MEF UNIVERSITY

ALTERNATIVE CREDIT SCORING MODEL FOR THIN FILE CUSTOMERS

Capstone Project

İstem Akca Korkmaz

İSTANBUL, 2019

MEF UNIVERSITY

ALTERNATIVE CREDIT SCORING MODEL FOR THIN FILE CUSTOMERS

Capstone Project

İstem Akca Korkmaz

Asst. Prof. Duygu TAŞ KÜTEN

İSTANBUL, 2019

MEF UNIVERSITY

Name of the project: Alternative Credit Scoring Model for Thin File Customers
Name/Last Name of the Student: İstem Akca Korkmaz
Date of Thesis Defense: …./…./2019

I hereby state that the graduation project prepared by İstem Akca Korkmaz has been completed under my supervision. I accept this work as a "Graduation Project".

…./…./2019
Asst. Prof. Duygu TAŞ KÜTEN

I hereby state that I have examined this graduation project by İstem Akca Korkmaz which is accepted by her supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

…./…./2019

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of İstem Akca Korkmaz and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
• Asst. Prof. Duygu Taş Küten	………………………..
• Prof. Dr. Özgür Özlük	………………………..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name	Date	Signature
İstem Akca Korkmaz	…./…/2019	

EXECUTIVE SUMMARY

ALTERNATIVE CREDIT SCORING MODEL FOR
THIN FILE CUSTOMERS

İstem Akca Korkmaz

Advisor: Asst. Prof. Duygu TAŞ KÜTEN

AUGUST, 2019, 21 pages

Credit scoring is a widely used tool for banks, financial institutions or corporations. Traditional credit score models are calculated from past financial history of users, and this may lead to exclude some people who have limited financial history from the credit system. Alternative credit scoring allows sector players to access to a larger portion of these customers. The credit scoring industry has expanded with an "all data is credit data" approach that combines traditional credit scoring systems with new data points.

In this study, we aim to build an alternative credit scoring model for customers who have limited financial historical data (thin file) by using alternative data points for a national bank in Turkey. Some of the alternative data points and variables have been gathered from one of the bank’s products: the authorized card for Turkish national league football tickets (Passolig). Using alternative data points combining with demographical and geographical information, we perform a comparison between the machine-learning approaches. We use logistic regression approach as a base model and perform a comparison between tree-based approaches: decision tree, random forest and XGBoost to select the most effective modelling approach.

Key Words: Alternative Credit Scoring, Thin File Customers, Binary Classification Techniques, Logistic Regression, Tree Based Algorithms

ÖZET

KREDİ GEÇMİŞİ AZ OLAN KİŞİLERE YÖNELİK ALTERNATİF KREDİ PUANLAMA MODELLERİ

İstem Akca Korkmaz

Tez Danışmanı: Asst. Prof. Duygu TAŞ KÜTEN

AĞUSTOS, 2019, 21 sayfa

Kredi puanlama yöntemleri bankalar, finansal kurumlar ve şirketler tarafından yaygın olarak kullanılır. Geleneksel kredi puanlama yöntemleri, finansal kullanıcıların geçmiş verilerine dayanarak hesaplanır ve bu durum, finansal geçmişi sınırlı olan kişilerin kredi sisteminin dışında kalmasına yol açabilir. Alternatif kredi puanlama yöntemleri, sektör oyuncularının bu kişilerin büyük bir kısmına erişmesine olanak sağlar. Geleneksel kredi puanlama yöntemlerini yeni alternatif veri kaynaklarıyla birleştiren kredi puanlama sektörü, "tüm veriler kredi verisidir" yaklaşımıyla genişlemektedir.

Bu çalışmada, Türkiye'deki bir ulusal bankanın kredi geçmişi az olan müşterilerine, alternatif veriler kullanarak bir kredi puanlama modeli oluşturmak amaçlanmaktadır. Alternatif veri kaynağı olarak bankanın ürünlerinden biri olan Türkiye ulusal futbol ligi yetkili kartı Passolig verileri kullanılmıştır. Demografik ve coğrafi verilerle birleştirilen bu alternatif veri farklı makine öğrenimi yaklaşımlarıyla modellenerek karşılaştırılmıştır. Lojistik regresyon yaklaşımı temel model olarak alınmış ve karar ağacı, rasgele orman ve XGBoost gibi ağaç tabanlı yaklaşımlarla karşılaştırılarak en etkili modelleme yaklaşımına ulaşılmaya çalışılmıştır.

Anahtar Kelimeler: Alternatif Kredi Puanlaması, Kredi Geçmişi Az Olan Müşteriler, İkili Sınıflandırma, Lojistik Regresyon, Ağaç Tabanlı Algoritmalar

TABLE OF CONTENTS

.

1. INTRODUCTION

Credit scoring is a widely used tool for banks, financial institutions or corporations that open a credit account for the customer while selling a product. The risk of nonpayment has led to lenders use a systematic credit scoring, so that make reliable decisions about whom to offer credit. Credit scores are not only used for lending decisions, many employers review credit reports when determining whom to hire, or when deciding whether to promote an existing employee (Hurley and Adebayo, 2016).

Traditional credit score models are calculated from past financial history of users. This traditional approach may lead to exclude some people who have limited or no financial history from the credit system (Pedro et al., 2015). Traditional credit scoring models do not cover a significant proportion of consumers globally, especially among those with thin or no files like millennials, members of Gen Z, refugees and immigrants (Stafferöd Westerlund, 2019).

Using traditional credit score models does not create a problem for only non-banked or thin file customers; it also results in a large amount of missed opportunity for banking sector and financial institutions. In the global economic conditions, with rates on the rise, banking sector seeks new strategies for the shifting lending landscape.

Alternative credit scoring allows sector players to lend more responsibly and help more qualified customers, with more accurate and expanding access to a larger portion of the global economy.

2. LITERATURE REVIEW

In this section, we present a review of the literature on alternative credit scoring and machine learning applications. In Section 2.1, we summarize some studies that represents the use of alternative data sources for credit scoring models. Section 2.2, we present the review of papers that use different machine learning models for credit scoring.

2.1. Alternative Data

Hurley and Adebayo (2016) discuss the current and future place of big data applications for credit scoring. The credit scoring industry has expanded with an "all data is credit data" approach that combines traditional credit scoring systems with new data points mined from consumers' offline and online footprints. The study presents an overview of techniques and methodologies that big data credit scoring likely use to design, test, and deploy machine-learning tools to assess creditworthiness. Scroll down pace of a loan applicant while scanning online terms and conditions or geographic location can be indicators of a high-risk borrower. These non-traditional data points can be used for alternative credit scoring models.

