**Training Set Overview**

In this project, we aim to understand the borrower’s profile, and build a model to predict whether a given loan would be charged off. We hope this project can make contribution to minimize the risk of future loan defaults.

The loan dataset is collected Mar 2019 for 1,827,125 of bank loans with loan statuses and 145 attributes, like loan amount, term, interest rate, grade, etc. 17 of the attributes have no data, while 58 of the attributes have a missing rate over 30%. Based on my own domain knowledge, 31 features would be available before the potential investors make an investment decision, like state, annual income, application type, etc.

According to the project guideline, I will set a new target label named ‘charged\_off’ (1-charged off; 0-fully paid/current), based on the original one ‘loan\_statuses’.

**Data Exploration & Insights**

* 133,526 loans are charged off, accounting for 7.3% of the total. The data is unbalanced.
* Number of loans continuously increase from 2013 to 2018, and charged-off rate reached its highest (0.12%) in 2015, then dropped sharply (see Figure 1).
* There are three pairs of variables highly correlated: ‘loan\_amnt’ and ‘installment’, ‘open\_acc’ and ‘total\_acc’, ‘pub\_rec’ and ‘pub\_rec\_bankruptcies’ (see Figure 2). I need to drop one of each pair to avoid collinearity.
* The dataset may not be complete, to a lack of educational loans (see Figure 3).
* The longer the term, the more likely the loan is to default (see Figure 4).
* Loans with high interest rate are more likely to be charged off (see Figure 5).
* The value of dti is mainly in the range of 0 to 40. We can clearly see that as the dti grows, charged-off rate also increase, so does the charged off percent (see Figure 6).
* The number of individual applications is pretty higher than joint application, as well as default rate (see Figure 7).
* Location also has influence on charge off rate. Maine has the lowest charged off ratio which is about 3.3%, while Arkansas has the highest ratio which is about 9.3% (see Figure 8).

**Summary**

This is an imbalance problem, because we have a lot more entries of people that fully paid their loans then people that did not pay back. During the EDA procedure, I inspected each feature, which is considered to be available would have been available to investors considering an investment in the loan. And find some useful features that would have relationship with target label, as well as some features that I need to drop. The feature selection job in the following modeling part will base on these insights.

