Credit Risk Prediction for Peer-To-Peer Lending Platforms: An Explainable Machine Learning Approach

[1*]Chong Pei Swee, [2]Farid Meziane, and [3]Jane Labadin

[1, 3]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[2]School of Computing and Engineering, University of Derby, UK

email: [1*]chongpeiswee@gmail.com, [2]F.Meziane@derby.ac.uk, [3]ljane@unimas.my

**Corresponding author*

Received: 23 Jun 2022 | Accepted: 30 August 2022 | Early access: 19 September 2022

Abstract - *Small and medium enterprises face the challenge of obtaining start-up fund due to the strict rules and conditions set by banks and financial institutions. The plight yields to the growth in popularity of online peer-to-peer lending platforms which are an easier way to obtain loan as they have fewer rigid rules. However, high flexibility of loan funding in peer-to-peer lending comes with high default probability of loan funded to high-risk start-ups. An efficient model for evaluating credit risk of borrowers in peer-to-peer lending platforms is important to encourage investors to fund loans and justify the rejection of unsuccessful applications to satisfy financial regulators and increase transparency. This paper presents a supervised machine learning model with logistic regression to address this issue and predicts the probability of default of a loan funded to borrowers through peer-to-peer lending platforms. In addition, factors that affect the credit levels of borrowers are identified and discussed. The research shows that the most important features that affect probability of default are debt-to-income ratio, number of mortgage account, and Fair, Isaac and Company Score.*

Keywords: Credit Risk Evaluation, Peer-to-Peer Lending, Logistic Regression; Explainable Machine Learning; Explainable AI.

1 Introduction

Peer-To-Peer (P2P) lending platforms are online services provided by financial institutions as an intermediary to initiate loans for private individuals (Bachmann et al., 2011). Loans for borrowers are funded by multiple investors, bound with agreed-upon terms and conditions, with profits generated from the interest made on the loans as the borrowers are given a certain duration to pay back the loan and interest. The higher the investment risk, the higher is the interest rate. Due to a reduction in loans to small businesses from banks, P2P lending has gained popularity for personal, small business start-ups and SMEs loans as these tend to have high failing rate to pay back their loans and with low credit scores. Indeed, P2P lending allows individuals and businesses to loan money directly from investors or lenders without going through the strict requirements and criteria of traditional banks and financial institutions. Although these platforms provide several instruments to assess and limit credit risks, they do not guarantee the repayment of loans (Meyer, 2007).

The most common credit score for risks assessment is the "Fair, Isaac and Company" (FICO) score. The FICO score is not suitable for P2P lending since these platforms are meant for relatively high-risk start-ups, and for those that failed to secure loans from banks due to their low credit scores. Small and medium-sized enterprises (SMEs) which are categorized as high-risk client by financial institution play an important role in many economies, and to encourage their growth, a reliable and accurate clients' credit risk evaluation is critical to build confidence among investors so that more funds are available on P2P lending platforms. This paper presents a supervised machine learning model that predicts the probability of default by considering more information related to the clients rather than just evaluating their credit score using FICO. The focus will be on solving the credit

evaluation problem for P2P lending marketplace and determine important features that contribute to the probability of default.

2 Literature Review

P2P lending has become an alternative to obtain loans from traditional financial institutions. Most of the middle-income population lost their creditworthiness as borrowers to obtain loans from traditional financial institutions after the financial crisis in 2008, causing P2P lending became the choice for getting a loan for many individuals (Namvar, 2013). According to Emekter et al. (2014), the lack of a physical contact between lenders and borrowers in an online P2P lending process has posed the problem of information asymmetry between lenders and borrowers. Hence, having an efficient and accurate credit risk evaluation method to decrease the investment risk without human intervention is critical to sustain the steady development of the P2P lending industry.

Setiawan et al. (2019) developed a P2P lending default loan classification model using data acquired from the Lending Club through the application of Extremely Randomised Tree (ERT) and RF methods and optimised their performance with Binary Particle Swarm Optimisation (BPSO) and SVM during the feature selection. BPSO is the binary version for particle swarm optimisation (PSO), a branch of swarm intelligence, that iteratively optimises the candidate solution by guiding it towards best known position and thus finally reaching to the best solution. The evaluation of the models revealed that the average performance of ERT can outperform RF.

Emekter et al. (2014) carried out a binary Logistic Regression model for classifying default and non-default loans. The forward stepwise iterative maximum likelihood method was implemented to determine variables that have strong influence on the model and was analysed by backward stepwise of iterative maximum likelihood method. Research stated that higher credit grade is associated with lower default risk. The researchers further evaluated the selection bias by taking two different population samples, one contains data of the United States national consumers, and another contains data of Lending Club consumers. Insignificant difference of default probability for two sample indicates the consideration of data beyond the Lending Club platform is unnecessary.

High-dimensionality and imbalance class of dataset from P2P lending platform is always the challenge for making accurate prediction of default probability. In research conducted by Zhou et al. (2019), gradient boosting decision trees (GBDT), extreme gradient boosting (XGBoost), and light gradient boosting machine (LightGBM) were integrated with heterogeneous ensemble learning technology to address the issue. The ensemble model of GBDT, XGBoost and LightGBM outperformed individual classifiers of their own, proving the ability of ensemble learning model to optimize prediction from a high dimension and imbalance dataset.

Dong et al. (2010) applied the logistic regression model with random coefficients (LRR) to develop credit scorecard. A dataset with 1000 samples was divided into 10 subsets with 9 of the subsets used as training sets while the remaining subset as the testing set. The random coefficients for 900 samples are generated using Gibbs sampling within the Bayesian inference starting with estimated coefficient of logistic regression with fixed coefficients (LRF). They performed empirical experiment to evaluate the prediction accuracy of LRF and LRR with Percent Correctly Classified (PCC) method. The LRR has the overall accuracy of 74% which outperform LRF with only 71% of overall accuracy. Dong et al. argue that the logistic regression is an optimal solution for credit scoring model for financial industry in term of result interpretability.

Wang et al. (2015) had implemented lasso-logistic regression ensemble (LLRE) learning algorithm to predict default probability based on a large imbalanced dataset. Researchers clustered the majority data into sub-groups based on variables similarity and applied bagging method to minority data. Weighted average was computed for aggregation of the ensemble model. Wang et. al. created the generated variables from the original variables by partitioning them into specific intervals. The generated variables successfully reduced noise and non-linearity, thus improving the performance of the Lasso-logistic regression model. LLRE outperforms all the compared models (LLR, RF and the Classification and Regression Tree (CART)) in modelling imbalanced large dataset with significantly higher average AUC value.

