Final Report for:

FINAL REPORT FOR TELECOM CHURN PREDICTION

Table of Contents

1. INTRODUCTION OF THE BUSINESS PROBLEM

1.1) Problem Statement

The senior management in a telecom provider organization is worried about the rising customer attrition levels. Additionally, a recent independent survey has suggested that the industry as a whole will face increasing churn rates and decreasing ARPU (average revenue per unit). The effort to retain customers so far has been very reactive. Only when the customer calls to close their account is when the company takes action. That has not proved to be a great strategy so far. The management team is keen to take more proactive measures on this front. You as a data scientist are tasked to derive insights, predict the potential behavior of customers, and then recommend steps to reduce churn.

Problem: The senior management in a telecom provider organization is worried about the rising customer attrition levels.

Constraints: The effort to retain customers so far has been very reactive. Only when the customer calls to close their account is when the company takes action. That has not proved to be a great strategy so far.

Scope: Predict the potential behavior of customers and analyze factors leading to the churn attrition.

Objectives: Derive insights, predict the potential behavior of customers, and then recommend proactive measures to reduce churn.

1.2) Need of the Project

Customer churn, also known as customer retention, it quantifies the number of customers who have left by cancelling their subscription or stopping paying for the services. This retention would cost business five times more to attract a new customer as it does to keep an existing one. Company needs to know what is their current attrition rate and what are the factors affecting this attrition.

Reducing churn rate leads to

- High revenue
- Increase in sales
- High Customer Lifetime Value
- Establishing Brand Value in the Market

1.3) Understanding Business Opportunity

Telephone service companies, Internet service providers, insurance firms, and alarm monitoring services, often use customer attrition analysis and customer attrition rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. Companies from these sectors often have customer service branches which attempt to win back defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients.

2.DATA REPORT

There are 3 Types of information available in the dataset:

1. Customer Usage: No: of calls, revenue reflecting from the customers calls, data of the minutes of the calls etc;

2. Service Quality: How long the customers have to wait, how many calls are made to customer service, number of calls dropped/blocked etc;

3. Customer Demographic: Customer information such as marital status, gender , number of cars they own, car types etc;

Renaming is not required as all the variables are being names properly

This dataset has been collected by documents and records method. Organization may already be having records of the customer's data, so this data would be used to examine and analyze the data.

It consists of various variables explaining the attributes of telecom industry and various factors considered important while dealing with customers of telecom industry. The target variable here is churn which explains whether the customer will churn or not. We can use this dataset to predict the potential behaviors of customers.

- The dataset consists of 26518 rows and 81columns.
- Out of 81, 21 columns are of object type and rest 60 are of either integer or float data type.
- The dataset contains a mix of data of types numeric, categorical. There are 47 continuous variables and 37 categorical variables.
- "Churn" is the target variable
- 0 indicates the number of customers who did not churn and 1 shows the number of people who have churned.

Below is sample of the dataset showing the first 5 rows.

	mou_Mean	totmrc_Mean	rev_Range	mou_Range	change_mou	drop_blk_Mean	drop_vce_Range	owylis_vce_Range	mou_opkv_Range
0	190.25	63.9400	26.00	43.0	-11.25	4.666667	8	20	28.44
1	443.00	39.9900	5.10	199.0	-78.00	4.333333	5	33	72.46
2	400.50	44.9900	13.88	172.0	-67.50	2.000000	1	7	93.60
3	53.50	34.6675	18.56	78.0	12.50	1.333333	2	0	4.70
4	37.00	21.0425	35.79	74.0	-33.00	5.666667	1	4	36.60

Table 1: Dataset Sample

	mou_Mean	totmrc_Mean	rev_Range	mou_Range	change_mou	drop_blk_Mean	drop_vce_Range	owylis_vce_Range	mou_o
count	26460.000000	26460.000000	26460.000000	26460.000000	26357.000000	26518.000000	26518.000000	26518.000000	265
mean	533.578283	47.189057	44.132108	378.747600	-10.210774	10.220366	5.519534	15.904895	1
std	541.184126	24.264305	71.822912	428.889022	255.202308	15.500170	8.827681	23.611921	1
min	0.000000	-26.915000	0.000000	0.000000	-2785.000000	0.000000	0.000000	0.000000	
25%	160.500000	30.000000	1.980000	114.000000	-83.000000	2.000000	1.000000	3.000000	
50%	368.500000	44.990000	15.990000	245.000000	-4.750000	5.333333	3.000000	9.000000	
75%	733.250000	59.990000	57.442500	485.000000	65.250000	12.666667	7.000000	20.000000	1
max	7667.750000	399.990000	1524.390000	6233.000000	3046.750000	411.666667	313.000000	542.000000	47

Table 2: Sample of descriptive statistics of first few variables.

- From Table 2, it is observed that there are missing values present in the dataset from the count variable.
- The maximum number of mean monthly use of minutes of the customers is 7667.75 minutes.
- The minimum number of dropped/failed calls is 0 and the maximum is 411 number of calls. The average number of mean dropped calls is 5. A huge variation is seen, and it is known to have outlier present in this variable "drop_blk_Mean"
- The maximum and minimum total revenue reflecting from the calls of customers are 13358 and 9 respectively.
- From "div" variable, it is observed that there are huge number of N/A values.
- There are no duplicate records present in the dataset.

3.Exploratory data analysis

3.1. Univariate analysis

Distribution of Target Variable "churn"

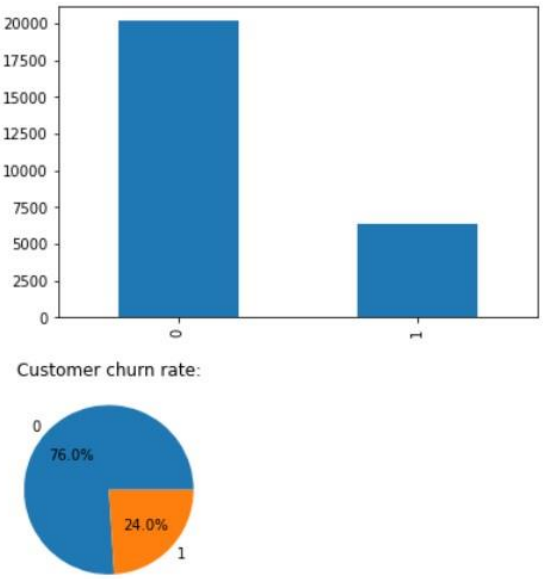


Fig 1.1 Pie chart and bar chart distribution of churn variable

The above bar chart shows us the number of telecom customers who have churned in comparison to who have not. There are 20154 customers who did not churn and 6364 customers who churned i.e. 76% of the customers who did not churn and 24% who did.

Distribution of categories for categorical variables.

```
crclscod : ['AA' 'EA' 'C' 'G' 'CA' 'E4' 'E' 'B' 'BA' 'O' 'A' 'DA' 'U' 'ZA' 'D4' 'A2'
 'D5' 'C2' 'GA' 'Z4' 'JF' 'IF' 'ZY' 'GY' 'Z' 'CC' 'I' 'B2' 'EM' 'E2' 'K'
 'M' 'EC' 'V1' 'CY' 'D' 'U1' 'Z5' 'C5' 'W' 'J' 'H' 'EF' 'Y' 'TP' 'A3' 'Z2'
 'Z1' 'D2']
asl_flag : ['N' 'Y']
prizm_social_one : ['T' 'U' 'C' 'S' 'R' nan]
area : ['ATLANTIC SOUTH AREA' 'LOS ANGELES AREA' 'MIDWEST AREA'
 'DC/MARYLAND/VIRGINIA AREA' 'NEW YORK CITY AREA' 'DALLAS AREA'
 'NORTHWEST/ROCKY MOUNTAIN AREA' 'NEW ENGLAND AREA' 'SOUTHWEST AREA'
 'CHICAGO AREA' 'OHIO AREA' 'NORTH FLORIDA AREA' 'CALIFORNIA NORTH AREA'
 'HOUSTON AREA' 'SOUTH FLORIDA AREA' 'CENTRAL/SOUTH TEXAS AREA'
 'TENNESSEE AREA' 'GREAT LAKES AREA' 'PHILADELPHIA AREA' nan]
refurb_new : ['N' 'R']
hnd_webcap : ['WCMB' 'WC' nan 'UNKW']
marital : ['M' 'U' 'B' 'S' 'A' nan]
ethnic : ['N' 'S' 'U' 'H' 'R' 'G' 'O' 'D' 'Z' 'F' 'I' nan 'J' 'B' 'P' 'M' 'X' 'C']
dwlltype : ['S' nan 'M']
dwllsize : ['A' nan 'C' 'B' 'K' 'J' 'D' 'L' 'I' 'N' 'E' 'M' 'G' 'O' 'F' 'H']
mailordr : ['B' nan]
occu1 : ['A' nan '1' '4' '6' 'E' '5' '3' '2' '7' '8' 'D' 'F' 'H' 'K' 'G' 'C' 'J'
 'B' 'Z' 'I' '9']
wrkwoman : ['Y' nan]
solflag : [nan 'N' 'Y']
proptype : ['A' nan 'D' 'B' 'E' 'M' 'G']
mailresp : ['R' nan]
cartype : [nan 'E' 'F' 'C' 'A' 'B' 'D' 'G']
children : ['Y' nan 'N']
csa : ['AIRCOL803' 'LAXCUL310' 'STLSTL314' 'LAXANA714' 'NCRGRB757' 'ATLATL678'
 'APCFCH703' 'NYCNEW201' 'OKCOKC405' 'SLCOGD801' 'OMADES515' 'NYCMAN917'
 'APCSIL301' 'NYCNAS516' 'LAXLAX213' 'HARNEW203' 'NYCBRO917' 'NYCQUE917'
 'NYCWHI914' 'PHXTUC520' 'CHIDAV319' 'SEASEA206' 'INDIND317' 'OHISAN419'
 'MINMIN612' 'BOSNSH603' 'LAXCOV626' 'FLNKIS407' 'SFRCBL408' 'HARSPR413'
 'HOUHOU281' 'APCBAL410' 'DALDAL214' 'SFRSCL408' 'BOSMAN603' 'NORZIM763'
 'NCRNWN757' 'AWISHE920' 'NYCNEW908' 'HOUGLV409' 'MIAMIA305' 'OHICAN330'
 'AIRSPA864' 'NYCSUF516' 'LAXBUR818' 'LAXONT909' 'SANKIL254' 'DENBOU303'
 'SFRSFR415' 'SFRSRO707' 'NCRVIR757' 'ATLMEM901' 'HWIMAU808' 'DENGLD303'
 'DETROS810' 'BOSBOS617' 'PHIPHI215' 'DENDEN303' 'MIAWPB561' 'ATLNOR678'
 'MIAFTL954' 'DETDET313' 'HARBRI203' 'NEVESC760' 'NMXELP915' 'HWIHON808'
 'SANMCA210' 'OHICOL614' 'BOSBOS978' 'PHIARD610' 'NEVLVS702' 'NYCFHD732'
 'FLNJAC904' 'MILMIL414' 'CHICHI773' 'DALFTW817' 'OMAOMA402' 'NCRRIC804'
 'HOUCON409' 'NYCNEW973' 'SDASFL605' 'OHIXEN937' 'DETPON248' 'NEVCOR619'
 'SANLAM512' 'DENCOL719' 'DETFER248' 'NMXLAR956' 'AWIGRE920' 'AIRMYR843'
 'CHINBK847' 'KCYKCM816' 'FLNCLR813' 'SANSAN210' 'NMXLUB806' 'DETJAC517'
 'VAHCHL804' 'KCYKCK913' 'NCRCHA704' 'LAXLAG949' 'OHIDEL740' 'SLCKAY801'
 'MIAFTM941' 'STLCRD618' 'NCRFAY910' 'ATLKNO423' 'SEACDA208' 'FLNORL407'
 'PHIWIL302' 'ATLANE678' 'APCANN443' 'NCRYOR803' 'SANAUS512' 'FLNWNP407'
 'LAXSAN714' 'LAXIRV949' 'NSHNSH615' 'DETTRO248' 'SHEMAR304' 'ATLCHA423'
 'HOUSPR832' 'VAHROA540' 'IPMSAG517' 'NYCTMR732' 'MIAMAR305' 'LAXVNY818'
 'NCRWIN336' 'SFROAK510' 'PHXPHX602' 'NCRPOR757' 'LAXRIV909' 'OKCMAN785'
 'LOUETN502' 'NEVSDG619' 'AIRGRE864' 'FLNTAM813' 'NCRCRY919' 'LAXPAS626'
 'FLNCOC407' 'PHICTR610' 'SEAALB541' 'NEVELC619' 'NYCNEW732' 'LOULEX606'
 'NYCWOO732' 'KCYLAW913' 'INHSBN219' 'NEVENC760' 'NVUREN775' 'NOLKEN504'
 'OHIDAY937' 'ATLMAC912' 'OHICIN513' 'APCWAS202' 'OHICLE216' 'FLNBEL352'
 'PHICHC215' 'PITHOM412' 'BOSPRO401' 'BOSBOS781' 'NYCMTK914' 'SHEFTR540'
```

Fig 1.2 Distribution of data in categorical variables

Distribution of working women category

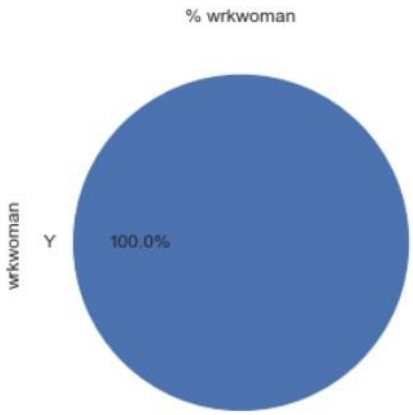


Fig 1.4 Pie chart of wrkwoman variable

Observation: All women customers who are using the service are working.

Distribution of Area category

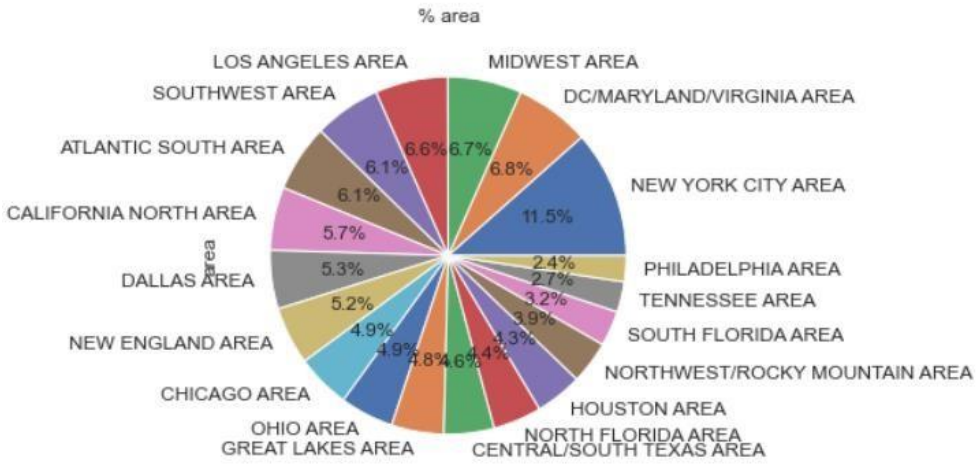


Fig 1.5 Pie chart of area variable

Observation: Maximum number of customers are from New York City followed by DC/Maryland/Virginia area. The lowest number of customers are from Philadelphia.

Distribution of Marital Status

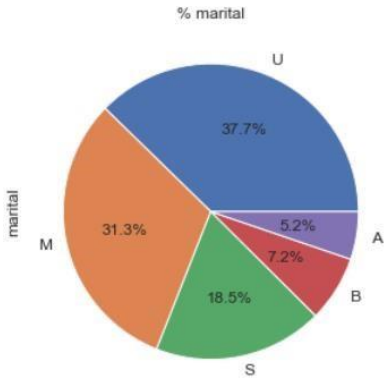


Fig 1.6 Pie chart of area variable

Observation: Unmarried/Single customers are using the telecom service more in comparison to the customers who are married.

Distribution of Models

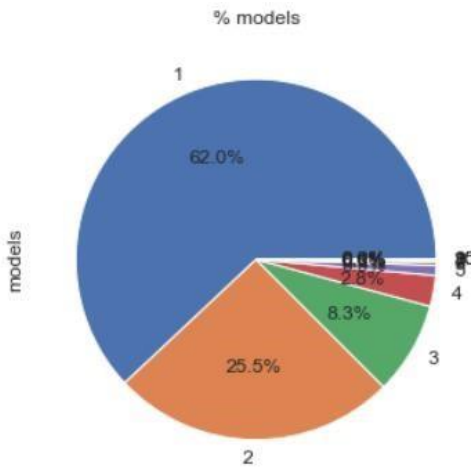


Fig 1.7 Pie chart of area variable

Observation: Customers use 62% of Model 1 category.

