Credit Risk Analysis Using Machine Learning

Panthangi Praharshitha[1], J Vinoj[2], Vurimi Venkata Krishna Vamsi[3], Pusuluri Pujitha[4], Thota Mohan Koteswar[5]

[1,2,3,4,5] *Department of CSE, Vignan's Foundation for Science, Technology and Research, Vadlamudi, Guntur, Andhra Pradesh, India*

Abstract: Credit risk assessment is a critical task for financial institutions when de- termining applicants' eligibility for credit products, such as credit cards. This study explores the use of machine learning techniques to predict credit card ap- proval outcomes based on applicants' demographic and financial information. The analysis uses the "Credit Card Approval Prediction" dataset from Kag- gle, comparing two machine learning models: Logistic Regression and Random Forest Classifier. After performing necessary data preprocessing, including han- dling missing values, encoding categorical variables, and addressing class im- balance through SMOTE, the models are evaluated using metrics like accuracy, precision, recall, and ROC-AUC. The results show that Random Forest Classi- fier outperforms Logistic Regression in prediction accuracy, demonstrating its potential to enhance decision-making in credit risk management.

Keywords: Credit Risk Assessment, Credit Card Approval, Machine Learn- ing, Logistic Regression, Random Forest Classifier, SMOTE, Data Preprocess- ing, Class Imbalance, Model Evaluation, Financial Risk

1. INTRODUCTION

Credit risk assessment is an integral part of the financial industry, helping institutions make informed decisions on lending. With the increasing volume of credit applica- tions, traditional manual methods of evaluation have become inefficient, prompting a shift toward automated solutions powered by machine learning. This study aims to apply machine learning algorithms to predict credit card approval based on various factors such as applicants' income, employment status, and credit history. By utiliz- ing the "Credit Card Approval Prediction" dataset, we compare the performance of Logistic Regression and Random Forest Classifier, two commonly used models. The research highlights the importance of preprocessing techniques, including addressing missing data and class imbalance, in achieving effective model performance. Ulti- mately, the study demonstrates the value of machine learning in streamlining and improving the accuracy of credit risk management processes.

2. LITERATURE SURVEY

Credit risk prediction has been a key area of interest for researchers, with numerous studies exploring various techniques and models to predict loan or credit defaults. Madaan et al. (2021) [1] presented a comparative study between Decision Trees and Random Forests in loan default prediction. They found that Random Forests provided superior prediction accuracy, a result that is consistent with the findings of this study, where the Random Forest Classifier outperformed other models in credit card approval prediction. Perera and Premaratne (2016) [2] explored the use of Artificial Neural Networks (ANNs) for predicting the payment behavior of leasing customers in Sri Lanka. Their approach highlighted the advantages of ANNs in capturing complex patterns in financial data. Similarly, this study demonstrates the utility of machine learning in predicting credit outcomes, though it focuses on simpler models like Lo- gistic Regression and Random Forests. Marqués et al. (2012) [3] delved into the behavior of base classifiers in credit scoring ensembles. Their research suggested that ensemble methods, such as Random Forests, significantly improve classification ac- curacy compared to single classifiers. This aligns with the current study's choice of Random Forest, which showed a better performance than Logistic Regression in predicting credit card approvals. Adewusi et al. (2016) [4] applied ANNs for loan recovery prediction, offering insights into the potential of neural networks in predict- ing credit risk outcomes. While this study employed a different type of model, the insights underscore the growing interest in using machine learning techniques to enhance financial decision-making. Choudhary et al. (2019) [5] focused on loan default identification and its effects on the financial sector. Their work emphasized the im- portance of accurate prediction models for mitigating the risks associated with loan defaults. This study also echoes the need for accurate risk assessment tools, particu- larly in the context of credit card approval. Atiya (2001) [6] provided a comprehensive survey on bankruptcy prediction using neural networks, which is a critical aspect of credit risk management. Although this research utilized neural networks, it highlights the ongoing trend of applying advanced machine learning techniques to predict finan- cial outcomes, a trend also explored in this study. Li et al. (2014) [7] investigated the comparative risk and return characteristics of high-yield and investment-grade bonds. While this study focused on bonds rather than credit cards, it contributes to the broader understanding of risk prediction models in financial markets, emphasizing the importance of accurately assessing risk. Pandit (2016) [8] used data mining techniques on loan approval datasets to predict defaulters. This dissertation pro- vided valuable insights into feature selection and data preprocessing, methods that were also applied in this study to preprocess the ”Credit Card Approval Prediction” dataset. Calcagnini et al. (2018) [9] examined the hierarchy of bank loan approval and performance. Their research contributes to understanding how various factors influence loan approval decisions, reinforcing the importance of accurate prediction systems in the lending process. Lastly, Sarma (2013) [10] focused on predictive mod- eling using SAS Enterprise Miner for business applications. This work contributed to the growing body of knowledge on predictive analytics in finance, offering practical solutions that could be applied to credit risk prediction.

