SME Credit Scoring Using Social Media Data

Master's Thesis

Septian Gilang Permana Putra

SME Credit Scoring Using Social Media Data

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE
TRACK DATA SCIENCE AND TECHNOLOGY

by

Septian Gilang Permana Putra
born in Kendal, Indonesia

Web Information Systems
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
http://wis.ewi.tudelft.nl

Exact
Molengraaffsingel 33
Delft, the Netherlands
www.exact.com

SME Credit Scoring Using Social Media Data

Author: Septian Gilang Permana Putra
Student id: 4609328
Email: septiangilangpermanaputra@student.tudelft.nl

Abstract

Credit analysis is required in a wide variety of decision of a modern economy. It includes understanding the credit risk of small-medium enterprises (SMEs), which today is the most significant contributor to the economy of almost every nation. Creditors usually use credit scoring as a tool to predict the probability of the SMEs to default in the future. The existing methods of SMEs credit scoring still rely on traditional data, which may require high cost and have low scalability. This thesis proposed an alternative approach of credit scoring for small-medium enterprises (SMEs), which incorporate a novel set of features extracted from social media data.

As a study case, we generate the credit scoring dataset which contains 20 traditional features and 35 social media features to quantify the creditworthiness of more than 20,000 SMEs. The social media features are formulated based on the previous studies in the adoption of social media data for personal credit scoring and the social media metrics for quantifying business social perception. To build the dataset, we develop the method to collect the information from some public websites and SMEs' Facebook page.

We conduct some experiments to develop credit scoring model for SMEs. We found that using only the social media features insufficient to model SMEs default in the future. However, by combining both social media features to build the credit scoring model, we will get better performance compared to the model developed using only traditional data.

Thesis Committee:

Chair:	Prof. dr. ir. G.J.P.M. Houben,, Faculty EEMCS, TUDelft
University supervisor:	Dr. ir. A. Bozzon, Faculty EEMCS, TUDelft
Company supervisor:	Dr. J.A. Redi, Exact

Preface

Alhamdulillah, praise and thanks to ***Allah SWT*** *for all the graces and blessings so that this study can be completed.*

This document concludes my two years as a Computer Science student at the Delft University of Technology. Throughout working on this project for almost a year, I have a great experience which allows me to grow both technically and personally. Working with them I am very proud that I can finish this thesis in which would never be possible without the valuable advice and support of many people that I would like to thank in this preface.

I would like to thank both of my supervisor, **Dr. Alessandro Bozzon** and **Dr. Judith Redi**, for the opportunity to work in this project. Because of this project, I got a chance of doing internship in a large enterprise and a comprehensive experience in conducting computer science research, which is invaluable. I also want to thank both of them for providing me with motivation, guidance, critics, and feedback throughout the course of this project. Thanks also to **Alex, Bikash, Jennifer, Jingting, Naim, Mark van Asten, Mark van Dijk, Marichelle, Pedro, Rajiv** and all of the people from Exact which always support and help me during my master thesis and internship at Exact.

Last but not least, I want to give my best thanks to my parents (**Eko Nowo** and **Sri Sugiyanti**) who always support me and pray for my success. To my fellow Indonesian comrades in Computer Science: **Reza, Ulin, Andre, Romi, Sindu, and Helmi**, thanks for the great moments we shared during the past two years and the supports in my studies and projects. I also want to give my huge appreciation to the participants who help me collecting the data from Facebook for this thesis. Without them, this thesis could not be done. Finally, special thanks to **Lembaga Pengelola Dana Pendidikan (LPDP)** for providing me the financial support and opportunity to study at this amazing university.

Septian Gilang Permana Putra
Delft, the Netherlands
September 19, 2018

Contents

List of Figures

Chapter 1

Introduction

Credit analysis is required in a wide variety of decision of a modern economy. Merchants, who exchange goods for promises to pay, need to evaluate the reliability of those promises. Commercial banks which lend the merchants the funds to finance their inventories, likewise need to calculate the probability of being repaid in full and on time. In both cases, merchants and commercial banks can use credit scoring as a tool to evaluate their customer's future loan performance. They use some indicators, such as a set of financial ratio, to predict whether their customers can repay the loan as promised. This thesis proposed an alternative approach of credit scoring for small-medium enterprises (SMEs), which incorporate a novel set of indicators extracted from the social media.

In this introductory chapter, we will introduces the background that motivate this thesis in section 1.1 followed by main research question and research sub-questions in 1.2. Then, the methods and contributions of this research are explained in section 1.3 and section 1.4. Finally, the structure of this document is elaborated in section 1.5.

1.1 Background

Small and medium-sized enterprises (SMEs) are the backbone of the economy in nearly every nation. In the Netherlands, for example, SMEs are estimated to represent 99.8 % of all enterprises, account for 66.6% of overall employment, and contribute to 62.9% value added of the non-financial business sector in 2016 [14]. Due to those significant contributions to the economy, it is obvious that supporting the financial needs of SMEs is crucial to a country's growth.

From the creditors perspective, SMEs is considered as high-risk clients. Besides having a higher failure rate compared to larger enterprises, their information also is much more opaque which makes it difficult for creditors to measure their creditworthiness, the likelihood that a borrower will be able to repay his or her debt in time. Due to those distinct characteristics, creditors need to manage SMEs credit risk separately from larger enterprises [4]. To mitigate the risk, many companies and banks use credit scoring which specifically built for SMEs as a guide in their credit-decision process. Credit scoring is the term used to describe the process of assigning a quantitative measure to a potential borrower as an estimate of how likely the default will happen in the future. [18].

There is a growing number of researches that developed various credit scoring model to measure the creditworthiness of SMEs. Although those researches vary in the type of the predictor used, most of them applied some machine learning method to build the scoring model. The most commonly used predictor is a set of financial ratios extracted from accounting data, e.g. current ratio, return on asset (ROA) and debt-to-equity ratio. For many years, financial ratios have been proven to have predictive power in explaining the future failure of SMEs [3, 16]. However, such detailed information is hard to obtain and often is not available since their financial reports are mainly for tax purposes. Also, analyzing financial data may not be sufficient as there might be another factor which determine SMEs' willingness and ability to repay their loan [19].

Besides the financial ratios, the knowledge about marketability, technology advantage, management quality, age, size, type of industry, and geographical area also often used in addition to the traditional financial data [6, 35]. The credit assessors should manage an on-site survey or have long enough historical relation with the borrowers to collect the data. Although the data could provide better risk assessment in SMEs lending, it takes time and costly. The subjectivity also involved in measuring some indicators which could affect the validity of the model. Thus, a more efficient approach is required in gathering and quantifying the information.

Each time SMEs and their customers use cloud-based services, make or accept digital payments, engage in social media, get rated online, ship packages, or manage their bookkeeping online, they left the digital footprints behind. From social media platforms such as Facebook, Twitter, Yelp, Foursquare, and TripAdvisor, creditors can gather information regarding business reviews, rankings, and other precious data regarding how SMEs interact with their customers. This alternative data can be mined for credit decision purposes via Application Programming Interfaces (APIs) or the use of scraping technology. With this alternative data, SME lenders are now able to develop a more comprehensive view of the SMEs' business and build a credit scoring model which more transparent, faster, easier, cheaper, and suitable for SMEs.

1.2 Research Questions

Motivated by previously elaborated background, this thesis proposes a new method in implementing credit scoring systems for SMEs which makes use of the social media data to build default prediction model. The objectives of this work can be summarized in the following main research question.

MRQ To what extent can social media data be used for predicting the loan default in SME credit scoring system?

In order to answer the main research question, the following sub-questions are defined:

RQ1 *How is the existing SME credit scoring system implemented?*

It has been mentioned previously that the credit scoring for SME is mostly built on top of financial data analysis So, we need to conduct literature studies on previous research which utilize financial data in their scoring systems Based on this study, we can identify common approaches and methods and use it as the basis of our framework.

RQ2 *How to build features to measure SMEs creditworthiness from social media data?*

Social media can be an alternative source of data for evaluating creditworthiness However, due to to the unstructured nature of social media data compared to the conventional data source (i.e financial report, loan history, application form), different techniques are required to collect, process, and analyze such knowledge The techniques should consider how the extracted features can be reliable to represent the actual condition of the firms.

RQ3 *How to incorporate social media features into SME credit scoring system?*

This research tries to develop a statistical model, which can help to assess the creditworthiness of SMEs using extracted social media features in addition to the financial features It includes the details such as, how the features are preprocessed, how specific methods are selected to build the model and how the model is evaluated.

RQ4 *Which social media features influence the performance of SME credit scoring systems?*

We hypothesize that the SME related information from social media may improve the performance of the default prediction model Thus, there may be some patterns that make an SME is riskier compared to other based on its social media features This research tries to identify the social media features of firms that may relate to its probability of default.

1.3 Methods

The purpose of credit scoring is to solve the problem of distinguishing between solvent (likely to repay) and delinquent (likely to default) borrower. In this project, we treat the credit scoring problem as a binary classification problem to achieve that purpose. To solve a binary classification problem, we need a dataset containing a label representing SMEs default status and a set of features representing SMEs' creditworthiness.

We conduct a literature survey on how the previous studies perform credit scoring, especially in SMEs case. We want to identify which definition of default is used as the label, which features are useful in representing the creditworthiness of SMEs and which methods are used to in traditional SMEs credit scoring. The result of this literature survey is also the answer to our RQ1, which we use to build a baseline model.

As we are interested in using social media as an alternative source of data, we also perform literature study related to the use of social media data in credit scoring and other related topics such as the business performance evaluation. From the literature survey, we expect to get the picture of how credit scoring model for SMEs can be built, and which features are required and relevant in measuring SMEs creditworthiness. The result of this survey is the answer for our RQ2, which then we use to develop credit scoring model for SMEs.

Next, we specify data requirement containing the specification of data that need to be collected to build SMEs credit scoring model. In this project, the initial dataset is provided by Exact which contains the name, Chamber of Commerce (KvK) registration number, address, website and also accounting data of Dutch SMEs. Some SMEs' information is missing and SMEs' default label is not available in the dataset. We complete that information by collecting the data from the public website. For the social media data source, we decided to use Facebook in this project due to its popularity in SMEs. Facebook also provides API which allows us to retrieve the information from the company's public page. The API also has search features which help us to find the company's public page of SMEs in our dataset.

With a set of features constructed from financial and social media data, we can perform some experiment to answer the RQ3. First, we define the evaluation metrics that will be used to compare the model. Then, we conduct an experiment which combine traditional features and social media features to build credit scoring model for SMEs. During the experiment, various methods will be examined in order to build the best credit scoring model for predicting SMEs' default.

Finally, we answer the RQ4 by analyzing the best model and our dataset to find which social features that may influence the creditworthiness of SMEs. Various visualization techniques are also used to provide better insight about the dataset and model. We also want to know whether adding social media features is useful in SMEs credit scoring model.

The summary of the research questions and the methods that described above can be seen in Figure 1.1.

1.4 Contribution

This thesis delivers the following contributions:

1. Literature survey on social media metrics and credit scoring related topic. We identify several methodologies and approaches used in previous research in credit scoring, specially for SMEs. We also present some research related to the use of social media data in predictive analysis and in business performance evaluation which feasible to be adopted in SMEs credit scoring.

2. Develop framework to generate a credit scoring dataset based on social media data We provide sets of procedures and steps required to build an alternative dataset for studying creditworthiness from social data. We hope that the frameworks can help further development, customization and extension of this topic by the research community.

3. Develop a trained machine learning model that can automatically predict SME loan default using social media data. This thesis provides more insights on how can the social media be incorporated in SME loan default prediction. Instead of using only the financial data as the traditional credit scoring system does, the social media data can also be an additional data source.

1.5 Outline

The remainder of this thesis is organized as follows. In chapter 2, the literature related to credit scoring and the use of social media metrics in business related topics are presented. In Chapter 3, we explain the credit scoring framework that we use to develop SMEs credit scoring system which incorporates social media data. Chapter 4 elaborates the data collection methods, including the data acquisition from social media and public website, and feature construction. Chapter 5 shows the results and evaluations of the experiments, as well as the investigation to find the features that may be related to the creditworthiness of SME. Then, the results and limitations are discussed, the studies are concluded, and some future works are proposed in chapter 6.

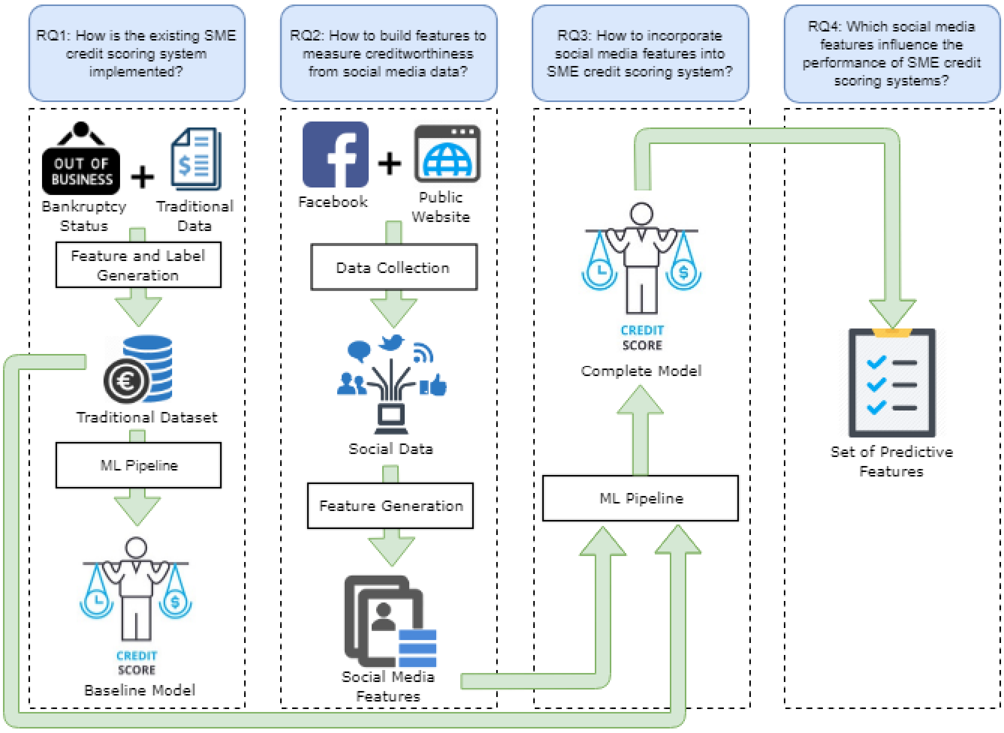

RQ1: How is the existing SME credit scoring system implemented?
RQ2: How to build features to measure creditworthiness from social media data?
RQ3: How to incorporate social media features into SME credit scoring system?
RQ4: Which social media features influence the performance of SME credit scoring systems?
OUT OF BUSINESS
Bankruptcy Status
Traditional Data
Feature and Label Generation
Traditional Dataset
ML Pipeline
CREDIT SCORE
Baseline Model
Facebook
Public Website
Data Collection
Social Data
Feature Generation
Social Media Features
CREDIT SCORE
Complete Model
ML Pipeline
Set of Predictive Features

Figure 1.1: Proposed Research Questions and Methods

Chapter 2

Related Works

This chapter discusses previous works related to both credit scoring and social media studies. In the first section, we provide a brief introduction and definition of credit scoring. The second section reviews the previous research on the implementation of the credit scoring for SMEs. In the third section, we present some literature related to the adoption of social media data in the personal credit scoring system, which may be adapted to develop a similar system for SMEs. The fourth section elaborates on existing studies which proposed some metrics to quantify company social perception using social media data, which may also be useful in credit scoring for SMEs. In the last part, these previous works are summarized, and their contributions related to this thesis are stated.

2.1 Fundamentals of Credit Scoring

The term credit scoring was first used in the banking industry, specifically within the context of lending money to those who are in need and collect interest on the payment the borrowers made. As some borrowers may fail to make their payments, there is a need to minimize the loss for the bank by accurately distinguish solvent (likely to repay) from delinquent (likely to default) borrower before the loan is granted. One method that can be used to evaluate the creditworthiness of potential borrowers is credit scoring.

Credit scoring is the term used to describe the process of assigning a quantitative measure to a potential borrower as an estimate of how likely the default will happen in the future. [18]. First, creditors build credit scoring model using statistical techniques to analyze various information of previous borrowers in relation to their loan performance. Afterward, the model can be used to evaluate a potential borrower who applies for a loan by providing the similar information which has been used to build the model. The result is either a score which represents the creditworthiness of an applicant or a prediction whether an applicant will default in the futures.

The definition of default in the context of credit scoring may vary between each financial institution as long as it complies with the Basel II Capital Accord [15]. Traditionally, credit scoring model is developed using bankruptcy status as the criterion of default. The bankruptcy status of company usually is available for public and can be obtained relatively easy. However, the creditors also suffer from losses before the event of bankruptcy occurs. For example when the payment is not being made in time, the debtors may lost the opportunity to make a transaction due to insufficient resources. Therefore, another definition of default can be used by creditors, which defined as the condition in which the borrower unable to repay in 90 days since the payment deadline. However, this definition of default is stricter and also harder to collect when building the model. If it is not available, previous research has shown that we can still apply the credit scoring model which is developed using the bankruptcy status as default definition to predict delay in repayment of more than 90 days without losing too much prediction power [20].

Prior adoption use of credit scoring, judgmental methods are used in which credit analyst conducted a manual assessment based on their previous experience and relationship with the borrower. The benefit of using credit scoring over judgmental methods in the approval process is it remove human bias and improve objectivity as the same criteria are applied to all borrowers [1]. Credit scoring can also greatly reduce the cost and time needed in the loan approval process compared to the judgmental method. In SME lending, the adoption of credit scoring leads to an increase in the number of loans approved for SMEs [9].

Beside those benefits, credit scoring also has some limitations that need to be addressed. In [30], Mester et al. found that the historical data used in credit scoring models does not necessarily contain enough information to estimate future performance. Another mentioned limitation is its accuracy depends on the sample used in the training process which assumed to be representative of the potential credit consumers of the future. When the assumption is incorrect or population changes, the static scoring model will fail to adapt which lead to inaccurate prediction. Regardless of those limitations, credit scoring does provide a method of quantifying the risk of the different group of potential borrowers which the previous method unable to do.

2.2 Developing Credit Scoring Model

Credit scoring models have been used in large companies and consumer lending for a few decades, but it is only recently adopted in SMEs lending. SMEs lending has distinct characteristic compared to the consumer or large firms lending which require creditors to build separate credit scoring model for it. Many studies investigated how to develop a credit scoring model specialized for SMEs. Those studies vary in the feature used to measure creditworthiness and also the technique used to build the scoring model. In this section, we try to highlight those features and techniques that have been used in the previous research to develop credit scoring model for SMEs.

2.2.1 Key Features in SMEs Credit Scoring

A lot of research related to credit scoring for SMEs try to find relevant features to estimate SMEs creditworthiness. Choosing relevant features is essential in order to produce high-quality credit scoring model. The most used features in those previous research is a set of financial ratios calculated from financial data.

Edmister was the first one to examine the predictive power of nineteen 19 financial ratios in explaining the future failure of SMEs [16]. The result of his study is a statistically significant discriminant function which confirms that financial ratios are valuable predictors of SMEs bankruptcy. However, not many people use this as it is not so stable due to many financial ratios used in the discriminant function.

Altman et al. [3] used four financial ratios calculated form accounting data to developed Z"-score, a bankruptcy prediction model for SMEs. The model is an enhanced version of the statistically proven Z-Score model used in large companies bankruptcy prediction. Ratios included are working capital/total assets, retained earnings/total assets, EBIT/total assets, and equity/total assets [1]. Due to its simplicity and stability, the model is popular and still used to evaluate SMEs creditworthiness.

However, there is a limitation to using only financial features in credit scoring for SMEs because such detailed information is hard to obtain and often not available. Also, there might be another factor which determines SMEs' willingness and ability to repay their loan because only between 30% and 50% of SMEs' default can be associated with those factors [19, 17]. As the financial data does not always available and satisfy the purpose of credit scoring for SMEs, many researchers have also tried to build the model by using a set of non-financial features, either as a substitute or complement for the financial features.

