Predictive Modelling Presentation

TELECOM CUSTOMER CHURN PREDICTION

TEAM – ISSSR

Identifying the indicators that will help predict if a customer will churn from the company or not

Background

Telecom Churn is becoming an increasingly significant problem today. With lots of carriers having promotions in terms of data, new phone and multiline it is becoming increasingly difficult to keep a customer engaged

Problem Statement

Around 7K customers have been identified where around 25% of customers churn. Based on information regarding customer related information , we need to predict if a customer will churn or not and determine what are the factors that cause him to churn

Data availability

The dataset was obtained from the IBM sample dataset repository that is commonly used for building models. The dataset is at a customer level , with 23 customer attributes. There is no out of time dataset to score for this problem.

Steps carried out for factor analysis of churned customers vs non churned customers

First Step

Data cleaning and Hypothesis building

- Treating Missing values and Invalid records
- Exploratory data analysis
- Outlier treatment

Second Step

Feature engineering and Model Generation

- Transform and adding new indicators
- Feature generation
- Baseline model
- Machine learning algorithms

Third Step

Model validation and fine tuning

- Scoring and Validating the model for best fit
- Attempt ways to minimise misclassification rate

Fourth Step

Insights Generation and next steps

- Provide Insights and Recommendation
- Identify ways to improve upon existing model

Exploratory data analysis was conducted to come up with high level insights about the business

Customer factors

- Online security
- Partner / Dependents
- Streaming TV / Movies
- Online backup
- Tech Support
- Payment method
- Paperless billing
- Multiple lines
- Tenure