Coenen et al. (2021) evaluates performance of machine learning methods from different families, namely the generalized linear models, support vector machines and gradient-boosted trees, under the context of spot factoring. They estimated the risk for spot factoring in terms of payment overdue using the machine learning methods mentioned earlier to achieve three tasks namely: binary classification of probability of default, prediction of days of overdue, and risk ranking with pre-defined labels. They found that the regression method shows higher consistency in getting high scores among all the method families in all the evaluation tasks.

The interpretability of the model is the major concern for financial institution since they are asked to provide evidence and reason for rejecting loan applications. Due to the regulation and transparency with regards to loan applications, "black-box" machine learning models (e.g., deep learning, tree-based model, and SVM) may not be a suitable approach for predicting the credit risk of borrowers. However, the logistic regression model provides good transparency on the relationship between predictors and the process of decision making. It is easier for the financial institution to interpret contributing factors to the default probability. An extension from default probability prediction, the dynamic behavioural scoring model, which predicts when the borrowers are likely to default (Wang et al., 2018), an advantage over classifications into default and non-default loans only. The logistic regression model is capable to provide probability outcomes to indicate the degree of influence from the variables on the loan default probability. Table 1 shows the summary of papers reviewed in this paper.

Table 1: Literature review summary.

Authors	Machine Learning Model	Summary
Setiawan et al. (2019)	Extremely Randomised Tree (ERT) and Random Forest (RF)	The evaluation of the models using Lending Club data revealed that the average performance of ERT can outperform RF.
Emekter et al. (2014)	Binary Logistic Regression	Selection bias evaluation with data of the United States national loan consumers, and data of Lending Club consumers shows insignificant difference of default probability. Consideration of data beyond the Lending Club platform is unnecessary.
Zhou et al. (2019)	Gradient Boosting Decision Trees (GBDT) with heterogeneous ensemble learning technology	Ensemble learning model optimized prediction from a high dimension and imbalance P2P lending platform dataset.
Dong et al. (2010)	Logistic Regression with Random Coefficients (LRR) and Fixed Coefficients (LRF)	The LRR outperformed LRF with higher overall accuracy. Logistic regression is an optimal solution for credit scoring model for financial industry in term of result interpretability.
Wang et al. (2015)	Lasso-logistic Regression Ensemble Learning algorithm (LLRE)	LLRE outperforms all the compared models of other families to predict default probability from imbalanced large dataset.
Coenen et al. (2021)	Generalized Linear Models, Support Vector Machines and Gradient-boosted Trees	Regression method shows higher consistency in getting high scores among all the method families in all the evaluation tasks.

3 Methodology

3.1 Model Formulation

Evident from our literature review, the logistic regression method is used of which the model equates the logit transform, the log-odds of the probability of a success, to the linear component as formulated in equation 1.

$$ln(\frac{p}{p-1}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k, \quad (1)$$

where p is the probability of loan to default and thus, 1-p is the probability of non-default loan occurred. The hypothesis function, is defined as: $h_\beta(x) = P(Y = 1/x\,;\beta)$, representing the predicted probability of loan, Y, to default corresponding to the loan information, x, as the independent variable and parametrised by β. In supervised learning, Y represents the label column with the value 1, representing a default loan and 0 indicating a non-default loan. Here, β represents the coefficients corresponding to each feature for fitting the model.

By rearranging equation (1), an expression for p is thus obtained as in equation (2):

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots+\beta_k x_k)}} \quad (2)$$

Equating $h_\beta(x)$ and p, then our hypothesis function is simplified to:

$$h_\beta(x) = \frac{1}{1+e^{\beta^T x}} \quad (3)$$

where β^T is the vector of coefficients corresponding to the independent variables x. The parameters β, of a logistics

regression function, were estimated using the maximum likelihood method. Employing the likelihood function:

$$L(\beta; y|x) = \prod_{i=1}^{n} (\frac{\pi_i}{1-\pi_i})^{y_i} \cdot (1-\pi_i)^{n_i} 1 \quad (4)$$

where,

$$\pi_i = \frac{e^{\sum_{k=0}^{K} x_k^{(i)} \beta_k}}{1+e^{\sum_{k=0}^{K} x_k^{(i)} \beta_i}},$$

leading to $L(\beta; y \mid x)$ to be

$$\prod_{i=1}^{n} (e^{y_i \sum_{k=0}^{K} x_k^{(i)} \beta_k}) \cdot (1 + e^{\sum_{k=0}^{K} x_k^{(i)} \beta_k})^{-n_i} .$$

Taking the logarithm of the likelihood function, resulting in

$$l(\beta) = \sum_{i=1}^{N} y_i (\sum_{k=0}^{K} x_k^{(i)} \beta_k) - n_i \cdot log(1 + e^{\sum_{k=0}^{K} x_k^{(i)} \beta_k}). \quad (5)$$

To determine the critical point of the likelihood function, the partial derivative of the likelihood function with respect to each β_k, where $k=1, 2, 3, \ldots, K$, are found and set them equal to zero. To simplify the process of finding derivative, the logarithm of likelihood function, $l(\beta)$ given in equation (5), is used.

3.2 Data

The dataset used is provided by the Lending Club, a peer-to-peer lending company from the United States of America (USA). Dataset was made available on Kaggle, an online community of data scientists and machine learning practitioners, by George (2018). The dataset contains approved loan records starting from year 2007 until the year 2018. There is total of 2,260,701 records with 151 columns, each record labelled with corresponding loan status which are 'Fully Paid', 'Current', 'Charged Off', 'In Grace Period', 'Late (31-120 days)', 'Late (16-30 days)', 'Default', 'Does not meet the credit policy. Status: Fully Paid' and 'Does not meet the credit policy. Status: Charged Off'.

3.3 Model Design

Our machine learning process follows the flow depicted in Figure 1.

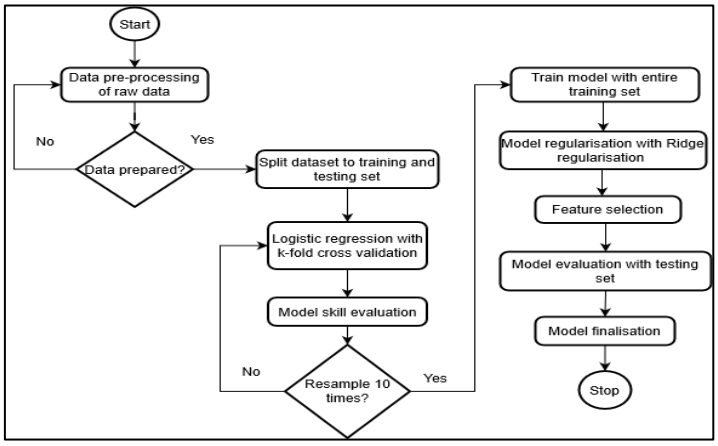

Figure 1: Machine learning process flow.

