Nishi Mishra
29 October 2019

Lending Club Loan Data Analysis

The Project:

Company Information:

Lending Club is a peer to peer lending company based in the United States, in which investors provide funds for potential borrowers and investors earn a profit depending on the risk they take (the borrowers credit score). Lending Club provides the "bridge" between investors and borrowers.

Problem Statement:

Questions to be Answered:

We will use data science and exploratory data analysis to take a peek Lending Club's loan data from 2007 to 2011, focusing on the following questions regarding this period:

Loan Absolute Variables Distribution: How does loan value, amount funded by lender and total committed by investors distribution looks like? Applicants income range: Range of Applicants income for both good and bad loans

Defaults Volume: How many loans were defaulted?

Average Interest Rates: What was the range of interest rate for the loans?

Loan Purpose: What were the most frequent Loan Purposes?

Loan Grades: Variation of interest rates for the different grades of loans

Delinquency Breakdown: How many loans were Charged Off(Bad loans)?

How does the loan data distribution look like? Using Data Science, we will paint a picture detailing the most important aspects related to the loans and perform EDA (Exploratory Data Analysis).

Analysis of loan Grades

Can we create a better, optimized model to predict credit risk using machine learning?

Can we increasee the precision of the model?

By analyzing these aspects, we will be able to understand our data better and also get to know a bit of Lending Club's story. The dataset contains 43K loan applications from 2007 through 2011 and it can be downloaded from the url www.lendingclub.com.

The Methodology:

Libraries:

- pandas for:

data loading, wrangling, cleaning, and manipulation, feature selection and engineering, descriptive statistics

- numpy for:

array data structure, the primary input for classifiers, model comparison ,matrix manipulation

- imblearn for:

oversampling imbalanced dataset

- scikit-learn for:

classifier models ,model evaluation

- matplotlib for:

data visualization

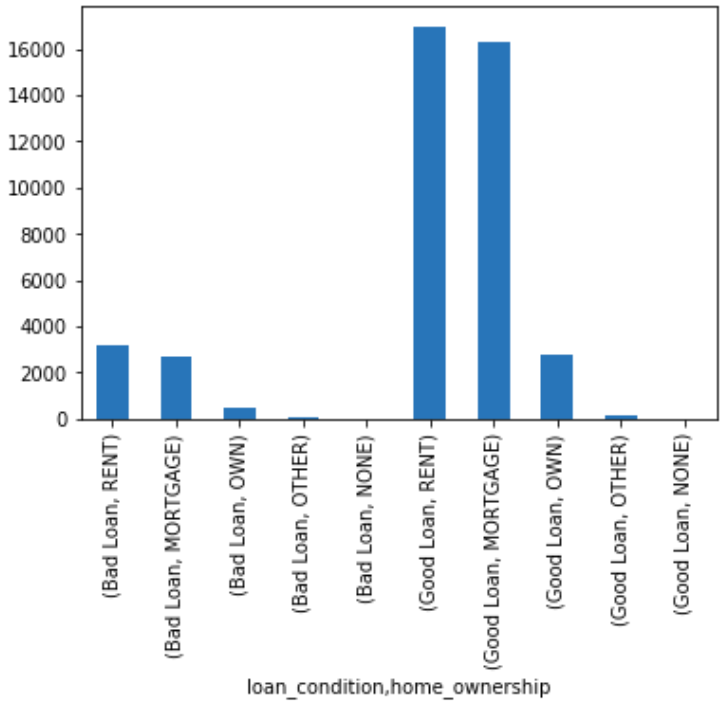
Data Wrangling and cleaning:

• The dataset was directly obtained from the url www.lendingclub.com .
Initially the shape of the data showed only 1 column which was due to a single heading present in the cvs file. After the deletion of that row the data set represented approx 43000 rows and 150 columns
• Several columns were renamed to make them readable
• The '%' sign from the interest rate column was removed and the column was converted into float data type for the ease of computation/visualisation.
• The employee length had two rows of object data type due to boolean / arithmatic operators(<,+) present in it. It was modified for the sake of calculations.
• In the complete dataset, 15% of the loans were classified as bad loans which gives us an imbalanced dataset.This problem will be dealt on later stage.
• "Charged Off", "Default", "Does not meet the credit policy" Charged Off", "In Grace Period","Late (16-30 days)", "Late (31-120 days)" were classified as Bad Loans and all others as Good Loans which was done under the column loan_condition . This column was our target variable.
• Logarithmic transformation of annual income was done due to large values.
• Their were 30,000 unique employee titles/designation which would be of minimal usage hence, this column was dropped.
• column "id" was made the index
• Columns having more than 30% of the values missing/Nan were dropped which resulted in 58 columns left for model training.
• The following histogram shows the incompleteness:
From the above histogram, we see there's a large gap between features missing "some" data (<20%) and those missing "lots" of data (>40%). Because it's generally very difficult to accurately impute data with more than 30% missing values, we drop such columns.
List other potential data sets you could use.
There were many other latest quarterly(Q1,Q2,Q3,Q4) dataset present on the lending club website but as the number of columns were very less hence the large dataset was used.
Also , several other cleaned datasets were also present on Kaggle but using the dataset directly from the website was much more of a realistic approach.

Initial findings.

• The initial findings seem to suggest a positive correlation between several features of the applicant with the loan status(target variable) which can be utilised for the model building.
• Also, the verification status did not have any significant impact on the loan target variable
• This shows that 41.3% of the bad loans were not verified.
• The fico_score mean of the good loans is more than 700 and most of the values of fico_score lies near 700 score hence the fico_score for good loans should be closer to 700.
• The majority of loans is either graded as B or C — together these

correspond to more than 50% of the loan population. While there is a considerable amount of A graded or “prime” loans (~17%), there is a small amount of E graded, or “uncollectible” loans (~0,06%). Which is a good sign for Lending Club.
• Number of bad loans were less in the category A jobs and it had more number of good loans as compared to the other category.



Most of the applicants had either rented or mortgaged conditions under home ownership.

Statistical inferences:

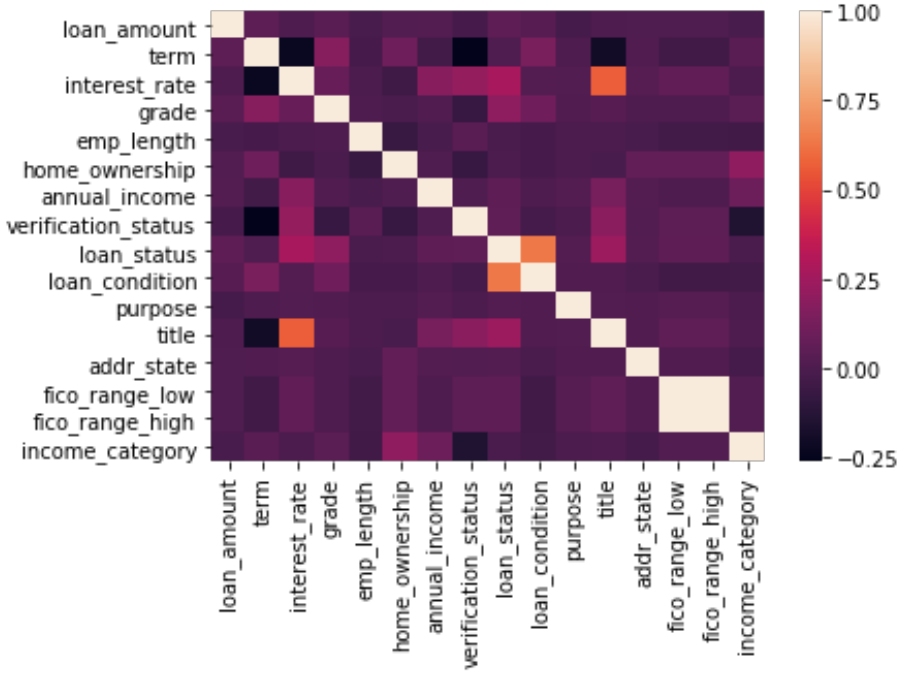
The direction to take the statistical analysis for this project was very clear once the data was obtained and visualised. It was very apparent that there was a positive correlation between the Loan Status and the following features:

- Interest rate
- employment length
- grade

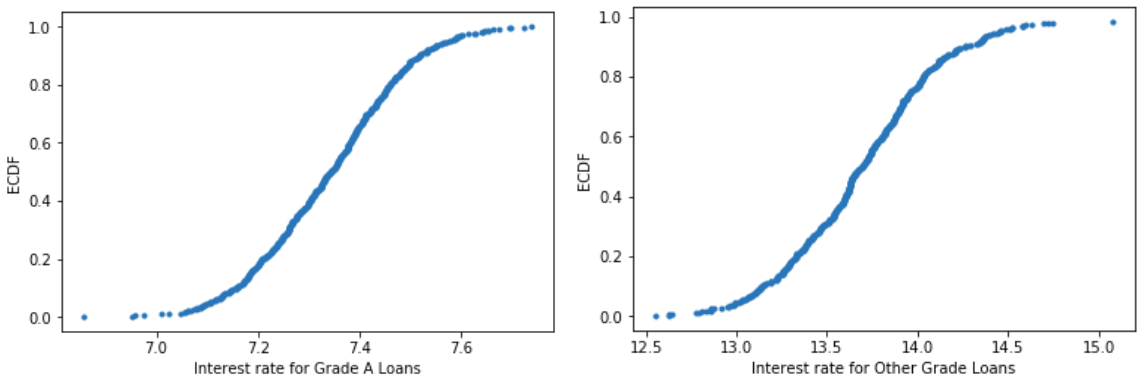
- fico range
- income
- purpose etc.

Employee length, annual income and fico scores have strong positive correlation with the loan amount .

Also the following heat map was plotted for the correlation among the features.



We also plotted the ecdf for interest rates for A grade loans and Interest rates for the other grades apart from a grade.



This observation made us to perform a Frequentist test on the interest rates.

The test was performed to check whether the interest rates offered for A grade loans were lesser than the other grades hence we did a single tail Welch's t-test as the variance is not equal.

$H0$: The interest rates offered for other grade loans is greater than the A grade loan.

H1: The interest rates offered for other grade loans is not greater than the A grade loan.

On running the test the following results were obtained:

- Ttest_indResult(statistic=324.96013788823643, pvalue=0.0)
- The confidence interval for A grade loans is: (7.339761200000001, 7.327202348671294, 7.352320051328707)
- The confidence interval for all other loan other than A grade loans is: (13.687281466395113, 13.651015426092169, 13.723547506698058)

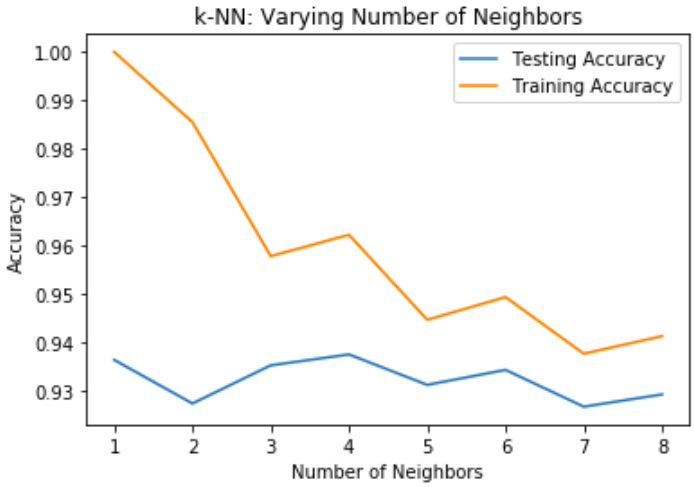
As the p value was less than 0.05 hence the null hypothesis was rejected .

<u>Machine Learning techniques used:</u>

We built our first model:

1)**<u>Knn Classifier:</u>** We got an accuracy of 0.93 . As our model has good loans dominating the sample we may need to over sample our data with bad sample.

We then now checked the effect on the accuracy by changing n in the model and the graph is plotted below:



From the above graph we can see that we get best result at n= 4 which gives us training accuracy as 0.96 and testing accuracy as 0.94

Due to imbalanced data out score results showed 1.0 while running RandomForest and DecisionTree Model as 85% of the loans were good(target variable).

Hence we used two oversampling techniques:

2) **imblearn(SMOTE)** for oversampling and **RandomForestClassifier** :

We see that the recall score and precision score comes out to be 1 even after oversampling the dataset with the minority class, which does not seen to be right. One of the reason can also be because of the dominating features.

We then checked the important features in the model and observed that few features such as: loan_status, loan_condition_int, recoveries, collection_recovery_fee, last_credit_pull_d, total_rec_ late_fee, total_rec_int, instalment, total_pymnt, last_pymnt_d, total_pymnt_inv are to be dropped due to there insignificance/ unavailability before the approval or similar terms to the loan_condition(target variable).

Hence we drop the above columns from the new data set and train our model again.

and got the following results using RandomForestClassifier(n_estimators=1000, random_state=123):

Validation Results

0.8879022147931467

0.9444308145240432

Test Results

0.8820985332831892

0.9388568896765618

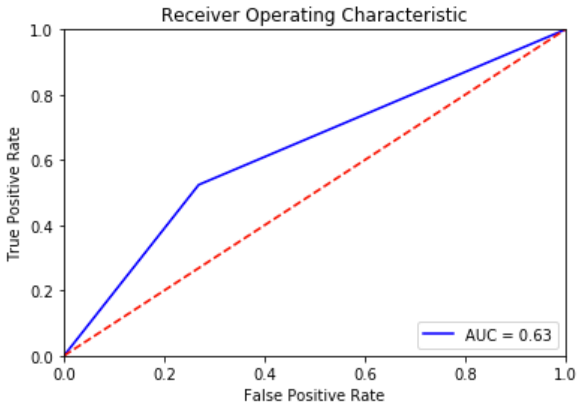
Oversampling improved our results to a good extent.

Next, we did **manual oversampling**:

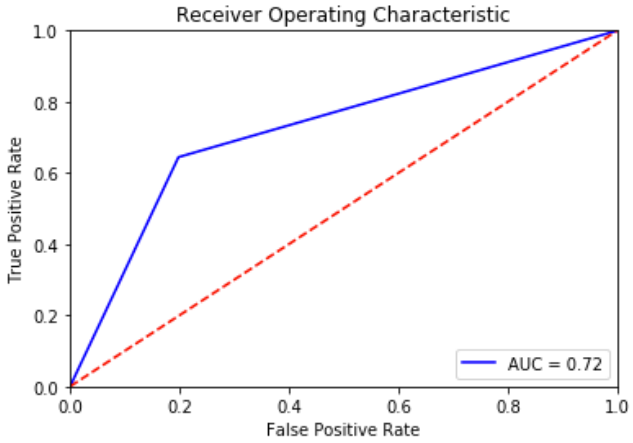
It shows that out of 43000 datasets 6431 data represents the bad loans . So to make our dataset balanced we will create a new dataset "balanced" comprising both the type in 3:2 ratio.

We used two models separately on our new (balanced)dataset to plot the ROC curve using the manually over sampled data.

3) KNN Classifier : Accuracy: 0.64



4)Logistic Regression: Accuracy score of :0.72



The results seems realistic but we can ry some more methods to improve our accuracy and hence the precision.

We will be proceeding for an Ensembled Learning Model to give much better result and hence we will now observe the performance of Six models together using a FOR loop.

The six models used below are:

1)LogisticRegression	Cv_score(mean)=0.72	recall_score=0.73	precision_score=0.74
2)DecisionTreeClassifier	0.88	0.88	0.89
3)LinearDiscriminantAnalysis	0.74	0.73	0.74
4)SVC	0.60	0.5	0.3
5)KNeighborsClassifier	0.64	0.62	0.63
6)MultinomialNB	0.53	0.53	0.56

It seems like **logistic regressio, Decision Tree and LDA models** agree to each other.Hence we will finally use this models into our Ensemble Techniques as our main aim is to increase the Precision and accuracy(decrease the False Positives which are more dangerous in our case).

5) Ensemble Model

This method combines the decisions from multiple models to improve the overall performance. This can be achieved in various ways.The first method which we would use is **Max Voting.**

a)Max Voting:

The accuracy score of the test model using the Max Voting Ensemble Method is:

0.77

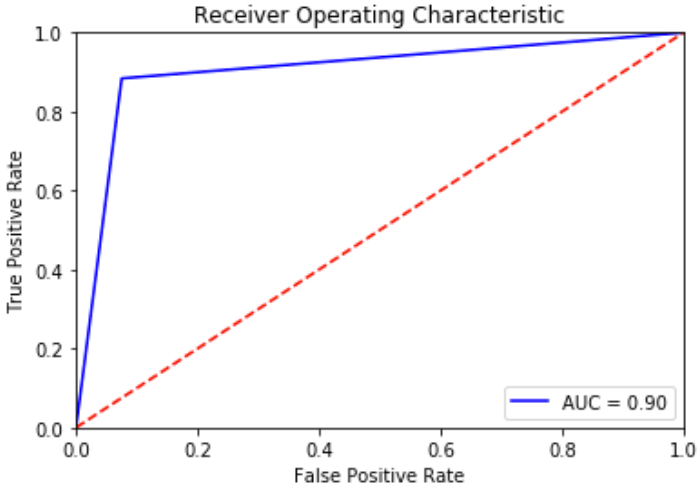
b)Advanced Ensemble Method of Bagging meta-estimator

Bagging meta-estimator is an ensembling algorithm that can be used for both classification (BaggingClassifier) and regression (BaggingRegressor) problems. It follows the typical bagging technique to make predictions. Following are the steps for the bagging meta-estimator algorithm:

The score using the Decision Tree bagging Classifier is:

0.91

The ROC Curve using above model is as follows:



Results:

We obtained good accuracy results using two models:

1) Advanced Ensemble Method of Bagging meta-estimator

	precision	recall	f1-score	support
0	0.9209	0.9341	0.9274	1608
1	0.8979	0.8784	0.8880	1061
accuracy			0.9120	2669
macro avg	0.9094	0.9062	0.9077	2669
weighted avg	0.9118	0.9120	0.9118	2669

2) Random Forest Classifier:

Validation Results

Score: 0.8879022147931467

Recall Score: 0.9444308145240432

Test Results

Score: 0.8820985332831892

Recall Score :0.9388568896765618

Limitations:

The dataset had imbalanced data in which the good loans were dominating. A lot of data was missing which might have affected our model. Various factors such as AGE, NET WORTH, EXISTING LOANS, EMI DEDUCTIBLE EVERY MONTH, etc. if were present we would have given better predictions . As the customer's repayng capacity depends on many other things which were not present in the data set hence model might not give 100% correct results.

Conclusion:

We applied machine learning methods to predict the probability that a requested loan on LendingClub will charge off/turn into a bad loan. After training and evaluating with different models (logistic regression, random forest, and Decision tree)and finally using ensemble method we found that in all the cases Decision tree performed the best.We selected Decision tree Bagging Classifier as our final model with AUC SCORE 0.9 on a test set.

Using this model can provide a somewhat informed prediction of the likelihood that a loan will charge off, using only data available to potential investors before the loan is fully funded.The major importance by far was found as the FICO score and the annual income. Hence, deviating from the basic prnciples of banking(FICO) would be risky for an investor.

We also found that, according to the Pearson correlations between the predictors and the response, the most important variables for predicting charge-off are the loan interest rate and term, and the borrower's FICO score and debt-to-income ratio.

Scope for Future work:

We can also try Boosting and Stacking method to give much better results. As our major concern was to increase precision these other models can help us to to the same.

Some more datasets of different quarter can be downloaded directly from the www.lendingclub.com website and the model can be tested for much better accuracy.

All of the important features can be examined for further correlations/patterns. A model predicting ROI and diversification of funds allocation based on the risk factor can be generated.

Risk assessment for specific customer types could be carried out using these same techniques.

Safest category of investment can be predicted and given to the clients so as to ensure good return on investment.

The project could be wrapped in a web application and tested on real world data, asking loan applicants to input their info and report back their results, which could be compared to the model's prediction. When new quarterly data are published, the model's predictive capabilities can be tested in earnest.

This projects also acts as a proof-of-concept for application analysis. It could easily be applied to other types of applications for which large data sets exist. For example, Credit card defaults.

Client Recommendations

To minimise the risk of a costly rejection, lending club should screen the applicants and choose the best qualified, with the lowest chances of rejection. Screened applicants should then have their applications prepared either in house or in the office itself

If all available candidates are risky, the company should look at the application features that are most significantly increasing risk and alter the application to reduce risk.

All the cases should be verified as we saw 41% of unverified loans turned Bad. Hence verification is highly recommended.