3. METHODOLOGY

The flowchart (Figure 1) presents a human-generated process for creating and de- ploying a predictive model. The pipeline starts with Data Collection from sources like application and credit datasets. This is followed by Data Preprocessing, which includes cleaning, merging, and feature engineering. In Model Training Evaluation, techniques like logistic regression and random forest are used. The Model Deploy- ment stage exposes the trained model through an API. Feedback Loop Monitoring then tracks performance and allows for retraining. Finally, a User Interface is devel- oped, providing a dashboard for user interaction. This systematic approach ensures efficient, iterative improvement and user accessibility.

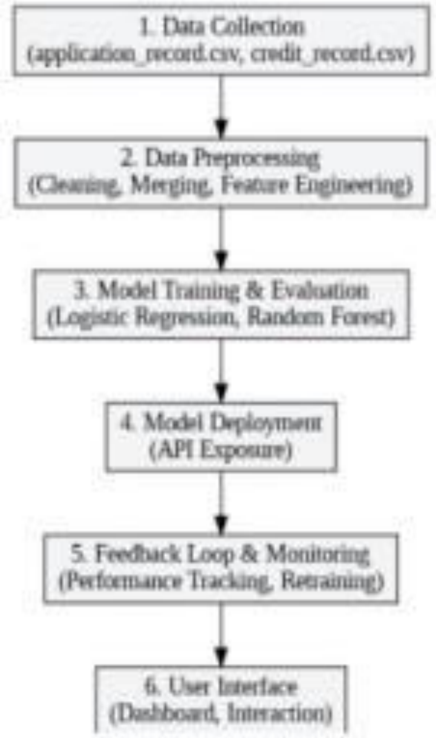

1. Data Collection
(application_record.csv, credit_record.csv)
2. Data Preprocessing
(Cleaning, Merging, Feature Engineering)
3. Model Training & Evaluation
(Logistic Regression, Random Forest)
4. Model Deployment
(API Exposure)
5. Feedback Loop & Monitoring
(Performance Tracking, Retraining)
6. User Interface
(Dashboard, Interaction)

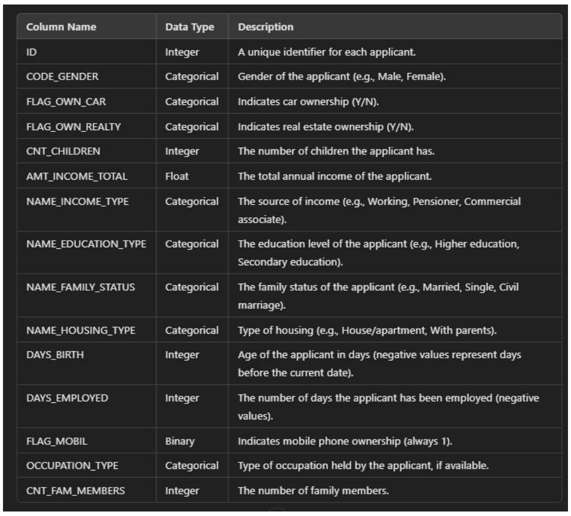
Figure 1: Overview of the Methodology for Credit Risk Analysis

3.1 Dataset Description

The analysis in this study is based on two key datasets: the *Application Record Dataset* and the *Credit Record Dataset*, which are outlined below.

3.2 Application Record Dataset

The *Application Record Dataset* (figure 2) serves as the foundational data source for understanding the demographic and socio-economic characteristics of credit card applicants. It contains 438,557 rows and 18 columns. The dataset includes critical features that help characterize each applicant’s financial profile, which is essential for predicting their creditworthiness.



