ANALYSIS OF LOAN DATASET
TASK-1

AIM:

The objective of the given task was to identify the risky loan applicants who have a tendency not to pay off the loan.

FACTORS:

The predictor variable to be used in the analysis was the ‘loan_status’ variable. It had 3 unique values:

1. Current
2. Fully Paid
3. Charged Off

We eliminated the rows containing ‘Current’ value because it is ambiguous. But the other two values clearly indicate that the loan applicant was able to pay the loan or was a defaulter at the end of the day.

APPROACH:

- Examined all the columns
- Determined the necessary features which could affect loan repayment
- Formed a Random Forest Classifier
- Analysed the ‘loan_status’ on multiple variables
- Plotted various inferences

DATA DESCRIPTION:

The dataset initially contained 39717 rows and 111 columns.
But, despite more than hundred columns, most of them had null values or had more than half of the values empty, so even imputing them with mean values would be of no use. That is why we dropped such type of columns.

In addition to this, we did further cleansing of dataset and selected columns which will be of our use and could affect the loan_status significantly. At the end, the shape of the dataset was reduced to 38577 rows and 13 columns.

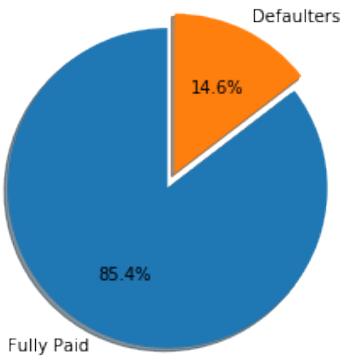
The columns which can significantly affect the loan_status are:

- loan_amnt
- term
- int_rate
- grade
- home_ownership
- annual_inc
- purpose
- addr_state
- dti
- emp_length

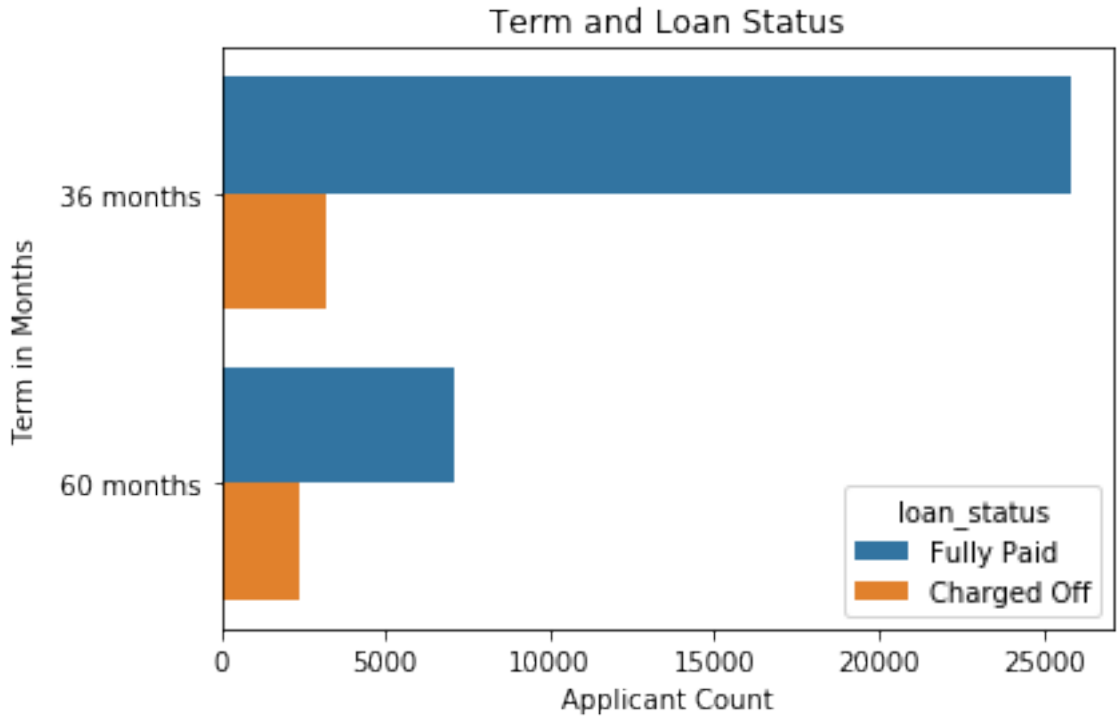
VISUALISATION AND INFERENCES

DEFAULTERS AND NON-DEFAULTERS:

We can observe in the pie chart below and observe basic inference that around 14.6% of total loan applicants turn out to be defaulters.



