Payoff Maximizing Orientated Loan Default Prediction Algorithms for Peer-to-Peer Lending Platforms: Evidence from Prosper

ECO353: Economics of Algorithms

Authors: Huiyang Chen, Yanqian Cao, Yimeng Gao
Instructor: Dr. Yao Luo

Abstract

As platforms that could potentially face higher-risk clients, peer-to-peer lending firms require exceptional precautions to prevent severe losses. Borrower credibility assessment and firm profit maximization are therefore recognized as aspects of significant attention. This research aims to investigate algorithms capable of accurate prediction of borrowers' loan default status and maximization of firm's expected payoff, with a comparative analysis between parametric and black-box approaches. Covering Prosper data with loan cases from 2005 to 2014, logistic regressions with and without regularization, as well as a random forest algorithm are adopted, each of which are trained and assessed separately prior and post to tuning targeting at expected payoff maximization. As compared to the random forest algorithm, logistic models in general yield less payoff, and the regularization term does not lead to significant changes regarding prediction accuracy and resulting firm payoff. Additionally, model tuning notably improves the profit brought by logistic regressions, yet at the same time weakens the outcome prediction accuracy. Finally, with remarkably significant enhancement in both prediction accuracy and resulting payoff, the tuned random forest algorithm performs the best among all model specifications. However, lack of interpretability along with extremely high cost are major problems for the black-box algorithm, in contrast, logistic regression models can reveal insightful findings that interest rate, number of delinquencies and credit inquiries are positively correlated with the probability of default.

1. Introduction

As a milestone of modern financial innovation, Peer-to-peer (P2P) lending directly connects borrowers with individual lenders through digital platforms. Unlike traditional

commercial banks, the P2P platform meets the needs of small lenders better because of the Internet's cross-regional, cross-temporal characteristics (Zhang and Sun, 2022). By enabling personalized lending and borrowing experiences tailored to specific financial needs, P2P grows to be an important channel for small loan customers to borrow money from. However, the inaccurate credit prediction of the borrower can potentially lead to loss to the lenders and damage to the reputation of the platform. Therefore, the accurate prediction of probable defaulters is crucial.

This research focuses on loan payment success prediction and payoff maximization for the second largest P2P company in America, Prosper, by constructing predictive models using logistic regression and random forest method on customer-specific data from this loan company. The result can help advance the risk assessment process and improve loan approval criteria based on predictive insights.

1.1 Literature Review

Logistic regression is a popular choice for classification problems, to predict and explain a binary categorical variable by using multiple independent variables, and it is broadly used in loan default prediction research (Sheikn et al, 2020; Maheswari and Narayana, 2020), with high accuracy rates in prediction results. The research conducted in 2021 to predict loan default by applying logistic regression and other three machine learning techniques used demographic predictors in their model, such as whether the borrower is a house owner, loan amount, and annual income (Zhu et al, 2023). The importance of considering credit grade as a predictor has been validated in another research paper by using logistic regression to predict the probability of default for borrowers in the P2P platform (Emekter et al. 2015). However, in one recent research

paper for loan default prediction, the random forest model shows the most astonishing performance in prediction accuracy among the three models, including logistic regression and Linear SVM (Victor and Raheem, 2021). The random forest model also shows the highest prediction accuracy rate in another P2P loan defaulting prediction research, among the four models(Zhu et al, 2019). Hence, both logistic and random forest models will be constructed in our essay, and a model comparison will be made between them.

Moreover, the above papers mainly focus on using classification algorithms to identify borrowers having high risks of default, with a lack of investigation on how to further adjust the model to maximize the firm's expected payoff. Therefore, we are going to design the classification algorithm with consideration of the expected payoff maximization problem, and make comparisons between the algorithms' performance to yield maximal payoff..

1.2 Data

The dataset contains the customer-specific loan statistics from Prosper, a peer-to-peer loan platform, with more than 110 thousand entries and 81 variables. Entries contain information regarding loan characteristics including loan amount, date, interest rate, as well as borrowers' risk assessment variables and demographic information, for instance, credit score, prosper rating, employment and residence region. For a direct comparison between predicted outcome and actual default status, only the observations indicating terminated loans are kept in the dataset for model fitting. In other words, only rows corresponding to a confirmed default status are reserved, with the target variable "loan status" showing either "completed" or "defaulted". After data cleaning, there are a total of 43092 rows left for investigation.

1.2.1 Exploratory Data Analysis

During the process, after discarding all non-informative variables, to find out the potential useful predictors, the data for the other variables were visualized in histograms, box plots, etc. It is noticeable that the geographical distribution shows that the loan default rate varies a lot across states in Figure 1, with particularly high rates in Lowa and North Dakota, which are above 12%. Therefore, dummy variables were added to states to incorporate state-fixed effects into our predictive model, capturing the regional economic conditions that potentially influence default rates. Since it is panel data with the earliest record showing a date in 2005, we also added a dummy to the years to incorporate the time-specific effect.

