Telecom Churn Cast Study

Problem Statement

- Identify customers who are at a high risk of churn along with the main indicators contributing to the churn. Its important for the revenue growth to retain highly profitable customers, so the main objective of the company is to be able to predict the customers who are likely to churn.

- The business objective is to predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months. To do this task well, understanding the typical customer behavior during churn will be helpful.

Overall approach

- Business Understanding
- Data Understanding
- Data Preparation
- Model Building
- Model Evaluation

Business Understanding

- 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.
- For many incumbent operators, retaining high profitable customers is the number one business goal.
- To reduce customer churn, telecom companies need to predict which customers are at high risk of churn

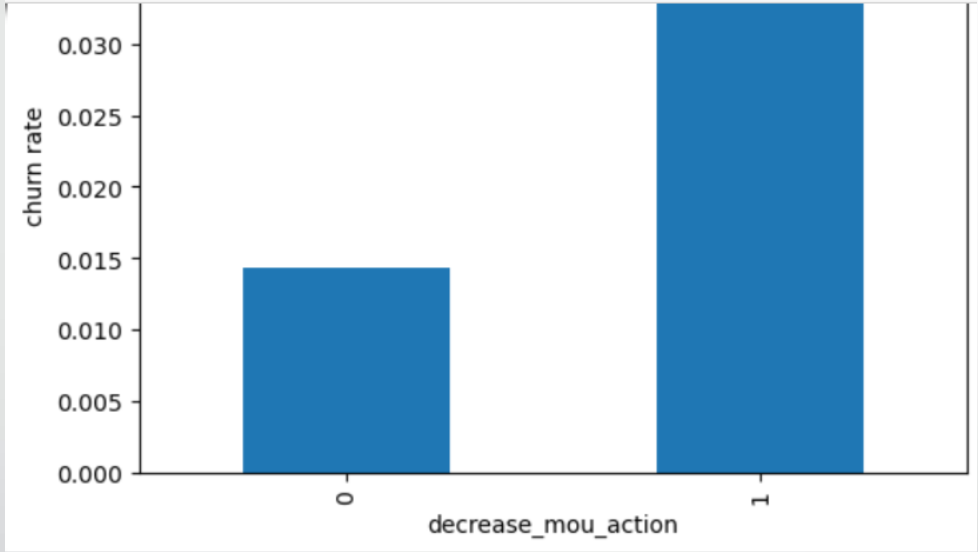
Data Understanding

- There are three phases of customer lifecycle :
- **The good phase**: In this phase, the customer is happy with the service and behaves as usual.
- **The action phase**: The customer experience starts to sore in this phase, for e.g. he/she gets a compelling offer from a competitor, faces unjust charges, becomes unhappy with service quality etc. In this phase, the customer usually shows different behaviour than the ‘good’ months. Also, it is crucial to identify high-churn-risk customers in this phase, since some corrective actions can be taken at this point (such as matching the competitor’s offer/improving the service quality etc.)
- **The churn phase**: In this phase, the customer is said to have churned.

Data preparation

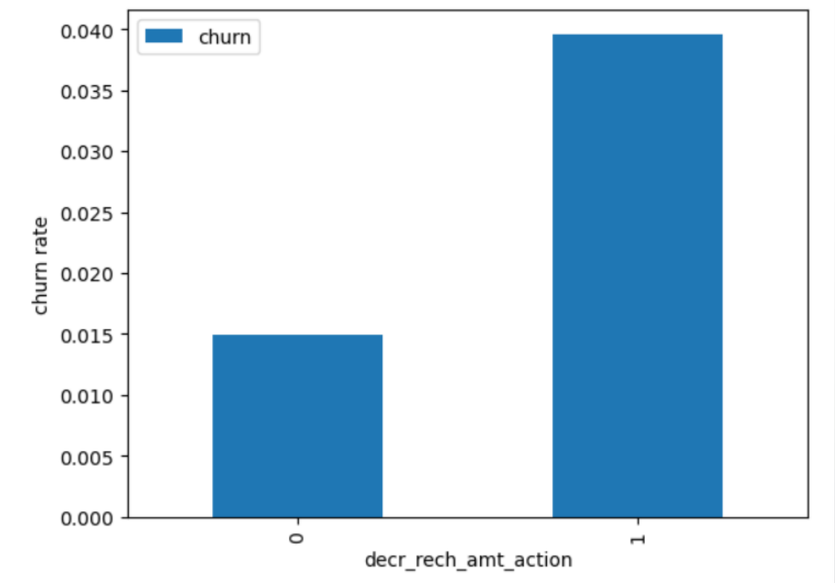
- Filter high-value customers
- Tag churners and remove attributes of the churn phase
- Null value treatment – Removed columns with more than 30% missing data, remove unwanted columns like date, mobile number,circle id etc. , deleted rows with missing values as it was a small % of the overall data.
- Outlier treatment : Outliers have been capped at 10th and 90th percentile for the lower and upper level respectively.
- Define new features : Derived new features using the existing ones for more intuitive analysis.

