TELECOM CHURN MANAGEMENT

NIRANJAN KUMAR	D18023
NARAYANA	D18030
SASIDHAR	D18014
ANIL P	D18035
SRAVANYA G TAYI	D18039
SATISH	D18036

Business Context – WHY?

- Telecom industry is one of the few industries that is hammered on both sides with increasing competition, higher CAPEX on one side and decreasing ARPUs (Average Revenue Per User) on the other side.
- In order to achieve sustainable profitability, it becomes imperative for a Service Provider to have an effective Customer Value Management initiative.
- Understanding the true value of a possible customer churn will help the company in it's customer relationship management.
- If we are able to predict the churn customers in advance, the attributes of customers whom we are going to loose in near future one can take corrective action so that we can minimize this problem.

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Business Questions

- Which factors are contributing to churn.
- Predict the probable churn customers.

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Project Objective

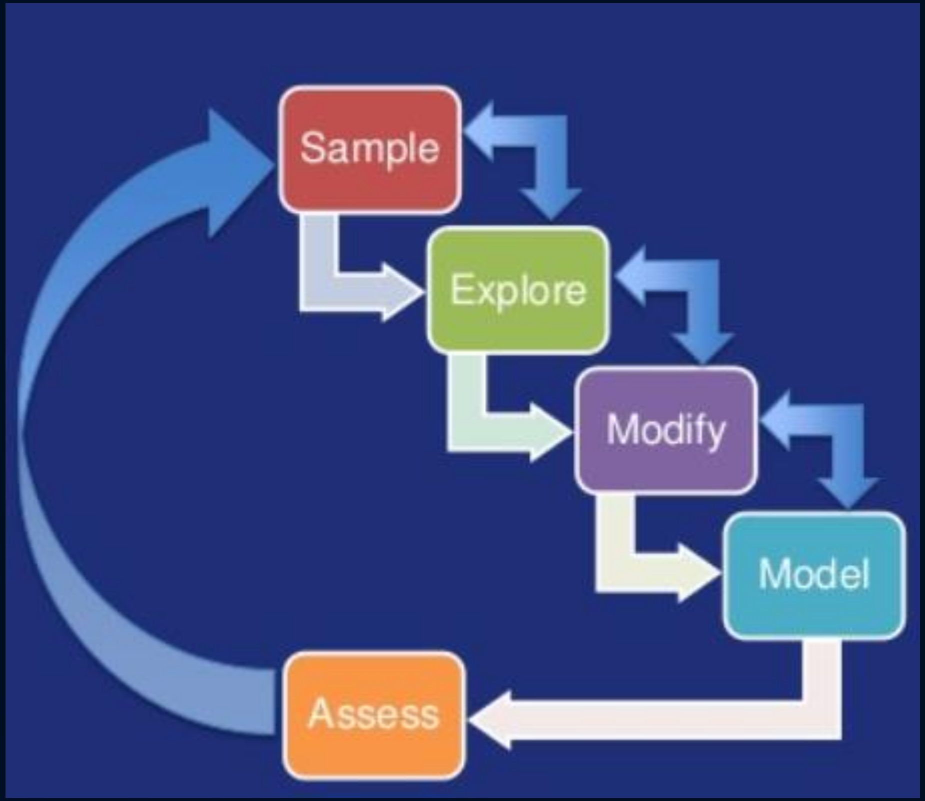
- The goal of the study is to apply analytical techniques to predict a customer churn and analyse the churning and non-churning customers.

- Develop a system through which the client could identify the customer churn rate and decide what should be the appropriate incentive for them.

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

SEMMA Framework

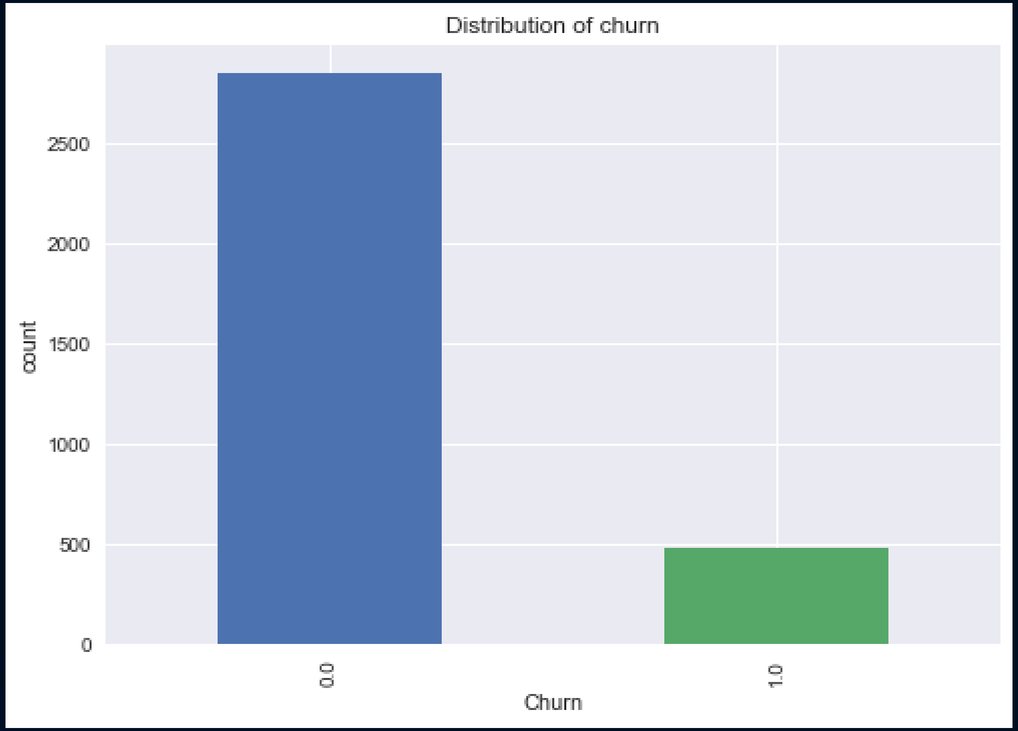
- ***SEMMA*** is an acronym that stands for *Sample*, *Explore*, *Modify*, *Model*, and *Assess.*
- It is a list of sequential steps developed by SAS Institute, one of the largest producers of statistics and business intelligence software.
- SEMMA offers an easy to understand process, allowing an organized and adequate development and maintenance of Machine Learning projects.



Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Challenges in Data

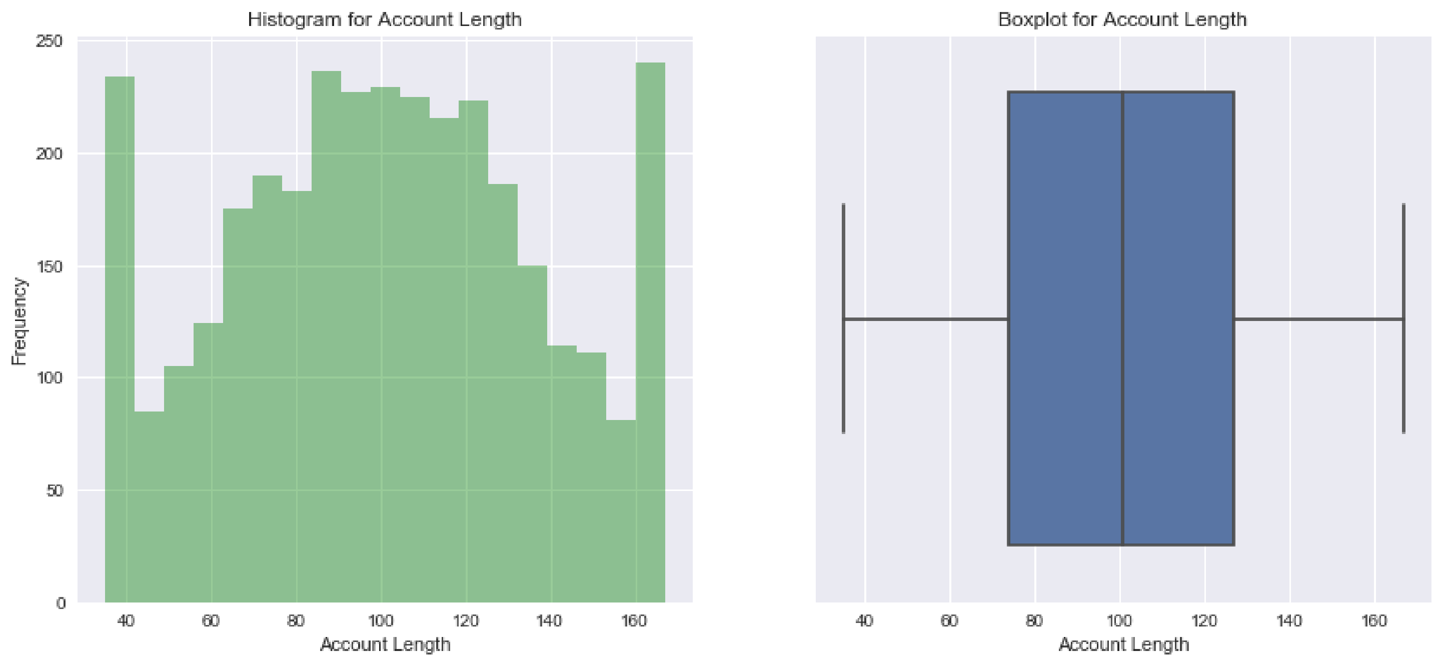
- It is the case of imbalanced data, so majority classes dominate over minority classes causing the machine learning classifiers to be more biased towards majority classes.
- This causes poor classification of minority classes.
- Used Stratified Sampling to reduce the effect of imbalanced data.