Attachment 1: Figure & Chart

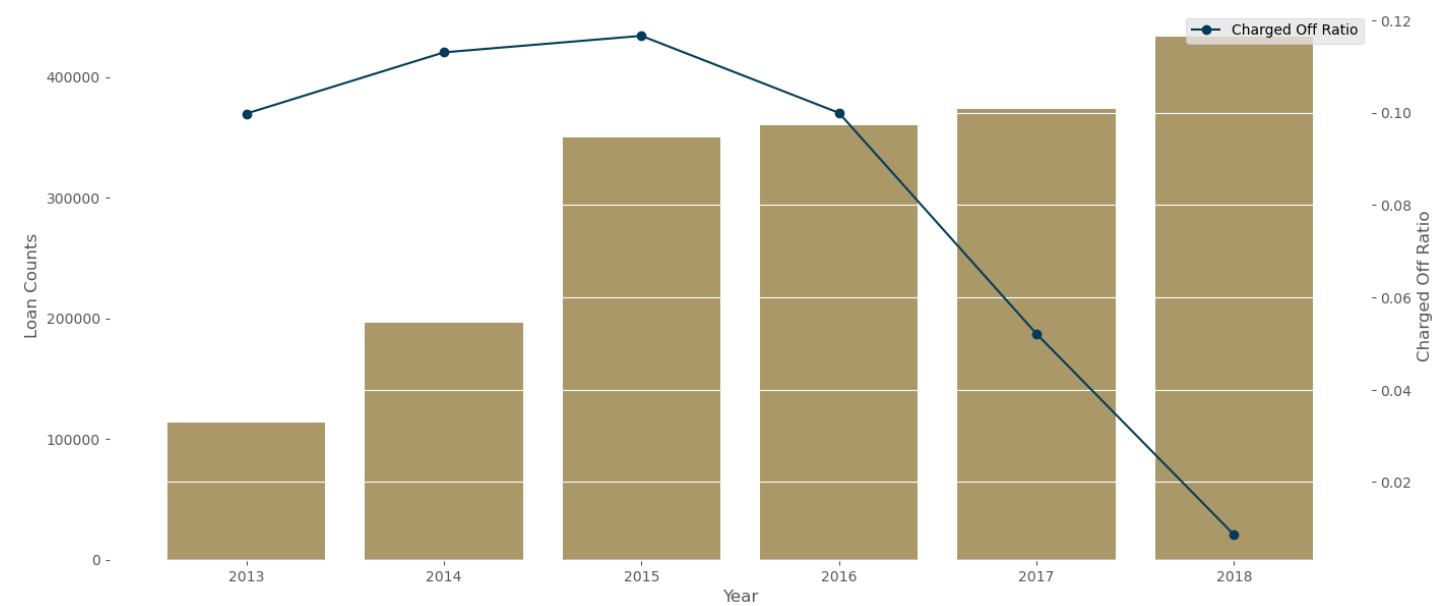


Figure 1. Loan Counts and Charged Off Ratio by Year

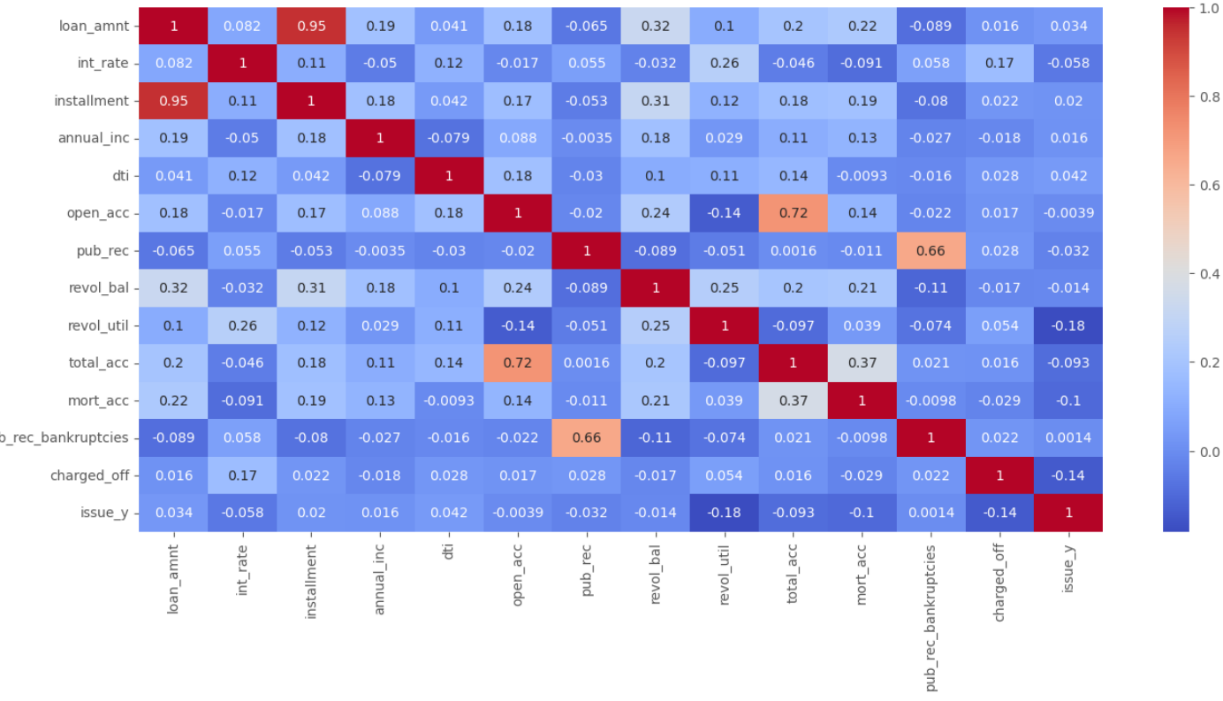