Exploratory Data analysis



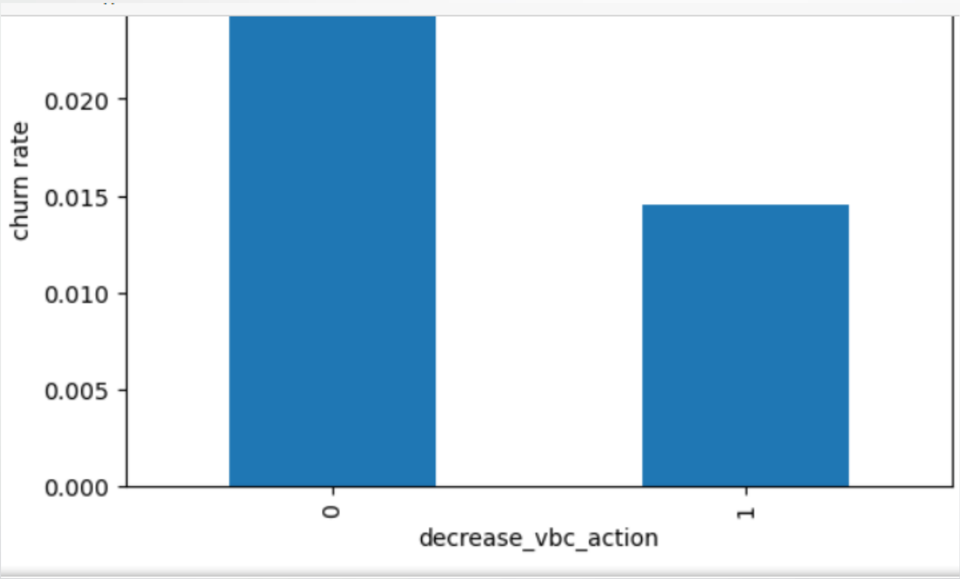
- Customers with a higher churn rate are those whose Minutes of Usage (MOU) decreased during the action phase compared to the good phase.

Exploratory Data analysis



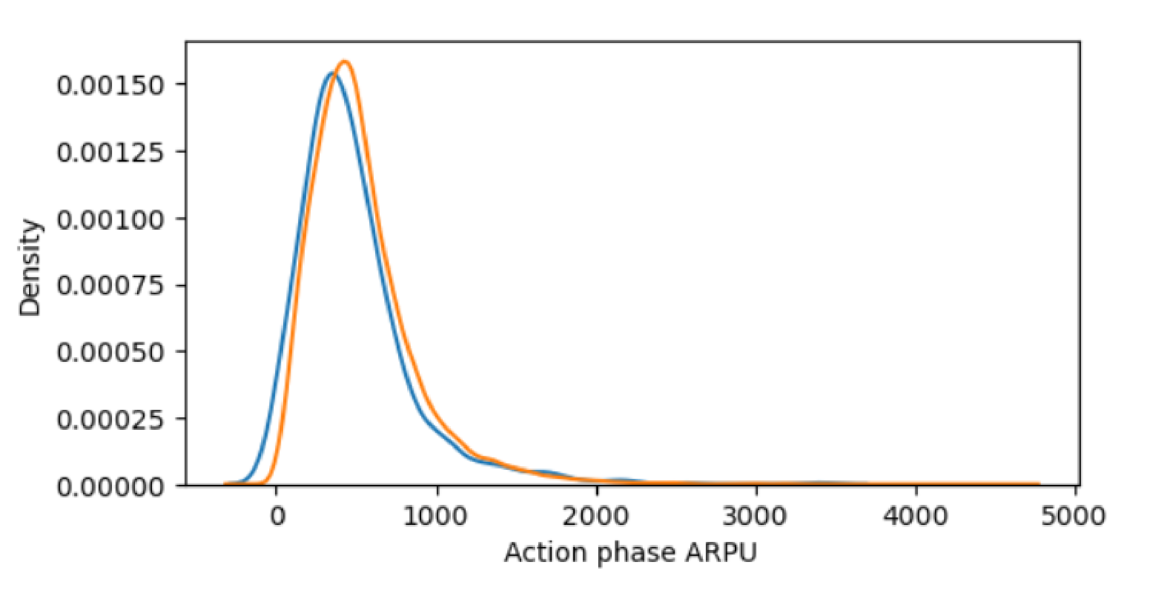
- Customers with a churn tendency are more prevalent among those whose recharge amount during the action phase is lower than the amount during the good phase.