As part of data pre-processing, columns with more than 49% of missing data were removed from the dataset. Data rows with missing value are labelled as default record (minority label) and the missing data is imputed with mean, median or mode. In addition, columns which cause data leakage and column of biographical data are also removed. Outliers are detected using box plot and are removed from the dataset. Data rows with loan status labelled as 'current' which do not indicate final status of loan are also removed. The labels are grouped into default or non-default, with the values 1 and 0 respectively. 1, 345, 350 records with 25 columns were the size of the dataset used in the experiments. Refer to Table A in appendix section for the description of columns selected. Once the dataset was pre-processed then it was split in the ratio of 80% for training and 20% for testing. Stratified sampling was implemented to ensure the default and non-default records were distributed evenly. There were 268,599 (19.96%) default records and 1,076,751 (80.04%) non-default records. Under-sampling is applied to majority class, the non-

default class with NearMiss-3 algorithm. M number of closest majority samples for each minority samples are kept. Then, majority samples with the largest average distance to k nearest-neighbours (the minority samples) are selected. NearMiss-3 ensures each default sample is surrounded with non-default samples and those samples which are more distinct are kept for model fitting. In model training, the 10-fold cross validation is implemented for finding the best tuned parameters. The 80% of training dataset is partitioned into 10 blocks, and validation is iterated for 10 times. For each iteration, 9 blocks are used as training sets and the remaining block is held out from training and used as a test set. Data resampling in each iteration is done by randomly choosing 9 data blocks from the original dataset and are combined into a training set for cross validation, while the remaining one block is used as testing set.

For feature selection, a null hypothesis, h_{null}, is used that stated that there is no relationship between the independent variables and the dependent variable with 0.05 significance level. A significant level of 0.05 is chosen based on the recommendation from Fisher (2022) where it is approximately twice the standard deviations from the mean of normal distribution. Coefficients with *p*-values of more than 0.05 falling within the confidence level are eliminate with backward elimination where significant test of the independent variables started with the full model. Least significant variable is removed from the model until only the remaining independent variables that have significant contribution to the dependent variable are left. This method can show the joint behaviour of all variables in a full model, thus avoiding removal of variable which is less significant when it is include independently into the model (Chowdhury & Turin, 2020). To avoid overfitting of the model, L2 regularisation (also known as Ridge regularisation) was adopted. Ridge is a method which shrinks the weight of less important coefficient towards zero without reaching the value zero.

Models trained using the best features selected is evaluated by plotting the Receiver Operating Characteristics (ROC) curve. Recall, precision, and F1-score for default loan (minority class) are used to evaluate the model performance and for fine-tuning decision threshold which gives the best model performance. Recall measures the fraction of correctly classified positive sample (true positive). Precision measures the fraction of correct predictions made among all the positive predictions. However, recall and precision are trade-off whereby the increase of recall causes decreases in precision and vice versa. Therefore, F-measure or F1-score is used to measure the harmonic mean of precision and recall. Logistic regression model with the highest F1-score will be chosen and used in the model finalisation stage. The evaluation of the model is done using unseen data.

4 Implementation and Testing

In this section, Pearson correlation is applied to analyse the correlation between numerical features and remove features which causes multicollinearity. Outliers for each numerical feature are removed based on the upper and lower inner fences of the data distribution. Categorical features are encoded and transformed into dummy variables. Missing values in the dataset are imputed with the corresponding median in the testing dataset. Medians for imputing the missing values in both training and testing were computed from the training set alone to avoid data leakage from the isolated testing set which causes the predictive model to know information of unseen dataset. Standardisation is applied to all numerical columns in the dataset using the formula given in equation (6),

$$X' = \frac{X-\mu}{\sigma} \quad . \tag{6}$$

Rescaling the features using standardisation allows fair comparison of impacts of independent variables on the dependent variable based on weight of coefficients. Table 2 shows the mean and variance for standardising each feature.

Table 2: Mean and variance for feature scaling.

Features	Mean (5 decimal place)	Variance (5 decimal place)
loan_amt	14333.05653	69883770.0
annual_inc_log	4.81408	0.04229
dti	18.29155	67.23525
pub_rec	0.20219	0.25033
revol_bal_log	4.04207	0.12949
revol_util	54.71481	518.36330
mo_sin_old_il_acct	123.36687	1750.75500
mo_sin_old_rev_tl_op	167.76806	5898.07600
mort_acc	1.52570	3.037950
num_rev_accts	13.84327	44.18611
FICO_mean	694.16450	676.26580

The full dataset of 1,060,604 records and 47 columns (with dummy variables created from categorical variables) was split in stratified fashion with respect to the label column (dependent variable) where 80% was taken as the training dataset and the 20% was the testing dataset. Stratify splitting ensures both training and testing set have the same ratio of default and non-default records. The distribution of records grouped with loan status is shown in Table 3. Table 4 shows the distribution of under-sampled records using NearMiss-3.

Table 3: Number of records in training and testing dataset grouped with loan status.

Datasets	**Training**	**Testing**
Non-default	679,397	169,850
Default	169,086	42,271

Table 4: Number of records in training and testing dataset grouped with loan status.

Datasets	**Training**	**Testing**
Non-default	169,086	169,850
Default	169,086	42,271

Receiver-Operator-Characteristic (ROC) area-under-the-curve (AUC) was used as test score in the cross-validation. Table 5 and Table 6 show the performance of the models fitted to imbalanced and under-sampled balance dataset in 10-fold cross validation.

Table 5: Logistic regression model 10-fold cross-validation performance with Ridge regularisation of $10^{-5} \leq \lambda \leq 10^{-2}$

Magnitude of Penalty Term	**Mean Model Fitting Time (s)**	**Mean ROC AUC**	**Performance Ranking**
0.00001	13.4482	0.6974	13
0.0000177828	7.9149	0.7002	10
0.0000316228	6.5548	0.7009	9
0.0000562341	7.0857	0.7043	6
0.0001	6.7144	0.7051	4
0.000177828	6.7424	0.7057	2
0.000316228	6.2994	0.7058	1
0.000562341	6.3834	0.7054	3
0.001	6.2227	0.7045	5
0.00177828	6.2059	0.7031	7
0.00316228	5.9686	0.7014	8
0.00562341	5.7796	0.6997	11
0.01	4.7390	0.6980	12

In Table 5, logistic regression model with λ=0.000316228 (actual value is 0.00031622776601683794) has the best ROC AUC score of 0.7058.

