TELECOM CHURN PREDICTION CASE STUDY

Manish Kumar
Radhika Kute
Sarswati Salunkhe

PROBLEM STATEMENT

OVERVIEW

- Telecom market is highly competitive market
- Average 15-25% annual churn rate
- New customer acquisition cost is 5-10 times higher than cost for retaining existing customer
- To reduce churn rate it is important to predict which customers are at high risk of churn

CHURN

Revenue Based Churn

- Customers who have not utilized any revenue-generating facility such as mobile, outgoing calls, SMS, etc. over given period of time.
- Can be aggregated as customers who generate less than INR 4 revenue
- This definition cannot separate low budget customers from churning customers

CHURN

Usage Based churn

- Customer who have not done any usage, either incoming or outgoing – in terms of calls, internet, etc. over a period of time
- If customer has already stopped using service for certain period of time it would too late to take any corrective action for retaining them
- Predicting churn based on usage would be useless if customer has already switched to another operator

CHURN

High Value Churn

- Approx 80% revenue comes from top 20% customers
- These are called high-value customers
- Reduction in churning of high-value customers will reduce significant revenue leakage

AIM

- The aim of this study is to predict High-value churn
- The high-value customer do not churn suddenly
- There is decrease in revenue generated by them before switching to another operator
- We will divide the given 4 month period into 2 months good, 1 month action and 1 month churning phase and analyse accordingly

