upGrad

Domain Oriented Case Study

(Telecom Churn Case Study)

Team Members (DS-C54) : Umal Kumar Bhole
Stuti Bhatt
Vedurla V S N Anjani Suchithra

Table Of Content

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.
    - For many incumbent operators, retaining highly 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 reading & Data understanding

- There are two main models of payment in the telecom industry - postpaid (customers pay a monthly/annual bill after using the services) and prepaid (customers pay/recharge with a certain amount in advance and then use the services).
- In the postpaid model, when customers want to switch to another operator, they usually inform the existing operator to terminate the services, and we directly know that this is an instance of churn.
- However, in the prepaid model, customers who want to switch to another network can simply stop using the services without any notice, and it is hard to know whether someone has actually churned or is simply not using the services temporarily (e.g. someone may be on a trip abroad for a month or two and then intend to resume using the services again).
- Thus, churn prediction is usually more critical (and non-trivial) for prepaid customers, and the term ‘churn’ should be defined carefully. Also, prepaid is the most common model in India and Southeast Asia, while postpaid is more common in Europe in North America.
- This project is based on the Indian and Southeast Asian markets.

Definition of churn

- There are various ways to define churn, such as:
  - **Revenue-based churn:** Customers who have not utilized any revenue-generating facilities such as mobile internet, outgoing calls, SMS, etc. over a given period of time. One could also use aggregate metrics such as ‘customers who have generated less than INR 4 per month in total/average/median revenue’.
  - The main shortcoming of this definition is that there are customers who only receive calls/SMSes from their wage-earning counterparts, i.e. they don’t generate revenue but use the services. For example, many users in rural areas only receive calls from their wage-earning siblings in urban areas.
  - **Usage-based churn:** Customers who have not done any usage, either incoming or outgoing - in terms of calls, internet, etc. over a period of time.
  - A potential shortcoming of this definition is that when the customer has stopped using the services for a while, it may be too late to take any corrective actions to retain them. For e.g., if we define churn based on a ‘two-month zero usage’ period, predicting churn could be useless since by that time the customer would have already switched to another operator.
- In this project, we will use the **usage-based** definition to define churn.

High-value churn

- In the Indian and the southeast Asian market, approximately 80% of revenue comes from the top 20% customers (called high-value customers). Thus, if we can reduce churn of the high-value customers, we will be able to reduce significant revenue leakage.
- In this project, we will define high-value customers based on a certain metric (mentioned later below) and predict churn only on high-value customers.

Understanding customer behavior during churn

- Customers usually do not decide to switch to another competitor instantly, but rather over a period of time (this is especially applicable to high-value customers). In churn prediction, we assume that there are three phases of the customer lifecycle :
    1. **The 'good' phase:** In this phase, the customer is happy with the service and behaves as usual.
    2. **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 behavior 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.)
    3. **The 'churn' phase:** In this phase, the customer is said to have churned. We define churn based on this phase. Also, it is important to note that at the time of prediction (i.e. the action months), this data is not available to us for prediction. Thus, after tagging churn as 1/0 based on this phase, we discard all data corresponding to this phase.
- In this case, since we are working over a four-month window, the first two months are the 'good' phase, the third month is the 'action' phase, and the fourth month is the 'churn' phase.

Dataset and Data dictionary

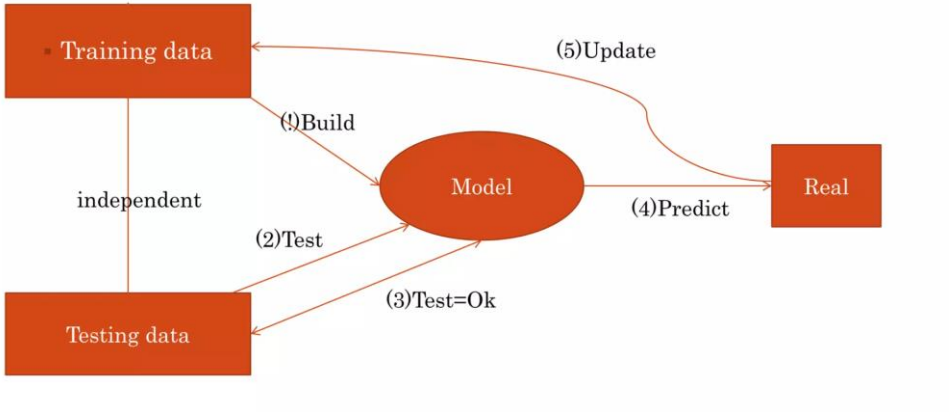
- The dataset can be downloaded from here. The source data is in a CSV file.
- Data dictionary is uploaded. The data dictionary contains meanings of abbreviations. Some frequent ones are loc (local), IC (incoming), OG (outgoing), T2T (telecom operator to telecom operator), T2O (telecom operator to another operator), RECH (recharge), etc.
- The attributes containing 6, 7, 8, and 9 as suffixes imply that those correspond to the months 6, 7, 8, and 9 respectively.
- Dataset contains 99999 rows and 226 columns
- We did not see any missing value in the dataset

