**CAPSTONE PROJECT #2**

**Lending Application Result Predictions**

|  |  |
| --- | --- |
| **Introduction** | **2** |
| **Data** | **2 & 3** |
| **Exploratory Data Analysis** | **3 & 4** |
| **Machine Learning** | **5 & 6** |
| **Conclusion** | **6** |

**Introduction:**

LendingClub is a pioneer in peer-to-peer lending founded in 2006. It is currently the largest P2P lender in the world.

In order for an applicant to determine whether they will qualify for a loan, he/she will be required to input a significant amount of personal information into the LendingClub website. Given the concern about personal privacy & data security, the goal of this project is to create a simple calculator that can predict whether the requested loan can be approved with minimal personal data needed. This calculator can be used as a first step in case potential customers don’t want to provide all the details a full application will require.

**Data:**

The lending data is sourced from Kaggle website. It is an agglomeration of monthly data sets from the LendingClub.com website. Data is from 2007 to 2018 (accepted\_2007\_to\_2018Q4.csv for approved data & rejected\_2007\_to\_2018Q4.csv for rejected data).

The rejection dataset has significantly less features than the approved dataset. As such, for this exercise all features that are not present in the rejection data set are removed. Over 140 features are removed. Features that are kept include: Loan\_Amount, Debt\_To\_Income\_Ratio, State, Employment\_Length, and Risk\_Score.

For the remaining features:

* State is categorized into regions based on the Bureau of Economic Analysis region split as each separate state is not expected to have significant influence on the prediction
* Employment\_Length is converted from string to int through removing the word “year” from the end, and also any length <1 is relabeled as 0 & 10+ is relabeled as 10.
* Risk\_Score is normalized because Reject data post November 5, 2013 uses Vantage score and FICO before the date. Approved data all use FICO score (mean of primary and secondary applicant when there is secondary applicant).
* Approved is labeled as 1 and Rejected is labeled as 0.
* Any Loan\_Amount more than $40k is removed as anything higher is considered non-standard loan and are treated as outliers.
* Any loan with NaN data sets are removed as the data set is large and this will not affect model result.
* Note that Debt\_To\_Income\_Ratio in the approved data set is the mean between primary applicant and secondary applicant. Reject dataset did not include secondary applicant info.
* Applications Debt\_To\_Income\_Ratio of greater than 0.5 have been removed as these are deemed to be outliers.
* Data are randomly shuffled to get more accurate cross validation results using StratifiedKFold (random state 42).

Once the separate data sets are cleaned up, both files are concatenated together based on column names. After concatenation, the entire data set goes through stratified sampling based on Loan\_Amount to get a sample pool of 5% of original. The reason for this significant data reduction is because of limits with computation power. Once the sample pool is created, we then apply stratified sampling based on Loan\_Amount again on the Rejected dataset in order to achieve the 30% approved and 70% rejected ratio for machine learning. Prior rebalancing, the ratio is in the 10% range.

The final data wrangling steps taken are to apply sklearn LabelEncoder & OneHotEncoder on region in order to break out the features into binary columns.