DATA PREPROCESSING

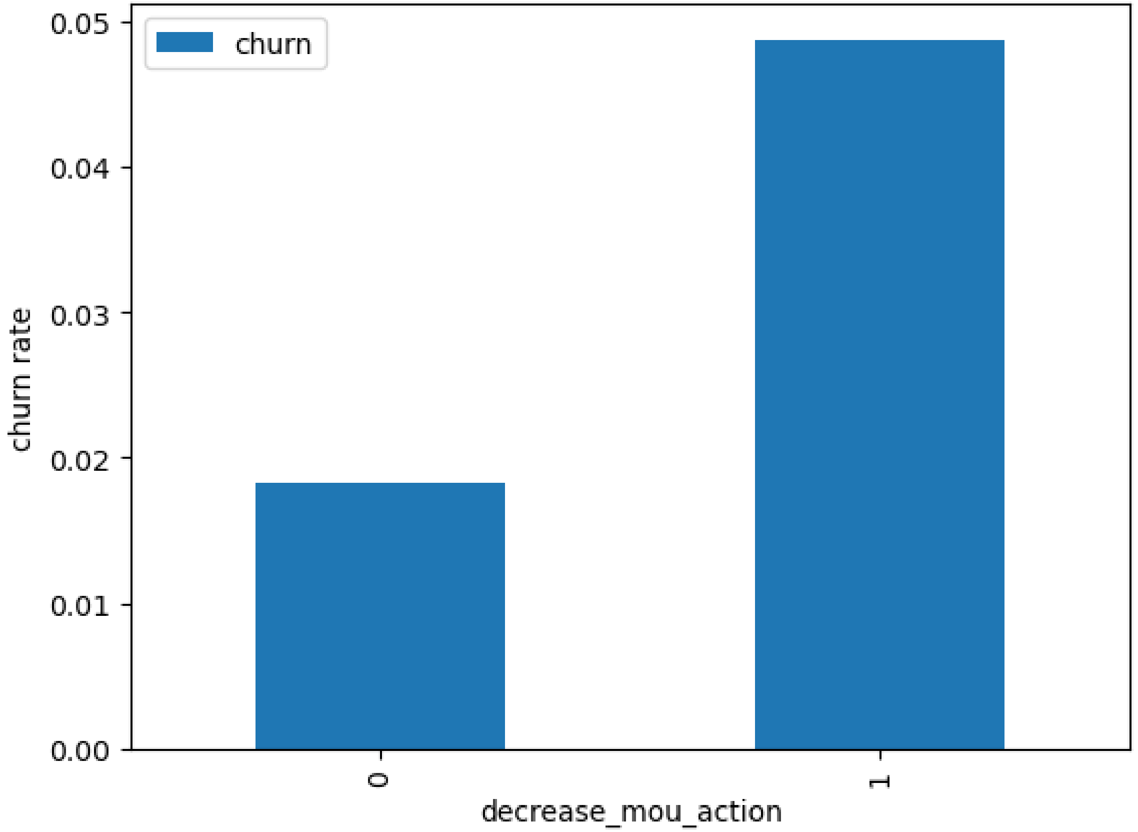
HANDLING MISSING VALUES & OUTLIERS

- Columns having more than 35 % missing values are dropped
- To find high-value customer we have shortlisted the customers having more than 70 percentile of average recharge value in good phase
- Rows having more than 50 % missing values are counted and dropped
- Rows having missing values in monthly use columns are dropped
- Checked if we get enough churn data.
- Churn data is found to be 3.39 % which is acceptable so proceeded further
- Removed outliers below 10 percentiles and above 90 percentiles

EDA

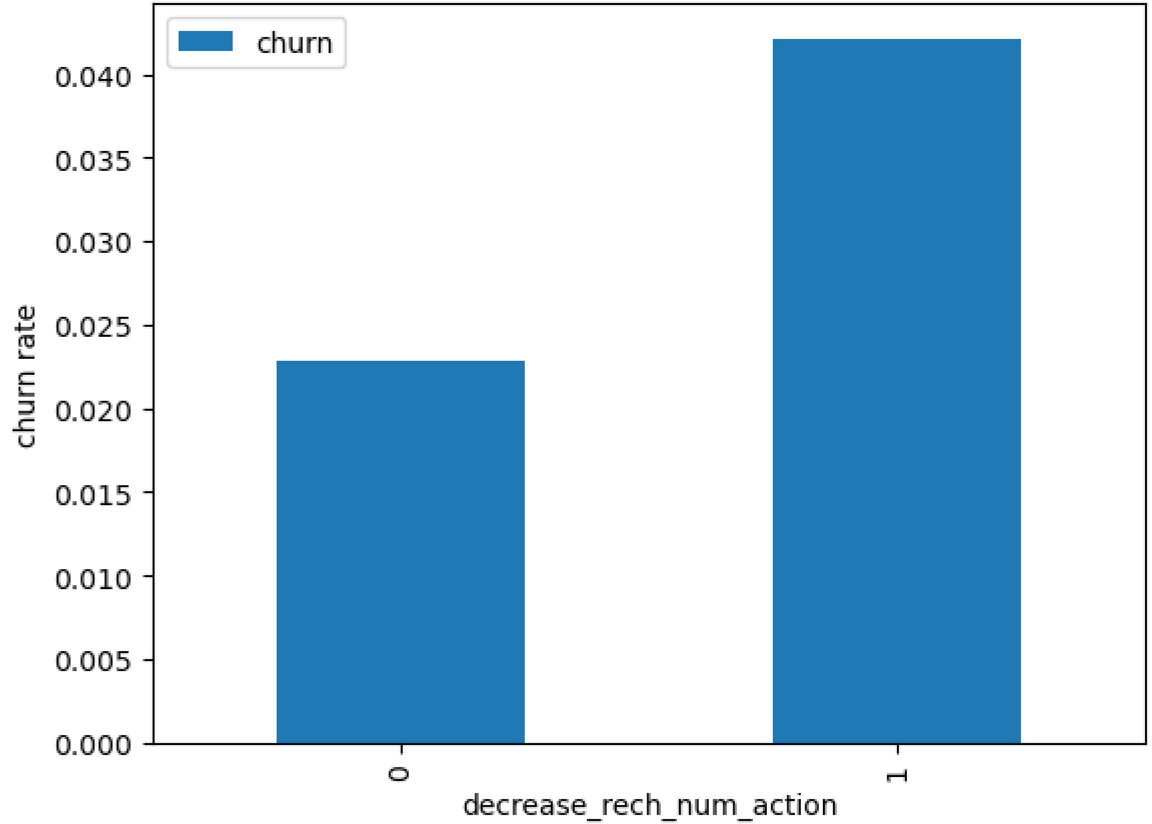
UNIVARIATE ANALYSIS

Churn rate is more for the customers whose MOU decreased in the action phase than good phase



