**Introduction**

In this report, we detail our analysis of a dataset related to finance. The dataset is a loan dataset consisting of loan information such as loan amount and loan term as well as application information such as employment, credit scores, home ownership, annual income, and other information. This dataset is synthetically created as part of a competition that is held at the conclusion of an online EDX course.

While the competition is over and the original dataset no longer obtainable, we obtained what seems to be a copy of that dataset from this GitHub page: <https://github.com/ZellW/MSCapstoneComp>

While working with the dataset, we address the data quality issues present in the dataset. We also explored the data to find interesting patterns in the data that can differentiate between fully paid loans and charged off (defaulted loans). We then attempt to fit a decision tree to create a prediction model.

Finally, we explain the challenges faced working with the dataset and discuss the accuracy of the prediction model based on the decision tree.

**About the dataset**

The dataset is a historic information about granted loans and their ultimate results, it has over ten thousand observation, with 19 features. The features are detailed below, which is obtained from the competition website:

• Loan ID: A unique Identifier for the loan information.

• Customer ID: A unique identifier for the customer. Customers may have more than one loan.

• Loan Status: A categorical variable indicating if the loan was paid back or defaulted.

• Current Loan Amount: This is the loan amount that was either completely paid off, or the amount that was defaulted.

• Term: A categorical variable indicating if it is a short term or long term loan.

• Credit Score: A value between 0 and 800 indicating the riskiness of the borrowers credit history.

• Years in current job: A categorical variable indicating how many years the customer has been in their current job.

• Home Ownership: Categorical variable indicating home ownership. Values are "Rent", "Home Mortgage", and "Own". If the value is OWN, then the customer is a home owner with no mortgage

• Annual Income: The customer's annual income

• Purpose: A description of the purpose of the loan.

• Monthly Debt: The customer's monthly payment for their existing loans

• Years of Credit History: The years since the first entry in the customer’s credit history

• Months since last delinquent: Months since the last loan delinquent payment

• Number of Open Accounts: The total number of open credit cards

• Number of Credit Problems: The number of credit problems in the customer records.

• Current Credit Balance: The current total debt for the customer

• Maximum Open Credit: The maximum credit limit for all credit sources.

• Bankruptcies: The number of bankruptcies

• Tax Liens: The number of tax liens.

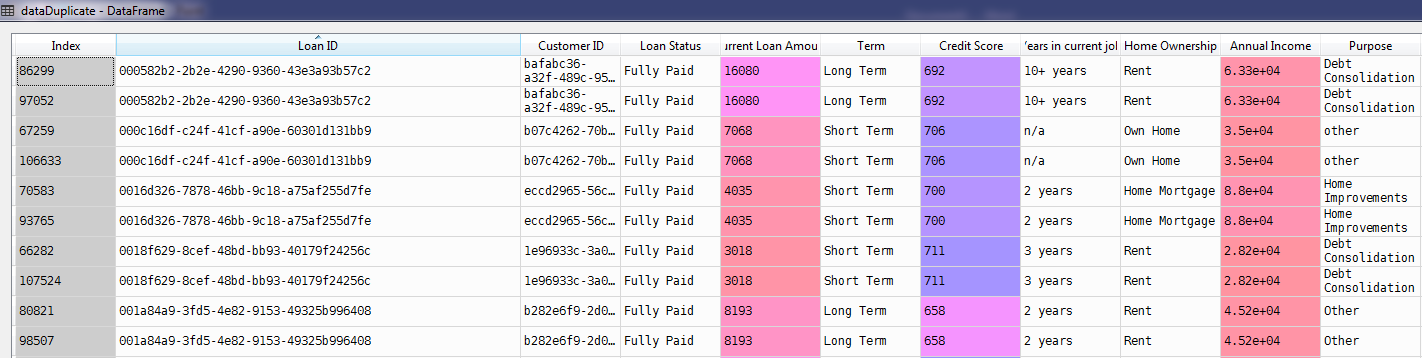
The question we want to have answered by researching this dataset are as follows:

* What is the ratio of fully paid loans and charged off loans?
* Does having a high credit rating lead to higher chances of a fully paid loan?
* What are the relationship between the term of the loan and the ratio of good and bad loans?
* Does the loan amount have any effect on the chances of the loan being paid back in full?
* What loan and applicant characteristics can best predict whether a loan will be paid back or not?

