Predicting Loan Payment Status

Group 7

Deepti Batra - MT18244
Diksha Sapra - MT18245
Arunav Dutta - MT18156
Pramil Panjawani - PhD19008

Problem Statement

In recent times the number of default loan payments has been increasing which has caused a sort of crisis in the banking setup and has impacted the economy. So in order to counteract it we have tried to come up with an algorithm which based on set of attributes determines if the customer would be able to pay back the loan.

The Datasets and Data Acquisition

We got the data from a firm called lending club.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.
We have selected this data set by creating an account on their site, and downloading the publicly available dataset.

Some of the columns available are Credit policy, purpose,interest rate, installment, log annual income, debt to income ratio, etc

The original data set had a lot of null value and blank spaces. Also not all the columns given in the dataset were useful. We check the frequency and position of null values and then treat it accordingly by substituting mean, median, mode or deleting the row/column entirely. We fill public records,last six month data,Delinquency 2 years,revol util with mode, log of annual income with mean and drop the row for the number of days borrower had a credit line.

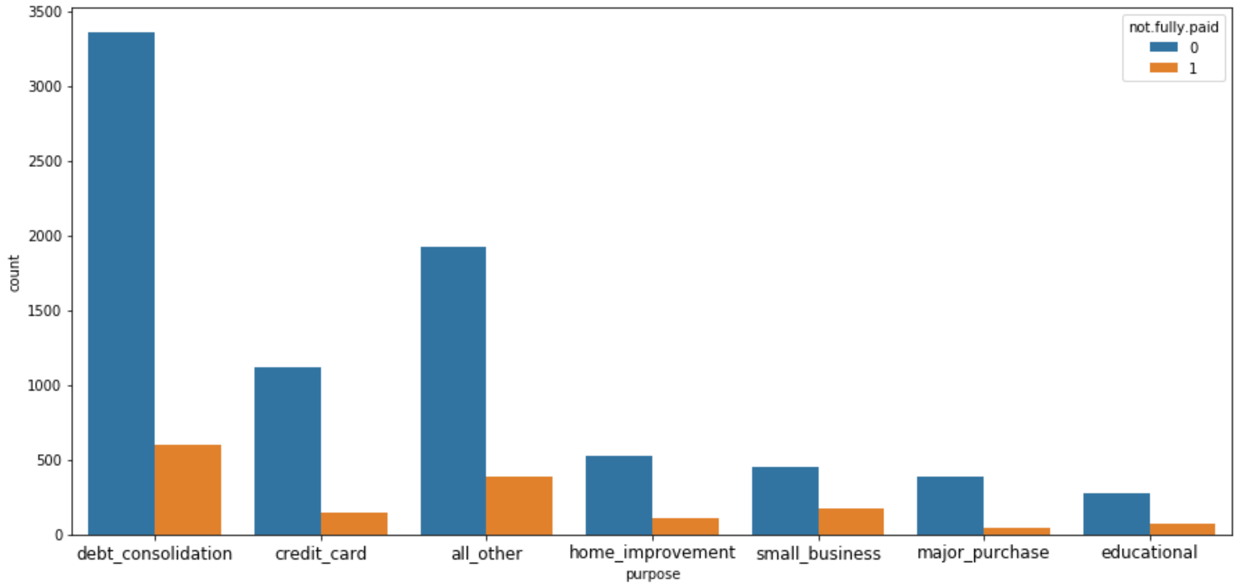
Hypothesis

We try to find a relationship between attributes and loan payment status
Some of the Hypothesis we propose are:-

1) Loans taken for Debt consolidation form a major chunk of defaulted payments again.
2) Higher interest rate is charged to people who have a history of bad loans.
3) Recent financial status should be given more importance while giving loans.
4) Higher the FICO Score, lower the interest rate.
5) Depending on all these factors we can easily identify who will be able to pay back the loan and might not.

