Churn Prediction Modeling

The churn model has been built for TELECOM, a major telecom provider in the middle-east. This model would help company to predict the customers who are most probable to abandon the services of the company in the case of both PREPAID and POSTPAID using the concepts of LOGISTIC REGRESSION

By Team F :

Amar Preet

Ishan Bagadiya

Ragini rathi

Sunny Chopra

EXECUTIVE SUMMARY:

The term churn is used to describe the customer attrition. Churn rate is the number of customers who have left the company divided by the total number of customers. All major business faces the problem of churn rate. A successful business model develops strategies to reduce the churn rate and try to retain as many customers as possible because it is very expensive to acquire a new customer than to keep your existing one from leaving .There is a huge cost involved to nullify the effect of a churning customers. In our project we are provided with the task of analyzing the data of a telecommunication firm who is facing a high churn rate of 15%. The data is provided by CrowdAnalytix to us and as a Data science consultants our task will be to provide recommendations and strategies to reduce the churn.

Why Churn is a problem?

Most of the telecom companies suffer from voluntary churn rather than involuntary churn. Churn rate has very strong impact on the life time of the value of customers. It is important because it significantly affects the future of the organization and the length of service. For example, if a company has a customer churn rate of 50%, then the average lifetime of its customers is 2 years. It is estimated that approximately 75 percent of the 17 to 20 million subscribers signing up with a new wireless carrier are coming from another wireless service provider. This clearly states that these customers had churned. It has been observed that Telecom companies are spending hundreds of dollars to acquire a new customer. But when that customer leaves, the company not only face decline in revenue but also end up losing all the resources it spend on the acquisition of that customer. So, churn leads to decline profitability of telecom organization.

Following are the steps taken by telecom companies so far to reduce Churn:

Telecom companies are using two approaches to address churn:

(a) **Untargeted Approach**: The untargeted approach relies on mass advertising and superior products to increase brand loyalty and thus retain customers.

(b) **Targeted Approach**: The targeted approach is to identify customers who are likely to churn. The next step is to design offers that will encourage customers to stay.

Role of Data Science: Data Science plays crucial role in identifying the customers who are likely to churn. The companies now a days want to identify in advance the customers who will churn. This timely analysis will help companies to come up with customer focused campaigns that will target customers with special offers. This process can bring in huge loss for the company in case the predictions are inaccurate. This is because if company is not targeting the right set of customers then it will result in giving away special offers to people who are not churning. As a result, the company will end up wasting its money.

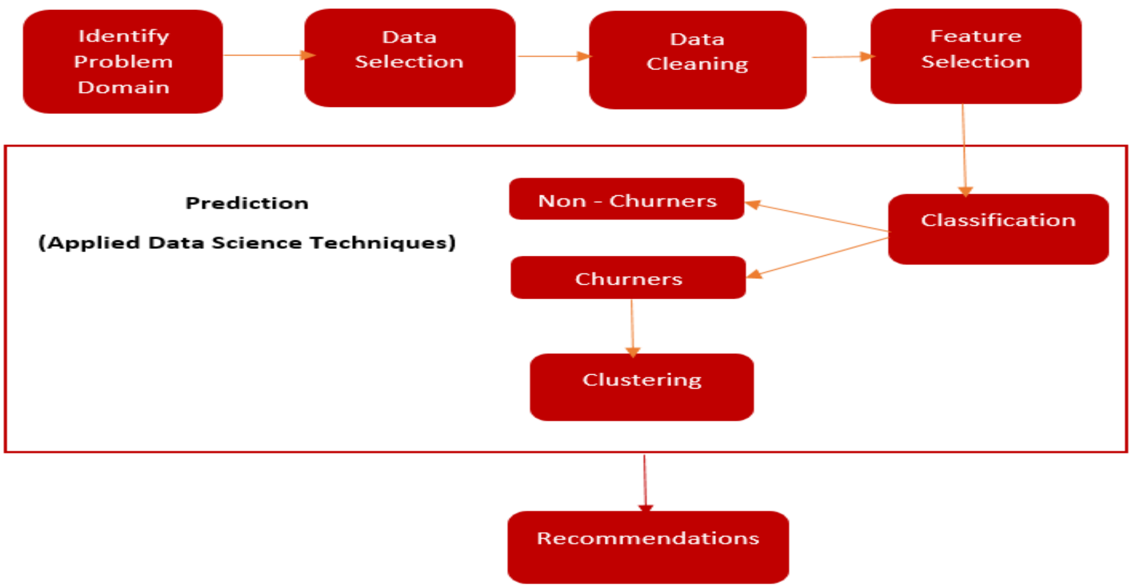
There are various predictive modeling technique for predicting customer churn such as Logistic regression, Random forest, Support Vector Machines. There are also various feature selection techniques such as information gain, stepwise regression that are being used to select the most important features.

