Domain Oriented Case Study – Telecom churn Case study

by

Akshay Tula

S Ravisasank Reddy

Problem Statement

In telecom, annual churn rates range from 15-25%. Industry shifts focus to retaining high-value customers, amid intense competition and acquisition costs.

Questions to be Answered

1. Define Business Objective: Predict churn in the ninth month using data from the preceding three months (June, July, August).
2. Analyze Customer Behavior: Understand the three phases of the customer lifecycle - 'good', 'action', and 'churn' phases. Identify high-churn-risk customers during the 'action' phase for proactive intervention.
3. Data Preprocessing: Encode months (June, July, August, September) as 6, 7, 8, and 9 respectively. Tag churn based on the 'churn' phase and discard corresponding data for prediction.
4. Predictive Modeling: Utilize features from the 'good' and 'action' phases to build predictive models for churn in the ninth month, focusing on high-value customers' behavior during churn.

Strategy Applied : Data Preparation

1.**High-Value Customer Selection:** Utilize the 70th percentile of average recharge amounts in the first two months to filter high-value customers, ensuring a focused analysis on those likely to generate significant revenue.

2.**Churn Identification:** Identify churners by observing specific criteria such as no call activity and zero mobile internet usage in the fourth month, enabling precise tagging of customers at risk of churn.

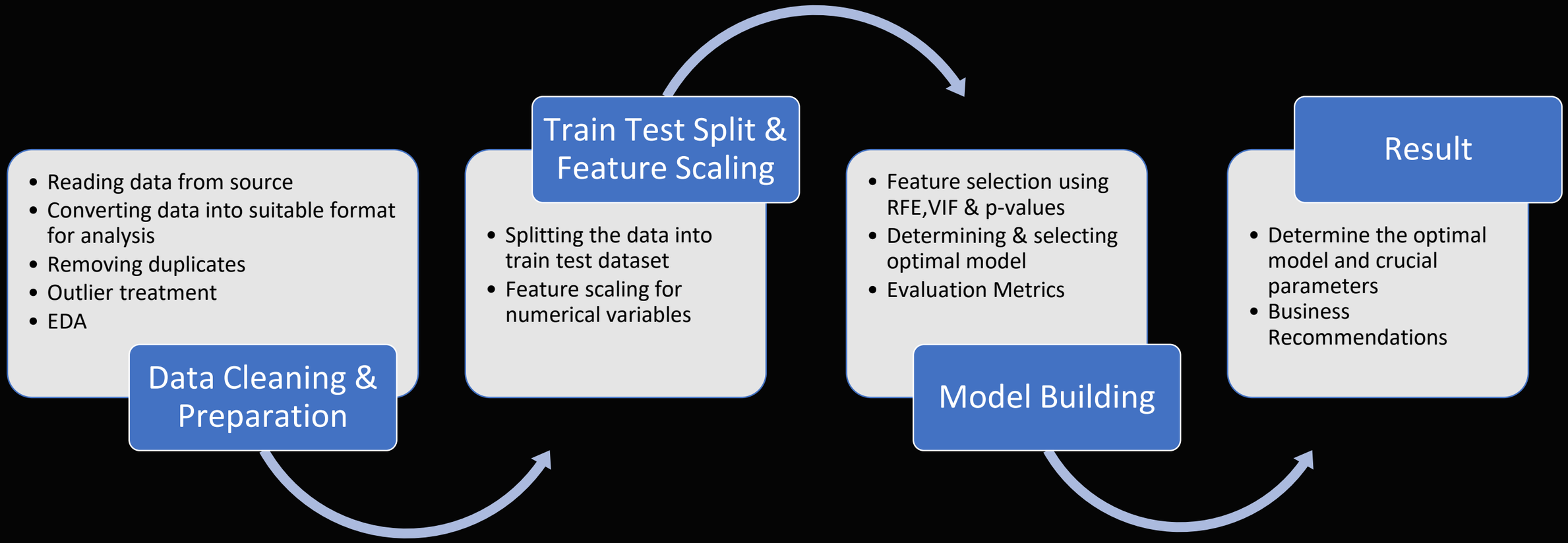
3.**Data Filtering:** Eliminate attributes related to the churn phase (month 9) to streamline the dataset and concentrate analysis on relevant 'good' and 'action' phases, optimizing predictive modeling efforts.

4.**Data Preprocessing:** Implement essential data preparation techniques such as feature engineering and handling missing values to enhance the dataset's quality and suitability for predictive modeling tasks.

