ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/371361738

Loan Default Prediction: A Complete Revision of LendingClub

Article *in* Revista Mexicana de Economía y Finanzas · June 2023

DOI: 10.21919/remef.v18i3.886

CITATIONS
0

READS
76

5 authors, including:

Pilar Madrazo-Lemarroy
Anáhuac University
8 PUBLICATIONS **38** CITATIONS

SEE PROFILE

All content following this page was uploaded by Pilar Madrazo-Lemarroy on 14 June 2023.

The user has requested enhancement of the downloaded file.

REVISTA MEXICANA DE ECONOMÍA Y FINANZAS
Nueva Época
REMEF
(THE MEXICAN JOURNAL OF ECONOMICS AND FINANCE)

Revista Mexicana de Economía y Finanzas, Nueva Época
Volumen 18 Número 3, Julio - Septiembre 2023, pp. 1-13, e886
DOI: https://doi.org/10.21919/remef.v18i3.886

(First Reception: November 22, 2021; Last Reception: July 25, 2022; Accepted: November 1, 2022; Published: June 6, 2023)

Loan Default Prediction: A Complete Revision of LendingClub

José Antonio Núñez Mora - Instituto Tecnológico y de Estudios Superiores de Monterrey, México
Pamela Moncayo[1] - Instituto Tecnológico y de Estudios Superiores de Monterrey, México
Carlos Franco - Instituto Tecnológico y de Estudios Superiores de Monterrey, México
Pilar Madrazo-Lemarroy - Universidad Anáhuac, México
Jaime Beltrán - Universidad Anáhuac, México

Abstract

The study aims to determine a credit default prediction model using data from LendingClub. The model estimates the effect of the influential variables on the prediction process of paid and unpaid loans. We implemented the random forest algorithm to identify the variables with the most significant influence on payment or default, addressing nine predictors related to the borrower's credit and payment background. Results confirm that the model's performance generates a F1 Macro Score that accomplishes 90% in accuracy for the evaluation sample. Contributions of this study include using the complete dataset of the entire operation of LendingClub available, to obtain transcendental variables for the classification and prediction task, which can be helpful to estimate the default in the person-to-person loan market. We can draw two important conclusions, first we confirm the Random Forest algorithm's capacity to predict binary classification problems based on performance metrics obtained and second, we denote the influence of traditional credit scoring variables on default prediction problems.

JEL Classification: C24, G23, O16.

Keywords: Random Forest, P2P lending, LendingClub, SMOTE, Fintech, Default Prediction.

Predicción del default: Una revisión completa de LendingClub

Resumen

El objetivo del estudio es determinar un modelo de predicción de default crediticio usando la base de datos de LendingClub. La metodología consiste en estimar las variables que influyen en el proceso de predicción de préstamos pagados y no pagados utilizando el algoritmo Random Forest. El algoritmo define los factores con mayor influencia sobre el pago o el impago, generando un modelo reducido a nueve predictores relacionados con el historial crediticio del prestatario y el historial de pagos dentro de la plataforma. La medición del desempeño del modelo genera un resultado F1 Macro Score con una precisión mayor al 90% de la muestra de evaluación. Las contribuciones de este estudio incluyen, el haber utilizado la base de datos completa de toda la operación de LendingClub disponible, para obtener variables trascendentales para la tarea de clasificación y predicción, que pueden ser útiles para estimar la morosidad en el mercado de préstamos de persona a persona. Podemos sacar dos conclusiones importantes, primero confirmamos la capacidad del algoritmo Random Forest para predecir problemas de clasificación binaria en base a métricas de rendimiento obtenidas y segundo, denotamos la influencia de las variables tradicionales de puntuación de crédito en los problemas de predicción por defecto.

Clasificación JEL: C24, G23, O16.

Palabras clave: Random Forest, Préstamos persona a persona, LendingClub, SMOTE, Fintech. Predicción del Default.

[1] Corresponding author. Av Carlos Lazo 100, Santa Fe, La Loma, Álvaro Obregón, 01389 Ciudad de México, CDMX. Email: A01490264@tec.mx

*No source of funding for research development.

This article is under the license CC BY-NC

1. Introduction

The fintech ecosystem is a living organism, growing and transforming along with technological development, driven by the consumers' demands for ubiquity, instantaneity, and user experience. The term fintech became utterly relevant around 2015, but the merge of finance and technology is not a novelty (Arner et al., 2015, 2016). We have experienced fintech ever since the creation of automated machine tellers, e-banking to streamline access to financial products and synergic connection between financial institutions and consumers. The Cambridge Centre for Alternative Finance (CCAF) has been tracking the fintech ecosystem since 2015, generating survey-based information about the industry´s development and valuable insights that have help as a benchmark to participants at a global level. For 2020, they collected information from 703 firms, compared to 205 European firms in 2015 and around three hundred firms in the Americas in 2016. These reports show how the growth of the fintech ecosystem represents a complex dynamic between individuals and the necessity of access to financing and liquidity.

