THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

T9 Outliers – Final Project

# Telecom Churn Prediction


- Pranav Chandaliya
- Pooja Chandrashekara
- Vaishnavi Nagarajaiah
- Sunisha Harish
- Kunal Inglunkar


# Problem Statement

A telecom company's postpaid business of voice-only plans is struggling to maintain its strong foothold in the local market due to:

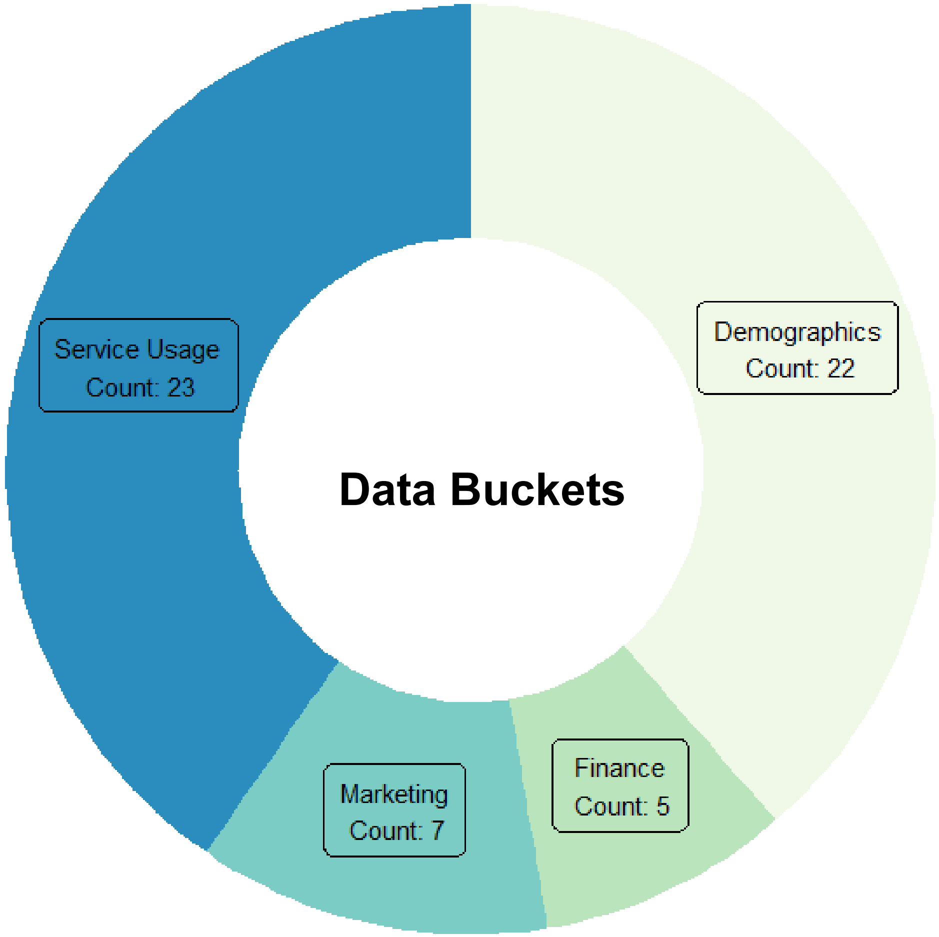
- High churn rate amongst customers leading to a revenue decline of ~500k USD every month.
- The decline in overall customer base (high churn rate combined with low acquisition rate), leading to a decline in total market share.

