TELECOM CUSTOMER CHURN PREDICTION

Ravi Vanani
Harleen
Abhishek

PROBLEM STATEMENT

In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, **customer retention** has now become even more important than customer acquisition.

BUSINESS OBJECTIVES

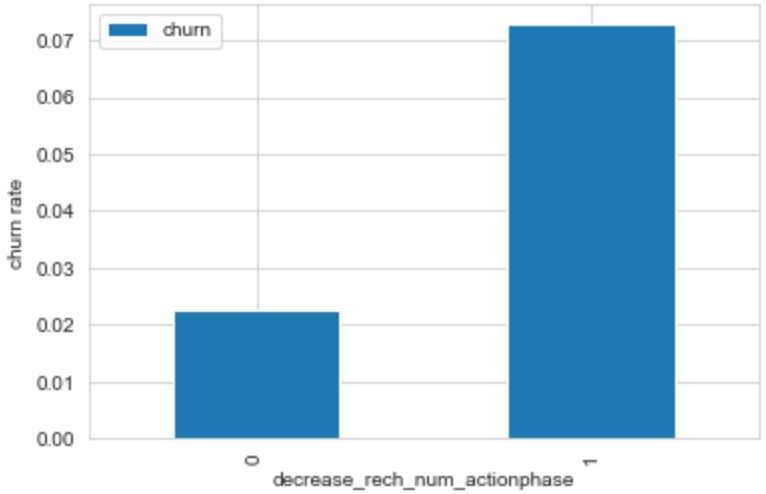
In this project, our business objective is to analyse customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn. Thus, our focus would be on

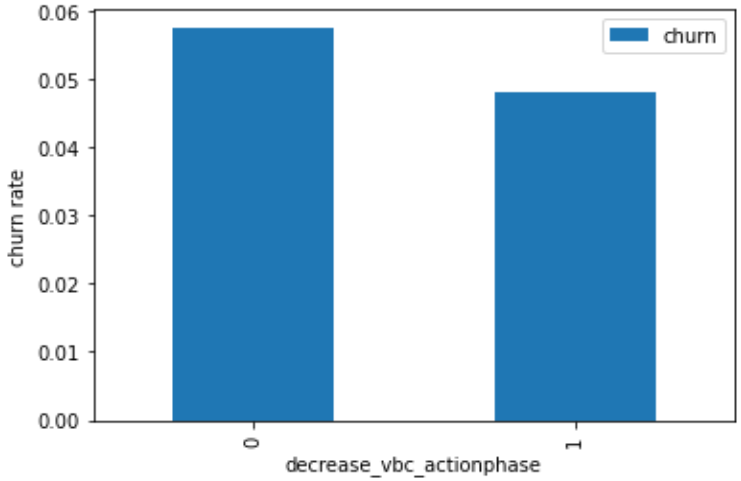
(i) Retaining high profitable customers.
(ii) Predicting which customers are at high risk of churn (in order to devise customer retention strategies accordingly).

