. However, the computation time of a model increases linearly with the number of trees, and at a certain number of trees, the gains in performance are very slight (Probst, Wright, & Boulesteix, 2019).

The maximum depth of each tree (max_depth) sets the maximum number of splits from a root node before stopping at a leaf node. Shallow trees with small depth are simpler and less likely to overfit. Maximum depth is also closely connected to the leaf node size hyperparameter, as a lower leaf node

size leads to deeper trees, and a higher leaf node size limits the depth of trees (Probst, Wright, & Boulesteix, 2019). For certain datasets, including larger ones with more noise variables, previous studies have shown that increasing the leaf or terminal node size (min_samples_leaf) can improve the model's performance, decreasing runtime (Probst, Wright, & Boulesteix, 2019).

Increasing the minimum samples to split a node (min_samples_split) reduces the creation of very small, specific branches, thus also potentially reducing overfitting. Nodes are split based on a split criterion which is Gini by default but can be set to entropy or log-loss instead.

The max_features hyperparameter determines how many features are considered at each decision node. The default value of sqrt means that number of features considered at each node is the square root of all features; these are randomly selected from all the features. Using fewer features can help to reduce overfitting. The maximum number of features can also be set to None, log2, an integer, or float value. Finally, bootstrapping means sampling with replacement, where random samples are drawn multiple times from the dataset, leading to some rows being chosen multiple times. Using bootstrapping can also reduce overfitting.

### 4.1.3 XGBoost

eXtreme Gradient Boosting (XGBoost) is a boosting tree machine learning algorithm that is widely used for classification tasks, including mode choice, due to its high accuracy, interpretability, flexibility, and scalability (Li, et al., 2024). It starts with a weak learner, or simple decision tree, and builds a new tree that learns from and reduces the residual errors in the weak one. Through each iteration, the model improves mistakes from previous trees, and the final prediction is a weighted total of all decision trees, weighted by the trees' predictive accuracy (Chen & Cheng, 2023). Bagging and boosting are two types of ensembled learning techniques. XGBoost is a boosting method as it learns from an ensemble of decision trees iteratively, meanwhile Random Forest is a bagging method because it constructs each tree independently and then combines them (Xu, et al., 2024). The two methods often perform similarly, although in some studies XGBoost models have a slightly higher accuracy than Random Forest (Zhang, Zhang, Liu, & Zhang, 2023) (Fatima, Hussain, Amir, Ahmed, & Aslam, 2023).

In Table 4.2 are some of the hyperparameters that are typically tuned in XGBoost models. These include hyperparameters tuned in a previous travel mode choice study (Chen & Cheng, 2023), as well as hyperparameters that are useful for controlling overfitting (DMLC XGBoost, n.d.). Two of these hyperparameters, number of trees and maximum depth, are the same as in Random Forest. The learning rate sets how much each tree contributes to the ensemble; a lower learning rate reduces the impact of each tree, slowing the learning process (DMLC XGBoost, n.d.) The larger the gamma, or min_split_loss, the more conservative the algorithm is. A larger min_child_weight means more weight is required in each leaf, which leads to simpler trees and less overfitting. Setting a lower subsample ratio means the model randomly samples a smaller amount of the total training data prior to growing trees. This introduces randomness to the model and helps to prevent overfitting. The max_delta_step is the maximum allowed weight for each tree; a lower maximum value makes the model more conservative, reducing how much the weights of each tree can change at a time. The default value of 0 means no limit is set on how much the weights can change.

*Table 4.2. XGBoost Default Hyperparameter Values*

| Hyperparameter | Description | Range | Default value |
|---|---|---|---|
| n_estimators | Number of trees in the forest | [0, ∞] | 100 |
| max_depth | Maximum depth of each tree | [0, ∞] | 6 |
| learning_rate (eta) | Step size shrinkage used to prevent overfitting | [0,1] | 0.3 |
| gamma (min_split_loss) | Minimum loss reduction required to make a further split on a leaf node of the tree | [0, ∞] | 0 |
| min_child_weight | Minimum sum of instance weight needed in a child | [0, ∞] | 1 |
| subsample | Subsample ratio of the training instances | [0,1] | 1 |
| max_delta_step | Maximum delta step allowed for each leaf output | [0, ∞] | 0 |

## 4.2 Model Training and Interpretation Process

For each algorithm, a model was trained using the variables in Table 3.4 from the data processing chapter as features. For each row in the data, the other mode costs for the alternatives not chosen are included to prevent data leakage from giving the model only the mode-dependent cost. The target, or y variable, is the transportation mode. The model training and interpretation workflow which is described in this section is also shown in Figure 4.2. The same features and training/test set split are used for all three algorithms, and the same CV folds are used during hyperparameter tuning.

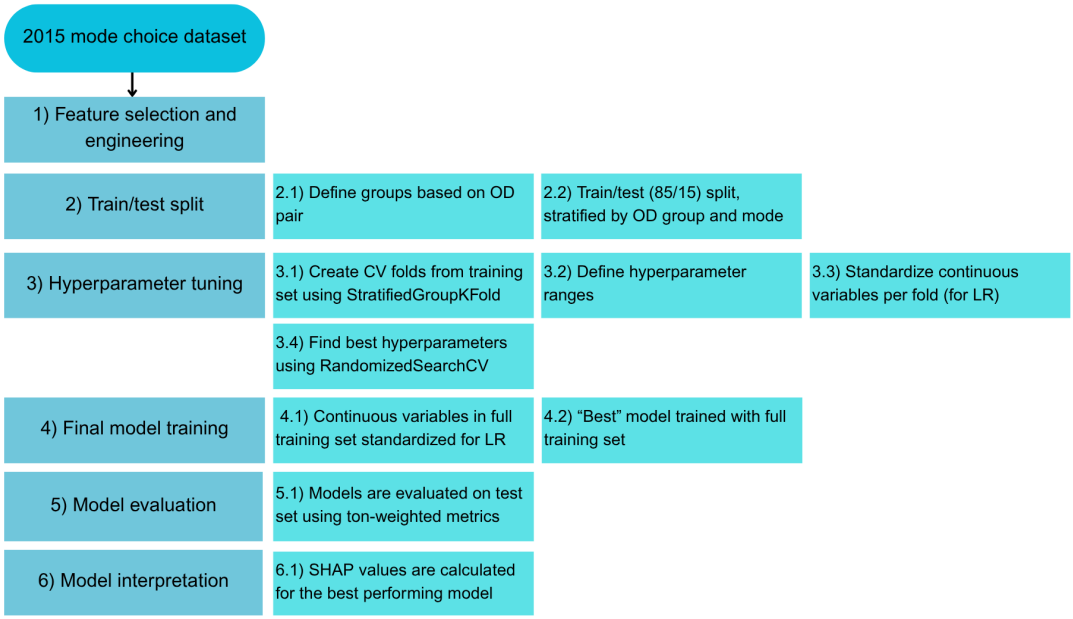


*Figure 4.2. Model Training and Interpretation Workflow*

The data is split into a training (85% of the data) and test set (15%), stratified by group and class. Groups are formed based on OD pairs, so all of the rows of data with the same origin and destination zone pair are placed in a single group. All data rows associated with one group are placed in the same subset of data during data splitting, either the training or test set. Splitting by group is important when

data rows are not independent, for instance when multiple samples are taken from the same patient (scikit-learn, n.d.). In this dataset, the rows associated with the same OD pair are not independent, as they have the same costs and distances. If some rows for the same OD pair were placed in the training set and others in the test set, the model would see the actual mode chosen for these rows during training and then would be tested with other similar rows. The model's predictive performance may thus be overestimated because some predictions may be based on the actual mode that the model saw during training for these rows, which is a form of data leakage. By stratifying by class as well, the distribution of modes by rows is mostly preserved in the training and test sets.

For logistic regression, the continuous variables of cost and terminal distance are standardized after the initial data splitting. These are not standardized for Random Forest and XGBoost.

To find the best hyperparameters for each algorithm for this dataset, RandomizedSearchCV is used. This is a cross-validation technique that does not require the creation of a separate validation set to test the hyperparameters, as it splits the training set in *k* smaller sets, or folds, trains the model on *k-1* folds, and holds the remaining part of the data for validation (scikit-learn, n.d.). The number of folds chosen is 5. These are created out of the 85% of the data that is allocated for training. In each iteration, four of the folds serve as the training data, and the fifth fold is used for validation. Thus, in each iteration, 68% of the total data is used for training during hyperparameter tuning and 17% for validation. The folds are created using StratifiedGroupKFold so that data rows with the same OD pair are kept in the same fold, and each fold maintains a similar distribution of modes based on the rows.

In random search, a range of values is pre-chosen for each hyperparameter, from which a random sample is taken to be tested. The number of iterations chosen is 20, so 20 combinations of hyperparameters are tested on each of the 5 folds. Different class weights for the minority classes rail and inland waterway were also tested in LR and RF. A class weight on rail for instance penalizes mistakes in rail predictions compared to road and inland waterway mistakes.

When finding the hyperparameters and training the final models, the tonnage weights in each row are applied as sample weights, increasing the importance of each data row proportionally to its transported tonnage. This leads the model to prioritize correctly predicting rows with higher tonnage amounts. The scorers for hyperparameter tuning, or the metric that is being optimized during the selection of hyperparameters, are the tonne-weighted log-loss and tonne-weighted rail and inland waterway recall. The final models are trained on the full training set.

As the models are trained with tonnage sample weights, they are also evaluated with tonne-weighted metrics. The final accuracy, precision, recall, F1-score, log-loss, and confusion matrices are all calculated per tonnes, rather than per rows. For instance, the precision for each mode is the percentage of correctly classified tonnes out of all the tonnes predicted for that mode, rather than the percentage of correctly classified rows. These evaluation metrics are calculated for the "best" model for each algorithm, or the one with the best hyperparameters found, on the test set which was not used during hyperparameter tuning. The log-loss of the training set and the test set are both calculated to assess potential overfitting, a much smaller training set log-loss compared to the test set being one indication of overfitting (Hawkins, 2004).

To explore model explainability, the Shapley Additive exPlanations (SHAP) values are computed for Random Forest and XGBoost. Shapley values explain the contribution of each feature to the model's output and can offer insights into the local (data instance-level) and global (model-level) importance of

each feature. SHAP assumes that features in the dataset are independent. SHAP can also be sensitive to outliers and noisy data; high correlations or noise in the data can lead to unreliable and misleading estimates of feature importance (Pan & Takefuji, 2025). For logistic regression, the estimated coefficients are also discussed in the Results chapter. The estimated coefficients for logistic regression and SHAP values for Random Forest and XGBoost are used to investigate whether the models' predictions and how the features are used to make these predictions align with what is expected from a theoretical perspective.

## 4.3 Methodology Conclusions

Logistic regression was chosen as a baseline machine learning method to compare to more complex algorithms. Random Forest and XGBoost were selected as previous studies have shown they offer the highest accuracy in freight mode choice prediction. These algorithms offer a range for which to compare not only predictive performance but also interpretability and other model characteristics. A limitation of this study is the number of models compared. There are many other machine learning algorithms for classification tasks, such as Artificial Neural Networks, Naïve Bayes, Support Vector Machine as well as other bagging and boosting decision tree methods which may offer other advantages not covered by the three models in this research. For instance, several previous studies have identified a trade-off between predictive accuracy and behavioral interpretability in machine learning algorithms; Random Forest and XGBoost often have higher accuracy than neural networks but produce less reasonable behavioral outputs such as elasticities and willingness to pay (Zhao, Yan, Yu, & van Hentenryck, 2020) (Martin-Baos, Lopez-Gomez, Rodriguez-Benitez, Hillel, & Garcia-Rodenas, 2023).

Finally, the model training and interpretation process was discussed, including the specific considerations made to split the data by group and class and weight the data samples and evaluation metrics by tonnage. The described model training and interpretation workflow could offer insights on how to handle aggregated freight flow data, potentially filling a research gap as previously only disaggregate-level datasets have been used to predict freight mode choice.

# 5 Results

Section 5.1 in this chapter presents the results of the logistic regression, Random Forest, and XGBoost models. Section 5.2 compares the results of the various machine learning algorithms, based on the criteria outlined in Section 2.4 in the literature review and against the current NEAC mode choice model.

## 5.1 Model Results

The test set accuracy, log-loss, and training/test set log-loss gap are presented in Table 5.1 below. The overall accuracies, or percentage of correctly classified tonnes in the dataset, are very similar across the different models, ranging from 89.1% in the logistic regression model to 92.1% in the XGBoost model. Similarly, the log-loss is highest in logistic regression and lowest in XGBoost, indicating that the latter fits the test set better. The high accuracies are largely due to the models' high accuracy in predicting road tonnes. Since the vast majority of tonnes are transported by road (79%), the overall predictive accuracy is high despite the models' comparatively poor ability to correctly identify tonnes transported by rail and inland waterway.

