TELECOM CHURN CASE STUDY

SHRAVANI R

SANDHYA

DHARMVIR

THE PROBLEM

What is the problem?	• To find customers who are about to churn from a telecom operator
Who has this problem?	• Telecom Operator
Why should this problem be solved?	• Cost of acquiring new customer is 5-10% more than retaining existing customers
How will I know this problem has been solved?	• By calculating the churn rate

BACKGROUND INFORMATION

Business problem overview

1. In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. 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.

2. For many incumbent operators, retaining high profitable customers is the number one business goal and they have a churn rate of around 15-25% annually.

3. So, for telecom operators retaining existing customers has become more important than acquiring new customers.

4. To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

5. In this project, we 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.

DEFINITIONS OF CHURN

Revenue-based churn:

- Customers who have not utilised 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’.

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.

STEPS

Step 1 :

- Data reading
- Data Understanding
- Data Cleaning
- Imputing missing values

Data Preparation (Step 2 & 3)

Step-2 :

- Need to Filter high value customers

Step-3 :

- Derive churn
- Need to Derive the Target Variable

Step-4 :

- Derived variable
- EDA
- Split data in to train and test sets
- Performing Scaling

Step-5 :

- Handling class imbalance
- Dimensionality Reduction using PCA
- Classification models to predict Churn (Use various Models)

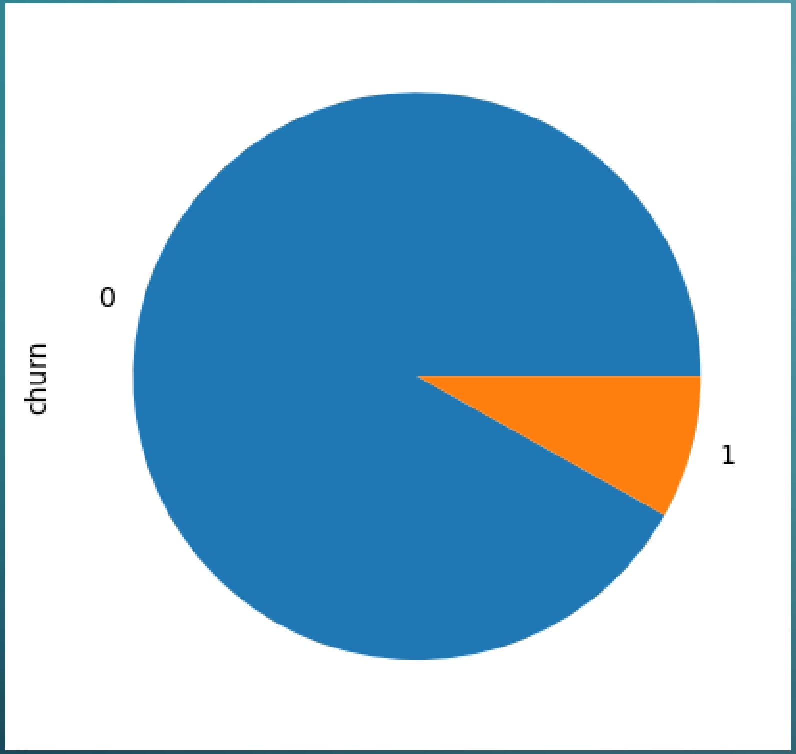
Step-6 :

- Model Evaluation
- Prepare Model for Predictor variables selection (Prepare multiple models & choose the best one)

There are three phases of customer lifecycle :

- The ‘good’ phase [Month 6 & 7]
- The ‘action’ phase [Month 8]
- The ‘churn’ phase [Month 9]
- In this case, since we are working over a four-month window, the first two months are the ‘good’ phase[6 & 7], the third month is the ‘action’ phase[8], while the fourth month is the ‘churn’ phase[9].