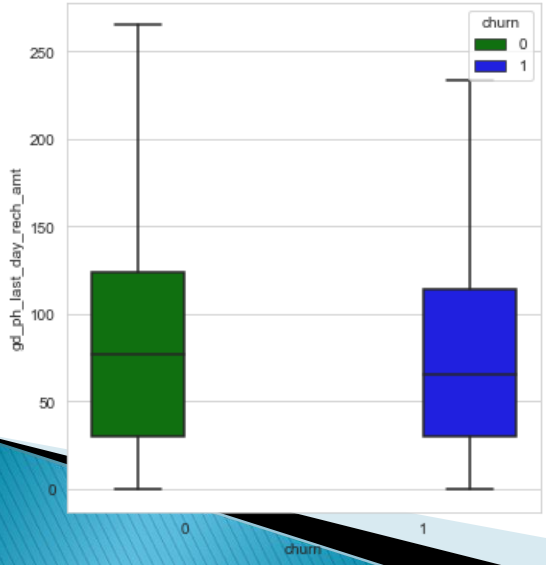
STEPS TO REACH OBJECTIVES

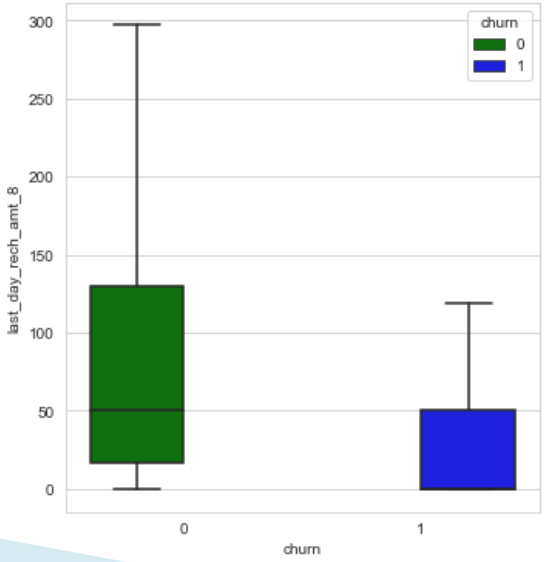
- Importing and inspecting data
- Data Cleaning and preparation
  - 1. Filtering the high-value customers
  - 2. Deriving Churn column
  - 3. Dropping columns with more than 40% null values
  - 4.Check for Duplicates
  - 5.Handling outlier
  - 6.Checking the Correlation between independent variables (multicollinearity)
- EDA
- Dummy variables creation
- Feature scaling
- Dealing with class imbalance
- Model Building & Evaluation
- Conclusion & Recommendations

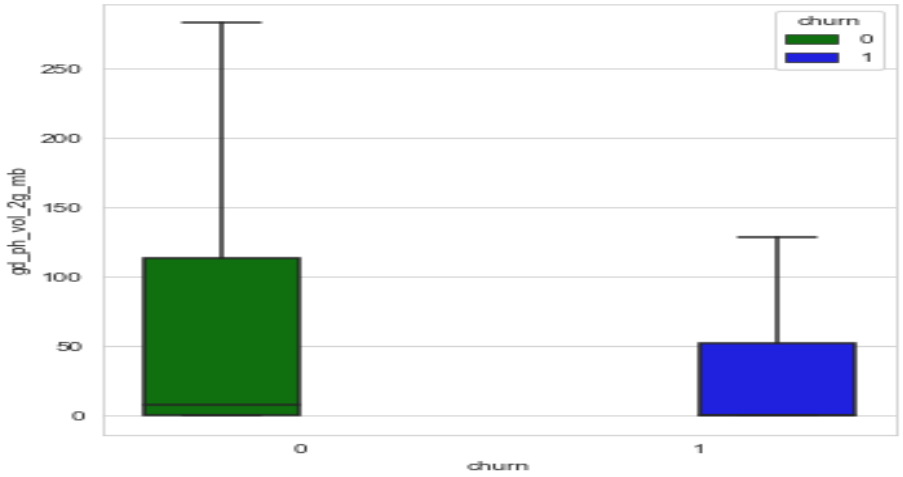
EDA (Univariant Analysis)