Dataset: Our dataset has 3333 instances. The churn is a predictor column. The predictor column contains 483 positive labels that are the customers who are churning and 2850 negative labels that are the customers who are not churning. Our dataset consists of 21 attributes out of which 15 attributes are numeric, 3 attributes are categorical and 3 variables are binary. Please refer the appendix for detailed variable names.

Problem Statement: The task is to build predictive model on the given data set to predict whether in the future the customer will churn or not. But the goal of our project is not predictive modeling, rather predictive modeling is the means to reach our goal. The major task here is to reduce churn by identifying the important features that lead to churn. And then come up with strategies and business models to achieve our goal. Also we will identify what extra information of the customer can be useful to perform better analysis and predictions.

Ref: https://www.crowdanalytix.com/contests/why-customer-churn/

Approach and plan of work:



Class Imbalance Problem:

Our data set contains only 15% of true labels. Based on the research and analysis done to avoid over fitting we decided to use Over Sampling using SMOT algorithm.

Our next approach will be to use Information Gain, Binary Search to determine the important features in our data. Then we will apply different classification models on the complete data as well as only on the important features. And we will compare the results obtained from them. Then we will perform customer segmentation on the churner population.

Based on this results we will provide a business model to the company to help them reduce churn rate.

Risks and Mitigations

Small set of TRUE data:

Our data contains only 15% of true data, which increases risk of over fitting. Even on using a default classifier we will be getting an accuracy of 85%.But our aim is to accurately predict the churn population.

Some important features missing:

In our data set we had 21 features. And our analysis was based on that set of features.

But it would have been very significant if we also had features like:

- Age.
- Profession.
- Owner/Rented Apartment.
- Internet Data Usage.
- Handling numeric and categorical data

Invalid classification of values based on decision boundaries.

Mitigations
Use of OverSampling.
Develop robust boundaries to ensure there is no error in classification.

Feature Selection:
PCA
Correlation
Information Gain:
Information Gain helps us determine the weight of every feature to determine the labels to be learned.It provides us the importance level of each attribute. That attribute will be used as the node factor for our decision tree.

Information Gain	Attribute #	Attribute
0.09806	17	CustServ Calls
0.03688	2	Intl Plan
0.0157	7	Day Charge
0.0157	5	Day Mins
0.01236	15	Intl Calls
0.00822	3	VMail Plan
0.00822	4	VMail Message
0.00686	12	Night Calls
0.00642	9	Eve Calls
0.00533	6	Day Calls
0	13	Night Charge
0	8	Eve Mins
0	14	Intl Mins
0	11	Night Mins
0	16	Intl Charge
0	10	Eve Charge
0	1	Account Length

Above is the rankings of the features based on the information gain.
The top features are Customer service calls,International plan and Day charge.And on doing analysis of the data we can see significant difference in these features for churners and non churners.Also it has to be noted that features like Night Mins, Intl Charge,Account Length have zero information gain.This is also due to the reason that there eists high correlation between these features and the features having high information gain.

Correlation:
Binary Tree:

We performed Binary Tree model on the complete data set,
The root node we obtained using binary tree was Day Minutes.The tree was split on this .
The important features obtained using Binary Tree were :
CustServ.Calls Day.Mins Eve.Mins Int.l.Plan Intl.Calls Intl.Mins
VMail.Plan
The confusion matrix we obtained from binary tree:
INCLUDE Matrix

Binary Tree Rules:
The most important rule obtained was:
Rule [Churn.=False. cover=1546 (66%) prob=0.02]
Day.Mins< 245.6
CustServ.Calls< 3.5
Int.l.Plan=no
Day.Mins< 221.9

Classification Models:

The output column of the data set, 'Churn', is a binary attribute having the value True or False. We applied Naïve Bayes, Bayes Net, AD Tree, Support Vector Machine, Logistic Regression and Random Forest to test the accuracy.