Table 6: Logistic regression model 10-fold cross-validation performance with Ridge regularisation of $10^{-5} \leq \lambda \leq 10^{-2}$ using NearMiss-3 under sampled training dataset.

Magnitude of Penalty Term	**Mean Model Fitting Time (s)**	**Mean ROC AUC**	**Performance Ranking**
0.00001	6.3202	0.6719	11
0.0000177828	5.0328	0.6704	13
0.0000316228	4.0227	0.6715	12
0.0000562341	3.5477	0.6747	9
0.0001	3.2301	0.6733	10
0.000177828	2.9689	0.6747	8
0.000316228	2.7428	0.6771	6
0.000562341	2.4805	0.6761	7
0.001	2.3890	0.6779	3
0.00177828	2.4373	0.6781	1
0.00316228	2.5117	0.6779	2
0.00562341	2.3430	0.6777	4
0.01	2.1376	0.6774	5

In Table 6, logistic regression model with λ=0.00177828 (actual value is 0.0017782794100389228) has the best ROC AUC score of 0.6781. L2 logistic regression model with the best performance penalty term, λ, is fitted to complete imbalanced and balanced training sets. Logistic regression model is fitted to imbalanced training dataset with λ=0.00031622776601683794 and is fitted to under-sampled training data with λ=0.0017782794100389228.

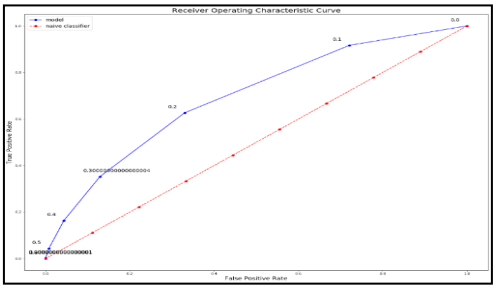

Figure 3: ROC AUC plot for logistic regression classifier of imbalanced dataset with threshold labels.

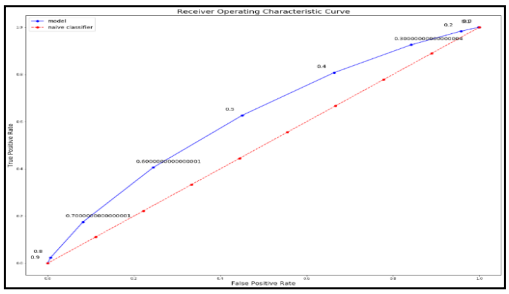

Figure 4: ROC AUC plot for logistic regression classifier of under-sampled dataset with threshold labels.

ROC curve shown in Figure 3 indicates that the threshold of 0.2 give reasonable classification result with high TPR but relatively low FPR for the model trained with imbalanced dataset. Figure 4 shows that threshold of 0.5 is the best threshold for model fitted to balanced dataset. To further evaluate the choice of suitable decision threshold, precision, recall and F1-score for default loan classification at each threshold was computed. F1-score is used to find the harmonic mean of recall and precision. F1-score ranges from 0.0 to 1.0 where 1.0 for perfect recall and precision. Figure 5 and Figure 6 show the changes of precision and recall against decision thresholds for model fitted to imbalanced and balanced dataset.

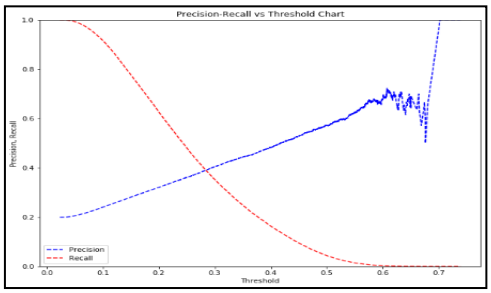

Figure 5: Default class's precision-recall curves of model trained with imbalanced data.

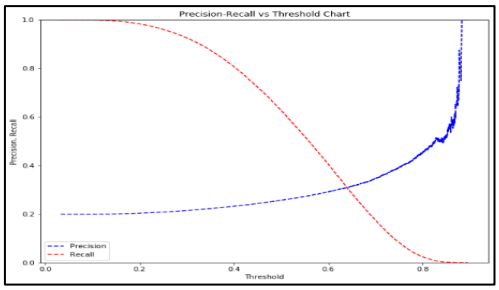

Figure 6: Precision-recall curves of model trained with under-sampled balance data.

High recall can trade off precision, therefore F1-score is used to seek balance between recall and precision. Thus, threshold which gives the highest F1-score is preferred.

Table 7: Default class's precision, recall, and F1-score of model trained with imbalanced data.

Threshold	Precision	Recall	F1-score
0.0	0.19928	1.00000	0.33233
0.1	0.24036	0.91649	0.38084
0.2	0.32084	0.62653	0.42437
0.3	0.40435	0.35242	0.37660
0.4	0.48275	0.16349	0.24426
0.5	0.57200	0.04369	0.08119
0.6	0.68000	0.00282	0.00561
0.7	1.00000	0.000047	0.000095

Result in Table 7 shows that logistic regression model fitted to imbalanced dataset gives the highest F1-score of 0.42437 with precision of 0.32084 and recall of 0.62653 at threshold of 0.2.

Table 8: Default class's precision, recall, and F1-score of model trained with under-sampled balance data.

Threshold	Precision	Recall	F1-score
0.0	0.19928	1.00000	0.33233
0.1	0.19960	0.99924	0.33274
0.2	0.20360	0.98306	0.33734
0.3	0.21480	0.92560	0.34868
0.4	0.23251	0.80736	0.36104
0.5	0.25702	0.62698	0.36458
0.6	0.29215	0.40586	0.33974
0.7	0.34672	0.17506	0.23265
0.8	0.45281	0.02418	0.04590

From Table 8, logistic regression model fitted to balanced dataset gives the highest default class's F1-score of 0.36458. The respective default class's precision is 0.25702 and 0.62698 for recall at threshold of 0.5.

5 Results and Discussion

Area under the Receiver-Operator-Characteristic (ROC) curve measures the ability of the classification model to distinguish between two classes. The larger the area under the curve (AUC), the better the proposed model to distinguish between the classes. The baseline of ROC curve is a straight diagonal line with AUC = 0.5, indicating a random classifier which makes a random guess on the distinction between the two classes.

