Capstone Project on

P2P lending risk analysis

Group 5| PGP-DSE July 2019 (Pune) | 28th Nov. 2019

Presented by:
Bhawesh Panchal
Rupesh Ghule
Siddharth Biswas
Surbhi Welekar
Sudhendu Awasthi

Under the guidance of:
Mr. Ankush Bansal

Problem Definition

Problem Statement:

Given transaction data from one of major US market player for over a decade, we need to:

Predict interest rate of potential borrowers.

Predict the probability of default for a potential loan

Objective:

One of the major risks involved in P2P lending is borrower defaults on the loan.

Lending money to a borrower, there's a risk the borrower might not be able to pay back the loan. This is called defaulting.

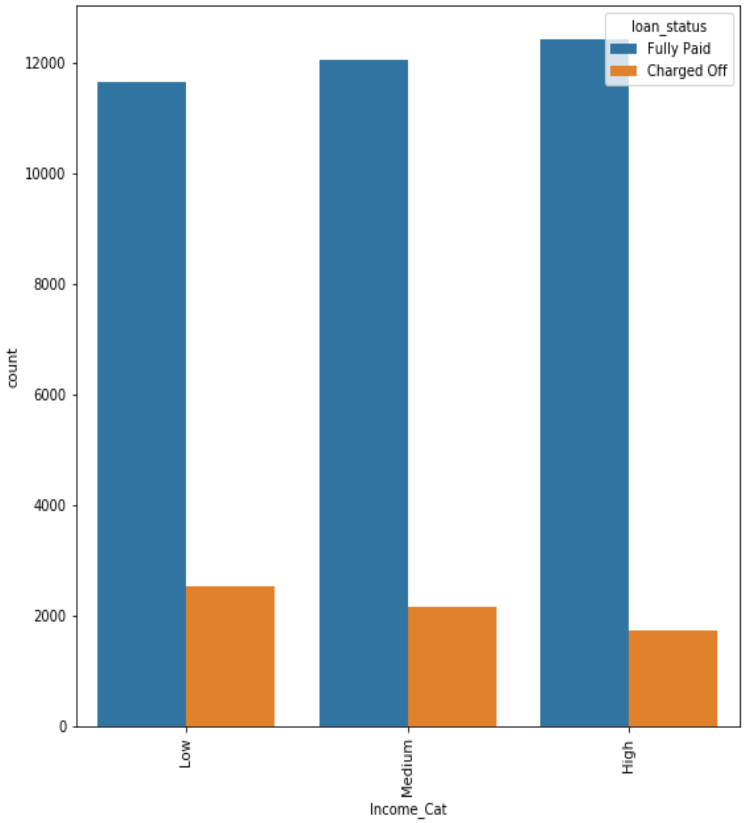
Data Preparation

- Our data had 42542 rows and 144 columns.
- Columns with More than 80 % missing values dropped
- Columns having nunique = 1 dropped
- 44 columns left
- Strings which can not be considered as categorical variables removed i.e. desc.
- desc - too much information
- Columns with many categories like emp_title removed
- Some na values, where data can be guessed easily, were replaced. i.e. 'emp_length'
- we replaced na values by 0. There was no experience, as the applicant was fresher
- We also removed unnnecessary characters like ‘%’, ‘years’, ‘yrs’
- some replacements by mode
- Removed columns having highly imbalanced data like tax_liens.
- 0.0 42429
- 1.0 1I

