Name: Muley, Tushar

Title: Week 11 – 12 Term Project

Date: June 5, 2021

Lending institution have a basic need to be able to accurately determine the riskiness of the individuals applying for a loan. It is one of their biggest concerns as the margin for return is low and cost of servicing loans is high. Automation has assisted in lowering costs for servicing loans but loan losses add to the cost. All financial lending institutions want to be able to lend to as many individuals as possible. They want to see a recognized return in the form of returned principal and income as interest. The goal is to identify, which attributes better determine the riskiness of the individuals that financial organization lending money to. I am proposing we determine if more features or different features are need to predict the riskiness of the individuals before lending money to them.

Data set: <https://www.kaggle.com/laotse/credit-card-approval?select=credit_card_approval.csv>

**Question:**

My proposal is to use Kaggle data that contains the below listed features and determine if they will assist in determining the riskiness of potential applicants.

**Data Exploration:**

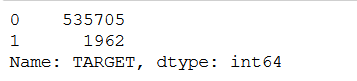
|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| ID | Client Number |
| CODE\_GENDER | Gender |
| FLAGOWNCAR | Is there a car |
| FLAGOWNREALTY | Is there a property |
| CNT\_CHILDREN | Number of Children |
| AMTINCOMETOTAL | Annual Income |
| NAMEEDUCATIONTYPE | Education Level |
| NAMEFAMILYSTATUS | Marital Status |
| NAMEHOUSINGTYPE | Way of Living |
| DAYS\_BIRTH | Age in days |
| DAYS\_EMPLOYED | Duration of work in days |
| FLAG\_MOBIL | Is there a mobile phone |
| FLAG\_WORKPHONE | Is there a work phone |
| FLAG\_PHONE | Is there a phone |
| FLAG\_EMAIL | Is there an email |
| JOB | Job |
| BEGIN\_MONTHS | Record month |
| STATUS | Status |
| TARGET | Target |

**Figure 1: Table of all column names present in the data**

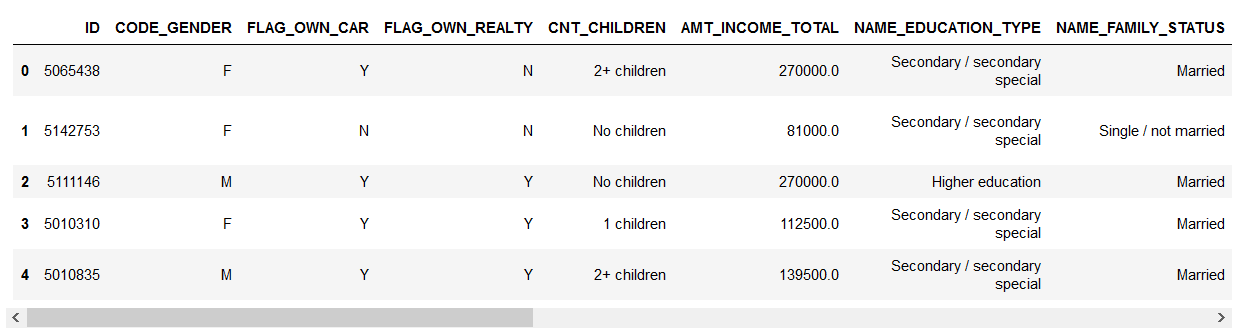
I have profiled the Kaggle data and found some note worthy item I want to share in this analysis. As part of my Milestone One (1) I profiled the data and performed graphical analysis to better understand what kind of data is available and how best to use this data.

Interesting facts about the data:

* 537,667 rows
* Zero (0) missing records
* 333,832 applicants are female
* 358,317 applicants have some type of secondary education
* 384,003 married applicants
* Target variable are indicators
  + - * 0 – Low Risk
      * 1 – High Risk

