Jordan Jobes

**Lending Tree Loan Success Predictor Analysis**

**Final Report**

1. **Background**

Economic growth is accelerated by consumers having the ability to borrow. It is a balancing act between allowing credit worthy consumers to borrow and rates that provide a return to the bank. This balancing act becomes essential for a healthy economy.

If consumers default on the loan (don’t pay back the full amount of the loan contract) banks projected revenue decreases and may lead to unstable profitability. This could have a negative impact on the economy due to less credit available for consumers. The ability to predict loan success is important for the bank revenue projections and could also be an indicator of a shrinking economy.

1. **Data Source**

The dataset used in this analysis is provided by Lending Tree on kaggle.com. The data consists of ~2.3 million loans, ~170 features for each row, ranging from 2003 – 2018.

*https://www.kaggle.com/wordsforthewise/lending-club*

1. **Problem Statement + Method**

The goal of this analysis is to build a model which predicts whether a loan will be successful for Lending Tree. Their current success rate is 80%. The goal is to improve this.

The dataset contains loans in a variety of in-process stages and complete (fully paid or charged off). For this analysis we are only concerned with fully paid vs charged to feed the model with a known ending.

A loan which is charged off means the bank has given up on trying to get payments from the lendee and has closed the loan with outstanding debt.

1. **Data Cleaning**

The dataset had a large number of features that were duplicates, added no value or NaN values that needed to be updated. This was also where I investigated each feature to get familiar with them. I wanted to maintain as many features as possible in order to maximize the model potential. The model would likely work well with fico score, debt to income ratio, and marital status alone, but decision tree-based models may find hidden patterns that I cannot intuitively understand.

Missing Values:

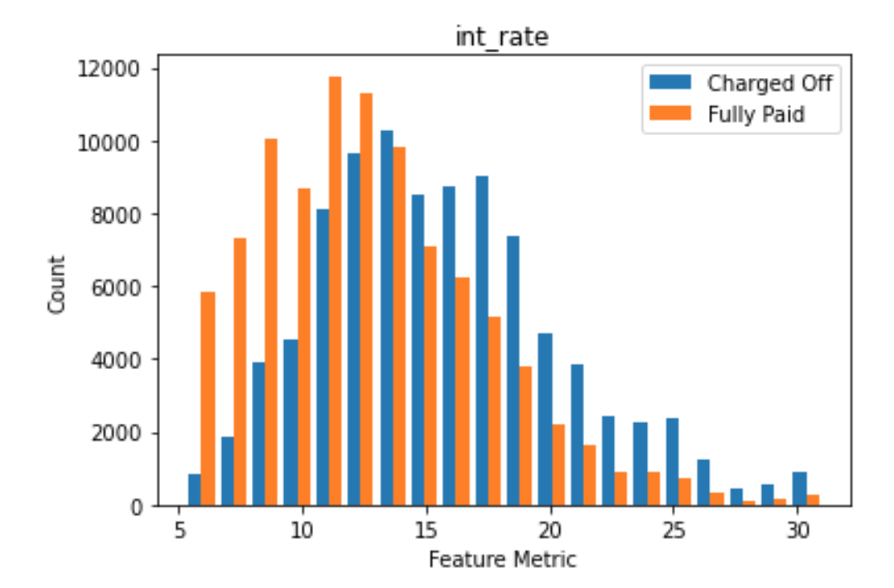
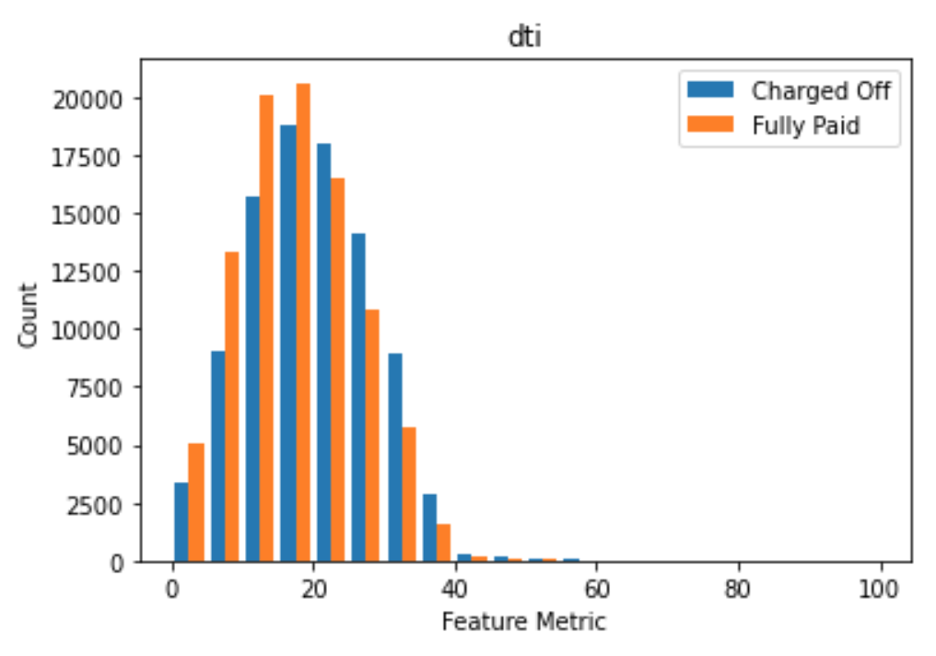
All numeric features with missing values were able to be replaced with zero and make sense. Many of these features were the number of bankruptcies, number of tax liens, so replacing all with zero made sense (if any poor metrics were present, Lending Tree would surely want to document this).

