Telecom Churn Case Study

By-

Prithvi Raj Chauhan
Dhruv Garg
Nikita Dandriyal

Business Problem Overview

- 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.**

- In this project, you will analyse customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn.

Understanding and Defining Churn

- 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 you 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 market.

Steps Followed

1. Reading, understanding and visualising the data
2. Preparing the data for modelling
3. Building the model
4. Evaluate the model

Reading, Understanding and Visualising the Data

➜ Importing DataSet
➜ Reading And Understanding the Dataset
   ◆ Data Shape - (99999,226)
   ◆ Data info
   ◆ Data describe

➜ Handling Missing Data
➜ Eliminating the data columns as the data columns are not required in our examination
   ◆ Dropping columns as the column has only 1 unique value.
➜ Filtering High Rise Customers
   ◆ Creating column avg_rech_amt_6_7 by summing up total recharge amount of month 6 and 7. Then taking the average of the sum.
   ◆ Obtaining the 70th percentile of the avg_rech_amt_6_7
   ◆ Customers who have recharged more than or equal to 10, Filter taht customers.

Data Shape = (30011, 178)

- ➔ Handling the Missing Values in Rows
  - ◆ Here we are checking various missing values in the columns and removing the columns accordingly to optimise the data for evaluation.
- ➔ Tag Churners
  - ◆ Presently label the beat clients (churn=1, else 0) in light of the fourth month as follows: The people who have not settled on any decisions (either approaching or friendly) AND have not utilized portable web even once in the stir stage.
  - ◆ Eliminating all the attributes corresponding to the phase in churn
  - ◆ Checking Churn Percentage = 3.39

- ➔ Outliers Treatment
  - ◆ In the separated dataset aside from mobile_number and stir segments every one of the sections are numeric sorts. Subsequently, changing over mobile_number and stir datatype to protest.
  - ◆ Removing the outliers which are below 10th and are above 90th percentile
  - ◆ Data Shape= (27705, 136)

➔ Taking in New features
 ◆ Deriving new column
    ● Decrease_mou_action
    ● Decrease_rech_num_action
    ● Decrease_rech_amt_action
    ● decrease_arpu_action
    ● decrease_vbc_action

# Exploratory Data Analysis (EDA)

Univariate Analysis

On the basis on the churn rate whether the customer decreased his MOU in action month

We can obviously see that the agitate rate is something else for the clients, whose minutes of usage(mou) diminished in the activity stage than the great stage.
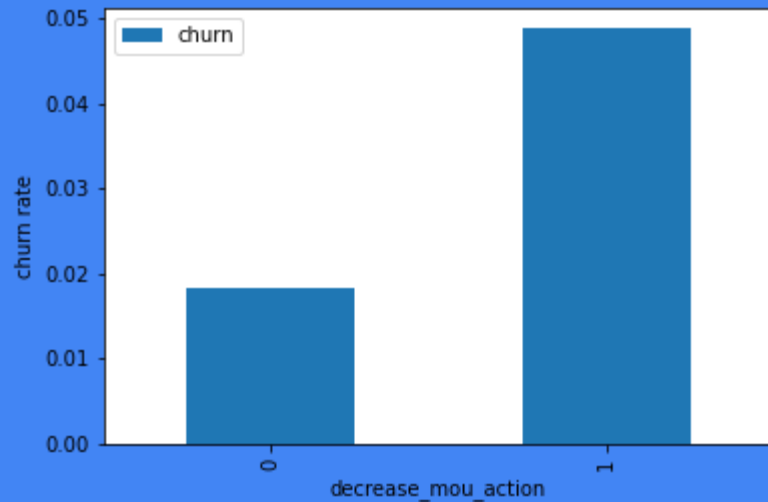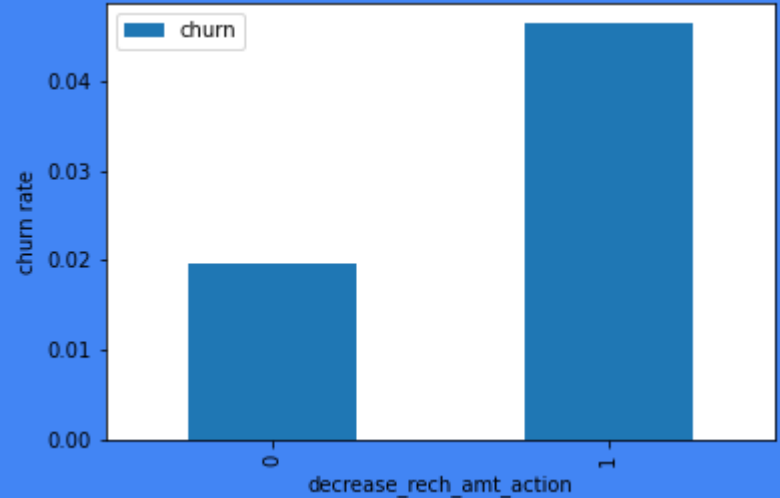
Agitate rate on the premise whether the client diminished his/her number of re-energize in real life month