Strategy Applied : Modelling

1. **Dual-Purpose Predictive Modeling:** Construct models to predict high-value customer churn, enabling proactive actions such as personalized offers or discounts, while also identifying key churn predictors for insights into customer behavior.
2. **Class Imbalance Handling: Address the challenge of class imbalance (typically 5-10%) in churn prediction models by employing techniques to balance the representation of churners and non-churners.**
3. **Attribute Importance Identification: Develop additional models, such as logistic regression or tree-based models, to uncover significant predictor attributes indicative of churn. Ensure consideration of multicollinearity issues in logistic regression.**
4. **Visual Representation of Insights: Present findings visually through plots and summary tables to effectively communicate the importance of predictor attributes, aiding decision-makers in understanding churn indicators.**
5. **Recommendation Development: Formulate actionable strategies to manage customer churn based on insights derived from predictive models, facilitating proactive retention efforts and business decision-making.**

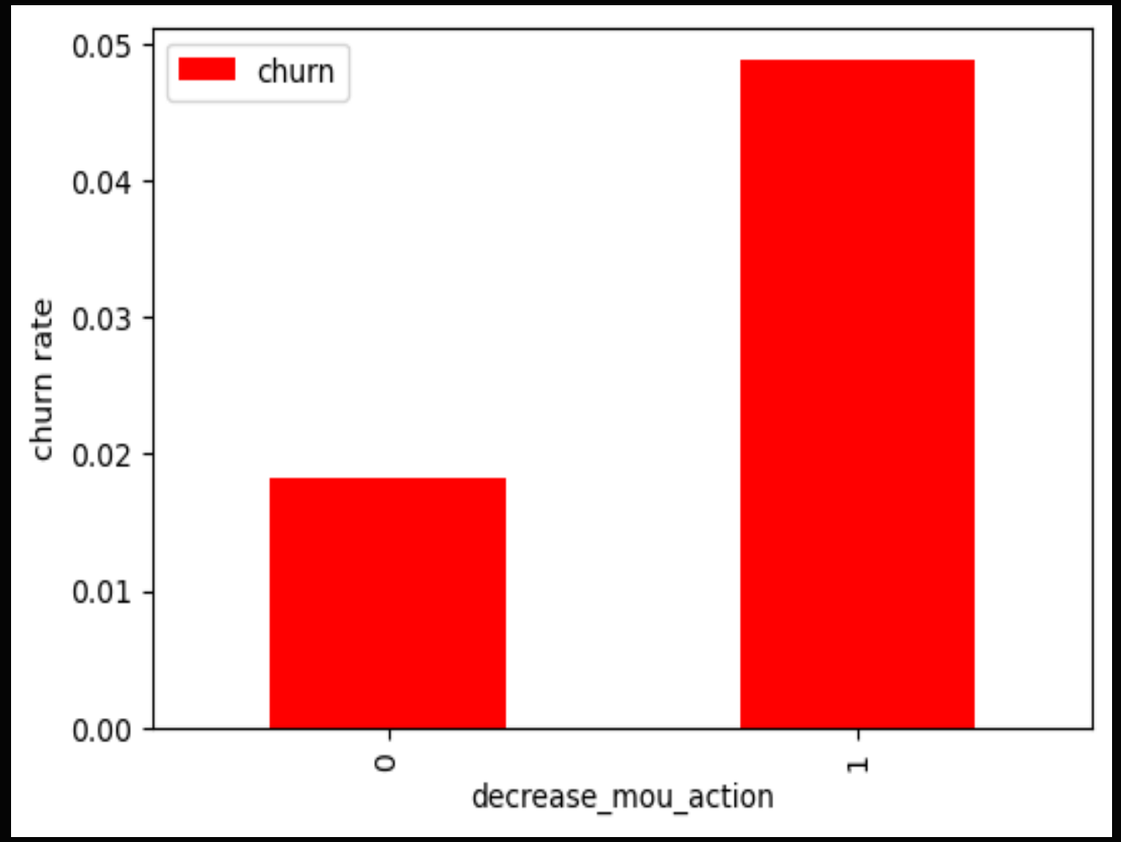
Problem Solving Methodology

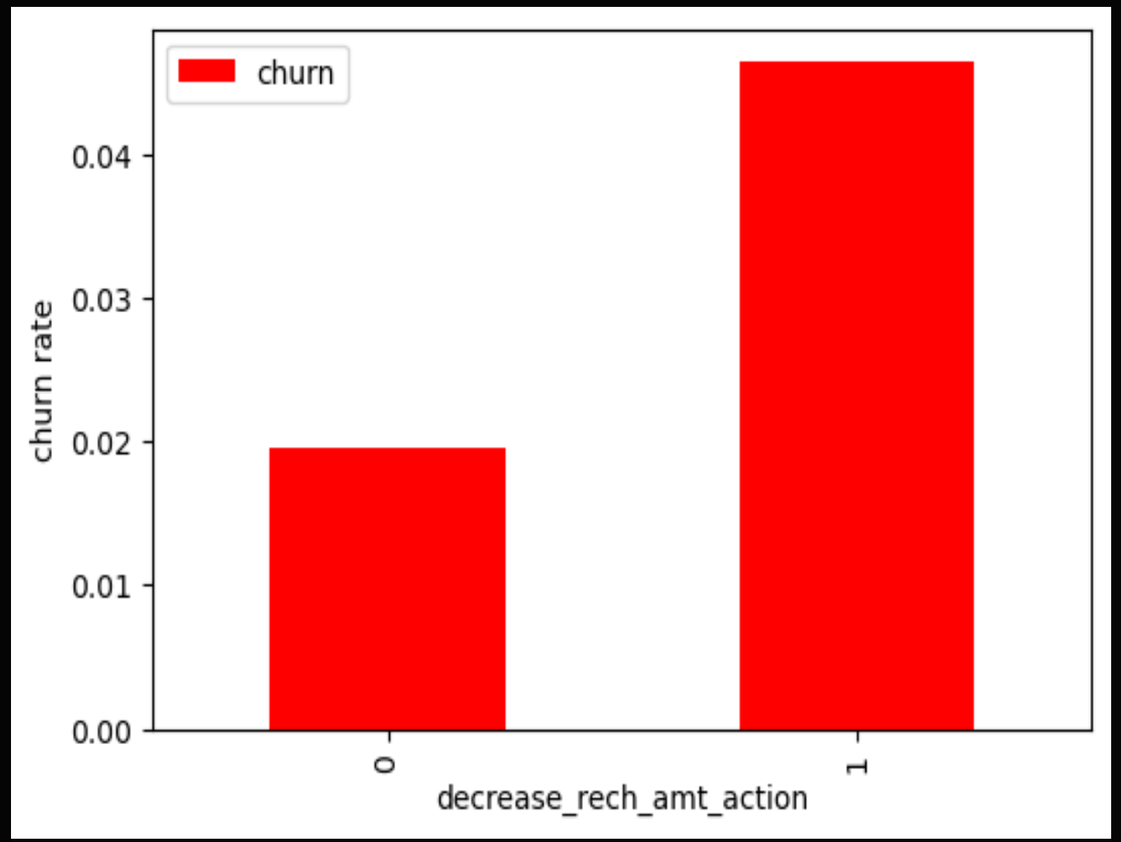


Data Handling Mechanism

1. Dropping columns with null values >30%
2. Filtering High Value Customers by summing recharge values for June & July and then taking avg
3. Dropping the rows having null %age in 6^{th} ,7^{th} & 8^{th} Month
4. Tagging churned customers as 1/0, 1 being churn 0 being no churn
5. Handling outliers by removing customers below 10^{th} and above 90^{th} percentile
6. Deriving New Feature columns : decrease_mou_action, decrease_rech_num_action etc.