According to the 2nd Global Alternative Finance Market Benchmarking report elaborated by the CCAF, the most representative market in the fintech ecosystem is crowdlending or peer-to-peer consumer lending (P2P). It has a global market volume of USD 3.5 billion, ahead of all other business models in the ecosystem. This business model's mechanism is about creating a virtual marketplace where investors and borrowers meet. While investors choose the loans, they will fund according to their risk appetite, the platform generates a credit scoring model for borrowers' creditworthiness assessment (Ziegler et al., 2021).

We can attribute the growth of this market to the credit access restriction subsequent to the 2007 Global Recession in the first world countries, and financial access impairs in emerging economies (Brunnermeier, 2009). These platforms seized the opportunity and developed an internet-based marketplace gathering borrowers and investors, creating a business model with high transaction rates. Likewise, the overall success of this marketplace is because it has allowed the unbanked population to participate in the P2P dynamic, achieving financial inclusion goals while constantly raising the number of financial services consumers.

The success of P2P lending markets is evident and justified. The platforms belonging to this market have forged a trusting relationship with users, offering transparent, fast, and convenient access to financing. First platforms such as ZOPA (UK based, founded in 2005), LendingClub, and Prosper (USA based, founded in 2007), inspired other countries to create more platforms. China is the best example. Since 2007 its P2P consumer lending market grew exponentially, being a significant competitor in the fintech ecosystem until 2018. China's Banking Regulatory Commission passed strict regulations for the P2P marketplace in 2016, and several platforms could not comply. Such requirements restricted platforms from pulling loans and offering credit services, so they only can act as intermediaries. The Chinese P2P lending market started to crash in 2018, facing fraud and delinquency scandals. By 2020 the CCAF omitted China from the global fintech report due to the dramatic fall in P2P lending market activities. (Stern et al., 2017).

Even though a regulatory effect caused the Chinese P2P lending market controversy, the rest of the platforms are not exempt from suffering difficulties in capital returning and loan recovery due to a more flexible regulation. Therefore, some regulatory requirements have included the platform to share the risk with investor (e.g., regulation in some countries now demand their participation in the loan funding).

Risk is the cornerstone in this mechanism and it is related, mainly, to the creditworthiness of borrowers and the platform's capacity to discern bad borrowers from good borrowers. Therefore, P2P lending platforms aim to mitigate or compensate the default risk through attractive yields that attract investors' trust. For risk mitigation, each platform develops its own credit model that determines the default probabilities and repayment capacities of potential borrowers. In order to achieve this, they usually go to historical data and design quantitative models. Access to such datasets is challenging, mainly for startups, who rely on benchmarks to construct their credit models.

This research aims to analyze a public dataset from a well-known P2P lending platform called LendingClub (https://www.lendingclub.com/) to identify a set of predictors of payment or default. Some of these predictors include credit history, loan characteristics, borrower characteristics, and the probability of default. We address other available research papers that study the LendingClub dataset in different periods and use different methodologies. In this study, we will extract the information for the whole operation of this platform (2007-2020) and analyze which of the 140 predictors available are suitable for a standard data-driven classification and prediction model. The methodology involves testing a Random Forest algorithm to identify feature importance in the default prediction.

Results show that we can develop a model with only a few variables and achieve accurate classification metrics. Therefore, from the 140 variables available, we selected nine according to feature importance provided by the Random Forest algorithm. We also address the class imbalance in the target variable with SMOTE oversampling for default observations. A marginal improvement compared to the analysis with the original class imbalance was identified, due to the Random Forest capacity to handle this condition.

This study's contribution is in the credit risk assessment for the fintech ecosystem. We evidence that credit history variables are determinants in the default prediction for the LendingClub dataset. The layout of this paper is as follows. Section 2 presents the literature review; section 3 explains the model; sections 4 and 5 describe the dataset and descriptive statistics, respectively. Results for the model are discussed in section 6, followed by section 7 for conclusions.

