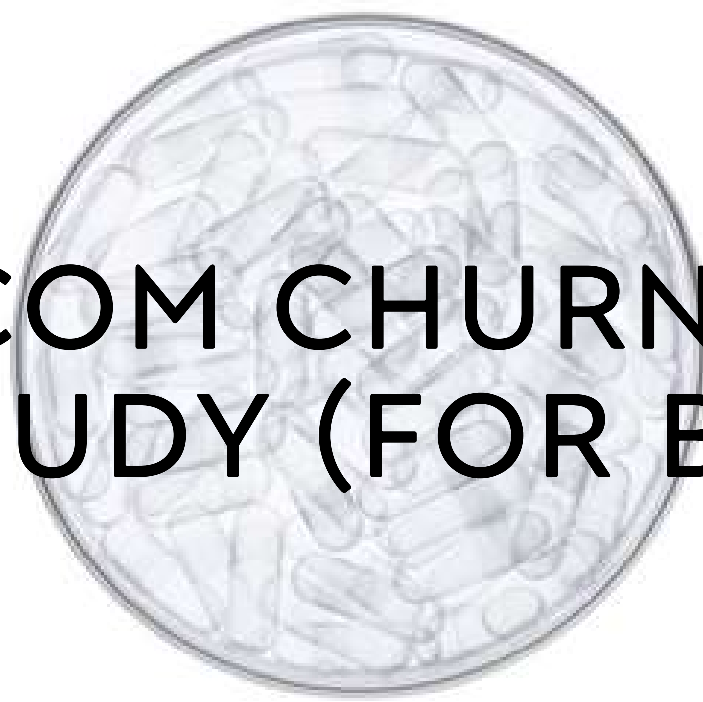


TELECOM CHURN CASE STUDY (FOR BA)

INTRODUCTION AND PROBLEM STATEMENT

- Churn consists of detecting customers who are likely to cancel a subscription to a service. churn is a problem for telecom companies because it is more expensive to acquire a new customer than to keep an existing one from leaving.

- 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.

OVERVIEW OF CASE STUDY

- Identifying Factors Contributing To Churn Via Exploratory Data Analysis (EDA)
- Data Preparation For Model Building
- Building A Prediction Model And Evaluating The Model
- Provide Insights And Recommendation On Churn And Customer Retention

EXPLORATORY DATA ANALYSIS (EDA)

- Pre-processed telecom data has 99999 rows and 226 columns.
- Columns with missing values were evaluated and columns having more than 30% missing data were dropped from analysis.
- Columns with single values and providing no insights to analysis were also dropped from analysis.
- Remaining rows and columns were used to focus on High Value Customers (HVC) (HVC are the ones 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).
- Based on the above HVC definition, 30000 rows were filtered for these customers

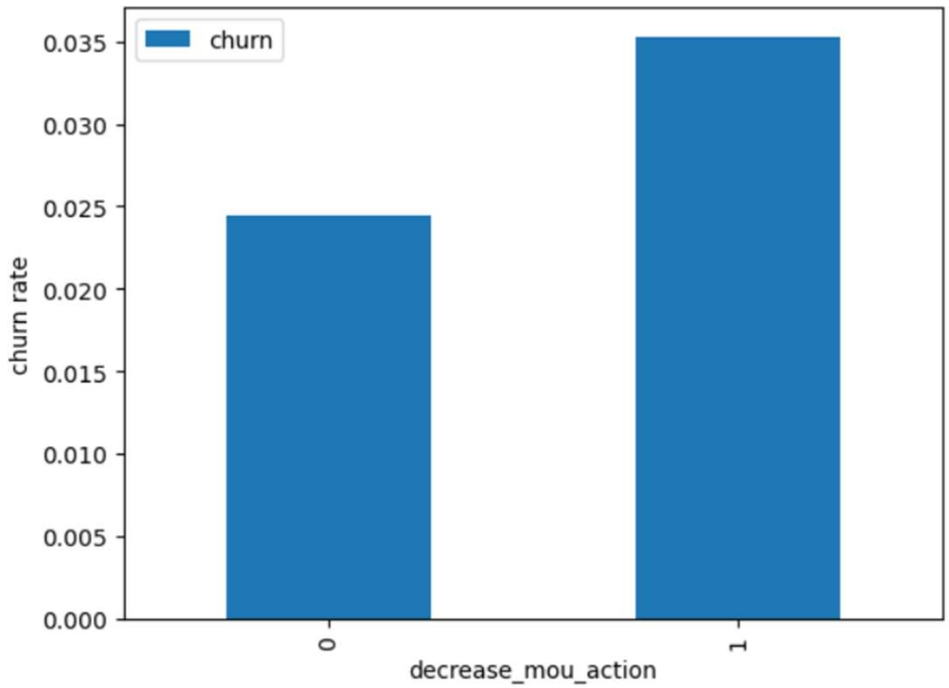
EXPLORATORY DATA ANALYSIS (EDA)