Pedro et al. (2015) present an approach to build a model of financial risk assessment from mobile phone usage detail records gathered from telecommunication companies. Every time a mobile phone is used, the communication event is logged into telecommunication companies' database as a CDR (Call Detail Record) entry. CDRs contain information about the details of the communication event: caller ID and dialed number, time and date of the call or SMS, duration and so on. BTS (Base Transceiver Station) connects mobile devices within a telecommunications network through set of cell towers and allows to get and receive signals. Records from BTS which provide geographical location in latitude and longitude of the communication events are also logged into the database. Pedro et al. (2015) combine these data points and apply supervised machine learning methods to build a new credit scoring model named "MobiScore". This approach allows authors to create an alternative credit scoring model for thin file customers in a Latin America country who cannot take part in the credit system because of the lack of past financial records.

Schoen et al. (2013) present a comprehensive review on models using social media data as a rich source of data for individuals. Researchers use the social media data for various prediction models such as stock market movement predictions, forecasts for movie boxoffice revenues, prediction of election outcomes and so on. The repository of accumulated data on social media (such as education status, the number of followers, work history, shares, activities, whom they are friends with) provides a lot of information about individuals without financial background. Wei et al. (2015) presents that there are advantages to collect information from an individual's network rather than only individualized data. Consumers have the above average chance of interaction with people who have similar creditworthiness, and thus network-based scoring can help lenders to reduce misjudgments about customers who have limited personal financial history.

There exists an adequate and expanding amount of literature on alternative credit scoring models and alternative data points. In the big data era, possibilities for alternative data sources is numberless; that's why; it is possible to determine alternative data points based on the specific business needs and the accessibility of the data. In this study, we use detailed data points of the customers who have the authorized card for Turkish national league football tickets (Passolig) whereas have thin financial history in the bank's database.

2.2. Modeling

Abdou et al. (2011) present a review of different statistical methods applied in building credit scoring models. Regression analysis, support vector machines, discriminant analysis, decision trees, logistic regression, neural networks, k-nearest neighbors are widely used examples for credit scoring models. Among these, it is not possible to talk about an approach that works best and always works well.

Munkhdalai et al. (2019) compare the results of several machine learning approaches, and FICO credit scoring system which is a human expert-based model for credit scoring. The authors also comprehensively review the most recent studies in credit scoring to determine the machine algorithms for using their comparative study. They encounter that most of the studies compare their recommended methods with logistic regression approach, based on the review of documentation. Louzada et al. (2016) represent a broad and also systematic review on studies related with theoretical and practical approaches on binary classification methods for credit scoring over the years. In this paper, the authors classify the methods used in certain aspects by covering researching studies made between the years 1992 and 2015; including 187 papers. As illustrated in Figure 1, the logistic regression is one of the most used binary classification techniques among all during the considered time period, on the comparison studies. One of the main reasons of using logistic regression widely for credit scoring models is interpretability. Logistic regression models are easy to interpret for knowledge extraction from the components of the model (Munkhdalai et al., 2019). If there is a rejection decision based on a credit scoring model, banks need to provide the reasons of rejection to the certain regulatory parties. Logistic regression models are transparent in terms of providing the functional relationship of the variables. (Dong et al., 2010)

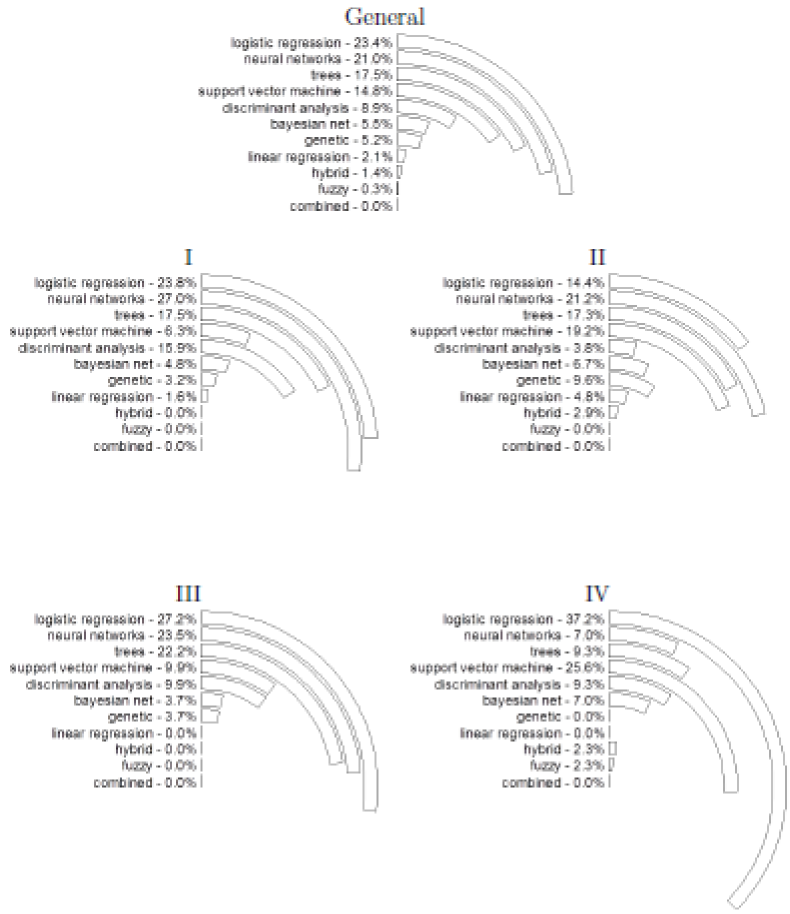


Figure 1: Circular bar plots concerning the techniques used in the paper's comparison studies. Reprinted from “Classification methods applied to credit scoring: Systematic review and overall comparison,” by F. Louzada, A. Ara, and G. B. Fernandes, 2016, Surveys in Operations Research and Management Science, 21(2), 117-134.

Mues et al. (2004) propose to use of decision diagrams based on the decision tree models in credit scoring to develop easily understandable and applicable models in daily practice. Decision tree models represent non-parametric statistical methods that provide high flexibility without assumption on data distribution (Jiang, 2009). Galindo and Tamayo (2000) address decision tree models as an example for the transparent models that can be conveniently interpretable by the local decision maker. Moreover, the set of attributes to be used for credit scoring models may contain missing values for some individuals. For example, the bank transaction information may be unavailable for a new bank customer. Missing value imputation approach can be chosen under these conditions; however, the real-time imputation will require additional computational power and time. Considering the credit allocation decisions need to be made instantly, decision tree-based methods which are not overly sensitive to the loss values can offer an effective solution. Jiang (2009) considers the advantage of decision trees as the background knowledge for the users is less required. As shown in Figure 1, decision tree algorithms are the third most used binary classification techniques in general, stated specifications of the decision tree-based models can seem to be the reasons for this.

