Predicting Loan Grade From Lending Club Dataset

Noah Wong
Dept. of Mathematics & Statistics
University of Minnesota Duluth
Duluth, United States
wongx565@d.umn.edu

Bipasha Kundu
Dept. of Electrical Engineeering
University of Minnesota Duluth
Duluth, United States
kundu020@d.umn.edu

Sudipta Paul
Dept. of Electrical Engineering
University of Minnesota Duluth
Duluth, United States
paul1086@d.umn.edu

***Abstract*—Predicted loan grades on peer-to-peer loans from Lending Club by implementing machine learning algorithms like random forest, decisions tree classifiers, neural network, logistic regression and k-nearest neighbor (KNN). Decision Tree algorithm gave the best accuracy and F1-scores, also was able to predict the classes with less data. It also ran quickly compared to all other algorithms.**

I. Introduction

In the field of finance, determining the risk of investments is key to successful trading. Being able to quantitatively assess the possibility of a borrower defaulting on a loan is essential for any investor to make money. To offset the risk, investors will set an appropriate interest rate to loans. A borrower's credit score have been effective in the past for determining risk, but in the age of computers and data mining many investors are searching for new approaches. Machine Learning techniques have been used to predict whether a borrower will default on a loan. These algorithms can better prepare investors to avoid risky investments. We are approaching this interaction of borrower and investor from the opposite side. This paper looks to aid borrowers in reducing their interest rates. Machine Learning algorithms are used to predict the interest rate given to a particular loan application. This research can assist borrowers in improving the interests rates that are offered to them. The models we created can predict the interest rates borrowers will receive, which in turn, can help them reduce interest rates on potential loans.

The dataset used in this project comes from Lending Club a peer-to-peer (P2P) lending service that connects borrowers to lenders. Lending Club will grade the loans based on a scale from A to G. Loans with grade A have the lowest interest rate and are therefore determined to be least risky to investors. Lower grades correspond to higher interest rates. Investors can use these grades to help determine which loans they want to back. Borrowers have to submit their financial information to receive the loan and respective interest rate.. Using these data we predicted the loan grades given a borrower's financial history. The models we built could be purposed for borrowers to learn what types of interest rates they will be offered and what parts of their credit score they should look to improve for higher loan grades. Most research on this topic benefits the investors in ways to deny loans or justify higher interest rates. New technologies, like data mining algorithms, in the hands of investor have rarely been used to lower interest rates and therefore make less money. Machine Learning and big data can be a powerful tool and also a dangerous one for exploiting borrowers. When put into the hands of wealthy investors and powerful corporations they can become weapons of injustice. It is a lofty goal, but we hope this work looks to empower an often forgotten community.

Our data represents small three-to-five year loans up to $40,000. These loans are mainly used by families and small businesses. The two most common intended use of these loans as described by the borrowers were ”debt consolidation” and ”credit card”. These two reasons make up 80% of the Lending Club data. Lending Club is advertised to the borrowers as a resources to obtaining financial security. It is our goal to create models that can aid the borrowers in their search to rebuild financial security and pay off their debts. Much of the research done in this topic empowers investors yet in this project, we seek to reduce the computation disparity by providing equitable resources to borrowers. People struggling with debt, likely don't have the time or resources to design machine learning algorithms to improve the interest rate they will receive. On the other hand, the use of these tools by investment firms and banking corporations is commonplace. We are by no means disparaging these practices by investors but, it's important to remember that these automated tools while effective in making money can have severe consequences for families unable to consolidate debt after being rejected for a loan by the hand of a machine learning algorithm.

In this paper we explore five different machine learning algorithms in a shotgun blast technique for solving this problem. Along the way we encountered roadblocks of computing power and class disparity. Given the pandemic that occurred during the time we were testing and building models, all work was done on personal laptops. Without external servers or support to find any other options we had to reduce the size of our data set. With long-distance learning and the reliance on laptops our team could not afford to leave our laptops running for hours so we cut down on the size of our data.