Data structure

80% Categorical

20% Numeric

Data imputation

99%

Missing Rows	Total Charges NA
0%	0.5%

- Total charges were null for customers who had 0 tenure. Hence, Imputed to 0

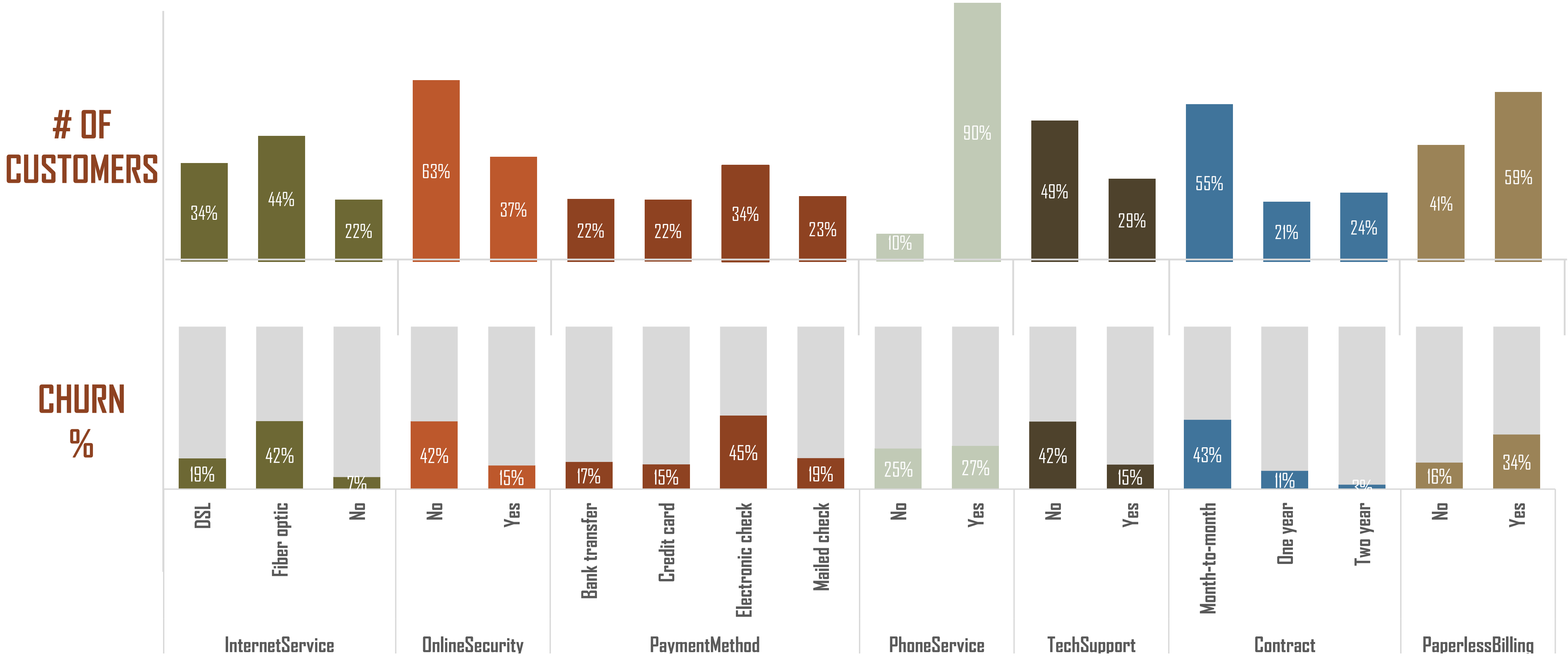
Data partitions

80% Training

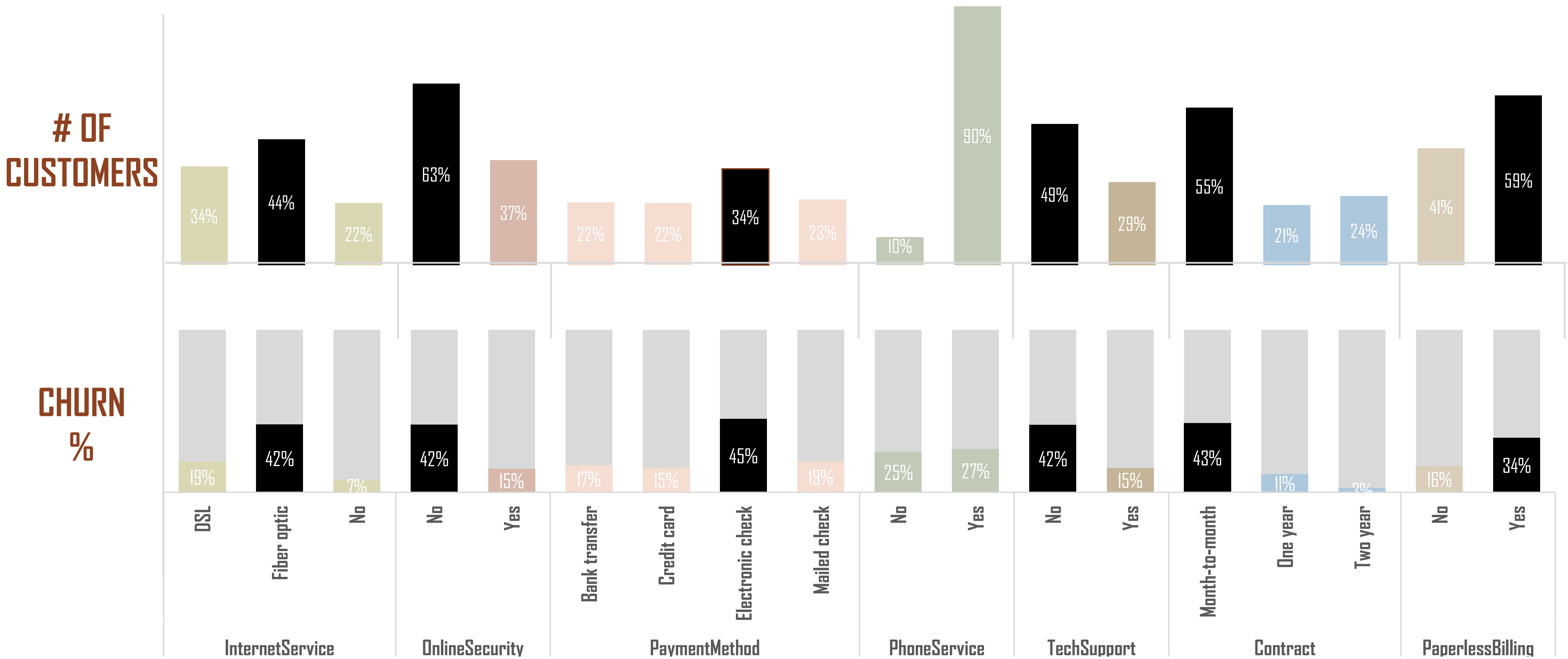
20% Testing

- Hypothesis and model building was done on training set
- Testing dataset was used to evaluate the model to remove overfitting and generalizability

Bivariate analysis gives an indication about some possible characteristics of churn behavior



Bivariate analysis gives an indication about some possible characteristics of churn behavior



- It is observed that Customers who don't opt for online security, have electronic check payment method, don't have tech support and are on a month to month contract are likely to churn
- These customers also represent a sizeable chunk of the customer base

Feature engineering proved to be important in a dataset with minimum variables. We validate using WOE and build our baseline model

Variable enhancement

- Categorical variables were converted to indicator variables in order to study if there are interaction effects. For ex. The effect of a particular variable occurring more in a year
- Math transformation was used to get best representation of features