Data Preparation

- The following data preparation steps are crucial for this problem:
  1. **Derive new features** This is one of the most important parts of data preparation since good features are often the differentiators between good and bad models. We will use our business understanding to derive features that we think could be important indicators of churn.
  2. **Filter high-value customers** As mentioned above, we need to predict churn only for the high-value customers. Define high-value customers as follows: Those who have recharged with an amount more than or equal to X, where X is the 70th percentile of the average recharge amount in the first two months (the good phase).
  3. **Tag churners and remove attributes of the churn phase** Now tag the churned customers (churn=1, else 0) based on the fourth month as follows: Those who have not made any calls (either incoming or outgoing) AND have not used mobile internet even once in the churn phase. The attributes we need to use to tag churners are:
     - total_ic_mou_9
     - total_og_mou_9
     - vol_2g_mb_9
     - vol_3g_mb_9
- After tagging churners, we need to remove all the attributes corresponding to the churn phase (all attributes having ‘ _9’, etc. in their names).

Model Building

- Build models to predict churn. The predictive model that we are going to build will serve two purposes:
  1. It will be used to predict whether a high-value customer will churn or not, in the near future (i.e. churn phase). By knowing this, the company can take action steps such as providing special plans, discounts on recharge, etc.
  2. It will be used to identify important variables that are strong predictors of churn. These variables may also indicate why customers choose to switch to other networks.



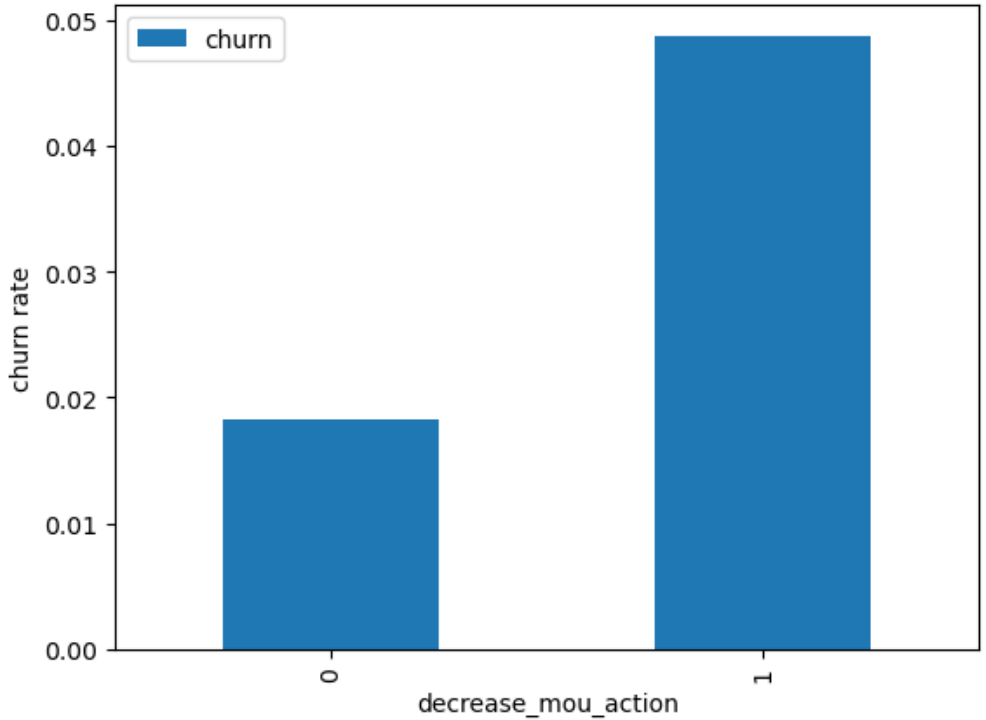
Exploratory Data Analysis (EDA)

- We can take the following suggestive steps to build the model:
    - Handle missing values(Drop all columns which have > 30% missing values)
    - Delete unwanted columns which will not used eg Date columns
    - Filter with high-value customers
    - Again, handling missing values in columns, and this time with 50%
    - Tag churners
    - Delete attributes for the churning phase
    - Understanding and treating outliers
    - Add New columns if required.