Distribution of Children

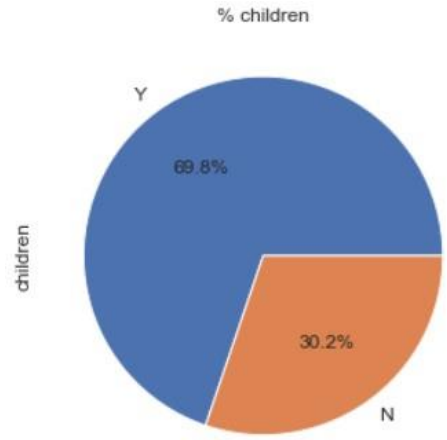


Fig 1.8 Pie chart of children variable

Observation: Approximately, 70% of the customers have children who use the service and 30% of the customers do not have children.

Distribution of Number of Cars

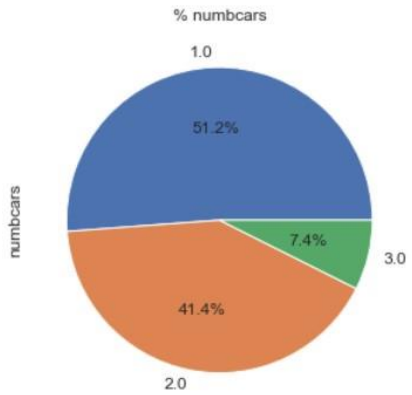


Fig 1.9 Pie chart of numbcars variable

Observation: 51% of the customers has only one car and 41.4% has 2 cars. Only 7.4% of the users have 3 cars.

Distribution of hnd_webcap

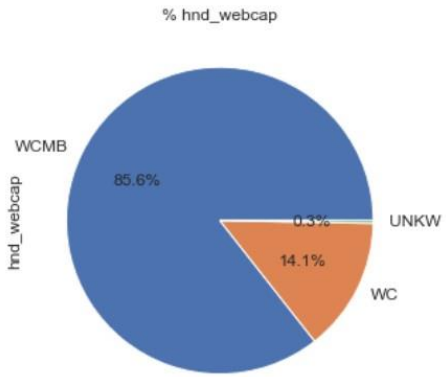


Fig 1.10 Pie chart of hnd_webcap variable

Observation: Most of the customers have a web compatible handsets.

Distribution of income category

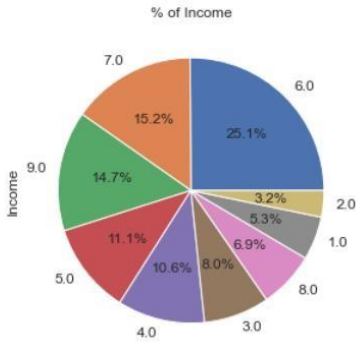


Fig 1.11 Pie chart of income variable

Observation: Majority of the subscribers belong to lowest class and middle class income.

Distribution of data for numeric variables

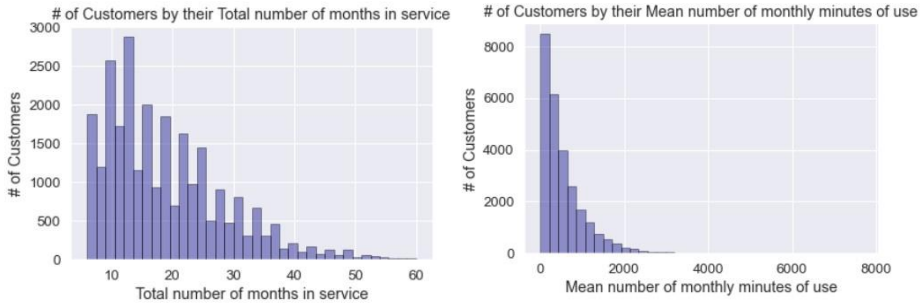


Fig 1.12 Dist plot of months and monthly mins variables

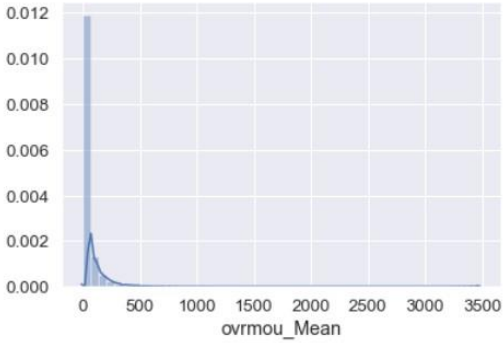


Fig 1.13 Dist plot of ovrmou_Mean variable

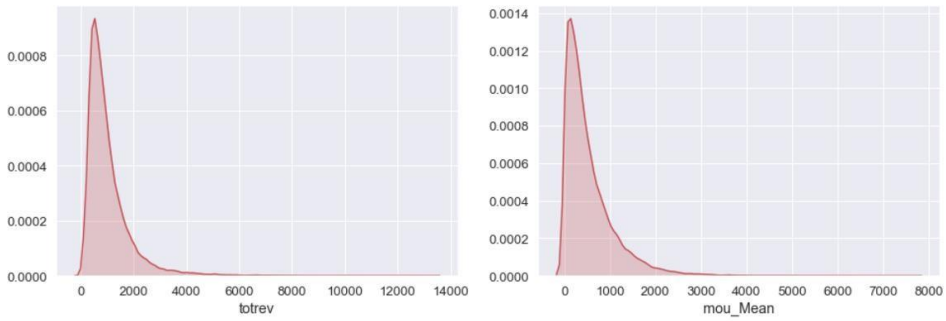


Fig 1.14 Distribution plot of totrev and mou_mean

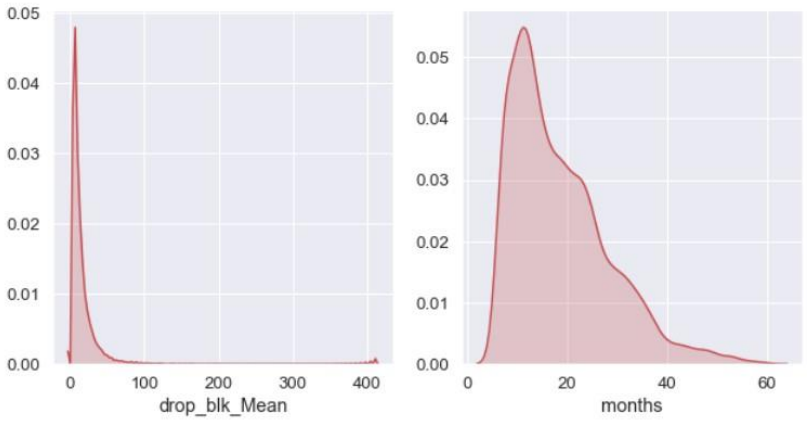


Fig 1.15 Distribution plot of drop_blk_Mean and months

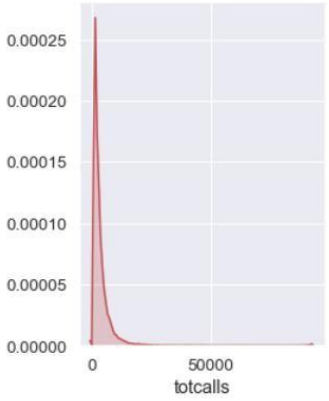


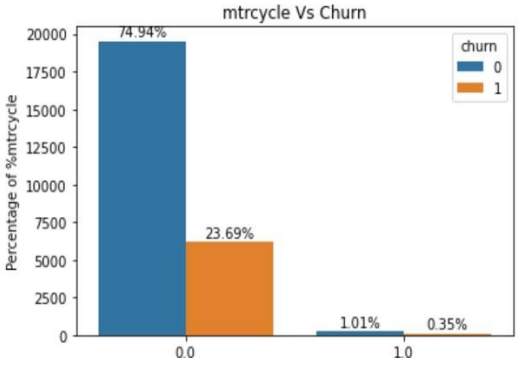
Fig 1.16. Distribution plot of totcalls

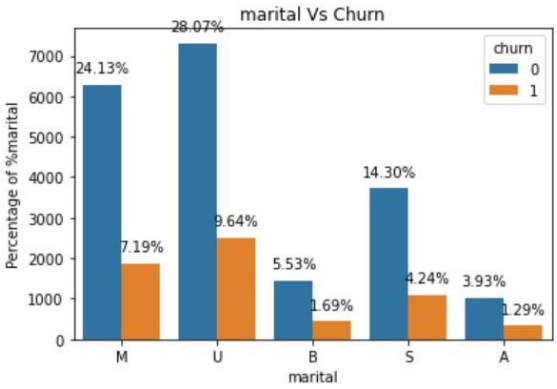
Observation: The above graphs show us that these continuous variables are highly skewed.

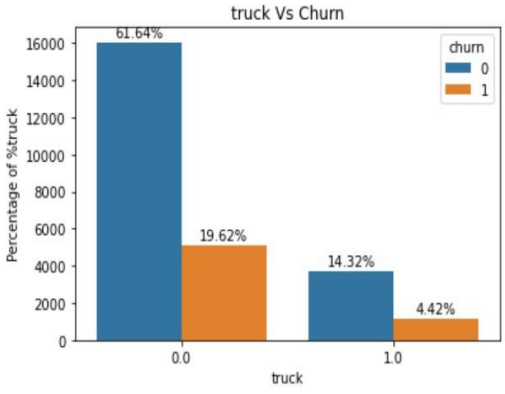
3.2. Bivariate Analysis

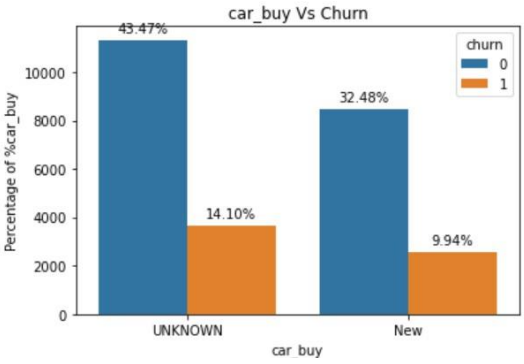
Bi-variate analysis is performed to find the correlations between various variables.

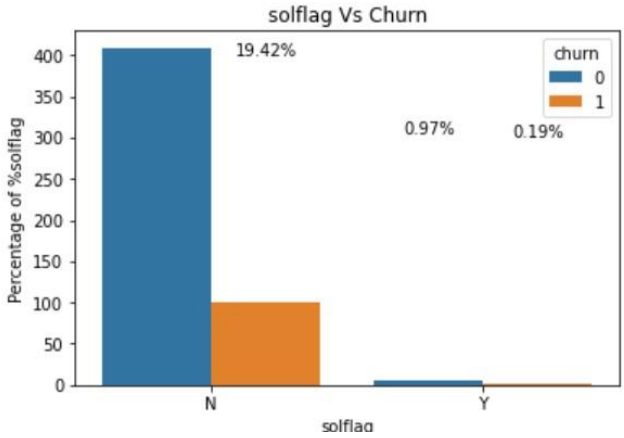
Categories with respect to churn

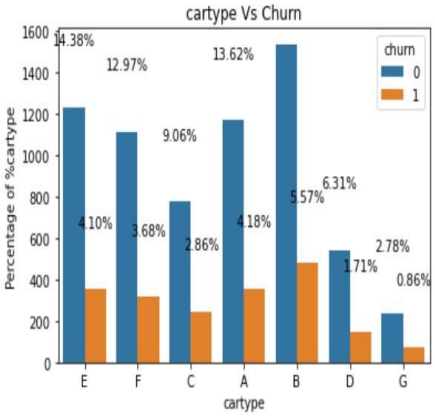












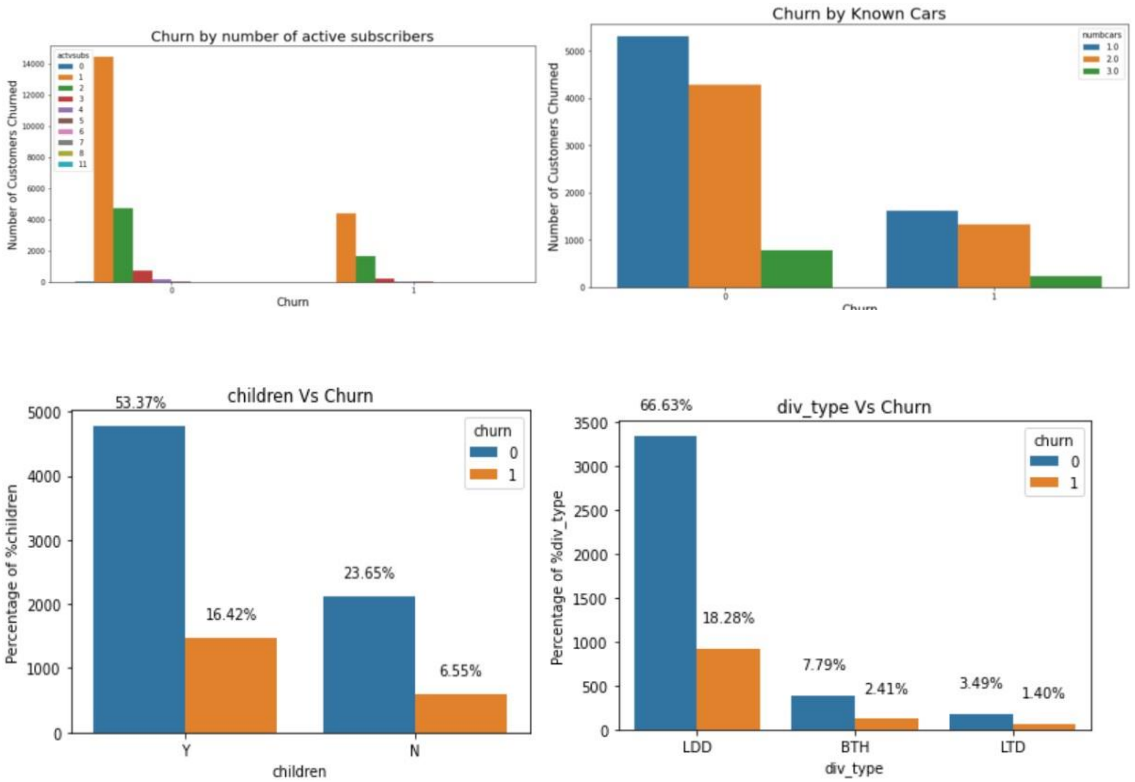


Fig 2.1 Bar plots of categorical variables w.r.t. churn

Observation:

- As observed from the above category vs churn graphs, It seems that there is more churn rate with customers who are single and do not have children.
- 66.3% of Customers who do not churn are in LDD division type.
- Customers with children do not churn frequently as customers who do not have children.
- Customers who have 3 cars churn the least.
- Customers who have a motorcycle do not churn and there is only 0.35% of customers who do not have motorcycle that churn.
- As already observed there are high number of customers from New York City Area and also with respect to churn rates, it also has high churn rate.

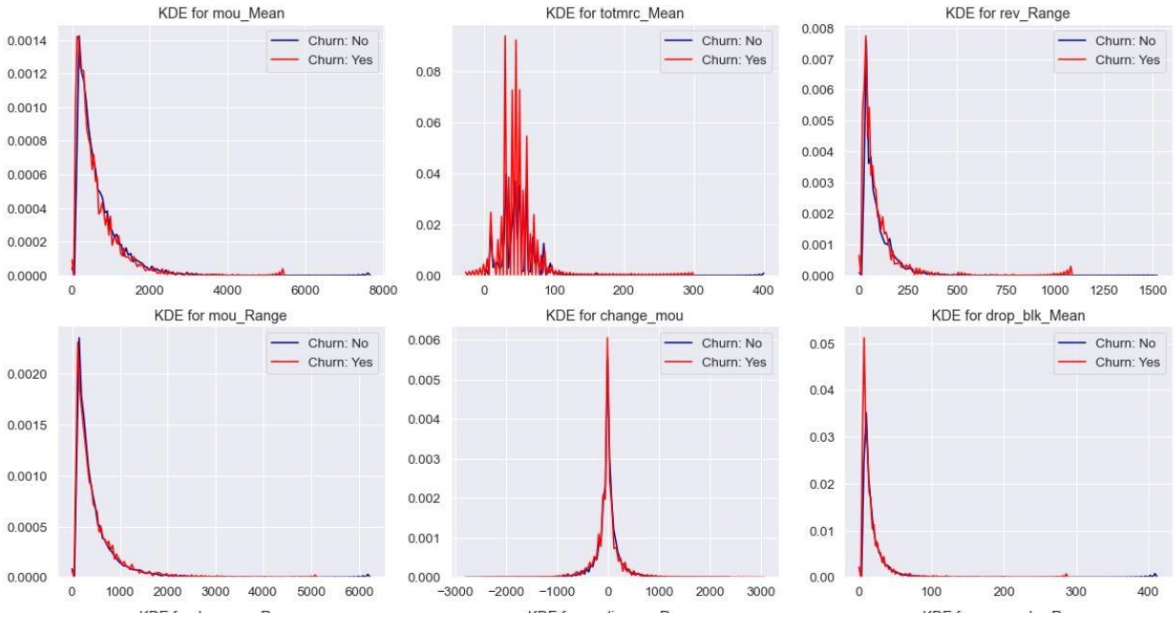


Fig 2.2 Dist plots of numerical variables w.r.t churn

Observation:

Refer to Appendix 1.4 , Churn and No churn rate is almost same for most of the continuous variable, in other words there is no significant difference.

