CHURN PREDICTION OF POSTPAID TELECOM CUSTOMERS

DSA 5103

Instructor: Dr. Charles Nicholson

Group #10

Maitrik Das (*Maitrik.das-1@ou.edu)*

Raj Saha *(raj.saha-1@ou.edu)*

Kaustubh Pande *(kaustubhpande@ou.edu)*

Rahul Bharadwaj *(Rahul.v.bharadwaj@ou.edu)*

Executive Summary

Concise problem statement

The objective of the problem is to perform churn prediction for the customers of a telecom company in the USA. This problem consists of the postpaid customers who are from 19 different segmented regions of the country and the variables describe numerous telecom sector characteristics such as different consumer call patterns, different revenues of the telecom sector for each consumer and also their information inclusive metadata such as credit status, marital status, income, types of phone model, handset prices etc. The churn is the actual target variable which is in binary form 0 and 1 which means if the customer churns it is shown as 1 and if the customer doesn't churn it is defined as 0.

List of Major Concerns /Assumptions

- From features like average consumer calls or revenue patterns it is assumed that these customers are mainly postpaid although this is not clearly mentioned in the problem statement.
- For features like credit status and marital status some labels are not clearly mentioned in the problem statement although they can be assumed as specific labels for the necessary of the solution without impacting in the form of noise to the whole data.

Summary Of Findings

- Most of the consumer call patterns such as the number of calls and the time of their conversations are positively correlated.
- Numerous telecom revenues and consumer call patterns are also positively correlated.
- There is a noticeable difference in behavior between churned and unchurned consumers for particular churn detective telecom features as the number of dropped calls, the number of customer care calls, or the overall conversation duration to customer care.
- For churn detection, personal information of the consumers such as credit card status or marital status are also deemed as important decisive factors.
- A number of features relating to customer calls are engineered before predictive modeling.
- Tree based models result in better outcomes over other linear models as this data is infested with correlations, outliers and skews.
- The length of time a customer utilizes the service is seen to be the most significant factor in determining the fundamental forecasts regarding churn, according to some main feature interpretation following the predictive modeling.
- This predominantly depicts that old customers mostly tend to stay with the telecom services, they are not likely to churn but the new customers tend to churn significantly.

Recommendations

- For the problem perspective it is recommended that special attention should be directed towards churned customers for better user experience, so that they won't face issues like dropped calls and consequently their number of calls or total minutes of calling to customer care won't get rambled up over the unchurned ones.
- Analysis shows that credit status is important for churned customers so it is highly recommended to choose new customers irrespective of their credit status whereas emphasis must be given to those who are having credit card.

Problem background

Problem description

Customer churn or customer attrition is the phenomenon where customers of a business no longer purchase or interact with the business. A high churn means that a higher number of customers no longer want to purchase goods and services from the business. Customer churn can prove to be a roadblock for an exponentially growing organization and a retention strategy should be decided to avoid an increase in customer churn rates. The objective of the problem is to perform churn prediction for the customers of a telecom company in the USA, This problem consists of post-paid customers who are from nineteen different segmented regions of the country and the variables describe numerous telecom sector characteristics such as different consumer call pattern, different revenues of the telecom sector for consumers, and their metadata. This dataset consists of many columns broadly classified into revenue, consumer call patterns and usages, and metadata such as ethnicity, credit status, marital status, handset used, etc. Revenue includes columns such as mean monthly revenue, total revenue generated by a customer, change in revenue for the past month, etc. Usage consists of calls, data, and SMS use patterns such as the total number of calls placed over the lifetime of the customer, Average monthly minutes of use over the previous three months, etc.
In the current era, If a company has to grow at a study pace, it has to invest in acquiring new customers. Every time a customer leaves, it represents a significant investment loss. Both time and effort need to be channelled into replacing them, and research has proven that acquiring a new customer is difficult compared to retaining one. Being able to predict when a client is likely to leave, and offer them incentives to stay, can offer huge savings to a business. Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are low.

Context

Telecom Churn (loss of customers to competition) is a problem for telecom companies because it is expensive to acquire a new customer and companies want to retain their existing customers. Most telecom companies suffer from voluntary churn. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers. Existing customers will often have a higher volume of service consumption and can generate additional customer referrals. For example, Netflix uses churn prediction to target customer retention by sending email recommendations to target customers.

Background

Churn rate (also known as attrition sometimes), in its broadest sense, is a measure of the number of individuals or items moving out of a collective group over a specific period. It is one of two primary factors that determine the steady state level of customers a business will support. Derived from the butter churn, the term is used in many contexts but is most widely applied in business with respect to a contractual customer base. Examples include a subscriber-based service model as used by mobile telephone networks and pay TV operators. The term is often synonymous with turnover, for example, participant turnover in peer-to-peer networks. Churn rate is input into customer lifetime value modelling, and can be part of a simulator used to measure return on marketing investment using marketing mix modelling. Churn rate, when applied to a customer base, refers to the proportion of contractual customers or subscribers who leave a supplier during a given time period. It is a possible indicator of customer dissatisfaction, cheaper and/or better offers from the competition, more successful sales and/or marketing by the competition, or reasons having to do with the customer life cycle.