Starting with Univariate analysis

FINDING

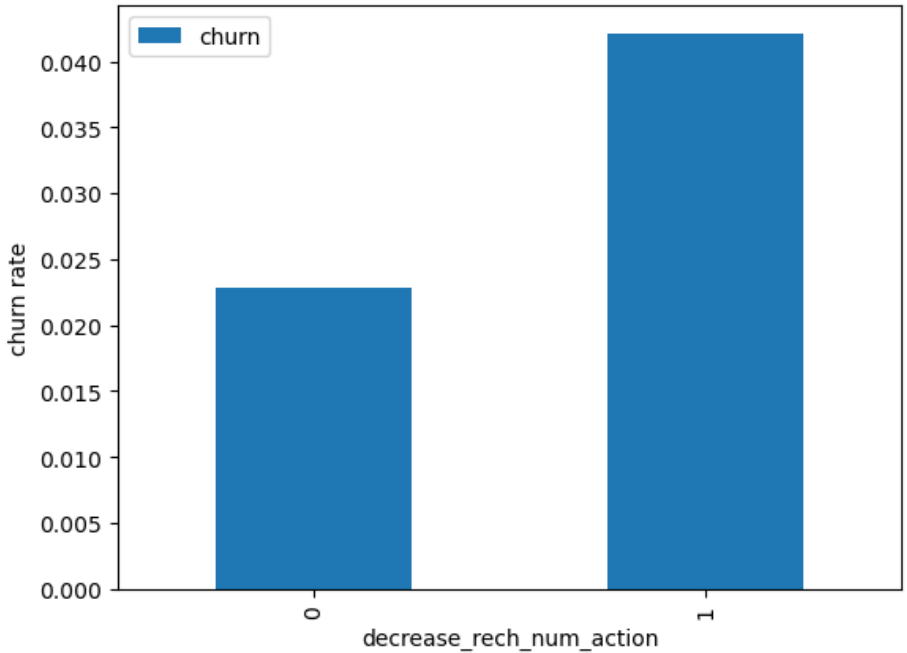
- The data indicates that customers who experienced a decrease in their minutes of usage (MOU) during the action phase had a higher churn rate.



Churn rate in regard to the customer decreased no. of recharge in action month

FINDING

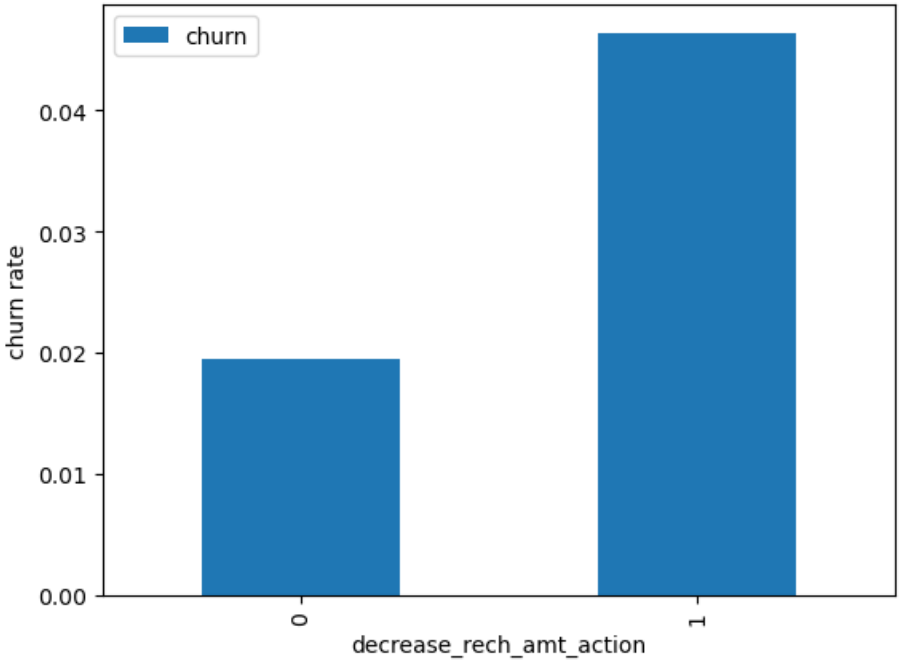
- As anticipated, the churn rate is higher for customers who made fewer recharges in the action phase compared to the good phase.



