Business Report

Name: Kriti Gupta

Batch: PGPDSBA Online Mar 20_A

Project: Capstone Project

Table of Contents

Introduction

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.

Need the of the Project

It has become the need of the hour to predict the potential behaviour of the customers so that proactive steps can be taken by the telecom provider to reduce churn rate. It has also become important to validate whether the independent survey stating that the whole telecom industry would face increasing churn rate and decreasing ARPU or not.

This project will be beneficial to highlight important factors which contribute to high attrition rate, it would be able to provide a retention strategy and recommendations to target priority customers in a timely manner.

Understanding business/social opportunity

The main business opportunities are as follows:

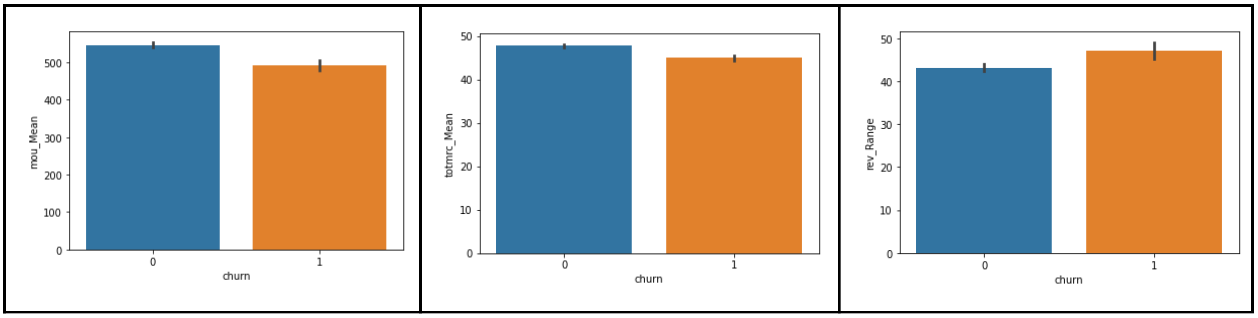
- ❖ Predicting Churn Rate
- ❖ Better Customer Retention Policy
- ❖ Proactive Customer service
- ❖ Prioritizing customers

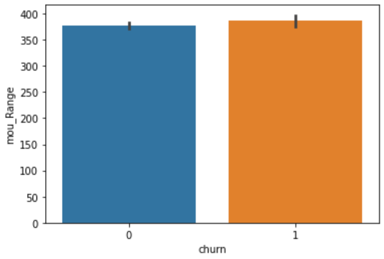
Exploratory Data Analysis

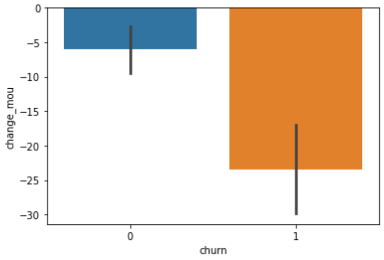
Univariate Analysis

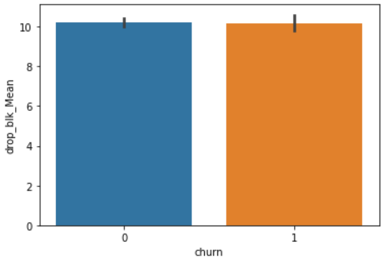
The graphs below show every variable in terms of Churn