True to form, the agitate rate is something else for the clients, whose number of re-energize in the activity stage is lesser than the number in great stage
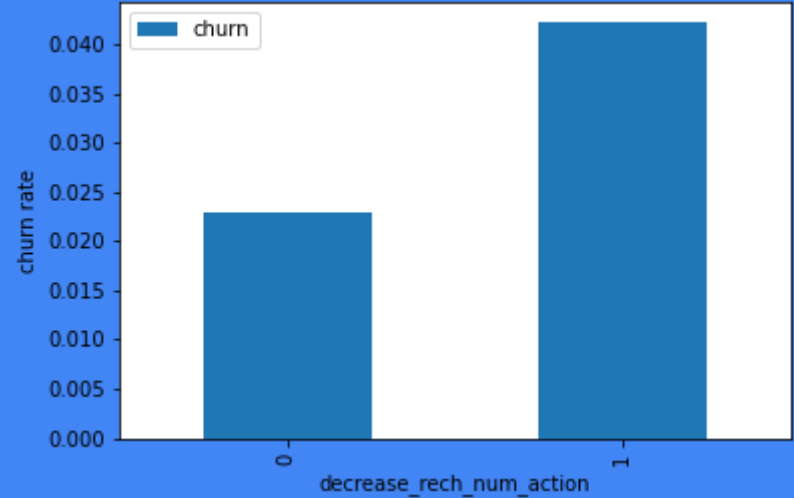
Agitate rate on the premise whether the client diminished his/her measure of re-energize in real life month

Here additionally we can plainly see that a similar way of behaving. The stir rate is something else for the clients, whose measure of re-energize in the activity stage is lesser than the sum in great stage

Churn rate on the premise whether the client diminished his/her volume based cost in real life month

The churn rate is something else for the clients, whose volume based cost in real life month is expanded. That implies the clients don't do the month to month re-energize more when they are in the activity stage. Here we can see the normal outcome.
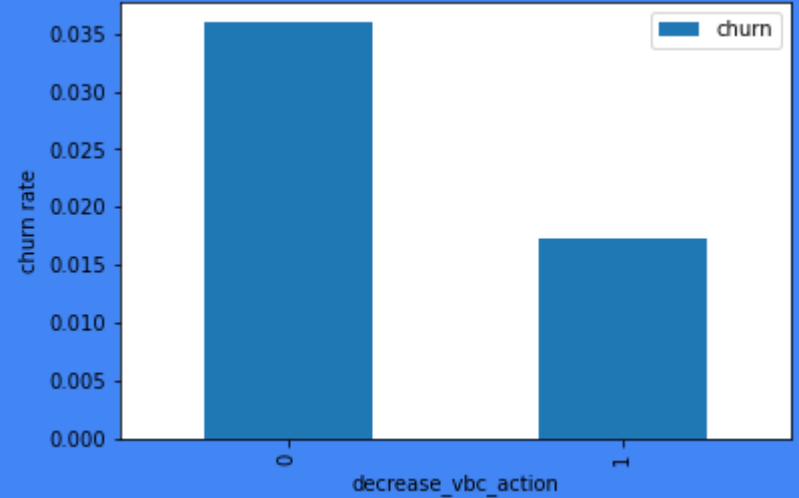
Examination of the typical income per client (churn and not churn) in the activity stage

Normal income per client (ARPU) for the churned clients is generally densed on the 0 to 900. The higher ARPU clients are more averse to be churned.