Naive Bayes - It is a classification algorithm based on Bayes rule that assumes conditional independence between the inputs attributes, given output. This assumption simplifies the representation of P(X/Y), and the problem of estimating it from training data. Naïve Bayes Classifiers can be built with real-valued inputs.

Bayes Net - A Bayesian network is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

Formally, Bayesian networks are DAGs whose nodes represent random variables that may be observable quantities, latent variables, unknown parameters or hypotheses. Edges represent conditional dependencies; nodes that are not connected represent variables that are conditionally independent of each other. Each node is associated with a probability function that takes, as input, a particular set of values for the node's parent variables and gives the probability of the variable represented by the node. For example, if m parent nodes represent m Boolean variables then the probability function could be represented by a table of 2^m entries, one entry for each of the 2^m possible combinations of its parents being true or false.

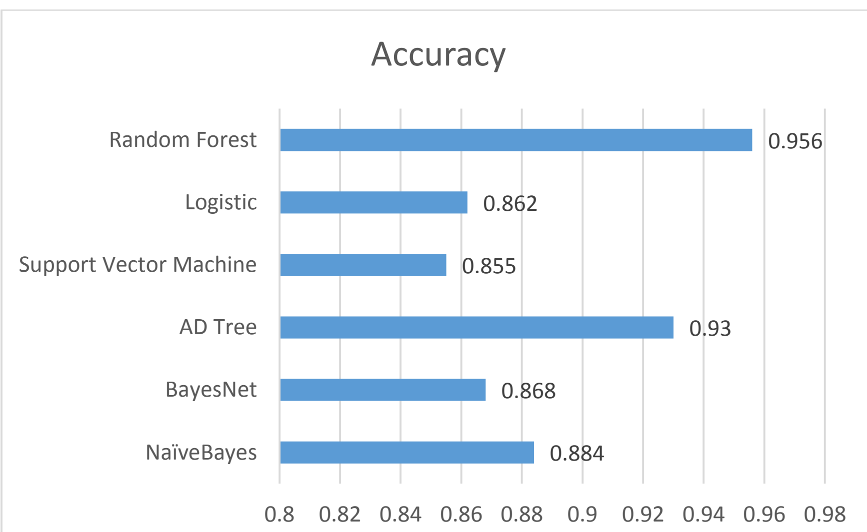
AD Tree - An alternating decision tree (AD Tree) is a machine learning method for classification. It generalizes decision trees and has connections to boosting. An alternating decision tree consists of decision nodes and prediction nodes. Decision nodes specify a predicate condition. Prediction nodes contain a single number. AD Trees always have prediction nodes as both root and leaves. An instance is classified by an AD Tree by following all paths for which all decision nodes are true and summing any prediction nodes that are traversed. This is different from binary classification trees such as CART (Classification and regression tree) or C4.5 in which an instance follows only one path through the tree.

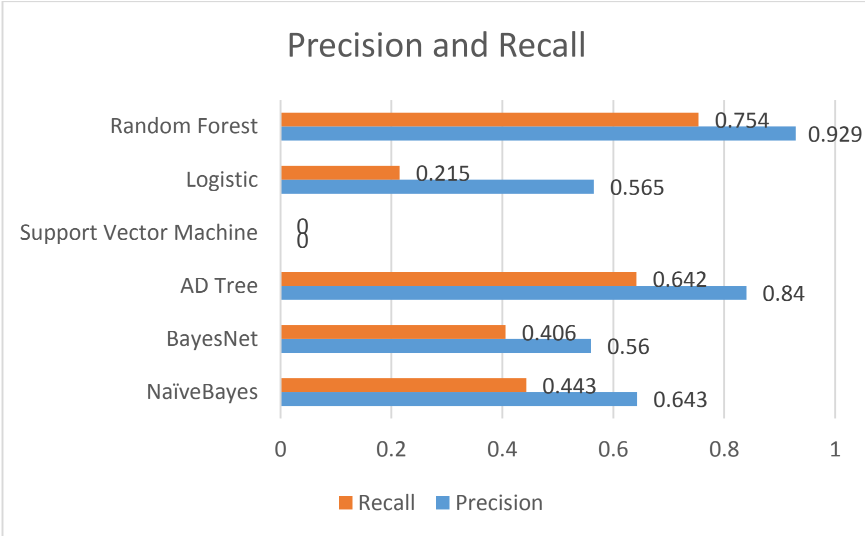
Logistic regression - Logistic regression, or logit regression, or logit model is a direct probability model. The binary logistic model is used to predict a binary response based on one or more predictor variables. That is, it is used in estimating the parameters of a qualitative response model. The probabilities describing the possible outcomes of a single trial are modeled, as a function of the explanatory (predictor) variables, using a logistic function.