Exploratory Data analysis



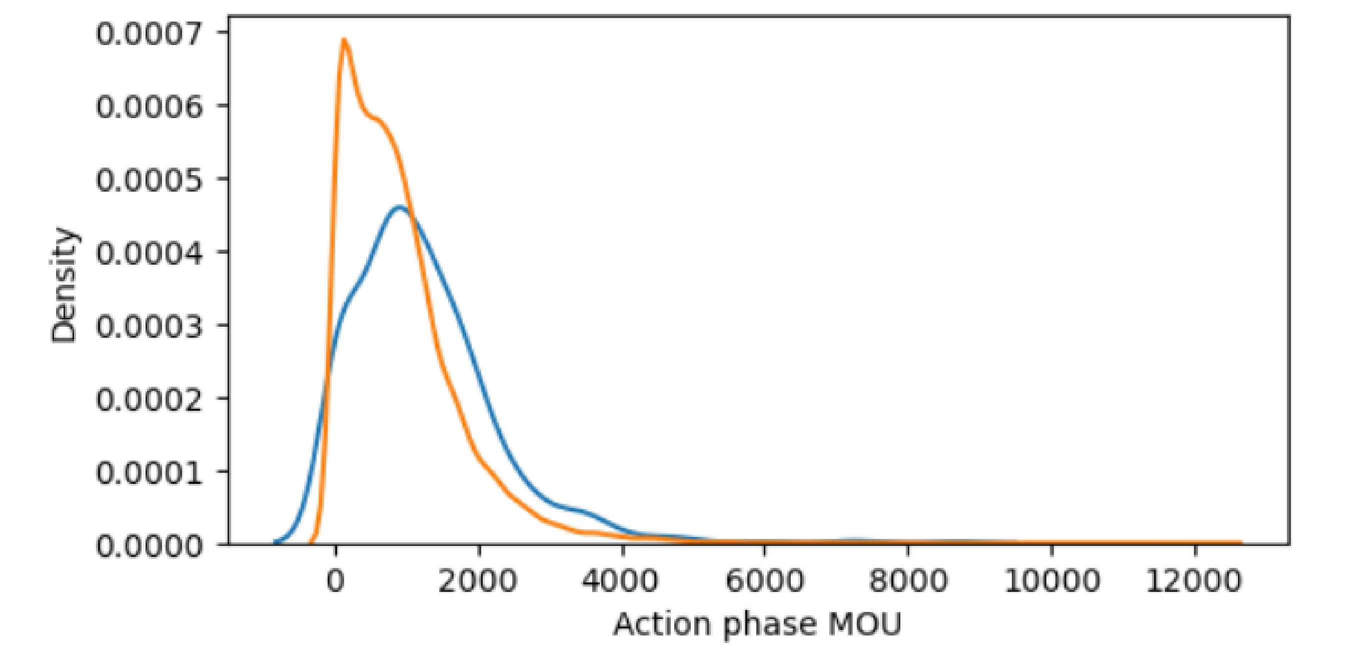
- The churn rate is higher among customers whose 'volume-based cost' experienced an increase during the action month. This suggests that customers are less likely to invest in higher monthly recharges during the action phase.