Data description

The data set consists of 100 variables and approx. 100 thousand records. Out of which 79 are numeric variables and 21 are categorical variables. This data set contains different variables explaining the attributes of the telecom industry and various factors considered important while dealing with customers of the telecom industry.

Numerical variables:

rev_mean - Mean monthly revenue
mou_Mean - Mean number of monthly minutes of use
totmrc_Mean - Mean total monthly recurring charge
da_Mean - Mean number of directory assisted calls
ovrmou_Mean - Mean overage minutes of use
ovrrev_Mean - Mean overage revenue
roam_Mean - Mean number of roaming calls
change_mou - Percentage change in monthly minutes of use vs previous three-month average
change_rev - Percentage change in monthly revenue vs previous three month average
drop_vce_Mean - Mean number of dropped voice calls
drop_dat_Mean - Mean number of dropped data calls
recv_vce_Mean - Mean number of received voice calls
recv_sms_Mean - Mean number of received sms
comp_vce_Mean - Mean number of completed voice calls
comp_dat_Mean - Mean number of completed data calls
custcare_Mean - Mean number of customer care calls
ccrndmou_Mean - Mean rounded minutes of use of customer care calls
cc_mou_Mean - Mean unrounded minutes of use of customer care calls
inonemin_Mean - Mean number of inbound calls less than one minute
threeway_Mean - Mean number of three way calls
mou_cvce_Mean - Mean unrounded minutes of use of completed voice calls
mou_cdat_Mean - Mean unrounded minutes of use of completed data calls
mou_rvce_Mean - Mean unrounded minutes of use of received voice calls
owylis_vce_Mean - Mean number of outbound wireless to wireless voice calls
mouowylisv_Mean - Mean unrounded minutes of use of outbound wireless to wireless voice calls
iwylis_vce_Mean - Mean number of inbound wireless to wireless voice calls
mouiwylisv_Mean - Mean unrounded minutes of use of inbound wireless to wireless voice calls
peak_vce_Mean - Mean number of inbound and outbound peak voice calls
peak_dat_Mean - Mean number of peak data calls
mou_peav_Mean - Mean unrounded minutes of use of peak voice calls
mou_pead_Mean - Mean unrounded minutes of use of peak data calls
opk_vce_Mean - Mean number of off-peak voice calls
opk_dat_Mean - Mean number of off-peak data calls
mou_opkv_Mean - Mean unrounded minutes of use of off-peak voice calls
mou_opkd_Mean - Mean unrounded minutes of use of off-peak data calls
months - Total number of months in service
uniqsubs - Number of unique subscribers in the household
actvsubs - Number of active subscribers in household
totcalls - Total number of calls over the life of the customer
totmou - Total minutes of use over the life of the customer
totrev - Total revenue
adjrev - Billing adjusted total revenue over the life of the customer
adjmou - Billing adjusted total minutes of use over the life of the customer
adjqty - Billing adjusted total number of calls over the life of the customer

avgrev - Average monthly revenue over the life of the customer
avgmou - Average monthly minutes of use over the life of the customer
avgqty - Average monthly number of calls over the life of the customer
avg3mou - Average monthly minutes of use over the previous three months
avg3qty - Average monthly number of calls over the previous three months
avg3rev - Average monthly revenue over the previous three months
avg6mou - Average monthly minutes of use over the previous six months
avg6qty - Average monthly number of calls over the previous six months
avg6rev - Average monthly revenue over the previous six months
hnd_price - Current handset price
phones - Number of handsets issued
models - Number of models issued
Lor - Length of residence
Adults - Number of adults in household
Income - Estimated income
Numbcars - Known number of vehicles
eqpdays - Number of days (age) of current equipment

Categorical data:

Customer_ID - Customer_ID
churn - Instance of churn between 31-60 days after observation date
prizm_social_one - Social group letter only
area - Geogrpahic area
dualband - dualband
refurb_new - Handset: refurbished or new
rv - RV indicator
truck - Truck indicator
ownrent - Home owner/renter status
dwlltype - Dwelling Unit type
dwllsize - Dwelling size
Marital - Marital Status
Infobase - Estimated income
HHstatin - Premier household status indicator
ethnic - Ethnicity roll-up code
creditcd - Credit card indicator
kid0_2 - Child 0 - 2 years of age in household
Kid3_5 - Child 3 - 5 years of age in household
Kid6_10 - Child 6 - 10 years of age in household
Kid11_15 - Child 11 - 15 years of age in household
Kid16_17 - Child 16 - 17 years of age in household
new_cell - New cell phone user
hnd_webcap - Handset web capability
crclscod - Credit class code
asl_flag - Account spending limit