Exploratory Data Analysis: Univariate Analysis

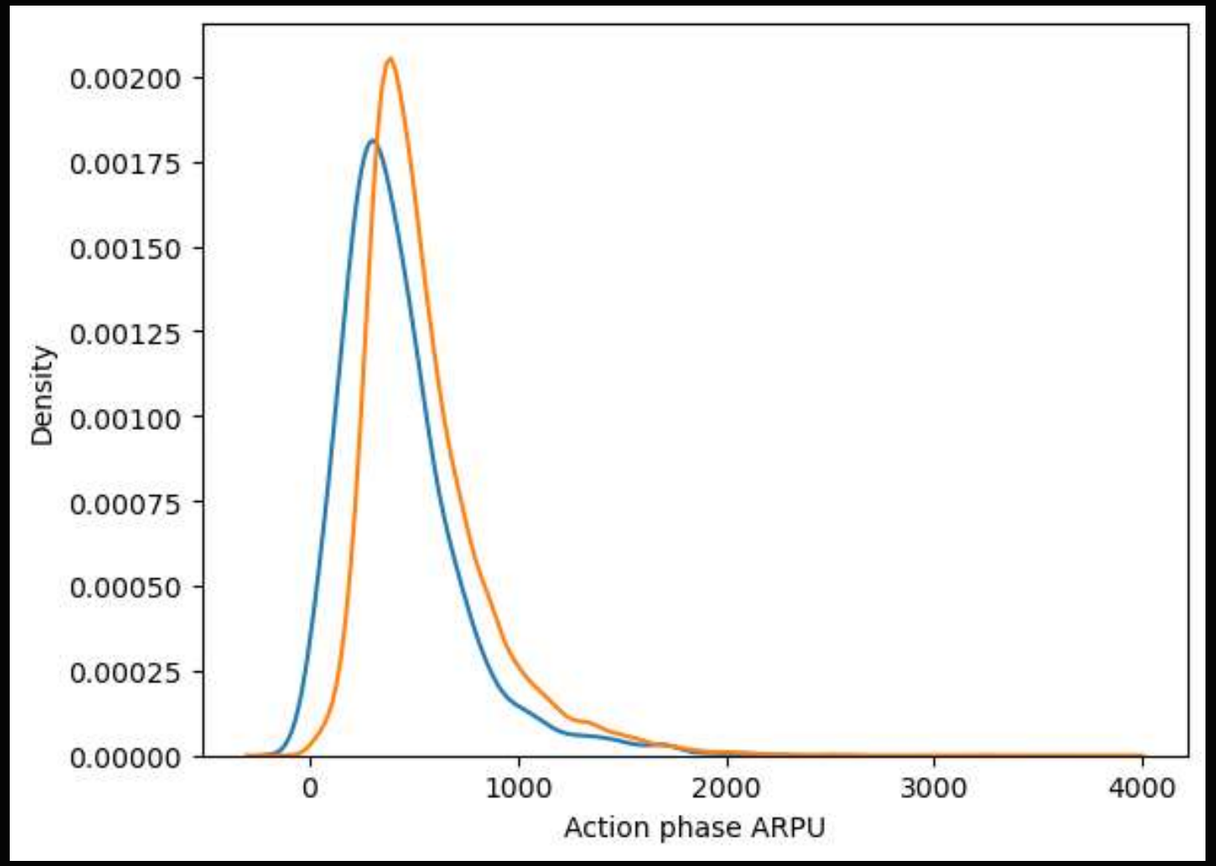


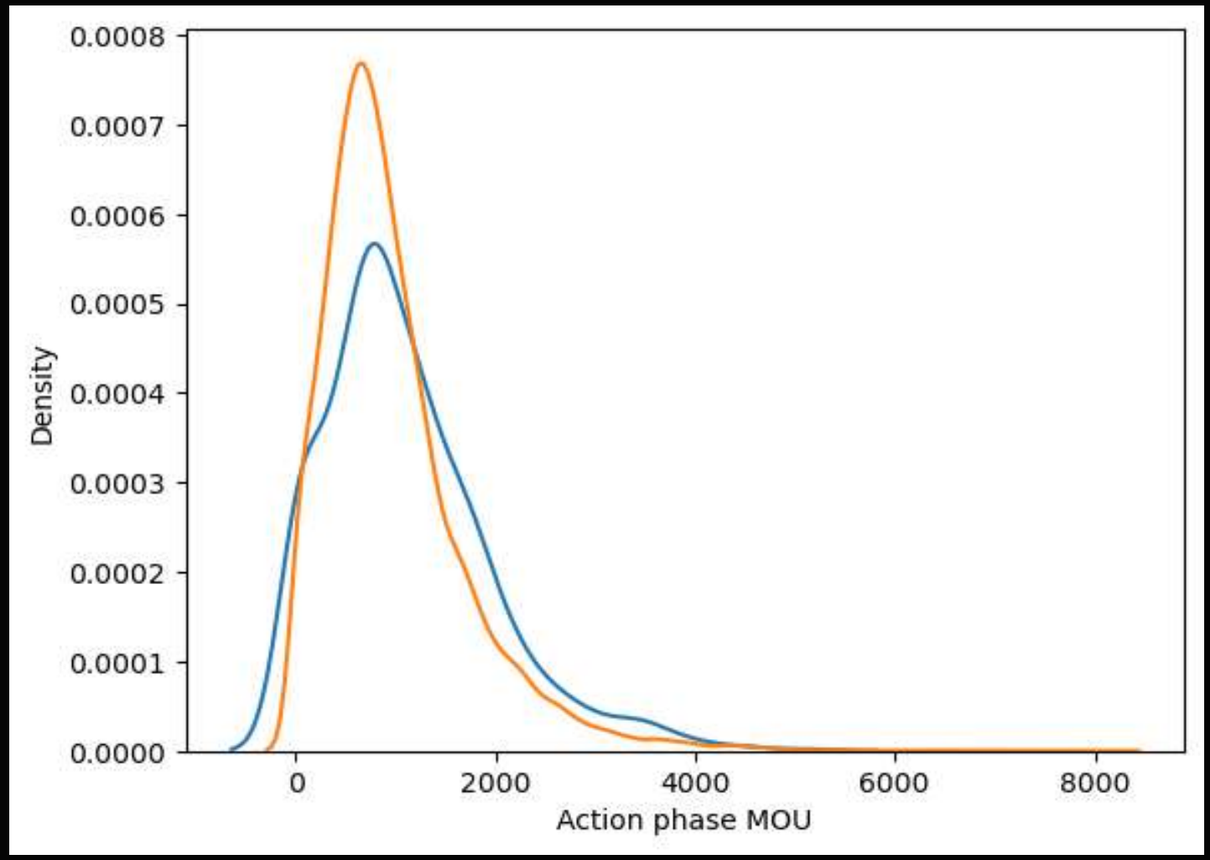


- Churn rate is more for the customers where mins of usage has decreased in the action phase

- Churn rate is more for the customers where volume based cost action month is increased i.e customers do no prefer to do monthly recharges when they are in action phase