Bensic et al. [8] examine some statistical techniques to build a credit scoring model based on the owner's profile, small-business activities, and loan information. They used non-financial features such as owner's age, owner's occupation, amount and length of the loan, interest rate, repayment method, location and business sector to build the model. The research emphasized the importance of loan information as well as owner's personal and business characteristics in SMEs credit scoring model.

Sohn et al. [35] aim to better represent SMEs with the high degree of potential growth in technology by developing the first technology-based credit scoring model. The model was built using logistic regression on four non-financial aspects: management ability, technology level, marketability of technology, and profitability of technology; and producing 93% recall and 68% precision.

Lee et al. [25] proposed an accounting ethics-based SME credit scoring model which can reduce the default rate resulting from the moral hazard associated with unethical accounting behaviors. The research analyzed 16 accounting behavior related features representing the internal control structure, financial transactions, related parties, business ethics, and other unethical conduct to build the model. From the final model, five features were reported to have significant predictive power in estimating SMEs default, including the accuracy of accounting records, amounts of cash in hand, loans to related parties, loans to affiliated companies, and compensation to the majority shareholder.

Altman et al. [5] use some non-financial features in addition to some financial ratios to develop a default prediction model for SMEs which have insufficient financial information. The non-financial features used include age, size, sector and other features to measure operational risk. The inclusion of those non-financial features leads to significant improvement in the prediction accuracy of the model by up to 13%.

[1] $Z" - score = 6.56X_1 + 3.26X_2 + 6.72X_3 + 1.05X_4$, where:
X_1 = working capital/total assets,
X_2 = retained earnings/total assets,
X_3 = EBIT/total assets, and
X_4 = book value equity/total assets

Pederzoli et al. [34] was the first one to combines innovation-related features, i.e R&D productivity and value of the patent, with financial features to predict the default event on SME lending. Five financial related features used are equity/total debt, cash/total asset, EBITDA/total assets, EBITDA/interest expense, and revenue, which each of them consecutively represents equity, liquidity, profitability, coverage, and sales. The research found that the addition of innovation related features measures to the financial data can increase the accuracy of the model by 4.5%.

Ciampi et al. [13] implemented SME credit scoring model by combining the geographical area, business sector and the size of SMEs with the financial features. For the financial features, they selected 11 financial ratios calculated from accounting data, including cash flow-to-debt ratio, debt-to-capital ratio, acid test ratio, interest expense/turnover, current ratio, equity/fixed assets, ROI, net margin, long term assets/number of employees, and interest expense/bank loans. They found that size, business sector, and geographical area are influential to SMEs credit scoring model as the model that include those features has significantly higher accuracy compared to the model developed only using the financial features.

2.2.2 Credit Scoring Techniques

The variety of techniques used to build the credit scoring model for SMEs are generally similar to techniques applied in to build credit scoring model for consumer and corporate. For many years, discriminant analysis was the prevalent statistical technique applied in the credit scoring model for SME. Edward Altman [2] pioneered the application of multivariate discriminant analysis (MDA) for predicting corporate bankruptcy using accounting data of healthy and bankrupt companies. He formulated the discriminant function to distinguish healthy companies from those that likely to bankrupt. The company is predicted to bankrupt if its score produced by the discriminant function exceeds a specific threshold; otherwise it is predicted as a healthy company. This method was initially designed for the medium and large company until 30 years later the discriminant function specifically for was developed [3].

Another conventional technique that also has been used widely in credit scoring is logistic regression. The technique is preferred because it is computationally cheap and unlike MDA, it does not require the data to be normally distributed. Altman et al. [4] used logistic regression to build SME default prediction model based on a set of financial ratios and compare the accuracy with the model built using MDA. The research reported that the logistic regression model produces higher accuracy compared to the MDA model. The performance of logistic regression model can also be boosted by performing a transformation to the predictors in order to make their distribution closer to normal.

Recently, more modern machine learning techniques, including support vector machines (SVM), decision trees, and artificial neural network (ANN), are adopted SME credit scoring. Huang et al. [22] used SVM with RBF kernel in evaluating the applicant's credit score. Compared to neural networks, genetic programming, and decision tree classifiers, the SVM classifier achieved a similar performance with relatively few input features. Ciampi et al. [13] implemented SME credit scoring model using various machine learning technique, including ANN, logistic regression, and MDA. The ANN model performs better compared to the other credit scoring model which implemented using logistic regression and discriminant analysis. Bastos [37] evaluated the credit scoring models based on a variant of gradient boosting machine, extreme gradient boosting (XGBoost), across five credit datasets against other models produced using current techniques, including logistic regression (LR), neural network (NN), decision tree (DT), support vector machine (SVM), AdaBoost, and random forest (RF). The proposed XGBoost-based credit scoring model achieves promising performances by producing highest accuracy in all of the datasets used.

Usually, the number of sample for default borrower is significantly lower than the number of sample for solvent borrower, which cause the model prone to overfit. Marques et al. [27] solve this problem using re-sampling techniques to balance the number of sample before applying logistic regression algorithm to build the model. The research benchmarked seven resampling algorithms, including various under-sampling and over-sampling techniques, such as One-Sided Selection (OSS), Neighborhood CLeaning rule (NCL), random under-sampling (RUS), under-Sampling Based on Clustering (SBC), random over-sampling (ROS), Synthetic Minority Over-sampling TEchnique (SMOTE), Safe-Level SMOTE, and combination of SMOTE and data cleaning, called SMOTE+WE. The experimental results demonstrate that in general, over-sampling techniques perform better than any under-sampling approach with the SMOTE+WE top the result.

Brown et al. [11] studied how different models based on techniques that commonly used in credit scoring application are affected by extremely imbalance dataset. They used some imbalance credit scoring dataset to compare the AUC of nine classification algorithms: linear discriminant analysis (LDA), logistic regression, neural network (NN), C4.5 decision tree, quadratic discriminant analysis (QDA), least square SVM, k-nearest neighbours, random forests ,and gradient boosting trees. Compared to the other algorithms, random forest and gradient boosting trees performed well in dealing with data where a massive class imbalance was present.

2.3 Social Media Usage in Credit Scoring

Creditors are using the same basic information they have used for decades to determine creditworthiness. While these valuable pieces of information will likely remain at the core of credit scoring, a wider array of data with the potential to provide more accurate risk scoring is now available. This alternative data is driven largely by the explosion of social media which provides unprecedented amounts of information that can support decisions about an applicant's creditworthiness. To the best of our knowledge, although previous researchers already used social media data in credit scoring, these model are limited only to evaluate personal or consumer lending.

Masyutin et al. [28] used the data collected from VKontakte, Russian's most popular social media platform, to discriminate between the solvent and delinquent borrowers in personal lending. Applicants demographic data and some social media metrics, such as the number of posted video, number of posted photos, and number of days since the last post, were used to build the scoring model. The research shows that social data is better in predicting fraudulent cases, which the borrower initially intends to receive a loan and not to pay any installments right from the start, and confirms that social media data may be used to enrich the classical application of credit scoring.

Tan et al. in their work [36] proposed a social media based credit scoring model for microloans. Demographics, preferences and network information collected from borrowers' Facebook profile pages were used as predictors to build the model. The demographic information is represented by age, gender, marital status, religion, education, and location; the preferences information is obtained from borrower's interests and groups; the network information is extracted through borrower's social interactions on Facebook. Even though the input data used were solely from the social media platform, the performance of the model is at the acceptable level.

Zhang et al. [38] construct a consumer credit scoring model by fusing social media information collected from social platform PPDai with the information that are used in the credit scoring traditional model. The model is able to catch 86.02% default customer with overall accuracy 84.86%. Both studies show that loan information, social media information, and credit information are important factors for predicting the default. The research provides empirical evidence that repayment prediction can be improved by incorporating social network metrics.

2.4 Social Media Metrics and Business Evaluation

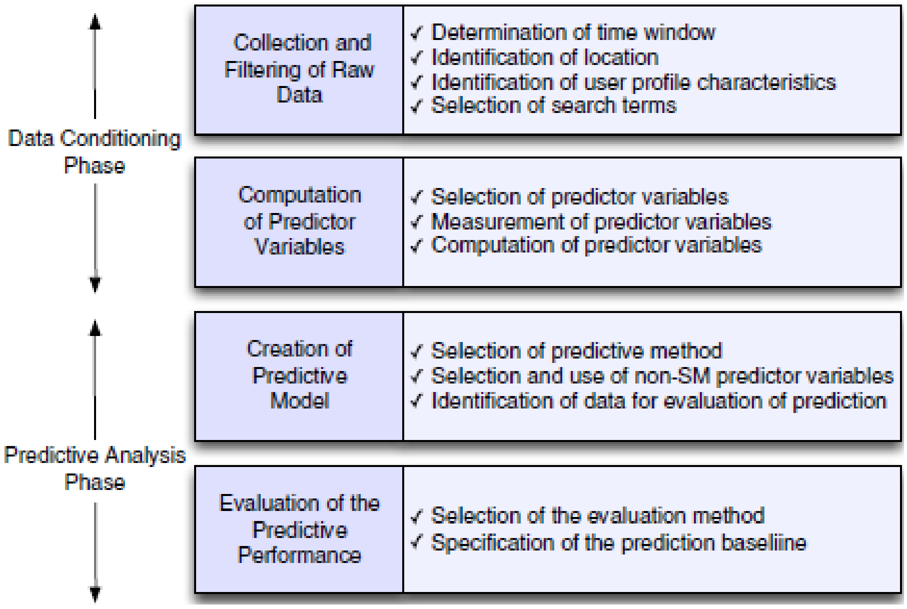

Data Conditioning Phase
Collection and Filtering of Raw Data
✓ Determination of time window
✓ Identification of location
✓ Identification of user profile characteristics
✓ Selection of search terms
Computation of Predictor Variables
✓ Selection of predictor variables
✓ Measurement of predictor variables
✓ Computation of predictor variables
Predictive Analysis Phase
Creation of Predictive Model
✓ Selection of predictive method
✓ Selection and use of non-SM predictor variables
✓ Identification of data for evaluation of prediction
Evaluation of the Predictive Performance
✓ Selection of the evaluation method
✓ Specification of the prediction baseliine

Figure 2.1: The social media data framework for predictive analytics [24]

It has been proven that social media can be used to improve traditional credit scoring. However, we have not found any studies which propose a set of metrics derived from social media data as predictors in the enterprise credit scoring model. In this section, we present previous research which utilizes social media data for evaluating the performance of a business entity.

Kalampokis et al. [24] review the use of social media data for predictions in various areas, including product sales and stock market volatility, and propose a conceptual framework in using social media data for predictive analytics. The framework consists of four stages as shown in Figure 2.1. The first two stages are called data conditioning phase, while the last two stages are called predictive analysis phase. The first stage is raw data collection and filtering, which includes determining the platform as the data source, time window, and search term to get the desired information. The second stage is computing predictor variable, which includes selecting social media metrics as predictor variables and calculate desired variables from raw social media data. Generally, the variables can be grouped into 3 categories: volume-related variables, such as the number of tweets or likes; sentiment-related variables; and profile characteristics of the users, such as the number of followers and the location. The third stage is predictive model creation, which includes selecting a set of methods used in modeling, combining the predictors with non social media predictors, and applying some statistical techniques to build a prediction model. The last stage is the model evaluation, which includes choosing evaluation metrics and defining the specification of the baseline model to compare.

Neiger et al. [32] proposed key performance indicators (KPIs) and related evaluation metrics to evaluate the use of social media in the promotion. There are 4 KPIs proposed in this study: (1) insight, consumer feedback which for example can be measured from its sentiment; (2) exposure, measured by the number of visit, rating, comments, and other metrics related to how many times the contents are viewed; (3) reach, measured by the amount of likes, unsubscribes, demographics of fans and other metrics related to quantity and diversity of people who have been in contact with; and engagement, which measured by the number of retweet, mention, like/dislike and other metrics related to the quantity of people who participate in creating, sharing, and using the content.

Bonson et al. in [10] proposed a set of metrics to assess the company's customer engagement and mood using the data collected from the company's Facebook page. They formulated engagement metrics based on three aspects: popularity, commitment, and virality. Popularity is measured by the number of likes on Facebook, commitment refers to the number of comments and virality is measured by the number of shares on Facebook. For all of those reactions, 3 social media metrics are derived: percentage of total post reacted, the number of reaction per post, and the number of reaction per post per 1000 fans. While customer mood is measured by conducting sentiment analysis to the comments and calculate the ratio between the positive, negative and neutral moods expressed.

Luo et al. [26] studied the predictive relationships between social media and firm equity value by analyzing customer review and rating from CNET.com and Lexis-Nexis. The research suggested that social media-based metrics are significant leading indicators in comparison to conventional behavioral metrics. As firm equity is also usually used to calculate some financial ratio in credit scoring, adopting social media metrics in credit scoring for SMEs can be beneficial. The research also reported that the social media metric has a shorter time to reach the peak of its predictive power, which means the data collected from social media will be obsolete faster than the other conventional behavioral metrics.

Oztamur et al. [33] analyzed the role of social media for SMEs from the perspective of firm marketing performance. In this study, the number of likes and followers, richness of content, interaction with customers and the use of language are chosen as criteria to asses the SMEs. In [29], McCann et al. compiled the measurement of social media in mainstream academic literature and other business-oriented publications. Some social media metrics are discussed, such as duration needed to resolve customer service request, number of click-throughs that lead to purchase from social media platform, the volume of mention across channels, and the amount of unique visitor.

2.5 Chapter Summary

Credit scoring is a process of assigning a quantitative measure to the potential borrower as an estimate of how likely the default will happen in the future. The most popular techniques used in credit scoring are discriminant analysis and logistic regression, even though nowadays the more modern machine learning techniques such as Artificial Neural Network (ANN) and Gradient Boosting Trees have been reported to outperform them. Credit scoring for SMEs traditionally uses financial ratios derived from accountancy data as predictors. Some non-financial information used to complement the financial ratios because the financial data often is unavailable and unable to satisfied the purpose of credit scoring. Many researchers have reported that the addition of non-financial data, for example business sector and business, as predictors can increase the performance of the credit scoring model for SMEs. We present the research related to credit scoring for SMEs in Table 2.1.

An alternative source to get non-financial information is from social media. It provides unprecedented amounts of information that can be used to estimate an applicant's creditworthiness, including SMEs lending. However, the question about what kind of information from social media that is relevant to be used in SMEs credit scoring is still unanswered as the adoption of social media in credit scoring nowadays is limited to the personal lending only. Some research formulated various metrics calculated from social media data to evaluate the performance of companies. The metric used in those previous studies can also be calculated for SMEs by collecting the data from their social media account and may also be good predictors in evaluating the creditworthiness of SMEs. The research related to social media metrics and features used in personal credit scoring and business evaluation is summarized in Table 2.2.

Table 2.1: Previous Work in SMEs Credit Scoring

Research	**Techniques**[2]	**Features**
Edmister (1972)	MDA	quick ratio, cash flow/current liabilities, inventory/sales, net working capital/sales, current liabilities/equity, and equity/sales
Altman et al. (2005)	MDA	net working capital/total assets, retained earnings/total assets, EBIT/total assets, and equity/total assets
Bensic et al. (2005)	LogR, ANN, DT	loan details, owner's personal and business related features
Sohn et al. (2005)	LogR	16 features measuring management ability, technology level, marketability of technology, and profitability of technology
Altman et al. (2007)	LogR	EBITDA/total assets, EBITDA/interest expenses, cash/total assets, short-term debt/equity, retained earnings/total assets
Altman et al. (2010)	LogR	cash/total assets, EBITDA/total assets EBITDA/interest, Retained earnings/total assets, short-term debt/equity, SME's age, size, sector and operational related feature
Ciampi et al. (2013)	MDA, LogR, ANN	cash flow/total debt, debt-to-capital ratio, acid test ratio, interest expense/turnover, current ratio, equity/fixed assets, net margin, long term assets/number of employees, ROI, interest expense/bank loans, size, business sector, and geographical area.
Pederzoli et al. (2013)	LogR	equity/total debt, EBITDA/total assets, cash/total asset, EBITDA/interest expense, revenue, R&D productivity, patent's value
Lee et al. (2017)	LogR	16 accounting behavior features related to internal control structure, financial transactions, related parties, business ethics, and other unethical conduct

ANN = Artificial Neural Network, DT = Decision Trees, LogR = Logistic Regression, MDA = Multivariate Discriminant Analysis

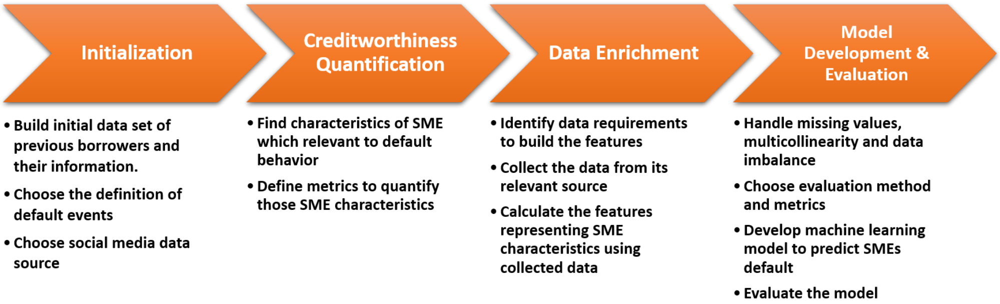
Table 2.2: Social Media Metrics and Features Used in Related Works

Research	Objective	Platform	Social Media Features
Masyutin et al. (2015)	Build credit scoring model for personal loans	VKontakte	• political affiliation • number of photos or videos post • number of days since last visit • number of days since last post
Tan et al. (2016)	Build credit scoring model for microloans	Facebook	• demographics information • followed page and preference • a set of network metrics
Zhang et al. (2016)	Build credit scoring model for personal loans	PPDAI	• amount of platform currency • contribution and reputation score • score of group belonged
Neiger et al. (2012)	Develop KPIs and metrics to evaluate the use of social media in promotion	Facebook, Twitter	• volume and type of feedback • number of views, comments, shares • number and growth rate of fans • ratings, number of likes or dislikes • number of retweets, mentions
Bonson et al. (2013)	Develop metrics to assess stakeholders' engagement and mood of a corporate Facebook page	Facebook	• percentage of liked, shared, commented post • average of likes, shares, comments • average of likes, shares, comments per fans • ratio positive and negative comment
Luo et al. (2013)	Predict firm equity value	CNET, LexisNexis	• volume and sentiment of review • volume and level of rating.
Oztamur et al. (2014)	Develop metrics for evaluating SMEs marketing.	Facebook, Twitter	• number of fans or followers • frequency of updates • content richness and relativeness • interactiveness

Chapter 3

Credit Scoring Framework

This chapter describe our credit scoring framework to build credit scoring model which incorporate social media data. The framework consists of four stages: initialization, creditworthiness quantification, data enrichment, and model development and evaluation. Here, we explain those stages in the four separated section. Figure 3.1 below shows the illustration of the framework that we design.



Initialization
Creditworthiness Quantification
Data Enrichment
Model Development & Evaluation

Initialization:

- Build initial data set of previous borrowers and their information.
- Choose the definition of default events
- Choose social media data source

Creditworthiness Quantification:

- Find characteristics of SME which relevant to default behavior
- Define metrics to quantify those SME characteristics

Data Enrichment:

- Identify data requirements to build the features
- Collect the data from its relevant source
- Calculate the features representing SME characteristics using collected data

Model Development & Evaluation:

- Handle missing values, multicollinearity and data imbalance
- Choose evaluation method and metrics
- Develop machine learning model to predict SMEs default
- Evaluate the model

Figure 3.1: Credit Scoring Framework

3.1 Initialization

3.1.1 Initial Dataset

The initial dataset is a primary requirement for our credit scoring framework as it can affect the decision making in the next stages of the framework. For example, it can influence which features can be used to measure SMEs creditworthiness. It is because the data that we can collect depends on the information that we have in the initial dataset. Having an initial dataset can also affect us in selecting the suitable source of our social media data. For example, if the SMEs in the initial dataset is working in e-commerce sector, the social media data source that provide us with the comments and review from their customer such as Trustpilot can be preferable.