2. Literature Review

Research on the P2P lending market concentrates on two big groups. The first group studies social and behavioral aspects of the mechanism to identify what motivates investors to fund a specific loan and participate in a P2P lending market. Several approaches find that the borrowers' characteristics, such as their photograph (Gonzalez & Loureiro, 2014), profile, and social networks, together with a description of loan purpose, are influential factors over investors' decisions. Other set of studies relate investors' decisions to herd behavior in P2P lending market platforms, demonstrating that investors prefer to fund loans with a certain percentage of funders to share the risk with (Lee & Lee, 2012; Zhang & Liu, 2012). Further studies try to assess the information asymmetry present in these marketplaces and mitigate it to avoid adverse selection in investment (Weiss et al., 2010). Authors find the inclusion of soft information to be helpful in the mitigation of adverse selection (Tao et al., 2017). From the social perspective, P2P lending is related to financial inclusion efforts and social capital development since these platforms have enabled access to financial services for unbanked individuals, and young people with little credit history.(Hasan et al., 2020; Maskara et al., 2021)

The second group of research focuses on the business model operation and the use of technological developments such as big data software and alternative data analytics in creating credit scoring models and credit risk analysis. Platforms access information such as consumers' payment history, insurance claims, and social networks and combine these with traditional data sources such as FICO rates to generate a better assessment of creditworthiness within users (Jagtiani & Lemieux, 2019). Among this group, there are some works that suggest using psychometric and demographic variables, as well as email usage information to generate a creditworthiness assessment when credit history is unavailable, generating sufficient accuracy proof in implementing statistical classifiers using these predictors (Djeundje et al., 2021).

With the increment of mobile devices usage, several studies have included metrics generated in these devices for credit risk assessment. Variables such as call records, mobile location, applications installed, and SMS activity prove to increase accuracy in default prediction when used along with credit bureau information (Agarwal et al., 2020; Björkegren & Grissen, 2018; Óskarsdóttir et al., 2019). Soft data and user-generated text are also employed to enhance predictive models for credit risk assessment (Netzer et al., 2019). Text is processed and categorized to determine creditworthiness according to spelling error rate, length of text, upper and lower cases, readability, and tone. These variables positively impact the creation of enhanced credit risk models (Berg et al., 2020). There is also a body of literature around credit default probability and overall credit risk assessment implementing traditional econometric alternatives and AI-based algorithms. Econometric models such as binary classifications or logistic regressions are used as benchmarks for estimating probabilities as they are highly interpretable.

Most P2P lending research is developed using the LendingClub dataset, one of the few public datasets available for such tasks. The platform has quarterly information about loans (2007-2020) available for investors to help them analyze borrowers' conditions[2]. Serrano-Cinca et al. (2015) used the LendingClub dataset and, through a logit model, they assessed determinant variables in default prediction for a period of six years (2008-2014). This work identified variables related to the characteristics of the loan and the borrower's credit history information and proposed a model using loan purpose, annual income, housing situation, credit history, and indebtedness levels. In a second approach, they used logistic regression for predictive assessment, concluding that the stronger predictors are platform risk grade assignation and indebtedness levels.

Despite the transparency of a logit model, researchers have applied other methodologies that are not restricted to linear assumptions and that can handle large data volumes, outperforming the results of logit classifications and offering better insights to Big Data methods. Research on the LendingClub dataset uses Machine Learning algorithms because it does not require extensive preprocessing and can handle multicollinearity conditions. Literature shows the application of algorithms such as support vector machines, neural networks, and ensemble methods. Cho et al.

[2] LendingClub retired the P2P marketplace in the last quarter of 2020 as they are currently developing new financial products as a neo-bank.

(2019) trained an Instance-Based Entropy Fuzzy SVM algorithm to identify default probabilities in P2P lending. They proposed investment decision models to maximize the expected return on non-defaulting loans. Kin et al. (2020) also trained an ensemble of four classifiers (neural networks, random forest, adaptive boosting, and extreme gradient boosting) considering five common characteristics for credit analysis.[3] Ensemble methodologies have trained several weak estimators to yield a unified, robust estimation looking for error rate reductions and better predictive accuracy (Dietterich, 2000). Deep learning algorithms such as convolutional neural networks were trained in Chengeta and Mabika (2021) to identify default and possible frauds in P2P lending. Authors propose a model where loan purpose, employment status, and credit scorings conveniently identify possible defaulters.

Several research papers use tree-structured algorithms for borrower classification and identify potential defaulters according to the importance of the dataset features (Breiman, 2001). In contrast with logit regressions, studies find that Machine Learning algorithms such as Random Forest (RF) build better prediction models for binary and multilevel classification (e.g., Jin et al., 2015; Li and Zengyi, 2020; Ye et al., 2018; Zhu et al., 2019). Zhu et al. (2019) use the LendingClub dataset for the 2019Q1 and perform RF using fifteen attributes belonging to credit characteristics such as loan amount, installment, and grade, concluding that this algorithm outperforms SVM, Decision Tree, and logistic regression. Li and Zengyi (2020) propose a model for lenders' profit evaluation, using LendingClub to validate the model. In contrast to Zhu et al., authors found relevant variables such as in debt to income and interest rate. Jin et al. (2015) employ RF for feature selection and a posterior evaluation of other Machine Learning models. The resulting variables selected are term, annual income, loan amount, debt to income ratio, credit grade, and revolving utilization. Ye et al. (2018) develop a profit score model using RF optimization genetic algorithm to study the maximization of lender profits.

