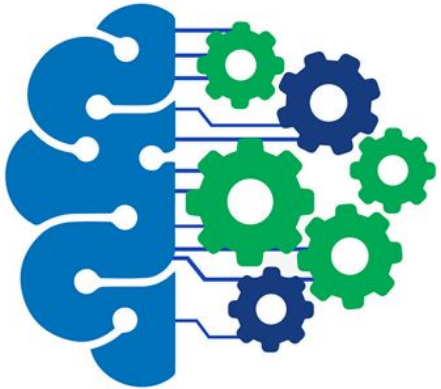
SPRINGBOARD CAPSTONE TWO: PRESENTATION

Project by Shubham Saurabh

Problem Statement

Goal was to create a model that trains on a set of loans to invest in and categorize them as DEFAULT or NOT. There are many ways to achieve this but for the purposes of this problem, we decided to create models based loan database of clients to predict whether a loan is worth investing in or not.

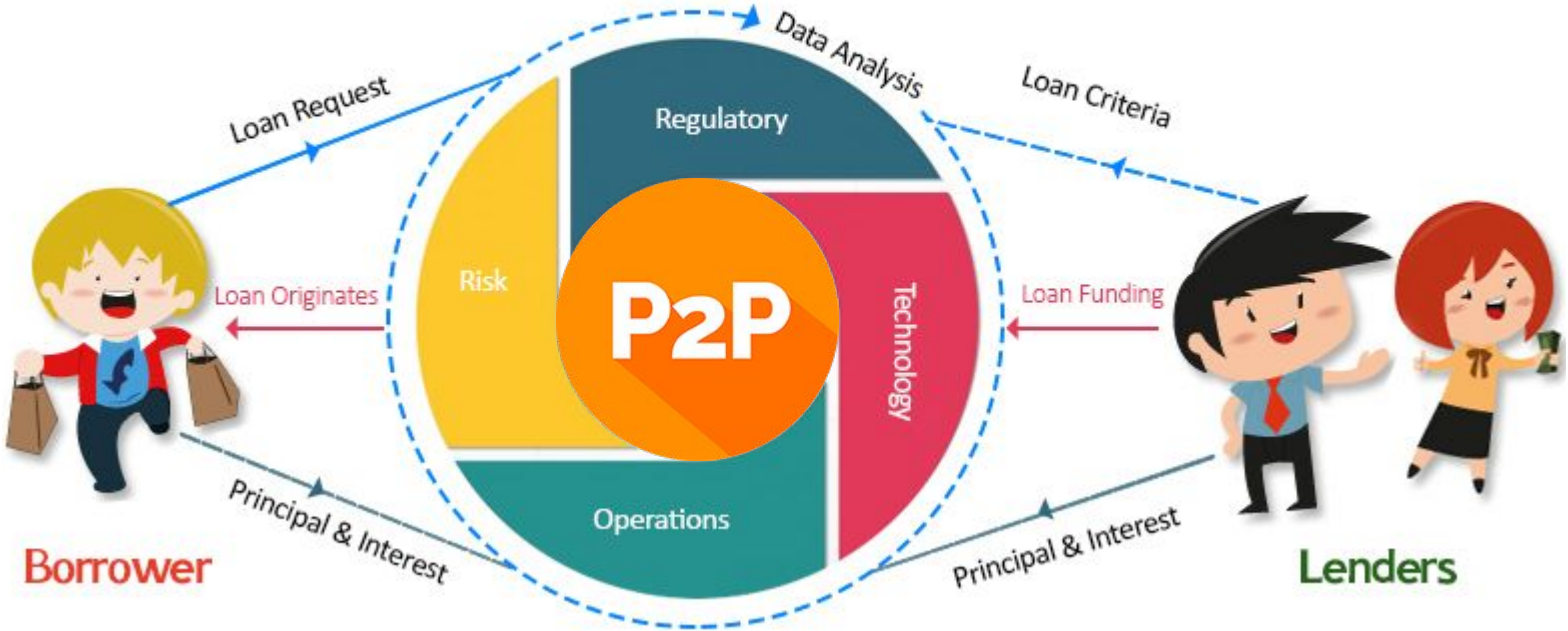


Stakeholders

WHO CARES?