We conducted this project in collaboration with Exact, who provided us with an initial dataset containing 218,778 Dutch SMEs. The general information regarding SMEs including the name, address, industry sector, size, website, and Chamber of Commerce (KvK) registration number are available in the initial dataset. Exact also provided us with the SMEs accounting data which comply with the Reference Classification System of Financial Information (RCSFI) or in Dutch ”Referentie GrootboekSchema” (RGS). The RCSFI is a standard classification in business reporting to help comparing the information between companies and reducing the number of reports that an entrepreneur needs to create [12]. Exact developed a system which map general ledger (G/L) accounts in their customers bookkeeping system into RGS classes.

3.1.2 Defining Default and Time Horizon

The SMEs credit scoring can be treated as a binary classification problem of predicting whether an SMEs will default in the future. Similar to other classification problem, first we need to define the label. In the context of credit scoring, the label is a condition when a debtor cannot meet the obligation to repay, which commonly referred to as default event. Traditionally, credit scoring models were developed using the bankruptcy as the definition of the default event. However, the creditors also suffer from losses before the bankruptcy occurs, for example when the borrowers unable to repay the debt in time. Therefore, some earlier events, such as failure to repay in 90 days since the payment deadline, are also often used to indicate a default event in credit scoring.

In this project, we use the bankruptcy event as the definition of default events, which is also the label for our classification problem. We choose bankruptcy as default definition because a lot of previous studies in SMEs credit scoring also use the same definition. Moreover, the credit scoring model which is created using bankruptcy event as the definition of default definition can also be applied to predict the default defined using earlier default event without losing too much its prediction power [20].

Besides the label, we also need to define the time horizon as the features that describe the creditworthiness of SMEs is changing over time (time series). There are two properties that we need to consider, input horizon and prediction horizon. Input horizon specify the past data that is still relevant to develop the model, while prediction horizon refers to the period over which the default probability is estimated in the future. Figure 3.2 illustrate the timing diagram in predictive analysis using time series data.

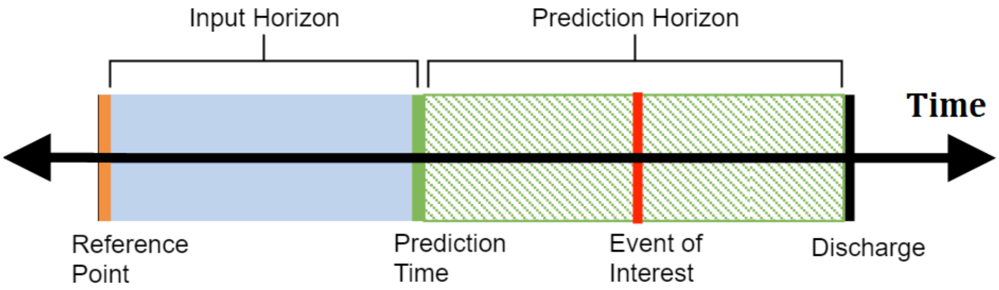


Figure 3.2: Timing diagram in predictive analysis using time series data

The choice of the time horizon is a key decision to build a credit scoring model, which varies depending on the objective for which the credit risk model was developed and also the data used to build the model. For example, banks usually use one year as their predicting horizon, which is long enough for them to take some actions in mitigating the credit risk. For the modeling horizon, it is really depends on the institution and the type of information used to build the model. We choose six month for both model horizon and prediction horizon in this project. The reason is social media features that will be used to build the model has a shorter time window in which it is still relevant compared to conventional behavior metrics [26].

Based on the definition above, the credit scoring model that we develop in this project will predict the likelihood of the SMEs to go bankrupt in the next six months using their data from previous six months.

3.1.3 Social Media Platform Selection

Every time SMEs and their customers use engage in social media, they create digital footprints which can be useful to determine both capacity and willingness to repay loans. There are a lot of social media platforms which contains extensive information related to SMEs, including Yelp, TripAdvisor, Twitter, LinkedIn, and Facebook. Every social media platform has their own characteristics that need to be considered when choosing it as our data source. The initial data that we have is also affect our choice of social media platform. Some criteria that can be used to select the suitable social media platform for credit scoring model are the popularity of social media platform especially related to the activity area of SMEs in our initial dataset, the possibility of adopting previous related research, and the availability of the information that we want to collect.

We decide to use Facebook as our social media data source for various reasons. In terms of popularity, Facebook is the most popular social media platform used by the SMEs, followed by LinkedIn and Twitter [31]. It is also still relevant for the population of SMEs in our dataset as it is not limited to any specific activity area. Also, Facebook is the most commonly used platform in credit scoring and business performance evaluation according to our literature survey. By using Facebook, we can adopt some approach used in those previous work, for example collecting the similar information or calculate similar features. In term of the availability of data, Facebook allow company to create their own business page, which is different from personal page. The information on the company business page can be collection using Graph API, a public API to read and write to the Facebook social graph. The API also support search function which help us to find and match the Facebook page with the SMEs in our initial dataset.

3.2 Creditworthiness Quantification

In the previous chapter, we surveyed various approach used by previous studies to build credit scoring model for SMEs and various social media features used in credit scoring and business performance evaluation. Here, we explain which features from those previous studies that we adopt to quantify creditworthiness and build credit scoring model for SMEs.

3.2.1 Traditional Features

In the previous chapter, we already present some previous studies related to credit scoring for SMEs. We refer the features which commonly used in SMEs credit scoring based on those studies as traditional features. Traditional features for SMEs credit scoring basically can be categorized into financial and non-financial features. The most commonly used non-financial features are business sector and size. For the financial features, a set of financial ratios derived from accounting data is the most commonly used. The researchers usually use financial ratio to gauge the following SME's condition:

Profitability, indicates the ability of a firm to generate a profit to develop their current business. A higher profitability ratio means that the SMEs can manage expenses effectively to create more profit.

Leverage, indicates how firm's assets and business operations are financed (using debt or equity). A high leverage ratio indicates that the SME may have incurred a higher level of debt than it can be reasonably expected to service with ongoing cash flows.

Liquidity, examine the ability of a business to pay off its short-term debt. The higher the liquidity ratio, the better the ability of an SME of pay off its obligations in a timely manner.

Interest Coverage, measures the ability of a business to pay the interest on its outstanding debt. A low interest coverage ratio indicates a strong indicator that an SME may default on its loan payments.

Table 3.1 shows the list of traditional features that we use in this project. We use some ratios for measuring the profitability, leverage, and liquidity condition due to the limited financial features that are available in our dataset to calculate those ratios. We calculate three profitability ratios (net margin, EBIT margin and return on asset), two leverage ratios (debt ratio and debt-to-capital ratio), and two liquidity ratios (current ratio and quick ratio) as the features to quantify SME creditworthiness. We also include the financial measures which are used to calculate those ratios in our traditional feature set because we suspect that they can also represent the SMEs financial condition.

We also adopt sector and size in our traditional feature set because they are also available in our initial dataset. The sector information in our dataset follows a standard of sector classification called the Dutch Standard Industrial Classification 2008 or in dutch "Standaard Bedrijfsindeling 2008"(SBI 2008). This standardized classification is based on the activity classification used by the European Union (NACE) and on the United Nations (ISIC). The industry sector in SBI 2008 is divided into 21 main categories, and each main categories is divided into several subcategories. We only use the main categories as a feature in developing our SMEs credit scoring model. The SME's size in our dataset measures the size based on the number of employees, which is categorized into eight categorical values. For more details description about the values of both size and sector variable can be read in Appendix A.

3.2.2 Social Media Features

To quantify the creditworthiness of SMEs using social media data, we proposed some new features and combined them with the other features that have been used in the previous credit scoring and business evaluation studies. Those features are derived

Table 3.1: Traditional features to quantify SMEs creditworthiness

Name	**Description**
equity	book value of equity
total assets	total assets owned
current assets	assets that can be converted into cash within a year
total liabilities	total obligations and debts owned
current liabilities	obligations that are due within one year
cost	amount of money that has been used up to produce revenue
tax	amount of tax paid
revenue	amount of money received from the sold products
EBITDA	earnings before interest, taxes, depreciation and amortization
EBIT	earnings before interest and taxes
net profit	earnings after all of the expenses
current ratio	current assets/current liability
quick ratio	current assets without inventory/current liability
net margin	net profit/revenue
EBIT margin	EBIT/revenue
return on asset	net profit/total assets
debt ratio	total liability/total assets
debt-capital ratio	total liability/(total liability+equity)
sector	SMEs category based on SBI 2008
size	classification based on number of employees

from the data types which available on the social media platform that we choose. Each of those data types contains some properties that we use to calculate the social media features. For example, comments data can consist of message, date created, reaction, and post id related to the comment. In our case, the Facebook data types which we use to derive social media features are user, post, visitor post, and comment.

The first set of features that we adopt are the number of photo posts, number of video posts, and the number of days since the last post. Those features are proposed by Masyutin et al. in [28] for building credit scoring model for personal client. We then propose some similar features to those, such as the number of story posts, number of days since the last comments, and number of days since the last visitor posts.

Next, we also use the number of fans, number of mentions, ratings volume and rating level, which are adopted from Neiger et al. [32] as social media features. We also follows the approach proposed by Bonson et al. [10] which calculate the average number of reactions, shares, and comments per post and the percentage of reactions, shared, commented posts, the percentage of negative comments and the percentage of positive comments on the Facebook page. Because currently Facebook has six different reaction(like, love, haha, wow, sad and angry.), we divide the reaction related features into positive and negative. We assume the like, love, haha, and wow as a positive reaction to a posted content on Facebook, while sad and angry as a negative reaction. In addition to that, we also use build similar features based on visitor posts such as the percentage of positive and negative visitor posts, and the number of visitor posts.

Lastly, we propose a set of trend based features that measures the gradient of some features related to post, share and comment during the six months of input horizon. We calculate the trends for the following features: posts, shares, comments, P reaction, N reaction, shared, commented, P reacted, N reacted, P comments, and N comments.

Table 3.2: Social media features to quantify SMEs creditworthiness

Feature Name	**Description**
fan counts	number of follower in Facebook [32, 33]
talking about count	number of content which mention the page [32]
rating count	number of rating submitted [32, 26]
overall star rating	average rating submitted [32, 26]
posts	number of content created [33]
shares	number of share in their posts [10, 32]
comments	average number of comments per posts [10]
P reaction	average number of positive reaction per post [10]
N reaction	average number of negative reaction per posts [10]
shared	percentage of post which is shared [10]
commented	percentage of post which has comments [10]
P reacted	percentage of post with positive reaction [10]
N reacted	percentage of post with negative reaction [10]
photo posts	percentage of posts which contains photo [28]
video posts	percentage of posts which contains type video [28]
story posts	percentage of posts which contains text only
visitor	number of content created by others in the page
P vpost	percentage of positive visitor posts
N vpost	percentage of negative visitor posts
P comments	percentage of positive comments in their posts [10, 26]
N comments	percentage of negative comments in their posts [10, 26]
SL post	number of day since their last post [28]
SL visit	number of day since last visitor post
SL comment	number of day since last comments
t-posts	trend of number of content created in 6 months
t-shares	trend of number of share in their posts in 6 months
t-comments	trend of average number of comments per posts in 6 months
t-P reaction	trend of average number of positive reaction in 6 months
t-N reaction	trend of average number of negative reaction in 6 months
t-shared	trend of percentage of post which is shared in 6 months
t-commented	trend of percentage of post which has comments in 6 months
t-P reacted	trend of percentage of post with positive reaction in 6 months
t-N reacted	trend of percentage of post with negative reaction in 6 months
t-P comments	trend of percentage of positive comments in 6 months
t-N comments	trend of percentage of negative comments in 6 months

3.3 Data Enrichment

The initial dataset that we already have can be insufficient to produce the features we want for quantifying the SMEs creditworthiness. The problem can be either the data related to the features is not available, the data is available but there are some missing values, or the data is completely available but we are not sure that all of the information is valid. If that the case, the data enrichment stage is required to validate the currently available data and collect more data to create the features for building SMEs credit scoring model. The data enrichment stage consist of three part: formulating the requirements, data collection, and feature construction.

3.3.1 Formulating Data Requirements

The data requirements specify the information that need to be collected in order to validate and complete the initial dataset. For each of those data, we also need to find relevant data source. Some ”intermediate” data may also be required to be collected first for collecting the other information. For example, to collect the bankruptcy label and non-financial traditional data, a valid information of KvK number, name and address is required. When the data requirements is fulfilled, we should be able to build the features specified in the creditworthiness quantification stage.

The information related to all of the legal entities that participate in economic transactions in the Netherlands, including SMEs, are available in the Dutch Chamber of Commerce website, `www.kvk.nl`. We can use it to complete and validate the SMEs' basic information, i.e. KvK number, name, address and website. Those valid basic information, especially KvK, can be used to collect bankruptcy label and non-financial traditional data (business size and sector) from other source. The information related to bankruptcy status of business is also available in some public websites, such as `faillissementsdossier.nl` and `drimble.nl`, which can be searched based on KvK number or name. We can also search brand names, business sector and size of based on KvK number, name and address in the following websites: `bedrijveninzicht.nl`, `opencompanies.nl`, and `drimble.nl`.

Other information that we need to collect are SMEs' social media data from the Facebook pages. Before we can collect SMEs' social media data from Facebook, we need to have knowledge about their Facebook pages, which represented by Facebook IDs. As mentioned earlier, Facebook provide us with a public API that can be used to find SMEs' Facebook ID and collect the data from their page. Alternatively, we can also use the SMEs websites as our data source to collect their Facebook ID because usually companies listed their contacts, including social media IDs, at their websites. Table 3.3 shows the details regarding what kind of information that need to be collected together with its source and other information is required to get them.

3.3.2 Data Collection

After the requirements is specified, the next step is collecting the data. The method used to collect the data depends on its data source. For example, if the data source provide an API to collect the data, such as Facebook, using API is preferred as it is easier to do. However, if the data is available in public website which does not provide

Table 3.3: Data collected from publicly available sources

Information	Source	Input
Basic Information (name, address, website)	`kvk.nl`	KvK number
Bankruptcy Label (status, date issued)	`faillissementsdossier.nl`, `drimble.nl`	KvK number
Other Information (sector, size, brand)	`opencompanies.nl`, `bedrijveninzicht.nl`, `drimble.nl`	KvK number name, address
Social Media ID	Facebook API website	name, address
Social Media Data (post, comments, etc.)	Facebook API	Facebook ID

API to access them, then the scrapping method can an option. Here, we will not explain the data collection part in detail as it will be explained in the next chapter.

3.3.3 Feature Construction

Once all of data has been collected, we can construct the features that we need to build credit scoring model for SMEs. If the data is really big to be processed in a single machine, some special data processing tools is needed. One of the alternative is using cloud computing platform, such as AWS EC2. Cloud computing platform can provide us with virtual machine with the capabilities of processing large amount of data in single machine. If the data is still too big for the highest virtual machine available in cloud platform, Apache Spark can be the solution. Apache Spark is a cluster computing platform which provides a faster and more general data processing.

3.4 Model Development and Evaluation

3.4.1 Handling Outliers and Missing Values

Credit scoring dataset can contain outliers that must be managed properly. Outliers are values that are located far from others for a certain characteristic and may negatively affect the modeling results. The easiest way of dealing with outlier is removing all extreme data that fall outside of the normal range, for example at a distance of more than two or three times the standard deviation. This solution is dangerous as it can remove default SMEs which are erroneously considered as outliers. Another technique which can also be used is winsorisation. The technique transforms and limits extreme values considered as outliers to a specified percentile of the data.

Several methods for dealing with missing values are available, such as removing all records with missing values or excluding records that have significant missing values (e.g., more than 50% is missing). However, this would result in too many data lost. Another way is substituting the missing values with corresponding mean or median values of overall observations. All of these methods assume that no further information can be gathered from analyzing the missing data, which is not necessarily true. Missing values may be part of a trend, may be related to other characteristics, or may indicate bad performance. Therefore, missing values should be analyzed first, and if they are found to be random, they may be excluded or imputed using statistical techniques; otherwise, if missing values are found to be correlated to the performance, it is preferable to include missing values in the analysis.

3.4.2 Feature Selection

Feature selection is the process of choosing a subset of the available features by eliminating those that are either redundant or possess little predictive information. By performing feature selection, we can produce a simple credit scoring model with optimal predictive power and avoid over-fitting because of the curse of dimensionality, which occurs when too many irrelevant and redundant features used but not enough instances is available to describe the pattern. Feature selection also can help us identifying predictive features, which can provide clearer insight and a better understanding of the credit scoring.

We combine three feature selection technique in this project. First, features with extremely low variance, weak predictive power or is illogical will be screened out. Next, we conduct correlation based features selection. It is necessary because the model will be biased if some highly correlated indicators are included in the model. We measure the correlation among features to identify some highly correlated groups of features and then choose one or more variables from each group, which can represent all of the information contained in those group. Finally, we perform recursive feature selection by iteratively building the prediction model and removing less important features. A set of features which produce the best performance is chosen to build the final model.

3.4.3 Handling Class Imbalance

Imbalanced class distribution is a common problem in the credit scoring applications. For example, it is common to find the number of default samples to be less than 10% of overall samples. The solutions to deal with the class imbalance problem can be implemented either at the algorithmic or data levels.

At the algorithmic levels, we can set more weight on minority samples which usually inversely proportional to the ratio of its population. By setting more weight on minority class, the classifier will penalize misclassifying the minority samples more than misclassifying the majority samples during training process. However, giving too much weight to minority class can result in over-fitting. As the classifier only learn from limited number of sample, it unable to generalize if a new minority data is present in the validation set or test set.

At the data level, we can apply various resampling algorithms to change the class distribution of the data. It can be done by either over-sampling the minority class or under-sampling the majority class until both classes are fairly represented. While under-sampling may potentially throw away useful information, oversampling worsens the computational load and creates noise. Usually, over-sampling is preferred as it outperforms under-sampling in most cases.

The most popular oversampling algorithm that mostly used classification problem is Synthetic Minority Over-sampling TEchnique (SMOTE). We decided to use a variant of this algorithm, SMOTE+Tomek Links, in this project to deal with the imbalance dataset. This algorithm generates artificial samples from the minority class by interpolating existing instances that lie close together. It finds the *k* nearest neighbors samples belonging to the minority samples and then generate the synthetic examples in the direction of some or all of those minority samples [21]. Afterwards, it minimize the class overlap by removing some instances based on the Tomek Links.

3.4.4 Selecting Classification Algorithms

Numerous classification algorithms or classifiers for developing credit scoring model have been presented in the previous literature. Those classifiers not only differ in their performance, but also differ in their characteristics. The characteristic that we usually consider in selecting classifier is the interpretability of the model. This interpretability of the model becomes more important due to the General Data Protection Regulation (GDPR) in the European Union (EU), which grant the subject of the model with the right to an explanation. Based on the interpretability of the model, we can categorize classifiers into two categories: ”white-box” and ”black-box” algorithm. In this project, we choose a white-box algorithm, logistic regression, and a black-box algorithm, XGBoost, to develop credit scoring models for SME and then compare their performance.

Logistic regression is the most commonly used algorithm in the credit scoring, which is the main reason why we choose it. People often use it to solve various binary classification problem due to its speed and the interpretability of the model that it produce. Logistic regression assumes that the probability of classes is logistically distributed. Similar to the other regression algorithm, logistic regression calculates a continuous response variable through the linear combinations of predictor variables. The response variable in logistic regression is the $log(odds\_ratio)$ as shown by the equation below.

$$log(odds\_ratio) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \tag{3.1}$$

The probability of classes can be calculated by transforming the equation above into the following equation.

$$p = \frac{1}{1 + exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)]} \tag{3.2}$$

XGBoost is an open source implementation of gradient boosting algorithm, which is fast, efficient, and scalable. The model for XGBoost is tree ensembles which consist of classification and regression trees, which can be used to solve any supervised learning problems. We are interested in using this algorithm because it achieves promising results on numerous standard classification benchmarks. Also, numerous winning solutions in machine learning competitions of Kaggle employed XGBoost to train their models .

3.4.5 Model Evaluation

Model evaluation is essential to the development process of a credit scoring model as it assess the performance of each modeling procedure. To evaluate the model, we need to choose the evaluation metrics and methods, which depends on the modelling purpose and also the data that we have.

