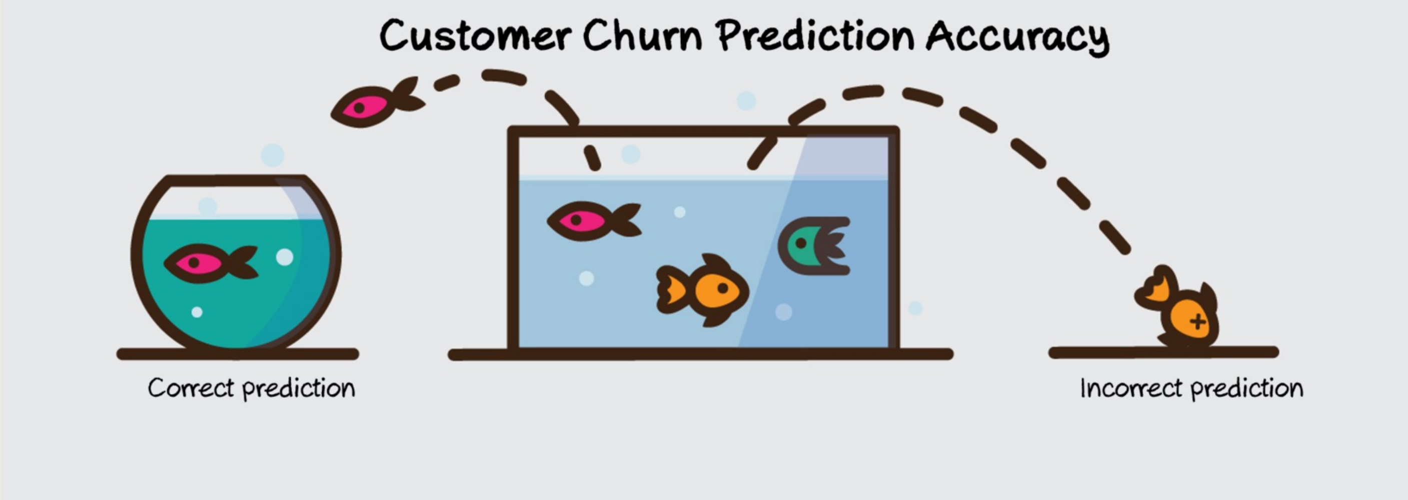
CUSTOMER CHURN – FINANCIAL SERVICES

Melissa Russell, Jonathan Voth, & Nick Wawee

Introduction



- **Problem:** Customer Churn
- **Primary Goal:** Fit model(s) to the data that will successfully predict if a customer will churn or not.
- **Secondary Goal:** Determine which features are most significant in predicting customer churn.

Predictions:

- Neural network, Random Forest, & ADABoost models will successfully predict customer churn.
- Age, marital status, and income will be the most significant features in predicting customer churn.

Dataset: *Credit Card Customers*

- Retrieved from Kaggle.com
- ~10,000 customers of a bank
- Response variable is attrition_flag – if the customer has left or not
- 16.07% of customers have churned
- 18 features

