A Feature Elimination Machine Learning Model for Credit Assessment and Repayment Behavior Prediction in Marketplace Lending

GEORGIOS RIGOPOULOS
Department of Economics,
National and Kapodistrian University of Athens,
GREECE

Abstract: - With the rapid development of the credit industry and the advent of marketplace lending, credit scoring models play a vital role in reducing the risk exposure for lenders. However, traditional credit scoring models like the FICO Score make it hard for people with weak credit history to acquire credit services. Credit scoring models based on machine learning can provide accurate assessments for such thin-credit people, but a lot of private data, like social media activities, are used during the evaluation procedure. In this work, a credit scoring approach with a focus on marketplace lending is proposed that combines machine learning with a novel feature selection method that follows a backward elimination approach. Thus, many irrelevant features are eliminated from the dataset during the feature selection, and private data are not used or remain limited. The model is trained and tested in a large loan dataset available in the public domain. It performs pretty well compared to traditional credit scoring method and can be used to provide credit assessment for thin-credit history individuals without using personal private data. The approach has also explanatory power, as the feature selection approach offers a perspective for understanding how each feature affects individual loan repayment behavior.

Key-Words: - Credit scoring, marketplace lending, random forest, machine learning, feature elimination, random forests, backward elimination.

Received: May 22, 2024. Revised: October 5, 2024. Accepted: November 6, 2024. Published: November 28, 2024.

1 Introduction

The rapid development of the credit industry in recent years has resulted in outstanding revolving customer credit, and the development of marketplace lending, however, lending money is still a risky business. For example, around 10% of the loans lent out by the Lending Club (LC), the largest marketplace lending platform in the US (based on volume issued per year) failed to be repaid on time, according to data released by LC, [1], [2]. In order to reduce the repayment failure ratio, various types of credit scoring techniques are widely adopted in the loan application procedure, to classify applicants depending on the applicant's information like asset ownership, credit history, and other relevant factors. A credible customer is more likely to repay its debt compared to a non-one, who is considered as highly risky. The FICO Score was first introduced in 1989 and became the widely adopted standard in US financial institutions for credit scoring. Although the FICO score made significant contributions to credit assessment in the past period, it has some weaknesses, as it depends on customer credit history or historical behavior. Therefore, customers with a thin credit history can hardly gain access to loan services, [3].

However, big data and machine learning techniques are changing the business since they make it possible to use more abundant data in credit risk predictions. These new technologies can base credit score predictions on a much broader range of features [3]. Hence, to make up for the lack of underserved customers, a lot of fintech companies have developed their unique algorithms for individual credit assessment using big data and machine learning technology. This allows fintech companies to gain information from a much broader range of sources, even with no apparent link to creditworthiness, for example, digital fingerprints and social network activities, [4]. Combining big data and machine learning models into credit assessment demonstrates superiority compared to traditional scoring models in many cases, [5]. Using big data and machine learning technology to process and analyze diverse sources of data, including social media, has become the core business for some fintech companies, however, there is questioning for its usefulness and its legitimacy due to privacy and confidentiality. Nowadays, with the rise of

awareness for individual privacy data protection, people are paying more attention to how their data is processed and a lot of countries and unions have published different data protection laws.

Many researchers have used AI technology for credit scoring, including nearest neighbors (KNN) [6], support vector machines (SVM) [7], and Random Forests (RF) [8]. Results show that the AI approach is superior in handling credit scoring problems compared to traditional statistical techniques. However, although the research results in well-performed models, almost all customer information is used without considering whether the selected features are valid or considering the differences in importance between features, [9]. Hence, to better understand individual repayment behavior and increase model performance, the feature selection process is necessary to be included during the data preprocessing procedure. By removing the irrelevant features, the model would not only be improved in accuracy but would also require less execution time.

Following the above, the main aim of this research is to build a machine learning credit assessment model that can predict whether an individual can repay the loan, focusing on marketplace lending, without processing sensitive private data like social media activities or contacts. The model focuses on identifying the features that have significant effects on individual repayment behavior. A novel feature selection method is proposed based on the idea of ablation study on the random forests model. Selecting the appropriate set of features is critical not only to increase the prediction accuracy but also to help understand the logic behind the individual loan repayment behavior. The main contribution of this research is that it constructs a machine learning model without using sensitive private data like social media activities and is not necessarily dependent on customer credit history. This paper focuses on the feature selection process that is executed in the preprocessing and during the training phase as well.