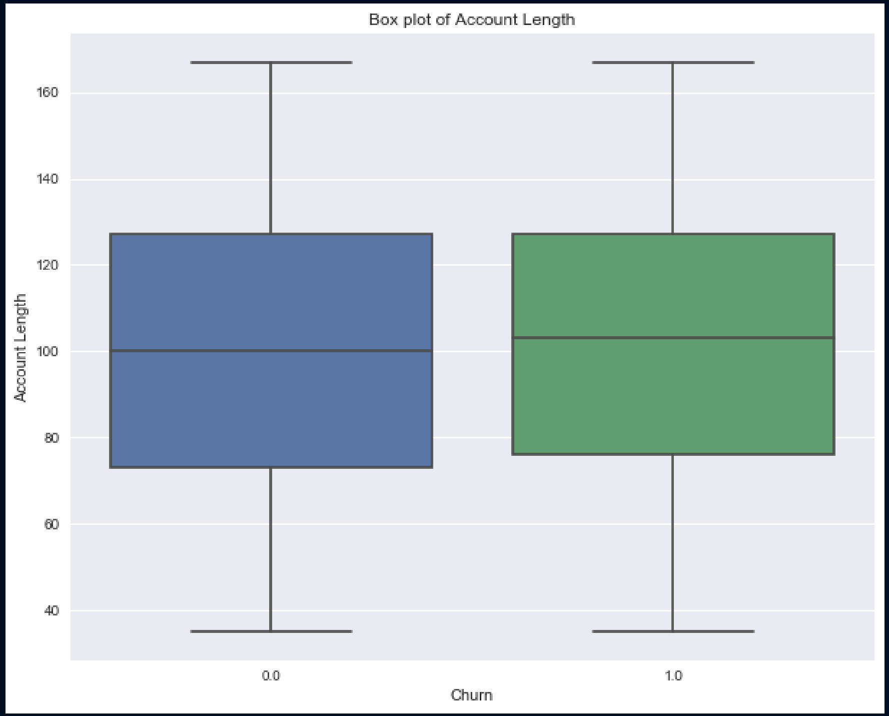


Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

EXPLORATORY DATA ANALYSIS

Significance of Account Duration

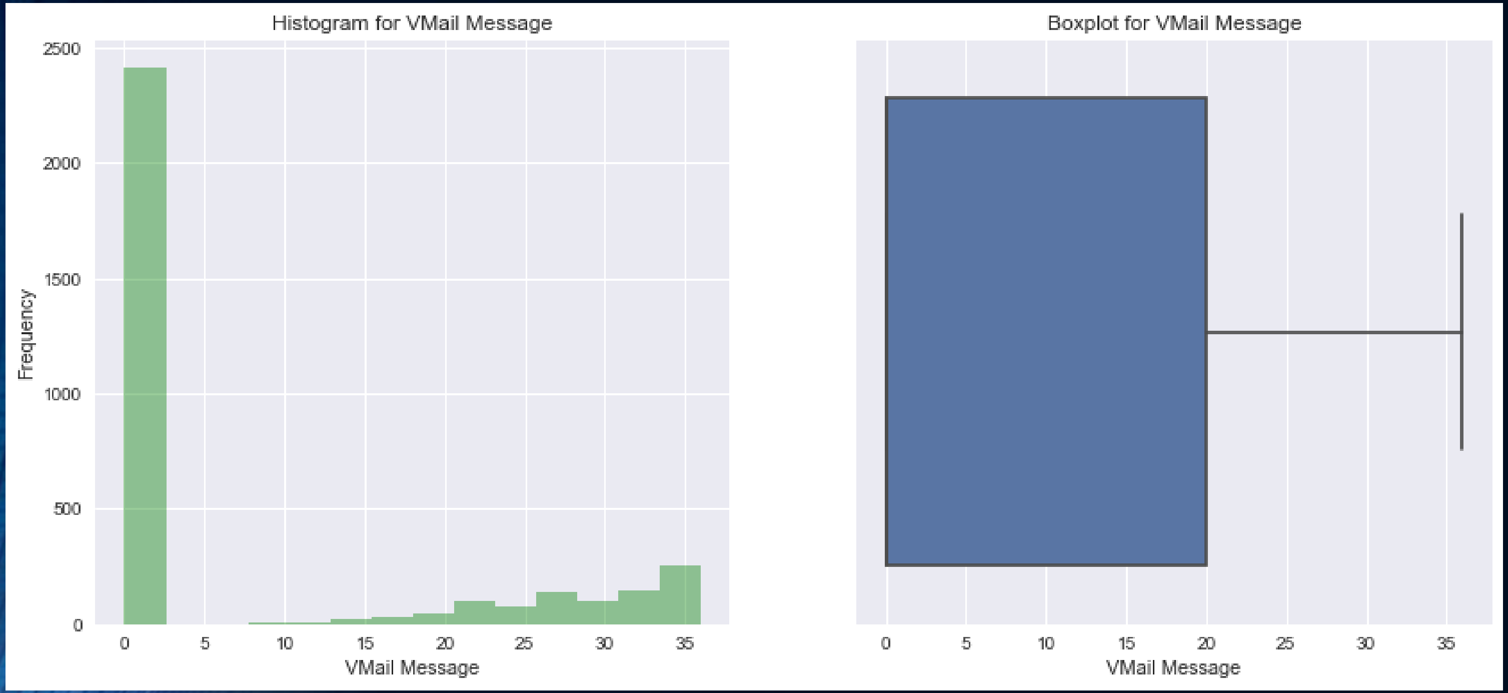


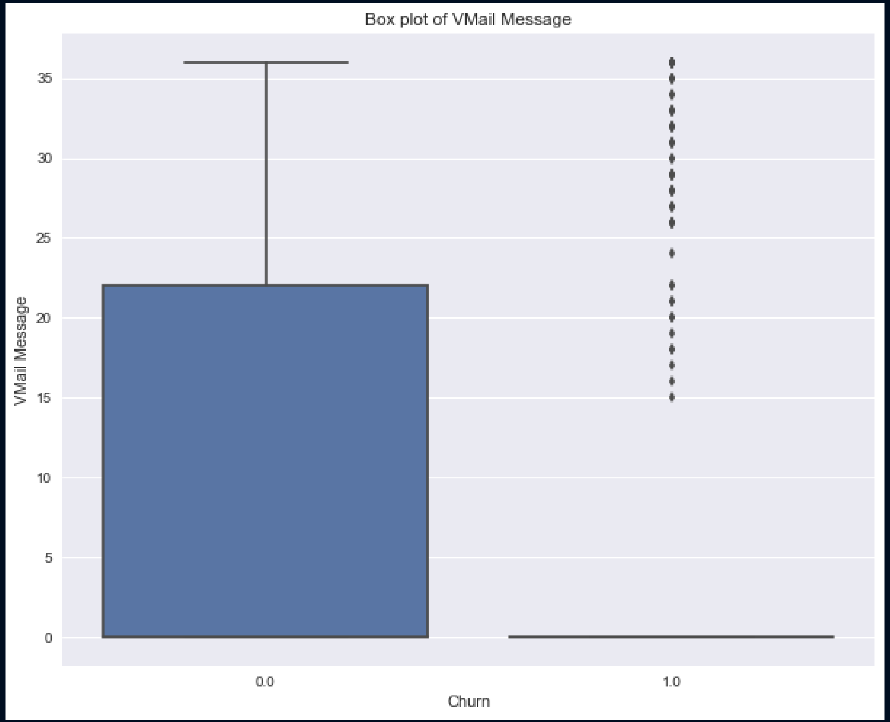


- Customer's Account duration almost follows normal distribution. It shows that the data has a mix of old and new customers.
- Account Length has no effect on the churn rate.

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Significance of VMail Message