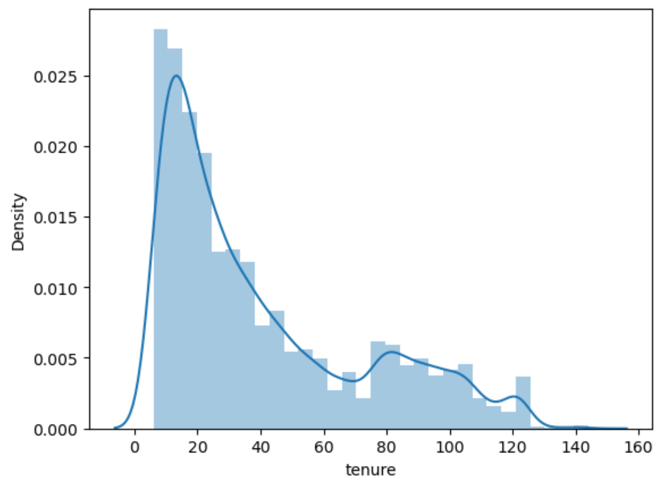
CHURN / NON CHURN Percentage

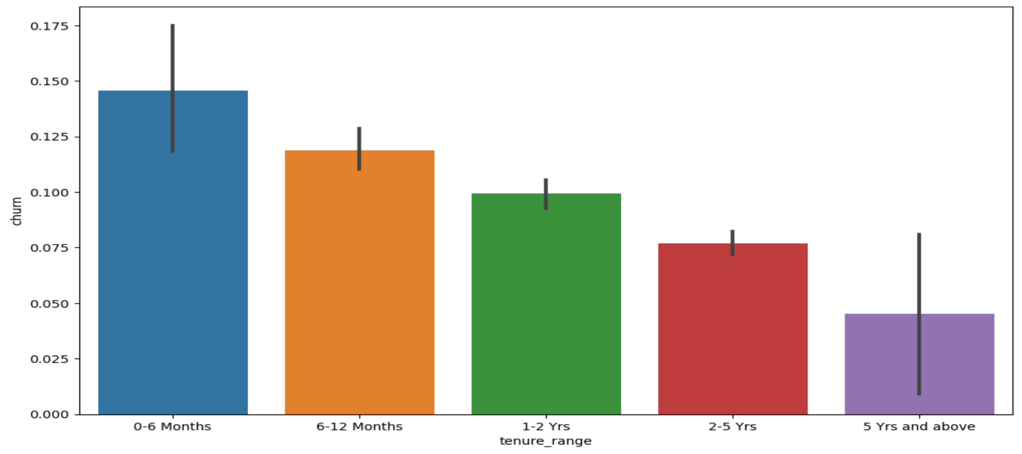


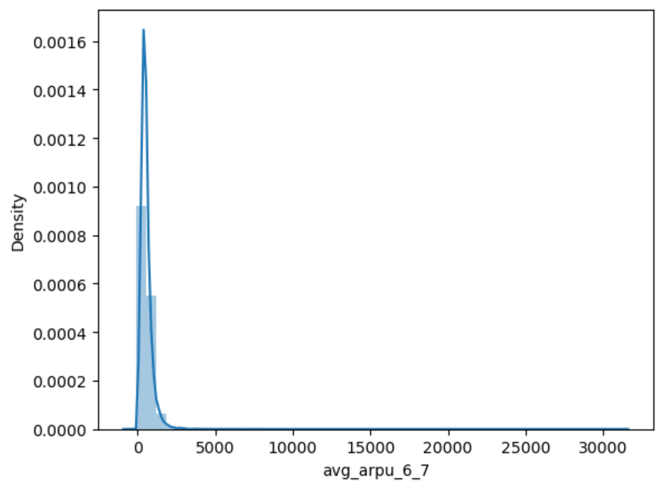
As we can see that the churn percentage is very less , we can assume that most of the cutomers do not churn.

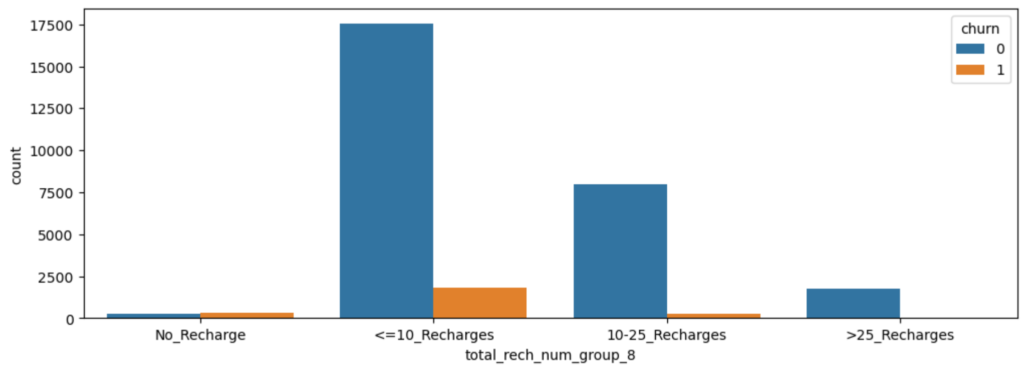
This typically shows a class imbalance.

