DOMAIN ORIENTED CASE STUDY

Telecom Churn

Telecom churn

A. BUSINESS UNDERSTANDING

1. Business Problem
2. Business Objective
3. Business Domain

B. ASSUMPTION FOR PREDICTION

C. ANALYSIS STEPS

1. Data Understanding/Cleaning
2. Filtering the high-value customers
3. Tagging churn and non-churn customers
4. EDA
5. Data Preparation for Model Building
6. Building a model
7. Model Evaluation and prediction

D. RECOMMENDATIONS

A. BUSINESS UNDERSTANDING

1. Business Problem

- **Competitive Market**: High churn rates (15-25%) in telecom industry.
- **Cost Dynamics:** it costs 5-10 times more to acquire a new customer than to retain an existing one

2. Business Objective

- **Predict the churn** in the last (i.e. the ninth) month using the data (features) from the first three months

  - *Define High-Value Customers & predict churn => Focus on them to prevent revenue leakage*
  - *Identify customers at high risk of churn and identify the main indicators of churn*

3. Business Domain

- **Payment Methods in Telecom**
  - *Postpaid: clear termination*
  - *Prepaid: ambiguous cessation*
  - ⇨ *Churn prediction is usually more critical for prepaid customers*
- **Definition of Churn**
  - *Revenue-Based Churn: Customers not generating revenue (e.g., less than INR 4 per month).*
  - *Usage-Based Churn: Customers with no usage (calls, internet, SMS) over a period.*
  - ⇨ *Chosen Definition: Usage-based definition for identifying churn.*
- **High Value Customers:** *top 20% of users generating 80% of revenue*
- **Region Specific**: *Indian and Southeast Asian markets where prepaid models are predominant.*

B. ASSUMPTION FOR PREDICTION

Customer Lifecycle Phases in Churn Prediction

Good Phase: Customer is satisfied and uses the service as usual.

Month 1 – 2
(June & July)

Month 3
(Aug)

Action Phase: Customer starts showing signs of dissatisfaction

(e.g., influenced by competitor offers, facing service issues).

Month 4
(Sep)

Churn Phase: Customer has churned

C. ANALYSIS STEPS

1. Data Understanding/Cleaning
2. Filtering the high-value customers

Steps	Our Comments	Actions	Remarks
1. Data Understanding/ Cleaning		**Importing necessary packages & libraries, loading the dataset into a data frame and review overall data frame (df)**	Original df: 99999 rows and 226 columns
		Checking and Handling Missing Value *Dropping columns with over 30% of the values missing*	df: 99999 rows and 186 columns
	We dropped the Date columns as we are trying to predict customer churn using Logistic Regression, not time series data We also dropped 'Circle ID as it is not affecting the data analysis in any way	**Checking and Dropping unnecessary columns** *Dropping the Date columns* *Dropping 'Circle ID' column*	df: 99999 rows and 178 columns df: 99999 rows and 177 columns
2. Filtering the high-value customers	We added the ***avrg_rechrg_amt_6_7*** columns to filter high-value customers up to the 70th percentile, resulting in around 30K rows as mentioned in the problem	**Filtering the high-value customers**	df: 30011 rows and 178 columns
	We found 114 rows with more than 50% missing values and decided to drop them To ensure data quality, we checked for missing values again and found that Minutes of Usage - Voice Calls in Jun, Jul, Aug, Sep had missing values. We then dropped rows where all related columns were missing.	**Checking and Handling Missing Value after filtering** *Dropping rows where all related Minutes of Usage - Voice Calls in Sep columns were missing.* *We also do similar and sequential tasks for the rows in August, June, July*	df: 29897 rows and 178 columns df: 28307 rows and 178 columns df: 27991 rows and 178 columns

C. ANALYSIS STEPS

1. Data Understanding/Cleaning

2. Filtering the high-value customers

→ Post data cleaning and filtering, our data frame of 27,991 rows and 178 columns retains ~93% of the high-value data subset (27,991/30,011)

→ Based on the definition of Churn in the Business Domain part and facts given with the understanding of the Data, we can infer that:

- High-value customers when identified can reduce revenue loss at large and is related to ARPU and RECH variable.
- Both high-value customers and churners will directly relate to the amount of revenue generated depending on the ARPU variable
- Churners will have low usage of both internet and calls, for the same reason churners will have a relation with variables like DATA, 3G, VOL, 2G, LOC, and STD

