Predicting Loan Default Risk in P2P Lending Platforms: A Study of Lending Club Borrowers

Vinay Singh

Oracle fusion Lead, McGraw Hill, Columbus, OH
Email: *vinay.ocp[at]gmail.com*

Abstract: ***The peer-to-peer lending industry has experienced rapid growth due to increasing demand from borrowers and lenders. These platforms have done well because of digital changes and the wider reach of the Internet, which connects people of all ages and backgrounds. Lending institutions encounter significant challenges in accurately predicting loan defaults. When large loan amounts are defaulted, it results in considerable business losses. This study focuses on loan defaults in online peer-to-peer lending. The dataset "BALANCED_Data_Predicting_Default. csv" used for this research was sourced from Carmen—Ohio State University. This dataset contains 58 variables on 20, 000 actual Lending Club loans issued in 2015. This dataset was loaded to Orange3, a data mining application. The loan status was selected as the dependent variable and categorized into two groups: "default" and "fully paid" loans. The dataset was preprocessed to remove any irrelevant data. We evaluated the variance and removed variables with little variation. Some attributes were excluded based on our judgment and business knowledge. Certain columns, such as "collection_recovery_fee" and "recoveries, " were considered irrelevant since they didn't provide useful insights into loan defaults. This research aims to apply AI and ML, specifically Decision Trees, logistic regression, Random Forests, SVM, Neural Networks, and gradient boosting, to predict the default probability of Lending Club borrowers. As part of this research, we will pick the best-performing Model and report its performance. If these models are used, Lending Clubs and loan companies can make data-driven decisions, enhance services, and predict customers' defaults. We have tried multiple machine learning models, including logistic regression, random forest, gradient boosting trees, support vector machine (SVM), and neural networks. We tuned the parameters of different models (e. g., the number of layers in neural networks). In this case, the gradient boosting tree performs well, as we achieved the best result, F1 0.883.***

Keywords: P2P Lending, Lending Club, Orange3, Imbalanced dataset, Loan-default, Prediction, Logistic Regression, Random Forest, Gradient Boosting tree, Support Vector Machine (SVM) and Neural Networks

1. Introduction

Financial institutions lend capital as a key strategy to generate revenue and stay competitive. However, there's a risk of losing money if borrowers don't repay their loans.

In many developing countries, the lending risk is so high that it's almost like a lottery. Lack of reliable borrower information and old technology makes it hard to keep track of people who do not repay loans. Lending organizations need a good history of loan repayments to stay in business. One way they try to ensure this is by increasing interest rates to cover potential losses, but this isn't a smart economic choice.

Certain factors can increase the likelihood of loan defaults, making it crucial to develop effective predictive methods. Many studies have been carried out on predicting defaults, and here are a few summaries. One such study by Cheng et al. (2019) focused on predicting loan repayment patterns based on mobile phone usage behavior.

The default rate is likely to be much higher in a peer-to-peer lending (P2P) platform, which operates without a third party. According to Byanjankar et al. (2015), various methods have been used for predictions, but machine learning and data mining techniques have been the most successful. This study follows the booming trend of improving prediction accuracy. Therefore, given a set of loan defaulters {l1, l2,. . ., ln}, the study aims to preprocess and balance the dataset before creating a prediction model to determine whether each instance results in a loan default or is fully paid.

We aim to improve prediction performance using an imbalanced dataset of loan defaulters "BALANCED_Data_Predicting_Default. csv. " The dataset used for this research was sourced from Carmen—Ohio State University.

We have observed significant variations in the results while using different models due to our imbalanced dataset. We could have balanced the data and improve predictions by utilizing various sampling strategies. However, time constraints prevented us from implementing these techniques at this stage.

This study will help financial institutions identify potential loan defaulters, allowing them to reduce anticipated losses or overhead costs.

2. Methodology

Data Description

The dataset obtained for the analysis was extracted from Carmen—Ohio State University, comprising loan default records from 2007-2015. The Lending Club is an online peer-to-peer lending platform headquartered in San Francisco, California. A sample of the dataset is provided in Figure 1. The offspring or new feature sets generated or selected for modeling are the borrower's loan amount, interest rate, installment, and annual income, all of which have a numerical datatype. Note that these five are the independent features or variables. The dependent or target variable used for the analysis is the loan status, a categorical data type classified into two main categories: "Default" and "Fully Paid. " From

International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

the figure, the classification identifies that 50% of loans are Fully paid, whereas 50% are defaulted.

