Is Verification Status Important?

Ramnath Kumar
Dept. of Computer Science and Information Systems
BITS Pilani, Hyderabad Campus
Hyderabad, India
ramnathkumar181@gmail.com

Hussain Yaganti
Dept. of Economics
BITS Pilani, Hyderabad Campus
Hyderabad, India
hussain@hyderabad.bits-pilani.ac.in

***Abstract*—LendingClub is an American peer-to-peer lending company, headquartered in San Francisco, California. We attempt to analyze their dataset and predict the behavior of a given individual using machine learning methods. We also investigate the feature importance of each feature and make valuable observations on the same. In this paper, we attempt to predict the expected returns for loans to a given borrower. Since a loan default will result in a loss of both principal and interest, we will maximize our profits by predicting the probability of default of the borrower to help avoid investment in those high-risk notes. Our model achieves an impressive performance score of 96.8% and considers the noise added due to multiple features.**

***Index Terms*—Credit Modelling, Machine Learning, Feature Importance**

I. Introduction

LendingClub is an American peer-to-peer lending company, headquartered in San Francisco, California. Lending club is one of the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC) and offer loan trading on a secondary market. Lending Club is a peer-to-peer lending company, the largest of its kind globally with $11.1 billion originated loans. It is an online platform where borrowers can obtain loans from lenders, and investors can purchase notes backed by payments based on credits. We attempt to analyze their dataset and predict whether the behavior of a given individual. We use machine learning methods such as gradient boosting to predict the same. We also investigate the feature importance of each feature and make valuable observations on the same. Credit risk modeling is the best way for lenders to understand how likely a particular loan is to get repaid. In other words, it's a tool to understand the credit risk of a borrower. Credit risk modeling is especially important because this credit risk profile keeps changing with time and circumstances. Credit risk refers to the chance that a borrower will not make their payments on time and default on their debt. It refers to the likelihood that a lender may not receive their interest due or the principal lent on time. Defaults on debt result in an interruption of cash flows for the lender and increase collection cost. In extreme cases, some parts of the loan or even the entire investment are dropped, resulting in a lender's loss. It is challenging and complex to pinpoint precisely how likely a person is to default on their loan. Simultaneously, accurately assessing credit risk can reduce the likelihood of losses from default and delayed repayment. Credit risk modeling refers to the process of using data models to find out two important things. The first is the probability of the borrower defaulting on the loan. The second is the impact on the financials of the lender if this default occurs. In this paper, we focus solely on answering the first issue. Financial institutions rely on credit risk models to determine the credit risk of potential borrowers. This will help us decide whether or not to sanction based on the credit risk model validation. Previous credit risk models are very rigorous and based on various assumptions to justify a given feature's inclusion. The addition of the feature is left to the intuition and cannot be justified if the model's premises fail. In this paper, we evaluate the feature importance of each feature used by our model, trained on the lending club dataset.

In this paper, we propose a system that utilizes several machine learning algorithms tailored to use the feature importance of the LendingClub data. We compare different approaches and conclude our findings. In Section 2, we discuss the related work in this domain. In Section 3, we define the various datasets we experimented on. In Section 4, we present our models and evaluate their performance. In Section 5, we discuss the results from our numerous experiments.

II. Related Works

Previous works on credit risk modeling can be broadly divided under two prominent subheadings: (i) The Models Based on Financial Statement Analysis and (ii) Machine Learning Models.

A. The Models Based on Financial Statement Analysis

The models described in this section are based on the analysis of financial statements of borrowing institutions. These models chiefly take into account well known financial ratios that can be useful in determining credit risk. These models can be used to determine the likelihood of a company going bankrupt. Models such as Altman Z score [1] and Moody's risk [4] calculation fall in this category. Other papers such as [5] and [6] also use financial statement analysis to determine credit risk.