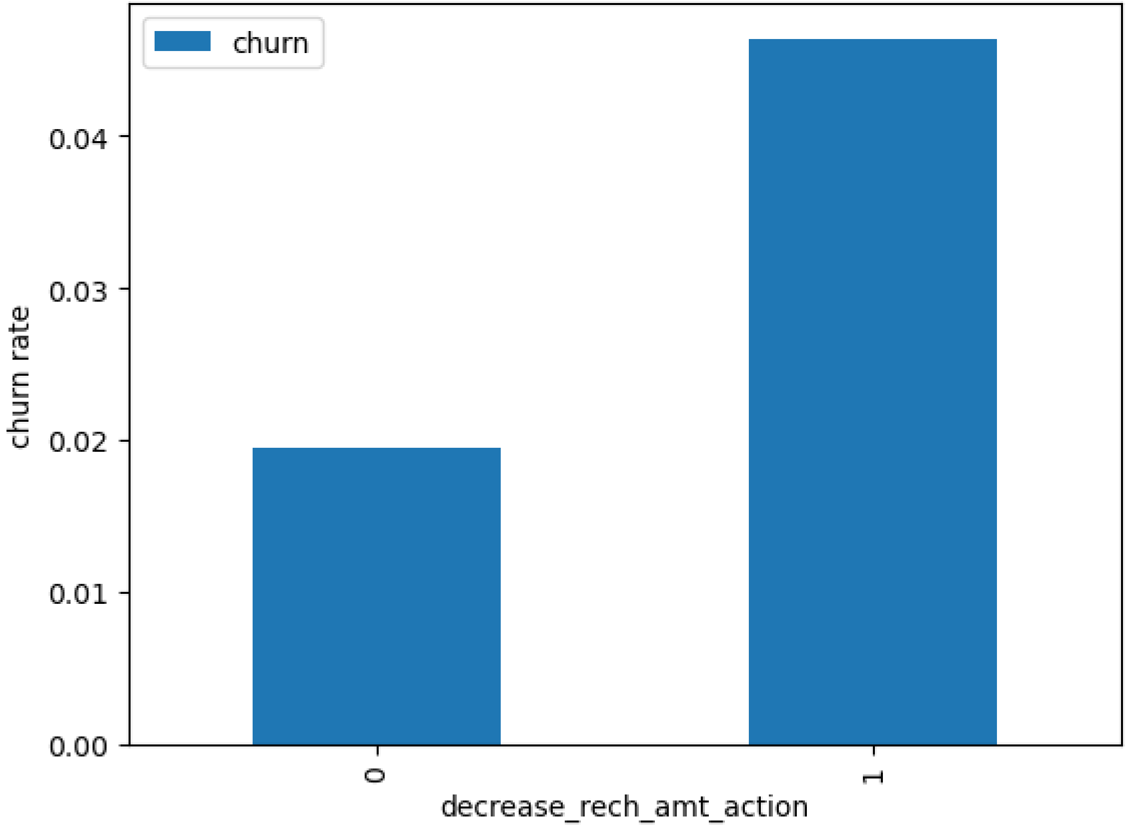
UNIVARIATE ANALYSIS

Churn rate is more for the customers whose number of recharge in the action phase is lesser than the number in good phase