P2P LENDING COMPANIES

HOW P2P LENDING WORKS



WHAT AFFECTS LOAN APPROVAL?

A lot of factors might affect loan approval beside credit scores like:

Income
Employment History
Debts owed
Collateral
Bank Accounts History
Delinquency
Trading Accounts
Credit cards
Bankruptcy
Many more Factors...

DATA INFORMATION:

SOURCE: https://www.lendingclub.com

INFORMATION:

Data acquired for Period: Jan 2012 - Dec 2018
Number of Fields: 182
Number of Records: 986,041

	0	1
id	68407277	68355089
member_id		
loan_amnt	3600.0	24700.0
funded_amnt	3600.0	24700.0
funded_amnt_inv	3600.0	24700.0
term	36 months	36 months
int_rate	13.99	11.99
installment	123.03	820.28
grade	C	C
sub_grade	C4	C1
emp_title	leadman	Engineer
emp_length	10+ years	10+ years
home_ownership	MORTGAGE	MORTGAGE
annual_inc	55000.0	65000.0
verification_status	Not Verified	Not Verified
issue_d	Dec-2015	Dec-2015
loan_status	Fully Paid	Fully Paid
pymnt_plan	n	n
url	https://lendingclub.com/browse/loanDetail.action?loan_id=68407277	https://lendingclub.com/browse/loanDetail.action?loan_id=68355089
desc		
purpose	debt_consolidation	small_business
title	Debt consolidation	Business
zip_code	190xx	577xx
addr_state	PA	SD

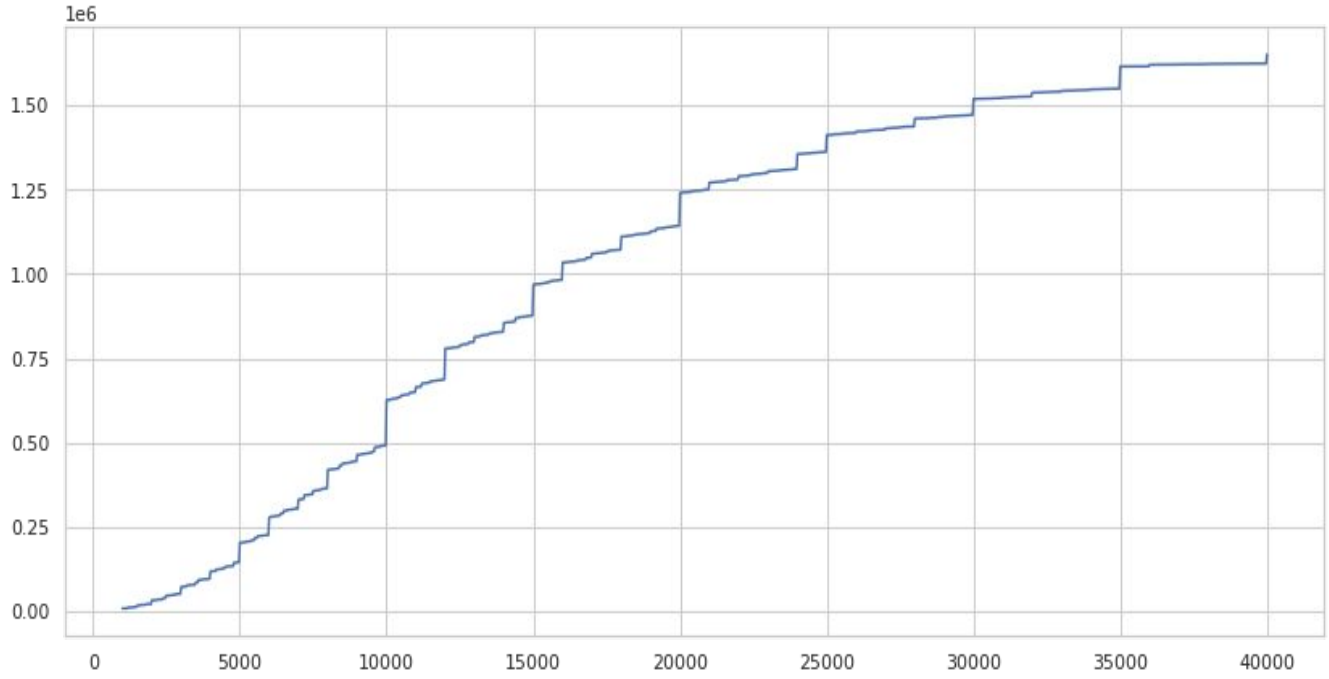
Two files:
Lending-club.csv
LCdataDictionary.xlsx