There were certain key features like dti (debt to income ratio) which is so important conceptually that I removed all rows without this information. This is justified because no bank would provide a loan without understanding this ratio.

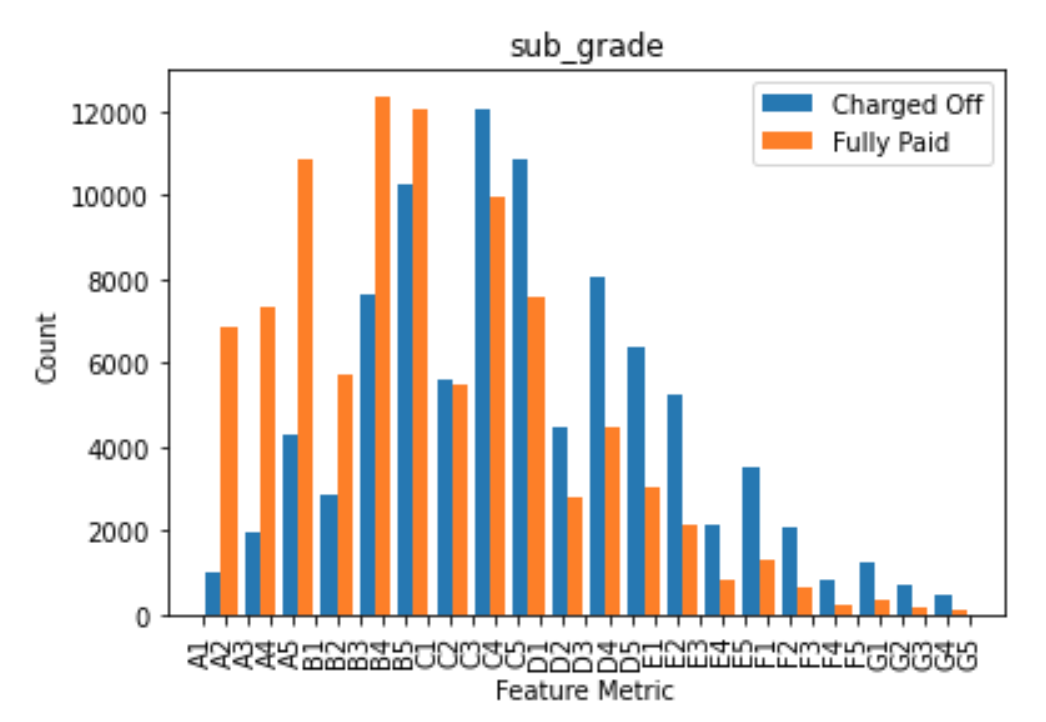
Luckily, the remaining object features with NaN values were minimal (<1%). I removed these from the database again under the assumption that the banks need this information to provide a loan. The missing values were likely due to how the data was extracted from Lending Tree.

