TELECOM CHURN CASE STUDY

PREPARED BY ANISH LAKHOTIYA, TEJINDER SINGH AND ABISHEK JONATHAN

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.
- To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

GOAL OF THE CASE STUDY:

- Retaining high profitable customers is the number one business goal
- To analyze customer-level data of the telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn.
- Business Problems to Address:
  - To predict high-value customer will churn or not.
  - To identify important variables that are strong predictors of churn
  - Build model with the main objective of identifying important predictor attributes which help the business understand indicators of churn.
  - Recommend strategies to manage customer churn

STEP 1: IMPORTING LIBRARIES AND DATA

- The following Python libraries were imported:

  - Data analysis– 1) Numpy 2) Pandas
  - Data visualization– 1) Matplotlib 2) Seaborn
  - Machine Learning– 1) Statsmodels 2) Scikit Learn

- The datafile “Leads.csv” was uploaded to start the EDA process. Upon uploading the dataframe was stored as “df”. The dataframe was inspected using head() function. The data frame contains 226 columns.

STEP 2: INSPECTING THE DATAFRAME

- The shape of the Dataframe (99999 rows, 226 columns) was examined using the shape attribute.
- To understand the data type of each column and number of missing values in each column, info() function was used.
- The data type of 179 columns
  - 179 columns is “float64” type
  - 35 columns are “int64” type
  - 12 columns are of “object” type

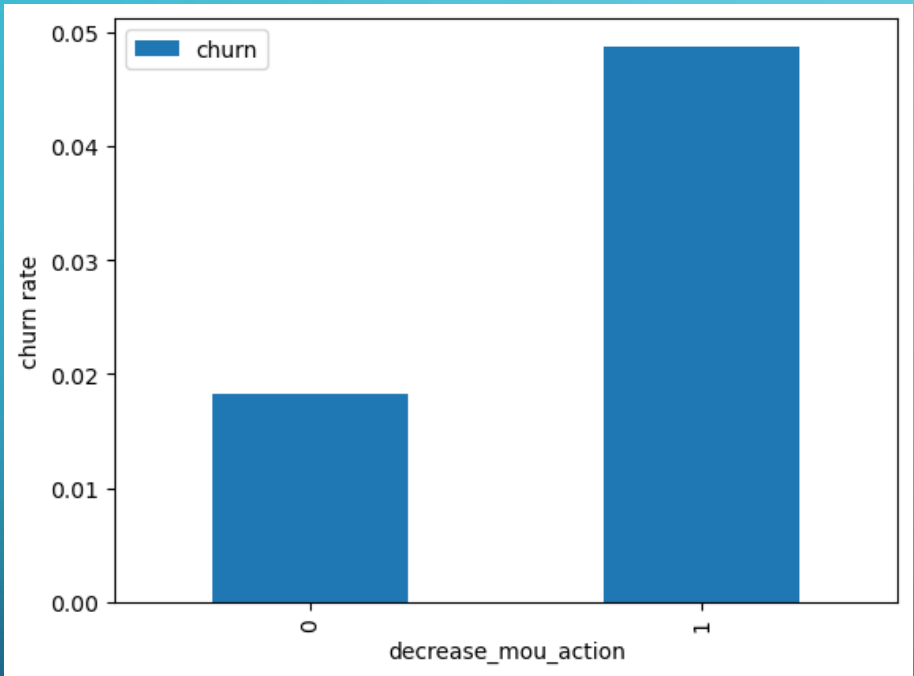
STEP 3: DATA PREPARATION

- Identifying the list of columns have more than 30% value missing and then dropping those columns.
- Next set of columns to be dropped are “date” as not required for analysis and “circle_id” has one unique value which will not have any impact. After dropping this columns left are 177.
- Find 70th percentile of the average recharge for 6^{th} & 7^{th} month.
- Filter the customers, who have recharged more than or equal to X(70^{th} percentile).
- We have 30078 rows and 178 columns
- Dropping rows having more than 50% value missing.
- Deleting the records for which MOU for Sep(9), Jun(6),Aug(8) & July(7) are null.
- We can see that we have lost almost 7% records. But we have enough number of records to do our analysis.