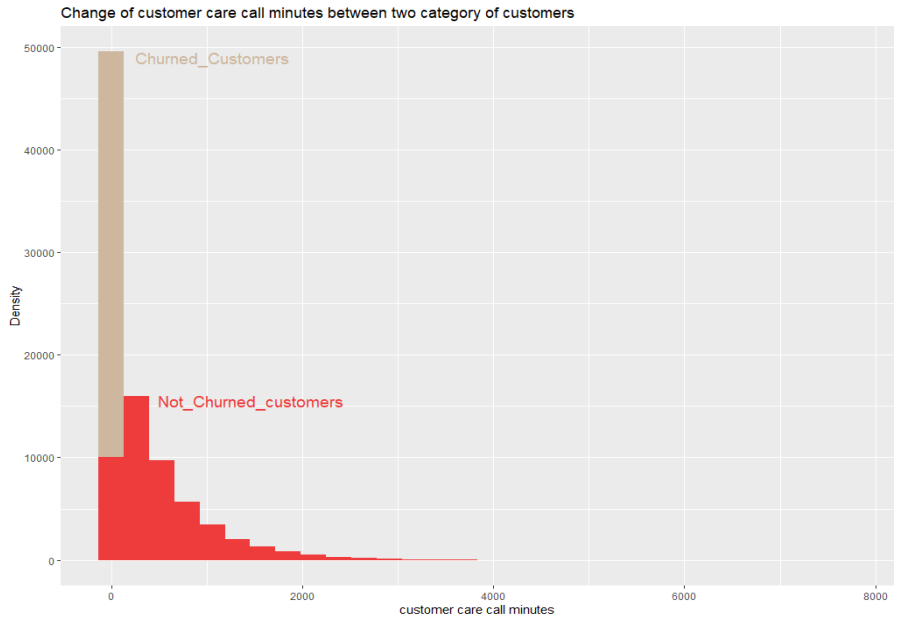
Exploratory data analysis

The roadmap of the whole data analysis lies into following paths.

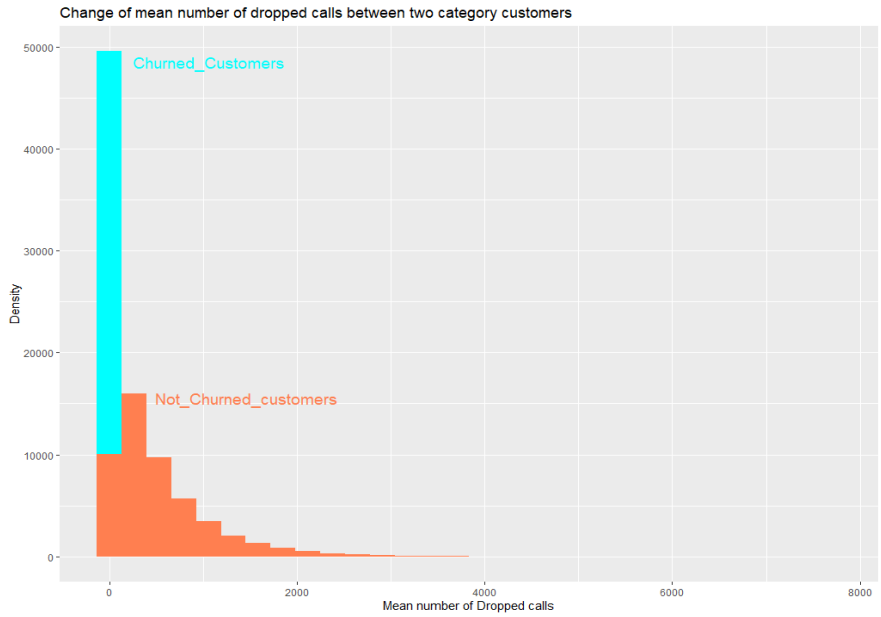
- **Drawing decent comparisons between two categories of customers (Churned and unchurned) for consumer call patterns.**
- **Showcasing inter correlations between numerous consumer call patterns and revenues.**

- **Analysis with customer metadata and showcasing if they may fit as proper decisive factors for predicting potential churn.**