EDA – Customer Attributes

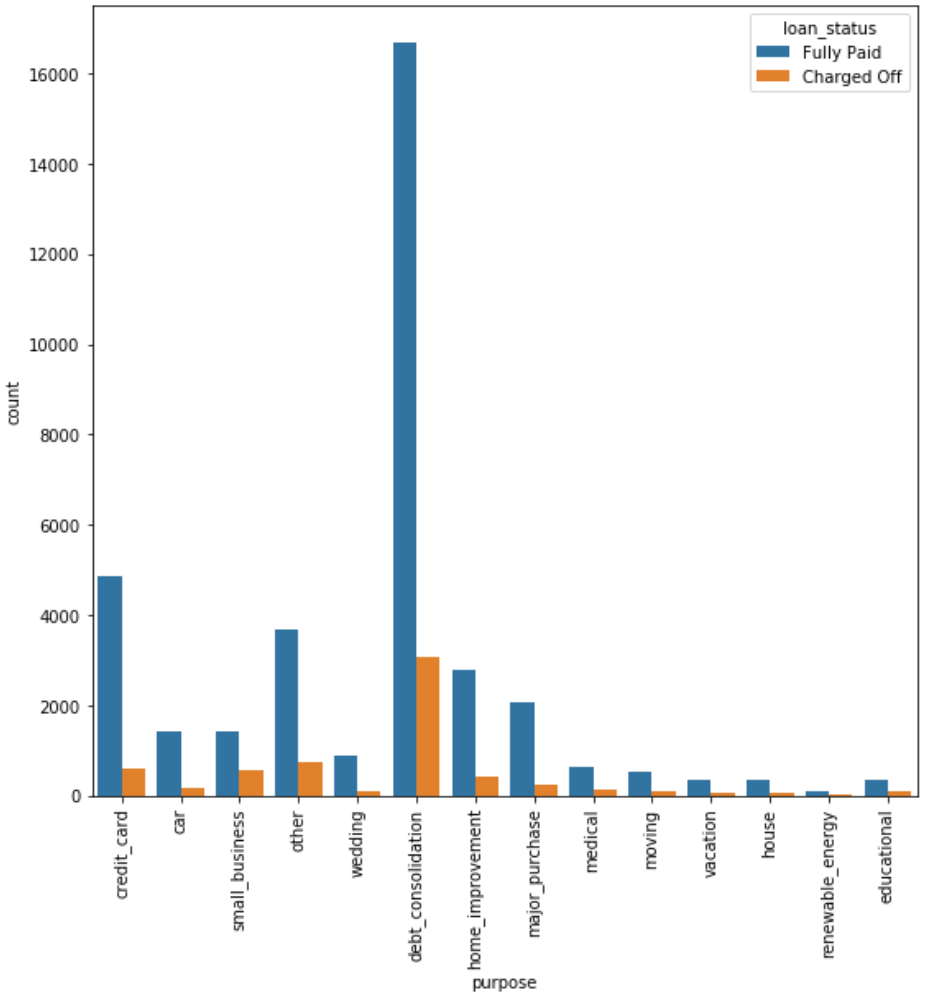
Income Category vs Loan Status

- High income lower bad loans
- Less income higher bad loans



Interest rate vs Loan Status

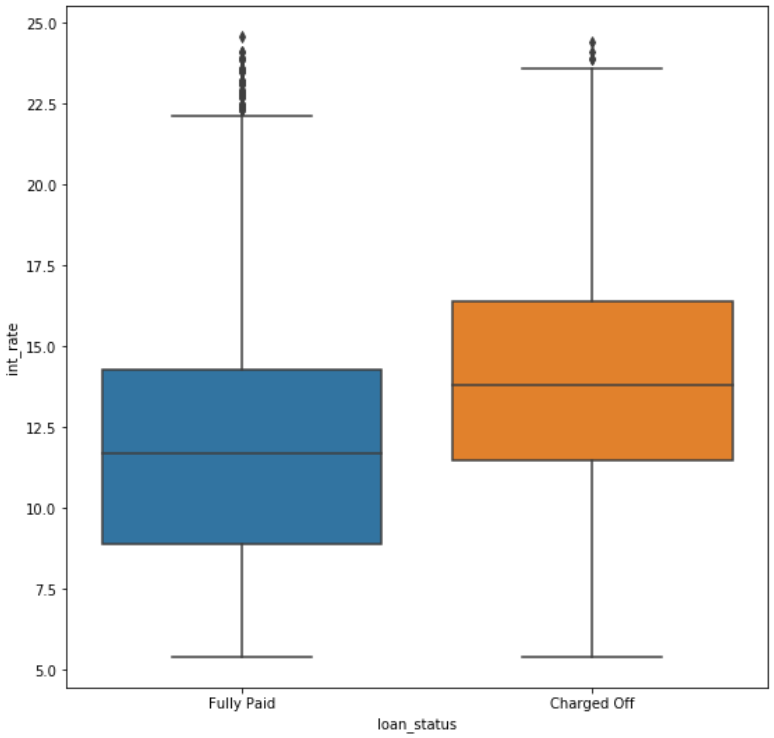
- Small business is the most risky purpose to grant loan



