Telecom Churn

Case Study

Steps for Model Building

- Reading, understanding and visualising the data
- Preparing the data for modelling
- Building the model
- Evaluate the model

Understanding Data

The dataset contains customer-level information for a span of four consecutive months - June, July, August and September. The months are encoded as 6, 7, 8 and 9, respectively.

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.

Data Dictionary File

Acronyms	Descriptions
MOBILE_NUMBER	Customer phone number
CIRCLE_ID	Telecom circle area to which the customer belongs to
LOC	Local calls - within same telecom circle
STD	STD calls - outside the calling circle
IC	Incoming calls
OG	Outgoing calls
T2T	Operator T to T, i.e. within same operator (mobile to mobile)
T2M	Operator T to other operator mobile
T2O	Operator T to other operator fixed line
T2F	Operator T to fixed lines of T
T2C	Operator T to it's own call center
ARPU	Average revenue per user
MOU	Minutes of usage - voice calls
AON	Age on network - number of days the customer is using the operator T network
ONNET	All kind of calls within the same operator network
OFFNET	All kind of calls outside the operator T network

The above image is the sample of the data dictionary which contains the 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.

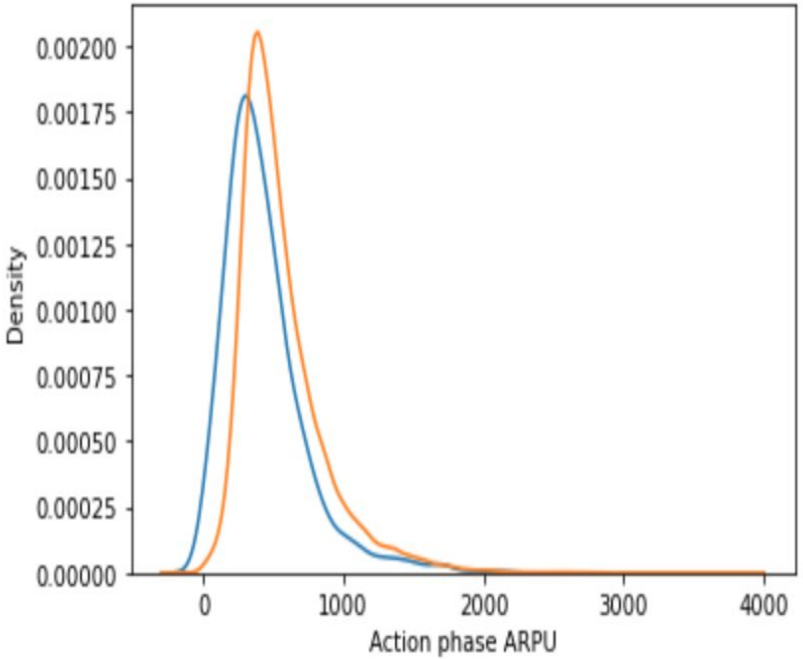
The attributes contains 6, 7, 8, 9 as suffixes implies that those correspond to the months 6, 7, 8, 9 respectively.

Data Preparation

The following data preparation steps are crucial for this problem:

1. **Filter high-value customers**
2. **Tag churners and remove attributes of the churn phase**