**Pre-processing the data**

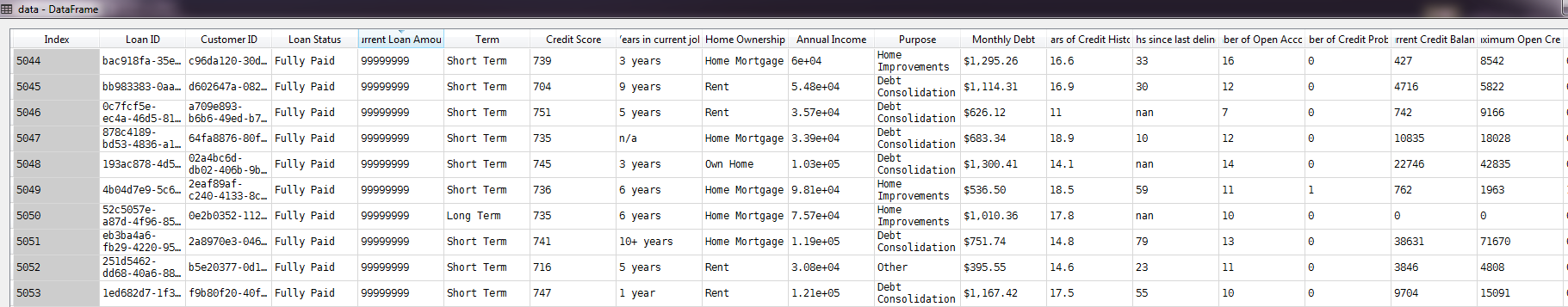
The dataset is stored in the form of a comma separated value file ( CSV ), with the first row being reserved for the column names.

After loading the data, we check for duplicate values and found the following data:



After looking through the duplicate data, it is clear that the rows have the exact same attributes, and thus can be safely removed, leaving only the rows with unique Loan ID values.

Next, we found that some rows have their Current Loan Amount value set at all nines:



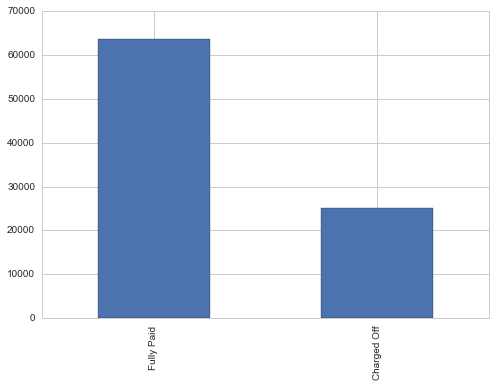
When we examined the dataset with only those rows, they are all fully paid loans. To reduce bias in the dataset during fitting the decision tree, we set the values to -1.

Among other cleaning steps we did:

* We found 2 rows that have invalid Maximum Open Credit values. We dropped them from the dataset altogether because the amount is statistically insignificant compared to the size of the dataset.
* We found credit score that defy the credit scoring rules as set by the competition website. Upon further inspection, we found that the values all have 0 appended at the end, possibly by accident. We fix the anomaly by dividing the values by 10.
* The loan purpose categorical feature contains the values “other” and “Other”. All instances of “other” are changed to “Other”.
* For home ownership, there are values of “HaveMortgage” and “Home Mortgate”. All instances of “HaveMortgage” are changed to “Home Mortgage”

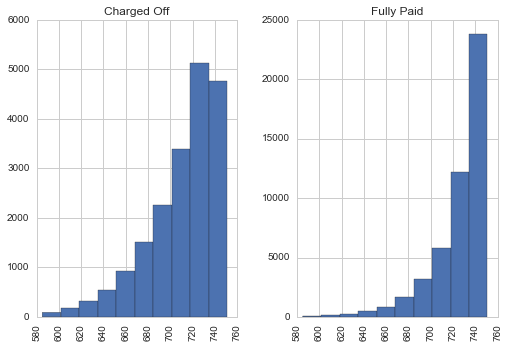
**Exploring the data**

First, we obtain the class distribution of the dataset:



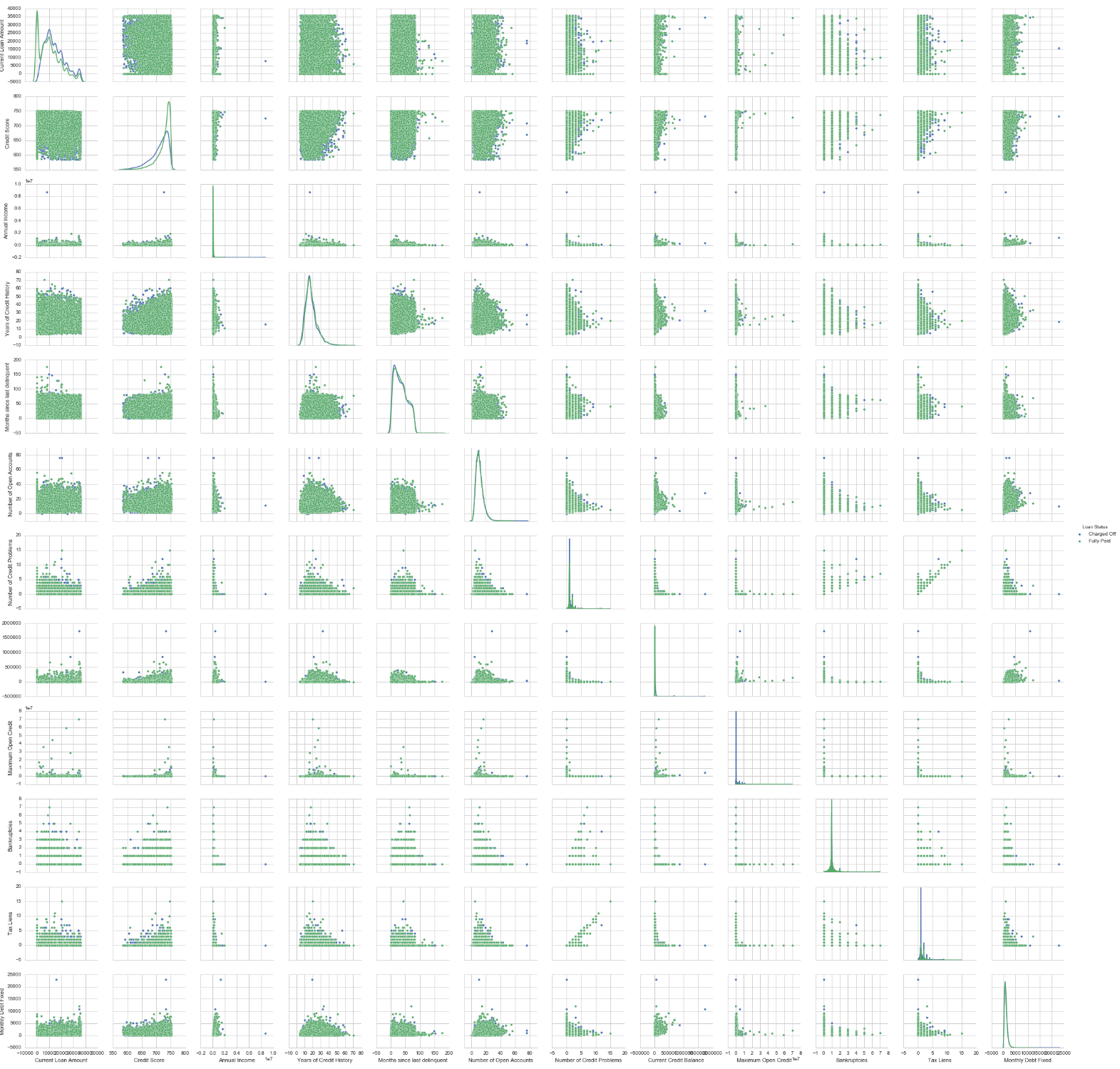
We see that most of the loans in the dataset are mostly fully paid, with a ratio of about 70% Fully Paid and 30% Charged Off loans.