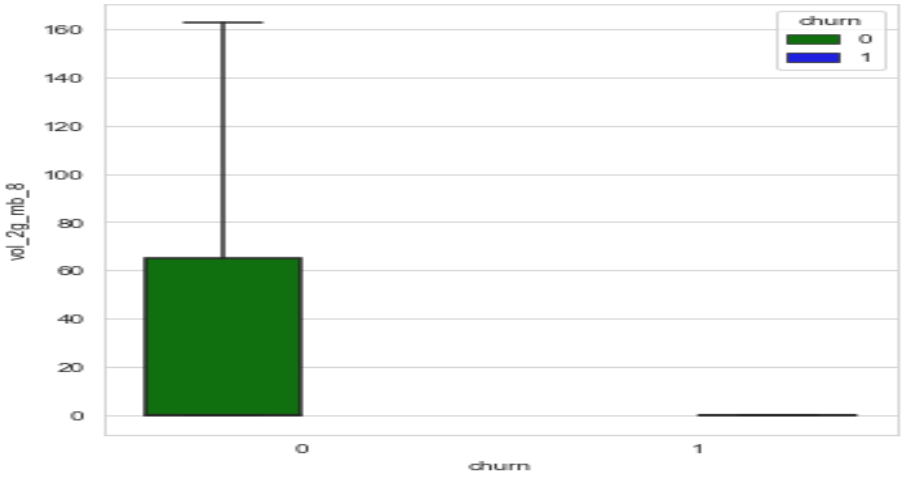


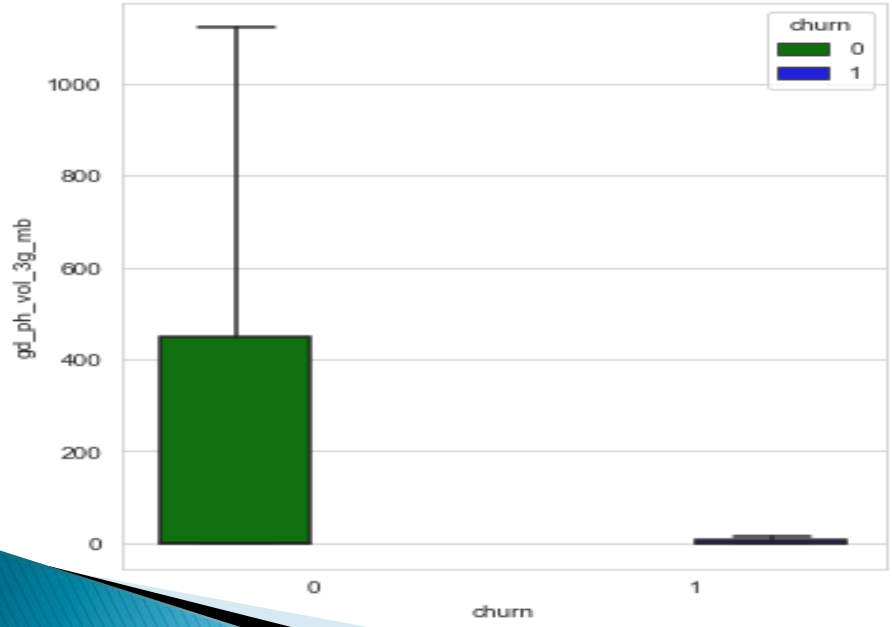


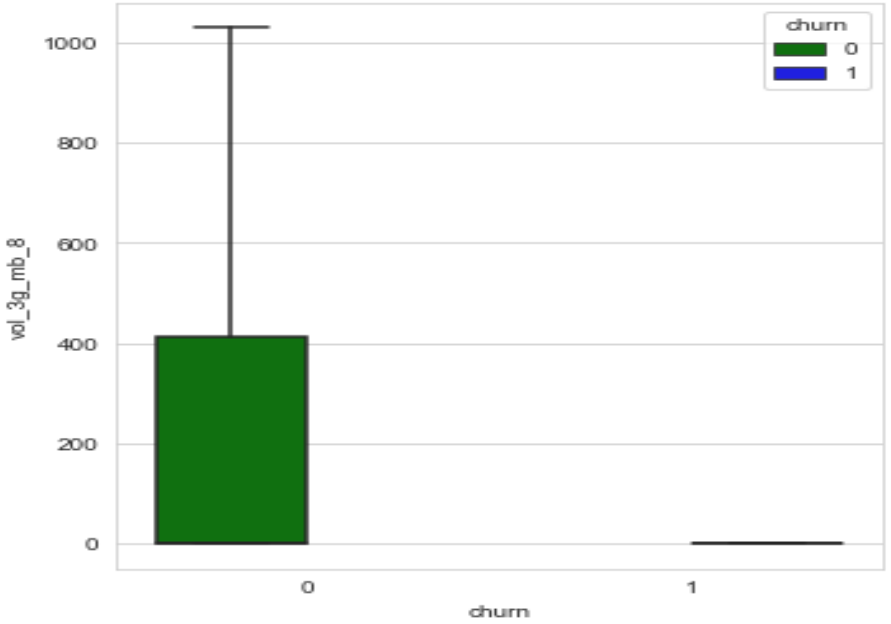


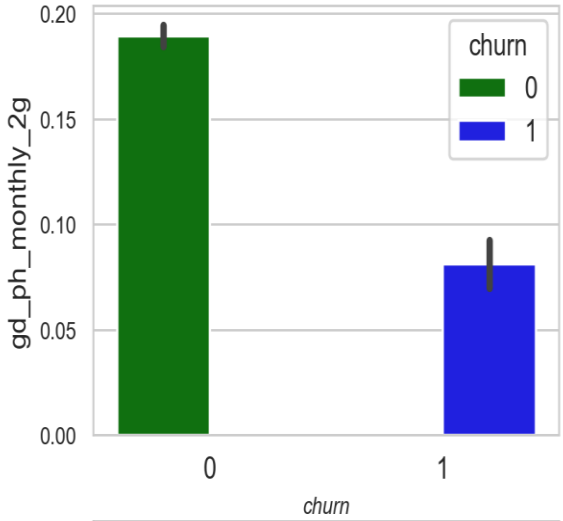


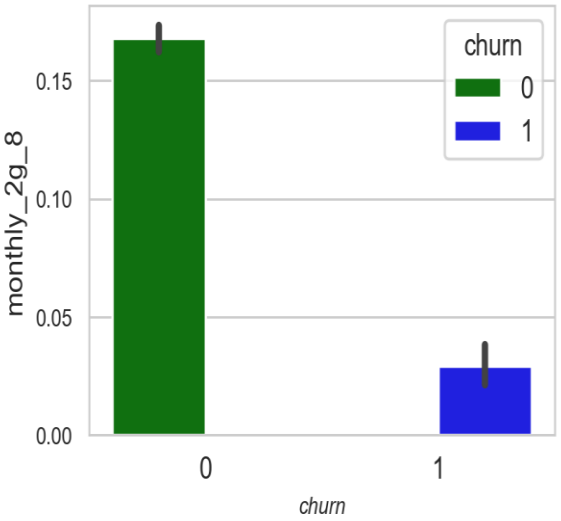

churn
0
1
gd_ph_vol_2g_mb
250
200
150
100
50
0
0
1
churn

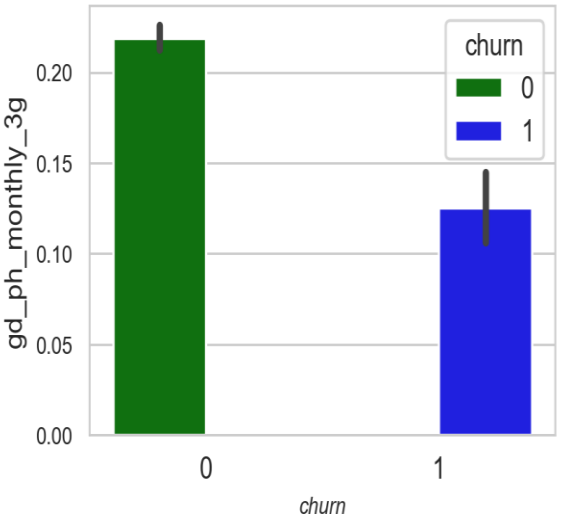