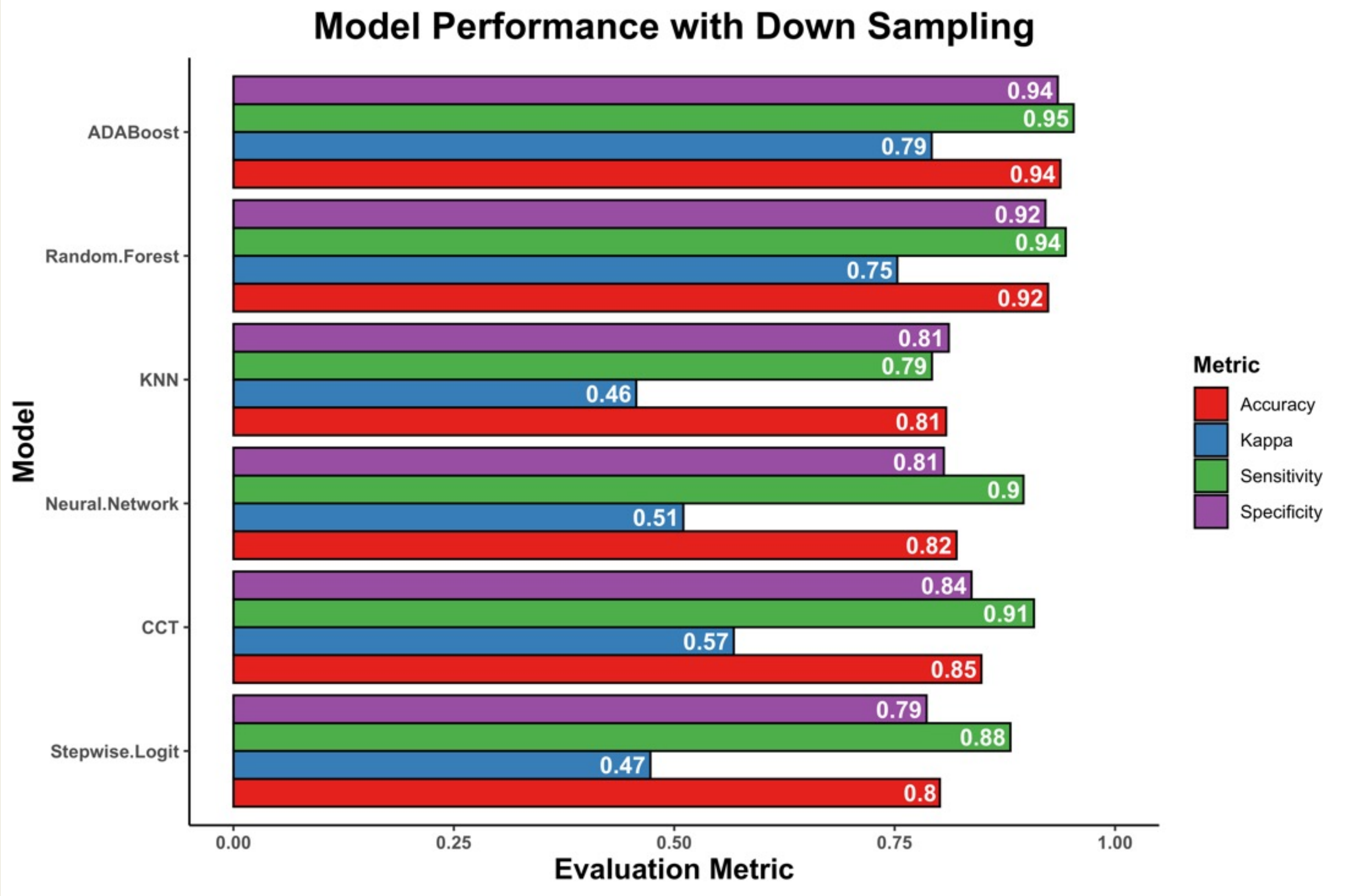
Name	Data Type	Type	Description
Attrition_Flag	Categorical (2, 16% Flagged)	Response	Did the customer leave?
Customer_Age	Integer	Feature	Age in years
Gender	Categorical (2)	Feature	Male or Female
Dependent_Count	Integer	Feature	How many dependents the customer has
Education_Level	Categorical (4)	Feature	How educated the customer is
Marital_Status	Categorical (3)	Feature	Marital Status
Income_Category	Categorical (4)	Feature	Income Bracket
Card_Category	Categorical (4)	Feature	What kind of card the customer posses
Months_on_book	Integer	Feature	How long the customer has had a relationship with the bank
Total_Relationship_Count	Integer	Feature	Total number of products a customer has
Months_Inactive_12_mon	Integer	Feature	Number of inactive months within the last 12
Contacts_Count_12_mon	Integer	Feature	How frequent the customer contacted the bank in the last 12 months
Credit_Limit	Integer	Feature	Credit limit in $
Total_Revolving_Bal	Integer	Feature	Total balance
Avg_Open_To_Buy	Integer	Feature	Limit - Balance
Total_Amt_Chng_Q4_Q1	Float	Feature	Change in transaction amount in Q4/Q1
Total_Trans_Amt	Integer	Feature	Total transaction amount
Total_Ct_Chng_Q4_Q1	Float	Feature	Change in transaction count in Q4/Q1
Avg_Utilization_Ratio	Float	Feature	Balance / Limit