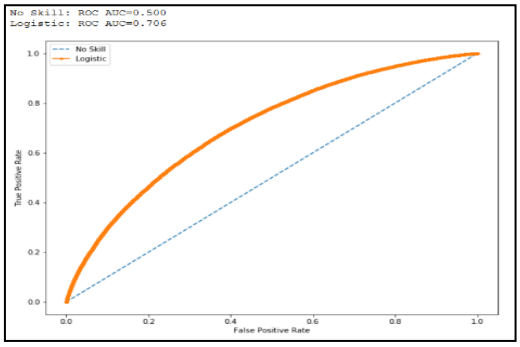

Figure 7: ROC AUC plot of model trained with imbalanced dataset.

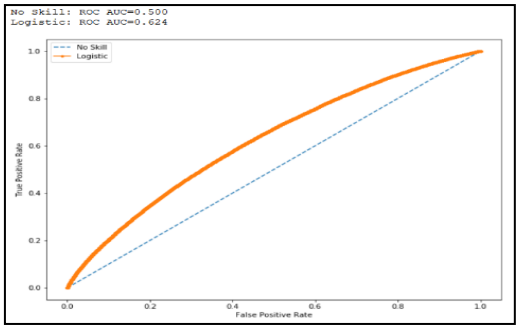

Figure 8: ROC AUC plot of model trained with under-sampled balance dataset.

Figure 7 depicts the model fitted to imbalanced dataset giving ROC AUC value of 0.706 whereas Figure 8 shows the model fitted to balanced dataset giving the value of ROC AUC to be 0.624. The difference in the ROC AUC values indicates that the model fitted to imbalanced dataset has better capability to differentiate between the default and non-default classes than the model fitted to balanced dataset. However, ROC AUC measures the overall classification performance of the model without considering the effect of majority class which cause the algorithm to be bias towards the non-default class. Due to the large skewed class distribution, ROC may give over-promising evaluation on an algorithm performance (Davis & Goadrich, 2006).

Precision-recall curve (PRC) is a better alternative of ROC for evaluating the performance of binary classifier on an imbalanced dataset. Unlike fixed baseline of ROC, baseline of PRC changes with the ratio of positive (P) and negative (N) class in the dataset. PRC baseline is defined as y = P/ (P+N) and AUC of no-skill classifier is identical to y position of PRC baseline (Saito & Rehmsmeier, 2015).

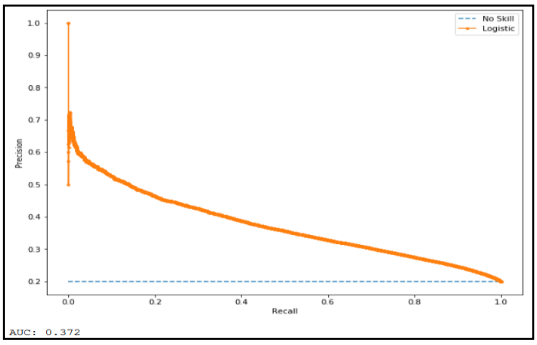

Figure 9: Precision-recall curve of model fitted to imbalanced dataset.

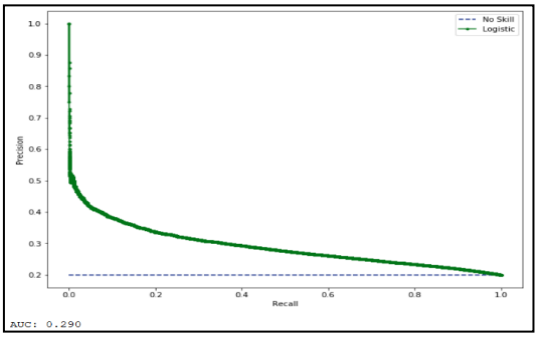

Figure 10: Precision-recall curve of model fitted to balanced dataset.

Based on the results shown in Figure 9 and Figure 10, AUC of no-skill classifier is found to be 0.2. Both models have AUC larger than the no-skill classifier which indicates they are not random classifier. Model fitted to the imbalanced dataset has larger AUC of 0.372 than model fitted to balanced dataset with AUC of 0.290. Therefore, the model fitted to an imbalanced dataset outperforms the model fitted to balanced dataset in distinguishing between two classes.

A total of 212,121 samples of the Lending Club loan records from isolated testing dataset were used to make predictions using two logistic regression models where one model is fitted to imbalanced dataset and the other one fitted to balanced dataset. The testing set contains 169,850 non-default samples and 42,271 default samples.

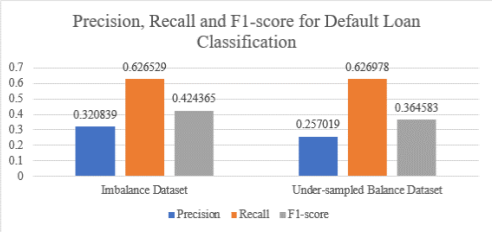

Figure 11: Comparison of default class's precision, recall, and F1-score of logistic regression models trained on imbalanced and under-sampled balance training dataset.

Based on the histogram shown in Figure 11, recalls of both models do not have significant difference where 0.626529 was found for imbalanced dataset and 0.626978 for balanced dataset. However, the model trained with imbalanced dataset has both higher F1-score and precision than model trained with balanced dataset. The result shows that NearMiss-3 under-sampling method does not improve the model performance on classifying default and non-default loan. NearMiss-3 ensures positive and negative samples with significant difference are selected while allowing positive samples to be surrounded by some majority samples. In exchange of keeping positive samples surrounded by negative samples, overlapping of both classes occurs causing a decrement for positive class precision. Precision evaluates fraction of exactly positive samples which are correctly classified as positive. Although high recall allows classification model to become more sensitive to positive class, high precision is important for avoiding misclassification of non-default loan and good clients. Hence, in comparison of the two models, it is found that the model trained with imbalanced dataset has better performance evident from the higher precision obtained and the evaluation of the F1-score. In feature selection, the logistic regression model fitted to the imbalanced dataset is employed. The model has 47 independent variables with a constant variable, which is the model intercept. Hypothesis testing with *p*-value computed from *t*-test is implemented to select statistically significant features. Defining a null hypothesis, H_0, of which the feature is insignificant to the default probability of client, tested with p-values of the feature at significance level, α of 0.05. Backward elimination was implemented for feature selection where the most insignificant feature is removed, and model is retrained with the remaining features before the next significance test is carried out. The steps are repeated until no insignificant features are left.