	Description
id	A unique LC assigned ID for the loan listing.
member_id	A unique LC assigned Id for the borrower member.

DATA EXPLORATION:

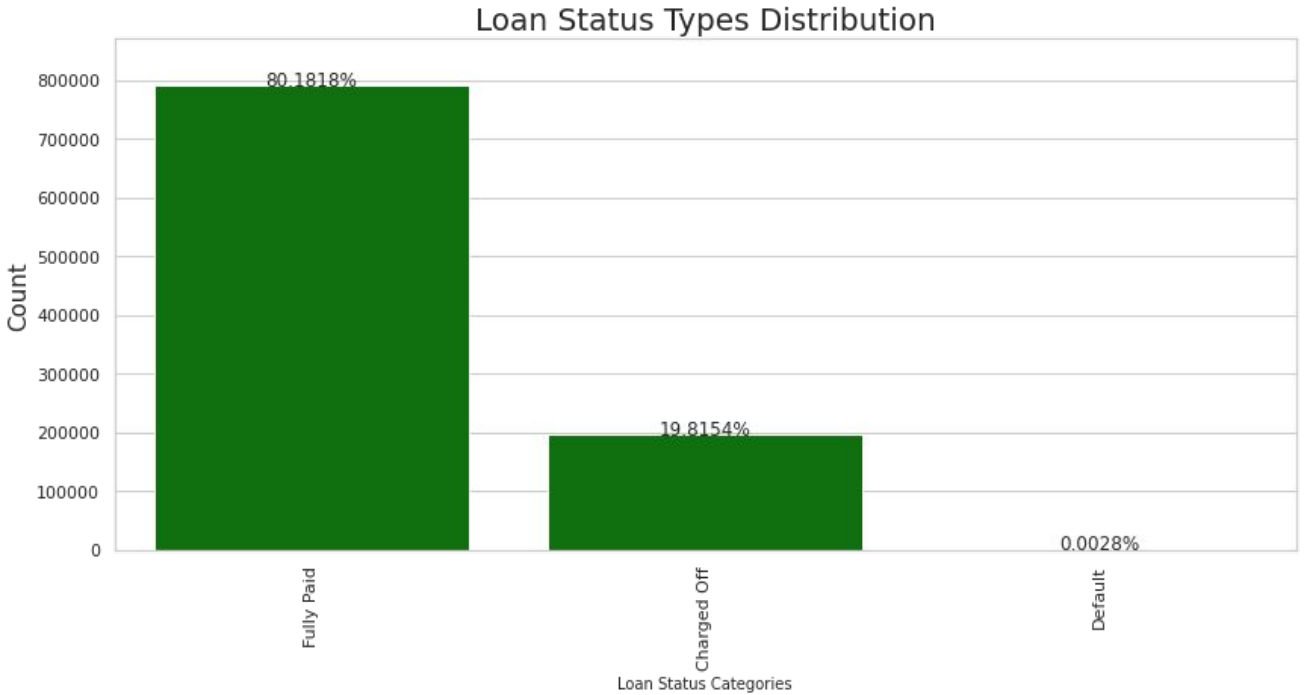
What's the distribution of Loan Status?
What's the distribution of Amount of loans?
What's the distribution of Interest Rate?
What's the % of Defaults in loans?
What's the most common grades?
What's the most common employer titles?
What's the most common Purpose that a client request a loan?
What's the difference between Terms?
And a lot of other questions that will raise through the exploration

LOAN DISTRIBUTION:



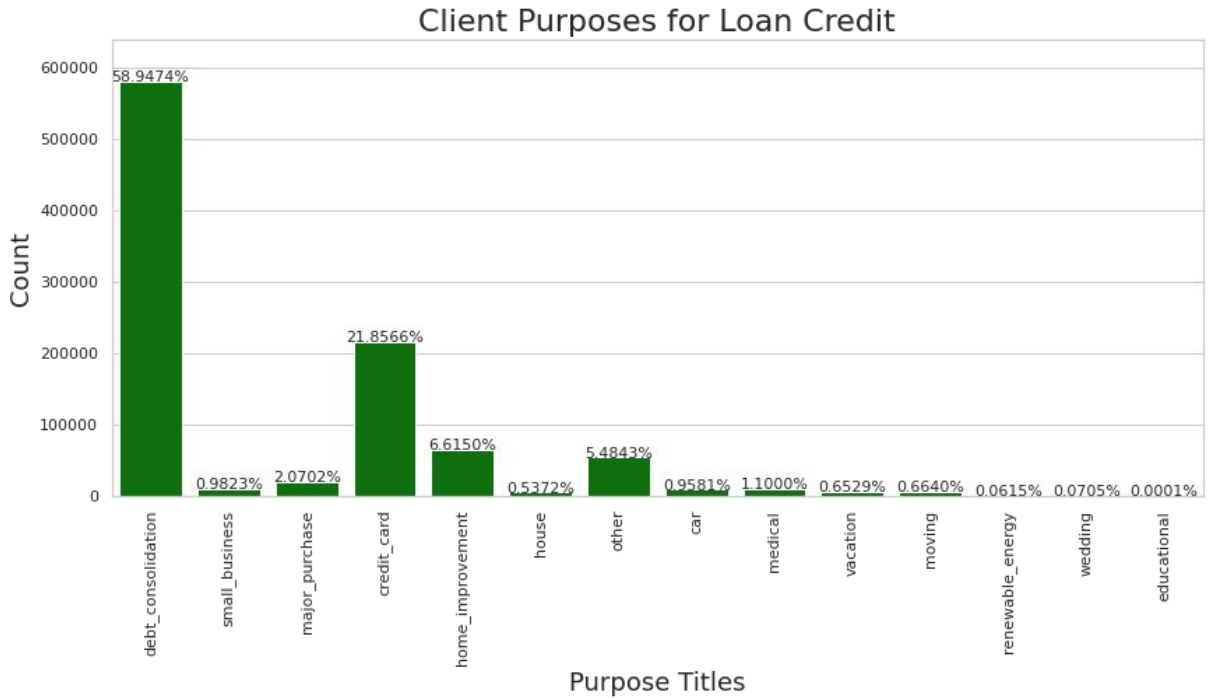
CDF showing total loan amount distribution

LOAN STATUS:



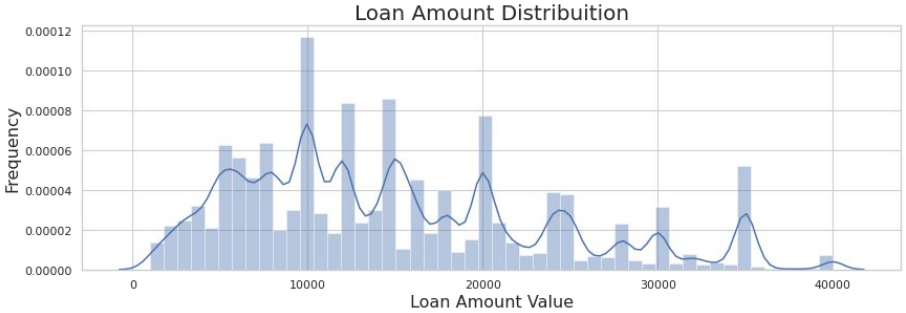
Three different Loan status are shown here among which Charged off and Defaults are clubbed together later on.