Churn rate is high for customers who have called the customer service. Quality of Customer service calls should also be improved.

It is also noticed that most of the customer retentions happen when the customers had subscribed to the telecom service after 1 year. So the company should pay attention to the customers who have been with them for almost a year.

Churn rate is high for the number of failed data calls made by the customer.

Clearly the data is skewed as we would expect a large majority of the customers to not churn. This is important to keep in mind for our modelling as skeweness could lead to a lot of false negatives.

First, we will analyze a correlation heatmap to interpret the direction and strength of relationship between any two numeric variables.

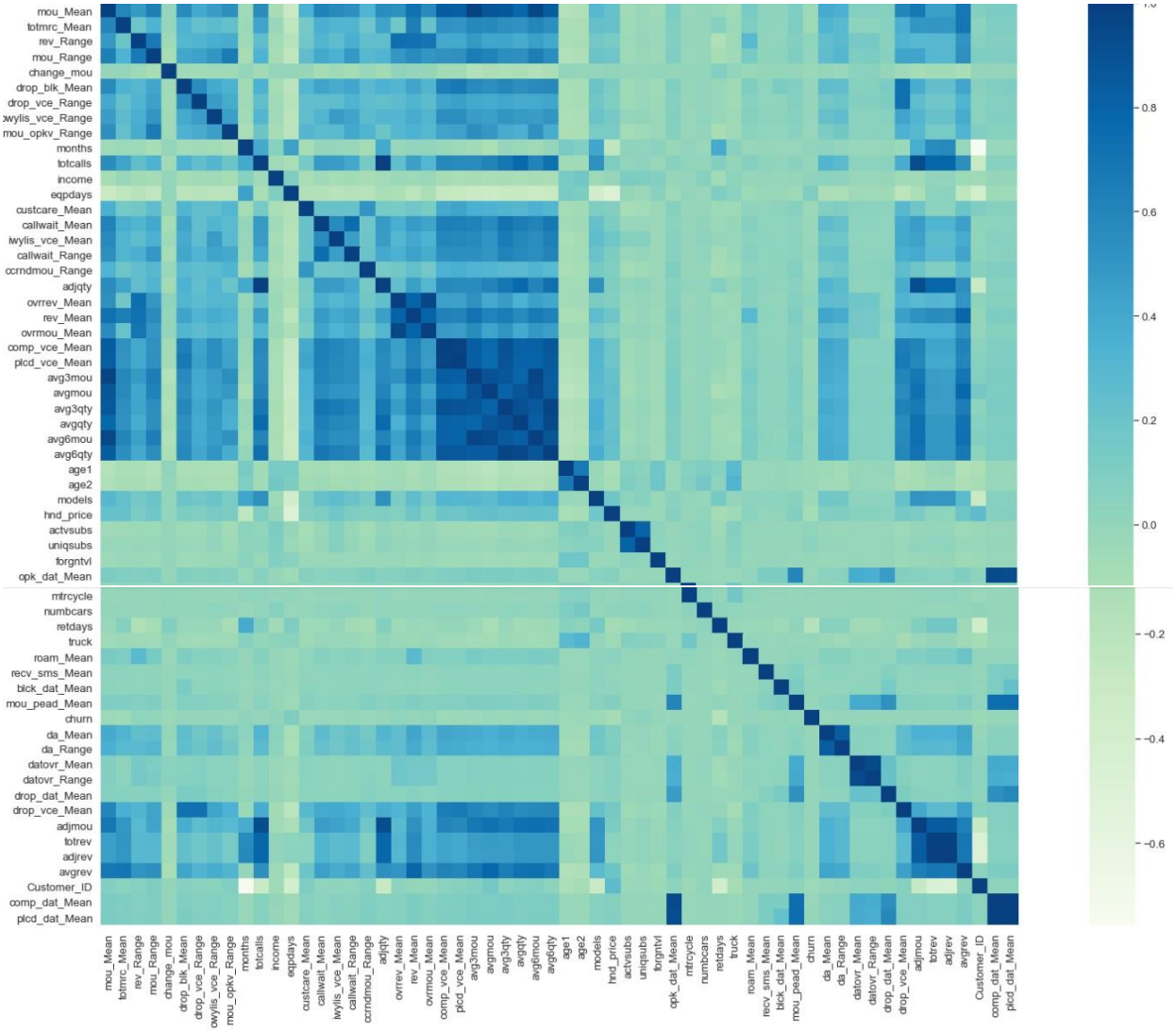


Fig 2.3 Heat Map

Several of the numerical data are highly correlated.

1. Totcalls and adjqty
2. Adjqty and totcalls
3. Adjrev and totrev and many more.

The top highly correlated variables are shown below:

```
totcalls       adjqty          0.999821
adjqty         totcalls        0.999821
adjrev         totrev          0.998556
totrev         adjrev          0.998556
comp_dat_Mean  plcd_dat_Mean    0.992201
plcd_dat_Mean  comp_dat_Mean    0.992201
mou_Mean       avg3mou         0.985201
avg3mou        mou_Mean        0.985201
plcd_vce_Mean  comp_vce_Mean    0.984325
comp_vce_Mean  plcd_vce_Mean    0.984325
avg6qty        avg3qty         0.970395
avg3qty        avg6qty         0.970395
avg6mou        avg3mou         0.965277
avg3mou        avg6mou         0.965277
ovrmou_Mean    ovrrev_Mean     0.956581
ovrrev_Mean    ovrmou_Mean     0.956581
avg6mou        mou_Mean        0.952443
mou_Mean       avg6mou         0.952443
avgqty         avg6qty         0.947467
avg6qty        avgqty          0.947467
avg6mou        avgmou          0.940460
avgmou         avg6mou         0.940460
datovr_Range   datovr_Mean     0.935367
datovr_Mean    datovr_Range    0.935367
opk_dat_Mean   comp_dat_Mean   0.934892
comp_dat_Mean  opk_dat_Mean    0.934892
adjmou         adjqty          0.930018
adjqty         adjmou          0.930018
adjmou         totcalls        0.929786
totcalls       adjmou          0.929786
dtype: float64
```

Fig 2.4 Top 30 variables with high correlation

These variables have a very high correlation and should likely to be removed. But in a real life business scenario, it is always recommended that the variables shouldn't be dropped without consulting the organization. So following the real life scenario approach, PCA is used later rather than dropping the variables. It helps us with multicollinearity and dimensionality reduction of the dataset.

3.3. Business insights from EDA

1. It is observed that most of the telecom customers generate low revenue, while it is only a few who are the high revenue generator. So these high revenue generator customers are deemed to be more important.
2. Customers who have more than one active subscriber in the household are less likely to churn and subscribers with only one active subscriber are more likely to churn. Therefore, the company could suggest or come up with family pack discounts etc; for multiple active subscribers to further decrease the churn rate.
3. As mentioned earlier in the report, relatively new customers who have been subscribers for a year with company tend to churn out more, therefore loyalty discounts etc; can be bought forward to these customers to reduce the churn rate.
4. Most of the customers who have churned have high number of failed calls.
5. Telecom company can also improve their network coverage in the areas where there are high churn rates. One such area that we found from EDA is New York city. Majority of the subscribers and churn rates happen in NYC, so this could be avoided by improving the network coverage quality in that area

4. Data Cleaning/Data Pre-processing

There are large number of missing values in several columns of the dataset.

Approach that has been used to treat the missing values:

1. High number of null values i.e. if the columns have more than 40% of null values present, these columns have been dropped.
2. If the percentage of missing values is between 1% and 40%, imputation has been done by median for continuous variables and mode has been imputed for the missing values of categorical variables.
3. There was no relation between the columns that have between 2% and 10% missing values.

Below is a list of columns with null values with their percentages.

```
solflag             98.057923
retdays             96.734294
wrkwoman            87.661211
div_type            81.080775
occu1               73.489705
proptype            71.600422
cartype             67.708726
children            66.226714
mailordr            64.145109
mailresp            62.681952
numbcars            48.868693
dwllsize            38.049627
dwlltype            31.604948
income              25.254544
hnd_webcap           8.982578
prizm_social_one     7.093295
avg6mou              3.069613
avg6qty              3.069613
age2                 1.764839
truck                1.764839
ethnic               1.764839
age1                 1.764839
marital              1.764839
forgntvl             1.764839
mtrcycle             1.764839
car_buy              1.764839
hnd_price            0.957840
change_mou           0.607135
rev_Mean             0.218719
mou_Range            0.218719
rev_Range            0.218719
ovrrev_Mean          0.218719
totmrc_Mean          0.218719
roam_Mean            0.218719
ovrmou_Mean          0.218719
datovr_Range         0.218719
datovr_Mean          0.218719
da_Range             0.218719
da_Mean              0.218719
mou_Mean             0.218719
area                 0.018855
csa                  0.018855
dtype: float64
```

Fig 2.5 Percentage of Missing values in columns list

D. Removal of unwanted variables

"CUSTOMERID" variable has been removed as it does not hold any importance.

E. Outlier Treatment

Boxplots are used to detect the outliers.

Boxplots for all numerical variables have been plotted and it is observed that all the continuous variables have outliers present.

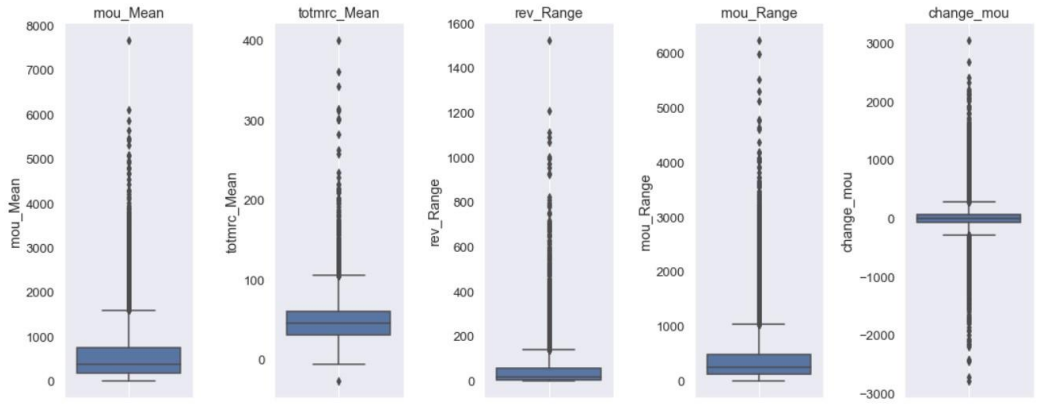


Fig 2.5 Sample of outliers present in first 5 columns

Outlier Treatment Approach:

1. Outlier treatment is done by calculating lower bound and upper bound of the column and then treating the outliers accordingly.
2. Created a function to calculate the lower bound and upper bound of the data and this is applied to all the continuous variables columns.

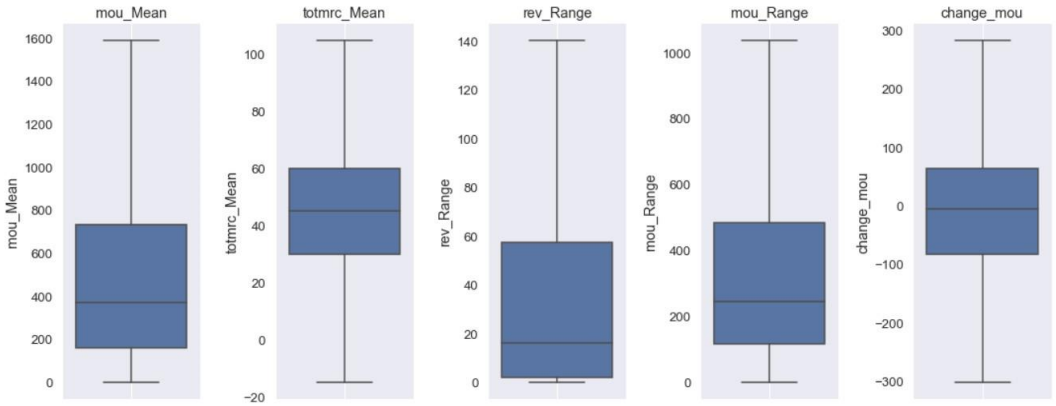


Fig 2.5 Boxplot of first 5 columns after outlier treatment

F. Variable Transformation

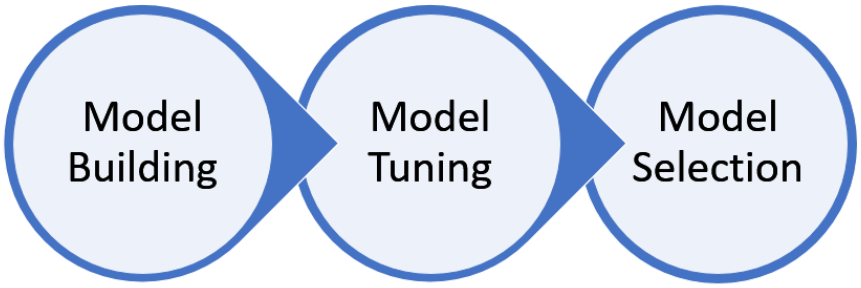
Linear and Tree based algorithms are applied on the dataset. It is observed that 3 types of categorical variables are present in the dataset: nominal, ordinal or Boolean types.

Hot encoding is applied to nominal variables for distance based algorithms such as Naïve Bayes, Logistic Regression etc;. And for the rest of the variables, Label Encoding has been performed.

Two datasets have been created and they are used separately for Distance and Tree based algorithms. There are 69 features in Tree dataset and 187 features in Distance Algorithms Dataset.

5.Model Building and Tuning

5.1. Model Approach



Various Models have been used on scaled and unscaled dataset. Hyperparamter tuning has been done wherever necessary. There is imbalance in the dataset and the models were built on SMOTE datasets as well. Most optimum model has been evaluated using F1-scores, Precision and Recall. These models were built using sklearn and statsmodel libraries.

Initially, the models were built on unscaled and scaled datasets. But the model's performance was very poor in terms of F1 Score. Models could not predict customers who are going to churn and could only predict the customers who would not churn with high F1 and recall scores.

After analyzing the dataset again, it has been found out that the model's poor F1 score performance was due to the imbalance present in the dataset. There is a minority class present.

Therefore, SMOTE datasets were created and the models were applied on these datasets and this approach has improved the F1 score to 20-30%, which is a moderate improvement.

5.2.Model Building on Unscaled and Scaled datasets

Step 1: Split of Dataset

The data set is split into 80% training data and 30% test data. The "Churn" column is defined as the class (y_train), the remaining columns as the features ("X_Train").

The following models are built on the dataset:

1. Naïve Bayes
2. Logistic Regression
3. KNN
4. Linear Discriminant Analysis
5. Random Forest
6. Support Vector Machine
7. XGBoost

.

Naïve Bayes:

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Naïve Bayes classifier from Sklearn is built on the dataset.

```
model = GaussianNB()
```

Fig 1.1 Naïve Bayes Sklearn function

This Algorithm has been used on an unscaled dataset.