B. Machine Learning Models

As technology has progressed, new ways of modeling credit risk have emerged, including credit risk modeling using R and Python. These include using the latest analytics and big data tools to model credit risk. Zhang et al. [10] have proposed

an enhanced multi-population niche genetic algorithm for credit scoring. Bao et al. [2] proposed a credit risk model by integrating both supervised and unsupervised models. Zhang et al. [9] proposed a novel approach of selecting classifiers using the Genetic Algorithm, keeping in mind both the accuracy and diversity of the ensemble. furthermore, unsupervised clustering was also integrated with a fuzzy assignment procedure in the model, to make more use of the data pattern and improve performance. Perez et al. [8] implemented a linear mixed model (LMM) to calculate the credit risk of financial companies in home equity loans. Other factors like the evolution of economies and the subsequent emergence of different credit risk types have also impacted how credit risk modeling is performed. In most of these works, feature selection is one of the most important steps. Feature selection is an NP-hard problem that can be solved by three types of methods, namely wrappers, embedded methods and filters [7]. To improve the performance of the prediction model, wrappers use an enumeration algorithm that searches the space of attribute subsets. This usually leads to higher accuracy compared to filters, achieved at the cost of computational load. Embedded methods evaluate attribute importance while training a prediction model in order to evaluate attribute importance while training a prediction model in order to avoid model overfitting. Filters evaluate the relevance of attributes based on data characteristics, without involving any learning algorithm. In our work, we use the embedded method where we use feature importance to evaluate the useful features. To our best knowledge, the work performed by Bao et al. [2] is very similar to ours. Their work also uses the idea of feature importance for cropping down the dimensionality of the feature space. The dataset they used was also from a P2P lending dataset from China. We perform a very similar experiment on a more public dataset, and try to draw various insights from the feature importance scores.

III. Datasets

Our dataset extracted from the lending club statistics, which comprised of more than 800,000 data points. We collected the loan data from the years 2007-2015 for our research. We believe that such a robust dataset could provide enough insight to learn about peer-to-peer lending and formulate an efficient credit model. The status of the loan was our dependent variable in our research. Each entry had 75 features provided, including the status of the loan. Using the information of the zip code, we collected the mean and median income of the households in the given states using the "Median Household Income and Mean Household Income" dataset. These features were added into our dataset to aide the credit model. We removed the attributes which have more than 20% values empty. We also removed other characteristics derived from loan status. For example, we removed attributes like recovery fee, etc. non-zero only for those users who did not pay back the loan. We also remove features such as ID, sex, and other features, which we believe have no semantic correlation with the loan status. We dropped features that semantically correlated with the loan state the number of attributes reduced to 23. We further create more datasets by disregarding the features which do not play a significant role in prediction. Such features would act as noise to our model and reduce the performance of our model.

TABLE I
Features

Feature	***Description***
loan_amt	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.
funded_amt	The total amount committed to that loan at that point in time.
funded_amt_inv	The total amount committed by investors for that loan at that point in time.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
int_rate	Interest Rate on the loan.
installment	The monthly payment owed by the borrower if the loan originates.
grade	Lending Club assigned loan grade.
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.
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN,MORTGAGE, OTHER.
annual_income	The self-reported annual income provided by the borrower during registration.
mean	Population weighted mean of a given zip code area.
median	Population weighted median of a given zip code area.
payment_plan	Indicates if a payment plan has been put in place for the loan.
dti	A ratio calculated using the borrowers total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrowers self-reported monthly income.
delinq_2year	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.
open_acc	The number of open credit lines in the borrower's credit file.
pub_rec	Number of derogatory public records.
revol_bal	Total credit revolving balance.
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
total_acc	The total number of credit lines currently in the borrower's credit file.
total_payment	Payments received to date for total amount funded.
total_payment_inv	Payments received to date for portion of total amount funded by investors.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified.
loan_status	Current status of the loan

Summary of the features in our dataset used in this paper.