In this project, we use some evaluation metrics to compare the model, they are F1 score, ROC curve, and Matthews correlation coefficient (MCC). Those metrics are chosen because they are suitable for binary classification where the class distribution is imbalance. For the method, we use 5-fold cross-validation to validate the model. We choose cross-validation because it can test how the model will generalize when predicting new data that was not used in estimating the model.

F1-score is the harmonic mean between the precision and recall. The purpose of using this metrics to choose the model is to get the larger number of correctly classified positive instances (true positive, TP) and also keep the number of incorrectly classified negative instances (false positive, FP) low. This metrics can be calculated using the following equation.

$$F1\text{-}score = 2\frac{Precision \cdot Recall}{(Precision + Recall)} = \frac{2TP}{(2TP + FP + FN)} \tag{3.3}$$

A receiver operating characteristic (ROC) curve is a plot which displays how the number of correctly classified positive instances varies with the number of incorrectly classified negative instances. Each point on the ROC curve represents a classification threshold that corresponds to particular values of the false positive rate, and true positive rate. To compare the result between two model, we usually compare the area undor the ROC curve. The larger area under the ROC curve of a model, the better its performance.

The Matthews correlation coefficient (MCC) is used as a measure of the quality of binary classifications, which takes into account the confusion matrix of the prediction result. The MCC returns a value between -1 and +1. A value of +1 represents a perfect prediction, 0 no better than random prediction and -1 indicates total disagreement between prediction and label. The MCC can be calculated using following formula.

$$MCC = \frac{(TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.4}$$

Chapter 4

Data Collection

This chapter elaborates how we collect our data, which is organized into five section. Section 4.1 elaborate the data model that we use to store and organize the data in our project. In Section 4.2, we describe how the data is collected from public website. Section 4.3 explain process to find Facebook ID for each SMEs. Section 4.4 is about social media data collection and social media feature construction. Lastly, we explain briefly about our final dataset for building SMEs credit scoring model in Section 4.5.

4.1 Data Model

We need to collect SMEs' information to enhance our initial dataset. Our goal in collecting the SMEs' data is to have a set of features for every SMEs in our initial dataset. The problem is the data required to build those features are collected from various data source. The SME's data collected from one source need to be linked with the data from other sources. Before we start collecting the data, we need to design our data model to better store and manage the SMEs' information.

Figure 4.1 illustrate the data model for storing SMEs information. In designing the data model, we started from the initial dataset that we already have. The initial dataset consist of the SMEs' basic information and also their financial records which comes separately. The SMEs' basic information is a static data while the financial records is time series data that can be aggregated to calculate the required financial features. We represent both information as two tables in our data model, company base and financial record. Then, we improved the data model by adding six more tables which are formulated based on our data requirements.

In implementing the data model above, we decided to use multiple CSV files. We choose the CSV file format to store the data because it can be opened and edited in various environment and programming language. Also it follows flat-simple schema and can clearly differentiate storing the numeric values and text. Each CSV files represent one table in the data model, which will be filled and updated by different data collection pipeline and sources.

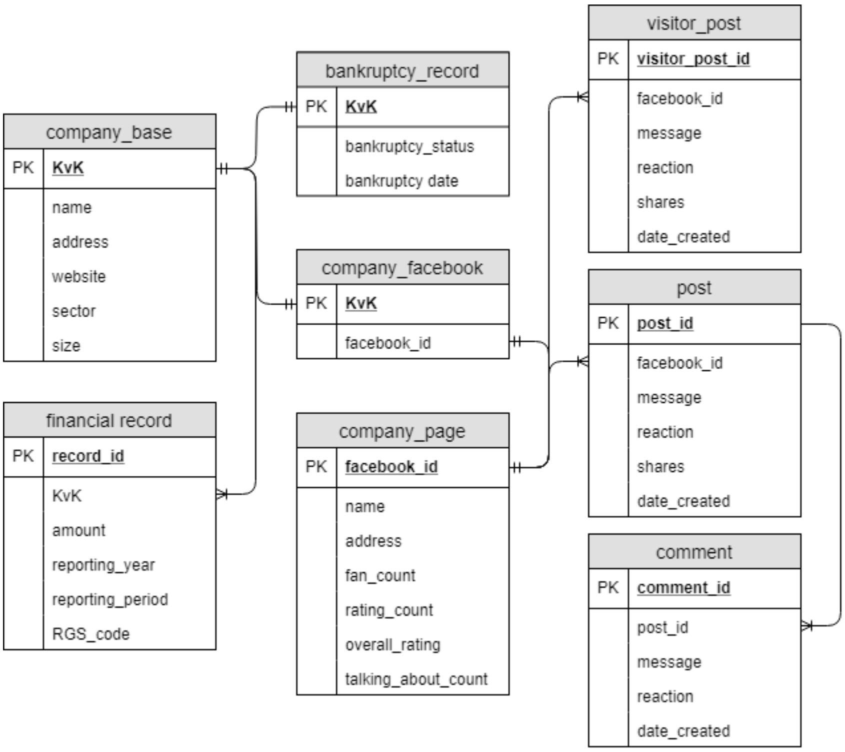


Figure 4.1: The data model diagram to store SMEs' information

4.2 Collecting Data from Public Website

We use some public websites as our data source to verify and complete some tables in our data model, including company_base and bankruptcy_record. The information that we collect from those public websites include name, address, website, bankruptcy label, bankruptcy date, business sector and business size from the public website. The acquisition of those data follows the same method which is by scraping using SMEs' KvK number as the input URL. The architecture of our web scraper are illustrated in the Figure 4.2 and explained below.

- **Seed URL**: A web scraping begins with a list of URLs called seed list. The URLs in seed list (Seed URLs) are constructed from the KvK number which can be different for each public website. For example, to collect the data of SME which has KvK number 23056300 from opencompanies, the Seed URL is `https://www.opencompanies.nl/23056300`.
- **Scheduler** The scheduler basically handle the scraping strategy and chooses the URL from the queue to be sent to the downloader module.In handling the scraping strategy, the scheduler decide some scheduling policies such as the interval seconds between connections to a single web site, the number of retries when the connection is failed, the number of simultaneous connections to different websites, etc. In this project, we implement it in the Python script which utilize Selenium.

- **Downloader**: In a web scraper, The HTML downloader send the HTTP request to the website and retrieve the content as an HTML file. The downloader also takes care Sessions, cookies, and authentication if its needed. In this project, we use Google's Chrome controlled by Selenium as our HTML downloader.
- **HTML Parser**: Typically, a web scraper has an HTML parser that analyzes the HTML of the web pages it crawls with the intention to extract information such title of the page, headings and paragraphs, links and different tags in the page. In this project, we implements it by creating Python script which utilize the Beautiful Soup.
- **Scrap Frontier**: This is the list to which the web scraper adds the URLs it discovers during its crawl. Sometimes the Seed URL does not provide the direct access to the web page which contains the information we are looking for, and another URL is need to be captured to make another request. The downloader and parser identifies hyperlinks in the web pages and adds them to this repository. All the URLs in the crawl frontier will be visited by downloader and its HTML is extracted to get the information we are looking for. For example, to verify the name, address, and websites information of SME with KvK number 23056300 from `www.kvk.nl`, the Seed URL is `https://diensten.kvk.nl/TST-BIN/ZS/ZSWWW01@?TYPE=NDNR&NDNR=23056300&NSDN=%3F`. However, accessing that URL provide us to all of the URL of pages related to the specified KvK number, which include its head office and branch. As we only interested in the head office, we need to find and store that URL in Scrap Frontier and then visit it to found the information that we need. Not all of our scraper contains Scrap Frontier as it depends on each website.
- **Storage**: The storage permanently stores the information collected from the scraping. The storage may also stores other information like link structure of the document space, the time each document was fetched etc. In this project, we use CSV files to store the information extracted from public websites.

The initial dataset that we have contains 218,778 Dutch SME. After collecting the data from public websites, we have the following information in our dataset.

Information	**Total**
Verified KvK	217280
Bankruptcy found	3264
Size	124261
Sector	132227
Website	46948

Table 4.1: Summary of the dataset after scrapping from public websites

4.3 Finding SMEs Facebook ID

The user of a social media platform can be referred by unique identifier such as social media ID or username. To collect the information from social media profile, the

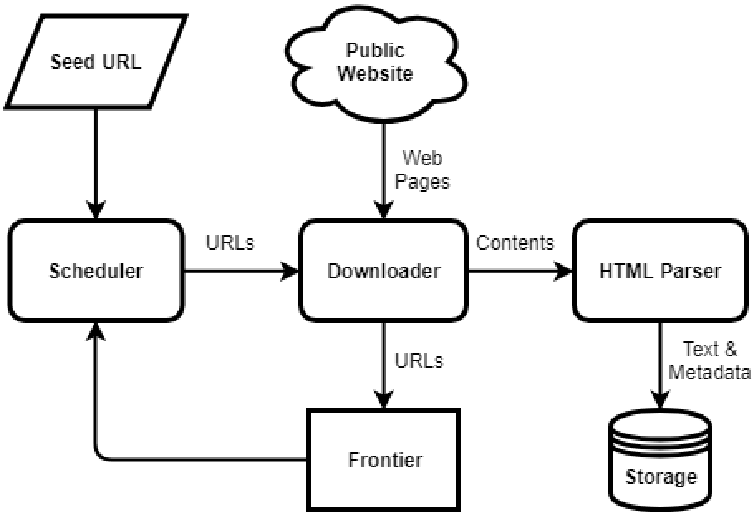


Figure 4.2: Functional diagram of Web Scraper

knowledge about their ID or username is a prerequisite. In our initial dataset, the information related to SMEs' Facebook ID or username is not available, which means we need to find it ourself using the data that we already have. To get the official Facebook ID for each SMEs, we combined two strategies: scraping SMEs official website and finding SMEs user ID through Facebook API. The combination of these strategies are illustrated by Figure 4.3.

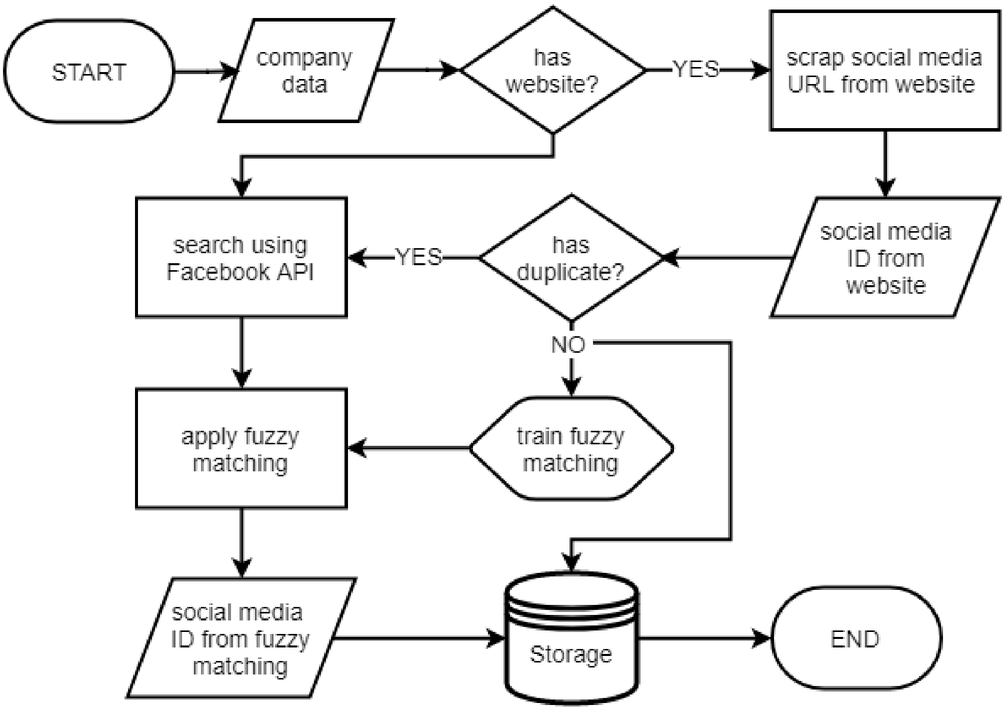

START
company data
has website?
YES
scrap social media URL from website
search using Facebook API
YES
has duplicate?
social media ID from website
NO
apply fuzzy matching
train fuzzy matching
social media ID from fuzzy matching
Storage
END

Figure 4.3: Strategy to find SMEs' Facebook ID

Scraping SMEs Website

The first strategy that we use is scraping from SMEs websites. We already have the information about 46,945 SMEs websites after scrapping it from the public websites. We utilize the knowledge of the SMEs website in our dataset to find out their social media accounts. Nowadays, many companies, including SMEs, often put the URL to their social media profile in the homepage of their website. By scrapping through HTML file of their homepage, it is possible to get their social media ID by looking for a specific URL pattern. For example, URL which is related to the Facebook page has pattern like `https://www.facebook.com/<username>`, `http://www.facebook.com/<ID_number>`, or `http://www.facebook.com/pages/<name>/<ID_number>`.

We scrapped the homepage of all SMEs' websites in our dataset and stored all of the URL which contains the URL pattern mentioned above. Then, we cleaned the URL by only capturing the username or userID part of the URL which represent the SMEs' Facebook accounts. Using SME's Facebook username or ID number, we can collect their Facebook data through Facebook API. However, sometimes we found that some Facebook accounts represent more than one SMEs in our dataset or one SMEs is represented by more than one Facebook accounts. There are a lot of possible cause for that, including some SME has the same parent company, the websites' URL redirect us to the web hosting sites because it is inactive, the SMEs have multiple Facebook accounts, etc. In that case, we put them as the subject for our second strategy (fuzzy matching). For this strategy, we only trust the SMEs which are represented by a unique Facebook account as the correct SME-Facebook account pairs.

Using this method, we can find 17,866 Facebook accounts of the SMEs in our dataset. This number is really low compared to the number of valid KvK number, which means we need another method to find the SMEs' Facebook accounts. Although the number is really low, but we have high confidence on the correctness of SME-Facebook account pairs that we found. We can use the knowledge of those SMEs-Facebook account pairs in our second strategy to tune our fuzzy matching technique.

Search API and Fuzzy Matching

In the second strategies, we utilize the page search features provided by Facebook's Graph API for finding more SMEs' Facebook account. Using SMEs' name or brand names as search query, the API can provide us with the list of Facebook page which has similar name to our query in JSON format as the result. Because the result can contains more than one accounts due to similarity in the names, we need to apply fuzzy matching technique to choose the relevant Facebook account for the SMEs from those list. The fuzzy matching technique will choose the relevant Facebook account from the list which has the most similar name and address with the actual companies in our dataset.

We define two features which measure how similar the name or address of a Facebook account with the SME in our dataset based on Levenshtein distance as a float in the range [0, 1]. The first feature, L_{full}, score the similarity between two string by calculating $2.0 \times M/T$, where T is the total number of elements in both string, and M is the number of matches. For example,

$$L_{full}(\text{"Delft"}, \text{"TU Delft"}) = 2.0 \times 5/13 = 0.77$$

While, the second feature, L_{token}, score the similarity between two string by calculating $M/\min(T1,T2)$, where $T1$ is the total number of elements in the first string, $T1$ is the total number of elements in the second string, and M is the number of matches after sorting the words. For example,

$$L_{token}(\text{"Delft"}, \text{"TU Delft Building"}) = 5/min(5,17) = 1.0$$

To do the fuzzy matching, we calculate both similarity features for name and address and define the probability of a Facebook account (p) is relevant as weighted score of those similarity feature, which can be described using the following formula.

$$p(x,y) = S_{name}(x,y) * w_{name} + S_{address}(x,y) * w_{address}$$

, while

$$S_i(x,y) = L_{full}(x_i,y_i) * k_{full} + L_{token}(x_i,y_i) * k_{token}$$

The problem is we need to decide the weight w and k from the equation above first before we can calculate the probability of a Facebook account is relevant given an SME in our dataset. One of the solution is by solving it as binary classification problem of predicting whether a Facebook account is relevant using those four features as the input features. To train the prediction model, we use the SMEs-Facebook account pairs which we already collected from the first strategy for training the model.

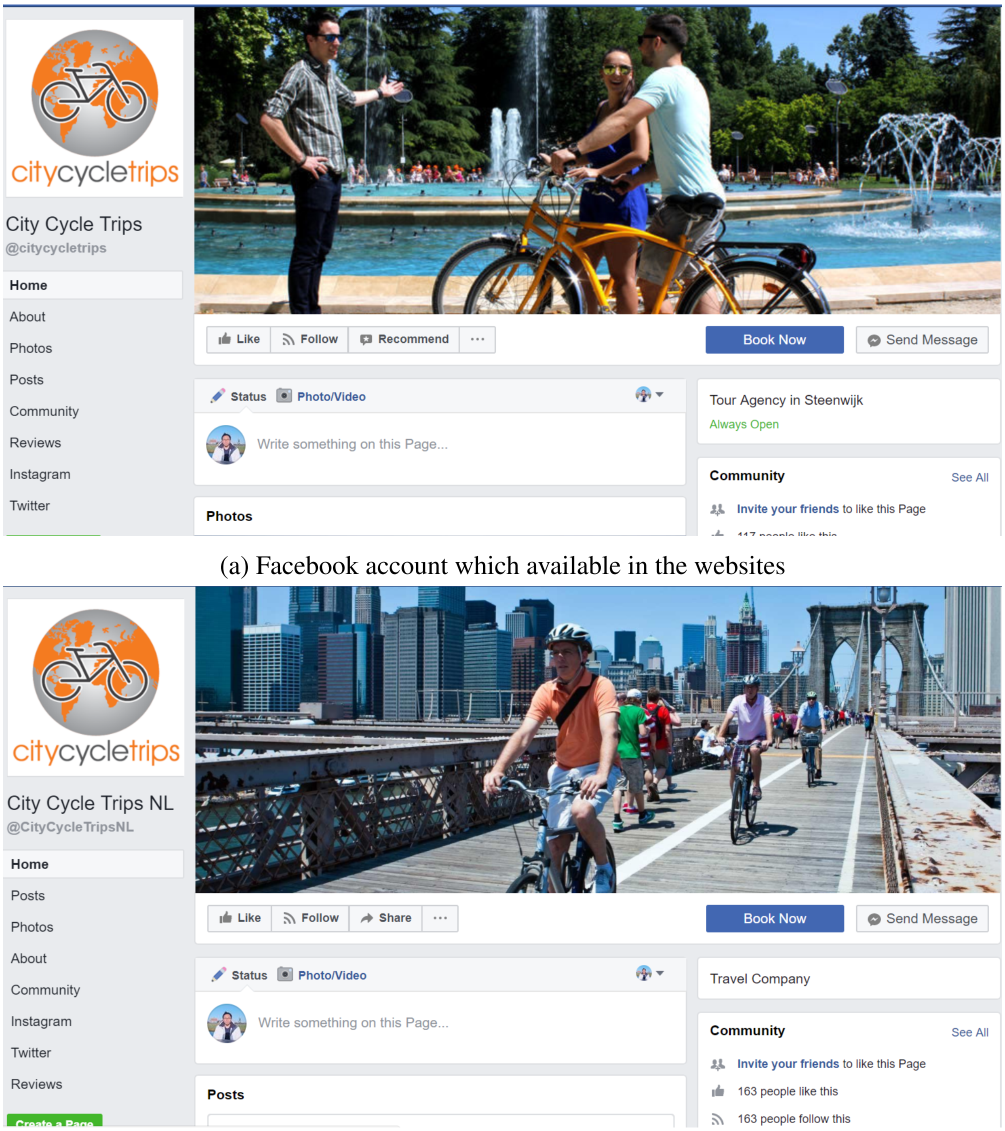
It is possible that the model gives us more than one relevant Facebook accounts for an SMEs, which can be caused by either the unofficial page of SMEs are created in Facebook or the SMEs have indeed multiple Facebook accounts. Figure 4.4 shows the multiple Facebook accounts which are marked as relevant for an SME. So, another layer of selection method is required to select the most relevant Facebook ID. The next selection method is using heuristic approach by picking the Facebook ID with the highest score in selection criteria. The selection criteria are chosen so that it can maximize the chance of the best matching Facebook ID is chosen. The overall work flow for this fuzzy matching strategy is shown in pseudocode below.