The rest of the paper is organized as follows: Section 2 reviews the literature on AI credit scoring models, black-box problem and model transparency, feature selection method, and the use of random forests models in credit assessment. In section 3 the methodology is outlined and in section 4 the model and feature selection procedure are presented in detail. Finally, a discussion and limitations along with a conclusion are presented.

2 Background

2.1 FinTechs and AI Models

To improve financial inclusion and provide loan services for credit-invisible individuals many fintech companies developed their own algorithms using AI and alternative data, [4]. These AI models make a substantial contribution to expanding access to financial services for underrepresented groups. Three issues, however, may raise concerns, even though they have demonstrated advantages over the traditional FICO Score.

The first problem is model transparency, also known as the black-box problem. The black-box problem of AI algorithms means that humans cannot understand the reason and logic during the AI algorithm decision-making procedure. The input values are often high-dimensional and non-linearly coupled, which makes it hard to track and understand how those inputs are transformed into the result, [10]. This opacity makes customers hard to trust whether the result is reasonable, or the data used is within the legal requirements. At the same time, the opacity makes it hard for regulatory authorities to determine whether an algorithm produces the result unbiased or violates local laws, [11]. One way of solving the problems is to develop explainable models by making the decision-making procedure understandable and building trust between the customer and the algorithms developer. This can also reduce algorithmic biases that are prejudicial to specific groups, [11].

The bias issue is the second problem. Even though every fintech company claims to have algorithms that analyze consumer creditworthiness fairly and accurately, AI algorithms don't have the ability to determine whether a decision is moral because they base their judgments only on data, [12]. So, preexisting stereotypes or other social issues could lead AI algorithms to make discriminatory decisions. When AI uses social media data, this issue might get worse and raise the possibility that the AI algorithm may choose racial, ethnic, or gender-discriminatory characteristics, [4].

The third problem is accountability, or the ability to determine whether a decision was made within legal and ethical standards and hold someone responsible if those standards are not met. Making mistakes is not unacceptable; however, there is no clear idea of who should take responsibility when the AI algorithm makes a mistake or how to create accountability for AI. Research suggests that the lack of clear accountability for AI in financial institutions might cause a shift from bank liability for human mistakes to bank liability for the problem

caused by reliance on AI algorithms, [13]. Researchers from Google proposed that external audits that are designed to identify these risks from outside the system can also serve as accountability measures for these deployed models and organizations that build and deploy AI algorithms should also internalize these risks, hold to account ethical principles, and try to identify problems before causing consequences, [14].

2.2 Machine Learning Models

Explainable models like linear regression and decision trees are frequently used in individual creditworthiness assessment in parallel to "black-box" AI algorithms, [15]. Logistic regression leads to comparably accurate predictions to neural network models, but neural networks cannot offer explanations for the predicted results, and it is difficult to explain the underlying reason for rejected customers, [15]. The decision tree model for credit risk assessment has also been used successfully reducing the rate of non-performing loans by 6%, [16]. The model used 11 factors, including gender, age, monthly income, monthly expenses, savings, collateral values, previous credit status/rating, etcetera. And model suggests that the factor "collateral values" has the highest correlation with the predicted result. It is also suggested in relevant studies that using the decision tree algorithm is a feasible solution for credit assessment, [17]. This model reduces the data complexity by proper variable selection and pre-pruning to generate understandable and interpretable results in fast training. Although the decision tree is a well-adopted algorithm used in creditworthiness assessment procedures, research suggests that the decision tree model does not perform significantly better than other models, [18]. Two main limitations that affect model performance are the available data and sample selection bias. Credit scoring models built using historical data of past applicants can potentially lead to a biased sample when used to evaluate present applicants. However, in practice, both the financial institution and the customers tend to accept the model that is easy and clear to understand.

Despite that the decision tree algorithm performs relatively well in the training-testing experiment; it is prone to bias and overfitting problems. To solve this problem, random forests algorithm was used in 2001, [19], which can be considered an extension of the decision tree algorithm. It has three significant advantages compared with the decision trees model. It reduces the risk of overfitting, is more flexible as it can be used for both regression and classification, and is easy to evaluate the factor contribution for the model. Another work compared the classification results of the random forests model and the decision trees models, [20]. The researchers used 20 data sets from the UCI repository, and the number of data points in those data sets varies from 148 to 20000 and used 10-fold cross-validation to avoid overfitting. Results show that for the same number of attributes, the random forests model has a better performance on large data sets and the decision tree model performs better on small data sets.

