P2P Loan Default Prediction

Tanya Nadkarni, Javier Recasens, Skyler Shasky, Suhas Tummalapalli

Introduction

In this project, we built a classification model to predict good loans for investment in Lending Club, the world's largest online credit marketplace for peer to peer lending. Ever wondered what makes you a good investment? To answer this kind of questions we also uncovered insights which will help educate individuals on what factors contribute positively and negatively to them being picked on the Lending Club. We will use R, Azure ML, R Shiny and Tableau to develop and deploy an intelligent web application capable of providing useful information for Lending Club users.

Problem definition

Peer to peer lending is a very simple process where the borrower applies for the loan and if he/she meets a certain criterion, their loan details are uploaded to an online platform. Investors browse through the platform and build their own portfolios depending on the expected risk and expected return. If the borrower fails to meet the minimum criterion, his/her details are removed from the platform.

In our project, we want to classify good and bad loans as a function of the relevant attributes of a loan applicant. Loans that indicate good repayment behavior are considered “good” and loans that indicate less than perfect repayment behavior are considered “bad”. Every day, Lending Club issue new approved loans into their platform, individuals interested in investing in those loans can use our solution to pick good loans by filtering the important attributes. From a borrower’s perspective, this project also provides factors contributing to default.

Compared to other approaches that try to solve this problem, we improved our analysis by using sentiment analysis on the description of individual loans. We found that the probability of a bad loan could be related to the attitude or emotions embodied in the description.

The dataset used to create the statistical models contain complete information on all the loans issued during a given period (from 2007-2016). Loan details include the current loan status and the latest payment information. The dataset also contains the demographic information of the loan applicants and features like credit scores, number of loan inquiries etc. There are 1,300k observations and 113 features in the dataset. Each observation represents a different loan. The total size of the dataset is 1 GB.

Literature Survey

Our literature review consists of the revision of various studies that assess the default risk of loans by combining Decision Trees, Bootstrap and Gradient Boosting Techniques [1], as well as an overview of the newly emerging people to people lending space [2]. We also reviewed other machine learning methods like Support Vector Machines (SVM) and Neural Networks [3] [4] [5].

There are several statistical techniques for credit scoring [6] and default prediction such as logistic regression, neural networks, and/or classification trees. Logistic regression is by far the most widespread due to high predictive capability and results not significantly different from more recent techniques. Our literature review consists of revision of various studies that asses the default risk of loans.

Rather than just a logistic regression model using train and test sets, several logistic regression models can be performed to analyze predictive capability of each variable. Model 1 considers only 1 explicative variable and adds variables till we reach a complete model. Significance of each variable and goodness of fit of can be measured using Hosmer-Lemeshow and Nagerkelke Statistics. Random Forest [7] [8] offers several advantages. It runs efficiently on large databases and does not require supervised feature selection to work well. However, data imputation is required for this dataset to improve performance and the choice of method of data imputation is likely to significantly affect outcome of prediction. The Gaussian Naïve Bayes model [9] was found to work well, possibly due to features such as credit scores and regional population being distributed in Gaussian and should be taken into consideration.

For the text sentiment analysis, we analyzed loan description data from applications guided by the framework of the LIWC [10]. To understand previous studies, we reviewed the usage of a Generalized multiple kernel learning model for credit risk evaluation using sentiment analysis [11]. Our mode of approach is to identify the key features and sentiments that we believe to be strong indicators of default rates, analyze the text for these sentiments and verify correlations to include in our models.

We reviewed literature about asymmetry of information [9] [12] which applies to the situation of P2P lending because lenders do not have much information about the borrower's credibility and ability to repay the loan.

Proposed Method

We propose using a data-driven approach when investing in Lending Club by leveraging statistical methods and self-service visualizations.

Innovations

The primary innovations we bring to this space are the addition of sentiment analysis for the loan description column, as well as an interactive user interface via an R Shiny app that provides borrowers with the ability to determine their likelihood of procuring a loan. By sharing on the web our model and results we will also serve as an investment recommender for new issued loans.