This is an indication of sound risk management in the lending club. Though prediction accuracy may be very high, the Model's output will be inaccurate due to overfitting, which will happen in the training phase. The loan dataset's loan status was highly imbalanced. This calls for preprocessing the data by using practical approaches such as cleaning and selecting appropriate features for classification.

loan_status	loan_amnt	int_rate	grade	sub_grade	home_ownership	annual_inc	verification_status	purpose	zip_code	addr_state	dti	delinq_2yrs	fico_range_high	inq_last_6mths	open_acc	pub_
Default	5625	14.65	C	C5	OWN	16000	Verified	credit_card	760xx	TX	13.5	0	669	5	20	
Default	5000	12.29	C	C1	RENT	42000	Not Verified	debt_consolidation	123xx	NY	18.17	0	669	0	13	
Default	5000	12.29	C	C1	OWN	40000	Verified	credit_card	310xx	GA	8.58	0	664	0	7	
Default	8000	14.65	C	C5	MORTGAGE	48000	Verified	debt_consolidation	219xx	MD	14.13	0	704	3	11	
Default	20000	9.99	B	B3	MORTGAGE	48000	Not Verified	credit_card	720xx	AR	13.2	0	709	2	7	
Default	4200	14.65	C	C5	RENT	52000	Verified	other	191xx	PA	24.37	0	674	2	22	
Default	2400	9.17	B	B2	RENT	35000	Not Verified	debt_consolidation	956xx	CA	18.52	0	714	1	10	
Default	10000	14.31	C	C4	RENT	70000	Verified	other	914xx	CA	13.44	0	704	3	7	
Default	4025	15.61	D	D1	RENT	40000	Not Verified	debt_consolidation	604xx	IL	19.2	0	669	1	6	
Default	13075	17.57	D	D4	MORTGAGE	128000	Verified	debt_consolidation	921xx	CA	18.74	0	669	1	13	
Default	27125	12.39	C	C1	RENT	145000	Verified	credit_card	967xx	HI	23.47	0	704	0	11	
Default	12000	12.69	C	C2	RENT	84000	Verified	moving	974xx	OR	17.74	0	704	0	9	
Default	14000	11.53	B	B5	RENT	42500	Verified	credit_card	100xx	NY	20.1	2	669	0	9	
Default	12800	13.33	C	C3	OWN	46000	Verified	debt_consolidation	251xx	WV	9.34	0	719	0	4	
Default	9525	23.99	F	F3	MORTGAGE	31800	Verified	debt_consolidation	851xx	AZ	26.92	0	664	0	7	
Default	15000	9.99	B	B3	RENT	75000	Verified	debt_consolidation	770xx	TX	5.71	0	669	1	10	
Default	16000	12.05	C	C1	RENT	62400	Verified	debt_consolidation	900xx	CA	33.67	0	684	1	22	
Default	5125	17.57	D	D4	RENT	16000	Verified	debt_consolidation	928xx	CA	14.85	0	689	0	4	
Default	5525	19.99	F	F4	RENT	28800	Verified	small_business	324xx	FL	22.92	0	679	0	7	

BALANCED_Data_Predicting_Defaul

Figure 1: Sample of dataset

Data Preprocessing

Data Cleaning

Any data with missing values were removed to prevent misclassifications. Orange provides various data cleaning tools, such as imputation, normalization, filtering, and outlier detection. Filtering was applied to eliminate irrelevant or redundant data, while outlier detection was used to identify and remove data points that significantly differed from the rest of the dataset.

When we loaded the data into Orange3, we defined the proper type and role for your variables of interest. The variables are defined below:-

- Discrete (Categorical)
- Continuous (Numerical)
- String
- Meta:-to Provide extra information.

Feature selection

The initial stage involves choosing the right column as the target for our Model. Our primary objective is to forecast loan repayment behavior, distinguishing between those who will fulfill their loans and those who will default. In the dataset "BALANCED_Data_Predicting_Default, " we identified the sole attribute representing the loan status. This attribute will serve as the target column. The data reveals an equal split, with approximately 50% categorized as non-defaulters (fully paid) and the remaining 50% as defaulters.