Table 1 provides a brief description of the features used by our models. For the status variable (dependent variable), we grouped current and fully paid as positive samples and the others as negative samples. We also removed any data points which did not have information for the status. Using the above features, we created three different datasets, each trying to make a point. Table 2 provides other statistics about the dataset. To train the model, we split the dataset into training

TABLE II
DATASET

Name	# Features	Description
Lending_Club_23	23	Used all features including verification status.
Lending_Club_22	22	Removed verification status feature from Lending_Club_23
Lending_Club_9	9	Fine tuned number of features to 9 to maximize feature importance

Summary of the dataset. Each of the datasets have 177479 data points of which 146340 are positive and 31139 are negative.

and testing with a split-ratio of 0.2. Hence, we train them and test the model on different datasets. This process allows us to test our model with unseen examples and can critically evaluate our model. In the subsequent sections, we discuss our methodology and our experiments, which lead us to our credit model for P2P lending sites.

IV. Methodology

A. Task Definition

Given a user u and set of features F as $\{f_1, ..., f_d\}$ where d is the number of features, we incorporate in our model. Our goal is to predict the probability of the user to default on his loan. However, choosing the right d is of paramount importance.

As stated by Pedro Domingos[3], it is crucial to cut down the number of features to as little as possible. This problem solves the curse of high dimensionality. Generalizing becomes exponentially harder as the examples' dimensionality grows because a fixed-size training set covers a dwindling fraction of the input space. The similarity-based reasoning that machine learning algorithms depend on breaks down in higher dimensions.

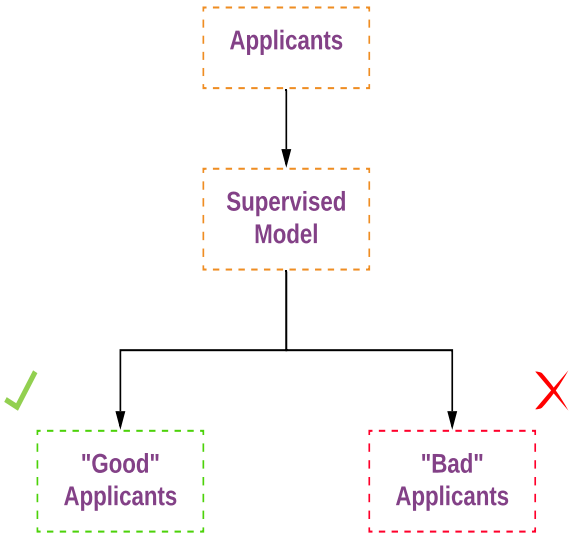

Fig. 1. Schematic Diagram of main idea in this work

B. Research Motivation

This work focuses on the creation of a supervised machine learning model which could aid the credit risk in the P2P sector. The main idea of our work is shown in Fig. 1.

In order to test the validity of the feature selection, we compare different combinations of experiments as shown in Fig. 2. As the No Free Lunch theorem indicates, there is no single ML algorithm can perform best on all practical learning problems. Therefore, we adopt a broad set of diverse, accurate and representative machine learning algorithms, namely gradient boosting (GB), logistic regression (LR),and random forest (RF) to build individual models.

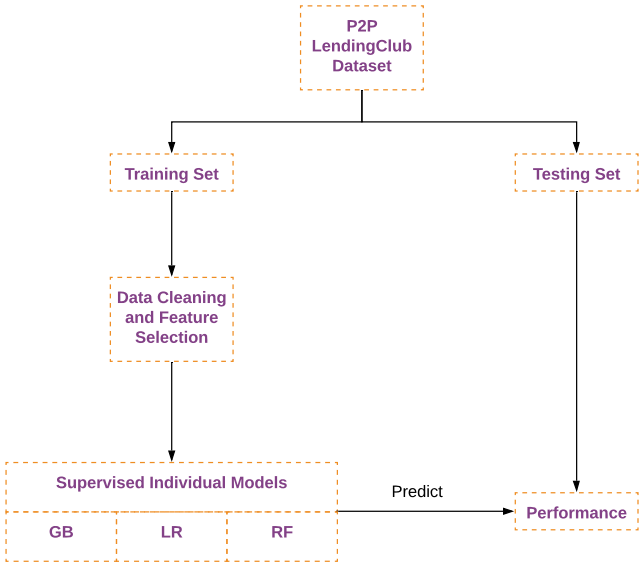

Fig. 2. The experimental process in this work

C. Summary of Proposed Approach

We run various machine learning models on the above datasets to study the distribution of features and draw different inferences on peer-to-peer lending. We used multiple algorithms, including gradient boosting, random forest, and logistic regression algorithms, to train our model and to plot the feature importance of each feature while predicting. We fix the parameters of our models so that we can effectively compare our models on different datasets.