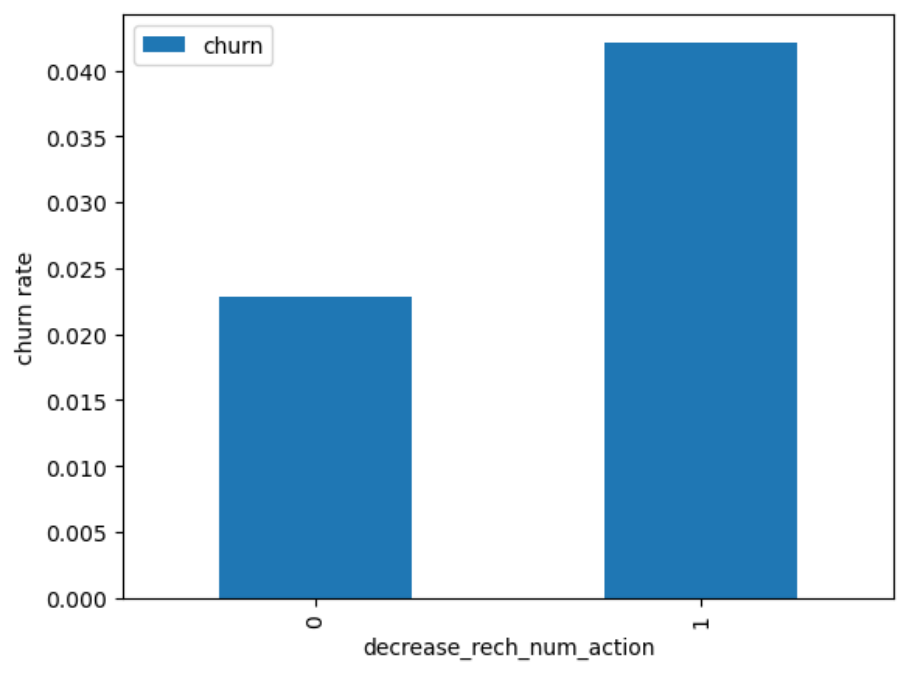
STEP 3: DATA PREPARATION (CONTINUED)

- Identify and tag customers (churn=1, else 0) for fourth month (Sep) who have not made any calls (either incoming or outgoing) AND have not used mobile internet even once in the churn phase.
- Check percentage after deleting the attributes for fourth month(Sep)
- Derive new columns for below decreased in the action phase than the good phase
  - Usage
  - Recharge
  - Average Revenue
  - Volume based cost

STEP 4: EDA - UNIVARIATE ANALYSIS

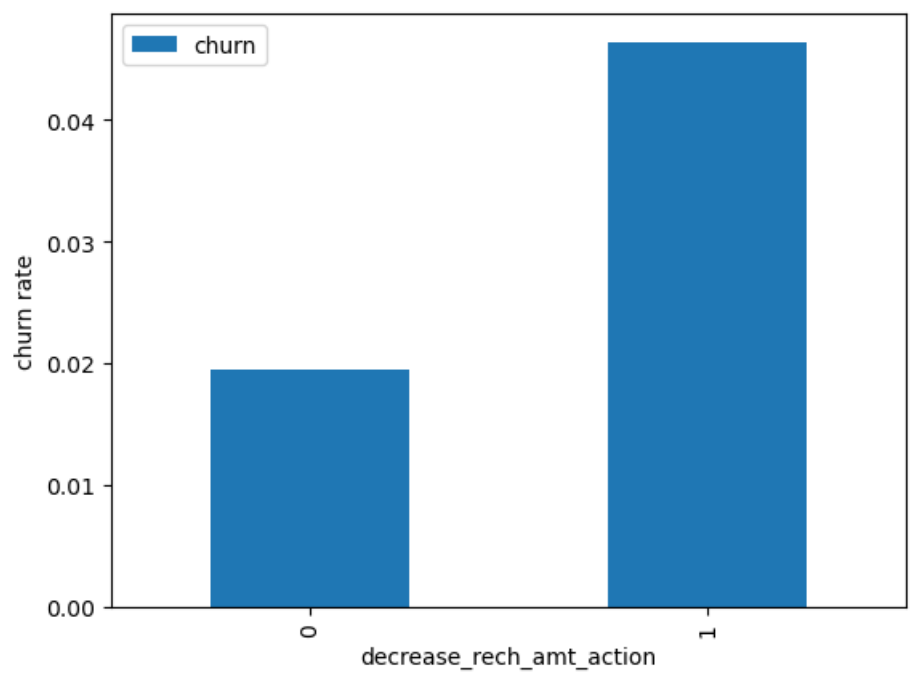


- Churn rate on the basis whether the customer decreased her/his MOU in action month