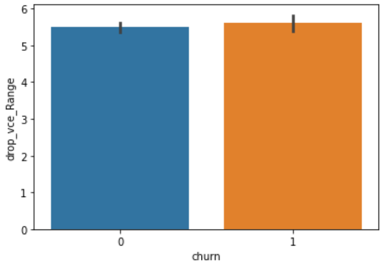
Figure 1: Univariate Analysis

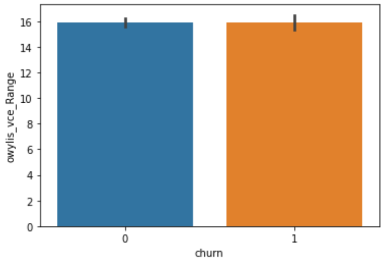


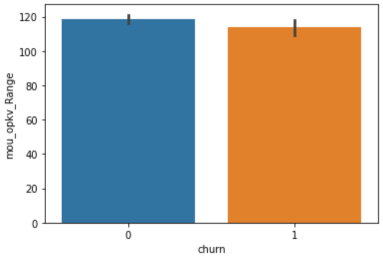

400
350
300
250
200
150
100
50
0
mou_Range
0
1
churn

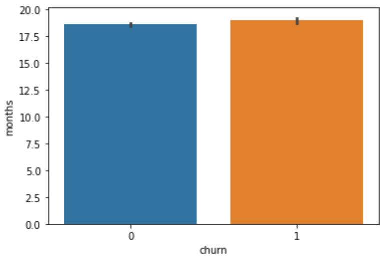

0
−5
−10
−15
−20
−25
−30
change_mou
0
1
churn

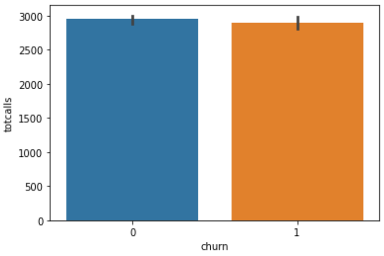

10
8
6
4
2
0
drop_blk_Mean
0
1
churn

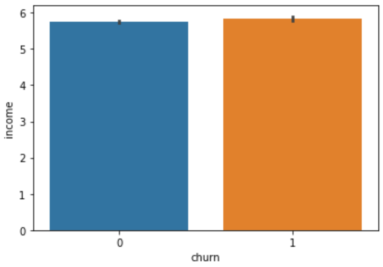

6
5
4
3
2
1
0
drop_vce_Range
0
1
churn

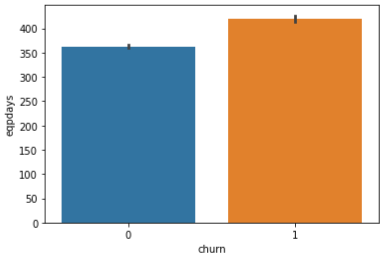

16
14
12
10
8
6
4
2
0
owylis_vce_Range
0
1
churn

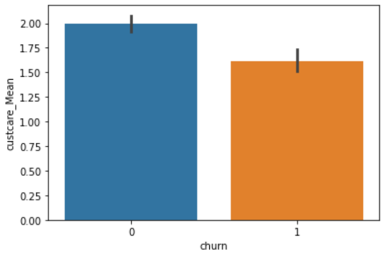

120
100
80
60
40
20
0
mou_opkv_Range
0
1
churn

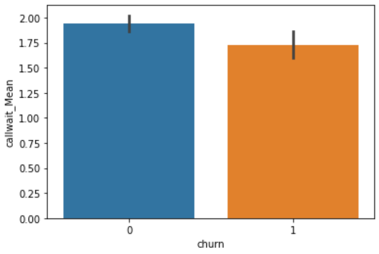

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
months
0
1
churn

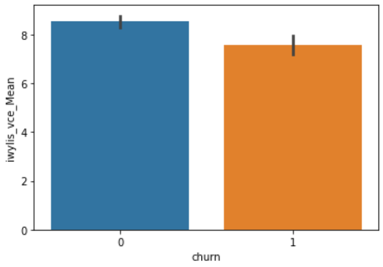

3000
2500
2000
1500
1000
500
0
totcalls
0
1
churn

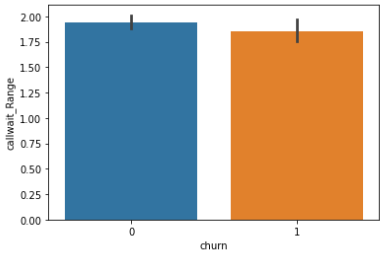

6
5
4
3
2
1
0
income
0
1
churn

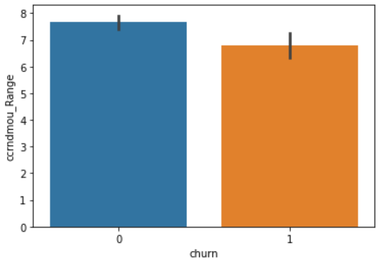

400
350
300
250
200
150
100
50
0
eqpdays
0
1
churn

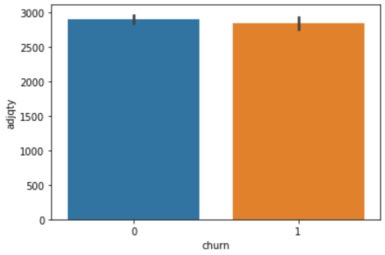

2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
custcare_Mean
0
1
churn

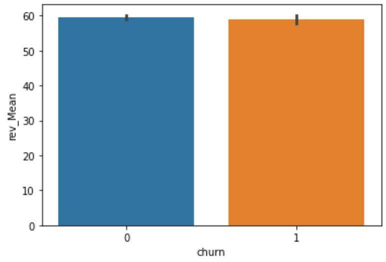

2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
callwait_Mean
0
1
churn

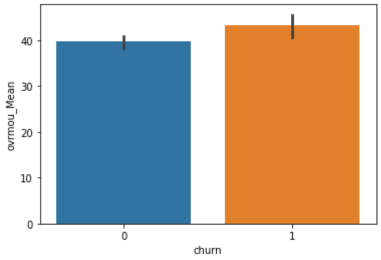

8
6
4
2
0
iwylis_vce_Mean
0
1
churn

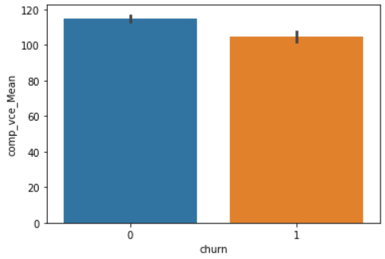

2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
callwait_Range
0
1
churn

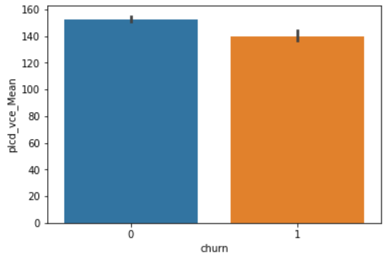

8
7
6
5
4
3
2
1
0
ccrndmou_Range
0
1
churn

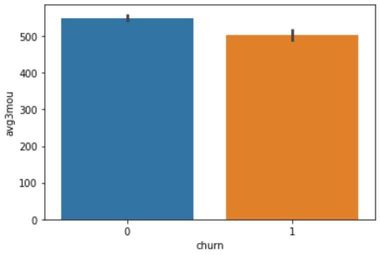

adjqty
0
500
1000
1500
2000
2500
3000
0
1
churn

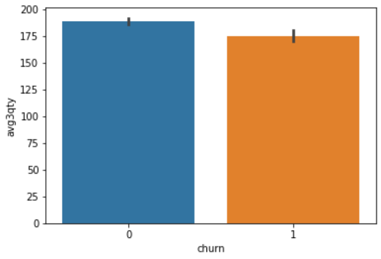

rev_Mean
0
10
20
30
40
50
60
0
1
churn

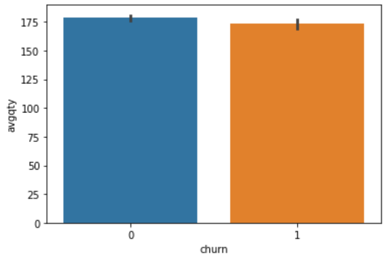

ovrmou_Mean
0
10
20
30
40
0
1
churn

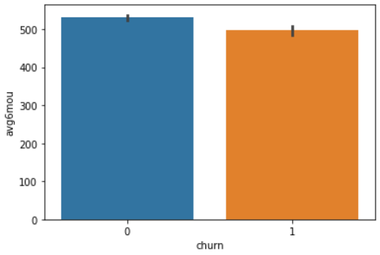

comp_vce_Mean
0
20
40
60
80
100
120
0
1
churn

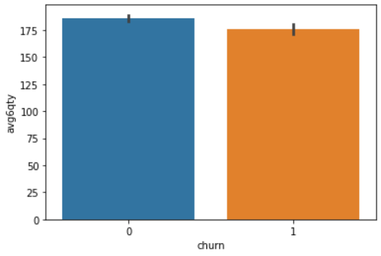

plcd_vce_Mean
0
20
40
60
80
100
120
140
160
0
1
churn

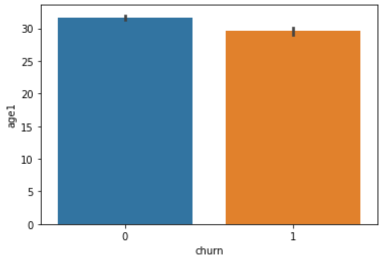

avg3mou
0
100
200
300
400
500
0
1
churn

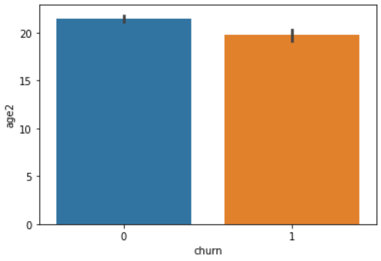

avg3qty
0
25
50
75
100
125
150
175
200
0
1
churn

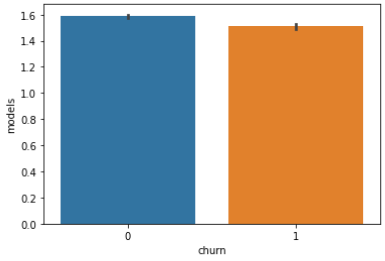

avgqty
0
25
50
75
100
125
150
175
0
1
churn

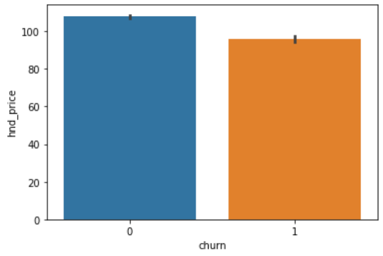

avg6mou
0
100
200
300
400
500
0
1
churn

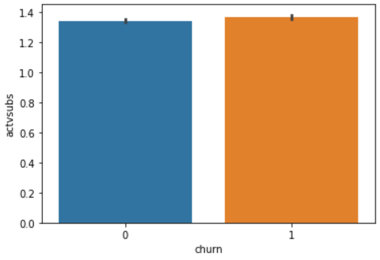

avg6qty
0
25
50
75
100
125
150
175
0
1
churn

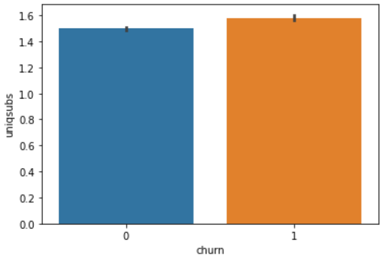

age1
0
5
10
15
20
25
30
0
1
churn

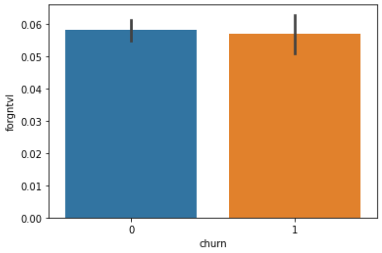

age2
0
5
10
15
20
0
1
churn

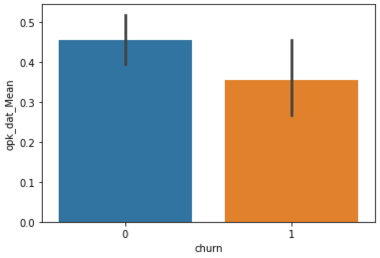

models