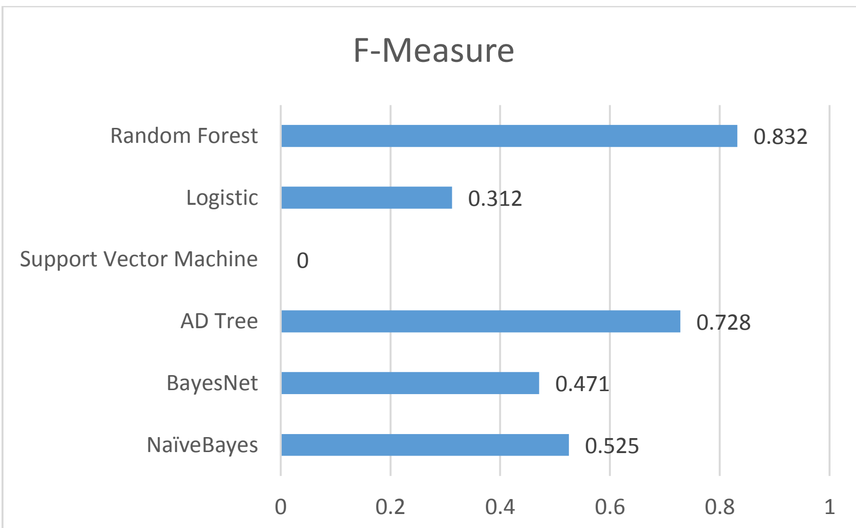
Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables, which are usually (but not necessarily) continuous, by estimating probabilities.