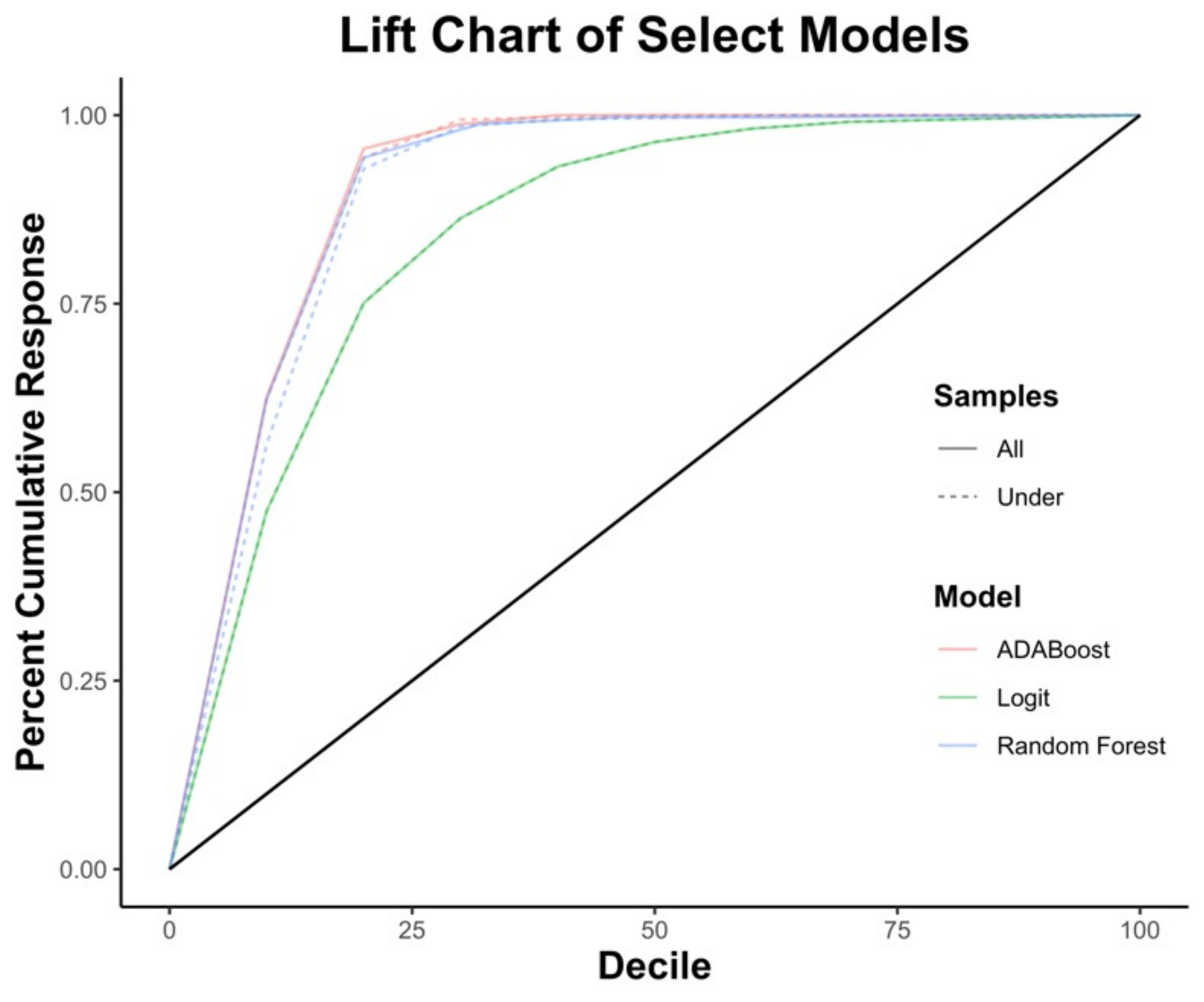
Methods

- Preprocess and clean the data
  - *Remove unknown values and card category feature*
- Response variable contained 15.7% customer attrition
- Partition data 70:30
- Logistic regression, classification tree, nearest neighbor, and neural network
- Ensemble methods (Random Forest and ADABoost)
- Under-sampling was used to address the uneven distribution in the response variable
- Evaluation metrics: accuracy, kappa, sensitivity, and specificity
  - *Focus on sensitivity as customer attrition was considered “positive”*

Results – Performance Metrics