2.3 Feature Selection

Feature selection is considered an essential procedure to reduce the model complexity and prevent model overfitting. Laborda and Ryoo test three feature selection methods (Chi-squared test and correlation coefficients, forward stepwise elimination, and backward stepwise elimination) on different credit scoring models (logistic regression, support vector machine, K-nearest neighbors and random forest), [21]. The experiment compared the model simplicity and model accuracy after each feature selection procedure. Results show that the forward and backward stepwise elimination significantly contributed to simplifying models than the Chi-squared test and correlation coefficients, and the random forests model showed the greatest improvement in the model accuracy after the feature selection procedure. Among the three feature selection methods, the forward stepwise elimination performed relatively better than the other two methods at both model simplicity and model accuracy.

The mean decrease in impurity is used to measure how the model accuracy decreased when excluding a particular factor, and the average reduction of accuracy calculates the mean decrease accuracy by randomly permutating the factors in the observed sample. Both the mean decrease in impurity and the mean decrease in accuracy are considered the critical evaluation for feature importance. Zhang, Yang, and Zhou designed a model that combined the random forests algorithm and feature selection for credit scoring using data from the UCI database of Machine Learning Database, [9]. The experiment results show that compared to the single random forests algorithm, combining the feature selection method increases the model accuracy by approximately 4%.

Unlike the current mainstream research, this research focuses on the feature selection procedure instead of model accuracy since this research aims not only to develop a model for individual credit assessment but also to use this model to understand

the logic behind loan repayment behavior. Although the random forests model is a non-explainable model, this work follows an ablation study approach to help better understand how each feature affects the prediction result.

3 Method and Data

This work proposes a credit scoring model based on a random forests algorithm, optimized by cross-validation and a novel feature selection procedure called Backward Ablation Selection. Although the non-explainable AI models increase the difficulty of regulation and undermine the mutual trust between financial institutions and customers, in this work an ablation study was performed to enhance the understanding of the random forests model used in this research. The Backward Ablation Selection starts with the model that includes every feature, removing the least important element to examine the influence of that feature. This step is repeated until the removed feature significantly reduces the model's accuracy. The aim of developing such a model is to provide creditworthiness assessment services for "thin-credit" customers without using sensitive and private data while providing data-driven decision-making support for financial institutions when facing these customers. Initially, preliminary data cleaning is performed to remove meaningless variables and replace null values. Then a combined feature selection method to filter the irrelevant variables and reduce the data dimension and future model training time follows.

The dataset used for the model is the Lending Club Loan Dataset which contains approximately two million observations and one hundred and forty-five features from all loans issued from 2007 to 2015 in the United States. A preprocessing step was performed initially with data cleaning by removing variables that have over 50% of the missing value rate since using those variables is highly likely to cause prediction errors. For the remaining variables, four statistical approaches were used to deal with missing values. Listwise deletion for null or missing values deletion, pairwise deletion to exclude variables with missing values, but still use them when analyzing other variables with non-missing values, multiple imputation to replace missing values with predicted values, and maximum likelihood to estimate the missing data based on existing data.

Then a preliminary variable selection was performed to exclude features generated after the loan decision because those variables cannot affect the previous decision-making procedure. There are three conventional general methods of feature selection: filters approach, wrappers approach, and embedded approach, [22].

Considering the large amount of loan credit data used in the research, a novel feature selection method was applied, which combines the chi-square test and variance test with the backward selection to select the best feature data subset. The benefits of using this combined method are that highly irrelevant features will be removed by the chi-square test, and variance hence can reduce the model training time. The backward selection procedure can help understand how each feature affects the prediction result.

Then a random forests model is built. After the model is built, the work continues to the backward selection for feature selection. The backward selection method starts with the model with all available features, then removes the least essential feature and compares the decreased accuracy of the new model with the previous model. If the removed feature does not significantly influence the model accuracy, the new model is treated as the complete model and the last procedure is repeated until removing the least essential feature greatly decreases the model accuracy. Besides the stepwise selection, a five-fold cross-validation was followed during the model training procedure.