On the other hand, Bastos (2007) and Zhang et al. (2008) claim that decision tree models have limitations on the stability of classification accuracy. Small variations on a variable may lead large changes in classification results. For instance, considering two features that have similar classification power on a dataset, if there is a small change in one of them, decision tree algorithm may split a node by using the other feature rather than the previous one. This tendency of the decision trees may create an entirely different split and tree structure than the classification based on the former feature (Bastos, 2007). Munkhdalai et al. (2019) and Louzada et al. (2016) summarize ensemble methods used for improving the performance of credit scoring models. Ensemble methods combines several decision trees to reach the better classification performance than a single decision tree. Bagging and boosting are two of most popular ensemble methods. The main idea of the ensemble methods is to use set of a weak learners to create a one strong learner using the same learning algorithm. Bagging method chooses each weak learner model independently, learns in parallel and combines the results by averaging the responses of the weak learners. On the other hand, boosting method chooses the weak learner models sequentially by taking into account the previous ones' success. Random Forest is an algorithm that uses bagging method based on decision trees, and XGBoost is an algorithm that can apply boosting technique on both linear model solver and tree learning algorithms.

Our primary goal in this study is to build an alternative credit scoring model using real consumer data and to provide machine-learning approaches that can serve as a baseline. Therefore, we use logistic regression approach as a base model and perform a comparison between tree-based approaches: decision tree, random forest and XGBoost to select the most effective modelling approach for our alternative data and features.

3. PROJECT DEFINITION

3.1. Project Objectives

This project aims to build alternative credit scoring model for customers who have limited financial historical data (thin file) by using alternative data points for a national bank in Turkey that currently uses only traditional scoring approach.

The breakdown of project objectives are as follows:

- Using alternative data points combining with demographical and geographical information,
- Building several binary classification algorithms with alternative data,
- Evaluating the best performing model.

With the alternative scoring model, the bank can expand the credit penetration in the national market and reach the customers who have limited financial historical data (thin file).

3.2. Project Scope

In this project, some of the alternative data points and variables have been gathered from one of the bank’s products: the authorized card for Turkish national league football tickets (Passolig). The card is mainly used to buy combined or single tickets for Turkish national football league, where monetary transactions can also be performed by the users. We have combined several spending data points from the authorized card, and demographic and geographical data. Additionally, we have added the past credit status data of the customers, who have the card, into dataset.

We consider this business problem as a binary classification problem where the target variable is credit status. We apply logistic regression, decision trees, random forest and XGBoost algorithms based on these past different data points, in order to calculate an alternative credit score for the future customer.

4. EXPLORATORY DATA ANALYSIS

4.1. Data Summary

In this project, we used a dataset shared by the bank including 142 variables and 40,370 unique customer records who both has a Passolig card and already got loan from the bank (See Appendix A). The variables contain different kinds of data types and various information that we have combined in following headings:

Variable Category	No of Variables
Bank Acquisition	4
Bank Service Transactions	5
Card - Shop	24
Card - Top up	6
Card - Transaction	12
Card - Withdrawal	6
Credit	20
Date	13
Demographical	21
Geographical	6
ID	2
Passolig - Football Tickets	20
Telecom Invoice	3
TOTAL	142

Table 1: Summary of Variables

- Bank acquisition and bank service transactions: Bank acquisition variables indicate the channel and product information of the customers. Bank service transaction variables provides information about the activities like password inquiry or payment of visa fee for the card.
- Card: This group of variables indicates shopping, transaction, withdrawal and top up activities such as top up TL to Passolig debit card or shopping transaction made with Passolig card excluding transactions related with national football league.
- Credit: Credit variables include data for credit status, application amount, interest rates and late payment credits. We have defined "credit status" as the target variable which is classified into this heading.

- Date: Date variables give information about the time at which the credit or card application is started and the beginning of legal proceedings.
- Demographical: These variables include age, gender, the place of birth, marital and educational status, the address and e-mail information.
- Geographical: These variables include specific late payment ratios by county and district calculated based on the past data in the bank's database.
- Passolig: The variables related with Passolig give information about tickets purchased. In addition, there are variables indicates the class of the tickets bought such as VIP, combined and so on.
- Telecom Invoice: These variables include late payment or legal proceedings for telecom invoices that are ordered as automatic payment by the user.

Credit status is our target variable and named as "KREDI_HESAP_DURUM" in the dataset. Distribution of the certain credit status types in the dataset can be found as below:

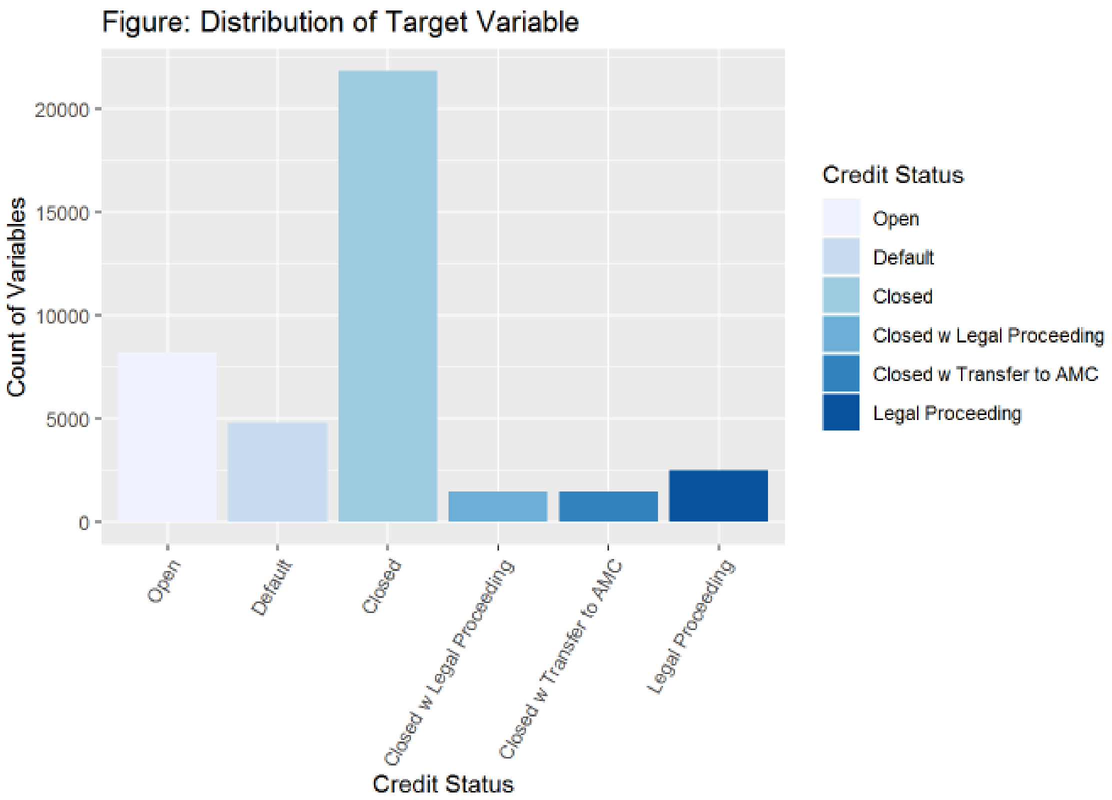


Figure 2: Distribution of Target Variable

“Closed” and “Open” status show that loans that are paid and closed, and still have installments, respectively. “Default” status implies loans that have late payments in the maturity terms. “Closed w Legal Proceeding” and “Legal Proceeding” status reveal the loans closed with legal proceedings or have an ongoing legal process. “Closed w Transfer to AMC” status shows the loans that are transferred to asset management companies because of the late payments. Categories rather than “Closed” and “Open” show the loans that have the problem in the installments. In the pre-processing stage, we thus have to manipulate these four categories into one category as “Default” and the remaining two categories combined as “Good” for binary classification models.

4.2. Pre-processing and Exploratory Analysis

During the pre-processing phase, we have excluded 15 out of 20 variables from “Credit” category since these variables indicate the results of the credit status and may cause multicollinearity for our models. Remaining variables in “Credit” category includes numerical variables like the day of the week or the hour of the day of the loan application. These variables may contribute to regression model as categorical indicators. Additionally, we have excluded the variables in “ID” and “Date” categories. “ID” category has distinct ID features which would be no contribution to the classification models. There are other separate columns rather than variables in “Date” category that show information about the duration of the membership or duration from last shopping date and so on. Thus, there is no need to make feature engineering on “Date” category to create new variables. There are 6 more variables that are excluded from the dataset that we have examined through visualizations and indicate no correlation with the target variable or same values for all rows.

After excluding certain variables, we have analyzed the ratio of missing variables for each remaining feature. 70 variables out of remaining 108 have missing values in different ratios.

NA % Range	No of Variables
>90	10
70-89	30
40-69	19
10-39	5
<10	6

Table.2: Summary Table of Ratio of Missing Variables

When we investigate missing values for each variable, 64 variables out of 70 variables show shopping, transaction, withdrawal and top up activities. Thus, missing values for these variables indicate that there is no transaction for this specific customer. Missing value ratios for these variables are between 90% and 10%. We have imputed missing values with zero for these 64 features, because all of them are numerical values and we would like to investigate their contribution to model in the modeling stage.

Remaining 6 variables indicates NPL values which shows specific late payment ratios by county and district calculated based on the past data in the bank’s database. These ratios can be strong indicators for regression model, and thus we have considered to impute missing values with mean/median imputation approach. We have calculated mean and median values for each column and replaced missing variables with both mean and median. We have compared mean imputed, median imputed and original distribution of variables through data visualizations (See Appendix B). We have not come across with significant differences on distributions. On the other hand, median value for “Ilce NPL – Tasıt” variable is calculated as zero which means that if we impute missing variables for this feature with median value, default probability for missing values would be 0. This may mislead our algorithms for calculating default probability. Therefore, we have move forward to model building with mean imputed variables.

We have grouped our target variable into two categories as “Good” and “Default”. Additionally, we have grouped some categorical variables by setting range like age or the application hour of the day, and created two different datasets with grouped and non-grouped categorical variables. Since all categorical variables should be converted into binary variables for regression model, we would like to decrease the number of variables so that we can save time on computing. However, this can create a certain risk on accuracy or classification effectiveness of the model, thus, we have created two different versions and compared the results of the models for both versions to find the most efficient and accurate model for credit scoring.

5. METHODOLOGY

5.1. Logistic Regression

Logistic regression builds a model based on the estimation of linear combination between the explanatory variables and the binary response variable, and transforms log-odds to probability with logistic function (Munkhdalai et al., 2019). Variation of the explanatory variables affect the classification performance of a regression model; thus, we evaluate regression models for the determination of optimal parameters. One of the measurement criteria for model selection with optimal parameters is AIC value. AIC is a relative measure of model parsimony and estimates information loss with respect to different models. As AIC value indicates the relative information loss among the variation of the variables, the model with a lower AIC value is healthier.

To build a successful credit scoring model, a classifier needs to be understandable, accurate and fast (Liu, 2002). Considering ease of interpretability of the model, we seek for a model as simple as possible with a combination of reliable accuracy rate and calculation speed. Therefore, we compare the AIC score of the different variations of regression models as well as their accuracy rate via confusion matrices and ROC curves.

We build two different logistic regression models as our initial models with two different datasets that we have created during the pre-processing phase: dataset with non-grouped categorical variables (Model 1) and dataset with grouped categorical variables (Model 2). We try many iterations with two base models by feature selection based on *p*-values of each independent variables. We summarize accuracy and AUC values of model iterations with comparative AIC scores in Table 3.

Model Iteration	AIC	Accuracy %	AUC %
Model 1.a	42,926	75.58%	69.10%
Model 2.a	42,932	75.50%	68.60%
Model 1.b	42,923	75.57%	69.10%
Model 2.b	42,930	75.49%	68.60%
Model 1.c	42,919	75.52%	68.70%
Model 2.c	42,912	75.51%	68.50%

Table 3: Evaluation Metrics of Logistic Regression Iterations

Model 1.a and Model 2.a corresponds to the initial model that is trained with all variables. According to the summary of the results of Model 1.a and Model 2.a, there are some variables that are strongly correlated with other variables. We thus have excluded these variables and built a revised model: Model 1.b and Model 2.b. Then, we have investigated *p*-values and selected the variables that have *p*-values very close to 1 and exclude them from the model. We have created a new model based this selection, that are Model 1.c and Model 2.c.