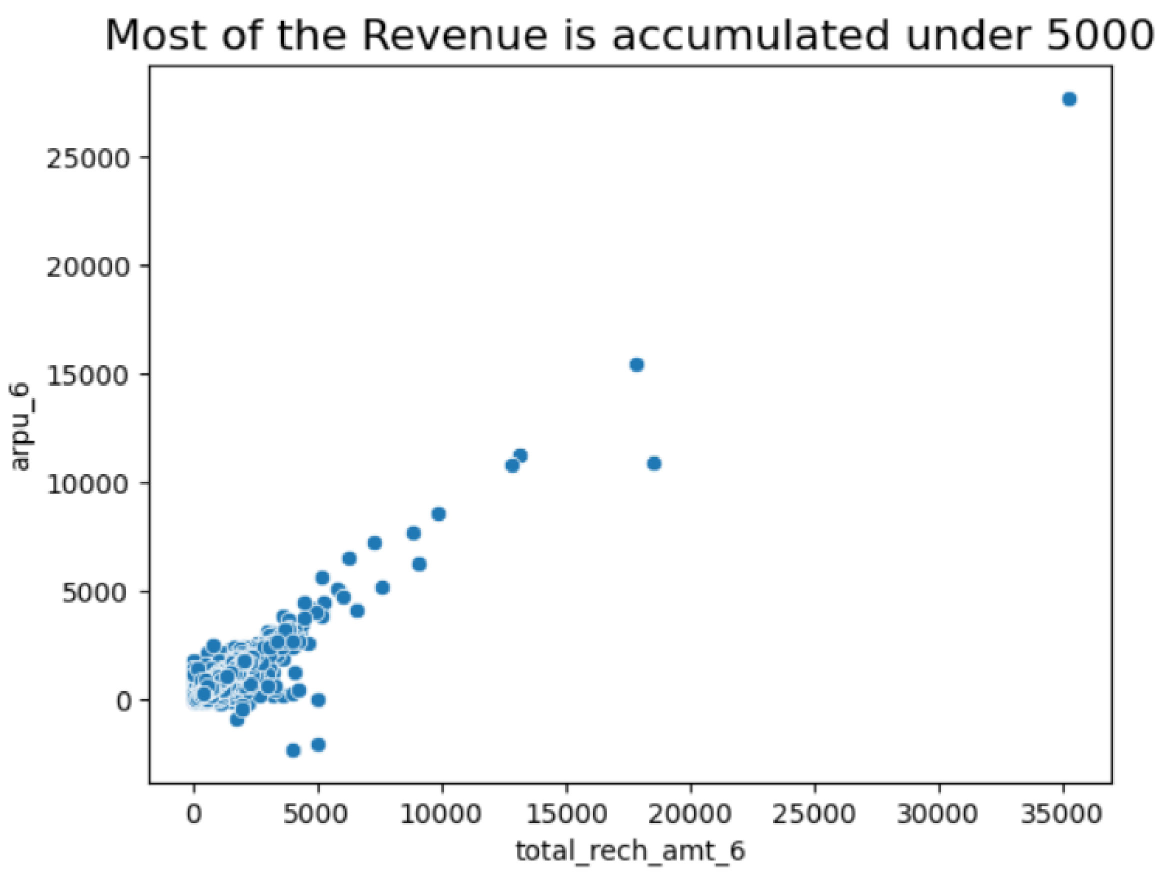
Note

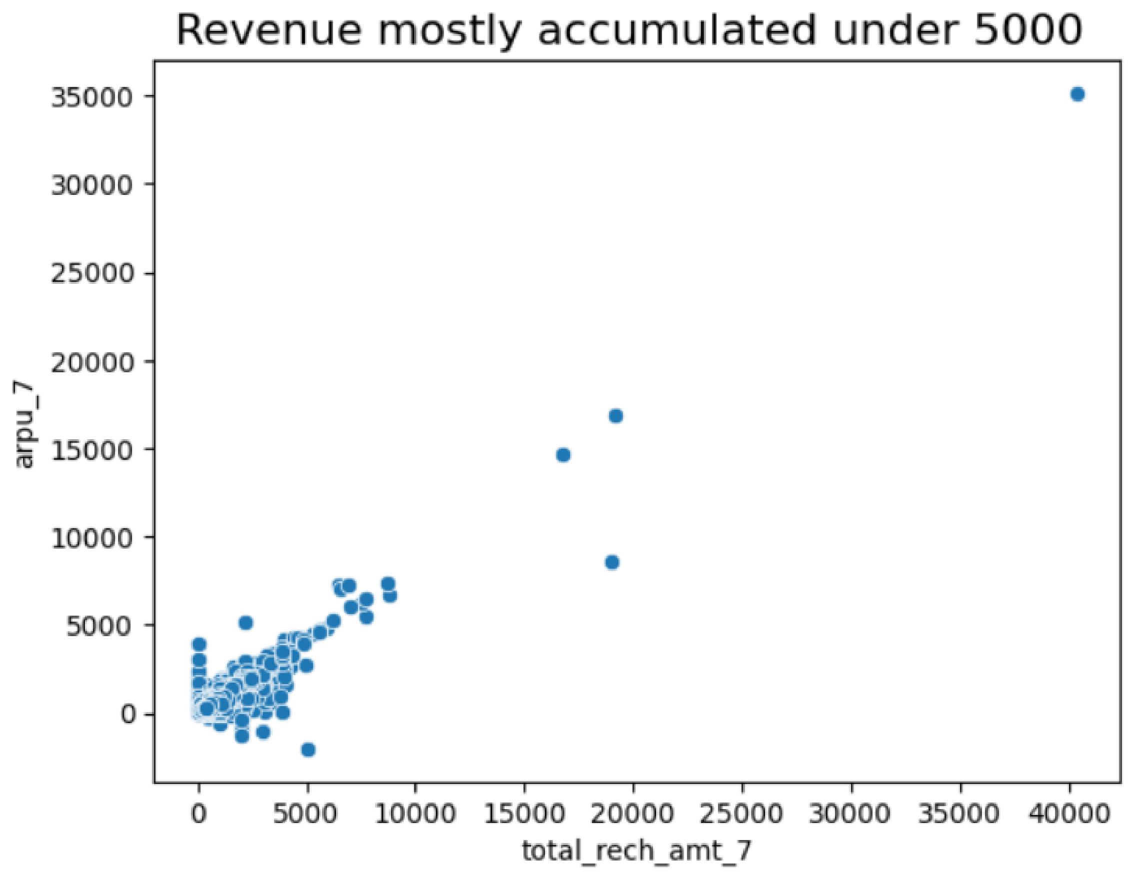
3. Business Domain

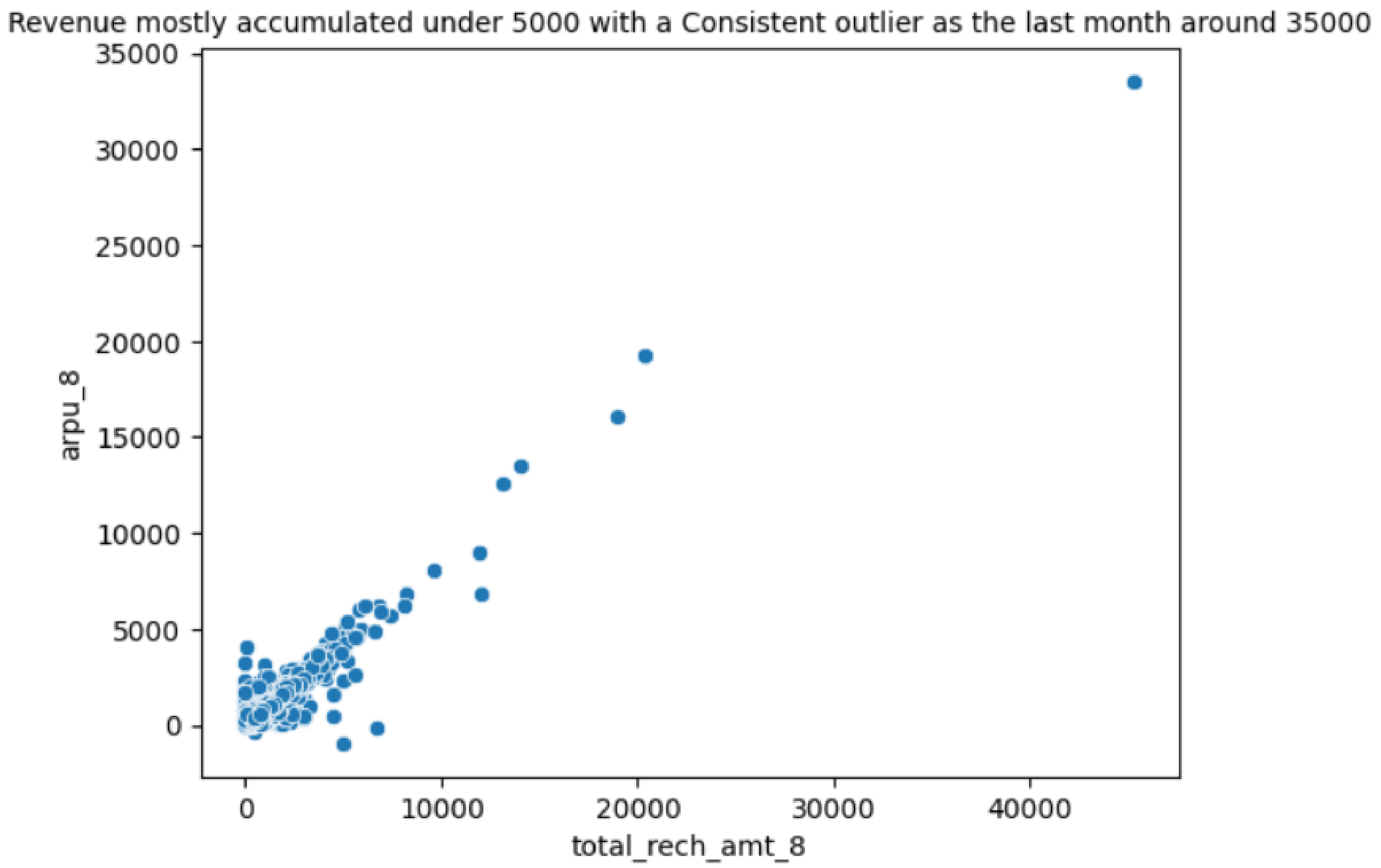
- **Payment Methods in Telecom**
  - *Postpaid: clear termination*
  - *Prepaid: ambiguous cessation*
  - ⇨ *Churn prediction is usually more critical for prepaid customers*
- **Definition of Churn**
  - *Revenue-Based Churn: Customers not generating revenue (e.g., less than INR 4 per month).*
  - *Usage-Based Churn: Customers with no usage (calls, internet, SMS) over a period.*
  - ⇨ *Chosen Definition: Usage-based definition for identifying churn.*
- **High Value Customers:** *top 20% of users generating 80% of revenue*
- **Region Specific:** *Indian and Southeast Asian markets where prepaid models are predominant.*

C. ANALYSIS STEPS

Check correlation between ARPU and RECH variable in June, July and August







➔ Based on the graphs, a linear relationship exists between total recharge amount and average revenue per user. we can see that most of revenue is accumulated under 5000 in the first month with an outlier going uptil 25000. But in the next 2 months we can see that the revenue slowly creeping into the 10000 as well with outliers lying around 3500

C. ANALYSIS STEPS

3. Tagging churn and non-churn customers

We also tagged the churned customers (churn=1, else 0) as the problem statement and removed all the attributes corresponding to the churn phase.