4 Data Preprocessing and Feature Selection

The original data set, which is publicly available, contained over 2.2 million observations and 145 features. Among the 144 explanatory features, there are 46 features that have over 50% of the missing cases, including loan case ID and customer ID. Since the number of missing values was too large to be substituted, the 46 features were removed from the original data set. Sensitive data like loan case ID and the customer ID were deleted from the data set. For the remaining features, observations containing missing values were removed instead of replaced to reduce bias. The features "employment title", "zip_code", and "address state" were also removed from the original data set since rejecting loans to some particular occupations or regions is highly unethical and discriminatory. After removing all missing values, the data set size was reduced to 96 features with approximately 1.1 million observations. After the preliminary data cleaning procedure, meaningless features that were generated after a loan decision had been made were dropped. A total of 18 features were identified as meaningless

during this procedure. The dropped features and their detailed explanation are listed below:

- funded_amnt: The total amount committed to that loan at that point in time.
- funded_amnt_inv: The total amount committed by investors for that loan at that point in time.
- Grade: Loan Club assigned loan grade.
- sub_grade: Loan Club assigned loan grade.
- issue_d: The month in which the loan was funded.
- earliest_cr_line: The date the borrower's earliest reported credit line was opened.
- out_prncp_inv: Remaining outstanding principal for a portion of the total amount funded by investors.
- total_pymnt: Payments received to date for the total amount funded.
- out_prncp: Remaining outstanding principal for total amount funded.
- total_pymnt_inv: Payments received to date for a portion of the total amount funded by investors.
- total_rec_prncp: Principal received to date.
- total_rec_int Interest received to date.
- total_rec_late_fee: Late fees received to date.
- recoveries post charge off gross recovery.
- collection_recovery_fee: post charge off collection fee.
- last_pymnt_d: Last month payment was received.
- last_pymnt_amnt: Last total payment amount received.
- last_credit_pull_d: The most recent month Loan Club pulled credit for this loan.

Next, the variance selection method was followed to examine the necessity for all features. If a feature has relatively low variance, it suggests that all values in that feature are very close to each other; hence this feature can be approximately considered as a constant number. The threshold for variance was set to be 0.01 since if the variance is lower than 0.01 suggests that the feature is not separate enough to support model training. Since the variance test only tests for feature value distribution, there is no need to consider the interactive effect for features with low variance as these features act approximately as constant under any situation. The removed features and their variance can be found in Table 1.

Table 1. Removed features and their variance

Feature Name	Variance
pymnt_plan	4.57070140e-04
policy_code	0
acc_now_delinq	4.13456292e-03
hardship_flag	5.95409284e-04
num_tl_30dpd	2.60408793e-03

Then the chi-square test was performed to examine the correlation between the predicted variable and explanatory variables. The chi-square test was performed on the "df_category" data set and applied label encoding to transform all non-numeric feature values into number values in order to calculate the p-values. The p-value threshold was set to 0.05 as a commonly accepted significance level in statistical research. The categorical features and their corresponding p-values are listed in Table 2. Only one feature, "purpose", is identified as having no significant effect on the predicted variable due to their p-values failing to pass the 0.05 threshold.

Table 2. Categorical features and their p-value

Feature name	P-value
Term	5.14624698e-087
emp_length	3.44970539e-101
home_ownership	5.92059625e-076
verification_status	0.00000000e+000
Purpose	1.09467496e-001
initial_list_status	4.35410063e-075
application_type	9.76401930e-018
disbursement_method	2.99584437e-046

After feature filtering, there were 72 remaining variables that were considered to have significant effects on the customer's loan status, and they were used for the random forests model for classification after label encoding to all categorical features. 80% of observations were used as the training data set and the rest 20% for the testing data set. During the model training procedure, in order to increase the classifier accuracy, 5-fold cross-validation was selected. The training data was partitioned into ten equally sized slices. For each validation, one slice of data was used for testing, and the remaining nine slices were used for training the random forests model.

4.1 Feature Selection Based on Random Forests, Ablation Study, and Backward Selection

Next, a fusion feature selection method was applied based on the ideas of a random forest algorithm, backward elimination selection, and ablation study

(Backward Ablation Selection). The random forests model can rank all features' importance by calculating their mean decreased impurity. Each feature's importance is how many impurities it reduces on average. The ablation study is used in the machine learning area that tests the performance of a model by removing certain components to understand the component's contribution to the overall system. The Backward Ablation Selection method starts with the full random forests model and removes the least important feature. After the least important feature has been removed, it examines whether removing this feature has a significant impact on the model's accuracy.