- Upon further processing of these 30k rows for HVC, further columns with missing values were dropped and a final data with rows close to 28k were used to analyse the churn.
- For this data, the new columns were derived for further analysis in the churn phase.
- New columns were based on minutes of usage, recharge amount and average revenue per customer.

EXPLORATORY DATA ANALYSIS (EDA)

- Univariate analysis

(churn=1, non-churn=0)



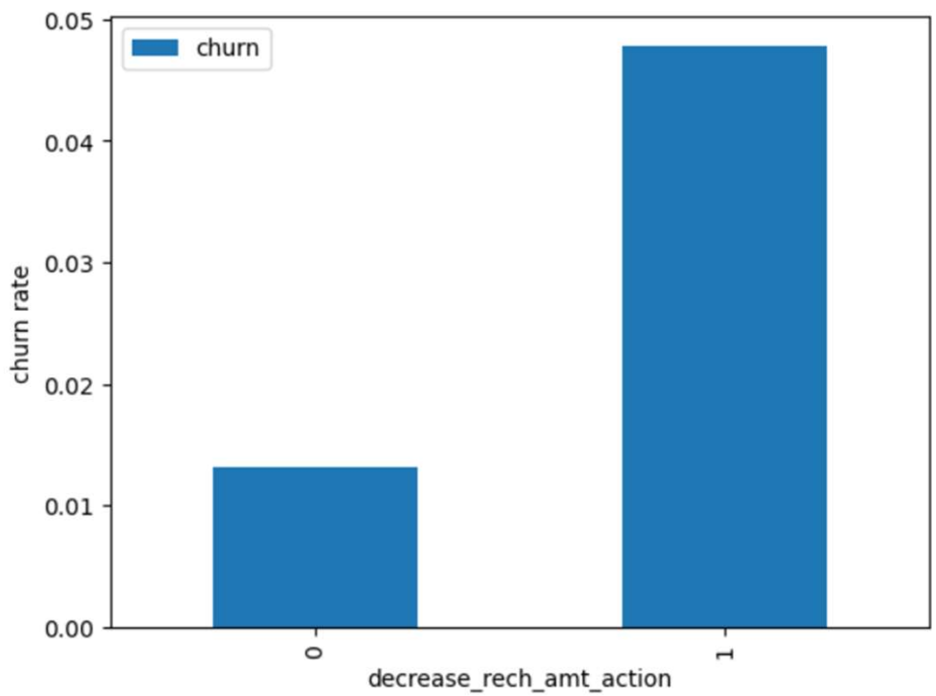
- Observation:

- Customers with higher churn rate have decreased minutes of usage

EXPLORATORY DATA ANALYSIS (EDA)