Lending Club updates their historical dataset every month. We have built our solution in such a way so that updated datasets could be easily processed with SQL stored procedures and its final output used to re-train our models. Lending Club issue new loans every day and provides this information via an API. Even though we propose an interactive solution for prediction, integration with the API is feasible and could dramatically improve the investment pattern in Lending Club.

Text Analysis

One of the new approaches we brought through our work is the analysis of text data. We found that other studies of data from Lending Club lack this feature in their analysis. Lending club typically require applications to write a short explanation of why they are seeking a loan, how they plan to use it and include any other relevant details that would help them make a case for themselves. We believe that there is a lot of valuable information hidden in this text, and there is a strong chance that this information contains predictive power.

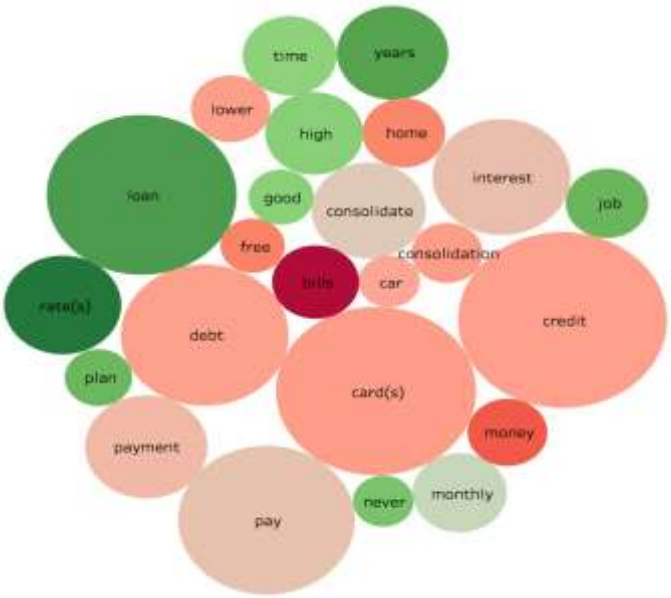
The text data was available to us as a field called "desc" in our data set, which we refer to as the loan description. The descriptions were cleaned in Excel and R to exclude any noise existing in the data. We wanted to choose simple yet precise methods that would extract the information embedded within the text.

We created variables for simple numeric measures of the data like the length of the description, the number of words in the description and the length of the average word. We created a dictionary of commonly occurring words in the description which we could use as predictors. We then also created dummy variables which check for the presence of these keywords in each potential borrower's description.

We then created a logistic model which included the most important variables resulting from a Random Forest Gini Index decrease accumulation. We found multiple variables to be significant predictors and we further studied them to quantify the effects they have on loan default rate. Our interactive interface allows users to input their loan description

and the interface tells them what particularly within their description contributes the most to positive and negative changes to their expected loan default rates.

The analysis was conducted entirely in R with the help of package such as "stringr", "stringi", "GGally", "ggplot2". The visualizations were created in Tableau.



The size of each bubble represents the frequency of the word in the training data.

The color of each bubble measures the numeric value of the effect of the presence of the word in the description.

Figure1: Text Analysis

User Interface/Visualizations

We used Tableau primarily for exploratory data visualization. This provided some key contextual plots that summarized the data and gave insight to various features of the peer to peer lending industry.

We created an R Shiny application where a potential borrower can adjust input variables to determine likelihood that they will get a loan. This will empower borrowers by giving them the knowledge of which factors are reducing their chances of finding a lender. This will also give lenders (investors) a place to evaluate investments. The output of the app displays a percentage chance that the borrower will end up as a bad loan, as well as the contribution of individual variables on that chance. There is also a “Model Info” tab that shows information about the predictive power of the model.