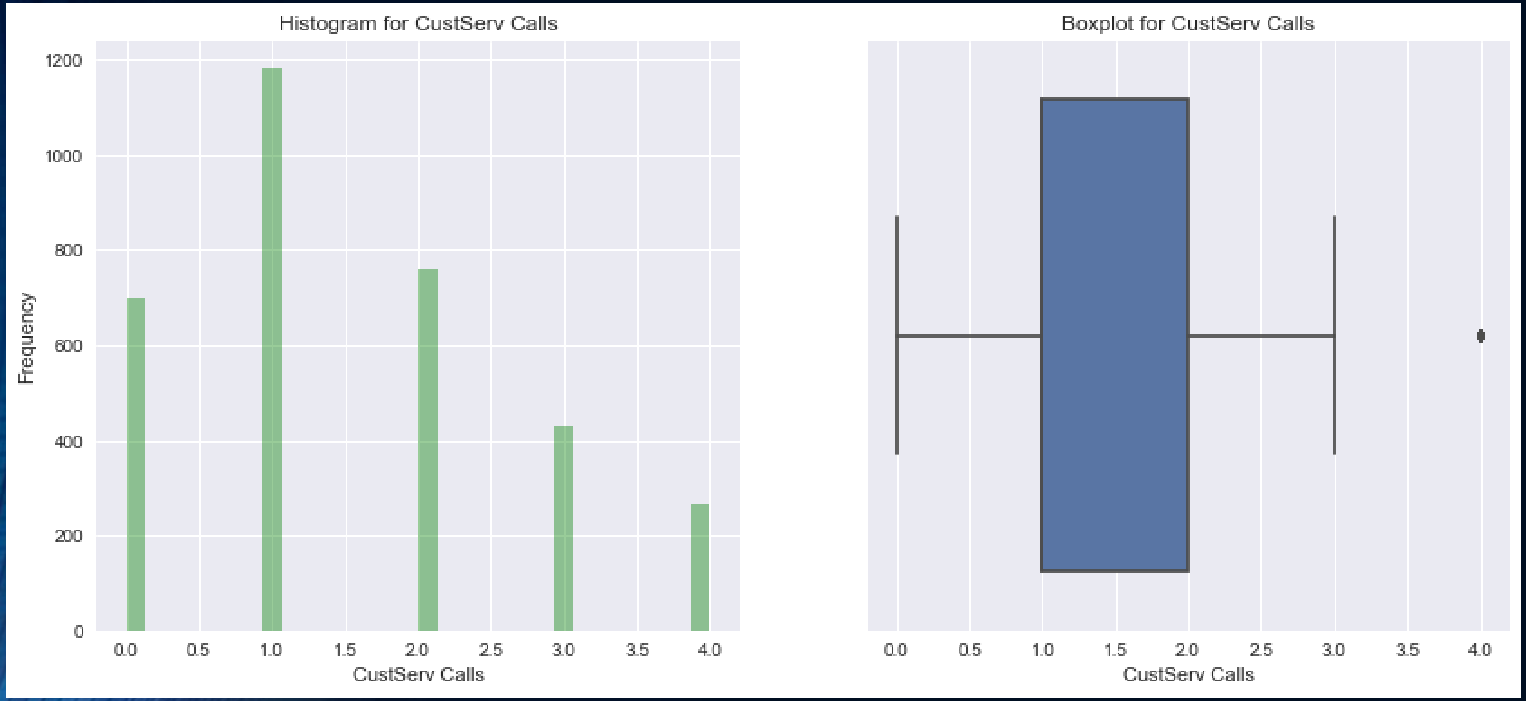


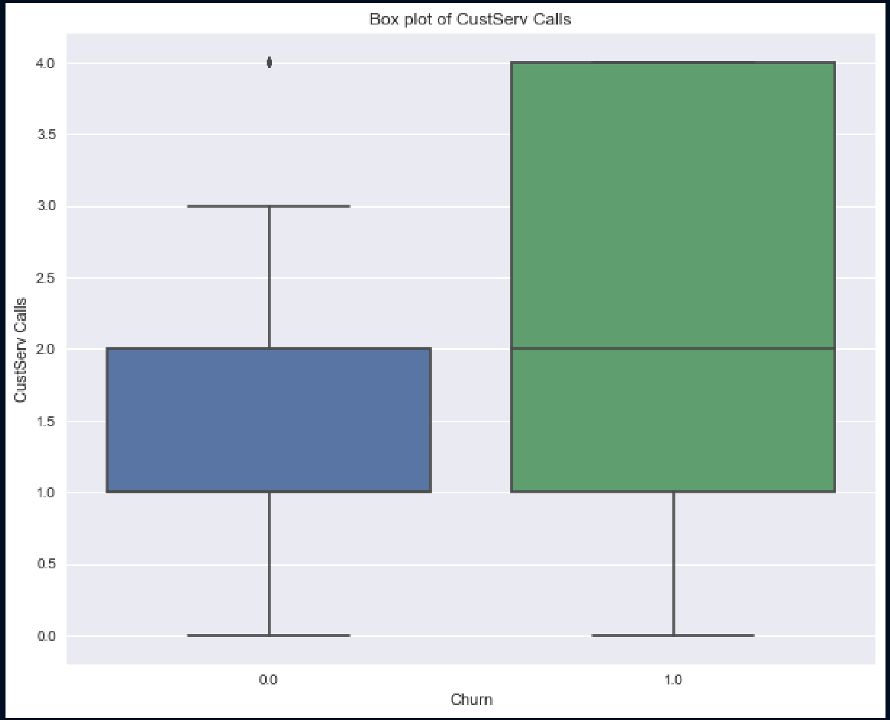


- 2411 customers out of 3333 is not at all using the Vmail Message service.
- Almost 72% of customers are not using this service.
- Customers who are using Vmail service is not leaving the company.

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Significance of Customer Service Calls

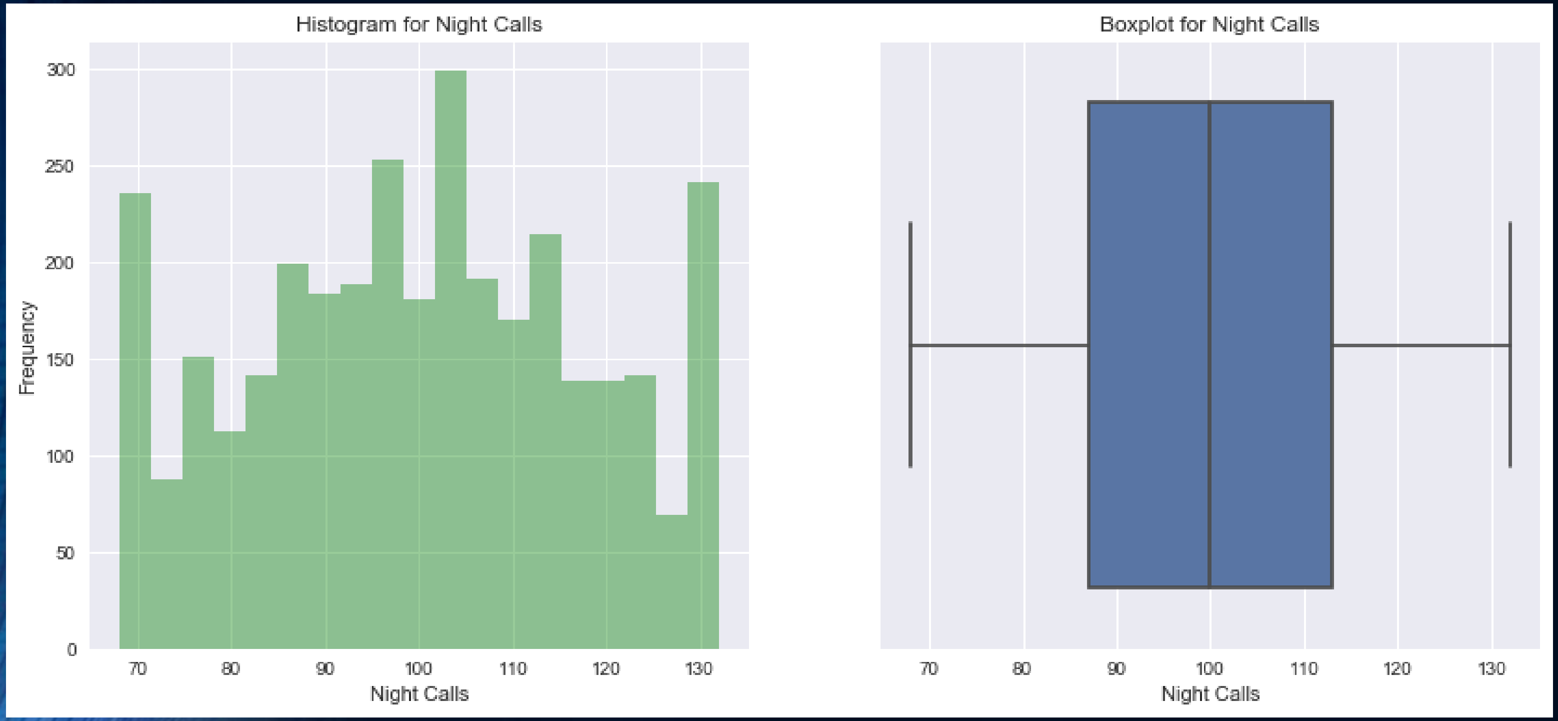


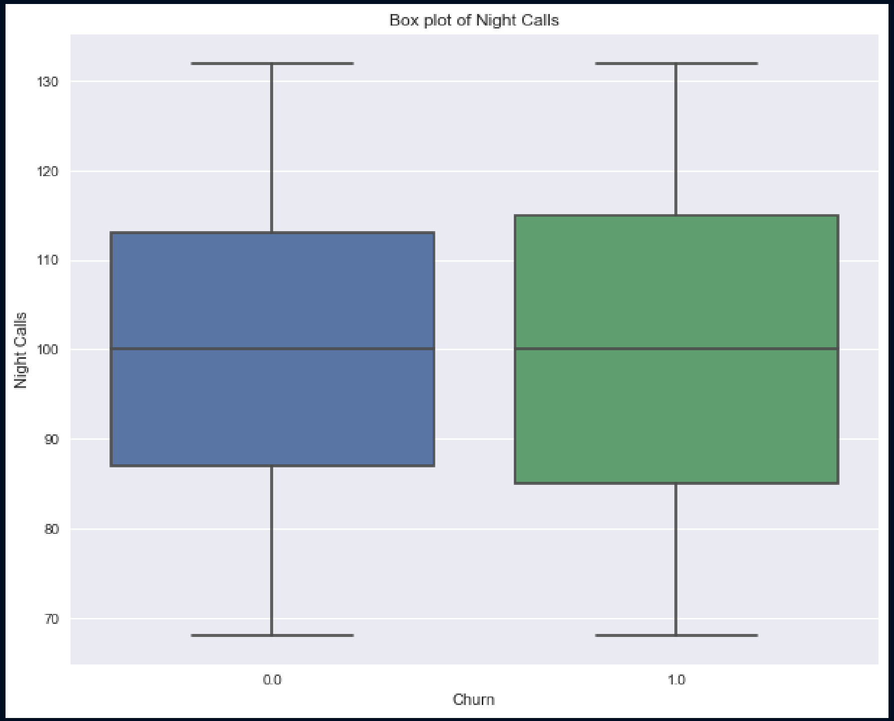


- Majority of customers (1200) has called customer service only once.
- 750 customers has called customer service twice.
- Customers who called customer service more than twice (on average) left the company.