EDA – Customer Attributes

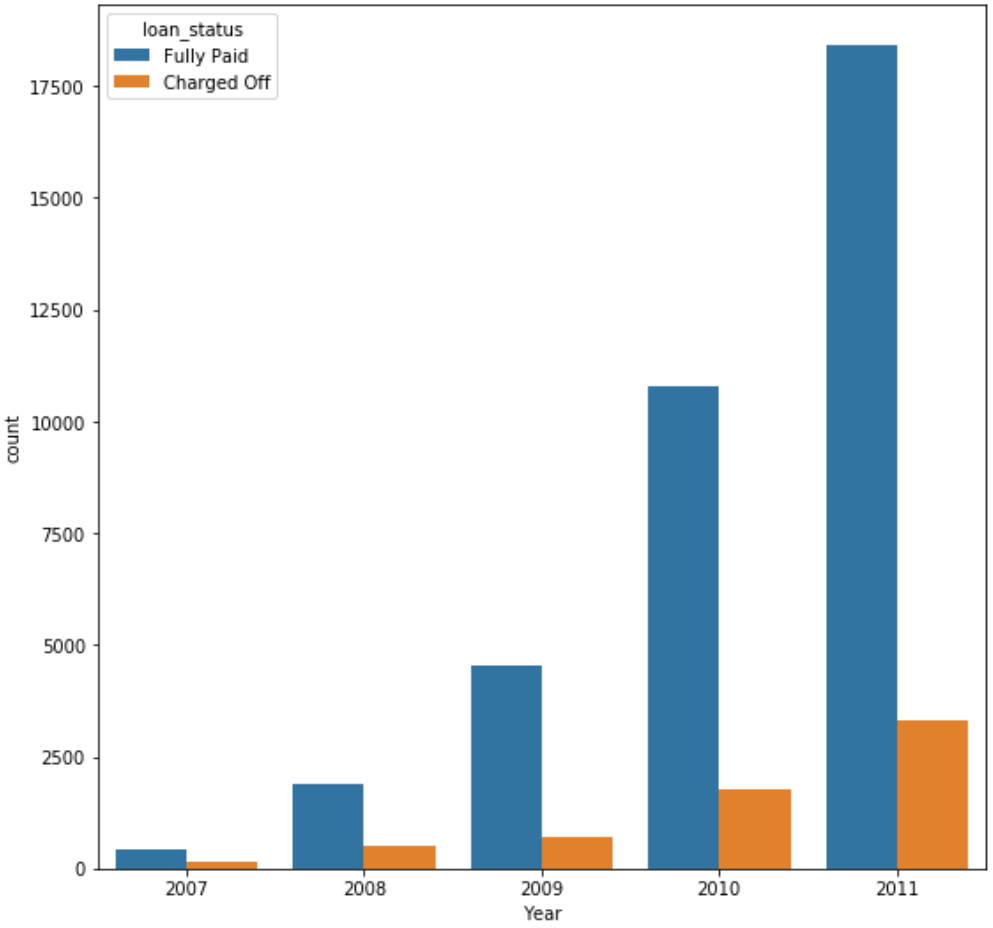
Interest rate vs Loan Status

- High int rate lower bad loans
- Less int rate higher bad loans



Year vs Loan Status

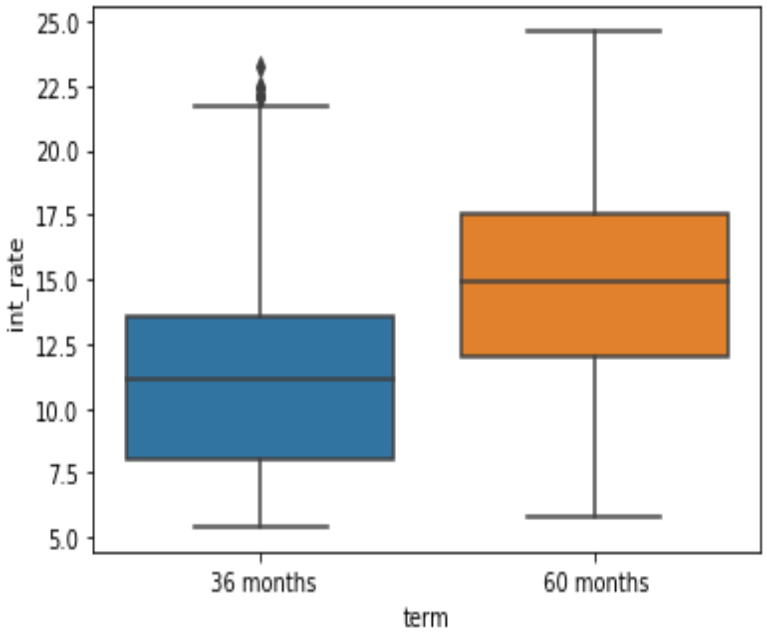
- Ratio on bad loans is decreasing