Exploratory Data Analysis: Univariate Analysis

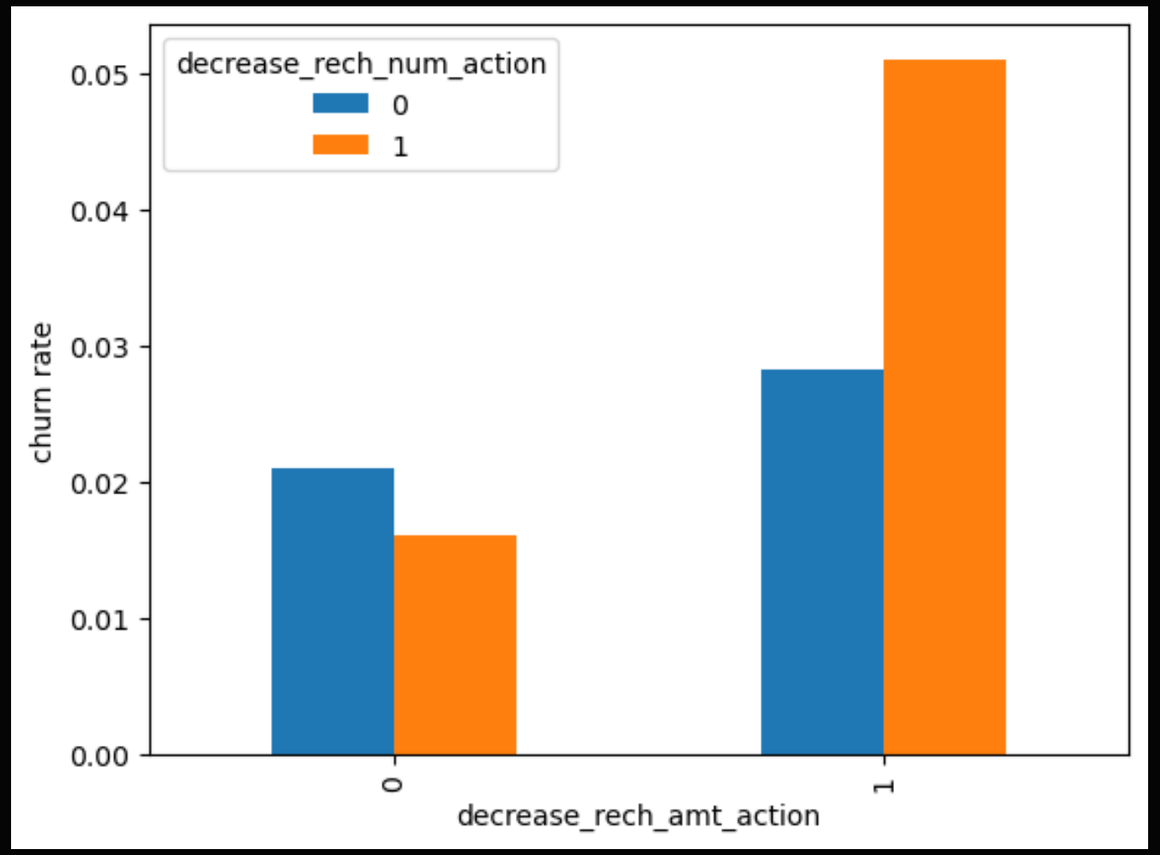


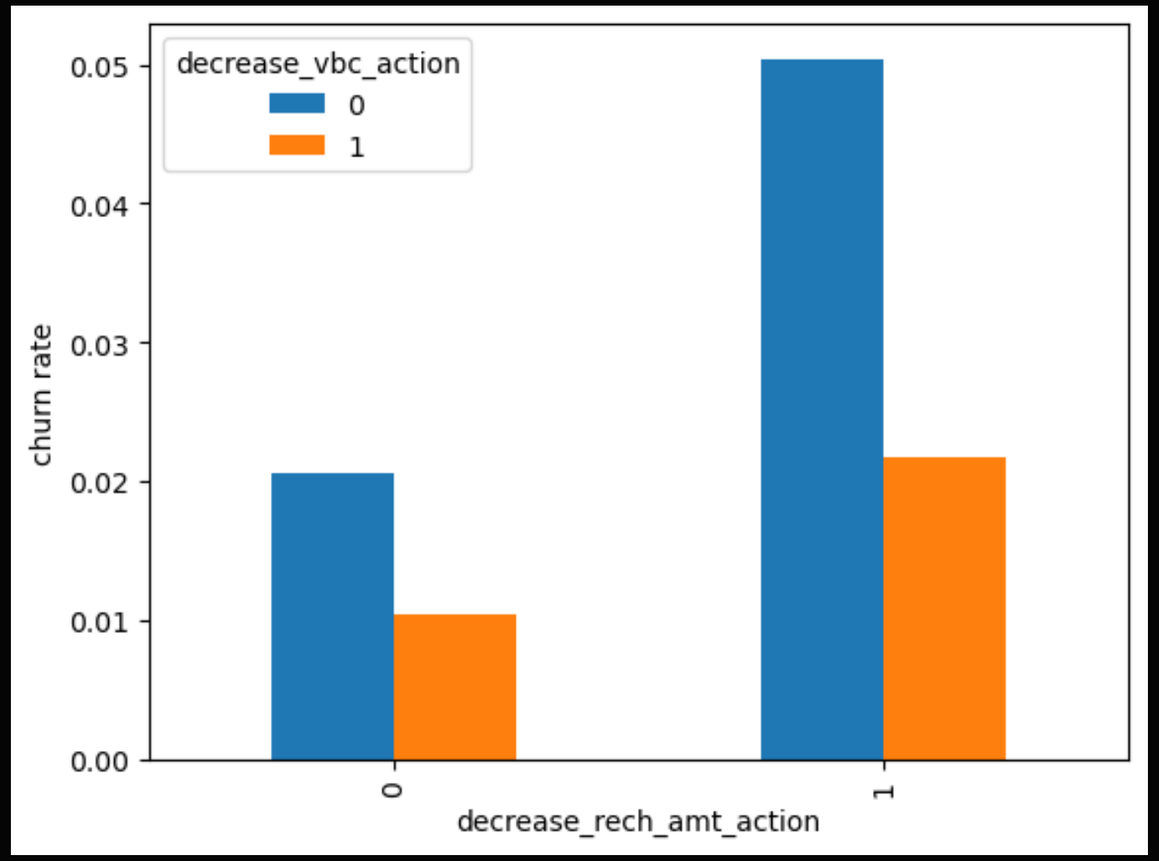


Average revenue per user (ARPU) for the churned customers is mostly densed on the 0 to 900. The higher ARPU customers are less likely to be churned.