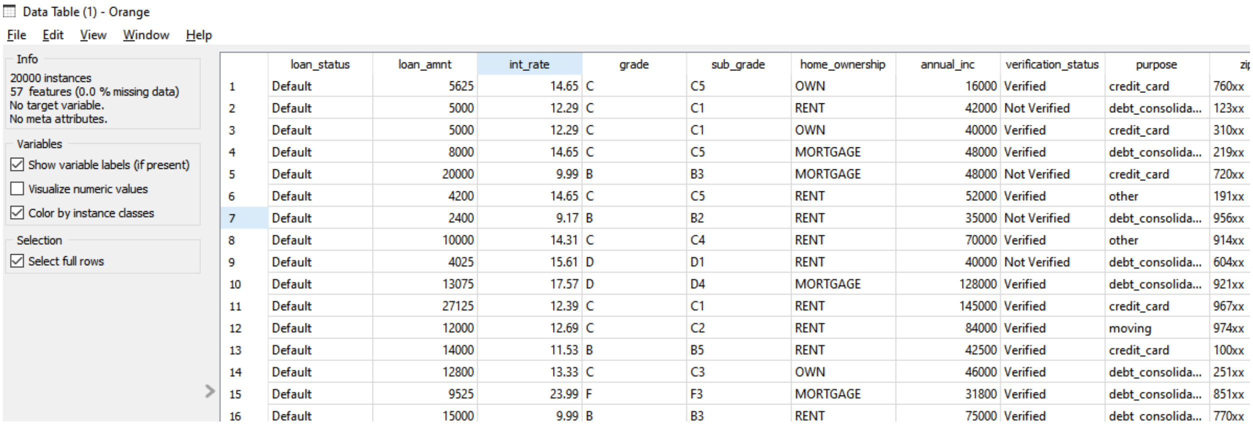
The non-informative predictors are removed to improve the prediction model's performance. The dependent feature is the loan status classified into default and fully paid loans; the remaining features are independent. Except for the member ID column, which was used as a unique identifier, all independent features chosen are numerical, whereas the dependent feature is categorical.

The dataset comprises default indicators, payment records, credit histories, and more. This dataset considers individuals classified as 'current' status as non-defaulters. Additionally, we've received a data dictionary offering detailed descriptions of the features included.

Feature Statistics: Figure 2 for instance shows Feature Statistics, Figure 3: without Imputing the data and without removing outliers & Figure 4: Impute the data and by removing outliers.