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

# About the Dataset

Our data is majorly classified into the following categories:

| Data Types | Examples |
|---|---|
| Demographics | occupation, age |
| Service Usage | monthly call minutes, roaming calls |
| Finance | credit rating, monthly revenue |
| Marketing | retention calls, referrals made by customer |




Rows : 51,047 Columns : 58

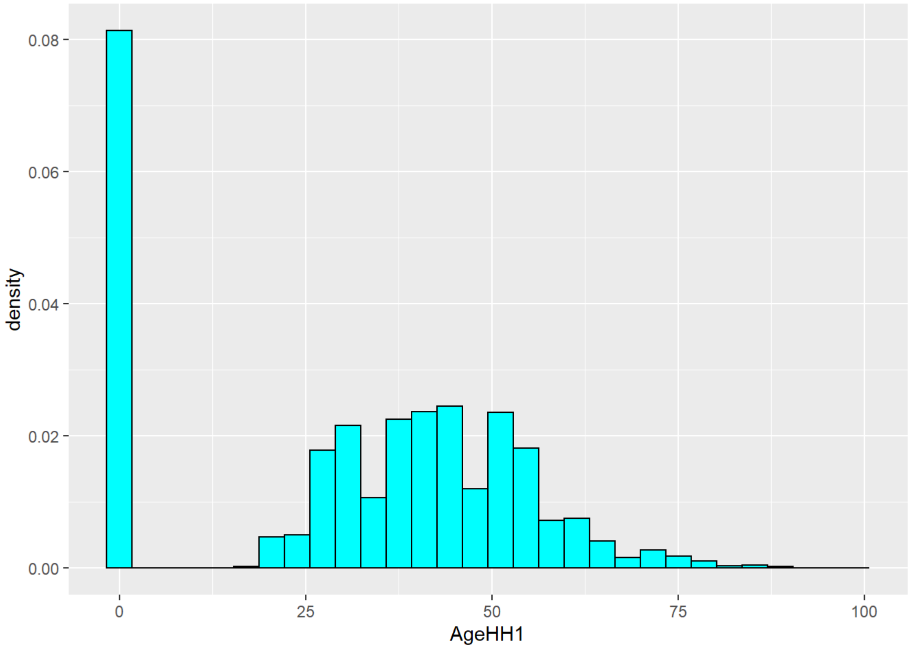
# Key takeaways from EDA

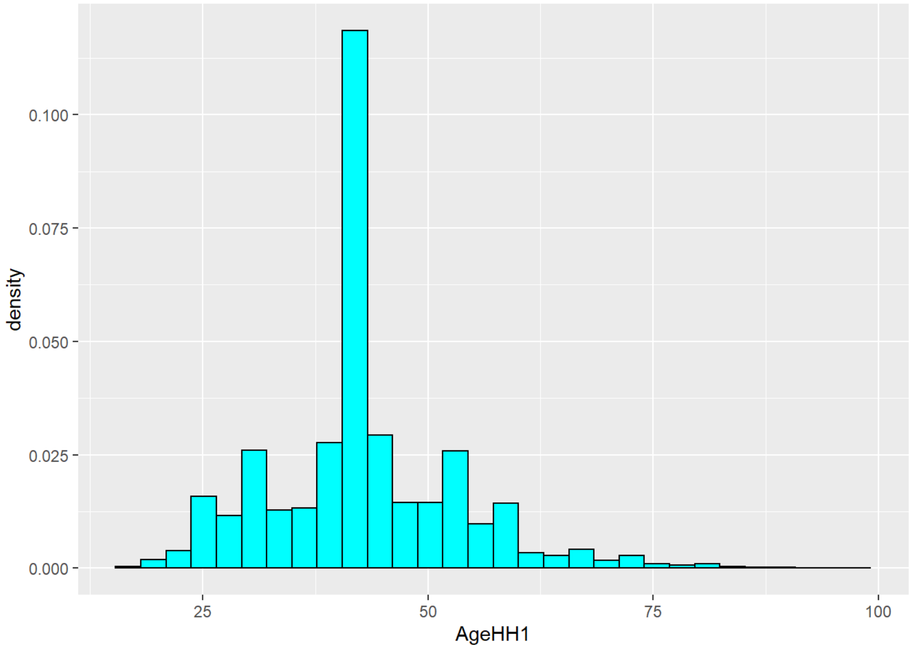
- We observed null values in few columns but it was less than 2%.
- Few customers have incorrect age as 0 which could be interpreted as null values. Hence imputation is necessary.
- There are too many features (57), hence feature selection is required.
- Few customers had < = 0 monthly revenue . These can be considered as outliers and it also shows that few customers are inactive.
- Target class is imbalanced.