EDA – Business Prospect Customer Attributes

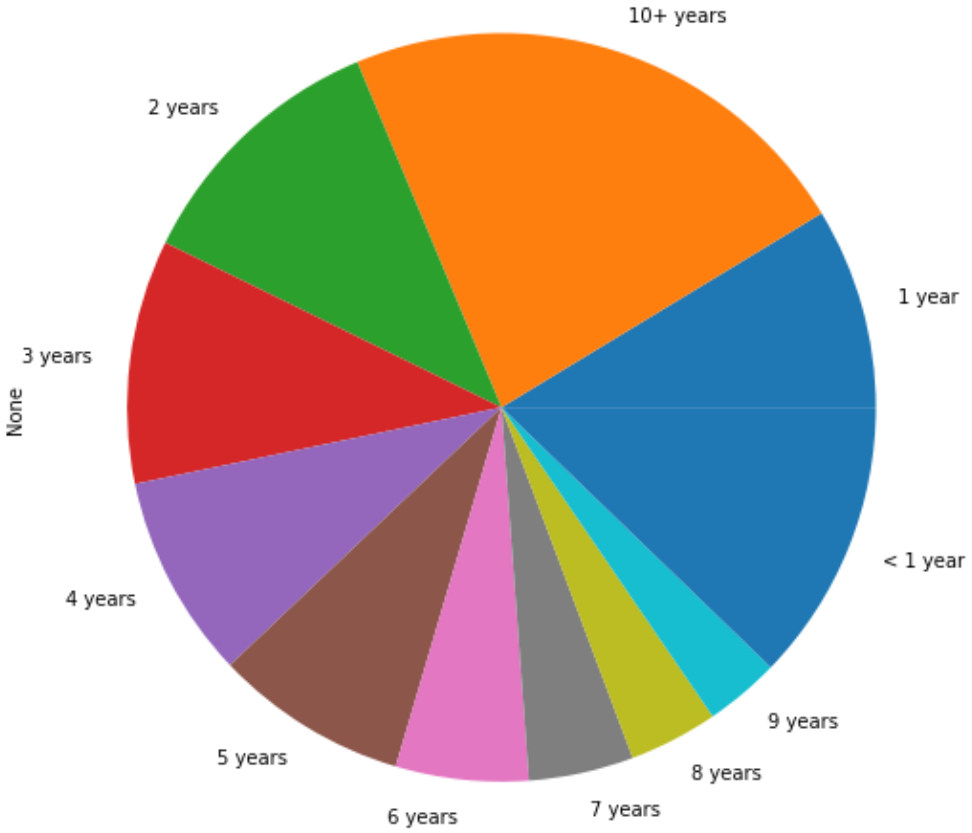
Term vs Loan Status

- Low term lower bad loans
- High term higher bad loans



Employment duration

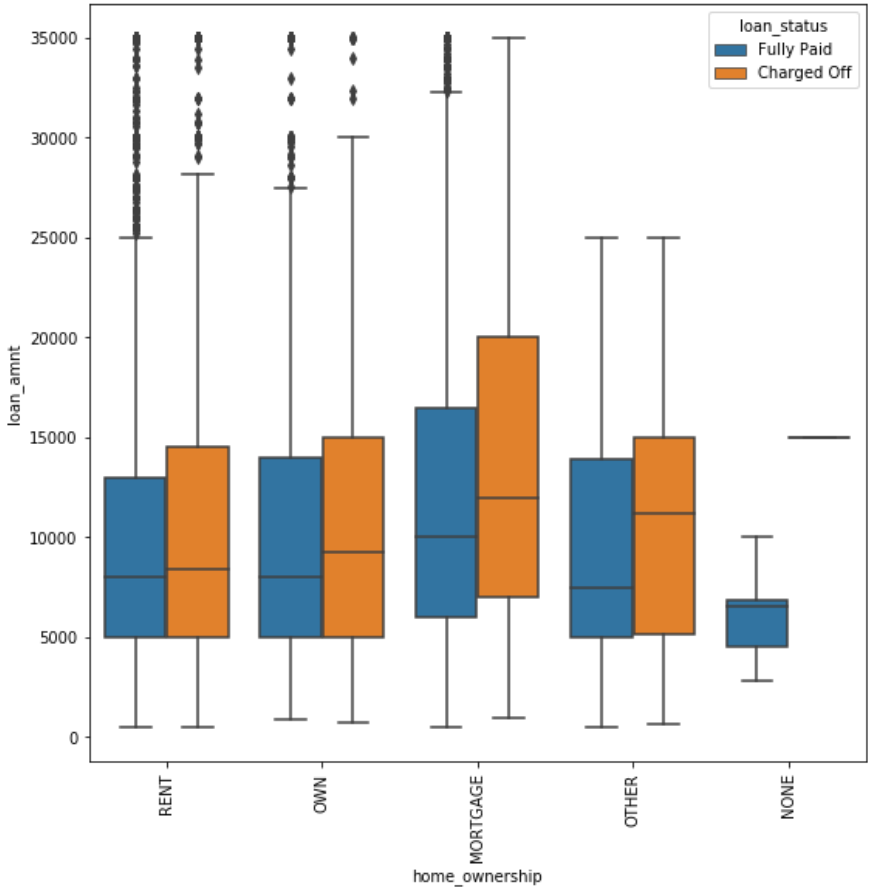
Nearly 50% of our customers are having 3 or more year of experience.