Performance Metrics of Naïve Bayes Model on Training Set

	precision	recall	f1-score	support
0	0.80	0.66	0.72	13512
1	0.31	0.49	0.38	4255
accuracy			0.62	17767
macro avg	0.56	0.57	0.55	17767
weighted avg	0.68	0.62	0.64	17767

Fig 1.2. Classification Report of Training Dataset

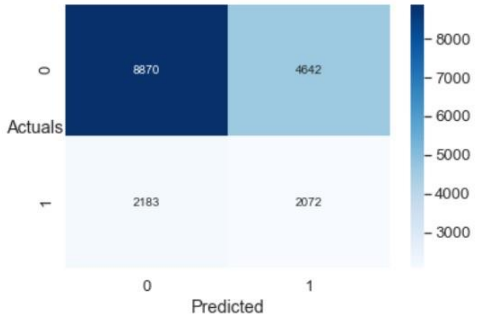


Fig 1.3. Confusion matrix of Training Dataset

Performance Metrics of Naïve Bayes Model on Testing Dataset

	precision	recall	f1-score	support
0	0.80	0.66	0.72	6642
1	0.31	0.48	0.38	2109
accuracy			0.62	8751
macro avg	0.55	0.57	0.55	8751
weighted avg	0.68	0.62	0.64	8751

Fig 1.4. Classification Report of Testing Dataset

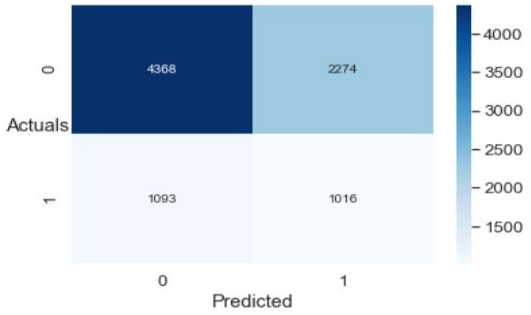


Fig 1.5. Confusion Matrix of Testing Dataset

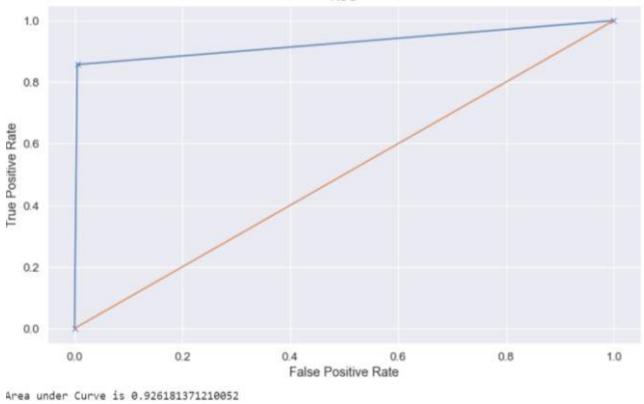


Fig 1.6. AUC /ROC score/plot of Testing Dataset

Interpretation: Naïve Bayes is built on the training dataset and further evaluated on the testing dataset. This model has achieved F1- Score of 0.38 and recall of 0.48 for predicting churners of the company.

However, compared to the churners prediction, it has predicted high percentage of nonchurners correctly.

Logistic Regression

It's a fast model to learn and effective on binary classification problems. Support Vector Machine (SVM) is a **supervised** learning algorithm and mostly used for classification tasks but it is also suitable for regression tasks.

Performance Metrics of Logistic Regression Model on Training Set

	precision	recall	f1-score	support
0	0.76	1.00	0.86	13512
1	0.49	0.01	0.03	4255
accuracy			0.76	17767
macro avg	0.63	0.50	0.45	17767
weighted avg	0.70	0.76	0.66	17767

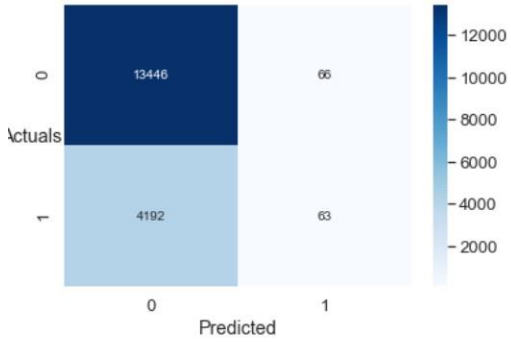


Fig 1.7. Classification Report of Training Dataset **Fig 1.8. Confusion Matrix of Training Dataset**

Performance Metrics of Logistic Regression Model on Testing Dataset

	precision	recall	f1-score	support
0	0.76	0.99	0.86	6642
1	0.44	0.01	0.03	2109
accuracy			0.76	8751
macro avg	0.60	0.50	0.44	8751
weighted avg	0.68	0.76	0.66	8751

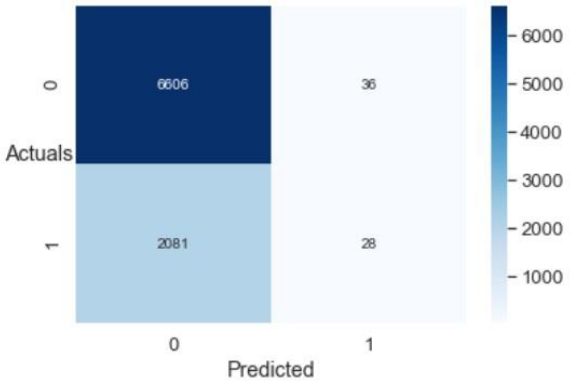


Fig 1.8. Classification Report of Testing Dataset **Fig 1.9 Confusion Matrix of Testing Dataset**

Interpretation: Sklearn's Logistic regression model is built on the training dataset and further evaluated on the testing dataset. This model has achieved F1- Score of 0.03 and recall of 0.01 for predicting churners of the company. This model's performance is very poor in predicting the churners. Hence, we will evaluate this model further by tuning the hyper parameters in the next section. However, compared to the churners prediction, it has predicted high percentage of non-churners correct.

KNN Model

Performance Metrics of KNN Model on Training Set

	precision	recall	f1-score	support
0	0.80	0.96	0.87	13512
1	0.65	0.25	0.36	4255
accuracy			0.79	17767
macro avg	0.73	0.60	0.62	17767
weighted avg	0.77	0.79	0.75	17767

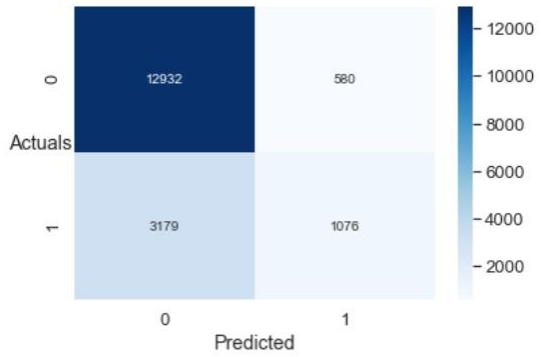


Fig 1.10. Classification Report of Training Dataset **Fig 1.11. Confusion Matrix of Training Dataset**

Performance Metrics of KNN Model on Testing Dataset

	precision	recall	f1-score	support
0	0.76	0.92	0.83	6642
1	0.30	0.11	0.16	2109
accuracy			0.72	8751
macro avg	0.53	0.51	0.50	8751
weighted avg	0.65	0.72	0.67	8751

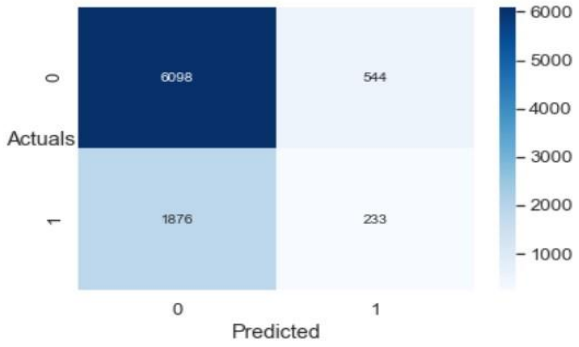


Fig 1.12. Classification Report of Testing Dataset **Fig 1.13. Confusion Matrix of Testing Dataset**

Interpretation: KNN model is built on the training dataset and further evaluated on the testing dataset.

This model has achieved F1- Score of 0.16 and recall of 0.11 for predicting churners of the company. This model's performance is poor in predicting the churners but better when compared to Logistic Regression's model's performance.

Hence, we will evaluate this model further by tuning the hyperparameters in the next section. However, compared to the churners prediction, it has predicted high percentage of nonchurners correct. And the overall accuracy rate seems to be good.

LDA Model

Performance Metrics of LDA Model on Training Set

	precision	recall	f1-score	support
0	0.76	0.99	0.86	13512
1	0.51	0.04	0.07	4255
accuracy			0.76	17767
macro avg	0.64	0.51	0.46	17767
weighted avg	0.70	0.76	0.67	17767

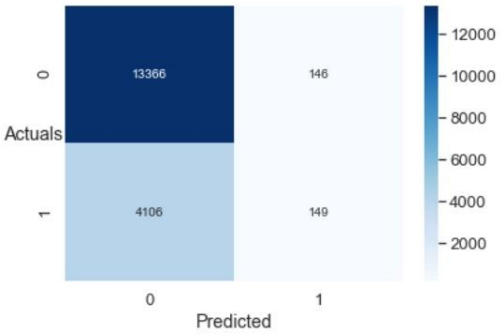


Fig 1.14. Classification Report of Training Dataset **Fig 1.15. Confusion Matrix of Training Dataset**

Performance Metrics of LDA Model on Testing Dataset

	precision	recall	f1-score	support
0	0.76	0.99	0.86	6642
1	0.44	0.03	0.06	2109
accuracy			0.76	8751
macro avg	0.60	0.51	0.46	8751
weighted avg	0.68	0.76	0.67	8751

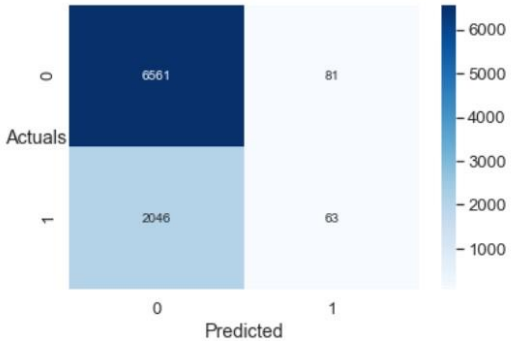


Fig 1.16. Classification Report of Testing Dataset **Fig 1.17. Confusion Matrix of Testing Dataset**

Interpretation: KNN model is built on the training dataset and further evaluated on the testing dataset.

This model has achieved F1- Score of 0.16 and recall of 0.11 for predicting churners of the company. This model's performance is poor in predicting the churners but better when compared to Logistic Regression's model's performance.

Hence, we will evaluate this model further by tuning the hyperparameters in the next section. However, compared to the churners prediction, it has predicted high percentage of nonchurners correct. And the overall accuracy rate seems to be good.

SVM Model

Performance Metrics of SVM Model on Training Set

	precision	recall	f1-score	support
0	0.76	1.00	0.87	13512
1	1.00	0.01	0.02	4255
accuracy			0.76	17767
macro avg	0.88	0.51	0.44	17767
weighted avg	0.82	0.76	0.66	17767

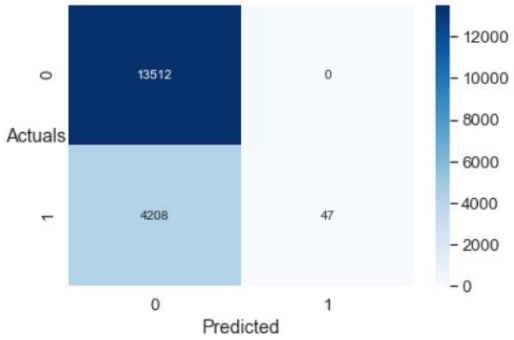


Fig 1.18. Classification Report of Training Dataset **Fig 1.19. Confusion Matrix of Training Dataset**

Performance Metrics of SVM Model on Testing Dataset

	precision	recall	f1-score	support
0	0.76	1.00	0.86	6642
1	0.00	0.00	0.00	2109
accuracy			0.76	8751
macro avg	0.38	0.50	0.43	8751
weighted avg	0.58	0.76	0.65	8751

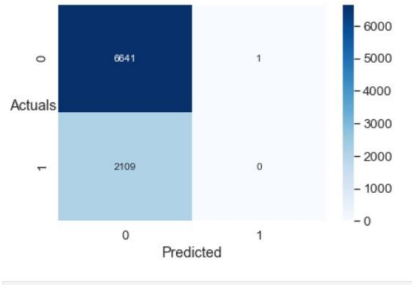


Fig 1.20. Classification Report of Testing Dataset **Fig 1.21. Confusion Matrix of Testing Dataset**

Interpretation: SVM model is built on the training dataset and further evaluated on the testing dataset.

This model has achieved F1- Score of 0 and recall of 0 for predicting churners of the company. This model's performance could not predict any of the churners and hence, this model will not be taken into consideration for this project and tuning the model wouldn't achieve a significant improvement in the scores

However, compared to the churners prediction, it has predicted high percentage of nonchurners correct. And the overall accuracy rate seems to be good but still it has been decided to not go further with this model.

Models are able to detect only non-churners and not the customers who are going to churn and hence we have used the models on SMOTE dataset.

5.3. Model Building on SMOTE unscaled datasets

Naïve Bayes:

Performance Metrics of Naïve Bayes Model on Training Set

	precision	recall	f1-score	support
0	0.65	0.13	0.22	13102
1	0.52	0.93	0.66	13102
accuracy			0.53	26204
macro avg	0.58	0.53	0.44	26204
weighted avg	0.58	0.53	0.44	26204

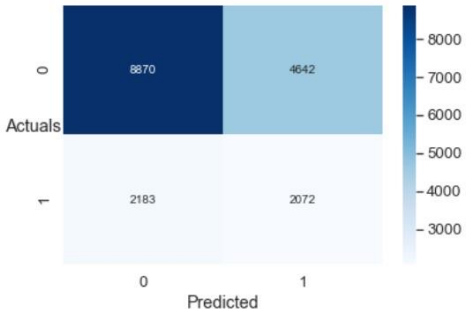


Fig 1.22. Classification Report of Training Dataset **Fig 1.23. Confusion matrix of Training Dataset**

Performance Metrics of Naïve Bayes Model on Testing Dataset

	precision	recall	f1-score	support
0	0.76	0.12	0.21	6391
1	0.25	0.88	0.38	2072
accuracy			0.31	8463
macro avg	0.50	0.50	0.30	8463
weighted avg	0.63	0.31	0.25	8463

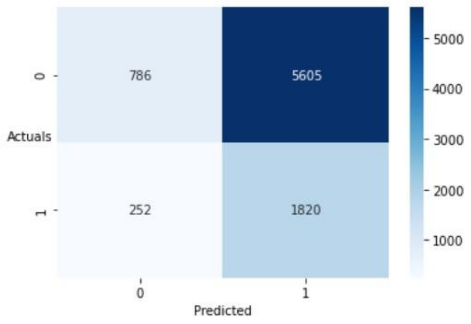


Fig 1.24. Classification Report of Testing Dataset **Fig 1.25. Confusion Matrix of Testing Dataset**

Logistic Regression

Performance Metrics of Logistic Regression Model on Testing Set

	precision	recall	f1-score	support
0	0.60	0.59	0.60	13102
1	0.60	0.61	0.60	13102
accuracy			0.60	26204
macro avg	0.60	0.60	0.60	26204
weighted avg	0.60	0.60	0.60	26204

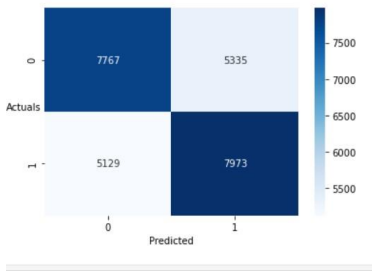


Fig 1.26. Classification Report of Testing Dataset **Fig 1.27. Confusion Matrix of Testing Dataset**

KNN Model

Performance Metrics of KNN Model on Training Set

	precision	recall	f1-score	support
0	0.95	0.69	0.80	13102
1	0.76	0.96	0.85	13102
accuracy			0.83	26204
macro avg	0.85	0.83	0.82	26204
weighted avg	0.85	0.83	0.82	26204

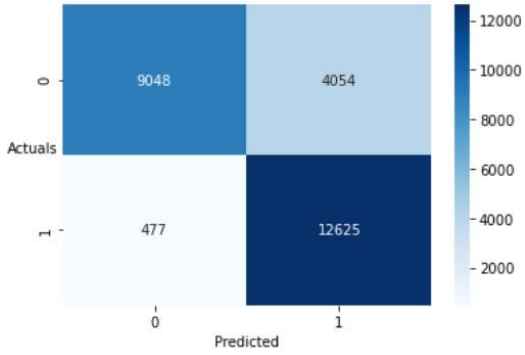


Fig 1.30. Classification Report of Training Dataset **Fig 1.31. Confusion Matrix of Training Dataset**