- Univariate analysis

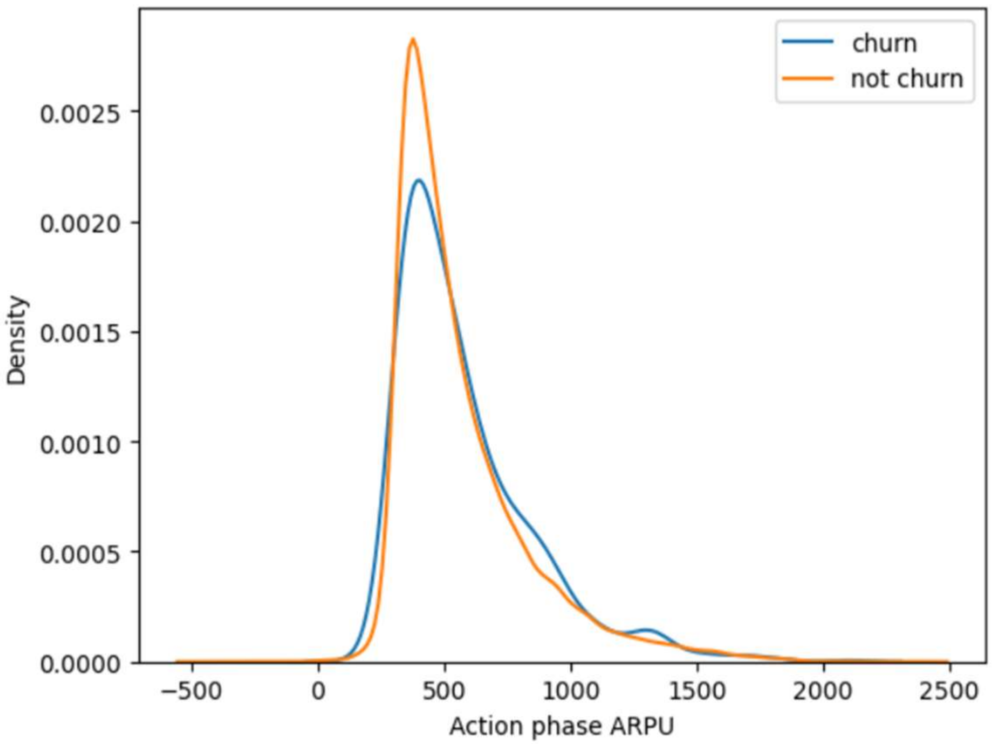
(churn=1, non-churn=0)



- Observation:

- Customers with higher churn rate have reduced the recharge amount in action phase compared to the good phase.

EXPLORATORY DATA ANALYSIS (EDA)



- Analysis of the average revenue per user (ARPU) in action phase:

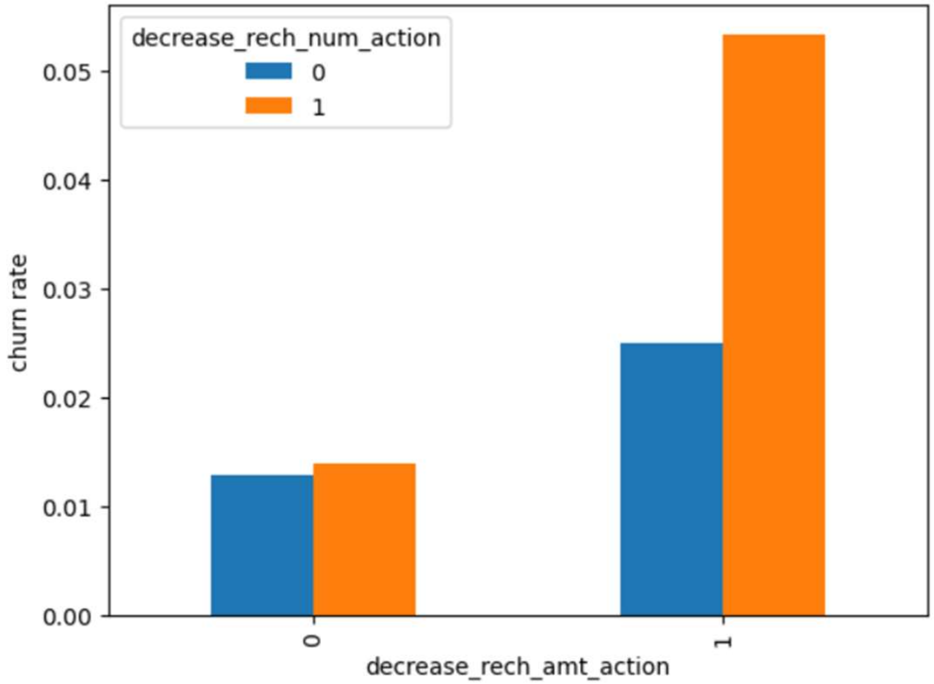
ARPU for the churned customers is mostly denser on the 0 to 900. The higher ARPU customers are less likely to be churned.