In implementing the *isRelevant* function above, some classifier including logistic regression, XGBoost, Random Forest, SVM, nearest neighbour have been evaluated to build the prediction model which predict the relevancy of Facebook ID. To evaluate the model, we use 5-fold cross validation which the result is presented in Table 4.2. From the result, we choose the XGBoost model for our fuzzy matching technique because it outperform other models in various metric, including accuracy and ROC_AUC. It also produce relatively high precision and recall compared to other models.

Table 4.2: Classifier evaluated for fuzzy matching

Classifier	**Accuracy**	**ROC_AUC**	**Precision**	**Recall**
XGBoost	**0.975298**	**0.976732**	0.820346	0.796832
Logistic Reg.	0.975285	0.962507	0.842223	0.765546
SVM	0.974542	0.965664	**0.843497**	0.749903
Random Forest	0.972044	0.953327	0.789696	0.780412
5-Nearest Neigh.	0.970322	0.937321	0.760712	**0.796930**

We also conduct some experiment to choose the best selection criteria for our heuristic model. If there are multiple relevant Facebook accounts are predicted by



(a) Facebook account which available in the websites

(b) Another Facebook account marked as relevant

Figure 4.4: Example of multiple Facebook accounts which are marked as relevant for an SME

the model for an SME, the criteria will be used to select one Facebook account from the list. The grid search are performed to choose the best criteria to be used in the heuristic approach. We calculate some scoring criteria from the product combination of log of fan_count (log_fans), probability of Facebook accounts is relevant (p_XGB), name similarity calculated using L_full (name_sim) and address similarity calculated using L_full (name_sim). Figure 4.5 shows the number of correct and wrong result produced by those criteria, and the best scoring is multiplication between log_fans, name_sim and p_XGB.

The default threshold for the binary classification model is 0.5. However, using that as our threshold of probability may not be optimal. The threshold are tuned in

Algorithm 1 Determining the correct social media ID for an SMEs, C

```
function FUZZYMATCHING($C, f, threshold$)
    $possibleID \leftarrow \{\}$
    for all query $Q$ in $C$ do
        $R_Q \leftarrow APIsearch(Q)$
        for all item $X$ in $R_Q$ do
            $p(X, relevant) \leftarrow isRelevant(X, Q, C, f)$
            if $p(X, relevant) > threshold$ then
                $possibleID.insert(X)$
    for all item $\chi$ in $possibleID$ do
        $criteria_\chi \leftarrow \log_{10}(fans + 10) * L_{phrase}(\chi_{name}, C_{name}) * p(\chi, relevant)$
    $bestMatch \leftarrow argmax_\chi(criteria)$
    return $bestMatch$

function ISRELEVANT($X, Q, C, f$)
    $L_{pn} \leftarrow L_{phrase}(X_{name}, Q)$
    $L_{tn} \leftarrow L_{token}(X_{name}, Q)$
    $L_{pa} \leftarrow L_{phrase}(X_{addr}, C_{addr})$
    $L_{ta} \leftarrow L_{token}(X_{addr}, C_{addr})$
    $p(X, relevant) \leftarrow f(L_{pn}, L_{tn}, L_{pa}, L_{ta})$
    return $p(X, relevant)$
```

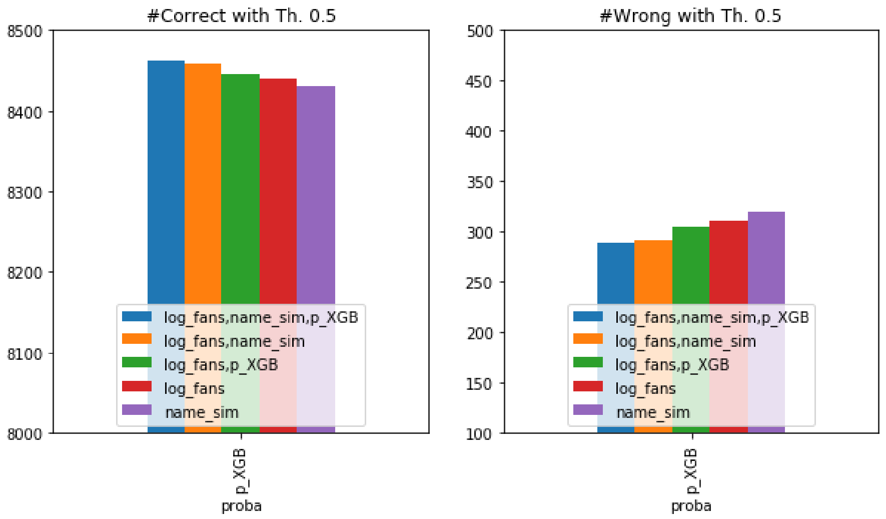


Figure 4.5: Threshold evaluated for fuzzy matching

order to allow the relevant Facebook ID that has probability slightly below 0.5 to be included when employing the heuristic approach. To select the best threshold, we performed an experiment using the complete fuzzy matching pipeline, which includes the best scoring criteria that we already found. We observe the output of our fuzzy matching technique by measuring the number of SMEs which have correct, wrong and empty or no relevant Facebook ID predicted in relation to the selected threshold. The result of this experiment are shown by Figure 4.6. Based on the figure, the threshold are chosen to be 0.25 as the fuzzy matching system begin to stable from this value.

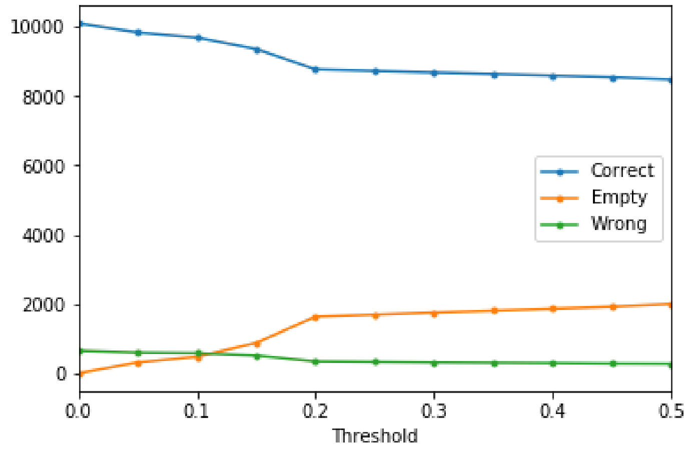


Figure 4.6: Criteria evaluated for fuzzy matching

We apply our tuned fuzzy matching technique to find the relevant Facebook ID of SMEs in or dataset after excluding the pairs from the first strategy. Out of 109,454 SMEs which have non-empty search result, the fuzzy matching only provide us with 26,306 SME-Facebook account pairs. After combining the result of both strategy to find SMEs' Facebook account, we are able to find 44,172 SME-Facebook account pairs.

4.4 Social Media Data Collection and Feature Construction

Once the Facebook ID of the SMEs are found, all of the data which available to public in their profiles can be collected using Facebook API. Those data include general information, e.g. fan count, and feed related information, e.g. posts, reactions, shares and comments. Figure 4.7 shows the work flow to get the data from company social media profiles. For each company's Facebook ID we found, the following information are collected.

- **posts**: feed created by the company during the last 6 months before time horizon (between 6 months and 12 months before the bankruptcy date if the company is bankrupt, or between 6 months and 12 months before the collection date of bankruptcy status if otherwise). The attributes collected from post are creation date, type, message, number of comments, number of shares, and reaction count.
- **visitor_posts**: feed created by customer during the last 6 months before time horizon. The attributes collected from visitor post are creation date, and message.
- **comments**: comments posted by customer during the last 6 months before time horizon. The attributes collected from comments are creation date, and message.
- **fan_counts**: the number of follower in the social media platform (Facebook).

- **overall_star_rating**: the average rating provided by customer.
- **rating_count**: the number of rating submitted by customer.
- **talking_about_count**: the number of mention in the content created by customer.

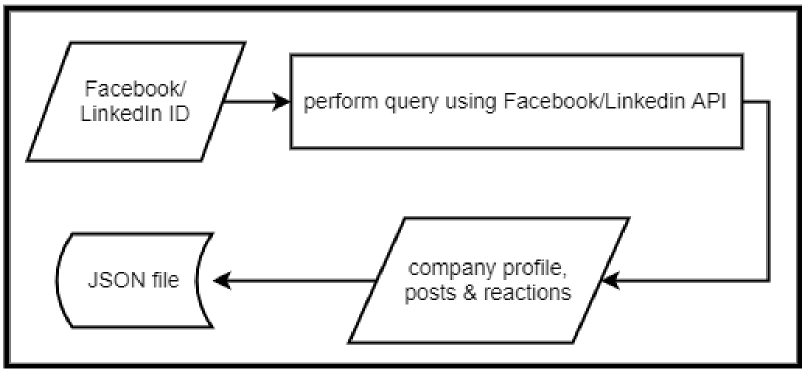


Figure 4.7: High level overview of data social media data collection

The data returned from Facebook API is in JSON format. We transform it into CSV files to make it easier for cleaning and pre-processing the data. We clean the posts, comments and visitor posts by removing the data, which are not inside the model horizon specified in this project. For preprocessing step, we also perform sentiment analysis to visitor post and comments to classify whether it has positive, negative or neutral mood. We focus on doing sentiment analysis on two language, Dutch and English. If the visitor posts or comments is written in Dutch or English, the sentiment analysis will be done by classify them into positive, negative or neutral. If the language are not detected, it will automatically assign the comments or visitor posts to neutral.

In this project, we used some pre-trained model for the language detection[1] and sentiment analysis[2]. After data cleaning and preprocessing has been done, we have more than 7GB of data containing Facebook page's general info, posts, visitor posts and comments of 44,172 SMEs in our dataset. Because the size of data is enormous, the Spark cluster are used to derive 35 social media features that we already specified in the previous chapter.

4.5 Final Dataset

We calculate the financial features by combining the initial data from exact and also the bankruptcy status of the SMEs. Similar to the social media data, we only use the financial data inside the input model horizon, which is between 6 months and 12 months before the bankruptcy date if the it is bankrupt, or between 6 months and 12 months before the collection date of bankruptcy status if otherwise. The financial dataset that we build consist of 40,051 SMEs, with 524 (1.308%) is bankrupt.

[1] https://github.com/aboSamoor/pycld2
[2] https://github.com/clips/pattern

To build the final dataset, we did inner join between the financial dataset, social media dataset, the bankruptcy record and company base table (to get sector and size information) using KvKs as the key. The result is the final dataset which contains the KvK as SMEs identifier, bankruptcy label, traditional features and social media features. The final dataset contains 20 financial features and 35 social media features of 25,654 SMEs, with 194 (0.756%) is bankrupt.

Chapter 5

Model Development and Analysis

In this chapter, we describe our experiments in building credit scoring model for SMEs using various features that we already collected. We organize this chapter into four section. Section 5.1 explain how we setup our experiments to build credit scoring model from three different set of features: traditional features only, social media features only and the combination of both traditional and social media features. In Section 5.2, 5.3 and 5.4, we report performance of the best model created from those three different set of features. Finally, we analyze those models and the features that gives significant influence to the models in Section 5.5.

5.1 Experimental Setup

We conduct some experiments to build credit scoring model using three different features set: traditional features only, social media features only and the combination of both traditional and social media features. The objective of the experiments is to build the best performing model in predicting the SMEs' bankruptcy status in the next 6 months and find which features are influential in the best performing model.

The dataset used in the experiments consists of 25,654 SMEs, with 194 (0.756%) of them are bankrupt. For each feature set, we employ ”white-box” and ”black-box” type of classifiers to build the model. The ”white-box” classifier is represented by logistic regression, while the ”black-box” classifiers represented by XGBoost. We will compare the performance of each model with the two non-classifier model, Random Guess and Weighted Guess (See Appendix C). The performance of both non-classifier model is shown by Figure 5.1.

Model	AUC	Accuracy	F1-score	MCC	Precision	Recall
Random Guess	0.5	0.5	0.0149	0	0.0076	0.5
Weighted Guess	0.5	0.9849	0.0076	0	0.0076	0.0076

Table 5.1: Performance of non-classifier models as baseline

5.1.1 Developing Traditional Model and Social Media Model

For building the model using the traditional only and social media only feature sets, we handle any missing values first by substituting them with the median of the feature where it is missing. We assume that any missing values in our dataset is just coincidence and does not have any relation to the SMEs creditworthiness. We chose median rather than mean because it will choose better representative value if the data is skewed. Then, we employ some feature selection technique to reduce the dimensionality of the data, they are variance based, correlation based, recursive feature selection. We first use variance based feature selection to drop any features which has extremely low in variance by choosing 0.001 as the minimum limit. Then, we calculate the correlation among the features and identify some highly correlated groups. From each group, one features is selected as a representative. Finally, we perform recursive feature selection by iteratively building the prediction model and removing less important features. The selected features then are standardized to get unit variance. We assume that the selected feature are independent from each other.

After selecting the model, the next step is handling the data imbalance. Handling data imbalance is not necessary if minority class are not rare in an absolute sense but are rare only relative to other objects. In our case, the rarity of minority class is absolute as we only have 194 samples of bankrupt SMEs, which makes handling the imbalance data is required to be done. We use SMOTE+Tomek Links Removal technique, which a combination of oversampling and undersampling method. This technique will generate some artificial minority data using SMOTE and later clean the noisy data and reduce class overlap with Tomek Links Removal technique [7].

The last step is training the model using the classifier that we choose. To get the best performing model from the selected features, we need to optimize the hyperparameter of classifier as well as the imbalance technique that we use. To select the hyperparameter which produce the best performing model, we use grid search with stratified 5-fold cross-validation to evaluate the model performance. The best performing model form the grid search are selected as our final model and the evaluation result are used as the estimation of the model performance if trained using all of the data.

5.1.2 Developing Combined Model

To develop the combination model, which use the traditional and social media feature, we experiment with two different approach: early combining and late combining. Figure 5.1 shows how the early combining and late combining work. In the early combining approach, the best features from both traditional and social media model are selected as the initial features. Afterwards, we employ the same techniques that previously used in traditional and social media model development above for selecting features, handling imbalance data, training, evaluation and selecting the best model. For the late combiner approach, we combine the best performing traditional and social media model using another classifier as a combiner. The probability of default predicted by both traditional and social media model are used as the features to train the combiner classifier.

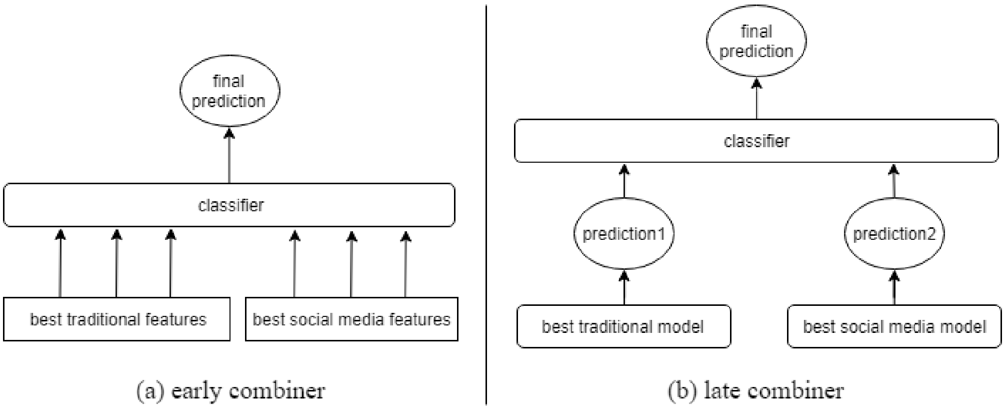


Figure 5.1: The combined model

5.2 Traditional Model

After applying one-hot encoding to the size and sector category, the dataset used to build the model contains 48 traditional features. Because there are some missing values in those features, we conduct data imputation to some financial ratio features, including net margin, EBIT margin, return-on-assets (ROA), and debt assets. Those financial ratios are calculated using revenue and total assets as the denominator which are not available in the financial data of some SMEs that we got from Exact.

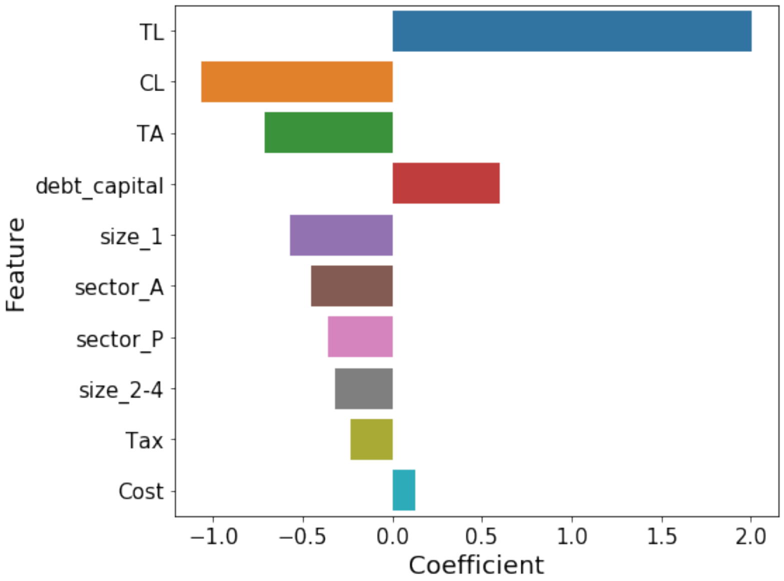
In the feature selection process, we remove some categorical features because of their extremely low variance, which are sector B, sector D, sector E, sector O, size 100-499, and size 50-99. We also remove some features due their extremely high correlation with the other features, they are return-on assets (ROA), current ratio, net margin, net profit, EBITDA, and revenue (See Appendix D for more details). The features **net profit** and **EBITDA** are correlated with **EBIT**, **net margin** is correlated with **EBIT margin**, **current ratio** is correlated with **quick ratio**, and **revenue** is correlated with **cost**.

The last feature selection process are recursive feature selection based on the importance of features. The final set of features then will be used to build the final model. The logistic regression and XGBoost classifier produce different set of final features, which are presented in Figure 5.2. The final features that we used to build logistic regression model are **total assets (TA), total liabilities (TL), current liabilities (CL), Cost, Tax, debt-to-capital ratio, sector A, sector P, size 1,** and **size 2-4**. While for the XGBoost, the traditional feature that we use are **total assets (TA), total liabilities (TL), current liabilities (CL), Tax, Cost, EBIT, debt-to-capital ratio, sector A, sector K, sector P, sector S, size 1,** and **size 2-4**.