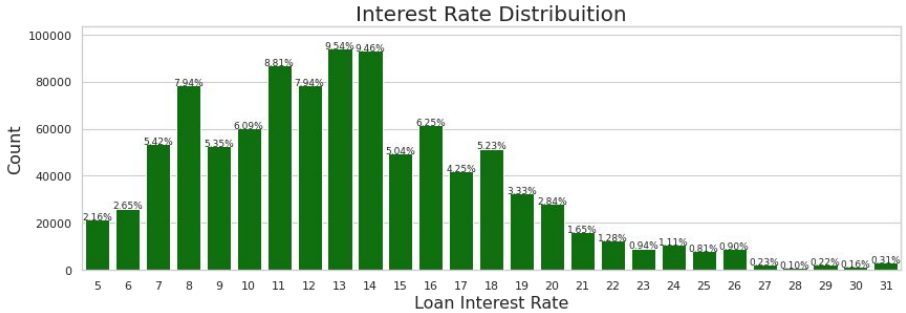
LOAN PURPOSE:



Debt consolidation seems to be one of the main loan purpose followed by Credit Cards and Home Improvement.

LOAN AMOUNT vs INTEREST:



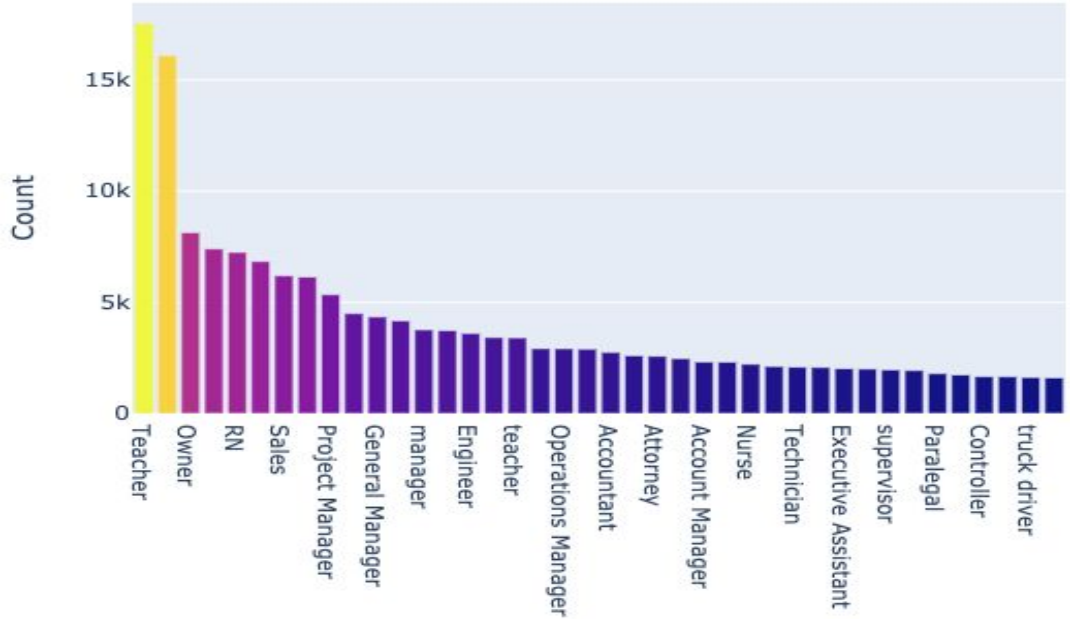


LOAN AMOUNT

Loan amount and interest had similar distribution as observed.