*Table 5.1. Test Accuracy and Log-Loss*

| | **Logistic regression** | **Random Forest** | **XGBoost** |
|---|---|---|---|
| Test accuracy | 0.891 | 0.915 | 0.921 |
| Test log-loss | 0.3490 | 0.2951 | 0.2454 |
| Train/test log-loss gap | 4.37% | 24.99% | 20.95% |

The log-loss gap is the test set log-loss minus the training set log-loss divided by the training set log-loss. While no specific thresholds were found in academic literature, some programming resources suggest that a gap larger than 5% is an indicator of overfitting (Ogbemi, M, 2023). Overfitting is problematic in this context because the models are intended to be used for predicting under changing conditions; if the models overfit, they will be less reliable in new, unseen scenarios. Based on the 5% threshold, the logistic regression model is not overfitting to the training set, whereas the RF and XGB models with a log-loss gap of 25% and 21% respectively are overfitting greatly.

The improved performance of the Random Forest and XGBoost models on the test set and the greater variance of their performance between the test and training sets compared to the logistic regression model reflect the bias-variance trade-off (DMLC XGBoost, n.d.). The RF and XGB models are better able to fit the test set and reduce bias but have a greater variance in performance across the different datasets, whereas the logistic regression model has higher bias but lower variance.

In all of the models, tonnes actually transported by inland waterway and rail are most mistaken for road tonnes. For road tonnes that are misclassified, these are most often classified as inland waterway tonnes; this is presented in the confusion matrices in Figure 5.1.

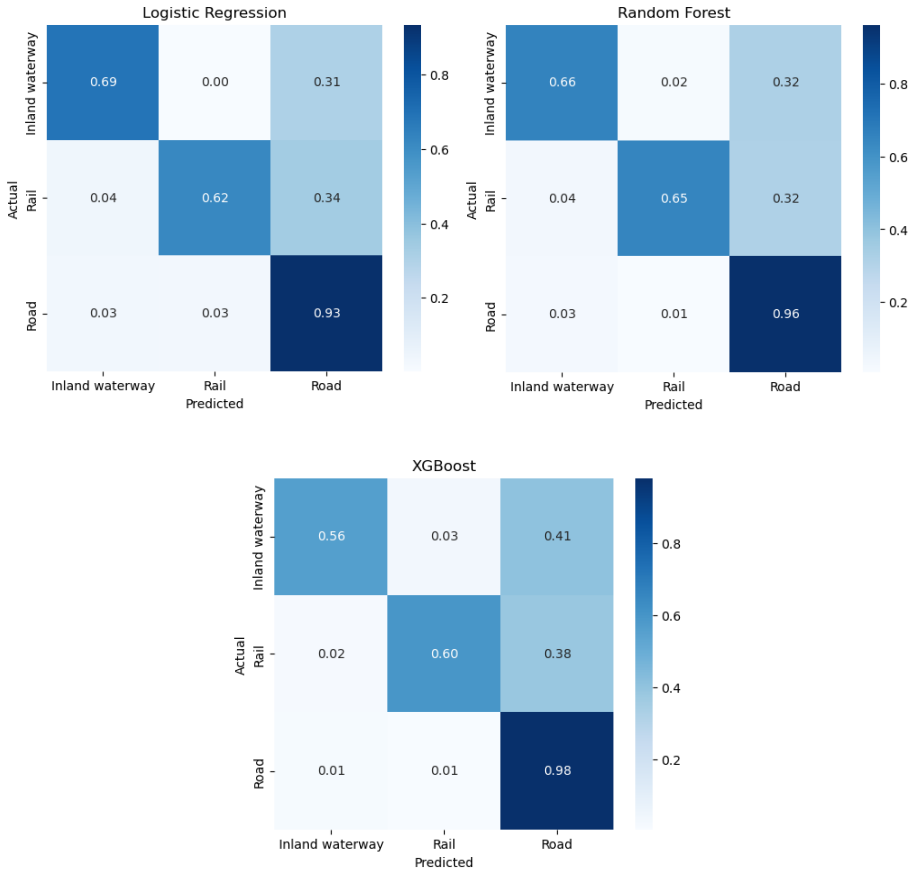


*Figure 5.1. Normalized Confusion Matrices*

Table 5.2 displays the precision and recall for each mode. Precision is the proportion of total tonnes predicted for a mode that were actually transported by that mode, or true positives over the sum of true positives and false positives. Recall is the proportion of all actual tonnes for a mode that the model correctly classified as that mode, or true positives over the sum of true positives and false negatives. The F1-score is the harmonic mean of precision and recall.

*Table 5.2. Classification Report*

| Mode | Model | Precision | Recall | F1-score | Support |
|---|---|---|---|---|---|
| Inland waterway | Logistic Regression | 0.52 | 0.69 | 0.59 | 35,582.00 |
| | Random Forest | 0.54 | 0.66 | 0.59 | |
| | XGBoost | 0.68 | 0.56 | 0.61 | |
| Rail | Logistic Regression | 0.69 | 0.62 | 0.65 | 65,200.87 |
| | Random Forest | 0.88 | 0.65 | 0.75 | |
| | XGBoost | 0.89 | 0.60 | 0.71 | |
| Road | Logistic Regression | 0.94 | 0.93 | 0.94 | 578,842.56 |
| | Random Forest | 0.95 | 0.96 | 0.95 | |
| | XGBoost | 0.93 | 0.98 | 0.96 | |

The road precision and recall is much higher than for the minority classes in all of the models. Compared to the logistic regression model, Random Forest improves the precision of the minority classes by 2 percentage points for inland waterway and 19 points for rail, while XGBoost improves precision by 16 and 20 points for inland waterway and rail respectively. The recall for inland waterway drops 3 and 13 points respectively for Random Forest and XGBoost compared to logistic regression. For rail, recall also drops 2 percentage points in XGBoost but increases by 3 in the Random Forest compared to the 0.62 recall in logistic regression. This largely demonstrates the rail-precision trade-off, where precision improves at the expense of recall, or vice versa. Based on the average F1-scores across inland waterway and rail, Random Forest has the best performance in predicting the minority classes with an average F1-score of 0.67, followed by XGBoost (0.66) and logistic regression (0.62).

Table 5.3 below presents the differences between actual and predicted mode shares in the test set for each model. Both logistic regression and Random Forest overpredict inland waterway and rail and underpredict road, with the latter model displaying greater differences from the actual mode shares. XGBoost has by far the smallest differences between actual and predicted mode shares.

*Table 5.3. Differences Between Actual and Predicted Mode Shares*

| | **Logistic Regression** | | **Random Forest** | | **XGBoost** | |
|---|---|---|---|---|---|---|
| | Difference | Relative error | Difference | Relative error | Difference | Relative error |
| Inland waterway | +2.79% | +53.24% | +3.43% | +65.46% | +0.50% | +9.54% |
| Rail | +4.78% | +49.84% | +4.68% | +48.80% | -0.05% | -0.52% |
| Road | -7.57% | -8.89% | -8.11% | -9.52% | -0.45% | -0.53% |

### 5.1.1 Logistic Regression Coefficients

As mentioned in Section 3.3.1, due to the high multicollinearity of some variables, the estimated coefficients may be unreliable. This is evident in several of the coefficients which have opposite signs from what might be expected, shown in Table 5.4: the ServiceRail coefficient for rail is negative, meaning that a higher rail service negatively affects rail utility. This could be due to interactions with other variables, such as rail cost, as both of these have high VIF values. Another possibility is that the rail service variable is capturing the effects of other variables that were not included in the model; other factors such as road infrastructure quality, trade flow characteristics, or physical barriers could be correlated with rail service and influence mode choice, leading to the counterintuitive negative coefficient.

The rail cost and OriginRail coefficients for inland waterway are both negative, meaning as the cost of rail and the distance to a rail terminal increases, the utility for inland waterway decreases. Despite potential multicollinearity, most coefficients have low standard errors and are statistically significant, with the exception of NST 5, 9, and 10 which are not significant.

The alternative-specific constants (ASCs) for rail and inland waterway are 6.18 and -8.65 respectively, with road as the base alternative, which are very different from the ASCs in the NEAC mode choice model which are on average -1.96 for rail and -1.89 for inland waterway across all commodity types. In the NEAC model, the dummy rail variables for East-West Europe and Eastern Europe are both positive,

meaning rail is favored compared to road and inland waterway. In the LR model, road is preferred in both East-West Europe and Eastern Europe over inland waterway and rail which both have negative coefficients.

*Table 5.4. LR Coefficients and Alternative-Specific Constants with Road as Base Alternative*

| | **Inland waterway** | | **Rail** | |
|---|---|---|---|---|
| | *Coefficient* | *Standard Error* | *Coefficient* | *Standard Error* |
| Alternative-specific constants | -8.654 | 0.197 *** | 6.177 | 0.121 *** |
| Road cost | 2.455 | 0.009 *** | 1.242 | 0.004 *** |
| Rail cost | -0.134 | 0.009 *** | -0.727 | 0.005 *** |
| IWW cost | -2.928 | 0.014 *** | 0.0927 | 0.002 *** |
| IWW available | 14.977 | 0.154 *** | -0.582 | 0.007 *** |
| NST 0 | 0.348 | 0.122 *** | -0.782 | 0.121 *** |
| NST 1 | -10.205 | 0.175 *** | -2.788 | 0.121 *** |
| NST 2 | 3.212 | 0.123 *** | 4.533 | 0.121 *** |
| NST 3 | 2.885 | 0.175 *** | 0.853 | 0.121 |
| NST 4 | 0.149 | 0.123 ** | 0.232 | 0.121 ** |
| NST 5 | -0.211 | 0.123 | 0.541 | 0.121 |
| NST 6 | 0.373 | 0.123 ** | 0.153 | 0.0121 ** |
| NST 7 | -5.994 | 0.196 *** | 3.599 | 0.121 *** |
| NST 8 | 0.122 | 0.123 ** | 0.031 | 0.121 ** |
| NST 9 | -0.705 | 0.122 | -0.731 | 0.121 *** |
| NST 10 | 1.427 | 0.124 *** | 0.465 | 0.122 |
| Origin IWW | -2.292 | 0.009 *** | 0.303 | 0.002 *** |
| Origin Rail | -0.564 | 0.003 *** | -0.202 | 0.002 *** |
| Dest IWW | -2.623 | 0.01 *** | 0.046 | 0.002 *** |
| East West | -4.892 | 0.032 *** | -3.06 | 0.014 *** |
| East Europe | -2.538 | 0.022 *** | -3.053 | 0.012 *** |
| Service Rail | -4.444 | 0.008 *** | -2.362 | 0.005 *** |
| Service Water | 0.345 | 0.07 *** | -0.267 | 0.003 *** |

Other coefficients are more in line with what is behaviorally realistic. The availability of inland waterway increases the utility of inland waterway, though this coefficient (14.98) is very high. The utility of both inland waterway and rail increases as the cost of road increases. Utility of rail decreases as the cost of rail and distance to rail terminal increases. As inland waterway costs increase, the utility of inland waterway decreases and the utility of rail increases. These are consistent with typical substitution patterns. Finally, a higher service quality of inland waterway increases inland waterway utility and decreases rail utility.

### 5.1.2 Random Forest and XGBoost SHAP Values

The beeswarm plots in Figure 5.2 show the SHAP values of the variables for each mode in the Random Forest model. The variables are listed along the y-axis with the strongest predictors at the top. Each

point represents one prediction, with the vertical spread representing the density of predictions with the same SHAP value. A negative SHAP value means the variable predicts a negative outcome, i.e. that the mode is not chosen. The colors indicate the feature values, with blue representing low values and red representing high values. For the binary variables (inland waterway availability, East Europe, East-West Europe), blue = 0 and red = 1.

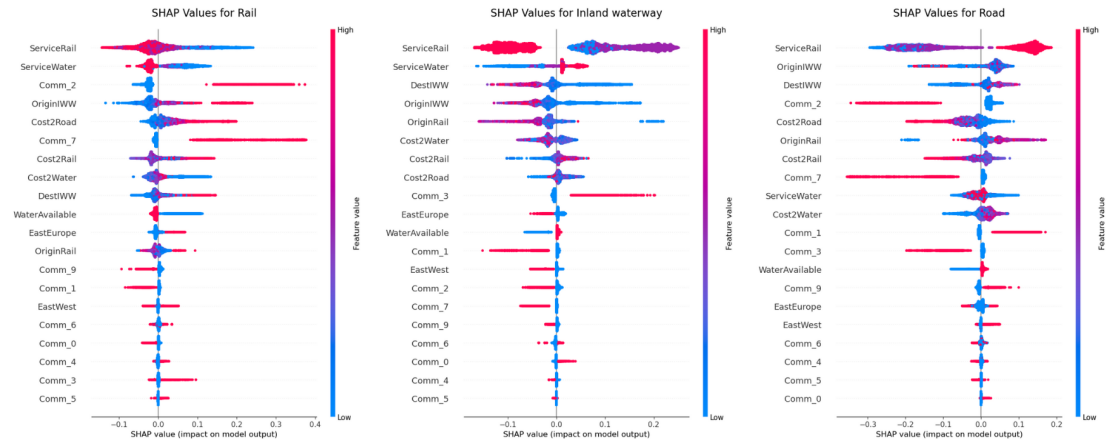


*Figure 5.2. RF Beeswarm Plots*

For all modes, the strongest predictor is rail service. The plots show that better quality rail service is associated with a lower probability of choosing rail and inland waterway. This is counterintuitive for rail, where a higher rail service might be expected to increase the probability of choosing rail. This contradiction was also evident in the Logit coefficients. A higher rail service is associated with positive road predictions. Based on these results, the model has learned that even when rail service is high, road transportation dominates, thus associating high rail service with a high probability of choosing road.