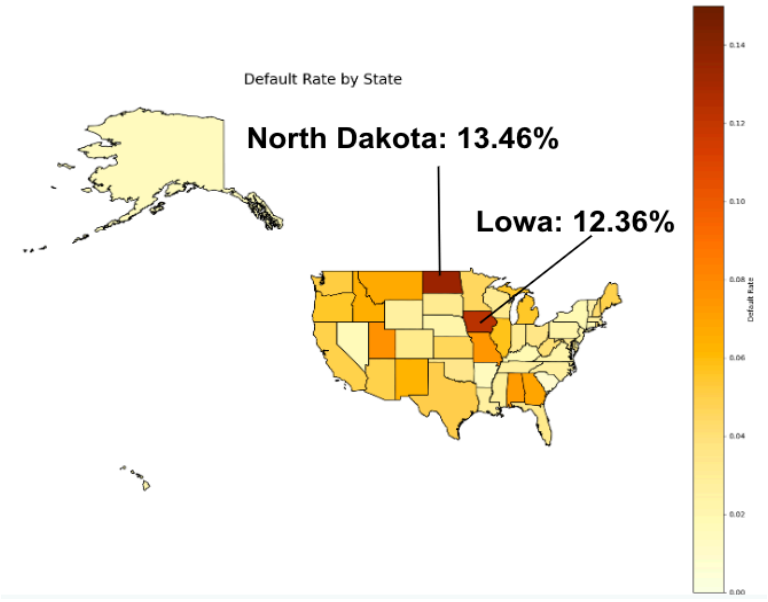

Figure 1: Loan Default Rate by State

When investigating the loan default rate by occupation, we found that the highest default rate falls in different types of students, as shown in Figure 2. Hence, to check if the default rate is influenced by occupation, this variable was added to our model with a dummy added to all student groups.

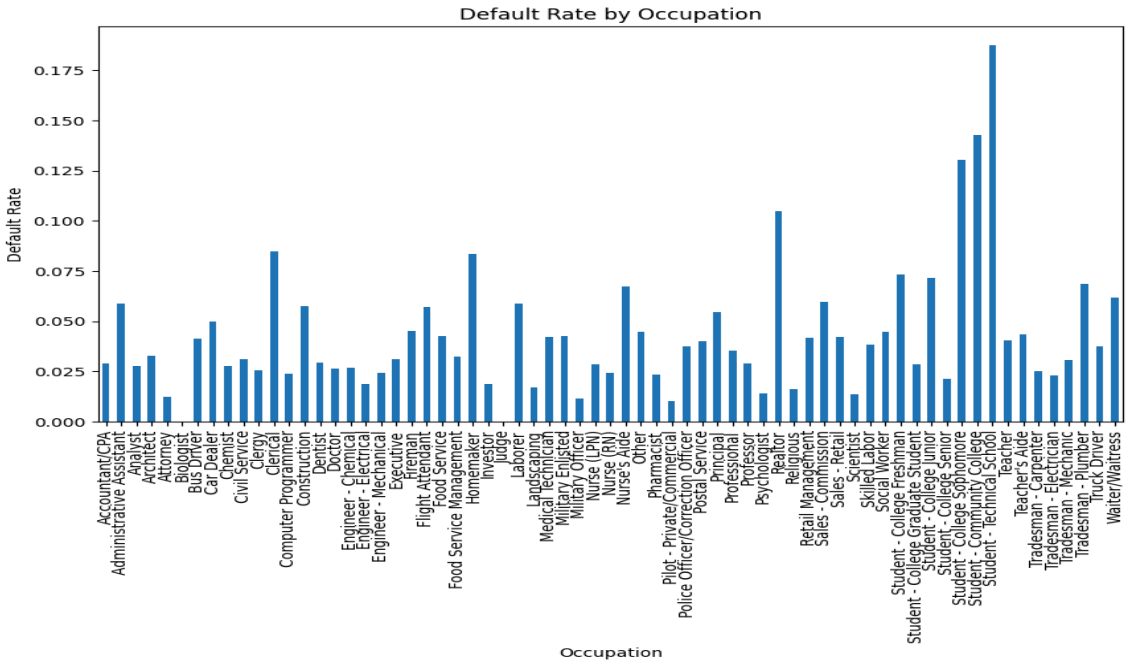

Figure 2: Loan Default Rate by Occupation

After visualizing the patterns of the data and considering the predictors used by the inspiring papers, 10 variables were kept as potential predictors for further investigation, including ProsperScore to represent credit score, LoanOriginalAmount to show the initial borrowing amount, isBorrowerHomeOwner to indicate whether the loan customer has their residential property, BorrowerRate for the interest rate to each borrower, and six others. 8 of the variables are numerical, the rest two categorical variables are the indicators of home owner status and income verifiability.

Correlation matrix was also used to avoid multicollinearity for the potential predictors we found above, and the plot is shown in Figure 4. Except for the high correlation between ProsperScore and BorrowerRate, at -0.72, all other correlations are small and acceptable. The strong negative correlation between those two variables might be because that the platform tends to offer a lower interest rate to a person with high credit score; however, both of them were kept

as potential predictor because the factors of loan customers used to calculate them are not the same and the algorithms used to calculate them are also different, so they could represent different aspects of a borrower.

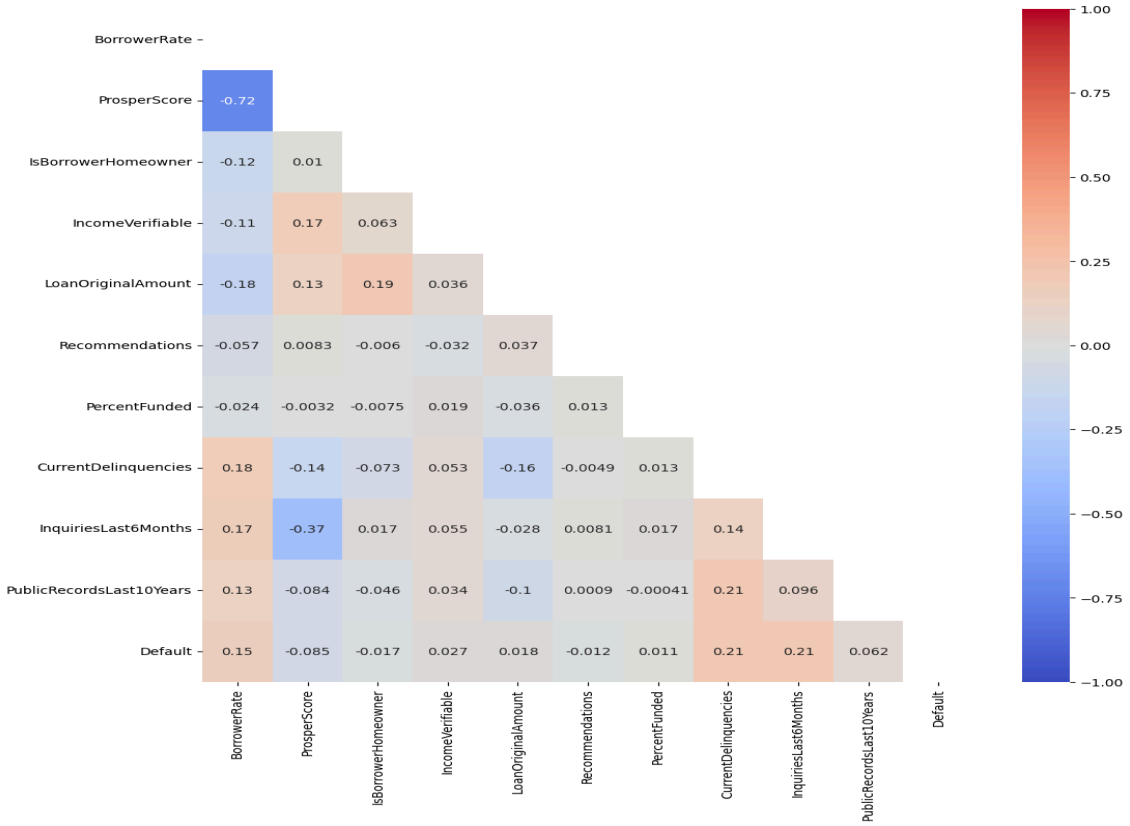

Figure 4: Correlation Matrix

	Interest Rate	Prosper Score	Loan Original Amount	Recommendations	Percent Funded	Current Delinquencies	Inquiries Last 6 Months	Public Records Last 10 Years
count	43092	20669	43092	43092	43092	42472	42472	42472
mean	0.383	0.549	0.1536	0.00232	0.960665	0.009304	0.02939	0.01346
std	0.162	0.230	0.1524	0.012048	0.054419	0.028452	0.04490	0.03287
min	0	0	0	0	0	0	0	0
25%	0.247	0.4	0.0456	0	0.96463	0	0	0
50%	0.362	0.6	0.1029	0	0.96463	0	0.01587	0
75%	0.518	0.7	0.2059	0	0.96463	0	0.03175	0
max	1	1	1	1	1	1	1	1

Table 1: Summary of Numerical Variables

1.2.2 Missing Data Imputation

Completing data processing, 75% of the data is randomly assigned to the training set, and the rest 25% is assigned to the test set. Missing values on two important predictors, credit grade and prosper score, which are both numerical variables for risk evaluation, are filled by the 5-nearest neighbors approach, which is widely used to deal with both quantitative and qualitative data (Radha and Visalakshi, 2014). Each missing value is imputed by the uniformly weighted average of values from 5 other observations that have minimum Euclidean distance to it, while such distance is calculated based on all other features included in the dataset. Importantly, the main reason for employing this tool is that variables being filled have rating properties, which are evaluations based on various perspectives. Individuals with similar values on such variables are rated at similar risk levels, potentially tending to share similar characteristics. Therefore, this study utilizes values from other most similar individuals to find the "best estimates" of the

missing ones. In contrast, missing values in individuals' regional information are not arbitrarily imputed. Instead, a separate category "unknown" is created and assigned to observations with no information in their state variable.

2. Methodology

Adhering to the existing literature, this study aims to investigate the respective strengths and shortcomings of the parametric method and the black box algorithm through comparing their classification capacity and the maximized expected payoff obtained after implementation. Two approaches, logistic regression and random forest algorithm, are employed respectively by this research. The two models are trained separately in predicting individual borrower's default status based on the explanatory variables that are considered to be the most effective. Besides, sharing the same objective, both models are tuned to achieve expected payoff maximization from the loan default prediction.

2.1 Classification Algorithms

2.1.1 Parametric Model - Logistic Regression

As a classic classification algorithm for binary outcome, a logistic regression returning the predicted log odds of defaulting is adopted, maintaining the following mathematical form:

$$log\left(\frac{p}{1-p}\right) = X\beta$$

where X and β are both in matrix form, representing predictor variables and their corresponding coefficients; p denotes the probability of default.

Two logistic regression models are separately fitted with and without the L2 (Ridge) regularization, which adds a penalty term involving the sum of squares of all the coefficient values to the objective function such that

$$L2\ regularization\ term\ =\ \lambda \sum_{j=0}^{J} \beta_j^{\ 2}$$

where λ represents the regularization coefficient that controls for the regularization strength, whereas β_j indicates a particular model coefficient. Given that

$$Cost\ =\ Logistic\ Loss\ Function\ +\ L2\ regularization\ term$$

, the regularization term penalizes large coefficients, resulting in a logistic regression in favor of simpler models with small coefficients, thereby preventing potential overfitting issues.

Predictor variables are determined by a likelihood-based approach, forward Akaike's Information Criteria (AIC) variable selection. Proposed by the statistician Hirotsugu Akaike, the AIC measuring model predictive performance follows the distribution that

$$AIC\ =\ -\ 2[ln\ (L(\widehat{\beta},\ \hat{\sigma}^{\ 2}|\ Y))\ -\ (p\ +\ 2)] \propto n\ ln(\frac{RSS}{n})\ +\ 2p$$

with p, n and RSS indicating the number of predictors, the sample size, and the residual sum of squares, respectively (Akaike, 1973). As an estimator of the prediction error, the AIC is aimed to be minimized during the selection process. Starting from a null model with a mere intercept, the forward selection algorithm adds predictor variables incrementally then examines the resulting AIC in each iteration, ultimately returning a best performing model with the smallest AIC (Akaike, 1973).

2.1.2 Black Box Algorithm - Random Forest

First introduced by William A. Belson in 1959, the decision tree classifies the data by splitting them into different divisions through each predictor node. The goal of such a tree is to minimize the impurity, which is quantified by the Gini index that

$$Gini\ Index\ =\ 1\ -\ \sum_{i=1}^{n}(P_i)^2\ =\ 1\ -\ [(P_+)^2\ +(P_-)^2]$$

where P_+ and P_- denote the probability of being classified as class 1 (default) and the probability of being classified as class 0 (not default) after a particular split (Belson, 1959). A decision tree aims to discover effective predictor nodes so that strong homogeneity, or low impurity, can be found among observations that are classified into the same division — in other words, achieving optimization through predictors with significant contribution to accurate classification.

Later in 2001, Leo Breiman and Adele Cutler introduced the random forest algorithm that incorporates multiple decision trees built based on random subsets of the entire dataset, where predictions by all trees are averaged and combined for better robustness (Breiman, 2001). After adopting the random forest algorithm, features are ranked by impurity-based feature importance, evaluated by means of each feature being able to help in the reduction of impurity.

2.2 Algorithm Tuning by Expected Payoff Maximization

Both models are tuned by choosing the best hyperparameter aiming to maximize the expected payoff of the firm, which can be expressed mathematically as the following:

$$\{\theta\}\ =\ argmax\ \beta\pi_{FN}+\ (1\ -\ \beta)\pi_{TP}+\ (1\ -\ \alpha)\pi_{TN}+\ \alpha\pi_{FP}$$

where θ represents the optimal set of hyperparameters of each model that yields maximum payoff. In this formula, α represents the type I error (false positive) rate, β represents the type II

error (false negative) rate, whereas π indicates the payoffs for each case including false negative, true positive, true negative and false positive. In this study, the payoff matrix is explicitly defined based on the direct earning and loss as the following:

Predicted / Actual	No Default (0)	Default (1)
No default (0)	$\pi_{TN} = Interest$	$\pi_{FN} = -(Loss\ of\ loan + Loss\ of\ Interest)$
Default (1)	$\pi_{FP} = 0$	$\pi_{TP} = 0$

With columns containing information regarding individual borrower's interest rate and loan amount in the dataset, the payoff is calculated at individual level to assess model performance. Specifically, the prediction result obtained by comparing the predicted and actual outcomes, either false negative, true positive, true negative or false positive, is incorporated into the calculation of individual payoff. For each specific observation, the payoff is the amount of gain or loss corresponding to its prediction result, calculated by the above matrix. In this way, with an entire column of individual payoff, the expected payoff is then computed as the weighted average of payoffs of all possible cases, associated with their corresponding probabilities.