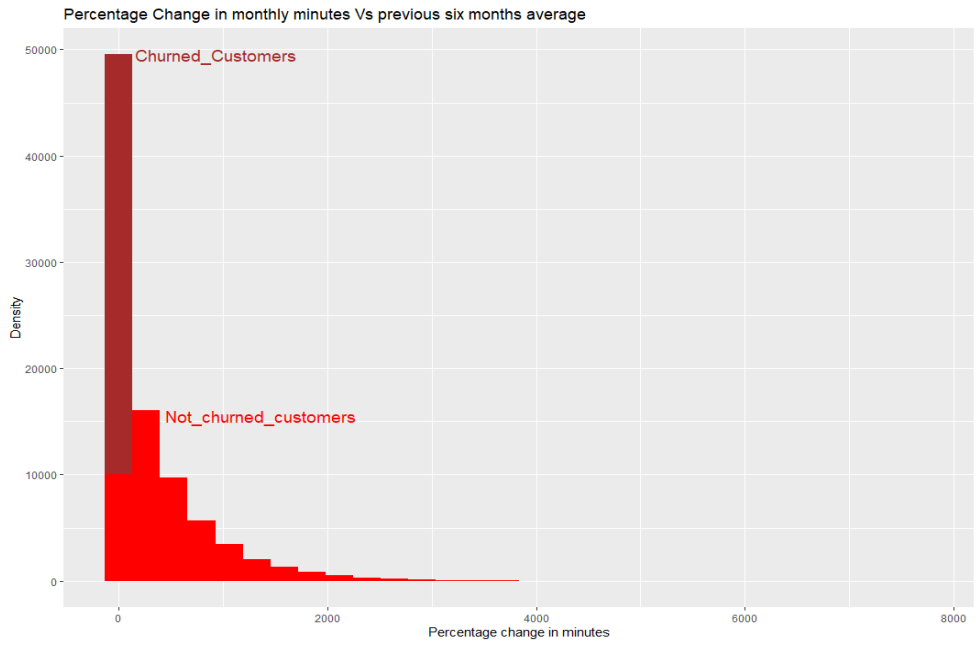
Some important features viz normal call patterns, dropped calls, peak hour calls, off-peak hour calls, customer care calls, less than one-minute calls, overage usage of minutes, last three months' change in calls, last six months change in calls are deemed to be significant and important in churn predictive analytics and comparative analysis are drawn between these features.



This plot demonstrates a clear discrimination between two categories of customers where pragmatically churned customers are ahead of those customers who haven't churned yet as churned customers face more issues, so they spend more time in calling customer care.

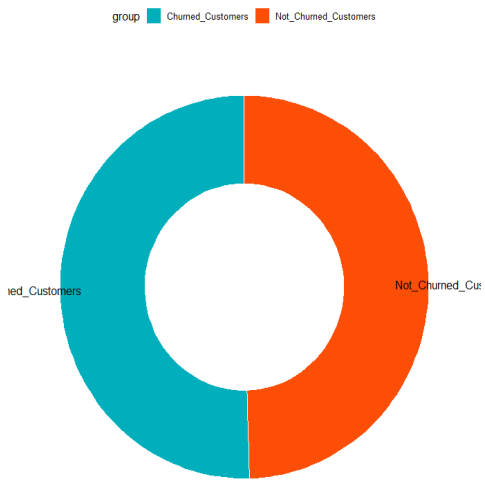


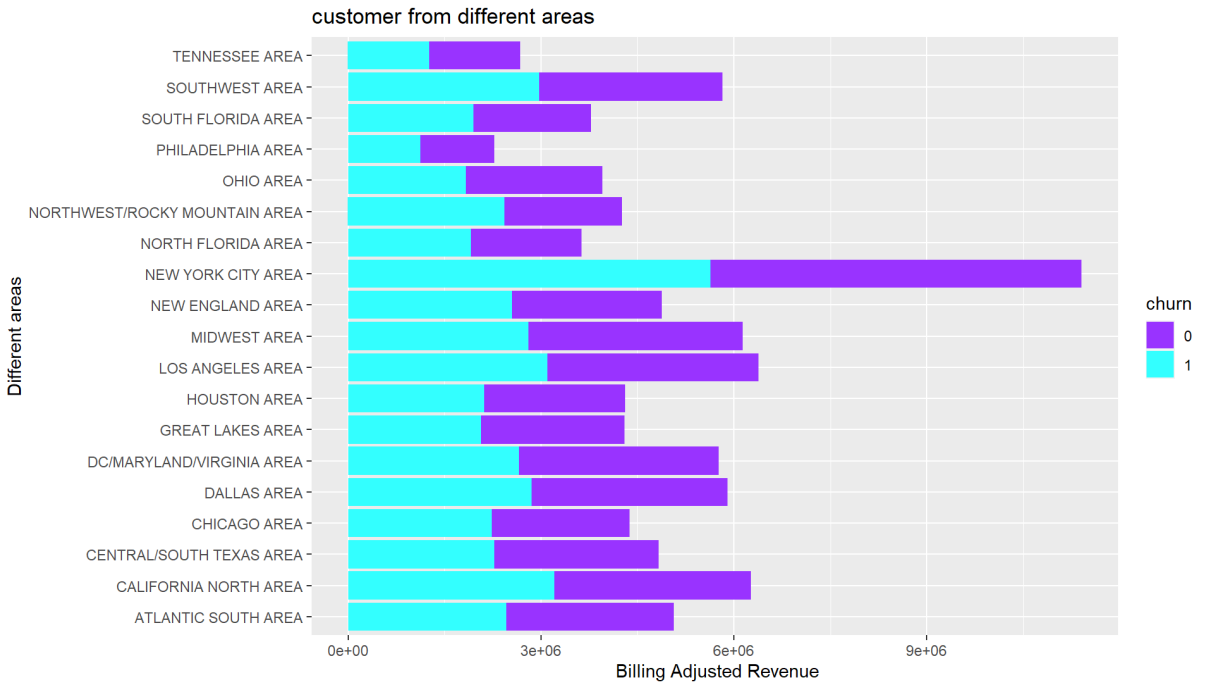
This visualization also makes justice into the fact that why churned customers are not happy with the company because according to this plot they face more dropped calls which is definite reason of making them annoyed.