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0
1
churn

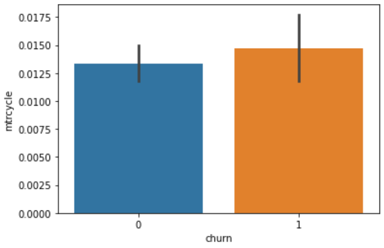

hnd_price
0
20
40
60
80
100
0
1
churn

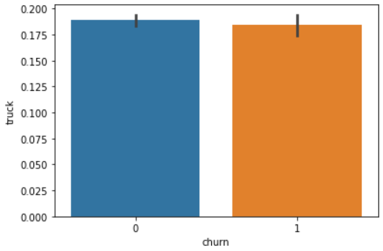

actvsubs
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
1
churn

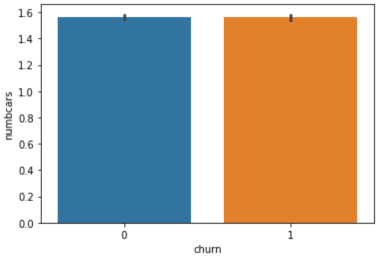

uniqsubs
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0
1
churn

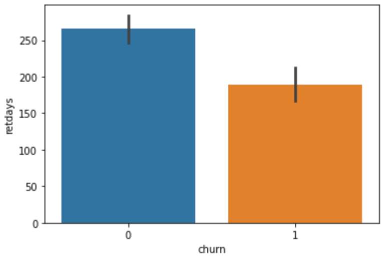

forgntvl
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
1
churn

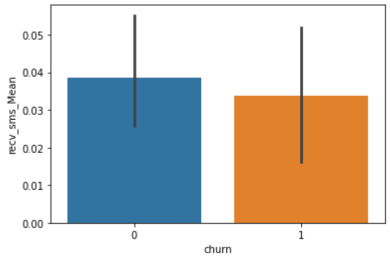

opk_dat_Mean
0.0
0.1
0.2
0.3
0.4
0.5
0
1
churn

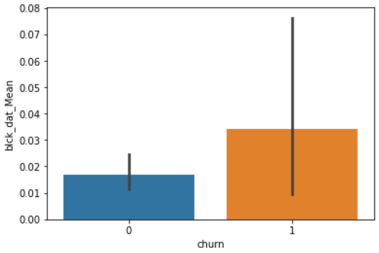

mtrcycle
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0
1
churn

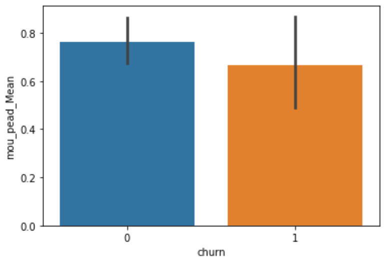

truck
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
0
1
churn

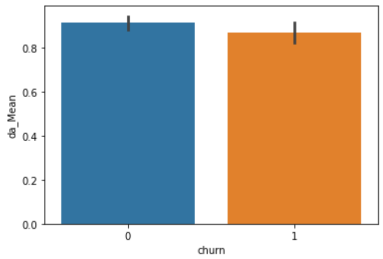

numbcars
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0
1
churn

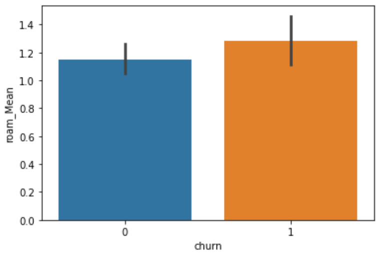

retdays
0
50
100
150
200
250
0
1
churn

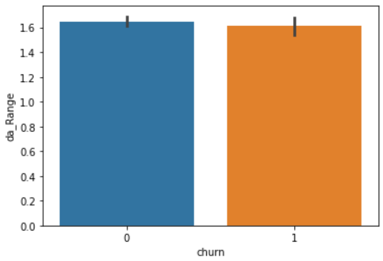

recv_sms_Mean
0.00
0.01
0.02
0.03
0.04
0.05
0
1
churn

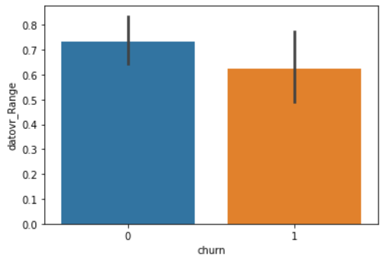

blck_dat_Mean
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0
1
churn

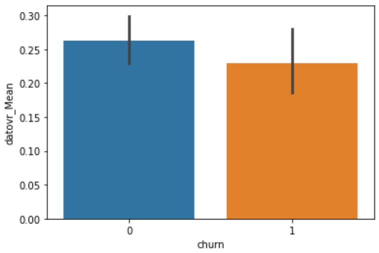

mou_pead_Mean
0.0
0.2
0.4
0.6
0.8
0
1
churn

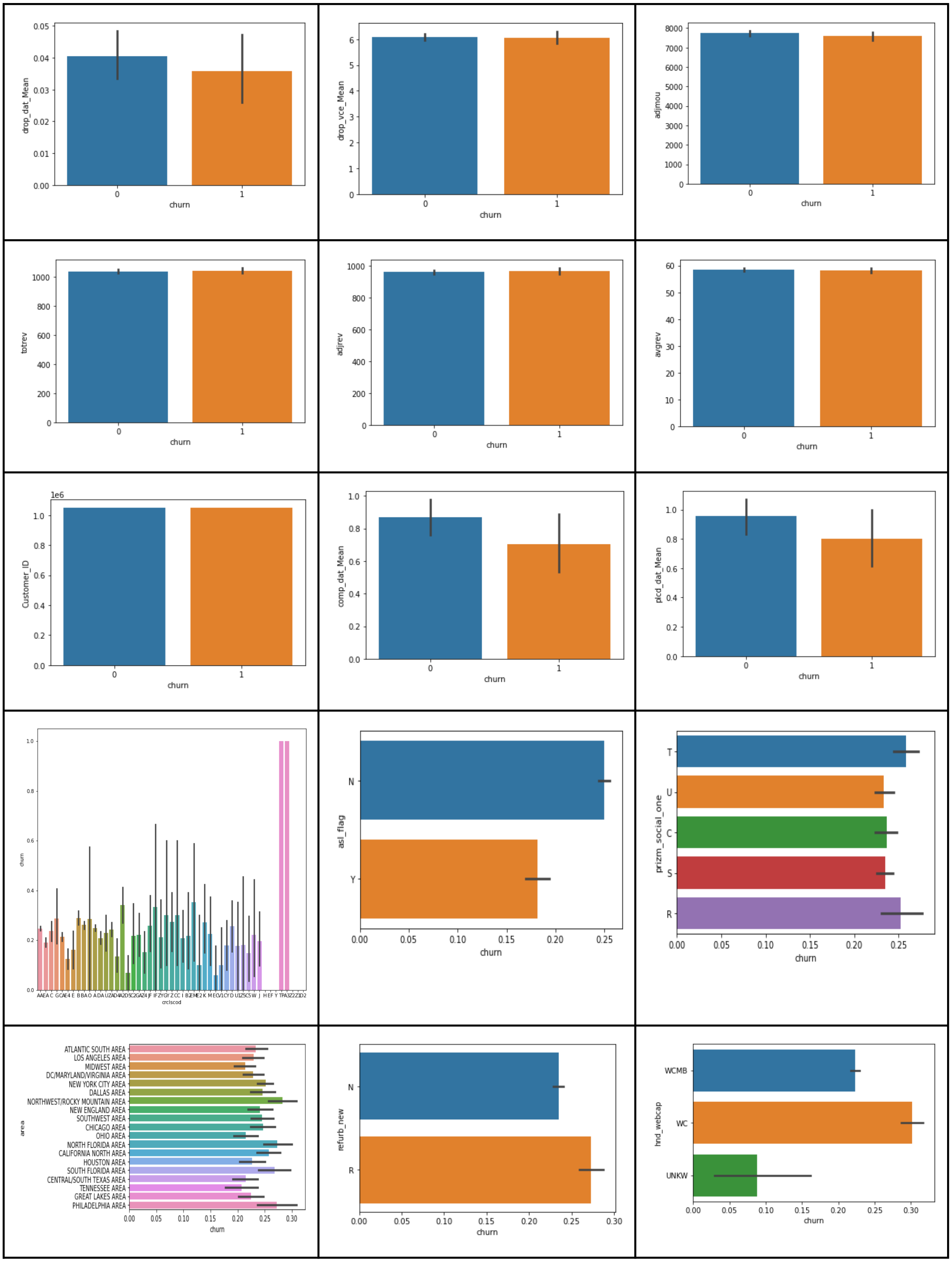

da_Mean
0.0
0.2
0.4
0.6
0.8
0
1
churn

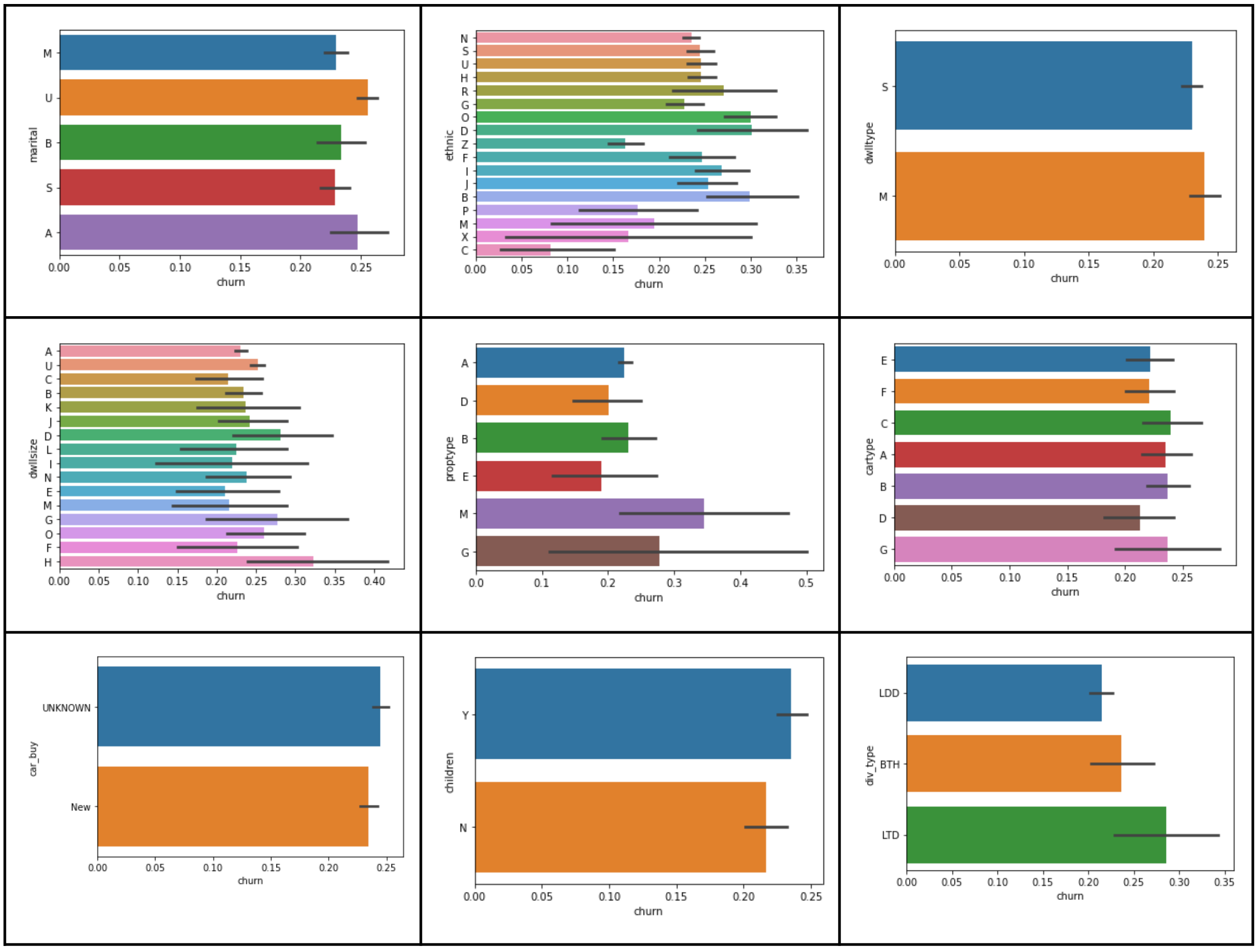

roam_Mean
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
1
churn


da_Range
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0
1
churn


datovr_Range
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
1
churn


datovr_Mean
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
1
churn


drop_dat_Mean
0.00
0.01
0.02
0.03
0.04
0.05
0
1
churn
drop_vce_Mean
0
1
2
3
4
5
6
0
1
churn
adjmou
0
1000
2000
3000
4000
5000
6000
7000
8000
0
1
churn
totrev
0
200
400
600
800
1000
0
1
churn
adjrev
0
200
400
600
800
1000
0
1
churn
avgrev
0
10
20
30
40
50
60
0
1
churn
Customer_ID
1e6
0.0
0.2
0.4
0.6
0.8
1.0
0
1
churn
comp_dat_Mean
0.0
0.2
0.4
0.6
0.8
1.0
0
1
churn
plcd_dat_Mean
0.0
0.2
0.4
0.6
0.8
1.0
0
1
churn
churn
0.0
0.2
0.4
0.6
0.8
1.0
AAEA C GCAE4 E B BA O A DA UZAD4A2D5C2GAZ4 JF IF ZYGY Z CC I B2EME2 K M ECV1CY D U1Z5C5 W J H EF Y TPA3Z2Z1D2
crclscod
asl_flag
N
Y
0.00
0.05
0.10
0.15
0.20
0.25
churn
prizm_social_one
T
U
C
S
R
0.00
0.05
0.10
0.15
0.20
0.25
churn
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
0.00
0.05
0.10
0.15
0.20
0.25
0.30
churn
refurb_new
N
R
0.00
0.05
0.10
0.15
0.20
0.25
0.30
churn
hnd_webcap
WCMB
WC
UNKW
0.00
0.05
0.10
0.15
0.20
0.25
0.30
churn