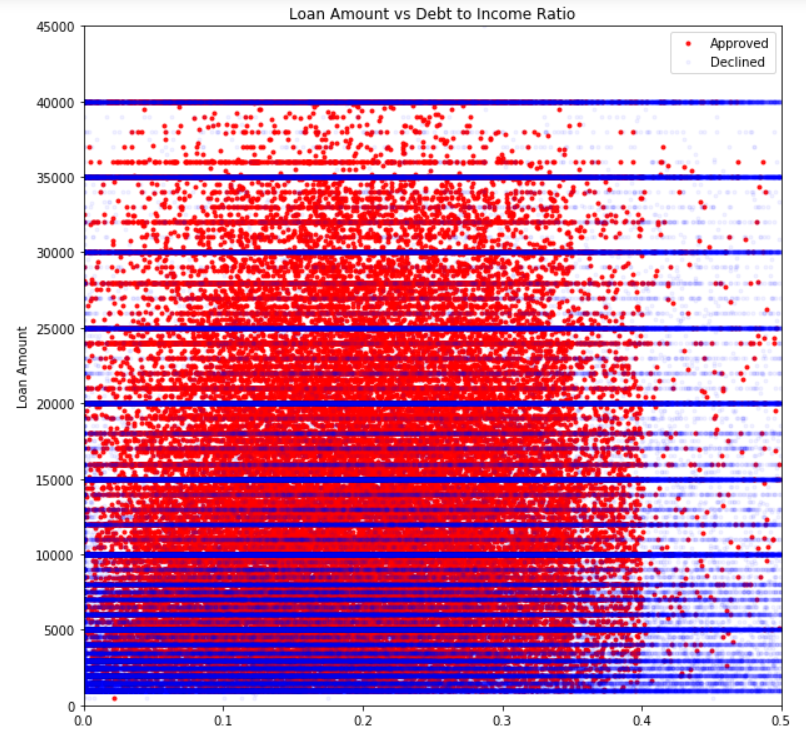
The final dataframe is saved down into MS\_Data.csv and reloaded into a new Jupyter notebook in order to open up more memory for machine learning calculations.

**Exploratory Data Analysis:**

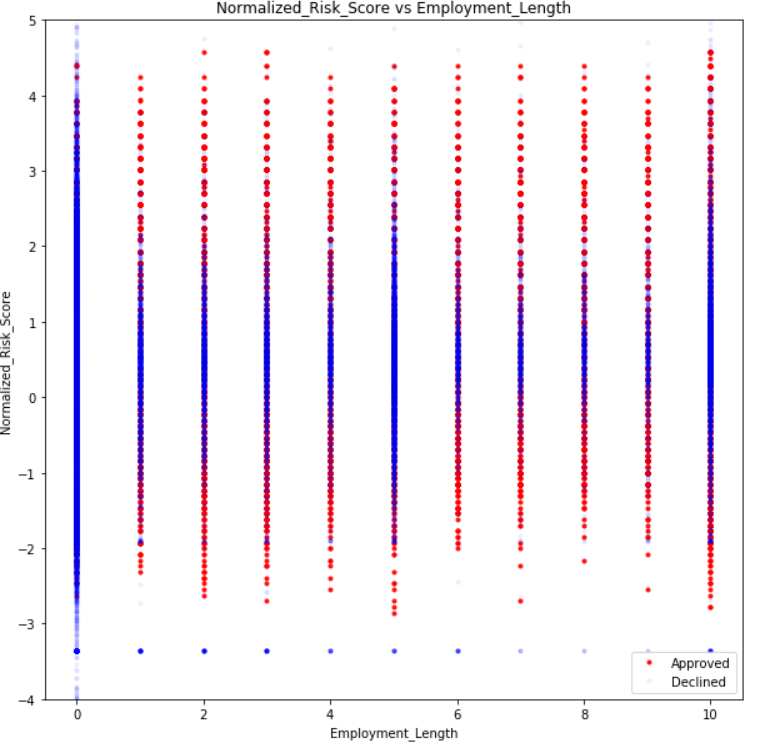
A correlation matrix is ran on the truncated data set. Employment\_Length has a much higher correlation and anticipated. Significantly higher than any other features.



Mapping out data points of Loan Amount vs DtI ratio shows an interesting spread for approved and rejected applications. Approved applications are more evenly spread out while the rejected data tend to focus heavily on particular loan\_Amounts. Also, what is interesting is that few approved transactions go beyond 0.4 DtI ratio while it is common place for rejected loan. It is noted that 0.4 DtI is the maximum limit set by Lending Club in order for loans to get approved (stated on their website).



From the graph below we can see that there is a significant concentration of rejected loans sitting in the 0 (<1year) Employment\_Length category.



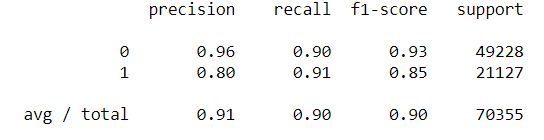
**Machine Learning**

Since the goal is to predict approve and decline, classified supervised learning is used.