ARPU for the non-churned customers is mostly denser on the 0 to 1200.

EXPLORATORY DATA ANALYSIS (EDA)

- Bivariate analysis (Recharge amt and number of recharge)

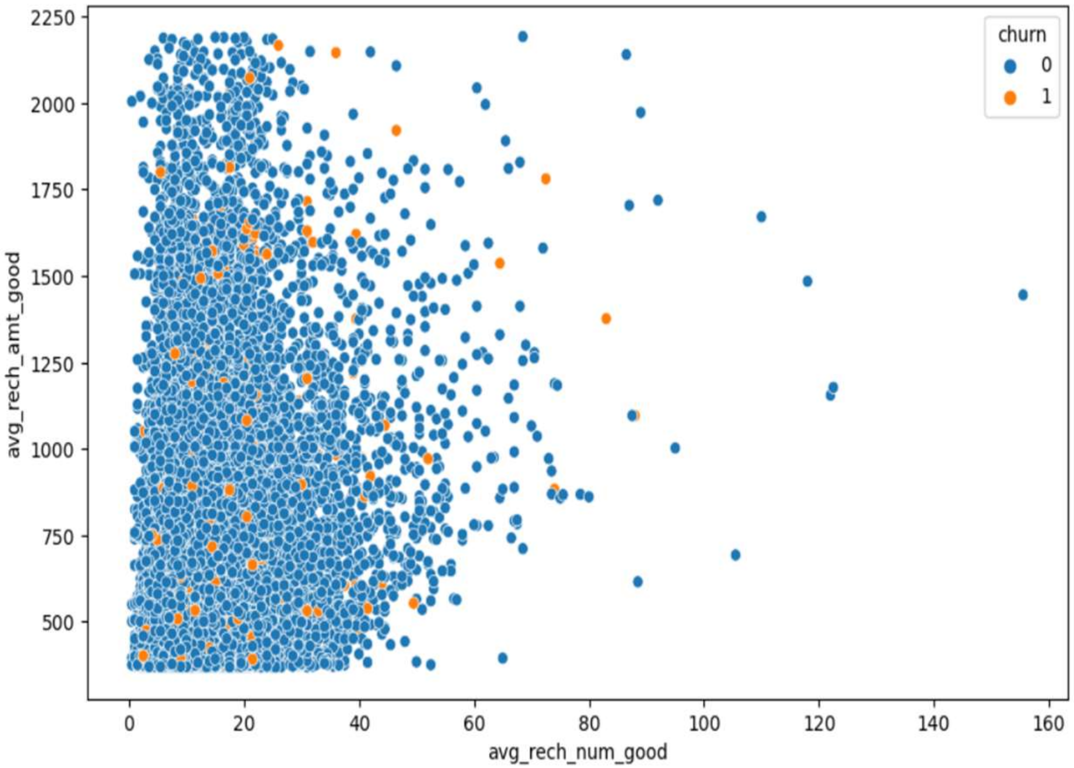
(churn=1, non-churn=0)



- Observation:

- 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 (EDA)



- Analysis of the average recharge amount vs number of recharge

The scatter plot shows a more proportional distribution indicating more number of recharges have more amount of recharge.