marital
M
U
B
S
A
0.00
0.05
0.10
0.15
0.20
0.25
churn
ethnic
N
S
U
H
R
G
O
D
Z
F
I
J
B
P
M
X
C
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
churn
dwlltype
S
M
0.00
0.05
0.10
0.15
0.20
0.25
churn
dwllsize
A
U
C
B
K
J
D
L
I
N
E
M
G
O
F
H
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
churn
proptype
A
D
B
E
M
G
0.0
0.1
0.2
0.3
0.4
0.5
churn
cartype
E
F
C
A
B
D
G
0.00
0.05
0.10
0.15
0.20
0.25
churn
car_buy
UNKNOWN
New
0.00
0.05
0.10
0.15
0.20
0.25
churn
children
Y
N
0.00
0.05
0.10
0.15
0.20
0.25
churn
div_type
LDD
BTH
LTD
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
churn

Important insights after the insights

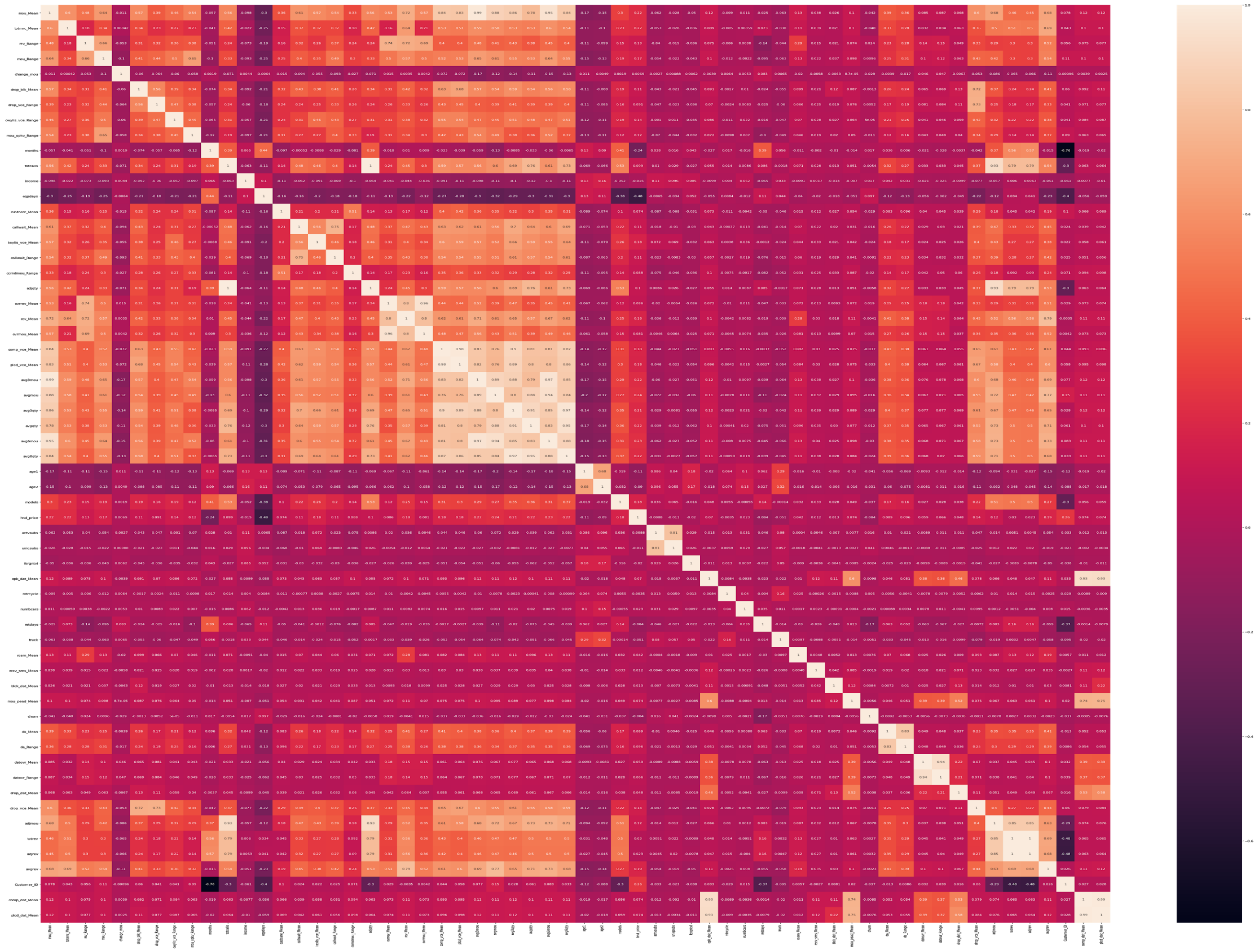
- **Churn rate is high when the following are also high:**
  - Range of revenue (charge amount)
  - Range of number of minutes of use
  - Range of number of dropped (failed) voice calls
  - Total number of months in service
  - Number of days (age) of current equipment
  - Mean overage minutes of use
  - Mean number of roaming calls
  - Mean number of blocked (failed) data calls
  - Number of active subscribers in household
  - Number of unique subscribers in the household
  - Foreign travel dummy variable

- **Churn rate is higher:**
  - In Northwest/ Rocky Mountain and Philadelphia Areas
  - When the Credit Class code is TP
  - When the Account spending limit is No
  - Customers from R Social group letter, Marital Status A and D Ethnicity have higher Churn rate
  - When the handset is refurbished and has WebCam capability of WC
  - Customers living in G Property type, M Dwelling unit, dwelling size is H and fall under LTD service area have higher churn rate
  - Customers with children (Y) and who have motorcycle
  - Customers where the Dominant vehicle lifestyle is G and the vehicle status is unknown.
  - Average Age of first household member is 30 years
  - Average Age of second household member 20 years

- **Churn rate is low when the following are also low:**
  - Mean number of monthly minutes of use
  - Mean total monthly recurring charge
  - Percentage change in monthly minutes of use vs previous three month average - **Negative in this case**
  - Range of unrounded minutes of use of off-peak voice calls
  - Total number of calls over the life of the customer
  - Mean number of customer care calls
  - Mean number of call waiting calls
  - Mean number of inbound wireless to wireless voice calls
  - Range of number of call waiting calls
  - Range of rounded minutes of use of customer care calls
  - Billing adjusted total number of calls over the life of the customer
  - Mean number of completed voice calls
  - Mean number of attempted voice calls placed
  - Average monthly minutes of use over the previous three months
  - Average monthly minutes of use over the life of the customer
  - Average monthly number of calls over the previous three months
  - Average monthly number of calls over the life of the customer
  - Average monthly minutes of use over the previous six months
  - Average monthly number of calls over the previous six months
  - Mean number of off-peak data calls
  - Number of days since last retention call
  - Mean number of received SMS calls
  - Mean revenue of data overage
  - Range of revenue of data overage
  - Mean number of completed data calls
  - Mean number of attempted data calls placed
  - Current handset price

Bivariate Analysis

Figure 2: Bivariate Analysis - Correlation



High Positive Correlation between the following variable

- Mean number of monthly minutes of use
- Mean monthly revenue (charge amount)
- Mean number of completed voice calls
- Mean number of attempted voice calls placed
- Average monthly minutes of use over the previous three months
- Average monthly minutes of use over the life of the customer
- Average monthly number of calls over the previous three months

High Positive Correlation between Range of revenue (charge amount) and the following variables:

- Mean overage revenue
- Mean monthly revenue (charge amount)

High Positive Correlation between Mean number of dropped or blocked calls and Mean number of dropped (failed) voice calls

High Positive Correlation between Range of number of dropped (failed) voice calls and Mean number of dropped (failed) voice calls

● Average monthly number of calls over the life of the customer ● Average monthly minutes of use over the previous six months ● Average monthly number of calls over the previous six months **High Positive Correlation between** the following variables: ● Total number of calls over the life of the customer ● Average monthly number of calls over the life of the customer ● Average monthly number of calls over the previous six months ● Billing adjusted total minutes of use over the life of the customer ● Total revenue ● Billing adjusted total revenue over the life of the customer ● Billing adjusted total number of calls over the life of the customer **High Positive Correlation between** Mean number of off-peak data calls and the following variables: ● Mean number of completed data calls ● Mean number of attempted data calls placed **High Positive Correlation between** Mean unrounded minutes of use of peak data calls and the following variables: ● Mean number of completed data calls ● Mean number of attempted data calls placed **High Negative Correlation between** Total number of months in service and Customer_ID	**High Positive Correlation between** Number of active subscribers in household and Number of unique subscribers in the household **High Positive Correlation between** Mean number of directory assisted calls and Range of number of directory assisted calls **High Positive Correlation between** Mean revenue of data overage and Range of revenue of data overage **High Positive Correlation between** Mean overage revenue and the following variables: ● Mean monthly revenue (charge amount) ● Mean overage minutes of use **High Positive Correlation between** Mean monthly revenue (charge amount) and the following variables: ● Mean overage minutes of use ● Average monthly minutes of use over the previous three months ● Average monthly revenue over the life of the customer **High Positive Correlation between** Mean number of call waiting calls and the following variables: ● Range of number of call waiting calls ● Average monthly number of calls over the previous three months **Negative Correlation between** Number of days (age) of current equipment and Current handset price **Negative Correlation between** Unique tournament specific customer ID for scoring purposes - Total revenue and Billing adjusted total revenue over the life of the customer

Business Insights from EDA

Balanced Data

Since the dataset is balanced 24% is when Churn = 1 and 76% when Churn = 0, there is no need for using SMOTE technique to balance it.