Churn rate regarding customer decreased her/his amount of recharge in the action month

FINDING

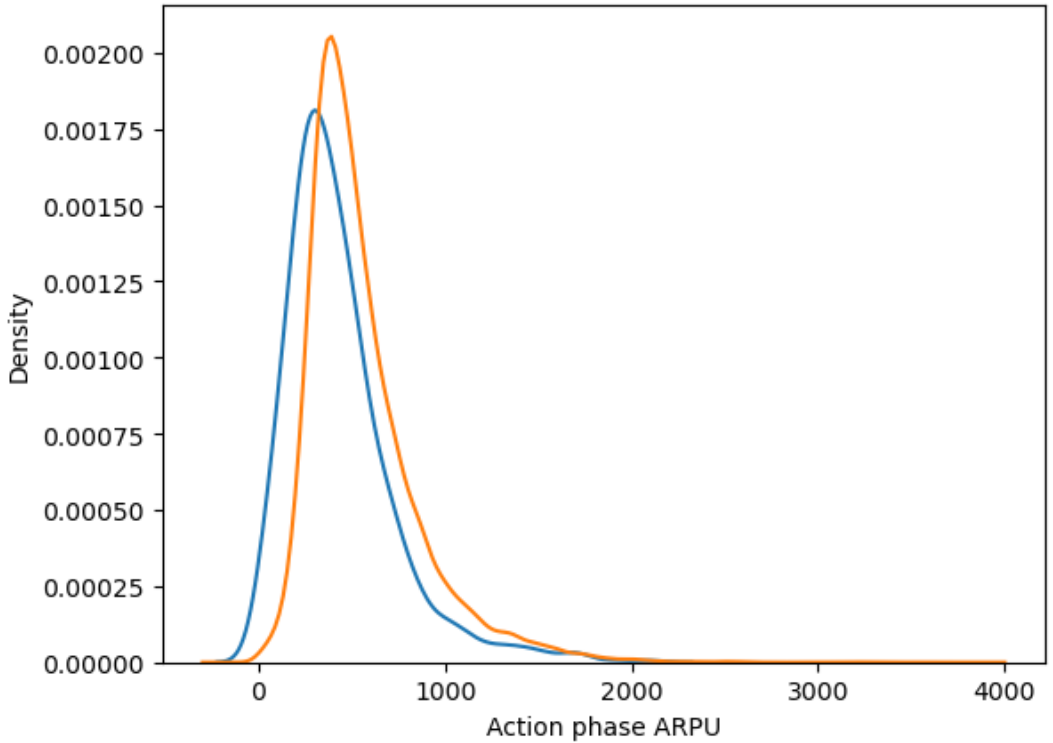
- Here, we observe a similar pattern. The churn rate is higher for customers whose recharge amount in the action phase is less than that in the good phase.