1. **Exploratory Analysis**

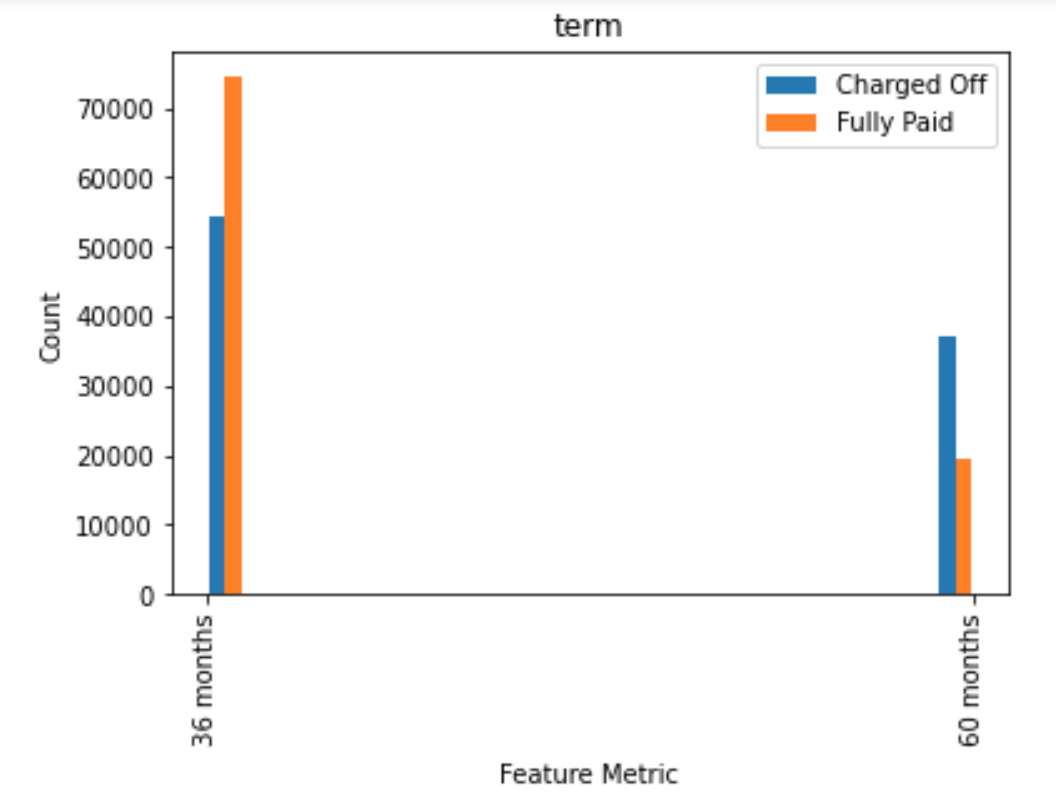
EDA yielded some inconclusive results. All of these charts are based on a 50/50 split of charged off vs fully paid so the charged off trends are more visible and comparable to fully paid. Even on critical features like debt to income ratio and interest rate (based mainly on credit health) there was no clear trend predicting a success loan payoff or not. Interest rate gives insight into Lending Tree’s guess of the risk of providing the loan. The larger the rate, the greater the risk in their eyes. Debt to Income is a big factor in calculating the interest rate.

****

The loan grade is another key variable in determining interest rates. There is a slight trend to A grades paying off vs F/G grades defaulting, but in both cases some A grades defaulted and some F grades paid off. Interesting!

****

The cleanest trends are shown below. The loan length (time to pay off loan) had an interesting trend: the shorter term lengths was a high indicator on full loan payment.

****

Overall, EDA gave some insight into the key features Lending Tree uses to calculate the loan risk. With no clear features to predict the loan success, we need to move to a machine learning model which can decipher the signal from the noise.