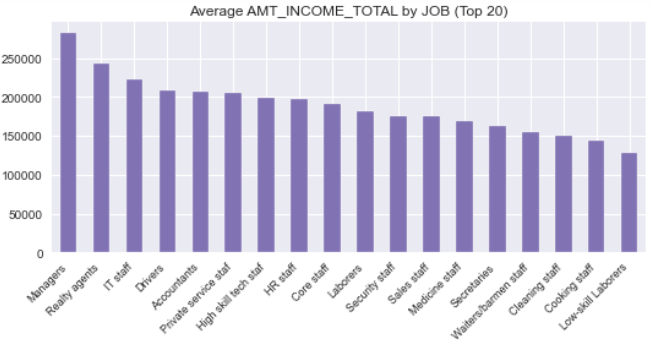


* Most of the data points are categorical which I will change to binary



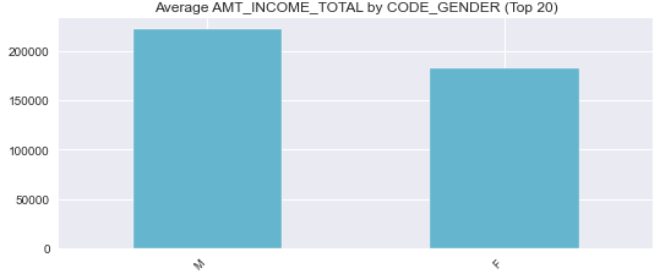
Below I have included a few key graphs from my graphic analysis.

**Graphs and Observations:**



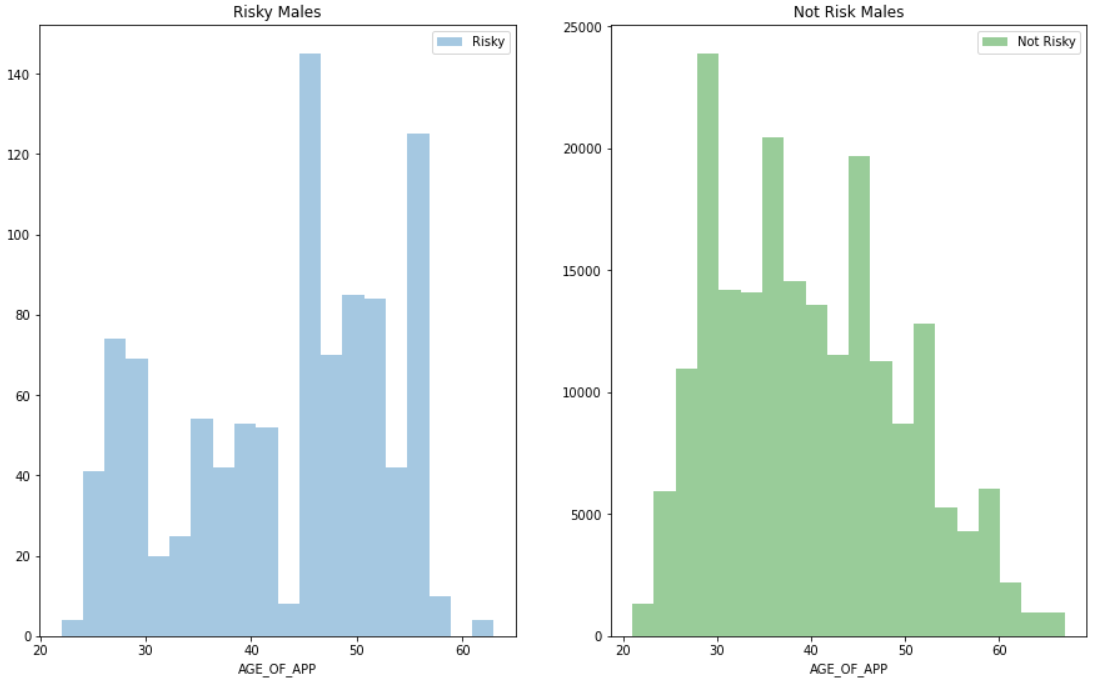
**Figure 2: Average income by job description**

The above graph shows the average income by job description provided by each applicant. Our top three applicants are in highly desirable job positions and average income is strong.