- **Gradient Boosting:** The model was run with 50 estimators, and each tree had a max depth of 8 and with a learning rate of 0.5.
- **Random Forest:** The model was run with 50 estimators, and each tree had a max depth of 8.
- **Logistic Regression:** The model was trained with L2 penalization, for 100 iterations.

We do not work with SVM due to the quadratic complexity of the algorithm. The SVM algorithm has a complexity of $O(n_{features} * n^2_{observations})$. We compute various metrics such as precision, recall, f score, and accuracy to compare different

models on each model. The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the classifier's ability not to label as positive a sample that is negative. The recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples. The support is the number of occurrences of each class in y_true. The traditional F-measure or balanced F-score (F 1 score) is the harmonic mean of precision and recall. Furthermore, we define a feature importance score in the upcoming sections, which compares the various features in our trained model. The inferences from this will enable us to propose a novel model for P2P lending.

D. Feature Importance

We derive the feature importance from the gradient boosting model. For each decision tree, we calculate the importance of each node using the Gini Importance. The tree constructed is a binary tree, with only two children for each parent node. Suppose, the weighted number of samples reaching the node j be w_j and the Gini Importance of the node j be I_j. The left and right children of node j are represented as left(j) and right(j), respectively. Let f_i be the frequency of label i at given node and C be the total number of unique labels, then the Gini Importance of the node j:

$$I_j = \sum_{i=1}^{C} -f_i(1 - f_i) \tag{1}$$

Then, the formula for the importance of node j N_j:

$$N_j = w_j I_j - w_{left(j)} I_{left(j)} - w_{right(j)} I_{right(j)} \tag{2}$$

We define a set M_i such that node j$\in M_i$ if node j splits on feature i. The feature importance of each feature F_i:

$$F_i = \sum_{j \in M_i} N_j / \sum_{k \in all\ nodes} N_k \tag{3}$$

The feature importance is now normalized to allow easy comparison between different models. The normalized feature importance $normf_{ij}$ for a given tree j:

$$normF_{ij} = F_i / \sum_{k \in all\ features} F_k \tag{4}$$

We then average out the feature importance for T trees in the gradient boosting algorithm to obtain $FinalF_i$ where j refers to index of each tree:

$$FinalF_i = \sum_j normF_{ij}/T \tag{5}$$

This definition allows us to quantitatively measure the importance of each feature in our trained model.

E. Experiments

1) On LendingClub_24 dataset: We train on three models on the base feature set to create a baseline. We train the models with the parameters mentioned before. This model performed best with the gradient descent model giving a training accuracy score of 0.972 and a validation accuracy score of 0.972. Other metrics of the models have been provided in Table-VI.

TABLE III
PERFORMANCE METRICS

Model	*Precision*	*Recall*	*f1-score*	*Accuracy*
Gradient Boosting	0.96	0.92	0.94	0.97
Random Forest	0.96	0.82	0.87	0.94
Logistic Regression	0.94	0.87	0.90	0.95

Classification report on Lending_Club_24 dataset

The ROC_AUC score of our model is 0.97. From the metrics mentioned in Table-VI, we can conclude that our model is not overfitting as both training and validation accuracy are considerably close. The feature importance of the gradient boosting model is shown in Fig-3.