➔ As the result, the percentage of churn and non-churn respectively is 3.39% and 96.61%.

4. Exploratory Data Analysis

We used to IQR score method to determine outlier in numeric columns *(except 'Churn', 'mobile number' columns)*

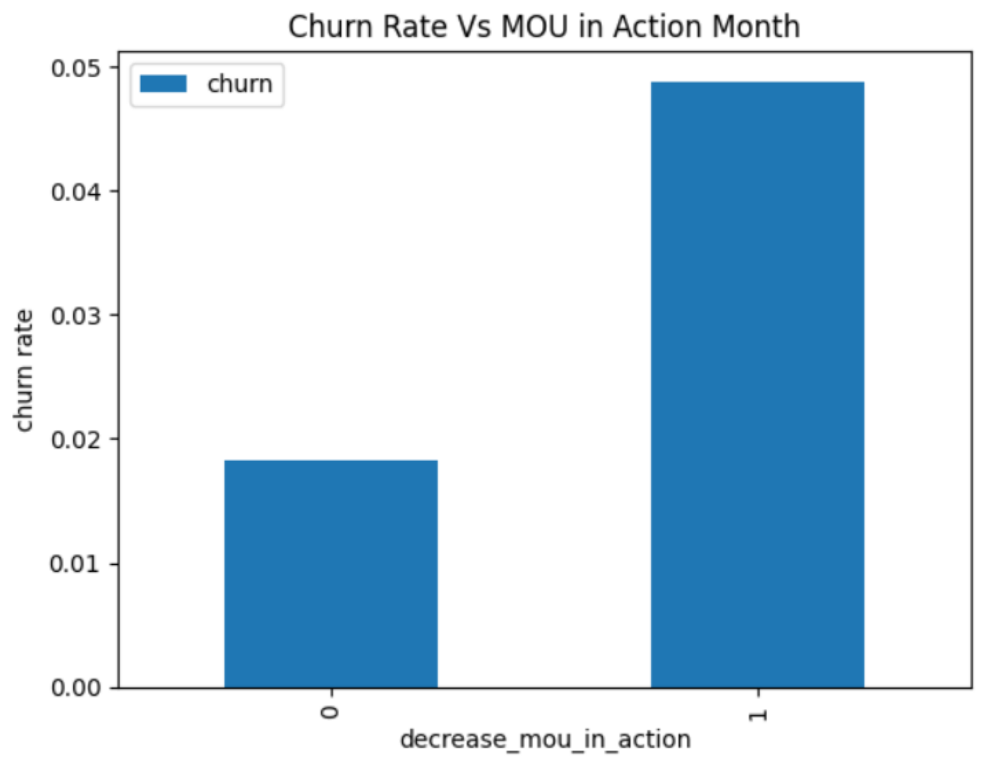
Also, adding new columns that indicate amount of decrease in the action phase compared to in the good phase for Attributes such as :

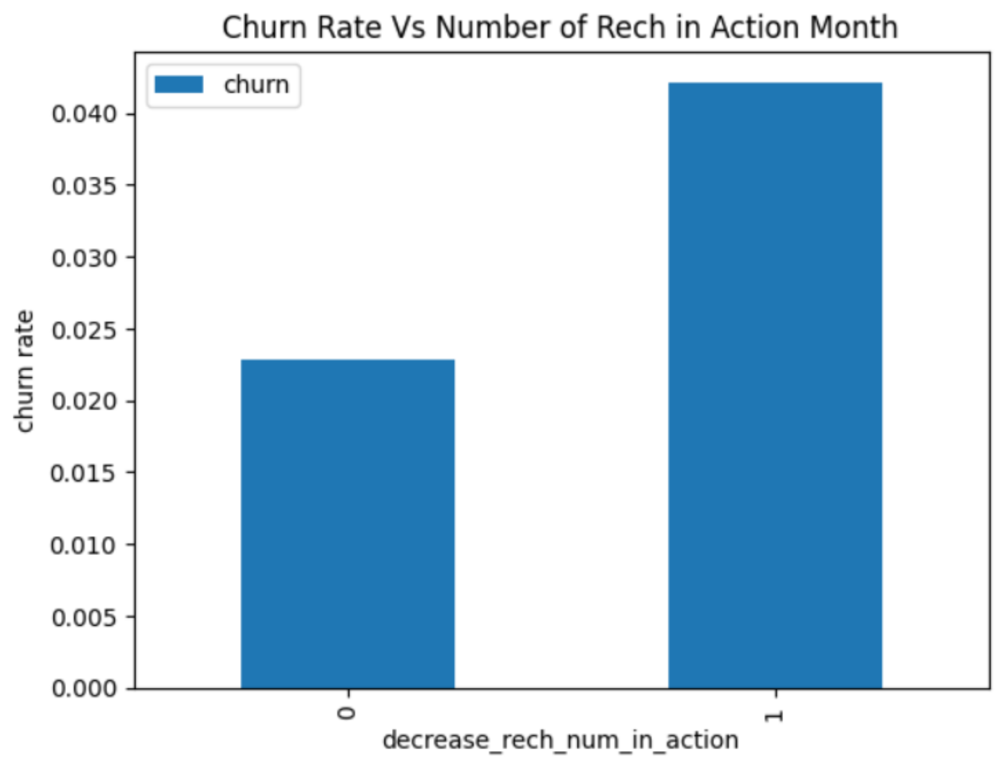
- MOU: decrease_mou_in_action (Decrease=1, else 0)
- RECH_NUM: decrease_rech_num_in_action (Decrease=1, else 0)
- RECH_AMT: decrease_rech_amt_in_action (Decrease=1, else 0)
- ARPU: decrease_arpu_in_action (Decrease=1, else 0)
- VOL: decrease_vbc_in_action (Decrease=1, else 0)