We compare the evaluation metrics of each model. While evaluating a model performance, accurate approach would interpret the combination of certain metrics and evaluate the tradeoff between them (Liu, 2002). Although accuracy score and AUC value are slightly lower than the rest of the models, Model 1.c and Model 2.c seem as healthier models in terms of AIC value. We consider this difference on accuracy and AUC value is acceptable. When we compare Model 1.c and Model 2.c, accuracy and AUC values are almost similar. Therefore, we consider to move forward with Model 2.c because of its better performance in calculation time. Also, we apply different algorithms to grouped dataset in order to compare the model performances, accordingly.

5.2. Tree-Based Models

Decision tree models basically creates classification models by creating a set of if-then determination conditions in tree-based structures. Decision tree algorithms can be used for both regression and classification problems. CART (Classification and Regression Tree) method uses Gini index for classification criterion, for our binary classification problem we use CART method (Louzada et al., 2016).

We use the grouped dataset for training of the model and use “rpart” package to build a classification tree. We have built our first model with the default parameters of the package and investigated the results. Our initial tree model gives 74.36 % accuracy rate. Then, we try iterations with the model by changing “minsplit” value which corresponds to the minimum number of observations should exist in a node for a split to be attempted. Default value for this parameter is given as 30 in the “rpart” package. We set the value as 10 in order to create a more detailed tree and, inherently to obtain more accurate classification. As it is known, CP value represents complexity parameter and is used to control the size of the tree. If the cost of adding a variable from the current node is higher than CP value, then decision tree structure stops growing. Default CP parameter value is 0.01, however we prefer to create a bigger tree and set value as 0.001. With the parameter tuning, our decision tree classification model gives accuracy rate as 76.15%. It's possible to make more iterations via parameter tuning; however, this may create the risk of overfitting and increase in computing time.

Random Forest algorithms use a set of decision trees created from subset data that are randomly selected from the train dataset. Building a random forest classifier for our credit scoring model, we use "randomforest" package in R. We build the initial model with default parameters and investigate the results. However, the calculation time of the random forest model is quite high. As we aim to create an alternative model with speed, it is not preferred to apply parameter tuning for random forest algorithm since it would increase the number of trees. The accuracy rate of the initial random forest model is 75.93%.

XGBoost algorithm is an ensemble boosting method that optimizes the size of the tree and objective function with regularization parameters. We create initial boosting model with default parameters, and we then investigate the results by iterations with parameter tuning. XGBoost algorithm is quite flexible and fast, that we can test various iterations in a short amount of calculation time. Nevertheless, in this study, we aim to compare different algorithms for the alternative credit scoring model, and thus we have set parameters of XGBoost algorithm similar to decision tree model. Final XGBoost model gives 76.10% of accuracy rate.

6. CONCLUSIONS

6.1. Comparison of Evaluation Results of Models

Models on credit scoring problems predict “Default” and “Good” credit status as binary variables based on the past data. It is widely agreed that accuracy score is a simple and direct measurement of the binary classification performance. Nevertheless, accuracy score may not be sufficient as standalone for evaluation of the model. Interpretability, simplicity and velocity of the model are other practical benchmarks of the credit scoring models.

Furthermore, a lending institution should assess the risk of misclassification errors in terms of loss and opportunity cost. Confusion matrix may serve this purpose for estimating the risk. Confusion matrix provides a comparison table for actual and predicted class labels as illustrated:

Table : Confusion matrix.

		M	
		$\{1\}$	$\{0\}$
D	$\{1\}$	TP	FN
	$\{0\}$	FP	TN

Table 4: Confusion Matrix

If a customer’s status is predicted as “Good” while the actual status is “Default; it will create a commercial risk for the lending institution. This type of error is named “Type II Error” or “False Positive Rate (FPR)”. In other case, if a customer is classified as “Default” status while the actual status is “Good”, this will cause opportunity loss for the lending institution. The latter error is named “Type I Error” or “False Negative Rate (FNR)”. Thus, evaluation criteria for credit scoring models should be combined with multiple benchmarks so that the financial risk would be assessed. Therefore, we use accuracy rate, Type I error, Type 2 error and the computation time to compare our models.

	Logistic Regression	Decision Tree	Random Forest	XGBoost
Accuracy %	75.51%	76.15%	75.93%	76.10%
Type I Error (FNR)	1.54%	2.04%	2.41%	3.22%
Type II Error (FPR)	22.97%	21.81%	21.71%	20.77%
Calculation Time	26.62 sec	29.20 sec	14.48 min	40.16 sec

Table.5: Summary of Evaluation Metrics of the Models

The primary finding on comparison of accuracy scores is that all models have marginally close rates. XGBoost algorithm shows the best performance in terms of Type II error among all models. The computation time of the random forest algorithm is quite long, and thus it is not suggested to apply random forest algorithm in the bank's alternative scoring implementation. Although the decision tree performance seems sufficient; due to the possible risk about the limitations on the stability of classification accuracy, it is cautious to monitor how it will perform with future data points. In the light of this study, our suggestion is to use XGBoost, logistic regression and decision tree models for the bank's alternative scoring implementation by comparing the results together with future data points.

6.2. Conclusions

It is a necessity that credit scoring models become more efficient, accurate, and more inclusive for the people like have thin financial history, in rapidly changing world. Innovative models based on alternative data sources and machine learning applications creates opportunities for new customers as well as creates efficiency for the financial institutions on risk management.

It is easier to create alternative models with reasonable amount of time and resources thanks to the enhancement of machine learning applications and growing transactional data. As alternative credit scoring models advance, their capacity to precisely assess the risk will increase thanks to the learning loop created by data science and sector players. This enable more customers to reach financial credit and create growth for financial lending sector.