ARPU for the not churned clients is generally densed on the 0 to 1000.
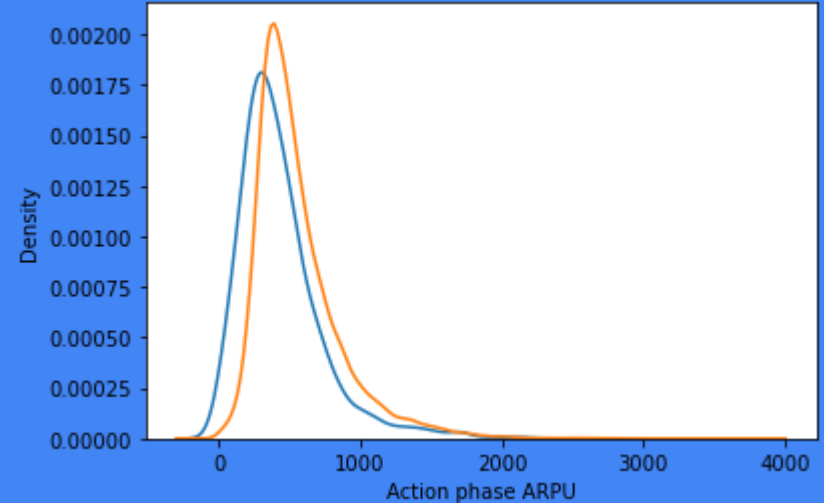
Examination of the minutes of use MOU (churn and not churn) in the activity stage

Examination of the minutes of use MOU (churn and not churn) in the activity stage

Minutes of usage(MOU) of the churn clients is for the most part populated on the 0 to 2500 territory. Higher the MOU, lesser the churn likelihood.
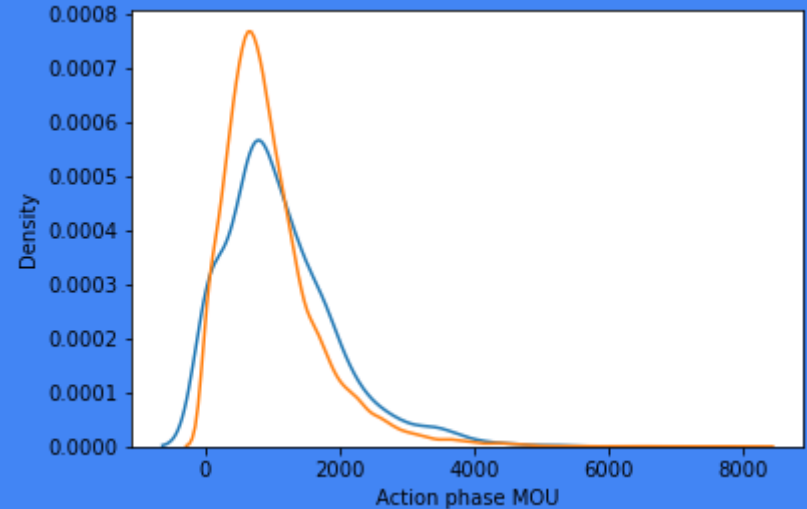
Bivariate analysis

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

As now we can see obviously from the above plot, that the churn rate is something else for the clients, whose re-energize sum as well as number of re-energize have diminished in the activity stage than the great stage.
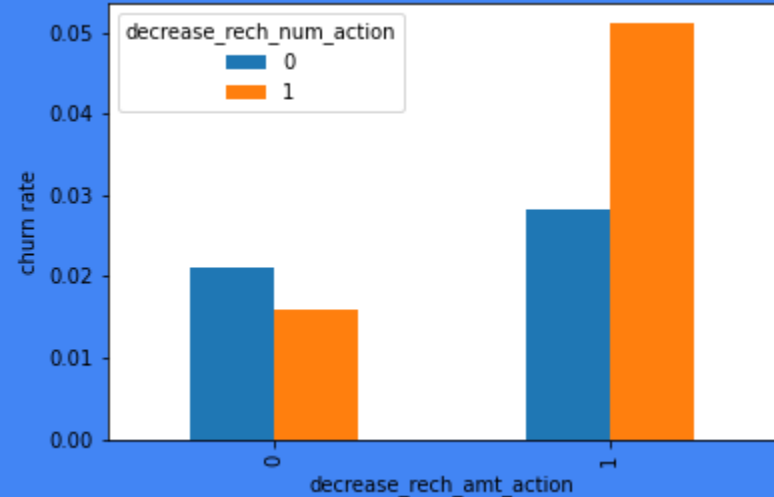
Further Examination of churn rate by the diminishing re-energize sum and volume based cost in the activity stage