- Under sampled models were chosen because of they had overall a better performance
- ADABoost and Random Forest performed the best
  - Sensitivity and Specificity in the mid-90th percentile
- Stepwise logistic regression used to interpret customer attributes relationship to attrition



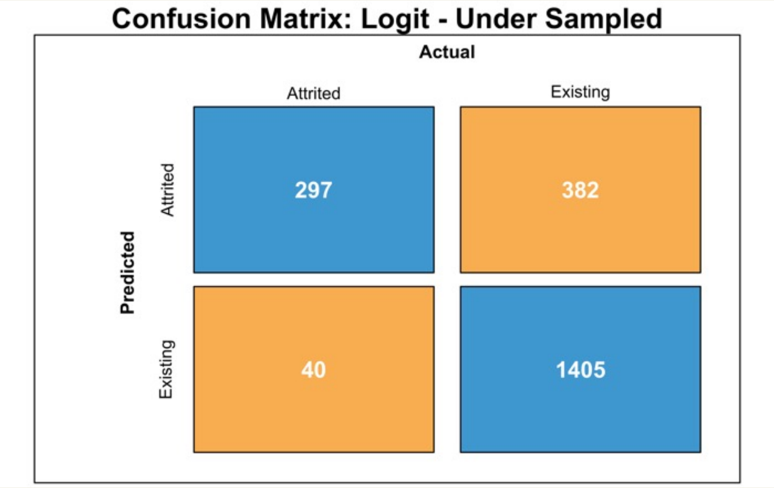


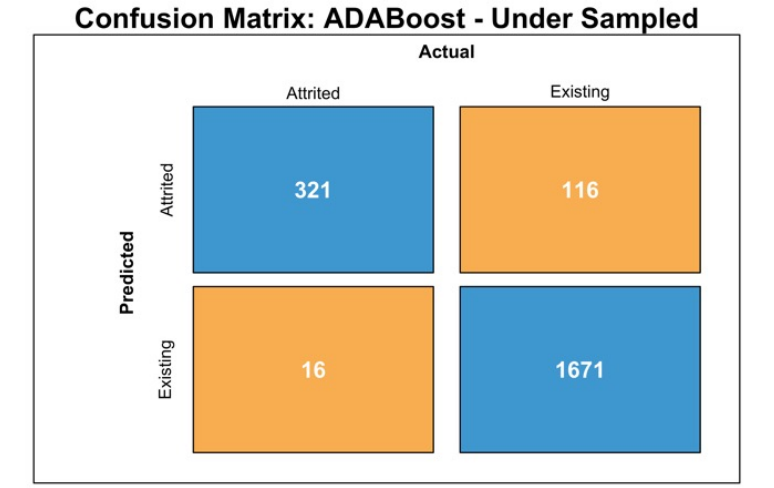
Lift Chart Performance

- Down sampling had little impact on performance
- Ensemble tree models clearly outperforms the logit model
  - *Tree model: Top 25%*
  - *Logit: Top 60%*
  - *Ensembles would aid in decreasing marketing costs*