Another variable whose SHAP values appear counterintuitive is OriginRail, where shorter distances from the origin zone to the nearest rail terminal have a positive influence on choosing inland waterway, and longer distances increase the probability of choosing rail. This is the opposite pattern as what was found in the LR estimated coefficients, where a greater rail terminal distance negatively affected the utility of both inland waterway and rail. The rail cost SHAP value is also counterintuitive, with a higher cost associated with a higher probability of choosing rail. This could be because rail is preferred even for long distance trips where the costs are higher.

Many of the SHAP values for inland waterway are largely behaviorally consistent: lower inland waterway service reduces the probability of choosing inland waterway and increases the probability of choosing rail. Lower distances to the nearest inland waterway terminal, lower inland waterway costs, and inland waterway availability are positively associated with choosing inland waterway.

The XGBoost beeswarm plots with SHAP values in Figure 5.3 appear quite similar to those for the RF model. ServiceRail is still one of the strongest predictors for all modes, but destination zone distance to inland waterway terminal and Commodity 2 have replaced rail service as the biggest predictors for inland waterway and rail respectively. NST 2 has the highest proportion of tonnes transported by rail out of all the commodity types, as described in Figure 3.1 in the Data Preprocessing chapter. Both RF

and XGB learned this pattern and rely on NST 2 for rail predictions, although this predictor is much more important for XGB. Likewise, Commodity 1 is a predictor for road in both RF and XGB, but it is a much stronger predictor in the latter. Many of the other variables' mean absolute SHAP values are also much larger in the XGB model than in Random Forest.

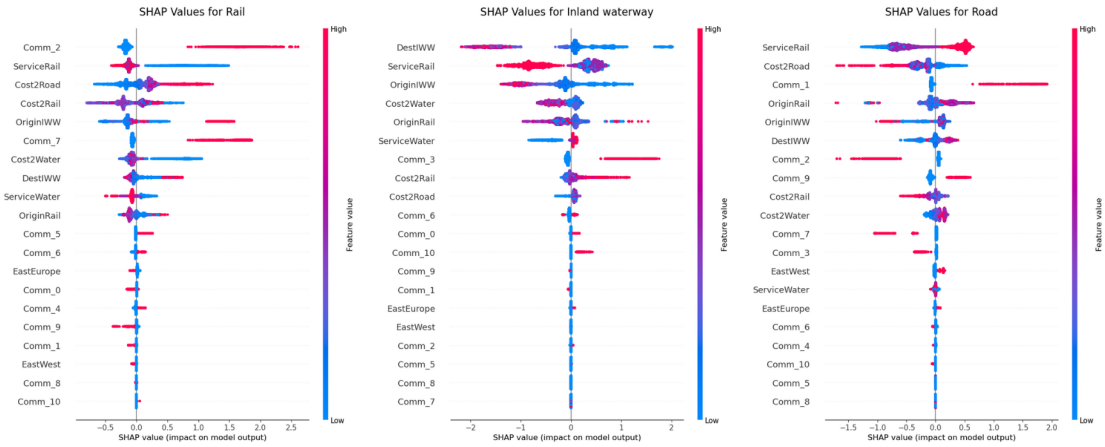


*Figure 5.3. XGB Beeswarm Plots*

The XGBoost beeswarm plots reveal some of the same counterintuitive relationships between rail service and rail, and rail service and road, as the Random Forest model. A lower rail service is used by the model to positively predict rail, and a higher rail service is associated with choosing road. The variable OriginRail for rail predictions is also still counterintuitive, with longer distances positively associated with choosing rail.

Other predictors appear more behaviorally realistic with XGB than with RF, such as OriginRail (for inland waterway predictions), rail cost, and road cost. Greater distances from the origin zone to nearest rail terminal are positively associated with choosing inland waterway, whereas in RF, short rail terminal distances predicted inland waterway. Lower rail costs predict choosing rail in the XGB model, compared to higher rail costs predicting rail selection in RF. ServiceWater was one of the top predictors for rail in the RF model; in XGB, this is a much smaller predictor, with rail cost and road cost as stronger predictors which intuitively should be strongly associated with choosing rail. Previous studies have also consistently demonstrated that cost is often considered by logistics decision-makers as the most important factor in mode choice (Tavasszy, Van de Kaa, & Liu, 2020). Likewise, inland waterway cost becomes a stronger predictor for inland waterway, and road cost for road, in the XGB model, which also appears more reasonable than the RF model where these are less important for predictions compared to other variables.

## 5.2 Model Comparison Against Evaluation Criteria

In this section, the various models are compared and evaluated against the evaluation criteria outlined in Section 2.4 in the literature review. These criteria are: predictive performance, interpretability, practicality, computation time, generalizability, robustness, and data efficiency. The current NEAC mode choice model will also be discussed in relation to these criteria. Since the NEAC MNL model was

not estimated as part of this research, it may not be as directly comparable for all evaluation criteria to the three machine learning models which were all trained on the same dataset.

### 5.2.1 Predictive Performance

XGBoost achieved the best predictive performance overall, with the highest accuracy, lowest log-loss, and the smallest differences between actual and predicted mode shares. Logistic regression performed only slightly worse in terms of accuracy than the more complex machine learning models; this could be because most of the relationships in the dataset are linear, and RF and XGB improve on the small number of data instances where nonlinearities are present. Since logistic regression has the same mathematical formulation as an MNL model, an MNL model estimated with this dataset would have the same accuracy as the LR model.

With the logistic regression and Random Forest models trained in this research, a trade-off between mode share accuracy and minority class performance was identified; adding class weights for the two minority classes improved their recall and precision but overestimated their mode shares. In a weighted Logit model predicting mode choice in the Rhine-Alpine corridor, the relative errors for predictions for inland waterway, rail, and road mode shares were between 0.28% and 1.16% (Ramos, et al., 2024). This shows that very accurate mode share predictions are possible to achieve with Logit models.

Logistic regression had the lowest train/test log-loss gap, indicating the least overfitting, while Random Forest achieved the highest average F1-score for the minority classes. In comparison, XGBoost better balanced predictive accuracy and realistic mode shares. Based on these results, the XGBoost model performs best in overall predictive performance compared to the other two machine learning models. Considering that logistic regression and Logit models have the same mathematical formulation, the XGBoost model is also better in terms of predictive performance than an MNL model would be with this dataset.

### 5.2.2 Interpretability

As discussed in Section 2.4 on evaluation criteria, there are several types of interpretability. The first is behavioral interpretability, where economic indicators such as willingness to pay and elasticities are derived from a model. As mentioned in the results section for logistic regression, the coefficients estimated in this model are not reliable due to high multicollinearities. For this reason, it is not recommended to use the coefficients to derive the behavioral insights that might normally be derived from Logit-based models. If behavioral interpretability is important, then an MNL or logistic regression model would be preferred over Random Forest, XGBoost, or other machine learning models. In this case, the variables should be adjusted or transformed to reduce multicollinearities, while taking precautions not to remove too much information from the model which could result in poor predictive performance.

The second type of interpretability relates to explanations about the impact that the variables have on a model's predictions, such as through feature importance analysis. Although there are several model-agnostic explainability tools, SHAP was used in this research to demonstrate how feature importance insights can be derived from a machine learning model for freight mode choice. The mean average SHAP values for each variable for predicting each mode were discussed for the Random Forest and XGBoost models in the results chapter. It is also possible to observe local feature importance (i.e., the impact that features have on a single observation), though these were not presented in the results.

The current NEAC MNL model is used not to derive behavioral insights but rather to observe the differences in mode shares between a base scenario and one or more other scenarios. The predicted mode shares per OD pair and commodity from the mode choice model are inputted into the traffic assignment model, with the Logit coefficients and alternative-specific constants remaining fixed across different scenarios. For this reason, the second type of interpretability is considered more relevant for this use case than behavioral interpretability.

SHAP is a model-agnostic tool and thus can be used for any machine learning model, though the SHAP results were not shown for the logistic regression model in this research. In previous freight mode choice studies, the variables with the greatest SHAP-based importance were shipment weight, shipment monetary value, shipping cost, distance, and travel time, with commodity type and industry type shown to be less important (Liu, et al., 2024) (Xu, et al., 2024). Many of the commodity types are less important than other variables to the models' predictions in this research, which is consistent with prior studies. Some differences were observed between the SHAP values in the Random Forest and XGBoost models, with the latter exhibiting more behaviorally realistic feature importance values. For example, in XGB, the generalized mode costs were ranked higher in terms of mean absolute SHAP values for their respective modes than in RF. This is more aligned with the known strong influence cost has on mode choice.

SHAP values can also be used to help validate a model or different scenarios in a model by observing whether the way the model uses each feature for its predictions aligns with domain knowledge. For instance, if the cost of rail is increased in a new scenario, then the SHAP values for rail cost for predicting rail might be expected to become more negative compared to the base scenario where the costs were lower. By already aligning more with behavioral expectations, this kind of validation might be easier with the XGBoost model than the Random Forest model.

### 5.2.3 Practicality

Practicality involves considering the usefulness of the prediction results for the model's end-users. In this case, the end-users are both the analysts who use the model to run different scenarios as well as the clients who commission Panteia to perform the analysis work. Practicality can be measured through more quantitative means, such as by assigning a penalty for prediction errors that are more impractical for end-users (Yong, Ma, & Palmqvist, 2025). In this section, practicality will be discussed more qualitatively in terms of how the models can be used in application.

One of the challenges with machine learning models is extrapolation. Machine learning models are not capable of predicting using features with values that go far outside the range of values in the training data (Gao, Yang, Zhang, Li, & Qu, 2021). This could potentially limit the types of projects that the models in this research could be used for. For instance, if one of the policy changes that is being evaluated (e.g. an increase in cost or change in a network) goes far beyond the range of values within the current dataset, an MNL model may be more appropriate. Likewise, machine learning models have not been used for predictions on new transportation alternatives or for forecasting over longer time periods of 10 or more years (van Cranenburgh, Wang, Vij, Pereira, & Walker, 2022), also because of their known issues with extrapolation. This could affect the practical use of a machine learning-based mode choice model for analysts, as it requires an additional step of considering the appropriateness of the model for a given use case and may require retraining the model with a wider range of values before analyzing different scenarios. In some cases, data with wider ranges may also not be available or realistic.

A second challenge with machine learning models is that it is not possible to determine causality. With Discrete Choice models, a causal relationship between policy changes and mode choice can be established based on theoretical assumptions. Although model explainability tools like SHAP can reveal the importance of different features to a machine learning model's predictions, they cannot be used to derive the effect features have on mode choice. SHAP values only reveal how the model uses each feature to make its predictions. For instance, if the SHAP values reveal that higher rail costs have negative SHAP values for predicting rail, it is not correct to assume that higher costs lead to lower rail use based on the SHAP values alone. As one of the purposes of a mode choice model like NEAC is to determine the appropriate policy to enact in order to produce the desired change in people's behavior, an MNL model appears more practical as it is able to more clearly describe the cause and effect of different policies. If a machine learning model is used to evaluate policy interventions, different language must be used beyond the more straightforward causal wording. The acceptability, trust, and understanding of such a model by policymakers has not been extensively studied.

Despite the higher predictive accuracy and other advantages gained from machine learning models, from a practicality standpoint, an MNL model provides results that may be more understandable to end-users. The applicability of a machine learning model should also be carefully considered on a project-by-project basis.

### 5.2.4 Computation Time

The training times for each model are shown in Table 5.5 below. It is assumed that the current NEAC mode choice model has a similar estimation time as the logistic regression model training time. The estimation time for an MNL model with this dataset would be much lower than all of the machine learning models as it does not include hyperparameter search time.

*Table 5.5. Models' Computation Time*

| | Logistic Regression | Random Forest | XGBoost |
|---|---|---|---|
| Hyperparameter search time (seconds) | 45.6 | 218.9 | 88.3 |
| Training time (seconds) | 3.8 | 2.2 | 2.1 |
| Total time (seconds) | 49.4 | 221.1 | 90.4 |

The parameter search for logistic regression takes less time than the other two models mainly because only two parameters are being searched, compared to 7 hyperparameters for both Random Forest and XGBoost. The hyperparameter search time is longer for Random Forest than XGBoost partly because of the max depth ranges: for Random Forest, a larger range of up to 10 depth is searched, whereas the tree depth was limited to 4 for XGBoost in order to reduce the high overfitting that was observed.

### 5.2.5 Generalizability

One way to assess a model's generalizability is through testing it on data from a different time period or geographical region than the data it was trained on. Generalizability to new time periods is more relevant for NEAC, as the mode choice model is used to estimate mode shifts over time given changes in costs, infrastructure, or other predictor variables. However, generalizability to different geographical regions is assessed and discussed in this section as this can be done with the data that has already been collected.

The models are trained on the full training dataset excluding rows originating or arriving in a certain country; these omitted data rows become the test set. This is repeated for all of the countries in the dataset, so that each country is removed from the training set and used as the test set data once for each algorithm. The hyperparameters are kept the same as described in the Results chapter for each algorithm.