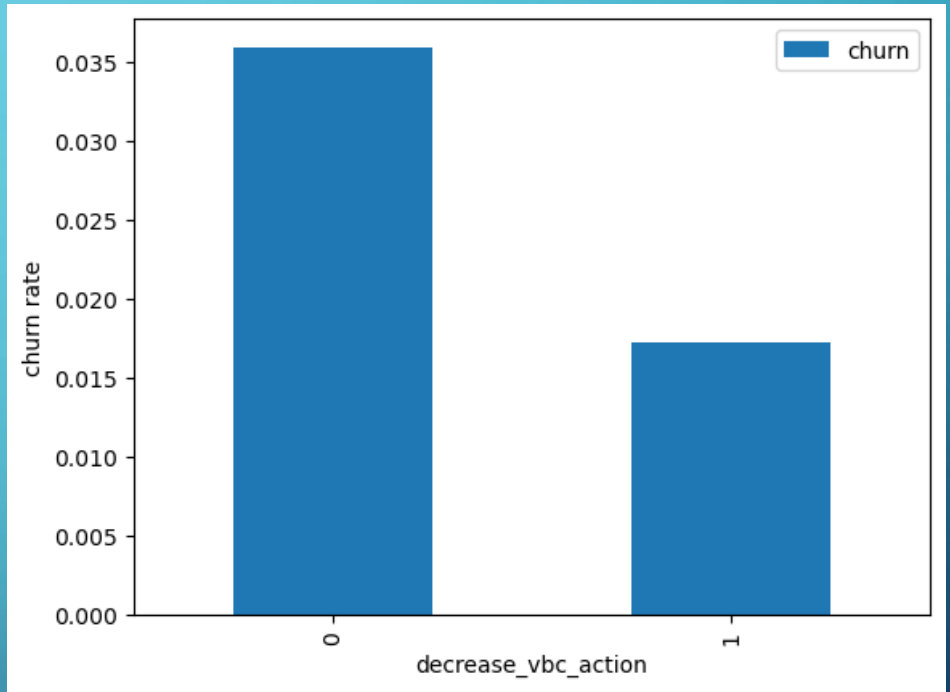


- Churn rate on the basis whether the customer decreased her/his number of recharge in action month

STEP 4: EDA - UNIVARIATE ANALYSIS

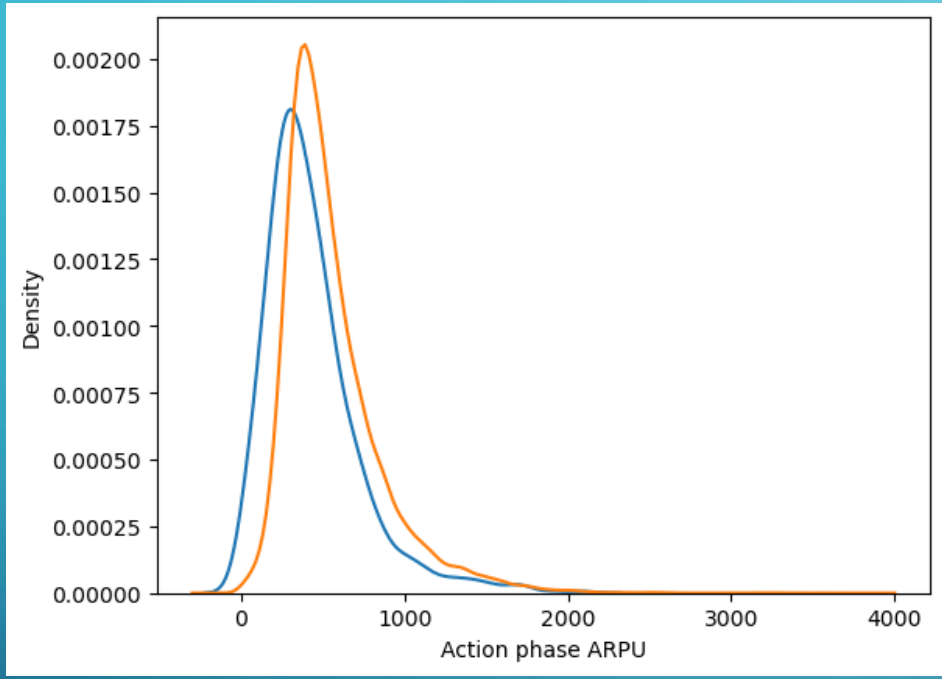


- Churn rate on the basis whether the customer decreased her/his amount of recharge in action month