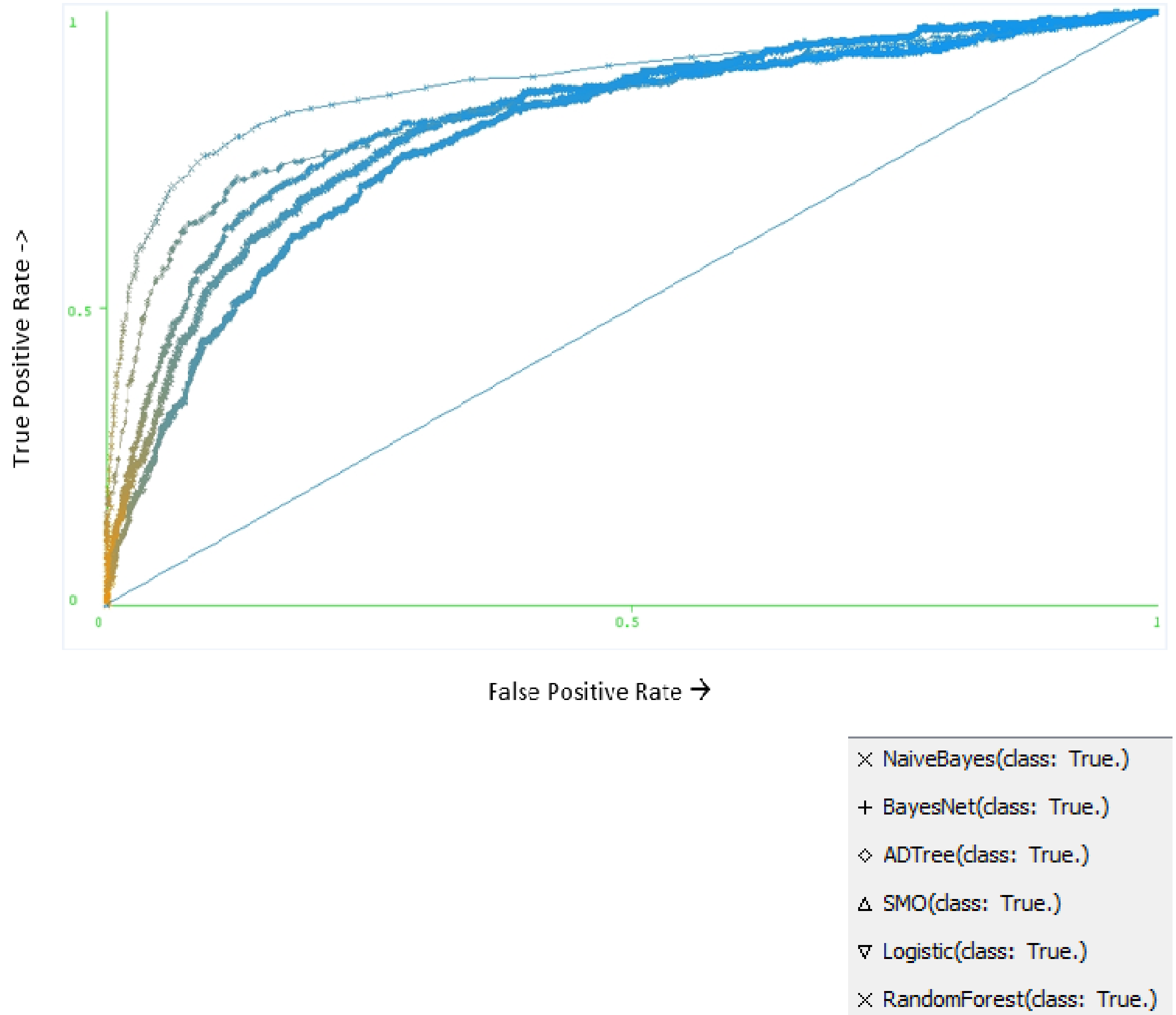
Random Forest - Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests correct for decision trees' habit of over-fitting to their training set.

We performed 10-Fold and 4-Fold cross-validation by taking all the features into account and taking only selected features. For the class imbalance problem, we performed oversampling so that the number of records to each class becomes roughly equal. The oversampling was performed using the SMOTE (Synthetic Minority Over-Sampling Technique) which generates the minority samples based on nearest neighbor algorithm. For cross-validation after oversampling, the test set was kept separate and only the training set was oversampled so that result is not over fitted. The performance measures are accuracy and F-measure values for each of the entries. Below are some of the results we obtained after doing 10 fold cross-validation.




Precision and Recall
Random Forest
Logistic
Support Vector Machine
AD Tree
BayesNet
NaïveBayes
0.754
0.929
0.215
0.565
0
0
0.642
0.84
0.406
0.56
0.443
0.643
0
0.2
0.4
0.6
0.8
1
Recall
Precision


F-Measure
Random Forest
Logistic
Support Vector Machine
AD Tree
BayesNet
NaïveBayes
0.832
0.312
0
0.728
0.471
0.525
0
0.2
0.4
0.6
0.8
1



Based on the metrics like accuracy, AUC and F-measure Random Forest is the best model. We achieved accuracy an accuracy of 95.6% with Random Forests. The next best model was AD Tree where we achieved an accuracy of 93%.

Information from the features

How can we reduce churn rate?