Table 5.6 presents the average test set accuracy and log-loss across all of the model versions where one country was excluded during training but used as the test data. As seen in the table, the average accuracy for each algorithm is much lower than the 89-92% the models previously discussed. The log-losses are also much higher than before. Using the average scores, XGBoost performs the best in terms of accuracy and Random Forest on log-loss.

*Table 5.6. Average Accuracy and Log-Loss for Leave-One-Country-Out Test*

| | **Average Accuracy** | **Average Log-Loss** |
|---|---|---|
| Logistic Regression | 0.628 | 0.856 |
| Random Forest | 0.653 | 0.780 |
| XGBoost | 0.668 | 0.828 |

For countries with a high road share, the models tend to perform well on this unseen data. They perform much worse for countries with a larger inland waterway share. For countries with a high rail share, the results are mixed, with models achieving a high accuracy for some of these countries and a low accuracy for others. The full results for all countries are shown in Appendix A.

Based on these results, none of the models are especially generalizable to new, unseen countries, as performance drops significantly when tested on countries the models were not trained on. Although the average accuracy and log-loss for logistic regression is lower than for the other two models, logistic regression's performance is either equal to or slightly worse than the other models for most of the countries. For three countries (Bulgaria, Czech Republic, and Spain), the accuracy and log-loss of logistic regression is significantly better than the other two models. For this reason, logistic regression may be considered the most generalizable to new countries of the three models.

### 5.2.6 Robustness

Robustness is a model's ability to perform well even with errors, or noise, in the data. Two types of noise are class noise, where a row in the data has the wrong class, and attribute noise, where the values of an attribute are incorrect, missing, or unknown (Fabra-Boluda, 2024). A model can be tested for robustness by introducing noise into the training set and comparing the model's test set performance with its performance when trained on a non-noisy training set.

Noise is added to the generalized cost and terminal distance variables in the training set according to a Gaussian or normal distribution. The noise levels used are 5% increments between 5% and 30%. The standard deviation for each cost and terminal distance variable is calculated. The noise level times this standard deviation becomes the new standard deviation for the Gaussian noise distribution, where the mean is 0. For example, if the standard deviation of rail cost is €50, then with a noise level of 5%, the standard deviation of the Gaussian noise distribution is €2.50. With 30% noise, the standard deviation is €15. Noise is added to each value in the three cost and three terminal distance variables

according to this distribution. Any values that become negative when noise is added are set to zero. The same training/test set splitting and hyperparameters are used as before.

The graph in Figure 5.4 below shows the tonne-weighted log-loss for each of the algorithms with different noise levels. For each model, increasing the noise increases the log-loss. For logistic regression, this increase is much smaller compared to RF and XGB.

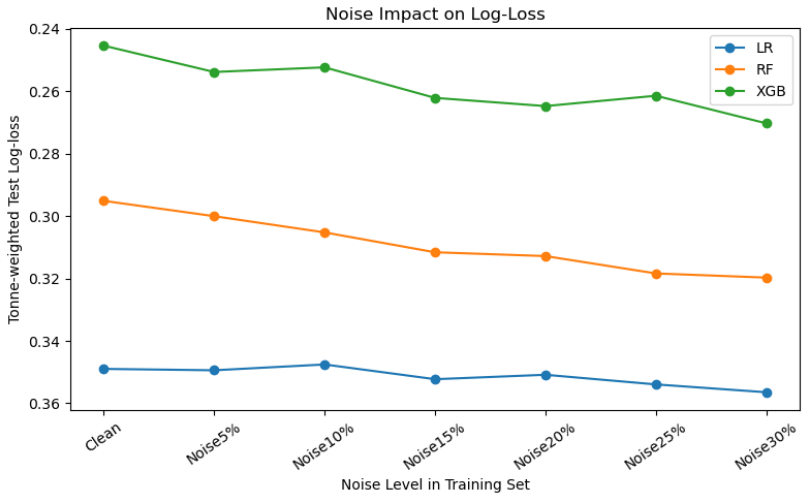


*Figure 5.4. Models' Log-Loss with Varying Noise Levels*

The accuracy of the models with varying noise levels is displayed in Table 5.7. For logistic regression, the accuracy steadily decreases as more noise is introduced. For Random Forest and XGBoost, the accuracy actually increases slightly with 20% and 10% noise respectively. Based on the accuracy and log-loss values with increasing noise, logistic regression is the most robust model, as the accuracy drops 1.6% and log-loss only 2% with 30% noise, compared to a similar drop in accuracy with RF and XGB but an 8.5% and 10.2% decrease in log-loss respectively. However, the overall log-loss and accuracy of RF and XGB are still better than LR despite this larger drop.

*Table 5.7. Models' Accuracy with Varying Noise Levels*

| | **No noise** | **5% Noise** | **10% Noise** | **15% Noise** | **20% Noise** | **25% Noise** | **30% Noise** |
|---|---|---|---|---|---|---|---|
| Logistic Regression | 0.891 | 0.888 | 0.885 | 0.885 | 0.884 | 0.880 | 0.877 |
| Random Forest | 0.915 | 0.915 | 0.899 | 0.897 | 0.916 | 0.910 | 0.906 |
| XGBoost | 0.921 | 0.916 | 0.919 | 0.916 | 0.912 | 0.913 | 0.910 |

### 5.2.7 Data Efficiency

Data efficiency refers to how well a model performs with small datasets. This is especially relevant for freight mode choice models as often there is not that much high quality data available. One way to measure data efficiency is through learning curves which show how a model's performance changes as the training set size increases. A data efficient model is able to perform well on a subset of the total training data. More complex, nonlinear models like decision trees have been found to perform better

with larger datasets, whereas simpler models like logistic regression tend to be better for small ones (Viering & Loog, 2023). The learning curves for the three machine learnings models in Figure 5.5 show the changes in training and test set accuracy and log-loss for different training set sizes.

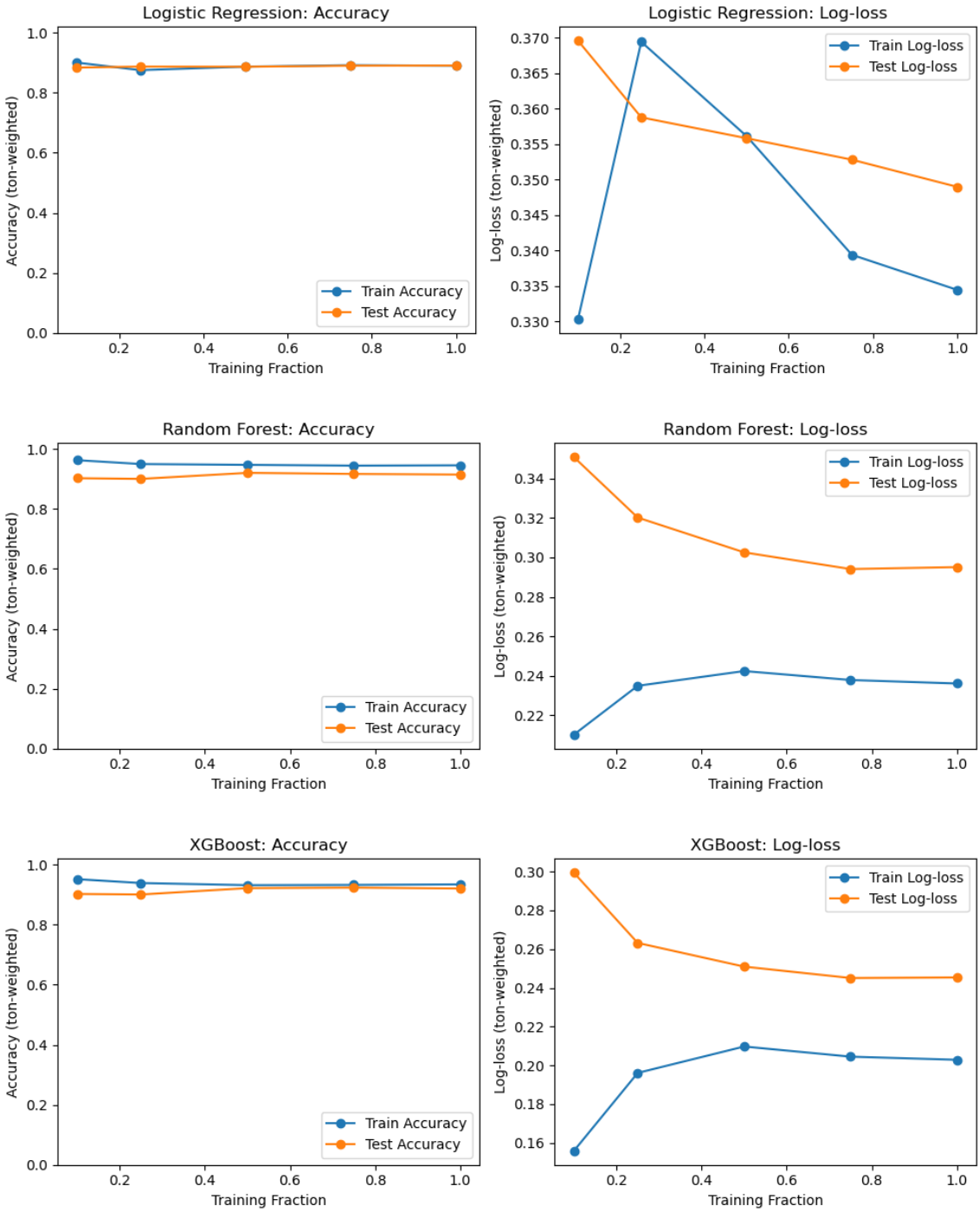


*Figure 5.5. Models' Learning Curves*

The learning curve plots reveal that the training and test set accuracy in the logistic regression model are very similar even with only 20% of the total training data. For Random Forest and XGBoost, the gap between the two accuracies is wider and closes as more training data is used. This aligns with

expectations about linear models such as logistic regression being more data efficient than more complex, nonlinear models such as Random Forest and XGBoost. In the log-loss learning curves, the test set log-loss starts higher with 20% of the training data for all of the algorithms and improves as more data is introduced. The gap between the training and test set log-losses closes with more data, although the size of this gap even with 100% of the training data for RF and XGB reveals the models' overfitting. The log loss improves the most with more training data in the XGBoost model and the least with logistic regression.

### 5.2.8 Model Comparison Summary

The logistic regression, Random Forest, XGBoost, and when possible, NEAC MNL models are compared using the seven criteria. Their relative performance is summarized in Table 5.8, where green (+) indicates the best performance, yellow indicates moderate, and red (-) indicates the worst.

*Table 5.8. Model Performance Against Evaluation Criteria*

| | NEAC MNL | Logistic Regression | Random Forest | XGBoost |
|---|---|---|---|---|
| Predictive performance | | | - | + |
| Interpretability | + | | - | |
| Practicality | + | | | |
| Computation time | + | + | - | |
| Generalizability | / | + | | |
| Robustness | / | + | | - |
| Data efficiency | / | + | | - |

The XGBoost model performed the best in terms of predictive performance, with the highest overall accuracy and smallest differences between actual and predicted mode shares, while maintaining balanced minority class precision and recall. The current NEAC MNL model is superior in terms of interpretability and the related practicality, although SHAP values can provide some level of interpretability for machine learning models. Between RF and XGB, the latter produced SHAP-based feature importance values that were more consistent with behavioral expectations.

The logistic regression model was shown to perform the best in the remaining criteria. LR had the lowest computation time and highest data efficiency compared to the other two machine learning models. It also had the best results in the generalizability and robustness tests. RF performed worse than one or both of the other machine learning models on all of the criteria. Given the high importance of predictive performance as well as XGBoost's close results to logistic regression in computation time and generalizability, XGBoost is considered the overall strongest machine learning model for this dataset and use case.

# 6 Discussion

In this research, three machine learning models were trained on a 2015 European freight flows dataset and evaluated based on various criteria including predictive performance and interpretability. In this chapter, the results of the models, namely the accuracy, precision, and recall, will be discussed in comparison to previous studies. Next, some of the limitations of this research are described. Finally, the application of the models used in this study is considered, including the types of projects where these models may be useful.

All of the machine learning models trained in this research had similar predictive accuracies, ranging from 89.1% in the logistic regression model to 91.5% and 92.1% for Random Forest and XGBoost respectively. In previous studies comparing RF and XGB for predicting freight mode choice, the accuracies of these models were within a few percentage points of each other, with RF performing better than XGB (Uddin, Anowar, & Eluru, 2021) (Xu, et al., 2024). In studies that also used logistic regression and/or MNL, the accuracies of these linear models were much lower than the more complex models they were compared to. In the Uddin et al., 2021 study, Random Forest accuracy was 75.4%, while the MNL model had an accuracy of 42%. Similarly, RF and XGB achieved accuracies of 91% and 87% respectively in a 2024 study using US commodity flow data, and the baseline MNL model estimated had an accuracy of 73% (Xu, et al., 2024). In another study, the RF accuracy was the highest with 72.9%, followed by XGB (71.3%) and LR (55.9%) (Liu, et al., 2024). Although the better performance of the more complex RF and XGB models in this study compared to the LR model aligns with previous research, the differences in accuracy between the various models is much smaller than in other studies.