churn
0
1
vol_2g_mb_8
160
140
120
100
80
60
40
20
0
0
1
churn

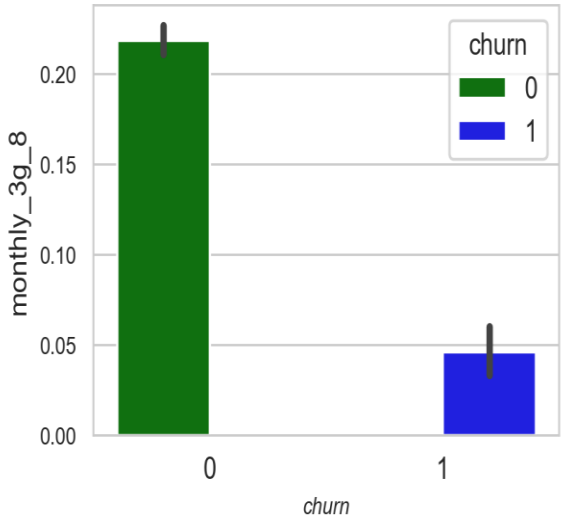

churn
0
1
gd_ph_vol_3g_mb
1000
800
600
400
200
0
0
1
churn

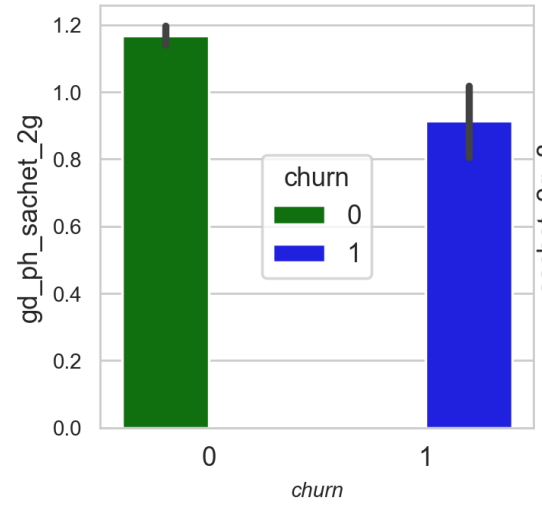

churn
0
1
vol_3g_mb_8
1000
800
600
400
200
0
0
1
churn

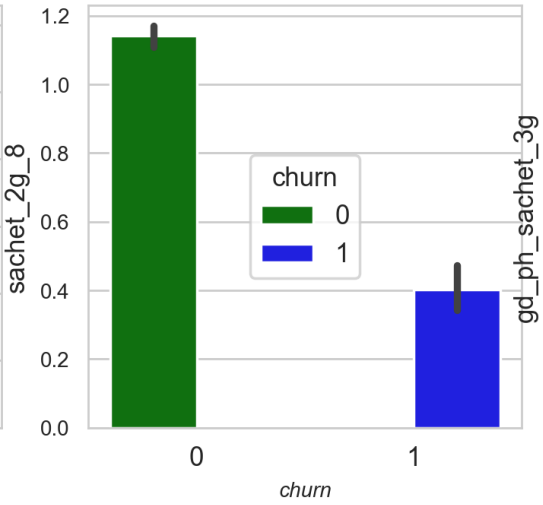