# Data Preprocessing

# How to deal with incorrect data entries and handle outliers?

- Imputed median for age column as we observed data was skewed
- Removed inactive customers as well (Monthly revenue and minutes  <= 0)




THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

# What are the important features that are impacting customer churn?

- Our dataset contains a lot of variables, hence it was necessary to select only the important features.
- For this purpose, we have done Chi Square tests on the categorical variables and ANOVA test on the numerical variables.
- We have selected the features based on the p value obtained.

# Feature selection for Categorical Features

| Variables | P-value ( <0.05) | Chi-square (X-squared) |
|---|---|---|
| IncomeGroup | 0.00020 | 32 |
| ChildrenInHH | 0.03163 | 5 |
| Homeownership | 0.00303 | 9 |
| PrizmCode | 0.00026 | 19 |

# Feature selection for Numerical Features

| Variables | P-value ( <0.05) |
|---|---|
| MonthlyMinutes | 0 |
| MonthlyRevenue | 0.0068 |
| TotalRecurringCharge | 0 |
| DirectorAssistedCalls | 0 |
| OverageMinutes | 2e-04 |
| RoamingCalls | 0.0141 |
| DroppedCalls | 5e-04 |
| UnansweredCalls | 0 |
| CustomerCareCalls | 0 |
| MonthsInService | 0 |
| UniqueSubs | 0 |
| ActiveSubs | 5e-04 |
| CurrentEquipmentDays | 0 |
| AgeHH1 | 0 |
| AgeHH2 | 0 |

# How to handle an imbalanced dataset for customer churn?

We have tried the following sampling techniques to balance our data before modelling:

- Under sampling
- SMOTE (Over Sampling)

We have tested our models on both the sampling techniques.

# Which model evaluation metrics should be considered to choose the best fit model?

**Actual**

**Predicted**

| | Churned | Not Churned |
|---|---|---|
| Churned | Well Done (TP) | Not that critical (FP) |
| Not Churned | Danger Zone (FN) | Well Done (TN) |

**Danger Zone** <- Customers who are going to churn but are not detected by the model (Recall)

**Not that Critical** <- Customers who aren't going to churn but model says churn

# Logistic Regression Model

- Logistic regression is one of the commonly used models for classification.
- For our model, we used all the features that we selected and we built a binomial logistic regression model.
- We built the model on pre-processed under sampled and over sampled data.

# Logit Model Evaluation

### Confusion Matrix for SMOTE

| | Prediction | | |
|---|---|---|---|
| Actual | | Not Churned | Churned |
| | Not Churned | 7,074(TN) | 106(FP) |
| | Churned | 5,516(FN) | 120(TP) |

Accuracy: 56.13%    Recall: 53.09%

### Confusion Matrix for Under Sampling

| | Prediction | | |
|---|---|---|---|
| Actual | | Not Churned | Churned |
| | Not Churned | 1,895(TN) | 1,047(FP) |
| | Churned | 1,441(FN) | 1,377(TP) |