Payment *Tenure *Multiline

+

Tanh Square Cube Log Exponential

(Tenure , Total charges, Monthly charges)

Weight of Evidence

Variable	Variable importance
Contract	1.188434
M2M	1.043467
tenure	0.797388
Contract_2Y	0.79628
NoFibreOptic	0.762879
TechSupport	0.710629
OnlineSecurity	0.709078
InternetService	0.673835
TechSupport_No	0.646902
Fiber optic	0.569696
OnlineBackup	0.538982
DeviceProtection	0.514396
PaymentMethod	0.450602
PM_ElectronicCheck	0.444976
MonthlyCharges	0.42982

- Contract type of a customer is crucial to determine if a customer will churn
- Customers who prefer electronic means (such as paperless billing, electronic means and fibre net) churn a lot

For logistic problems, categorical variables are said to have highest predictive power. Additional features will be useful for different models

MODEL RESULTS

<table>
<tr><th></th><th>Estimate</th><th>Std. Error</th><th>z value</th><th>Pr(>|z|)</th></tr>
<tr><td>(Intercept)</td><td>-2.8509</td><td>0.229323</td><td>-12.432</td><td>< 2e-16</td></tr>
<tr><td>tenure</td><td>-0.01623</td><td>0.004351</td><td>-3.731</td><td>0.000191</td></tr>
<tr><td>Contract with Month to Month</td><td>1.877343</td><td>0.236394</td><td>7.942</td><td>2.00E-15</td></tr>
<tr><td>FiberOptic Internet</td><td>1.240954</td><td>0.083523</td><td>14.858</td><td>< 2e-16</td></tr>
<tr><td>PaperlessBilling_Yes</td><td>0.515963</td><td>0.084045</td><td>6.139</td><td>8.30E-10</td></tr>
<tr><td>Electronic check Payment</td><td>0.577782</td><td>0.078116</td><td>7.396</td><td>1.40E-13</td></tr>
<tr><td>Tenure + Contract_MM</td><td>-0.02058</td><td>0.004942</td><td>-4.163</td><td>3.14E-05</td></tr>
</table>

Confusion Matrix

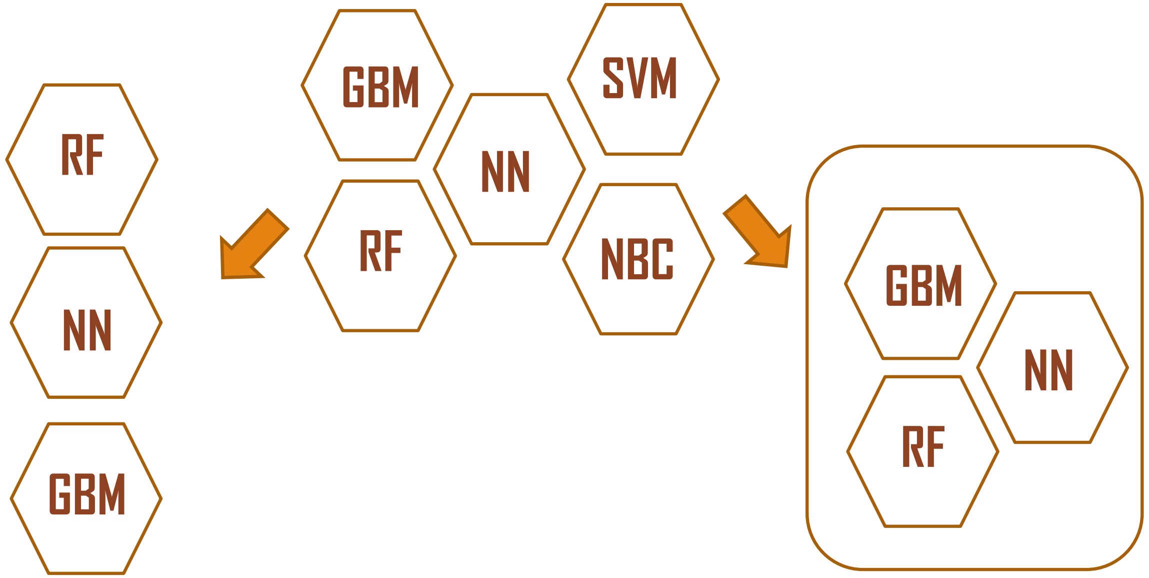
	Predict Churn	Predict Active
Actual Churn	156	311
Actual Active	82	1212

Misclassification rate 20%

AUC Statistic 0.833

Alternate models

- Apart from logistic regression, we have with us , the disposal to employ a wide variety of machine learning algorithms, in order to boost accuracy



- Machine learning models can be evaluated individually and as well as combined into an ensemble

