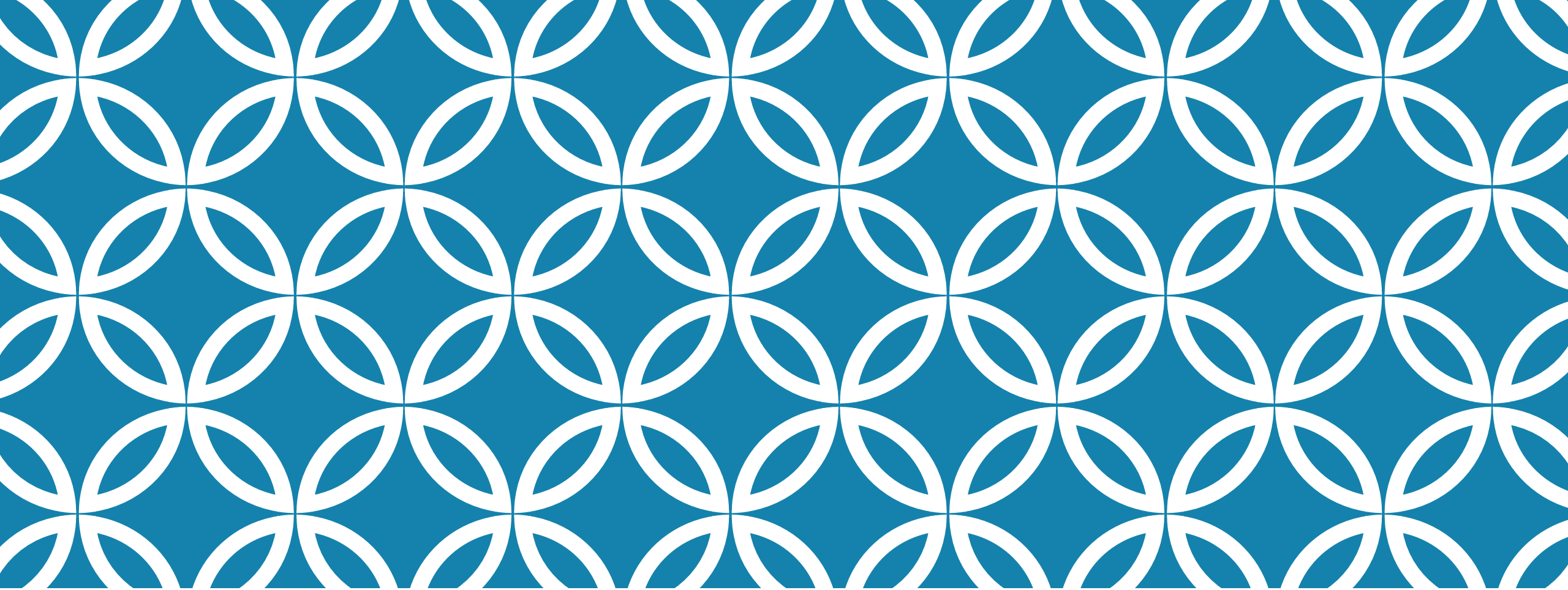


DOMAIN CASE STUDY TELECOM CHURN

By Swati, Sowmya & Chandana

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 behaviour 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

1. **Filter high-value customers**

2. **Tag churners and remove attributes of the churn phase**

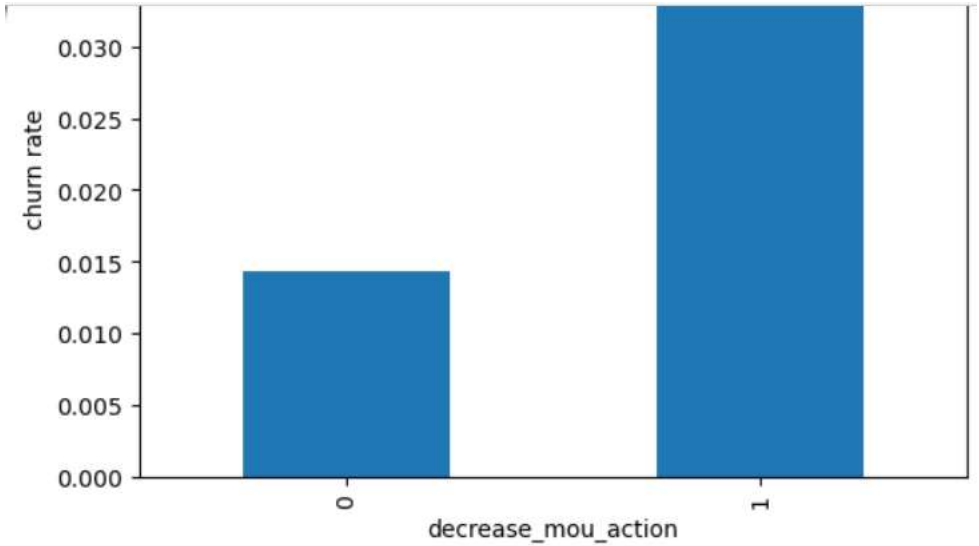
3. Null value treatment – Removed columns with more than 30% missing data, remove unwanted columns like date, mobile number,circle id etc. , deleted rows with missing values as it was a small % of the overall data.