One reason for these smaller differences between LR and RF/XGB could be that most of the relationships in the dataset are linear and thus capable of being correctly captured by logistic regression. In this case, the small gains in accuracy with RF and XGB could be from the few areas in the data where there are nonlinear interactions. Secondly, previous studies used disaggregate shipment-level data, whereas this research uses aggregated data of goods transported between regional zones. It is possible that with aggregation, some detailed information and variation in the data was lost where more complex models would have had a greater advantage over a linear model. Additionally, due to the initial very high overfitting observed in RF and XGB, the range of hyperparameters included in the hyperparameter search was reduced. These limited ranges constrained the complexities of the models in order to control overfitting. The moderately simple nature of the RF and XGB models thus further shrinks the differences in performance between these models and LR.

Each of the models is much better at predicting road correctly than the two minority classes, rail and inland waterway. This is a common occurrence with imbalanced datasets, where the model learns to predict the majority class well and fails to accurately predict minority classes (Abdelhamid & Desai, 2024). However, this is not always the case in other machine learning models predicting mode choice. For instance, in the US commodity flow data, air transportation has a mode share of only 0.7%, but in a study comparing RF, XGBoost, and CatBoost, the precision and recall for air were both 1.0 (Xu, et al., 2024). In a study predicting passenger mode choice with different machine learning algorithms, the F1-score for buses was 0.84, despite having a mode share of only 3% based on a household travel

survey dataset in China (Zhang, Zhang, Liu, & Zhang, 2023). Taxis, bike sharing, and subways were also predicted poorly, and this was attributed to the class imbalance and limited amount of data for these modes. In the models trained in this research, adding a class weight for inland waterway and rail in the LR and RF models improved their precision and recall, but the highest F1-scores were still only 0.59 and 0.75 respectively, compared to road which had a high F1-score of 0.95.

Besides a class imbalance in the dataset, other data quality characteristics such as data completeness and feature accuracy have been shown to have a high effect on machine learning model performance (Mohammed, et al., 2025). Data completeness is defined as an absence of missing or unknown values. Feature accuracy is the extent to which feature values are equal to their "respective ground truth values" (Mohammed, et al., 2025). In this dataset, there are some data rows with an unknown commodity type (NST 10) and some possible errors in calculating some of the feature values, including generalized costs and terminal distances; these may be lowering data completeness and feature accuracy which could also be negatively affecting the model's performance. Due to limited time, a more extensive data validation was not undertaken to address these data quality issues.

Another limitation is that the comparison of models across the evaluation criteria was conducted qualitatively. This assessment could be strengthened by, for instance, applying a weight to each criterion based on its importance for freight mode choice models used in policy analysis. As these weights were not found in existing literature, they could be derived in future studies from interviews with policymakers or domain experts. Additionally, gains in predictive accuracy have been achieved in previous studies by using ensemble learning techniques such as stacking or voting methods which combine the results from multiple algorithms (Liu, et al., 2024), though this research was limited to only single-model approaches.

The machine learning models in this research, as well as the NEAC MNL mode choice model, are suitable for regional analyses. As the machine learning model with the strongest overall performance, the XGBoost model is considered the most suitable for use in the NEAC framework. The types of projects where the XGBoost model could be used may differ from those of the current NEAC MNL model. Firstly, the dataset used in the MNL model contains NUTS 3 zone information, whereas the dataset for training the machine learning models uses the larger NUTS 2 zones. For this reason, the machine learning models are not able to predict mode choice on a more local level with the dataset as it was constructed. Secondly, as mentioned previously, machine learning models do not generalize well to new scenarios with data values far outside the range present in the training data. Thus, the current NEAC MNL model may be better suited for policy analyses where the proposed changes are very different from the current situation. Lastly, the XGBoost model has a recall of 0.56 and 0.60 for inland waterway and rail respectively, and a precision score of 0.68 for inland waterway. Therefore, it should be considered whether these scores are acceptable depending on the purpose and scope of the scenario analysis.

# 7 Conclusion

This chapter summarizes the main contributions of this research and answers the research questions. The main research question was on the role machine learning-based approaches can play in freight mode choice modeling for policy analysis. To answer this question, three machine learning models were trained using a 2015 aggregated freight flow dataset. These models were evaluated and compared to an MNL mode choice model in the NEAC framework using seven criteria to evaluate machine learning models for freight transport policy applications. Previous studies have only used disaggregate data to predict freight mode choice and have not considered the models' application to real-world policy analysis. Based on the results in this research, using aggregated data can provide high predictive accuracy, although the models struggle more with predicting the minority classes of rail and inland waterway compared to road. Using the NEAC model as a case study, the advantages and disadvantages of machine learning models compared to Logit models were considered. Depending on the policy analysis case, SHAP values can provide adequate interpretability for machine learning models.

The main contributions of this study are thus: 1) a demonstration of the predictive results that can be achieved using aggregated freight flow data, even when the data has some data quality issues, 2) a workflow for how to train and evaluate machine learning models using aggregated freight flow data, including using tonnes as sample weights and splitting the data into training/test sets stratified by OD pair and mode, 3) an exploration of the types of explanatory variables that can be useful with aggregated freight data, both through a literature review and through the inclusion of some of these in the trained models, and 4) a discussion of how SHAP values can be used to help validate and explain machine learning models intended for use in freight policy analyses. In the remainder of this chapter, the five sub questions will be addressed and answered.

1. **What are the criteria a machine learning mode choice model should meet to be suitable for freight transport policy analysis?**

Through a literature review, seven criteria were identified to evaluate machine learning models used in freight transport policy analysis. These criteria are predictive performance, interpretability, computation time, data efficiency, generalizability, robustness, and practicality. The three models trained in this research were compared to each other and to the current NEAC MNL model using these seven criteria.

2. **Which machine learning methods are most suitable for modeling freight mode choice?**

Previous studies have compared the predictive performance of various machine learning classification models for mode choice. These models include logistic regression, Naïve-Bayes, K-Nearest Neighbors, Support Vector Machine, Artificial Neural Networks, decision trees, Random Forest, and various gradient boosting algorithms such as XGBoost and CatBoost. In many studies, Random Forest and XGBoost have been shown to produce the highest accuracies for mode choice modeling. This is supported by this research, as Random Forest and XGBoost had higher accuracies than the logistic regression model.

3. **What additional explanatory variables and external datasets can enhance model performance?**

The explanatory variables used in the models were included based on their known influence on mode choice decisions as well as the availability of the data. Several highly correlated variables were excluded from the models in order to help with model interpretation. The final set of variables covers mode-specific shipping characteristics (i.e., cost), commodity characteristics (i.e., commodity type), infrastructure characteristics (i.e., availability of inland waterway, distance to terminals, rail and inland waterway service level), and other regional characteristics (i.e., whether the origin and destination are in Eastern Europe or Western Europe). These variables provided sufficient information to the models which resulted in high overall predictive accuracy. However, more mode-specific variables, specifically for rail and inland waterway, would help the models to distinguish these minority classes from the majority class of road.

4. **How does the performance of machine learning models compare to that of an MNL model?**

This question is answered in the Model Comparison section of the Results chapter. The three machine learning models and current NEAC MNL model were compared as mentioned using the seven criteria described in the literature review. XGBoost performed the best in predictive performance. The NEAC MNL model is the most interpretable and practical. Logistic regression performed the best in the other remaining criteria: computation time, data efficiency, generalizability, and robustness. Random Forest did not perform better in any of the criteria than the other models. Due to the high importance of predictive performance and XGBoost's similar results to logistic regression for some of the other criteria, XGBoost was considered the best machine learning model.

5. **Based on the results, should a machine learning mode choice model be incorporated into the NEAC framework, and if so, under what conditions?**

The machine learning models trained in this research are suitable for scenario analyses on how freight transport policies could affect changes in mode shares for large regional areas in the European Union. Ideally, the time frame of these changes would be short- to medium-term. As mentioned in the Results and Discussion chapters, these models cannot determine causality. Depending on the specific policy analysis required, the machine learning models could be preferred over the MNL model due to the former's higher predictive accuracy. However, since the MNL model is more interpretable, it cannot be completely substituted by the machine learning models. In the following chapter, additional details are provided on how the XGBoost model could be incorporated into the NEAC framework.

# 8 Recommendations

In this chapter, recommendations are provided both for areas of future research and for the company on the implementation of the models in this research.

### 8.1.1 Future Research

A limitation of this study is that only three machine learning models were trained and assessed. In future studies, other algorithms could be compared in order to potentially improve the current accuracy achieved with the three algorithms used in this research. In addition to testing different algorithms, using ensembled learning techniques where two or more models are combined may also improve predictive performance for machine learning freight mode choice models using aggregated data.

Secondly, adding other explanatory variables to the models and using class imbalance techniques could help improve minority class recall. Adding other variables would give the models more information to help differentiate rail and inland waterway from road. Based on the literature review, other variables that have been shown to influence freight mode choice include physical locations of shippers and buyers and mode characteristics such as flexibility, safety, and security. Including other country-specific variables may also help the models to identify mode share patterns by country, potentially improving predictive accuracy as well. To overcome the lower predictive accuracy with minority classes, testing different class imbalance techniques such as SMOTE may also help to improve rail and inland waterway accuracy.

Due to the MNL models' greater interpretability and practicality for a greater variety of freight transport policy studies, it cannot be completely replaced by the machine learning models used in this research. Depending on the specific analysis or project, one of the two methods may be preferred over the other. For this reason, another area of future research is in hybrid models, where the two methods could be combined into one single model that could be used for all types of projects and would combine the benefits from both methods.

Lastly, future studies could seek input from policymakers and additional domain experts on the relative importance of the evaluation criteria described in this research, as well as on the perceived trustworthiness and acceptability of machine learning models for different types of freight policy applications.

### 8.1.2 Company Recommendations

Out of the three machine learning models in this study, the XGBoost model is the most suitable for implementation in the NEAC framework due to its higher predictive performance and better or similar performance than the other models in the other criteria. This model is recommended as a complement, rather than a substitute, of the current NEAC MNL model. Section 5.2.3 and Chapter 6 contain additional details about the types of projects that the XGBoost model may and may not be suitable for.

Prior to using the model for policy analysis, the 2015 freight flows dataset should be adjusted to fill in missing OD data using estimates and checked for errors in costs and other variable calculations. After

these adjustments, the model will have to be retrained using the same workflow as outlined in Section 4.2. The ranges for the hyperparameters may also have to be changed.

XGBoost performed the best out of the three models on this dataset, but this does not guarantee that it will perform the best on a validated, less biased dataset. If the model's performance on an adjusted dataset is much worse than the performance presented in this report, it is recommended to use logistic regression, Random Forest, or another algorithm not trained as part of this research.

Once the performance of XGBoost has been validated using the workflow from this report, the model can be retrained using 100% of the training data. To estimate the changes in mode shares based on different scenarios, the values of the features can be adjusted and the mode shares between the base scenario and other scenarios compared. Changes can only be incorporated into features that already exist in the trained data. Therefore, in order to introduce a new policy change, this must be done on one of the existing features used in this research, or a new feature must be added to the full dataset and the model retrained prior to scenario analysis. SHAP can be used to help validate the predictions in the base and new scenarios. Sections 5.1.2 and 5.2.2 contain more information on interpreting predictions using SHAP.

# 9 Bibliography


Abdelhamid, M., & Desai, A. (2024). *Balancing the Scales: A Comprehensive Study on Tackling Class Imbalance in Binary Classification.* Ready Tensor, Inc.

Aboutaleb, Y., Danaf, M., Xie, Y., & Ben-Akiva, M. (2021). Discrete Choice Analysis with Machine Learning Capabilities. *arXiv:2101.10261*.

Ahmed, U., & Roorda, M. (2022). Modeling Freight Vehicle Type Choice using Machine Learning and Discrete Choice Models. *Transportation Research Record*, 541-552.

Baak, M., Koopman, R., Snoek, H., & Klous, S. (2020). A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Computational Statistics & Data Analysis*.

Benjdiya, O., Rouky, N., Benmoussa, O., & Fri, M. (2023). On the use of machine learning techniques and discrete choice models in mode choice analysis. *Scientific Journal of Logistics*, 331-345.

Chen, H., & Cheng, Y. (2023). Travel Mode Choice Prediction Using Imbalanced Machine Learning. *IEEE Transactions on Intelligent Transportation Systems*, 3795-3808.

DMLC XGBoost. (n.d.). *DMLC XGBoost*. Retrieved from XGBoost Parameters: https://xgboost.readthedocs.io/en/stable/parameter.html

DMLC XGBoost. (n.d.). *Notes on Parameter Tuning*. Retrieved from DMLC XGBoost: https://xgboost.readthedocs.io/en/stable/tutorials/param_tuning.html

Duranton, S., Audier, A., Hazan, J., Langhorn, M., & Gauche, V. (2017, April 18). *The 2017 European Railway Performance Index*. Retrieved from Boston Consulting Group: https://www.bcg.com/publications/2017/transportation-travel-tourism-2017-european-railway-performance-index

European Commission. (2017). *An Overview of the EU Road Transport Market in 2015.* European Commission.

European Commission. (2024). *Transport in the European Union: Current trends and issues.* Brussels: European Commission.

Eurostat. (2020, April 25). *77% of inland freight transported by road in 2020.* Retrieved from Eurostat: https://ec.europa.eu/eurostat/web/products-eurostat/news/-/ddn-20220425-2

Eurostat. (2022). *Key figures on European transport.* Luxembourg: Publications Office of the European Union.