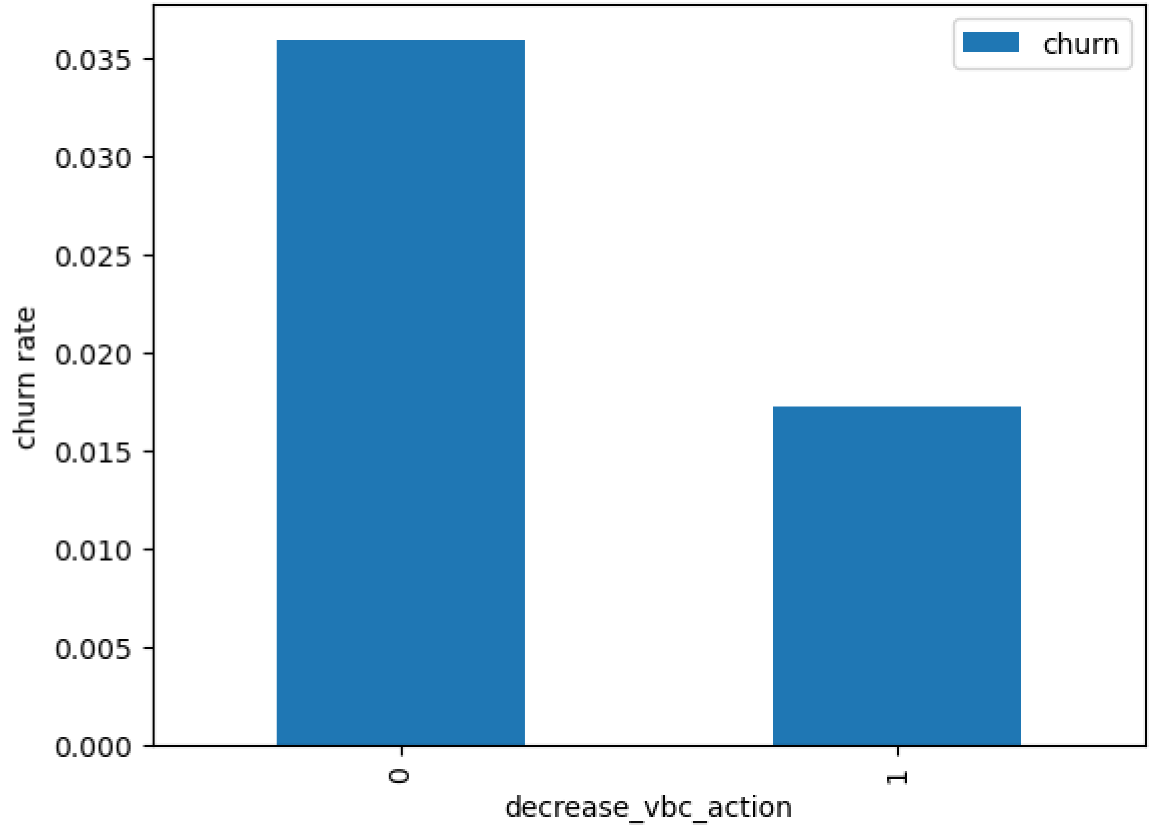
UNIVARIATE ANALYSIS

Churn rate is more for the customers whose amount of recharge in the action phase is lesser than the number in good phase