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

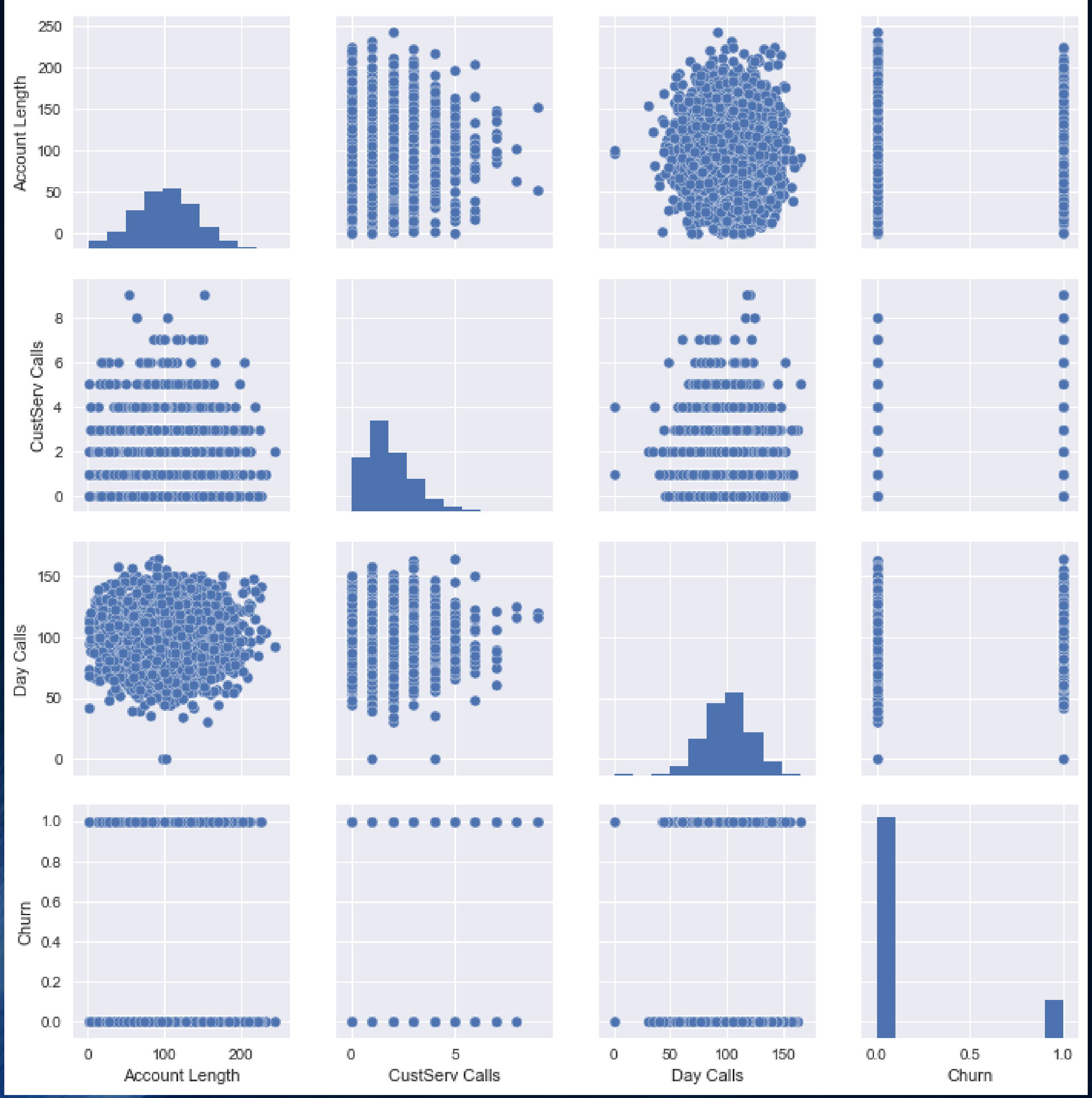
Significance of Night Calls





- Distribution of night calls show that it is a bit right skewed – More calls at late nights.
- Customers who are making more calls at Night are leaving the service provider.

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn



- From the scatter plot, we can see that there is no linear relationship between the predictors and target variable.

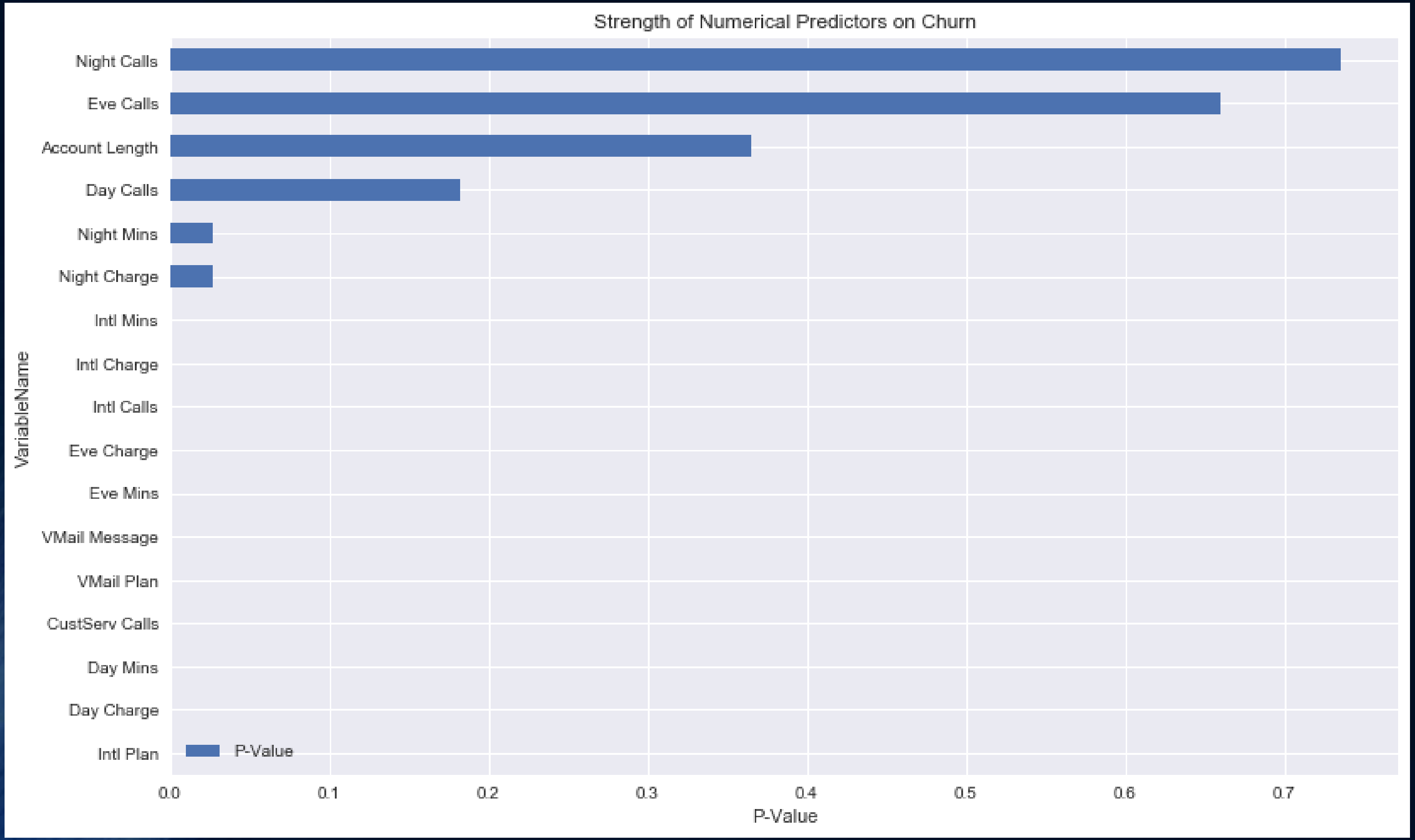
- There is no pattern observed between the account duration and calls made during the day.

- From the scatter plot between Account Length and Customer service calls, we can see that the relatively new customers are calling the customer service more number of times.