EDA – Business Prospect Customer Attributes

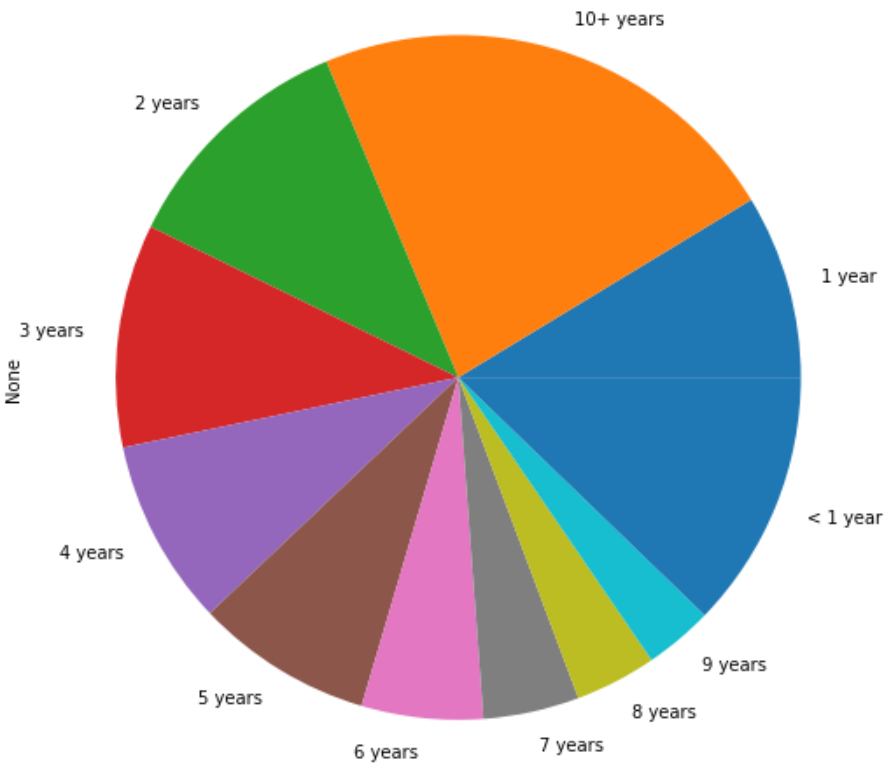
Home owner vs Loan Status

High loan amount granted to customers having mortgaged is a risk



Purpose of loan

Around 40% of our customers are having purpose as debt consolidation.