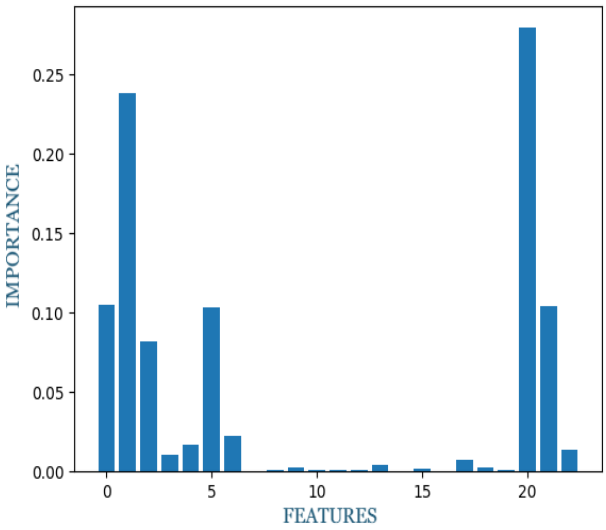

Fig. 3. On Lending_club_24 dataset

We can see that the feature importance of verification status is relatively low. The above inference leads us to question the relevance of verification status in P2P lending. Thus, it motivated us to create Lending_Club_23 dataset and studied the distribution of features. Many other elements also have low feature importance and need to be addressed.

2) On LendingClub_23 dataset: We train various models on the new feature set after removing validation status. We train the models with the parameters mentioned before. This model performed best with the gradient descent model giving a training accuracy score of 0.972 and a validation accuracy score of 0.965. Other metrics of the models have been provided in Table-IV.

TABLE IV
PERFORMANCE METRICS

Model	*Precision*	*Recall*	*f1-score*	*Accuracy*
Gradient Boosting	0.96	0.92	0.94	0.97
Random Forest	0.97	0.88	0.88	0.94
Logistic Regression	0.94	0.87	0.90	0.95

Classification report on Lending_Club_23 dataset

The ROC_AUC score of our model is 0.97. From the metrics mentioned in Table-IV, we can conclude that our model is not overfitting as both training and validation accuracy are considerably close. The feature importance of the gradient boosting model shown in Fig-4.

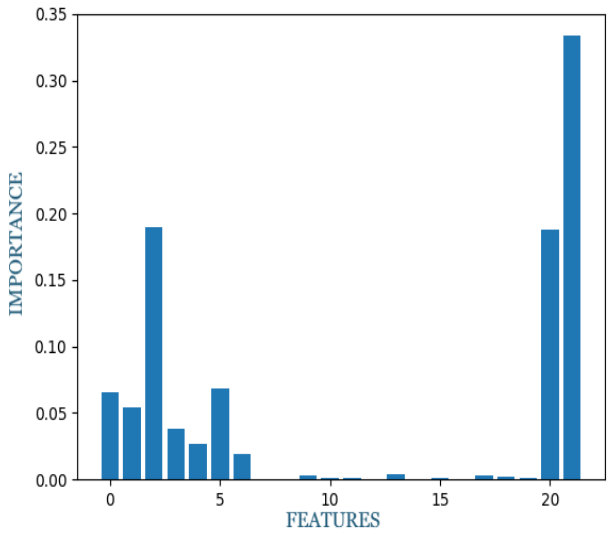

Fig. 4. On Lending_club_23 dataset

We can see that the feature importance roughly remains the same with some features with very low feature importance. The above inference motivated us towards the creation of the Lending_Club_9 dataset and study the distribution of features. We use only the top 9 features in our new dataset and removed features 7-19.

3) On LendingClub_9 dataset: We train various models on the new feature set after removing features of lower importance. We train the models with the parameters mentioned before. This model performed best with the gradient descent model giving a training accuracy score of 0.972 and a validation accuracy score of 0.968. Other metrics of the models have been provided in Table-V.

The ROC_AUC score of our model is 0.97. From the metrics mentioned in Table-V, we can conclude that our model is not overfitting, as both training and validation accuracy is considerably close. The feature importance of the gradient boosting model is shown in Fig-5.