Performance Metrics of KNN Model on Testing Dataset

	precision	recall	f1-score	support
0	0.76	0.56	0.64	6391
1	0.25	0.47	0.33	2072
accuracy			0.53	8463
macro avg	0.51	0.51	0.49	8463
weighted avg	0.64	0.53	0.57	8463

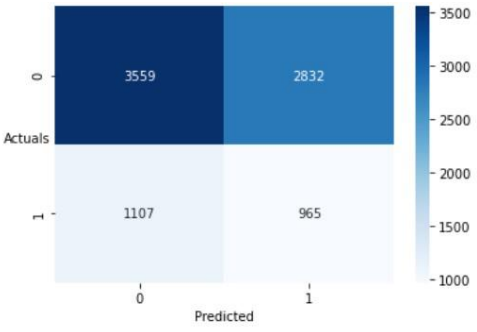


Fig 1.32. Classification Report of Testing Dataset **Fig 1.33. Confusion Matrix of Testing Dataset**

LDA Model

Performance Metrics of LDA Model on Training Set

	precision	recall	f1-score	support
0	0.71	0.68	0.70	13102
1	0.69	0.73	0.71	13102
accuracy			0.70	26204
macro avg	0.70	0.70	0.70	26204
weighted avg	0.70	0.70	0.70	26204

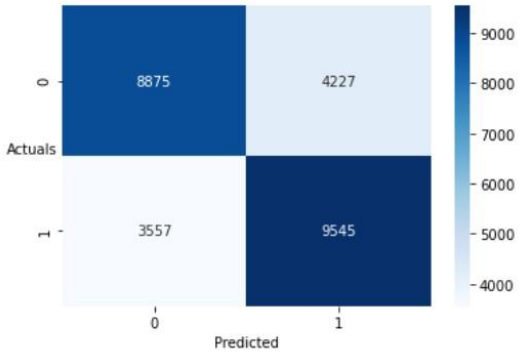


Fig 1.34. Classification Report of Training Dataset **Fig 1.35. Confusion Matrix of Training Dataset**

Performance Metrics of LDA Model on Testing Dataset

	precision	recall	f1-score	support
0	0.81	0.59	0.68	6391
1	0.31	0.56	0.40	2072
accuracy			0.58	8463
macro avg	0.56	0.57	0.54	8463
weighted avg	0.68	0.58	0.61	8463

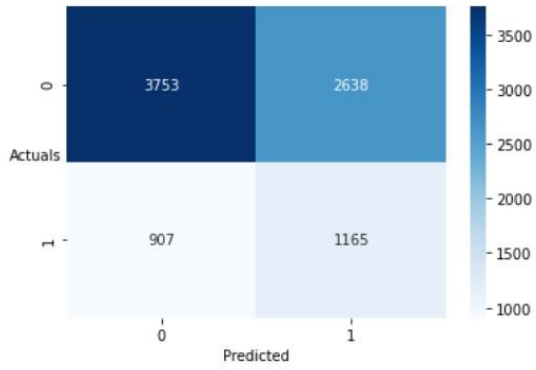


Fig 1.36. Classification Report of Testing Dataset **Fig 1.37. Confusion Matrix of Testing Dataset**

5.4. Model Tuning

The 3 model tuning techniques used are:

- Ensemble modelling
- Gridsearch CV
- SMOTE

Ensemble Modelling:

Random Forest, XGBBoost,AdaBoost ensemble algorithms are used. These models use multiple weak learners to make better predictions. But, the results were not that great.

Gridsearch CV:

Sklearn's GridSearchCV was used for hyper tuning the model parameters.

SMOTE:

To overcome the problem of High imbalance in the dataset, SMOTE was used. So far, this has brought comparatively better results.

All of the performance metrics of each model tuning method are shown above in the report.

Ensemble Modelling

RANDOM FOREST

Performance Metrics of Random Forest Model on Training Set

	precision	recall	f1-score	support
0	1.00	1.00	1.00	13512
1	1.00	1.00	1.00	4255
accuracy			1.00	17767
macro avg	1.00	1.00	1.00	17767
weighted avg	1.00	1.00	1.00	17767

Fig 1.38. Classification Report of Training Dataset

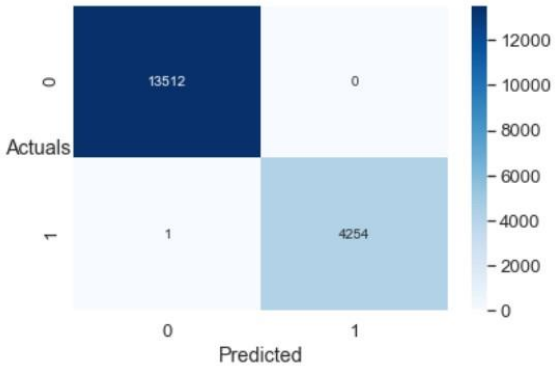


Fig 1.39. Confusion Matrix of Training Dataset

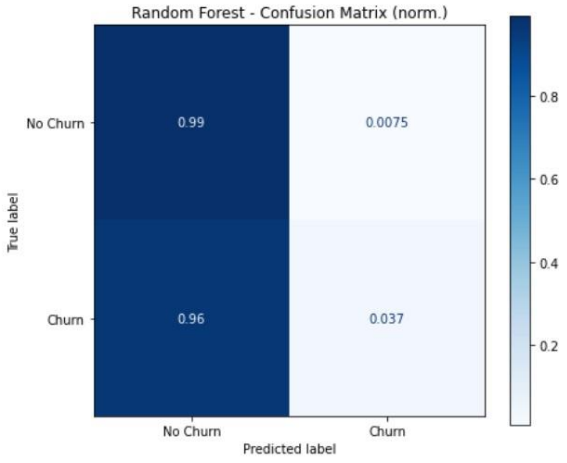


Fig 1.40. Confusion Matrix of Testing Dataset

Interpretation: Random Forest model is built on the training dataset and further evaluated on the testing dataset.

It is observed that the F1 score and precision scores has achieved a perfect score on training dataset. But however, Performance metrics on testing dataset is very poor. F1 score is very low i.e. 0.045. (Refer to Appendix)

This model could barely predict any of the churners correctly and hence, this model will not be taken into consideration for this project However, compared to the churners prediction, it has predicted high percentage of non-churners correct.

XGBoost Model

Performance Metrics of XGBoost Model on Training Set

	precision	recall	f1-score	support
0	0.77	1.00	0.87	13512
1	0.90	0.05	0.10	4255
accuracy			0.77	17767
macro avg	0.83	0.53	0.48	17767
weighted avg	0.80	0.77	0.68	17767

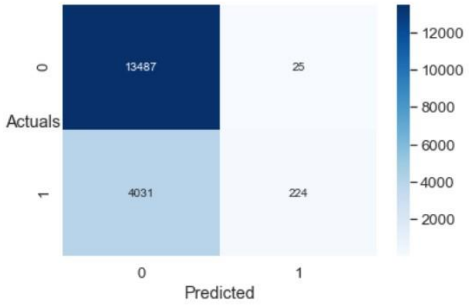


Fig 1.41. Classification Report of Training Dataset **Fig 1.42. Confusion Matrix of Training Dataset**

Performance Metrics of XGBoost Model on Testing Dataset

	precision	recall	f1-score	support
0	0.76	0.99	0.86	6642
1	0.52	0.02	0.05	2109
accuracy			0.76	8751
macro avg	0.64	0.51	0.45	8751
weighted avg	0.70	0.76	0.67	8751

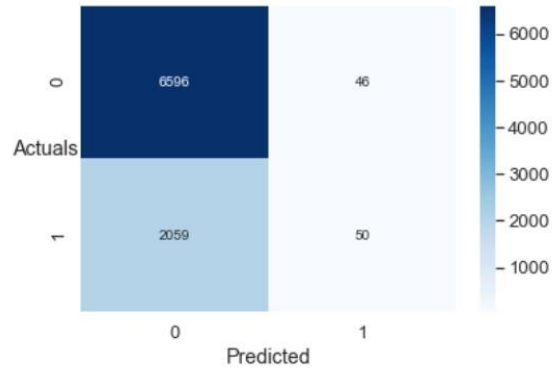


Fig 1.43. Classification Report of Testing Dataset **Fig 1.45. Confusion Matrix of Testing Dataset**

Interpretation: XGBoosting model is built on the training dataset and further evaluated on the testing dataset.

It is observed that the F1 score and precision scores is very low on training dataset and hence expected that the Performance metrics on testing dataset is very poor. F1 score is very low i.e.
0.05.

This model could barely predict any of the churners correctly and hence, this model will not be taken into consideration for this project However, compared to the churners prediction, it has predicted high percentage of non-churners correct.

- **Ada Boosting**

This algorithm has also been built. Refer to Appendix 14 for further reference of the evaluation results of this model. It has achieved a low score just like the other ensemble models.

6. Model Evaluation and Validation

- Precision: It is the number of positive prediction values divided by the total number of positive class values predicted. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative
- Recall: Recall is another metric that evaluates the total number of positive predictions divided by the number of positive class values. It's the intuitively the ability of the classifier to find all the positive samples.
- F1-Score: conveys the balance between the precision and the recall

Model Validation

Models comparison on Unscaled Dataset

	Naive Bayes	Logistic Regression	LDA	KNN	SVM	XGB	ANN	ADA
Precision	0.30	0.55	0.48	0.31	0.0	0.63	0.32	0.54
Recall	0.44	0.01	0.03	0.12	0.0	0.02	0.27	0.07
F1 Score	0.36	0.02	0.06	0.12	0.0	0.05	0.29	0.12
AUC Score	0.56	0.50	0.51	0.52	0.5	0.51	0.54	0.52

Accuracy rates of all these models are moderately good. F1-score of Naïve Bayes is the highest followed by ANN and KNN. All the other evaluation metrics for Naïve Bayes is high as well.

Models comparison on SMOTE dataset

Models	F1 Score	Recall	Precision	Accuracy
Naïve Bayes	0.38	0.88	0.25	0.31
Logistic Regression	0.41	0.61	0.31	0.59
KNN	0.33	0.47	0.25	0.53
LDA	0.40	0.56	0.31	0.58

Naïve Bayes has highest recall rate and F1-score comparated to other algorithms isn't low. But Naïve Bayes cannot be considered as an optimal model as it does not have a balance between the evaluation metrics. It has the lowest accuracy rate.

Therfore Naïve Bayes is not taken into consideration.

Most Optimal Model

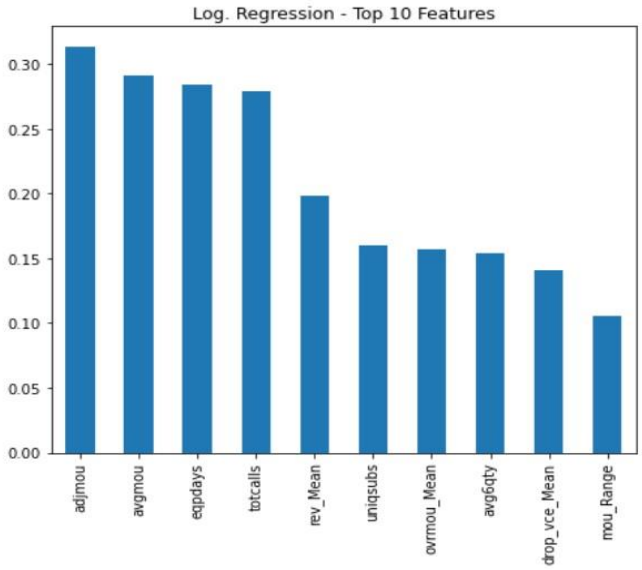
- Logistic Regression is considered to be the optimal model and it has the highest F1, Recall and accuracy scores.
- It has a good balance between the evaluation metrics.
- It is able to predict 1110 correct customer churns out of 1797 actual churns in the test data. Which means out of all the customers who will actually churn this model is able to predict 61% of them correctly

- This model is believed to the company predict the customers who are going to churn and will thus lead to customer retention. Which will in turn affect companies ARPU.

7. Business Insights and Recommendations

Top Line Questions of Interest to Senior Management:

1. What are the top five factors driving the likelihood of churn?



1. Adjmou: Billing adjusted total minutes of use over the life of the customer
2. Avgmou: Average monthly minutes of use over the life of the customer

3. Eqpdays: Number of days (age) of current equipment
4. Totcalls: Total number of calls over the life of the customer
5. Rev_Mean: Mean monthly revenue (charge amount)

All of these factors play an important role in customer churn rate. For example, if the customers who have old technology equipment in their household are more likely to churn.

As there is a chance that the telecom industry might not be able to give a good coverage and encounter problems in using the services.

In terms of avgmou factor, more the number of minutes of use of the services provided by the company means the customers are not likely to churn and the same goes with total number of customers.

In terms of rev_Mean, if the charge amount is high, customers are more likely to churn compared to others. This would lead to cost and billing issues as well and there are other factors that have high feature importance which suggests that cost and billing is an important factors influencing the customer churn.

2. Validation of independent survey's findings for the telecom industry. a) Whether "cost and billing" and "network and service quality" are important factors influencing churn behavior?

mou_mean, change_mou, avgmou,rev_range, totmrc_mean, avgrev and ovrmou_Mean are among the 15 most important features. Customers having higher usage, hence higher bill amount are more likely to churn compared to the others

drop_vce_Range, drop_blk_Mean features are not important features and hence network and service quality is not the most important factor.

3. Would you recommend rate plan migration as a proactive retention strategy?

- Rate plan migration strategy is recommended for the telecom operator.

Following this strategy will help the operator reduce customer churn and will also help arrest the dropping ARPU.

4. What would be your recommendation on how to use this churn model for prioritization of customers for a proactive retention campaign in the future?

- You can use the most optimal model to predict the churners and save/extract the them to a target customers file for prioritization of customers for a proactive retention campaign in the future

5. What would be the target segments for proactive retention campaigns?

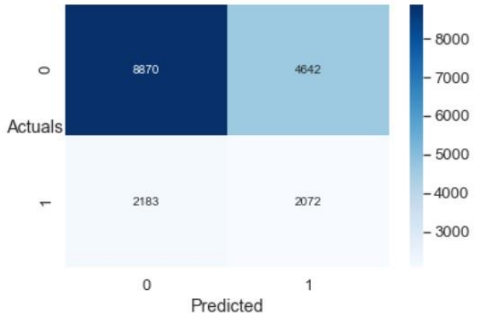
Below are few customer target segments for proactive retention campaigns:

- Customers who have the highest usage.
- Customers group that have been with the company for more than a year or so could be retained as well by offering loyalty programs.
- crclscod is within the top few factors influencing churn. Customers belonging to credit class 2 are more likely to churn, therefore can be retained.