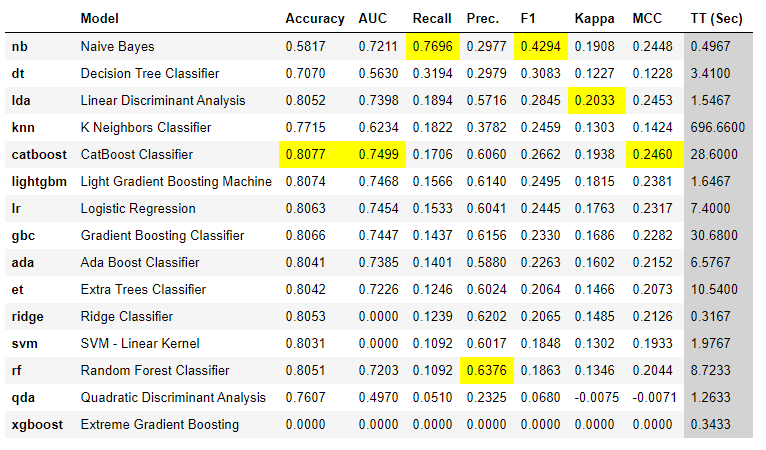
1. **Model Selection**

Due to the large amount of features and loans in this dataset I decided to go with pycaret to get a quick snapshot on which models performed the best.

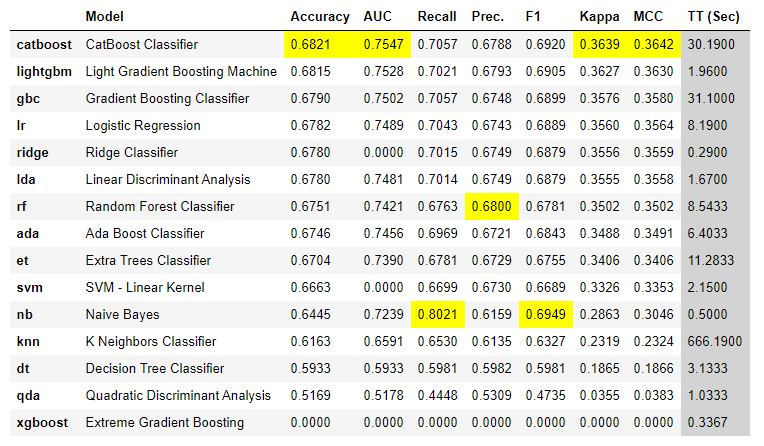
Normalizing Strategy

The goal of this model is to reduce False Positives (loans which defaulted but were predicted to succeed). False negatives (loans that succeed but were predicted to fail) are not a concern. Due to this, I performed training/testing with 2 sets of data: one with a 50/50 ratio of fully paid vs charged off and one with an 80/20 ratio.

80/20 split results:

****

50/50 split results:



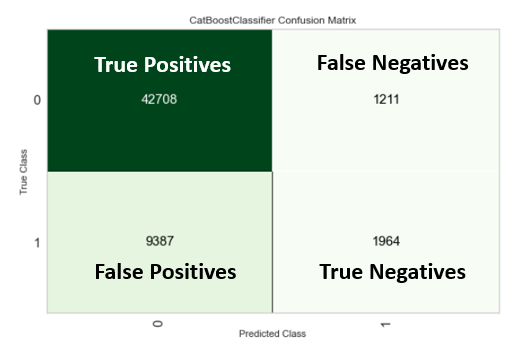
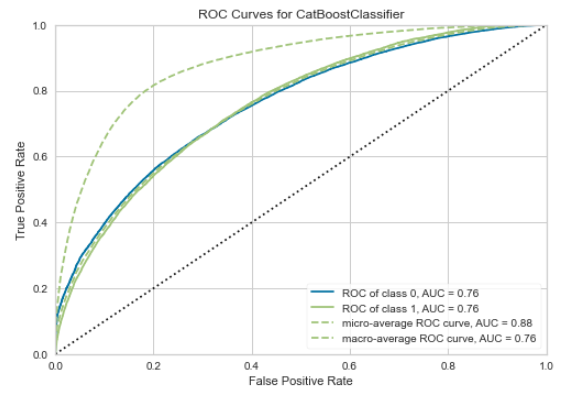
The 80/20 split produced higher accuracy while the 50/50 split produced higher recall. This makes sense because the 50/50 split has less negatives to classify incorrectly and more positive loans to classify correctly.