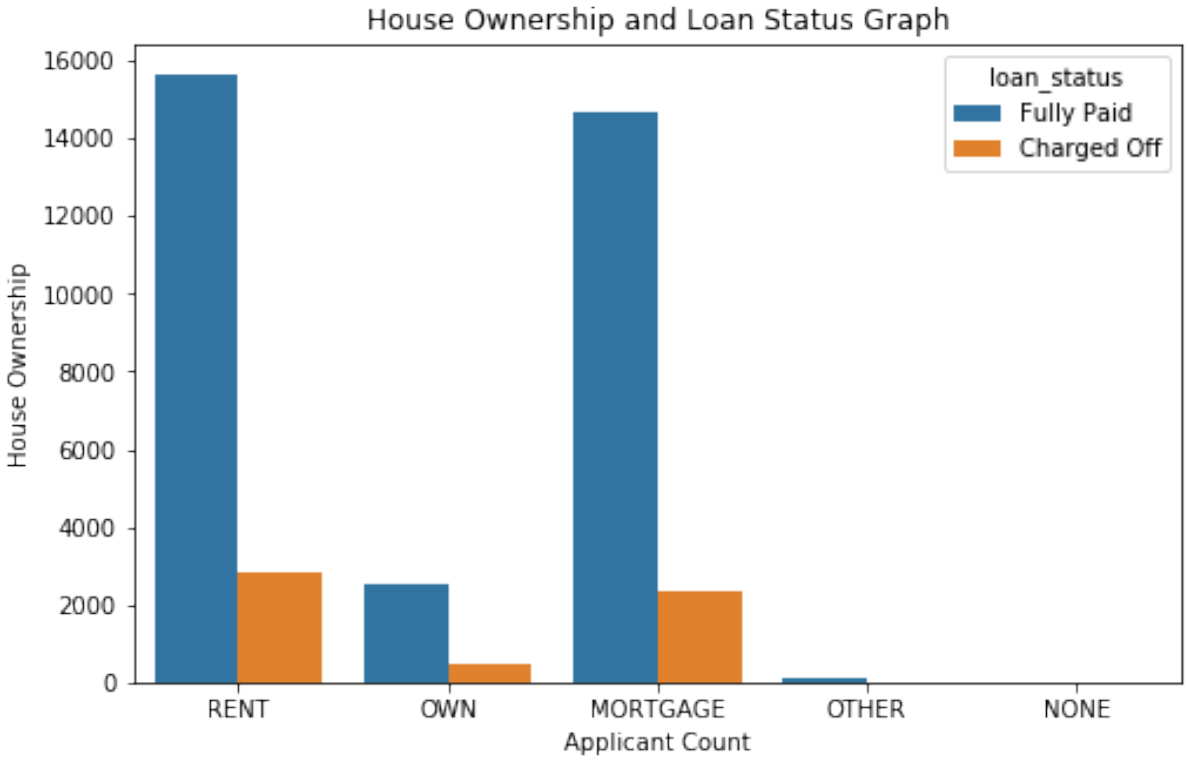
Term and Loan Status



By observing the graph between term and loan status, we can identify that even though the number of defaulters are higher in 36 month term loan, we have to keep in mind that in the same term tenure, number of people who paid off the loan completely is quite efficient as compared to the 60 month period.

That means, the **ratio of defaulters to non-defaulters is much higher in 60 month period than 36 month period. That means, the applicant with loan term of 60 month turn out to be more risky.**

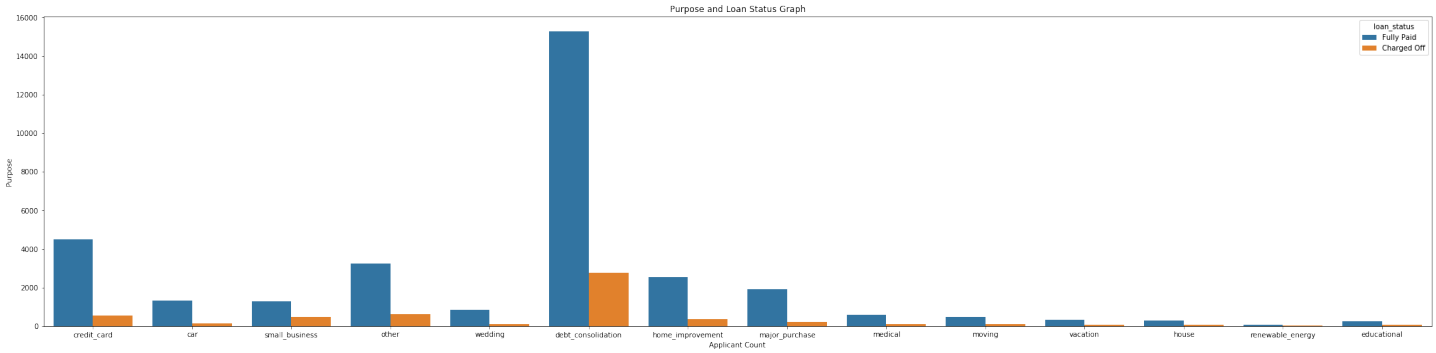
Home Ownership and Loan Status



It can be clearly derived from the above plot that the people living on Rent or Mortgage have a tendency of not repaying the loan amount as compared to the number of people having their home houses.