Appendices

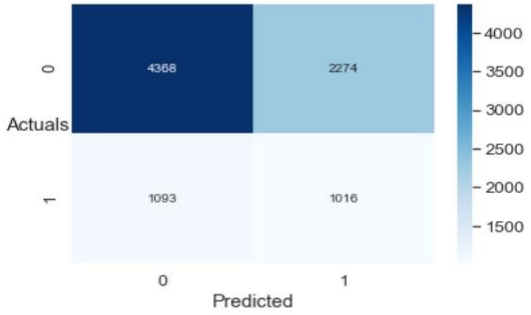
Appendix 1: Performance metrics of Training dataset: Naïve Bayes

	precision	recall	f1-score	support
0	0.80	0.66	0.72	13512
1	0.31	0.49	0.38	4255
accuracy			0.62	17767
macro avg	0.56	0.57	0.55	17767
weighted avg	0.68	0.62	0.64	17767



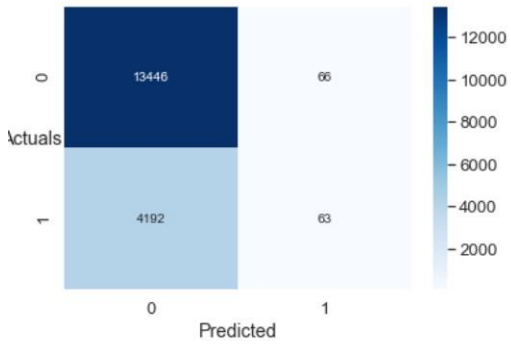
Appendix 2: Performance metrics of Testing dataset: Naïve Bayes

	precision	recall	f1-score	support
0	0.80	0.66	0.72	6642
1	0.31	0.48	0.38	2109
accuracy			0.62	8751
macro avg	0.55	0.57	0.55	8751
weighted avg	0.68	0.62	0.64	8751



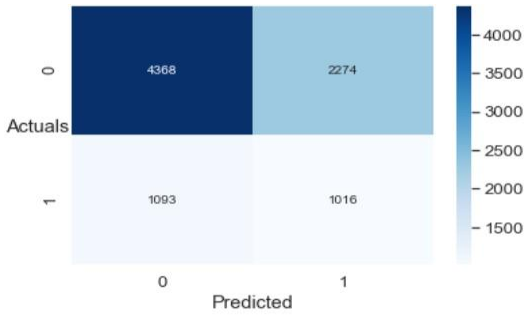
Appendix 3:Performance Metrics of Logistic Regression Model on Training Set

	precision	recall	f1-score	support
0	0.76	1.00	0.86	13512
1	0.49	0.01	0.03	4255
accuracy			0.76	17767
macro avg	0.63	0.50	0.45	17767
weighted avg	0.70	0.76	0.66	17767



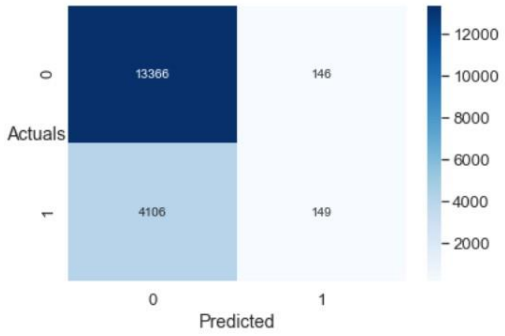
Appendix 4: Performance Metrics of Logistic Regression Model on Testing Dataset

	precision	recall	f1-score	support
0	0.76	0.99	0.86	6642
1	0.44	0.01	0.03	2109
accuracy			0.76	8751
macro avg	0.60	0.50	0.44	8751
weighted avg	0.68	0.76	0.66	8751



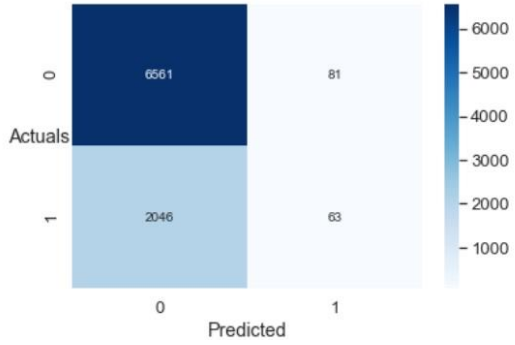
Appendix 5: Performance Metrics of LDA Model on Training Set

	precision	recall	f1-score	support
0	0.76	0.99	0.86	13512
1	0.51	0.04	0.07	4255
accuracy			0.76	17767
macro avg	0.64	0.51	0.46	17767
weighted avg	0.70	0.76	0.67	17767



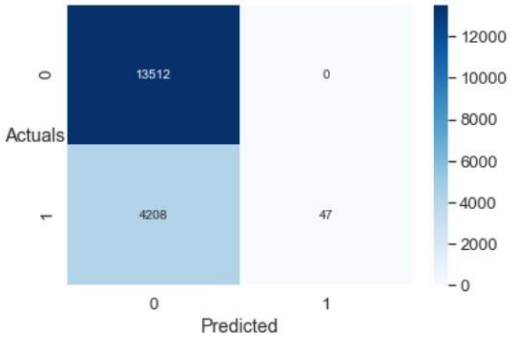
Appendix 6: Performance Metrics of LDA Model on Testing Dataset

	precision	recall	f1-score	support
0	0.76	0.99	0.86	6642
1	0.44	0.03	0.06	2109
accuracy			0.76	8751
macro avg	0.60	0.51	0.46	8751
weighted avg	0.68	0.76	0.67	8751



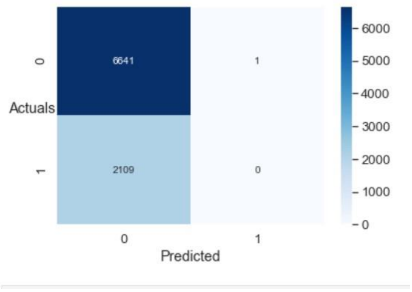
Appendix 7: Performance Metrics of SVM Model on Training Set

	precision	recall	f1-score	support
0	0.76	1.00	0.87	13512
1	1.00	0.01	0.02	4255
accuracy			0.76	17767
macro avg	0.88	0.51	0.44	17767
weighted avg	0.82	0.76	0.66	17767



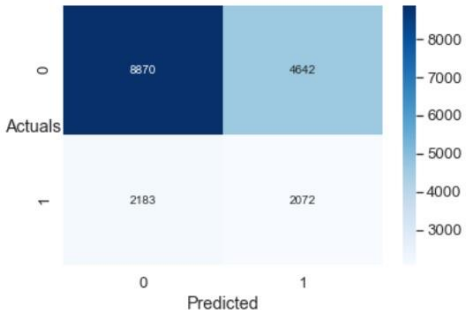
Appendix 8: Performance Metrics of SVM Model on Testing Dataset

	precision	recall	f1-score	support
0	0.76	1.00	0.86	6642
1	0.00	0.00	0.00	2109
accuracy			0.76	8751
macro avg	0.38	0.50	0.43	8751
weighted avg	0.58	0.76	0.65	8751



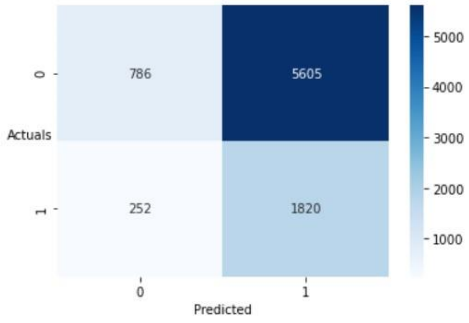
Appendix 9: Performance Metrics of Naïve Bayes (smote)Model on Training Set

	precision	recall	f1-score	support
0	0.65	0.13	0.22	13102
1	0.52	0.93	0.66	13102
accuracy			0.53	26204
macro avg	0.58	0.53	0.44	26204
weighted avg	0.58	0.53	0.44	26204



Appendix 10: Performance Metrics of Naïve Bayes Model (SMOTE)on Testing Dataset

	precision	recall	f1-score	support
0	0.76	0.12	0.21	6391
1	0.25	0.88	0.38	2072
accuracy			0.31	8463
macro avg	0.50	0.50	0.30	8463
weighted avg	0.63	0.31	0.25	8463



Appendix 11: Performance Metrics of LDA Model on Training Set

	precision	recall	f1-score	support
0	0.71	0.68	0.70	13102
1	0.69	0.73	0.71	13102
accuracy			0.70	26204
macro avg	0.70	0.70	0.70	26204
weighted avg	0.70	0.70	0.70	26204

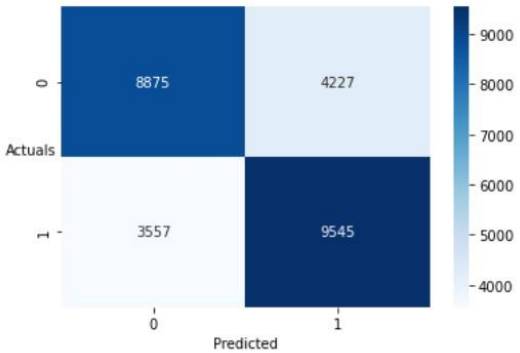


Fig 1.2. Classification Report of Training Dataset

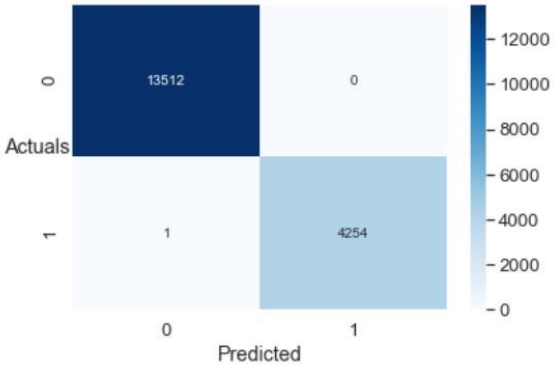
Fig 1.2. Confusion Matrix of Training Dataset

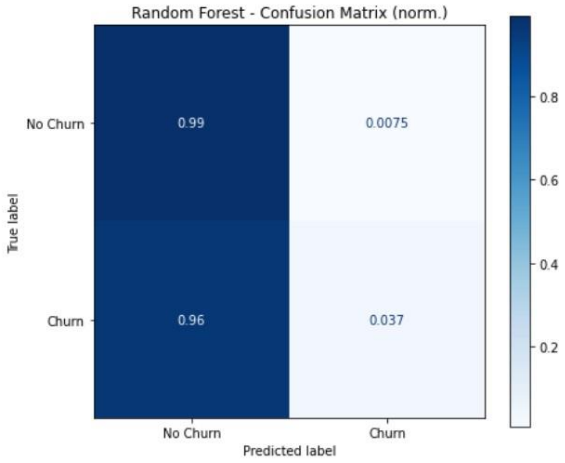
Appendix 12: Performance Metrics of LDA Model on Testing Dataset

	precision	recall	f1-score	support
0	0.81	0.59	0.68	6391
1	0.31	0.56	0.40	2072
accuracy			0.58	8463
macro avg	0.56	0.57	0.54	8463
weighted avg	0.68	0.58	0.61	8463

Appendix 13: Performance Metrics of Random Forest Model on Training Set

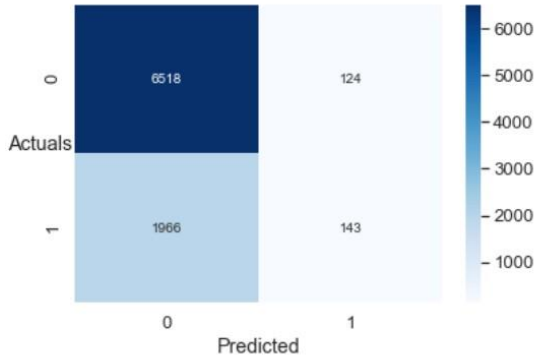
	precision	recall	f1-score	support
0	1.00	1.00	1.00	13512
1	1.00	1.00	1.00	4255
accuracy			1.00	17767
macro avg	1.00	1.00	1.00	17767
weighted avg	1.00	1.00	1.00	17767



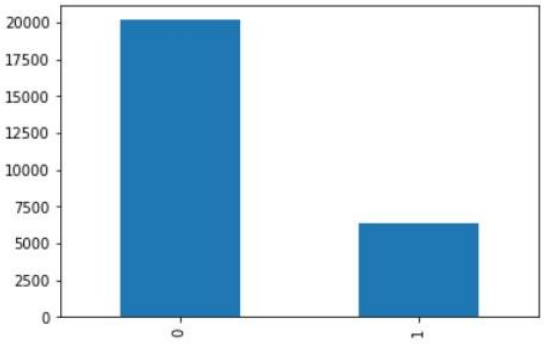


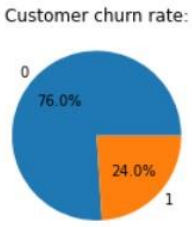
Appendix 14: Ada Boosting Results

	precision	recall	f1-score	support
0	0.77	0.98	0.86	6642
1	0.54	0.07	0.12	2109
accuracy			0.76	8751
macro avg	0.65	0.52	0.49	8751
weighted avg	0.71	0.76	0.68	8751



