



Telecom Churn Case Study

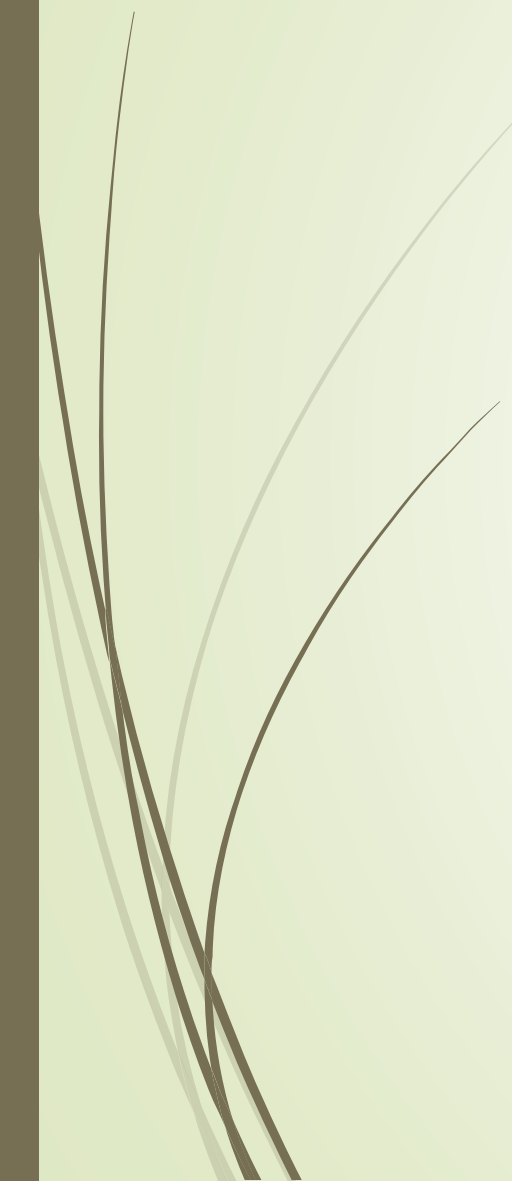
Analysis and Inferences

Submitted by

Raviraj Kangle
Mutyala Sridevi
Manish Gehani



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.
 - Customer retention has now become more important than customer acquisition.
 - To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.
- 

Goal

- Analyze 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


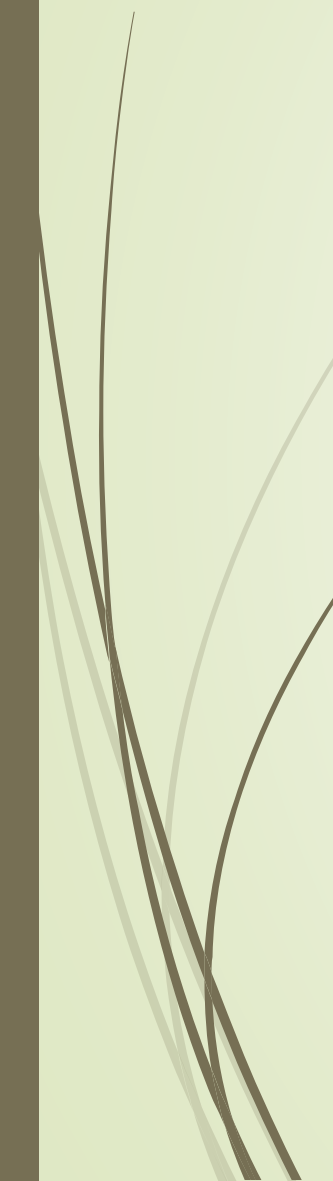
- There are two main models of payment in the telecom industry - **postpaid** (customers pay a monthly/annual bill after using the services) and **prepaid** (customers pay/recharge with a certain amount in advance and then use the services).
- Churn prediction is usually more critical (and non-trivial) for prepaid customers, and the term 'churn' should be defined carefully.
- The data pertains to 4 months June-September tagged by 6-9 numbers.
- June-July are indicated as 'Good' phase, August as the 'Action' phase and September as the 'Churn' phase during which the data was collected under various activities.



Defining 'Churn'

- **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.
- Approximately 80% of revenue comes from the top 20% of customers (called high-value customers).
- If we can reduce the churn of high-value customers, we will be able to reduce significant revenue leakage.
- In this task, we will define high-value customers based on 'Good' phase average recharge amount being greater than 70th percentile and predict churn only on high-value customers.