Average revenue per customer (churn and not churn) in the action phase

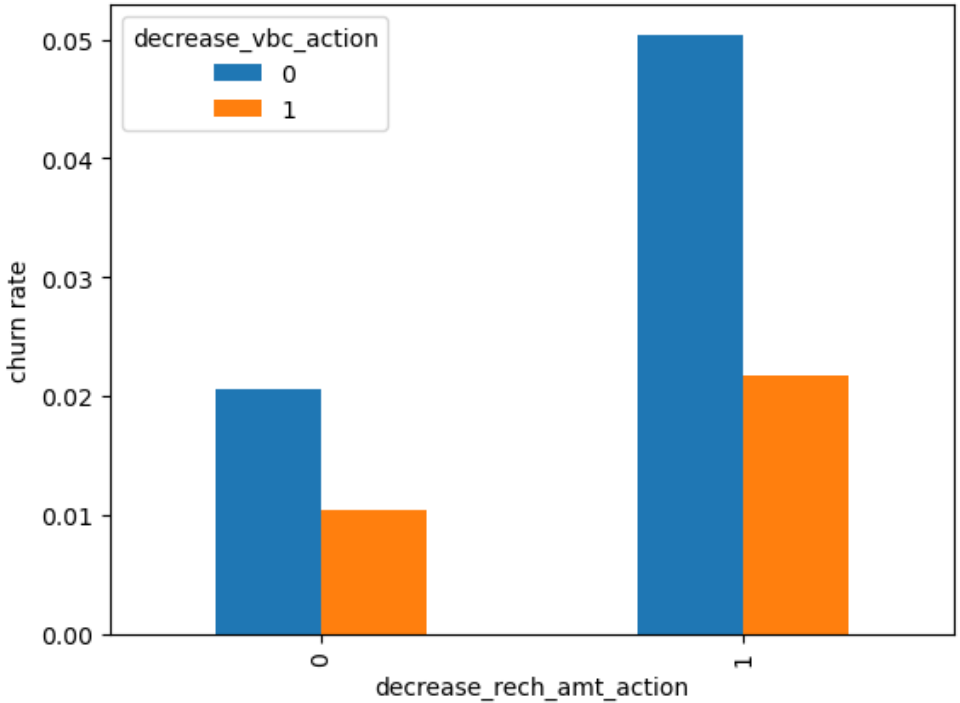
FINDING

- The average revenue per user (ARPU) for churned customers is primarily concentrated in the 0 to 900 range. Customers with higher ARPU are less prone to churn. For non-churned customers, ARPU is predominantly concentrated in the 0 to 1000 range.



Bivariate analysis

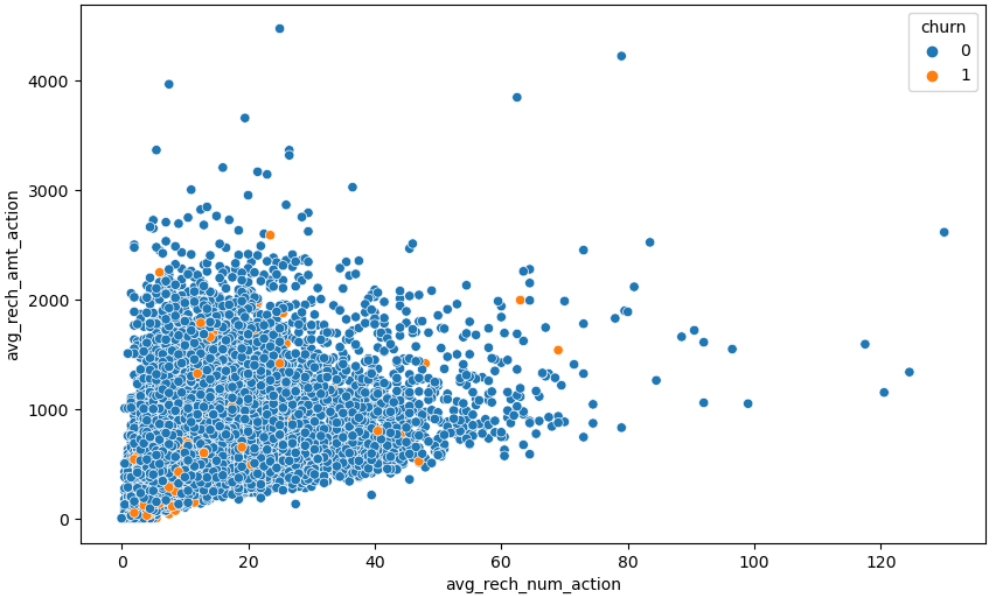
- Analyze churn rate against decreasing recharge amount and volume-based cost in the action phase.



Analyze recharge amount against a number of recharges in action month using scatter plot for better understanding

FINDINGS

- Similarly, we observe a higher churn rate among customers whose recharge amount decreases while the volume-based cost increases in the action month.
- The plotted data above highlights a greater churn rate among customers who experience a decrease in both recharge amount and the number of recharges during the action phase compared to the good phase.
- The pattern depicted above indicates a strong correlation between the number of recharges and the total recharge amount. More recharges generally result in a higher total recharge amount.

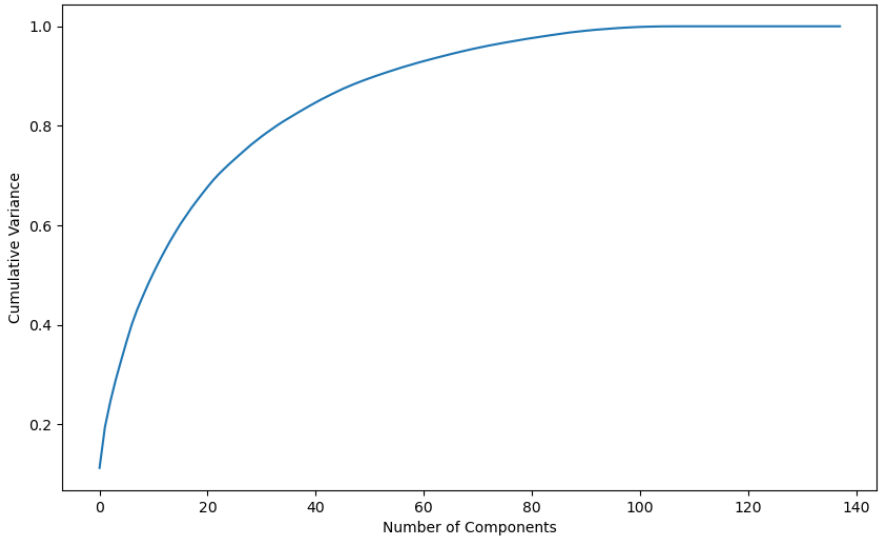


Principal Component Analysis (PCA)

