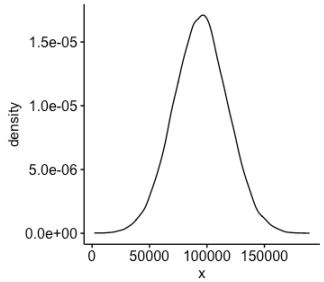
Finding the best model for Wells Fargo Credit Approval
By Daniel Kim

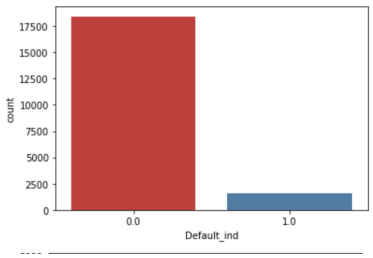
Abstract:
Approving lines of credit is one of the core functions of a bank, and one of the requirements of providing a line of credit is to be able to intelligently assess the risk being taken by loaning money to an individual. Given previous financial history and information of an individual, it is possible to construct a statistical model to predict the likelihood of whether the bank should continue to give credit to the customer (no default), or stop giving credit (default). This paper explores logistic regression and random forest models as options to construct such a model. In the early stages it focuses on how the data should be formatted and cleaned to optimize the effectiveness of the final model. Models are then constructed and analyzed to determine which features of an applicant's financial history are most relevant in determining whether they will default on a loan. Out of the 20 predictors in the dataset, eleven predictors has a relationship with the default, and three/four have the most effect on the chances of getting a default. They are: average card debt, utilization of all credit card accounts, total credit debt, credit age, number of credit inquiries in the last 12 months, credit due to age, and past delinquency. By using these predictors, it was determined that a random forest model is more suitable for this application than a logistic regression model. Although random forest models are less transparent, it is worth the more in-depth analysis of interdependence between variables in an applicant's financial history.

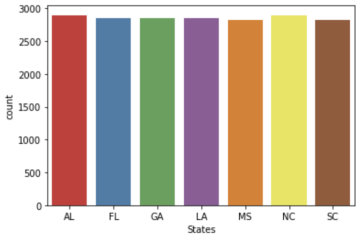
Data Exploratory/Data Processing/ Cleaning Section
When observing the three datasets, the first thing to note is how all of the predictors except the average monthly debt have a normal distribution, observations, or datapoints gathering together near the average amount. The average monthly debt has a right-skewed distribution, but upon closer inspection it seems that the 99999 in its observations is meant to be written as a missing value instead. By replacing the 99999 with NA instead, the distribution changed into a normal distribution again. Three predictors, customer's annual income, percentage of open credit cards with over 50% being used, and average monthly debt, has missing values in it. Having too many missing values could reduce the statistical significance and reduce the quality of our results due to bias. To reduce these negative effects mean imputation, the replacement of missing values with the mean of the affected predictor, was implemented in those three predictors. In some cases using this technique could negatively change our results, but considering how all of the predictors have a normal distribution, the technique is justified.



After processing the data, we moved forward by quickly visualizing each part of the variables. The first thing to note is our response variable the default is wildly unbalanced, 92% has no default, while 8% has a default. This can cause our models to have a huge difficulty predicting the ones with the default because of the lack of information there is compared to the no default one, resulting in our model having an increased chance of having an error finding a default. This problem is later addressed in the paper. The second thing to note is that the predictor States have no relevance to predicting the default account. By

looking at the count and the variance among the states, there's little difference between each other so its relevance seems very little for the model.





Logistic Regression

When running the logistic regression, the stepwise regression method, where the function determines which predictors are the best for the model by picking the ones that best reduce the prediction error, is used. The method revealed that total credit debit, average card debit, credit age, Number of non-mortgage credit-product accounts by the applicants with delinquency in the past 12 months, Number of non-mortgage credit-product accounts by the applicants with delinquency in the past 6 months, Number of credit inquiries in last 12 months, Number of credit cards opened by applicant in last 36 months, ratio of balance divided by credit limit on all credit card accounts (called utilization), if the person already has a bank account, and their annual income. To optimize the model, a decision boundary needs to be made to reduce the amount defaulting clients by the model. Decreasing the threshold is the same as reducing the amount of risk a firm is willing to take. To find the threshold the accuracy and precision is compared to one another, and the movement they cross one another is when the cutoff level has been found. From this test it is determined that 0.15 would be the optimal decision boundary for the model. Once all of these tunes were implemented, the results were interesting. According to the results, the logistic regression achieved a 89% accuracy when it comes to identifying who does not have a default in their credit. Our sensitivity, or how many people were correctly predicted to have no default within our no default prediction, was around 93%. It seems that the model is very good at predicting those that should have no default in their record. However the same cannot be said for those with default. According to the specificity, or the number of people that were correctly predicted to have a default, the model prediction was only 48% correct. There is a possibility that the reason why

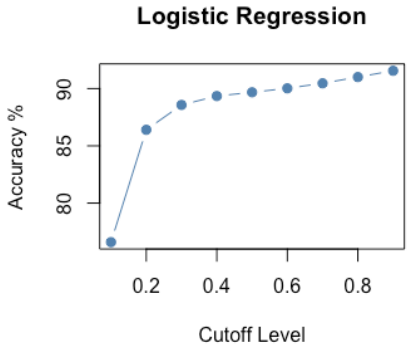
```
               Accuracy : 0.8984
                 95% CI : (0.8897, 0.9066)
    No Information Rate : 0.9198
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3766

 Mcnemar's Test P-Value : 5.403e-05

            Sensitivity : 0.9348
            Specificity : 0.4813
         Pos Pred Value : 0.9538
         Neg Pred Value : 0.3915
             Prevalence : 0.9198
         Detection Rate : 0.8598
   Detection Prevalence : 0.9014
      Balanced Accuracy : 0.7080

       'Positive' Class : 0
```



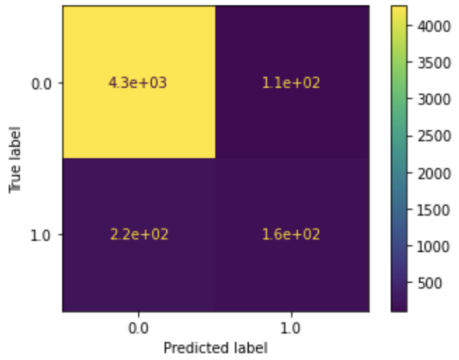

```
Confusion Matrix and Statistics

          reference
data        0    1
     0   4299  208
     1    300  193
```

Random Forest Model

To prepare the model, the categorical columns in the data had to be replaced with dummy variables, or replacing words with 1 and 0, and the numerical columns were transformed with standard scaler to speed the model process. Once the model was made, a confusion matrix was used to the measure the performance of Random Forest Model. According to the test, it seems that the accuracy has improved to 90%, but the precision or quality of positive prediction made by model, reduced to 60%. The recall of the model is 43%, but the F1 score, which is the weighted average of precision and recall, is 50%. This score is more useful indicator to determine if the model is effective or not because of the imbalance class distribution of the customers that have defaults. Models under normal circumstance treat their response variable with equal weights, as if there equal amount of yes and no. Normally accuracy would be a good indicator on whether the model is effective or not. However if the dataset is imbalanced, this could cause the model to misinterpret the data, and cause problems in the future. F1 score can handle the data imbalance by getting the average of the quality of positive predictions (precision and recall), thus a better indicator of the model performance. While the positive rate, or predicting people who do not have default, seems to be high in logistic regression, the random forest model has the edge in finding the negative rate. The negative rate, or the predicting if the customer has a default, is higher with the random forest model. The false negative rate is around 36%, much lower than the 48% rate from the logistic regression. Having a false negative, or falsely giving credit to a customer even though they should not receive it, has a much bigger negative impact on the bank than false positive scenario.

```
0.872712723681341
Training Accuracy is:  0.9841875
Testing Accuracy is:  0.9322105263157895
The f1 score for the model is: 0.5061349693251533
The accuracy for the model is: 0.9322105263157895
The recall score for the model is: 0.4330708661417323
The Precision score for the model is: 0.6088560885608856
```



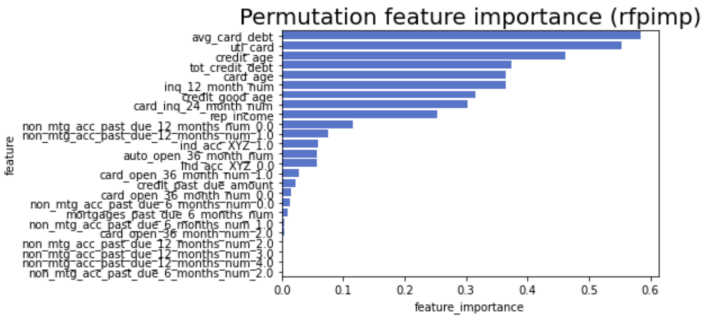

```
FNR.mean()
```

```
0.364218075612092
```

Do existing bank customers have an advantage?

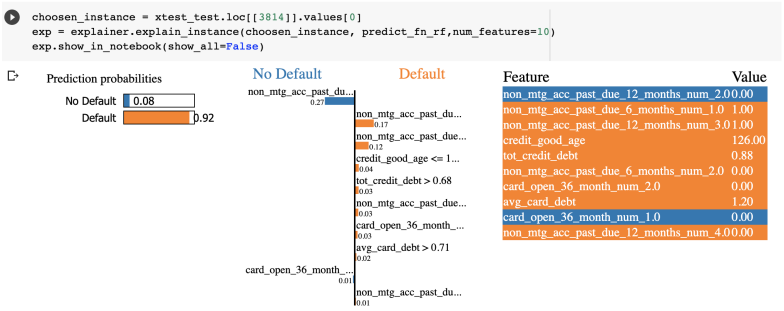
Using the random forest model, the permutation technique is used to determine the importance of each predictor. The permutation works by shuffling each predictor randomly (ex. Average card debt is added and removed) to see if the metrics say accuracy, for example, have changed or not. With this technique, the existing bank customer predictor was able to be observed, and according to the results, it seems that

they do have little to no effect on getting credit. The importance of the existing customer scale is minuscule compared to other predictors, so it seems that will not have any special benefits.



How to explain results to the customer

If the customer get's defaulted and needs an explanation, it'll be best to use the LIME explanation to provide a visual cue on why they received it. A LIME, or Local Interpretable Model Agnostic Explanations is a technique that approximates any black box machine learning model with visuals (like a bar graph) to explain what each variable/columns reasons are for having that prediction. Best of all, the bank employees can specifically find that customer's information by selecting a row, and all of the variables that affect the model's decision are shown. In case the customer asks for advice on avoiding the default, the tool allows the employees to pick a variable, say delinquency for example, and explain how having less of it would decrease the chance that the customer will get a default.



Conclusion:

According to the models, it seems that the most important variables that determine whether a person receives a default were delinquency (non_mtg_acc_past_due_6_months, non_mtg_acc_past_due_12_months), credit good age, total credit debt, average card debt, and the number of times a credit card is open within 36 months. Delinquency was the biggest factor in determining the customer default status, so the bank can advise their customers in advance in order to reduce the amount of defaults a customer will have. Between the two model, logistic regression and random forest, the latter is more effective in predicting defaults. This is due to the model being resistant to outliers and missing data, as well as having more resilience to imbalanced data than the logistic regression. Most importantly, the random forest is relatively easier to tune and use than other machine learning models, so the potential to further improve the model is available. The random forest model is the desired victor in this situation.