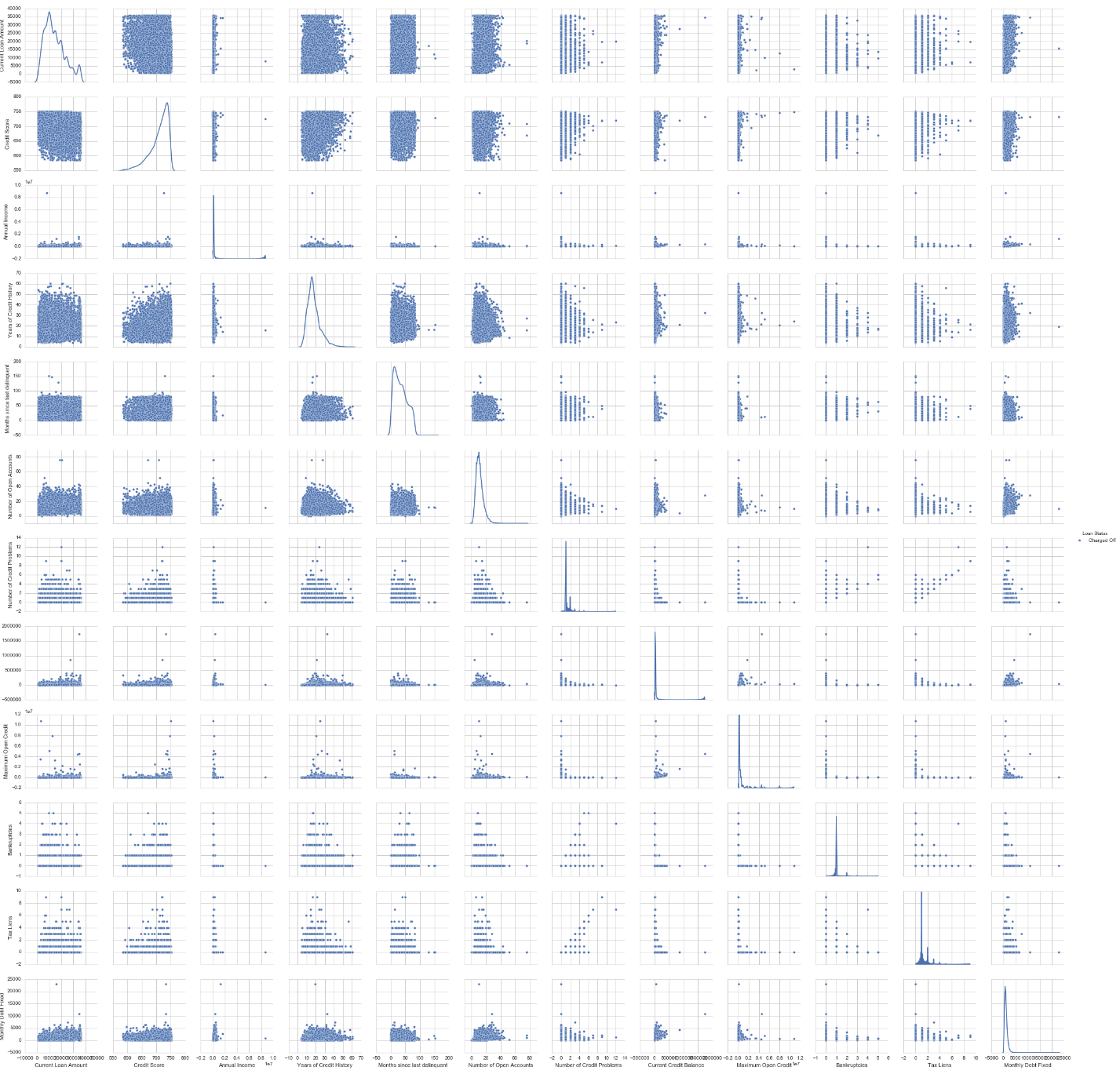
Next, we plot a histogram of credit scores, divided by loan status:



Here, we the differences between the distribution between credit scores when comparing between fully paid and charged off loans. Most of the loans are at the top of the credit score scale for both types of loans, matching the narrative where most of the granted loans depend of credit score.

Finally, in order to find interesting features that could be used as predictors, a scatter matrix of all numeric columns in the dataset is done. The first image uses both types of loan, while the second image plots only the charged off loans.





Looking at all the possible relationship between any two numeric features, there is little in ways we can differentiate between good and bad loans.

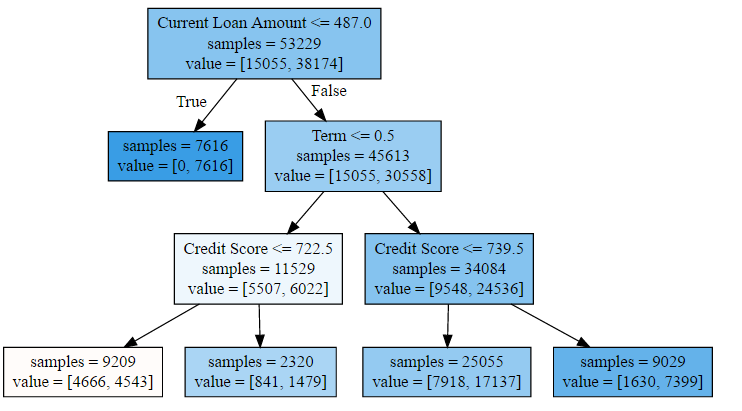
**Prediction Model**

To hopefully get some way to determine bad loans, we employ a decision tree, which is a machine learning algorithm that uses attribute tests to split samples into classes.