- Instantiate PCA
- Fit train data set on PCA
- Check the Principal components
- Check the Cumulative variance of the Principal Components
- So as per analysis we perform PCA with 60 components
- Logistic regression with Principal Components Analysis
- Tuning hyperparameter
- Logistic regression with optimal C
- Prediction on the train data set
- Check the prediction on the test data set
- Hyperparameter tuning
- Plotting the accuracy with various C and gamma values

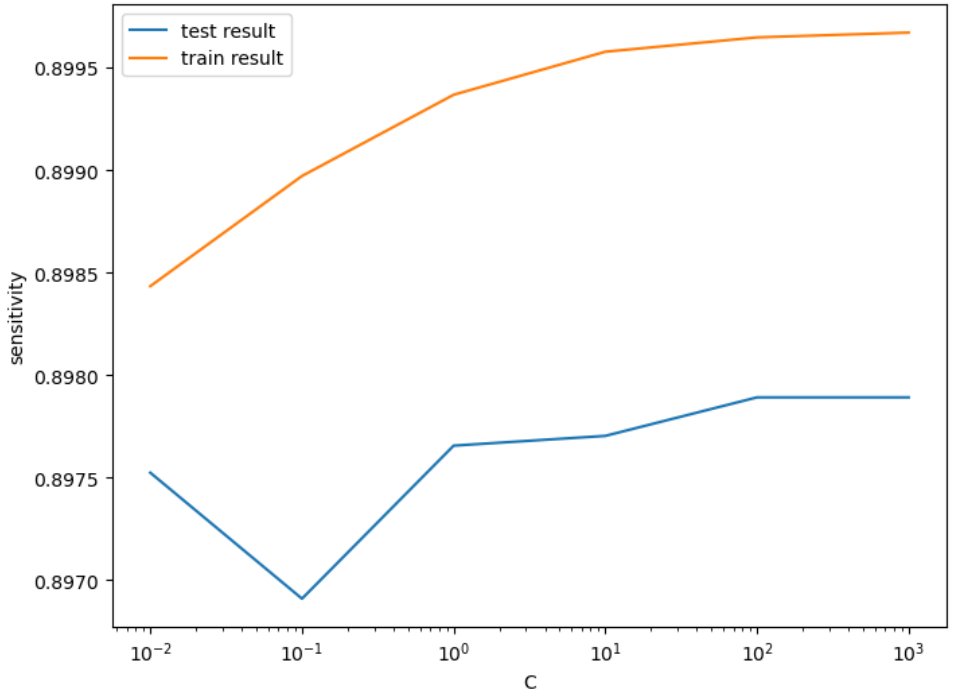
Sensitivity Recall

- Our primary focus lies on achieving a higher Sensitivity/Recall score rather than emphasizing accuracy. This strategic choice is rooted in the need to prioritize churn cases over non-churn cases. Our main objective is to retain customers who might be at risk of churning. In this context, it's acceptable to occasionally misclassify non-churn customers as potential churners and offer them incentives to ensure customer retention. Consequently, the sensitivity score holds greater importance in this scenario.



Overall Model Overview

- 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.



Decision tree with PCA

- Hyperparameter tuning
- Model with optimal hyperparameters
- Prediction on the train set
- Check the prediction on the test set

Model Overview

•Train set

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

•Test set

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

The accuracy and specificity is quite good in the test set.

Random forest with PCA

- Prediction on the train data set
- Prediction on the test data set

Model Overview

•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 been decreased while evaluating the model on the test set. The accuracy and specificity is quite good in the test set.