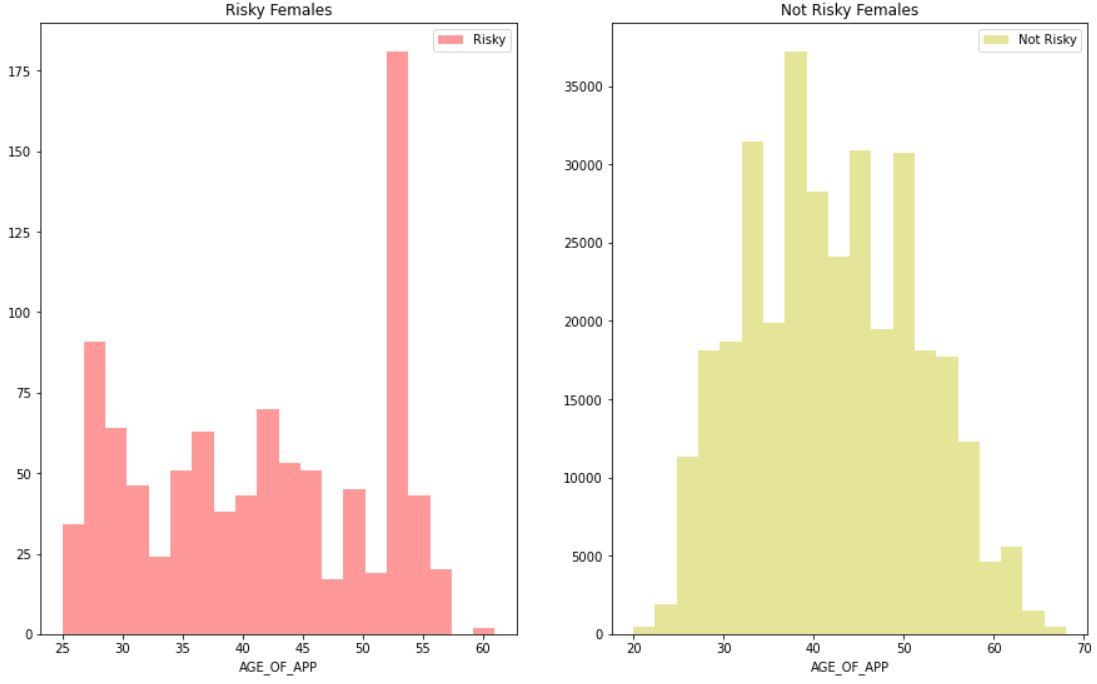


**Figure 3: Split between male and female applicant as show by average income**

The above figure show what we have seen in recent trends that males tend to make more than females on average

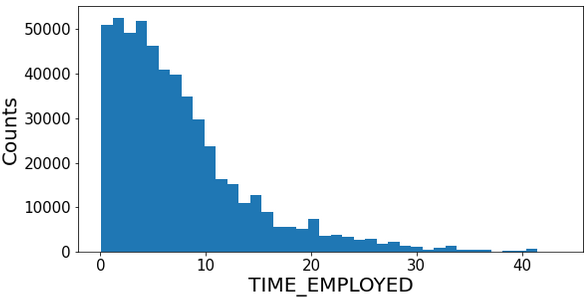
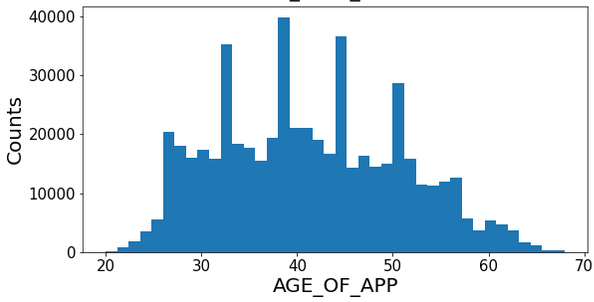


**Figure 4: Number of Males by Risk and Age**



**Figure 5: Number of females by Risk and Age**

The above two figures (Figure 4 and 5) show our applicant by gender, age and riskiness of default.

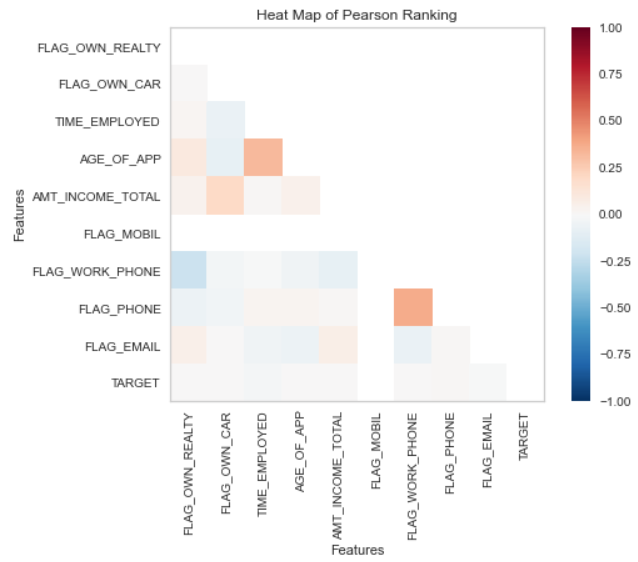


**Figure 6: Left– Applicants by Age Right– Applicants by Time of Employment**

****

**Figure 7: Scatter plot of Age of Applicants and Time Employed**

The above graphs show applicant by age and time of employment. All these graphs tell us a lot about the population of the dataset. For example, applicants are mostly females, high earning professionals, spreads across wide age group who are early in their employment history. Risk exposure is low but I will speak about that in a later section. The above scatter plot indicates we have a good mix of individuals from all age groups and time of employment. This gives a good mix that is not skewed in any one direction. I can always review and possibly remove the older age groups and time of employment to make sure that demographic part of the data is not lending a bias to the data.