We can see that the feature importance roughly remains uniform. This inference suggests that we have successfully

TABLE V
PERFORMANCE METRICS

Model	*Precision*	*Recall*	*f1-score*	*Accuracy*
Gradient Boosting	0.97	0.92	0.94	0.97
Random Forest	0.97	0.88	0.92	0.96
Logistic Regression	0.94	0.87	0.90	0.95

Classification report on Lending_Club_9 dataset

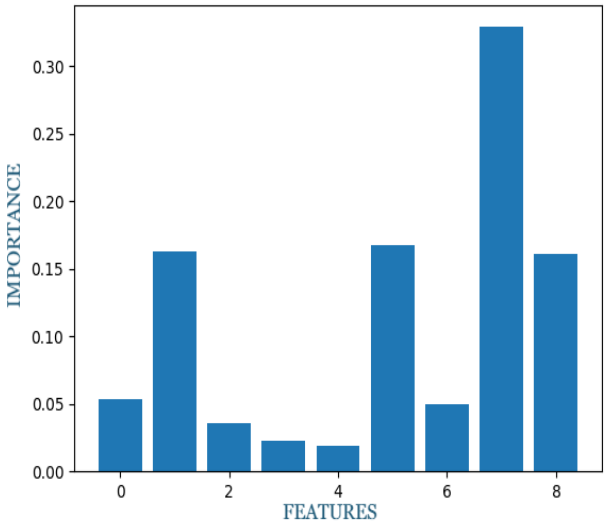

Fig. 5. On Lending_club_9 dataset

removed most of the noisy features from our feature set, and all our features are of approximately equal importance. We end up with only 9 features namely: loan_amt, funded_amt, funded_amt_inv, term, int_rate, installment, total_payment and total_payment_inv.

V. RESULTS

As stated before, we successfully created a model that allows us to predict the loan status of peer-to-peer loans with almost 96.8% accuracy. As shown from the above experiments, we can make a few logical inferences with certainty. One of our study's paramount references is that verification status is not an essential attribute in predicting the loan status. The above conclusion could bring a tremendous change in choosing the people to give the loans to (simplifying this sentence is essential, as well as expanding it). Another surprising finding is that the annual income has relatively no importance. Regardless of how much an individual earns, his consumption depends on his taste and environment. Furthermore, a risk-averse individual may save a lot more than a risk-loving individual(expanding this would help add all of the above features to explain).

TABLE VI
PERFORMANCE METRICS

	Precision	*Recall*	*f1-score*	*ROC_AUC score*	*Acc.*
Model 1	0.9	0.97	0.96	0.97	0.972
Model 2	0.97	0.97	0.96	0.97	0.965
Model 3	**0.97**	**0.97**	**0.97**	**0.98**	0.968

Comparison of the three models

VI. CONCLUSION

REFERENCES

[1] Edward I Altman. "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy". In: *The journal of finance* 23.4 (1968), pp. 589–609.

[2] Wang Bao, Ning Lianju, and Kong Yue. "Integration of unsupervised and supervised machine learning algorithms for credit risk assessment". In: *Expert Systems with Applications* 128 (2019), pp. 301–315.

[3] Pedro M Domingos. "A few useful things to know about machine learning." In: *Commun. acm* 55.10 (2012), pp. 78–87.

[4] Eric G Falkenstein, Andrew Boral, and Lea V Carty. "RiskCalc for private companies: Moody's default model". In: *As published in Global Credit Research, May* (2000).

[5] Hyeongjun Kim, Hoon Cho, and Doojin Ryu. "An empirical study on credit card loan delinquency". In: *Economic Systems* 42.3 (2018), pp. 437–449.

[6] Trust R Mpofu and Eftychia Nikolaidou. "Determinants of credit risk in the banking system in Sub-Saharan Africa". In: *Review of development finance* 8.2 (2018), pp. 141–153.

[7] Stjepan Oreski and Goran Oreski. "Genetic algorithm-based heuristic for feature selection in credit risk assessment". In: *Expert systems with applications* 41.4 (2014), pp. 2052–2064.

[8] A Pérez-Martín, A Pérez-Torregrosa, and M Vaca. "Big Data techniques to measure credit banking risk in home equity loans". In: *Journal of Business Research* 89 (2018), pp. 448–454.

[9] Haoting Zhang, Hongliang He, and Wenyu Zhang. "Classifier selection and clustering with fuzzy assignment in ensemble model for credit scoring". In: *Neurocomputing* 316 (2018), pp. 210–221.

[10] Wenyu Zhang, Hongliang He, and Shuai Zhang. "A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm: An application in credit scoring". In: *Expert Systems with Applications* 121 (2019), pp. 221–232.