Figure 2. Correlation between Loan Status and Numeric Features

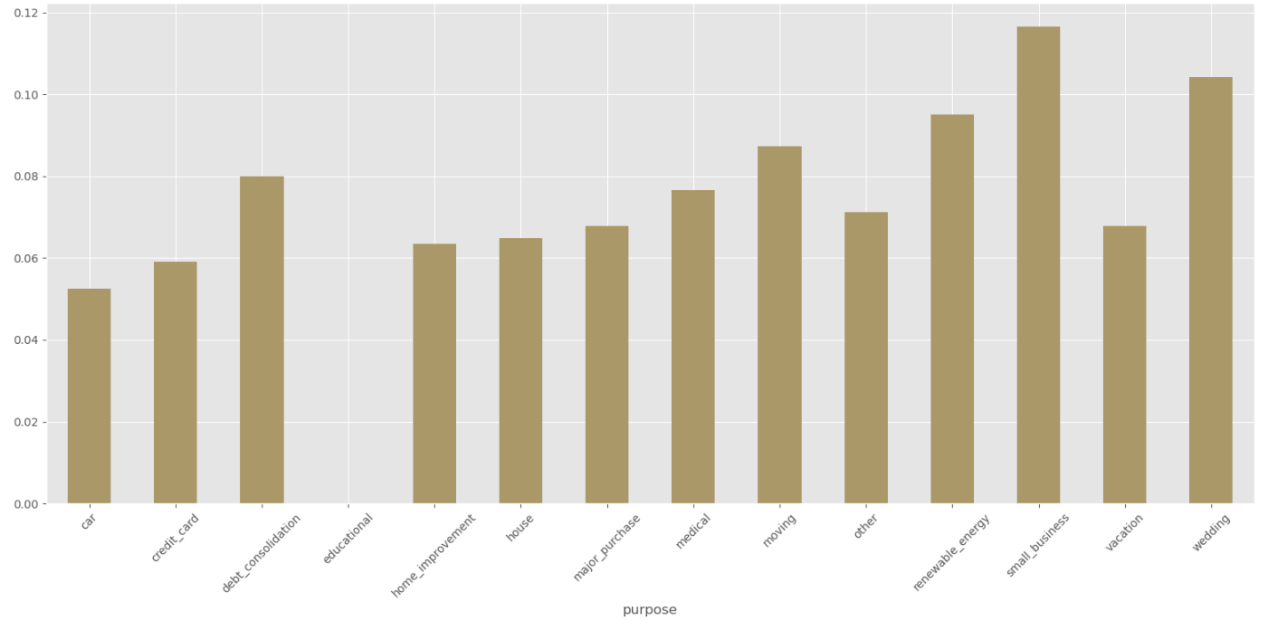


Figure 3. Charged Off Ratio by Loan Purpose

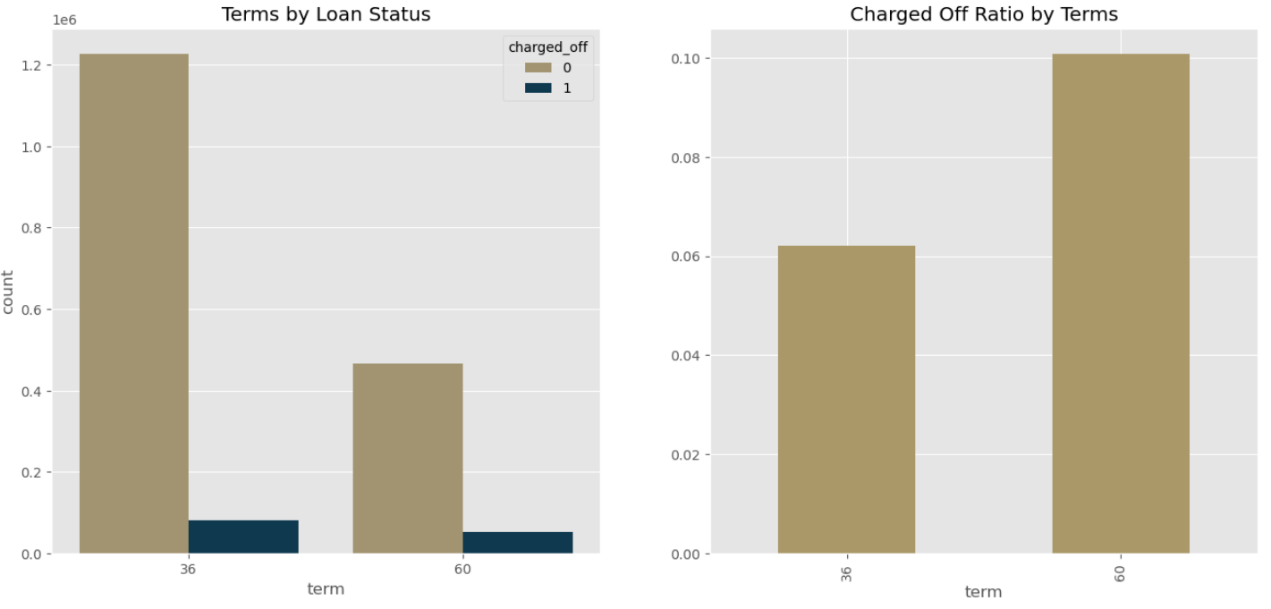


Figure 4. Charged-off Count and Ratio by Term