In terms of percentage distribution

```
0    0.760012
1    0.239988
Name: churn, dtype: float64
```

Clustering
The customers were bundled into 4 categories after scaling the data using KMeans Clustering, the silhouette_score for 4 clusters was the least (0.003).

Kmeans_clusters	**Freq**
Cluster 1	6772
Cluster 2	5609
Cluster 3	10223
Cluster 4	3914

These four clusters can be further used to target customers using proactive strategies and better retention policies.

- Cluster 4 has 3914 customers are the top tier (Tier I) customers with highest revenue, with highest call minutes and have the lowest equipment days.
- Cluster 3 has 10223 customers are Tier IV with the Lowest Revenue generating customers, with lowest call minutes and have the highest equipment days and income.
- Cluster 1 (6772) and Cluster 2 (5609) are Tier II and Tier II customers with nominal revenue, call minutes and income.

Data Cleaning and Preprocessing

Data Collection
Data was provided by Great Learning in a csv format file to predict Churn rate in Telecom Industry.

Data Processing
After using Jupyter Notebook to read the csv file, it was found that it has 26518 rows and 81 columns out of which approximately 26% was categorical data and 80% was non-categorical data. The data type of variables were different, 65 variables (float), 15 variables (int) and 21 variables (object).

Removing Unwanted Variables
The variables(12) were dropped since the missing values in these variables was more than 60%, mostly were categorical except Number of days since last retention call which was numeric variable:

- Mail order buyer
- Occupation of first household member
- Working woman in household

- Infobase no phone solicitation flag
- Property type detail
- Mail responder
- Dominant vehicle lifestyle
- Children present in household
- Division type code
- Number of days since last retention call

The missing value in Known number of vehicles was 49% was dropped

Unique tournament specific customer ID for scoring purposes (Customer ID)was also dropped since it is just like an ID which will not be required ahead.

Missing Value Treatment

There were around 42 variables which had missing values. The categorical variables (5) with missing value of less than 5% - Communications local service area, Geographic area, Marital status, Ethnicity, New or used car buyer were imputed with Mode values respectively. The numerical variables (21) where the missing values were less than 5% were imputed with median.

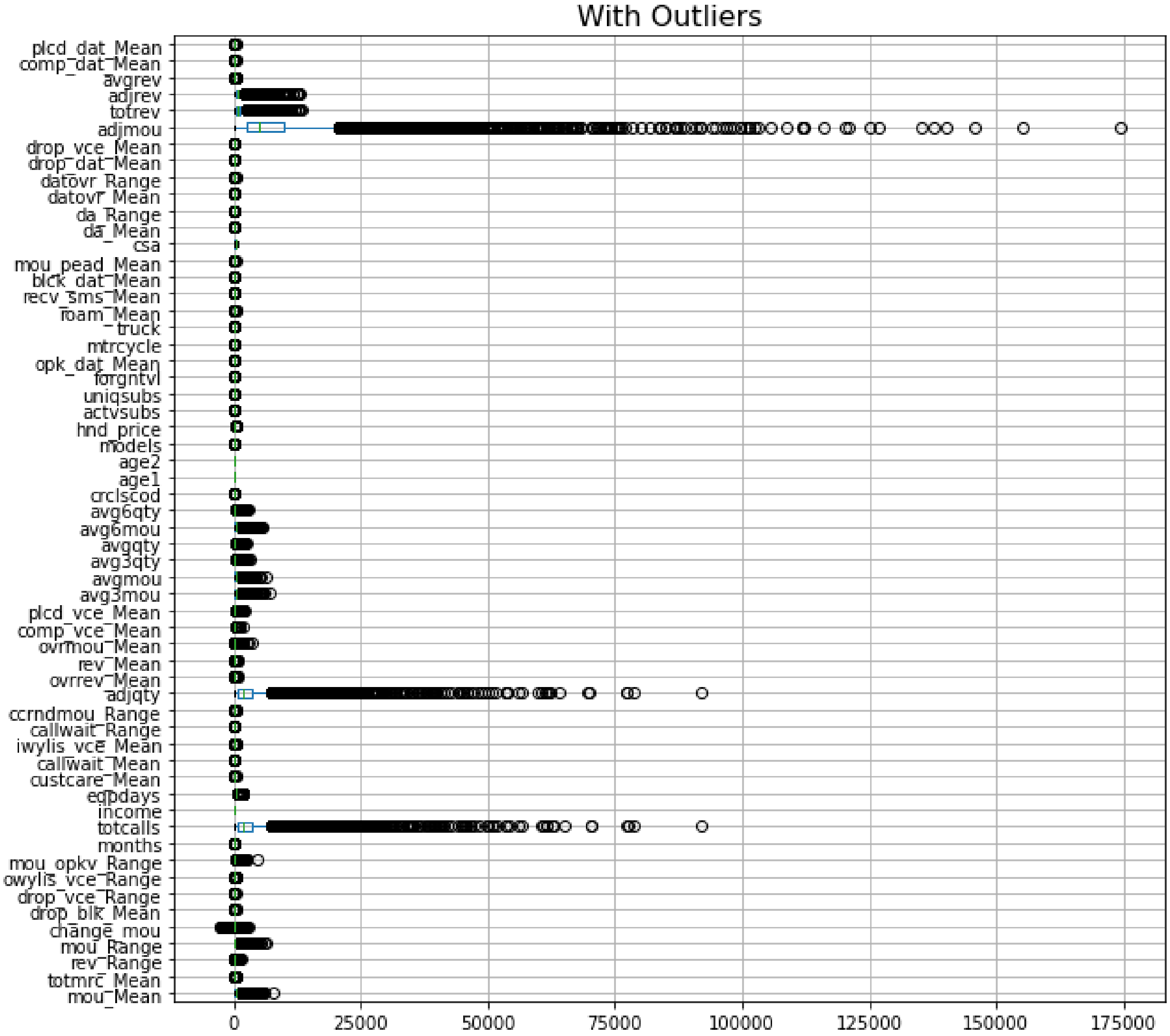
For 4 Categorical variables new categories were created:

- New Social group X was created for missing values (7%)
- The missing values for Handset web capability (9%) were clubbed with previous category of UNKW (unknown)
- New Dwelling unit type U was created for missing values (31%)
- New Dwelling size type U was created for missing values (38%)

For income variable - the missing values (25%) were imputed using the ‘Pad’ interpolation method.

Outlier Treatment

Figure 3: Presence of Outliers



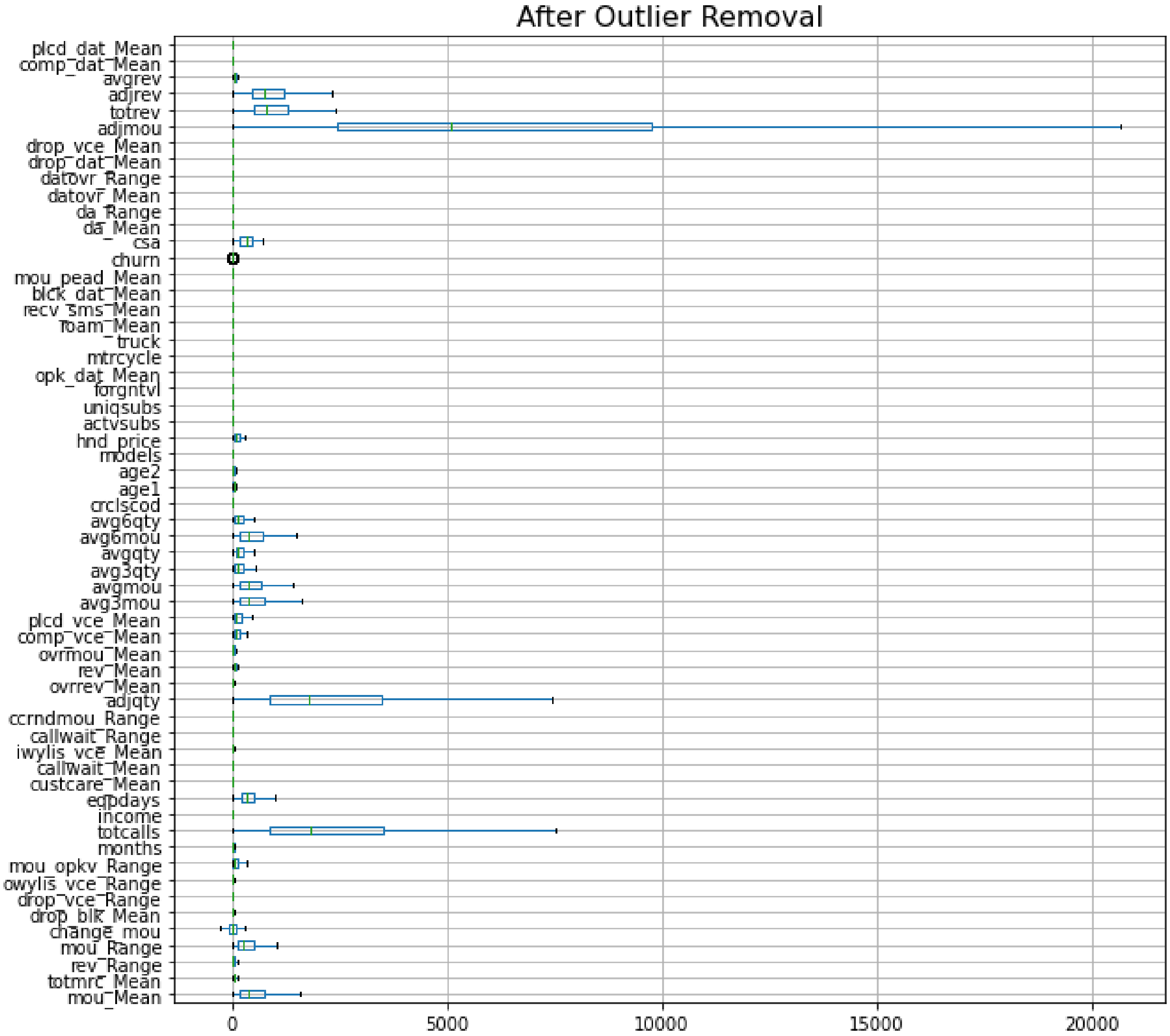
The data had outliers present and were identified using box plot, conventionally these are identified based on the inter quartile range (IQR = Q3 - Q1, where Q3 is 75th percentile and Q1 is 25th percentile). These outliers were treated with the following rules:

If value <1.5*IQR, then replace with 25th percentile

if value> 1.5*IQR, then replace with 75th percentile

The outliers are now capped

Figure 4: After Outlier Treatment



Variable Transformation

All the categorical variables were converted:

- Communications local service area and Credit class code were converted using label encoding. Communications local service area (CSA) had 694 unique values, if it was converted by using one hot encoding a lot of variables would have been created which would have increased dimensions of the dataset.

  Credit class code was converted using label coding since classes are present likert scale can be used.
- All the other categorical variables were converted using one hot encoding.

No duplicate values were present in the dataset.

Addition of new variables

Since the categorical variables were converted one hot encoding new variables were created based on Categorical variables' categories.

New shape of the dataset is - 26518 rows and 124 columns

Model Building

In order to begin with model building, the entire data was split into X (independent variables) and y has only one dependent variable i.e Churn. After this the next step was to split the data into train and test dataset (X_train, y_train, X_test and y_test) in a 70:30 ratio.

After splitting the data various models were applied on the dataset, such as:

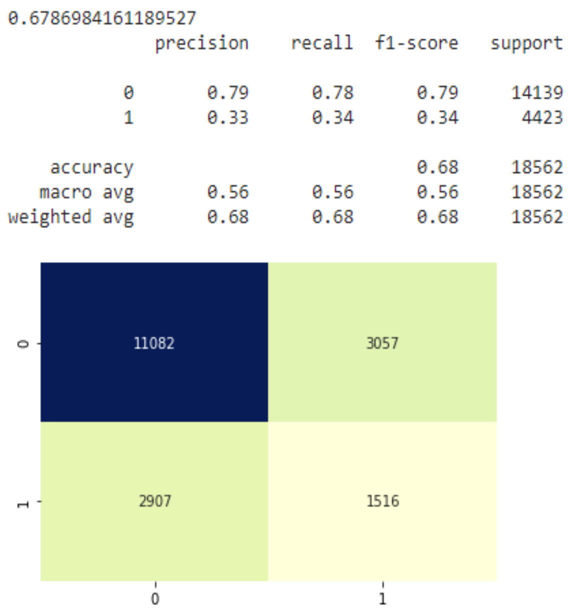
- ❖ Linear Discriminant Analysis - LDA
- ❖ Logistic Regression
- ❖ Random Forest
- ❖ Naive Bayes
- ❖ K-Nearest Neighbor(KNN)
- ❖ Bagging
- ❖ ADA Boosting
- ❖ Gradient Boosting

Table1: Models Used and their score

	Model	AUC Score	Accuracy	Precision	Recall	F1-score
Train Dataset	Linear Discriminant Analysis - LDA	0.644	0.76	0.55	0.03	0.05
Test Dataset		0.625	0.75	0.44	0.02	0.04
Train Dataset	Logistic Regression	0.637	0.76	0.54	0.02	0.05
Test Dataset		0.619	0.75	0.43	0.02	0.03
Train Dataset	Random Forest	0.998	0.85	1.00	0.37	0.54
Test Dataset		0.645	0.76	0.65	0.02	0.04
Train Dataset	**Naive Bayes**	**0.599**	**0.68**	**0.33**	**0.34**	**0.34**
Test Dataset		**0.589**	**0.67**	**0.33**	**0.32**	**0.32**
Train Dataset	K-Nearest Neighbor (KNN)	0.803	0.79	0.65	0.26	0.38
Test Dataset		0.546	0.72	0.31	0.11	0.16
Train Dataset	Bagging	1.00	1.00	1.00	1.00	1.00
Test Dataset		0.639	0.76	0.55	0.06	0.12
Train Dataset	ADA Boosting	0.719	0.77	0.61	0.10	0.18
Test Dataset		0.648	0.75	0.49	0.09	0.14
Train Dataset	Gradient Boosting	0.743	0.77	0.77	0.06	0.11
Test Dataset		0.659	0.76	0.58	0.04	0.08

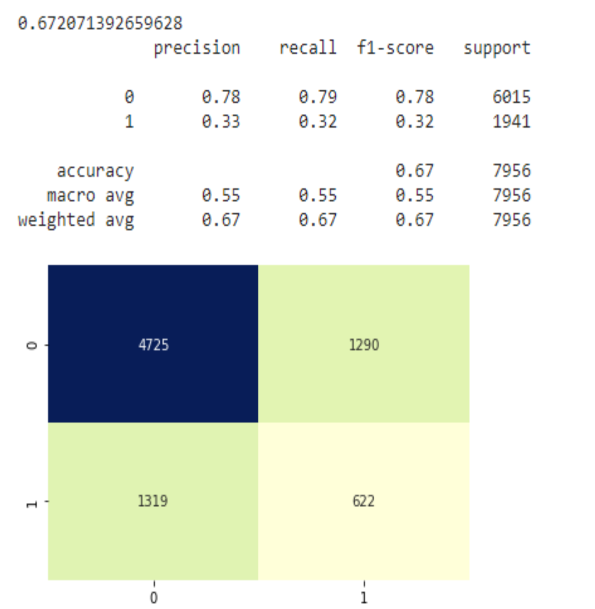
Naive Bayes Model

Figure 5: Confusion Matrix - Train Dataset



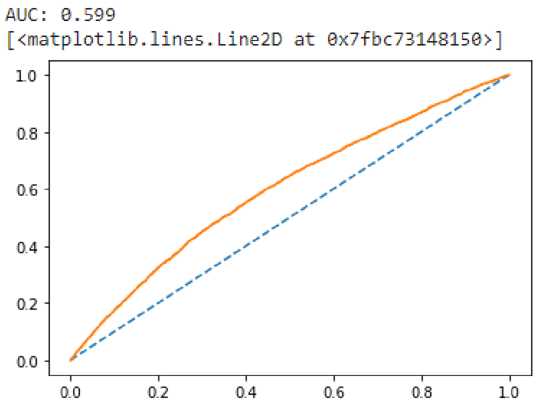
```
0.6786984161189527
              precision    recall  f1-score   support

           0       0.79      0.78      0.79     14139
           1       0.33      0.34      0.34      4423

    accuracy                           0.68     18562
   macro avg       0.56      0.56      0.56     18562
weighted avg       0.68      0.68      0.68     18562
```

Figure 6: Confusion Matrix - Test Dataset



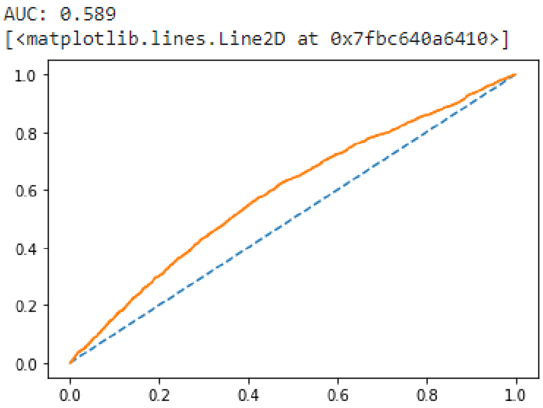
```
0.672071392659628
              precision    recall  f1-score   support

           0       0.78      0.79      0.78      6015
           1       0.33      0.32      0.32      1941

    accuracy                           0.67      7956
   macro avg       0.55      0.55      0.55      7956
weighted avg       0.67      0.67      0.67      7956
```

Figure 7: ROC Curve - Train Dataset



```
AUC: 0.599
[<matplotlib.lines.Line2D at 0x7fbc73148150>]
```

Figure 8: ROC Curve - Test Dataset



```
AUC: 0.589
[<matplotlib.lines.Line2D at 0x7fbc640a6410>]
```

Based on the above analysis it is clear that **Naive Bayes Model is performing better** than any other Model on test dataset with **accuracy of 67%, recall of 32%, precision of 33% and f1-score of 32% the AUC score is also 0.589.**

There is a great need to control the Type II Error i.e. when the model predicts that the customer will not churn but in reality it does churn. Inorder to do so Recall is to be increased, here with Naive Bayes Model Recall is at 32%.