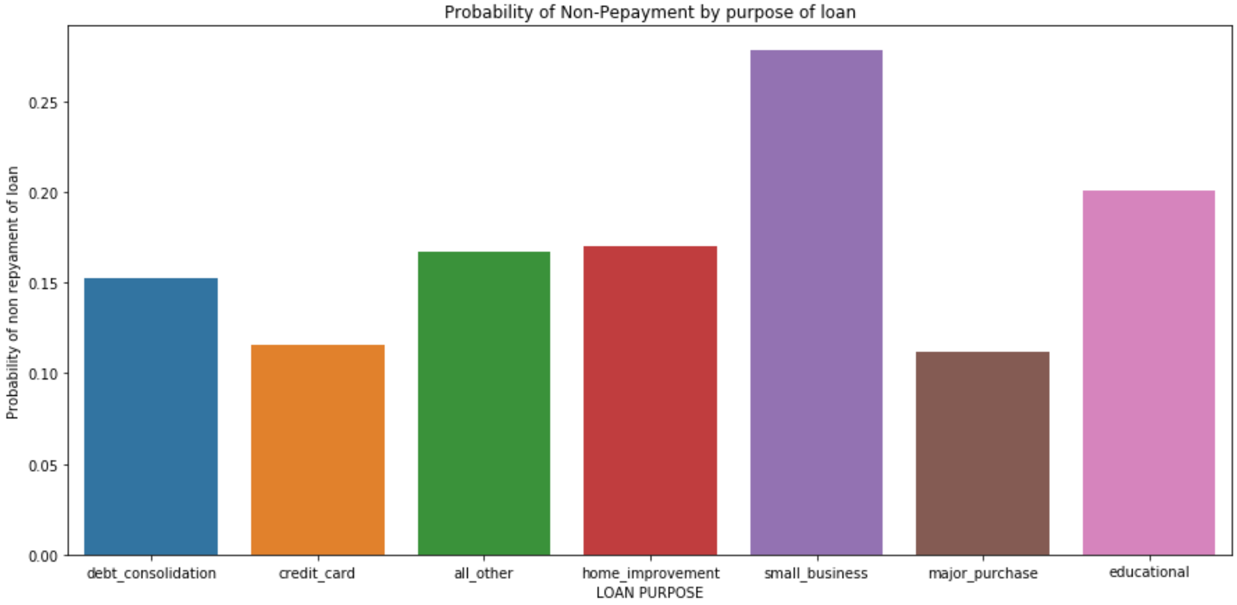
Visualization and Statistics

1) Loan status depending on the purpose



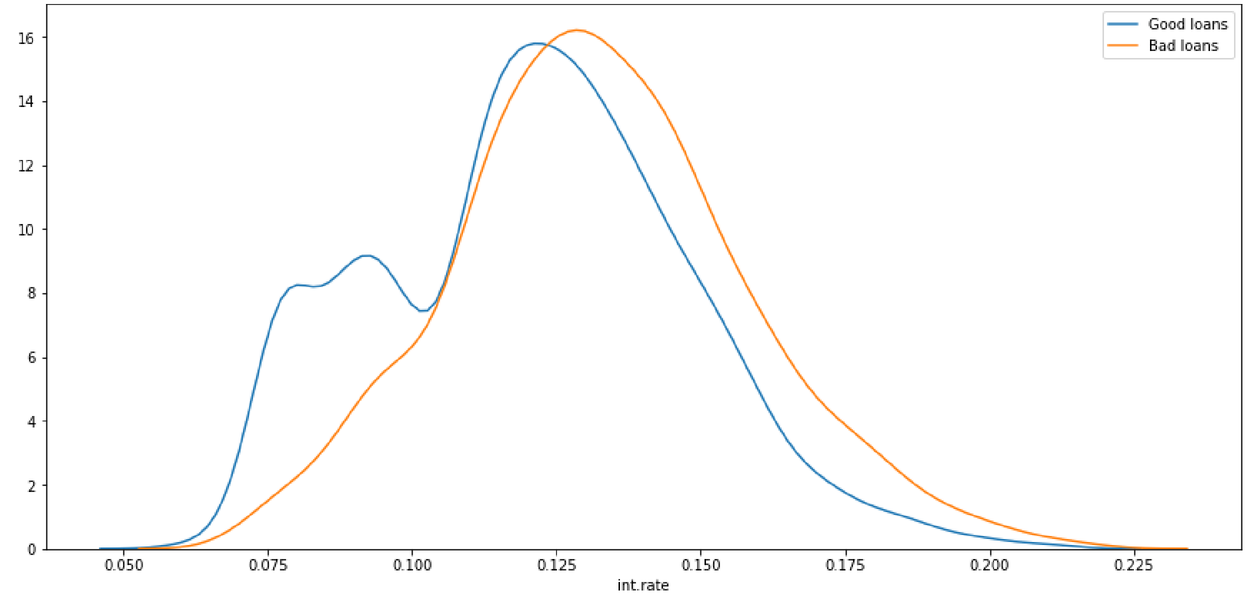
We can see most of the loans are taken for debt consolidation and credit cards.