Importantly, this design of the payoff matrix makes an extra assumption that the earnings and losses from the lenders are reflected in the payoff of the company. In other words, the company acts as if it represents the lenders, which is not necessarily true in the real world where Prosper acts like an intermediate agent and makes revenue by charging service fees from the lenders and the borrowers. The payoff matrix is designed this way for two major reasons. The first reason is that if the company is regarded as an intermediate agent, then in the case of false

negatives, the company will not experience any direct loss other than annoying or potentially losing the lenders as their clients, which is difficult to quantify. The second reason is that this design allows retrieval of all of the variables involved in calculating the payoffs per transaction from the dataset, thus making the best possible use of the available data to construct the payoff matrix. In particular, with the extra assumption made, the payoffs for false positive and true positive are zero because the loan will not be offered if the algorithm decides that the borrower is highly likely to default, so there is no direct financial loss.

2.2.1 Logistic Regression

For the logistic regression classification model, the hyperparameter is a single variable as the threshold for classification. After fitting the logistic regression model on the training set, we can use the model to produced predicted probability of default for each borrower given the covariates, and the default classification threshold is 0.5, meaning that those with predicted probability of default being greater than or equal to 0.5 will be classified as default, and no default otherwise. Tuning this threshold involves trying all possible values from 0 to 1 with 0.01 interval and calculate the expected payoff per transaction for each possible value, and choose the threshold value that yields the maximized expected payoff per transaction. With the optimal threshold computed, we then use this threshold alongside with the logistic model trained from the training set to carry out classification on the testing set, and report the prediction accuracy and expected payoff per transaction on the testing set.

2.2.2 Random Forest

To address any potential overfitting issue, the random forest model is tuned by retaining selective predictors with top 30 impurity-based feature importance, while its hyperparameters are optimized through grid searching, incorporating a custom score function that returns the

expected payoff calculated based on the classification outcome. Maintaining the ultimate goal of this study, the algorithm preserves hyperparameters delivering the maximum expected payoff.

Three hyperparameters are optimized, including the number of trees involved in the model, the maximum tree depth and the minimum sample size required to split an internal node. In their investigation involving the balance between Area Under the Curve (AUC) gain, processing time and memory usage, Oshiro, Perez and Baranauskas conclude an appropriate range of the number of trees in a forest should be between 64 and 128 (Oshiro et al., 2012). Therefore, three parity points within this interval, 64, 96, 128, are set as available choices. On the other hand, referring to Sandeep Ram's 2020 article, three of the most common options, 3, 5, 7, are set for the maximum tree depth (Ram, 2020). Finally, options for the minimum sample size required to split, 100, 300, 500, are determined with a combined consideration on the sizes of the training and test sets, where such a parameter generally varies from 1%-10% of the entire data volume. Lastly, a 5 fold cross-validation is used to test the algorithm's generalization power on new data.

3. Result

Note that prior to model fitting, we conducted a feature scaling (min-max normalization) to all of the relevant variables identified from exploratory data analysis. Such transformation would make interpretation from the logistic regression model less straightforward, but could significantly save computation time and avoid convergence issues induced by various spread of different variables during model fitting.

Feature scaling (min-max normalization):

$$X_{scaled} = \frac{X_{original} - min(X)}{max(X) - min(X)}$$

3.1 Preliminary Prediction Results

3.1.1 Logistic Regression Model

According to the forward AIC variable selection output, essential factors on loan default status include the interest rate charged to the borrower, the risk score determined by borrower's historical prosper record (prosper score), number of inquiries during six months prior to borrowing, the credit grade, income, information on whether the borrower reported the purpose of the loan, year and state. Categorical variables are also included, though not all levels are identified as the most influential: credit grade is selected only for AA, representing the highest credit, and HR, indicating the worst, while income is only selected for the $75k-$100k category. Regarding time and region fixed effects, year 2006 along with State Idaho, Massachusetts, Michigan, New York are chosen. Additionally, the unknown state is also considered to be significant.

With predictors determined by forward AIC variable selection, the classification performance of the logistic regression prior to model tuning is evaluated on the test set. Among a total of 10773 cases, 185 true positive (TP), 9513 true negative (TN), 77 false positive (FP) and 998 false negative (FN) are obtained, reaching a test accuracy of 0.9002 — approximately 90.02% of the the loan status in the test set are correctly predicted by this logistic algorithm. With L2 regularization, the pre-tuned logistic model results in an expected payoff of $405.53. Importantly, although the test accuracy seems fairly high, the true negative rate goes up to 84.36%, and such a remarkably problematic misclassification of actual default cases can potentially lead to very serious losses. Consequently, model tuning towards expected payoff maximization is needed to achieve the objective in preventing serious losses via default status prediction.