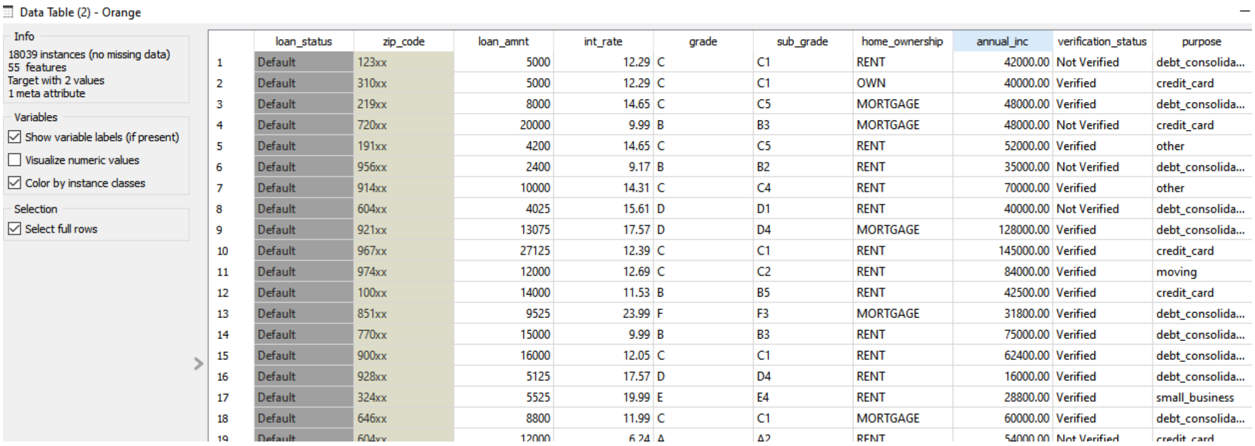
Figure 2: Feature Statistics

International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942



	loan_status	loan_amnt	int_rate	grade	sub_grade	home_ownership	annual_inc	verification_status	purpose	zi
1	Default	5625	14.65	C	C5	OWN	16000	Verified	credit_card	760xx
2	Default	5000	12.29	C	C1	RENT	42000	Not Verified	debt_consolida...	123xx
3	Default	5000	12.29	C	C1	OWN	40000	Verified	credit_card	310xx
4	Default	8000	14.65	C	C5	MORTGAGE	48000	Verified	debt_consolida...	219xx
5	Default	20000	9.99	B	B3	MORTGAGE	48000	Not Verified	credit_card	720xx
6	Default	4200	14.65	C	C5	RENT	52000	Verified	other	191xx
7	Default	2400	9.17	B	B2	RENT	35000	Not Verified	debt_consolida...	956xx
8	Default	10000	14.31	C	C4	RENT	70000	Verified	other	914xx
9	Default	4025	15.61	D	D1	RENT	40000	Not Verified	debt_consolida...	604xx
10	Default	13075	17.57	D	D4	MORTGAGE	128000	Verified	debt_consolida...	921xx
11	Default	27125	12.39	C	C1	RENT	145000	Verified	credit_card	967xx
12	Default	12000	12.69	C	C2	RENT	84000	Verified	moving	974xx
13	Default	14000	11.53	B	B5	RENT	42500	Verified	credit_card	100xx
14	Default	12800	13.33	C	C3	OWN	46000	Verified	debt_consolida...	251xx
15	Default	9525	23.99	F	F3	MORTGAGE	31800	Verified	debt_consolida...	851xx
16	Default	15000	9.99	B	B3	RENT	75000	Verified	debt_consolida...	770xx

Figure 3: Without Impute and removing outliers:-20000 instances & 57 features



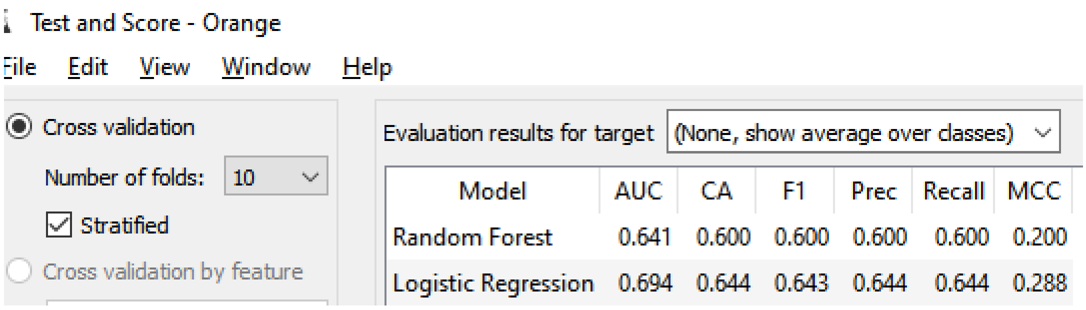
	loan_status	zip_code	loan_amnt	int_rate	grade	sub_grade	home_ownership	annual_inc	verification_status	purpose
1	Default	123xx	5000	12.29	C	C1	RENT	42000.00	Not Verified	debt_consolida...
2	Default	310xx	5000	12.29	C	C1	OWN	40000.00	Verified	credit_card
3	Default	219xx	8000	14.65	C	C5	MORTGAGE	48000.00	Verified	debt_consolida...
4	Default	720xx	20000	9.99	B	B3	MORTGAGE	48000.00	Not Verified	credit_card
5	Default	191xx	4200	14.65	C	C5	RENT	52000.00	Verified	other
6	Default	956xx	2400	9.17	B	B2	RENT	35000.00	Not Verified	debt_consolida...
7	Default	914xx	10000	14.31	C	C4	RENT	70000.00	Verified	other
8	Default	604xx	4025	15.61	D	D1	RENT	40000.00	Not Verified	debt_consolida...
9	Default	921xx	13075	17.57	D	D4	MORTGAGE	128000.00	Verified	debt_consolida...
10	Default	967xx	27125	12.39	C	C1	RENT	145000.00	Verified	credit_card
11	Default	974xx	12000	12.69	C	C2	RENT	84000.00	Verified	moving
12	Default	100xx	14000	11.53	B	B5	RENT	42500.00	Verified	credit_card
13	Default	851xx	9525	23.99	F	F3	MORTGAGE	31800.00	Verified	debt_consolida...
14	Default	770xx	15000	9.99	B	B3	RENT	75000.00	Verified	debt_consolida...
15	Default	900xx	16000	12.05	C	C1	RENT	62400.00	Verified	debt_consolida...
16	Default	928xx	5125	17.57	D	D4	RENT	16000.00	Verified	debt_consolida...
17	Default	324xx	5525	19.99	E	E4	RENT	28800.00	Verified	small_business
18	Default	646xx	8800	11.99	C	C1	MORTGAGE	60000.00	Verified	debt_consolida...
19	Default	604xx	12000	6.24	A	A2	RENT	54000.00	Not Verified	credit_card