Model Tuning

Only Naive Bayes Model was working better in comparison with the rest of the models but it is always better to tune the models and see their performance by changing the Threshold.

Table 2: Naive Bayes Model Used at different Thresholds

	Naive Bayes Model	AUC Score	Accuracy	Precision	Recall	F1-score
Train Dataset	**At Threshold 0.07**	**0.599**	**0.39**	**0.26**	**0.83**	**0.39**
Test Dataset		**0.589**	**0.39**	**0.26**	**0.83**	**0.40**
Train Dataset	**At Threshold 0.05**	**0.599**	**0.37**	**0.25**	**0.86**	**0.39**
Test Dataset		**0.589**	**0.37**	**0.26**	**0.86**	**0.40**

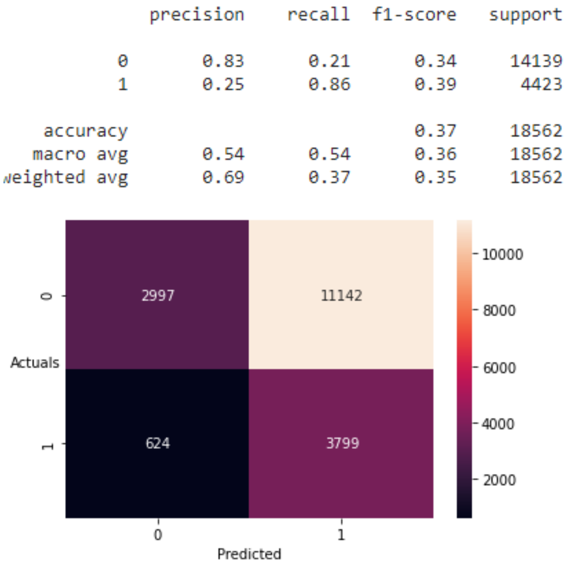
After tuning Naive Bayes Model and adjusting the threshold of 0.05 the Model performance on test dataset is the best with a recall of 86% thereby reducing Type II error i.e. 278 where predicted will not churn but actually churned, hence it is recommended to use Naive Bayes Model at a threshold of 0.05 to predict Churn or Customer Behavior.

Model Validation

The model is validated by comparing its performance on train and test data, which can be seen by the below graphs i.e. Confusion Matrix and ROC curve given below:

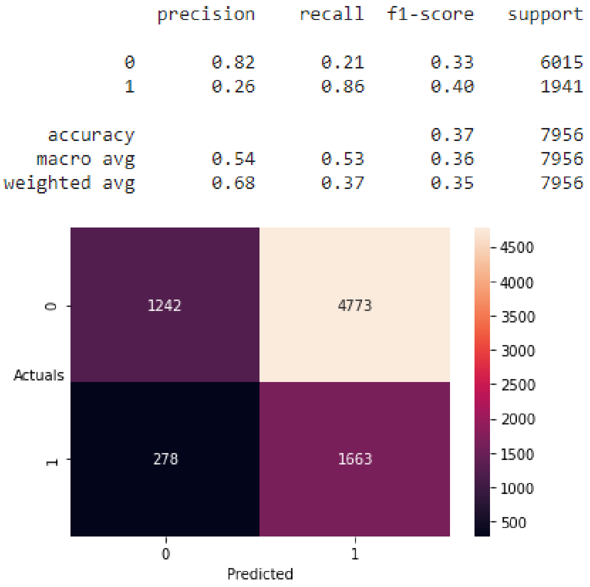
Naive Bayes Model at Threshold 0.05

Figure 9: Confusion Matrix - Train Dataset



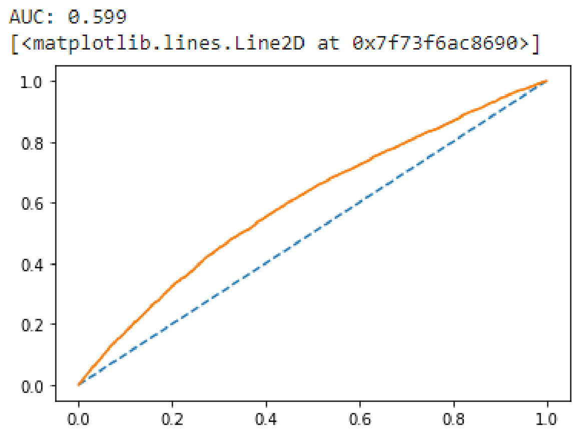
```
              precision    recall  f1-score   support

           0       0.83      0.21      0.34     14139
           1       0.25      0.86      0.39      4423

    accuracy                           0.37     18562
   macro avg       0.54      0.54      0.36     18562
weighted avg       0.69      0.37      0.35     18562
```

Figure 10: Confusion Matrix - Test Dataset



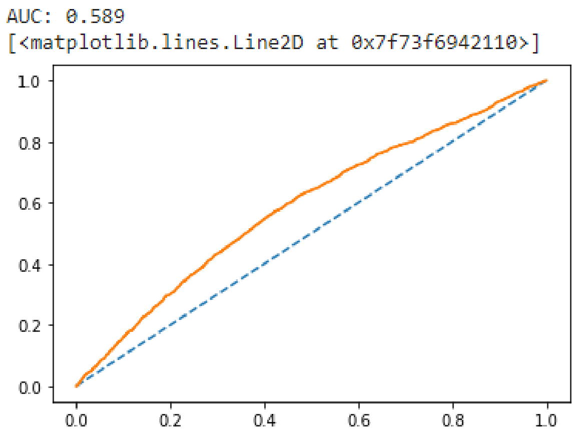
```
              precision    recall  f1-score   support

           0       0.82      0.21      0.33      6015
           1       0.26      0.86      0.40      1941

    accuracy                           0.37      7956
   macro avg       0.54      0.53      0.36      7956
weighted avg       0.68      0.37      0.35      7956
```

Figure 11: ROC Curve - Train Dataset



```
AUC: 0.599
[<matplotlib.lines.Line2D at 0x7f73f6ac8690>]
```

Figure 12: ROC Curve - Test Dataset



```
AUC: 0.589
[<matplotlib.lines.Line2D at 0x7f73f6942110>]
```

Based on the above analysis it is clear that **Naive Bayes Model's performance** at a threshold of 0.05 on the test dataset is with **accuracy of 37%, recall of 86%, precision of 26% and f1-score of 40% and the AUC score is also 0.589.**

While Accuracy refers to how well the model classifies datapoints correctly.
Accuracy = (TN+TP)/ (TN+FN+FP+TP)

Precision is the ability to find all relevant instances, the % of customers predicted to churn that actually churned.
Precision = TP /(TP+FP)

Recall is also known as sensitivity and as the True Positive Rate because it refers to how well the model classifies true values. Also described as the % of the customers who will churn and that the model is able to predict as churn.
Recall = TP/ (TP+FN)

In this case the primary focus was to predict the potential behaviour of the customers so that proactive steps can be taken by the telecom provider to reduce churn rate. It became a necessity to validate the model on the basis of recall inorder to reduce Type II error.

When the threshold is taken as 0.05 (Model Tuning) Naïve Bayes Model performance on the test dataset is the best with a recall of 86% thereby reducing Type II error i.e. 278 where predicted will not churn but actually churned.

Business Insights

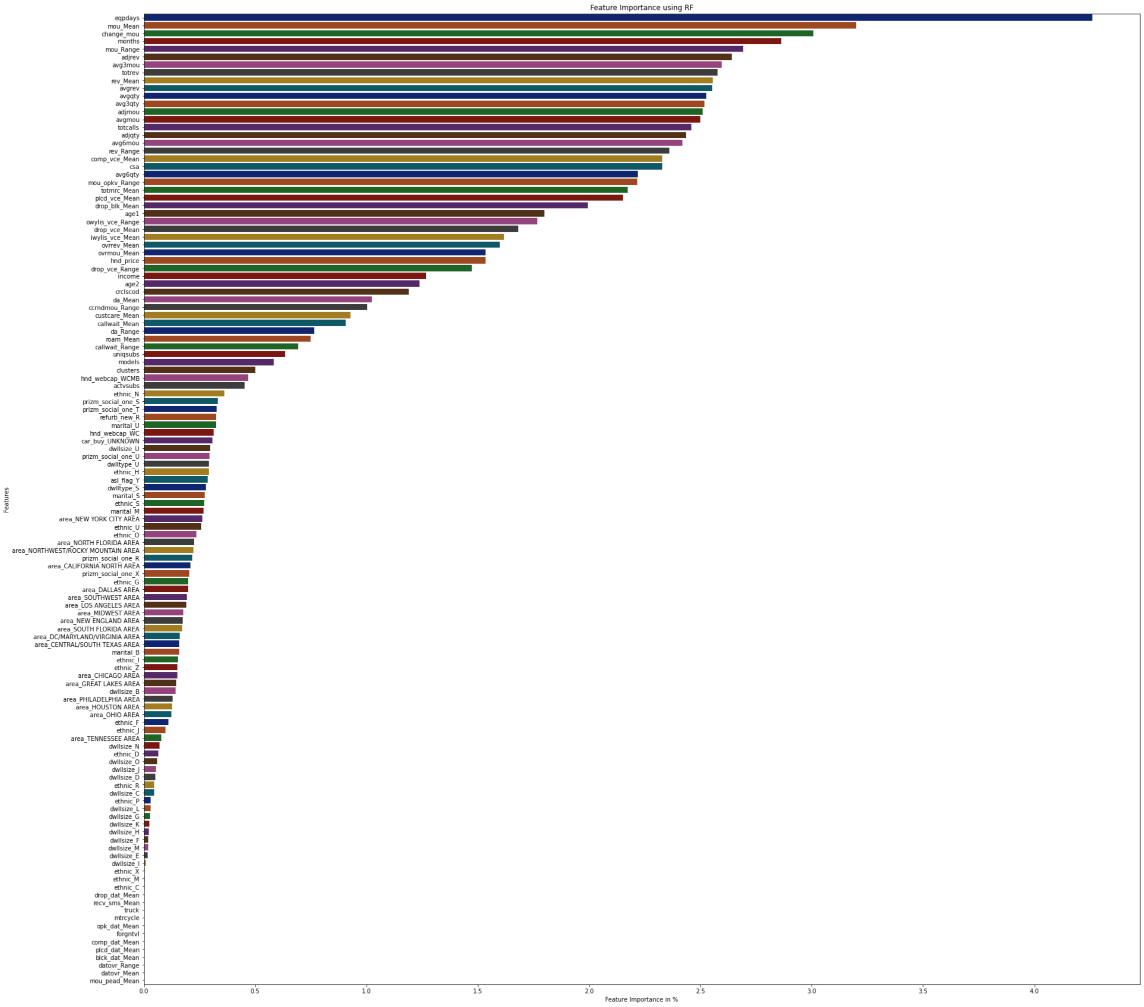
Clusters/ Customer Segmentation

These four clusters can be further used to target customers using proactive strategies and better retention policies.