The applicant with a house on rent or mortgage can be more risky for the bank.

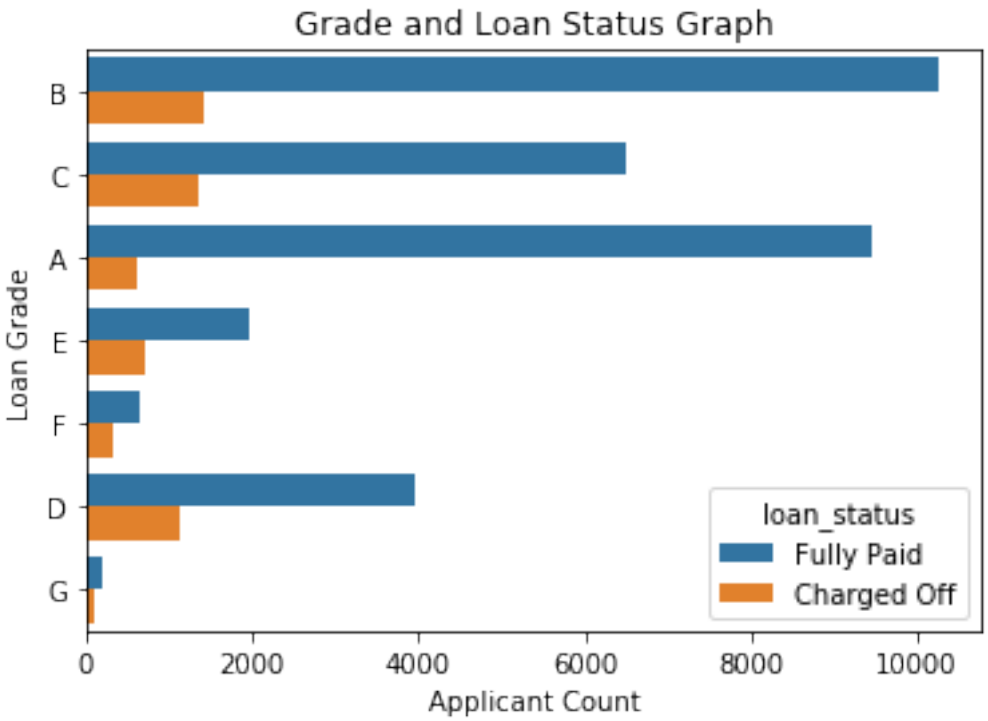
PURPOSE AND LOAN STATUS:



(The plot is visible clearly on the jupyter notebook)