Approach followed

- 
- 
- 1- Data Understanding
 - - Importing Data and Check Statistics
 - 2- Data Cleaning
 - - Check missing values/checking outliers and fix those by checking their statistics
 - 3- Exploratory Analysis
 - - Uni-Variate, Bi-Variate and Correlation or pair plots
 - 4- Data Preparation
 - - Filter high-value customers
 - - Feature Scaling
 - 5- Build Model
 - - Split the data in train and test, features scaling, check correlation matrix,
 - - Features Selection using RFE and manual
 - 6- Model Evaluation
 - - Confusion Matrix
 - - Accuracy , Sensitivity, Specificity, Precision and recall, ROC Curve
 - 7- Prediction of test Data



Data Pre-processing

- Null values have been handled appropriately using the imputation, retaining the missing values as special category and deletion methods
- Features having same values throughout are dropped as they do not contribute in the prediction and may actually lead to bias
- Churn customers are tagged as per no activity in the 'Churn' month
- Data scaled to prevent the effect of outliers



Exploratory Data Analysis



Inferences from uni and bi-variate analysis

(performed on derived variables)

- ▶ We can see that there is only 3.39 percentage churn which is very low. This indicates target class imbalance
- ▶ The churn rate is more for the customers, whose minutes of usage(mou) decreased in the action phase than the good phase
- ▶ The churn rate is more for the customers, whose number of recharge in the action phase is lesser than the number in good phase
- ▶ The churn rate is more for the customers, whose amount of recharge in the action phase is lesser than the amount in good phase
- ▶ The churn rate is more for the customers, whose volume based cost in action month is increased. That means the customers do not do the monthly recharge more when they are in the action phase



Inferences from uni and bi-variate analysis

(performed on derived variables)

- The higher is the Average Revenue Per Customer (ARPU) – those customers are less likely to be churned
- Higher the Minutes of Usage (MOU), lesser the churn probability
- The 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
- The recharge number and the recharge amount are mostly proportional. More the number of recharge, more the amount of the recharge



Dealing with Class Imbalance

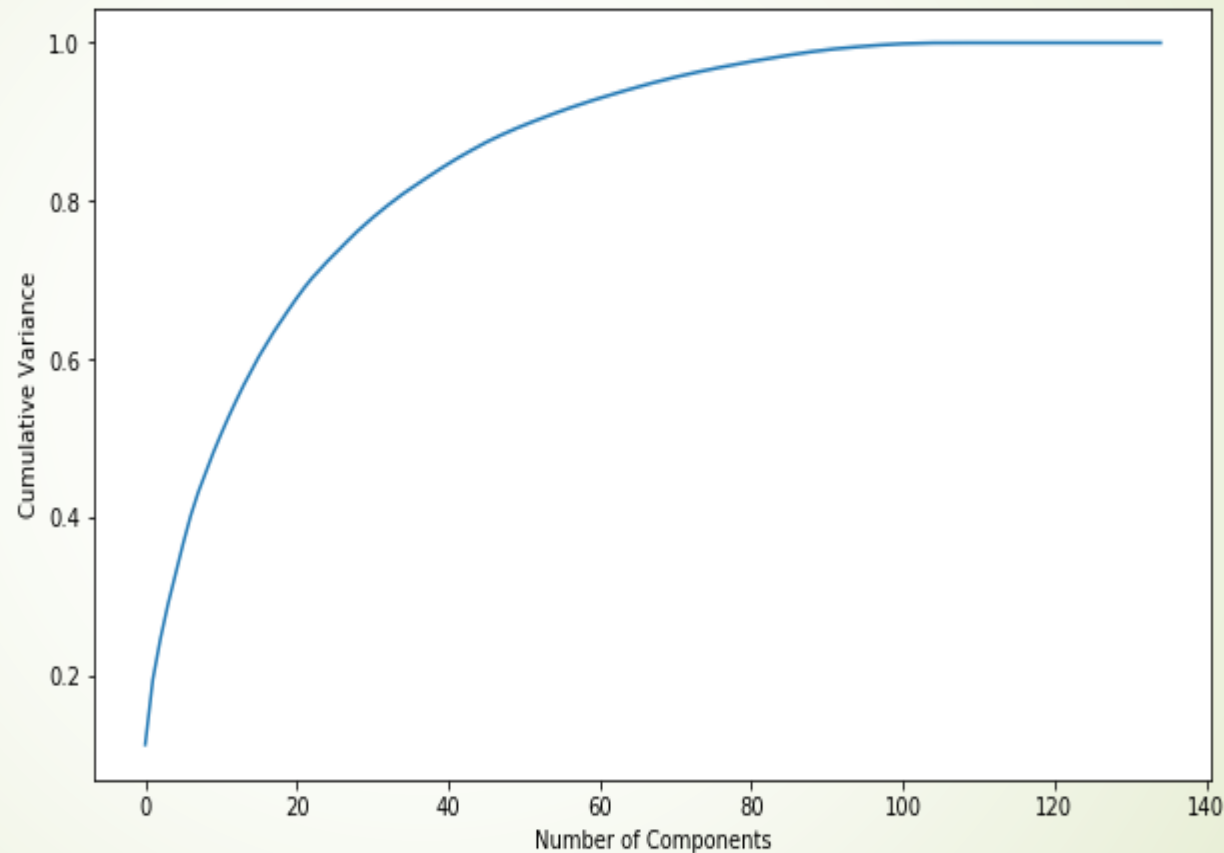
- Synthetic samples were created by doing up sampling using SMOTE (Synthetic Minority Oversampling Technique) as the Churn customers have very low representation in the dataset.