- ❑ Cluster 4 has 3914 customers are the top tier (Tier I) customers with highest revenue, with highest call minutes and have the lowest equipment days.
- ❑ Cluster 3 has 10223 customers are Tier IV with the Lowest Revenue generating customers, with lowest call minutes and have the highest equipment days and income.
- ❑ Cluster 1 (6772) and Cluster 2 (5609) are Tier II and Tier II customers with nominal revenue, call minutes and income.

Important Features

Figure 13: Important Features



The top 30 Important features are as under:

Table 3: Top 30 Important Features

Variable	Description
eqpdays	Number of days (age) of current equipment
mou_Mean	Mean number of monthly minutes of use
change_mou	Percentage change in monthly minutes of use vs previous three month average
months	Total number of months in service
mou_Range	Range of number of minutes of use
adjrev	Billing adjusted total revenue over the life of the customer
avg3mou	Average monthly minutes of use over the previous three months
totrev	Total revenue
rev_Mean	Mean monthly revenue (charge amount)
avgrev	Average monthly revenue over the life of the customer
avgqty	Average monthly number of calls over the life of the customer
avg3qty	Average monthly number of calls over the previous three months
adjmou	Billing adjusted total minutes of use over the life of the customer
avgmou	Average monthly minutes of use over the life of the customer
totcalls	Total number of calls over the life of the customer
adjqty	Billing adjusted total number of calls over the life of the customer
avg6mou	Average monthly minutes of use over the previous six months
rev_Range	Range of revenue (charge amount)
comp_vce_Mean	Mean number of completed voice calls
csa	Communications local service area
avg6qty	Average monthly number of calls over the previous six months
mou_opkv_Range	Range of unrounded minutes of use of off-peak voice calls
totmrc_Mean	Mean total monthly recurring charge
plcd_dat_Mean	Mean number of attempted data calls placed

drop_blk_Mean	Mean number of dropped or blocked calls
age1	Age of first household member
owylis_vce_Range	Range of number of outbound wireless to wireless voice calls
drop_vce_Mean	Mean number of dropped (failed) voice calls
iwylis_vce_Mean	Mean number of inbound wireless to wireless voice calls
ovrrev_Mean	Mean overage revenue

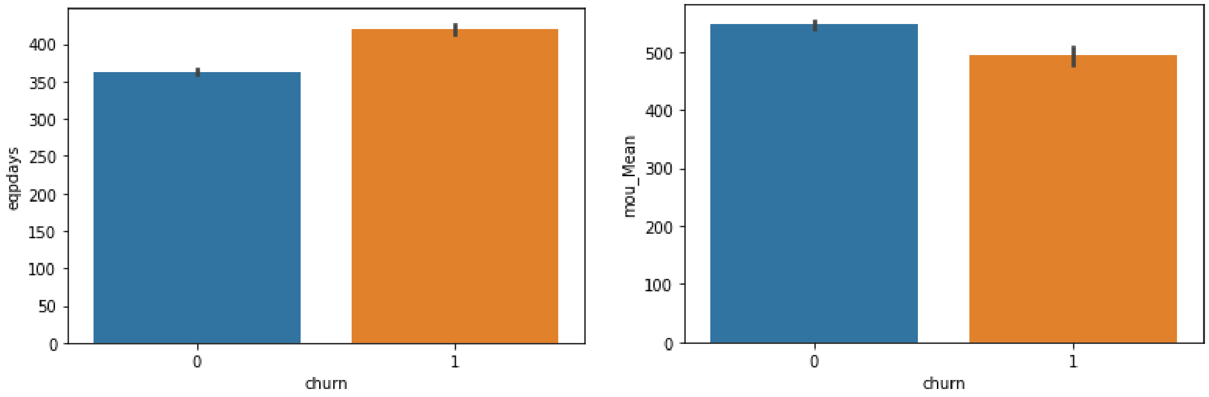
These top 30 **important features** are to be looked at to predict Churn rate in a telecom industry.

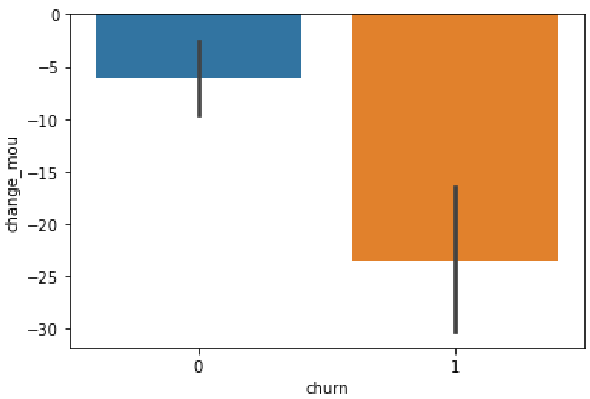
- Out of these 30 the top most important feature is Number of days (age) of current equipment' - when on an average the Number of days (age) of current equipment' is 400 or more then churn is higher.

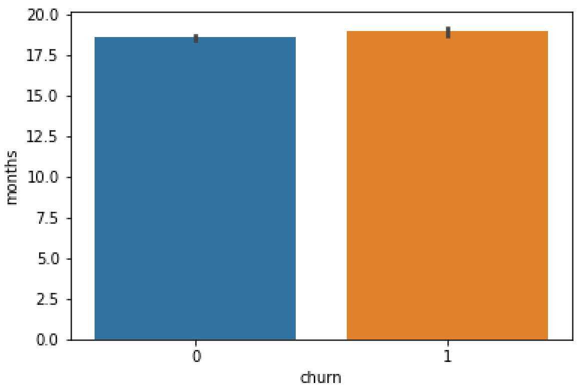
Other top four features are**:**

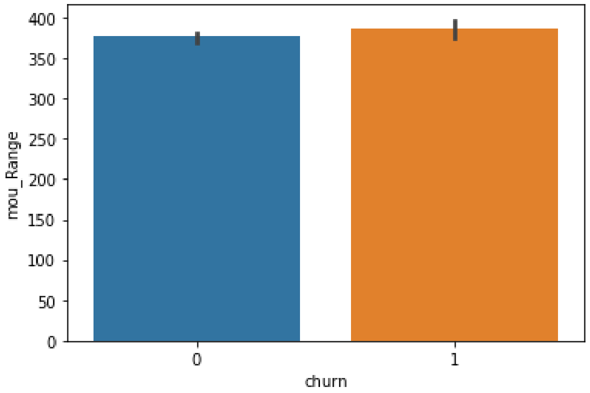
- When the Mean number of monthly minutes of use is 500 or more then churn is lower.
- When the Percentage change in monthly minutes of use vs previous three month average is ranges from 0 to -30 then churn is higher.
- When the Total number of months in service on an average is 18 or more then churn is higher.
- When the Range of number of minutes of use on average is above 350 then churn is higher.

Figure 14: Top 5 Important Features Vs Churn









Recommendations

❑ In Northwest/ Rocky Mountain; Philadelphia Areas and Customers falling under the LTD service area, since the churn rate is higher, better customer services can be provided by increasing customer service executives, reducing TAT to solve customer's problems and by putting in more cell towers.

❑ Customers from Social group - R , Marital Status - A and D - Ethnicity group have a higher Churn rate - these groups can be kept in mind before hand and be proactively reached by customer care.

❑ Customers living in G - Property type, M - Dwelling unit type , Dwelling size - H have a higher Churn rate - better facilities should be provided in these area/ types to reduce churn rate.

❑ When the Average Age of first household member is 30 years and/or Average Age of second household member 20 years churn rate is higher, better packages can be provided to customers falling under these age groups.

❑ Customers with refurbished handsets and WebCam capability of WC churn so package offers can be given to provide phone and call pack/ internet connection on a monthly rental scheme.

❑ Customers with children and who have motorcycle churn - better children friendly packages and connectivity (roaming packs) should be provided.

❑ The Naïve Bayes Model is able to predict Churn with 86% accuracy and predicted 672 (train dataset), 278 (test dataset) will not churn but actually churned. (reducing Type II error).

❑ Tier I customers account for approx. 15% (3914) of the customers are high revenue customers and are to be prioritized for a proactive retention campaign – by giving competitive rate plan, quicker customer service.

❑ Tier IV customers account for approx. 39% (10223) of customers are with the lowest Revenue generating but since the proportion is high customers are to be managed better – by providing better rate plans; monthly rental plans for handset, providing a better combined offer for handsets and rate plan, better customer services.