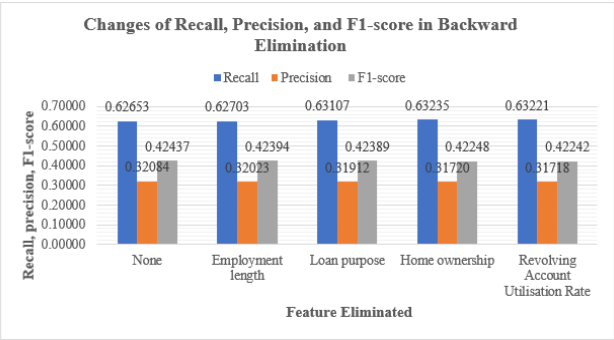

Figure 12: Changes of Recall, Precision, and F1-score in Backward Elimination.

Figure 12 shows positive class' recall increased from 0.62653 to 0.63235 after the elimination of insignificant categorical features. However, the precision decreases from 0.32084 to 0.31720 and the F1-score decreases from 0.42437 to 0.42248. Based on the three evaluation criteria, it was found that none have shown significant changes, and thus, further support the hypothesis test for which employment length, loan purpose and home ownership are insignificant to the probability of client to default in loan. The elimination of insignificant categorical features has revealed that revolving account utilisation rate, "revol_util", is insignificant with p-value of 0.866 which is larger than 0.05 significance level. Elimination of revolving account utilisation rate from the model causes the recall to drop from 0.63235 to 0.63221. There is also insignificant drop for precision and F1-score which changes from 0.31720 to 0.31718 and 0.42248 to 0.42242 respectively. Upon the implementation of the feature selection with backward elimination, 19 features were selected.

In model finalisation phase, all data samples available are used for model fitting including those from testing set which is isolated from model fitting previously. A list of new coefficients correspond to each feature is obtained. Table 9 shows the coefficients obtained by fitting the logistic regression model to the full dataset.

Table 9: Coefficients of logistic regression model fitted to complete dataset of Lending Club loan record from year 2007 until the fourth quarter of year 2018.

Features	**Coefficients, β**	**Exp(β)**
Intercept	-2.3733	0.0932
Loan amount	0.1592	1.1726
Log-transformed borrowers' annual income	-0.0730	0.9296
Debt to income ratio	0.1942	1.2143
Public derogatory records	0.0160	1.0161
Log-transformed total credit revolving balance	-0.1002	0.9047
Months since oldest bank instalment account opened	-0.0271	0.9733
Months since oldest revolving account opened	-0.0491	0.9521
Number of mortgage accounts	-0.1919	0.8254
Number of revolving accounts	0.0435	1.0445
Mean FICO score	-0.1782	0.8368
64 months payment term	0.5531	1.7386
Credit rating: Grade B	0.3205	1.3778
Credit rating: Grade C	0.7255	2.0658
Credit rating: Grade D	0.9880	2.6859
Credit rating: Grade E	1.1781	3.2482
Credit rating: Grade F	1.2641	3.5399
Credit rating: Grade G	1.2147	3.3693
Income source verified by borrower's employer	0.1353	1.1449
Income source verified by Lending Club	0.1041	1.1097
Joint application	-0.0811	0.9221

The discourse on features is separated into two parts which are discussion on continuous numerical features and the other on categorical features represented in dummy variables. All continuous numerical features are standardised to obtain standardised regression coefficients which allow comparison of absolute values to determine their relative importance in the logistic regression model. According to the absolute value of standardised coefficients, the top three most important numerical features are found to be debt to income ratio, total number of mortgage account and FICO score.

Debt-to-income ratio (DTI) with coefficient of 0.1942 has positive relation with the dependent variable which indicates that the higher the DTI, the higher the chance of borrower to default on loan. DTI is the ratio calculated by dividing monthly debt obligation with monthly gross income. Therefore, DTI reflects the ability of borrower to secure a loan whereby high DTI indicates the borrower is less likely to afford extra debt with the current income. Standardised coefficient of the total number of mortgage account is -0.1919 which defines borrower with more mortgages has lower loan default rate. Mortgage is a secured loan with real asset as collateral, and the evaluation on the ability of applicant to afford the real asset is used for mortgage application from financial institution, as it is suspected that borrower with several mortgage indicates that their credit records are good enough to fulfil the requirements of getting the mortgage loan. This explains the research outcome that the borrower with more mortgage account has lower probability of default (POD) than those with less mortgage record. The third important feature for predicting POD is the mean FICO score. Negative coefficient of -0.1782 indicates that borrowers with high FICO score tends to pay off the loan. Formula behind FICO credit score is kept secret from customers, but there are five key factors for FICO score credit report disclosed by FICO which are payment history, account owed, credit history, credit mix, and new credit. According to a study carried out by Avery, Brevoort and Canner (2012) on the effect of credit history length on credit score of foreign-born individuals in U.S. made, short credit history has caused lower credit score in this population. The result supports the consideration of length of credit history in FICO credit report where individual with longer credit history tends to have better credit score.

Next, loan amount has positive coefficient of 0.1592 which indicates that loan of higher amount has greater POD. The larger the amount loan offered by the Lending Club, the higher the interest charged which indicates higher risk. Credit revolving balance is the next important independent variables for predicting POD with negative coefficient of -0.1002. Revolving balance is the carried forward unpaid balance after each payment cycle of a credit account. In general, higher debt owed leads to higher POD, but the occurrence of negative coefficient of credit revolving balance shows that amount of debt owed does not directly reflect the POD of a borrower. It is shown that the amount of outstanding credit balance is positively correlated to income and amount of real asset owned by an individual (Kim & Devaney, 2001). High-income population have higher credit limit, hence rising

their purchasing power and increase the revolving balance subsequently. With the same reason, it explains the finding that an increase in mortgage account implies a decrease in the probability of default, since better financial status allow a person to afford more mortgages. In fact, one can have high debt amount but with large available credit, whereas an individual who owes less debt may have less credit available or even max out credit card. Due to this reason, credit scoring model such as FICO score will consider credit utilisation rate which provide more informative debt to credit limit ratio. Number of months since oldest bank instalment account opened and number of months since the oldest revolving account was opened, have negative coefficient of -0.0271 and -0.0491 respectively. These two coefficients are complementing to the fact that individuals with longer credit history have better credit. Credit scoring models available in market always consider length of credit account since it is opened and used, as well as average age of all account owned by a borrower for credit evaluation since all this information reflect the attitude of the individual towards their credit. Among all numerical features, the number of public derogatory records is the least important predictor for POD with positive coefficient of 0.0160. Public derogatory records include tax liens, public bankruptcy record and any financial obligation which are not paid as agreed. It is reasonable that individual with more public derogatory record tends to have bad credit history resulting in loan repayment failure.