One of the crucial advantages for a parametric logistic regression model compared to a black-box model (like a random forest model) is that the model can produce estimates of regression coefficients, allowing one-to-one relationship between predictors and the outcome to be interpreted. Such interpretation can be a reliable reference for future decision making of the firm. Table 2 displays the numerical output from the logistic model without any regularization fitted on the training set transformed by min-max scaling, and we will make one example of interpretation here: The estimate for borrower rate is 3.0457, which means that, while holding all other variables constant, when the interest rate increases from 0 percent (the minimal value in the training set before scaling) to 1 percent (the maximal value in the training set before scaling), the log-odds of default will increase by 3.0457. This reveals the positive correlation between interest rate and the probability of default. From table 3, we can identify such positive correlation for interest rate, prosper rating score, number of current delinquencies (past due loans), number of credit inquires in the last 6 months, year 2006, and credit grade of HR (the lowest credit grade). In contrast, credit grade of AA (the highest credit grade), the income range of 75k to 100k, whether the borrower has reported a purpose for the loan, and whether the borrowers are in certain states, are negatively correlated with the probability of default.

Variable	**Coefficient**	**Std. Error**	**z**	**P>|z|**	**Lower CI**	**Upper CI**
Intercept	-2.9196	0.1449	-20.1447	2.99E-90	-3.2037	-2.6356
BorrowerRate	3.0457	0.1771	17.1929	3.00E-66	2.6985	3.3929
ProsperScore	0.3227	0.153	2.1088	3.50E-02	0.0228	0.6227
Current Delinquencies	5.5782	0.549	10.1605	2.98E-24	4.5021	6.6542
Inquiries Last 6 Months	5.932	0.3624	16.3678	3.25E-60	5.2217	6.6424
Creditgrade AA	-0.1592	0.0919	-1.7315	8.34E-02	-0.3393	0.021
Creditgrade HR	0.3658	0.0662	5.5297	3.21E-08	0.2362	0.4955
Income 75k-100k	-0.1951	0.0676	-2.8876	3.88E-03	-0.3275	-0.0627

Reported purpose	-1.2754	0.0499	-25.5428	6.59E-144	-1.3732	-1.1775
Year 2006	0.5742	0.0565	10.1551	3.15E-24	0.4634	0.685
State ID	0.0259	0.2233	0.1159	9.08E-01	-0.4118	0.4635
State MA	-0.4657	0.1744	-2.6697	7.59E-03	-0.8076	-0.1238
State MI	-0.0404	0.1008	-0.4012	6.88E-01	-0.238	0.1571
State NC	-0.3792	0.1403	-2.702	6.89E-03	-0.6542	-0.1041
State NY	-0.5067	0.1156	-4.3847	1.16E-05	-0.7333	-0.2802
State Unknown	-1.2212	0.0653	-18.6997	4.97E-78	-1.3492	-1.0932

Table 2: Output of Logistic Regression without Regularization

However, when implementing the logistic regression model with L2 regularization, we can no longer produce an informative table for statistical inference because the loss function is no longer a straightforward likelihood function, but with an additional penalty term. This feature causes estimates of regression coefficients to be biased toward small values, so the model actually sacrifices parts of interpretability to predictive performance (avoiding overfitting). Therefore, we will not be interpreting the output from the logistic regression model with L2 regularization.

3.1.2 Random Forest Model

As indicated in figure 3 from below, variables selected by impurity-based feature importance for the random forest model are found to be highly similar when compared to that from the AIC-based selection: borrower's interest rate, number of inquiries, proper score and information on borrower's reporting status on the loan purpose are again identified as strong contributors to the prediction, while additional significant predictors that are not selected by the AIC approach include the amount of the loan, number of borrower's previous accounts delinquent and the number of borrower's public records in the past 10 years. State and year are

similarly of high importance, yet more dummies are included as top influencers when compared to the selection made by forward AIC approach.

Incorporating all features in the cleaned dataset in the pre-tuned random forest model, a test accuracy of 0.8988 is achieved, which is slightly lower than that of the logistic regression. This might potentially be affected by an excessive amount of involved predictors. The expected payoff calculated using the test set yields $491.85, which is higher than that obtained from a pre-tuned logistic regression model. However, the black-box algorithm has relatively weak interpretability, not supporting analyses regarding statistical inferences of individual predictor effects on the outcome variable. Additionally, adjustments dedicated to profit maximization are indeed necessary considerations that should be incorporated into the model tuning process.