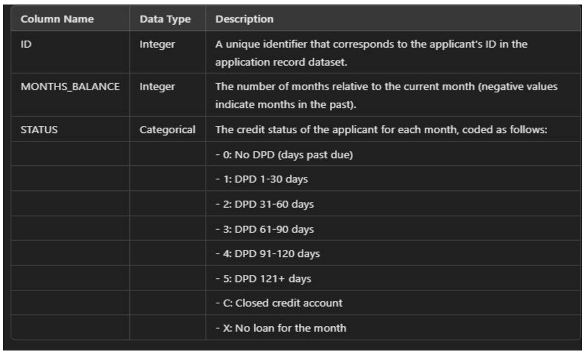
Column Name	Data Type	Description
ID	Integer	A unique identifier for each applicant.
CODE_GENDER	Categorical	Gender of the applicant (e.g., Male, Female).
FLAG_OWN_CAR	Categorical	Indicates car ownership (Y/N).
FLAG_OWN_REALTY	Categorical	Indicates real estate ownership (Y/N).
CNT_CHILDREN	Integer	The number of children the applicant has.
AMT_INCOME_TOTAL	Float	The total annual income of the applicant.
NAME_INCOME_TYPE	Categorical	The source of income (e.g., Working, Pensioner, Commercial associate).
NAME_EDUCATION_TYPE	Categorical	The education level of the applicant (e.g., Higher education, Secondary education).
NAME_FAMILY_STATUS	Categorical	The family status of the applicant (e.g., Married, Single, Civil marriage).
NAME_HOUSING_TYPE	Categorical	Type of housing (e.g., House/apartment, With parents).
DAYS_BIRTH	Integer	Age of the applicant in days (negative values represent days before the current date).
DAYS_EMPLOYED	Integer	The number of days the applicant has been employed (negative values).
FLAG_MOBIL	Binary	Indicates mobile phone ownership (always 1).
OCCUPATION_TYPE	Categorical	Type of occupation held by the applicant, if available.
CNT_FAM_MEMBERS	Integer	The number of family members.

Figure 2: Columns of the Application Record Dataset

3.3 Credit Record Dataset

The *Credit Record Dataset* (figure 3) provides a time-series view of each individual's credit behavior, capturing their payment status over several months. It contains 1,048,575 rows and 3 columns. This dataset is crucial for identifying patterns in repayment behavior and detecting signs of financial distress.

The combination of the *Application Record Dataset* and the *Credit Record Dataset* forms a robust foundation for conducting credit risk analysis. By utilizing a blend of socio-economic data and historical credit behavior, researchers can develop models that accurately predict the likelihood of default. This can aid financial institutions in making informed lending decisions, ultimately reducing risk and improving financial outcomes.



Column Name	Data Type	Description
ID	Integer	A unique identifier that corresponds to the applicant's ID in the application record dataset.
MONTHS_BALANCE	Integer	The number of months relative to the current month (negative values indicate months in the past).
STATUS	Categorical	The credit status of the applicant for each month, coded as follows:
		- 0: No DPD (days past due)
		- 1: DPD 1-30 days
		- 2: DPD 31-60 days
		- 3: DPD 61-90 days
		- 4: DPD 91-120 days
		- 5: DPD 121+ days
		- C: Closed credit account
		- X: No loan for the month

Figure 3: Columns of the Credit Record Dataset

3.4 Data Pre-processing

In any machine learning project, data pre-processing ensures model accuracy, reli- ability. For the credit risk prediction system, the dataset from Kaggle undergoes a rigorous pre-processing stage to handle inconsistencies and prepare the data for model training. This process involves several key steps, including:

- Handling Missing Values: Missing values were handled by imputing numeri- cal columns with the median and categorical columns with the mode to preserve data integrity and minimize bias.
- Encoding Categorical Variables: Categorical variables were transformed using one-hot encoding for non-ordinal features and label encoding for the target variable to make them compatible with machine learning algorithms.
- Outlier Detection: Outliers were identified using box plots and IQR meth- ods. Extreme values were either capped or removed to avoid distorting the model's predictions. .
- These pre-processing steps are essential for ensuring that the models receive clean, well-structured data, thereby improving their ability to accurately predict credit risk.