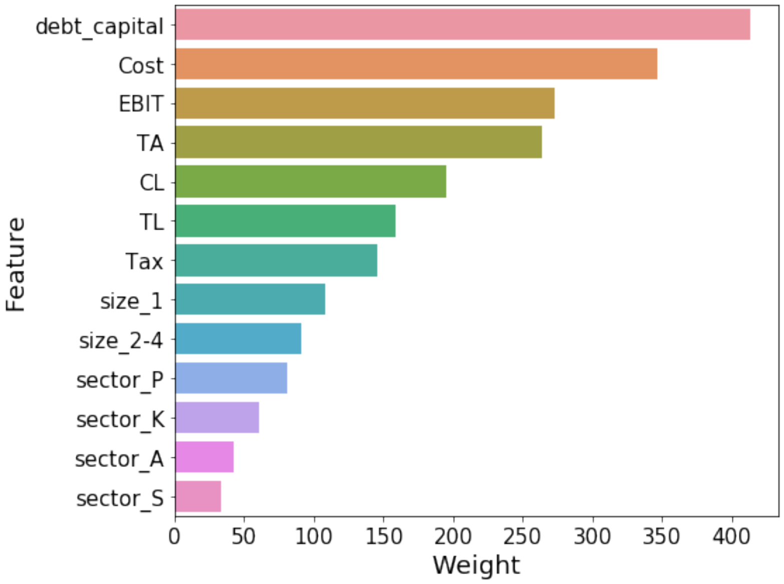
Model	AUC	Accuracy	F1-score	MCC	Precision	Recall
Traditional, LR	0.7031	0.9815	0.0871	0.0816	0.0747	0.1134
Traditional, XGB	0.7933	0.9715	0.1252	0.1343	0.0834	0.2574

Table 5.2: Performance of model developed using traditional features

The best performing traditional model for both logistic regression and XGBoost



(a) Logistic regression features



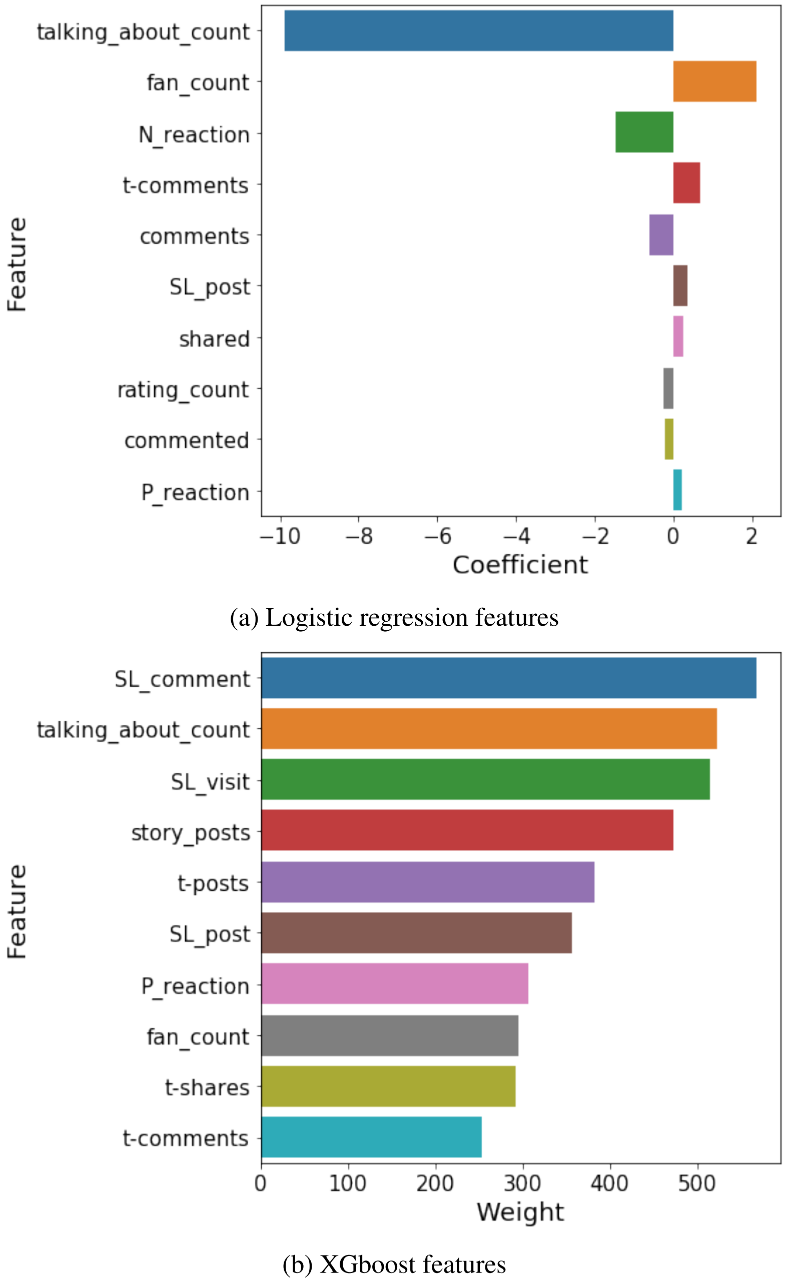
(b) XGboost features

Figure 5.2: Features used in traditional credit scoring model

are presented in Table 5.2. If we compare the result with the random and weighted guess, the performance of both our traditional credit scoring model is a lot better in term of AUC, F1-score, precision and MCC. Our model can detect more bankrupt SMEs while also maintain the accuracy really high with 98.15% and 97.15%. Both model has AUC above 70% which means the models are quite good in differentiating between bankrupt and healthy company.

5.3 Social Media Model

In our experiment to develop model based on only the social media data, we use our dataset which contains 35 social media features. All of those social media features are numerical and without any missing values, which does not require any data imputation nor one-hot encoding to be done.



(a) Logistic regression features

(b) XGboost features

Figure 5.3: Features used in social media credit scoring model

During the feature selection process, we drop percentage of negatively reacted post (N reacted) and trend of the percentage of negative comment (N comments) from the features set, because of their extremely low variance. We also remove some features due their extremely high correlation with the other features, including trend of the percentage of positively reacted post (t P reacted), trend of the average of positive reaction per post (t P reaction), trend of the average of negative reaction per post (t N reaction), and the number of negative reaction (See Appendix D for more details). Interestingly, the number of negative reaction (**N reaction**) is highly correlated with the average number of shares per post (**shares**).

In the last feature selection process, recursive feature selection, the logistic regression and XGBoost classifier also produce different set of final features for social media dataset. The selected features for logistic regression and XGBoost are presented in Figure 5.3. The final social media features that we used in our logistic regression model are **fan count, talking about count, rating count, comments, P reaction, N reaction, shared, commented, SL post,** and **t comments**. And for the XGBoost, the social media feature that we use are **fan count, talking about count, P reaction, story posts, SL post, t shares, t comments, t posts**, and **SL visit**.

Model	AUC	Accuracy	F1-score	MCC	Precision	Recall
SocMed, LR	0.6648	0.8685	0.0307	0.0388	0.0169	0.2784
SocMed, XGB	0.6751	0.9853	0.0364	0.0294	0.0382	0.0359

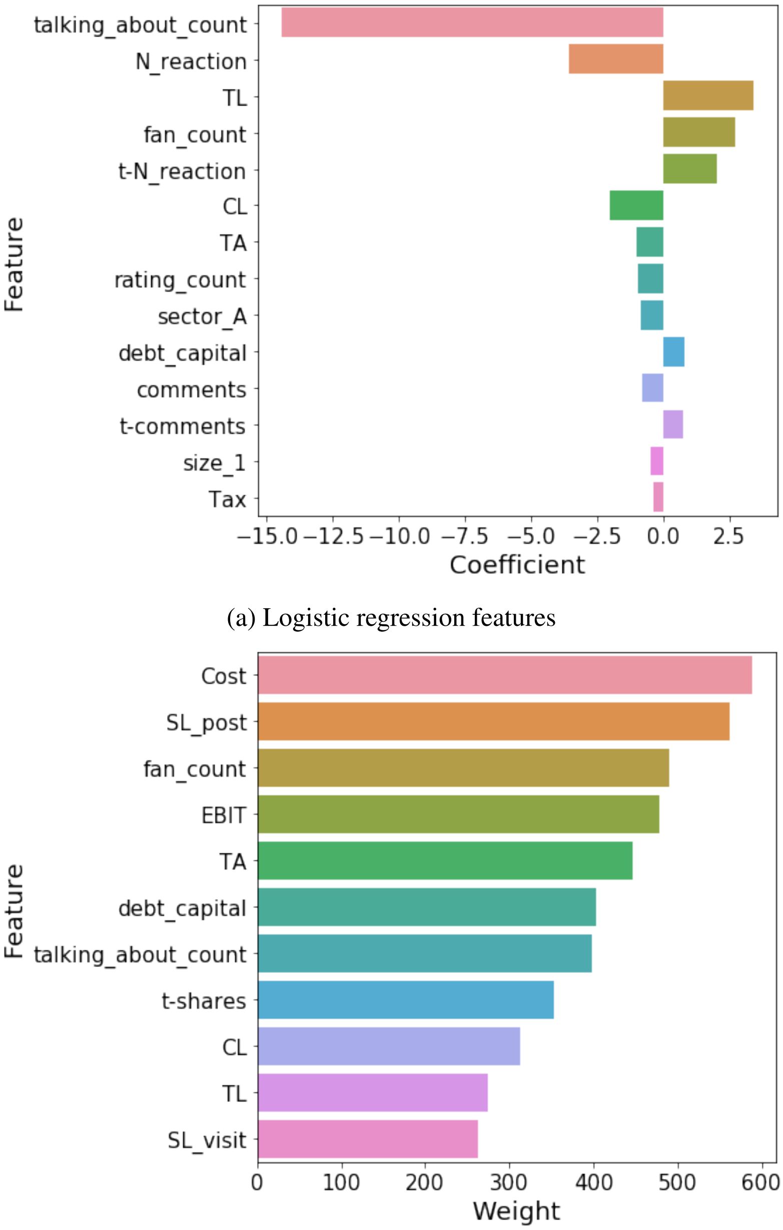
Table 5.3: Performance of model developed using social media (SocMed) features

The best performing social media model for both logistic regression and XGBoost are presented in Table 5.3. Although the performance of both social media model is better compared to random and weighted guess for F1 score, MCC, and precision, however it is far below the traditional model. Both social media model also has AUC around 66%, while the random and weighted guess have both 50%. It means that both social media model cannot differentiate really well between bankrupt and healthy company. The AUC shows that only around 66% chance that a random picked bankrupt company will have higher predicted probability of bankrupt to a random picked healthy company.

5.4 Combined Model

To build the early combining model, we select the features used to develop the best performing traditional and social media model. The initial features set for for building logistic regression model consist of 20 features, and for the XGBoost is 17 features. Afterwards, we select the final set of features for both logistic regression and XGBoost using the recursive feature selection and use them to build the early combined model.

The final features used to build early combined model for logistic regression and XGBoost are presented in the Figure 5.4. The final features for logistic regression model are **total assets (TA), total liabilities (TL), current liabilities (CL), Tax, debt-to-capital ratio, fan count, talking about count, rating count, comments, N reaction, t comments, t N reaction, sector A,** and **size 1**. For XGBoost, the final feature are **total assets (TA), total liabilities (TL), current liabilities (CL), Cost, EBIT, debt-to-capital ratio, fan count, talking about count, SL post, t shares,** and **SL visit**.

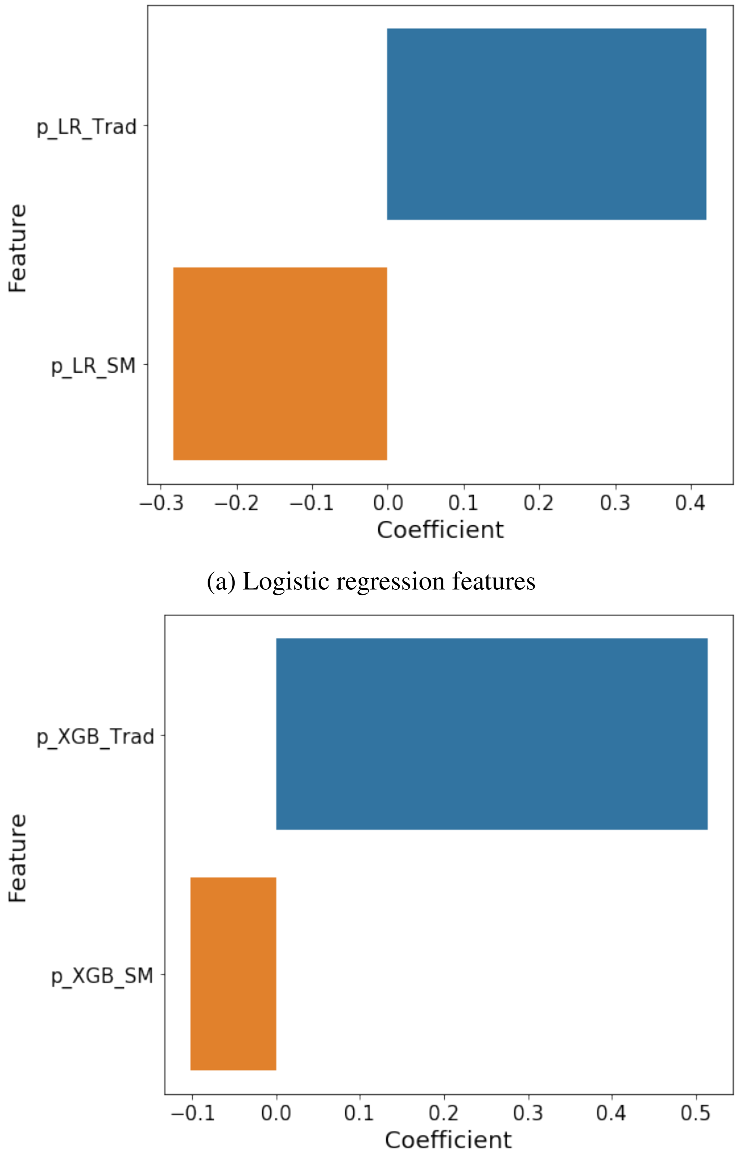


(a) Logistic regression features

(b) XGboost features

Figure 5.4: Features used in early combining credit scoring model

For the late combining model, we use logistic regression to combine probability of default predicted by traditional model and social media model. In this experiment, we combine the probability from the same classifier together. So, the traditional-logistic regression model is combined with social media-logistic regression model, and the traditional-XGBoost model is combined with social media-XGBoost model. Figure 5.5 shows the coefficient produced by the combiner classifier (logistic regression) for both logistic regression and XGBoost combination. The figure shows the probability of social media (p_LR_SM and p_XGB_SM) in both model have negative coefficients, which can be explained as the social media features alone is not a good predictor. Based on that, we can conclude that using late combining method to accommodate the social media features in credit scoring model is not a good choice if our social media model is not performing really well.



(a) Logistic regression features

(b) XGboost features

Figure 5.5: Features used in late combining credit scoring model

The performance for the combined model are presented in Table 5.4. It shows that generally the performance of early combining model are better than the late combining model. All of the performance metrics also shows that the combining models are generally better than the traditional and social media model which developed using the same technique. The best ”white-box” model can be achieved by early combining method using logistic regression, while the best ”black-box” model can be achieved by early combining method using XGBoost. Both combined models have AUC above 73% and the accuracy above 95%, which means they can differentiate the bankrupt and healthy SMEs better compared to the traditional and social media model while still maintaining good accuracy.

Model	AUC	Accuracy	F1-score	MCC	Precision	Recall
Early, LR	0.7518	0.9596	0.1030	0.1230	0.0624	0.3036
Early, XGB	0.8283	0.9815	0.1506	0.1494	0.1161	0.2160
Late, LR	0.7344	0.9809	0.0917	0.0883	0.0783	0.1287
Late, XGB	0.8115	0.9745	0.1418	0.1492	0.0990	0.2628

Table 5.4: Performance of early and late combining model

5.5 Feature Analysis

In the logistic regression, the importance of a feature is represented by its coefficients. The coefficient of features in logistic regression model is related to the $log(odds\_ratio)$. Positive coefficients means if the features increase, the probability of SMEs to bankrupt will also increases. The absolute value of coefficients shows how big the change in the feature will affect the change in $log(odd\_ratio)$.

For the XGBoost, the importance of features are more complicated and cannot be related to the probability of bankrupt. XGBoost model consist of multiple trees which trained sequentially. The importance of features in XGBoost is defined as the number of times a feature is used to split the data across all of the trees. The more a features used in to split, the more important it is. In XGBoost model, the features can be split anywhere, which make the model does not satisfy the monotonicity constraints. So, there is no information on whether a feature will affect the probability of default positively or negatively in the XGBoost.

5.5.1 The Influence of Traditional Features

From the model explained above, we can see that the set of traditional features that are involved to build the best performing model in both logistic regression and XGBoost, either in traditional only model or combined model, have a lot of similarity or subset. In fact, all of the features used to build traditional-logistic regression model are also used to build the traditional-XGBoost model. The features like **total assets (TA), total liabilities (TL), current liabilities (CL), debt-to-capital ratio** are always present in all of the traditional and combined model. It can indicates that the traditional features are good estimator for SMEs bankruptcy.

By analyzing the coefficients of the traditional logistic regression model in Figure 5.2, we can see how each traditional features affect the probability of SMEs to bankrupt. The figure shows that the total liabilities has positive coefficient, but the current liabilities has negative coefficient. If we looking at the context of bookkeeping, it make a lot of sense as the total liabilities is the total (short and long term) debt that they owe, while current liabilities is the debt which need to be paid on shorter cycles and used to run the business. For example, if a coffee shop buy a kiosk by borrowing the money from the bank, there is a risk that they cannot paid the installment in the future, which increase the probability of bankrupt. However, if they order a lot of coffee to sell, it can be an indicator that the business is going well, which decrease the probability of bankrupt.

The total assets and tax have negative coefficients which mean an increase to those features will reduce the probability of default. The larger total assets means that the SMEs has better chance to pay their debt as they can still liquidate their assets if it is necessary. Tax can also be an estimate of profit as the higher profit. The higher tax that SMEs paid also show the higher profit that they can generate.

The debt-to-capital ratio and cost have positive coefficients, which can also be explained. The debt-to-capital ratio measure the portion of total capital which belong to debt. The higher debt-to-capital ratio means more portion of debt and less portion of equity in the total capital, which lead to the higher risk of bankruptcy. The cost is related to how much money that SMEs need to spend to provide their service. It comes from various elements, including salary, raw material, electricity, etc. The SMEs which can minimize their cost means they will have more money to spend. So reducing the cost can also reduce the probability to bankrupt.

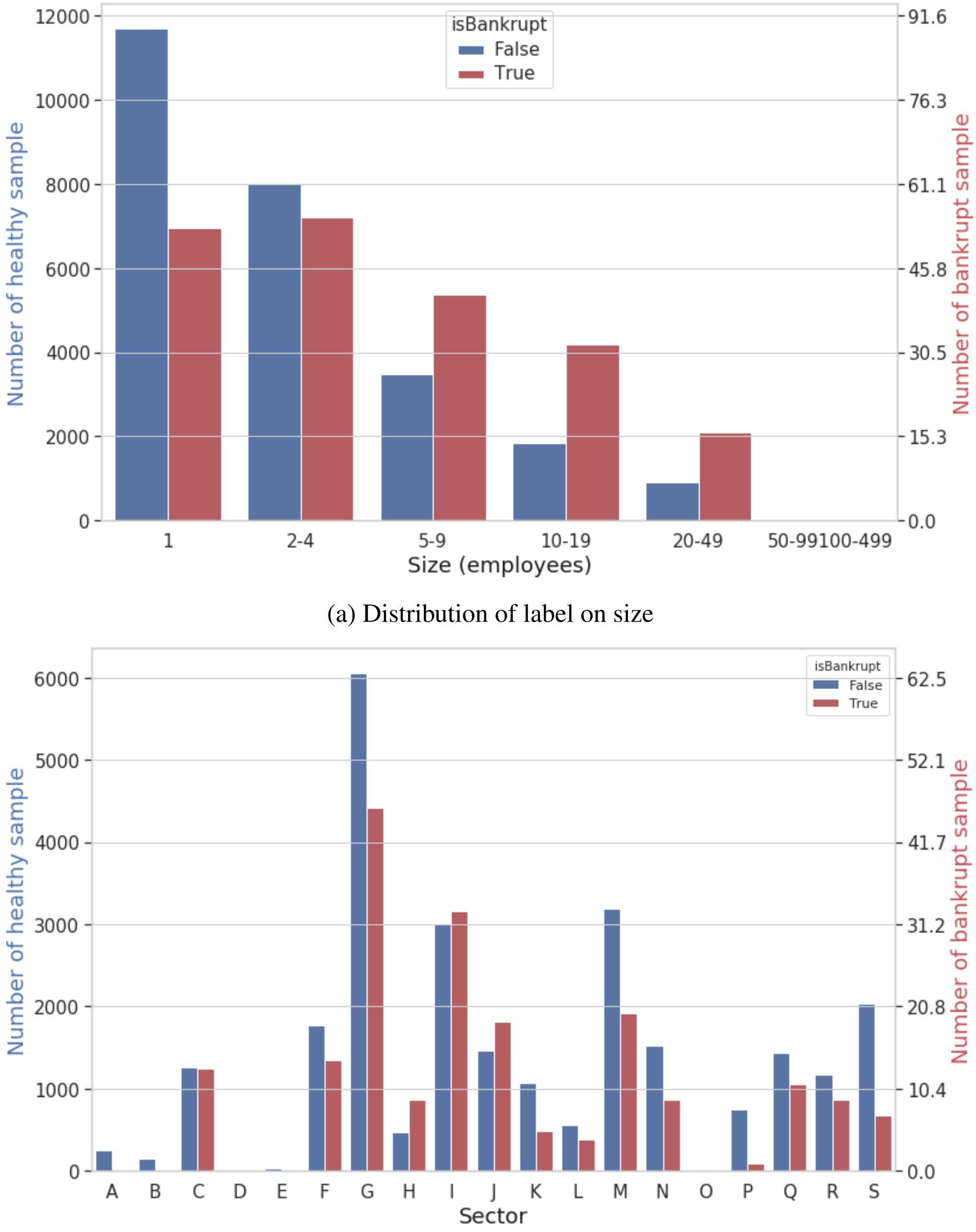
For the categorical variable, the micro enterprise with the size 1 and 2-4 has negative coefficients which means they are less likely to bankrupt compared to the bigger size SMEs. Our models also claims that the sector A (Agriculture, forestry and fishing) and sector P (Education) is less likely to bankrupt compared to other sector. However, we cannot justify this claims yet as in our current model we assume that the categorical features has small interaction only use small sample of data to build the model. As we can see from the Figure 5.6, some sector does not have the bankrupt sample, which means more data are needed to see whether that claims is valid.