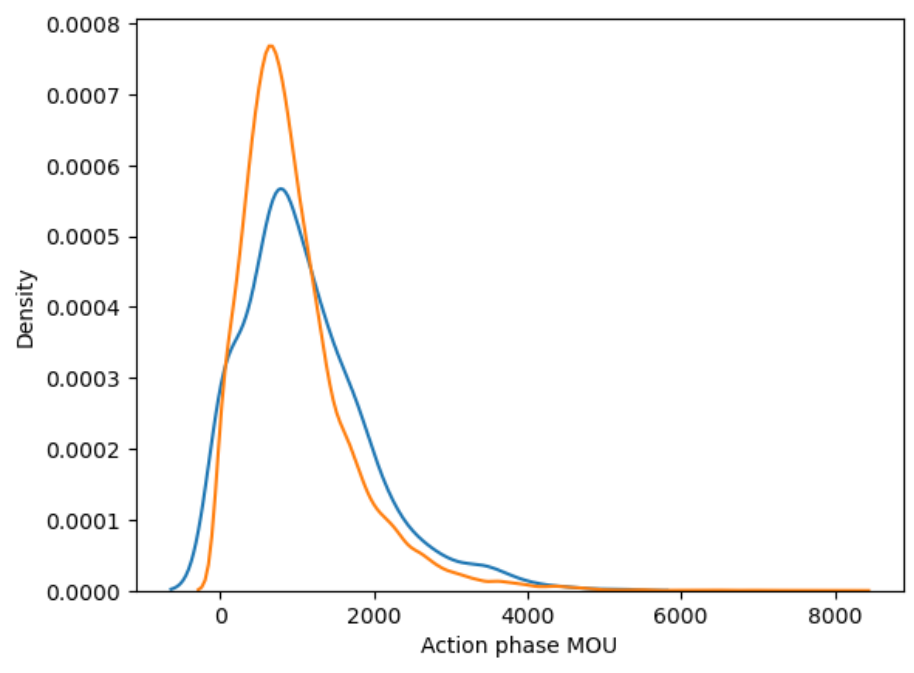


- Churn rate on the basis whether the customer decreased her/his volume based cost in action month

STEP 4: EDA - UNIVARIATE ANALYSIS

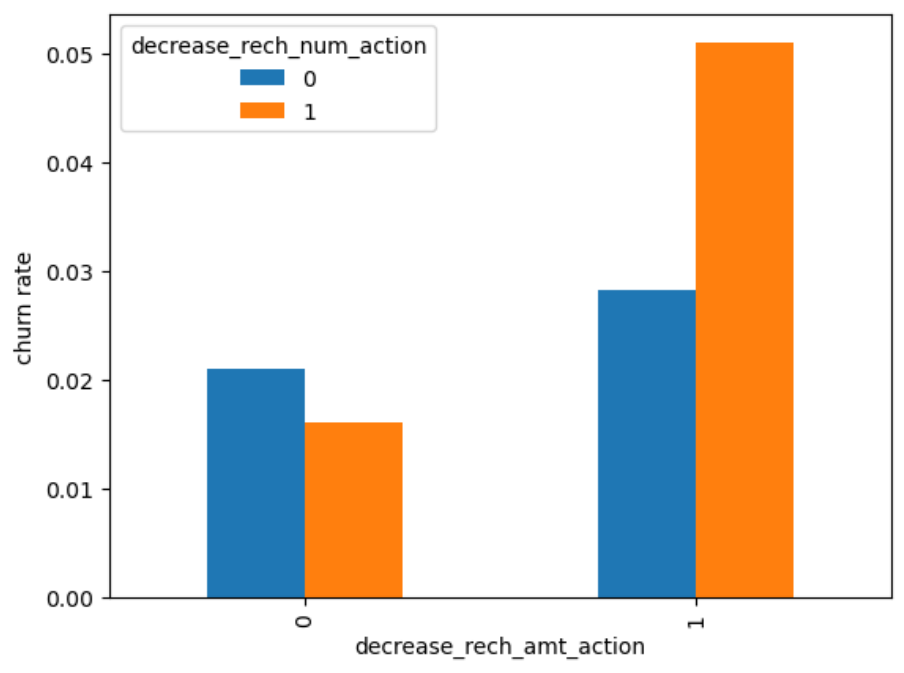


- Average revenue per user (ARPU) : Higher ARPU customers are less likely to be churned.