This is an important visualization in terms of predicting churn because analysis shows that churned customers are having a good time initially but after some time their mentality gets changed as they get lured by attractive offers by other competitive companies so consequently their behaviours of using telephonic attributes get changed .This plot shows that churned customers show more changes in their present behaviours as compared to their last six months' behaviours over unchurned customers which has been really practical in churn analytics point of view .

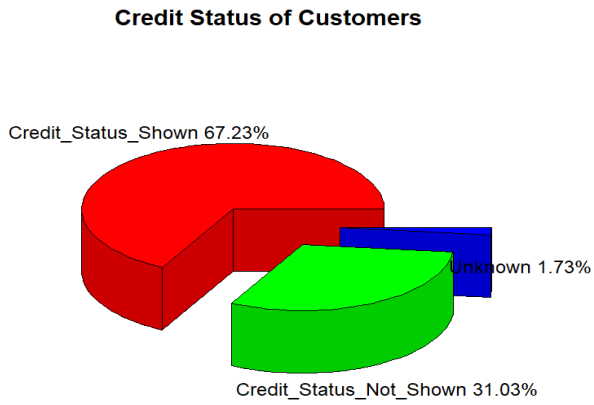
<u>Showcasing customers and their metadata inclusive information</u>



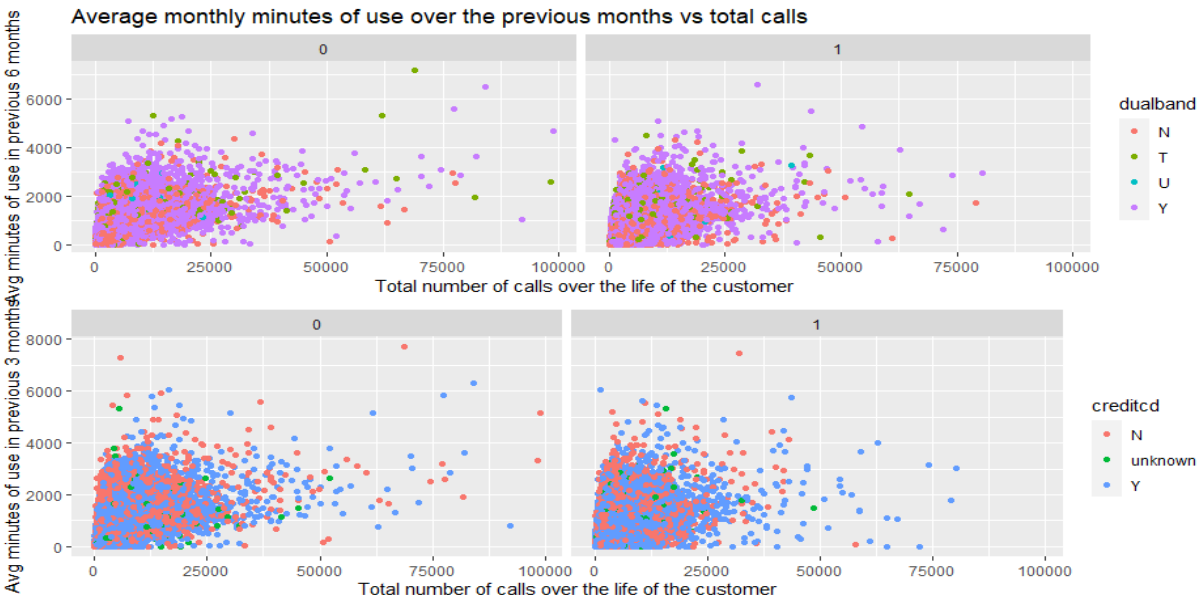


This data is evenly balanced with customers which is a positive attribute for the whole data analysis and that evenness is also shown inside the bar plot of the customers visualized from all nineteen regions. Churn rate is almost equal for all regions .

Customer metadata information :