Grid search was employed in order to get the best possible hyper parameters for Random forest and Gradient boosting machine

Random Forest

max_depth	mtries	ntrees	logloss	AUC
7	7	500	0.410149	0.83452
8	5	1000	0.41023	0.83345
7	7	1000	0.410277	0.83138

AUC Statistic

0.834

- Grid search was applied to get the variable sampling rate and maximum depth. Number of trees were based on user specification

Gradient Boosting Machine

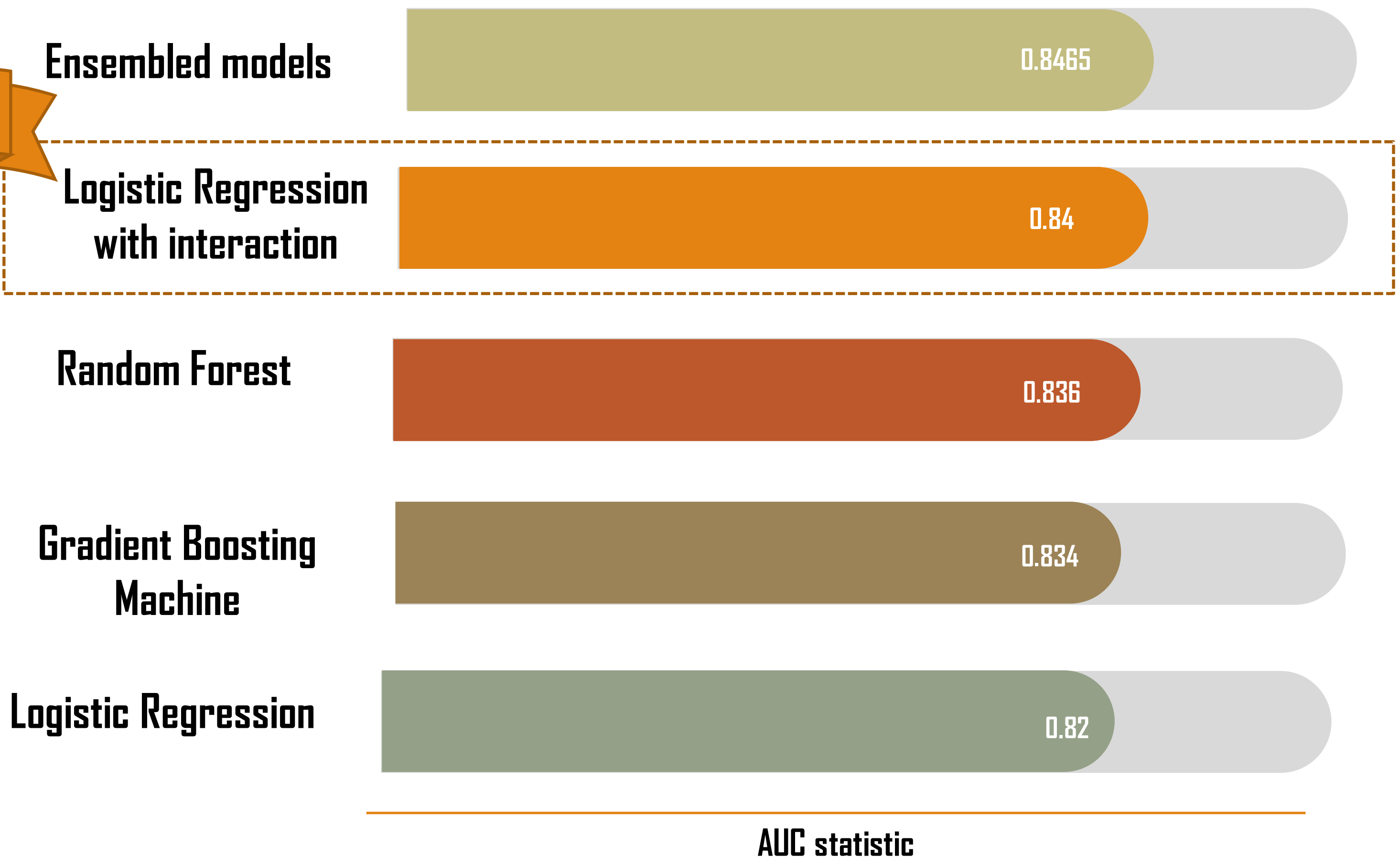
max_depth	min_rows	nbins	sample_rate	AUC
3	8	512	0.55	0.836802
6	256	64	0.65	0.83543
6	512	16	0.5	0.835386

AUC Statistic

0.836

- H2O package in R uses a validation dataset to optimize for number of trees and learning rate. The rest of the parameters are user inputs