APPENDIX A

#	Columns	Data Type	Explanation
1	CREDIT_ID	Numerical	Application ID
2	BASVURU_TARIHI	Date	Application Date
3	KREDI_HESAP_DURUM	Categorical	Status of the credit
4	DONEM	Numerical	Application Date as "Year&Month"
5	CUST_ID	Numerical	Customer ID Number
6	MUSTERI_OLMA_TARIHI	Date	First Signature Date of Customer Service Contract
7	CARD_EMBOSS_DATE	Date	Last Print Date of Active Passolig Card (renewal, loss etc.)
8	FIRST_EMBOSS_DATE	Date	Last Print Date of Active Passolig Card
9	KART_TIPI	Categorical	Debit or Credit Card
10	TAKIM	Categorical	Footbal Team Name of Passolig Card
11	TEAM_ID	Categorical	Footbal Team ID of Passolig Card
12	KAYIT_TARIHI	Date	First Application Date before Signature of Customer Service Contract
13	SHOP_COUNT	Numerical	Count of Debit Card Shopping (excl. Single & Combined Football Tickets)
14	MAX_SHOP_AMOUNT	Numerical	Maximum Amount of Debit Card Shopping (excl. Single & Combined Football Tickets)
15	MIN_SHOP_AMOUNT	Numerical	Minimum Amount of Debit Card Shopping (excl. Single & Combined Football Tickets)
16	AVG_SHOP_AMOUNT	Numerical	Average Amount of Debit Card Shopping (excl. Single & Combined Football Tickets)
17	FIRST_SHOP_DAYS	Numerical	Duration from Debit Card Issue Date to 1st Shopping Date (excl. Single & Combined Football Tickets)
18	LAST_SHOP_DAYS	Numerical	Duration from Last Shopping Date to Credit Issue Date (excl. Single & Combined Football Tickets)
19	ONSHOP_COUNT	Numerical	Count of Debit Card Shopping - Online (excl. Single & Combined Football Tickets)
20	MAX_ONSHOP_AMOUNT	Numerical	Maximum Amount of Debit Card Online Shopping (excl. Single & Combined Football Tickets)
21	MIN_ONSHOP_AMOUNT	Numerical	Minimum Amount of Debit Card Online Shopping (excl. Single & Combined Football Tickets)
22	AVG_ONSHOP_AMOUNT	Numerical	Average Amount of Debit Card Online Shopping (excl. Single & Combined Football Tickets)
23	FIRST_ONSHOP_DAYS	Numerical	Duration from Debit Card Issue Date to 1st Online Shopping Date (excl. Single & Combined Football Tickets)
24	LAST_ONSHOP_DAYS	Numerical	Duration from Last Online Shopping Date to Credit Issue Date (excl. Single & Combined Football Tickets)
25	WD_COUNT	Numerical	Count of Debit Card Cash Withdrawal
26	MAX_WD_AMOUNT	Numerical	Maximum Amount of Debit Card Cash Withdrawal
27	MIN_WD_AMOUNT	Numerical	Minimum Amount of Debit Card Cash Withdrawal
28	AVG_WD_AMOUNT	Numerical	Average Amount of Debit Card Cash Withdrawal
29	FIRST_WD_DAYS	Numerical	Duration from Debit Card Issue Date to 1st Cash Withdrawal
30	LAST_WD_DAYS	Numerical	Duration from Last Cash Withdrawal to Credit Issue Date
31	INS_COUNT	Numerical	Count of Debit Card Cash Top up
32	MAX_INS_AMOUNT	Numerical	Maximum Amount of Debit Card Cash Top up
33	MIN_INS_AMOUNT	Numerical	Minimum Amount of Debit Card Cash Top up
34	AVG_INS_AMOUNT	Numerical	Average Amount of Debit Card Cash Top up
35	FIRST_INS_DAYS	Numerical	Duration from Debit Card Issue Date to 1st Cash Top up
36	LAST_INS_DAYS	Numerical	Duration from Last Cash Top up to Credit Issue Date
37	DC_SIFRE_COUNT	Numerical	Count of password transactions - Debit Card
38	DC_SORGU_COUNT	Numerical	Count of inquiries - Debit Card
39	DC_HARCAMA_COUNT	Numerical	Count of spending transactions - Debit Card (excl. Single & Combined Football Tickets)
40	DC_ALISVERIS_COUNT	Numerical	Count of shopping transactions - Debit Card (excl. Single & Combined Football Tickets)
41	DCTXN_COUNT	Numerical	Count of total transactions - Debit Card (excl. Single & Combined Football Tickets)
42	CC_ALISVERIS_COUNT	Numerical	Count of shopping transactions - Credit Card (excl. Single & Combined Football Tickets)
43	CC_HARCAMA_COUNT	Numerical	Count of spending transactions - Credit Card (excl. Single & Combined Football Tickets)
44	CC_FAIZ_COUNT	Numerical	Count of transactions on interest - Credit Card (excl. Single & Combined Football Tickets)
45	TXN_INSTALL_TYPE	Boolean	Installment Transactions or not? (Y:1, N:0) (excl. Single & Combined Football Tickets)
46	INSTALL_CNT	Numerical	Number of installment transactions (excl. Single & Combined Football Tickets)
47	CCTXN_COUNT	Numerical	Count of total transactions - Credit Card (excl. Single & Combined Football Tickets)
48	AMOUNT_PAID_SELF	Numerical	Average price of single footbal tickets bought for user by user
49	BOUGHT_OWN_TOTAL	Numerical	Count of single football tickets bought for user by user
50	AMOUNT_PAID_FOR_OTHERS	Numerical	Average price of single footbal tickets bought for others by user
51	BOUGHT_FOR_OTHER_TOTAL	Numerical	Count of football single tickets bought for others by user
52	AMOUNT_PAID_BY_OTHERS	Numerical	Average price of single footbal tickets bought for user by others
53	BOUGHT_BY_OTHER_TOTAL	Numerical	Count of single football tickets bought for user by others
54	COMBINED_PAID_SELF_TOTAL	Numerical	Total price of combined footbal tickets bought for user by user
55	COMBINED_OWN_TOTAL	Numerical	Count of combined footbal tickets bought for user by user
56	COMBINED_PAID_BY_OTHERS	Numerical	Total price of combined footbal tickets bought for user by others
57	COMBINED_BY_OTHERS_TOTAL	Numerical	Count of combined footbal tickets bought for user by others
58	COMBINED_PAID_FOR_OTHERS_	Numerical	Total price of combined footbal tickets bought for others by user
59	COMBINED_FOR_OTHERS_TOTA	Numerical	Count of combined footbal tickets bought for others by user
60	ISLEM_TARIHI	Date	Final Date for last single or combined football ticket purchase
61	TRIBUN_BILET	Boolean	Classification of Single Football Ticket purchased : VIP or not? (VIP:1)
62	TRIBUN_COMBINE	Boolean	Classification of Combined Football Ticket purchased : VIP or not? (VIP:1)
63	TRIBUN	Boolean	Is there any VIP status on total footbal tickets purchased? (Y:1, N:0)
64	TOTAL_PASSES	Numerical	Total number of entrances to stadium for games
65	VISA_FEE_FLAG	Boolean	Visa Fee for the card is paid or not? (Y:1, N:0)
66	VISA_START_DATE	Date	Start Date for Visa Fee Duration
67	VISA_END_DATE	Date	End Date for Visa Fee Duration
68	VISA_PAYMENT_DATE	Date	Date of Visa Fee Payment
69	CARD_HOLDER_DURATION	Numerical	Duration of Card Ownership
70	CARD_USABLE	Boolean	Visa Fee paid or not while getting credit? (Y:1, N:0)
71	VERILIS_TARIHI	Date	Date of issue of Passolig Card