2) Where do most loan defaulters come from



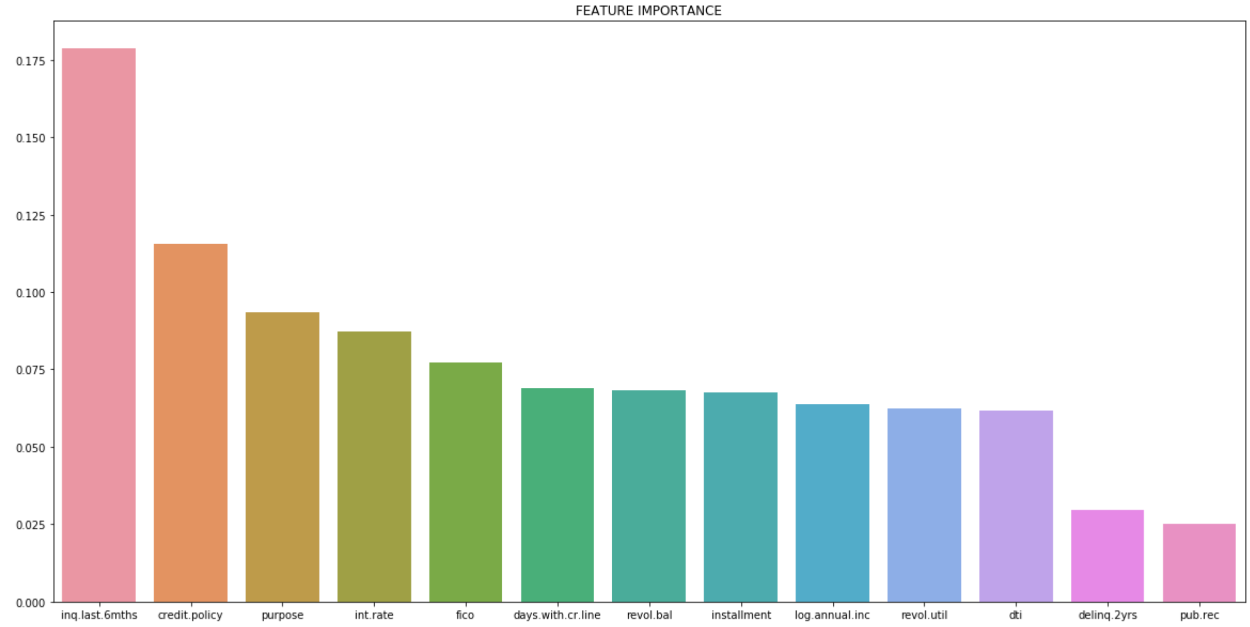
Most of the defaulters come from Small Businesses, then education loans followed by home improvement.

3) Relation between bad loans and interest rates



We can see the graph is skewed towards the left(higher rates) bad loans meaning people with already a poor history have to pay higher interest rates.

4) What 5 factors affect loan repayment the most?