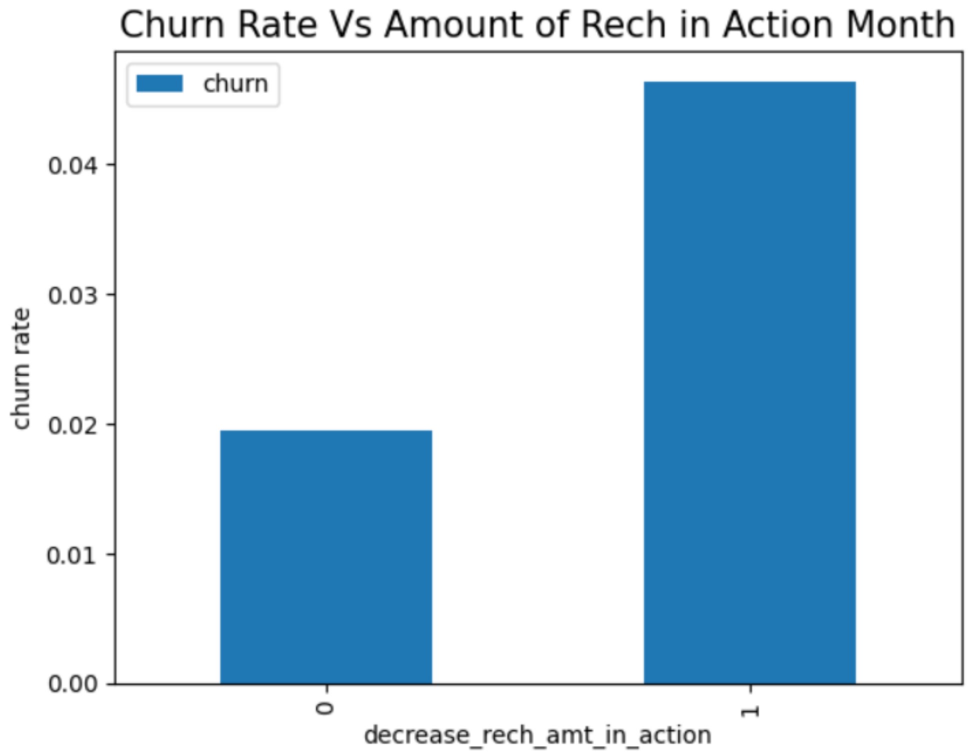
C. ANALYSIS STEPS

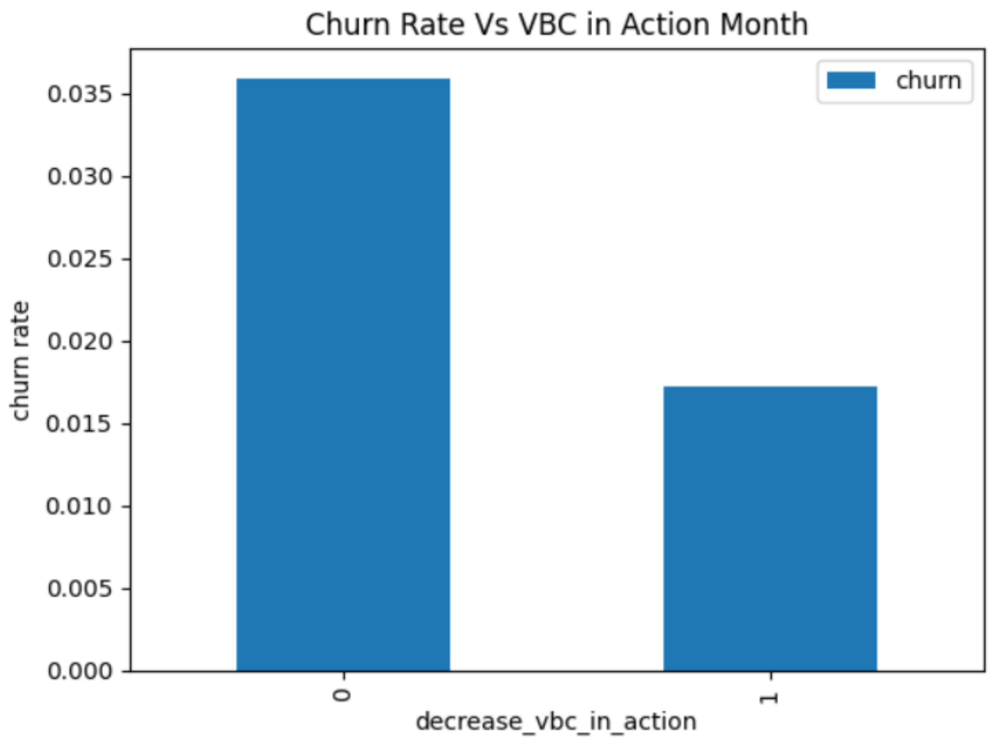
4. Exploratory Data Analysis

Univariate Analysis









➔ From these graphs in comparison Churn Rate vs MOU/Number of Rech/ Amount of Rech/ VBC, we can see that:

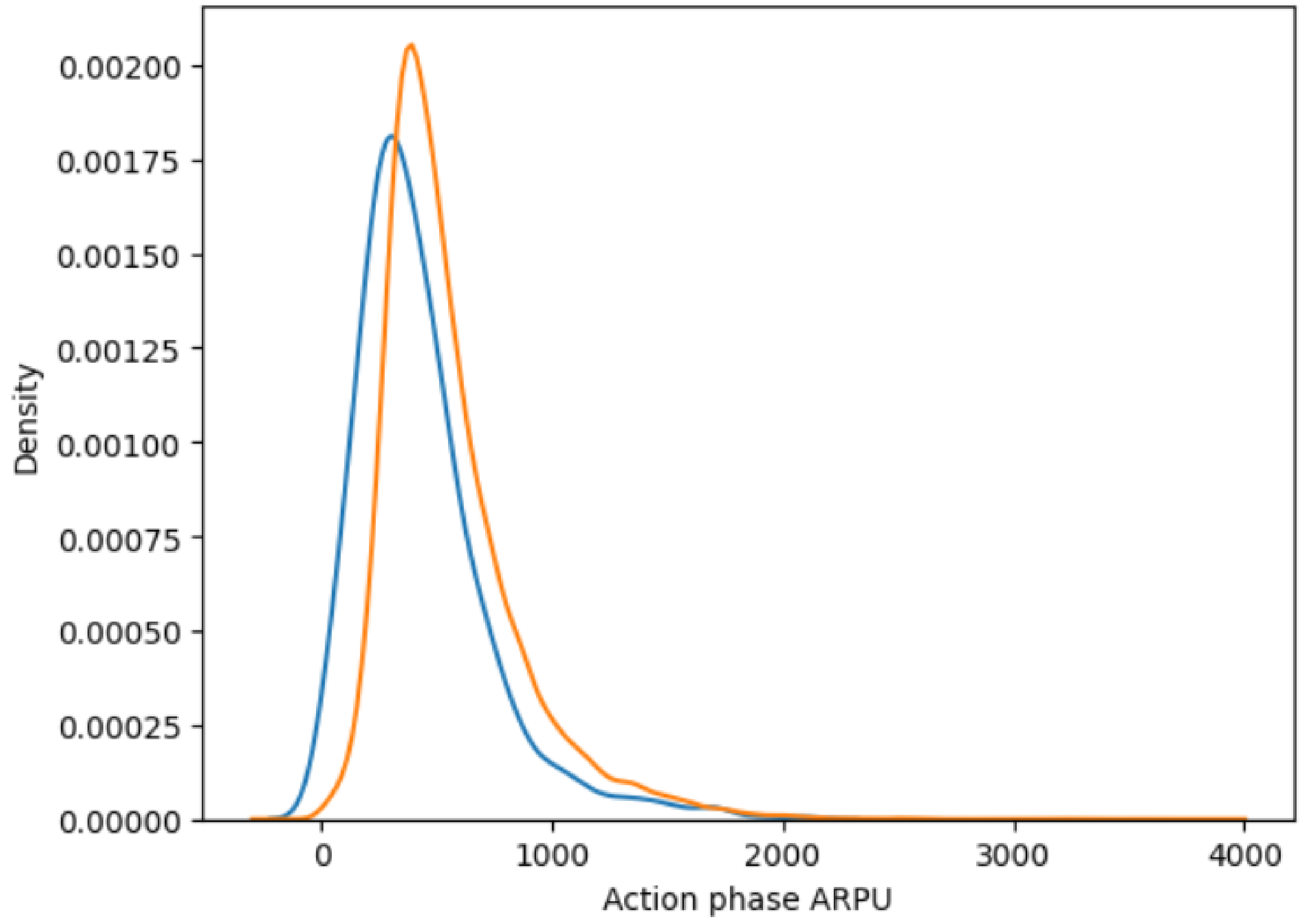
Churn rate is higher for the customers:

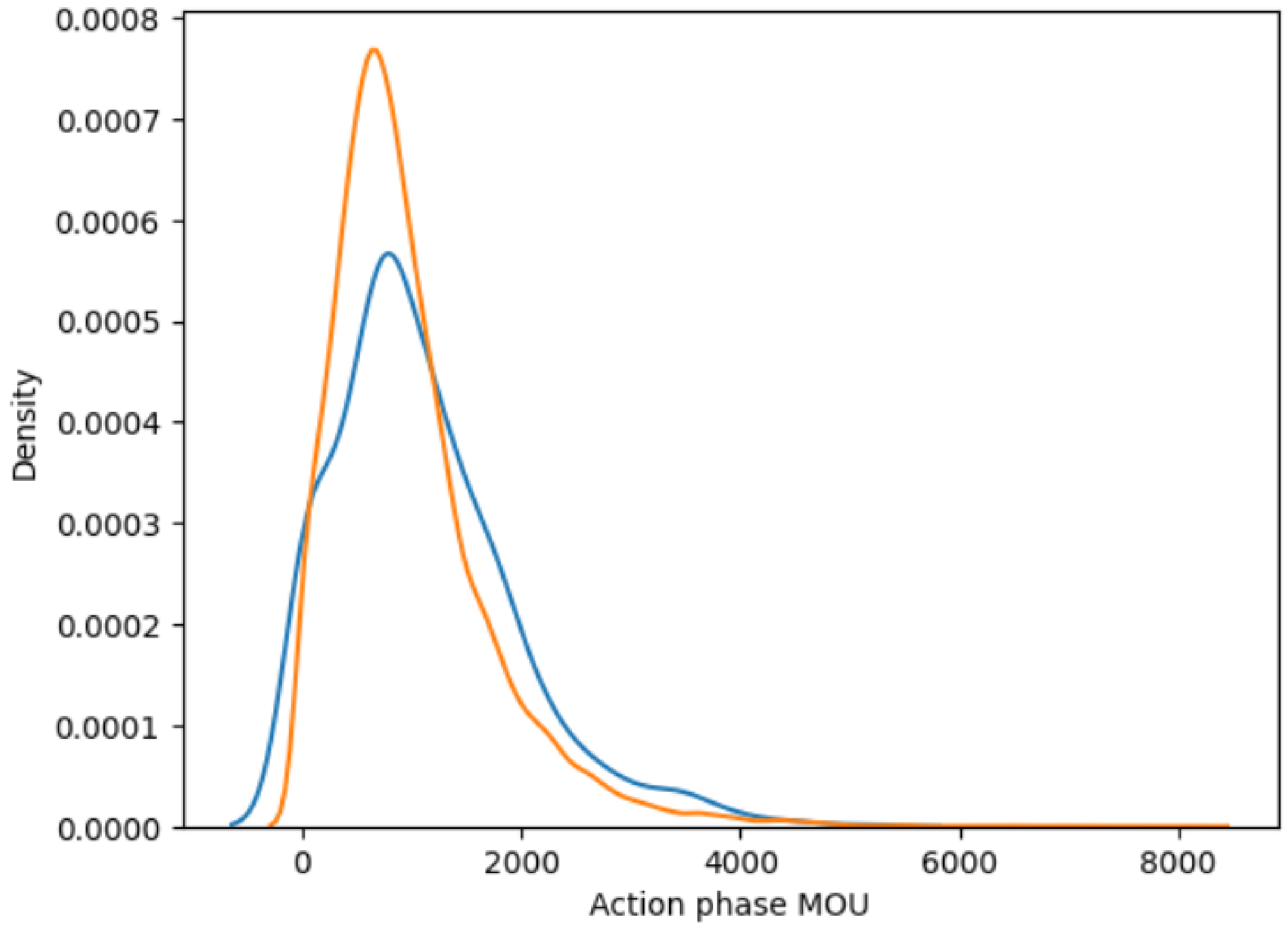
- whose MOU decreased in the action month
- whose number of recharge decreased in the action month
- whose amount of Recharge is decreased in the action month
- whose VBC is increased in the Action month, which means that customers are not prone to doing monthly recharge.

C. ANALYSIS STEPS

4. Exploratory Data Analysis

Univariate Analysis: ARPU





➔ From these graphs:

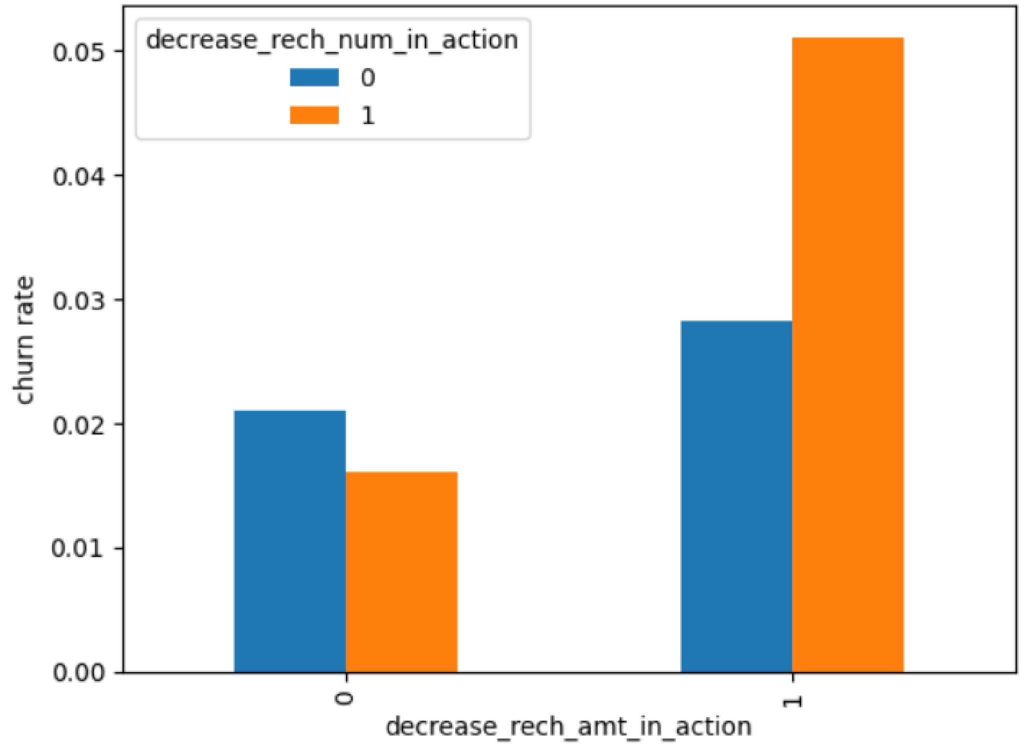
- Analysis shows our previous inference to be true that Higher the rate of ARPU means lower the rate of churn.
- We can see and count our inference to be true that Higher the rate of MOU lesser the rate of churn will be.