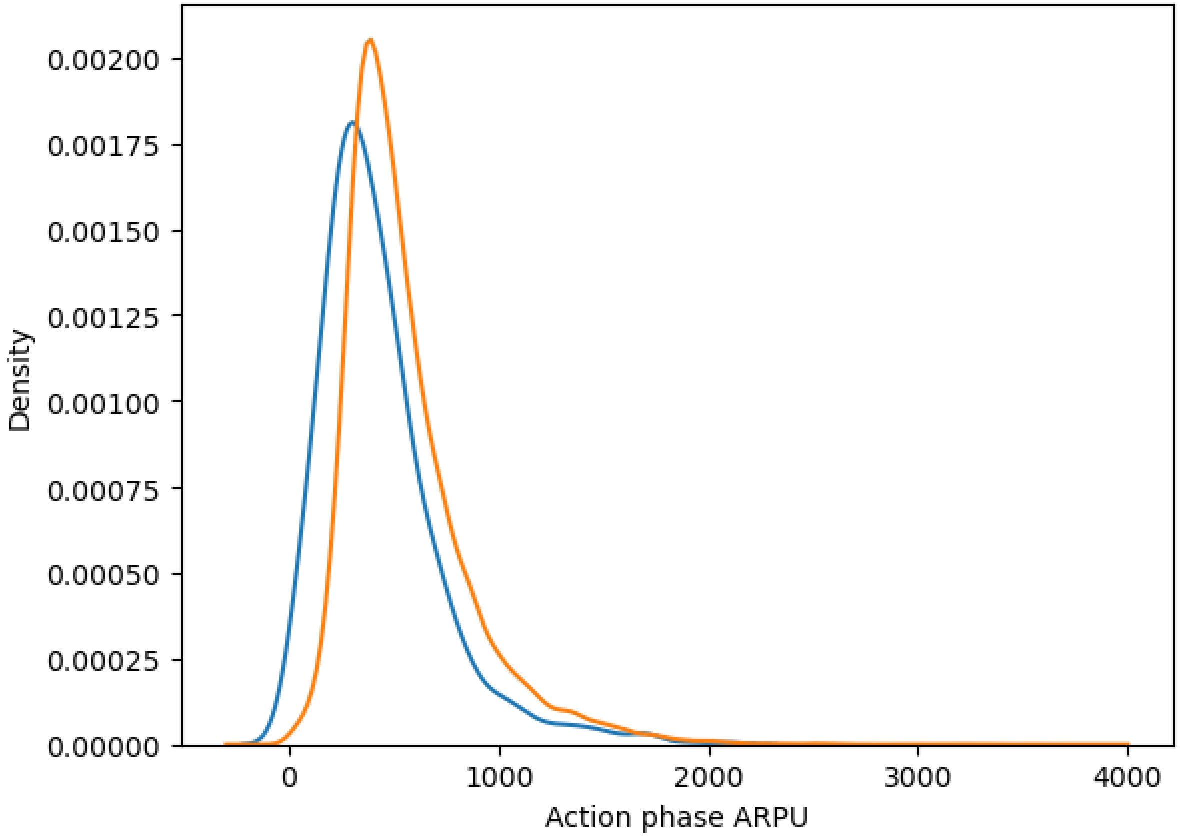
UNIVARIATE ANALYSIS

Churn rate is more for the customers whose volume based cost in action month is increased when they are in action phase