DATA PROCESSING FOR BUILDING A MODEL

- Further cleaning the data to remove unique values and columns which do not aid in churn prediction models
- Splitting data into Train-test data set (80:20)
- Data imbalance was noted which was addressed by using SMOTE (Synthetic Minority Oversampling Technique)
- Feature scaling was done by importing standardscaler from sklearn.preprocessing library.

MODEL BUILDING

- Logistic Regression with RFE (Recursive Feature Elimination) was used to build a suitable model for churn prediction
- Model 1 –Variance Inflation Factor (VIF) and p-value were evaluated

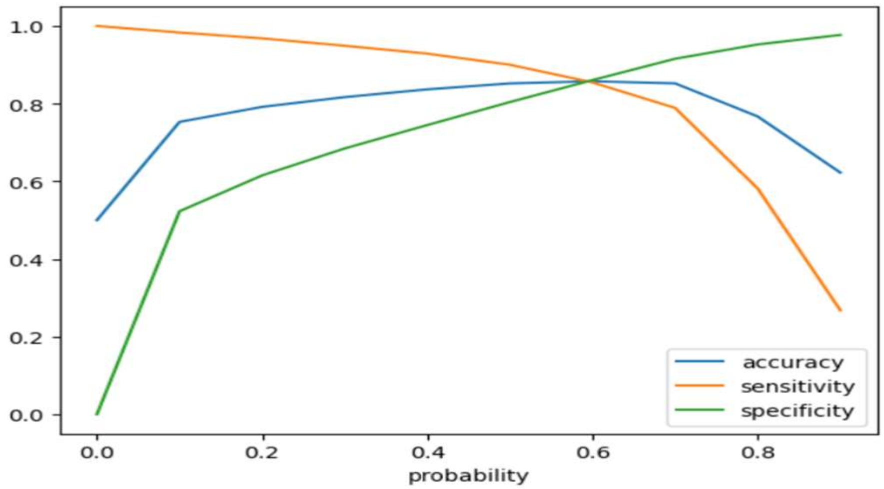
  Based on the VIF and high p-values “og_others_8” was dropped.
- Model 2 – Model without the “og_others_8” variable

The VIF value for offnet_mou_8 showed 10.11 (greater than cut off of 5), hence we can remove the feature as it has multicollinearity with other features.

MODEL BUILDING

- Model 3 –Model post removal of "offnet_mou_8" variable

The VIF and p-values were significant and showed no multicollinearity among the variables. Thus, model-3 was finalized for further churn prediction using train data.



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 intersect each other, we can see that there is a balance between sensitivity and specificity with a good accuracy.

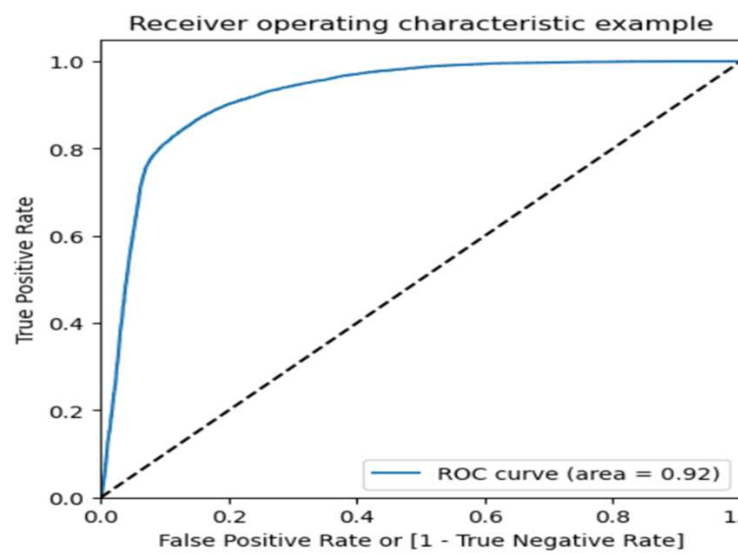
Here we are intended to achieve better sensitivity than accuracy and specificity. Although we should be considering 0.6 as per the above curve as optimal probability cut-off, we are considering 0.5 to achieve higher sensitivity.