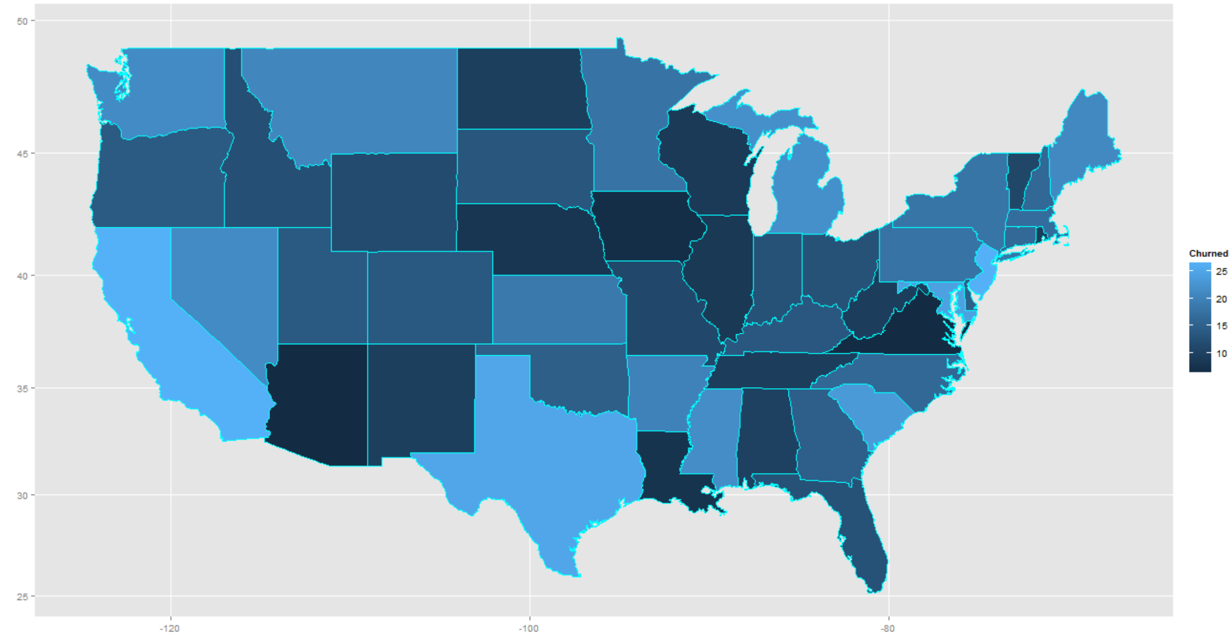
Customer Segmentation:

Our data included information of customers from all the 50 states from the U.S.

We considered the state feature as an important factor to see the churn rate of the customers.The number of people in each state highly varied , so instead of looking at the number of customers who are churners we did our analysis on the percentage of churners for each state.

Based on the results obtained we saw that in the states like Texas and California the churn rate is 25 ,21 percent respectively.While the churn rate for states like Hawaii,IOWA the churn rate is as low as 5-

6%.This factor is a distinguishing one.



As expected in States like California and Taxes the competition is high and thus is the churn rate.While in the states like Pennsylvania ,Florida where there is a significant amount of retired population the churn rate is low.

Based on these our recommendation will be that while planning their strategy,the company should keep into account the states ,the type of customers their and the competition in the state.

Recommendations to reduce churn rate:

Prototype based on recommended call rates:

While classifying the customer as a churner or non-churner we identified that Day charge and Night charge as important features. From the data we calculated the (Call Rates/Min) for day ,evening and night charges.

The call rates were the same for each and every customer irrespective of his account length or the number of minutes he talked during a particular period of that day.

We obtained the following result:

Time	Mean Minutes	Mean Minutes for Churners	Rates/Min	Suggested Rates/Min
Day	180	242	5.88	5.12
Evening	200	188	11.76	12.7
Night	210	265	22.22	21.5