Exploratory Data analysis



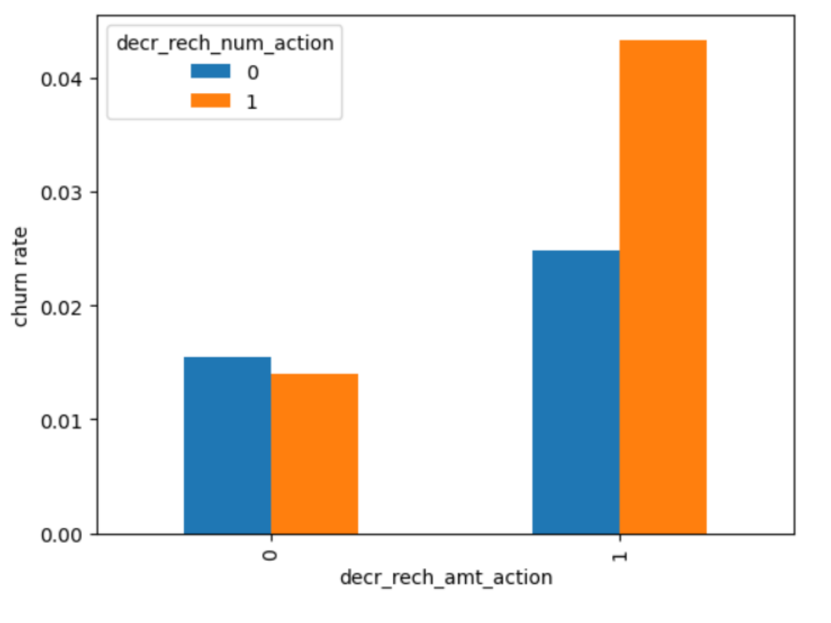
- Churned customers tend to have a higher Average Revenue Per User (ARPU), particularly within the 0 to 900 range. The likelihood of churn diminishes for customers with higher ARPU. On the other hand, for non-churned customers, ARPU is predominantly high in the 0 to 1000 range.

Exploratory Data analysis



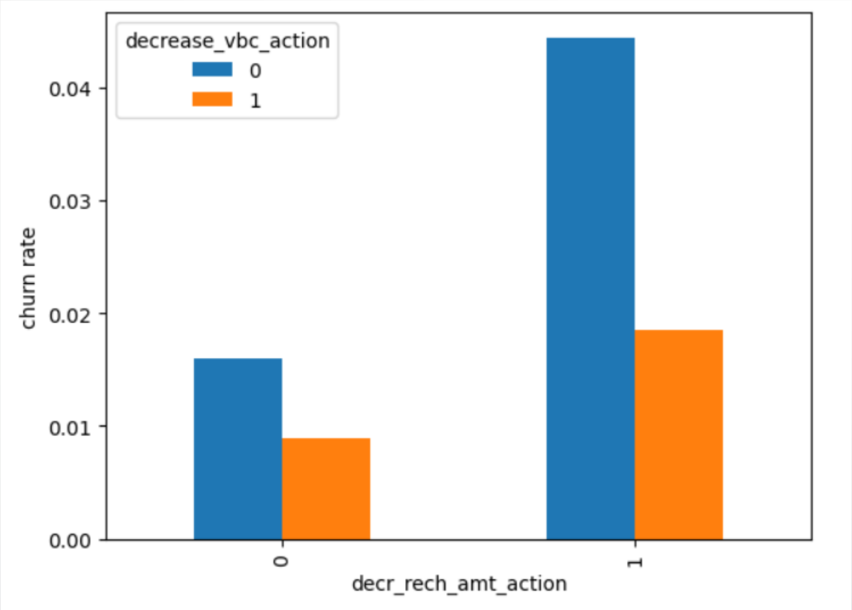
- Churned customers primarily exhibit Minutes of Usage (MOU) concentrated within the 0 to 2500 range. There's an inverse relationship: as MOU increases, the likelihood of churn decreases.