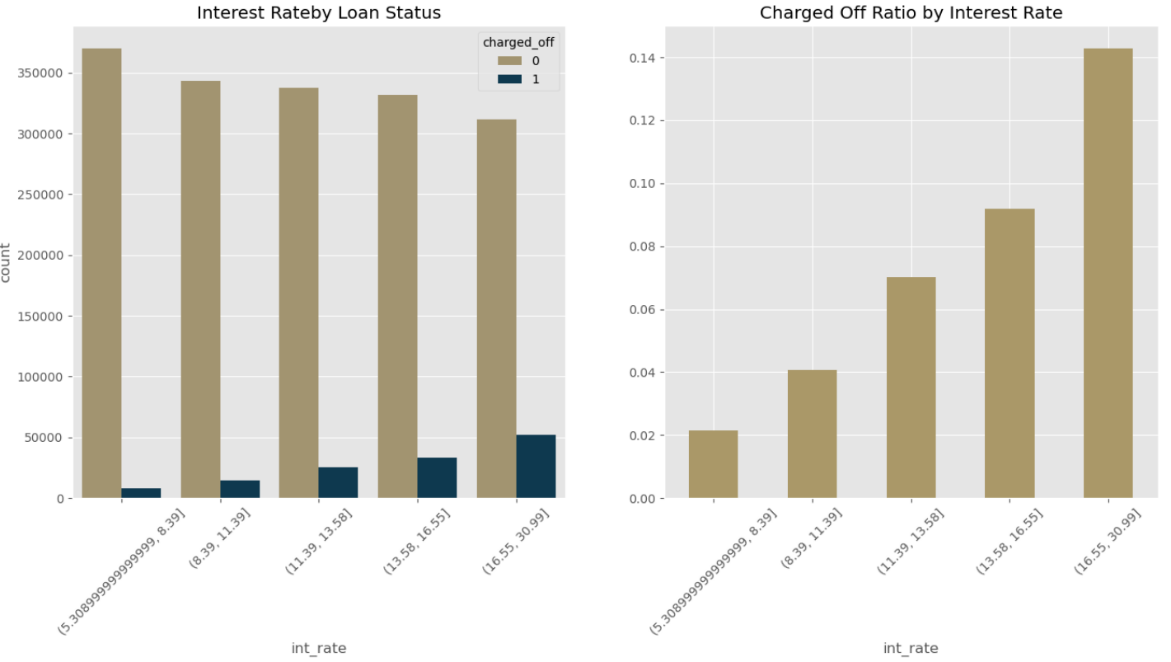


Figure 5. Charged-off Count and Ratio by Interest Rate

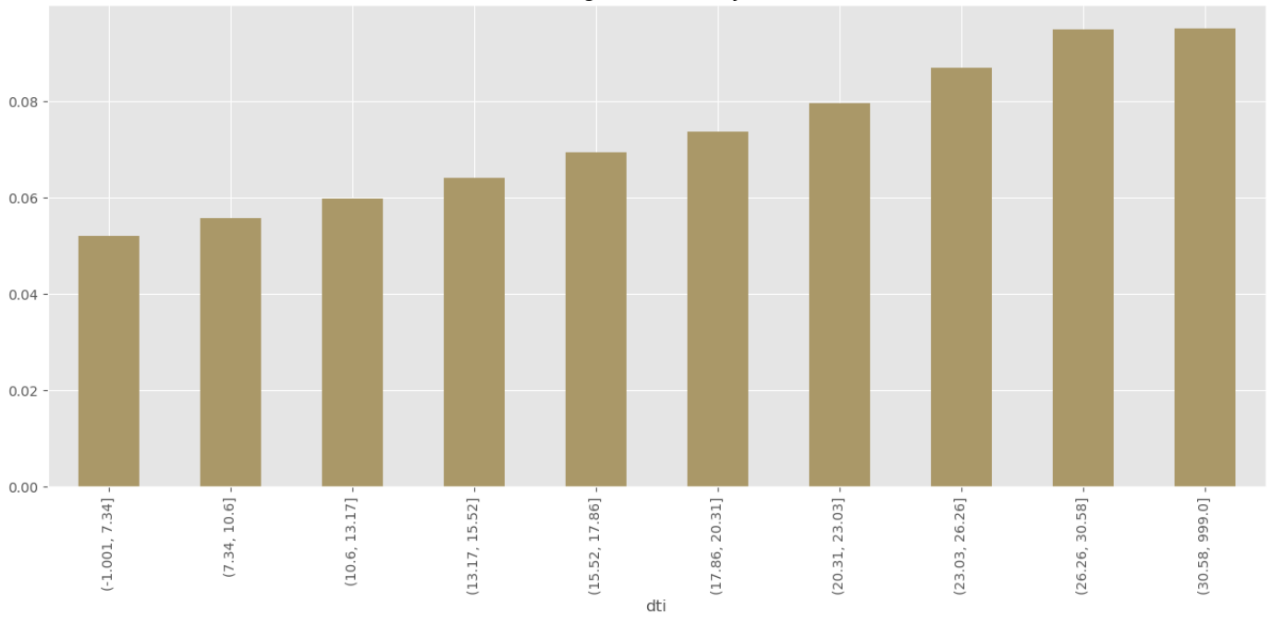


Figure 6. Charged-off Ratio by dti