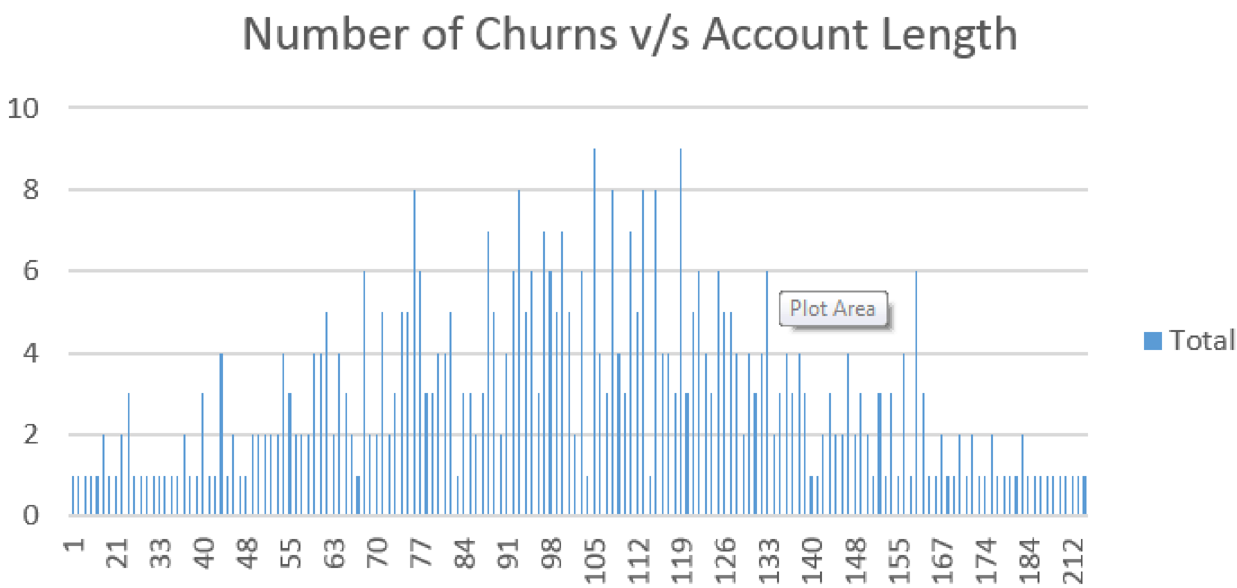
According to the result the churner population had a much higher mean minutes for day and night relative to the non-churners. But they also had to pay the same charge. They didn't have any advantage of being a high calling customer and was a major reason for their churning.

Thus we suggest the company to provide them with alternative rates, as suggested in the table above.

We also suggest that the company can increase night charge for this set of customers as there calling minutes are less during this period. And thus company can also compensaate for the revenue loss due to the decrease in day and night charges.

Result: When we used the suggested rates as our train data,we were able to reduce the churn rate from 15% to 11.5 %.

Relation with Account Length:



Account Length is the number of days the customer is part of the network.

From the graph above we can see that the number of churners increases when the account length ranges from 91 to 119.This provides a very useful insight.

A customer spents atleast 2 months with the company.And then his probability to churn increases,as he is not satisfied with the services of the firm.And on the other hand if the company is able to retain the customer for around 120-130 days i.e 4 months ,the chances of churn reduces significantly.

Thus the company must be very critical about the customers that falls between 2-4 months range.And must come up with strategies and offers to retain them .

Conclusions & Further Study:

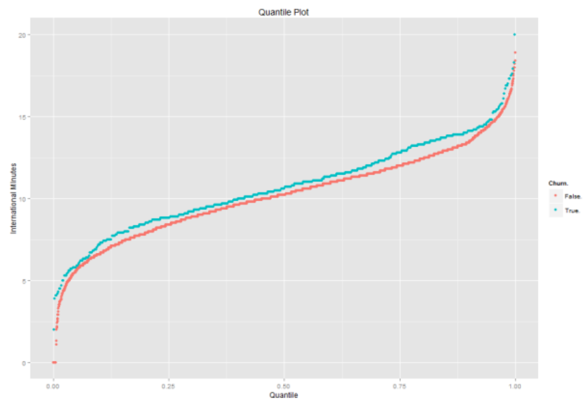
Making use of data science and machine learning algorithms for performing oversampling,feature selection ,classification and clustering we were able to identify the important features and come up with a model to predict churn.We also performed customer segmentation based on the demography.And came up with recommendations based on that.

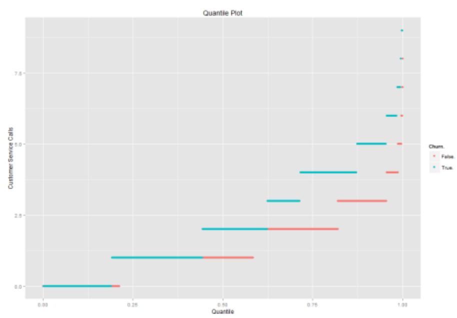
We were also successful in coming up with ideas and prototype to reduce churn.