Eurostat. (2024, November 13). *Inland waterway transport by type of cargo and country/region of loading and unloading (iww_go_atycafl).* Retrieved from Eurostat: https://ec.europa.eu/eurostat/databrowser/view/iww_go_atycafl/default/table?lang=en

Eurostat. (2024, July 4). *Railway transport - national and international railway goods transport by loading/unloading NUTS 2 region (tran_r_rago).* Retrieved from Eurostat: https://ec.europa.eu/eurostat/databrowser/view/tran_r_rago/default/table?lang=en

Eurostat. (2024, July 6). *Road freight transport by region of loading (t, tkm, journeys) - annual data (road_go_ta_rl).* Retrieved from Eurostat: https://ec.europa.eu/eurostat/databrowser/view/road_go_ta_rl/default/table

Eurostat. (2024, August 6). *Road freight transport by region of unloading (t, tkm, journeys) - annual data (road_go_ta_ru).* Retrieved from Eurostat: https://ec.europa.eu/eurostat/databrowser/view/road_go_ta_ru/default/table?lang=en

Eurostat. (2024, August 6). *Road freight transport by type of operation and type of transport (t, tkm, vehicle-km) - annual data.* Retrieved from Eurostat: https://ec.europa.eu/eurostat/databrowser/view/ROAD_GO_TA_TOTT/default/table?lang=en

Eurostat. (2025, May 27). *Goods transported.* Retrieved from Eurostat: https://ec.europa.eu/eurostat/databrowser/view/RAIL_GO_TOTAL/default/table?lang=en

Eurostat. (2025, April 16). *Modal split of inland freight transport (tran_hv_frmod).* Retrieved from Eurostat: https://ec.europa.eu/eurostat/databrowser/view/tran_hv_frmod__custom_16939825/default/table?lang=en

Eurostat. (2025, May 16). *Transport by type of good (from 2007 onwards with NST2007).* Retrieved from Eurostat: https://ec.europa.eu/eurostat/databrowser/view/iww_go_atygo/default/table?lang=en

Eurostat. (n.d.). *NUTS - Nomenclacture of territorial units for statistics*. Retrieved from Eurostat: https://ec.europa.eu/eurostat/web/nuts

Fabra-Boluda, R. (2024). Unveiling the robustness of machine learning families. *Machine Learning Science and Technology*.

Fatima, S., Hussain, A., Amir, S., Ahmed, S., & Aslam, S. (2023). XGBoost and Random Forest Algorithms: An In-Depth Analysis. *Pakistan Journal of Scientific Research*, 26-31.

Freiesleben, T, & Grote, T. (2023). Beyond generalization: a theory of robustness in machine learning. *Synthese*.

Gao, K., Yang, Y., Zhang, T., Li, A., & Qu, X. (2021). Extrapolation-enhanced model for travel decision making: An ensemble machine learning approach considering behavioral theory. *Knowledge-Based Systems*.

Han, Y., Camara Pereira, F., Ben-Akiva, M., & Zegras, C. (2022). A neural-embedded discrete choice model: Learning taste representation with strengthened interpretability. *Transportation Research Part B: Methodological*, 166-186.

Hawkins, D. (2004). The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.*, 1-12.

Hillel, T., Bierlaire, M., Elshafie, M., & Jin, Y. (2021). A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of Choice Modelling*.

Jensen, A., Thorhauge, M., de Jong, G., Rich, J., Dekker, T., Johnson, D., Ojeda Cabral, M., Bates, J., & Nielsen, O. (2019). A disaggregate freight transport chain choice model for Europe. *Transportation Research Part E: Logistics and Transportation Review*, 43-62.

Kashifi, M., Jamal, A., Kashefi, M., Almoshaogeh, M., & Rahman, S. (2022). Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. *Travel Behaviour and Society*, 279-296.

Koppelman, F. (1981). Non-linear utility functions in models of travel choice behavior. *Transportation*, 127-146.

Leest, E., Duijnisveld, P., & Hilferink, P. (2006). *Update of the NEAC Modal-Split Model.* Association for European Transport and contributors.

Li, X., Shi, L., Shi, Y., Tang, J., Zhao, P., Wang, Y., & Chen, J. (2024). Exploring interactive and nonlinear effects of key factors on intercity travel mode choice using XGBoost. *Applied Geography*.

Liu, D., Lim, H., Uddin, M., Liu, Y., Han, L., & Hwang, H. (2024). Improving the accuracy of freight mode choice models: A case study using the 2017 CFS PUF data set and ensemble learning techniques. *Expert Systems with Application*.

Martin-Baos, J., Lopez-Gomez, J., Rodriguez-Benitez, L., Hillel, T., & Garcia-Rodenas, R. (2023). A prediction and behavioural analysis of machine learning methods for modelling travel mode choice. *Transportation Research Part C*.

Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., & Naumann, F. (2025). *The Effects of Data Quality on Machine Learning Performance on Tabular Data.* Hasso Plattner Institute.

Newton, S., Kawabata, Y., & Smith, R. (2015). *NEAC-10 Model Description 2015.* Zoetermeer: Panteia/NEA.

Ogbemi, M. (2023, October 16). *What is Overfitting in Machine Learning?* Retrieved from freecodecamp: https://www.freecodecamp.org/news/what-is-overfitting-machine-learning/#:~:text=The%20accuracy%20gap%20is%20a,what%20you%20should%20look%20for.

Oluwadare, O. (2020). A Simulation Study on the Performance of Bayesian and L2 Regularization Methods in Multicollinearity Problem. *Anale Seria Informatica*.

Pan, H., & Takefuji, Y. (2025). Enhancing feature importance analysis with Spearman's correlation with p-values: Recommendations for improving PHLF prediction. *European Journal of Surgical Oncology*.

Probst, P., Wright, M., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining Knowledge Discovery*.

Ramos, C., Burgess, A., Van der Geest, W., Hendriks, I., Van Hassel, E., Shobayo, P., Samuel, L., Nicolet, A., Atasoy, B., van Doorser, C., Bijlsma, R., & Hofman, P. (2024). *Novel inland waterway transport concepts for moving freight effectively: D2.7 assessment of future scenarios.* NOVIMOVE.

Roßbach, P. (2018). *Neural Networks vs. Random Forests - Does it always have to be Deep Learning?* Retrieved from Frankfurt School Blog: https://blog.frankfurt-school.de/neural-networks-vs-random-forests-does-it-always-have-to-be-deep-learning/

Samimi, A., Kawamura, K., & Mohammadian, A. (2011). A behavioral analysis of freight mode choice decisions. *Transportation Planning and Technology*, 857-869.

scikit-learn. (n.d.). *3.1 Cross-validation: evaluating estimator performance*. Retrieved from scikit-learn: https://scikit-learn.org/stable/modules/cross_validation.html#

Sifringer, B., Lurkin, V., & Alahi, A. (2020). Enhancing discrete choice models with representation learning. *Transportation Research. Part B: Methodological*, 236-261.

Statistics Netherlands (CBS). (n.d.). *Eastern and Western European Countries*. Retrieved from Statistics Netherlands: https://www.cbs.nl/en-gb/news/2025/05/freight-transport-by-eastern-european-lorries-increased-sharply/eastern-and-western-european-countries

Tavasszy, L., Van de Kaa, G., & Liu, W. (2020). Importance of freight mode choice criteria: An MCDA approach. *Journal of Supply Chain Management Science, 1*(1-2).

Uddin, M., Anowar, S., & Eluru, N. (2021). Modeling Freight Mode Choice Using Machine Learning Classifiers: A Comparative Study Using the Commodity Flow Survey (CFS) Data. *Transportation Planning and Technology*, 543-559.

van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., & Walker, J. (2022). Choice modelling in the age of machine learning - Discussion paper. *Journal of Choice Modelling*.

van de Riet, O., de Jong, G., & Walker, W. (2012). Drivers of freight transport demand and their policy implications. In V. Himanen, & M. Lee-Gosselin, *Building Blocks for Sustainable Transport: Obstacles, Trends, Solutions* (pp. 73-102). Emerald Group Publishing Limited.

Viering, T., & Loog, M. (2023). The Shape of Learning Curves: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6).

Wang, S., Mo, B., & Zhao, J. (2020). Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions. *Transportation Research Part C: Emerging Technologies*, 234-251.

Wang, S., Wang, Q., & Zhao, J. (2020). Deep neural networks for choice analysis: Extracting complete economic information for interpretation. *Transportation Research Part C: Emerging Technologies*.

Xu, X., Yang, H., Jeong, K., Bul, W., Ravulaparthy, S., Laarabi, H., Needell, Z., & Spurlock, C. (2024). Teaching freight mode choice models new tricks using interpretable machine learning methods. *Frontiers in Future Transportation*.

Yi-Le Chan, J., Mun Hong Leow, S., Thye Bea, K., Khuen Cheng, W., Wai Phoong, S., Hong, Z., & Chen, Y. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics*.

Yong, T. K., Ma, Z., & Palmqvist, C. (2025). AP-GRIP evaluation framework for data-driven train delay prediction models: systematic literature review. *European Transport Research Review*.

Zhang, H., Zhang, L., Liu, Y., & Zhang, L. (2023). Understanding Travel Mode Choice Behavior: Influencing Factors Analysis and Prediction with Machine Learning Method. *Sustainability*.

Zhao, X., Yan, X., Yu, A., & van Hentenryck, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society*, 22-35.

# Appendix A: Generalizability Results

The full results for the generalizability test are shown in the table below. The road, rail, and inland waterway mode shares are calculated based on tonnage amounts across all data rows for a particular country. The data rows for a country are rows where the origin, destination, or both is within that country. The accuracy and log-loss are also tonnes-weighted.

| Country | Data Rows | Road % | Rail % | IWW % | Accuracy | | | Log-Loss | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | *LR* | *RF* | *XGB* | *LR* | *RF* | *XGB* |
| AT | 8678 | 96.96 | 2.62 | 0.41 | 0.898 | 0.866 | 0.903 | 0.417 | 0.484 | 0.275 |
| BE | 8566 | 85.08 | 0.94 | 13.98 | 0.326 | 0.304 | 0.436 | 1.936 | 1.469 | 1.418 |
| BG | 1098 | 91.42 | 6.32 | 2.26 | 0.527 | 0.066 | 0.376 | 1.094 | 1.835 | 1.352 |
| CH | 2321 | 97.31 | 1.26 | 1.43 | 0.954 | 0.960 | 0.956 | 0.169 | 0.308 | 0.211 |
| CZ | 7756 | 89.31 | 10.69 | 0.01 | 0.691 | 0.135 | 0.255 | 0.667 | 1.498 | 1.473 |
| DE | 50202 | 86.72 | 8.17 | 5.11 | 0.823 | 0.816 | 0.830 | 0.873 | 0.647 | 0.604 |
| DK | 688 | 63.73 | 36.27 | 0.00 | 0.741 | 0.762 | 0.751 | 0.693 | 0.593 | 0.748 |
| ES | 981 | 1.61 | 98.39 | 0.00 | 0.894 | 0.583 | 0.319 | 0.424 | 0.717 | 1.670 |
| FR | 5245 | 71.43 | 5.84 | 22.74 | 0.644 | 0.722 | 0.700 | 1.093 | 0.791 | 0.793 |
| HR | 206 | 0.00 | 99.67 | 0.33 | 0.406 | 0.950 | 0.855 | 0.807 | 0.601 | 0.559 |
| HU | 1534 | 48.57 | 31.08 | 20.36 | 0.548 | 0.655 | 0.619 | 1.072 | 0.720 | 0.927 |
| IE | 9 | 0.00 | 100.00 | 0.00 | 0.001 | 1.00 | 1.00 | 1.792 | 0.165 | 0.174 |
| IT | 953 | 20.58 | 79.42 | 0.00 | 0.838 | 0.218 | 0.366 | 0.386 | 0.985 | 1.232 |
| LU | 347 | 100.00 | 0.00 | 0.00 | 0.642 | 0.998 | 0.998 | 0.980 | 0.372 | 0.141 |
| NL | 12711 | 13.68 | 4.52 | 81.80 | 0.363 | 0.211 | 0.228 | 1.168 | 1.529 | 1.985 |
| NO | 198 | 0.00 | 100.00 | 0.00 | 0.719 | 0.967 | 0.961 | 0.330 | 0.342 | 0.305 |
| PL | 5373 | 12.74 | 86.86 | 0.41 | 0.697 | 0.756 | 0.794 | 0.624 | 0.490 | 0.433 |
| RO | 1057 | 0.31 | 70.63 | 29.06 | 0.232 | 0.324 | 9.386 | 1.769 | 1.347 | 1.471 |
| SE | 688 | 5.80 | 94.20 | 0.00 | 0.844 | 0.948 | 0.849 | 0.285 | 0.212 | 0.314 |
| SK | 1876 | 38.73 | 59.73 | 1.54 | 0.761 | 0.820 | 0.789 | 0.541 | 0.490 | 0.469 |

# Appendix B: Scientific Paper

## B.1. Abstract

Freight mode choice models until recently have been developed using discrete choice models such as Multinomial Logit (MNL) models. These models have advantages in transparency and interpretability but are often limited in their predictive performance and ability to capture complex, nonlinear relationships between variables. Although machine learning has been applied to modeling freight mode choice, a knowledge gap exists in the performance of these models when trained with aggregated freight data, as opposed to the more detailed, disaggregate shipment-level data, and in the role they could play in real-world policy analyses compared to MNL and other Logit-based models. To support accurate predictions of EU freight flows, and by extension, provide more reliable policy recommendations, it is important to use aggregate data, as this is the form most commonly available at the European level, and to evaluate models holistically beyond just their predictive performance. To fill this gap, three machine learning models are trained using EU aggregate freight data. These are compared to the NEAC MNL model, an EU freight transport model developed and maintained by Panteia, using seven criteria identified for evaluating machine learning models for freight transport policy evaluations. The results show that XGBoost achieves the highest predictive performance (92.1%), while logistic regression demonstrates advantages in generalizability, robustness, and data efficiency. The analysis highlights the trade-off between predictive performance and interpretability/practicality, demonstrating that machine learning models can complement but not replace Logit-based models in freight policy applications.