EDA



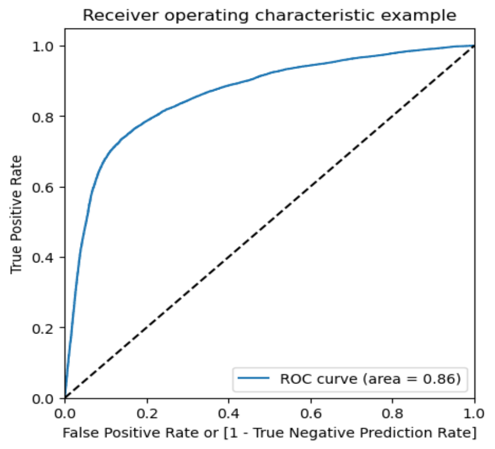


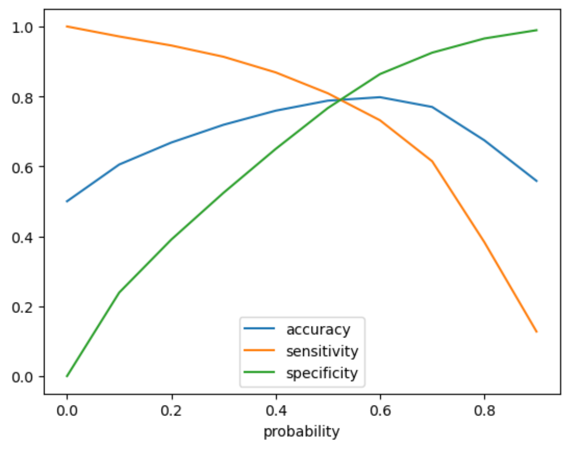


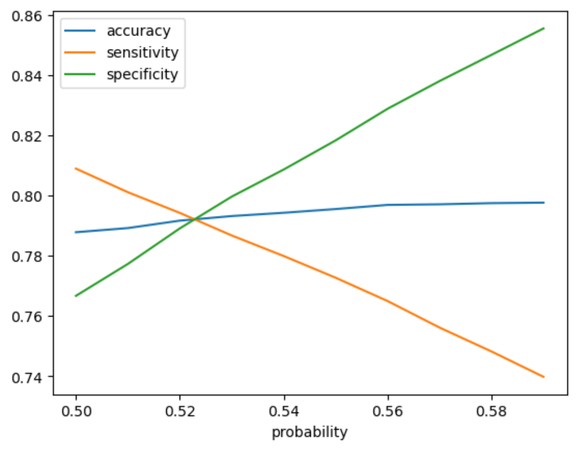


It is noticed that the maximum churns occur in 0-6 months period.
Post which there is stability as the customer retains in the network.
The average revenue per user in good phase of customer is given by arpu_6 and arpu_7.
It is evident that churn rate is high when there is no recharge or when the number of recharges are less than 10
Churn rate is inversely proportional to number of recharges

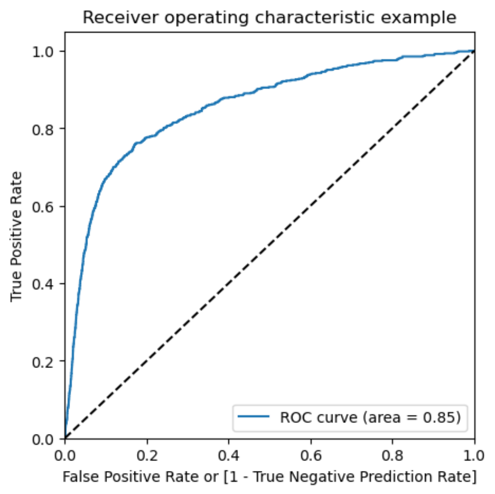
PREDICTIONS ON TRAIN SET







PREDICTIONS ON TRAIN SET



The optimal cutoff point in the probability to define the predicted churn variabe converges at 0.53

The accuracy of the predicted model is: 81.0 % The sensitivity of the predicted model is: 77.0 % As the model created is based on a sentivity model, i.e. the True positive rate is given more importance as the actual and prediction of churn by a customer

Summary:

- Note that the best parameters produced the accuracy of 91% which is not significantly deterred than the accuracy of original random forest, which is pegged around 92%

Conclusion :

- The best model to predict the churn is observed to be Random Forest based on the accuracy as performance measure.
- The incoming calls (with local same operator mobile/other operator mobile/fixed lines, STD or Special) plays a vital role in understanding the possibility of churn. Hence, the operator should focus on incoming calls data and has to provide some kind of special offers to the customers whose incoming calls turning lower.

Details:

- After cleaning the data, we broadly employed three models as mentioned below including some variations within these models in order to arrive at the best model in each of the cases.

FINAL RESULT

Logistic Regression :

- Logistic Regression with RFE Logistic regression with PCA Random Forest For each of these models, the summary of performance measures are as follows:

Logistic Regression

Train Accuracy : ~79%
Test Accuracy : ~80%

Logistic regression with PCA

Train Accuracy : ~91%
Test Accuracy : ~92%

Decision Tree with PCA:

Train Accuracy : ~93%
Test Accuracy : ~92%

Random Forest with PCA:

Train Accuracy :~ 91%
Test Accuracy :~ 92%

THANK YOU