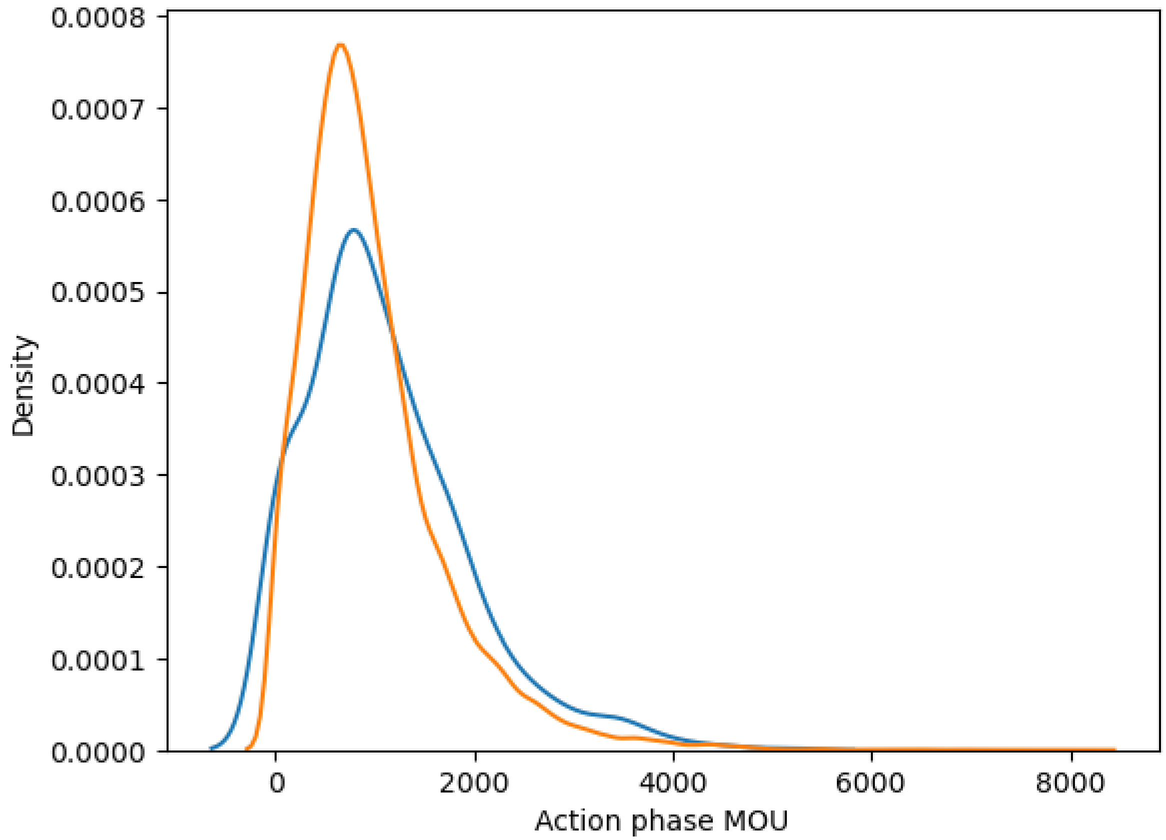
UNIVARIATE ANALYSIS

Average revenue per user(ARPU) for the churned customer is mostly dense on the 0 to 900, whereas for non churned customer it is mostly dense on the 0 to 1000