Appendix 1.1 Univariate Analysis : Categorical





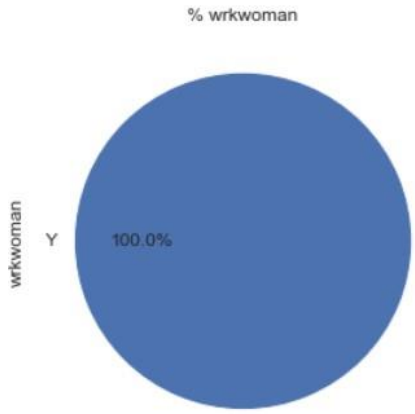


Fig 1.4 Pie chart of wrkwoman variable

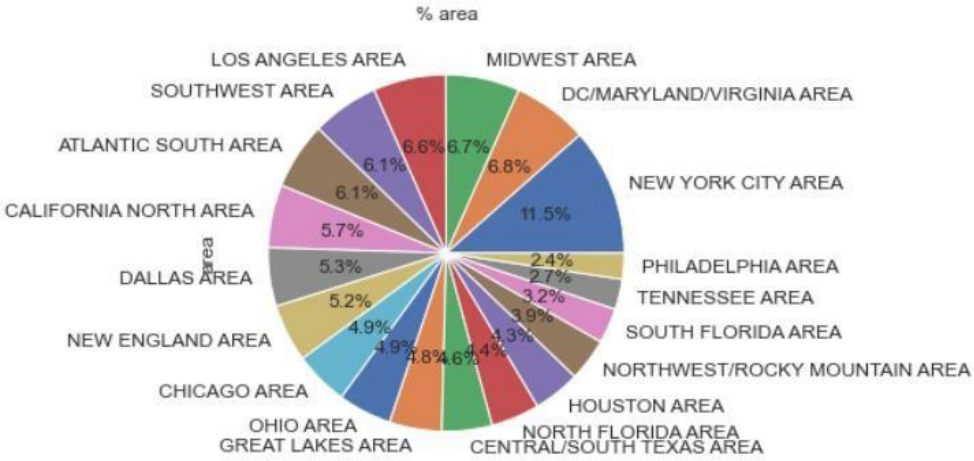


Fig 1.5 Pie chart of area variable

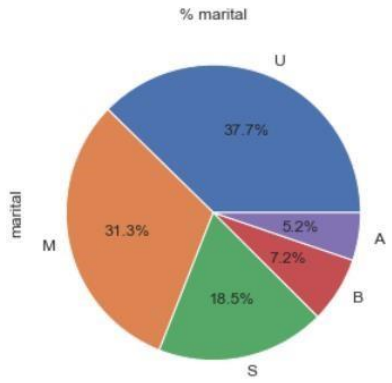


Fig 1.6 Pie chart of area variable

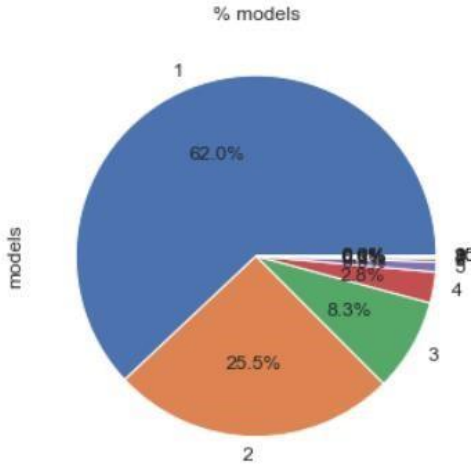


Fig 1.7 Pie chart of area variable

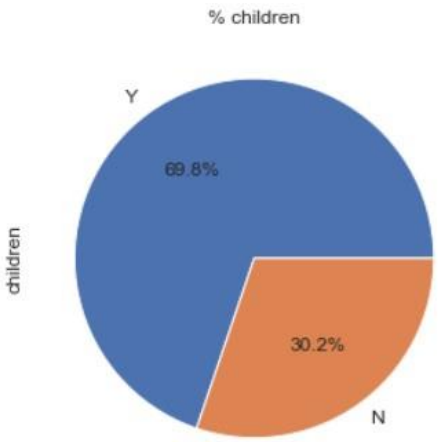


Fig 1.8 Pie chart of children variable

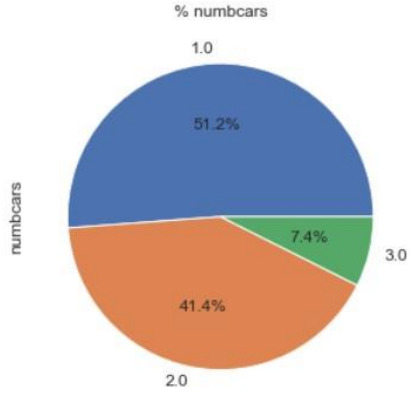


Fig 1.9 Pie chart of numbcars variable

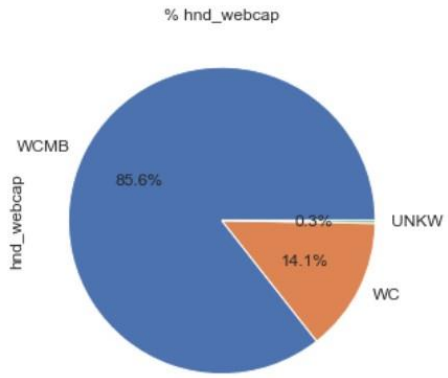


Fig 1.10 Pie chart of hnd_webcap variable

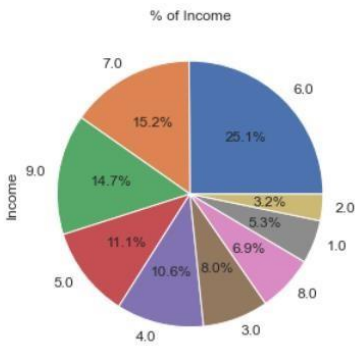


Fig 1.11 Pie chart of income variable

Appendix 1.2 Univariate Analysis : Continuous variables

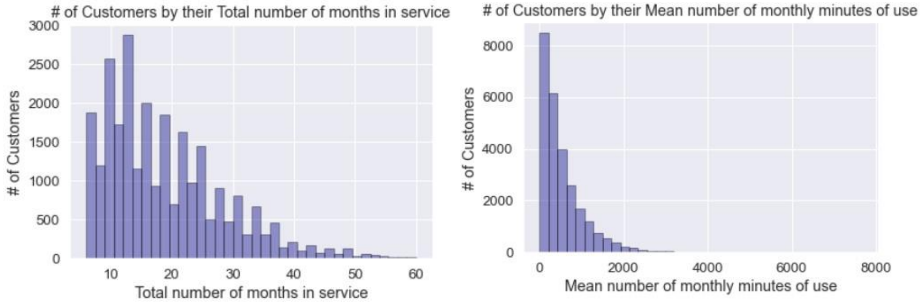


Fig 1.12 Dist plot of months and monthly mins variables

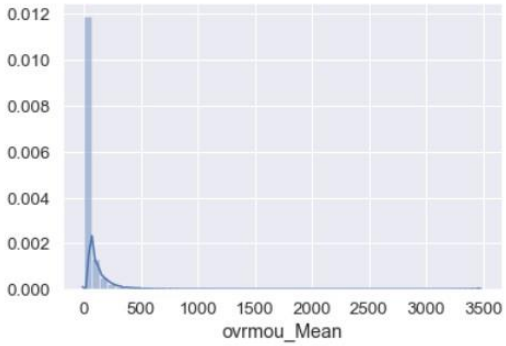


Fig 1.13 Dist plot of ovrmou_Mean variable

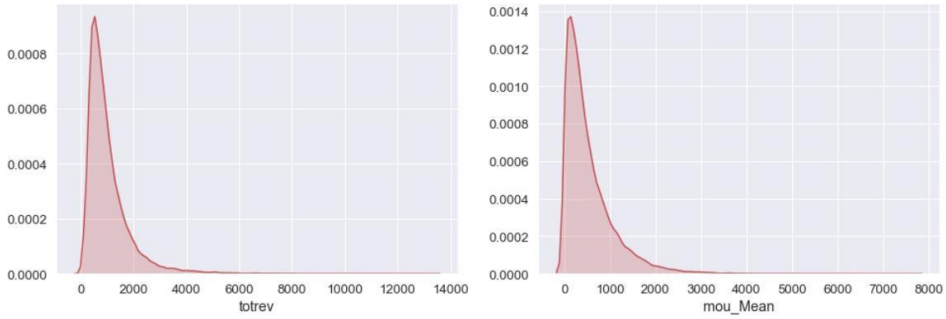


Fig 1.14 Distribution plot of totrev and mou_mean

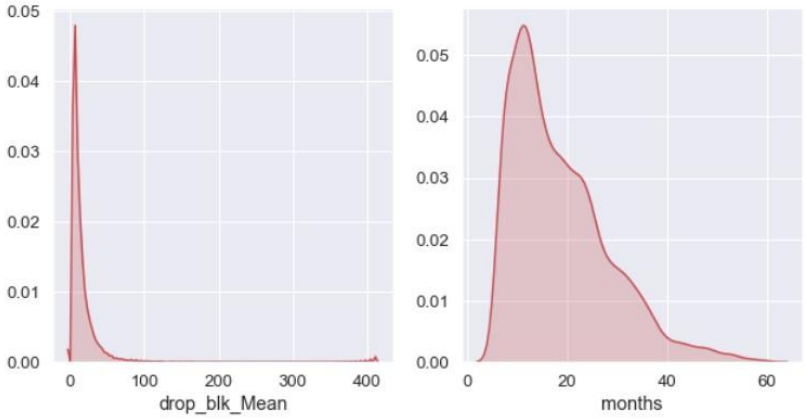


Fig 1.15 Distribution plot of drop_blk_Mean and months

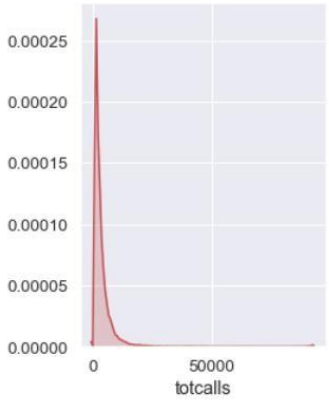
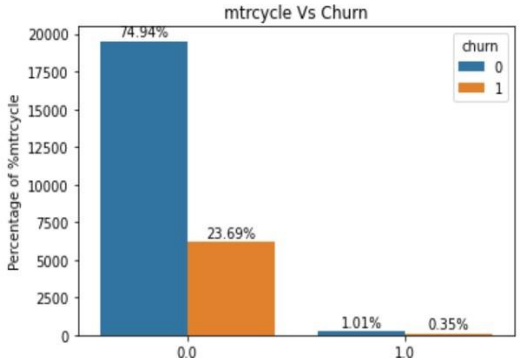
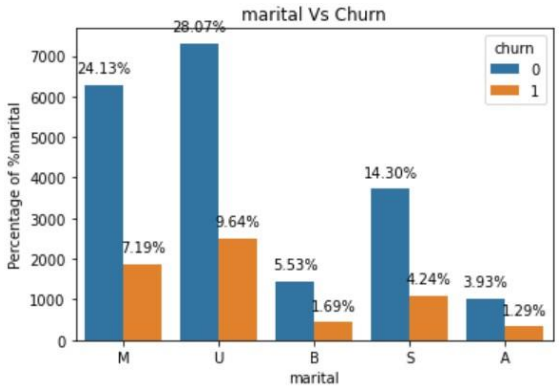
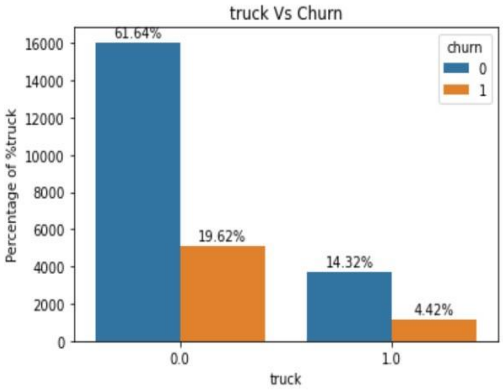


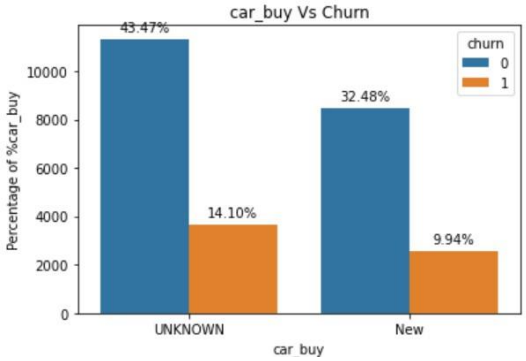
Fig 1.16. Distribution plot of totcalls

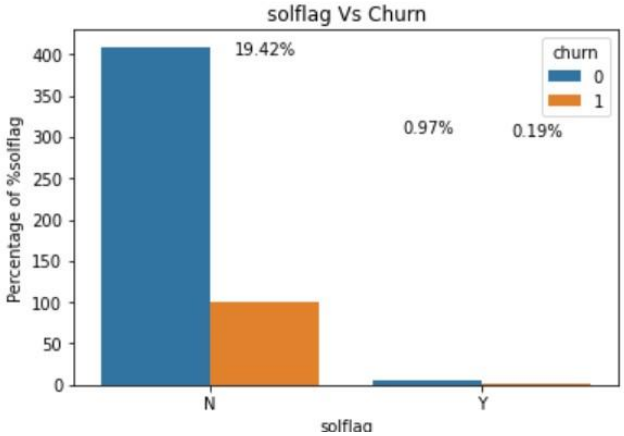
Appendix 1.3 Bivariate Analysis : Categorical variables

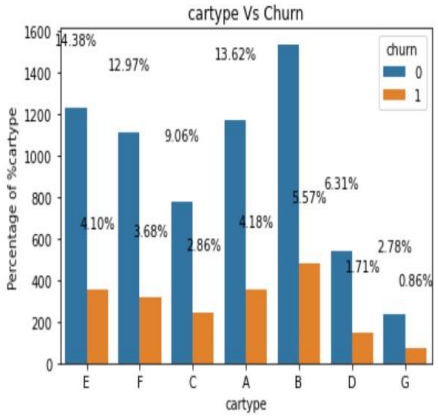

mtrcycle Vs Churn
Percentage of %mtrcycle
churn
0
1
74.94%
23.69%
1.01%
0.35%
0.0
1.0

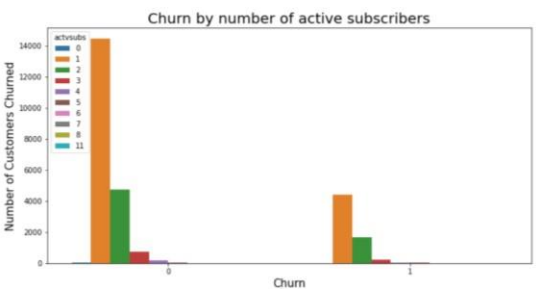

marital Vs Churn
Percentage of %marital
churn
0
1
24.13%
7.19%
28.07%
9.64%
5.53%
1.69%
14.30%
4.24%
3.93%
1.29%
M
U
B
S
A
marital

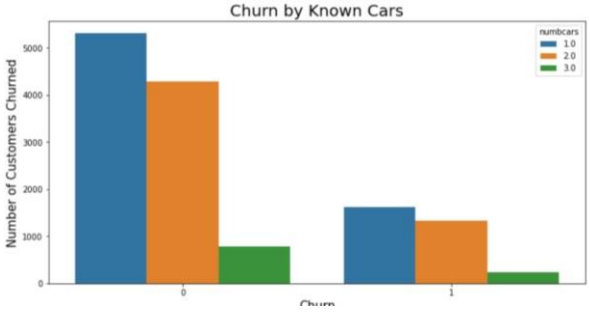

truck Vs Churn
Percentage of %truck
churn
0
1
61.64%
19.62%
14.32%
4.42%
0.0
1.0
truck

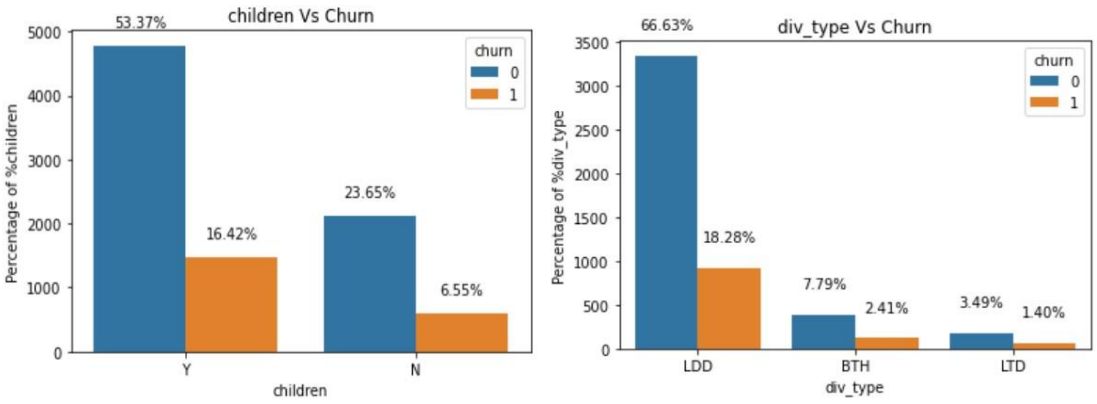

car_buy Vs Churn
Percentage of %car_buy
churn
0
1
43.47%
14.10%
32.48%
9.94%
UNKNOWN
New
car_buy


solflag Vs Churn
Percentage of %solflag
churn
0
1
19.42%
0.97%
0.19%
N
Y
solflag

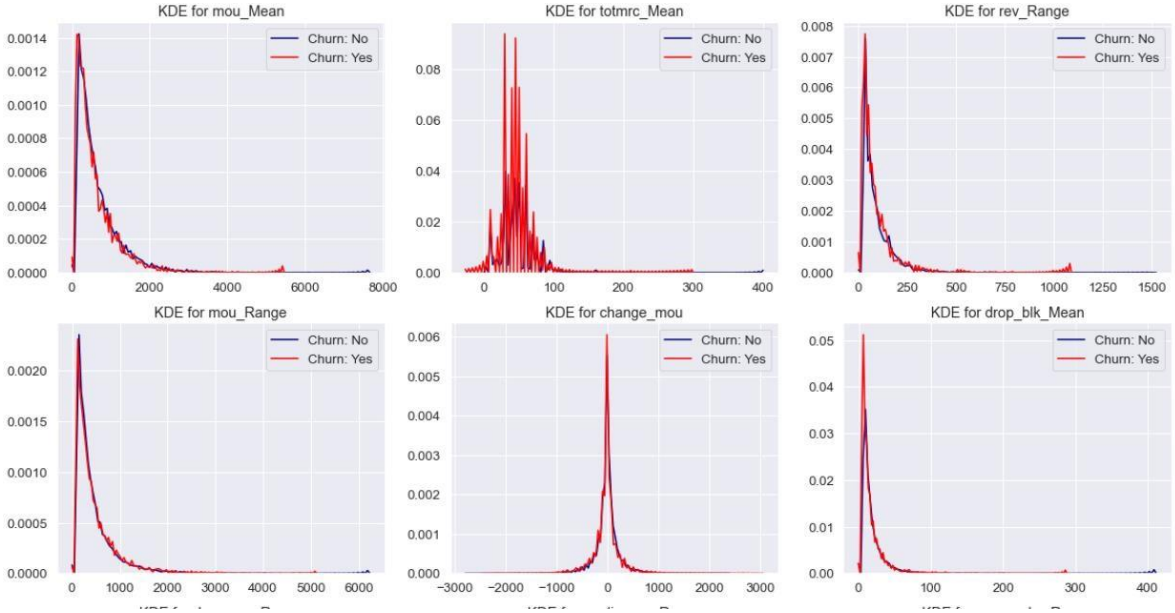

cartype Vs Churn
Percentage of %cartype
churn
0
1
14.38%
4.10%
12.97%
3.68%
9.06%
2.86%
13.62%
4.18%
5.57%
6.31%
1.71%
2.78%
0.86%
E
F
C
A
B
D
G
cartype

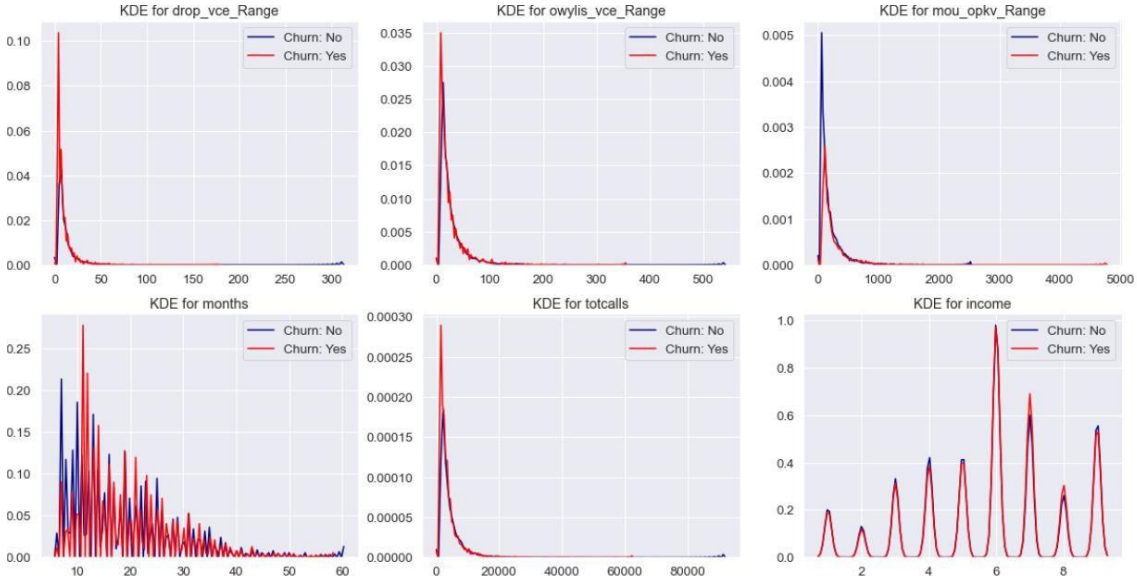

Churn by number of active subscribers
Number of Customers Churned
actvsubs
0
1
2
3
4
5
6
7
8
11
Churn