5 most important features are: inq.last.6mths, credit.policy, purpose, int.rate, fico

5) Best and 5 worst borrowers from the data

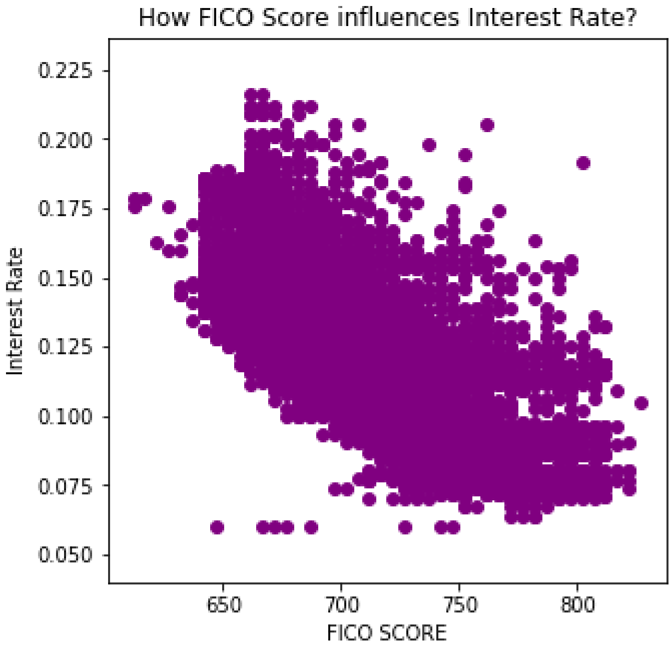
Using our models we can even produce which customers have a higher probability of returning the loan

Index	977	1670	1262	1289	1535
Probability of repaying	1.0	1.0	0.99833	0.99833	0.99833

Similarly we can find customers least likely to be able to repay

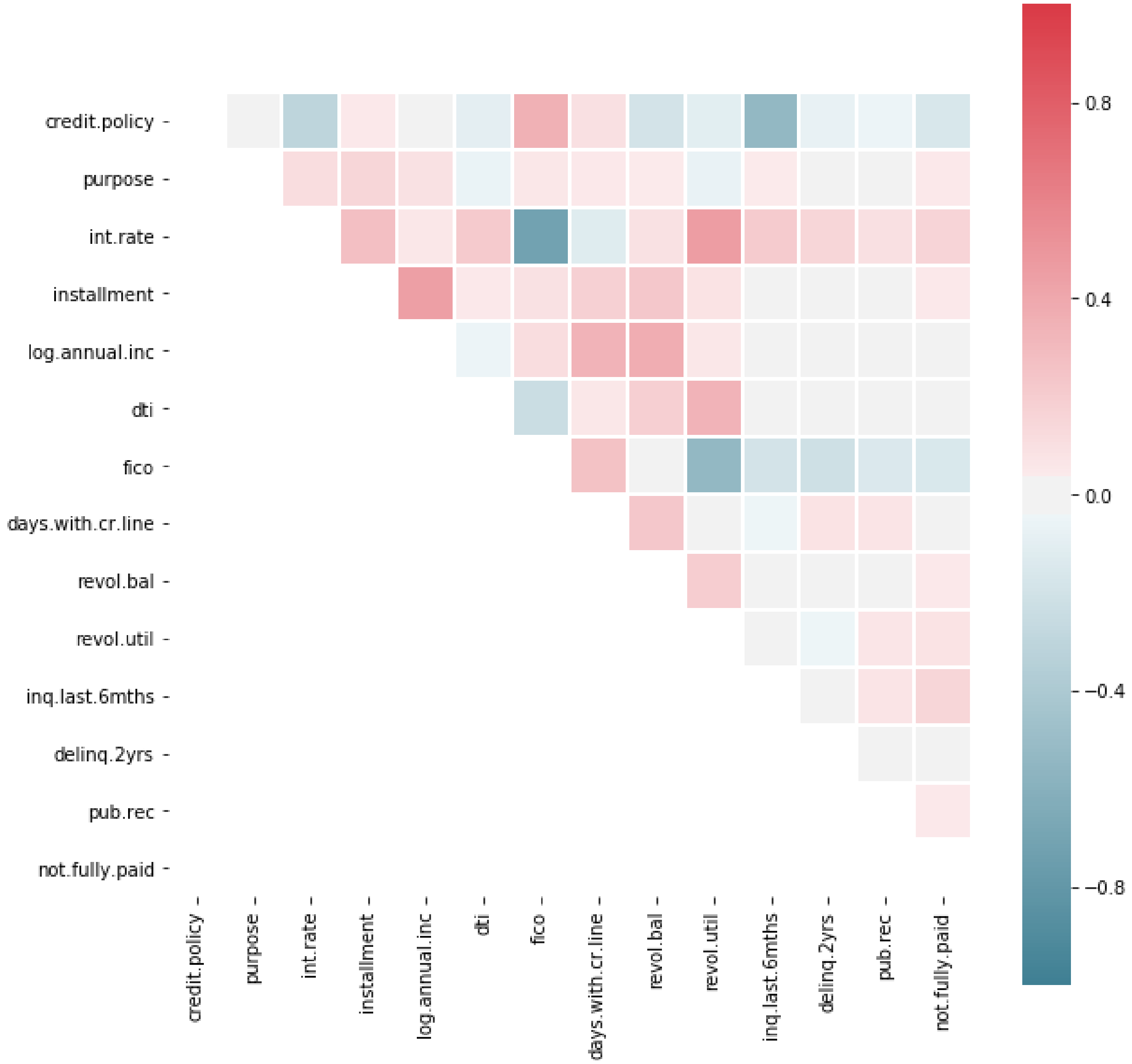
Index	43	1328	1820	393	520
Probability of defaulting	0.9316	0.93	0.8933	0.8916	0.0.8916