ARPU for the not churned customers is mostly dense on the 0 to 1000.

- Minutes of usage(MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability.

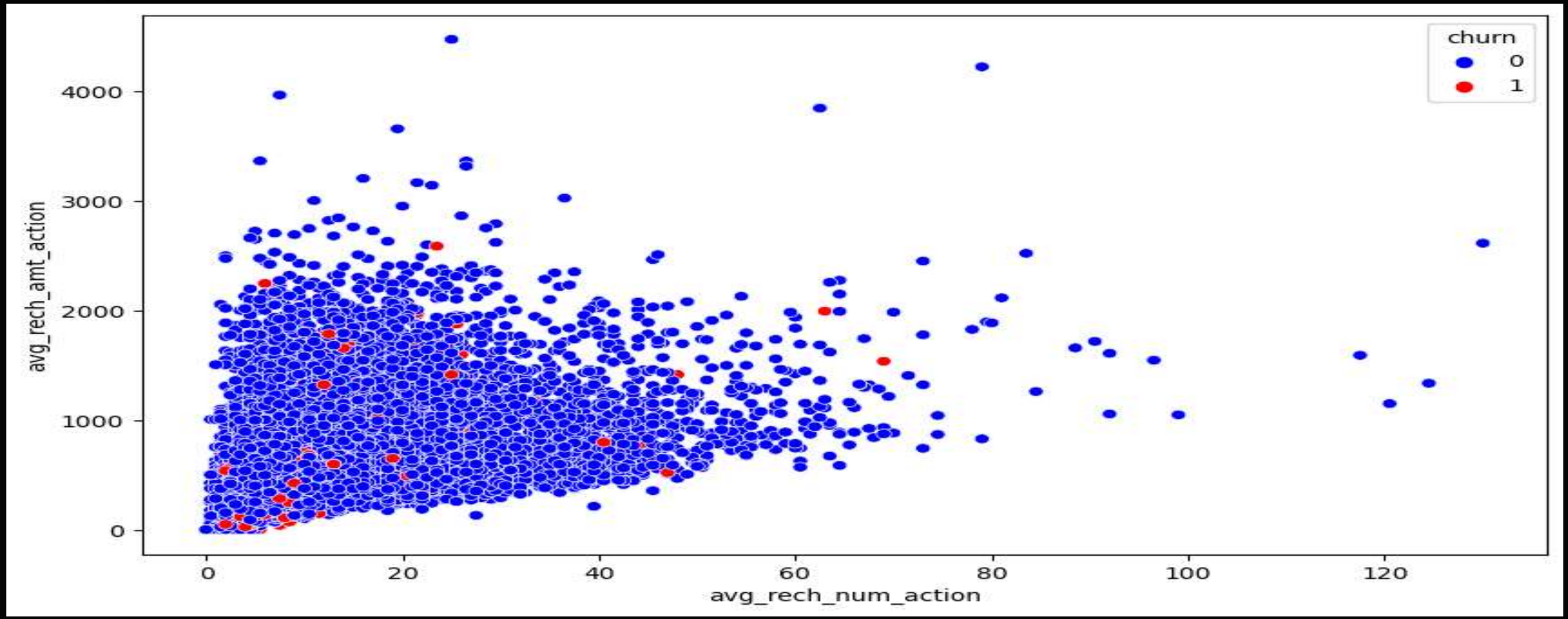
Exploratory Data Analysis: Bivariate Analysis