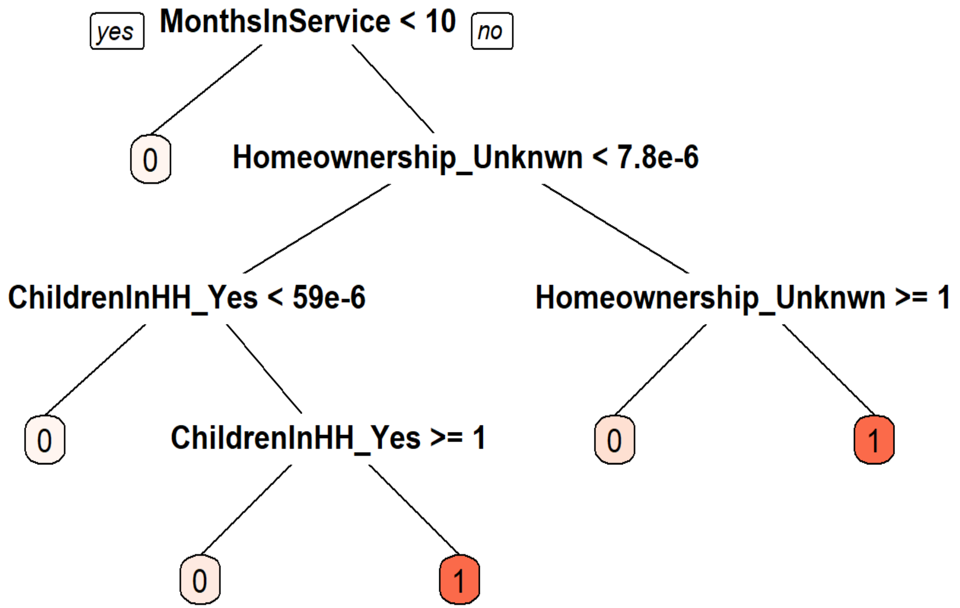
Accuracy: 56.8%   Recall: 56.8%

# Decision Tree Model

- A decision tree is a **decision support tool** that uses a tree-like model to make decisions and their possible consequences.
- We have taken the pre-processed under sampled data to fit in to the decision tree model.
- In our model we are trying to find the variables that is affecting the churn.

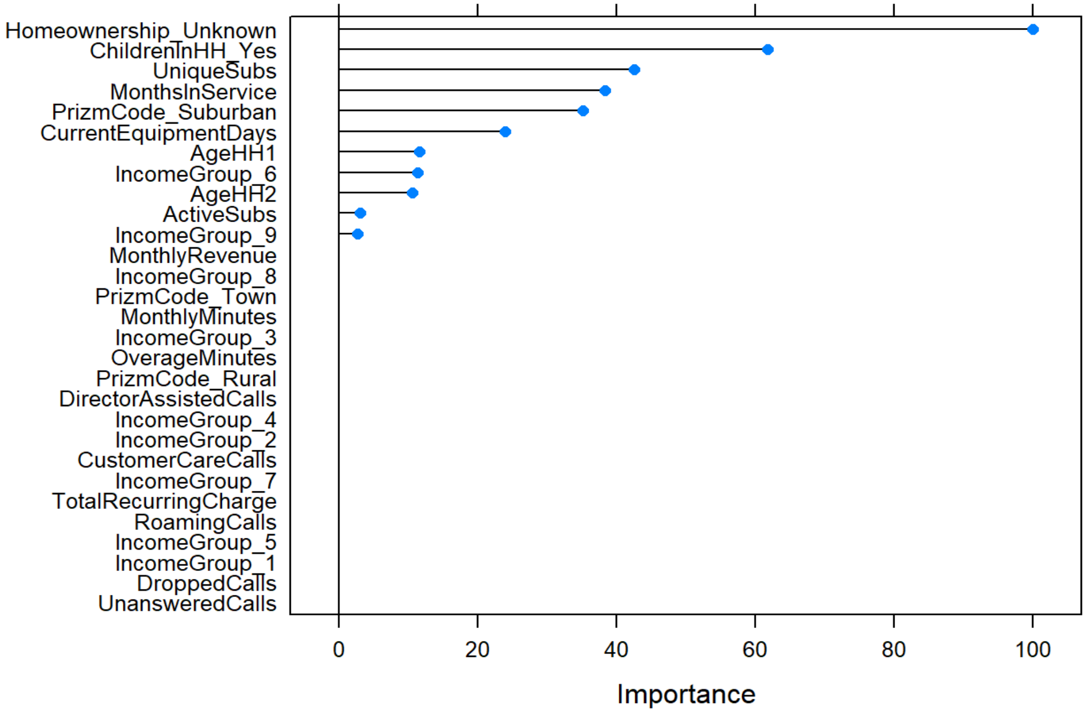
# Decision Tree Model Evaluation

- Figure shows the final decision tree.
- If the customer has less than 10 months in service then the customer is unlikely to be churned.
- If primary holders have no children, then the customer is unlikely to be churned.
- Customers unknown of owning a home are likely to be not churned.