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

STRENGTH OF PREDICTORS

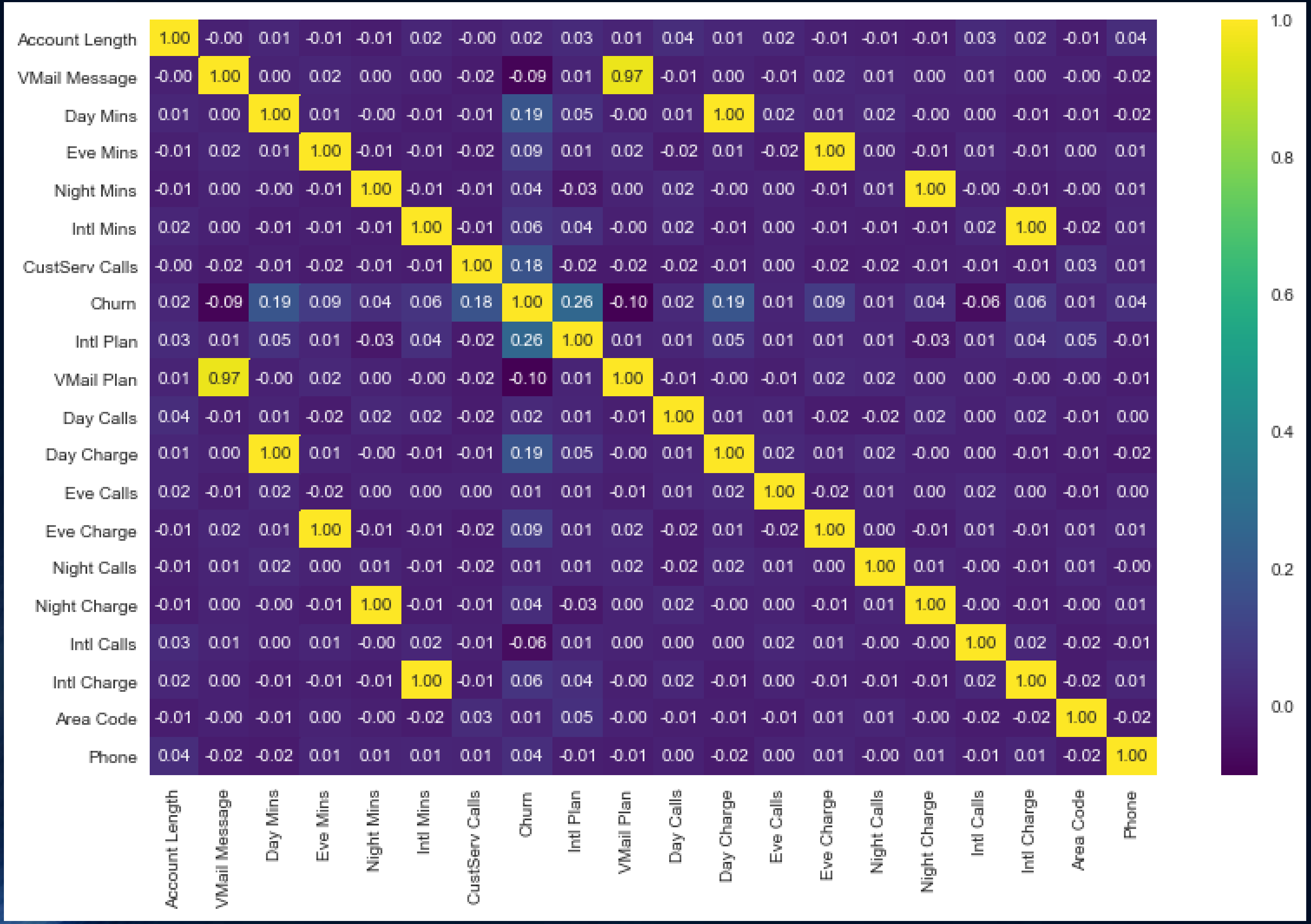
Strength of Predictors



Performed T-Test to find the strength of numerical predictors on churn.

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Correlation Plot



	Account Length	VMail Message	Day Mins	Eve Mins	Night Mins	Intl Mins	CustServ Calls	Churn	Intl Plan	VMail Plan	Day Calls	Day Charge	Eve Calls	Eve Charge	Night Calls	Night Charge	Intl Calls	Intl Charge	Area Code	Phone
Account Length	1.00	-0.00	0.01	-0.01	-0.01	0.02	-0.00	0.02	0.03	0.01	0.04	0.01	0.02	-0.01	-0.01	-0.01	0.03	0.02	-0.01	0.04
VMail Message	-0.00	1.00	0.00	0.02	0.00	0.00	-0.02	-0.09	0.01	0.97	-0.01	0.00	-0.01	0.02	0.01	0.00	0.01	0.00	-0.00	-0.02
Day Mins	0.01	0.00	1.00	0.01	-0.00	-0.01	-0.01	0.19	0.05	-0.00	0.01	1.00	0.02	0.01	0.02	-0.00	0.00	-0.01	-0.01	-0.02
Eve Mins	-0.01	0.02	0.01	1.00	-0.01	-0.01	-0.02	0.09	0.01	0.02	-0.02	0.01	-0.02	1.00	0.00	-0.01	0.01	-0.01	0.00	0.01
Night Mins	-0.01	0.00	-0.00	-0.01	1.00	-0.01	-0.01	0.04	-0.03	0.00	0.02	-0.00	0.00	-0.01	0.01	1.00	-0.00	-0.01	-0.00	0.01
Intl Mins	0.02	0.00	-0.01	-0.01	-0.01	1.00	-0.01	0.06	0.04	-0.00	0.02	-0.01	0.00	-0.01	-0.01	-0.01	0.02	1.00	-0.02	0.01
CustServ Calls	-0.00	-0.02	-0.01	-0.02	-0.01	-0.01	1.00	0.18	-0.02	-0.02	-0.02	-0.01	0.00	-0.02	-0.02	-0.01	-0.01	-0.01	0.03	0.01
Churn	0.02	-0.09	0.19	0.09	0.04	0.06	0.18	1.00	0.26	-0.10	0.02	0.19	0.01	0.09	0.01	0.04	-0.06	0.06	0.01	0.04
Intl Plan	0.03	0.01	0.05	0.01	-0.03	0.04	-0.02	0.26	1.00	0.01	0.01	0.05	0.01	0.01	0.01	-0.03	0.01	0.04	0.05	-0.01
VMail Plan	0.01	0.97	-0.00	0.02	0.00	-0.00	-0.02	-0.10	0.01	1.00	-0.01	-0.00	-0.01	0.02	0.02	0.00	0.00	-0.00	-0.00	-0.01
Day Calls	0.04	-0.01	0.01	-0.02	0.02	0.02	-0.02	0.02	0.01	-0.01	1.00	0.01	0.01	-0.02	-0.02	0.02	0.00	0.02	-0.01	0.00
Day Charge	0.01	0.00	1.00	0.01	-0.00	-0.01	-0.01	0.19	0.05	-0.00	0.01	1.00	0.02	0.01	0.02	-0.00	0.00	-0.01	-0.01	-0.02
Eve Calls	0.02	-0.01	0.02	-0.02	0.00	0.00	0.00	0.01	0.01	-0.01	0.01	0.02	1.00	-0.02	0.01	0.00	0.02	0.00	-0.01	0.00
Eve Charge	-0.01	0.02	0.01	1.00	-0.01	-0.01	-0.02	0.09	0.01	0.02	-0.02	0.01	-0.02	1.00	0.00	-0.01	0.01	-0.01	0.01	0.01
Night Calls	-0.01	0.01	0.02	0.00	0.01	-0.01	-0.02	0.01	0.01	0.02	-0.02	0.02	0.01	0.00	1.00	0.01	-0.00	-0.01	0.01	-0.00
Night Charge	-0.01	0.00	-0.00	-0.01	1.00	-0.01	-0.01	0.04	-0.03	0.00	0.02	-0.00	0.00	-0.01	0.01	1.00	-0.00	-0.01	-0.00	0.01
Intl Calls	0.03	0.01	0.00	0.01	-0.00	0.02	-0.01	-0.06	0.01	0.00	0.00	0.00	0.02	0.01	-0.00	-0.00	1.00	0.02	-0.02	-0.01
Intl Charge	0.02	0.00	-0.01	-0.01	-0.01	1.00	-0.01	0.06	0.04	-0.00	0.02	-0.01	0.00	-0.01	-0.01	-0.01	0.02	1.00	-0.02	0.01
Area Code	-0.01	-0.00	-0.01	0.00	-0.00	-0.02	0.03	0.01	0.05	-0.00	-0.01	-0.01	-0.01	0.01	0.01	-0.00	-0.02	-0.02	1.00	-0.02
Phone	0.04	-0.02	-0.02	0.01	0.01	0.01	0.01	0.04	-0.01	-0.01	0.00	-0.02	0.00	0.01	-0.00	0.01	-0.01	0.01	-0.02	1.00

Direct correlation between

- Day Charge & Day Mins
- Night Charge & Night Mins
- Intl Charge & Intl Mins

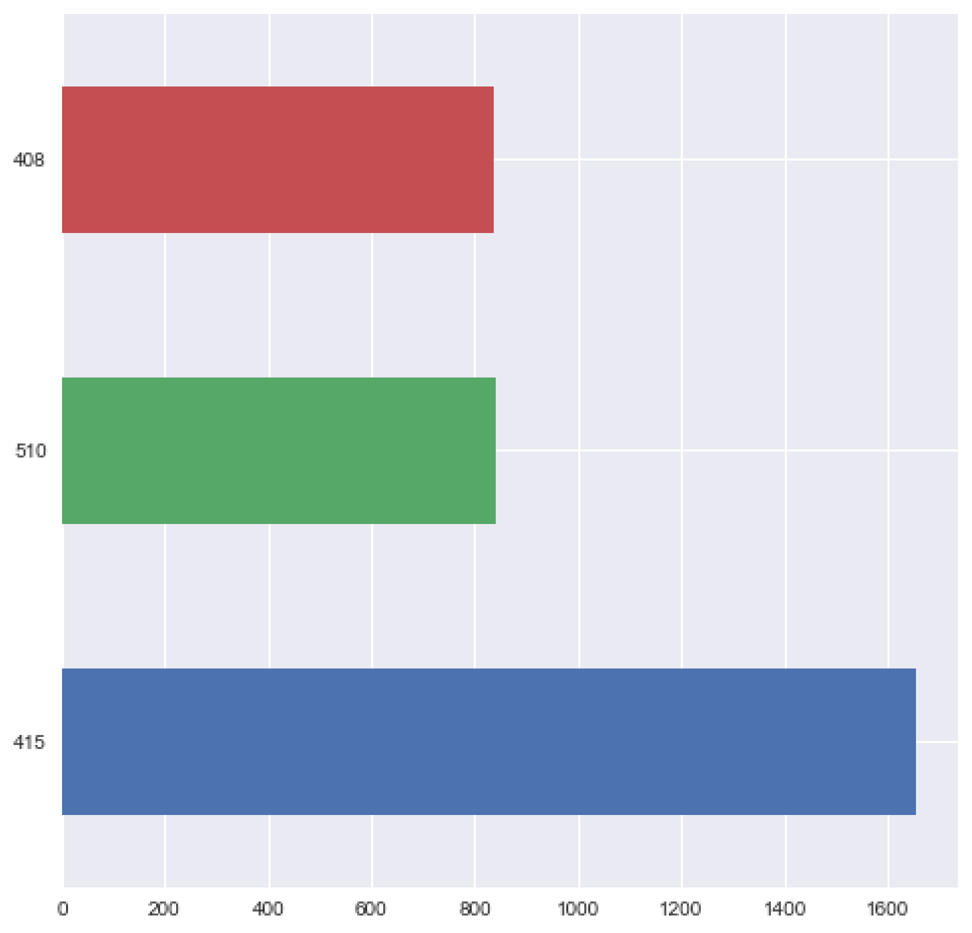
So we have dropped these variables from model:

- Day Charge
- Night Charge
- Intl Charge

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Feature Engineering

- This is often the most important step in applied machine learning
- Transformed the “Area Code” variable to three different variables.
- Created three different flag variables to indicate the Area code.



Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Model Building

- As there is no linear relationship between dependent and independent variables parametric techniques will perform very poorly.

- Tree based methods like decision tree is simple and useful for interpretation and Nonlinear relationships between parameters do not affect decision tree performance. However decision trees typically are not competitive with some of the ensemble techniques like boosting.

- The problem with decision trees is that for every small change in input data, it can lead to high variability in predictions. This high variability would qualify it as unstable classifier.

- To reduce the high variance associated with unstable model like decision tree, we have used Adaboost technique.

- Boosting helps to reduce both bias and variance. Thus boosting offers way to build a better model by balancing the bias – variance trade off.

- Adaboost algorithm penalizes the observations if the predictions are misclassified and it gives more importance to observations with correct predictions.

Model Building

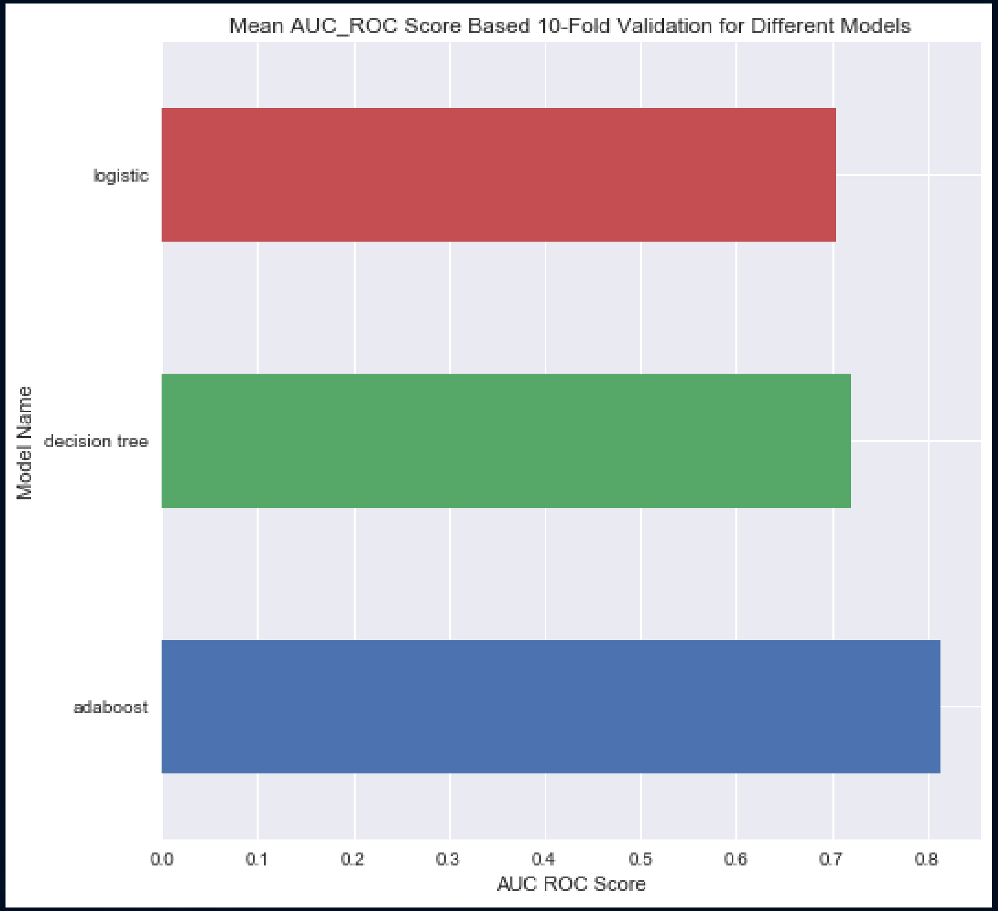
- We have tried 3 different modelling techniques with stratified sampling to reduce the affect of imbalanced dataset.
  - Logistic Regression
  - Decision Tree Classifier
  - AdaBoost Classifier

- The data was split into train and test in the ratio of 80: 20.
  - Training data has 2333 observations & 65 features.
  - Testing data has 1000 observations & 65 features.

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Model Selection

- We have evaluated all 3 models using ROC Score and selected Adaboost as the final model.
- Performed 10 Fold Cross Validation to identify how well model performed without any overfitting.



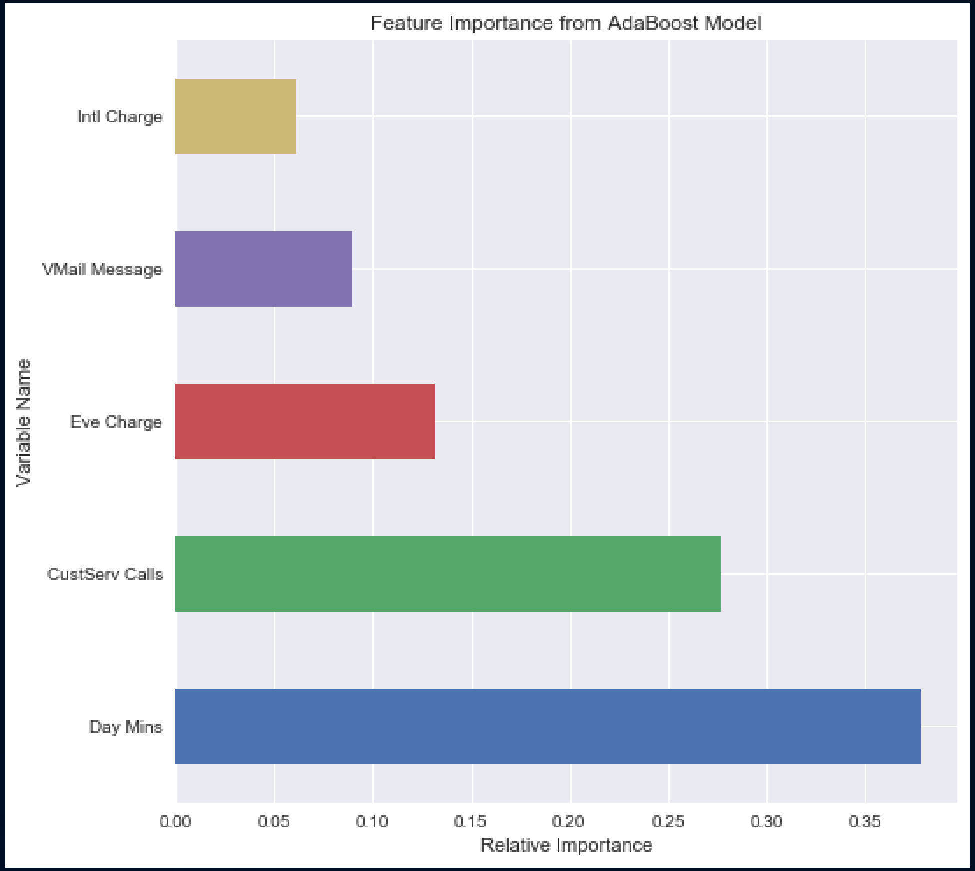
Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Feature Importance

- The advantage of Adaboost Classifier is that we can directly get the variables which are affecting the churn rate. This is an another reason of selecting the Adaboost Model.

- Important features mean that these features are more closely related with the churn rate.

- By know the factors which are driving the churn, we can take some preventive measures to reduce the churn rate.



Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Model Validation

- Model Validation is an critical phase in SEEMA Framework before they are deployed for use in the field.