INTEREST RATE

EMPLOYMENT-WISE BORROWERS RANK



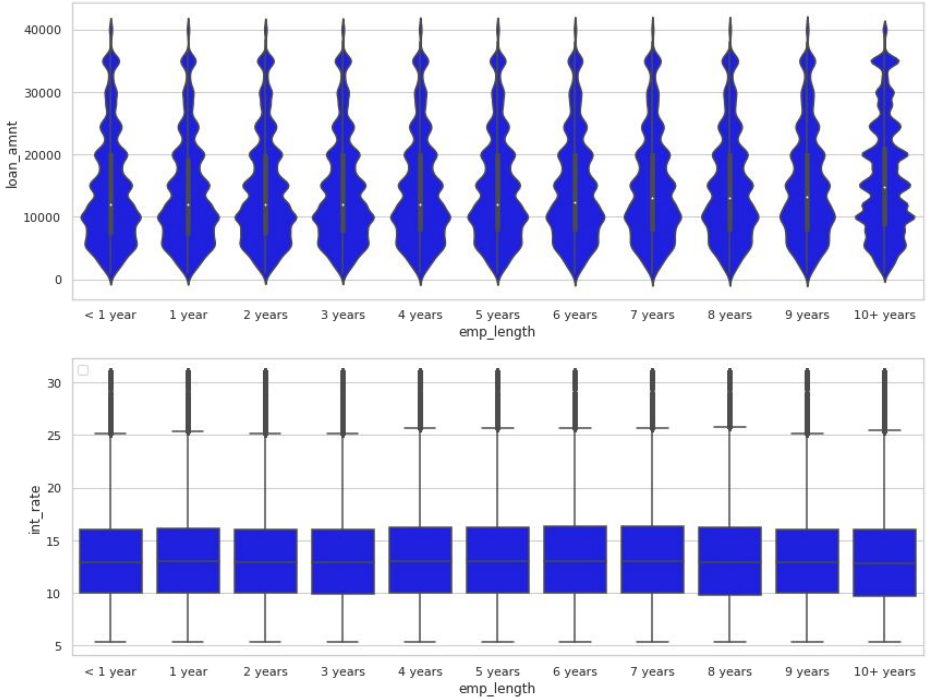
Teachers seems to be highest count for borrowers followed by Manager and Owner.

Loan Status state-wise:

addr_state	Charged Off	Default	Fully Paid
AK	0.000479696	0	0.001973549
AL	0.002827469	0	0.009525973
AR	0.001844751	0	0.005829372
AZ	0.004718871	1.01E-06	0.019903838
CA	0.027550579	0	0.113928325
CO	0.003082022	0	0.016451649
CT	0.002570887	1.01E-06	0.012291578
DC	0.000327573	0	0.001988761
DE	0.000552715	0	0.002340673

Partial Data Shown

Effect of Employment years:



A minimal difference was observed on loan amount and interest rate wrt number of employment years.

Loan grades wrt Loan status

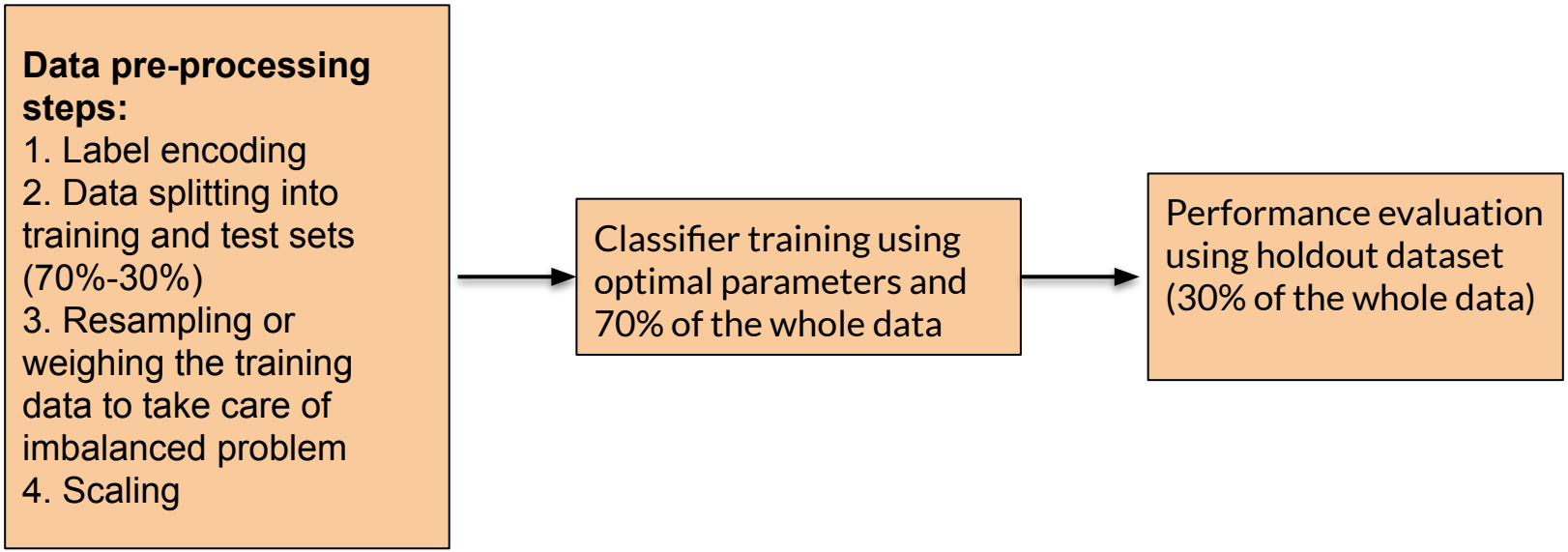
loan_status	A	B	C	D	E	F	G
Charged Off	9307	36463	62861	44757	27288	11107	3605
Default	4	6	5	9	4	0	0
Fully Paid	154854	247683	222989	104465	43535	13488	3611