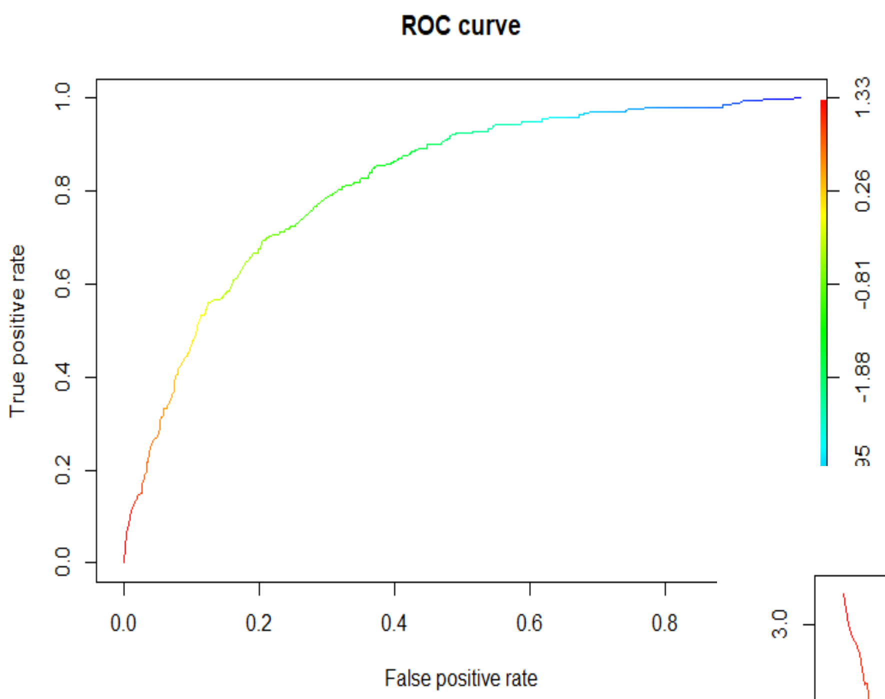
Model improvement was done by considering additional modelling techniques and ensembling different models together