Figure 4: With Impute and removing outliers:-18039 instances & 55 features

Performance valuation
Analysis of loan default predictions using different methods.

Regression Models

Random forest
A random forest is a meta-estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve predictive accuracy and control overfitting. The sub-sample size is always the same as the original input sample size, but the samples are drawn with replacement if bootstrap=True (default). With this Model, we got an accuracy score of F1 (0.60), as shown in figure 5.



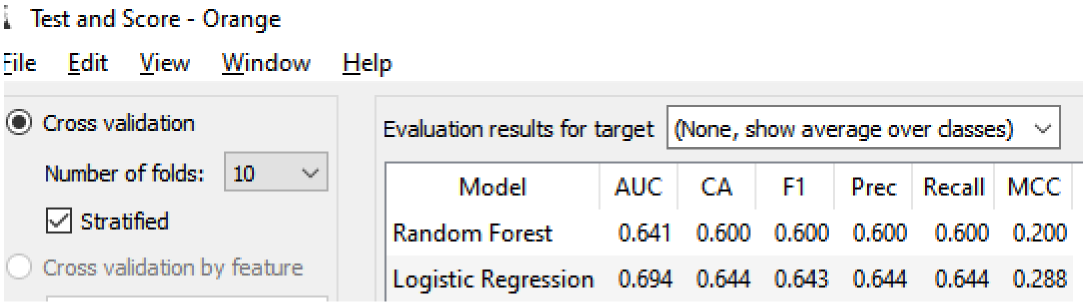
Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.641	0.600	0.600	0.600	0.600	0.200
Logistic Regression	0.694	0.644	0.643	0.644	0.644	0.288

Figure 5: Test and Score

Logistic regression
Logistic regression predicts the likelihood of an event happening using an equation based on specific input features. The Model learns the connection between these features and the outcome and then uses this relationship to estimate the probability of the event occurring with new data.

The logistic Model shows favorable Accuracy and Precision scores. However, due to our imbalanced dataset, we observe significant variations in results among different models. While employing diverse sampling techniques could balance the data and improve predictions, time constraints prevented us from implementing these techniques at this stage.

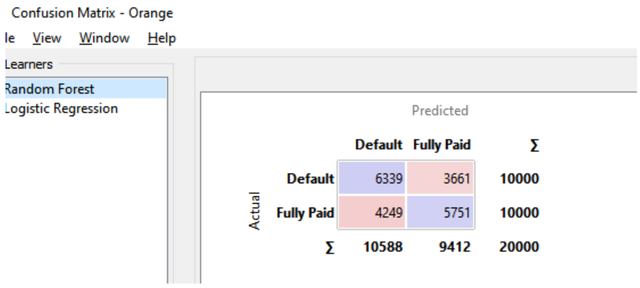
The Lending Club should prioritize "Grade" as a significant factor in their loan provision process. Additionally, the likelihood of default rises as annual income decreases, reaching a peak within the salary range of 0 to 25000. To address this, the Lending Club might consider commencing with lower principal loan amounts or conducting thorough credibility checks for applicants falling within this income bracket. Moreover, as interest rates increase, the probability of default also rises. Lending clubs should consider narrowing their interest rate range for self-employed applicants with less than a year of experience, as they exhibit a higher probability of default. With this Model, we got an accuracy score of 64, as shown in figure 6.



Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.641	0.600	0.600	0.600	0.600	0.200
Logistic Regression	0.694	0.644	0.643	0.644	0.644	0.288

Figure 6: Test and Score

Confusion Matrix: We cannot conclude that this is the best Model. We will have to perform cross-validation tests to check the results, as shown in figure 7.



	Default	Fully Paid	Σ
Default	6339	3661	10000
Fully Paid	4249	5751	10000
Σ	10588	9412	20000

Figure 7: Confusion Matrix

To pick the best-performing Model and report its performance, we used the "Test and Score" widget and examined different performance metrics, such as the "F1" score.

In addition to logistic regression, SVM, Neural networks, and random forest, we tried gradient boosting, particularly extreme gradient boosting. We tuned the parameters for extreme gradient boosting, including the number of trees, learning rate, lambda, and tree depth. The best F1 we could get from gradient boosting is 0.883. The following table shows the parameter study and the corresponding performance, as seen in figure 8:

extreme gradient bossting													
#trees	lr	lambda	depth	training insta	tree features	level feature	split feature	AUC	CA	F1	Prec	recall	MCC
200	0.3	1	6	1	1	1	1	0.917	0.872	0.872	0.882	0.872	0.754
300	0.3	1	6	1	1	1	1	0.916	0.873	0.872	0.882	0.873	0.755
100	0.3	1	6	1	1	1	1	0.918	0.878	0.877	0.891	0.878	0.769
50	0.3	1	6	1	1	1	1	0.921	0.833	0.882	0.898	0.883	0.781
20	0.3	1	6	1	1	1	1	0.923	0.884	0.883	0.903	0.884	0.786
20	0.1	1	6	1	1	1	1	0.923	0.885	0.883	0.906	0.885	0.791
20	0.5	1	6	1	1	1	1	0.921	0.880	0.879	0.893	0.880	0.773
20	0.1	5	6	1	1	1	1	0.918	0.880	0.879	0.895	0.880	0.775
20	0.1	0.1	6	1	1	1	1	0.919	0.878	0.877	0.891	0.878	0.769

Figure 8: Test and Score

We also tried some other strong models, including SVM and neural networks (with necessary parameter tunings such as several layers and hidden dimensions in neural networks), and the following table shows their performance, which is not as good as the gradient boosting tree, as seen in figure 9:

Model	AUC	CA	F1	Prec	Recall	MCC
Gradient Boosting	0.924	0.885	0.883	0.906	0.885	0.791
SVM	0.585	0.564	0.563	0.564	0.564	0.128
Neural Network	0.873	0.799	0.799	0.801	0.799	0.600

Figure 9: Performance

"Prediction" widget to generate predicted loan status:

The cross-validation scores suggest the gradient boosting tree is the best Model. Below is the prediction from gradient boosting model, shown in figures 10 & 11:

Paper ID: SR231114083515 DOI: https://dx.doi.org/10.21275/SR231114083515

International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

loan_status	Gradient Boosting	ient Boosting (De	t Boosting (Ful	Fold	loan_amnt	int_rate	grade	sub_grade	home_ownership	annual_inc
Fully Paid	Fully Paid	0.103774	0.896226	1	16000	5.32	A	A1	MORTGAGE	150000
Fully Paid	Fully Paid	0.105689	0.894311	5	20000	6.24	A	A2	MORTGAGE	87000
Fully Paid	Fully Paid	0.105689	0.894311	5	26500	5.32	A	A1	MORTGAGE	88500
Fully Paid	Fully Paid	0.105689	0.894311	5	13000	5.32	A	A1	MORTGAGE	70000
Fully Paid	Fully Paid	0.105689	0.894311	5	18100	6.03	A	A1	MORTGAGE	105000
Fully Paid	Fully Paid	0.105689	0.894311	5	14000	5.32	A	A1	MORTGAGE	122000
Fully Paid	Fully Paid	0.106213	0.893787	5	15000	5.32	A	A1	MORTGAGE	85000
Fully Paid	Fully Paid	0.106722	0.893278	1	28000	5.93	A	A1	MORTGAGE	136000
Fully Paid	Fully Paid	0.106722	0.893278	1	28000	5.32	A	A1	MORTGAGE	250000
Fully Paid	Fully Paid	0.106722	0.893278	1	22000	5.32	A	A1	MORTGAGE	190000
Fully Paid	Fully Paid	0.106722	0.893278	1	12000	6.89	A	A3	MORTGAGE	100000
Fully Paid	Fully Paid	0.106722	0.893278	1	22150	6.24	A	A2	OWN	123023
Fully Paid	Fully Paid	0.106722	0.893278	1	27600	5.32	A	A1	MORTGAGE	125000
Fully Paid	Fully Paid	0.108255	0.891745	5	28000	6.68	A	A3	MORTGAGE	115000
Fully Paid	Fully Paid	0.108255	0.891745	5	19500	6.24	A	A2	MORTGAGE	95000
Fully Paid	Fully Paid	0.108255	0.891745	5	28000	6.92	A	A4	MORTGAGE	160000
Fully Paid	Fully Paid	0.108285	0.891715	5	10000	6.24	A	A2	MORTGAGE	80000
Fully Paid	Fully Paid	0.108285	0.891715	5	6500	5.32	A	A1	MORTGAGE	70000
Fully Paid	Fully Paid	0.108285	0.891715	5	14000	6.89	A	A3	MORTGAGE	70000
Fully Paid	Fully Paid	0.108285	0.891715	5	17500	5.32	A	A1	MORTGAGE	95000
Fully Paid	Fully Paid	0.108285	0.891715	5	20000	5.32	A	A1	MORTGAGE	100000
Default	Fully Paid	0.108285	0.891715	5	19600	5.32	A	A1	MORTGAGE	125000
Fully Paid	Fully Paid	0.108374	0.891626	3	8000	7.26	A	A4	MORTGAGE	80000
Fully Paid	Fully Paid	0.108374	0.891626	3	28000	5.32	A	A1	MORTGAGE	163134

Figure 10: Prediction

loan_status	Gradient Boosting	ient Boosting (De	t Boosting (Ful	Fold	loan_amnt	int_rate	grade	sub_grade	home_ownership	annual_inc
Default	Default	0.934347	0.0656527	2	27125	12.39	C	C1	RENT	145000
Default	Default	0.934347	0.0656527	2	17325	13.99	C	C4	RENT	41302
Default	Default	0.934347	0.0656527	2	32900	16.99	D	D3	RENT	71000
Default	Default	0.934347	0.0656527	2	8000	12.59	C	C2	OWN	35000
Default	Default	0.934347	0.0656527	2	2500	16.99	D	D3	RENT	57283.2
Default	Default	0.934347	0.0656527	2	6000	13.33	C	C3	RENT	35525
Default	Default	0.934347	0.0656527	2	6400	13.99	C	C4	RENT	63200
Default	Default	0.934347	0.0656527	2	20000	13.67	C	C4	MORTGAGE	65000
Default	Default	0.934347	0.0656527	2	8500	13.33	C	C3	MORTGAGE	30000
Default	Default	0.934347	0.0656527	2	17025	18.99	E	E1	OWN	62000
Default	Default	0.934347	0.0656527	2	15000	17.57	D	D4	RENT	200000
Default	Default	0.934347	0.0656527	2	32000	13.99	C	C4	RENT	67200
Default	Default	0.934347	0.0656527	2	17475	17.86	D	D5	RENT	35000
Default	Default	0.934347	0.0656527	2	8000	11.99	B	B5	RENT	50000
Default	Default	0.934347	0.0656527	2	31825	14.65	C	C5	RENT	75000
Default	Default	0.934347	0.0656527	2	20125	13.67	C	C4	MORTGAGE	58000
Default	Default	0.934347	0.0656527	2	16000	11.53	B	B5	RENT	145000
Default	Default	0.934347	0.0656527	2	8000	14.31	C	C4	RENT	30189
Default	Default	0.934347	0.0656527	2	15000	13.18	C	C3	RENT	68471
Default	Default	0.934347	0.0656527	2	8575	19.99	E	E4	RENT	44000
Default	Default	0.934347	0.0656527	2	2875	17.57	D	D4	RENT	24000
Default	Default	0.934347	0.0656527	2	8000	13.33	C	C3	RENT	70000
Default	Default	0.934347	0.0656527	2	3700	23.99	F	F2	RENT	77000
Default	Default	0.934347	0.0656527	2	20000	18.25	E	E1	RENT	50000
Default	Default	0.934347	0.0656527	2	7200	12.69	C	C2	RENT	41000

Figure 11: Prediction

3. Conclusion

We have tried different models: logistic regression, random forests, gradient-boosting trees, support vector machines, and neural networks. In this case, the gradient-boosting tree performs well. The results indicate that the models used are effective in improving the accuracy of predicting loan defaults. However, the study only considers fully paid and default loans without accounting for loans that are considered risky.