72	ADRES_H_ISYERI	Boolean	Work Adress for confimation status H is taken or not? (Y:1 , N:0)
73	ADRES_E_ISYERI	Boolean	Work Adress for confimation status E is taken or not? (Y:1 , N:0)
74	ADRES_H_EV	Boolean	Home Adress for confimation status H is taken or not? (Y:1 , N:0)
75	ADRES_E_EV	Boolean	Home Adress for confimation status E is taken or not? (Y:1 , N:0)
76	TOTAL_SHOP_COUNT	Numerical	Count of Debit Card Shopping (incl. Single & Combined Football Tickets)
77	TOTAL_MAX_SHOP_AMOUNT	Numerical	Maximum Amount of Debit Card Shopping (incl. Single & Combined Football Tickets)
78	TOTAL_MIN_SHOP_AMOUNT	Numerical	Minimum Amount of Debit Card Shopping (incl. Single & Combined Football Tickets)
79	TOTAL_AVG_SHOP_AMOUNT	Numerical	Average Amount of Debit Card Shopping (incl. Single & Combined Football Tickets)
80	TOTAL_FIRST_SHOP_DAYS	Numerical	Duration from Debit Card Issue Date to 1st Shopping Date (incl. Single & Combined Football Tickets)
81	TOTAL_LAST_SHOP_DAYS	Numerical	Duration from Last Shopping Date to Credit Issue Date (incl. Single & Combined Football Tickets)
82	TOTAL_DC_ALISVERIS_COUNT	Numerical	Count of shopping transactions - Debit Card (incl. Single & Combined Football Tickets)
83	TOTAL_DC_HARCAMA_COUNT	Numerical	Count of spending transactions - Debit Card (incl. Single & Combined Football Tickets)
84	TOTAL_DCTXN_COUNT	Numerical	Count of total transactions - Debit Card (incl. Single & Combined Football Tickets)
85	TOTAL_CC_ALISVERIS_COUNT	Numerical	Count of shopping transactions - Credit Card (incl. Single & Combined Football Tickets)
86	TOTAL_CC_HARCAMA_COUNT	Numerical	Count of spending transactions - Credit Card (incl. Single & Combined Football Tickets)
87	TOTAL_CC_FAIZ_COUNT	Numerical	Count of transactions on interest - Credit Card (incl. Single & Combined Football Tickets)
88	TOTAL_INSTALL_CNT	Numerical	Number of installment transactions (incl. Single & Combined Football Tickets)
89	TOTAL_TXN_INSTALL_TYPE	Boolean	Installment Transactions or not? (Y:1, N:0) (incl. Single & Combined Football Tickets)
90	TOTAL_CCTXN_COUNT	Numerical	Count of total transactions - Credit Card (incl. Single & Combined Football Tickets)
91	Basvuru Kanal	Categorical	Application Channel
92	BASVURU TUTAR	Numerical	Credit Application Amount
93	ONAY TUTAR	Numerical	Approved Credit Amount
94	TOPLAM KULLANDIRIM_TUTAR	Numerical	Used Credit Amount
95	YASAL_TARIH	Date	Start Date of legal proceedings for the credit
96	YASAL_BAKIYE	Numerical	Amount of the credit on legal proceedings
97	Takip Tutar	Numerical	Amount on legal proceedings file
98	Basvuru Tarihi	Date	Application Date for Credit
99	Basvuru Günü	Numerical	Application Day for Credit
100	Basvuru Ayi	Numerical	Application Month for Credit
101	BasvuruHaftaGün	Numerical	Application Week&Day for Credit
102	Basvuru Saati	Numerical	Application Hour for Credit
103	DURUM_KODU	Categorical	Status of the credit
104	Faiz Orani	Numerical	Interest Rate
105	VADE	Numerical	Maturity
106	Kampanya Açiklama	Categorical	Description of Credit Campaign
107	Cinsiyet	Categorical	Gender
108	medeni_hal	Categorical	Marital Status
109	Yas	Numerical	Age
110	email_hotmail	Boolean	E-mail extension Hotmail or not? (Y:1, N:0)
111	email_gmail	Boolean	E-mail extension Gmail or not? (Y:1, N:0)
112	email_msn	Boolean	E-mail extension MSN or not? (Y:1, N:0)
113	email_outlook	Boolean	E-mail extension Outlook or not? (Y:1, N:0)
114	email_mynet	Boolean	E-mail extension Mynet or not? (Y:1, N:0)
115	email_wlive	Boolean	E-mail extension Wlive or not? (Y:1, N:0)
116	email_dottr	Boolean	E-mail extension Dottr or not? (Y:1, N:0)
117	email_icloud	Boolean	E-mail extension ICloud or not? (Y:1, N:0)
118	email_yahoo	Boolean	E-mail extension Yahoo or not? (Y:1, N:0)
119	email_upper	Boolean	E-mail extension Upper or not? (Y:1, N:0)
120	Calisma sekli	Categorical	Occupation Category
121	Meslek	Categorical	Occupation
122	Egitim	Categorical	Education
123	Dogum yeri	Categorical	Place of Birth
124	Kazanim Ürünü	Categorical	Acquisition Product
125	Kazanim Kanali	Numerical	Acquisition Channel Code
126	Kazanim	Categorical	Acquisition Channel
127	Kredi Gecikme Gün Sayisi 5	Numerical	Number of Default Days (default amount higher than 5 TL)
128	Geç Ödenmis Taksit Sayisi	Numerical	Number of Late Payment Installments
129	Ödenmemis Taksit Sayisi	Numerical	Number of Unpaid installments
130	Ödenmis Taksit Sayisi	Numerical	Number of Paid installments
131	Geç Ödenmis Taksit Sayisi 5	Numerical	Number of Late Payment Installments (default amount higher than 5 TL)
132	Gecikmedeki Taksit Adeti 5	Numerical	Number of Default Installments (default amount higher than 5 TL)
133	Bankamiz Müsterisi Mi	Boolean	Is the customer is regular bank customer?
134	Mahalle NPL	Numerical	Late Payment Probability rate by District
135	ILCE NPL	Numerical	Late Payment Probability rate by County
136	ILCE NPL - BAYi	Numerical	Late Payment Probability rate by County for "BAYİ" credits
137	ILCE NPL - TAsIT	Numerical	Late Payment Probability rate by County for "TASIT" credits
138	ILCE NPL - WEB	Numerical	Late Payment Probability rate by County for "WEB" credits
139	ILCE NPL - PTT	Numerical	Late Payment Probability rate by County for "PTT" credits
140	Max_Gecikmeli_Fatura_ToplamTutar	Numerical	Total amount of late payment telecom invoices
141	Max_Takip_Tutar	Numerical	Total amount of legal proceeding telecom invoices
142	Max_Gecikmeli_Fatura_ToplamAdedi	Numerical	Total number of late payment telecom invoices