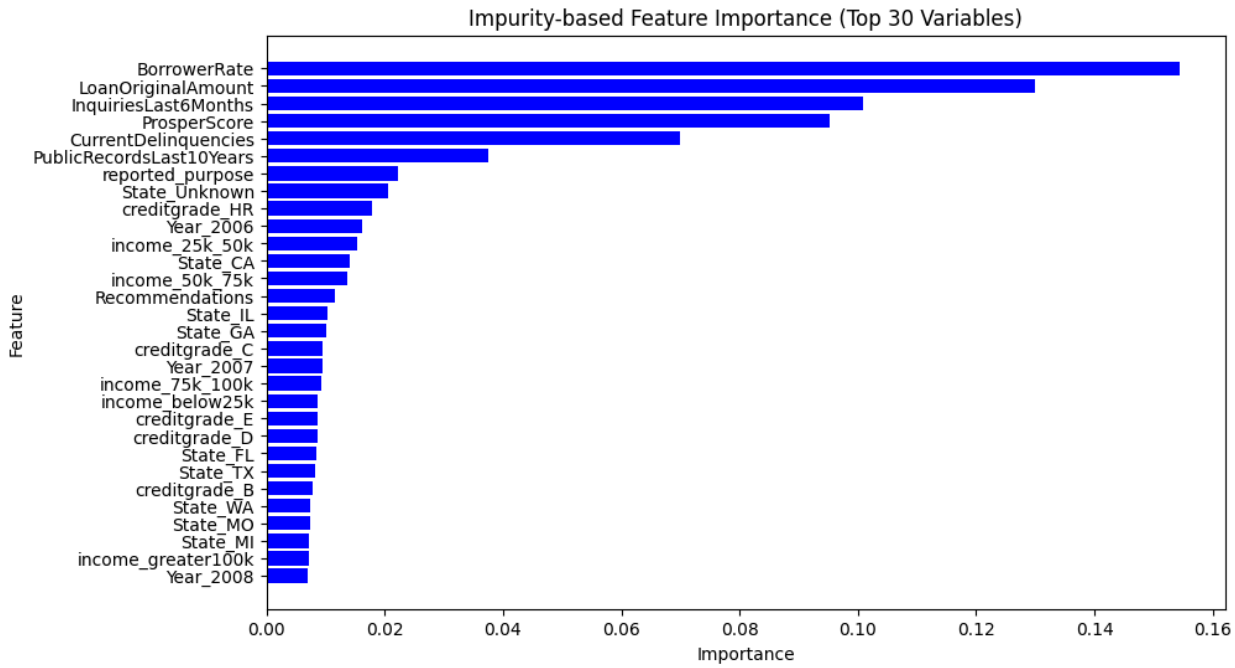

Figure 3: Feature Importance Matrix (Top 30 Variables)

3.2 Enhanced Performance with Model Tuning

After model fitting and model tuning on the training set to find the optimal threshold for classification, we used the model to predict the probability of default for each transaction in the testing set, and used the optimal threshold from the training set to carry out classification. The optimal threshold was 0.18, with optimal expected payoff per person as $513.98. After tuning the threshold, the accuracy of prediction reduced from 0.9007 to 0.8379.

Model performance is then evaluated for the tuned random forest algorithm. Having tested all possible combinations that can be composed with available hyperparameter choices, the expected payoff is computed using the test data for each iteration in the searching process. Maintaining the purpose of maximizing payoff, the algorithm ultimately reserves the hyperparameter combination yielding the highest expected payoff as the optimal choice. In this way, a number of trees of 96, a maximum tree depth of 3, and a minimum sample size to split of 100 are obtained, resulting in an optimal expected payoff of $973.33. With a huge increase of approximately 97.9% in the resulting payoff as compared to that obtained from a pre-tuned model, such result exemplifies the strong power of tuning in this black-box algorithm. Similarly, the test accuracy after model tuning rises to 0.9969, reflecting a substantial enhancement in predictive performance.

4. Discussion

4.1 Conclusion & Limitation

In this project, classification algorithms are developed to predict whether borrowers' default status on their loans with real data from Prosper. From the preliminary result section, it is noticed that the prediction accuracy of a logistic model with default threshold for classification

of 0.5 is slightly higher than that of a random forest model with default hyperparameters. However, after model tuning that aimed to maximize the payoff of the firm, the random forest model significantly outperforms the logistic classification model in terms of maximizing expected payoff and prediction accuracy, as shown in table 3. Note that implementing L2 regularization that aimed to reduce overfitting could achieve better prediction accuracy, but the optimal payoff per person is lower compared to an unadjusted logistic regression model.

Tuning	Model	Threshold	Prediction Accuracy	Optimal Payoff per Person in $
Pre-tuned	Logistic Regression	0.5 (default)	0.9007	402.92
	Logistic Regression (L2)	0.5 (default)	0.9002	405.53
	Random Forest	default	0.8988	491.85
Tuned	Logistic Regression	0.18 (tuned)	0.7705	549.62
	Logistic Regression (L2)	0.21 (tuned)	0.7927	541.74
	Random Forest	Number of trees = 96 Max tree depth = 3 Min samples to split at a node = 100	0.9969	973.33

Table 3 : Overall Model Comparison

Nevertheless, although the tuned random forest algorithm performs exceptionally well in terms of prediction accuracy in the test set along with the capacity of yielding a remarkably high expected payoff, a major problem is that it takes extremely high cost to search for optimized hyperparameters as exhausting all possible combinations is both time-consuming and expensive. In particular, it took over four hours to grid search for merely three optimal hyperparameters.