- Model Validation Metrics
  - Confusion Matrix
  - ROC Curve
  - Gain Chart
  - Lift Chart
  - KS Chart

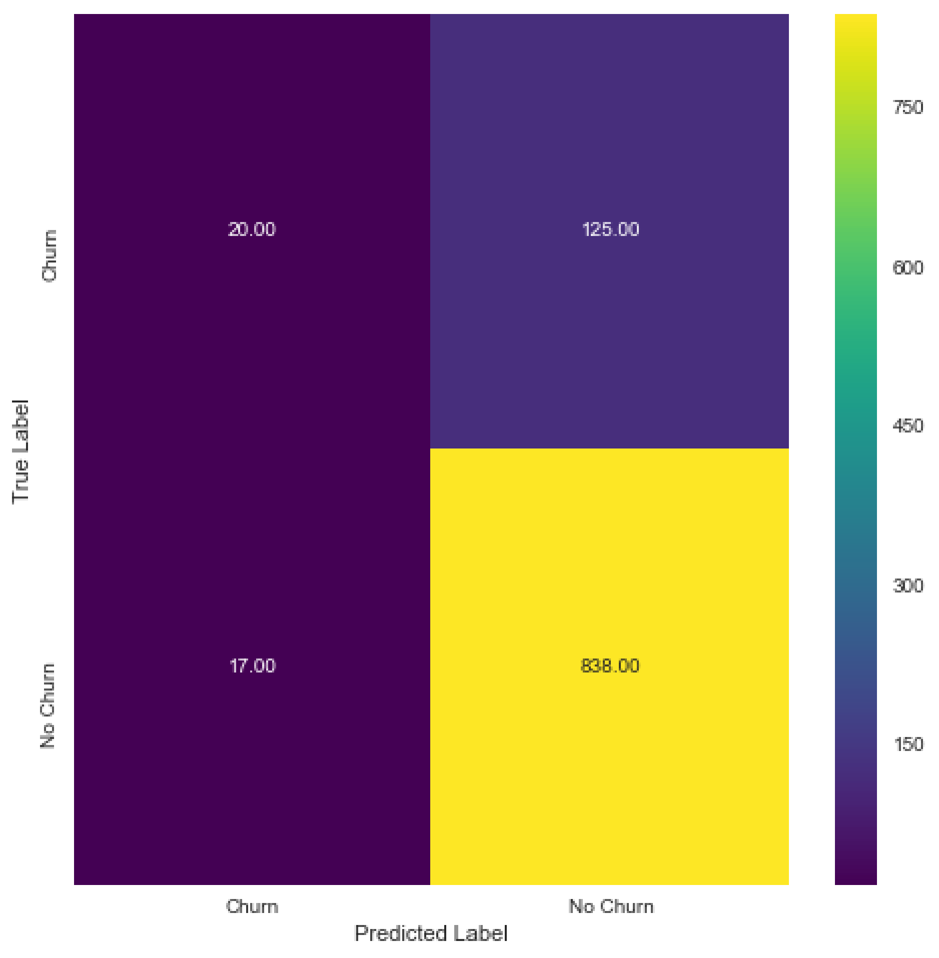
Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Confusion Matrix

- Accuracy may not be the right measure at times, especially if your Target class is not balanced.

- Specificity is high and sensitivity is low, primality driven by threshold value we have choose.

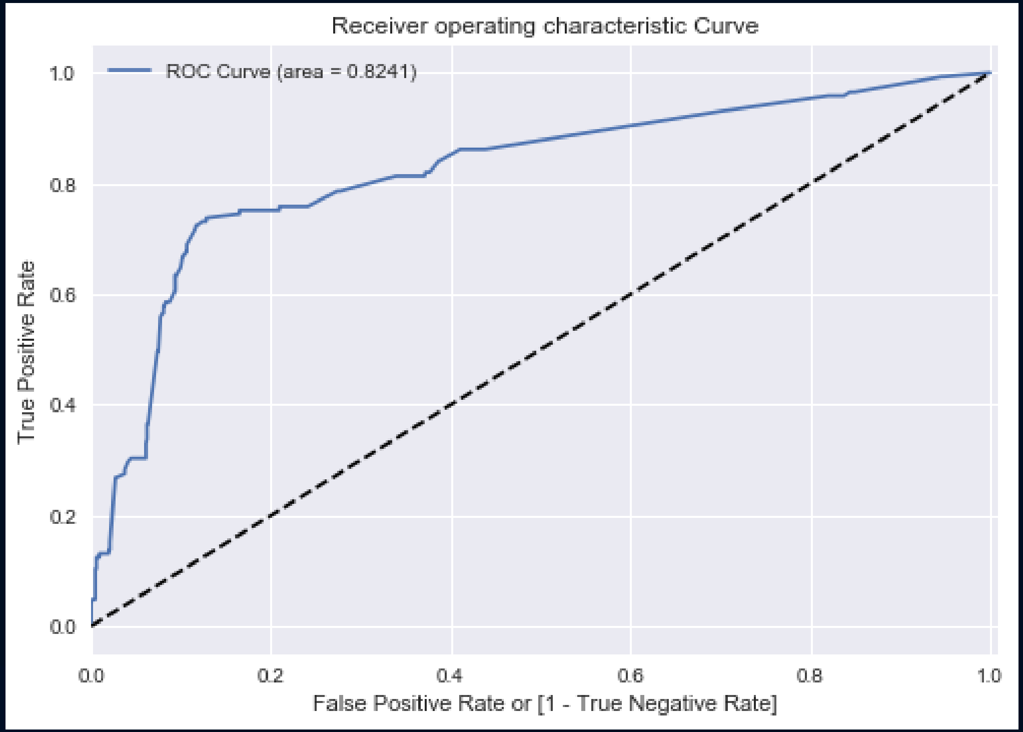
- Precision = 82%
- Recall = 86%



Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Receiver Operating Characteristic curve

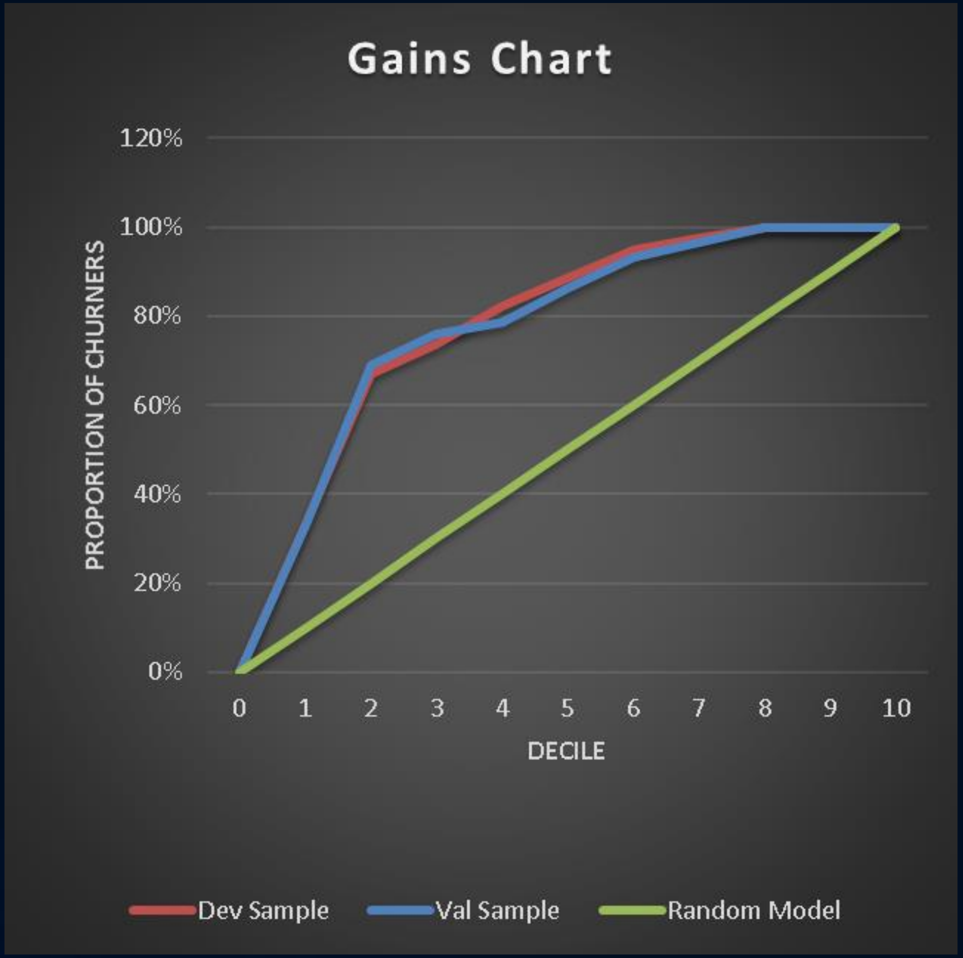
- The plot of 'True Positive Rate' (Sensitivity/Recall) against the 'False Positive Rate' (1-Specificity) at different classification thresholds.
- The area under the ROC curve (AUC) measures the entire two-dimensional area underneath the curve.
- It is a measure of how well a parameter can distinguish between two diagnostic groups.



Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Gain Chart

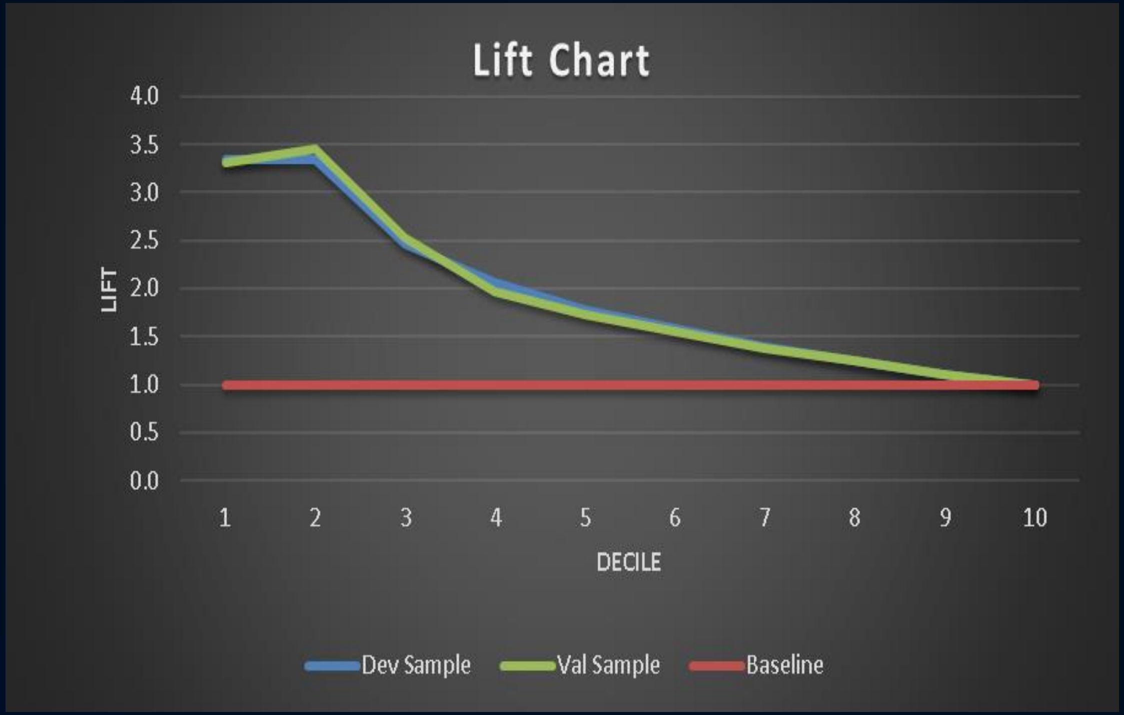
- Gain chart tells you how well is your model segregating responders from non-responders.
- The 'random' line in the corresponds to the case of capturing the responders ('Ones') by random selection.
- By targeting the first 40% of the data, the model will be able to capture 82.2% of the churners.



Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Lift Chart

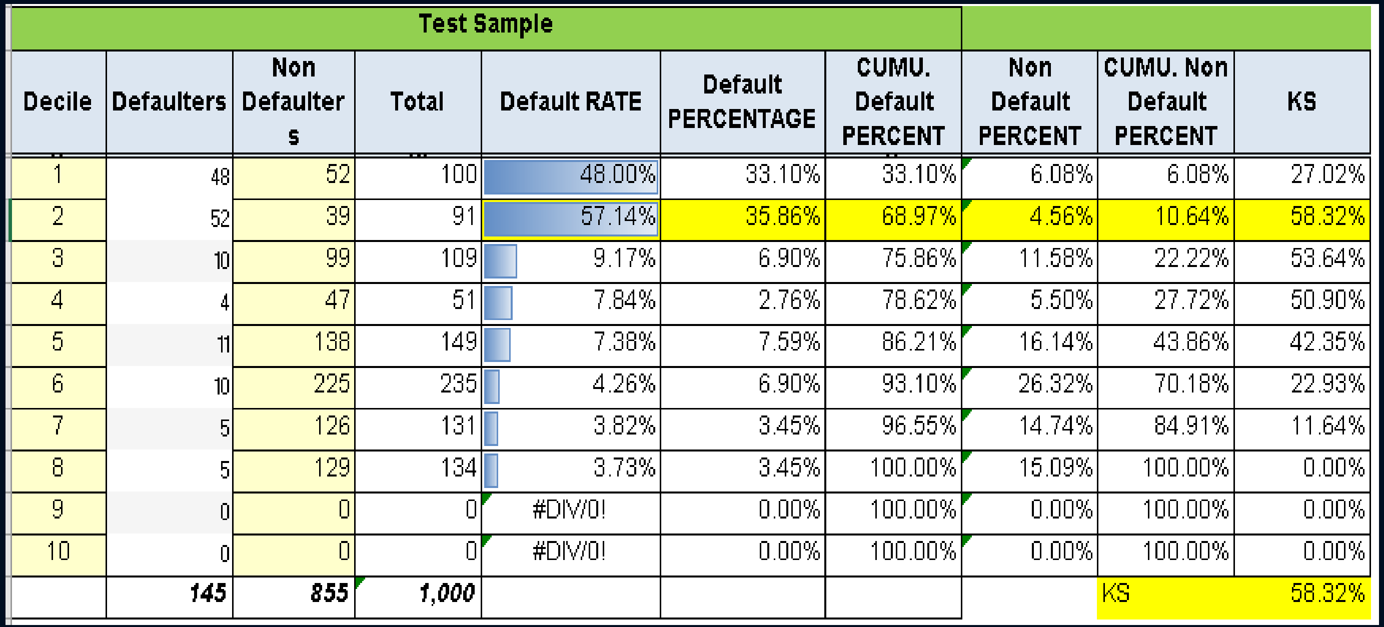
- Lift chart compares the response rates with and without using the classification model.
- It is used to evaluate the usefulness of the model.
- The Cumulative Lift for top two deciles is 3.45.
- we can expect 3.45 times the total number of targets (churners) to be found than by randomly selecting 20% of the data without a model.



Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Kolmogorov Smirnov chart

- K-S is a measure of the degree of separation between the positive and negative distributions.
- If the K-S is 100, model is able to separate groups in which one group contains all the positives and the other all the negatives.
- KS is maximum at second decile and KS score is 58.32%.



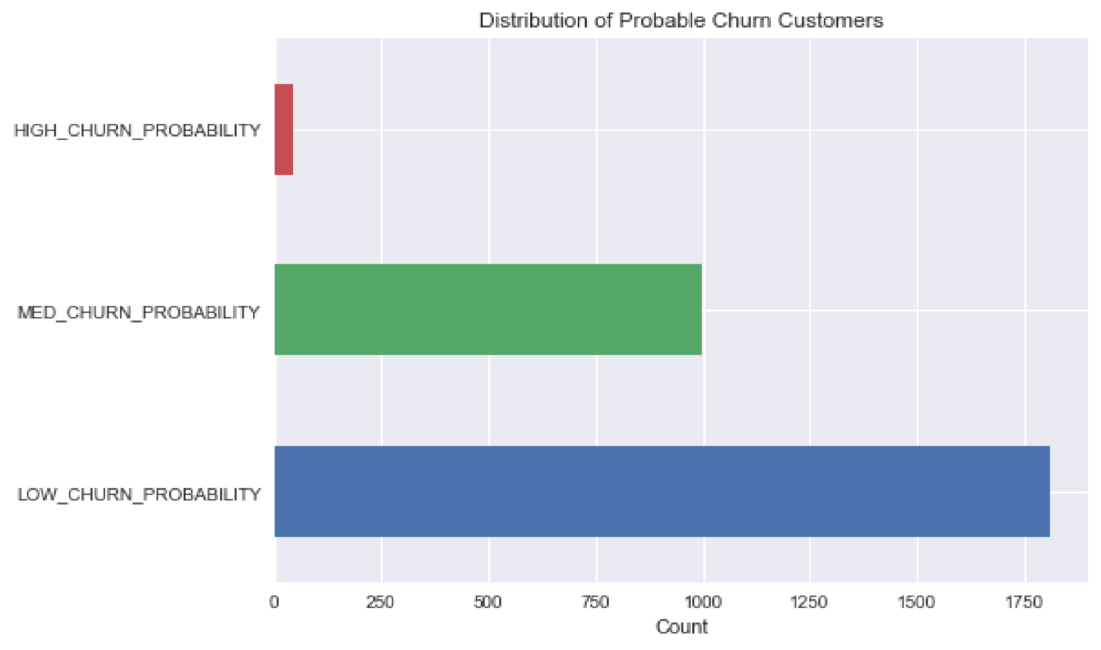
Test Sample									
Decile	Defaulters	Non Defaulters	Total	Default RATE	Default PERCENTAGE	CUMU. Default PERCENT	Non Default PERCENT	CUMU. Non Default PERCENT	KS
1	48	52	100	48.00%	33.10%	33.10%	6.08%	6.08%	27.02%
2	52	39	91	57.14%	35.86%	68.97%	4.56%	10.64%	58.32%
3	10	99	109	9.17%	6.90%	75.86%	11.58%	22.22%	53.64%
4	4	47	51	7.84%	2.76%	78.62%	5.50%	27.72%	50.90%
5	11	138	149	7.38%	7.59%	86.21%	16.14%	43.86%	42.35%
6	10	225	235	4.26%	6.90%	93.10%	26.32%	70.18%	22.93%
7	5	126	131	3.82%	3.45%	96.55%	14.74%	84.91%	11.64%
8	5	129	134	3.73%	3.45%	100.00%	15.09%	100.00%	0.00%
9	0	0	0	#DIV/0!	0.00%	100.00%	0.00%	100.00%	0.00%
10	0	0	0	#DIV/0!	0.00%	100.00%	0.00%	100.00%	0.00%
	145	**855**	**1,000**					KS	58.32%

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Actionable Business Insights

- Create Segments of customers based on probability of churn.
  - 0 – 0.4: Low churn probability
  - >0.4 – 0.5: Medium churn probability
  - >0.5: High churn probability

- The threshold value of greater than 0.5 was chosen because the same value is used by the Adaboost model for classifying churners and non-churners.



Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Actionable Business Insights

- For customers with High Churn probabilities will require an immediate call to understand their grievances/complains.
- May be a new relationship manager will be helpful to understand their choices in terms of payment, language and complaints.
- For customers with Medium Churn probabilities may require souvenirs for their time with the service provider.
- Keep in touch and do monitor their grievances and complains for better contribution

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Actionable Business Insights

- By targeting the customers who are present in the 40% of the data, we will be able to prevent 82% of churners.

- Improve the customer service experience, so that the number of calls to the customer care will be reduced. May be introduce 24/7 chat bot to deal with customer complaints without the hassle of calling.

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

Further Improvements

- Once the assessment of the model was completed i.e.. the last cycle in SEMMA framework. We can again go back to the Sample cycle, where we can get better sampling data.
- If we can get feedback given by the customers for the service provider, we could have come up with better model.
- Feedback at various levels like:
  - Feedback on network coverage
  - Feedback on call drops
  - Feedback on customer service
  - Feedback on Internet Coverage

Refer to github link for code: https://github.com/Niranjankumar-c/TelecomChurn

THANK YOU