Exploratory Data analysis



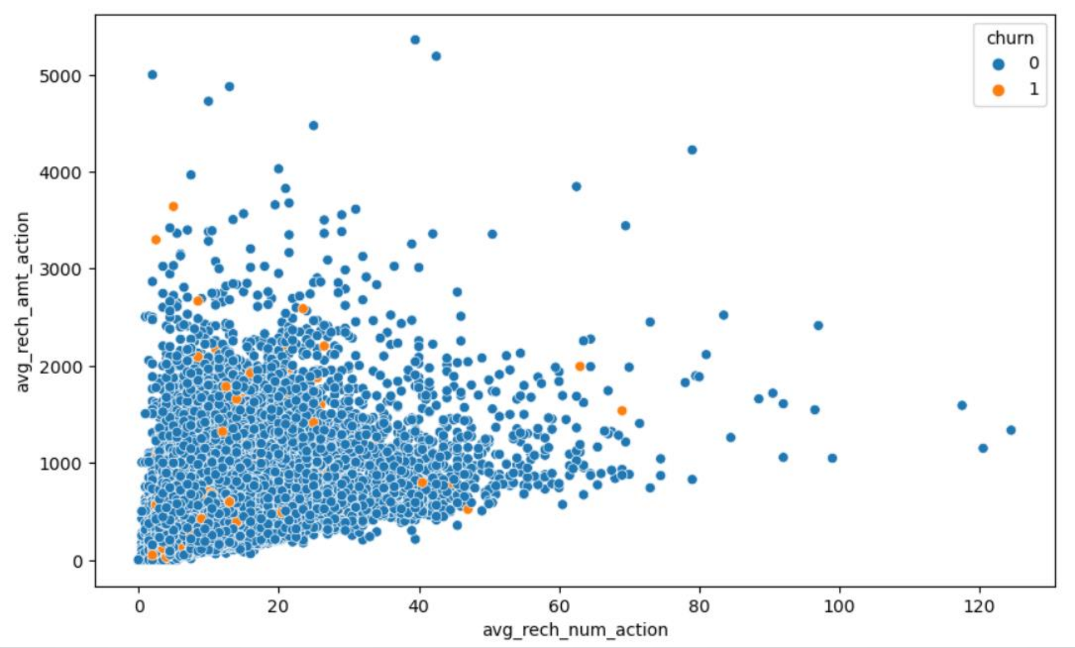
- It is evident from the above plot, that the churn rate is more for the customers, whose recharge amount as well as number of recharge have decreased in the action phase than the good phase

Exploratory Data analysis



- The data reveals a higher churn rate among customers who experience a reduction in recharge amounts, coupled with an increase in volume-based costs during the action month.

Exploratory Data analysis



- The data indicates a noticeable correlation: as the frequency of recharges increases, there is a corresponding rise in the total recharge amount. In simpler terms, when users engage in more recharge activities, they tend to spend more money.

Churn Prediction –
Model Building approach

- Train – test split : Used 70-30 split for train & test data.
- Class imbalance : As the churn is about 5-8% hence there is a class imbalance. Used SMOTE to balance it.
- Feature scaling : Used standard scaling.
- Feature selection : Used PCA for dimensionality reduction as we have a large number of features. Used RFE.
- Hyperparameter tuning
- Evaluation metrics : Built several models like logistics regression, SVM, decision tree, random forest. Highlighted the correct metric that gives the correct prediction for churn cases.

Logistic regression
Model summary

- Train set
  - Accuracy = 0.86
  - Sensitivity = 0.89
  - Specificity = 0.83
- Test set
  - Accuracy = 0.83
  - Sensitivity = 0.81
  - Specificity = 0.83
- Overall, the model is performing well in the test set, what it had learnt from the train set.

Support Vector Machine(SVM) with PCA Model summary

- We can achieve comparable average test accuracy (90%) with gamma=0.0001 as well, though we'll have to increase the cost C for that. So to achieve high accuracy, there's a tradeoff between:
- High gamma (i.e. high non-linearity) & average value of C
- Low gamma (i.e. less non-linearity) & high value of C
- The model will be simpler if it has as less non-linearity as possible, hence we choose gamma=0.0001 and a high C=100.

- 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
Model summary

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

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

- As per the model performance that the Sensitivity has been decreased while evaluating the model on the test set. However, the accuracy & specificity is quite good in the test set.