References

[1] Cheng, D., Zhang, Y., Yang, F., Tu, Y., Niu, Z., & Zhang, L. (2019, November). A dynamic default prediction framework for networked-guarantee loans. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp.2547-2555). doi: 10.1145/3357384.3357804

[2] Byanjankar, A., Heikkilä, M., & Mezei, J. (2015, December). Predicting credit risk in peer-to-peer lending: A neural network approach. In 2015 IEEE Symposium Series on Computational Intelligence (pp.719-725). IEEE. doi: 10.1109/ SSCI.2015.109

[3] "Prediction of LendingClub loan defaulters. " https: //www.kaggle. com/code/deepanshu08/prediction-of-lendingclub-loan-defaulters/notebook.

[4] "Orange Data Mining Library. " https: //orange3. readthedocs. io/projects/orange-data-mining-library/en/latest/.

[5] "Making Predictions. " https: //orangedatamining. com/blog/making-predictions/.

[6] "Predictions. " https: //orange3. readthedocs. io/en/3.4.0/widgets/evaluation/predictions. html.

[7] "Test and Score" https: //orange3. readthedocs. io/en/3.4.0/widgets/evaluation/predictions. html.

[8] "Feature Statistics" https: //orange3. readthedocs. io/projects/orange-visual-programming/en/latest/widgets/data/featurestatistics. html

Appendix

Lending Club variables with description:

Variables	Description
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
acc_open_past_24mths	Number of trades opened in past 24 months.
addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers

avg_cur_bal	Average current balance of all accounts
bc_open_to_buy	Total open to buy on revolving bankcards.
chargeoff_within_12_mths	Number of charge-offs within 12 months
collection_recovery_fee	post charge off collection fee
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
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.
mths_since_earliest_cr_line	The month the borrower's earliest reported credit line was opened
fico_range_high	The upper boundary range the borrower’s FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower’s FICO at loan origination belongs to.
grade	LC assigned loan grade
grade	LC assigned loan grade
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
int_rate	Interest Rate on the loan
issue_d	The month which the loan was funded
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan
mo_sin_old_rev_tl_op	Months since oldest revolving account opened
mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened
mo_sin_rcnt_tl	Months since most recent account opened
mort_acc	Number of mortgage accounts.
mths_since_recent_bc	Months since most recent bankcard account opened.
num_accts_ever_120_pd	Number of accounts ever 120 or more days past due
num_actv_bc_tl	Number of currently active bankcard accounts
num_actv_rev_tl	Number of currently active revolving trades
num_bc_sats	Number of satisfactory bankcard accounts
num_bc_tl	Number of bankcard accounts
num_il_tl	Number of installment accounts
num_op_rev_tl	Number of open revolving accounts
num_rev_accts	Number of revolving accounts
num_tl_30dpd	Number of accounts currently 30 days past due (updated in past 2 months)
num_tl_90g_dpd_24m	Number of accounts 90 or more days past due in last 24 months
num_tl_op_past_12m	Number of accounts opened in past 12 months
open_acc	The number of open credit lines in the borrower's credit file.
pct_tl_nvr_dlq	Percent of trades never delinquent
pub_rec	Number of derogatory public records
pub_rec_bankruptcies	Number of public record bankruptcies
purpose	A category provided by the borrower for the loan request.
recoveries	post charge off gross recovery
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub_grade	LC assigned loan subgrade
tax_liens	Number of tax liens
tot_coll_amt	Total collection amounts ever owed
tot_hi_cred_lim	Total high credit/credit limit
total_acc	The total number of credit lines currently in the borrower's credit file
total_bal_ex_mort	Total credit balance excluding mortgage
total_bc_limit	Total bankcard high credit/credit limit
total_il_high_credit_limit	Total installment high credit/credit limit
total_rev_hi_lim	Total revolving high credit/credit limit
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.

DATASET

BALANCED_Data_Predicting_Default.csv

Paper ID: SR231114083515 DOI: https://dx.doi.org/10.21275/SR231114083515