## B.2. Introduction

Freight transportation makes up 5% of the Gross Domestic Product (GDP) and 25% of total greenhouse gas emissions in the European Union (EU). In order for the EU to meet its emission reduction goals, inland freight transportation, 77% of which is currently by road, must shift more to inland waterway and rail (Eurostat, 2020) (Eurostat, 2022). Mode choice models are necessary to evaluate how well transport policies can affect this desired change. Although these are typically estimated using a Multinomial Logit (MNL) or other Logit-based model, machine learning models have gained more popularity recently due to their often higher predictive accuracy.

Previous studies applying machine learning models for modeling freight mode choice demonstrate that many machine learning algorithms produce more accurate predictions compared to Logit models, with tree-based methods such as Random Forest achieving the highest accuracy (Liu, et al., 2024) (Ahmed & Roorda, 2022) (Xu, et al., 2024). In some studies, this difference in accuracy between the best machine learning model and the Logit model can be as much as 33 percentage points (Uddin, Anowar, & Eluru, 2021), whereas other studies show discrete choice models have only 3-4% lower accuracy than the top machine learning classifiers (Wang, Mo, & Zhao, Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions, 2020). Finally, to obtain the improvement in predictive accuracy with machine learning while maintaining the interpretability of discrete choice models, other studies have developed hybrid approaches wherein for example a part of the utility function is specified using neural networks and the other part with an MNL model (Sifringer, Lurkin, & Alahi, 2020) (Han, Camara Pereira, Ben-Akiva, & Zegras, 2022).

The purpose of this study was to explore the role that machine learning models can play in freight mode choice modeling for transport policy analyses. This was done through a case study of the NEAC model, a European freight transport MNL model developed and maintained by Panteia. The existing knowledge gap was in the use of machine learning models trained on aggregated freight data, as previous studies have only used shipment-level disaggregated data. It was therefore unclear whether machine learning models trained with the less detailed aggregate data could achieve a comparable level of predictive accuracy. Because aggregated data is more easily accessible and available across European countries, it is necessary to utilize this type of data in order to develop a European-wide model capable of predicting mode shares across the region.

The second research gap was in the suitability of machine learning for freight transport policy analyses, considering other evaluation criteria beyond predictive performance. Previous studies focused on improving the predictive accuracy of machine learning models and did not more comprehensively consider whether and how these models could be useful in real-world policy applications. In response to this gap, seven evaluation criteria were defined, and several machine learning models were trained and compared to each other and to the NEAC MNL model using these criteria.

Based on a literature review and the input from a potential end-user of the machine learning models developed in this study, the evaluation criteria were defined and ordered in term of importance as follows: predictive performance, interpretability, practicality, computation time, robustness, generalizability, and data efficiency. The algorithms included in this study were logistic regression, Random Forest, and XGBoost, the former chosen as a baseline of comparison to the more advanced algorithms and the latter two chosen due to their high accuracy demonstrated in previous freight mode choice studies (Xu, et al., 2024) (Liu, et al., 2024). A dataset was constructed using datasets with freight tonnage amounts in the EU in 2015 and other external data sources. Lastly, recommendations were offered on whether and how machine learning models can be used in a freight transport model for policy applications.

## B.3. Data and Methodology

### B.3.1 Data Sources

The dataset was constructed from Eurostat freight flow data for road, rail, and inland waterway transport at a NUTS-2 level for 2015. NUTS-2 zones represent basic regions suitable for regional policy use (Eurostat, n.d.). Real data was used where available and supplemented by estimates from the NEAC base year database. The final dataset contains 67,210 data rows, with freight transported in tonnes between Origin-Destination (OD) pairs by mode and commodity type. In the data, 499,574 thousand tonnes are transported by inland waterway, 616,377 thousand tonnes by rail, and 4,241,777 by road. Due to estimations and adjustments that were necessary when constructing the dataset, inland waterway and rail tonnes are overestimated and road underestimated relative to Eurostat's aggregate statistics, but the dataset retains several key structural features of EU freight flows including the dominance of road transport and the main international corridors.

Explanatory variables such as costs, commodity type, and regional characteristics were included. Generalized cost variables were calculated from fixed (time-based) and variable (distance-based) values for each mode. Commodity types were incorporated from Eurostat data for inland waterway and from estimates for rail and road. Other variables include inland waterway availability, travel time and travel distance, distance from zones to rail and inland waterway terminals, rail and inland waterway service, and dummy variables for Eastern and Western Europe. Phik correlations and

Variance Inflation Factors (VIF) were calculated to assess multicollinearity. Variables with Phik correlations above 0.8 were excluded to help with model interpretability, while others with high VIF values were retained to preserve explanatory power. The final feature set is shown in Table B.1.

*Table B.1. Explanatory Variables*

| Variable | Description | Type of Variable |
|---|---|---|
| Generalized cost | Cost in euros per mode and OD pair | Continuous |
| Commodity type - NST 0 | Agricultural products and live animals | |
| Commodity type - NST 1 | Other food products and animal feed | Nominal |
| Commodity type - NST 2 | Solid mineral fuels | Nominal |
| Commodity type - NST 3 | Petroleum and petroleum products | Nominal |
| Commodity type - NST 4 | Ores, metal waste, roasted iron oxide | Nominal |
| Commodity type - NST 5 | Iron, steel, and non-ferrous metals | Nominal |
| Commodity type - NST 6 | Crude minerals and manufactured products; building materials | Nominal |
| Commodity type - NST 7 | Fertilizers | Nominal |
| Commodity type - NST 8 | Chemical products | Nominal |
| Commodity type - NST 9 | Vehicles, machinery, and other goods | Nominal |
| Commodity type - NST 10 | Other/unknown | Nominal |
| Inland waterway availability | Whether a viable path exists between the two zones by inland waterway | Binary |
| Distance from terminals | Distance in kilometres from zones to nearest inland waterway or rail terminal | Continuous |
| Eastern Europe | Whether both origin and destination are in Eastern Europe | Binary |
| East West | Whether origin is in Eastern Europe, destination is in Western Europe, or vice versa | Binary |
| Rail service | Rail service quality by country | Ordinal |
| Inland waterway service | Inland waterway service quality by country | Ordinal |

### B.3.2 Model Selection

Three algorithms were selected: logistic regression, Random Forest, and XGBoost. Logistic regression has the same mathematical formulation of MNL but applies regularization automatically through L1 (lasso) or L2 (ridge) regularization. Random Forest is an ensemble of decision trees built on random subsets of data. XGBoost is a gradient boosting algorithm that builds on simpler decision trees iteratively to reduce residual errors.

While artificial neural networks and support vector machines are widely used in choice modeling and have been shown to produce more accurate behavioral outputs such as willingness to pay (Martin-Baos, Lopez-Gomez, Rodriguez-Benitez, Hillel, & Garcia-Rodenas, 2023), they were not included because deriving and comparing behavioral outputs is outside the scope of this research and, due to time constraints, only a limited number of algorithms could be trained.

### B.3.3 Model Training and Evaluation

The data was split into training (85%) and test (15%) sets, stratified by OD pair and mode. Data instances with the same origin and destination are not independent, so they were kept together in either the training set or the test set to prevent data leakage. The continuous variables were standardized only in the logistic regression model. Hyperparameters were tuned using RandomizedSearchCV with 5-fold stratified group cross-validation. The models were trained with the tonnage amounts as sample weights, leading the models to prioritize correctly predicting rows with higher tonnage amounts.

The evaluate metrics included tonne-weighted accuracy, precision, recall, F1-score, log-loss, and confusion matrices as well as differences between actual and predicted mode shares in the test set. Additional evaluation tests were done to assess generalizability, robustness, and data efficiency. For generalizability, a country was taken out of the training set and used as the test set. This was done

for every country in the dataset, and the average accuracy and log-loss was taken to compare the models' performance on unseen regions. To assess robustness or sensitivity to measurement error, noise was injected to the cost and distance variables. To measure data efficiency, the learning curves were plotted to observe the change in performance with different amounts of training data. Shapley Additive exPlanations (SHAP) values were calculated for Random Forest and XGBoost to assess feature importance and interpretability. For logistic regression, interpretability was discussed based on the Logit coefficients and alternative-specific constants.

## B.4. Results

### B.4.1 Model Results

Logistic regression achieved the lowest accuracy (89.1%) but also the smallest overfitting, with a log-loss gap of 4.4%, shown in Table B.2. Random Forest and XGBoost produced higher accuracies (91.5% and 92.1%) and lower log-loss values, but both exhibited greater overfitting (25% and 21% log-loss gaps, respectively). These outcomes reflect the bias-variance trade-off: LR has higher bias but lower variance, whereas RF and XGB reduce bias at the expense of greater variance.

*Table B.2. Test Accuracy and Log-Loss*

| | Logistic Regression | Random Forest | XGBoost |
|---|---|---|---|
| Test accuracy | 0.891 | 0.915 | 0.921 |
| Test log-loss | 0.349 | 0.2951 | 0.2454 |
| Train/test log-loss gap | 4.37% | 24.99% | 20.95% |

Class-level results (Table B.3) show that all models predicted road with high precision and recall (>0.93), while performance on inland waterway and rail was weaker. RF achieved the highest average F1-score for the minority classes (0.67), slightly more than XGB (0.66) and LR (0.62). XGB, however, reproduced aggregate mode shares most accurately (Table X), with differences between actual and predicted mode shares below 1% for all modes, whereas LR and RF overpredicted inland waterway and rail shares.

*Table B.3. Precision, recall, F1-scores by Mode and Differences between Actual and Predicted Mode Shares*

| Mode | Model | Precision | Recall | F1-score | Mode Share Difference |
|---|---|---|---|---|---|
| Inland waterway | LR | 0.52 | 0.69 | 0.59 | +2.79% |
| | RF | 0.54 | 0.66 | 0.59 | +3.43% |
| | XGB | 0.68 | 0.56 | 0.61 | +0.50% |
| Rail | LR | 0.69 | 0.62 | 0.65 | +4.78% |
| | RF | 0.88 | 0.65 | 0.75 | +4.68% |
| | XGB | 0.89 | 0.6 | 0.71 | -0.05% |
| Road | LR | 0.94 | 0.93 | 0.94 | -7.57% |
| | RF | 0.95 | 0.96 | 0.95 | -8.11% |
| | XGB | 0.93 | 0.98 | 0.96 | -0.45% |

Coefficient estimates from the LR model, shown in Table B.4, highlighted possible multicollinearity problems: several signs were counterintuitive, such as a negative effect of rail service on rail utility. Another possibility is that the rail service variable is capturing the effects of other variables that were not included in the model; other factors such as road infrastructure quality, trade flow characteristics, or physical barriers could be correlated with rail service and influence mode choice,

leading to the counterintuitive negative coefficient. Other coefficients are more in line with behaviorally realistic substitution patterns: higher road costs increase the utility of rail and inland waterway, and inland waterway utility increases with inland waterway availability.

*Table B.4. LR Coefficients and Alternative-Specific Constants with Road as Base Alternative*

| | **Inland waterway** | | **Rail** | |
|---|---|---|---|---|
| | *Coefficient* | *Standard Error* | *Coefficient* | *Standard Error* |
| Alternative-specific constants | -8.654 | 0.197 *** | 6.177 | 0.121 *** |
| Road cost | 2.455 | 0.009 *** | 1.242 | 0.004 *** |
| Rail cost | -0.134 | 0.009 *** | -0.727 | 0.005 *** |
| IWW cost | -2.928 | 0.014 *** | 0.0927 | 0.002 *** |
| IWW available | 14.977 | 0.154 *** | -0.582 | 0.007 *** |
| NST 0 | 0.348 | 0.122 *** | -0.782 | 0.121 *** |
| NST 1 | -10.205 | 0.175 *** | -2.788 | 0.121 *** |
| NST 2 | 3.212 | 0.123 *** | 4.533 | 0.121 *** |
| NST 3 | 2.885 | 0.175 *** | 0.853 | 0.121 |
| NST 4 | 0.149 | 0.123 ** | 0.232 | 0.121 ** |
| NST 5 | -0.211 | 0.123 | 0.541 | 0.121 |
| NST 6 | 0.373 | 0.123 ** | 0.153 | 0.0121 ** |
| NST 7 | -5.994 | 0.196 *** | 3.599 | 0.121 *** |
| NST 8 | 0.122 | 0.123 ** | 0.031 | 0.121 ** |
| NST 9 | -0.705 | 0.122 | -0.731 | 0.121 *** |
| NST 10 | 1.427 | 0.124 *** | 0.465 | 0.122 |
| Origin IWW | -2.292 | 0.009 *** | 0.303 | 0.002 *** |
| Origin Rail | -0.564 | 0.003 *** | -0.202 | 0.002 *** |
| Dest IWW | -2.623 | 0.01 *** | 0.046 | 0.002 *** |
| East West | -4.892 | 0.032 *** | -3.06 | 0.014 *** |
| East Europe | -2.538 | 0.022 *** | -3.053 | 0.012 *** |
| Service Rail | -4.444 | 0.008 *** | -2.362 | 0.005 *** |
| Service Water | 0.345 | 0.07 *** | -0.267 | 0.003 *** |

In both RF and XGB, the SHAP values shown in Figure B.1 revealed the same counterintuitive relationship between rail service and the probability of choosing rail as in the Logit coefficients. Other predictors were more behaviorally realistic, especially in XGB: rail and road costs ranked higher in

importance compared to RF which is more aligned with the known strong influence cost has on mode choice.