Here, we can likewise see that the churn rate is something else for the clients, whose re-energize sum is diminished alongside the volume based cost is expanded in the activity month.
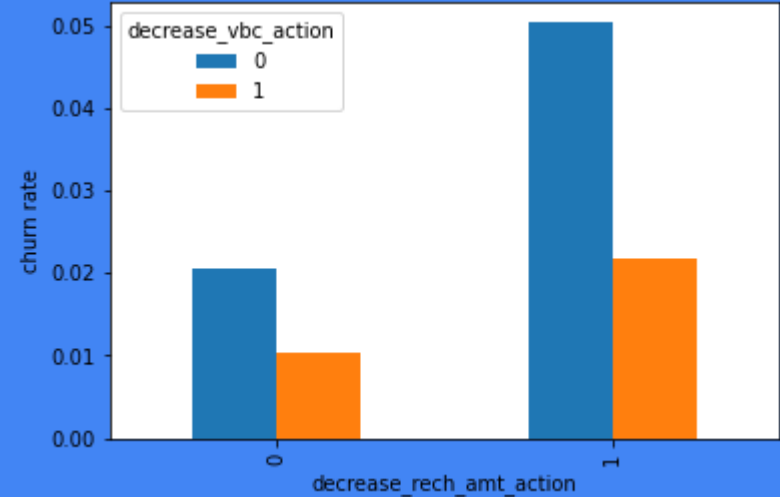
Examination of re-energize sum and number of re-energize in real life month

Presently we can obviously see from the above design that the re-energize number and the re-energize sum are generally proportional. More the quantity of re-energize, more how much the re-energize.
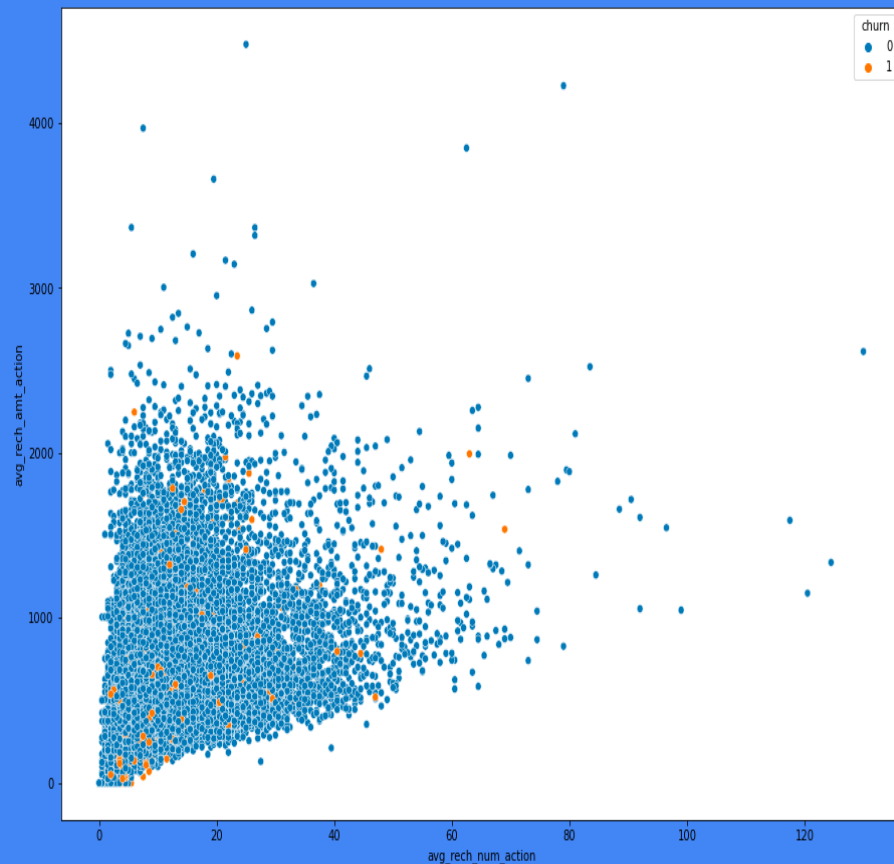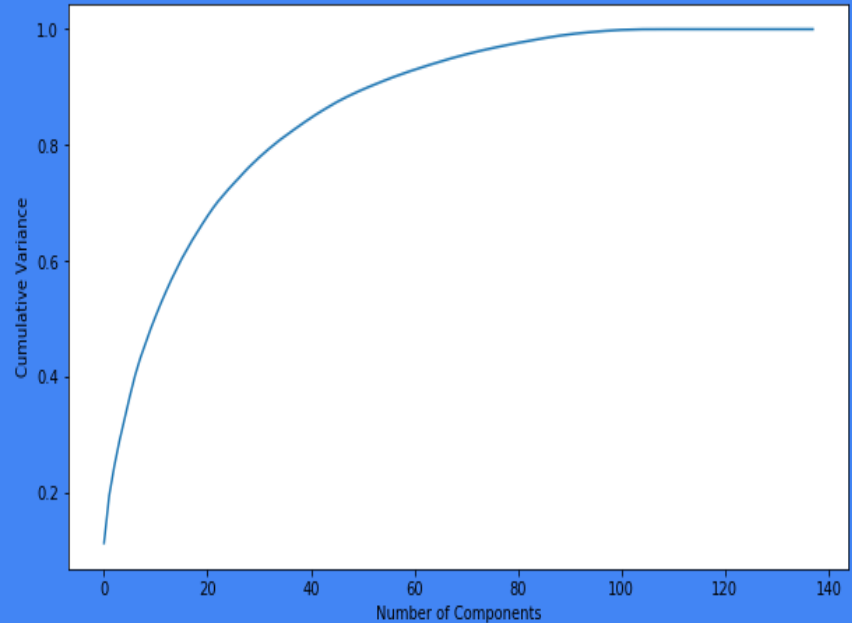
Train-Test Split