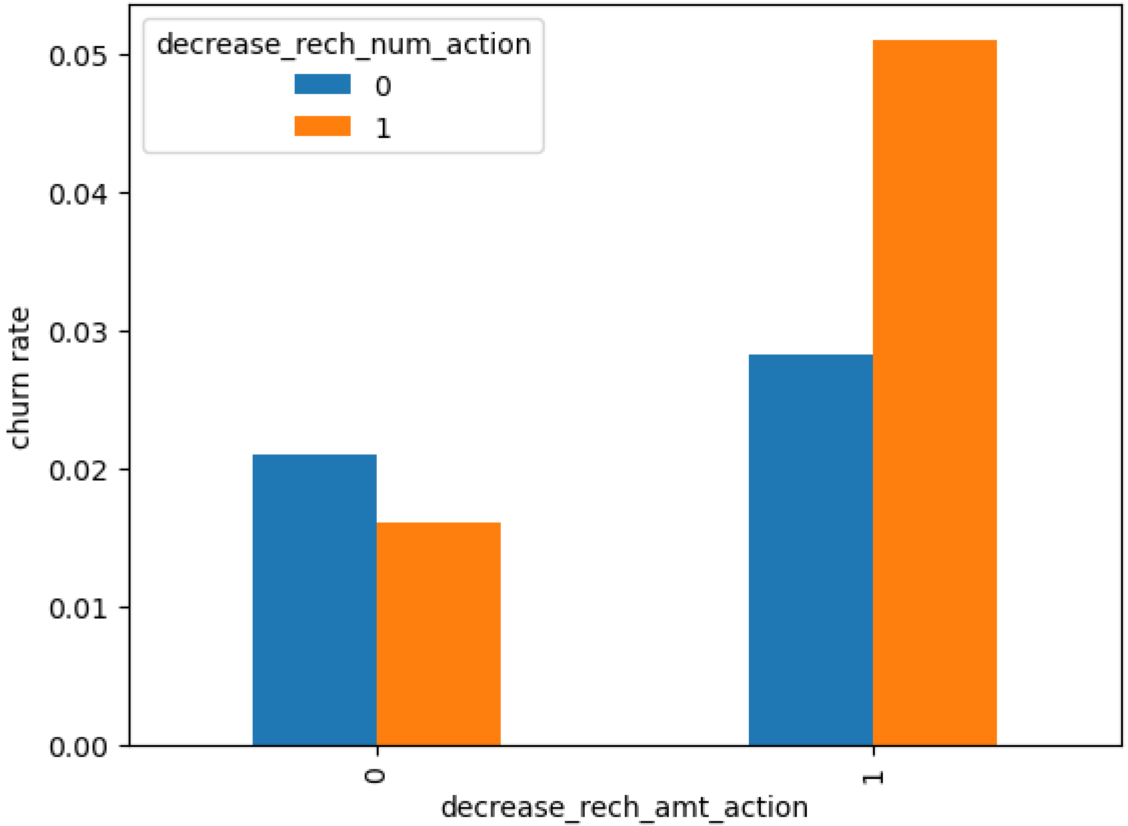
UNIVARIATE ANALYSIS

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



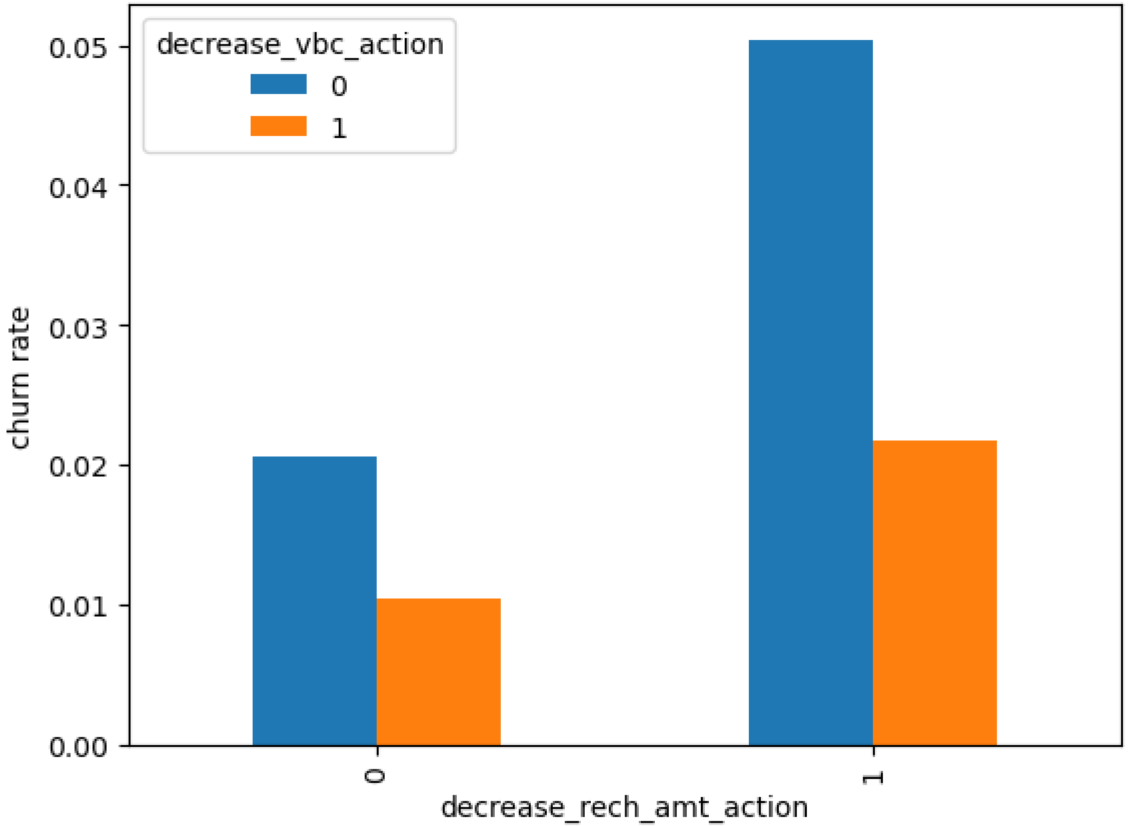
BIVARIATE ANALYSIS

Churn rate is more for the customers whose recharge amount as well as number of recharge have decreased in the action phase than in good phase



BIVARIATE ANALYSIS

Churn rate is more for the customers whose recharge amount is decreased along with the volume based cost is increased in the action month



TRAIN TEST SPLIT, DATA IMBALANCE & FEATURE SCALING

- Data is split into training set and test set in 80:20 ratio
- To remove imbalance in data SMOTE(Synthetic Minority Oversampling Technique) is used
- Data is scaled with the standard scaler

MODEL SELECTION

MODELLING WITH PCA

Logistic Regression (Model Summary):

	Train Set	Test Set
Accuracy	0.86	0.83
Sensitivity	0.89	0.83
Specificity	0.83	0.83

MODELLING WITH PCA

Decision Tree (Model Summary):

	Train Set	Test Set
Accuracy	0.9	0.86
Sensitivity	0.92	0.7
Specificity	0.88	0.87

MODELLING WITH PCA

Random Forest (Model Summary):

	Train Set	Test Set
Accuracy	0.84	0.8
Sensitivity	0.88	0.75
Specificity	0.8	0.8

CONCLUSION FOR MODELLING WITH PCA

After testing several models, we can observe that the Logistic regression model performs well for achieving the best sensitivity (around 81%), which was our final goal. We also got good level of accuracy (about 83%).

MODELLING WITHOUT PCA

Logistic Regression (Model Summary):

	Train Set	Test Set
Accuracy	0.84	0.89
Sensitivity	0.89	0.82
Specificity	0.79	0.78

CONCLUSION FOR MODELLING WITHOUT PCA

- Model seems to have good performance since the accuracy and sensitivity are relatively close between the training and test sets