5.5.2 The Influence of Social Media Features

Unlike the traditional features, the set of social media features that are involved to build the best performing model in both logistic regression and XGBoost, either in Traditional only model or combined model, have a lot of difference. If we look at only social media-logistic regression and social media-XGBoost model, only talking about count, t-comments, fan count and SL post which are used in both model. And if we look at the whole model which use social media features, only **talking about count** and **fan count** are present in all of them. It could be an indicators that the other social media features does not have significant influence in predicting bankruptcy status.

Figure 5.3 shows the coefficients of the social media logistic regression model. As we mentioned earlier, the coefficients also indicates how each social media features affect the probability of SMEs to bankrupt. According to the figure, the most dominant feature is talking about count, followed by fan count and the number of negative reaction. For the other social media features, their contribution in the social media logistic regression model and early combining logistic regression model are not significant as they have really small coefficient compared to those three features.

The talking about count measures the number of people mention them in a post. It has a negative coefficients in social media logistic regression model which means the higher number of talking about count, the less probable the SMEs will bankrupt.



(a) Distribution of label on size

(b) Distribution of label on sector

Figure 5.6: Distribution of label based on our data

That seems reasonable because the talking about count indicates how many people are engaged with Facebook page.

The fan count has positive coefficients which is quite interesting. This means that the number of fans has positive correlation with the probability of bankrupt. We suspect that the various sector of SMEs in our dataset affect this result. The higher number of fan count is relevant to indicates the whether a business is doing better if both the SMEs is from the same sector. Figure 5.7 shows that the sector can affect the distribution of fan count. Some business sector can have higher or lower fan count which also depend on the sector. However, we are unable to verify it yet due to our limited number of bankrupt sample.

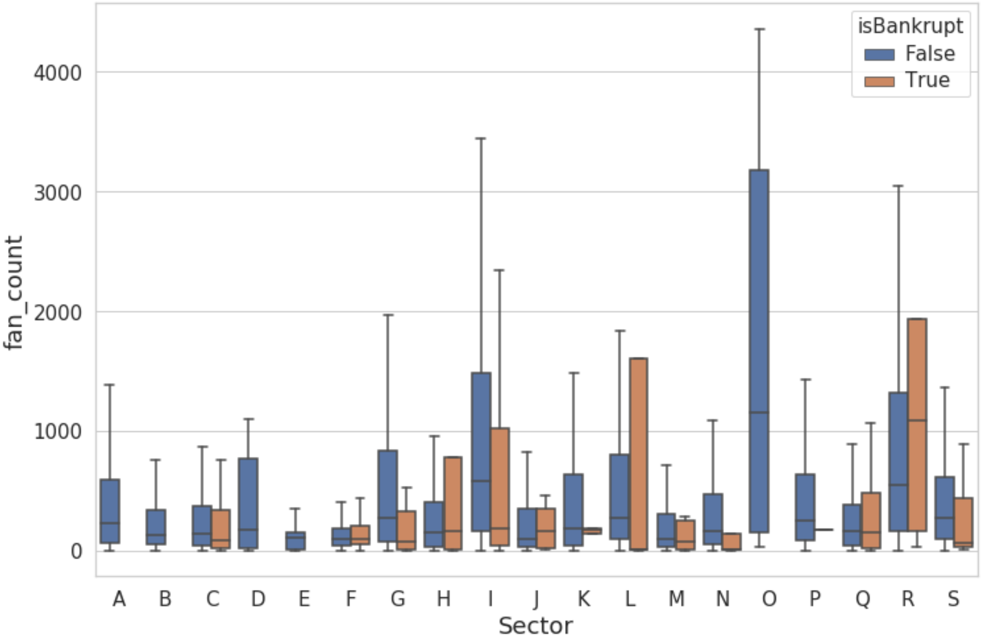


Figure 5.7: Box plot of fan count across various sector

Another interesting result is from the average number of negative reaction, as it has negative coefficients. It could be that the negative reaction indicates that a lot of people care about the page. It can also be supported by the high correlation between the average number of shares (shares) with the number of negative reactions (N reaction) per post (See Appendix D for more details). Based on those fact, we can think of the negative reaction as a sign of popularity instead of negative feedback.

Chapter 6

Discussions and Conclusions

In this chapter, we will discuss the result and the threat to validity of our study. Afterwards, we conclude the study by answering our research question and describe the possible future work.

6.1 Discussion

The reflection and analysis of each process and result in this project are discussed as follows.

Dataset Generation

In this project, we generate the our dataset by combining various data source, including Exact, public websites, and Facebook. Out of the initial data source that we have which contains 218,778 Dutch SME, in the end we only able to build the dataset which contains 25,654 SMEs. The significant drop in the number of SMEs are caused by some SMEs data cannot be found either in the Facebook or in the Exact's financial data. As we collected the data after the prediction horizon, this may also affect the availability of the data. Some data can be inaccessible because it has been a long time since it was created.

The performance of our fuzzy matching also affect the significant decrease in the number of SMEs in our final dataset. The Facebook search API provide us with 109,454 non-empty SMEs result as the input to our fuzzy matching technique that we already tuned beforehand. Hoewever, our fuzzy matching can only give us with 26,306 SMEs-Facebook account pairs, which only 24% the input. This result is also significantly lower if we compared to the fuzzy matching output during the tuning process. During the tuning process, the fuzzy matching output 8993 (8686 correct, 307 wrong) SMEs-Facebook account pairs out of 17,866 SMEs as the input, which is about 50%. We suspect that the poor performance of this fuzzy matching is because we only use address and and names to find the best match of Facebook page. While in reality, there are a lot of official Facebook page which does not contains address information which make them unable to be correctly matched. If that is the case, then by developing the new fuzzy matching technique which incorporates other information available in the Facebook page, such as phone number or websites, may increase the number of SMEs-Facebook account returned.

Model Development

In the previous chapter, we found that the performance of ”black-box” model is generally better than the ”white-box” model. Also, our best ”white-box” and ”black-box” SMEs credit scoring model can be achieved by combining traditional and social media features. Due to the multi-disciplinary nature of this study and the lack of domain experts, it is difficult to say to what extent the framework is fit to reality. Previously, we only compare their performance to the random and weighted guess which is unrealistic. To understand better how good is our best model, assume we have one million SMEs with the same ratio of bankrupcy (0.76%), which means there will be 7,600 bankrupt SMEs. Our model can detect 1642 bankrupt SMEs, missed 5,958 bankrupt SMEs, missclassified 12,497 SMEs as bankrupt, and correctly classified 979,903 healthy SMEs.

While there is no study which develop credit scoring using the similar information, we found some study which develop credit scoring model using the similar condition of extremely imbalance data set [11]. The study also use logistic regression and gradient boosting to developed the credit scoring model using various dataset which each of them contains only 1% and 2.5% default samples. The models which developed using 1% default sample produce the highest of AUC of 64.7% for logistic regression model and 74.5% for gradient boosting model. While the model developed using 2.5% default sample produce maximum AUC of 73.9% for logistic regression model and 88.3% for the gradient boosting model. Compared to the result above, our model which developed using 0.76% default sample is considered good, with the AUC of 75.2% for logistic regression model 82.8% for gradient boosting model. If we can improve the number of default or bankrupt sample in our dataset, we can expect to have the higher performance as has been proved by that previous study.

6.2 Thread to Validity

Following are some limitation and threats to the validity of this research.

1. Limited Sample of Bankrupt SMEs

   The generated credit scoring dataset in this project contains 25,654 SMEs, with only 194 (0.76%) of them are bankrupt which occurs in the last 5 years (2014-2018). Compared to the other datasets on credit scoring, e.g. Bene1, Bene2, Australian, Behav and German [11], our dataset is bigger as other datasets have less than 10,000 samples. However, if we compare the default sample and its ratio, then our dataset has more rare default sample and the distribution of label is also more imbalance compared to those public dataset. The dataset with the smallest default sample and the most imbalance dataset is Behav (240 default sample, 20%). Compared to the data of SMEs bankruptcy in the Netherlands, our sample also has smaller ratio of bankrupt sample. The number of bankrupt SMEs in the Netherland in 2016 is 3730 (0.849%) and in 2017 is 3018 (0.687%), while our dataset contains 194 (0.76%) from last 5 years.

   There is a chance that the model we produce does not cover all of possible characteristics and patterns of the bankrupt SMEs in the Netherlands, cannot generalize when evaluated on a new dataset. Despite that issue, the assessed

SMEs are also scattered over various sector and size which also has imbalance distribution. Some sectors in our dataset can be also underrepresented in our models as they do not have any representative of bankrupt sample. Although it may not cover all characteristic and patterns of bankrupt SME, but the white-box model that we produce still provide good performance and is reasonable. Some analysis of this dataset also provided similar result to some past research.

2. The Validity of Past Information

   In our project, we use the time series data to build our prediction model but the data is collected long after the event occurs, while ideally, the data is collected in a real-time or at least in the end of the input model horizon. By collecting the data long after the event occurs, we assume the data that represent the condition of SMEs' that bankrupt in the past can also be collected today and the data that we can collect today represent the condition in the past, which may not be true.

   For example, Facebook does not maintain the history of data which related to user basic information, such as fan count, talking about count and rating. So, when we collect past information of bankrupt SMEs from their Facebook page, we assume that those information are still relevant with the condition of the SMEs when bankruptcy occurs. In reality, it is possible that the SMEs information keep changing after the bankruptcy occurs.

   Another example is related to the bankruptcy status. We collect the bankruptcy status which happen in the past from public websites. We assume that all of the bankruptcy event in the past are still monitored by those websites and the SMEs that we cannot found in those websites are never bankrupt in the past. In reality, the websites that we use as bankruptcy data source can give different information for an SME. We found some case where a website does not have bankruptcy data of an SME while other websites have it. There is a chance that some bankrupt SMEs information are not available after the termination of their bankruptcy and we mislabel it as the solvent company.

3. The Performance of Other Pre-Trained Model,

   In this project, we use several pre-trained model in our framework as they are not in the focus of our study. We use pre-trained of language detection model and sentiment classification model to pre-process the comments and visitor posts data. The RGS classification code in our financial data is also predicted from Exact's RGS Mapping model. As we use them to build the features for developing the credit scoring model, the performance of those pre-trained model may also affect our result.

6.3 Conclusions

RQ1 *How is the existing SME credit scoring system implemented?* In order to answer this question we conduct a literature survey on how the previous studies build credit scoring model for SMEs. The previous studies usually solve the SME credit credit scoring as a binary classification problem which predict whether an SMEs will default. Most of the models for SMEs credit scoring are developed using traditional features (financial features, business size, business sector, and owner characteristics) from traditional data source (accounting data, survey, and questionnaire). The technique used to build the model usually is logistic regression due to its relatively good performance and also its interpretability. But some researchers also used more advanced classification technique, such as XGBoost, ANN, and Random Forest. Some of them also employ the method for handling the imbalance data. From those studies, we use the state-of-the-art techniques and the traditional features that they use to build our baseline model.

RQ2 *How to build features to measure SMEs creditworthiness from social media data?*

No study has incorporated social media data to build credit scoring model for SMEs. The adoption of social media in credit scoring nowadays is limited to the personal lending only. To answer this question, we conduct the literature survey on adoption of social media data in the personal credit scoring and on the social media metrics that are used to quantify business social perception. We propose 35 features features which are either similar or previously used by those studies to measure creditworthiness of SMEs.

RQ3 *How to incorporate social media features into SME credit scoring system?*

We did some experiments by building credit scoring models for SMEs which incorporate social media features. Our finding is the use of only social media features is insufficient to model SMEs default in the future. However, if we combine the social media features with the traditional features using early combining method, the performance of the model is better compared to the model developed only using traditional features and the same technique.

RQ4 *Which social media features influence the performance of SME credit scoring systems?*

From our experiments, we found that talking about count and fan count are present in all of the model that was developed using social media features. Both of them are also have relatively large coefficients in the both logistic regression model that use them (social media only and early combined). However, more studies are needed to justify this claim as the predictive ability of the social media features in SMEs credit scoring are still in question due their poor performance.

6.4 Future work

This thesis is a pilot research to study the credit scoring for SMEs from the perspective of social media. Given the reflection on our framework in the previous section, there are numerous rooms for improvement to continue this research, which is elaborated as the following.

1. Limited Sample of Bankrupt SMEs

   This research has already generated a credit scoring dataset of more than 20,000 SMEs, however we still have insufficient sample for the bankrupt SMEs. The initial number of bankrupt SMEs in our dataset is about 3000, but because there is missing information about those bankrupt company, such as financial record, size, sector, or Facebook ID, we need to discard them from our dataset. To get more information of those bankrupt company, we can try use premium data that we can purchase offered by some company. Another way is developing a special data collection method for the bankrupt company which involve human supervision, such as crowd-sourcing. By having more data of bankrupt SMEs, the applied analysis will be more accurate and the training of machine learning model may achieve better performance, which can mirror the pattern of SMEs bankruptcy in reality.

2. Improvement of the machine learning model

   Besides increasing the training dataset size, various other approaches can be applied to improve the performance of the credit scoring model. Currently, the model are developed under assumption that the features used to build credit scoring model are independent to the categorical features, such as size and sector. We use that assumption due to the limited sample that we have in for the bankrupt SMEs. However, if we have more sample. we can try to model also the interaction between those categorical features and other features that we use in our model development so that we can get better model.

3. Implementation of the methods to quantify creditworthiness of SMEs

   In this project, we proposed some features to quantify SMEs creditworthiness which derived from social media data. Our feature may cover the indicator that usually used to measure the performance of corporate in social media platform, which mostly related to marketing purpose. However, there are still a lot of information that can be collected from social media data which have not covered by those but also can be an indicator of SME creditworthiness. For example, we can analyze the activity of SMEs from their timeline which related to corporate social responsibility (CSR) as it is previously reported may affect the credit rating [23]. Another social media information that may relevant for SMEs credit scoring can also be collected from their owner social media profile. It has been reported that owner characteristic is influential features in SME credit scoring [8].

Bibliography

[1] Hussein A Abdou and John Pointon. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance and Management*, 18(2-3):59–88, 2011.

[2] Edward I Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609, 1968.

[3] Edward I Altman. An emerging market credit scoring system for corporate bonds. *Emerging markets review*, 6(4):311–323, 2005.

[4] Edward I Altman and Gabriele Sabato. Modelling credit risk for smes: Evidence from the us market. *Abacus*, 43(3):332–357, 2007.

[5] Edward I Altman, Gabriele Sabato, and Nicholas Wilson. The value of non-financial information in small and medium-sized enterprise risk management. *The Journal of Credit Risk*, 6(2):1–33, 2010.

[6] Silvia Angilella and Sebastiano Mazzù. The financing of innovative smes: A multicriteria credit rating model. *European Journal of Operational Research*, 244(2):540–554, 2015.

[7] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

[8] Mirta Bensic, Natasa Sarlija, and Marijana Zekic-Susac. Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance and Management*, 13(3):133–150, 2005.

[9] Allen N Berger, Adrian M Cowan, and W Scott Frame. The surprising use of credit scoring in small business lending by community banks and the attendant effects on credit availability, risk, and profitability. *Journal of Financial Services Research*, 39(1-2):1–17, 2011.

[10] Enrique Bonsón and Melinda Ratkai. A set of metrics to assess stakeholder engagement and social legitimacy on a corporate facebook page. *Online Information Review*, 37(5):787–803, 2013.

[11] Iain Brown and Christophe Mues. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3):3446–3453, 2012.

[12] G Buiten, R Boom, M Roos, and G Snijkers. Issues in automated financial data collection in the netherlands. In *Proceedings of the Fifth International Conference of Establishment Surveys: Statistics (ICESV) Switzerland*, pages 20–23, 2016.

[13] Francesco Ciampi and Niccolò Gordini. Small enterprise default prediction modeling through artificial neural networks: An empirical analysis of i talian small enterprises. *Journal of Small Business Management*, 51(1):23–45, 2013.

[14] European Commission. 2017 sba fact sheet netherlands. *European Commision*, 2017.

[15] Comitato di Basilea per la vigilanza bancaria. *International convergence of capital measurement and capital standards: a revised framework*. Bank for International Settlements, 2004.

[16] Robert O Edmister. An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial and Quantitative analysis*, 7(2):1477–1493, 1972.

[17] Jim Everett and John Watson. Small business failure and external risk factors. *Small Business Economics*, 11(4):371–390, 1998.

[18] Ron J Feldman. Small business loans, small banks and big change in technology called credit scoring. *The Region*, (Sep):19–25, 1997.

[19] Martin S Fridson and Fernando Alvarez. *Financial statement analysis: a practitioner's guide*, volume 597. John Wiley & Sons, 2011.

[20] Evelyn Hayden. Are credit scoring models sensitive with respect to default definitions? evidence from the austrian market. 2003.

[21] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, (9):1263–1284, 2008.

[22] Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, 33(4):847–856, 2007.

[23] Pornsit Jiraporn, Napatsorn Jiraporn, Adisak Boeprasert, and Kiyoung Chang. Does corporate social responsibility (csr) improve credit ratings? evidence from geographic identification. *Financial Management*, 43(3):505–531, 2014.

[24] Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis. Understanding the predictive power of social media. *Internet Research*, 23(5):544–559, 2013.

[25] Bo Kyeong Lee and So Young Sohn. A credit scoring model for smes based on accounting ethics. *Sustainability*, 9(9):1588, 2017.

[26] Xueming Luo, Jie Zhang, and Wenjing Duan. Social media and firm equity value. *Information Systems Research*, 24(1):146–163, 2013.

[27] Ana Isabel Marqués, Vicente García, and José Salvador Sánchez. On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society*, 64(7):1060–1070, 2013.

[28] AA Masyutin. Credit scoring based on social network data. *Business Informatics*, (3):33, 2015.

[29] Margaret McCann and Alexis Barlow. Use and measurement of social media for smes. *Journal of Small Business and Enterprise Development*, 22(2):273–287, 2015.

[30] Loretta J Mester et al. What's the point of credit scoring? *Business review*, 3(Sep/Oct):3–16, 1997.

[31] Nina Michaelidou, Nikoletta Theofania Siamagka, and George Christodoulides. Usage, barriers and measurement of social media marketing: An exploratory investigation of small and medium b2b brands. *Industrial marketing management*, 40(7):1153–1159, 2011.

[32] Brad L Neiger, Rosemary Thackeray, Sarah A Van Wagenen, Carl L Hanson, Joshua H West, Michael D Barnes, and Michael C Fagen. Use of social media in health promotion: purposes, key performance indicators, and evaluation metrics. *Health promotion practice*, 13(2):159–164, 2012.

[33] Dilhan Öztamur and İbrahim Sarper Karakadılar. Exploring the role of social media for smes: as a new marketing strategy tool for the firm performance perspective. *Procedia-Social and behavioral sciences*, 150:511–520, 2014.

[34] Chiara Pederzoli, Grid Thoma, and Costanza Torricelli. Modelling credit risk for innovative smes: the role of innovation measures. *Journal of financial services research*, 44(1):111–129, 2013.