Figure 2: R Shinny Visualization tool

Experiments / Evaluation

Exploratory Analysis

Visualization in Tableau provided an initial perspective of our data. One of the more interesting charts showed that the data was unbalanced with a large percentage a loans that were fully paid and in good status. We also found that the largest type of loan by far was debt consolidation, with credit card refinancing and home improvement in 2nd place and 3rd place respectively.

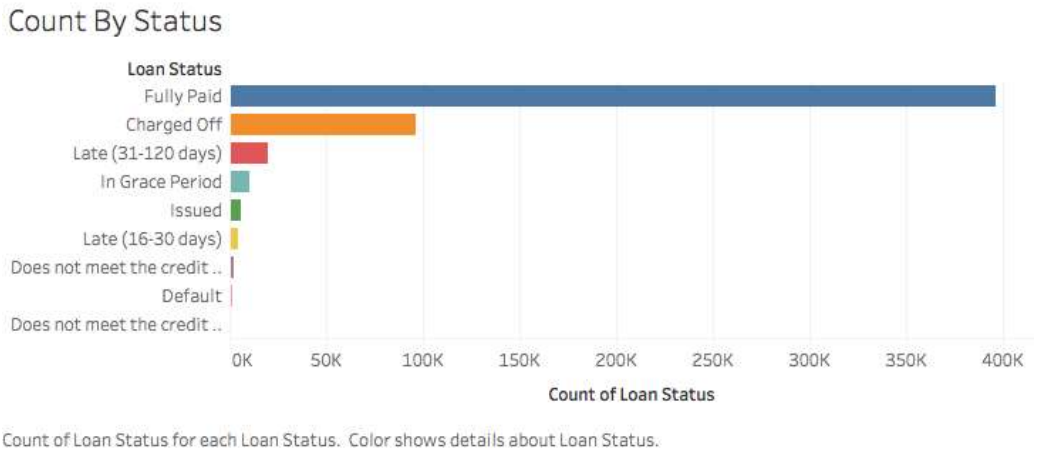


Figure 3: Count By Status

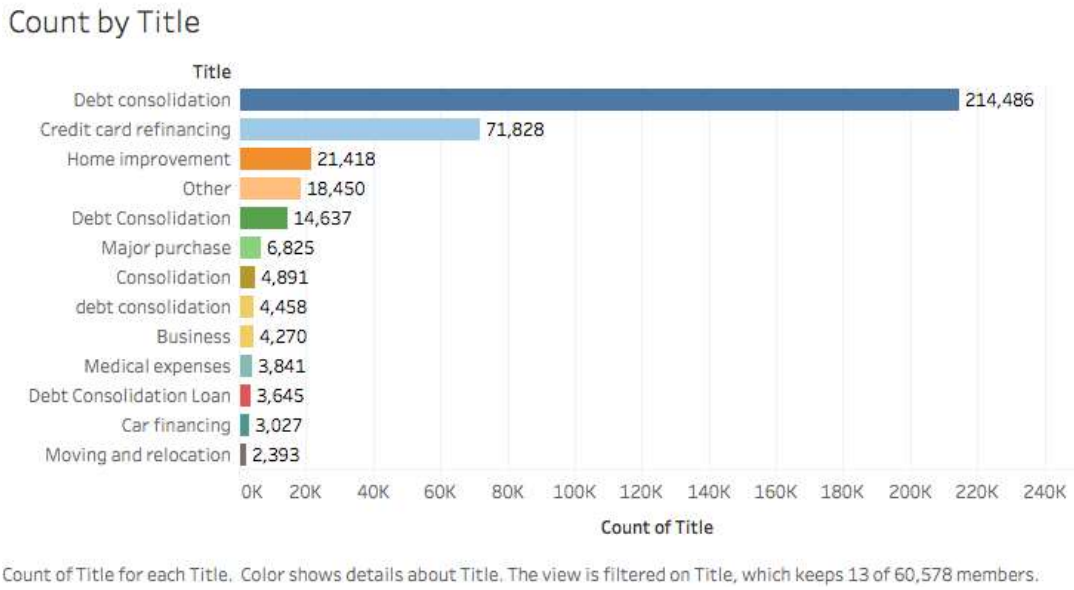


Figure 4: Count by Title

To visualize which predictors could be useful to discriminate between good and bad loans we plotted response distribution for numerical variables and stacked bar charts for categorical variables. As an example, we can clearly see below that a lower interest rate is associated with good loans (Response equal to 0) and the longer the term (60 months) the higher the chances of having a bad loan (Response equal to 1).