As it is observed even loans with high category status seems to be default and charged off often.

Machine learning Modeling:

Type: Supervised Learning
Binary Classification: 1 for “FULLY PAID” and 0 for “DEFAULT”
Data Imbalance: 1:5 Ratio
Tools: Python Scikit Learn

Modeling:


Data pre-processing steps:
1. Label encoding
2. Data splitting into training and test sets (70%-30%)
3. Resampling or weighing the training data to take care of imbalanced problem
4. Scaling
Classifier training using optimal parameters and 70% of the whole data
Performance evaluation using holdout dataset (30% of the whole data)

Model and Resampling method:

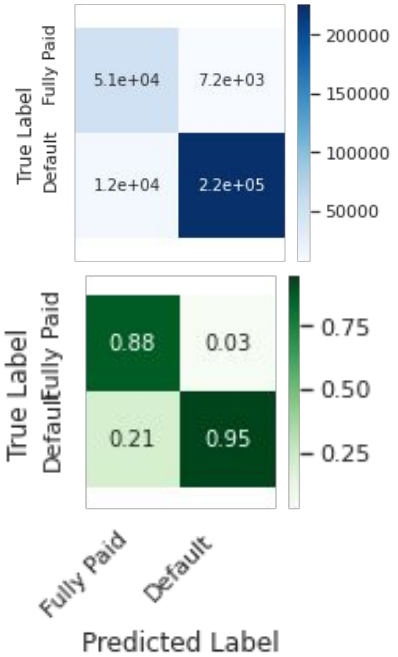
Models used:

1. Logistic Regression
2. Random Forest Classifier
3. Decision Trees

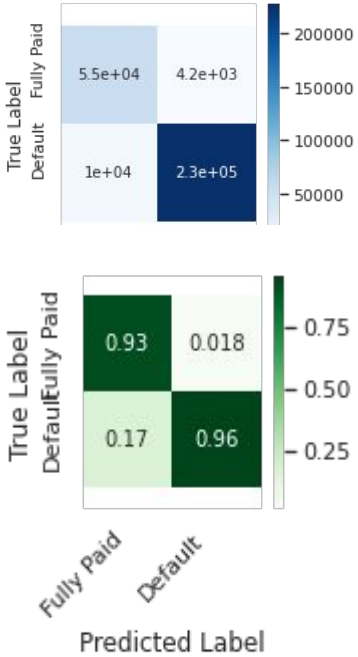
Resampling Method:

1. Random under-sampling (RUS)

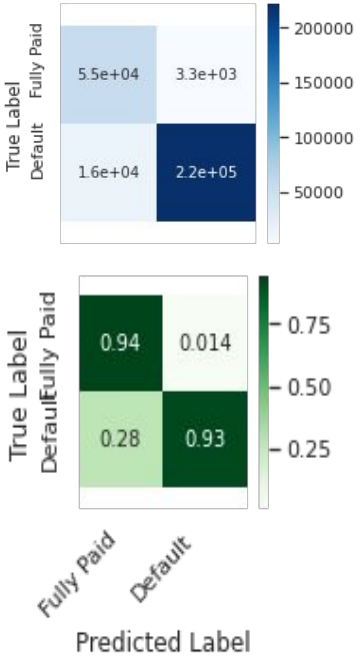
Model Comparison:



Logistic Regression



Random Forest

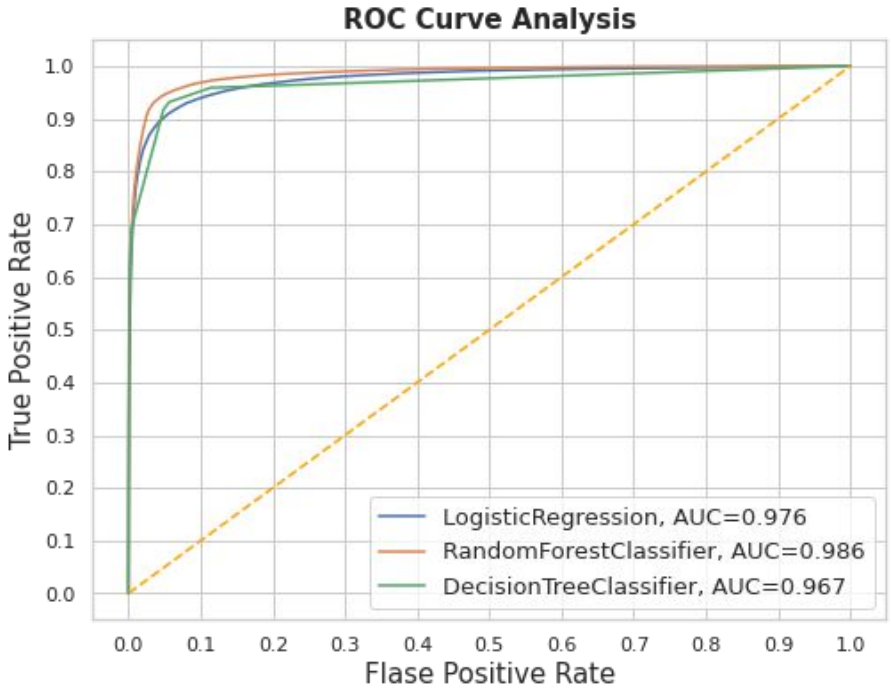


Decision Trees

Classification report:

Classifier: LogisticRegression		0	1	accuracy	macro avg	weighted avg
	precision	0.81	0.97		0.89	0.94
	recall	0.88	0.95		0.91	0.93
	f1-score	0.84	0.96	0.93	0.90	0.94
	support	58726	237087	295813	295813	295813
Classifier: RandomForestClassifier		0	1	accuracy	macro avg	weighted avg
	precision	0.85	0.98		0.91	0.95
	recall	0.93	0.96		0.94	0.95
	f1-score	0.89	0.97	0.95	0.93	0.95
	support	58726	237087	295813	295813	295813
Classifier: DecisionTreeClassifier		0	1	accuracy	macro avg	weighted avg
	precision	0.77	0.99		0.88	0.94
	recall	0.94	0.93		0.94	0.93
	f1-score	0.85	0.96	0.93	0.9	0.94
	support	58726	237087	295813	295813	295813

ROC Curve Comparison:



Conclusion:

Three different models were used to predict Defaulters. As expected Random Forest Classifier performed the best followed by Logistic Regression and Decision Tree.

Though it was assumed initially Decision Tree would out perform Logical Regression based on the high number of features available. It is important to note that the Default Prediction were the main criteria for the performance here since we would like to avoid a single default loan upsetting the profits gained by several fully paid loans. So as per this criteria Forest Tree performed the best followed by Decision Tree and Logistic Regression.

Future Work Ideas:

A new parameter could be introduced which could take in consideration the profits and loss from the sets of loan applicants and then by providing a risk factor on the net profit we could use it to approve lenders for better risk management from a list of several applicants to minimize loss.