Dealing with data imbalance

We are making engineered tests by doing upsampling utilizing MOTE(Synthetic Minority Oversampling Technique).

PCA Model

We can see that 60 parts make sense of amost over 90% difference of the information. Along these lines, we will perform PCA with 60 parts.

Applying transformation on the test set

We are only doing Transform in the test set not the Fit-Transform. Because the Fitting is already done on the train set. So, we just have to do the transformation with the already fitted data on the train set.

Emphasize Sensitivity/Recall than Accuracy

We are more centered around higher Responsiveness/Review score than the exactness.

Because we really want to think often more about beat cases than the not stir cases. The primary objective is to retain the clients, who have the possibility to stir. There ought not be an issue, in the event that we think about not many not stir clients as stir clients and give them a few motivating forces to holding them. Thus, the awareness score is more significant here.

Logistic regression with PCA

Tuning hyperparameter C

C is the opposite of regularization strength in Strategic Relapse. Higher upsides of C relate to less regularization.

The highest test sensitivity is 0.8978916608693863 at C = 100
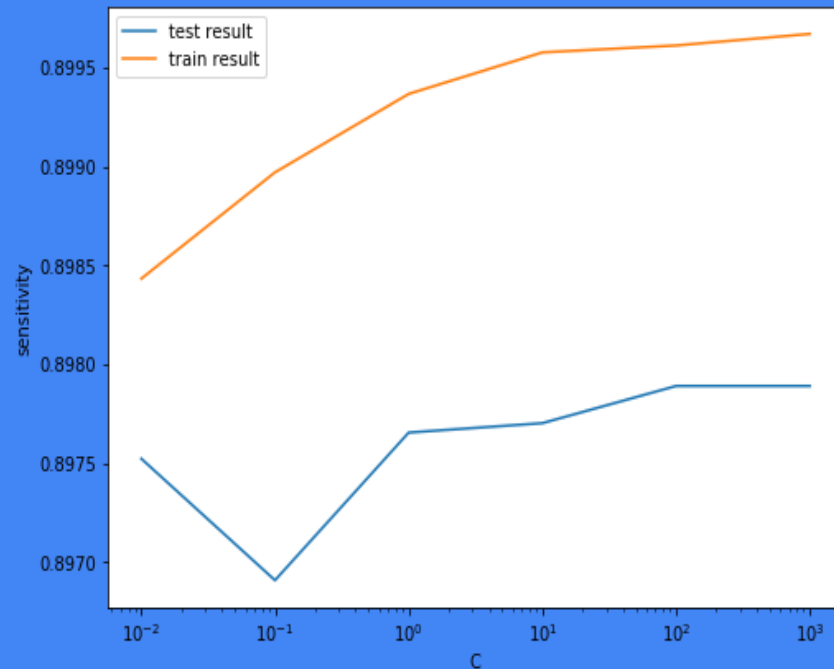
Model summary

