Project 1: Loan approval and rejection based on historic borrower data

Contents

1. Introduction

The most critical challenges in the lending industry are as below.

- The credibility of a borrower.
- Will borrower able to pay the entire loan amount with interest.

Investors provide loans to borrowers in exchange for repayment of the original amount and interest. The lender makes a profit with interest amount. If borrowers do not repay the loan, then the lender loses money. Many of the loans are not entirely paid off on time. However, some borrowers default on the loan.

We are going to use publicly available historical data from LendingClub.com and try to address the above issue by cleaning data and create a model to predict whether borrowers are likely to pay of default on their loans.

Data Source:

- https://www.lendingclub.com/info/download-data.action
- https://www.dataquest.io/blog/machine-learning-preparing-data

2. Data Set Description and Cleaning

I. The Dataset

Lending Club site, you can select different year and range to download dataset in CSV for both approved and rejected loans.
Data Dictionary is also available on same site with detail description of each features (columns).
The approved loans dataset contains information on current loans, complete loans and defaulted loans. Working data extracted from approved loans for the year 2007 to 2011.

lcdatadictionary.csv

II. Data Wrangling and Cleaning.

Load CSV file without first row on Pandas Data Frame because It contains text instead of columns details.
Drop any Columns having 50% row with Null value. We got data frame with 42538 Row and 58 Columns.
Load Data dictionary CSV to get description of columns to understand the approval dataset columns required or not required. Some columns may be meaning less or future data leak .

```
[177]: 1 Data_Dict.style.set_properties(subset=['description'], **{'width': '1000px'})
```

[177]:

	loanstatnew	description
0	acc_now_delinq	The number of accounts on which the borrower is now delinquent.
1	acc_open_past_24mths	Number of trades opened in past 24 months.
2	addr_state	The state provided by the borrower in the loan application

Below are several steps used for data cleaning

- Remove columns (features) which not required in machine learning like 'desc','url' 'zip code' etc.
- Remove columns (features), which values are unique or appear less than 4 times.
- Handles missing value by dropping rows if those columns have 3 or 4 missing row.
- Handles missing values by mean values if missing row counts is more than 4 or 5.
- Select object columns and worked on data conversation required for machine learning like 'emp_length',' grade' etc.
- Converted categorical value in binary for machine learning 'loan_status'

Loan Status feature on approval dataset.

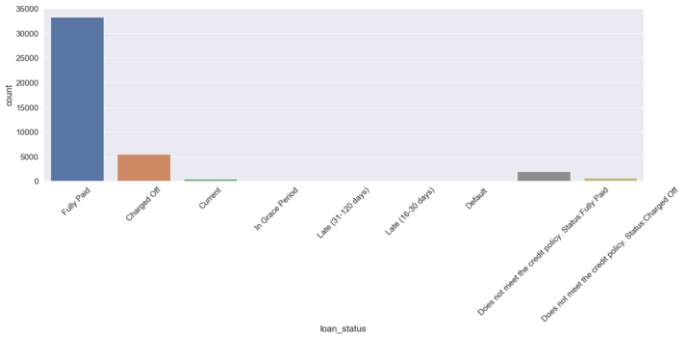


Fig 2.2.1 Bar graph of loan status

Loan Status feature after converting in binary

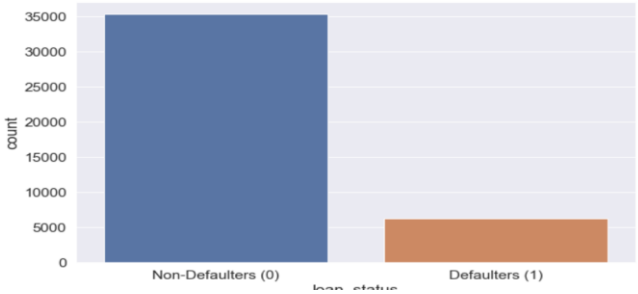


Fig 2.2.2 Count for non-defaulter and defaulter

- Finding outliers and removing from dataset because it can affect model prediction like annual income. So remove value above 99.5% of quartile.

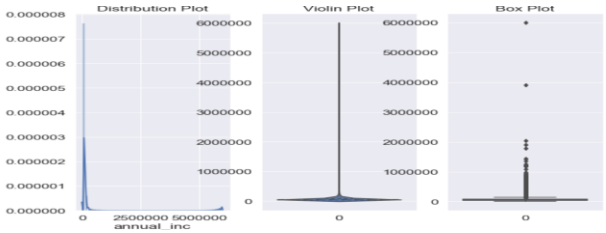


Fig 2.2.3 Annual income plot for outliers

3. Exploratory data analysis.

- Created Correlation heat map, pair plot and cluster map to identify correlation between different features with loan status.
- Boxplot between purpose of loan vs loan amount with loan status

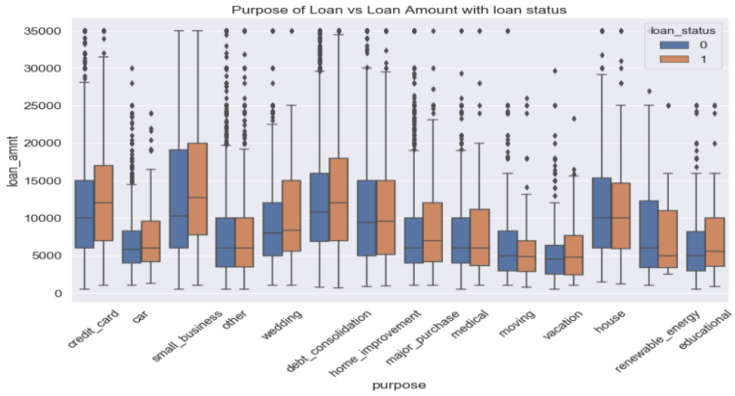


Fig 3.1 box plot purpose of loan vs loan amount with loan status

We can observe from above graph that wedding and major purchases are more towards defaulter.

- Seaborne count plot between purpose and loan Status

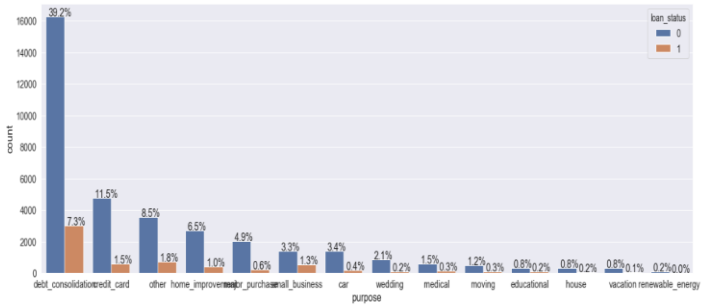


Fig 3.2 bar plot purpose of loan vs count of non-defaulter and defaulter

We can observe from above graph that 46.5% loans were to repay the previous debt and defaulter percentage is more.

- Seaborne count plot between payment term and loan Status

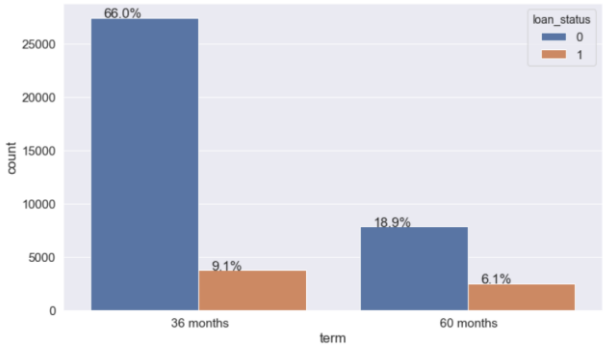


Fig 3.3 bar plot loan term vs count of non-defaulter and defaulter

We can observe from above graph that defaulter percentage is more in 60 months loan payment term.

- Seaborne count plot between employment length and loan Status

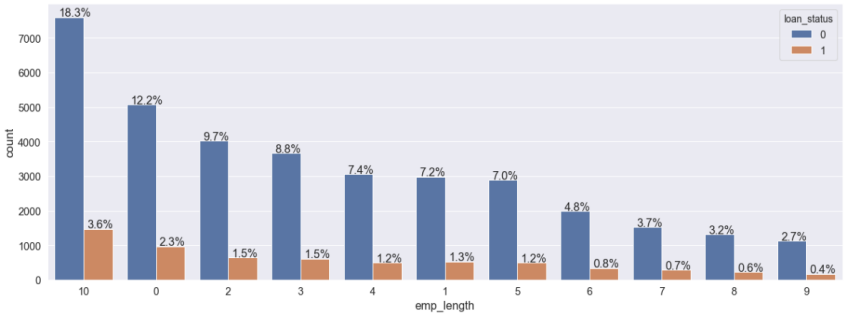


Fig 3.4 bar plot employment lenght vs count of non-defaulter and defaulter

We can observe from above graph that defaulter percentage is more for 5 to 9 years employment length.

After the EDA on approved dataset for different columns with target columns 'loan status' below are impotent features should consider Emp_Length; term; loan_amnt; grade; int_rate; purpose; annual_inc

4. Results and In-depth analysis using machine learning

Split data set in training and test data set. Training data set is 80% and test data 20%. Scaled target features using SMOTETomek.

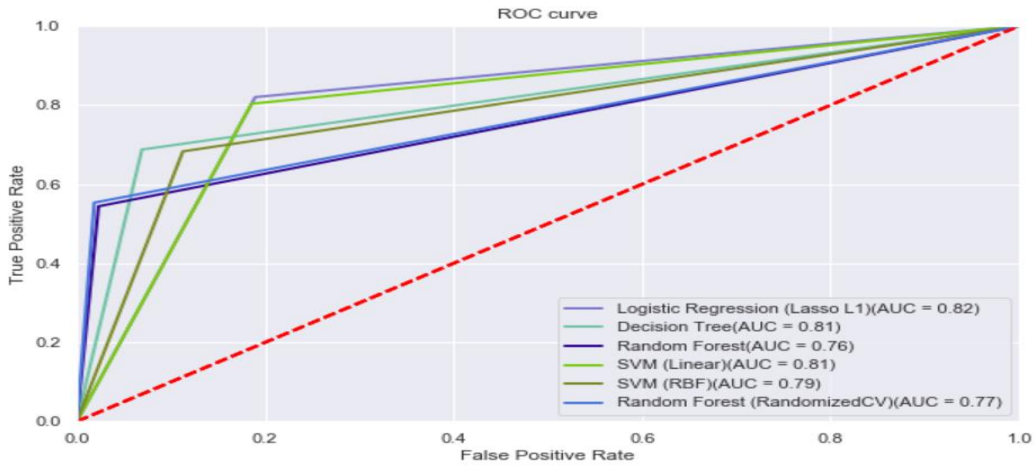
The process of model building is not complete without evaluation of model performance. Suppose we have the prediction from the model. How we can decide whether the prediction are accurate? We can plot the result and compare them with the actual values. Calculate the distance between the prediction and actual values. Lesser this distance more accurate will be the prediction. Since this is classification problem, we can evaluate our models using metrics. Like Accuracy, Precision, ROC curve etc.

Below six machine learning classification models have been used for prediction

- Logistic Regression
- Decision Tree
- Random_Forest_n=100
- SVM (Linear)
- SVM (RBF)
- Random Forest Random CV

.

	Model	Accuracy	Precision	Recall Score	F1 Score
0	Logistic Regression	0.812102	0.444398	0.820216	0.576464
1	Decision Tree Classifier	0.892939	0.647529	0.687500	0.666916
2	Random_Forest_n=100	0.909419	0.813149	0.543981	0.651872
3	SVM (Linear)	0.812703	0.444302	0.803241	0.572135
4	SVM (RBF)	0.855527	0.528358	0.682870	0.595759
5	Random_Forest_RandomCV	0.914712	0.846517	0.553241	0.669155



Chose Model.

By Looking score report for all six model and ROC curve, we can say random forest random CV is best-fit model for prediction.

We used test data on random forest random cv model and prediction result is similar quiet similar with actual loan status.

GitHub repository for code.

https://github.com/sunilww/CapstoneProject

5. Conclusion

From a proper analysis of positive points and constraints on the component, it can be safely concluded that the product is a highly efficient component. This application is working correctly and meeting to all Banker requirements. This component can be easily plugged in many other systems. There have been numbers cases of computer glitches, errors in content, and most prominent weight of features is fixed in automated prediction system, So in the near future, the so-called software could be made more secure, reliable, and dynamic weight adjustment. In the near future, this module of prediction can be integrated with the module of the automated processing system. The system is trained on old training dataset in future software can be made such that the new testing date should also take part in training data after some fixed time.

6. References

https://towardsdatascience.com/predicting-loan-repayment-5df4e0023e92

https://www.lendingclub.com/info/download-data.action

https://www.dataquest.io/blog/machine-learning-preparing-data