3. Model

Random forest combines multiple machine learning models to explain a wide range of data effectively. This model is an ensemble method that helps us with classification and regression problems. At the beginning of this century, random Forest appeared as an idea related to trees' natural differences, building some randomness to select variables, and voting for the most popular class (classification) or averaging the cases (regression). Breiman (2001) introduced random forest as a new predicted tool to compete with boosting and adaptive bagging.
Assuming a sample,

$$\mathcal{L}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

Where the couples (X, Y) come from a common distribution of n number of independent and identically distributed (*i.i.d.*) observations.
The coordinates of X are the input variables, such that $X \epsilon \mathcal{X}$, where $\mathcal{X}$ is a space of p dimension of the total number of variables.

[3] Character, capacity, capital, condition, and collateral in the evaluation of credit customers.

Y refers to the explained or dependent variable, such that $Y \epsilon \mathcal{Y}$, where the nature of the space $\mathcal{Y}$ relates for a regression problem as $\mathcal{Y} = \mathbb{R}$, and for a classification problem with $\mathcal{C}$ classes as $\mathcal{Y} = \{1, \dots, \mathcal{C}\}$.

So, for a learning sample $\mathcal{L}_n$ a predictor $\hat{h}$ is constructed, such that,

$$\hat{h}: \mathcal{X} \to \mathcal{Y}$$

From each prediction $\hat{y}$ of the explained variable corresponds to a given input observation $x \epsilon \mathcal{X}$. Where in classification, the focus of this paper, the probability of misclassification is $P\left(Y \neq \hat{h}(X)\right)$ with a misclassification rate of $\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{Y_i' \neq \hat{h}(Y_i')}$ such that we have a sample test $\mathcal{T}_m = \{(X_1', Y_1'), \dots, (X_m', Y_m')\}$ drawn from the distribution of (X, Y).

Therefore, let the random Forest $\left(\hat{h}(.,\Theta_1), \dots, \hat{h}(.,\Theta_q)\right)$ be a collection of tree predictors, with q *i.i.d* random variables $\Theta_1, \dots, \Theta_q$ independent of $\mathcal{L}_n$, where the random forest predictor $\widehat{h_{RF}}$ is obtained by collecting random trees aggregation. There is a majority vote among individual tree predictions in classification problems, such that $\widehat{h_{RF}}(x) = \underset{1 \leq c \leq \mathcal{C}}{\arg\max} \sum_{l=1}^{q} \mathbf{1}_{\hat{h}(x,\Theta_l)=c}$. A more detailed description of random Forest could be found in Breiman (2001) and Genuer and Poggi (2020).

4. Data

We use the LendingClub snapshot dataset for the 2007-2020Q3 period, the last dataset available since LendingClub retired the investment notes from the platform in the last quarter of 2020. This dataset is available in Kaggle, a data science repository. The original dataset contains 140 variables and 2,925,493 individual loans observations. It is noteworthy to mention that the dataset has a high percentage of missing values for several variables. The dataset has loans and borrowers' characteristics, such as their credit history, credit risk scores, and credit-issuing conditions. The loan status variable discloses the current loan repayment stage for individual borrowers, where the status is "fully paid," "charged off," "late," "in grace period," or "current."

5. Methodology

We performed data preprocessing before algorithm implementation. We eliminated variables presenting above 50% missing values and discarded individual observations that present blanks for any feature. We adopt this approach to avoid treating missing values with any strategy since this would bias the information. This dataset is not a time series; individuals do not necessarily meet the same conditions, especially regarding the variables that represent credit history. Finally, we encode categoric variables as dummies.

The dependent variable for this dataset is loan status. We select fully paid and charged off loan status to represent non-defaulted and defaulted loans. Under these conditions, we can perform binary classification algorithms to predict default probabilities based on the features presented in the LendingClub dataset.

We employ an RF algorithm with the preprocessed dataset to find the most representative variables according to the classification prediction objective. We selected a 60% -40% randomized split for training and testing subsets, respectively. We apply the feature importance approach for a dimensionality reduction of the dataset, based on how the model rates the input variables' relevance in the testing phase. Feature importance is between zero and one (0 = no influence over the target, and 1 = perfect target prediction).