APPENDIX B

Figure App B.1: Distributions “Mahalle NPL” with Mean & Median imputed values

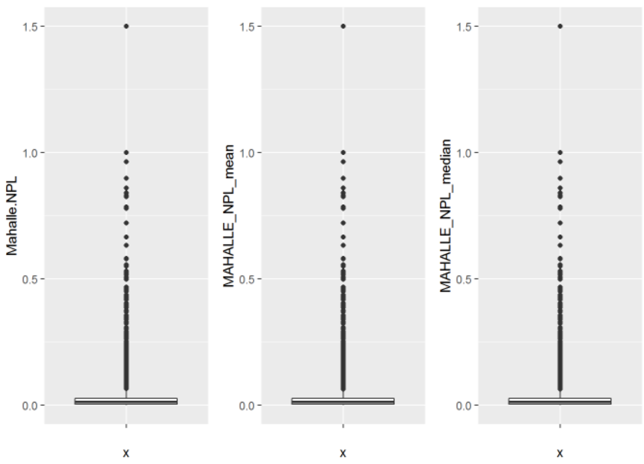


Figure App B.2: Distributions “İlce NPL” with Mean & Median imputed values

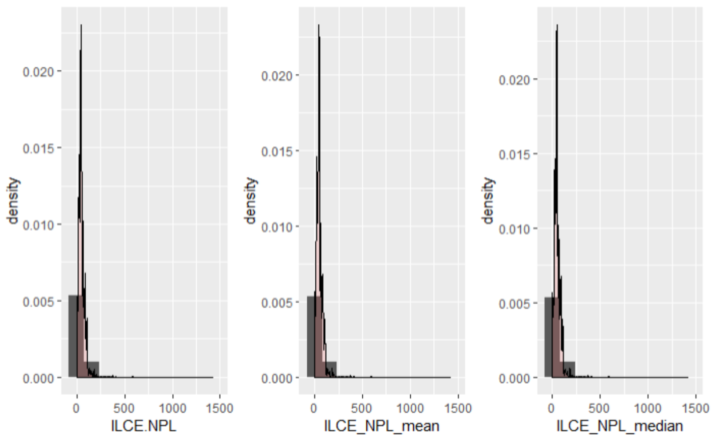
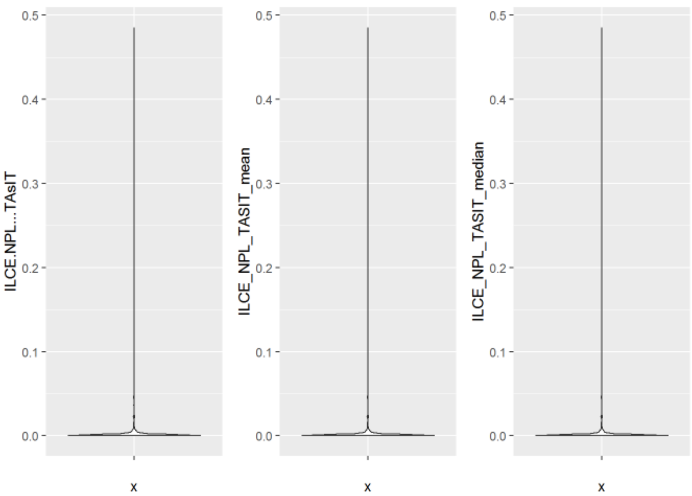


Figure App B.3: Distributions “İlce NPL - Tasıt” with Mean & Median imputed values



REFERENCES

[1] M. Hurley and J. Adebayo (2016), Credit Scoring in the Era of Big Data. Yale Journal of Law and Technology, 18(5), 149-202.

[2] San Pedro, J., Proserpio, D., and Oliver, N. (2015, June). MobiScore: towards universal credit scoring from mobile phone data. In International Conference on User Modeling, Adaptation, and Personalization (pp. 195-207). Springer, Cham.

[3] Stafferöd Westerlund, H. (2019, Feb 28). Transactional data: the future of credit scoring. [Web log post] Retrieved May 19, 2019, from https://blog.instantor.com/transactional-data-the-future-of-credit-scoring

[4] Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., and Gloor, P. (2013). The power of prediction with social media. Internet Research, 23(5), 528-543.

[5] Wei, Y., Yildirim, P., Van den Bulte, C., and Dellarocas, C. (2015). Credit scoring with social network data. Marketing Science, 35(2), 234-258.

[6] Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intelligent Systems in Accounting, Finance and Management, 18(2-3), 59-88.

[7] Munkhdalai, L., Munkhdalai, T., Namsrai, O. E., Lee, J. Y., & Ryu, K. H. (2019). An empirical comparison of machine-learning methods on bank client credit assessments. Sustainability, 11(3), 699.

[8] Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. Surveys in Operations Research and Management Science, 21(2), 117-134.

[9] Dong, G., Lai, K. K., & Yen, J. (2010). Credit scorecard based on logistic regression with random coefficients. Procedia Computer Science, 1(1), 2463-2468.

[10] Mues, C., Baesens, B., Files, C. M., & Vanthienen, J. (2004). Decision diagrams in machine learning: an empirical study on real-life credit-risk data. Expert Systems with Applications, 27(2), 257-264.

[11] Jiang, Y. (2009, March). Credit scoring model based on the decision tree and the simulated annealing algorithm. In 2009 WRI World Congress on Computer Science and Information Engineering (Vol. 4, pp. 18-22). IEEE.

[12] Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. Computational Economics, 15(1-2), 107-143.
[13] Bastos, J. (2007). Credit scoring with boosted decision trees.
[14] Zhang, D., Leung, S. C., & Ye, Z. (2008, November). A decision tree scoring model based on genetic algorithm and k-means algorithm. In 2008 Third International Conference on Convergence and Hybrid Information Technology (Vol. 1, pp. 1043-1047). IEEE.
[15] Liu, Y. (2002). The evaluation of classification models for credit scoring. Institut für Wirtschaftsinformatik, Georg-August-Universitat Göttingen.