Discussion on categorical features is made by comparing how each dummy variables or category level contribute to the probability of default. To measure the contribution of reference category to POD, only one categorical variable remained in the model each time. POD of reference category is determined by intercept or constant of the model while all numeric predictors remained zero. Since the logistic regression model calculates the log-odds of loan default, equation (7) is used for the transformation to POD:

$$p = \frac{1}{1+e^{-\beta_x^T}} \tag{7}$$

Table 10: Probability of default for each category in loan payment term.

Category	**Coefficients, β**	**Exp(β)**	**Probability of Default**
Intercept (36 months)	-1.7533	0.1732	0.1476
64 months	0.8733	2.3948	0.2932

Payment term (term) feature has two level of categories namely 36 months and 64 months. 36 months payment term is the reference category and "term_code_1" indicates 64 months loan payment period. From Table 10, POD of 36-months payment period is determined by intercept value. Loan with longer payment period, which is 64 months, has higher POD of 29.32% than loan with 36 months payment term (POD = 14.76%). The major loan purpose in Lending Club is debt consolidation, which is a type of loan of combining 2 and more loans into single mortgage. Delinquency of mortgage loan is closely related to income volatility even for high-income profile (Diaz-Serrano, 2005). Since income volatility may increase over time, thus extending the loan payment period can subsequently increase the risk of loan.

Table 11: Probability of default for each category in credit grade.

Category	**Coefficients, β**	**Exp(β)**	**Probability of Default**
Intercept	-2.2680	0.1035	0.0938
Grade B	0.3743	1.4540	0.1308
Grade C	0.8880	2.4302	0.2010
Grade D	1.2103	3.3545	0.2577
Grade E	1.5213	4.5782	0.3215
Grade F	1.6699	5.3116	0.3548
Grade G	1.5878	4.8930	0.3362

According to Table 11, the reference category, Grade A has the lowest POD of 9.38% and the POD are 13.08%, 20.1%, 25.77%, 32.15%, 35.48%, and 33.62% for grade B, C, D, E, F, and G respectively. Although Grade F is having the higher POD than the worst credit rating of Grade G, but the risk of default increase as the risk goes higher. It is reasonable to conjecture that Lending Club rating system is reliable reference for other financial institution while evaluating borrower's credit.

Table 12: Probability of default for each category in income source verification status.

Category	Coefficients, β	Exp(β)	Probability of Default
Intercept	-1.6672	0.1888	0.1588
Source Verified, income source verified by borrower's employer	0.2205	1.2467	0.1905
Verified, income source verified by Lending Club	0.2542	1.2894	0.1958

According to the Table 12, loan without income source verification has the lowest POD of 15.88% while POD of loan with "source verified" and "verified" are 19.05% and 19.58% respectively. The label "income source verified" defines that Lending Club had contacted the borrower's employer to verify his or her claim on the amount of earning; "income verified" defines the situation when Lending Club verified that the earning amount claimed by the borrower is within an acceptable range. According to Lending Club's company data obtained by Bloomberg, only 35.6% of income sources for application of popular loan types are verified in 2016 (Scully, 2017). As explained by Lending Club, verification of income is not applied to initial application which already passed their screening model, and the applicant is considered by Lending Club as lower risk borrower. However, Blackburn and Vermilyea (2012) found out that misstated income from borrower is one of the major causes for default on mortgage loan. Thus, the low POD of unverified income is inappropriate to explain credit level of borrower who does not passed Lending Club screening model.

Table 13: Probability of default for each category in loan application type.

Category	Coefficients, β	Exp(β)	Probability of Default
Intercept	-1.4981	0.2236	0.1827
Joint application	-0.1387	0.8705	0.1629

Lending Club allows joint application for single loan. Lending Club considers information from one of the applicants or both as factors to decide whether to approve or reject the loan. Both co-borrowers have the obligation to pay loan payment once the loan is approved. From Table 13, POD of reference category, individual application is 18.27% which is riskier than joint application with POD of 16.29%. Joint application for loan usually offered to population with short and incomplete credit history especially to those in undeveloped region, and it is proven to outperform individual application in term of repayment performance (Zhou & Wei, 2020).

The objectives of this research are to create a machine learning model which can predict probability of default and classifies the client's based on their ability to pay the loan. In this research, loan applicant with POD higher or equal to 20% is classified as default while POD lower than 20% is classified as non-default class.

6 Conclusions and Recommendations

The objectives of this research were to create a less bias solution that not only define client's credit through FICO score, but also a comprehensive evaluation that considers other factors related to the client for predicting probability of default (POD) using machine learning model. Logistic regression model fitted to imbalanced dataset outperforms model fitted to balanced dataset. Evaluation using area under precision-recall curve validates the model built for default loan classification is not a random classifier. Decision threshold value which achieves maximum balance between model recall and precision is selected with the highest F1-score.

Top 3 important features that affect POD are debt-to-income ratio, number of mortgage account, and FICO score. High debt-to-income ratio significantly contributes to the rise of POD. Revolving balance feature provides evidence to support the fact that the amount of outstanding payment does not reflect credit of an individual. However, high buying power due to high credit limit and good financial status can cause more revolving balance owed by an individual. This suggests that the utilisation rate of credit account and debt-to-income ratio are better evaluation factors for credit risk. The research result shows that FICO score is still an important factor for credit evaluation in P2P lending platform. The number of months since oldest bank instalment account opened and the number of months since oldest revolving account opened both have negative coefficient, which support the consideration of credit history length as effective factor for credit evaluation. Both mortgage number and credit history length explained the low credit score of inexperience borrower with short credit history and the reason why financial institutions prefer to allocate more resource to experienced borrowers.

The result of our model suggests that lenders should take extra precaution while dealing with borrowers who are having more public derogatory records and offering higher amount of loan is riskier. Besides, lenders should also beware of longer loan term which can increase the risk due to uncertainty causes by borrower's income volatility.

Nevertheless, credit rating model from the Lending Club is proven to have significant contribution on determining the risk of borrower in P2P lending platform, where the lower the grade of borrower the riskier he or she is. The model suggests that income source claimed by borrowers should be further verified with their employers to avoid misstate of income source which can increase the risk. Lack of credit score such as FICO score among SME entrepreneurs also posed difficulties while applying for loan. Thus, joint application of loan with better repayment performance is suggested as an alternative to offer loan to high-risk borrowers in online P2P lending platform.