Resulting features are: “recoveries”, “total_rec_prncp”, “collection_recovery_fee”, “last_fico_range_high”, “last_fico_range_low”, “last_pymnt_amnt”, “total_pymnt_inv”, “total_pymnt”, “funded_amnt”, “installment”, “loan_amnt”, “funded_amnt_inv”, “debt_settlement_flag_Y”, “debt_settlement_flag_N”, “total_rec_int”, “term”, “total_rec_late_fee”, “int_rate”, “issue_d”, “grade_A”.

We eliminate recovery-related variables because they are a trivial explanatory for defaulted or delinquent loans. LendingClub charges fees for recovery of principal and interest rate for late or no payments. Also, we delete redundant features with a correlation coefficient over 0.80. We maintain the following variables for further analysis:

Table 1. Selected features description.

Feature name	Description	Type
last_fico_range_high	The upper boundary ranges the borrower's last FICO pulled belongs to.	Numeric
last_pymnt_amnt	Last total payment amount received	Numeric
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.	Numeric
debt_settlement_flag_Y	Flags whether or not the borrower, who has charged-off, is working with a debt-settlement company.	Categoric: 0 for 'YES,' 1 for 'NO'
total_rec_int	Interest received to date	Numeric
term	The number of payments on the loan. Values are in months and can be either 36 or 60.	Numeric
int_rate	Interest Rate on the loan	Numeric
issue_d	The month which the loan was funded	Numeric
grade_A	LC assigned loan grade	Categoric: 1 for grade A, 0 for other grades

Source: From the LendingClub data dictionary

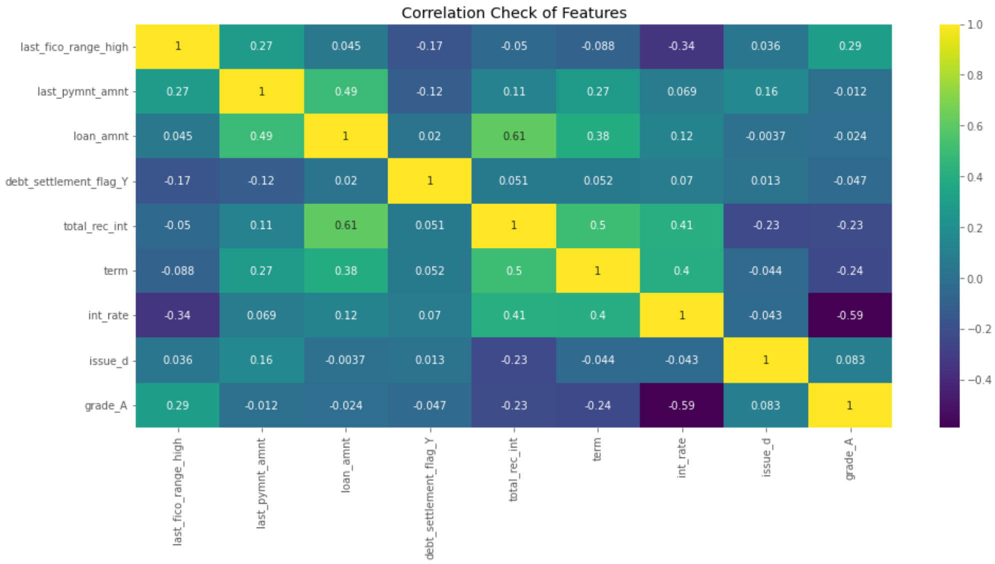

Figure 1. Correlation matrix for selected features.

Figure 1. displays the Pearson correlation matrix for the selected features. In Table 2., we present the descriptive statistics.

Table 2. Descriptive statistics for selected features.

	last_fico_ range_high	last_pymnt _amnt	loan_amnt	debt_ settlement_flag_Y	total_ rec_int	term	int_rate	issue_d	grade_A
mean	678.24	5531.04	14962.89	0.03	2579.96	42.34	13.29	2015	0.18
std	81.91	7283.07	9027.74	0.16	2809.52	10.58	4.87	1.63	0.38
min	0	-400	1000	0	0	36	5.31	2012	0
0.25	624	400.12	8000	0	796.42	36	9.75	2015	0
0.5	694	2026.8	12900	0	1651.48	36	12.74	2016	0
0.75	734	8437.4	20000	0	3293.22	60	16.02	2017	0
max	850	42192.05	40000	1	31714.37	60	30.99	2020	1

Furthermore, we use the Synthetic Minority Oversampling Technique (SMOTE) to handle the class imbalance problem present in this dataset. The number of fully paid loans is 1,121,412 and 269,193 charged-off loans (80.64% and 19.36%, respectively). We oversample the charged-off loan status, so the fully paid/charged-off ratio is 40%. Additionally, we apply 5-fold cross-validation to evaluate performance stability with F1-macro score and accuracy score.