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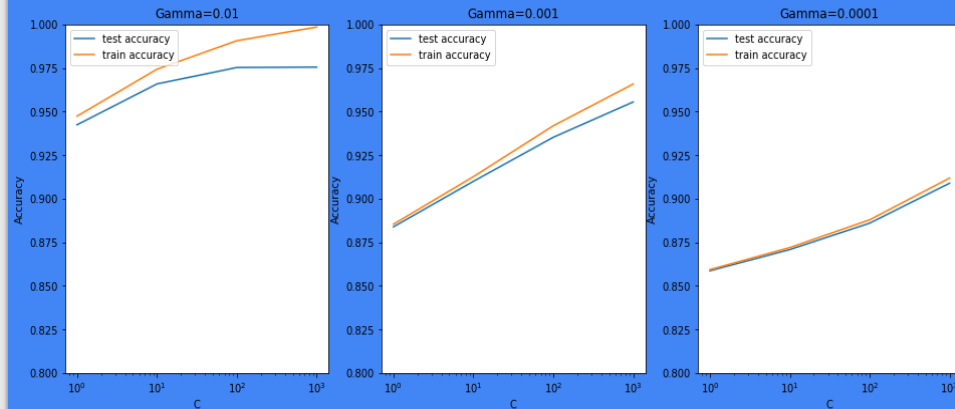- Sensitivity = 0.89
- Specificity = 0.83

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

Plotting the accuracy with various C and gamma values

The best test score is 0.9754959911159373 corresponding to hyperparameters {'C': 1000, 'gamma': 0.01}

- From the above plot, we can see that higher worth of gamma prompts overfitting the model. With the least worth of gamma (0.0001) we have train and test precision practically same.
- Likewise, at C=100 we have a decent exactness and the train and grades are equivalent.
- However sklearn proposes the ideal scores referenced above (gamma=0.01, C=1000), one could contend that it is smarter to pick a less complex, more non-straight model with gamma=0.0001. This is on the grounds that the ideal qualities referenced here are determined in view of the typical test precision (yet not considering abstract boundaries like model intricacy).
- We can accomplish practically identical normal test exactness (~90%) with gamma=0.0001 also, however we'll need to build the expense C for that. So to accomplish high exactness, there's a tradeoff between:
- High gamma (for example high non-linearity) and normal worth of C
- Low gamma (for example less non-linearity) and high worth of C
- We contend that the model will be less complex on the off chance that it has as less non-linearity as could really be expected, so we pick gamma=0.0001 and a high C=100.

Build the model with optimal hyperparameters

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

We can see from the model presentation that the Sensitivity has been diminished while assessing the model on the test set. In any case, the precision and explicitness is very great in the test set.

Model summary*

- Train set
    - Accuracy = 0.90
    - Sensitivity = 0.91
    - Specificity = 0.88
- Test set
    - Accuracy = 0.86
    - Sensitivity = 0.70
    - Specificity = 0.87

Random forest with PCA

We can see from the model presentation that the Sesitivity has been diminished while assessing the model on the test set. Notwithstanding, the precision and particularity is very great in the test set.

Model summary*

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

Final conclusion with PCA

In the wake of attempting a few models we can see that for achieving the best responsiveness, which was our definitive objective, the exemplary Strategic relapse or the SVM models performs well. For both the models the responsiveness was approx 81%. Likewise we have great precision of approx 85%.

Without PCA

Model analysis

We can see that there are not many highlights have positive coefficients and few have negative.

Many highlights have higher p-values and consequently became immaterial in the model.

Feature Selection Using RFE

Model-3

Presently from the model outline and the VIF list we can see that every one of the factors are huge and there is no multicollinearity among the factors.

Consequently, we can conclude that *Model-3 log_no_pca_3 will be the last model*.

Model-1 with RFE selected columns

Eliminating segment og_others_8, which is insignificant as it has the most noteworthy p-esteem 0.99.

Model-2

As we can see from the model rundown that every one of the factors p-values are huge and offnet_mou_8 section has the most elevated VIF 7.45. Subsequently, erasing offnet_mou_8 segment.

Model performance on the train set

Examination of the above bend
Accuracy - Becomes steady around 0.6

Sensitivity - Diminishes with the expanded probability.