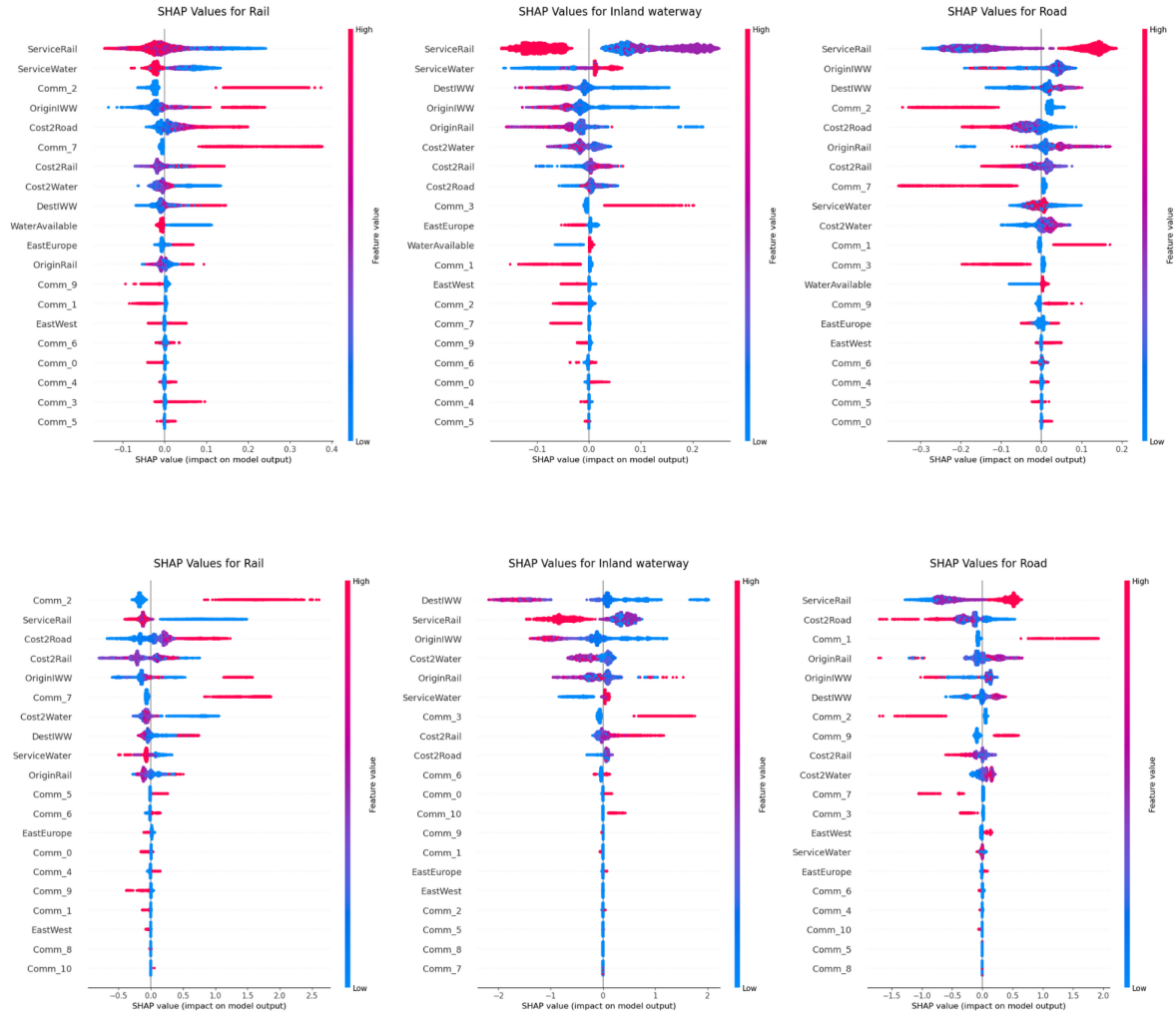


*Figure B.1. RF (top) and XGB (bottom) Beeswarm Plots*

### B.4.2 Model Comparison by Criteria

The three machine learning models were compared against the seven evaluation criteria. Where possible, the results were also compared with the NEAC MNL model, though the latter was not estimated as part of this research.

*Predictive performance.* XGBoost achieved the best overall predictive performance, with the highest test accuracy (92.1%), lowest log-loss (0.245), and smallest errors in predicted mode shares. Logistic regression performed only slightly worse in terms of accuracy, likely reflecting the mostly linear relationships in the dataset. Random Forest produced the best F1-scores for the minority classes, but this also resulted in overpredicting inland waterway and rail shares. Overall, XGBoost provided the best balance of predictive accuracy and mode share realism.

*Interpretability.* Interpretability can be divided into behavioral and explanatory dimensions. Behavioral interpretability, including willingness-to-pay and elasticities, was limited in this dataset due to multicollinearity; several logistic regression coefficients had counterintuitive signs. Explanatory interpretability, based on SHAP values, provided insights into the importance of each variable in RF and XGB's predictions. XGBoost produced feature importance rankings that aligned more closely with behavioral expectations in that generalized costs were stronger predictors than in Random Forest. The NEAC MNL model remains superior for deriving behavioral measures, though SHAP offers a useful complement for machine learning models.

*Practicality:* Practicality relates to the usability of results by analysts and policymakers. Machine learning models are limited by their inability to extrapolate beyond the training range and by their lack of causal inference. MNL models remain preferable in projects requiring scenario analysis with large deviations from the base year. Nevertheless, for short- to medium-term applications where explanatory causality is less critical, machine learning models may provide reliable predictive performance.

*Computation time:* Logistic regression required the least computation time (49.4 seconds), with only two hyperparameters to search. XGBoost (90.4 seconds) was more efficient than Random Forest (221.1 seconds) due its restricted tree depth. While the differences are small, they may become more important when using larger datasets or when extensive hyperparameter tuning is required.

*Generalizability:* The leave-one-country-out testing revealed a marked drop in performance for all models when tested on countries excluded from the training set. Average accuracies fell to 62.8% (LR), 65.3% (RF), and 66.8% (XGB). Logistic regression outperformed the other two models in some cases (Bulgaria, Czech Republic, Spain), while only being slightly less accurate in most other countries, suggesting it may be slightly more generalizable to new regions. Overall, none of the models demonstrated strong generalizability to unseen countries.

*Robustness:* Noise was introduced into the cost and terminal distance variables to test robustness. Logistic regression showed the least sensitivity to measurement errors, with accuracy falling by only 1.6% under 30% noise, while Random Forest and XGBoost displayed larger relative increases in log-loss (8.5% and 10.2%). Despite this, RF and XGB maintained higher absolute accuracies overall. Due to its smaller drop in performance with greater noise, LR is the most robust.

*Data efficiency:* Logistic regression achieved stable accuracy even with just 20% of the training data, while Random Forest and XGBoost continue to improve more in predictive performance with larger datasets. This aligns with expectations that linear models are more data efficient, while tree-based models benefit from larger, richer datasets.

The results across all seven criteria are summarized in Table B.5, where a plus sign (+) indicates the best performance, a minus sign (-) indicates the worst.

*Table B.5. Model Performance Against Evaluation Criteria*

| | NEAC MNL | Logistic Regression | Random Forest | XGBoost |
|---|---|---|---|---|
| Predictive performance | | | - | + |
| Interpretability | + | | - | |
| Practicality | + | | | |
| Computation time | + | + | - | |
| Generalizability | / | + | | |
| Robustness | / | + | | - |
| Data efficiency | / | + | | - |

XGBoost outperformed the other models in predictive performance and produced more behaviorally realistic SHAP values compared to Random Forest. Logistic regression performed best in terms of generalizability, robustness, computation time, and data efficiency. Random Forest did not outperform either other model across any criterion. The NEAC MNL model remains superior for interpretability and practicality due to its ability to produce behavioral indicators. Overall, XGBoost was considered the strongest candidate for integration into the NEAC framework, though logistic regression remained a valuable baseline, particularly when interpretability, robustness, or data efficiency are prioritized.

## B.5. Discussion

The predictive accuracies of the models in this research (89.1-92.1%) are largely consistent with earlier studies, though the relative differences between models are smaller. Prior studies found Random Forest and XGBoost outperformed logistic regression and MNL by a larger margin, with gaps exceeding 15-30 percentage points (Uddin, Anowar, & Eluru, 2021) (Xu, et al., 2024) (Liu, et al., 2024). In contrast, the models in this study showed only marginal improvements of the tree-based methods over logistic regression. One explanation is that the dataset used contains primarily linear relationships that logistic regression can capture effectively, leaving fewer gains for non-linear models. Another factor is the use of aggregated rather than disaggregate data. It is possible that with aggregation, some detailed information and variation in the data was lost where more complex models would have had a greater advantage over a linear model. Finally, the search ranges for RF and XGB hyperparameters were constrained to reduce overfitting, resulting in models of moderate complexity.

All three algorithms performed better at predicting road than the minority classes of rail and inland waterway, reflecting the imbalance in mode shares. This reflects a common challenge in transport mode choice modeling (Abdelhamid & Desai, 2024). In this study, adding class weights improved minority class recall, but this was still far below road recall. By comparison, other studies using different datasets have sometimes achieved strong performance even for minority modes (Xu, et al., 2024). The lower minority class performance in this research may be the result of data quality issues in the dataset or insufficient distinguishing features between road and the minority classes.

One of the limitations of this study is firstly the possible data quality issues in data completeness, as some flows lacked commodity classifications, and several OD pairs were excluded due to missing or confidential values. Secondly, generalized cost and terminal distance variables may contain estimation errors, resulting in lower feature accuracy. Previous work has emphasized that both completeness and accuracy strongly influence machine learning model performance (Mohammed, et al., 2025). Addressing these issues through more extensive data validation could improve minority mode predictions. Another limitation is that model evaluation across criteria was primarily qualitative; a weighted multi-criteria framework would enable more systematic assessment. Lastly, this study considered only single-model approaches, while ensembled learning techniques have been shown to improve predictive accuracy in other freight mode choice studies (Liu, et al., 2024).

The XGBoost model demonstrated the strongest overall performance and appears most suitable for integration into the NEAC framework. However, its potential applications differ from those of the current NEAC MNL model. Machine learning models cannot extrapolate well to scenarios with values outside the training range, such as long-term forecasts. For such cases, the MNL model remains more appropriate. While XGBoost predicts overall mode shares with high accuracy, its recall for inland waterway and rail is moderate (0.56 and 0.60 respectively), making it potentially less suitable for analyses where minority modes are of central interest. Based on these findings, the XGBoost model is determined to be a strong complement rather than substitute to the NEAC MNL model.

## B.6. Conclusion

This study examined the role of machine learning in freight mode choice modeling for policy applications. Three models were trained on a 2015 EU aggregated freight flows dataset and evaluated against an MNL model using seven criteria: predictive performance, interpretability, practicality, computation time, generalizability, robustness, and data efficiency. Previous studies used disaggregate shipper data and emphasized mainly predictive accuracy; this research demonstrates that aggregated data can still yield high predictive accuracy, while also offering an assessment of how model characteristics beyond accuracy shape their suitability for policy use.

The evaluation across criteria highlights trade-offs between models. XGBoost achieved the strongest predictive performance and the most behaviorally plausible SHAP feature rankings, while Random Forest improved minority class precision at the expense of mode share accuracy. Logistic regression outperformed the tree-based methods in computation time, generalizability, robustness, and data efficiency. The current NEAC MNL model remains more interpretable and practical. The results indicate that machine learning and MNL approaches should be seen as complementary: XGBoost enhances predictive performance within observed data ranges, while MNL maintains interpretability and extrapolation capabilities.

Future studies could compare additional algorithms, including neural networks and support vector machines, which may improve the current accuracy achieved and also offer additional advantages in the other criteria besides predictive performance. Secondly, adding other explanatory variables such as physical locations of shippers and buyers and other mode characteristics including flexibility, safety, and security could improve minority class recall. Including other country-specific variables may also help the models to identify mode share patterns by country, potentially improving predictive accuracy as well.