6. Results

We performed the RF algorithm on the dataset restricted to the resulting features selected in the methodology section. We demonstrate that the number of variables in the original dataset may not be essential for a credit risk analysis. From the 140 variables available, most models only use ten to fifteen variables for class prediction, as seen in the literature. In this section, we prove that the RF algorithm yields robust results for class prediction. We trained 60% of the dataset, left the rest for testing, and performed k-fold cross-validation on train and test samples. We repeat the process for the oversampled train set to address improvements in classification metrics. Random Forest results are compared to logit classification to show the performance improvement of the RF model. Table 3 presents the cross-validation results for train and test samples.

Table 3. Cross-validation results

	5-Fold Cross-Validation						
Classifier	**F1 - Macro Score**						
	1st Fold	**2nd Fold**	**3rd Fold**	**4th Fold**	**5th Fold**	**Mean**	**Std.Dev**
RF	0.96652641	0.96583453	0.96610146	0.96632976	0.96521389	0.96600121	0.000510266
RF (SMOTE)	0.97124209	0.9718793	0.97129974	0.98113967	0.98021147	0.97515445	0.005056883
LOGIT	0.90917119	0.90903796	0.91266451	0.91066891	0.91063253	0.91043502	0.001312509

The F1-Macro Score is a classification metric that averages each class F1 Score; this metric is helpful for skewed data and measuring classification performance in class-imbalance situations because it treats all classes as equals regardless of support values. The confusion matrix is a tangible representation of the predictive performance of this algorithm; it presents the number of correctly classified and misclassified observations. In Table 4, we present the confusion matrix results for the test set, and in Table 5, we present the classification report.

Table 4. Confusion matrix

Random Forest prediction for Classification (Test set)							
Confusion Matrix		**Predicted Class**		**Confusion Matrix (SMOTE)**		**Predicted Class**	
		0	**1**			**0**	**1**
True Class	**0**	443731	4834	**True Class**	**0**	442452	6113
	1	6581	101096		**1**	5421	102256

Logit prediction for Classification (Test set)			
Confusion Matrix		**Predicted Class**	
		0	**1**
True Class	**0**	433375	15190
	1	15746	91931

Table 5. Classification report

	Classification Report			
Classifier	**Accuracy**	**F1-Score**		**H-Score**
		0	**1**	
RF	0.98	0.99	0.95	0.88
RF (SMOTE)	0.98	0.99	0.95	0.88
Logit	0.94	0.97	0.86	0.69

The H-Score proposed by Hand (2009) is a Bayesian approach that specifies a prior distribution for each class loss independent of the algorithm. This measure replaces ROC – AUC scores since they present a dependency relationship with the algorithm used (Hand, 2009; Hand & Anagnostopoulos, 2013). The H-Score allows determining a cost of misclassification as a severity ratio. In this case, we penalize misclassification symmetrically, selecting a severity ratio of one. We observe the performance dropped in both RF and RF SMOTE predictions by an average of 10% compared to the F1-Score and Accuracy metrics. We consider evaluating several metrics because we propose a model for default prediction; from the business perspective, misclassification is problematic as it leads to unnecessary or unwanted risks.

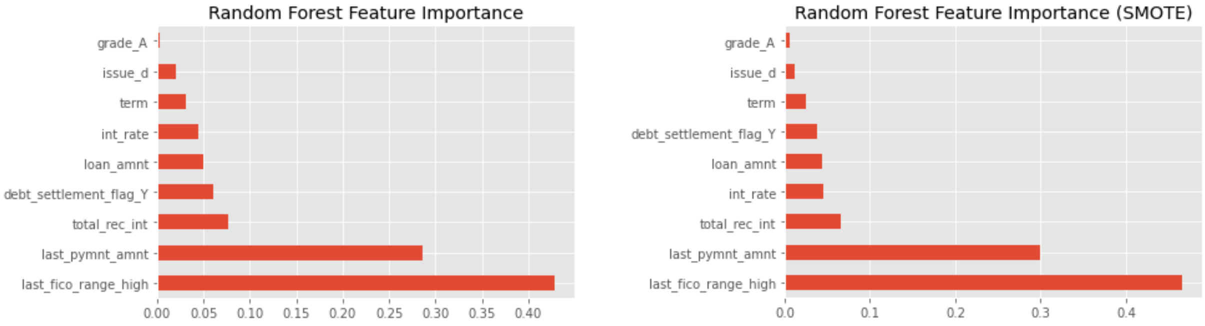

Figure 2. Feature importance for selected features (RF and RF SMOTE)

Figure 2 displays how selected features behave for RF and RF SMOTE. We observe a slight change in the debt settlement flag, interest rate, and loan amount position. This result indicates how the alternative SMOTE technique ponders the features differently, assigning more importance to interest recovery and interest rate. Nevertheless, both techniques' top three variables are consistent. FICO score, last payment amount, and total interest received to date give important insights about the loan and borrower information. FICO scores result from a proprietary algorithm for credit scoring based on credit history, while the interest received to date and the last payment amount represent the borrower's behavior in the LendingClub platform.