4. Outlier treatment : Outliers have been capped at 10^{th} and 90^{th} percentile for the lower and upper level respectively.

5. Define new features : Derived new features using the existing ones for more intuitive analysis.

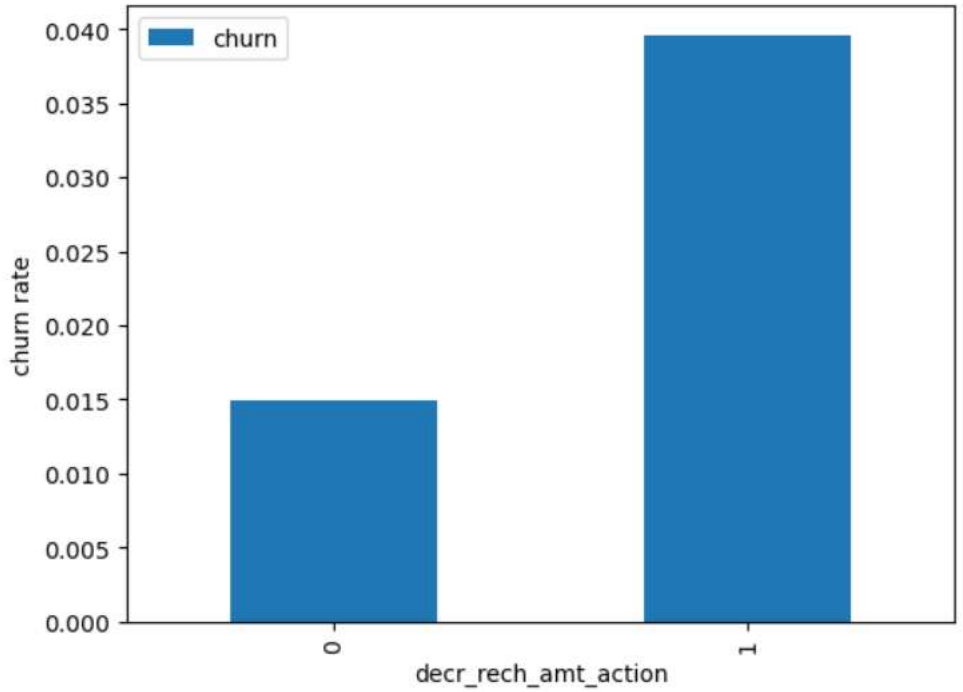
6.

EXPLORATORY DATA ANALYSIS



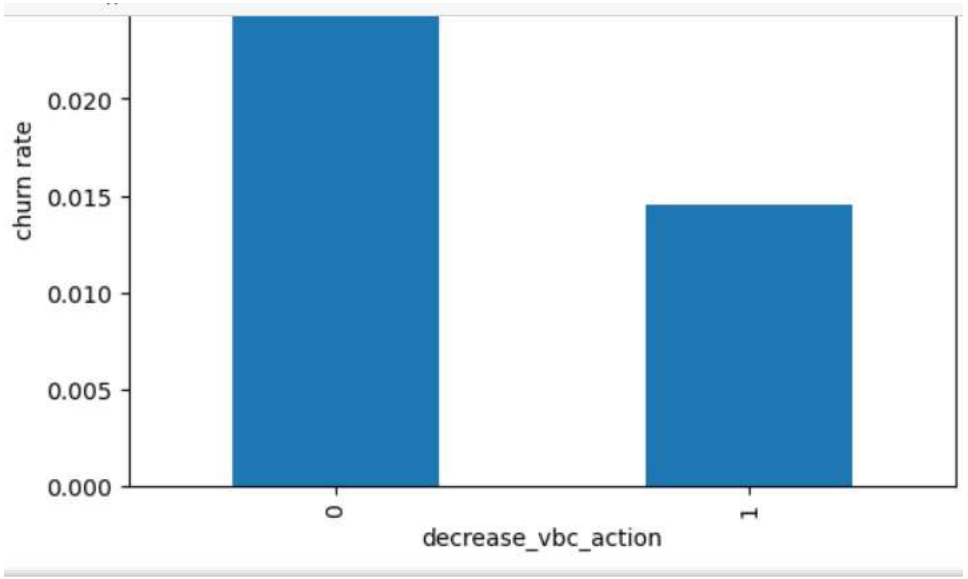
The churn rate is more for the customers whose minutes of usage(mou) decreased in the action phase than the good phase.

EXPLORATORY DATA ANALYSIS



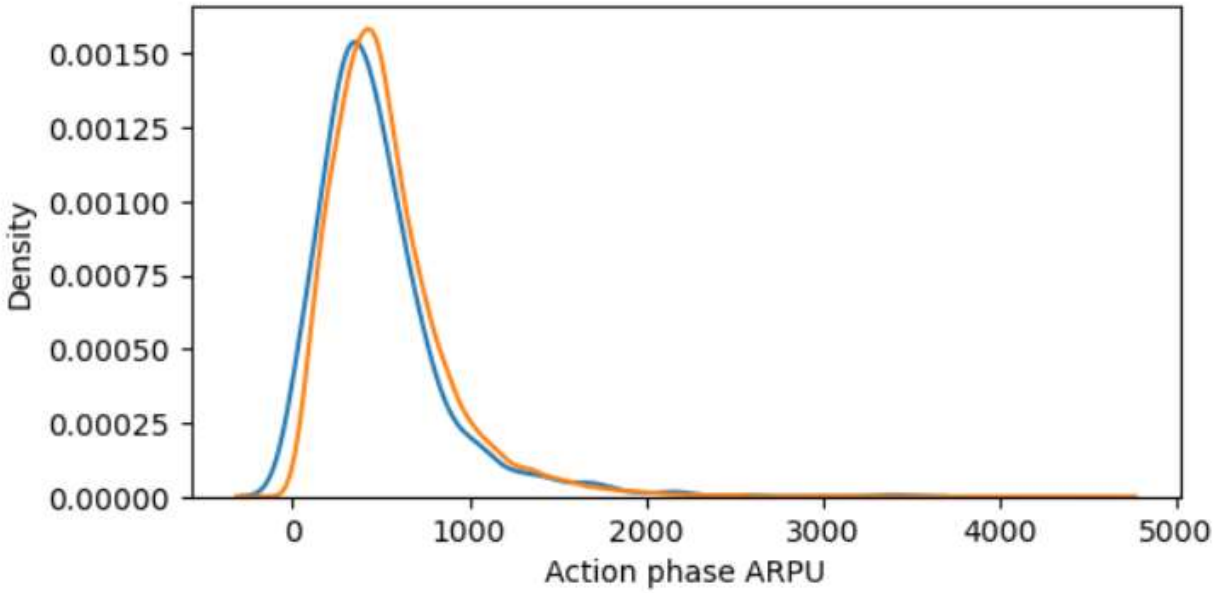
The churn rate is more for the customers, whose amount of recharge in the action phase is lesser than the amount in good phase.

EXPLORATORY DATA ANALYSIS



The churn rate is more in customers, whose ‘volume based cost’ increased in the action month. This indicates that the customers do not invest more monthly recharge when they are in the action phase.

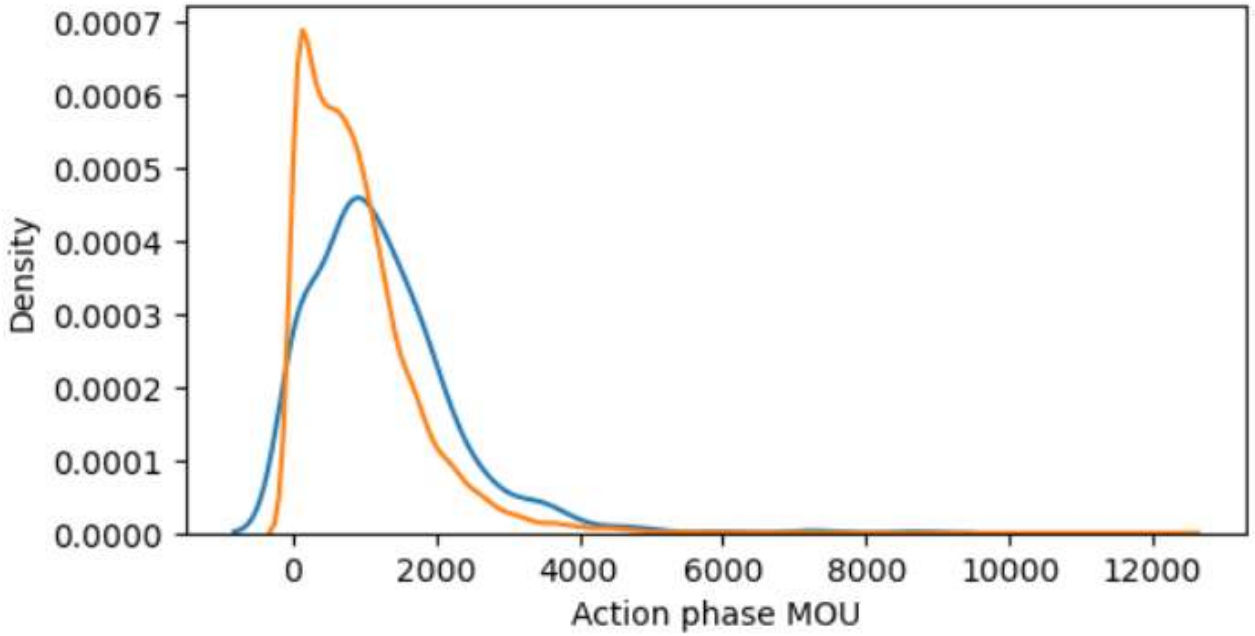
EXPLORATORY DATA ANALYSIS



Average revenue per user (ARPU) for the churned customers is mostly high on the 0 to 900. The higher ARPU customers are less likely to be churned.

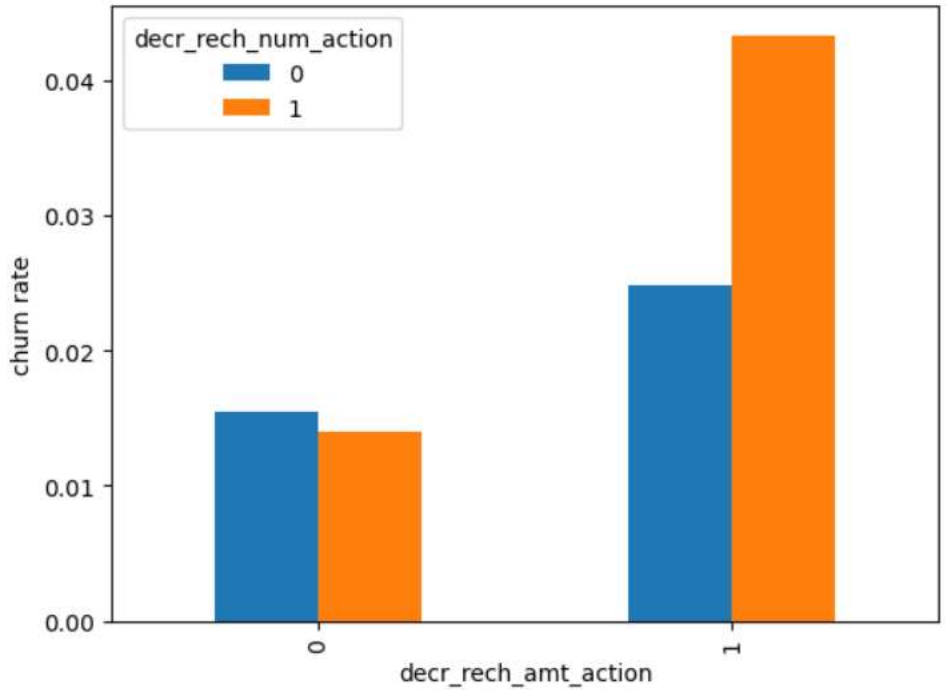
ARPU for the not churned customers is mostly high on the 0 to 1000.

EXPLORATORY DATA ANALYSIS



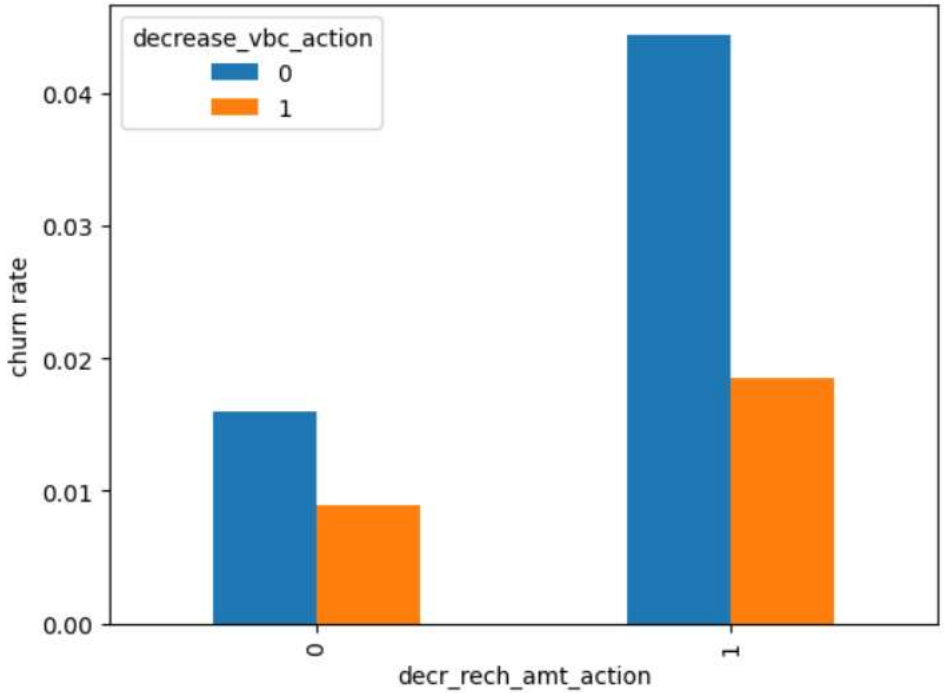
Minutes of usage(MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability.

EXPLORATORY DATA ANALYSIS



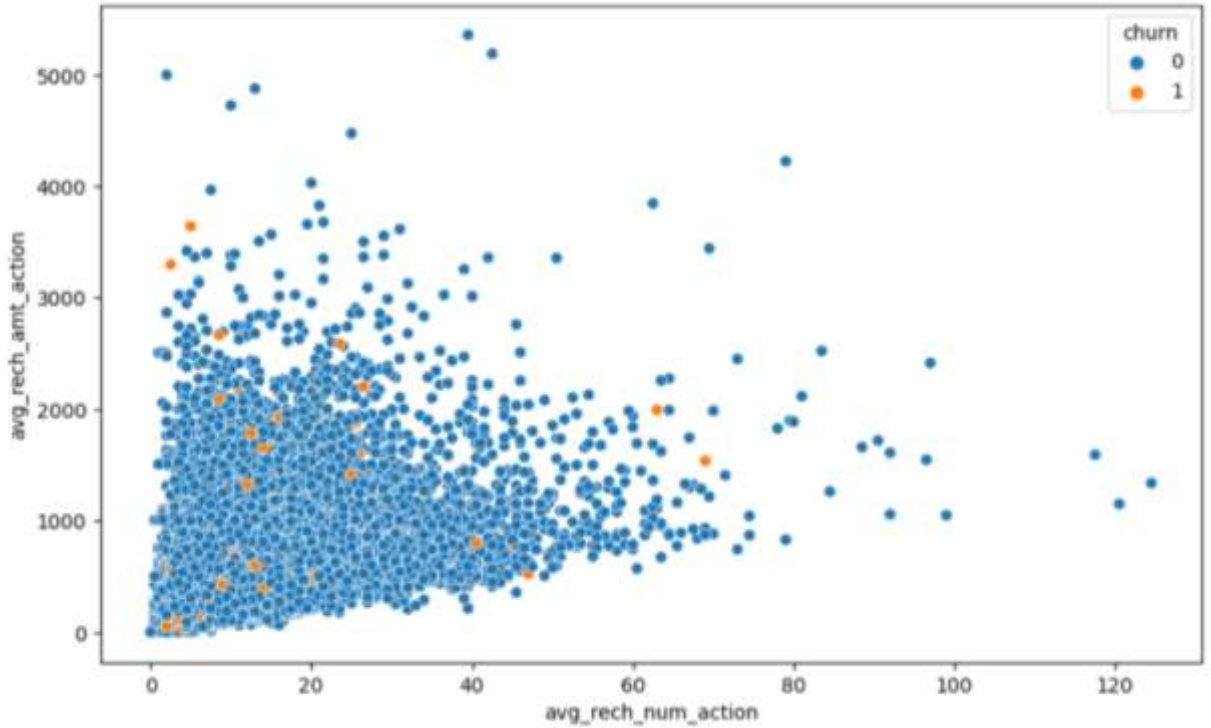
It is evident from the above plot, that the churn rate is more for the customers, whose recharge amount as well as number of recharge have decreased in the action phase than the good phase

EXPLORATORY DATA ANALYSIS



We can see that the churn rate is more for the customers, whose recharge amount is decreased along with the volume based cost is increased in the action month.

EXPLORATORY DATA ANALYSIS



It is shown From the above pattern that the recharge number & the recharge amount are mostly propotional. More the number of recharge, more is the amount of the recharge.

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.

Model summary

- 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

Model summary

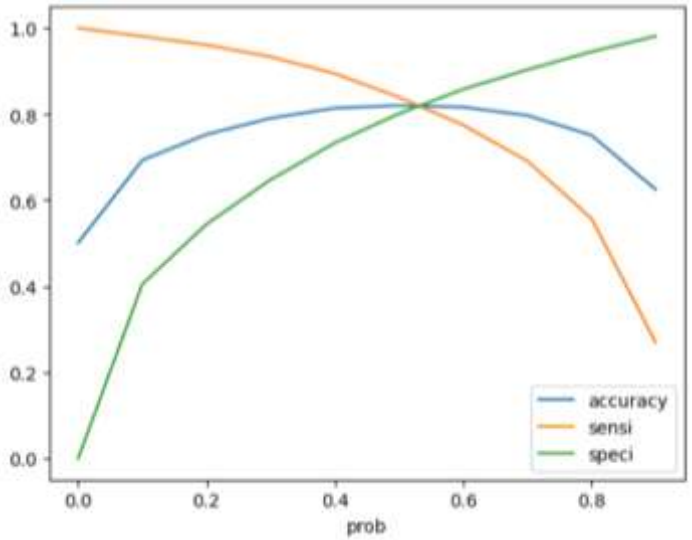
- Train set
  - Accuracy = 0.84
  - Sensitivity = 0.88
  - Specificity = 0.80
- Test set
  - Accuracy = 0.80
  - Sensitivity = 0.75
  - Specificity = 0.80

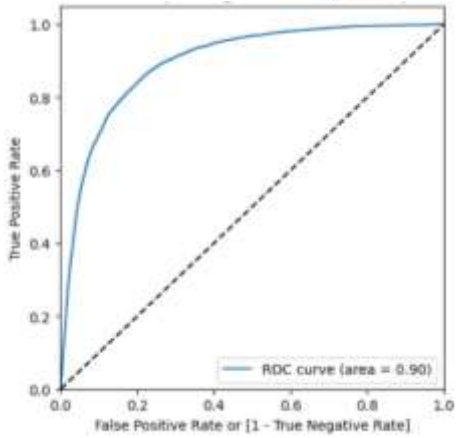
The Sensitivity has decreased while evaluating the model on the test set. However, the accuracy & specificity is quite good in the test set.

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

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 model with no PCA has good sensitivity and accuracy as per the analysis so far, which are comparable to the models with PCA. Hence, we can go for the more simplistic model such as logistic regression without PCA which explains the important predictor variables as well as the significance of each variable. The model also helps us to identify the variables which should be act upon for making the decision on to be churned customers. Hence, the model is more relevant in terms of explaining 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

1. Target the customers, whose MOU of the incoming local calls & outgoing ISD calls are lesser in the action phase that mostly in the month of August
2. Target the customers, whose outgoing others charge in July & incoming others on August are less.
3. Target the customers having value based cost(VBC) in the action phase increased are more likely to churn compared to the other customers. Hence, these customers may be a good target to provide offer and retain.
4. Customers with higher monthly 3G recharge in August is more are likely to be churned.
5. Customers having decreasing STD incoming MOU for operators for the month of August are also more likely to churn.
6. Customers decreasing monthly 2G usage for August are also expected to churn.
7. Customers having decreasing incoming MOU for August are more likely to churn.
8. roam_og_mou_8 variables have positive coefficients (0.7135) which is a sign that whose roaming outgoing minutes of usage is increasing are more likely to churn.