AUC gives a good representation of the balance between both true positives and false negatives. Both splits were around 75%.

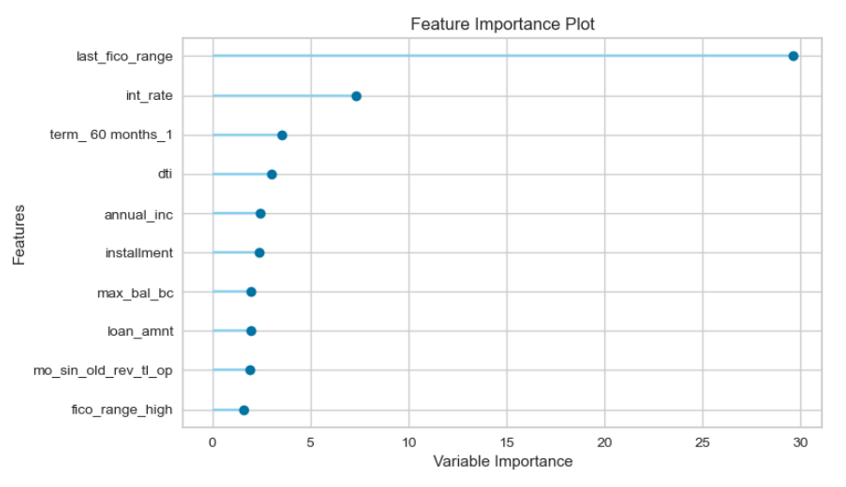
The Catboost Classifier and the Light Gradient Boosting Machine had the best AUC and Precision but Naive Bayes had the largest Recall by far at 80%. While this initially looks promising the high recall is due to a dilution of true positives and false negatives.

**Chosen model:**

The out of the box Lightgbm was chosen over catboost due to a lower false negative score and much less computing power required. This also performed very well in both splits which means it is more robust to handle varying customer conditions.



The model found the most important conceptually impactful features as well:

****

1. **Key Take-aways**

Reviewing the feature importance (in order) for Lending Tree:

**#1 Last FICO Range (last\_fico\_range)** -

**#2 Interest Rate (int\_rate)** - Being the #2 feature shows Lending Tree is not too far off on their calculations for predicting loan success. Optimizing the variables going into this calculation would be the next step.

**#3 Loan Amount (loan\_amnt)** - The smaller loans are more likely to succeed. Focus on smaller loan amounts at higher volumes vs less loans at larger lending amounts.

**#4 Debt to income ratio (dti)** - This basically tells LendingTree how much of their income is available for daily expenses + debt payoff. 40% max dti is the maximum banks would like to see in order to approve a mortgage.

**#5 Annual Income (annual\_inc)** - This ties into debt to income ratio.

The remaining key features tie to the above 5 mentioned.

This analysis is a perfect example of how machine learning can filter the noise and yield some interesting results. Most features had a 50/50 split between charged off and fully paid as seen in the EDA section. When the features are combined with machine learning, I was able to match the Lending Tree algorithm success rate used. Maybe they use Lightgbm!

While this exercise did not meet the goal of improving their success rate, a lot was learned. The key features in predicting success could be used to quickly vette loans that are not worth investigating further. This would lead to more efficient loan approval decision making which is very valuable.

1. **Further Investigation**

* The grade of loans are not representative of loan success. This seems to be the biggest error in Lending Tree’s calculation for interest rate. Investigate why this is.
* Investigate the loans LendingTree did not approve to see which loans may have been worth approving. This would shed light on their full confusion matrix (true negatives and false positives).
* Review the threshold used in the model to further reduce false negatives. This has to be weighed against losing overall revenue to improving loan success accuracy.