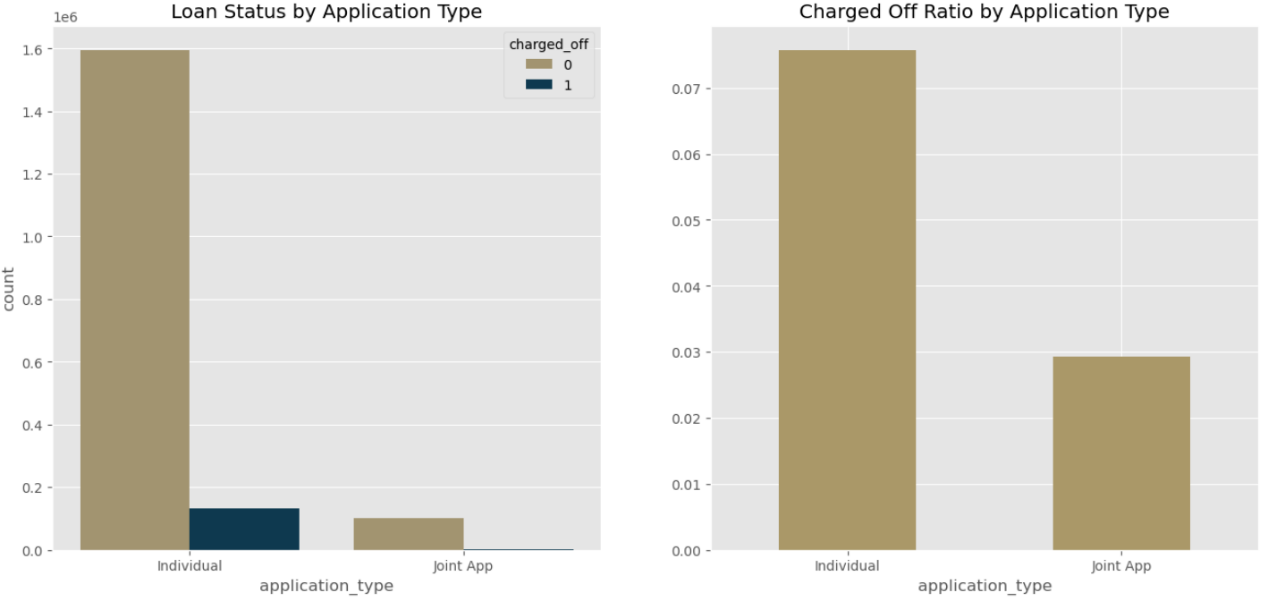


Figure 7. Charged-off Count and Ratio by Application Type

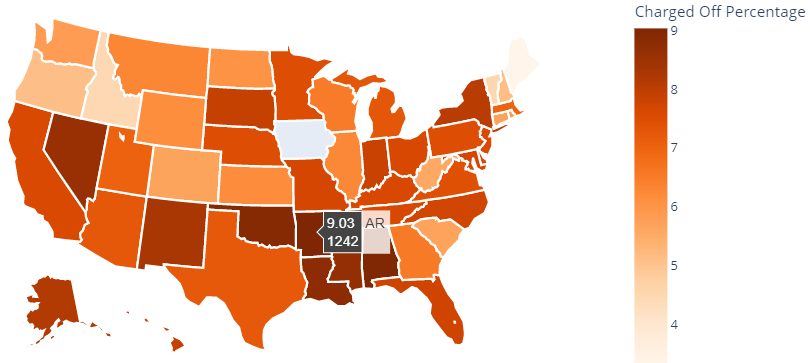


Figure 8. Charged Off Percentage of Each State