CONFUSION MATRIX

- For the train set data the confusion matrix showed a good accuracy, sensitivity and specificity based on the model-3 prediction.

Accuracy:- 0.8520420070011668
Sensitivity:- 0.9000700116686114
Specificity:- 0.8040140023337223

- Also, the Receiver Operating Characteristic (ROC) curve was closer to 1.



MODEL SUMMARY

- The model was run for the test data:
- Performance was good with similar results as seen in the training data with details as seen below.
- The model provided a brief insight on the top variables effecting the churn rate, thus we were able to draw conclusion on the churn probability for this dataset.

Model Summary

Train Set :

Accuracy:- 85.20%

Sensitivity:- 90%

Specificity:- 80.40%

Test set :

Accuracy:- 80%

Sensitivity:- 83.93%

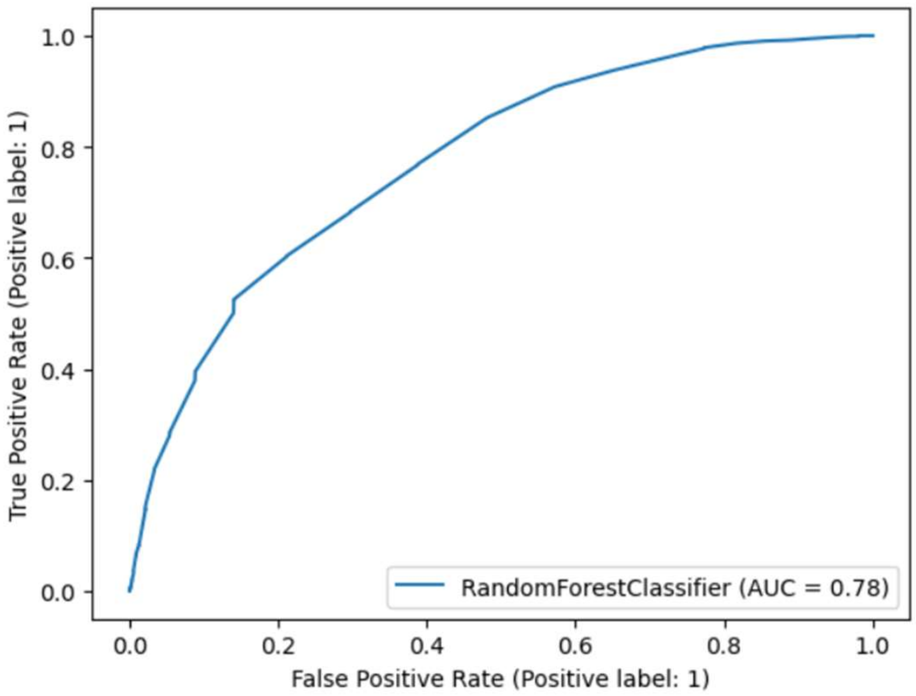
Specificity:- 79.88%

- From business point of view, this model is very easy to interpret and target the churn cases based on these 13 variables (as noted in python notebook) mainly focusing on August month.

RANDOM FOREST CLASSIFICATION

- The important features were drawn, however the churn prediction rate was not known for these features .

- The are under the curve (AUC) was close to 0.8.



INFERENCES

- Minutes of Usage (MoU):

For Off-network MoU, Roaming outgoing MoU, Std outgoing t2t (same network)/t2m (to other mobile networks) MoU : for all these as the MoU increased, the probability to churn also increases.

For Incoming MoU/International Outgoing/Outgoing other networks: if the MoU has reduced, the probability to churn increases.

- Internet Data packages:

For 2g/3g usage: If the usage had reduced, the probability of the churn increases.

CONCLUSION

Based on the inferences taken from the good and action phase months.

- Telecom vendor needs to provide better packages in outgoing and internet (2g/3g) packages
- Reduce charges on the incoming and promote more usage of their network
- Stay competitive by providing attractive plans and providing long term plans of 6months or an annual plan to customers in the good phase itself to avoid churn

THANK YOU