# Variable Importance in Decision Tree

- Figure shows the variable importance in the decision tree model.
- Customers unknown of owning a home is given high importance while plotting the decision tree followed by chances of children being primary holder, unique subscribers, months in service of the customers.




THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

# Confusion Matrix

- Confusion matrix shows the performance of the classification algorithm.
- There are 7180 True negatives, 4557 false positives , 0 false positives and 1079 true positives.
- Model gave the accuracy of 64.4%.

| | | Prediction | |
|---|---|---|---|
| | | Not Churned | Churned |
| Actual | Not Churned | 7,180(TN) | 4,557(FP) |
| | Churned | 0(FN) | 1,079(TP) |

# Naïve Bayes Model

- The Naive Bayesian classifier is based on Bayes' theorem with independent assumptions between predictors.
- Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors.
- Naïve Bayesian model is easy to build and fast to predict class of test data set, requires less training and performs well in case of categorical input variables.
- We have taken pre-processed under sampled and smote data for our model building.

# Naïve Bayes Model Evaluation

**Confusion Matrix for SMOTE**

| Prediction | | | |
|---|---|---|---|
| | | Not Churned | Churned |
| Actual | Not Churned | 754(TN) | 330(FP) |
| | Churned | 6426(FN) | 5306(TP) |

Accuracy: 47.3%    Recall: 30.05%

**Confusion Matrix for Under Sampling**

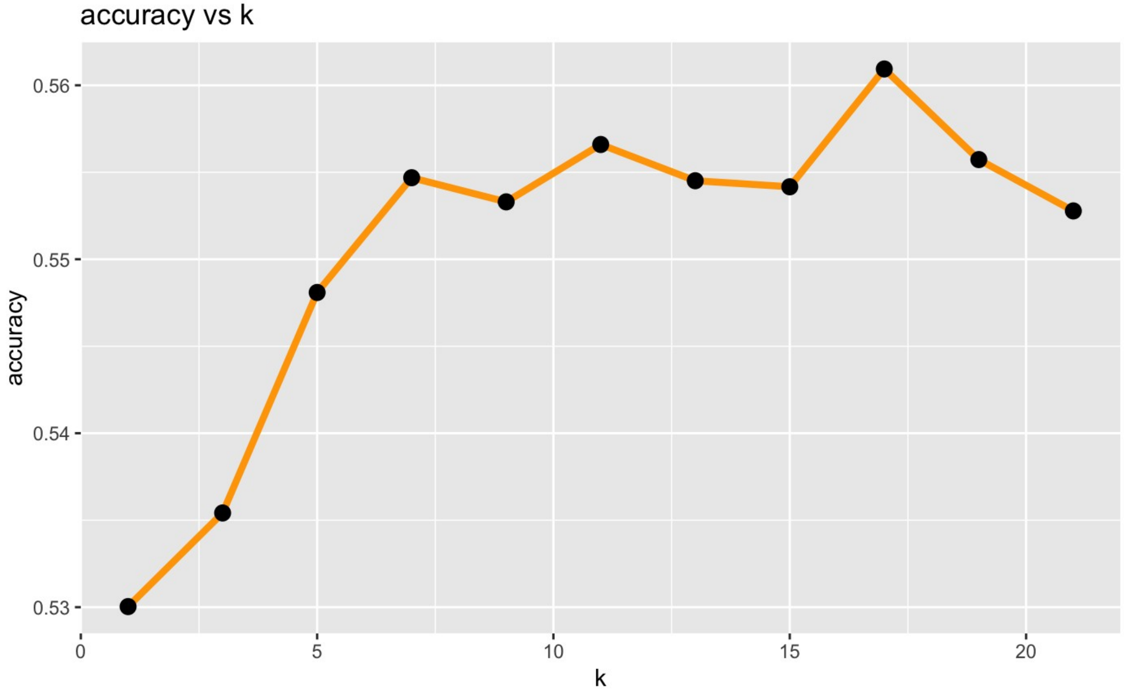
| Prediction | | | |
|---|---|---|---|
| | | Not Churned | Churned |
| Actual | Not Churned | 1528(TN) | 1089(FP) |
| | Churned | 6426(FN) | 1729(TP) |