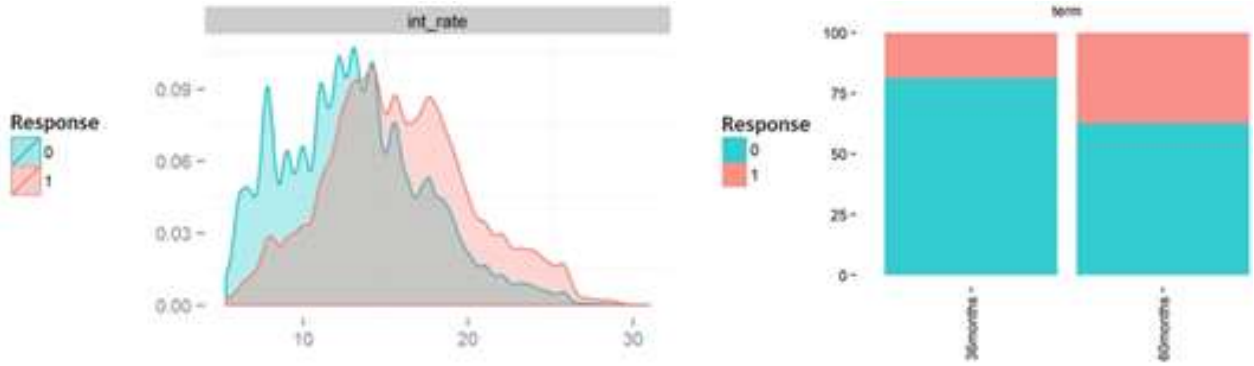


Figure 5: Important discriminant variables

Data Preparation

To prepare our dataset we take advantage of Microsoft SQL Server 2016 Developer Edition. We used SQL and R script for the data manipulation and transformations. We have created a table called "LoanStats" with all column definitions to store all 113 features. Then we inserted via stored procedures data coming from the 11 csv files.

For all features that represent a percentage we removed the % signed, transformed to a numeric and divided by 100. We did a de-normalization of the data to code the categorical variables.

We also removed rows where loan_status column is empty. "Current loan" value will be removed from the dataset.

Response creation

We created the response variable from the "loan_status" column. There are 7 loan statuses: Charged Off, Current, Default, Fully Paid, In Grace Period, Late (16-30 days) and Late (31-120 days). The status flow of a loan is shown below:

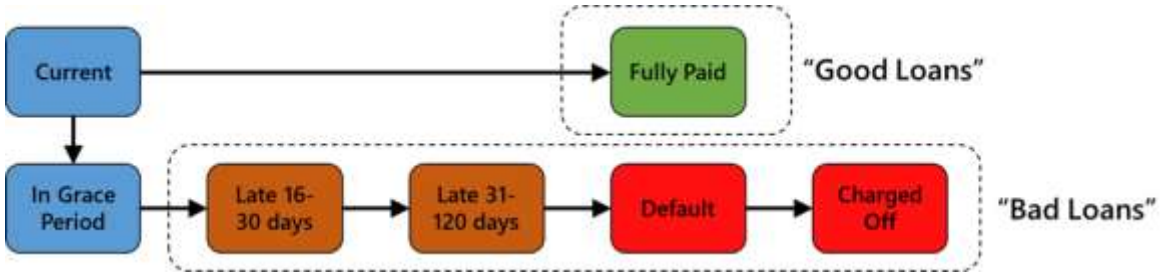


Figure 6: Loan Status flow

Since our objective is to classify good and bad loans we grouped them to create a binary response called "is_bad". We categorize Charged Off, Default and Late to a single category of bad loan which takes the value 1. Fully paid loans are categorized as good loans which take the value 0. We are not interested in "current" or in "Grace Period" loans because we don't know if those will behave poorly.