The three methods tried are Logistic Regression, Random Forest, and XGBoost.

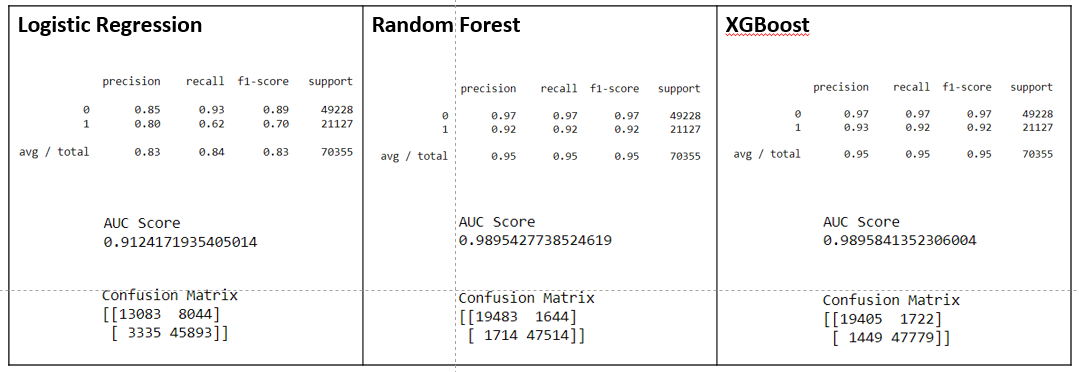
**Baseline**

Baseline prediction is based on non-Machine Learning techniques. Assumption if an application has higher than 40% debt to income ratio and employment length less than 1 year, the loan is rejected. The 40% maximum debt to income ratio is based on the recommendation from LendingClub but obviously based on the graph from above, there are exceptions. Employment less than 1 year is also based on the graph from above which shows rejected loans heavily skewed in the sub 1 year category. Note that loans over $40k (maximum standard loan amount) have already been removed in the data wrangling part so this is not applied in the baseline prediction.



The results are actually very strong with the non-MS method.

**Supervised Learning**



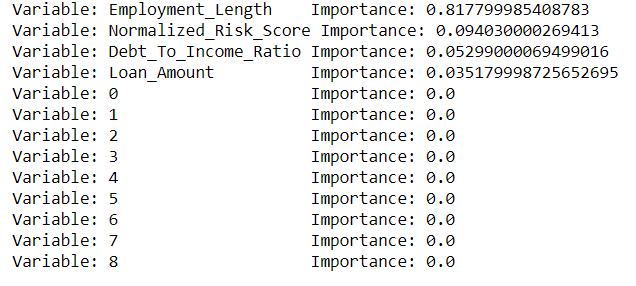
The AUC scores are calculated using cross-validation method applying 3 folds and random state of 42.

XGBoost has the highest AUC Score demonstrating it had the best performance.

Random Forest’s results are very close to that of XGBoost while Logistic Regression actually performed worse than Baseline.

**XGBoost – Deep Dive**

Taking a closer look at XGBoost we see that Employment\_Length has an unusually large influence relative to the other features. This result is similar to the correlation table seen previously. Random Forest has arrived with similar results but a slightly lower importance assigned to Employment \_Length.



**Conclusion:**

Based on the results, we can conclude that prediction is heavily dependent on Employment Length. Risk Score and Debt to Income ratio also have influence on the prediction. Unfortunately, the region features do not add any predicative ability to the models.

Due to the heavy reliance on Employment Length, non-machine learning method is also fairly accurate. This raises the question about if the baseline method can be used in certain circumstances as it is much easier to explain to the customers than machine learning methods.

However, in order to get the optimal prediction based on the limited data provided per applicant, XGBoost and Random Forest approaches are still far more accurate.

With AUC score close to 99%, both RF and XGBoost can be used as a preliminary application model by LendingClub.

Additional areas of expansion would be to get more features for rejection dataset as there are very few features available relative to the approved dataset.