Accuracy: 56.5%   Recall: 51.9%

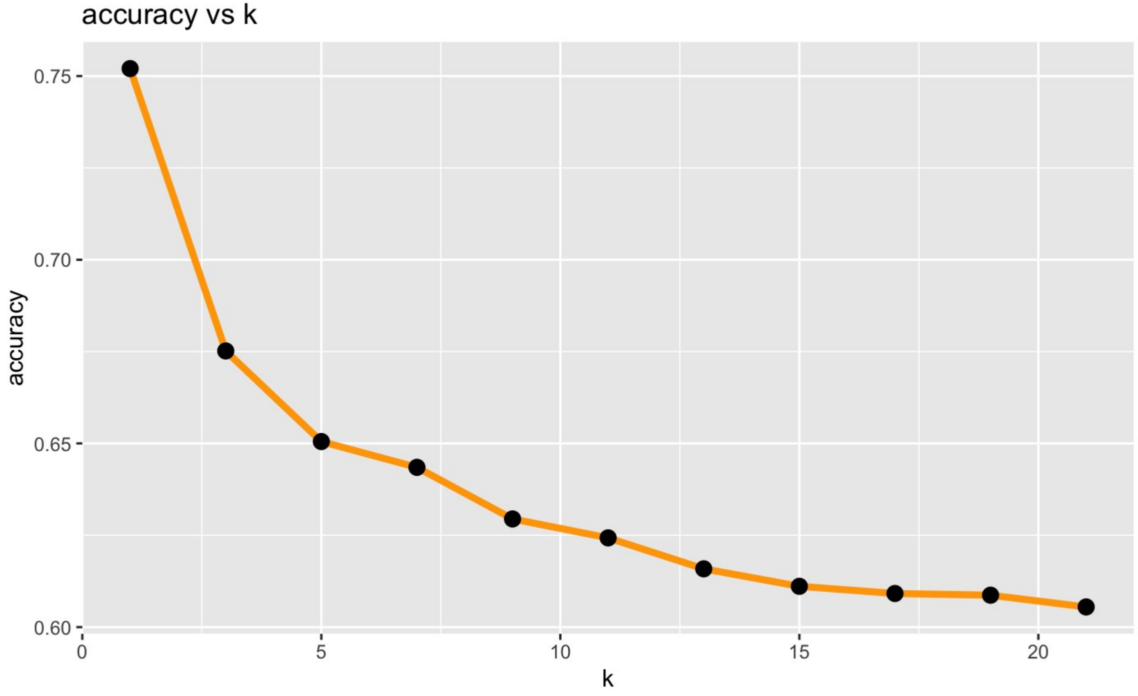
# KNN Model

- Churn variable will be used as the dependent variable. All other selected features will be used as predictor variables.
- Categorical data fields have been converted from Yes or No to either 0 or 1.
- We found out optimal K value by finding accuracy for each K value and selected the best one.