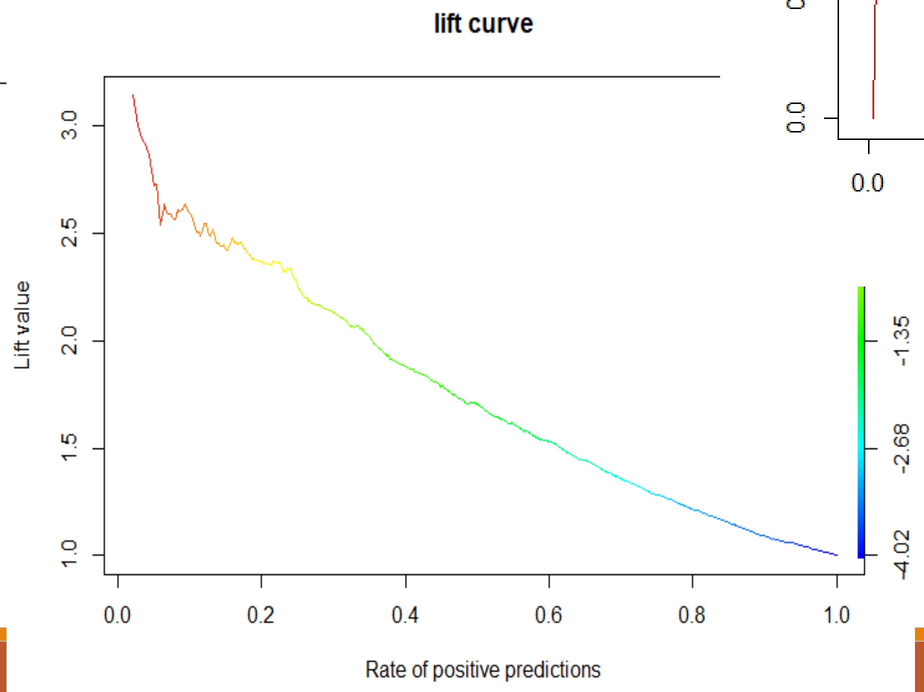


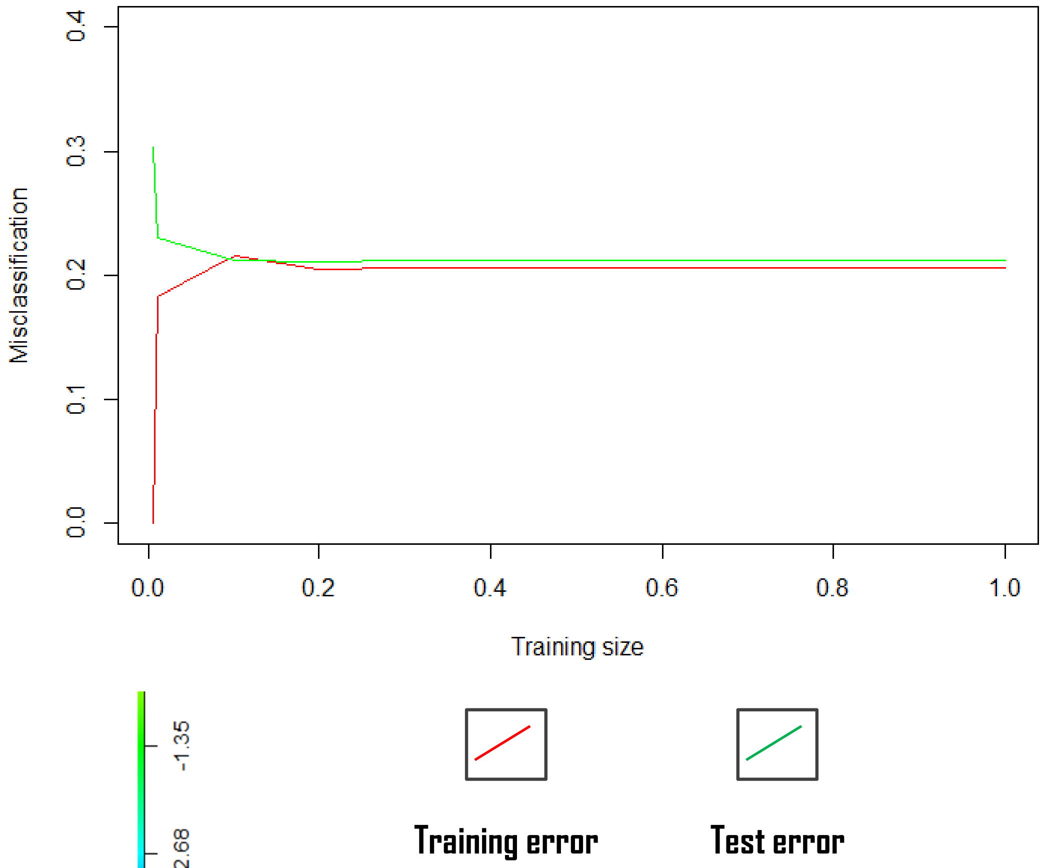
- The models whose input was considered for ensembling
  1. Logistic Regression
  2. KNN algorithm
  3. Random Forest

- The best model that can be formulated is an ensemble model and that is very comparative to the performance of a logistic model
- **Selecting a logistic model has many advantages**