Limitation of machine learning model proposed is that the model is fitted to imbalanced dataset. This causes the decision threshold for classifying default loan is set low to 0.2 for achieving highest F1-score. Low threshold value leads to higher false positive rate and causes loss of potential excellent borrowers. More appropriate resampling method can be applied for creating balanced dataset. As suggested by Yen and Lee (n.d.), different clusters in a dataset have their own characteristic where clusters with more majority samples than minority will behave like majority class, and a cluster will pose characteristic of minority class if it has more minority class samples. Under-sampling method based on clustering can be carried out to select majority class sample which may help the machine learning algorithm to better classifying default and non-default loan. Moreover, dataset from Lending Club contains loan records range from year 2007 until the fourth quarter of year 2018 which also includes records during the 2008 US financial crisis. Thus, it is suggested to select subset of data for model training based on economic situation such as economic downturn and economic upswing.

In conclusion, logistic regression model proposed provides human-interpretable information of how borrower's information and loan type affect the probability of default of loan on online P2P lending platform. Logistic regression model ensures the transparency of decision-making for loan approval and rejection which satisfy the requirement of Central Bank of Malaysia. it is hoped that the result obtained in this research can help local P2P lending platform in Malaysia to improve their credit screening process, hence provide a reliable online financial platform for both lenders and SME entrepreneurs.

Acknowledgement

The authors wish to thank Universiti Malaysia Sarawak for the facilities provided during the running of this research project.

References

Avery, R. B., Brevoort, K. P., & Canner, G. (2012). Does Credit Scoring Produce a Disparate Impact? *Real Estate Economics, 40*. doi:10.1111/j.1540-6229.2012.00348.x

Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., & Tiburtius, P. (2011). Online Peer-to-Peer Lending – A Literature Review. *Journal of Internet Banking and Commerce, 16*(23).

Blackburn, M. L., & Vermilyea, T. (2012). The prevalence and impact of misstated incomes on mortgage loan applications. *Journal of Housing Economics*, *21*(2), 151–168. https://doi.org/10.1016/j.jhe.2012.04.003

Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, *8*(1), e000262. https://doi.org/10.1136/fmch-2019-000262

Coenen, L., Verbeke, W., & Guns, T. (2021). Machine learning methods for short-term probability of default: A comparison of classification, regression and ranking methods. *Journal of the Operational Research Society*, *73*(1), 191–206. https://doi.org/10.1080/01605682.2020.1865847

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*. https://doi.org/10.1145/1143844.1143874

Diaz-Serrano, L. (2005). Income volatility and residential mortgage delinquency across the EU. *Journal of Housing Economics*, *14*(3), 153–177. https://doi.org/10.1016/j.jhe.2005.07.003

Dong, G., Lai, K. K., & Yen, J. (2010). Credit scorecard based on logistic regression with random coefficients. *Procedia Computer Science*, *1*(1), 2463–2468. https://doi.org/10.1016/j.procs.2010.04.278

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2014). 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

Fisher, R. A. (2022b). *Statistical Methods for Research Workers, 12th Ed. Rev.* (Twelfth Edition). Oliver and Boyd.

George, N. (2018). *All Lending Club loan data 2007 through current Lending Club accepted and rejected loan data*. Kaggle. https://www.kaggle.com/wordsforthewise/lending-club?select=accepted_2007_to_2018Q4.csv.gz

Kim, H., & Devaney, S. A. (2001). The Determinants of Outstanding Balances Among Credit Card Revolvers. *Journal of Financial Counseling and Planning, 12*(1).

Meyer, T. (2007, July 10). *Online P2P lending nibbles at banks' loan business*. Retrieved from http://www.venturewoods.org/wp-content/uploads/2007/11/p2p-lending.pdf

Namvar, E. (2013). An Introduction to Peer to Peer Loans as Investments. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2227181

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Scully, M. (2017, June 14). Biggest online lenders don't always check key borrower data. Retrieved August 29, 2022, from https://www.bloomberg.com/news/articles/2017-06-14/biggest-online-lenders-don-t-always-check-key-borrower-details

Setiawan, N., Suharjito, & Diana. (2019). A Comparison of Prediction Methods for Credit Default on Peer to Peer Lending using Machine Learning. *Procedia Computer Science*, *157*, 38–45. https://doi.org/10.1016/j.procs.2019.08.139

Wang, H., Xu, Q., & Zhou, L. (2015). Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble. *PLOS ONE*, *10*(2), e0117844. https://doi.org/10.1371/journal.pone.0117844

Wang, Z., Jiang, C., Ding, Y., Lyu, X., & Liu, Y. (2018). A Novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. *Electronic Commerce Research and Applications*, *27*, 74–82. https://doi.org/10.1016/j.elerap.2017.12.006

Yen, S. J., & Lee, Y. S. (2006). Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset. *Intelligent Control and Automation*, 731–740. https://doi.org/10.1007/978-3-540-37256-1_89

Zhou, J., Li, W., Wang, J., Ding, S., & Xia, C. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. *Physica A: Statistical Mechanics and Its Applications*, *534*, 122370. https://doi.org/10.1016/j.physa.2019.122370

Zhou, Y., & Wei, X. (2020). Joint liability loans in online peer-to-peer lending. *Finance Research Letters*, *32*, 101076. https://doi.org/10.1016/j.frl.2018.12.024

Appendix

Table A: Description of attributes from Lending Club 2007 to 2018 fourth quarter approved loan dataset.

Attributes	Description	Datatype
annual_inc	The self-reported annual income provided by the borrower during registration.	float64
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers	Object
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.	float64
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.	float64
fico_range_high	Highest FICO score value.	float64
fico_range_low	lowest FICO score value.	float64
grade	LC assigned loan grade	Object
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report.	Object
int_rate	Interest Rate on the loan	float64
loan_amnt	The listed amount of the loan applied for by the borrower.	float64

loan_status	Current status of the loan	Object
mo_sin_old_il_acct	Months since oldest bank instalment account opened	float64
mo_sin_old_rev_tl_op	Months since oldest revolving account opened	float64
mort_acc	Number of mortgage accounts.	float64
num_bc_tl	Number of bankcard accounts	float64
num_rev_accts	Number of revolving accounts	float64
open_acc	The number of open credit lines in the borrower's credit file.	float64
pub_rec	Number of derogatory public records	float64
pub_rec_bankruptcies	Number of public record bankruptcies	float64
purpose	A category provided by the borrower for the loan request.	Object
revol_util	Revolving line utilisation rate, or the amount of credit the borrower is using relative to all available revolving credit.	float64
tax_liens	Number of tax liens	float64
revol_bal	Total credit revolving balance	float64
term	The number of payments on the loan. Values are in months and can be either 36 or 60.	Object
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified	Object