[35] So Young Sohn, Tae Hee Moon, and Sanghoon Kim. Improved technology scoring model for credit guarantee fund. *Expert Systems with Applications*, 28(2):327–331, 2005.

[36] Tianhui Tan and Tuan Phan. Social media-driven credit scoring: the predictive value of social structures. 2016.

[37] Yufei Xia, Chuanzhe Liu, YuYing Li, and Nana Liu. A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, 78:225–241, 2017.

[38] Yuejin Zhang, Hengyue Jia, Yunfei Diao, Mo Hai, and Haifeng Li. Research on credit scoring by fusing social media information in online peer-to-peer lending. *Procedia Computer Science*, 91:168–174, 2016.

Appendix A

Categorical Features

Here, we describe the categorical features that we use to build credit scoring model in more detail.

Sector The feature indicates the industry sector the company is working at based on the Dutch Standaard Bedrijfsindeling (SBI 2008). Only top sector will be used here which represented by an alphabet from 'A' to 'T' for each category. Table A.1 shows the description for each sector category.

Table A.1: The description of sector features

Sector	Description
A	Agriculture, forestry and fishing
B	Mining and quarrying
C	Manufacturing
D	Electricity, gas, steam and air conditioning supply
E	Water supply; sewerage, waste management and remediation activities
F	Construction
G	Wholesale and retail trade; repair of motor vehicles and motorcycles
H	Transportation and storage
I	Accommodation and food service activities
J	Information and communication
K	Financial institutions
L	Renting, buying and selling of real estate
M	Consultancy, research and other specialised business services
N	Renting and leasing of tangible goods and other business support services
O	Public administration, public services and compulsory social security
P	Education
Q	Human health and social work activities
R	Culture, sports and recreation
S	Other service activities
T	Activities of households as employers; undifferentiated goodsand service-producing activities of households for own use

Size The feature measures the size of the SMEs based on the number of employees they have. Table A.2 shows the possible value of this features and the meaning behind those values.

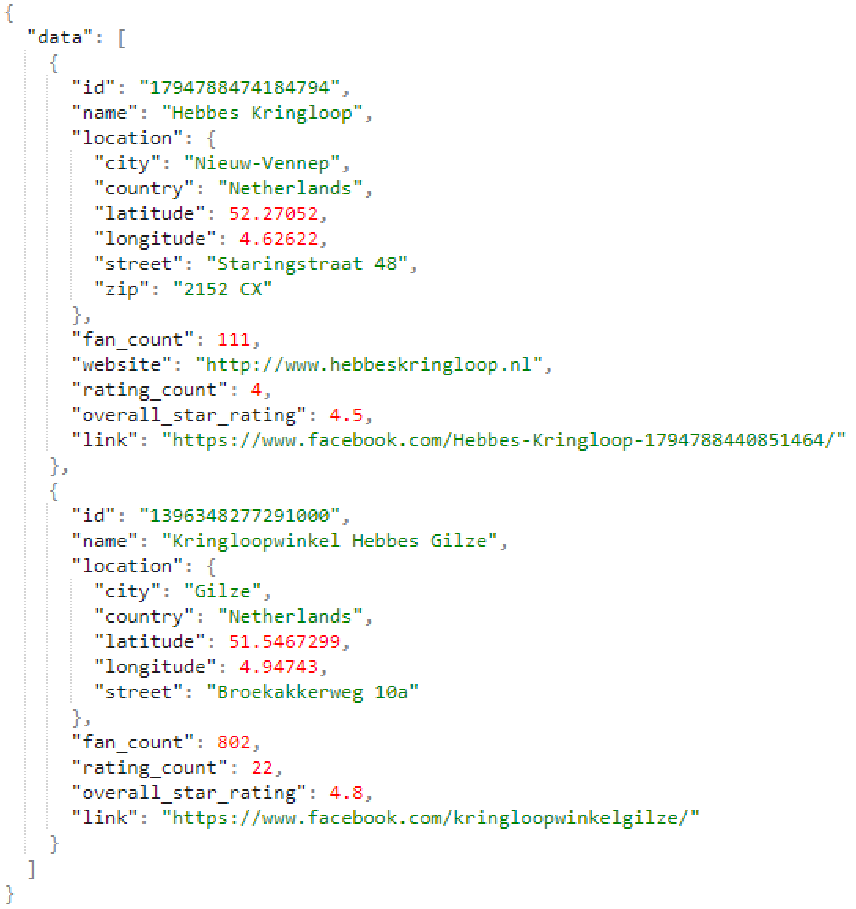
Table A.2: The description of size features

Size	**Description**
1	Has only one employees or self employed
2-4	Has 2-4 employees
5-9	Has 5 to 9 employees
10-19	Has 10 to 19 employees
20-49	Has 20 to 49 employees
50-99	Has 50 to 99 employees
100-499	Has 100 to 499 employees
Vanaf 500	More than 500 employees

Appendix B

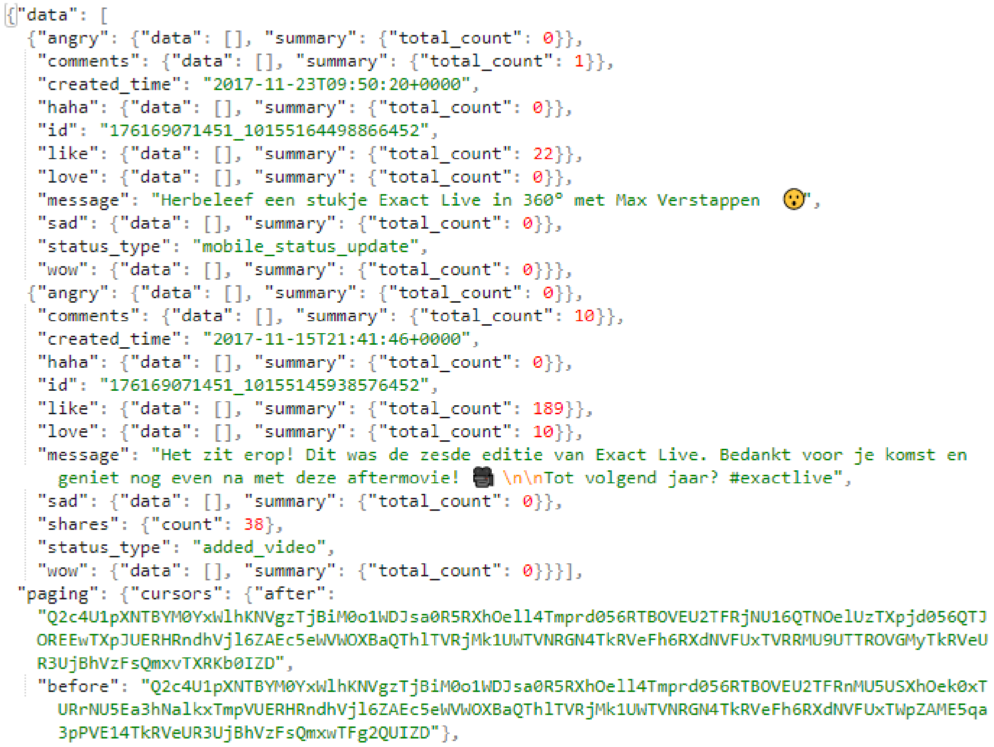
Facebook Graph API Example

The following are the example of result provided by Facebook's Graph API when we search the page based on name and the example of result when we query the post based on user ID.



```
{
  "data": [
    {
      "id": "1794788474184794",
      "name": "Hebbes Kringloop",
      "location": {
        "city": "Nieuw-Vennep",
        "country": "Netherlands",
        "latitude": 52.27052,
        "longitude": 4.62622,
        "street": "Staringstraat 48",
        "zip": "2152 CX"
      },
      "fan_count": 111,
      "website": "http://www.hebbeskringloop.nl",
      "rating_count": 4,
      "overall_star_rating": 4.5,
      "link": "https://www.facebook.com/Hebbes-Kringloop-1794788440851464/"
    },
    {
      "id": "1396348277291000",
      "name": "Kringloopwinkel Hebbes Gilze",
      "location": {
        "city": "Gilze",
        "country": "Netherlands",
        "latitude": 51.5467299,
        "longitude": 4.94743,
        "street": "Broekakkerweg 10a"
      },
      "fan_count": 802,
      "rating_count": 22,
      "overall_star_rating": 4.8,
      "link": "https://www.facebook.com/kringloopwinkelgilze/"
    }
  ]
}
```

Figure B.1: Example output of page search using keywords "Hebbes Kringloop"



```
{"data": [
  {"angry": {"data": [], "summary": {"total_count": 0}},
   "comments": {"data": [], "summary": {"total_count": 1}},
   "created_time": "2017-11-23T09:50:20+0000",
   "haha": {"data": [], "summary": {"total_count": 0}},
   "id": "176169071451_10155164498866452",
   "like": {"data": [], "summary": {"total_count": 22}},
   "love": {"data": [], "summary": {"total_count": 0}},
   "message": "Herbeleef een stukje Exact Live in 360° met Max Verstappen  😮",
   "sad": {"data": [], "summary": {"total_count": 0}},
   "status_type": "mobile_status_update",
   "wow": {"data": [], "summary": {"total_count": 0}}},
  {"angry": {"data": [], "summary": {"total_count": 0}},
   "comments": {"data": [], "summary": {"total_count": 10}},
   "created_time": "2017-11-15T21:41:46+0000",
   "haha": {"data": [], "summary": {"total_count": 0}},
   "id": "176169071451_10155145938576452",
   "like": {"data": [], "summary": {"total_count": 189}},
   "love": {"data": [], "summary": {"total_count": 10}},
   "message": "Het zit erop! Dit was de zesde editie van Exact Live. Bedankt voor je komst en
     geniet nog even na met deze aftermovie! 🎥 \n\nTot volgend jaar? #exactlive",
   "sad": {"data": [], "summary": {"total_count": 0}},
   "shares": {"count": 38},
   "status_type": "added_video",
   "wow": {"data": [], "summary": {"total_count": 0}}}],
 "paging": {"cursors": {"after":
   "Q2c4U1pXNTBYM0YxWlhKNVgzTjBiM0o1WDJsa0R5RXhOell4Tmprd056RTBOVEU2TFRjNU16QTNOelUzTXpjd056QTJ
   OREEwTXpJUERHRndhVjl6ZAEc5eWVWOXBaQThlTVRjMk1UWTVNRGN4TkRVeFh6RXdNVFUxTVRRMU9UTTROVGMyTkRVeU
   R3UjBhVzFsQmxvTXRKb0IZD",
   "before": "Q2c4U1pXNTBYM0YxWlhKNVgzTjBiM0o1WDJsa0R5RXhOell4Tmprd056RTBOVEU2TFRnMU5USXhOek0xT
     URrNU5Ea3hNalkxTmpVUERHRndhVjl6ZAEc5eWVWOXBaQThlTVRjMk1UWTVNRGN4TkRVeFh6RXdNVFUxTWpZAME5qa
     3pPVE14TkRVeUR3UjBhVzFsQmxwTFg2QUIZD"},
```

Figure B.2: Example output of posts query from a Facebook page

Appendix C

Random and Weighted Guess

In this appendix, we explain two non-machine learning classifier, random guess and weighted guess, that we use as baseline for evaluating our credit scoring model. Given n is total sample and x is ratio of positive items, P is the number of actual positive sample or xn, and N is number of actual negative sample or $(1-x)n$, both classifier can be defined as follows:

- **Random Guess**: randomly assign half of items to positive and the other half as negative.
- **Weighted Guess** : randomly assign x of items to positive, and the remaining $(1-x)$ items to negative.

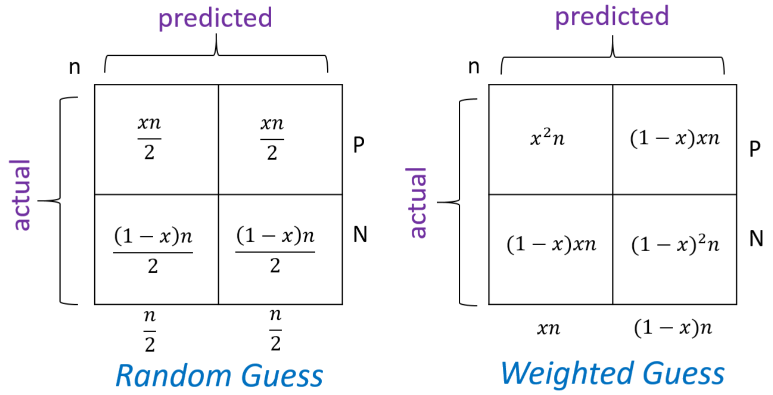


Figure C.1: Confusion matrix for random guess and weighted guess classifier

Based on the definition above, we can visualize the confusion matrix produced by those classifier as shown by Figure C.1. Based on the information about TP, FP, FN, and TN from the confusion matrix, we measure the performance of model including area under curve of ROC (AUC), accuracy, precision, recall, f1-score, and Matthews correlation coefficient (MCC). In the context of credit scoring, those evaluation metrics can be estimated using following formula:

- **AUC** is the probability that a randomly selected positive item has a higher score than that of a randomly selected negative item. As both does random scoring, the AUC is **0.5**.

- **Accuracy** is ratio of correctly predicted items to the total items.

$$Accuracy = \frac{TP+TN}{(TP+FP+TN+FN)}$$

- **Precision** is ratio of correctly captured relevant items among the items predicted as relevant.

$$Precision = \frac{TP}{(TP+FP)}$$

- **Recall** is ratio of correctly captured relevant items among the total of relevant items (P).

$$Recall = \frac{TP}{(TP+FN)}$$

- **F1-score** is harmonic mean between precision and recall.

$$F1\text{-}score = 2\frac{Precision \cdot Recall}{(Precision + Recall)} = \frac{2TP}{(2TP+FP+FN)}$$

- **MCC** is a measure of the quality of binary classifications, defined as:

$$MCC = \frac{(TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

The performance of both classifier based on evaluation metrics above are summarized in Table C.1 below.

Table C.1: Performance of Random Guess and Weighted Guess

Metrics	**Random Guess**	**Weighted Guess**
AUC	0.5	0.5
Accuracy	0.5	$x^2 + (1-x)^2$
Precision	x	x
Recall	0.5	x
F1-score	$\frac{x}{x+0.5}$	x
MCC	0	0

Appendix D

Modelling Details

This appendix contains the more details information of modelling process mentioned in Chapter 5. Figure D.1 and D.2 below shows the correlation matrix for features after the variance based feature selection for traditional and social media features.

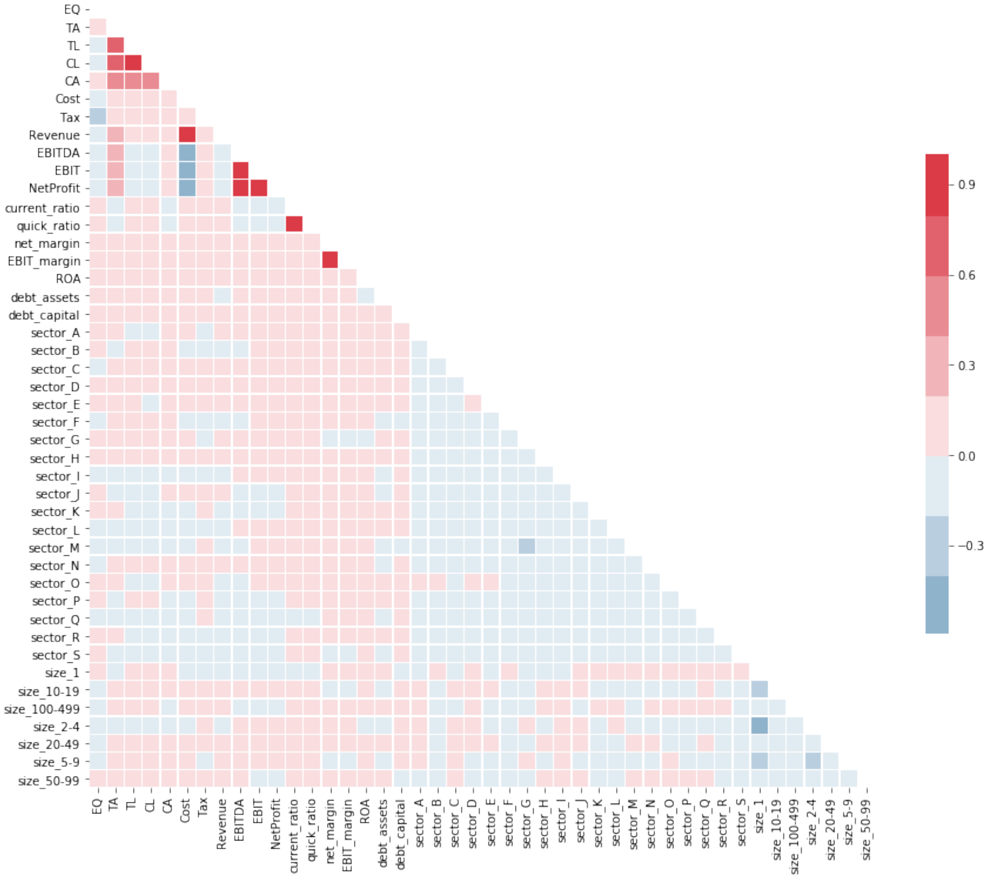


Figure D.1: Correlation matrix for traditional features

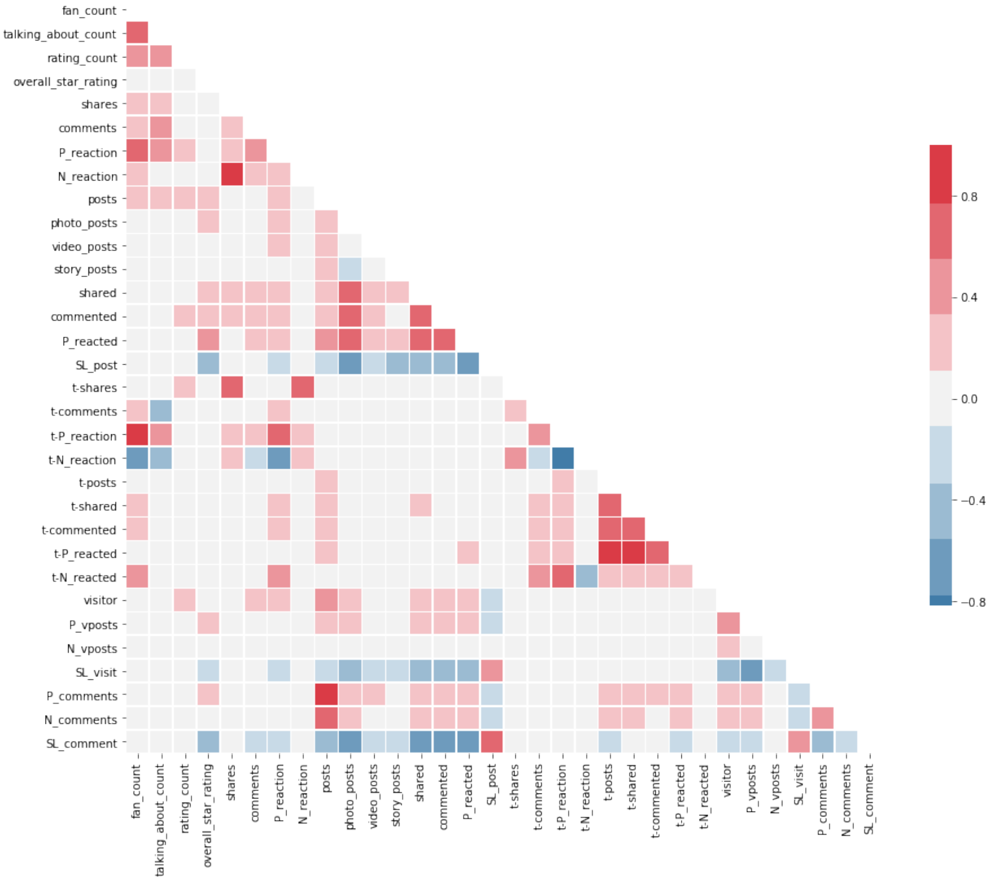


Figure D.2: Correlation matrix for social media features