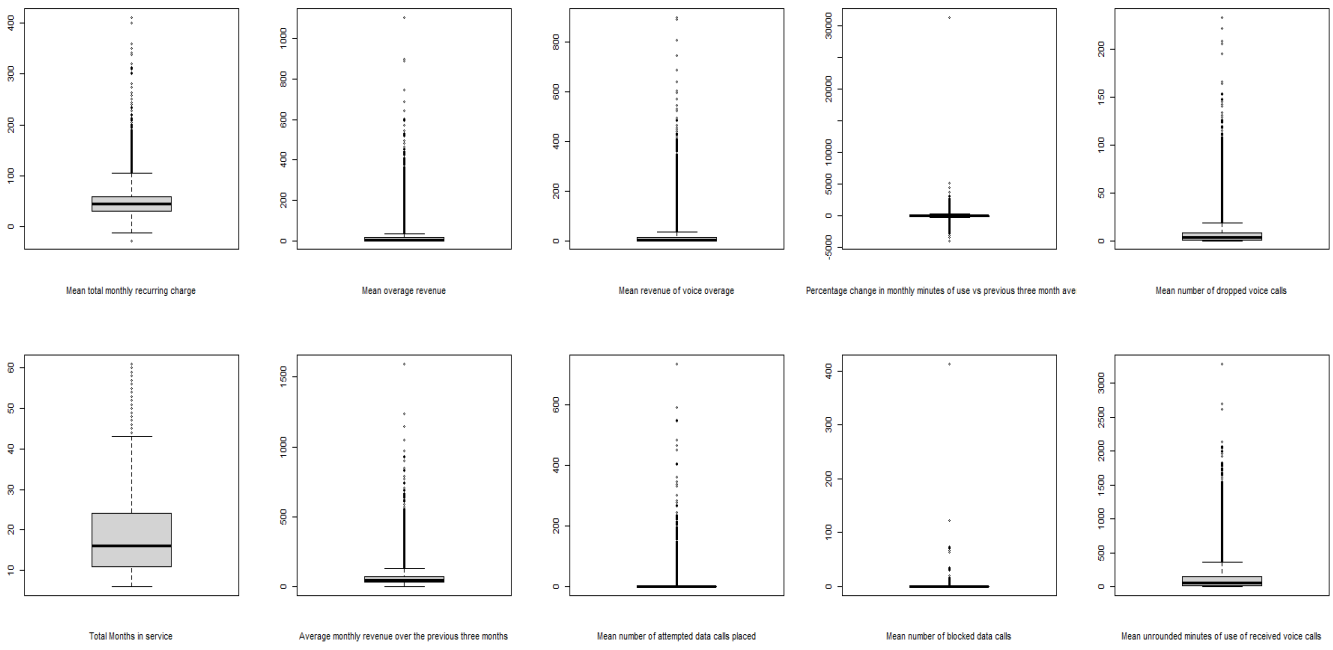
This credit card status is another decisive factor for potential churn prediction as from the pie-chart we can conclude that around 68% of customers are having credit card and 31% customers are not using credit card .



This visualization shows pattern amongst two categories of customers irrespective of their credit status. This demonstrates that both red and blue data points lie uniformly within the range of 25000 number of calls which means both of these customers either of having credit status belong to this range but when it comes to higher range i.e. 50000 calls , we can only see blue data points for churned customers. This clearly indicates that churned customers are dependent on having credit cards. On the contrary for unchurned customers, we can see both red and blue data points in the higher call range which clearly defines that for customers who may not churn are not really bothered about their credit status for attempting more calls or higher number of calls .

Methodology

Outlier Processing



From the boxplot analysis, it is evident that the outliers for the following features are to be removed:

1. Mean total monthly recurring charge,
2. Mean overage revenue
3. Mean revenue of voice overage
4. The Percentage change in monthly minutes of use vs previous three-month average
5. Mean number of dropped voice calls
6. Average monthly revenue over the previous three months
7. Mean number of attempted data calls placed
8. Mean number of blocked data calls
9. Mean unrounded minutes of use of received voice calls
10. Mean monthly revenue

Missing Value Imputation

For data preparation, the missing values from 43 out of 90 columns were dropped. This was achieved by deleting 30% of all missing values and imputing the rest by the following methods:

- Imputing with Median

  It is observed that this data is heavily infested with outliers or extreme values so for imputation of numerical variables it is widely imputed with the median as it is more robust towards handling outliers than the mean without impacting the whole distributions as a very minimal percentage of missing values are imputed.

- Imputing with Modes

  Since it won't alter the overall frequencies and distributions for specific labels or categories, the columns where it is observed that missing values are minimum in percentage relative to whole data samples those have been imputed by using modes.

  Creating a separate label for Missing values

  In this case, a separate category is created out of all the labels to prevent the data from becoming unintentionally biased whenever it is determined that the imputation by modes will affect the frequency of all labels or categories by giving that particular label an undesired bias.

Feature Transformation

From the above histograms, it is evident that there is some significant skewness in the following features:

1. Mean monthly revenue,
2. Mean number of monthly minutes of use,
3. Mean total monthly recurring charge,
4. Mean number of completed calls,
5. Mean number of attempted calls,
6. Mean number of inbound and outbound peak voice calls,
7. Mean number of completed voice calls,
8. Mean number of peak data calls,
9. Mean unrounded minutes of use of peak voice calls,
10. Mean number of outbound wireless to wireless voice calls

The skewness of these features can really impact the performance of the model. Hence, by taking the log, we have transformed these features into its logarithmic form, which helps to reduce the skewness of these features.

Feature Selection:

Some variables amongst all the features are excluded since they don't perfectly match the entire problem. In addition, variables with more than 30% of missing values are also discarded since they scarcely provide any context with the data.