Feature Scaling

- The data variance is too high and hence standardization technique is used for feature scaling to bring it into common range

Principal Component Analysis (PCA)



As per above graph 60 components explain almost more than 90% variance of the data. So, we will perform PCA with 60 components.

Logistic Regression with PCA (with parameter tuning resulting in taking $C=1$)

Summary

Train set

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

Test set

- Accuracy = 0.83
- Sensitivity = 0.81
- Specificity = 0.83

The model is performing well in the test set, what it had learnt from the train set.

Decision Tree with PCA (with max_depth=10, min_sample_leaf=50, min_samples_split=100)

Summary

Train set

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

Test set (varying slightly with each run)

- Accuracy = 0.85
- Sensitivity = 0.64
- Specificity = 0.86

The model is performing well in the test set, what it had learnt from the train set. The Sensitivity has been decreased while evaluating the model on the test set. However, the accuracy and specificity is quite good in the test set.

Random Forest with PCA

(with max_depth=5, min_samples_leaf=50, min_samples_split=100, max_features=20, n_estimators=300)

Summary

Train set

- Accuracy = 0.84
- Sensitivity = 0.88
- Specificity = 0.80


Test set (varying slightly with each run)

- Accuracy = 0.79
- Sensitivity = 0.75
- Specificity = 0.79

The Sensitivity has been decreased while evaluating the model on the test set. However, the accuracy and specificity is quite good on the test set.



Conclusion of models with PCA

- After trying the 3 models namely Logistic Regression, Decision Tree and Random Forest, we can see that for achieving the best sensitivity, which was our ultimate goal, the classic Logistic regression with PCA performs well. Sensitivity drops with Decision Tree though the accuracy metric is good enough. Random Forest accuracy metric is at par with logistic regression, but the sensitivity dropped.
 - Keeping these in view and the interpretability of the models, logistic regression with PCA can be selected as the final model for Telecom Churn prediction.
- 



Logistic Regression without PCA (using Recursive Feature Elimination)