To do this, we have to convert the categorical variables into dummy features. Some, like loan term and employment length, are encoded into a single variable with ordinal values. Others, such as home ownership and loan purpose, and split into dummy variables.

Using the scikit python library, the data was split -- 60 to 40 -- into training sets and test set. Then, a decision tree was fitted to the training set. We set the decision tree’s maximum depth to 3.

The following diagram shows the decision tree:

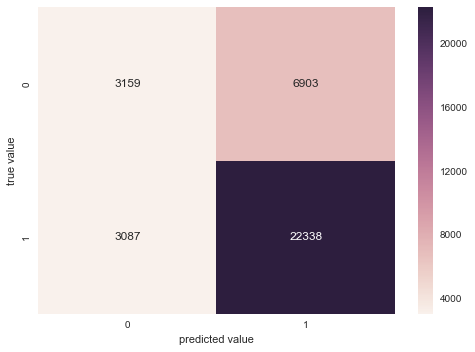


The only pure sampling was the one with all the loan amount set to -1. As explained earlier, all the loans with the outlier loan amounts are paid off loans, and thus the decision tree reflects that.

The most important thing is that the leaf nodes are incredibly impure; there are no combination of feature description that can reliably differentiate between the loans.

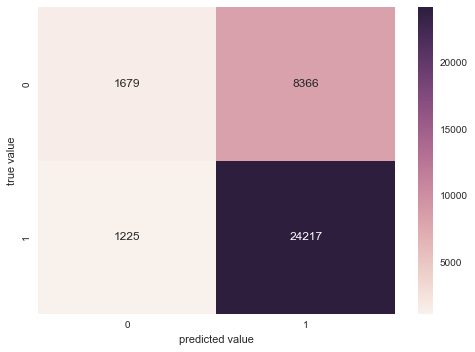
Initial scoring show the decision tree to be 73% accurate. However, since the initial dataset shows a skewed class distribution, this is no better than guessing that all loans will be fully paid. We need to find the ratio of True Negative, and False Positive.

The following image shows the confusion matrix:

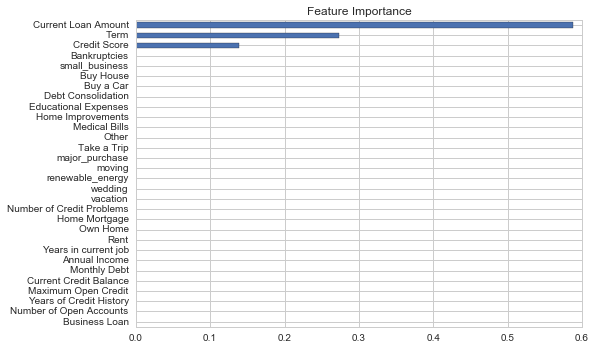


( 1 denotes fully paid, while 0 denotes charged off )

Here, there are 3159 rows correctly predicted as being charged off loans, while 6903 rows are incorrectly labelled. Hence, there is only about 40 percent accuracy for determining charged off loans.

Increasing the depth of the tree to 5 gives little effect on the confusion matrix:

The following graph shows the feature importance determined by the decision tree:



From this graph, it is clear that the various categorical features are not useful at all at determining the loan types. The only features that have some effect are the loan amount ( given the -1 values ), loan term, and credit score.

**Conclusion**

Using the decision tree, we have not been successful in obtaining the features that can differentiate between fully paid loans and charged off loans.

According to existing articles on loan prediction, the best way to predict this kind of data is by using random forest, which is a collection of decision trees created to predict labels based on some function across all trees.

Another way to process this dataset is to reduce the distribution bias between loans by introducing additional charged off loans, created synthetically.

**Code Location**

The python script used to process the dataset can be found in *work.py*, located in the same archive as the softcopy of the report.

REFERENCES

* Loan Granting Binary Classification - <https://gallery.cortanaintelligence.com/Competition/Loan-Granting-Binary-Classification->1