gd_ph_monthly_2g
churn
0
1
0.20
0.15
0.10
0.05
0.00
0
1
churn

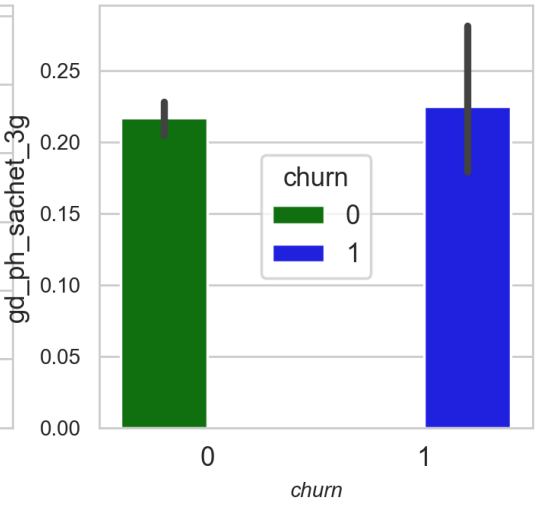

monthly_2g_8
churn
0
1
0.15
0.10
0.05
0.00
0
1
churn

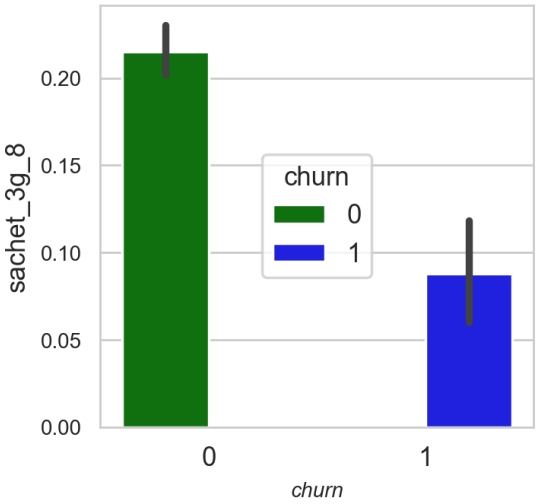

gd_ph_monthly_3g
churn
0
1
0.20
0.15
0.10
0.05
0.00
0
1
churn