  1. It can give a under-the-hood look of which variables are important and what is its contribution to risk
  2. It is more stable provided the Out of time validation dataset follows same distribution of training and validation dataset

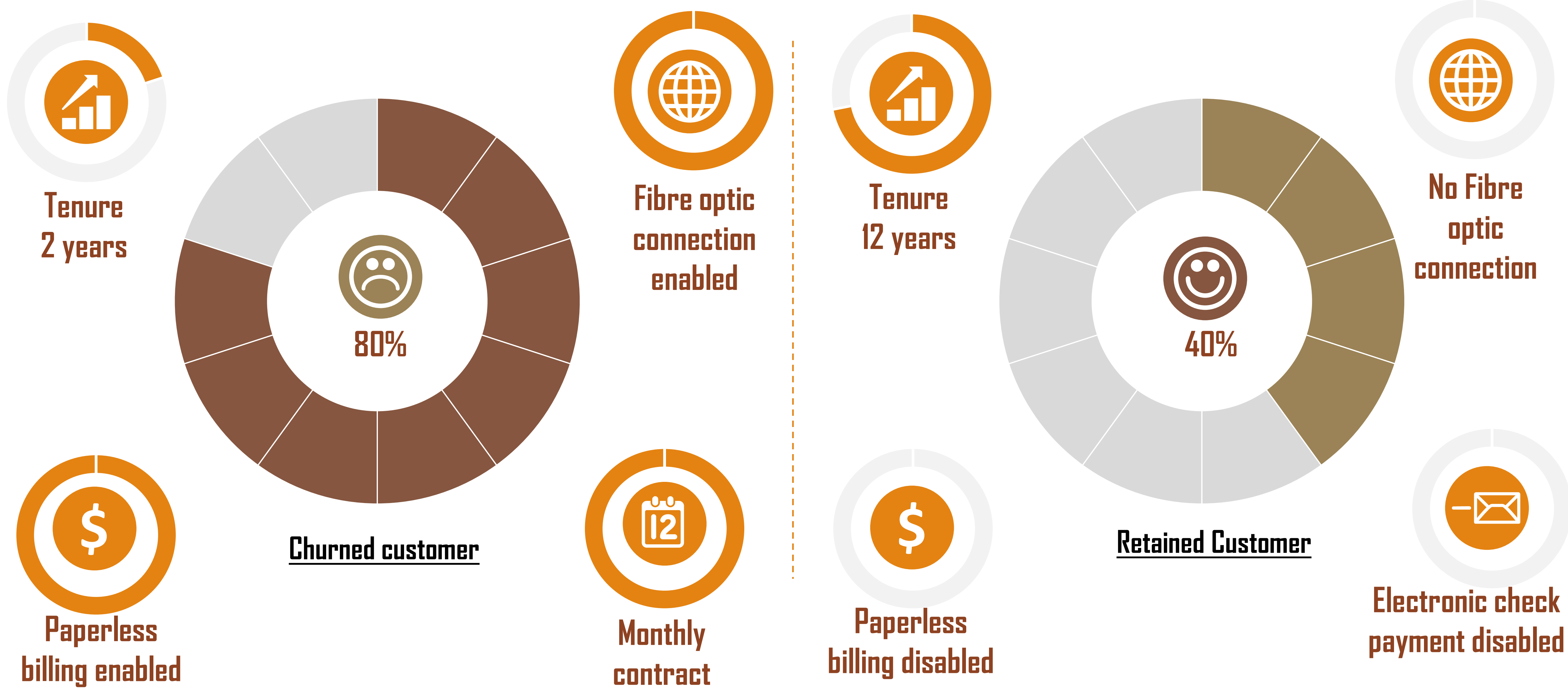
Using the logistic regression model, We generate the ROC , lift as well as learning curve for this model