Feature Engineering

Mentioned below are the features incorporated from the existing variables :

- Percentage of completed calls = (Mean number of completed data calls/ Total number of calls overthe life of the customer) *100.
- Percentage of dropped calls = (Mean number of dropped calls/ Total number of calls over the life ofthe customer) *100.
- Mean number of incompleted calls = (Mean number of completed calls - Mean number of attempted calls)
- Percentage of the mean number of incompleted calls = (Mean number of incompleted calls/ Totalnumber of calls over the life of the customer) *100.

Modelling Choices

The models implemented for this project are:

1. Logistic Regression,
2. Decision Tree,
3. Random Forest,
4. GBM,
5. Neural Net

Model Validation Plan

For model validation, the entire dataset is used to implement 5-fold Cross Validation. The primary evaluation metric chosen is log loss.For the model implementation part, first the performance of the model without considering the outliers is evaluated and later it is done considering the outliers to check whether there are any differences in the performance of the models.

Results

For the results, we only include our primary metric in the table, which is log loss.

Model Performance Summary:

Model	Method	Package	Hyperparameter	Selection	LogLoss
Logistic Regression	glm	h2o	family lambda alpha standardize	Binomial 0 0.1 TRUE	0.68203
Decision Tree	gbm	h2o	ntrees min_rows sample_rate col_sample_rate	1 1 1 1	0.67474
Random Forest	randomforest	h2o	ntrees col_sample_rate_change_per_level max_depth nbins_top_level	100 0.8 3 512	0.645126
GBM	gbm	h2o	ntrees col_sample_rate	250 0.8	0.61863

			sample_rate max_depth nbins_top_level	0.8 3 256	
Neural Net	deep learning	h2o	Hidden epochs activation	C(2) 50 RectifierWithDropout	0.65201

Table 1 : Performance of models without considering outliers

Retraining the above models, while considering the outliers

Model	Method	Package	Hyperparameter	Selection	LogLoss
Logistic Regression	glm	h2o	family lambda alpha standardize	Binomial 0 0.1 TRUE	0.7002
Decision Tree	gbm	h2o	ntrees min_rows sample_rate col_sample_rate	1 1 1 1	0.6612
Random Forest	randomforest	h2o	ntrees col_sample_rate_change_per_level max_depth nbins_top_level	100 0.8 3 512	0.6452
GBM	gbm	h2o	ntrees col_sample_rate sample_rate max_depth nbins_top_level	250 0.8 0.8 3 256	0.5988
Neural Net	deep learning	h2o	Hidden epochs activation	C(2) 50 RectifierWithDropout	0.65201

Table 2: Performance of models considering outliers

Key Findings of Analysis

- From the above performance of all the modes that have been implemented, GBM outperforms all the other models and have been selected as the main model.
- For this particular dataset, tree-based models perform better than linear models.
- Although the performance of the logistic regression and neural network has somewhat declined when outliers are taken into account, the performance of the tree-based models is unaffected, indicating that outliers have no impact on the performance of the tree-based models.

Feature Importance

The table below shows the top 5 features of each model that have been implemented:

Plots of the top 5 features of the Bagging, Boosting and Neural Network models are given below:

Logistic regression	**Decision Tree**	**Random Forest**	**GBM**	**Neural Net**
1. Mean rounded minutes of customer care calls. 2. Billing adjusted total number of calls over the life of the customer 3. Mean unrounded minutes of use of customer care calls 4. Total number of calls over the life of the customer 5. Dualband	1.Total number of months in service 2. Current handset price 3. Mean total monthly recurring charge 4. Average monthly revenue over the previous three months 5. Mean number of monthly minutes of use	1.Total number of months in service 2. Current handset price 3. Mean total monthly recurring charge 4. Mean number of monthly minutes of use 5. Average monthly revenue over the previous 3 months.	1.Total number of months in service 2. Average monthly revenue over the previous three months 3. Percentage change in monthly minutes of use vs previous three month average 4. Mean number of monthly minutes of use 5. Current handset price	1. Mean monthly revenue 2. Average monthly revenue over the previous three months 3. Current handset price 4. Percentage change in monthly revenue vs previous three month average 5. Total number of calls over the life of the customer

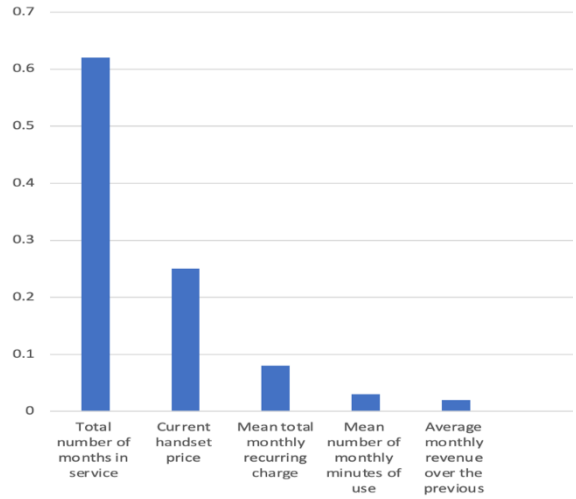


Fig 1: Random Forest (Bagging)

For the Bagging Model as shown above, the feature, "the total number of months in service for the customers", has the highest importance in terms of predicting towards the customer churn. Also, variables like "the current handset price", "mean total monthly recurring charge", "mean number of monthly minutes of use" and "Average monthly revenue over the previous 3 months", follow the priority of importance respectively.