C. ANALYSIS STEPS

4. Exploratory Data Analysis

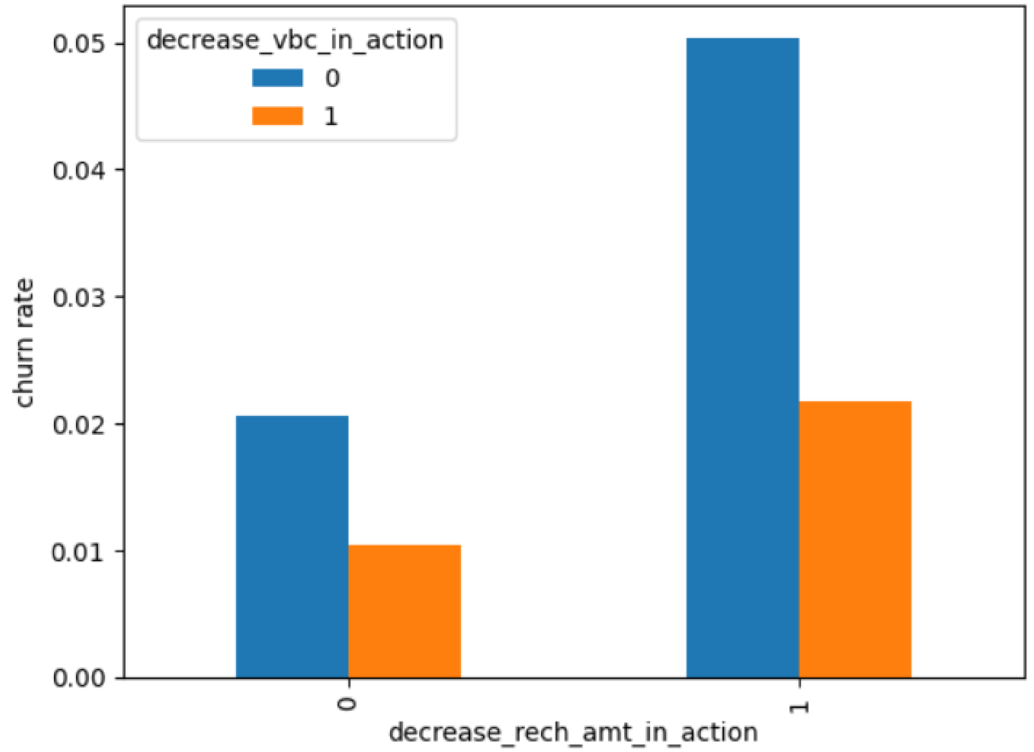
Bivariate analysis

Churn Rate vs Decrease in Amt of RECH and Num of Rech

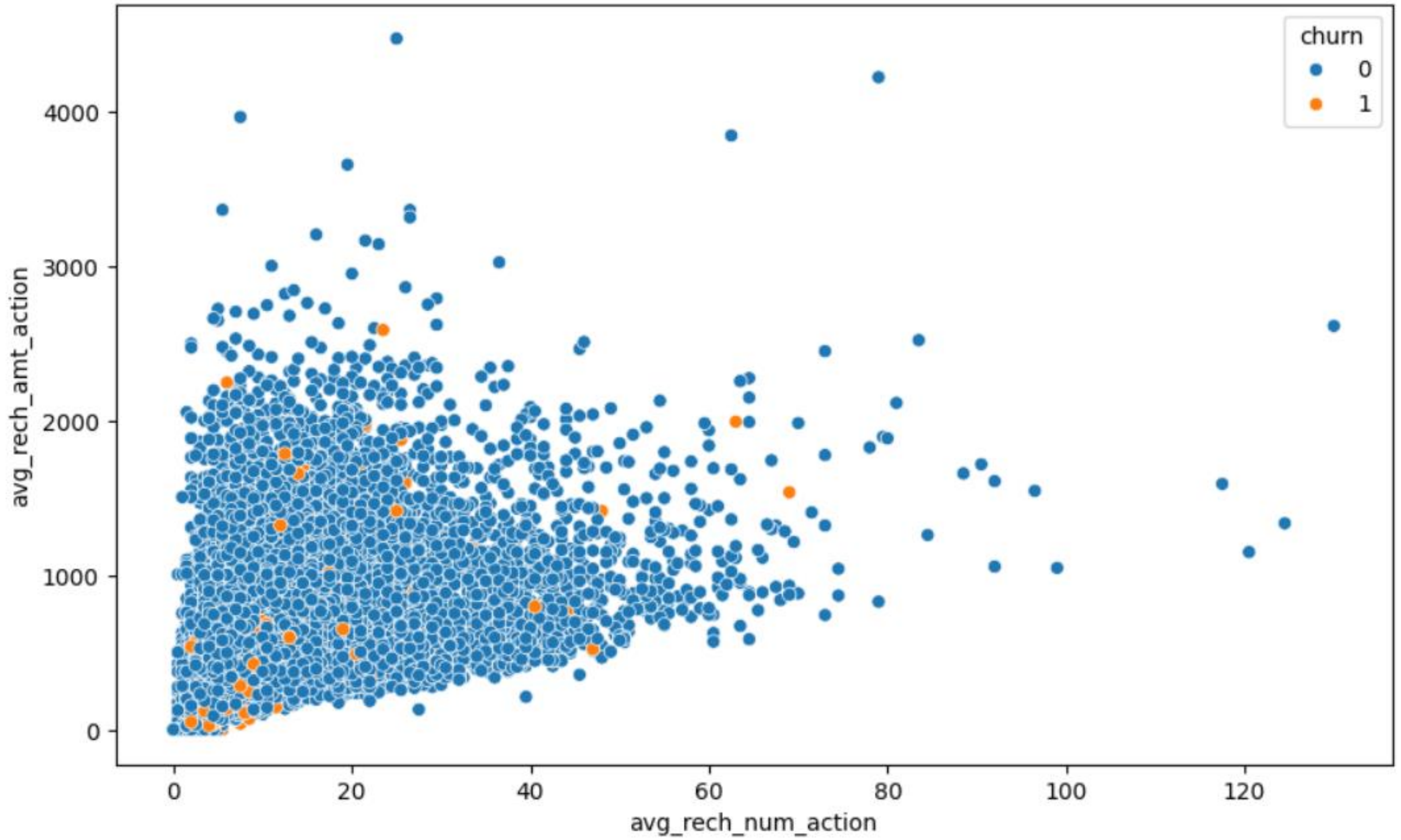


➔ The churn rate is higher for customers whose recharge amount and number of recharge have decreased in the action phase compared to the good phase

Churn Rate VS Decrease in RECH amt and VBC



➔ Churn rate is higher for the customers, whose recharge amount is decreased along with the volume-based cost is increased in the action month



➔ From the scatterplot, we can see that the percentage of churners is quite low, and it also displays a distinct proportionality

C. ANALYSIS STEPS

5. Data Preparation for Model Building

Once we have completed our data cleaning, dropped the necessary columns, and conducted our Univariable/Bivariable analysis, we can begin preparing our dataset for model building. This involves the following sequential steps:

Feature Engineering | **Train Test Split** | **Feature Scaling**

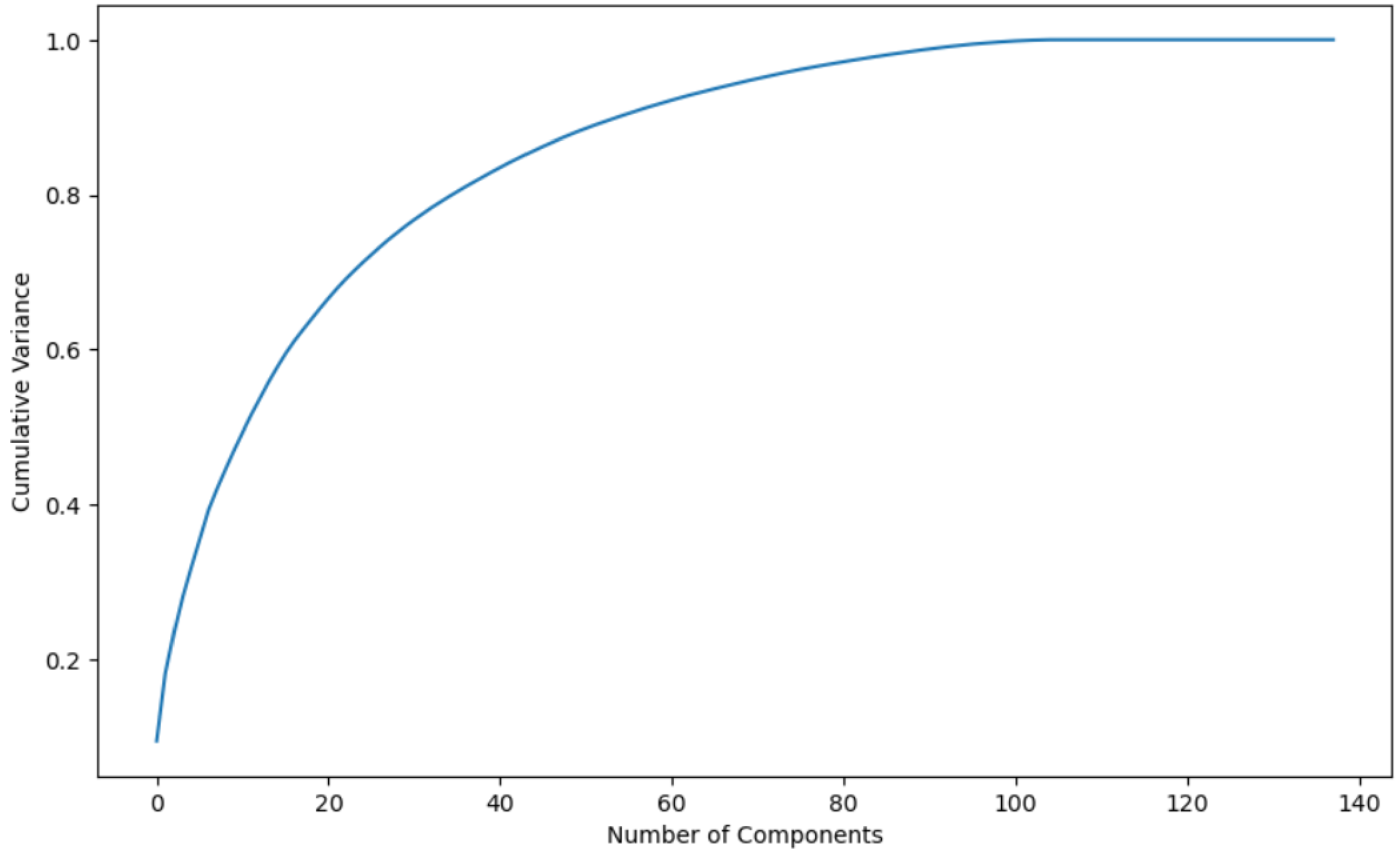
Besides, to select the model that best meets the business objectives, we implemented steps taken *(implement with PCA and w/o PCA)*:

- Model Training and Prediction
- Model Evaluation
- Model Comparison
- Final Model Selection

C. ANALYSIS STEPS

5. Data Preparation for Model Building

Before running the models, we implement PCA to reduce the dimensionality of the dataset and expectation resulting in a more efficient and potentially more effective model.



- ➔ We can see that 60 components explain almost more than 90% of the variance of the data.
- ➔ We will perform PCA with 60 components

C. ANALYSIS STEPS

6. Building a model

<u>Model : Implement Several Models with PCA</u>

Model summary

	Logistic Regression		Support Vector Machine(SVM)		Decision Tree		Random Forest	
Metric	*Train Performance*	*Test Performance*	*Train Performance*	*Test Performance*	*Train Performance*	*Test Performance*	*Train Performance*	*Test Performance*
Accuracy	0.96	0.96	0.96	0.96	0.96	0.96	0.84	0.80
Sensitivity	0.064	0.082	0.002	0.00	0.10	0.07	0.88	0.75
Specificity	0.99	0.99	1.00	1.00	0.99	0.99	0.81	0.80

<u>Final conclusion with PCA</u>

- After trying several models, all models seem to perform reasonably well on the training data, but there are differences in how well these models generalize to the test data.
- For achieving the best sensitivity, which was our ultimate goal, the classic Logistic regression or the SVM models preforms well.

C. ANALYSIS STEPS

6. Building a model

Model : Implement Logistic Regression with No PCA

Model	Description	Results and Action from model
Model - 1	We implement Feature Selection using RFE with 15 columns and ran Model 1	Removing **og_others_8 column**, which is insignificant as it has the highest p-value 0.99
Model - 2	We ran Model 2 after removing og_others_8 column	All the variables p-values are significant, but offnet_mou_8 column has the highest VIF 7.45 => Removing **offnet_mou_8 column**
Model - 3	We ran Model 3 after removing offnet_mou_8 column	all the variables are significant and there is no multicollinearity among the variables

➔ After running Models and checking the p-value and VIF, we can conclude that **Model - 3 log_no_pca_3 will be the final model**