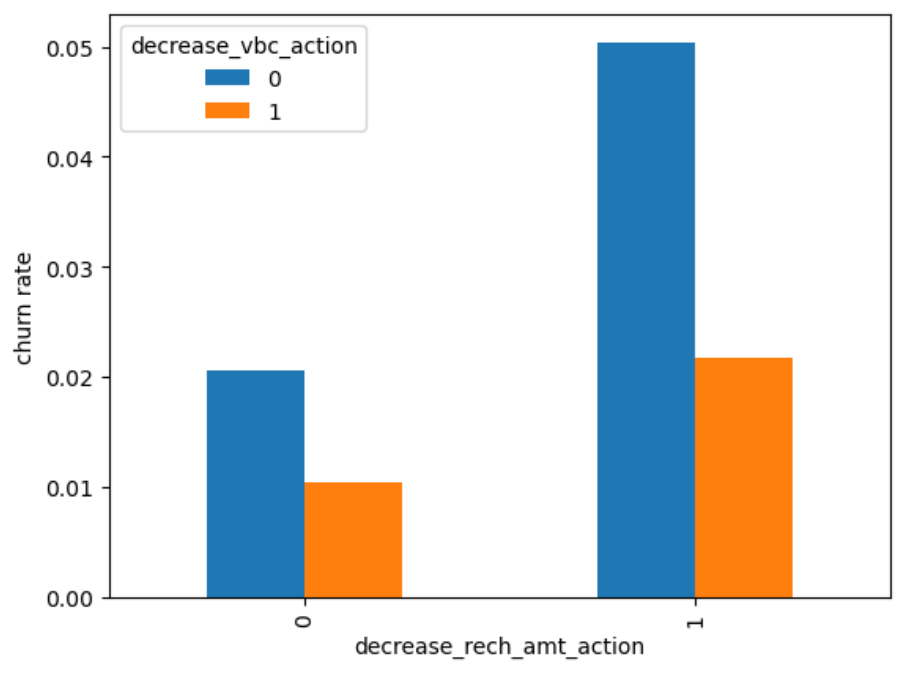


- Minutes of usage(MOU) : Higher the MOU, lesser the churn probability.