# Accuracy vs k value plot




**For Under Sampling**




**For SMOTE**

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

# KNN Model Evaluation

**SMOTE**

| Prediction | | | |
|---|---|---|---|
| Actual | | Not Churned | Churned |
| | Not Churned | 4427(TN) | 1,393(FP) |
| | Churned | 2,753(FN) | 4,243(TN) |

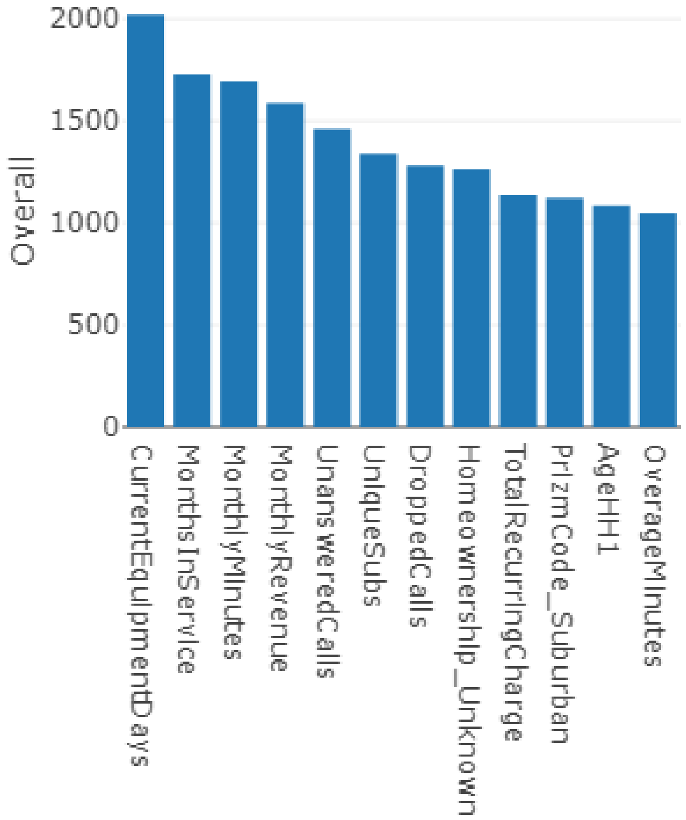
Accuracy: 67.6% Recall: 61.7%

**Under Sampling**

| Prediction | | | |
|---|---|---|---|
| Actual | | Not Churned | Churned |
| | Not Churned | 4818(TN) | 1653(FP) |
| | Churned | 2362(FN) | 1165(TN) |

Accuracy: 56.1% Recall: 59.7%

# Random Forest Model




- Figure shows the Feature importance for the Random Forest model.
- Current Equipment , months in service, Monthly minutes and Unanswered calls are important features in predicting churn.

.

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

# Random Forest Model Evaluation

**SMOTE**

| Prediction | | | |
|---|---|---|---|
| Actual | | Not Churned | Churned |
| | Not Churned | 5,026(TN) | 1,328(FP) |
| | Churned | 2,154(FN) | 4,254(TP) |

Accuracy: 72.7% Recall: 66.3%

**Under Sampling**

| Prediction | | | |
|---|---|---|---|
| Actual | | Not Churned | Churned |
| | Not Churned | 1,786(TN) | 1,328(FP) |
| | Churned | 1,127(FN) | 1,691(TP) |