Predict interest rate on basis of borrower's application data

- Algorithms Considered:

1.) Linear Regression Approach:

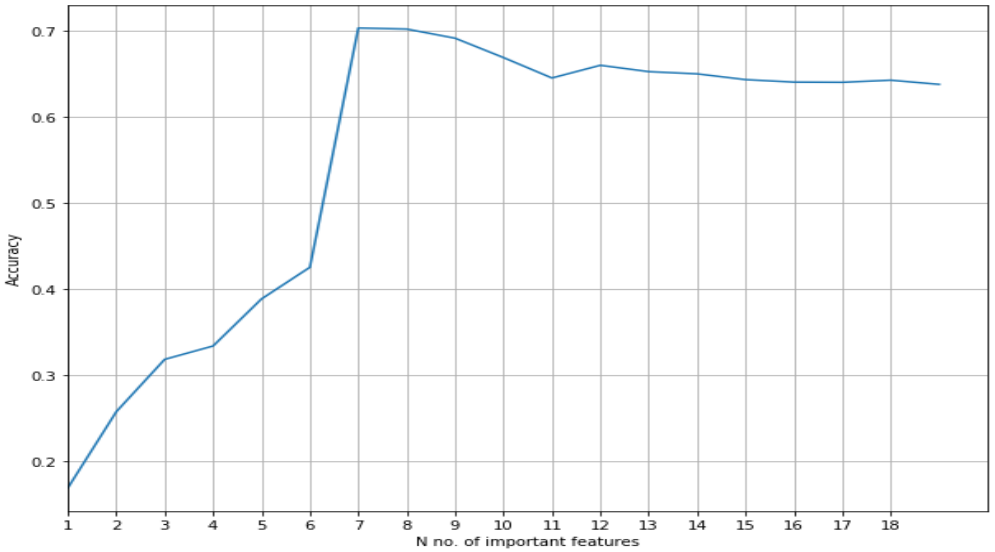
- Step 1 : On removing multi-collinearity. 95 out of 99 features were obtained
- Step 2 : Considering statistically relevant features. 45 out of 95 features were obtained
- Step 3 : Comparing Linear Regression Model:

Adjusted R squared value with multicollinearity	58.60 %
Adjusted R squared value without multicollinearity	52.40 %
Adjusted R squared value with statistically relevant features	52.40 %

- Based on p-values, relevant features were found to be term, installment, revol_util, delinq_2yrs, inq_last_6mths
- Accuracy needs to be improved

2.) RandomForest Regressor Approach:

- Cross Validation approach was used to evaluate the accuracy of prediction for interest rate. Feature importance was then calculated for all the available features
- The following graph was obtained on basis of feature importance:



- Most important features obtained were, revol_util, term, installment, inq_last_6mths, revol_bal, open_acc, funded_amnt

RandomForest Regressor	Accuracy
All features	62.60%
Important features – [Top 7]	70.33%