Feature Selection

Nearly half of the columns hold very small amounts of data are dropped from the dataset.

We create an array of columns that will not be considered to build the model. This includes purpose and loan descriptions which will be tackled in the text analysis and correlated variables (e.g. loan amount, funded amount, funded amount invested) identified by building a correlation matrix. For variables, such as zip code, expanding into Boolean values does not make sense and these are also dropped. They can be explored later if we are not satisfied with model performance. One of the papers we came across suggests merging the dataset with census data using zip code and using variables that offer better insight such as median income for that state. We were left with 24 predictors (18 numerical and 6 categorical).

Feature Engineering

We created a few new predictors which increased the total potential features to 31. The list of new predictors is the following:

- is_desc: We transformed description into a numeric binary predictor that takes the value 1 if it has a description or 0 it not.
- fico_avg: The average of the lowest and highest FICO borrower score for the last time it was pulled.
- years_w_credit_hist: Proxy for years of applicant. How many years since he opened the first credit line.
- avg_income_per_return: To leverage zip code, we joined external data such as the Tax Statistics from the IRS. We created a predictor that calculates the average gross income per tax return for each zip code (in thousands).

Classification Model

We approached this project as a binary classification problem to predict two classes of loans. The modeling strategy considers putting more emphasis on avoiding bad loans since they can wreak havoc the overall investment of an individual. We built different classification models. Specifically, we tested Stochastic Gradient Boosting, Random Forest, Decision Trees, Logistic Regression and a Naive Bayes classifier. We checked the assumptions of each of these models before implementing them on the given dataset. We found the accuracy and AUC scores of each of the models mentioned.

We decided to use binomial logistic regression as our classification model since the performance metric we considered were very competitive. Furthermore, through our literature survey we found it to be relatively less complicated and at the same time, quite effective. The intuition behind using logistic-regression are the benefits on inference as we want to be able to clearly explain the drivers and reasons for a good/bad loan.

To save time and resources, we have used only one of the large datasets for building the model. Once improved, it can be applied across all available Lending Club datasets.

We use the *caret* package to partition the dataset into test and training sets and the *glm()* function to fit the actual model. We arrive at the final model by checking for statistical significance and multicollinearity among variables.

Performance Measures

In our dataset, the proportion of good to bad loans is about 3:1 (76% of good loans). This means that we would gain 76% accuracy by just labelling all loans as good loans.

In a real situation, since we are not including current loans, the ratio of fully paid loans is usually much higher. Hence, the standard - net accuracy metric - is not a useful performance measure. Instead, we focused on a trade-off in identifying a bad loan at the expense of mislabeling some good loans. We considered the ROC curve and paid

focus on AUC when we train our models. We also focused on sensitivity as we are interested on minimizing the false negatives (bad loans predicted as good loans).

Since the levels in the response variable "is_bad" are unbalanced, we did an under sampling of the good loans to balance the proportion of class labels. We used the SMOTE R function which oversamples the bad loan event by using bootstrapping and k-nearest neighbor to synthetically create additional observations of that event and get a 1:1 proportion. Using SMOTE, the performance increased and so we decided to train our models on this modified dataset. Our model resulted in a AUC of 0.72 and Sensitivity of 0.6.

Conclusion

The problem of predicting loan default is one that has been studied many times, but we believe that we brought additional insight and creativity to the problem. The analysis of text gave us insights into how certain ways that applicants write their descriptions could have predictive power on their actual rates of default. Due to the large size of the data set and the high rates of confidence in the results, we believe our analysis to have real information that can assist lenders and borrowers alike. We recognize that the effects observed could be caused by underlying variables such as education levels, geographic areas associated with the vernacular, etc. This is something that may be interesting to study further.