II. Related Work

Research into the topic gave us more details on how peer-to-peer tradings works and how it became a popular funding technique. Lending Club started in 2007 and grew due to many factors including the distrust with commercial banks during the financial crisis in 2008, the use of the an internet trading platform to connect borrowers to investors and inability of borrowers to obtain loans with decent interest rates[2]. Also P2P has the potential of mutual profitability, borrowers will have smaller interest rates while lender can make more return on investment than through a traditional bank. These factors account for Lending Club becoming a major force in personal loan funding. As of 2019, Lending Club has reached $ 56 billion in total loans issued [3].

The P2P format will take borrowers loan applications and evaluate risk assigning a loan grade to it. Lenders then can fund the loans using the loan grade as a marker of risk. The ability to grade loans accurately based off a borrowers financial history is crucial for any P2P. Lending Club describes this process as creating a model rank ”which analyzes the performance of borrower members and takes into account FICO score, credit attributes, and other application data”[4]. This algorithm is hidden from both investors and borrowers. The model rank is then modified based on the loan amount and the loan term, 36 or 60 months. The longer 5 year loans as well as loans smaller than $5,000 and above $25,000 reduces the model. The final model rank determines the loan grade and subgrade. The loan grades range from A to G with each grade having a subgrade 1 to 5. The subgrade further breaks the interest rate into smaller pieces. We choose to predict the loan grade over the interest rate, because it gives a range to the possible interest rate and the loan grade will affect how lender see the loan.

Researchers from Istanbul have used the same machine learning algorithms of random forest, k-nearest neighbors, logistic regression and support vector machines to predict whether a loan will default or not. They also used Lending Clubs data, but they predicted the borrower status without using the FICO credit score. They determined that the random forest algorithm was most effective in prediction and performed better than Lending Club’s own loan grades. This research shows us that machine learning techniques can be effective, but we predicted the loan grade instead of whether a borrower shall default or not. Also the inclusion of the FICO score in our models is crucial since it is used by Lending Club in their assessment of the loan grade[5].

Research focused on small businesses found that while small businesses were twice as likely to be funded, when controlled for quality of application, the interests rates offered were two times higher than loans from traditional sources.They used logistic regression to predict whether a loan from a small business would be funded with good accuracy. However small business loans only make up 1% of the total loans. So their results don’t account for different reasons to apply for a loan. This paper does provide good resources and a simple model that could help small business get funding, but it focuses on a small population of who borrows from Lending Club[2].

III. Proposed Work

We have divided our works into several parts.

- Pre-processing: The Lending Club data set was very high dimensional. We have processed the data in order to train our machine learning algorithms.
- Implementation of four different algorithms which include Random Forest(RF), Neural Network (NN), K nearest neighbor(KNN) and Decision Tree(DT).
- Implementation of logistic regression from scratch.
- Finally compare all of them to identify which model performs better.

We used the data set of Lending Club Loan Data, which is publicly available on Kaggle[6]. This data set contains complete loan data for all loans issued through the 2007-2018, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the ”present” contains complete loan data for all loans issued through the previous completed calendar quarter. Additional important features include credit scores, number of finance inquiries, total annual income,home ownership, average current balance and collections among others. The file is a matrix of about 2.2 million observations and 151 variables. In order to reduce the dimensionality of the data set we used a number of techniques detailed below.

A. Data Preprocessing

To preprocess our data, we followed some steps. First, we dropped features which were not related to predict the load grades. Some of them were addr_state,zip code, charge_off_within_12_mths,emp_title,funded_amnt_inv & so on. We also removedsome columns that provided same information. These include settlement_status, settlement_date, settlement_amount etc. There were some rows with missing values in our data set. So we dropped all the rows with missing values. We set up a threshold of 80% and dropped all the rows that had more than 80% missing values in them. We also removed some rows with null values. All joint accounts were dropped, the joint accounts had two people’s financial history and thus would be messy to accommodate into the normal account. We found some features that are strongly correlated with each other.Some of them include fico_range_low,open_acc,total_pymnt,num_actv_rev. So we further reduced dimensionality by plotting the correlation graph. We used Pearson correlation graph to find the strongly correlated features. In this stage of pre-processing we were able to clear almost all unnecessary columns and rows from our data set. We had some categorical features in our data set like home_ownership,verification_status, loan_status. As we decided to implement RF,NN,KNN & DT, so we converted the categorical variables into dummy variables. This was particularly useful to our project as our data set had a number of attributes belonging to different data types. Finally,we were left with 159587 rows and 66 columns to the classifier.

Implementation Platform

- **Data Processing :** Pandas, Numpy
- **Machine Learning Libraries:** Keras,Tensorflow, Scikit-learn
- **Data Visualization:** Matplotlib, Seaborn

B. *Machine Learning Algorithms*

After processing the data set,we were left with reduced columns and samples. We have an imbalanced data set. To deal with imbalanced data set, there are several methods like up sampling, down sampling, generate synthetic samples, penalize models and apply different algorithms. In our project, we have tried to find the best algorithm without corrupting the data set. To maintain the same proportion as the input data set, we also used stratify while splitting as train and test set. Stratification means that the train_test_split method returns training and test subsets that have the same proportions of class labels as the input data set. To train our different algorithms, We used python Machine learning libraries.We also performed grid Search CV to get the best parameters because it improves the ability to accurately predict results.
As KNN works on distance metrics so it is advised to perform normalization of data set before its use. So we normalized the data using Min_Max_scaler. For Neural Network, we primarily performed our tests and experiments with 4 hidden layers. As we are predicting 7 grades, so in this case softmax was used as activation in our final layer and relu for other layers.

C. *Model from Scratched*

We implemented logistic regression from scratch. As our data set was an imbalanced data set, we used oversampling on training data. To implement logistic regression, we have divided our programs into five different parts which is sigmoid, cost, fit, predict and score.

D. *Compare All Model*

In order to verify our hypothesis, We will finally compare the accuracy and computational time among all five algorithms to see which algorithm performs better.We will summarize our results in a tabular and graphical form in the next section to draw comparisons and insights between each implementation.

IV. Experimental Evaluation

After dropping the unnecessary rows and column like missing values, noise or outliers, duplicate, and wrong data, we are left with 1.6 million rows and 83 features. To reduce the number of features further we examined the correlations between all the features. We dropped features that were highly correlated since these features would give us redundant information. After dropping correlated features the dataset remains with 49 columns. The correlation plot of the remaining 49 features is shown in Fig. 1 where the correlation scale is presented at the right side of the graph.

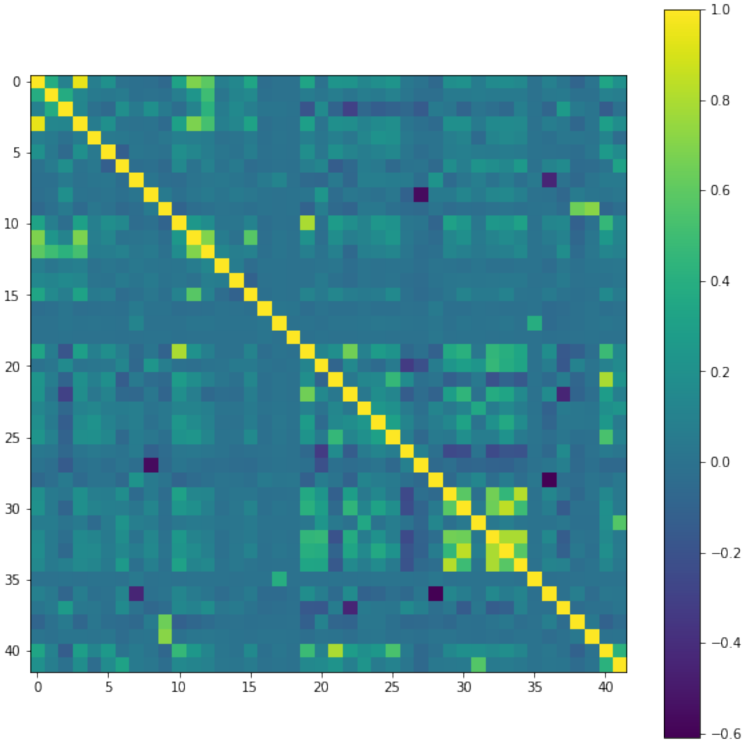

Fig. 1. Correlation Plot

Now, before feeding this dataset into our machine learning algorithm we need to consider our limitations in using such huge dataset. None of us in the group have GPUs in our computers. So, running this huge dataset can take us a full day with our CPU. That's why we have decided to use only 10% of our full dataset with 159587 rows and 49 features while keeping the same proportion of the 7 classes with the main dataset. The reduced dataset was then divided into train and test sets of 70% and 30% respectively.

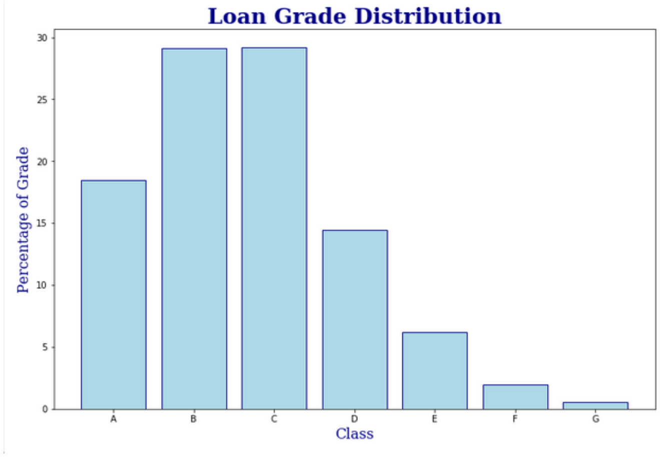

Fig. 2. Class distribution

From the class distribution plot presented in Fig. 2, it is found that the dataset has an imbalanced class distribution.

TABLE I
Best Parameters of the Algorithms

Algorithm	Parameters
K Nearest Neighbors	neighbors=10
Decision Tree	criterion= entropy, max_depth= 18
Random Forest	criterion=gini, max_depth=14, n_estimators=40
Neural Network	layers = 4, neurons = 72 (55, 5, 5, 7), optimizer = adam
Logistic Regression (Scratch)	learning rate=0.1, iteration=20000

Here, the loan grades of A, B, C and D are much more common than E, F and G. To deal with such an imbalanced class classification we can take several approaches like oversampling or undersampling of data, generate synthetic samples, apply different algorithms, penalized models etc. For this project our first priority was to apply different machine learning algorithm and find a suitable one that will give us high accuracy and high F1 score with low computation time. We also applied the oversampling technique, SMOTE (Synthetic minority oversampling technique) to our train dataset and implemented it with logistic regression algorithm.

The parameters of the 5 algorithms implemented in this project were varied. For example, the decision tree algorithm, both the maximum depth of the tree and the criterion for splitting had to be determined. We used grid search validation (GridSearchCV) to find the best parameters for all cases. The algorithms and their best parameters are presented in Table. 1.

After implementing the algorithms, we have compared them with respect to accuracy, computation time and F1-scores. Table II and Table III represents the comparison among the 5 algorithms we have implemented. Also, Fig. 3 shows the graphical representation of the F1 scores for all the 7 classes after implementing the algorithms.

TABLE II
Algorithm Comparison

Algorithm	Accuracy	Computation time (minutes)
K Nearest Neighbors (KNN)	0.55	100
Decision Tree (DCT)	0.96	5
Random Forest (RF)	0.87	40
Deep Neural Network (DNN)	0.87	6
Logistic Regression (LR) [Scratch]	0.68	170

TABLE III
F1-score Comparison

Class	DCT	RF	KNN	LR	DNN
A	0.99	0.96	0.72	0.81	0.93
B	0.97	0.92	0.59	0.63	0.89
C	0.96	0.91	0.56	0.62	0.89
D	0.91	0.89	0.37	0.54	0.81
E	0.84	0.53	0.28	0.52	0.68
F	0.89	0.34	0.18	0.45	0.53
G	0.92	0.14	0.04	0.47	0.47

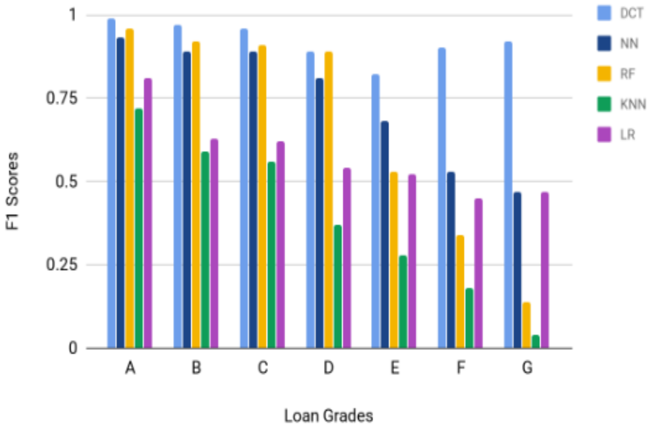

Fig. 3. F1 scores for loan grades with different algorithms

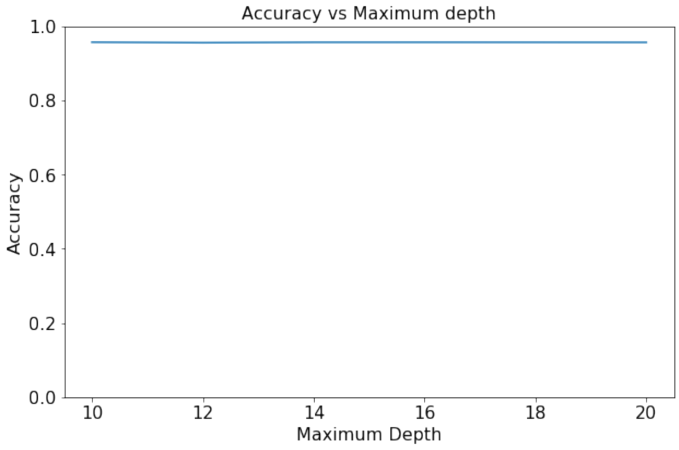

Fig. 4. Accuracy for decision tree classifier in terms of maximum depth of the tree

From Table II, it is found that the best performance in terms of accuracy and computation time is obtained in case of Decision Tree classifier (DCT) which is 96% accurate and ran in only 5 minutes. It not only gives us an excellent performance in terms of accuracy and computation time, most importantly, it provides us good F1-scores for all 7 classes which is shown in the first column of Table III. This means, despite the small amount of data for some classes, the DCT is able to accurately predict these classes. The accuracy from decision tree classifier in terms of maximum depth is plotted in Fig. 4. It is found that accuracy is not changing significantly with the change of the maximum depth.

The random forest (RF) also gives us satisfactory result in terms of accuracy (87%) and computation time (40 minutes). Unfortunately, it doesn't provide good performance in terms of F1-scores for all 7 classes. Especially, for class G, it has an F1-score of 14%. So, we can say that random forest is not a good candidate for predicting loan grades with this dataset.

The worst performance (accuracy: 55%, computation time: 100 minutes) is obtained in terms of K Nearest Neighbors (KNN). It also provides bad performance in terms of F1-scores for almost all our classes. So, KNN should be out of consideration for this particular dataset.

Deep Neural Network (DNN) is showing the second best performance after decision tree classifier (DCT) with an accuracy of 87% and a computation time of 6 minutes. In terms of F1 scores, it provides results better than RF or KNN, but not as good as DCT. We are considering DNN as our second preference. Fig. 5 and Fig. 6 shows the accuracy (train and validation) and loss (train and validation) plot respectively with respect to epochs for DNN.

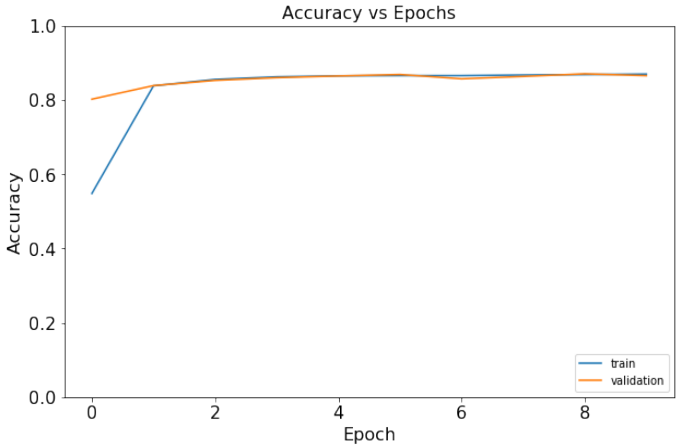

Fig. 5. Accuracy for neural network in terms of epochs

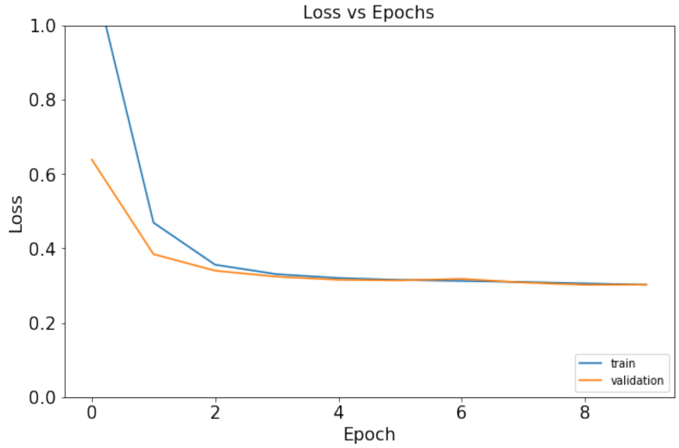

Fig. 6. Loss in terms of epochs

We implemented the logistic regression (LR) model from scratch. Here, we broke down our algorithm into 5 separate functions- sigmoid, cost, fit, predict, and score. Then we used our functions whenever we needed them and ran our program for 20,000 iterations. The major thing we did in case of applying logistic regression is that, we implemented the SMOTE (Synthetic Minority Oversampling Technique) on our training dataset. SMOTE created non-duplicated synthetic samples for the minority classes. The class distribution of the training dataset before and after applying SMOTE is presented in Fig. 7.

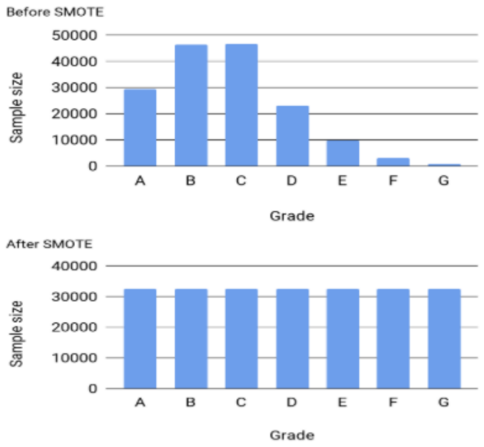

Fig. 7. Class distribution of the training set before and applying SMOTE

Even though, we implemented the SMOTE, logistic regression can not improve the performance significantly. With LR, we get poor accuracy (68%) and F1 scores, and very high computation time (100+ minutes) with respect to other models. Fig. 8 shows the loss vs iteration number for LR model. It can be found that the loss decreases rapidly upto 5000 iterations, and then decreases at a slow rate between 5000 to 20000.

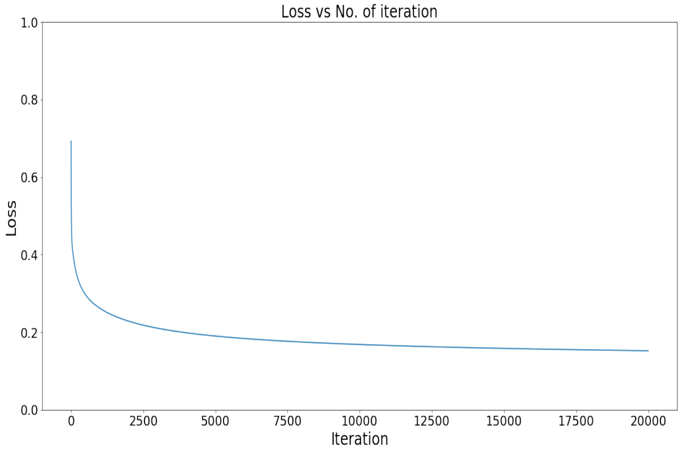

Fig. 8. Loss in terms of epochs

Finally, Table II, Table III and Fig. 3 clearly shows that the DCT is outperforming all other algorithm to predict the loan grades with respect to accuracy, F1 score, and computation time. The DNN has appeared to be the second best candidate for this project.

V. Conclusion and Future Work

From our experiments on the Lending Club dataset, we found that the decision tree algorithm predicted the loan grade with highest accuracy and F1-score. It was 9 % more accurate than the next best algorithm DNN. This algorithm also was quick to run given our computational restraints compared to the other algorithms presented. It had a comparable time to

DNN, but was at least 30 minutes fastest than LR, RF and KNN. One other benefit of the decision tree algorithm is its high F1-scores. Despite the class imbalance, the decision tree predicted the loan grades for classes E, F and G at a much higher rate than the other machine learning algorithms. We also found which algorithms not to use, both LR and KNN had poor accuracy's below 70% as well as extremely high computation times both over an hour and a half. It took an extremely long time to run these algorithms despite only using 10% of the full data set.

We hope that with this research and the models we have could constructed could benefit borrowers to obtaining lower interest rates. These models could give the individual borrowers the tools to better prepare themselves in their quest for getting out of debt. As we mentioned before about 80% of the lending club loans were for debt consolidation and credit card. It is our hope that if this models were used they could better prepare these individuals in the uphill battle of conquering their debt.

For our future work on the topic there are many directions we could take this project. If we had more time, we would want to use more techniques for solving the class imbalance of our data set. We would apply upsampling and downsampling to the algorithms with poor f1 scores to see if they can be improved. Research done on the Lending Club data shows that class imbalance techniques have been successful in improving accuracy. We also had an idea of predicting FICO score from the data set.

References

[1] “Anahita Namvar, Mohammad Siami, Fethi Rabhi , Mohsen Nader-pour”, ”Credit risk prediction in an imbalanced social lending environment”,,”International Journal of Computational Intelligence Systems 11”,”April 2018” , ”1-11”, pdf

[2] ”Traci L. Mach,Courtney M. Carter,Cailin R. Slattery,”,”Peer-to-Peer Lending to Small Businesses”,”6 Feb 2014”

[3] Lending Club Statistic, https://www.lendingclub.com/info/statistics.action

[4] Note Trading Platform, https://www.lendingclub.com/foliofn/rateDetail.action

[5] “MiladMalekipirbazari, VuralAksakalli”, “Risk assessment in social lending via random forests”, “Expert Systems with Applications: An International Journal,June 2015,”Pages 4621-4631”.

[6] Kaggle,”All Lending Club loan data”, https://www.kaggle.com/wordsforthewise/lending-club