7. Conclusions

We can draw two important conclusions from this study. First, we confirm the Random Forest algorithm's capacity to predict binary classification problems based on performance metrics obtained. We highlight the interpretation transparency achieved using this algorithm. The feature importance result allowed us to perform a dimensionality reduction that reproduced a robust model for default prediction using only nine variables. These results can be compared to other research articles using other Machine Learning based algorithms with similar performance reports. And second, we denote the influence of traditional credit scoring variables on default prediction problems.

Therefore, P2P lending platforms still have to consider credit bureau information to assess the credit risk of potential borrowers as a principal requirement. The resulting model presented in this study is data-driven; it is not strictly conclusive for all P2P platforms. Nonetheless, LendingClub is a consolidated corporation considered as a benchmark for the P2P lending market. Startups and active platforms can benefit from this study to generate credit risk models.

References

[1] Agarwal, S., Alok, S., Ghosh, P., & Gupta, S. (2020). Financial inclusion and alternate credit scoring for the millennials: role of big data and machine learning in fintech. Business School, National University of Singapore Working Paper, SSRN, 3507827. DOI: https://doi.org/10.2139/ssrn.3507827

[2] Arner, D. W., Barberis, J., & Buckley, R. P. (2015). The evolution of Fintech: A new post-crisis paradigm. Geo. J. Int'l L., 47, 1271. DOI: https://doi.org/10.2139/ssrn.2676553

[3] Arner, D. W., Barberis, J., & Buckley, R. P. (2016). 150 years of Fintech: An evolutionary analysis. Jassa, 3, 22–29.

[4] Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of fintechs: Credit scoring using digital footprints. The Review of Financial Studies, 33(7), 2845–2897. DOI: https://doi.org/10.1093/rfs/hhz099

[5] Björkegren, D., & Grissen, D. (2018). Behavior revealed in mobile phone usage predicts loan repayment. Available at SSRN 2611775. DOI: https://doi.org/10.2139/ssrn.2611775

[6] Breiman, L. (2001). Random Forest. Machine Learning, 45(1), 5–32. DOI: https://doi.org/10.1023/A:1010933404324.

[7] Brunnermeier, M. K. (2009). Deciphering the liquidity and credit crunch 2007-2008. Journal of Economic Perspectives, 23(1), 77–100. DOI: https://doi.org/10.1257/jep.23.1.77

[8] Chengeta, K., & Mabika, E. R. (2021). Peer to Peer Social Lending Default Prediction with Convolutional Neural Networks. In S. Pudaruth & U. Singh (Eds.), 4th International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems, icABCD 2021. Institute of Electrical and Electronics Engineers Inc. DOI: https://doi.org/10.1109/icabcd51485.2021.9519309

[9] Cho, P., Chang, W., & Song, J. W. (2019). Application of Instance-Based Entropy Fuzzy Support Vector Machine in Peer-To-Peer Lending Investment Decision. IEEE Access, 7, 16925–16939. DOI: https://doi.org/10.1109/access.2019.2896474

[10] Dietterich, T. G. (2000). Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems, 1–15. DOI: https://doi.org/10.1007/3-540-45014-9_1

[11] Djeundje, V. B., Crook, J., Calabrese, R., & Hamid, M. (2021). Enhancing credit scoring with alternative data. Expert Systems with Applications, 163, 113766. DOI: https://doi.org/10.1016/j.eswa.2020.113766

[12] Genuer, R., & Poggi, J.-M. (2020). Random Forest. In Random Forest with R (pp. 33–55). Springer. DOI: https://doi.org/10.1007/978-3-030-56485-8_3

[13] Gonzalez, L., & Loureiro, Y. K. (2014). When can a photo increase credit? The impact of lender and borrower profiles on online peer-to-peer loans. Journal of Behavioral and Experimental Finance, 2, 44–58. DOI: https://doi.org/10.1016/j.jbef.2014.04.002

[14] Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning, 77(1), 103–123. DOI: https://doi.org/10.1007/s10994-009-5119-5

[15] Hand, D. J., & Anagnostopoulos, C. (2013). When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? Pattern Recognition Letters, 34(5), 492–495. DOI: https://doi.org/10.1016/j.patrec.2012.12.004