- We can observe that logistic regression without PCA has equivalent sensitivity and accuracy to the models with PCA
- We chose model with PCA which explains the important predictor variables as well as the relevance of each variable
- The model also assists us in identifying the variables that should be considered when making the choice to churn clients
- Hence the model is more applicable in terms of explaining to the business

TOP PREDICTORS IN THE LOGISTIC REGRESSION MODEL

As we can see, the majority of the top variables have negative coefficients meaning these variables are inversely correlated to the churn probability.

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

RECOMMENDATIONS

RECOMMENDATIONS

- Target customers who use less minutes of incoming local and outgoing ISD calls during the action phase (primarily in August).
- Target customers with lower outgoing and incoming charges in July and August.
- Additionally, consumers with higher value-based costs in the action phase are more likely to churn than other customers. As a result, these customers may be an ideal target for providing an offer.
- Customers whose monthly 3G recharge in August is higher are more likely to be churned.

RECOMMENDATIONS

- Customers with declining STD incoming minutes of usage for operators T to fixed lines of T in August are more likely to churn.
- Customers who reduce their monthly 2g usage for August are more likely to churn.
- Customers with decreased incoming minutes of usage for operators T to fixed lines T in August are more likely to churn.
- The coefficients for the roam_og_mou_8 variables are positive (0.7135). Customers whose roaming outbound minutes consumption is increasing are more likely to churn.