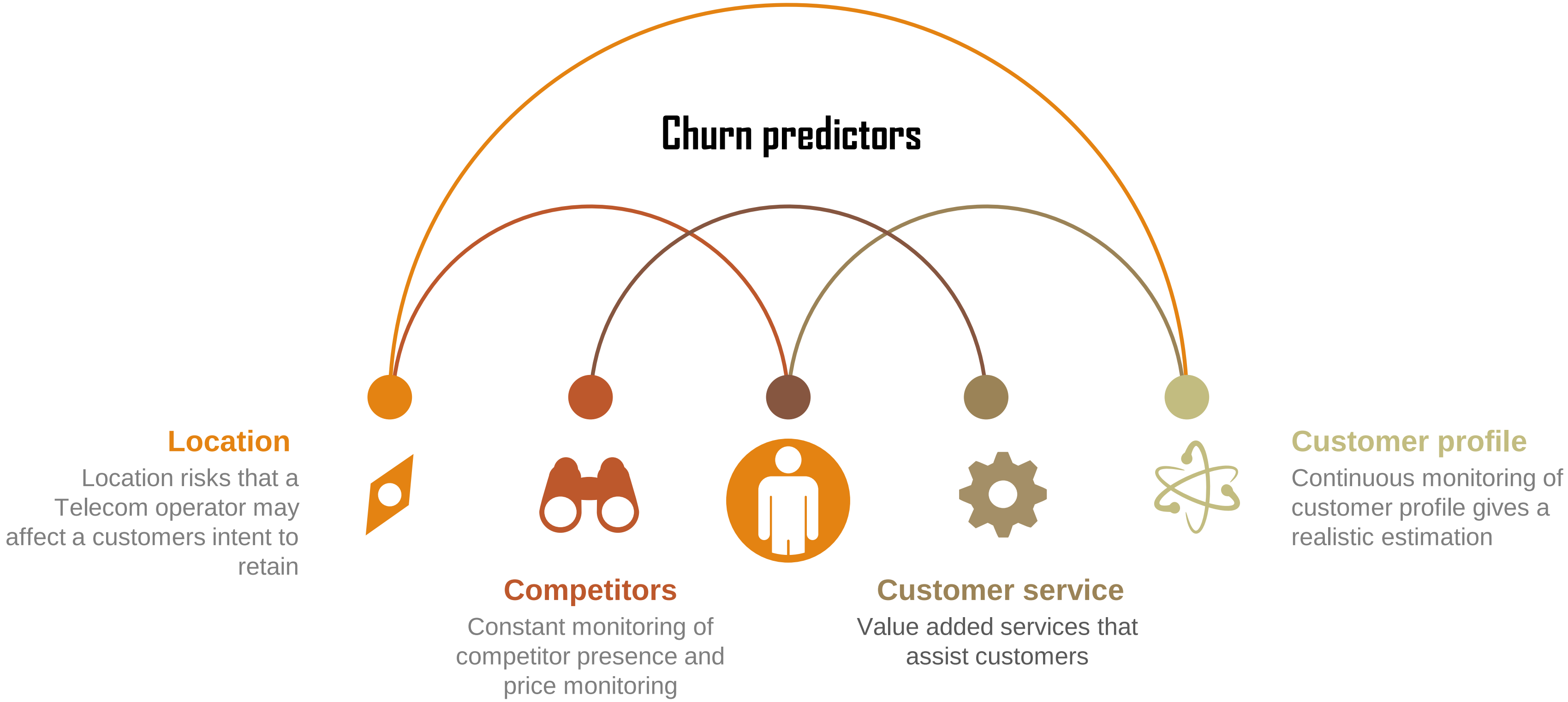


Based on the model building exercise, a specific set of profiles have been constructed for a churned customer and non churned customer



Next steps for better model accuracy statistics and having a detailed customer profile for churn

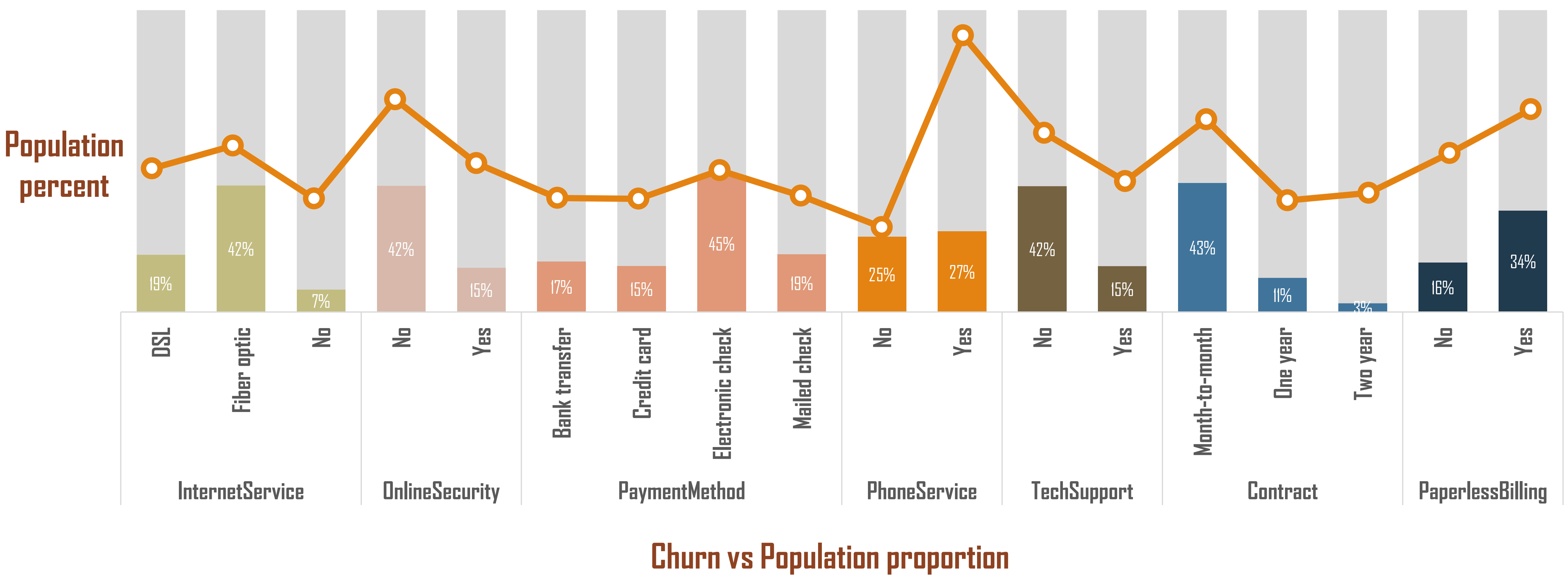
- While customer specific attributes are good indicators of churn, they represent only a part of the puzzle
- Customer experience with the company will be a firm determiner whether he will churn or not
- Misclassification rate is high, and tendency to predict churn isn't there. Therefore we need strong indicators of churn



Thank You

Appendix

Bivariate analysis gives an indication about some possible characteristics of churn behavior



Selecting the right interaction terms were done using 3 different scenarios

Scenario 1-

Consider all variables in the model → Create interaction terms for all variables → Run stepwise regression to judge significant interaction

Scenario 2 -

Consider all variables in the model → Create interaction terms for all variables → Run lasso regression to judge significant interaction

Scenario 3 -

Consider all variables in model → Run chaid tree to judge significant interaction

Selecting the right interaction terms were done using 3 different scenarios