Northeastern University
Khoury College of
Computer Sciences

Loan Default Prediction

DS 4400 Final Project

Matthew Martin

Khoury College of Computer Sciences, Boston, MA, U.S.A
Submitted 23 April 2019

Abstract

In this paper, I explored various machine learning architectures, such as logistic regression, support vector machines, random forests, and feed forward neural networks in hopes to classify whether a given loan will default. The data used was collected from *Lending Club*, which is comprised of over 1.3 million data points from loans issued from 2007 to 2018. The csv was found from a kaggle kernel[1]. I tested classification with various features. Two sets of results were found one with features after a loan has been issued such as total payment and the other only using features that are known at time of loan application. This data was used in a Kaggle competition in which the top competitor achieved an accuracy of 70%. My models had varying performance. Through experimental results it was found that certain models such as ensemble methods yielded better results than simple linear models.

I. Introduction

Each year, millions of people apply for loans for various needs. These needs can range from debt to major purchases. When loans are given there is the risk of the loan defaulting, when a borrower fails to repay the loan, and is officially denoted as money lost once the loan is "charged off". A "paid off" loan is when a borrower fully pays back the amount lended. According to *Lending Club* over $44 billion of loans were issued from 2011 to December 2018.

1. Data link: https://www.kaggle.com/wendykan/lending-club-loan-data#loan.csv

Lenders face risk of losing money when issuing a loan, so to mitigate this potential risk I explored if it could be predicted as to whether or not or not a loan will default with a certain borrower. This problem at hand is a binary classification problem. Both continuous numerical features and categorical features will be used.

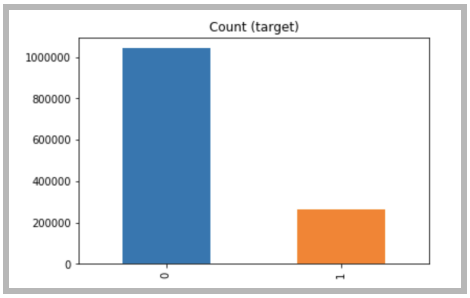
II. Dataset

A. Data Collection

The data comes from *Lending Club*, an agency that sets up borrower and lender partnerships. The dataset consists of loans from 2007 to 2018. There are 145 features and over 1.3 million data points, after only including two classes *paid off* and *charged off*.

B. Data Statistics

The feature of interest for classification was "loan status". For simplicity only loans with loan status of "Paid Off" or "Charged Off" were included for analysis and modeling. The dataset consists of over 1,400,000 paid off (0) loans and over 260,000 charged off (1) loans.

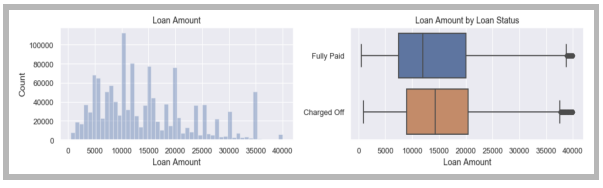


Label Distribution of Loan Status

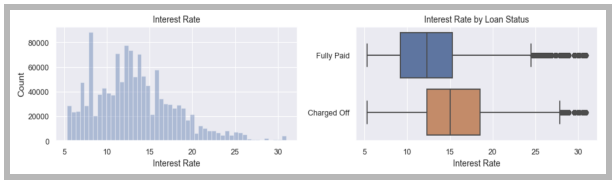
There is an imbalance of labels about a 5:1 paid off to charged off ratio.

Many of the features in this dataset had high counts of missing values. To deal with this I removed features that had more than 30% of values missing. After this the number of features was reduced to 87. Feature exploration was then conducted on this subset of features.

Accompanying the data was a feature dictionary provided by *Lending Club*. This allowed me to get a better domain knowledge of the features. I chose to explore features that are know before a loan is approved for brevity due to the high number of features. After some exploration it seems that loans of higher amounts and interest rates seemed to be charged off more.

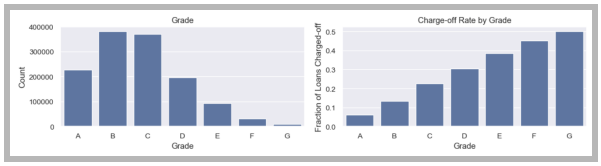


Distribution of loan amount



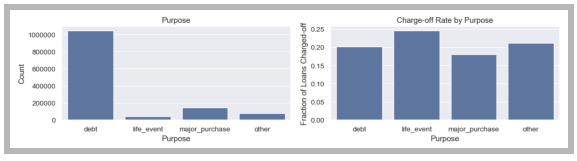
Distribution of interest rate

Loans of lower grade (D and E) seemed to be charged off more than higher grade loans even though they are less represented in the dataset. The grade of a loan is a quality score of a loan based on a borrower's likelihood of repayment based upon historical factors.



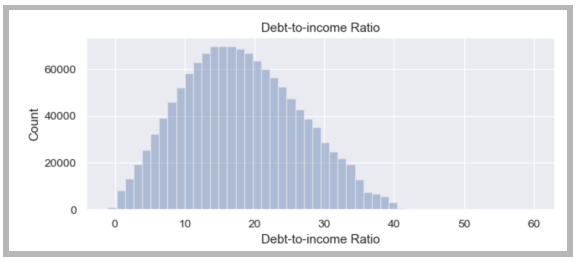
Distribution of loan grade

In terms of loan purpose majority of the data was due to the purpose of debt. Charged off loans were occurred most when life event was the purpose, such as moving or a wedding, followed by other, then debt, then major purchase.



Distribution of loan purposes

The borrowers contained in this dataset had an average debt to income (DTI) ratio of about 18.3. A distribution of DTI across all borrowers can be seen below.



Distribution of borrower DTI

It was also observed that borrowers who opened credit lines more recently seem to be charged off more.

C. Correlations

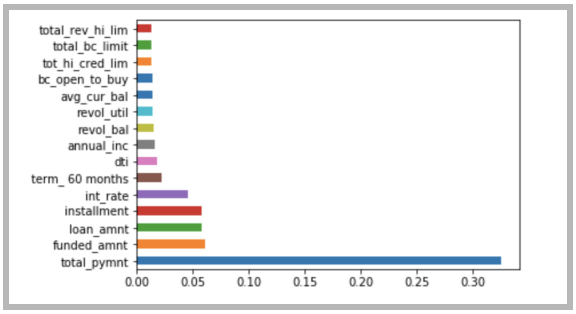
Due to multicollinearity I removed features that were more than 95% correlated with another feature. These features included num_stats, num_rev_tl_bal_gt_0, total_pymnt_inv, funded_amnt_inv, installment, and tot_hi_cred_lim. In addition features with zero correlation to loan status were also removed. These included out_prncp, out_prncp_inv, and policy_code. The top five strong positive and negative correlations can be seen on the next page.

Feature	Correlation
total_pymnt_inv	-0.318038
grade_B	-0.106295
bc_open_to_buy	-0.077103
avg_cur_bal	-0.071585
mort_acc	-0.069613
acc_open_past_24mths	0.100731
grade_D	0.108647
grade_E	0.127284
term_ 60 months	0.175899
int_rate	0.258412

Correlations between features and label.

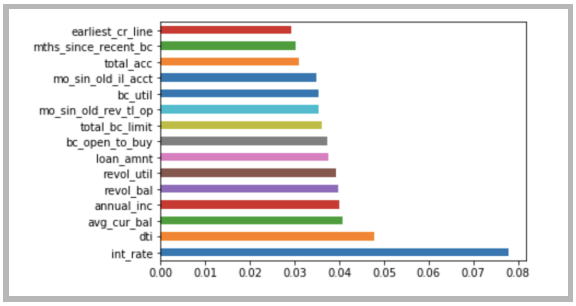
D. Feature Importance

To investigate what features were considered the most important I decided to run a Random Forest Classifier with 250 estimators. The results yielded that one feature was much more important than all others and that was total payment.



Feature importance from Random Forest

After discovering this it was tested to see the importance of features without this highly predictive feature.



Feature importance with only prior known features

It was shown that at the time off application the most important features are interest rate of the loan, DTI, average current balance, annual income, and revolving line utilization rate.

III. Feature Engineering and Selection

A. Feature Engineering

For the sake of feature dimensionality some feature's values were put into groups. For example loan purposes were grouped into debt, life event, major purchase, and other. Another example is state address of borrower which was deducted into five regions: Northeast, Southeast, Midwest, Southwest, and West.

Once these new features were defined all categorical features were turned into dummy encodings. Thirteen features were turned into dummy encodings. Since the data is grouped in years that the loans were issued the issue date was converted into a datetime object. This feature is used to split the data into the training and testing sets and will not be used in any model.

Before running any models features were standardized using z-score normalization.

B. Feature Selection

For feature selection, features that contained too many values, such as employment title and loan title, in which nothing could be inferred were removed. In addition, features that consisted of only one value were also removed. Features that only applied to one class, that would make the data totally separable upon these features. These highly deterministic features were found due to interpretation of the feature dictionary. Some of these features include anything involving recoveries. Recoveries are only associated with charged off loans.

Two sets of features were used in modeling. One set contains all features of the dataset that were kept after data preprocessing. The other set does not contain high feature importance variables that are identified after the loan is accepted. This set will be used to see how well a charged off loan can be predicted at time of application. The subsetted data set contained 41 features that were used in the models.

IV. Models and Results

A. Train/Test Split

This dataset consisted of data that is time series sensitive. Due to this, the issue date of the loans was converted to a datetime object. The dataset was split 80/20 train and test. The testing set contained loans from 2007 up to October 2016 and contains over 1 million samples. The testing set contains loans from October 2016 up to December 2018 and contains over 260,000 samples. For training and testing purposes the paid off class was subsampled for the testing and training sets were balanced.

B. Model Selection

For this classification tasks various models were used to see which yielded the best performance. These models consisted of

linear models, ensembles, and neural network architectures.

First, I wanted to see how the data could do with simple linear models. For this I chose logistic regression, linear discriminant analysis (LDA), and linear support vector machine. With logistic regression I am also able to gain insight as to how the features affect the prediction to get a highly interpretable result. I first ran plain logistic regression, but then used lasso regression to get rid of any noisy features. From this I realized that the data is not linearly separable so I needed to test with nonlinear models. I chose LDA since I prepocessed my data to be gaussian. SVM was chosen to try and play with hyperparameter C to see if an ideal margin could be set.

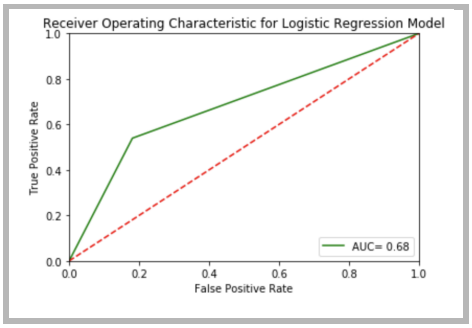
Next, since I needed to implement nonlinear models I chose to implement random forest classifier, adaboost with boosted decision tree, and a feed forward neural network with one hidden layer. I chose these three because I wanted to compare how ensemble methods perform in comparison to complex neural networks.

C. Model Results

For the model results, I will discuss the outcomes of the models when used with the subset of data that known is before loan acceptance. This is allows the models to be more realistic, since every applicant might not have a loan payment history. For this task recall of the positive class (charged off loans) is most important. False negatives are more detrimental in this case. It would be better that someone who applies for a loan is denied , due to being labeled as charged off, and can apply again rather than approving someone who will not actually pay off the loan. Due to this we want to try to elevate recall for charged off loans without losing too much precision.

a. Logistic Regression

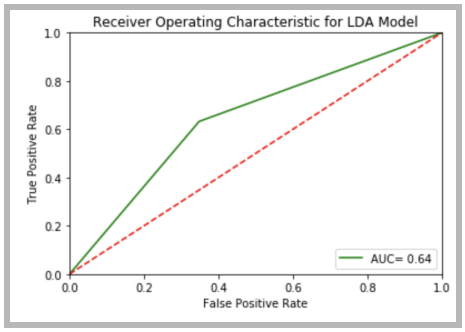
For logistic regression, four thresholds for prediction were used (0.25, 0.50, 0.75, 0.90). Out of these the best performing model had a test accuracy of 63.5%. The recall for charged off loans for this model was 0.67. This model also had the highest precision that did compromise recall with a value of 0.63. The accuracy of this model was 64.2%. That area under the curve (AUC) was 0.68. On the next page the ROC curve can be seen.



ROC curve for logistic regression with 0.5 threshold

b. Linear Discriminant Analysis

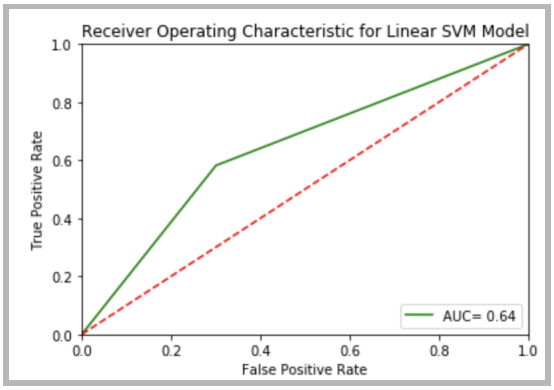
The LDA model had a test accuracy of 64.1%. The recall and precision for charged off loans were 0.68 and 0.63 respectively. The model had an AUC of 0.64.



ROC curve for LDA

c. Linear SVM

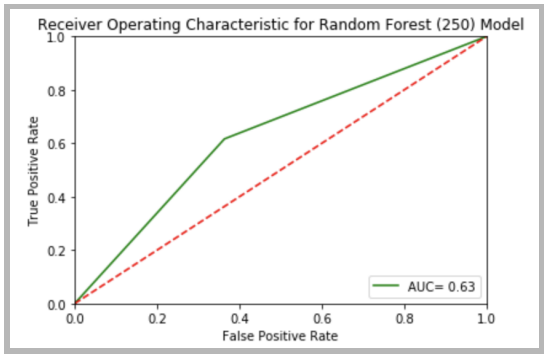
Three linear SVM models were run with C values of 1, 5, and 10. A weight of 2 was given to the charged off class. The SVM with the best results was that with C of 5. The model had a test accuracy of 61.6%. The recall and precision for charged off loans were 0.58 and 0.82 respectively. The model had an AUC of 0.64.



ROC curve for Linear SVM

d. Random Forest

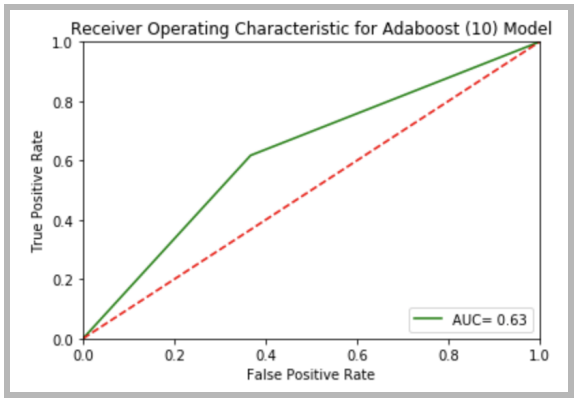
At first with the default settings from sklearn the random forest models were overfitting. To prevent this I made the max depth of each tree 10 and increase the max features to 35. This enables the model to have more selection since many features are not too important. In terms of precision and recall for charged off loans all models performed at the same rate. The precision and recall was 0.63 and 0.66 respectively. The model with 250 estimators had the highest accuracy at 0.6371. The AUC for this model was 0.626. The ROC curve for random forest can be seen on the next page.



ROC curve for Random Forest

e. Adaboost

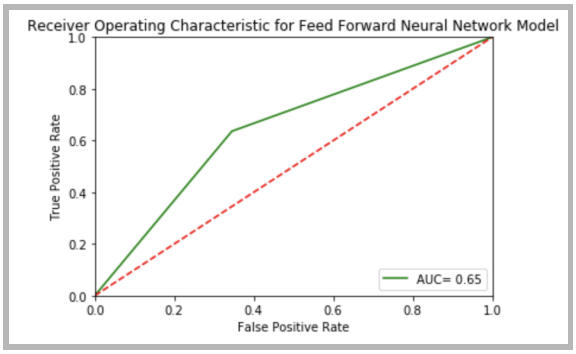
The adaboost models with a decision tree as the weak classifier were run with estimators of 10, 50, and 100. Each decision tree had a max depth of 10 and max features of 35. The best model was that of adaboost with 10 estimators. The precision and recall for the charged off class was 0.62 and 0.66 respectively. The accuracy of this model was 0.625. The AUC of this model was 0.625.



ROC curve for Random Forest

f. Feed Forward Neural Network

The architecture of the neural network consisted of an input size of 54 nodes, followed by a dense layer of 45 nodes, then another dense layer of 28 nodes, and finally a sigmoid to classify each datapoint. ReLu activation was used throughout the layers. Binary cross entropy was used for loss and stochastic gradient descent was used to optimize the model. The precision and recall for charged off loans was 0.64 and 0.68 respectively. The accuracy of the model was 0.645 on the test set. The model had an AUC of 0.65.



ROC curve for Feed Forward Neural Network

D. Model Comparison

When running the model with features known after the loan is accepted such as payment amounts, late fees, and total amounts paid the models perform very well with all models having an accuracy above 90% on the test set. These models are not as realistic though since all this information may not be available at the time of application. Below is the results of all models with both training sets.

Model	Train Accuracy	Test Accuracy	Train Accuracy (Subset)	Test Accuracy (Subset)
Logistic Regression (0.25)	92.9%	98.5%	57.4%	57.0%
Logistic Regression (0.50)	94.4%	98.2%	65.7%	64.1%
Logistic Regression (0.75)	92.8%	96.7%	56.3%	55.8%
Logistic Regression (0.90)	90.2%	94.7%	50.1%	50.4%
LDA	89.6%	94.8%	65.7%	64.1%
SVM (1)	93.9%	97.6%	61.3%	60.1%
Random Forest (10)	99.5%	91.8%	67.3%	63.5%
Random Forest (50)	99.9%	97.8%	67.4%	63.6%
Random Forest (100)	1.0%	98.3%	67.5%	63.7%
Random Forest (250)	1.0%	98.6%	67.5%	63.7%
Adaboost (10)	94.5%	98.8%	70%	62.5%
Adaboost (50)	95.4%	99.2%	76.7%	60%
Adaboost (100)	95.8%	99%	67.4%	58.3%
Neural Network	95.4%	99%	66.7%	64.5%

IV. Discussion/ Future Work

A. Challenges

There were various challenges that appeared throughout the course of this project. First, the size of the dataset was hard to manage and interpret. In addition, the domain knowledge of the task at hand took awhile to interpret and deciding on important features was difficult due to my inexperience in the field of loan lending.

The imbalanced dataset made it difficult to rely on metrics such as accuracy. A balance between recall and precision was difficult to obtain. Subsampling the major class (paid off) may have resulted in some information on the major class. This is not the more detrimental class, so it may not have had much of an impact.

During modeling, it was hard to determine if certain features primarily decided the prediction. This lead to separating the data into two subsets. Another challenge was that there wasn't much of a difference between feature values for loans that were paid and those that were charged off. This made optimizing models very difficult and resulting in the conclusion that it is very difficult to determine if an applicant will be likely to default at the time of application.

B. Insights

Some results were surprising. It was very interesting how well the models can predict what loans will charged off if you use features from after the application is accepted. This can be very beneficial if lenders are able to give the loan in portions and can end the contract if they feel they will not get the full amount back. With the same hyperparameters for the weak classifiers random forest outperformed adaboost. May be that the two models need different hyperparameters to be able to perform at the same rate. It was also interesting how linear and nonlinear models performed at about the same rate. SVM had the best recall and precision for the more detrimental class, but heavily misclassified paid loans as not paying. If the task did not care to about falsely classifying loans that will be paid off, such as a lender looking for more risky loans this model would be the best. For the general problem this would not be ideal even though it is very good at predicting loans that will default before you cannot just deny majority of applicants since paid loans are more likely than charged off loans in the real

world. In the end the neural network had the best performance, but only subtly. In terms of performance and interpretability logistic regression would be the best model to use since a lot of insight can be gained from the model and it performs similarly to the more complex models.

From the logistic regression model, it was able to be inferred the impact certain features had on the determination of whether a loan with be charged off. The features that indicate a charged off loan at lower values are total bank card limit, total number of accounts, average current balance, and annual income. The features that indicate a charged off loan at higher values are interest rate, accounts opened in past 24 months and debt to income ratio. Also, grade D and C loans, loan terms of 60 months and home renters tend to lead to charged off loans. Borrowers from the West are more likely to pay of their loans while borrowers from the Southeast are more likely to default.

Training and test time seemed to have a wide range between all of the models. Linear SVM had the longest runtime for linear models. The increasing amount of estimators made random forest and adaboost extremely time consuming. It may not be worth the time to use these more complex models if they have about the same performance as those that take substantially less time such as logistic regression. A table of training and test times for each model can be seen below. These times, in seconds, are of the model with only features before loan acceptance.

Model	Train Time	Test Time
Logistic Regression	12.8	0.012
LDA	3.62	0.007
SVM	223.07	.014
Random Forest (10)	41.40	0.17
Random Forest (50)	188.64	0.74
Random Forest (100)	353.60	1.39
Random Forest (250)	880.96	3.50
Adaboost (10)	61.18	0.19
Adaboost (50)	288.97	0.69
Adaboost (100)	583.93	1.40
FFNN	353	1.00

Train and test times for each model

C. Conclusion/Future Work

After feature exploration and interpreting model results can be deduced that predicting whether a loan will default or not is a very difficult task at the time of application. Many of the features are very similar between classes. In addition, in the real world there is a high imbalance on classes. Most of the time a loan will be paid off, so for this problem it is more important to be able to catch when a loan will default. Based on the results the neural network had the best performance, but due to its lack of interpretability in practice I would go with logistic regression due to its high interpretability. The interpretability of this model would allow employees to explain to both lenders and borrowers the findings in a real world setting. These model predictions can be flags for investors looking to invest in the right loans. Investors should watch out for lower grade loans (determined by Lending Club), borrowers from the Southeast of the U.S.A., loans with high interest rates, and borrowers with higher than average debt to income ratios because these indicate a higher probability of the loan defaulting.

In the future, it would be interesting to see if it is possible to create custom loan programs for borrowers based on his or her credentials. Another, direction would be to explore a regression analysis of the loan data to calculate interest rates for loans based upon their features and borrowers.