Table 3. Useful explanatory variables and their description

Feature Name	Feature description
int_rate	Interest Rate on the loan
Installment	The monthly payment owed by the borrower if the loan originates.
Dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
bc_open_to_buy	Total open to buy on revolving bankcards.
avg_cur_bal	The average current balance of all accounts
tot_hi_cred_lim	Total high credit/credit limit
max_bal_bc	Maximum current balance owed on all revolving accounts
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
mo_sin_old_rev_tl_op	Months since the oldest revolving account opened
total_rev_hi_lim	Total revolving high credit/credit limit
mo_sin_old_il_acct	Months since the oldest bank installment account opened
total_bc_limit	Total bankcard high credit/credit limit
debt_settlement_flag	Flags whether or not the borrower, who has charged off, is working with a debt settlement company.

If the feature being deleted is irrelevant, repeating the previous procedure until deleting the least essential feature will dramatically decrease the model's accuracy. The method also fully takes into account the potential influence of interaction terms. If a feature has a huge interactive effect together with other features but not itself, deleting such a feature will immediately result in a substantial reduction in the model accuracy. The initial full model with 72 features had reached 91.26% accuracy. The model accuracy fluctuated when less critical features were removed. The model accuracy reached the peak when only 13 features were left; further cutting resulted in a sharp drop in model accuracy. The model accuracy rose to 91.27% using 16 explanatory variables. The 13 most important features and their meanings are listed in Table 3.

4.2 Model Parameter Fine-tuning

Next, there was an investigation of how parameters influenced the model performance. There were two parameters, n_estimators, and max_depth, which were considered to impact the model significantly. The parameter n_estimators represents the number of trees in the forest, and the parameter max_depth limits the maximum depth of the tree. In order to further improve the model performance, several experiments were performed, but on average with n_estimator ranging from 1 to 100 trees in increments of 5 trees, the model accuracy increases rapidly when the number of trees increases from 1 to 10 and keeps subtle growth until the number of trees reaches 81 (Figure 1). While, when the number of trees exceeds 81, the model accuracy starts to decrease. This is also in line with the characteristics of random forests: within a certain range, the more sub-trees, the better the model accuracy. The accuracy will fluctuate when the number of sub-trees becomes larger and larger.

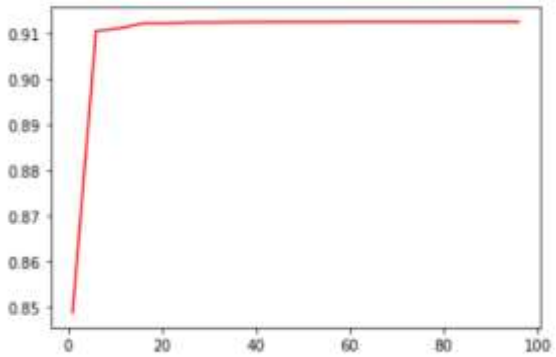

Fig. 1: The change of model accuracy when the n_estimator shifts from 1 to 100

Then a grid search was performed to identify the best fit for the parameter max depth of the random forests model. The optimal number for the maximum depth of the tree in our model was 57, and the final prediction accuracy was 91.28% after feature selection and parameter fine-tuning. The process was repeated with random sampling of the dataset observations, but the accuracy was converging to the values above.

Table 4 shows the changing trend of model accuracy and running time during the feature reduction procedure in the data preprocessing and feature selection phase.

Table 4. The model accuracy during the backward ablation selection

Number of features	Model accuracy
77 features	91.27%
67 features	91.26%
59 features	91.25%
44 features	91.27%
34 features	91.27%
29 features	91.25%
22 features	91.27%
16 features	91.27%
13 features	91.27%
10 features	90.13%

The model accuracy fluctuates within a small range before the number of features used in the model is reduced to 13. Further eliminating the least important feature leads to a significant reduction in the model accuracy. However, removing the top seven important features, int_rate, installment, dti, bc_open_to_buy, avg_cur_bal, tot_hi_cred_lim and max_bal_bc, does not significantly affect the random forests model accuracy. After removing the top seven critical features, the model accuracy increased from 91.27% to 91.28%. Although this tiny increase in the model accuracy does not necessarily suggest that deleting the seven features will improve the model performance, the randomness of the random forests algorithm might be the reason that caused this phenomenon. However, it does show that those seven features might not be as vital as we thought. The random forests model suggests that when trying to predict loan repayment behavior, any six features from the list below will provide sufficient support for decision-making:

- interest rate
- installment amount
- dti ratio
- Total open to buy on revolving bank cards
- The average current balance of all accounts
- Total high credit/credit limit
- Maximum current balance owed on all revolving accounts
- The revolving line utilization rate
- Months since the oldest revolving account opened
- Total revolving high credit/credit limit
- Months since the oldest bank installment account opened
- Total bankcard high credit/credit limit

Although the random forests model allows one to choose from a flexible range of features for credit scoring, one feature, the debt settlement flag, which is a flag that indicates whether or not the borrower has a charged-off history, cannot be ignored at any time. During the experiment, removing the debt settlement flag at any time immediately caused a significant reduction in model accuracy. Hence, whether or not the customer has a charged-off history is a necessary explanatory condition that must be considered.

5 Discussion and Limitations

The random forests model proposes a different calculation logic than the traditional FICO score. However, the random forests model performs relatively well even after removing the feature "Average current balance of all accounts" and the feature "Maximum current balance owed on all revolving accounts". The average model accuracy after removing those two features remains approximately 91.26%. The factors "length of credit history" and "types of credit that customers have" account up to 25% of the traditional FICO score calculation. However, both do not provide sufficient decision-making support for the random forests model to predict loan repayment behavior. Despite the differences, both the FICO Score and the present model agree on the importance of one aspect, which is the customer repayment history. The credit repayment history accounts for 35% of the FICO Score, which is the most significant component. During the feature selection procedure of the random forests model, it was suggested that repayment history significantly affects customer repayment behavior.

The random forests model showed a different approach for assessing individual loan repayment risk, as lack of credit history is no longer a severe defect as long as the customer has not failed to repay the loan in the past. For example, if a loan applicant is applying for credit for the first time, he/she will undoubtedly have no credit history.

However, by showing himself/herself is in good financial condition (low dti rate and certain level of current balance) and set affordable repayment conditions (proper interest rate and installment amount) the risk of this applicant failing to repay the credit can be determined. Under the most constrained conditions, the model can make predictions about applicants' repayment behavior with 91.27% accuracy using only six features. Financial institutions are able to assess and reduce their risk from multiple dimensions, and individuals would be more accessible and have more opportunities to receive the loans they need.

5.1 Limitations

This research has two key limitations, namely transparency and potential biases. The random forests model has demonstrated outstanding adaptability and accuracy in the experiments, however, the model itself is an ensemble algorithm. So, the model randomly constructs a variety of decision trees and decides on the most optimal. Hence it is almost impossible to understand how the prediction result was derived. Although the ablation study was used to examine the influence of factors and resulted in 13 factors that are considered to have significant effects on predicting repayment behavior, it is not feasible to find how exactly a factor will affect the predicted result. Further research could focus on designing or applying an explainable model to enhance trust between financial institutions and their customers.

Another limitation has to do with the data set and training, as it only contains information for customers who have been granted a loan. In other words, individuals who have been rejected by the loan club are excluded from the dataset. In general, there is no guarantee that individuals who are excluded from the loan service will have the same expected repayment behavior as the population in the dataset. Moreover, the data set includes applicants from the United States, so the model might not apply to other countries or regions due to different characteristics and thresholds during the repayment behavior prediction. However, the model can be trained in various datasets and be adapted to regional and business-specific needs.

6 Conclusion

Lending is risky and as the need for borrowing increases, financial institutions will face more challenges. Credit scoring, as a vital risk management technique, can not only accurately measure the customer default risk but also improve risk management efficiency. This research aims to build a credit assessment model focusing on marketplace lending without processing private personal data. Furthermore, to find a different perspective to understand loan repayment and default behavior other than simply using their credit length or credit history.

This research is trying to identify what features are actually affecting the prediction result. A random forests model was used with 5-fold cross-validation to increase the model performance and prevent over-fitting and under-fitting. After model training, ablation test was used to determine each feature's effects on the prediction result. The features that passed the ablation test are considered to significantly affect the prediction results. A grid search and recursive training were applied to determine the best parameter for the random forests algorithm. The experiment is based on the Lending Club open data set containing 145 features and over 2.2 million loan records. After selection, 13 features proved to be non-negligible. Fewer features mean that financial institutions can focus on collecting relevant data, and the workload of credit evaluators has been reduced compared to the original 145 features. The model has prediction accuracy rate of 91.27% and the experiment suggests the random forests model proved a relatively accurate prediction in distinguishing applicants. The model also opens up more opportunities for people who have been denied a credit service before due to their lack of credit history. Financial institutions could also reduce their risk by setting more appropriate interest rates and installment amounts. Further research could be carried out to analyze more data. Future research can also focus on the analysis of different types of data, for instance, using natural language processing techniques to process customers' application writing to determine the potential helpful message.