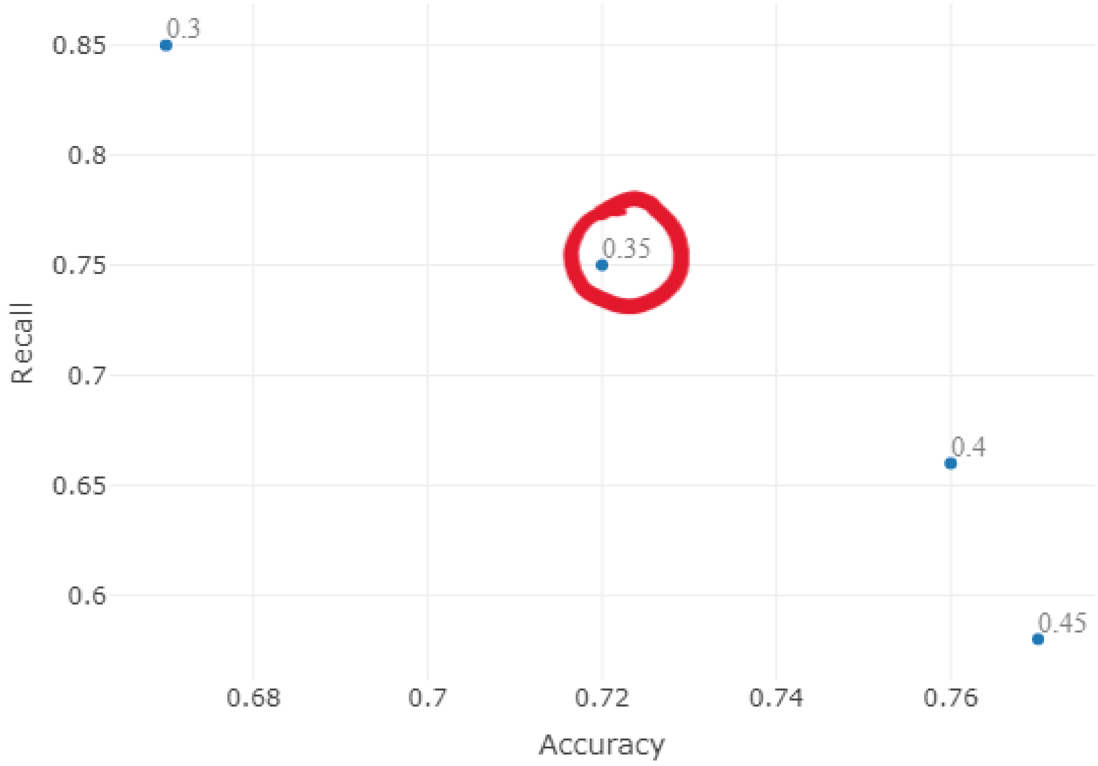
Accuracy: 58.61% Recall: 60.00%

# Model Evaluation

| | SMOTE | | UNDER SAMPLING | |
|---|---|---|---|---|
| **Model** | **Recall** | **Accuracy** | **Recall** | **Accuracy** |
| **Logistic Regression** | 0.53 | 0.56 | 0.56 | 0.56 |
| **Decision Tree** | 0.61 | 0.64 | 0.54 | 0.56 |
| **Naïve Bayes** | 0.10 | 0.47 | 0.51 | 0.56 |
| **KNN** | 0.61 | 0.67 | 0.59 | 0.56 |
| **Random Forest** | 0.66 | 0.72 | 0.60 | 0.58 |

# Improving Random Forest for better Recall

Trying out different threshold cut-off values for better recall and accuracy
Threshold values : 0.30,0.35,0.40,0.45




| | | Prediction | |
|---|---|---|---|
| | | Not Churned | Churned |
| Actual | Not Churned | 5,026(TN) | 1,328(FP) |
| | Churned | 2,154(FN) | 4,254(TP) |

**Best Fit Model**

| Accuracy | Recall |
|---|---|
| 0.72 | 0.75 |

# Model Interpretation

| Customer types | Counts | % |
|---|---|---|
| High risky customers<br>Churn Probability > 0.80 | 2379 | 73% |
| Moderate risky customers<br>Churn Probability  > 0.60 | 864 | 23% |

# Conclusion

- Logistic regression and Naïve Bayes didn’t perform well due to the complexity of the relation between the target and features.
- Reduced 57 features to around 20 by using Chi square and ANOVA test
- Data Imbalance technique **SMOTE** worked well.
- All the models had a low recall, it's difficult to detect customer which are going to churn.
- Random forest performed well hence the model was tuned for better Recall.
- We further divided the outcome of the model into high-risky customers and moderately risky customers for a better interpretation of the model.

# Thank you