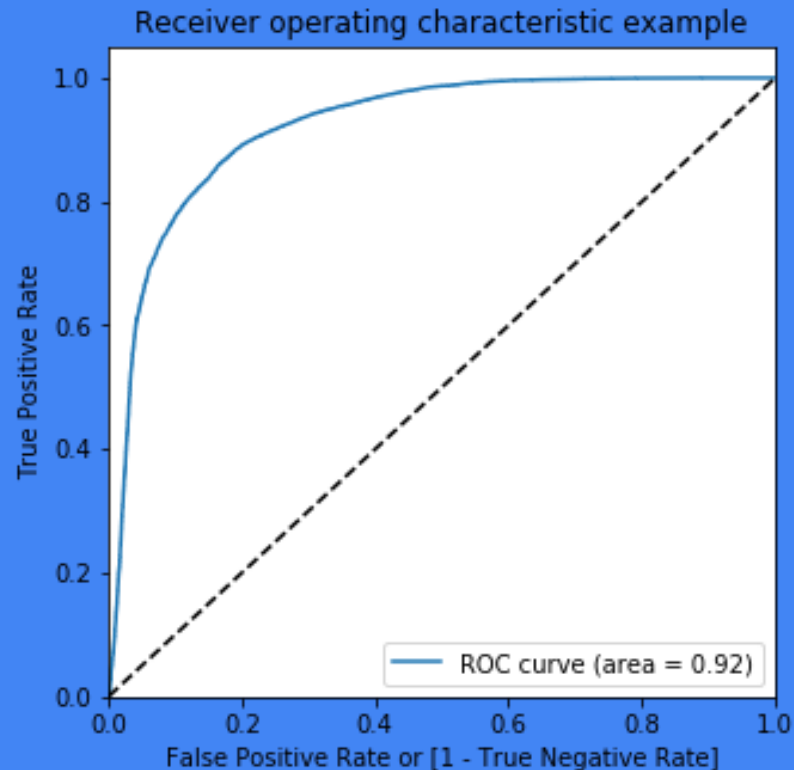Specificity - Increments with the rising probablity.

At point 0.6 where the three boundaries cut one another, we can see that there is an equilibrium between responsiveness and particularity with a decent precision.

Here we are expected to acheive preferred awareness over precision and particularity. However according to the above bend, we ought to take 0.6 as the ideal likelihood cutoff, we are taking *0.5* for achieving higher responsiveness, which is our fundamental objective.
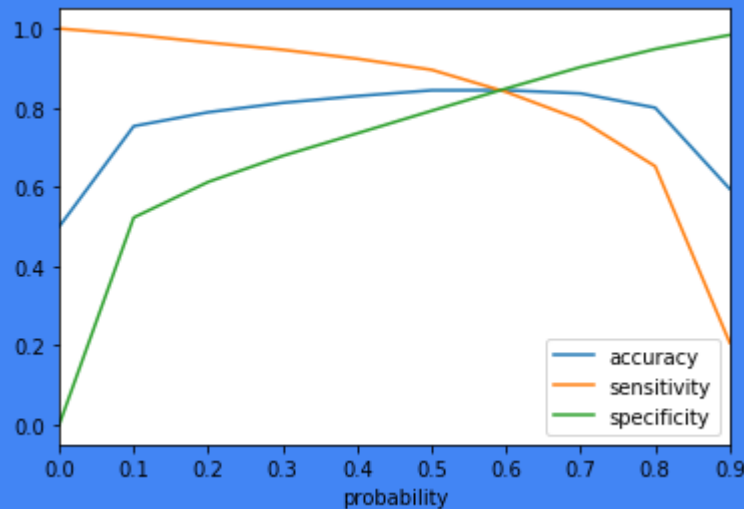
Plotting the ROC Curve (Trade off between sensitivity & specificity)

We can see the region of the ROC bend is more like 1, which is the Gini of the model.

Receiver operating characteristic example

True Positive Rate vs False Positive Rate or [1 - True Negative Rate]

ROC curve (area = 0.92)

Testing the model on the test set

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 learnt from the train set.

Final conclusion with no PCA

We can see that the strategic model with no PCA has great responsiveness and exactness, which are practically identical to the models with PCA. Thus, we can go for the more oversimplified model, for example, calculated relapse with PCA as it explains the significant indicator factors as well as the meaning of every variable. The model additionally helps us to distinguish the factors which ought to be follow up on for settling on the choice of the to be stirred clients. Consequently, the model is more pertinent as far as clarifying for the business.