[16] Hasan, I., He, Q., & Lu, H. (2020). The impact of social capital on economic attitudes and outcomes. Journal of International Money and Finance, 108. DOI: https://doi.org/10.1016/j.jimonfin.2020.102162

[17] Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: evidence from the LendingClub consumer platform. Financial Management, 48(4), 1009–1029. DOI: https://doi.org/10.1111/fima.12295

[18] Jin, Y., Zhu, Y., & Ltd., I. G. S. C. S. I. Pvt. (2015). A data-driven approach to predict default risk of loan for online peer-to-peer (P2P) lending. In G. S. Tomar (Ed.), 5th International Conference on Communication Systems and Network Technologies, CSNT 2015 (pp. 609–613). Institute of Electrical and Electronics Engineers Inc. DOI: https://doi.org/10.1109/csnt.2015.25

[19] Kun, Z., Weibing, F., & Jianlin, W. (2020). Default Identification of P2P Lending Based on Stacking Ensemble Learning. 2nd International Conference on Economic Management and Model Engineering, ICEMME 2020, 992–1006. DOI: https://doi.org/10.1109/icemme51517.2020.00203

[20] Lee, E., & Lee, B. (2012). Herding behavior in online P2P lending: An empirical investigation. Electronic Commerce Research and Applications, 11(5), 495–503. DOI: https://doi.org/10.1016/j.elerap.2012.02.001

[21] Li, X., & Zengyi, Z. (2020). Research on P2P Credit Assessment Based on Random Forest — from the Perspective of Lender's Profit. 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering, ICBAIE 2020, 242–244. DOI: https://doi.org/10.1109/icbaie49996.2020.00057

[22] Maskara, P. K., Kuvvet, E., & Chen, G. (2021). The role of P2P platforms in enhancing financial inclusion in the United States: An analysis of peer-to-peer lending across the rural-urban divide. Financial Management, 50(3), 747–774. DOI: https://doi.org/10.1111/fima.12341

[23] Netzer, O., Lemaire, A., & Herzenstein, M. (2019). When words sweat: Identifying signals for loan default in the text of loan applications. Journal of Marketing Research, 56(6), 960–980. DOI: https://doi.org/10.1177/0022243719852959

[24] Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. Applied Soft Computing, 74, 26–39. DOI: https://doi.org/10.1016/j.asoc.2018.10.004

[25] Serrano-cinca, C., Gutiérrez-nieto, B., & López-palacios, L. (2015). Determinants of Default in P2P Lending. PLOS One, 1–22. DOI: https://doi.org/10.1371/journal.pone.0139427

[26] Stern, C., Makinen, M., & Qian, Z. (2017). FinTechs in China–with a special focus on peer-to-peer lending. Journal of Chinese Economic and Foreign Trade Studies. DOI: https://doi.org/10.1108/jcefts-06-2017-0015

[27] Tao, Q., Dong, Y., & Lin, Z. (2017). Who can get money? Evidence from the Chinese peer-to-peer lending platform. Information Systems Frontiers, 19(3), 425–441.DOI: https://doi.org/10.1007/s10796-017-9751-5

[28] Weiss, G. N. F., Pelger, K., & Horsch, A. (2010). Mitigating adverse selection in p2p lending–Empirical evidence from Prosper.com. Available at SSRN 1650774. DOI: https://doi.org/10.2139/ssrn.1650774

[29] Ye, X., Dong, L.-A., & Ma, D. (2018). Loan evaluation in P2P lending based on Random Forest optimized by genetic algorithm with profit score. Electronic Commerce Research and Applications, 32, 23–36. DOI: https://doi.org/10.1016/j.elerap.2018.10.004

[30] Zhang, J., & Liu, P. (2012). Rational herding in microloan markets. Management Science, 58(5), 892–912. DOI: https://doi.org/10.1287/mnsc.1110.1459

[31] Zhu, L., Qiu, D., Ergu, D., Ying, C., Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. 7th International Conference on Information Technology and Quantitative Management, ITQM 2019 (Vol. 162, pp. 503–513). Elsevier B.V. DOI: https://doi.org/10.1016/j.procs.2019.12.017

[32] Ziegler, T., Shneor, R., Wenzlaff, K., Suresh, K., Ferri, F., Paes, C., Mammadova, L., Wanga, C., Kekre, N., Mutinda, S., Wang, B. W., Closs, C. L., Zhang, B., Forbes, H., Soki, E., Alam, N., & Knaup, C. (2021). Global Alternative Finance Market Benchmarking The 2nd Global Alternative Finance Market Benchmarking Report. DOI: https://doi.org/10.2139/ssrn.3957488

View publication stats