6) FICO rate and its influence on Interest rate



correlation: -0.7150089610860461

7) Correlation amongst all the variables



Data Modeling

We use Random Forest, Extra tree classifier with bootstrapping,SVM and logistic regression to predict weather or not the customer will be able to payback the loan.For repayment detection, we have to have highly sensitive data, since we have to decrease the false negatives more than false

positives: i.e.: No repayment calls not detected wrongly are more costly than non correct repayment calls wrongly detected.
Accuracy is not an accurate predictor, since data is highly imbalanced, even if the model is bad, it will predict 0 more easily and will have high accuracy..
Cohen's kappa and Mathews correlation both are good metrics. Mathew's correlation tells how good a binary classifier is. Our score (0.8) is a good score.

	Accuracy	Sensitivity	Specificity	ROC AUC	Kappa Score	Mathews Correlation Coeff
Extra tree classifier	91.1	87.4	94.6999	91.1	0.822	0.824
Random forest	91.0	84.6	97.3	91.0	0.82	0.827

Conclusion

From the evaluation metrics, it is clear that tree based classifiers fit the dataset in the best possible manner as compared to other classifiers. Among the tree based classifiers, Extra tree classifier fits in the best way. This can be ascertained from the fact that from the above table, the accuracy, sensitivity and more importantly the ROC AUC parameters for the Extra tree classifier have a higher value as compared to the Random forest classifier.

Reproducibility

The dataset is available in the public domain. It is open source and is available from the website datalendingclub.com. Lending club is a company based in the United States. It is a peer to peer lending company and basically provides the connecting medium between the investors and the borrowers. The Exploratory Data Analysis and the evaluation metrics methods used by us is also open source. We have extensively used libraries such as Pandas, Numpy,Seaborn and Sklearn.

Future Work

While seeing feature importances, we are seeing features from the training set. The lime package for interpretability allows us to see what features are important for each of the row of our testing data. What features may have been responsible for non repayment of one person might not be the same for the other.

GitHub link

https://github.com/CaptainPramil/Predicting-Loan-Payment-Status