Business Recomendation

- Focus on the clients, whose minutes of use of the approaching neighborhood calls and active ISD calls are less in the activity stage (for the most part in the period of August).
- Focus on the clients, whose active others charge in July and approaching others on August are less.
- Additionally, the clients having esteem based cost in the activity stage expanded are bound to beat than different clients. Consequently, these clients might be a decent objective to give offer.
- Customers, whose month to month 3G re-energize in August is more, are probably going to be agitated.
- Clients having diminishing sexually transmitted disease approaching minutes of use for administrators T to fixed lines of T for the period of August are bound to agitate.
- Customers diminishing month to month 2g utilization for August are generally plausible to agitate.
- Clients having diminishing approaching minutes of use for administrators T to fixed lines of T for August are bound to agitate.
- roam_og_mou_8 factors have positive coefficients (0.7135). That implies for the clients, whose meandering active minutes of use is expanding are bound to stir.

Plots of important predictors for churn and non churn customers

We can see that for the beat clients the minutes of use for the period of August is generally populated on the lower side than the non stir clients.
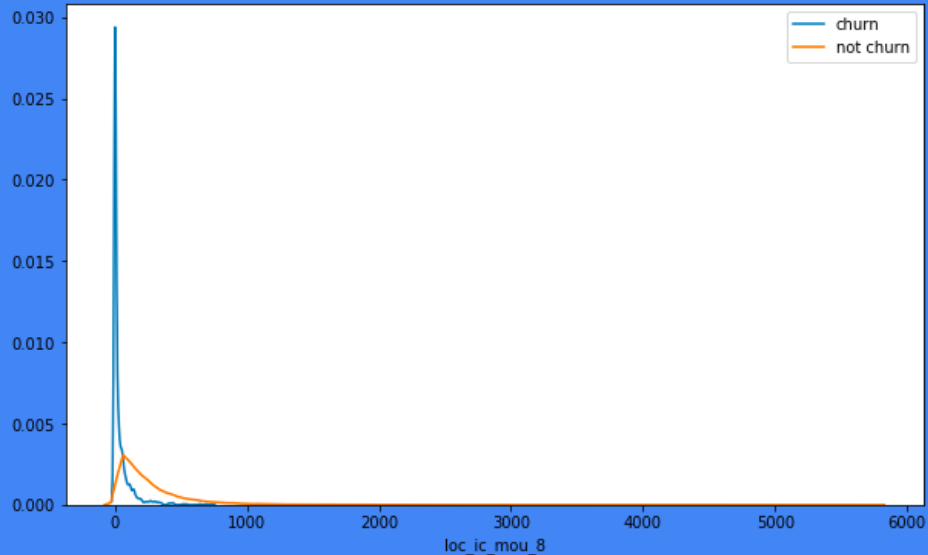
We can see that for the beat clients the minutes of use for the long stretch of August is generally populated on the lower side than the non stir clients.
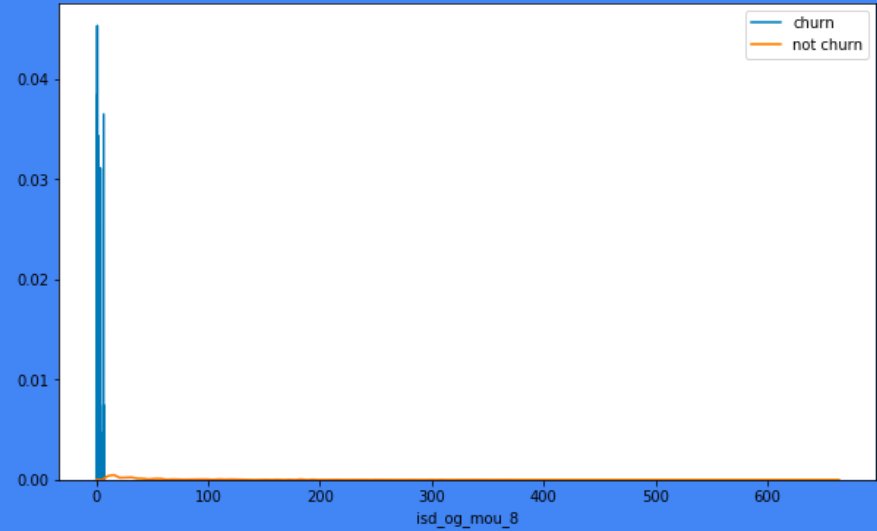
The quantity of mothly 3g information for August for the stir clients are especially populated aroud 1, though of non agitate clients it spreaded across different numbers.

Comparably we can plot every factors, which have higher coefficients, stir appropriation.