Random Forest with PCA

Model summary

- Train set
  - Accuracy = 0.84
  - Sensitivity = 0.88
  - Specificity = 0.80

- Test set
  - Accuracy = 0.80
  - Sensitivity = 0.75
  - Specificity = 0.80

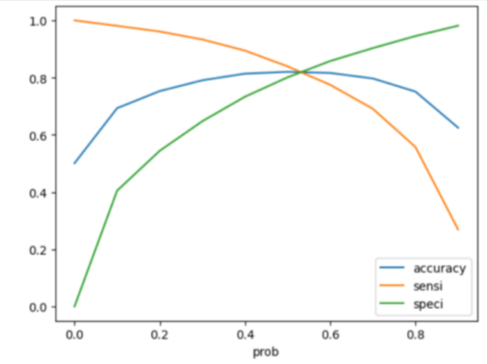
- The Sensitivity has decreased while evaluating the model on the test set. However, the accuracy & specificity is quite good in the test set.

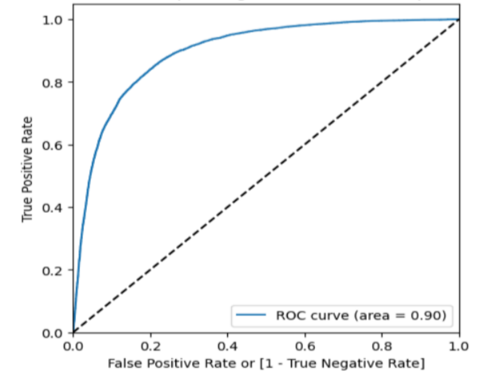
Conclusion with PCA

- After trying different models we can see that for achieving the best sensitivity, which is the goal, the classic Logistic regression or the SVM models performs well. For both the models the sensitivity was approximately 81%. Also we have good accuracy of approximately 85%.

Logistic regression

- Analysis of the curve
- Accuracy - Becomes stable around 0.6
- Sensitivity - Decreases with the increased probability.
- Specificity - Increases with the increasing probability.
- At point 0.6 the three parameters cut each other, we can see that there is a balance between sensitivity & specificity with a good accuracy.
- 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 can take 0.5 for achieving higher sensitivity, which is the Ultimate goal.





ROC curve is closer to 1, which is the Gini of the model.

Logistic regression
Model summary

- Train set
  - Accuracy = 0.84
  - Sensitivity = 0.81
  - Specificity = 0.83

- Test set
  - Accuracy = 0.78
  - Sensitivity = 0.82
  - Specificity = 0.78

- Overall, the model is performing well in the test set, basis the learning from the train set.

Conclusion without PCA

- The logistic regression model without Principal Component Analysis (PCA) demonstrates favorable sensitivity and accuracy, akin to models incorporating PCA, according to the current analysis. Consequently, choosing the simpler logistic regression model without PCA is warranted. This model not only elucidates key predictor variables but also underscores the significance of each variable. It assists in pinpointing the vital variables essential for decision-making regarding potential churned customers. Thus, this model holds greater relevance in conveying insights to the business.

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

Here are few top variables selected in the logistic regression model.

- We can see most of the top variables have negative coefficients, which means the variables are inversely correlated with the churn probability.
- E.g.:- If the local incoming MOU (loc_ic_mou_8) is lesser in the month of August than any other month, there are higher chances that the customer is likely to churn.

Recommendation

- Focus on customers whose Minutes of Usage (MOU) for incoming local calls and outgoing ISD calls have decreased, particularly in the action phase, mostly observed in the month of August.
- Identify customers with lower outgoing charges for other services in July and reduced incoming charges for other services in August.
- Target customers experiencing an increase in Value-Based Cost (VBC) during the action phase, as they are more likely to churn. Consider offering incentives to retain this segment.
- Customers with a higher monthly 3G recharge in August are at a higher likelihood of churning.
- Look out for customers witnessing a decline in STD incoming Minutes of Usage (MOU) for operators in August, as they are more prone to churn.
- Customers with a decreasing monthly 2G usage in August are also expected to churn.
- Identify customers with decreasing incoming MOU in August, as they show a higher likelihood of churning.
- Pay attention to the "roam_og_mou_8" variable, where positive coefficients (0.7135) indicate that customers with increasing roaming outgoing minutes of usage are more likely to churn.