References:

[1] Milne, Alistair K. L. and Parboteeah, Paul, The Business Models and Economics of Peer-to-Peer Lending (May 5, 2016). *ECRI Research Report*, 2016, No. 17, http://dx.doi.org/10.2139/ssrn.2763682.

[2] Anh, N. T. T., Hanh, P. T. M., & Le Thu, V. T. (2021). Default in the US peer-to-peer market with covid-19 pandemic update: An empirical analysis from lending club platform. *International Journal of Entrepreneurship*, 25(7), 1-19.

[3] Djeundje, V. B., Crook, J., Calabrese, R., & Hamid, M. (2021). Enhancing credit scoring with alternative data. *Expert Systems with Applications*, 163, 113766.

[4] Sadok, H., Sakka, F., & El Maknouzi, M. E. H. (2022). Artificial intelligence and bank credit analysis: A review. *Cogent Economics & Finance*, 10(1), 2023262.

[5] Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of fintechs: Credit scoring using digital footprints. *The Review of Financial Studies*, 33(7), 2845-2897.

[6] Marqués, A.I., García, V. and Sánchez, J.S. (2012) 'Exploring the behaviour of base classifiers in credit scoring ensembles', *Expert Systems with Applications*, 39(11), pp. 10244–10250.

[7] Tomczak, J.M. and Zie˛ba, M. (2015) 'Classification Restricted Boltzmann Machine for comprehensible credit scoring model', *Expert Systems with Applications*, 42(4), pp. 1789–1796.

[8] Van Sang, H., Nam, N. H., & Nhan, N. D. (2016). A novel credit scoring prediction model based on Feature Selection approach and parallel random forest. *Indian Journal of Science and Technology*, 9(20), 1-6.

[9] Zhang, X., Yang, Y. and Zhou, Z. (2018) 'A novel credit scoring model based on optimized random forest', *2018 IEEE 8th Annual Computing and Communication Workshop and Conference, CCWC 2018*, 2018-January, pp. 60–65.

[10] Zednik, C. (2021) 'Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence', *Philosophy and Technology*, 34(2), pp. 265–288.

[11] von Eschenbach, W.J. (2021) 'Transparency and the Black Box Problem: Why We Do Not Trust AI', *Philosophy & Technology*, 2021 34:4, 34(4), pp. 1607–1622.

[12] Penny Crosman (2016) Before AI Runs Amok, Banks Have Some Hard Decisions to Make, American Banker, [Online]. https://www.americanbanker.com/news/before-ai-runs-amok-banks-have-some-hard-decisions-to-make (Accessed Date: June 1, 2024).

[13] Costello, A.M., Down, A.K. and Mehta, M.N. (2020) 'Machine + man: A field experiment on the role of discretion in augmenting AI-based lending models', *Journal of Accounting and Economics*, 70(2–3), p. 101360.

[14] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *In Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44), https://doi.org/10.48550/arXiv.2001.00973.

[15] 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.

[16] Mandala, I.G.N.N., Nawangpalupi, C.B. and Praktikto, F.R. (2012) 'Assessing Credit Risk: An Application of Data Mining in a Rural Bank', *Procedia Economics and Finance*, 4, pp. 406–412.

[17] Chern, C. C., Lei, W. U., Huang, K. L., & Chen, S. Y. (2021). A decision tree classifier for credit assessment problems in big data environments. *Information Systems and e-Business Management*, 19, 363-386.

[18] Yap, B.W., Ong, S.H. and Husain, N.H.M. (2011) 'Using data mining to improve assessment of credit worthiness via credit scoring models', *Expert Systems with Applications*, 38(10), pp. 13274–13283.

[19] Breiman, L. (2001) 'Random Forests', *Machine Learning*, 2001 45:1, 45(1), pp. 5–32.

[20] Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5), 272.

[21] Laborda, J. and Ryoo, S. (2021) 'Feature selection in a credit scoring model', *Mathematics*, 9(7).

[22] Kumar, V. (2014) 'Feature Selection: A literature Review', *The Smart Computing Review*, 4(3).

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The author developed the present research in its entirety.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare.