Credit Card Attrition Prediction

Ruby Link, Ellie Strande, Jonathan Vergonio

MGSC 310 - Business Analytics

Background - Business Use Case

- **Goal:** Create a model that can **accurately predict which customers are likely to close their credit card accounts**

- This will allow bank to **reduce churn** and **improve retention rates** by proactively reaching out to customers likely to leave

Investigation: Customer Attrition

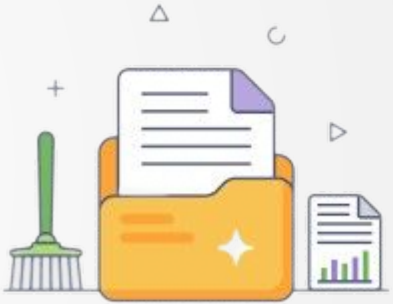
Target Variable: **Attrition_Flag**

Question: **Is it possible to accurately predict the behaviors of customers who will be attrited (have a closed credit card account)?**

Dataset consists of information on **~10,000** customers, including **age, salary, and credit card limit**

Cleaning of Data

- Mutates **'Gender', 'Education Level', 'Marital Status',** and **'Income Category'** to **factor variables**
- **Removes 14 unnecessary variables** from the data frame
- Turns the target variable, **'Attrition Flag'** to a **binary variable**
- Create **'Attrition_Flag_num'** for a **numerical** version



Feature Engineering: Log Transformation of Credit Limit (discussed in future slides)

About our Variables

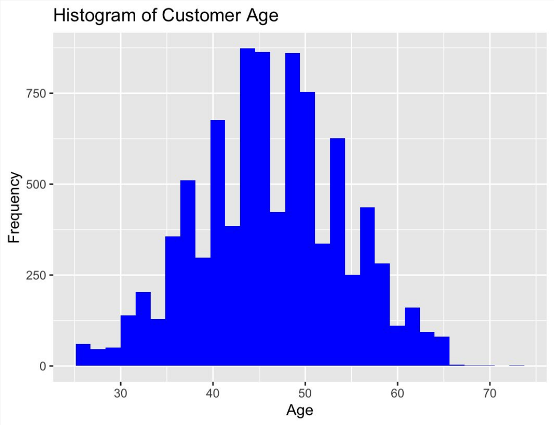
Quantitative Variables

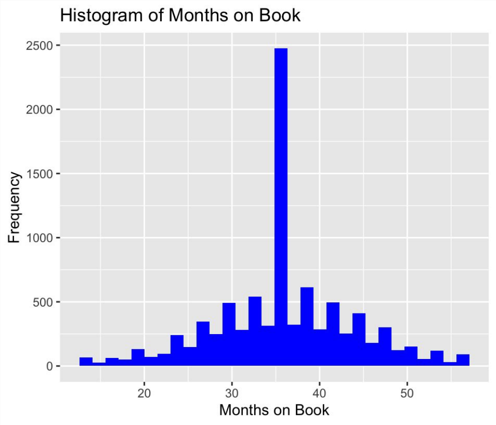
- **Customer Age**
- **Months on Book**
  - Period of relationship with bank
- **Months inactive 12 months**
  - Number of months inactive in last 12 months
- **Credit limit**
- **Total Revolving Balance**
- **Total Transaction Count**
  - Total transaction count in last 12 months

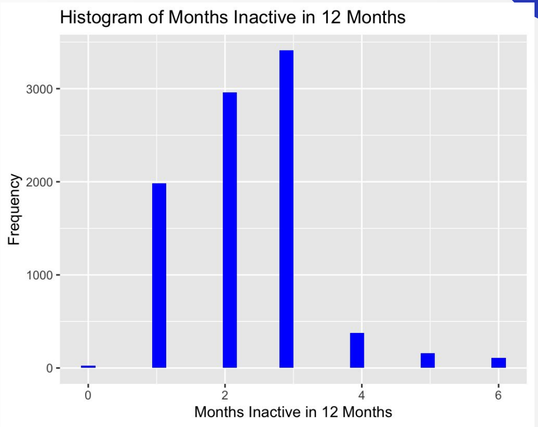
Factor Variables

- **Gender**
- **Income Category**
- **Attrition Flag (target)**
  - Whether the customer is existing (0) or attrited/left (1)

Summary Statistics: Quantitative Variables







Customer Age:

Min: **26**

Mean: **46.33**

Max: **73**

Months on Book:

Min: **13**

Mean: **35.95**

Max: **56**

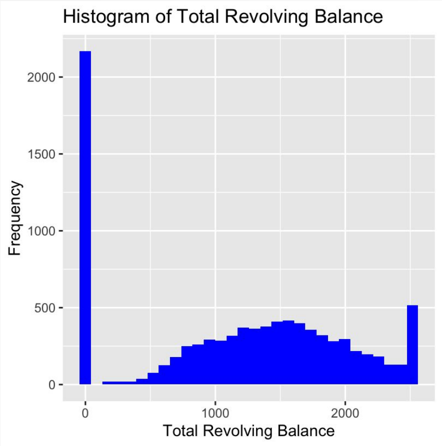
Months Inactive:

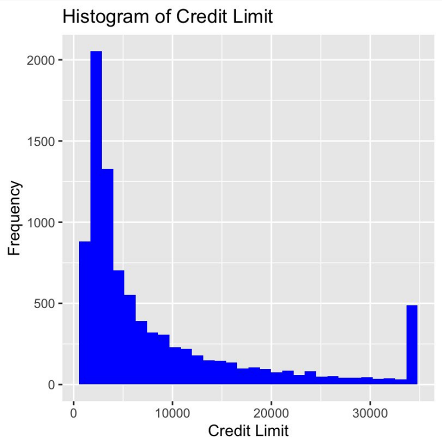
Min: **0**

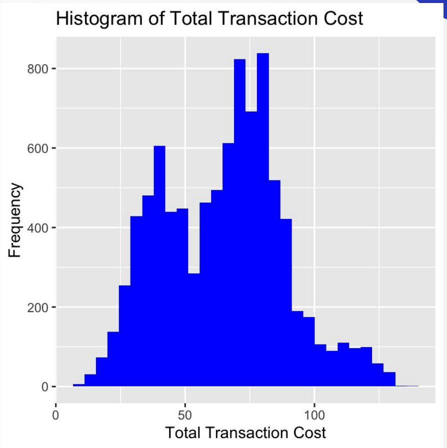
Mean: **2.337**

Max: **6**

Summary Statistics: Qualitative Variables







Total Revolving Balance:

Min: **0**

Mean: **1169**

Max: **2517**

***Credit Limit:**

Min: **1438**

Mean: **8523**

Max: **34516**

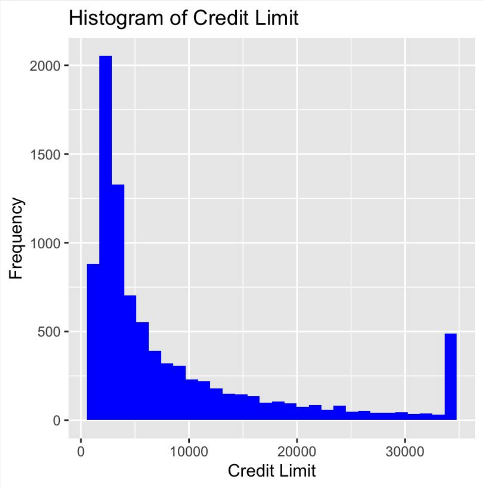
Total Transaction Cost:

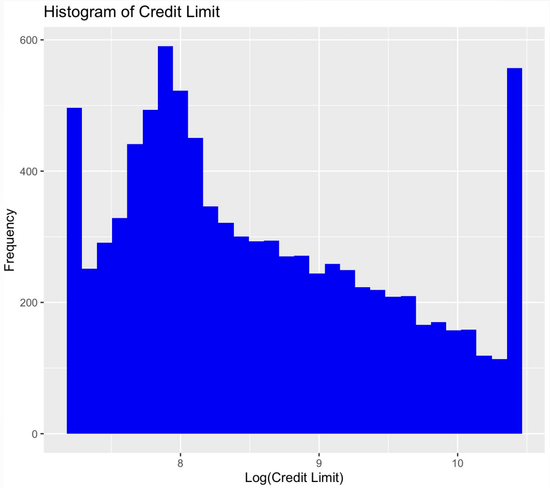
Min: **10**

Mean: **64.69**

Max: **139**

Log Transformation: Credit Limit





Credit Limit:

Min: 1438

Mean: 8523

Max: 34516

Credit Limit:

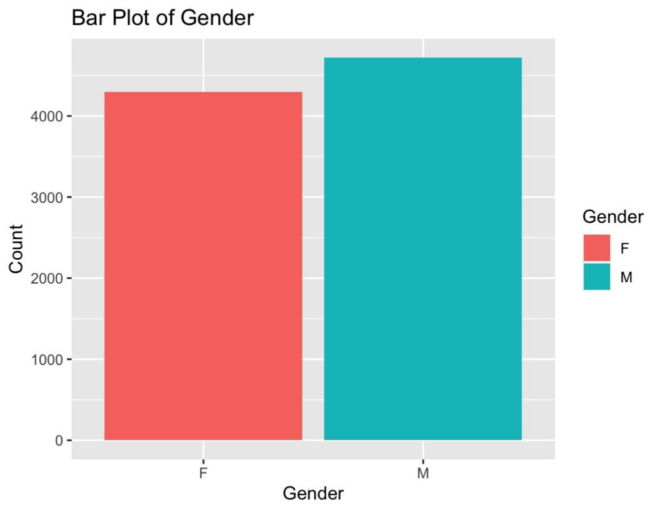
Min: 7.271

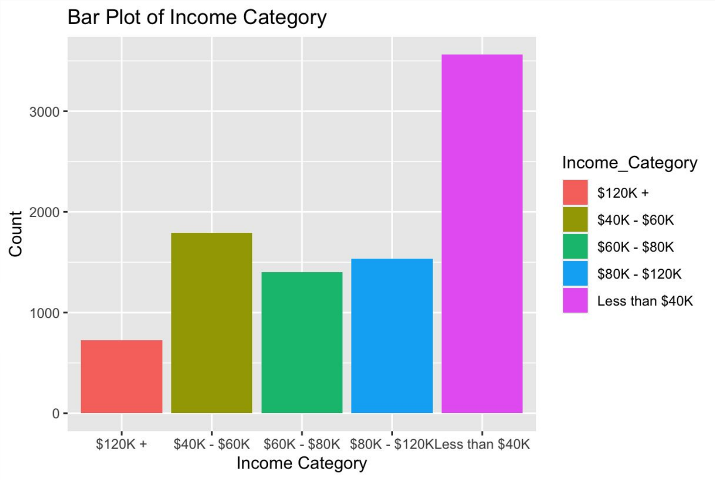
Mean: 8.581

Max: 10.449

Log Transformation is helpful with features that have high variance because it can help to reduce the impact of extreme values or outliers.

Summary Statistics: Qualitative Variables





From the charts above, we can see the **spread** of each qualitative variable.

The **income category with the most observations is less than 40k**, while the **age group with the least observations is 120k+**

There are **more observations of males than females** in the dataset

Outcome #1: Logistic Regression Model

Outcome #1: Logistic Regression

Selected variables & Significance - **Predicting Attrition_Flag**:

- Customer Age
- **Gender*****
- Income Category (**<40•**, **40-60***, **60-80***,80 -120, base: 120+)
- Months on book
- **Months Inactive 12 months*****
- **Credit Limit•**
- **Total Revolving Balance *****
- **Total Transaction Count *****

At alpha = 0.1, there are **8 statistically significant variables**

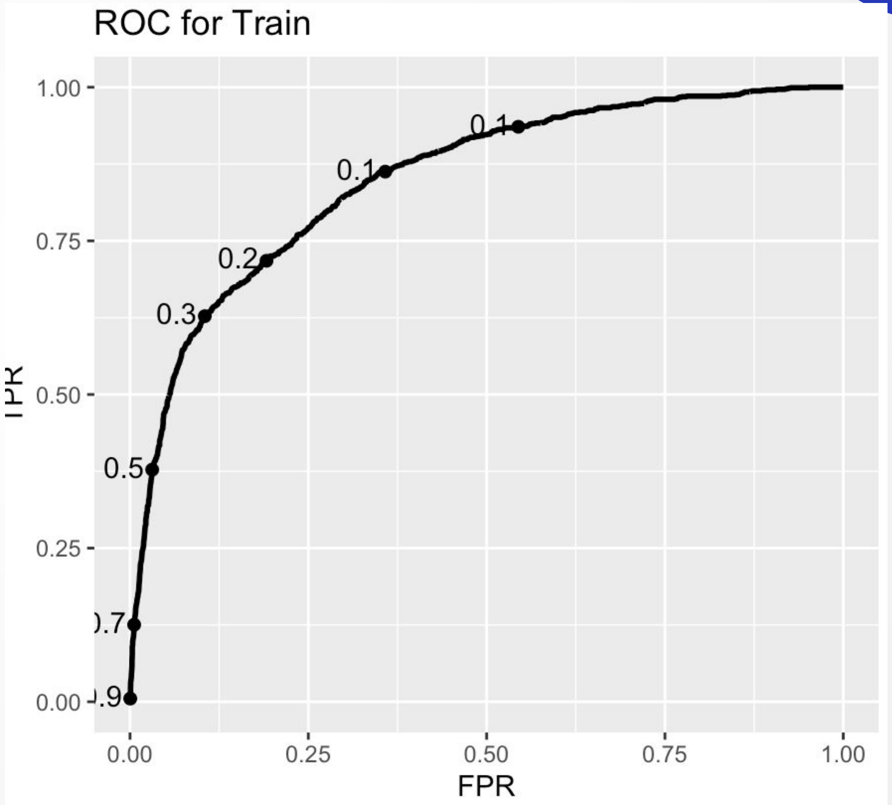
Because the majority of the Income dummies are significant, the Income variable was left in the model

Total Transaction Count is the most significant variable: for every additional transaction, the customer is **5.85% less likely** to be an Attrited Customer

Outcome #1: Choosing Model Params

- **Decided to use the decision cutoff at 0.3**
  - **Maximizes the TPR without increasing the FPR a significant amount**

** The test ROC yielded the same results



AUC = 0.8502706

Logistic Regression Results

Train

	Actual Negative (0)	Actual Positive (1)
Predicted Negative (0)	5413	435
Predicted Negative (1)	634	730

Test

	Actual Negative (0)	Actual Positive (1)
Predicted Negative (0)	1371	96
Predicted Negative (1)	157	179

Logistic Regression Analysis

Train

Accuracy: 0.8517748

Sensitivity: 0.6266094

Specificity: 0.8951546

Test

Accuracy: 0.8596783

Sensitivity: 0.6509091

Specificity: 0.8972513

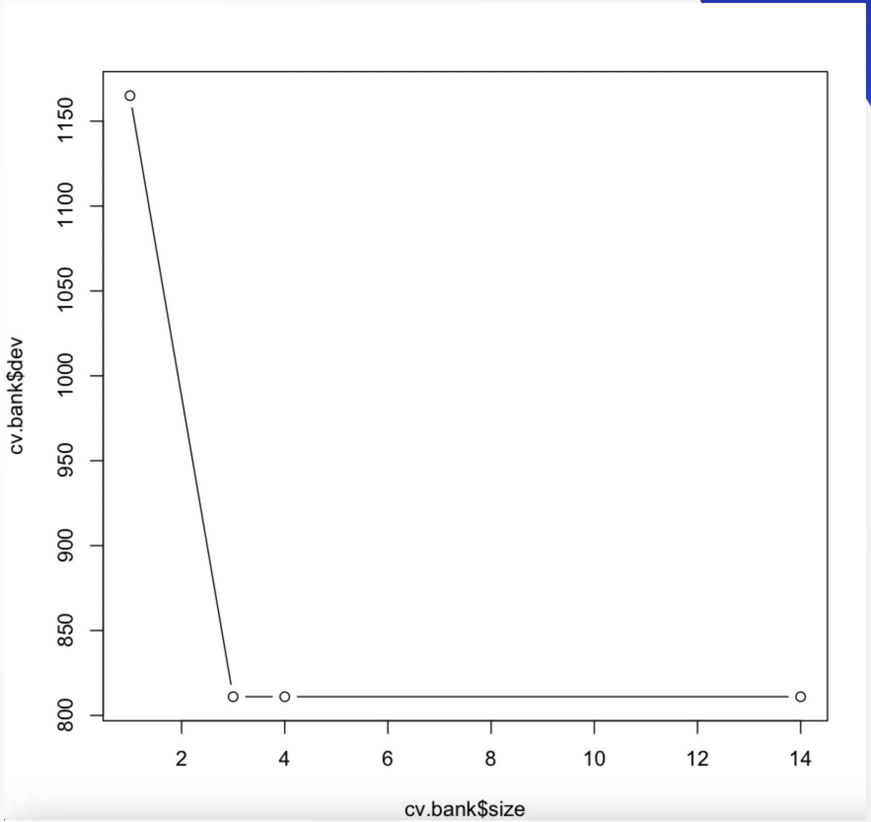
- TPR is low (sensitivity)
    - Severe **class imbalance**
    - Tendency to predict way more negatives
    - Implies **overfitting**
- Potential fixes:
    - Change the **classification threshold** to classify more positives
    - **Downsample** some of the overbearing negative cases
    - **Upsampling** the positives

Outcome #2: Decision Tree Model

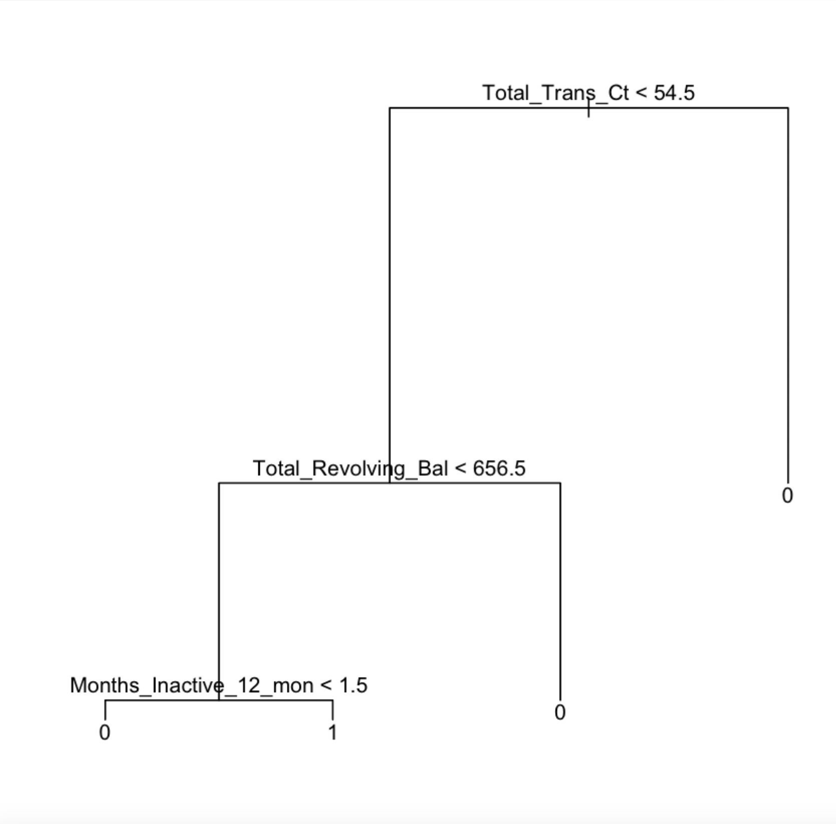
Outcome #2: Decision Tree Parameter Tuning

For the Decision Tree Model, we found that **2-4 minimum leaf nodes** (DT size) was the best.

We observed that 2 leaf nodes was too simple visually, so we went with **4 minimum leaf nodes**.



Outcome #2: Decision Tree



Prediction Results – Pruned Tree

Path leading to Attrition

(Credit Card Close):

1) Total Transaction Counts in the last 12 months is **under ~54.5.**
2) Total Revolving Balance is **under $656.6**
3) Inactive for **greater than 1.5 Months**

Outcome #2: Decision Tree Results

Train Set

Accuracy: 0.8929

Sensitivity: 0.476

Specificity: 0.973

Test Set

Accuracy: 0.9045

Sensitivity: 0.50

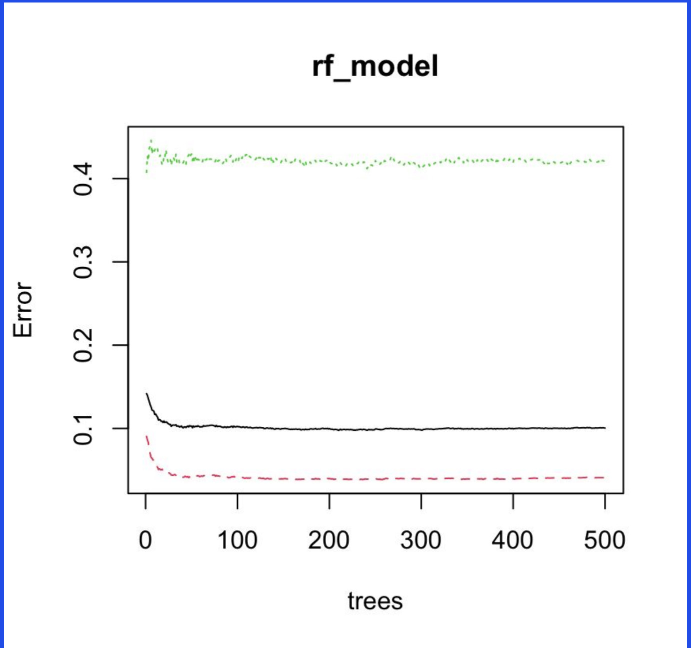
Specificity: 0.977

For the decision tree, the accuracy has increased and the model is great at predicting True Negatives, but accurate predictions on True positives is low.

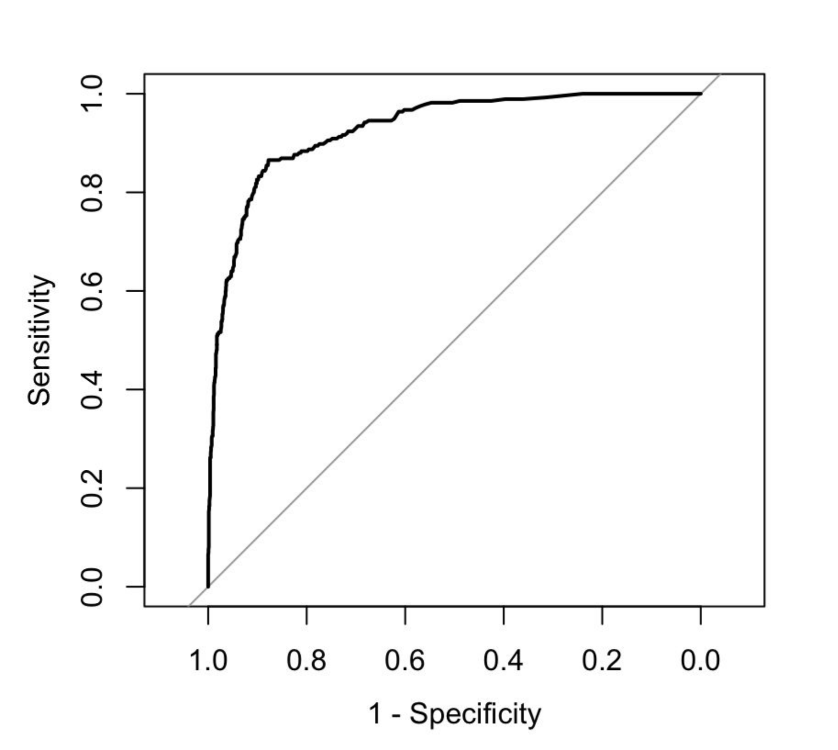
Outcome #3: Random Forest Model

Outcome #3: Random Forest Development

- **80/20** Train-Test Split
- All variables used
- **Ntree = 500**
    - The function will build a forest of 500 decision trees
- **Mtry = 5**
    - The algorithm will randomly select 5 predictor variables at each split



Outcome #3: Random Forests - Classification Results



Train Set

Accuracy: **0.901**

Sensitivity: **0.749**

Specificity: **0.923**

Test Set

Accuracy: **0.908**

Sensitivity: **0.752**

Specificity: **0.9299**

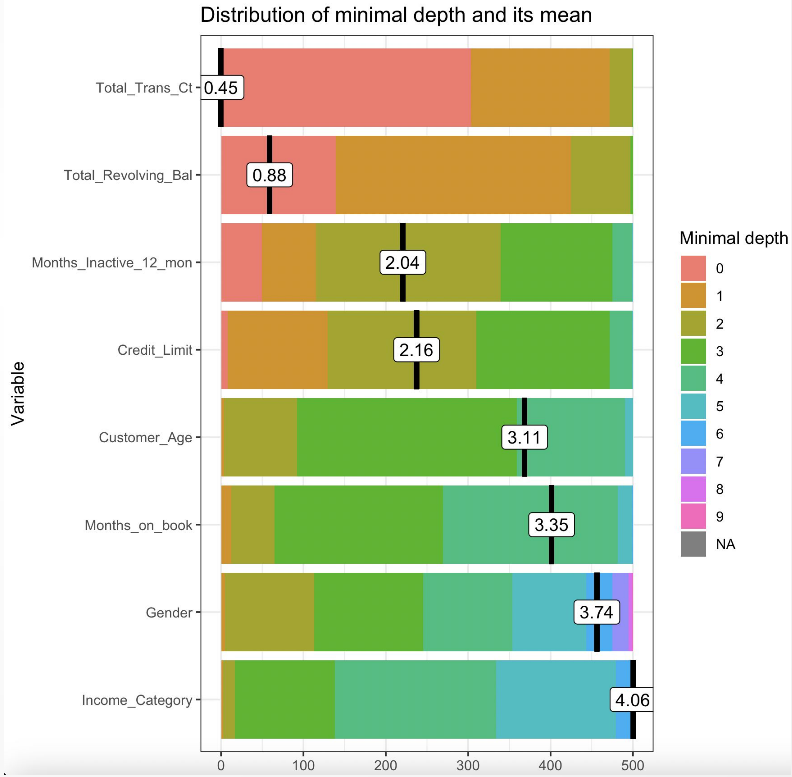
Outcome #3: Random Forests - Classification

Plotting Minimum Depth of Distribution

Most Important Variables:

1) Total Transaction Count
2) Total Revolving Balance
3) Months inactive 12 Months
4) Credit Limit

These variables show up at top of decision tree most of the time



Conclusion & Which Model Should Be Implemented

<u>Best model</u>: Random Forest of Decision Trees

- Generated **highest overall accuracy, specificity, and sensitivity scores** when compared to the logistic regression and decision tree scores
- **Sensitivity score is comparatively the highest**

Test Accuracy for Each Model:

Logistic Regression: 0.8597

Decision Tree: 0.9045

<u>Random Forest: 0.908</u>

Test Sensitivity for Each Model:

Logistic Regression: 0.651

Decision Tree: 0.50

<u>Random Forest: 0.752</u>

Conclusion: Recommendations to Businesses

It is important to look at these variables to reduce customer attrition:

1) If transaction counts are under ~50.
2) If total revolving balance is under $600.
3) Inactivity for greater than 1.5 months.

Solution(s): Offer Credit Card exclusive deals and rewards toward these customer populations.

Thanks!

Any questions?