3.5 Scaling

To ensure that all features contribute equally to the model, Standardization was applied to the dataset. This involved scaling the features so that they have a mean of 0 and a standard deviation of 1. This is important because many machine learning algorithms, such as Logistic Regression and Random Forest, perform better when the features are on a similar scale. Standardization helps prevent features with larger values from dominating the model's learning process, ensuring balanced importance across all features.

3.6 Dealing with Class Imbalance

To address this imbalance, we employed several techniques:

- Resampling Techniques: We implemented oversampling of the minority class (high-risk applicants) using the SMOTE. To create a more balanced dataset.
- Using Class Weights: For models like Random Forest we adjusted the class weights inversely proportional to class frequencies

3.7 Correlation Analysis

Correlation analysis was performed to identify relationships between numerical features in the dataset. A correlation matrix was generated(figure 4) , revealing the strength and direction of associations between variables. To facilitate interpretation, the correlation matrix was visualized using a heatmap, which highlighted strong pos- itive and negative correlations. This analysis helped in detecting multicollinearity, guiding feature selection, and ensuring that only the most relevant features were used for model training. Highly correlated features were either removed or combined to enhance model performance and avoid overfitting.

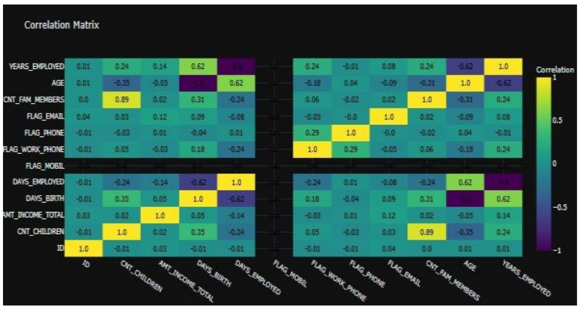


Figure 4: Correlation Matrix Heatmap

3.8 Feature Engineering

For this project, we implemented several techniques to create new features that better represent the underlying patterns in the data:

- Creating Interaction Features: We generated interaction terms between key features to capture non-linear relationships. For instance, combining credit history length with the applicant's age provided insights into the applicant's financial behavior over time.
- Binning Continuous Variables: Continuous features, such as income and age, were transformed into categorical variables through binning. This ap- proach allows the models to capture non-linear relationships more effectively, particularly for algorithms like Decision Trees and Random Forests, which can benefit from categorical inputs.
- Encoding Categorical Variables: We applied one-hot encoding to convert categorical features into numerical format, making them compatible with al- gorithms like K-Nearest Neighbors (KNN) and Neural Networks (MLP). This transformation is essential for effectively utilizing categorical data in our models.
- Scaling Engineered Features: After creating new features, we ensured that they were appropriately scaled, particularly for models sensitive to feature mag- nitudes. This step helps in achieving better model performance and conver- gence.
- Training and Testing: The training and testing phase is critical in evaluating the performance of our models. This project, adopted a systematic approach to ensure robust model development.
- Data Splitting: We divided the dataset for training, testing sets, typically following an 80-20 split. The training set was utilized to fit the models, while the models' generalization capabilities on unseen data.
- Model Training: Each selected models was trained on the training set. During this phase, hyperparameter tuning was conducted using technique such as Grid Search to optimize the model performance.
- Model Evaluation:After training, the models were evaluated on the testing set using performance metrics like accuracy, precision, recall, and F1-score. This step is crucial in understanding how well each model predicts loan defaults and identifying any overfitting or underfitting issues.
- Cross-Validation: To further validate model performance, we implemented k-fold cross-validation. This technique allows us to assess the stability of the model performance across different subsets of the data, providing a more reliable estimate of how the model is expected to perform in real-world scenarios.

3.9 Evaluation

- Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

- Recall

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

• F1-Score

$$F1 = \frac{2 \cdot \text{Precision:} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot TP}{2TP + FP + FN} \quad (4)$$

• Specificity:

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

3.10 Random Forest Model Performance

Random Forest is an ensemble learning method that aggregates the predictions of multiple decision trees to produce more robust and accurate results. It is widely used for classification and regression tasks due to its ability to reduce overfitting and handle high-dimensional data effectively.

The formula for the Random Forest prediction can be described as:

$$f_{\text{RF}}(x) = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \quad (6)$$

Where:

- $f_{\text{RF}}(x)$ is the final prediction for an input x,
- T is the number of decision trees in the forest,
- $f_t(x)$ is the prediction of the t-th tree in the forest.
- Handles missing data well by averaging over multiple trees.
- Reduces overfitting by using random subsets of features for each tree.
- Can model complex relationships in data through the ensemble of trees.

3.11 Results and Discussion

The Logistic Regression model showed the following performance metrics:

- Accuracy: 70%
- Precision: 70% for both classes (0 and 1)
- Recall: 71% for class 0 and 69% for class 1
- F1-score: 70% for both classes

The Logistic Regression model demonstrated balanced precision and recall across both classes (approved and rejected applicants), with an overall accuracy of 70%. The model is relatively simple, and while it performs decently, it struggles to achieve high recall, especially for the minority class (1: approved applicants).

The Random Forest Classifier model outperformed Logistic Regression with the following metrics:

- Accuracy: 98%
- Precision: 97% for class 0 and 100% for class 1
- Recall: 100% for class 0 and 97% for class 1
- F1-score: 98% for both classes

The Random Forest Classifier achieved an exceptional accuracy of 98%, signif- icantly outperforming Logistic Regression. It also had higher precision and recall, especially for the minority class (1: approved applicants). The model achieved a perfect recall for class 0 (rejected applicants), indicating that it correctly identified almost all rejections. Additionally, its F1-score of 98% shows that it strikes a good balance between precision and recall, making it highly suitable for credit approval predictions.

3.12 Discussion

Top Performing Models:

The comparison of model performance reveals that the Random Forest Classifier significantly outperforms the Logistic Regression model in predicting credit card ap- provals. With an accuracy of 98%, the Random Forest model demonstrated a strong ability to correctly classify both approved and rejected applicants. Its precision and recall metrics for both classes were notably higher, especially for the minority class (approved applicants), where it achieved a perfect recall of 100% for rejected applicants. In contrast, the Logistic Regression model, while providing balanced precision and recall across both classes, lagged behind with an accuracy of only 70%. This disparity highlights the Random Forest's strength in capturing complex patterns and its robustness in handling class imbalance. Overall, the results suggest that Random Forest Classifier is a more reliable and effective model for credit approval prediction, particularly in scenarios where high accuracy and balanced performance across classes are crucial.

4. CONFUSION MATRIX SUMMARY

The confusion matrices for both (figures 5,6) models reveal notable differences in their classification performance. The Logistic Regression model had a relatively balanced distribution between true positives (71%) and true negatives (70%) but struggled with classifying the minority class, as evidenced by a recall of 69% for approved applicants. On the other hand, the Random Forest Classifier achieved near-perfect classification, with a recall of 100% for rejected applicants and 97% for approved applicants, indicating its excellent ability to identify both classes correctly. The Random Forest model’s superior precision and recall across both classes, especially for the minority class, highlight its ability to manage class imbalance more effectively than Logistic Regression.

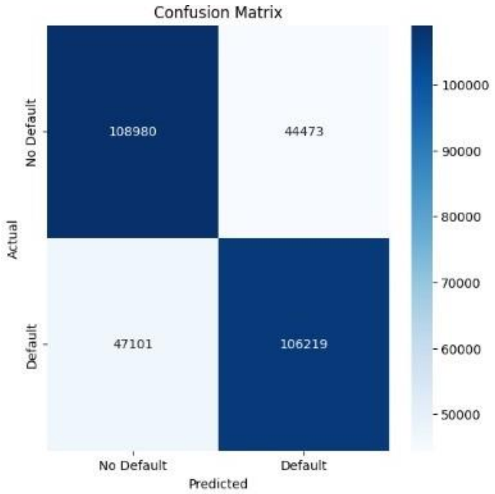


Figure 5: Confusion Matrix for Logistic Regression Model

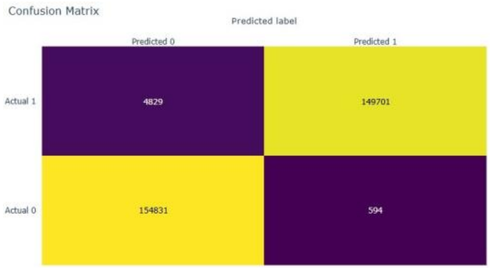


Figure 6: Confusion Matrix for Random Forest Classifier Model

5. CONCLUSION

In this study, we applied machine learning models—Logistic Regression and Ran- dom Forest Classifier—to predict credit card approval outcomes based on applicants’ demographic and financial data. The Random Forest Classifier significantly outper- formed Logistic Regression in terms of accuracy, precision, recall, and F1-score. With an accuracy of 98%, it demonstrated superior classification performance, particularly in handling class imbalance, where it exhibited near-perfect recall for both classes. In contrast, the Logistic Regression model, while decent, achieved a lower accuracy of 70% and struggled with detecting approved applicants. Overall, the Random For- est Classifier proved to be a more robust and reliable model for predicting credit card approval, making it a valuable tool for automated decision-making in credit risk assessment.