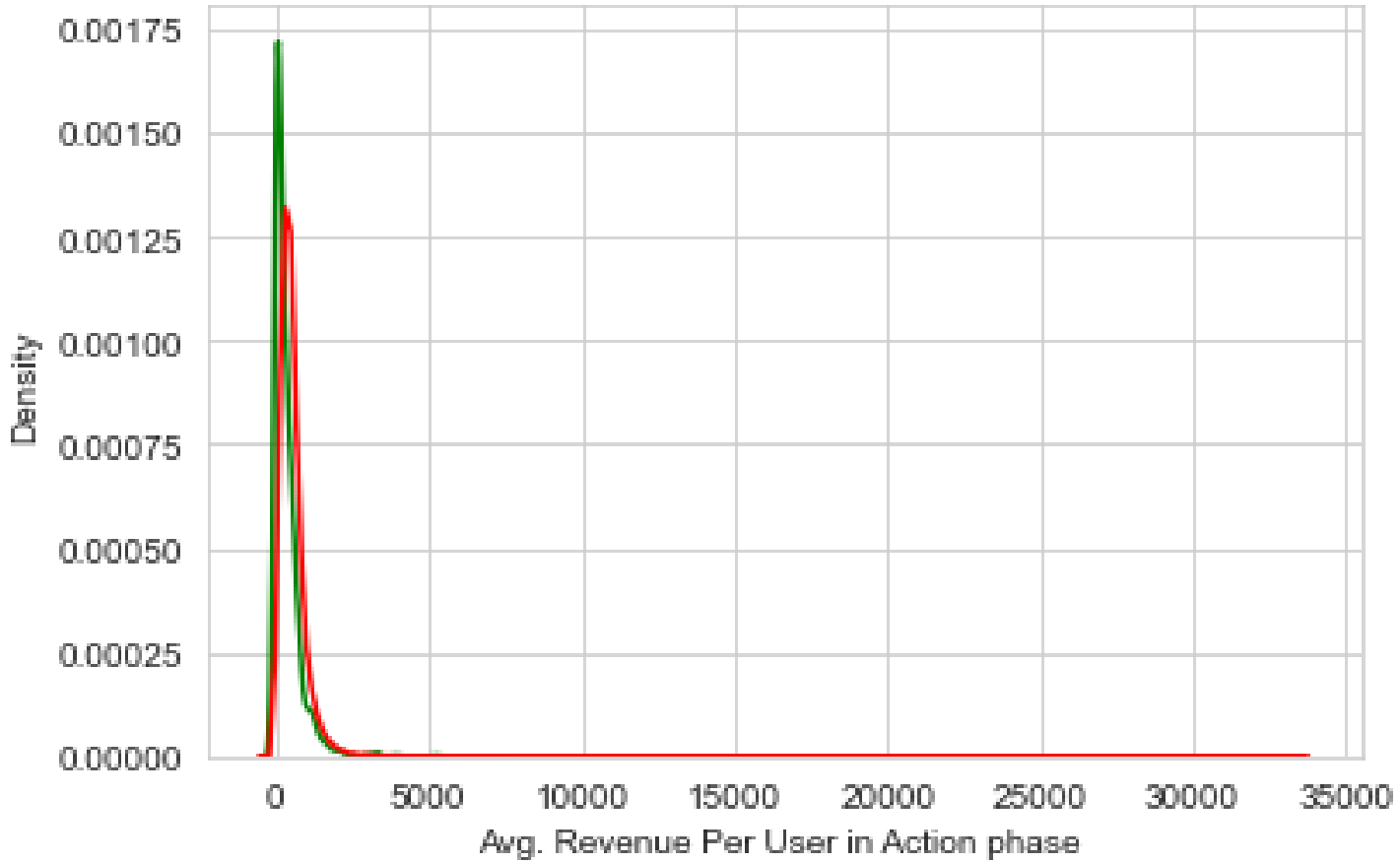

monthly_3g_8
churn
0
1
0.20
0.15
0.10
0.05
0.00
0
1
churn

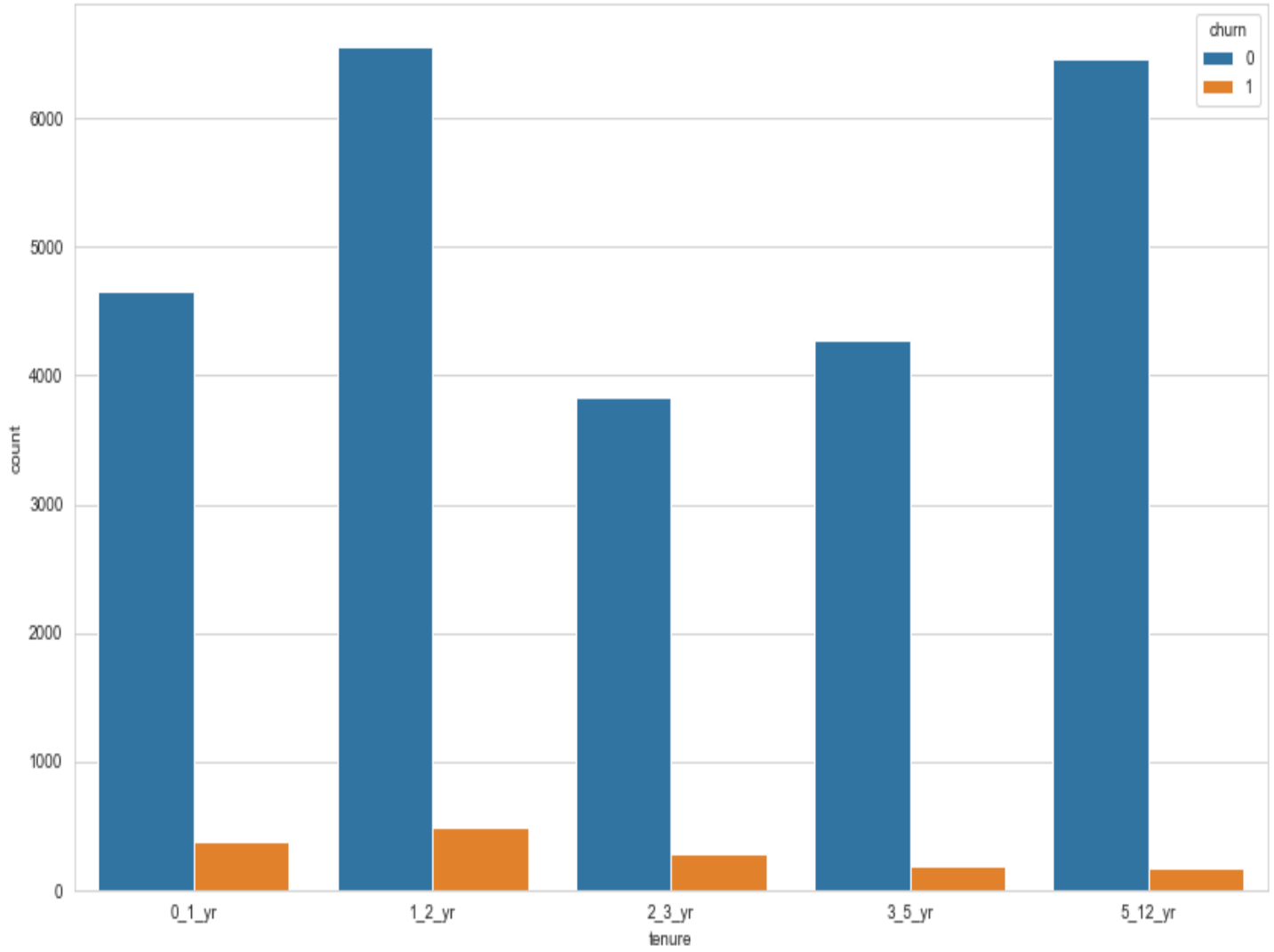

gd_ph_sachet_2g
churn
0
1
churn


sachet_2g_8
churn
0
1
churn


gd_ph_sachet_3g
churn
0
1
churn

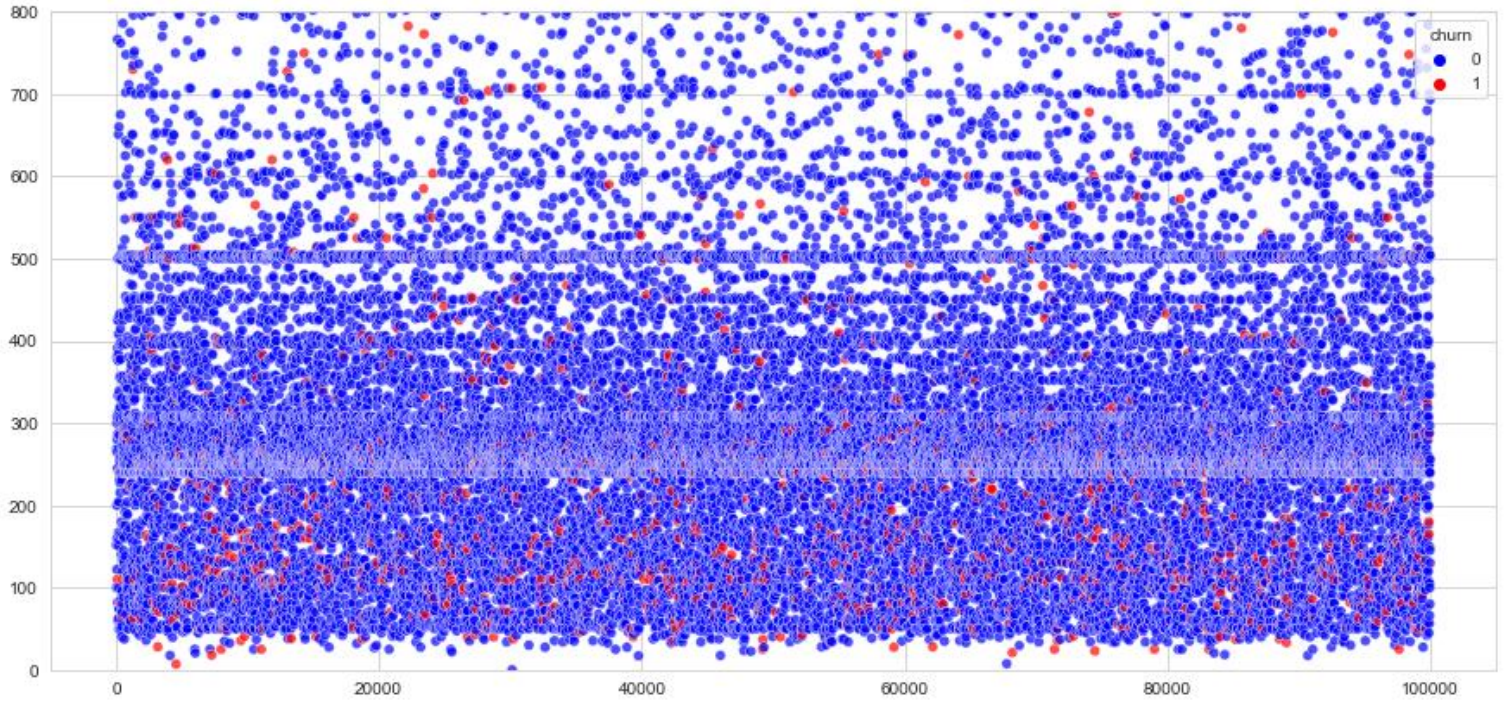

sachet_3g_8
churn
0
1
churn


0.00175
0.00150
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
Density
0
5000
10000
15000
20000
25000
30000
35000
Avg. Revenue Per User in Action phase


churn
0
1
6000
5000
4000
3000
2000
1000
0
count
0_1_yr
1_2_yr
2_3_yr
3_5_yr
5_12_yr
tenure

EDA (Bivariant Analysis)

Maximum recharge amount:



EDA - INFERENCES

- Decrease_rech_num_actionphase:
  The churn rate is more for the customers, whose number recharge in the action phase is lesser than the amount in good phase.
- Decrease_vbc_action:
  The churn rate is more for the customers, whose volume based cost in action month is increased. That means the customers do not do the monthly recharge more when they are in the action phase.
- Last_day_rech_amt:
  Noticeable drop in the last day recharge amount in the 8th month (action phase) for churned customers.
- Volume of data usage:
  i. Drop in volume of 2G and 3G data usage substantially in the 8th month i.e action phase for churned customers.
  ii. Usage of 3G data is comparatively lesser than that of 2G data, though the drop seems to follow similar pattern0
- 'monthly_2g' and 'monthly_3g' usage:
- Drop in monthly 2G and 3G subscriptions for churned customers in 8th Month i.e action phase.
- 2G (sachet_2g') and 3G('sachet_3g'):'sachet_2g' and 'sachet_3g' schemes we largely used in the good phase(6th and 7th months) and then the trend shows a sudden drop in the usage as we approach the 8th month i.e action phase for churned customers.

EDA - INFERENCES

- Maximum recharge amount:
  Users who had the max recharge amount less than 200 churned more.
- AON(Age on Network)/Tenure:
  As the number of years on network (i.e AON) increases, the number of people churned out (i.e. churn: 1) are reduced. Churn rate high in 0 - 2 years.
- average revenue per customer(ARPU) (churn and not churn) in the action phase
- ARPU for the churned customers is mostly densed on lower range than ARPU for the not churned customers. The higher ARPU customers are less likely to be churned.

Handling Imbalanced Dataset:

- Churn rate is less than 10% of the overall data available. This indicates that we would need to handle the class imbalance in this classification problem.
- We can handle imbalanced classes by balancing the classes by increasing minority or decreasing majority. These are following few techniques.
  I. Random Under-Sampling
  II. Random Over-Sampling
  III. SMOTE - Synthetic Minority Oversampling Technique
  IV. ADASYN - Adaptive Synthetic Sampling Method
  V. SMOTETomek - Over-sampling followed by under-sampling
  VI. Tomek link: is established based on a distance between instances from two different classes which are further used for removing majority class instance.

we ran three different algorithm (Logistic Regression, Decision tree, Random forest)with above techniques to handle imbalanced classes.

We pick up **Logistic Regression with Over-Sampling method as best performing model** (based on Highest Recall value).

MODEL BUILDING

- Splitting data into train and test set (70:30 ratio)
- Use RFE to choose top 25 feature variables for model building
- Repeated model building by removing variable whose p-value is more than 0.05 and VIF value greater than 5
- Balancing all metrics, 0.55 is taken as optimal cut off point
- With 0.55 as cut off point, below are performance evaluation metrics for train & test data set.
- Final model is built and with probability threshold value of 0.55

Train set	**Test set**
Accuracy = 0.77	Accuracy = 0.77
Sensitivity/Recall = 0.77	Sensitivity/Recall = 0.77
Specificity = 0.77	Specificity = 0.77

CONCLUSION & RECOMMEDATIONS

Top 15 Features Based on Feature Importance:

1. roam_og_mou_8
2. gd_ph_vol_2g_mb
3. gd_ph_sachet_3g
4. gd_ph_spl_og_mou
5. loc_og_t2c_mou_8
6. gd_ph_isd_ic_mou
7. offnet_mou_8
8. spl_ic_mou_8
9. vol_3g_mb_8
10. gd_ph_monthly_2g
11. sachet_3g_8
12. vol_2g_mb_8
13. loc_og_t2m_mou_8
14. total_rech_num_8
15. spl_og_mou_8

From above features of importance we can conclude that, the maximum features are from action phase minutes of usage (mou).

Feature Analysis:

i. Minutes of usage is one of the most important feature in churn prediction in action phase(i.e 8th month). Minutes of usage (Voice) in the action month is a very good indicator for customer churn. As the MOU in 8th month decreases, the chances of the customer churning out increases.
ii. Drop in Last day recharge amount in the action month(8th month) is an important indicator for churn of high value customers.
iii. Churn customers seems to have a relatively High ARPU (Average Revenue Per User) in the good phase, indicating that a sudden downturn in the ARPU from good to action phase is a major indicator of the customer churn.

Business Recommendations:

i. To recently joined customers (Tenure: 0-2 years) can be provided add on incentives for certain period.
ii. Provide incentives on recharge of data or voice to high value customers in the action phase to increase customer retention.
iii. Incentives based on usage can be provided to Customer having high ARPU in good phase to drive up the ARPU in action phase and retain the customer.
iv. Can provide free or discounted local onnet and mobile usage voice minutes during the action phase.
v. To retain customers, we need higher recall. As giving an offer to an user not going to churn will cost less as compared to loosing a customer and bring new customer, we need to have high rate of correctly identifying the true positives, hence recall.

Thank You