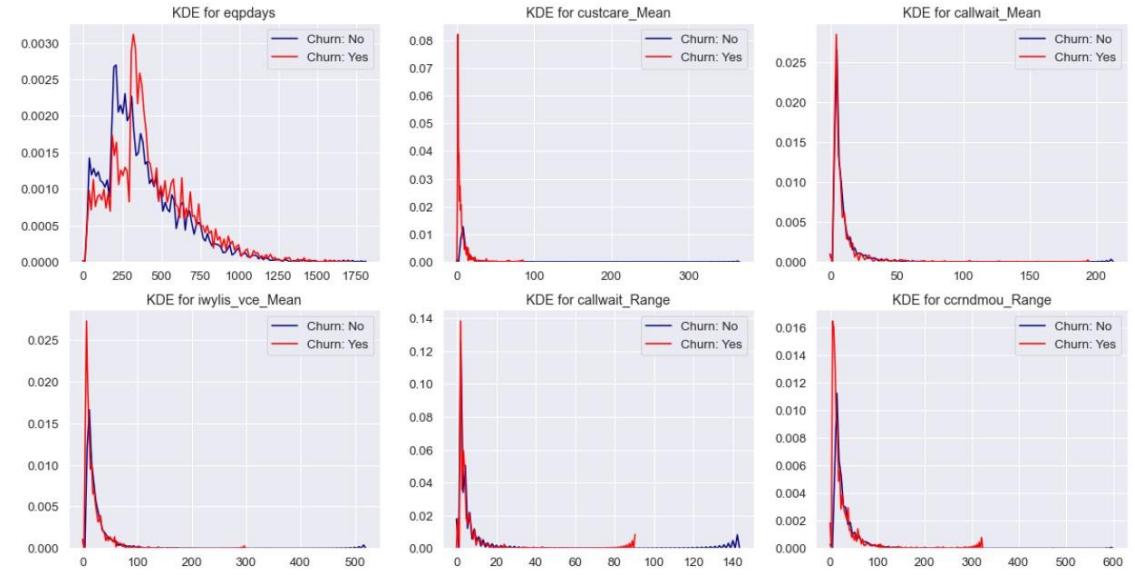

Churn by Known Cars
Number of Customers Churned
numbcars
1.0
2.0
3.0

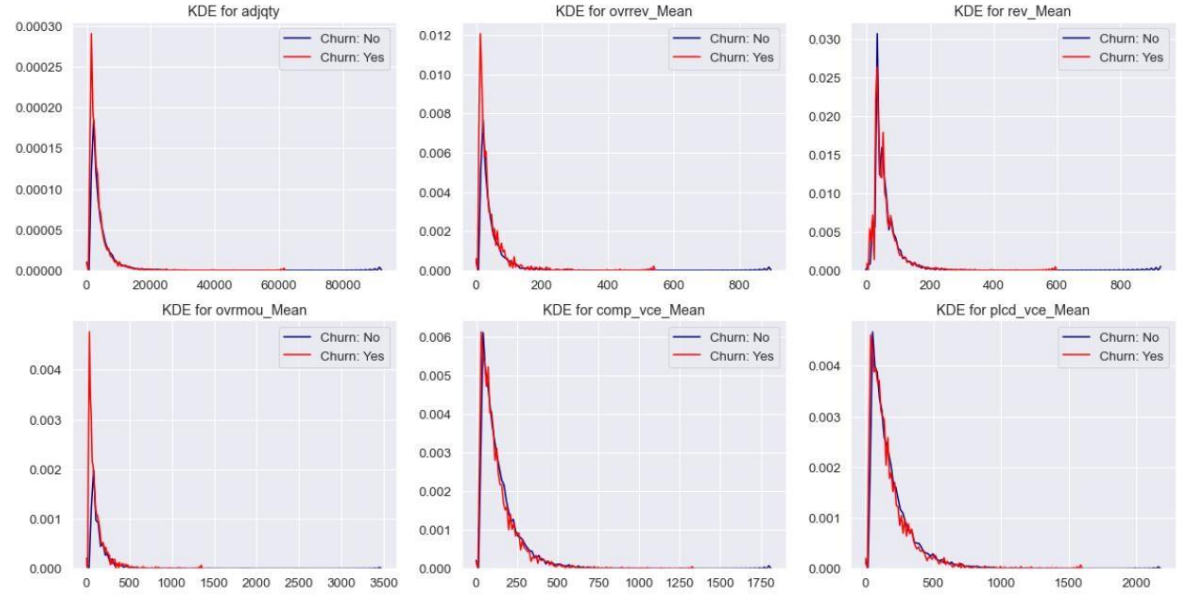


Appendix 1.4 Bivariate Analysis : Continuous variables




KDE for drop_vce_Range
KDE for owylis_vce_Range
KDE for mou_opkv_Range
KDE for months
KDE for totcalls
KDE for income
Churn: No
Churn: Yes


KDE for eqpdays
KDE for custcare_Mean
KDE for callwait_Mean
KDE for iwylis_vce_Mean
KDE for callwait_Range
KDE for ccrndmou_Range
Churn: No
Churn: Yes


KDE for adjqty
KDE for ovrrev_Mean
KDE for rev_Mean
KDE for ovrmou_Mean
KDE for comp_vce_Mean
KDE for plcd_vce_Mean
Churn: No
Churn: Yes

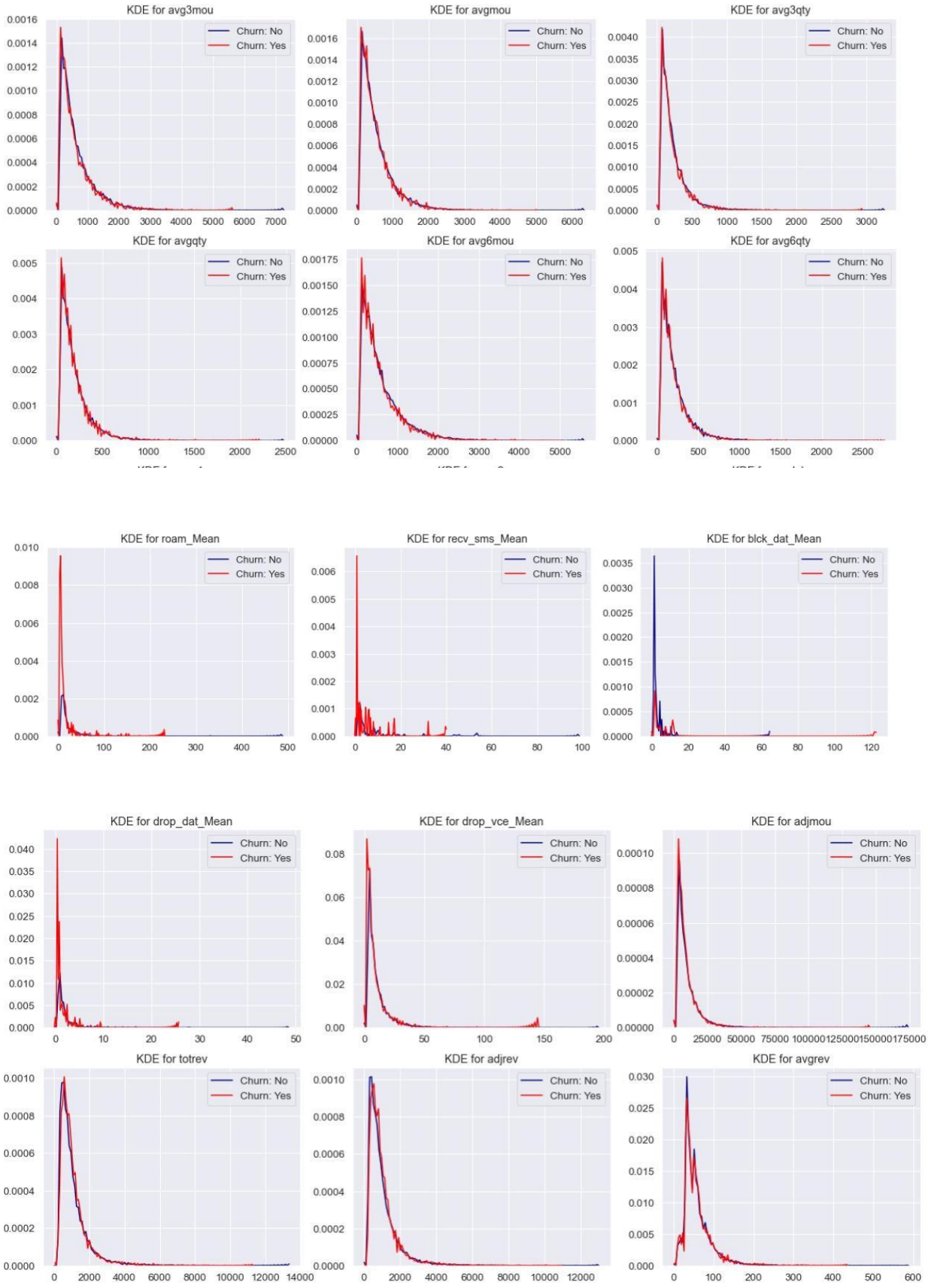


Fig 2.2 Dist plots of numerical variables w.r.t churn

Appendix 1.5 Heat map

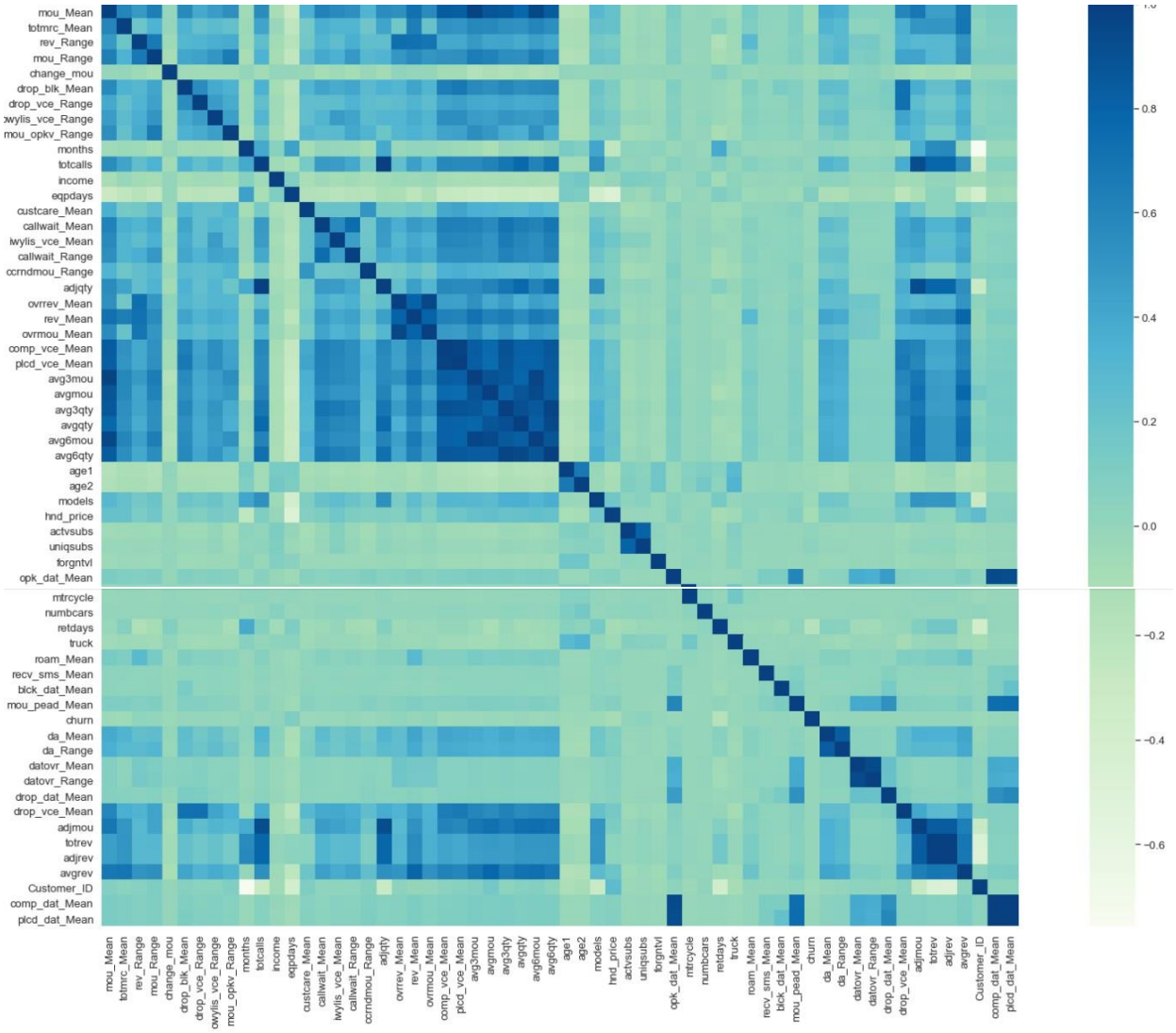


Fig 2.3 Heat Map

Appendix 1.6 Outliers

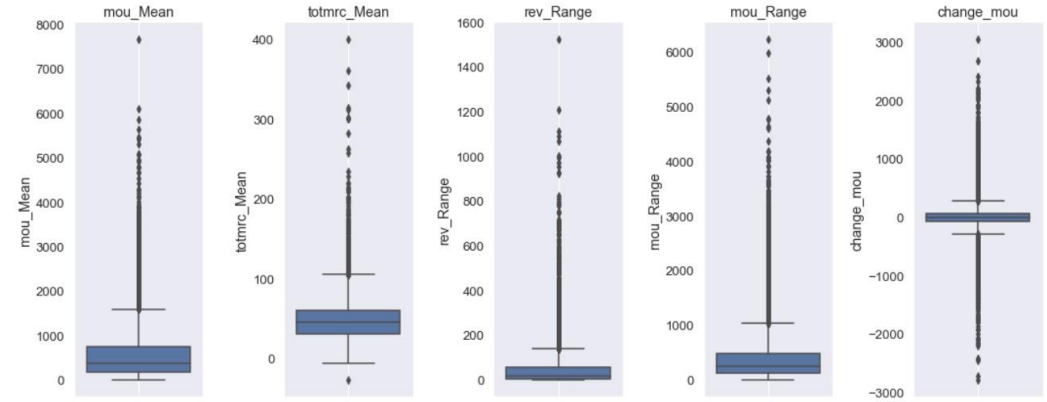


Fig 2.5 Sample of outliers present in first 5 columns

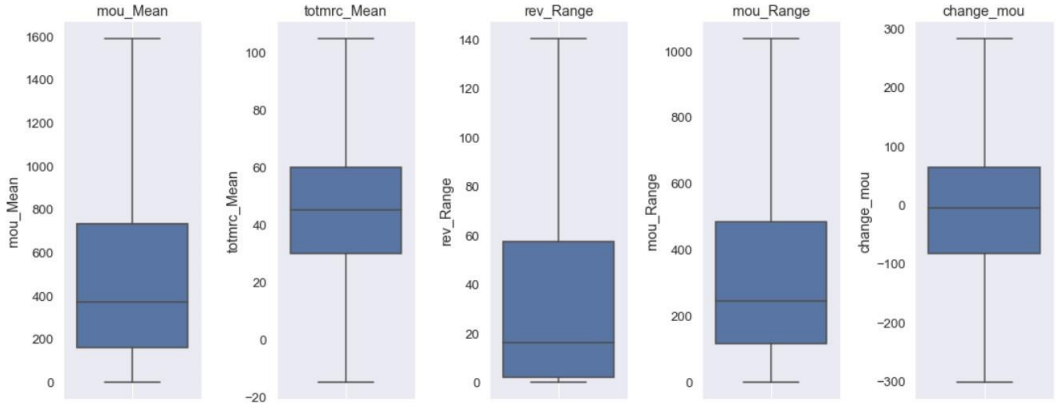


Fig 2.5 Boxplot of first 5 columns after outlier treatment