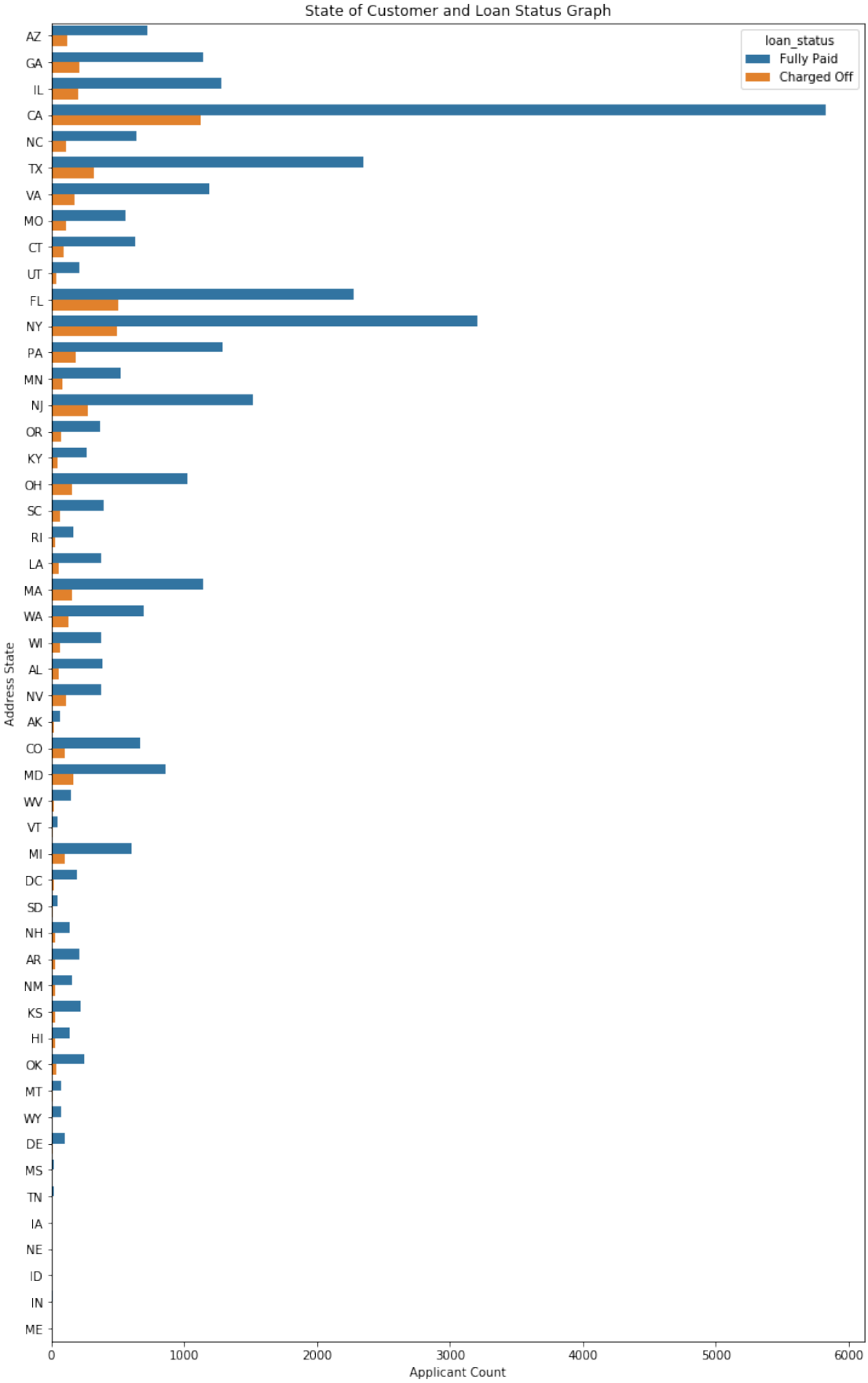
The applicants with purpose of debt_consolidation tend to become defaulters at the end. The risky purpose following it are credit_card and small_business.

GRADE AND LOAN STATUS



It can be concluded by observing the relationship between loan status and loan grade that the people with 'F' grade are more probable to turn out to be defaulters at the end. It is followed by G, E and D. **Almost half of the loan_applicants with 'F' grade turned out to be defaulters.**

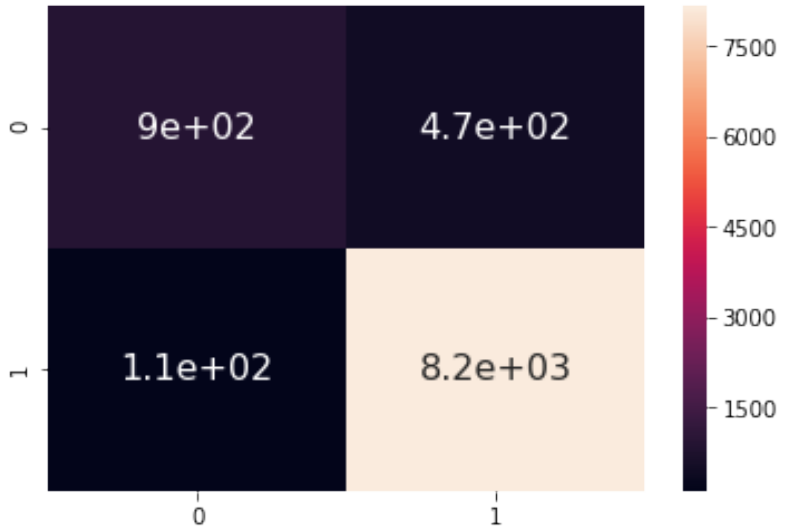
ADDRESS STATE AND LOAN STATUS



Loan applicants of California and Florida are seen to have more tendency of becoming defaulters as compared to other states. Even though New York also has the same number of defaulters as that of Florida, but considering the ratio of non defaulters and defaulters, we can't consider NY as risky because the of the linear statistical dependence.

California and Florida loan applicants can be classified as risky which turn out to be defaulters at the end.

CONFUSION MATRIX



Confusion matrix was obtained after fitting the data in the random forest classifier. The accuracy obtained was around 94%.

FINAL CONCLUSION

As per the analysis the loan applicant coming from California/Florida with a house on Rent/Mortgage with 'F' grade loan demanding a loan of 60 month period for debt consolidation is most risky loan applicant for the bank.

This means that the bank should sanction the loan for a shorter duration of time and the grade should not tend towards 'F' or 'G'. In addition to this, it should also examine the purpose of the loan carefully and look towards the annual income of the applicant.