In future we will try to collect more information on the data.

Try to achieve better scores

Appendices: Graphs Maps:





Correlation churn.csv using Pearson

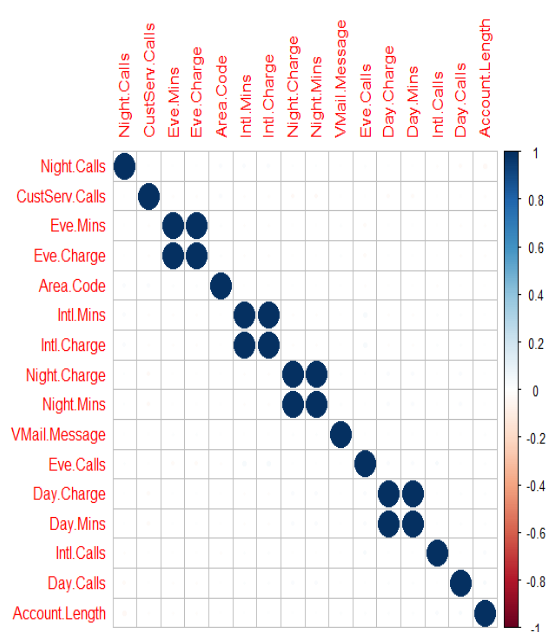
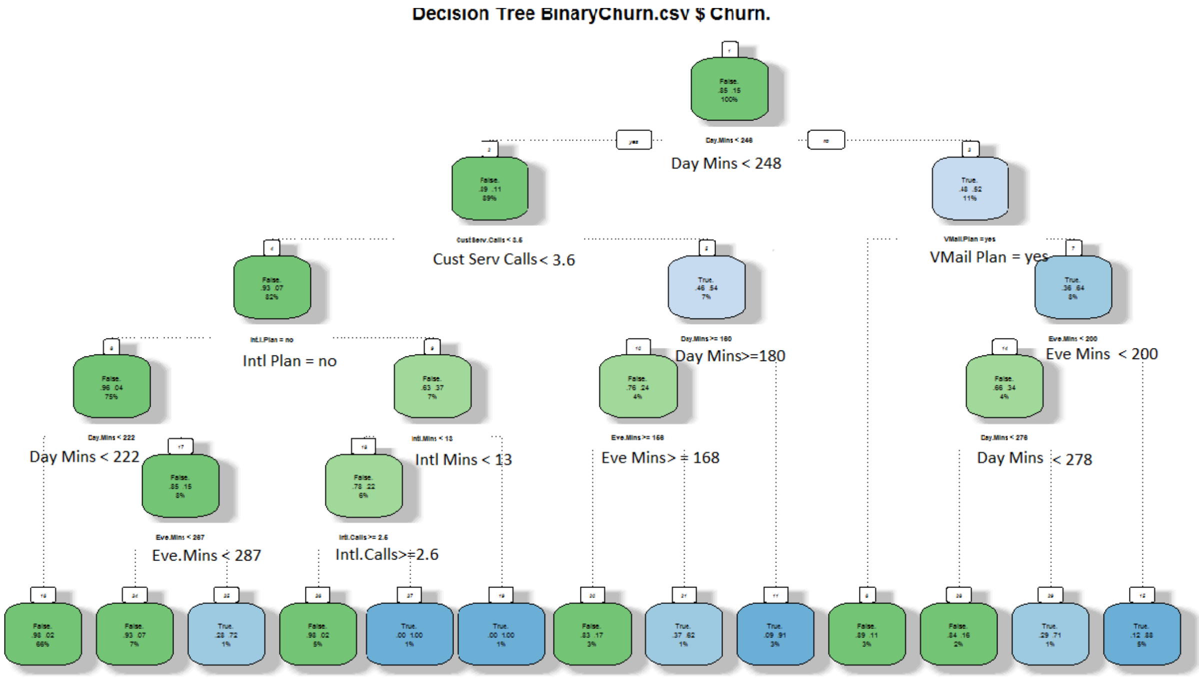


Fig: 1



Binary Decision Tree

Fig: 2