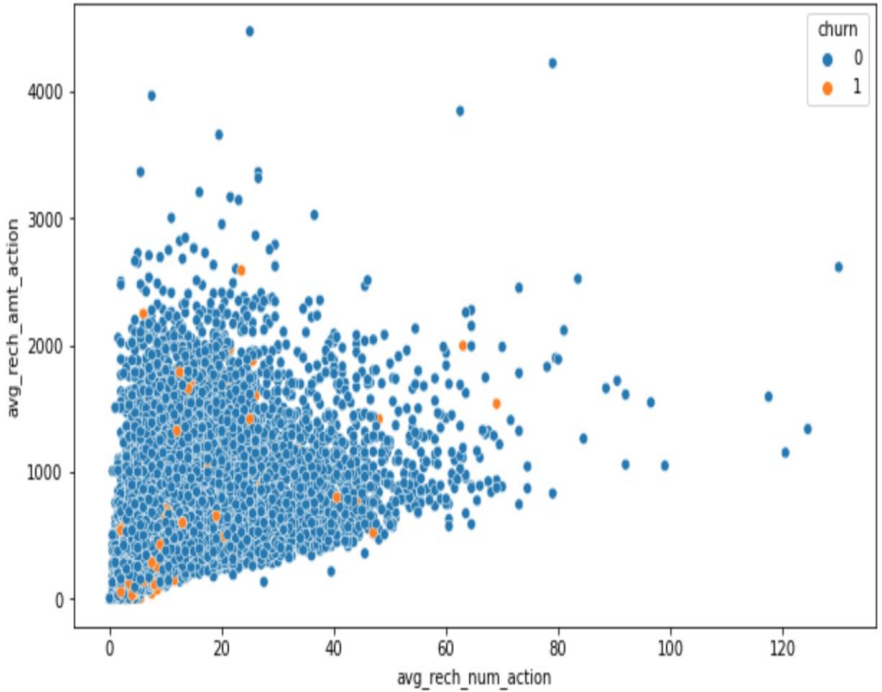
Distribution Plot



Average revenue per user (ARPU) for the churned customers is mostly densed on the 0 to 900.

The higher ARPU customers are less likely to be churned.

Scatterplot



We can see from the above pattern that the recharge number and the recharge amount are mostly proportional.

More the number of recharge, more the amount of the recharge.

Without PCA (Logistic regression with No PCA)

Through this step, we got to know that -

- We know that there are few features which have positive coefficients and few have negative.

- Many features have higher p-values and hence became insignificant in the model.

Final conclusion with PCA

After trying several models we can say that for achieving the best sensitivity, which was our ultimate goal, the classic Logistic regression or the SVM models performs well.

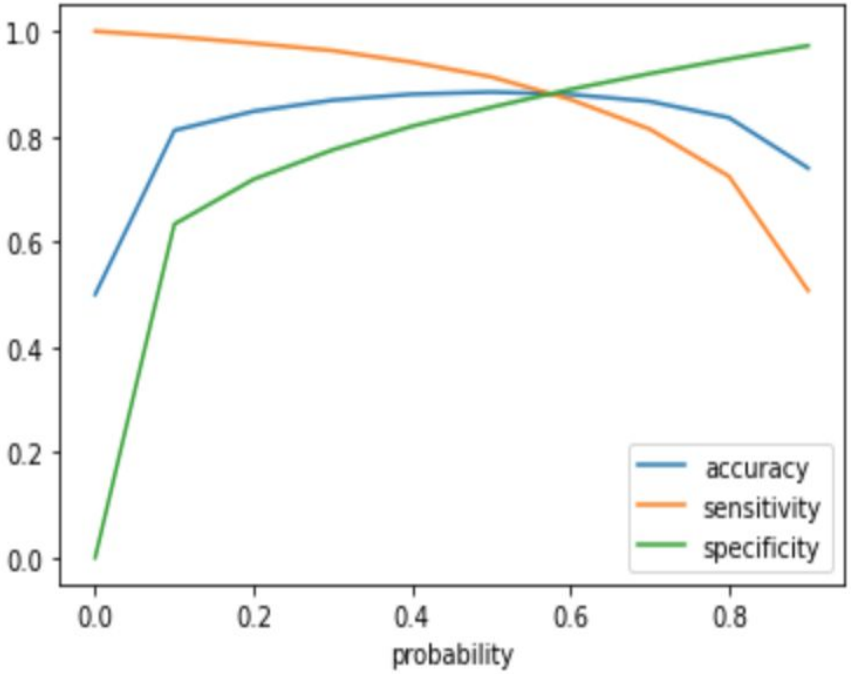
For the models the sensitivity was approximately 81%. Also we have good accuracy of approximately 85%.

Plotting (Accuracy, Sensitivity and Specificity) for different probabilities.

Analysis of the probability curve

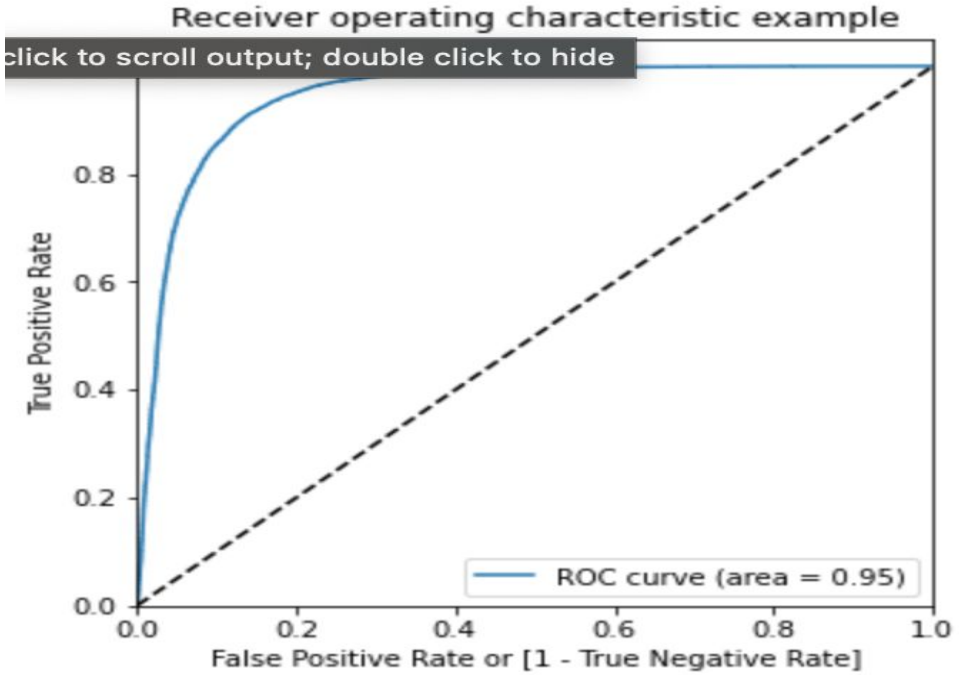
- Accuracy - Becomes stable around 0.6
- Sensitivity - Decreases with the increased probability.
- Specificity - Increases with the increasing probability.

At point 0.6 where the three parameters cut each other, we saw that there is a balance between sensitivity and specificity with a good accuracy.



ROC Curve

We saw the area of the ROC curve is closer to 1, which is the Gini of the model



Important predictors for churn and non churn customers

We can say that for the churn customers the minutes of usage for the month of August is mostly populated on the lower side than the non churn customers.

We can see that the ISD outgoing minutes of usage for the month of August for churn customers is dimmed to approximately 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 are very much populated around 1, whereas of non churn customers are spreaded across various numbers.

Similarly we can plot each variables, which have higher coefficients, churn distribution.

Recommendations

1. Target the customers, whose minutes of usage of the incoming local calls and outgoing ISD calls are less in the action phase (mostly in the month of August).
2. Target the customers, whose outgoing others charge in July and incoming others on August are less.
3. The customers having value based cost in the action phase increased are more likely to churn than the other customers. Hence, these customers may be a good target to provide offer.
4. Customers, whose monthly 3G recharge in August is more, are likely to be churned.
5. Customers having decreasing STD incoming minutes of usage for operators T to fixed lines of T for the month of August are more likely to churn.
6. Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.
7. roam_og_mou_8 variables have positive coefficients (0.7135). That means for the customers, whose roaming outgoing minutes of usage is increasing are more likely to churn.