The creation of an interactive user dashboard provided a practical application for our predictive model. We successfully created an application that can allow both lenders and borrowers to understand default likelihood and shine a light on the key factors contributing to that likelihood. Both sides of the equation can benefit from this. Investors will be able to make more informed decisions about where to lend their money, and borrowers will be able to optimize their chances at receiving a loan.

We believe that this space could certainly benefit from further research and development. By integrating with the Lending Club API, automated investment on new released loans could improve the response time. Our solution requires some time to fill out the attributes and this is a drawback as the best loans do not last too long on the Lending Club platform. By automating this process with an API integration, investors could dramatically improve the performance of their portfolios by quickly selecting the best loans.

In this project, we allocated focus towards variable interpretation in the predictive model because we wanted to understand the factors contributing to default. However, there are other models that could improve on ours from a pure prediction accuracy standpoint. Also, we focused on bad loans in a general sense, but it may be interesting to see if there are differences when predicting defaults versus late payments.

Appendix

Loan Status definition

Loan Status	Meaning
Charged Off	Loan for which there is no longer a reasonable expectation of further payments. Generally, Charge Off occurs no later than 30 days after the Default status is reached.
Current	Loan is up to date on all outstanding payments.
Default	Loan has not been current for 121 days or more.
Expired	Loan request did not receive full funding or did not pass final review and was not issued.
Fully paid	Loan has been fully repaid, either at the expiration of the 3- or 5-year year term or as a result of a prepayment.
In Funding	Loan request is listed on the site and is still receiving funding from investors.
In Grace Period	Loan is past due but within the 15-day grace period.
In Review	Loan request is pending a final review by our Credit Department to verify certain information in the application before the loan is issued.
Issued	New loan that has passed all Lending Club reviews, received full funding, and has been issued.
Issuing	Loan has been originated by our banking partner. Notes corresponding to the loan will be issued to investors within 2–5 days.
Late (16-30)	Loan has not been current for 16 to 30 days.
Late (31-120)	Loan has not been current for 31 to 120 days.
Not Yet Issued	Includes loan requests that are In Funding, loan requests that are In Review, and Notes that are Issuing.
Partially Funded	Loan did not receive full funding and is pending borrower acceptance.
Removed	Loan listing was removed based on a credit decision, or the inability to verify certain borrower information.