- From the above plot it is observed the churn rate is more where recharge amt as well as no. of recharges have decreased in action phase
- It is observed, the churn rate for customers is more where recharge amt is decreased along with vol based cost increased in action mth

Exploratory Data Analysis: Bivariate Analysis

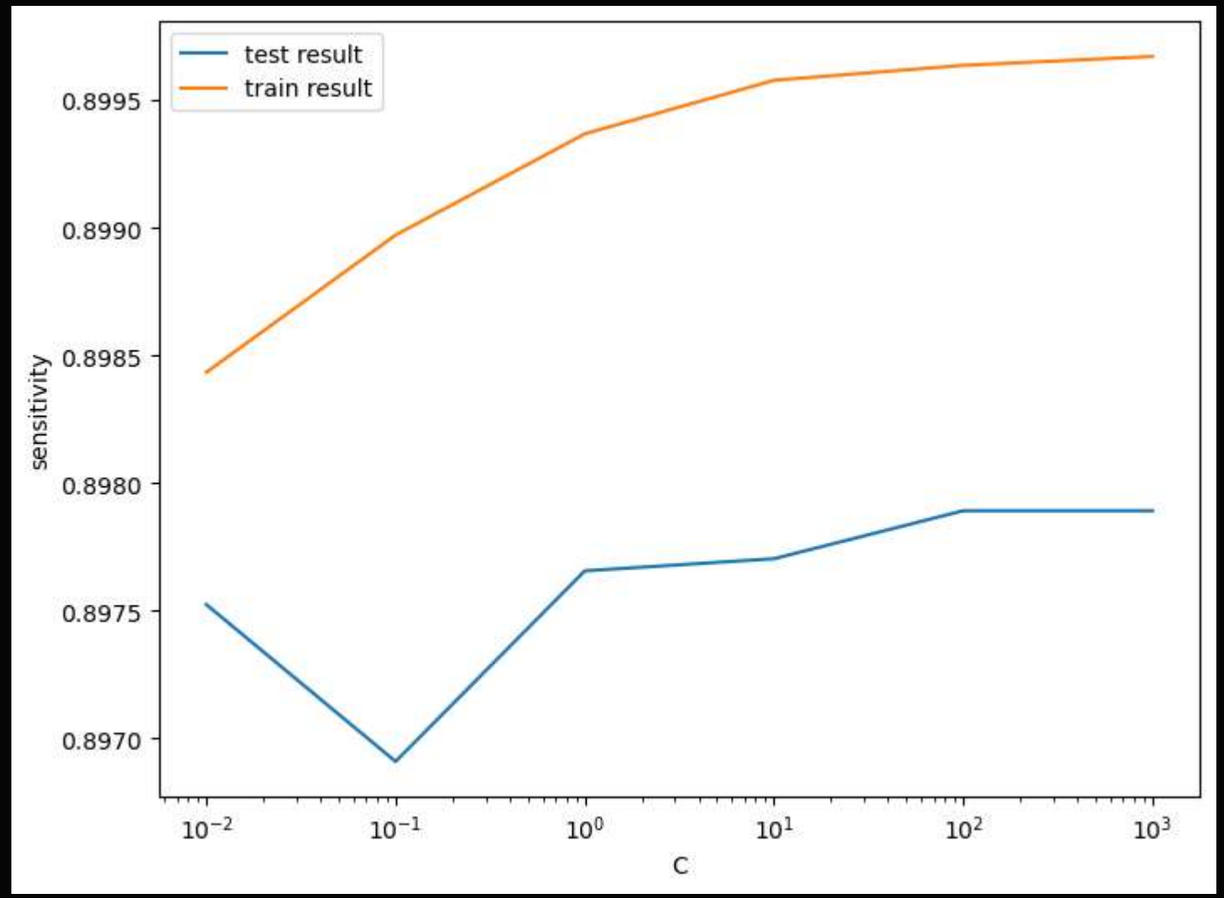


❖ From the scatterplot, it is observed that rchrg no & recharge amt is mostly proportional. More the no. of recharge more the amt of recharge

Model Building

Models Built

- Model with PCA
- Logistic Regression with PCA
- Decision Tree with PCA
- Random Forest with PCA
- Logistic Regression without PCA with fine & coarse tuning through RFE



Train set:

1.Accuracy: 0.87
2.Sensitivity: 0.89
3.Sepcificity: 0.83

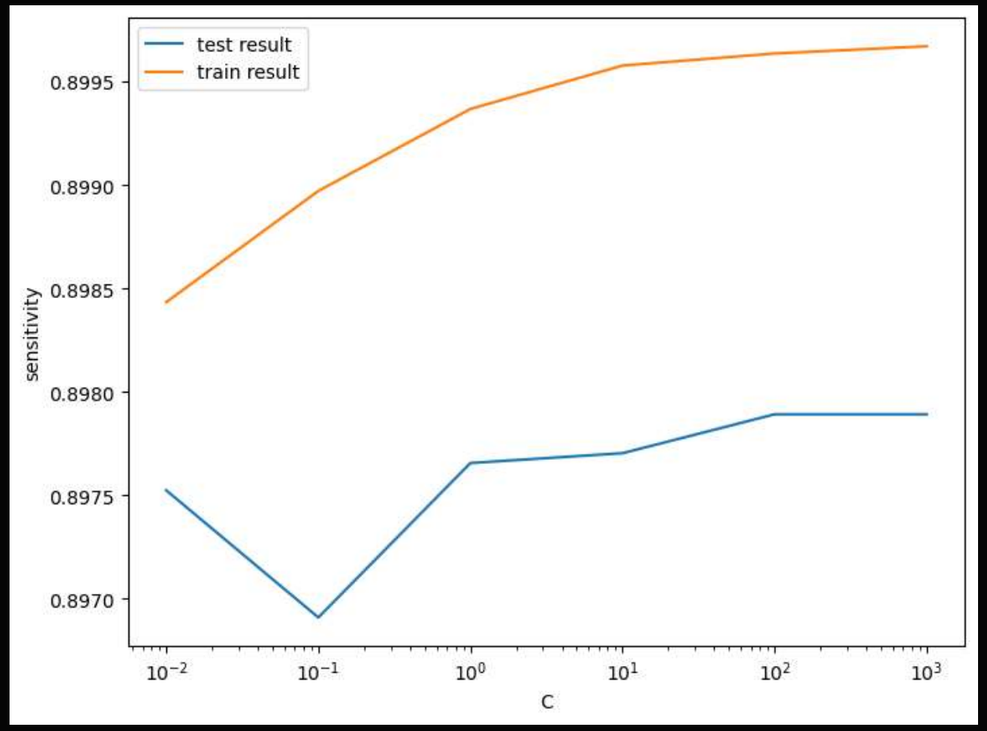
Test Set:

1.Accuracy: 0.83
2.Sensitivity: 0.81
3.Sepcificity: 0.83

Model Evaluation & Selection

Strategy Applied

- Calculated accuracy, sensitivity & specificity for the all the models
- After careful examination model, it is concluded that simple Logistic Regression model with PCA is viable



Confusion Matrix -Train Set

	Predicted Positive	Predicted Negative
Actual Positive	17,908	3,517
Actual Negative	2,154	19,271

Model Performance

Accuracy	86.76%
Sensitivity	89.94%
Specificity	83.58%

Model Prediction

	coef
const	-53.0128
offnet_mou_7	0.6096
offnet_mou_8	-3.2532
roam_og_mou_8	1.2482
std_og_t2m_mou_8	2.4408
isd_og_mou_8	-1.0212
og_others_7	-1.1915
og_others_8	-3780.7239
loc_ic_t2f_mou_8	-0.7547
loc_ic_mou_8	-1.9744
std_ic_t2f_mou_8	-0.7922
ic_others_8	-1.4913
total_rech_num_8	-0.4840
monthly_2g_8	-0.9031
monthly_3g_8	-0.9871
decrease_vbc_action	-1.3078

Top Features

Confusion Matrix -Test Set

	Predicted Positive	Predicted Negative
Actual Positive	4,452	896
Actual Negative	36	157

Model Performance

Accuracy	83.17%
Sensitivity	81.34%
Specificity	83.24%

Conclusions

Prediction

- Out of all the models, logistic Regression model with PCA is viable since it exhibits good accuracy, sensitivity and specificity as well as is a light model.

Model Performance

- Accuracy For Train Data- 86.76%
- For Test Data- 83.17%

Top 3 Variables

- Offnet_mou_7
- Offnet_mou_8
- roam_og_mou_8

Business Recommendations

1.Focus on customers with lower usage of incoming local calls and outgoing ISD calls during the action phase, primarily in August.

2.Target customers with reduced outgoing charges for others in July and incoming charges for others in August.

3.Customers experiencing an increase in value-based costs during the action phase are more prone to churn, making them potential targets for offers.

4.Customers with higher monthly 3G recharge amounts in August are at higher risk of churning.

5.Customers exhibiting a decrease in STD incoming minutes from operator T to fixed lines of T in August are more inclined to churn.

6.Customers with declining monthly 2G usage in August are highly likely to churn.

7.Customers showing a decrease in incoming minutes from operator T to fixed lines of T in August are more susceptible to churn.

8.The positive coefficient of the roam_og_mou_8,std_og_t2m_mou_8 variable suggests that customers with increasing roaming outgoing minutes are more prone to churn.