Moreover, the random forest model is essentially a black-box model that does not provide direct interpretation of the relationship between predictors and the outcome, whereas for a parametric logistic model, such interpretation is available. We concluded that the borrowers with higher number of current delinquency (past due payment) history and credit profile inquiries (number of times when a credit profile is checked) in their accounts are much more likely to default. Also, borrowers are more likely to default when the interest rate is higher. Conversely, we found that the borrowers who report their purposes of loan or those with better credit scores are less likely to default. In practice, it would be up to the company to decide which algorithm to apply. If the company places a strong emphasis on making short-term revenues, then a black-box model (a random forest model, in this case) is preferable to adopt in order to maximize their payoff. On the other hand, if the company is more interested in understanding the relationship between factors to make future decisions, a transparent logistic model is more preferable.

This study does involve some limitations. Firstly, missing values in credit grade and prosper score are imputed using the 5-nearest neighbors approach, which is reasonable but not a perfect approach since the missing data are estimated rather than observed. The estimation may be inaccurate and sensitive to outliers. Secondly, the assumption that the firm's payoff being directly associated with the earnings and losses of lenders might not be a realistic representation of the payoff matrix encountered by firms in the real world. Thirdly, peer-to-peer loan companies often maximize their payoff by choosing the optimal interest rate for each consumer, whereas models in this study mainly focus on classification — identifying which borrowers are more likely to default and, consequently, suggesting the lenders to not offer the loan. Finally, the proportion of defaulted observations is much smaller than those who did not default, resulting in an unbalanced dataset to carry out statistical modeling, which could induce bias in estimation.

4.2 Future Investigation

This research utilizes a dataset obtained from an open source, therefore might lack access to high sensitive information. Nevertheless, such information has great potential to affect borrowers' credibility. In the future, interested researchers can seek for a more updated and complete version of data to possibly retrieve more insightful variables and avoid imputation of missing data if possible. One may also develop a different pricing algorithm that finds the optimal interest rate to charge for each borrower, which would be more realistic than this study. Furthermore, one can investigate ways to fix the unbalanced distribution of default in the dataset, or obtain a more balanced dataset to make the modeling result more generalizable. Finally, one may also develop or apply other machine-learning or parametric models and compare the performance and interpretability of these models to yield more informative conclusions.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In: Petrov, B.N. and Csaki, F., Eds., International Symposium on Information Theory, 267-281.

Belson, W. A. (1959). Matching and Prediction on the Principle of Biological Classification. Journal of the Royal Statistical Society. Series C (Applied Statistics), 8(2), 65–75. https://doi.org/10.2307/2985543

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32. http://dx.doi.org/10.1023/A:1010933404324

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Applied Economics, 47(1), 54–70. https://doi.org/10.1080/00036846.2014.962222

Maheswari, P.,& Narayana, C. (2020). Predictions of Loan Defaulter - A Data Science Perspective. 2020 5th International Conference on Computing, Communication and Security(ICCCS). Patna, India, 2020, pp. 1-4. https://doi.org/10.1109/ICCCS49678.2020.9277458

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest? Machine Learning and Data Mining in Pattern Recognition (MLDM 2012). 154–168. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-31537-4_13

Ram, S. (2020, October 18). Mastering random forests: A comprehensive guide. Medium. https://towardsdatascience.com/mastering-random-forests-a-comprehensive-guide-51307c129cb1#:~:text=max_depth%3A%20The%20number%20of%20splits,shown%20to%20each%20decision%20tree

Sheikh, M. A., Goel, A. K., & Kumar, T. (2020). An Approach for Prediction of Loan Approval using Machine Learning Algorithm. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Jul. 2020, pp. 490–494. https://doi.org/10.1109/ICESC48915.2020.9155614.

Victor, L., & Raheem, M. (2021). Loan Default Prediction Using Genetic Algorithm: A Study Within Peer-To-Peer Lending Communities. Int. J. Innov. Sci. Res. Technol., vol. 6, no. 3 1195–1205.

Visalakshi, S., & Radha, V. (2014). KNN with Pruning Algorithm for Simultaneous Classification and Missing Value Handling. International Journal of Innovative Research in Science, Engineering and Technology Volume 3, Special Issue 3, March 2014. ISSN No: 2319 - 8753

Zhang, T., & Sun, W. (2022). Research on P2P Default Risk Prediction Based on Logistic Regression. https://doi.org/10.1109/IIP57348.2022.00036

Zhu, L., Qiu, D., Ergu,D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. Procedia Computer Science,Volume 162, pp. 503-513, ISSN 1877-0509. https://doi.org/10.1016/j.procs.2019.12.017.

Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L. (2023). Explainable prediction of loan default based on machine learning models. Data Science and Management, 6, 123–133. https://doi.org/10.1016/j.dsm.2023.04.003