Summary

Train set

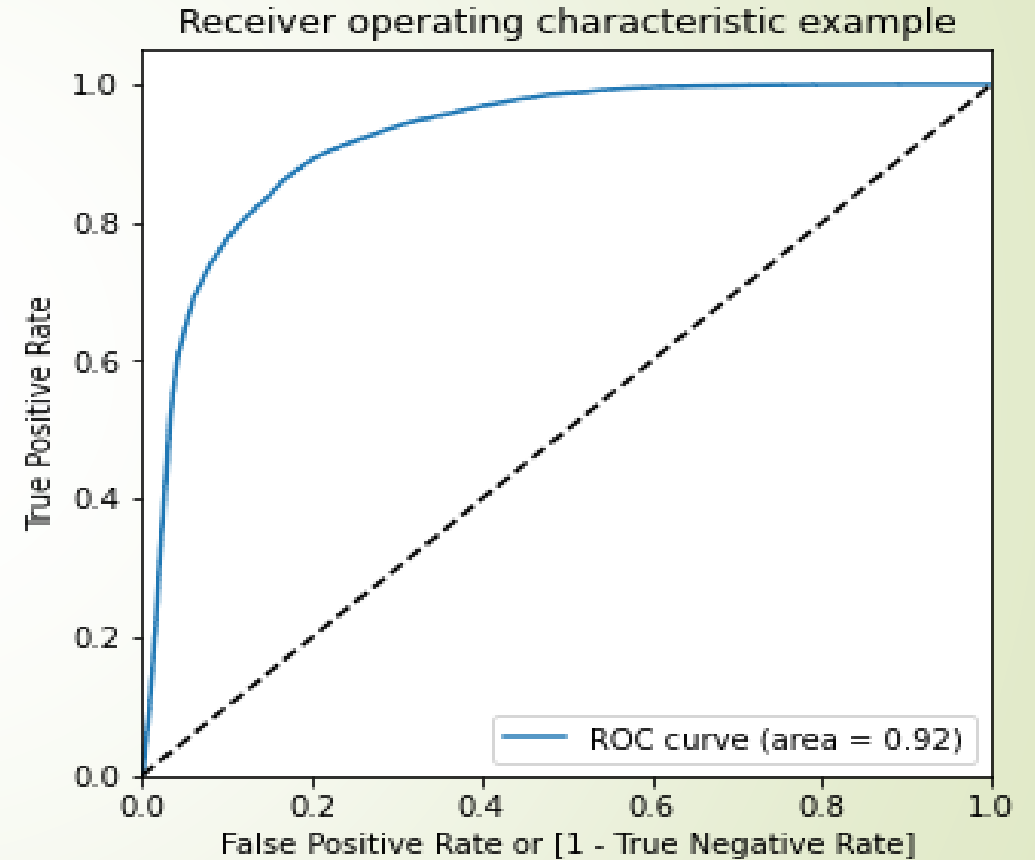
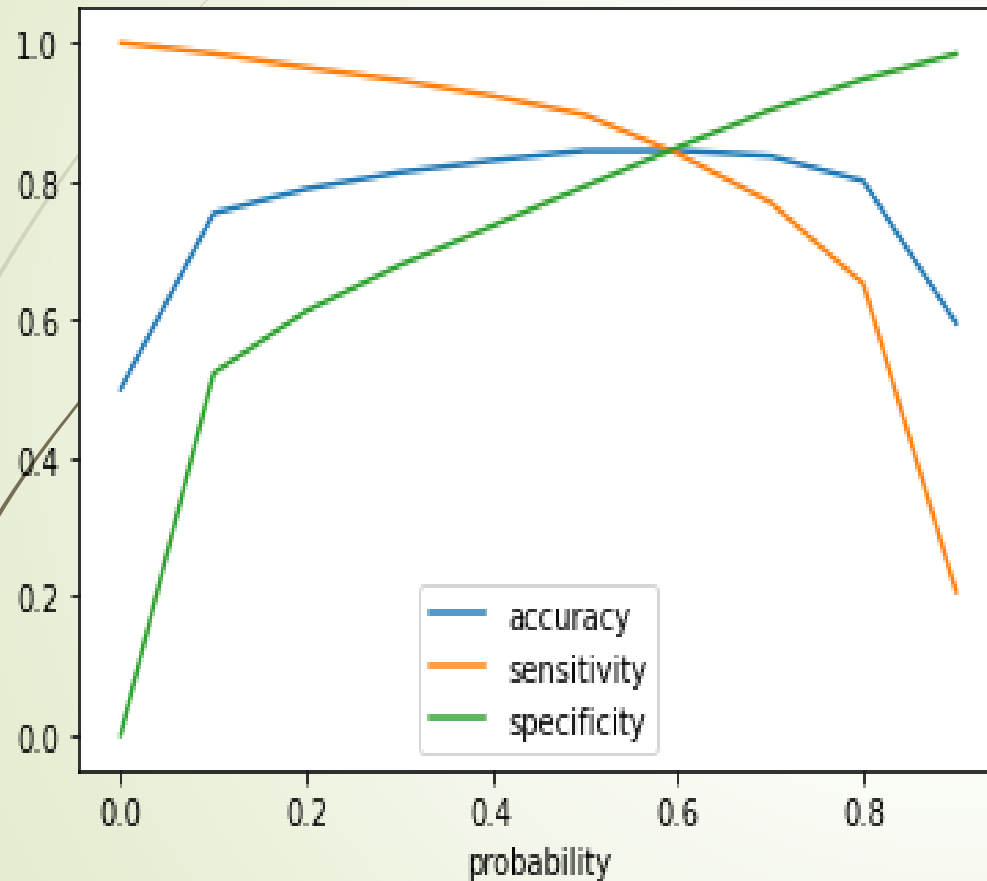
- Accuracy = 0.84
- Sensitivity = 0.89
- Specificity = 0.79

Test set

- Accuracy = 0.78
- Sensitivity = 0.82
- Specificity = 0.78


Overall, the model is performing well on the test set, what it had learnt from the train set.

Visualization of Optimal Cut-Off and ROC Curve





Final conclusion without PCA

- We can see that the logistic model with no PCA has good sensitivity and accuracy, which are comparable to the model with PCA.
 - We can go for the more simplistic model such as logistic regression with PCA as it explains the important predictor variables as well as the significance of each variable.
 - The model also helps us to identify the variables which should be acted upon for making the decision on the probable churn customers.
 - Hence, the model is more relevant in terms of explaining to the business clients.
- 



Business Recommendations

- If the local incoming minutes of usage (loc_ic_mou_8) is lesser in the month of August than any other month, then there is a higher chance that the customer is likely to churn
- 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).
- Target the customers, whose 'outgoing others' charge in July and 'incoming others' in August are less.
- Also, 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 offers.
- Customers whose monthly 3G recharge in August is more, are likely to be churned.



Business Recommendations

- 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.
- Customers with decreasing monthly 2g usage for August are most probable to churn.
- Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.
- '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.



Thank You