STEP 4: EDA -BIVARIATE ANALYSIS



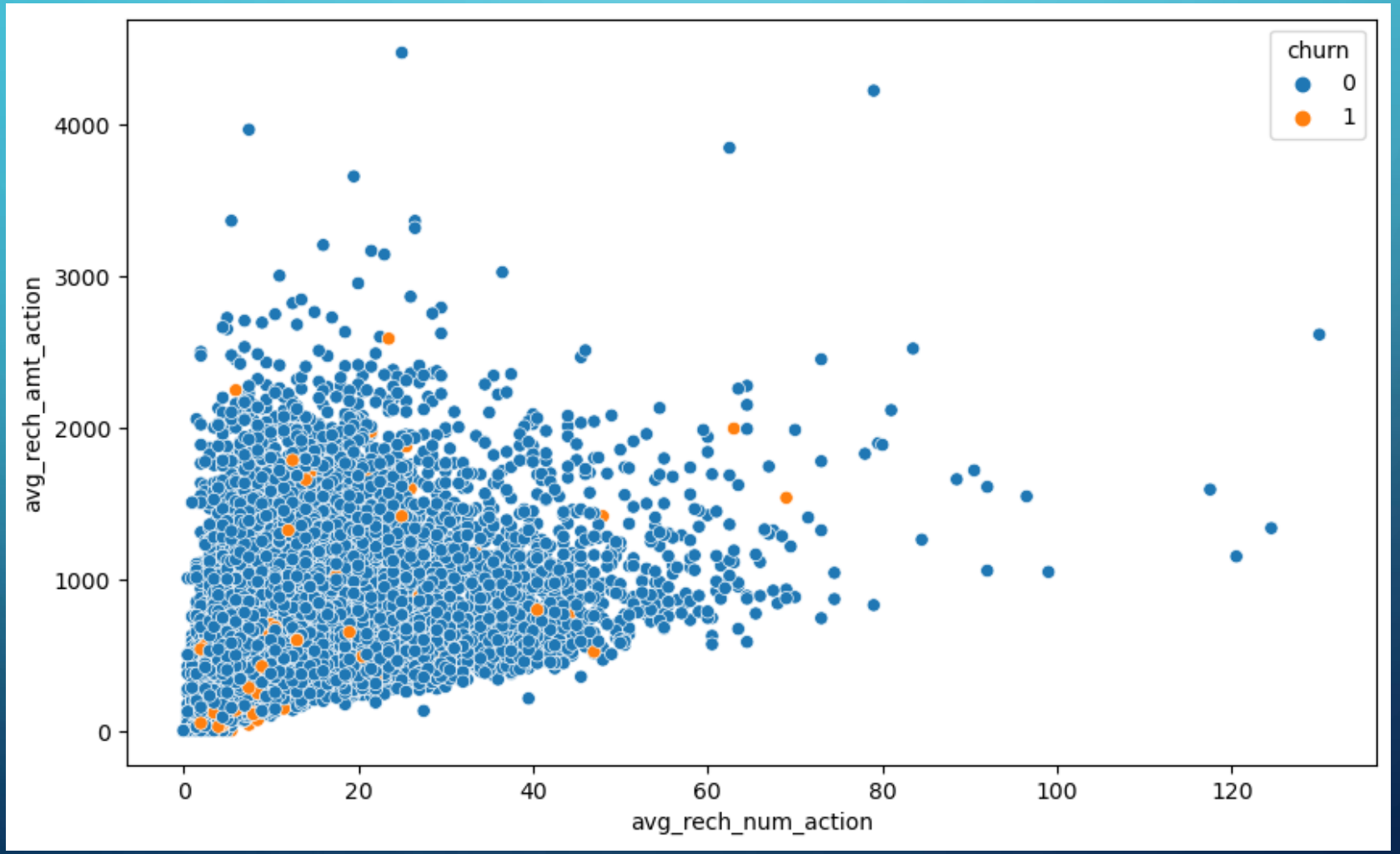
- Analysis of churn rate by the decreasing recharge amount and number of recharge in the action phase



- Analysis of churn rate by the decreasing recharge amount and volume based cost in the action phase

STEP 4: EDA -BIVARIATE ANALYSIS

- Analysis of recharge amount and number of recharge in action month



STEP 5: TEST-TRAIN SPLIT OF DATASET

- The dataset was split into train and test sets using train_test_split.
- 80% of the data was allocated for training, and 20% for testing the model performance.
- random_state was et to 100 for result consistency across runs.

STEP 6: MODEL SELECTION

Model with PCA

Logistic regression with PCA

Train set
- Accuracy = 0.86
- Sensitivity = 0.89
- Specificity = 0.83

Test set
- Accuracy = 0.83
- Sensitivity = 0.81
- Specificity = 0.83

SVM with PCA

Train set
- Accuracy = 0.89
- Sensitivity = 0.92
- Specificity = 0.85

Test set
- Accuracy = 0.85
- Sensitivity = 0.81
- Specificity = 0.85

Decision tree with PCA

Train set
- Accuracy = 0.90
- Sensitivity = 0.91
- Specificity = 0.88

Test set
- Accuracy = 0.86
- Sensitivity = 0.70
- Specificity = 0.87

Conclusion : After trying several models we can see that for achieving the best sensitivity, which was our ultimate goal, the classic Logistic regression or the SVM models preforms well. For both the models the sensitivity was approx. 81%. Also we have good accuracy of approx. 85%.

STEP 6: MODEL SELECTION

Model with no PCA

Logistic regression with no PCA

Train set
- Accuracy = 0.88
- Sensitivity = 0.91
- Specificity = 0.85

Test set
- Accuracy = 0.84
- Sensitivity = 0.76
- Specificity = 0.84

Final Conclusion between PCA and 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
- Going with more simplistic model such as logistic regression with PCA
- Logistic regression with PCA help us to identify the variables.
- Hence, the model is more relevant in terms of explaining to the business.

STEP 7: RECOMMENDATIONS

- Below are few top variables selected in the logistic regression model.

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

STEP 7: RECOMMENDATIONS

- Target the 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).
- Target the customers, whose outgoing others charge in July and incoming others on August are less.
- Also, the 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.
- Customers, whose monthly 3G recharge in August is more, are likely to be churned.
- 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.
- Customers decreasing monthly 2g usage for August are most probable to churn.
- Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.
- 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..