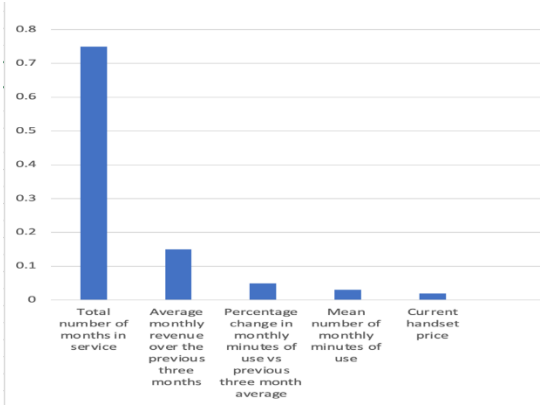


Fig 2: GBM (Boosting)

For the Boosting Model, similar to the Bagging Model, "Total number of months in service", carry the highest importance. But, the 2nd ("Average monthly revenue over the previous three months") and 3rd ("Percentage change in monthly minutes of use vs previous three month average") features for this model are different to that of the bagging model when compared with the priority of the features. The "Current handset price" feature which was at the 2nd priority for the Bagging Model is the 5th priority for the Boosting Model.

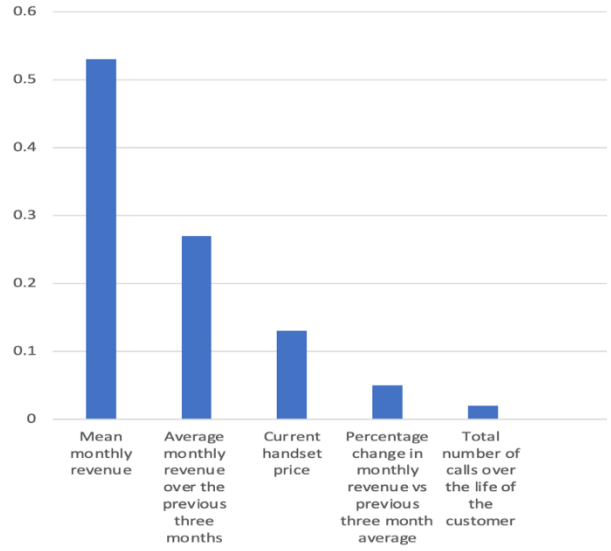


Fig 3: Neural Network

For the Neural Network model, the “Mean monthly revenue” feature shows of the highest importance for the prediction of customer churn. After comparing with the Boosting Model, the 3rd (“Percentage change in monthly minutes of use vs previous three month average”) and 5th (“Current handset price”) features are changed to 4th and 3rd place of priority respectively. Additionally comparing with both the Bagging and the Boosting models, a new feature, “Total number of calls over the life of the customer” has been added .

Conclusion

- Predicting churn is not just identifying at risk customers, i.e., customers who may churn, but it is also identifying the pain points leading up to the churn and helping increase overall customer retention and satisfaction.
- So, by identifying customers, customer success teams may be asked to start a reach-out campaign for these high-risk customers to provide support and re-engagement.
- It is presumed that these customers are primarily postpaid based on characteristics like typical consumer calls or revenue patterns even if this is not expressly stated in the problem statement. Some labels for features like credit status and marital status are not explicitly mentioned in the issue statement, but they can be presumed to be specialized labels for the necessary of the solution without having an impact on the entire dataset in the form of noise.
- The following methodologies were implemented before fitting the data to 5 different models:
    - Imputing the missing values for the numerical variables with median as mean can be impacted with outliers and Imputing the missing values for the categorical variables with mode.
    - Removal of outliers through boxplot analysis,
    - Removing unnecessary features,
    - Incorporating some new features from the existing features.
- The 5 different models implemented are:
    - Logistic Regression
    - Decision tree
    - GBM
    - Random Forest
    - Neural Network
- 5-fold Cross Validation was performed on the entire dataset with the models mentioned above in two different ways, one with considering outliers and another without considering outliers.
- After evaluation of all the models, GBM shows the best result with the tree based models outperforming the linear models for this particular dataset. Additionally, it is evident that outliers do not make any impact on the performance of the tree based models.
- The important features required for identifying at-risk customers are the same to those which are identified in the best working model evaluation i.e., the GBM .
- Focusing on these features from the model will help reduce churn.
- For this problem, by considering the GBM model, it is expected that it would reduce churn by a certain amount over the next few quarters in the long run.
- As a consequence of reduction in churn, it will lead to the increase in revenue as well.