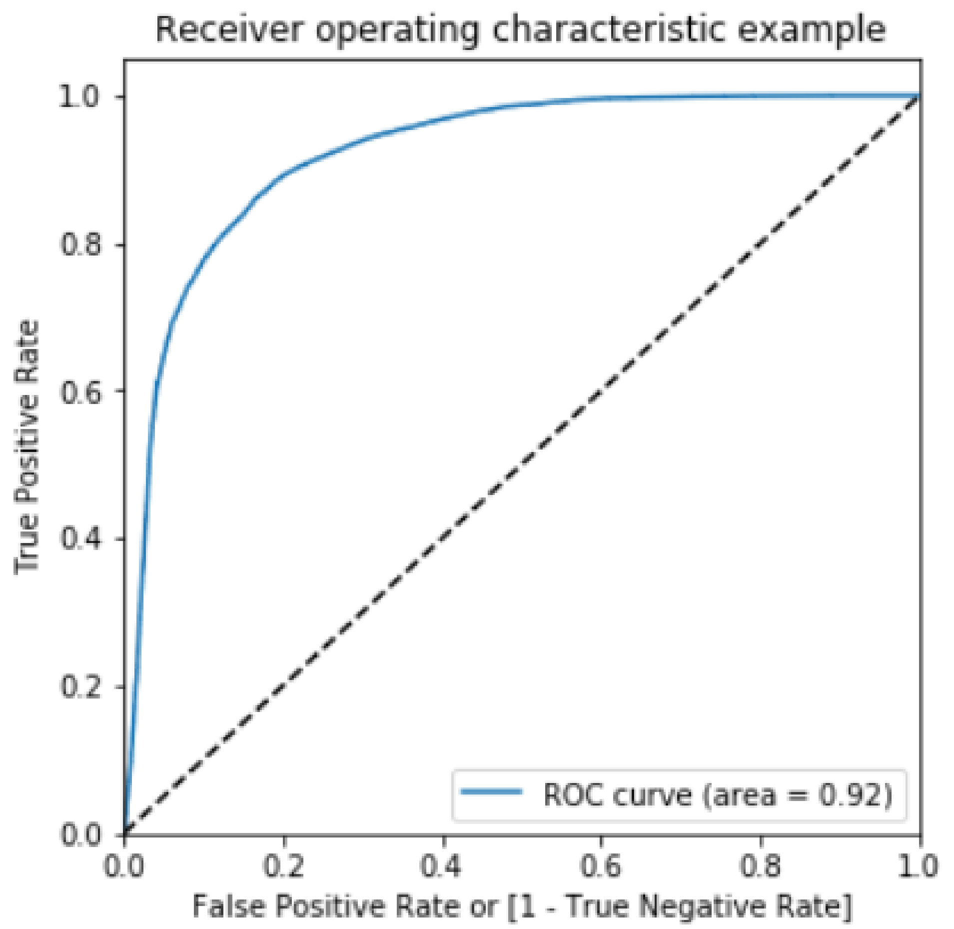
Model - 3 summary

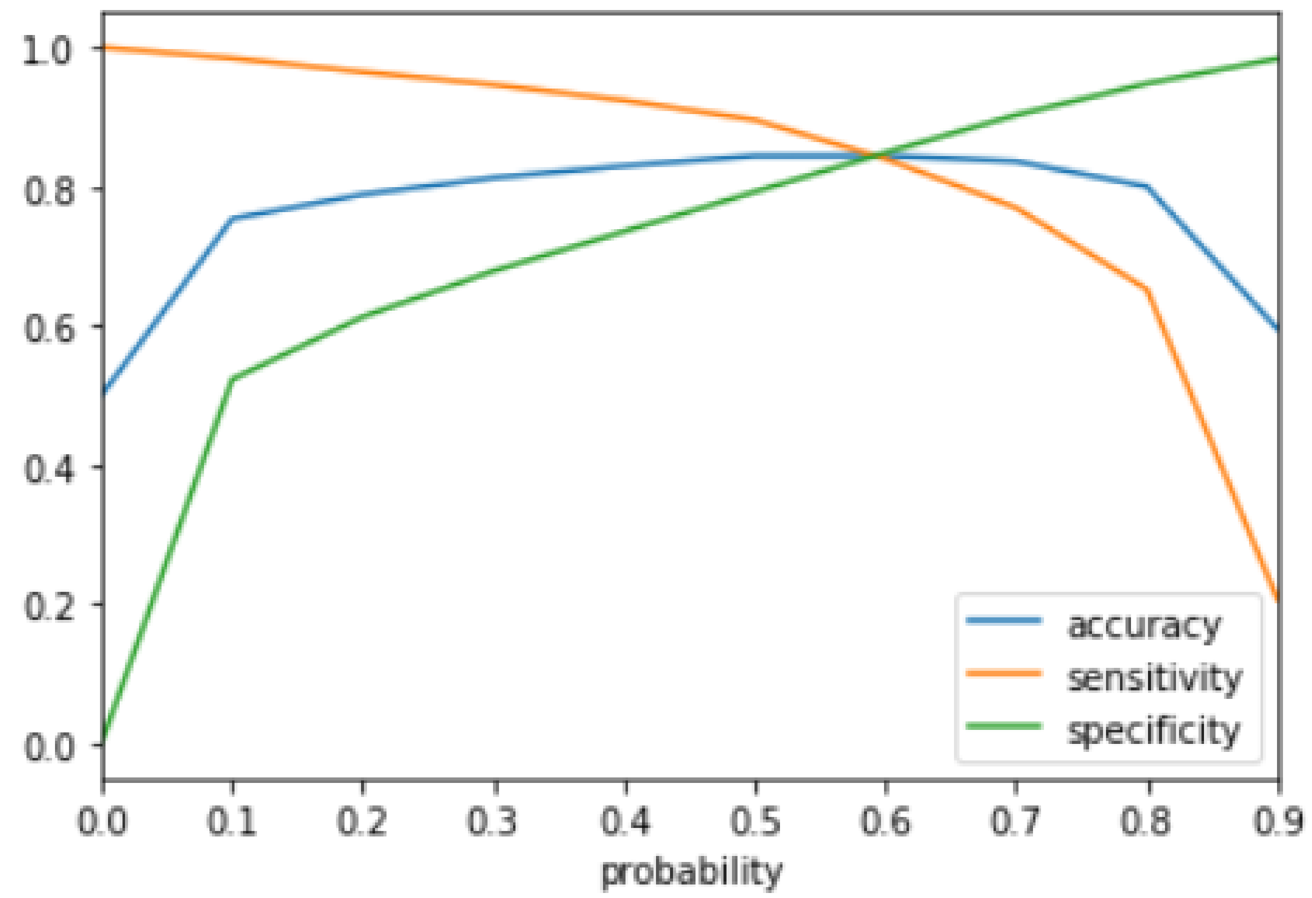
Dep. Variable:	churn	No. Observations:	42850
Model:	GLM	Df Residuals:	42836
Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-15720.
Date:	Sat, 16 May 2020	Deviance:	31440.
Time:	18:07:30	Pearson chi2:	3.92e+06
No. Iterations:	11		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2058	0.032	-37.536	0.000	-1.269	-1.143
offnet_mou_7	0.3665	0.022	16.456	0.000	0.323	0.410
roam_og_mou_8	0.7135	0.024	29.260	0.000	0.666	0.761
std_og_t2m_mou_8	-0.2474	0.022	-11.238	0.000	-0.291	-0.204
isd_og_mou_8	-1.3811	0.212	-6.511	0.000	-1.797	-0.965
og_others_7	-2.4711	0.872	-2.834	0.005	-4.180	-0.762
loc_ic_t2f_mou_8	-0.7102	0.075	-9.532	0.000	-0.856	-0.564
loc_ic_mou_8	-3.3287	0.057	-58.130	0.000	-3.441	-3.216
std_ic_t2f_mou_8	-0.9503	0.078	-12.181	0.000	-1.103	-0.797
ic_others_8	-1.5131	0.129	-11.771	0.000	-1.765	-1.261
total_rech_num_8	-0.5060	0.018	-28.808	0.000	-0.540	-0.472
monthly_2g_8	-0.9279	0.044	-21.027	0.000	-1.014	-0.841
monthly_3g_8	-1.0943	0.046	-23.615	0.000	-1.185	-1.004
decrease_vbc_action	-1.3293	0.072	-18.478	0.000	-1.470	-1.188

C. ANALYSIS STEPS

7. Model Evaluation and prediction





- **ROC curve:** We can see the area of the ROC curve is closer to 1, which is the Gini of the model.
- **Sensitivity – Specificity – Accuracy Plot**:

At a probability cut off 0.6:

- Accuracy: Becomes stable around 0.6
- Sensitivity - Decreases with the increased probability
- Specificity - Increases with the increasing probability.

'At point 0.6' where the three parameters cut each other, we can see that there is a balance between sensitivity and specificity with a good accuracy.

- **Cutoff Probability Selection:** Here we are intended to achieve better sensitivity than accuracy and specificity. Though as per the above curve, we should take 0.6 as the optimum probability cutoff, we are taking **0.5** for achieving higher sensitivity, which is our main goal.

C. ANALYSIS STEPS

Results of our final model

Metric	Train Performance	Test Performance
Accuracy	0.84	0.78
Sensitivity	0.81	0.82
Specificity	0.83	0.78

➔ Overall, the model is performing well in the test set, what it had learnt from the train set.

Final conclusion with no PCA

We can see that the logistic model with no PCA has good sensitivity and accuracy, which are comparable to the models with PCA. So, we can go for the more simplistic model such as logistic regression with no PCA.

D. RECOMMENDATION

Variables	Coefficients
loc_ic_mou_8	-3.3287
og_others_7	-2.4711
ic_others_8	-1.5131
isd_og_mou_8	-1.3811
decrease_vbc_action	-1.3293
monthly_3g_8	-1.0943
std_ic_t2f_mou_8	-0.9503
monthly_2g_8	-0.9279
loc_ic_t2f_mou_8	-0.7102
roam_og_mou_8	0.7135

'top variables' selected in the logistic regression model *(as picture)*

We can see most of the top variables have negative coefficients. That means, the variables are inversely correlated with the churn probablity.

E.g.: If the local incoming minutes of usage (loc_ic_mou_8) is lesser in the month of August than any other month, then there is a higher chance that the customer is likely to churn.

D. RECOMMENDATION

1. Target customers, whose minutes of usage of the incoming local calls and outgoing ISD calls are less in the action phase (mostly in the month of August).
2. Target customers, whose outgoing others charge in July and incoming others in August are less.
3. Customers having value-based cost in the action phase increased are more likely to churn than the other customers. Hence, these customers may be a good target to provide offer.
4. Customers who have a higher monthly 3G recharge in August are likely to churn.
5. Customers having decreasing STD incoming minutes of usage for operators T to fixed lines of T for the month of August are more likely to churn.
6. Customers decreasing monthly 2g usage for August are most probable to churn.
7. Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.
8. *roam_og_mou_8 variables* have positive coefficients (0.7135). That means for the customers, whose roaming outgoing minutes of usage is increasing are more likely to churn.

THANK YOU