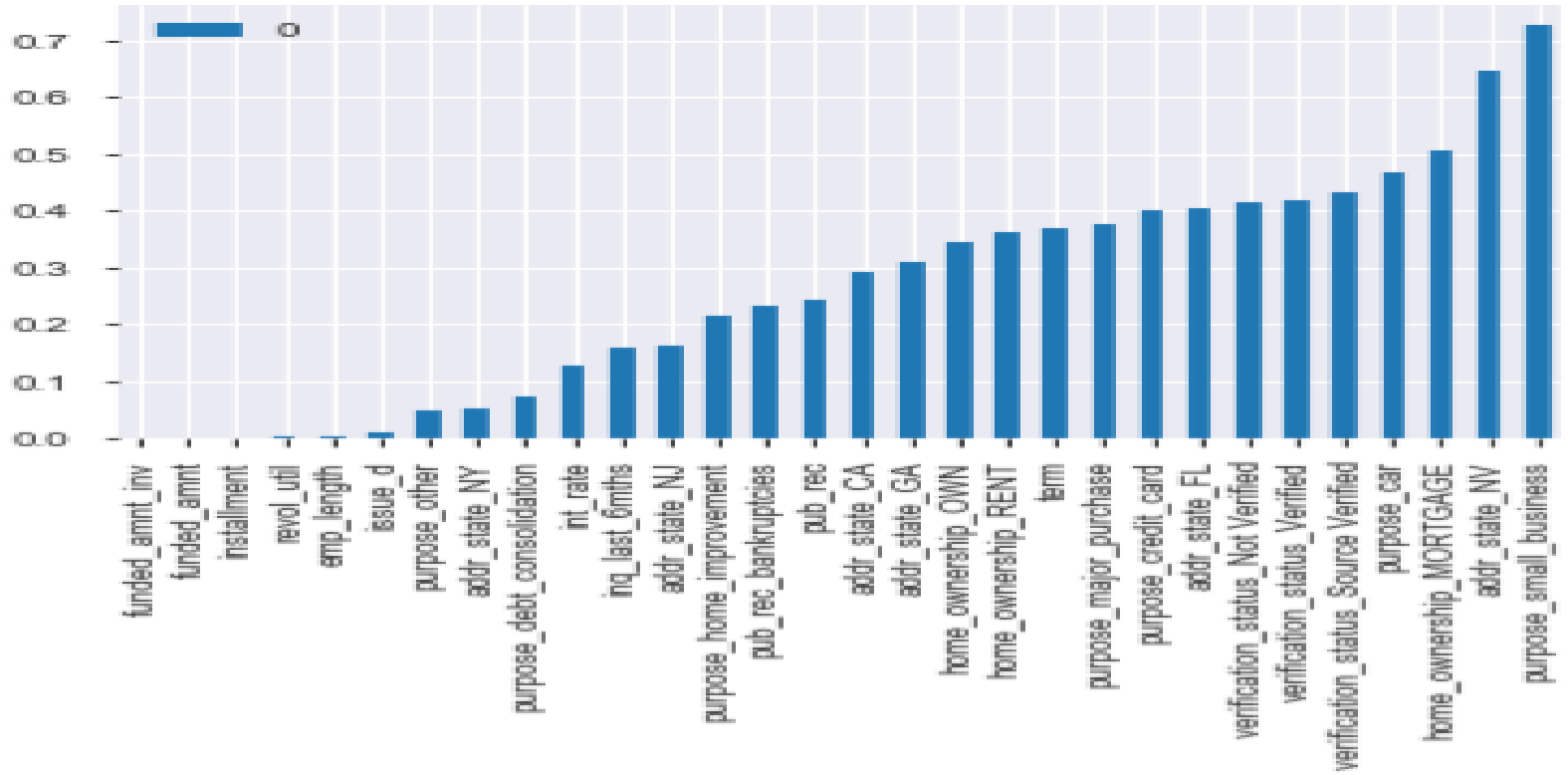
Predict default rate on basis of borrower's application data

- Algorithms Considered:
  - Features were selected on the basis of Random Forest Classifier

Algorithm	Train Score	Test Score	Recall
Logistic Regression (0.5)	65.04%	66.29%	0.64
Decision Tree	71.58%	83.96%	0.06
Random Forest	88.23%	83.96%	0.02
KNN	77.07%	83.96%	0.46
Adaboost Classifier	77.07%	83.96%	0.04
Logistic Regression (0.53)	65.04%	66.29%	0.64

Impact of Different Features

- At last , we applied Logistic Regression with 0.53 threshold which provided us 0.66 recall for bad loan case.
- Area Under Curve is 0.72



Conclusion

- The people having small business, mortgage.
- The people taking loan for more interest rate and for more term.
- The purpose of loan should be considered and verification of source is must.

Business Insights:-

- We have to make specific changes in terms and interest rate for people differently respectively as per their purposes.
- Home owners should be prefers while giving loans.
- People from Nevada, New Jersey and Florida should be examined properly before giving loan.

Thank you.