Confusion Matrices

- Low number of FN in comparison to TP ➔ high sensitivity
- Low number of FP in comparison to TN ➔ high specificity
- Logit model has twice the number of false negatives and positives

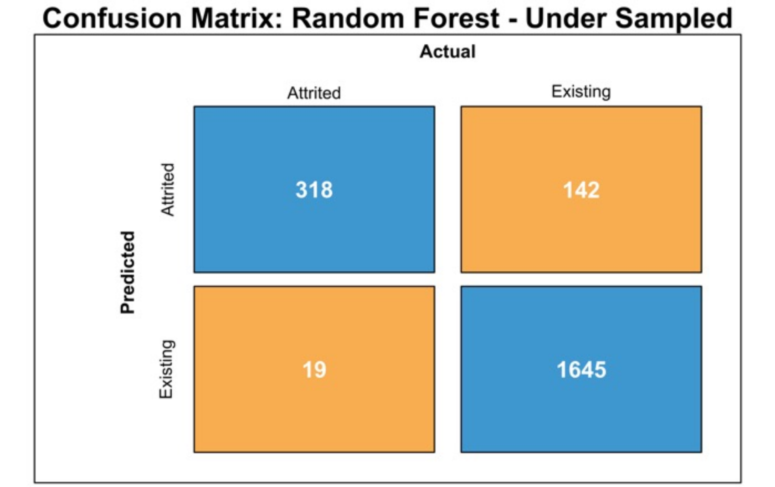




RF Optimal Parameters:

Number of Predictors = 8

Number of Trees = 50

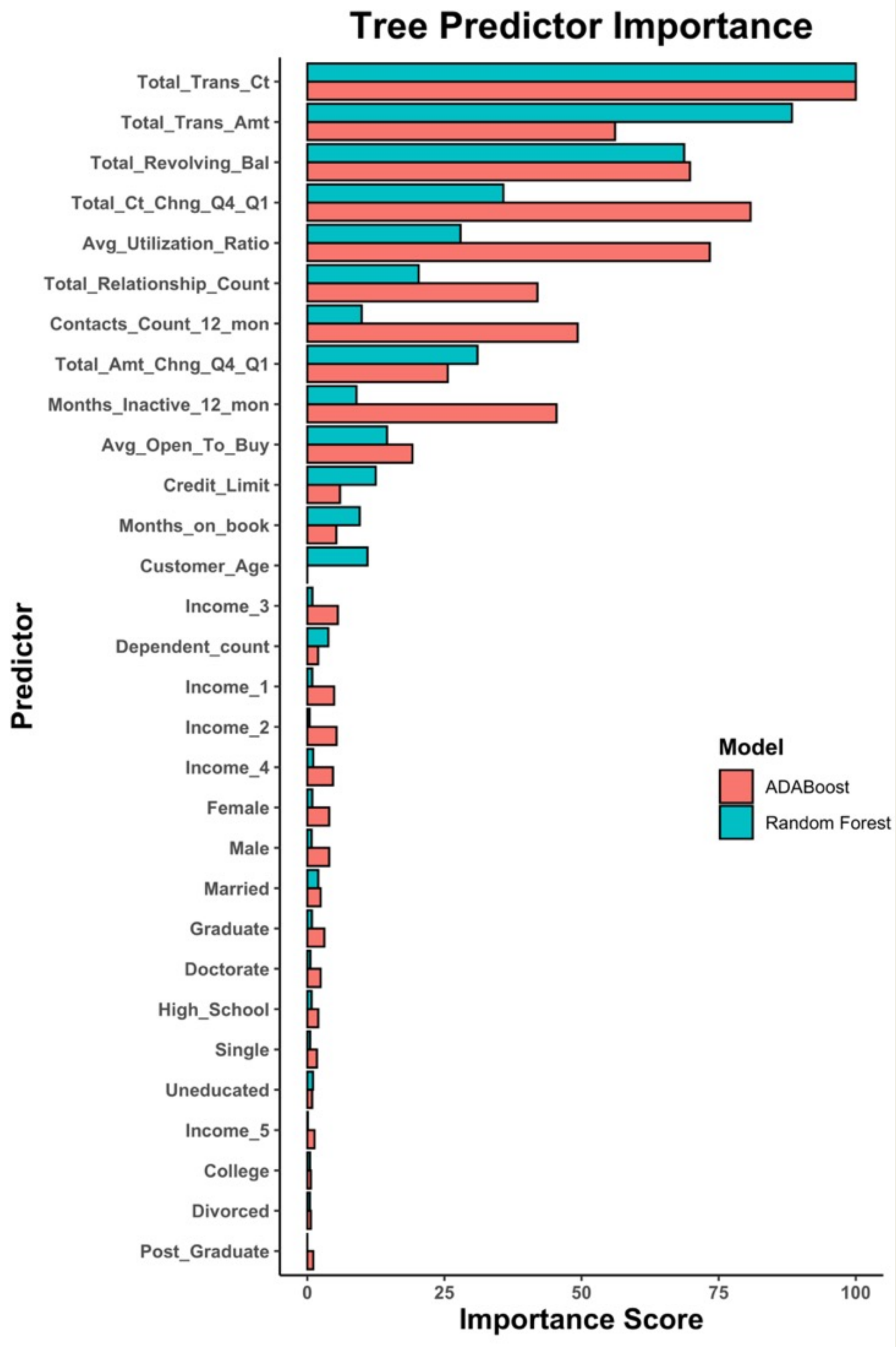


ADABoost Optimal Parameters:

Iterations = 150

Maximum Tree Depth = 6

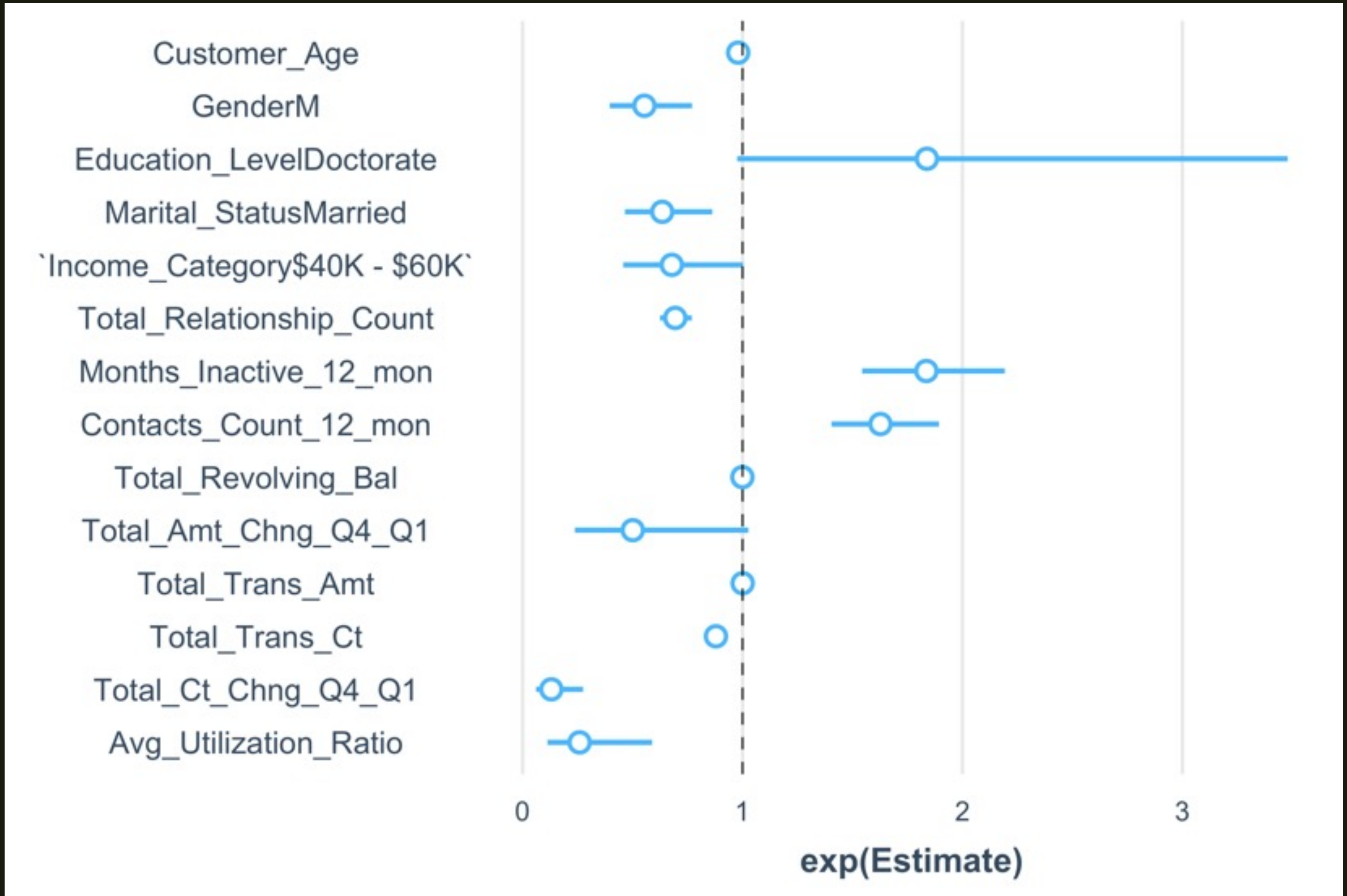
Learning Rate = 0.15



Ensemble Variable Importance

- Total transaction Number and Amount, Revolving Balance ➔ Most Important
- Several variables are more important in ADABoost compared to RF
- Age is somewhat important in RF but not in ADABoost

LOGIT VARIABLE IMPACT



Attrited Customer = 1, Existing Customer = 0

- Inactivity within the past twelve months and total transaction amount increase odds ratio (OR) of attrition to existing
- Characteristics that decrease OR
  - Transaction count ratio Q4/Q1
  - Balance / Limit
  - Male Gender
  - Married Customer
  - Total # Products
  - Total # of Transactions
- Four predictors that had questionable significance
  - **Total Amount Change Q1 – Q4**
  - **Doctorate Education Level**
  - **$40K - $60K Income category**
  - **Customer Age**

Discussion

Predictions:

- Neural network, Random Forest, & ADABoost models will successfully predict customer churn.
- Age, marital status, and income will be the most significant features in predicting customer churn.

Results:

- Neural network, Random Forest, & ADABoost models all successfully predict customer churn.
  - NN Sensitivity: 0.90
  - RF Sensitivity: 0.94
  - ADA Sensitivity: 0.95
- The number of transactions and amount of transactions are the most significant features in predicting customer churn.

Limitations & Future Work

- Limitations:
  - *Lack of context, did not know about data collection process and which bank it came from*
  - *More features– some additional customer attributes may be more telling such as automatic debit transfers*
- Future Work:
  - *Clustering analysis of attrited customers to define their characteristics*
  - *Dimensionality reduction prior to model fit*
  - *Usage of advanced down sampling (ENN) or oversampling techniques (SMOTE)*
  - *Generate profit models by simulating marketing costs and profit per customer*

THANK YOU! QUESTIONS?

Code Publicly Available:

https://gitlab.com/nickwawee/customer_churn

https://github.com/nickwawee/customer_churn