Preliminary Model Results

```
Coefficients: (1 not defined because of singularities)
                                  Estimate Std. Error    z value Pr(>|z|)
(Intercept)                      5.354e+14  1.699e+07  3.151e+07   <2e-16 ***
loan_amnt                        2.441e+11  1.338e+02  1.824e+09   <2e-16 ***
int_rate                         3.627e+13  2.270e+05  1.598e+08   <2e-16 ***
emp_length                       4.823e+13  1.380e+05  3.494e+08   <2e-16 ***
dti                             -5.947e+12  6.338e+04 -9.382e+07   <2e-16 ***
fico_range_low                   2.321e+12  2.085e+04  1.113e+08   <2e-16 ***
revol_bal                       -2.162e+09  1.852e+01 -1.168e+08   <2e-16 ***
revol_util                      -1.847e+12  1.778e+04 -1.038e+08   <2e-16 ***
out_prncp                        3.522e+12  1.641e+05  2.146e+07   <2e-16 ***
out_prncp_inv                           NA         NA         NA       NA
total_pymnt                      4.597e+16  1.463e+08  3.143e+08   <2e-16 ***
total_pymnt_inv                  3.386e+08  1.635e+02  2.071e+06   <2e-16 ***
total_rec_prncp                 -4.597e+16  1.463e+08 -3.143e+08   <2e-16 ***
total_rec_int                   -4.597e+16  1.463e+08 -3.143e+08   <2e-16 ***
total_rec_late_fee              -4.596e+16  1.463e+08 -3.142e+08   <2e-16 ***
recoveries                      -4.597e+16  1.463e+08 -3.142e+08   <2e-16 ***
collection_recovery_fee         -2.665e+12  4.737e+03 -5.626e+08   <2e-16 ***
last_pymnt_amnt                 -1.331e+11  1.313e+02 -1.014e+09   <2e-16 ***
last_fico_range_high            -4.789e+12  1.000e+04 -4.787e+08   <2e-16 ***
last_fico_range_low              1.922e+11  6.361e+03  3.022e+07   <2e-16 ***
Term 60 months                  -5.332e+14  1.296e+06 -4.115e+08   <2e-16 ***
HomeOwnershipNONE               -2.550e+15  3.357e+07 -7.596e+07   <2e-16 ***
HomeOwnershipOTHER              -4.541e+14  6.636e+06 -6.843e+07   <2e-16 ***
HomeOwnershipOWN                -2.524e+14  1.579e+06 -1.598e+08   <2e-16 ***
HomeOwnershipRENT               -8.877e+13  8.795e+05 -1.009e+08   <2e-16 ***
VerificationStatusSource Verified -3.450e+14  1.047e+06 -3.294e+08   <2e-16 ***
VerificationStatusVerified      -1.478e+14  1.038e+06 -1.424e+08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

References

[1]	**M.-L. Charpignon, E. Horel and F. Tixier, *Prediction of consumer credit risk,* CS229 Stanford, 2014.**
[2]	H. Wang, M. Greiner and J. Aronson, "People-to-People Lending: The Emerging E-Commerce," *AMCIS,* no. 802, 2009.
[3]	Y. Jin and Y. Zhu, "A data-driven approach to predict default risk of loan for online Peer-to-Peer(P2P) lending," *2015 Fifth International Conference on Communication Systems and Network Technologies,* 2015.
[4]	D. Olson, D. Delen and Y. Meng, "Comparative analysis of data mining methods for bankruptcy prediction," *Decision Support Systems,* no. 52, p. 464–473, 2012.
[5]	M.-C. Tsai, S.-P. Lin, C.-C. Cheng and Y.-P. Lin, "The consumer loan default predicting model – An application of DEA–DA and neural network," *Expert Systems with Applications,* vol. 36, p. 11682–11690, 2009.
[6]	S. Chang, S. D.-o. Kim and G. Kondo, "Predicting Default Risk of Lending Club Loans," *Stanford CS229: Machine Learning - Autumn 2015-2016.*

[7]	A. Chakraborty and G. Chakraborty, "Know your Interest Rate," *SESUG,* 2016.
[8]	M. Malekipirbazari and V. Aksakalli, "Risk assessment in social lending via random forests," *Expert Systems with Applications,* vol. 42, no. 10, p. 4621–4631, 2015.
[9]	C. Serrano-Cinca, B. Gutiérrez-Nieto and L. López-Palacios, "Determinants of Default in P2P Lending," *University of Zaragoza,* 2015.
[10]	P. Roshan, Y. Douglas and S. B. Elson, *Using Social Media to Gauge Iranian Public Opinion and Mood After the 2009 Election,* Santa Monica: US: RAND Corporation, 2012.
[11]	D. Zhang, W. Xu and Y. Zhu, "Can Sentiment Analysis Help Mimic Decision-Making Process of Loan Granting? A Novel Credit Risk Evaluation Approach Using GMKL Model," *System Sciences (HICSS),* no. 48th Hawaii International Conference, 2015.
[12]	Z. T. ABIR, "How Your VantageScore Credit Report Is Calculated," 2017. [Online]. Available: https://www.academia.edu/4413109/How_Your_Vantage-Score_Credit_Report_Is_Calculated.

Team Member Distribution

All team members have contributed equally.