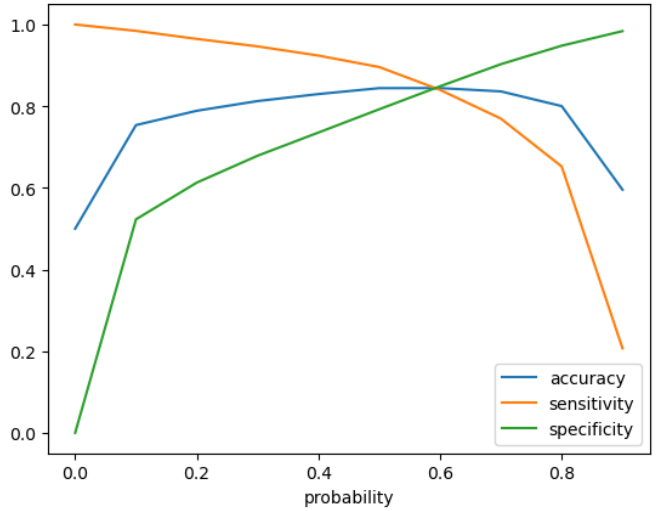
Logistic regression with No PCA

- Analysis of the Model
- It is evident that some features exhibit positive coefficients, while others display negative coefficients. Several features have elevated p-values, rendering them insignificant within the model.
- Coarse Tuning (Combining Auto and Manual Approaches)
- To begin, we will employ Recursive Feature Elimination (RFE) to remove a subset of features. Once we have refined our feature set, we will proceed with manual feature elimination. This involves assessing p-values and Variance Inflation Factors (VIFs) to make informed decisions about which features to retain.

Steps to follow:

- Selecting features with RFE
- Model-1 with RFE 15 selected columns
- Check Variance Inflation Factors
- Remove column highest p-value 0.99
- Model-2
- Check VIF - Model-2
- Model-3
- VIF - Model-3
- Checking the Model-3 performance on the train set
- Optimal Probability Cutoff Point
- Calculate accuracy, sensitivity, and specificity for various probability cutoffs.

	probability	accuracy	sensitivity	specificity
0.0	0.0	0.500000	1.000000	0.000000
0.1	0.1	0.753629	0.984411	0.522847
0.2	0.2	0.788751	0.964714	0.612789
0.3	0.3	0.812509	0.946371	0.678646
0.4	0.4	0.829638	0.923874	0.735403
0.5	0.5	0.844131	0.895823	0.792439
0.6	0.6	0.844271	0.839860	0.848681
0.7	0.7	0.836173	0.769522	0.902824
0.8	0.8	0.800163	0.652275	0.948051
0.9	0.9	0.595426	0.207001	0.983851



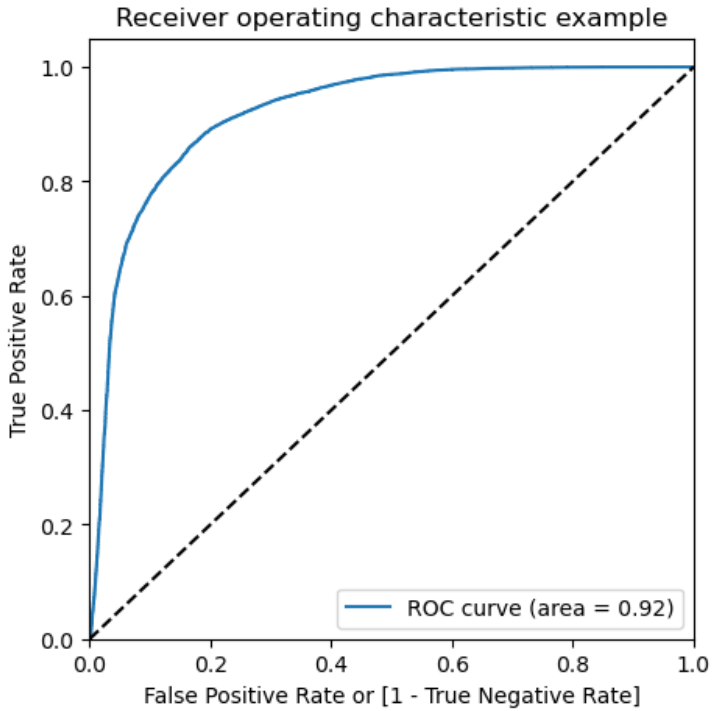
Analysis of the above curve

- Accuracy - Becomes stable around 0.6
- Sensitivity - Decreases with the increased probability.
- Specificity - Increases with the increasing probability.
- At point 0.6 there is a balance between sensitivity and 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 are taking 0.5 to achieve higher sensitivity, which is our main goal.

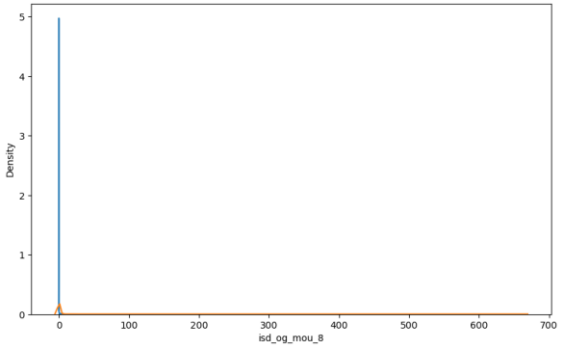
Plotting the ROC Curve

- We can see the area of the ROC curve is closer to 1, which is the Gini of the model.

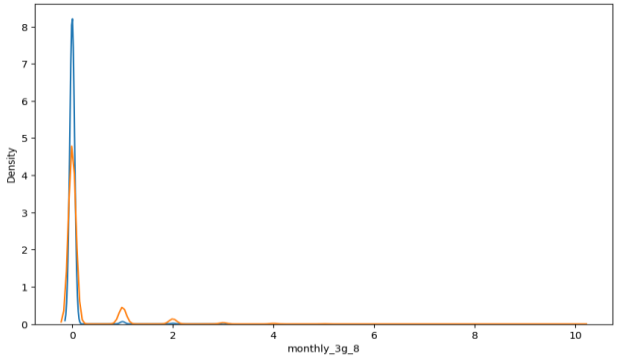
- Model summary
  - Train set
    - Accuracy = 0.84
    - Sensitivity = 0.81
    - Specificity = 0.83
  - Test set
    - Accuracy = 0.78
    - Sensitivity = 0.82
    - Specificity = 0.78
- Overall, the model is performing well in the test set, what it had learned from the train set.



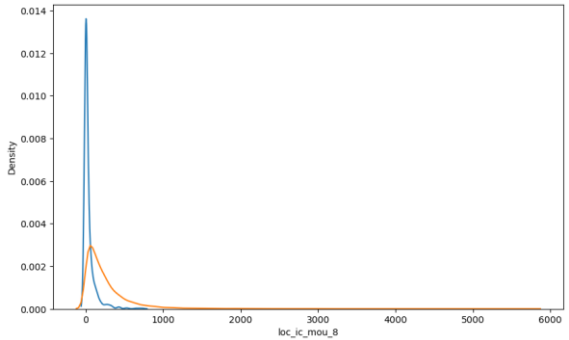
Important predictors for churn and non-churn customers



- We can see that the ISD outgoing minutes of usage for the month of August for churn customers is dense approximately to zero. On the other hand for the non-churn customers, it is little more than the churn customers.



- The number of monthly 3g data for August for the churn customers is very much populated around 1, whereas of non-churn customers it spread across various numbers.
- Similarly, we can plot each variable, which has higher coefficients, and churn distribution.



- We can see that for the churn customers, the minutes of usage for the month of August are mostly populated on the lower side than the non-churn customers.

Conclusion

- **Final conclusion with PCA**
  - The classic Logistic regression or the SVM models perform well. For both the models the sensitivity was approx 81%. Also, we have a good accuracy of approx 85%.

- **Final conclusion without PCA:**
  - The logistic model without PCA demonstrates strong sensitivity and accuracy, which are on par with the models that include PCA. Therefore, opting for a simpler model like logistic regression without PCA is advisable. This model effectively highlights the key predictor variables and their significance, aiding in the identification of variables crucial for determining potential churned customers. Consequently, this model offers greater relevance when it comes to explaining its implications to the business.

Business recommendation

- **Top predictors**
  - Below are a few top variables selected in the logistic regression model.
  - A majority of the right-side table variables exhibit negative coefficients, indicating an inverse correlation with churn probability. For instance, if the local incoming minutes of usage (loc_ic_mou_8) are lower in August compared to other months, there is a higher likelihood of customer churn.

- **Recommendations**
  - Prioritize customers with reduced usage of incoming local calls and outgoing ISD calls during the action phase, especially in August. Target customers with lower outgoing charges for other services in July and decreased incoming charges for the same in August. Customers experiencing an increase in value-based costs during the action phase are more prone to churn and can be a suitable target for offers. Pay attention to customers with increased 3G recharge in August, as they are more likely to churn. Customers with declinin STD incoming minutes of usage for operators T to fixed lines of T in August have a higher likelihood of churn. Customers reducing their monthly 2G usage in August are at a greater risk of churning. Customers with decreased incoming minutes of usage for operators T to fixed lines of T in August are more susceptible to churn. Customers with rising roaming outgoing minutes of usage (roam_og_mou_8) are more likely to churn, indicated by the positive coefficient (0.7135).

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

Thank You!