**Figure 8: Pearson Ranking Heat Map**

**Dimensionality Reduction**

**Converting Variable Types and Feature Reduction**

* I dropped these fields: mobile phone, work phone, email, number of months loan was open, number of children,
* Converted days of birth into years, as well as length of employment in months.
* Checked for missing data

In my Milestone Two (2) I want to briefly explain what enhancements I performed on the data to make it more model ready. I dropped the following fields as they did not provide high correlation to other features. Having a mobile phone, work phone, email did not give any correlation to riskiness of the individual. I also dropped number of months the loan has been open since we are interested in understanding if we are making loans to applicants who will return our funds with interest. I drop number of children as that has no impact on the loan or riskiness of the applicant.

**Using One Hot Encoding:**

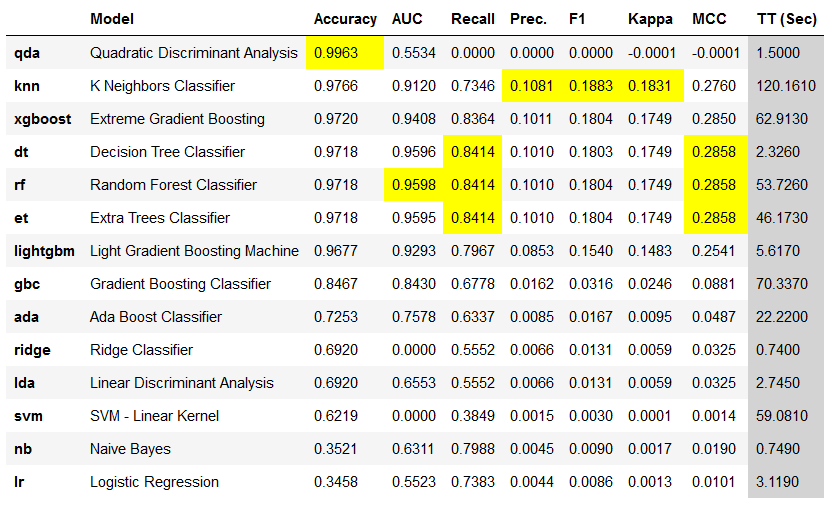
I converted the some of the remaining categorical indicators to binary so the model could utilize the data. Some I simply did a replace others I used a process called One Hot Encoding to generate multiple columns with dummy variables so the model would not interpret the information as a ranking of categories.

* These include gender, education type, job type, and housing type.

In the next milestone, Milestone Three (3) I made an incorrect decision to run a linear regression Random Forest Model, which produce incorrect results. During additional evaluation of the data in Milestone Four (4) I made the decision to do run a classification model. Since the end result of the data is to determine if an applicant is at high risk or low risk the model needs to classify the results. Since our data set has a higher balance of low-risk users it required the data be better balanced. The data was balanced before moving forward.

**Model:**

Below is a chart outline a compare done on different classifier models. I’ll highlight the two strongest models in Figure 10.



**Figure 9: Classifier Model Comparison**

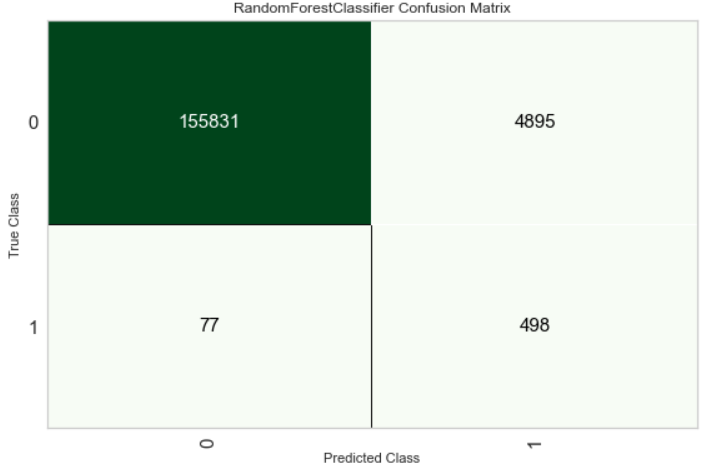
The first model selected was K Nearest Neighbor Classifier (KNN). That one was selected due to accuracy score of 0.9766 or 97.66%. The other scores are not the best but adequate with Recall being 0.7346 and Precession at 0.1081. The next model that was considered it Random Forest Classifier. It has a slightly lower Accuracy percentage of 97.18%, but the other measures like Recall is 84.14% and Precision is 0.1010 which are slightly lower than KNN but the I feel are better than KNN.

I setoff to balance the data since there are less risky individuals compared to low-risk individuals. Using the above two classifier models below are the results of the KNN Classifier model and the Random Forest Classifier model.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **AUC** | **Recall** | **Precision** | **F1** | **Kappa** | **MCC** |
| **KNN** | 0.9766 | 0.9120 | 0.7346 | 0.1081 | 0.1883 | 0.1831 | 0.2760 |
| **Random Forest** | 0.9718 | 0.9598 | 0.8414 | 0.1010 | 0.1804 | 0.1749 | 0.2858 |

**Figure 10: Comparison between KNN and Random Forest**

Taking a look at the two outputs the Random Forest Classifier performs a little better in Recall, AUC and Matthews Correlation Coefficient (MCC). I believe the Random Forest has a better balance compared to KNN Classifier model. The key metrics I am considering are Accuracy, AUC Recall, F1 and MCC. Those are either better in the Random Forest model or just slightly below the KNN model to not make a big difference. Below I have included the Confusion Matrix which show the amount of correctly identified low risk and high-risk individuals found by the Random Forest model.



**High risk**

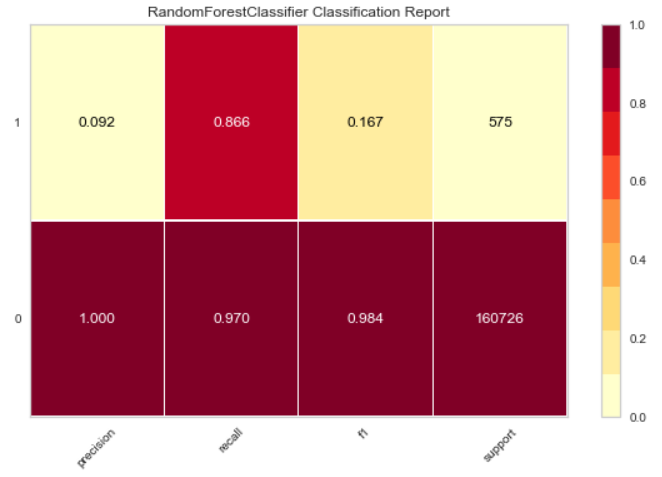
**Low risk**

**High Risk**

**Low Risk**

**Figure 11: Random Forest Confusion Matrix**

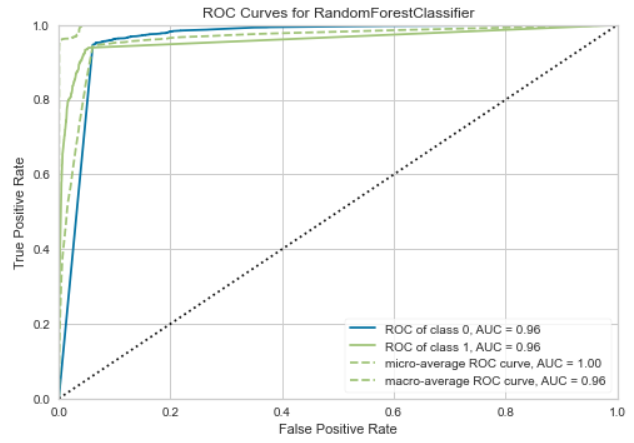
The Confusion Matrix above show the correct prediction of Low Risk of 155,831 compared to 4,895 that were predicted as Low Risk but were High Risk individuals. Similarly, 77 that were predicted as High Risk but where Low Risk and 498 that were predicted as High Risk. I believe there is a definite opportunity to improve the model with better balanced data.

****

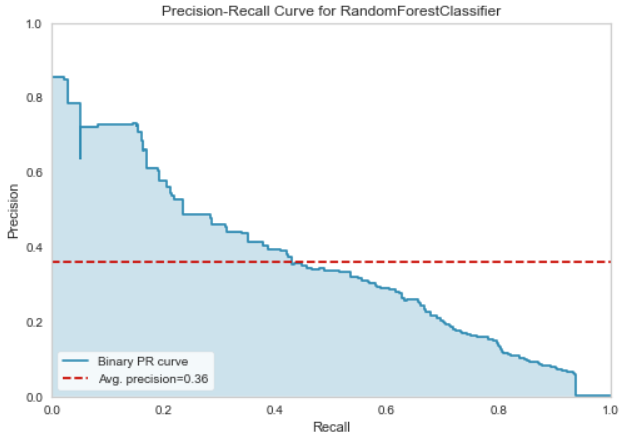
**Low risk**

**High risk**

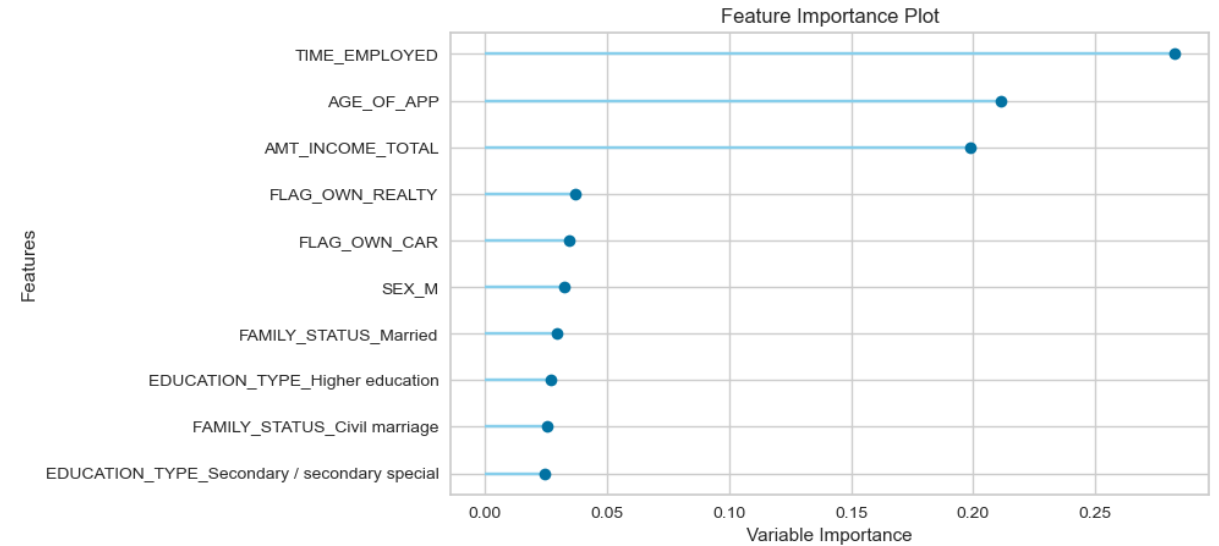
**Figure 12: Random Forest Plot from Classification Report**



**Figure 13: ROC Curve**



**Figure 14: Precession Recall Curve**

****

**Figure 15: Random Forest Feature Importance Plot**

To quickly go through the above figures, I want to point the how each of the classes High Risk (Class 1) compared to Low Risk (Class 0) in the Classification Report look very promising. I believe the imbalance of the classes is causing bias in the data towards identifying more Low Risk individuals and missing Risk High individuals or classifying them as Low Risk incorrectly. The ROC curve makes the data look great. The Feature Importance Plot show which Features have the greatest impact. It appears to be the tried-and-true features like time of employment, age of applicant and amount of income. The other features have an impact, but the top three have a greater impact on risk.

**Conclusion:**

The Random Forest Classifier model does the best job of identifying the Lowest Risk individuals at 96.61%. It also identified 3.03% as Low Risk but were in fact High Risk. The imbalance in the data between the number of High Risk and Low Risk does a make a very noticeable Confusion Matrix. We don’t want to make very risky loans that cannot be paid back as they will impact returns on loans and increasing cost of service existing loans. In other words, reducing the interest made on the financial institutions money. The top three important Feature remains time employed, age of applicant and amount of income I believe there are other variables that might provide more significant weight to getting better results.