REFERENCES

[1] Madaan, M., Kumar, A., Keshri, C., et al. (2021). Loan Default Prediction Us- ing Decision Trees and Random Forest: A Comparative Study. IOP Conference Series: Materials Science and Engineering, 1022, 012042.

[2] Perera, H.A.P.L., & Premaratne, S.C. (2016). An artificial neural network ap- proach for the predictive accuracy of payments of leasing customers in Sri Lanka.

[3] Marqués, A.I., Garc´ıa, V., & Sánchez, J.S. (2012). Exploring the behaviour of base classifiers in credit scoring ensembles. Expert Systems with Applications, 39(11), 10244–10250.

[4] Adewusi, A.O., Oyedokun, T.B., & Bello, M.O. (2016). Application of artificial neural network to loan recovery prediction. International Journal of Housing Markets and Analysis, 9(2), 222–238.

[5] Choudhary, G., Garud, Y., Shetty, A., Kadakia, R., & Borase, S. (2019). Loan default identification and its effect.

[6] Atiya, A.F. (2001). Bankruptcy prediction for credit risk using neural net- works: A survey and new results. IEEE Transactions on Neural Networks, 12(4), 929–935.

[7] Li, H., McCarthy, J., & Pantalone, C. (2014). High-yield versus investment-grade bonds: Less risk and greater returns? Applied Financial Economics, 24(20), 1303–1312.

[8] Pandit, A. (2016). Data mining on loan approved dataset for predicting defaulters (Doctoral dissertation, Rochester Institute of Technology).

[9] Calcagnini, G., Cole, R., Giombini, G., & Grandicelli, G. (2018). Hierarchy of bank loan approval and loan performance. Economia Politica, 35(3), 935–954.

[10] Sarma, K.S. (2013). Predictive modeling with SAS Enterprise Miner: Practical solutions for business applications. SAS Institute.

[11] 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.

[12] Schneider, A. (2018). Studies on the impact of accounting information and as- surance on commercial lending judgments. Journal of Accounting Literature, 41, 63–74.

[13] Arun, K., Ishan, G., & Sanmeet, K. (2016). Loan approval prediction based on machine learning approach. National Conference on Recent Trends in Computer Science and Information Technology (NCRTCSIT-2016), 18–21.

[14] Nehrebecka, N. (2018). Predicting the default risk of companies: Comparison of credit scoring models: LOGIT versus support vector machines. Econometrics, 22(2), 54–73.

[15] Vimala, S., & Sharmili, K.C. (2018). Prediction of loan risk using Naive Bayes and support vector machine. International Conference on Advanced Computing Technology (ICACT), 4, 110–113.

[16] Pérez-Mart´ın, A., & Vaca, M. (2017). Computational experiment to compare techniques in large datasets to measure credit banking risk in home equity loans. International Journal of Computational Methods and Experimental Mea- surements, 5(5), 771–779.

[17] Pérez-Mart´ın, A., Pérez-Torregrosa, A., & Vaca, M. (2018). Big data techniques to measure credit banking risk in home equity loans. Journal of Business Re- search, 89, 448–454.

[18] Nalic, J., & Švraka, A. (2018). Using data mining approaches to build credit scor- ing model: Case study—Implementation of credit scoring model in microfinance institution. 2018 17th International Symposium Infoteh-Jahorina (INFOTEH), IEEE, 1–5.

[19] Baesens, B., Roesch, D., & Scheule, H. (2016). Credit risk analytics: Measurement techniques, applications, and examples in SAS. John Wiley & Sons.

[20] Han, J.T., Choi, J.S., Kim, M.J., & Jeong, J. (2018). Developing a risk group predictive model for Korean students falling into bad debt. Asian Economic Journal, 32(1), 3–14.