Telecom Churn Case Study

1. Sharon Benison
2. Sheik Alauddeen
3. Shiwani Jamdagni

Contents

Problem Statement

- 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, 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.

Problem Statement

- Understanding and Defining Churn
  - In the telecom industry, two primary payment models exist: postpaid and prepaid. In the postpaid model, customers settle their bills after using the services, typically on a monthly or annual basis. Conversely, in the prepaid model, customers pay in advance or recharge with a certain amount and then utilize the services.
  - In the postpaid scenario, customers usually notify their current operator when switching to another provider, making it clear-cut to identify this as churn. However, in the prepaid model, customers can simply cease using the services without formal notice, making it challenging to discern between actual churn and temporary non-usage (e.g., due to travel).
  - Therefore, predicting churn becomes more critical and complex for prepaid customers, necessitating a precise definition of the term 'churn.' It's worth noting that the prepaid model predominates in India and Southeast Asia, whereas postpaid is more prevalent in Europe and North America.

Problem Statement

- Definitions of Churn
- There are various ways to define churn, such as:
  - 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’. The main shortcoming of this definition is that there are customers who only receive calls/SMSes from their wage-earning counterparts, i.e. they don’t generate revenue but use the services. For example, many users in rural areas only receive calls from their wage-earning siblings in urban areas.
  - 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. A potential shortcoming of this definition is that when the customer has stopped using the services for a while, it may be too late to take any corrective actions to retain them. For e.g., if we define churn based on a ‘two-months zero usage’ period, predicting churn could be useless since by that time the customer would have already switched to another operator. In this project, we will use the usage-based definition to define churn.

Problem Statement

- High-value Churn
- In the Indian and Southeast Asian markets, about 80% of revenue is generated by the top 20% of customers, known as high-value customers. Therefore, by focusing on reducing churn among these high-value customers, we can effectively minimize revenue loss. This project aims to identify high-value customers based on specific metrics (to be detailed later) and predict churn exclusively among this segment.
- Understanding the Business Objective and the 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 Preparation

- The following data preparation steps were mainly used for this problem:

  1. Derived new features

  This is one of the most important parts of data preparation since good features are often the differentiators between good and bad models. We used our business understanding to derive features that we think are important indicators of churn

  2. Filtered high-value customers

  As specified, high-value customers were defined as those who had recharged with an amount equal to or greater than X, where X represented the 70th percentile of the average recharge amount during the initial two months (the good phase).

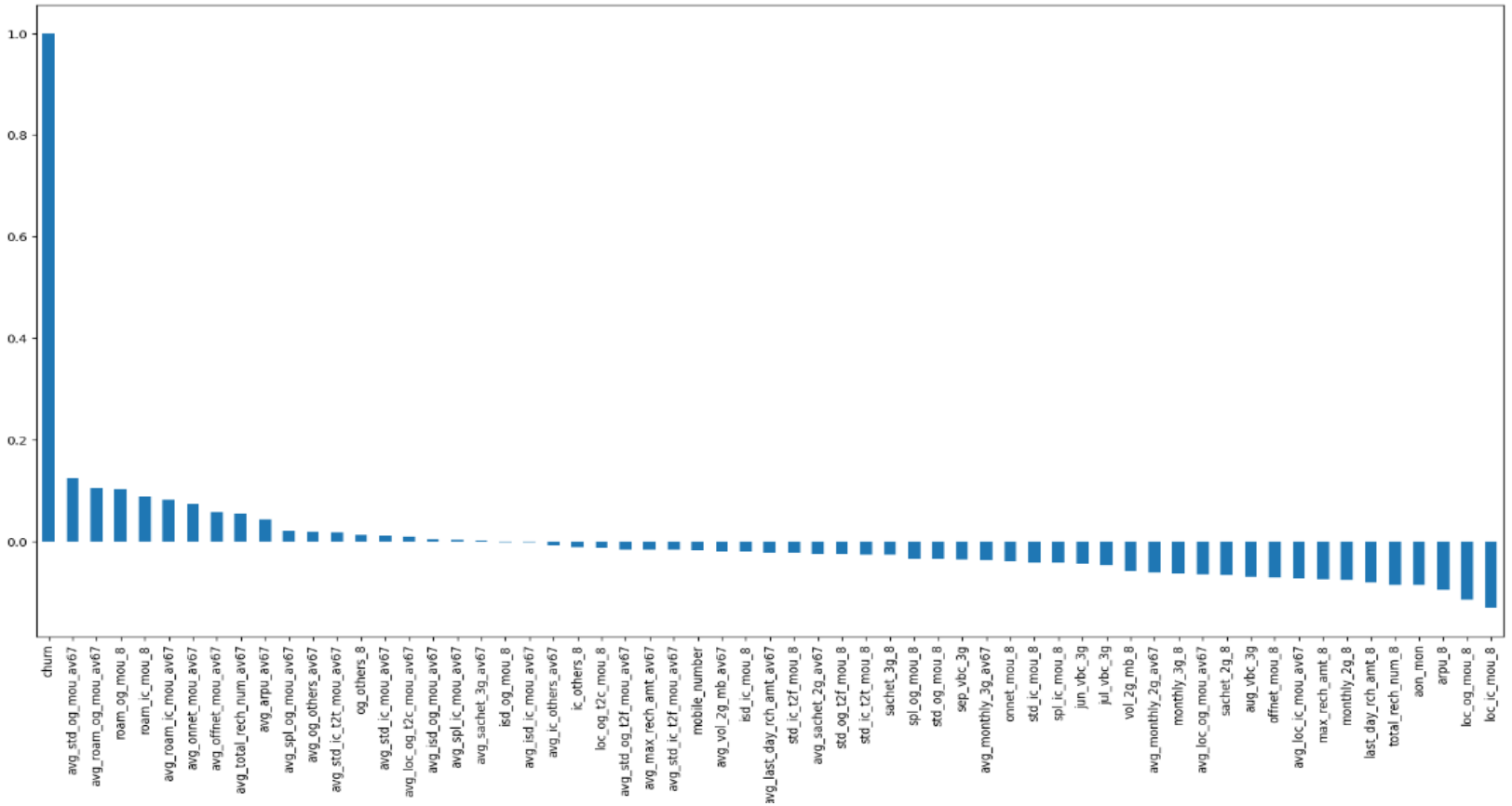
  3. Tagged churners and removed attributes of the churn phase

  In the tagging process, churned customers were identified (churn=1) based on their behavior during the fourth month. Specifically, customers who hadn't made any calls (incoming or outgoing) and hadn't used mobile internet even once during the churn phase were tagged as churned. Customers who didn't meet these criteria were tagged as non-churned (churn=0).

Data Preparation

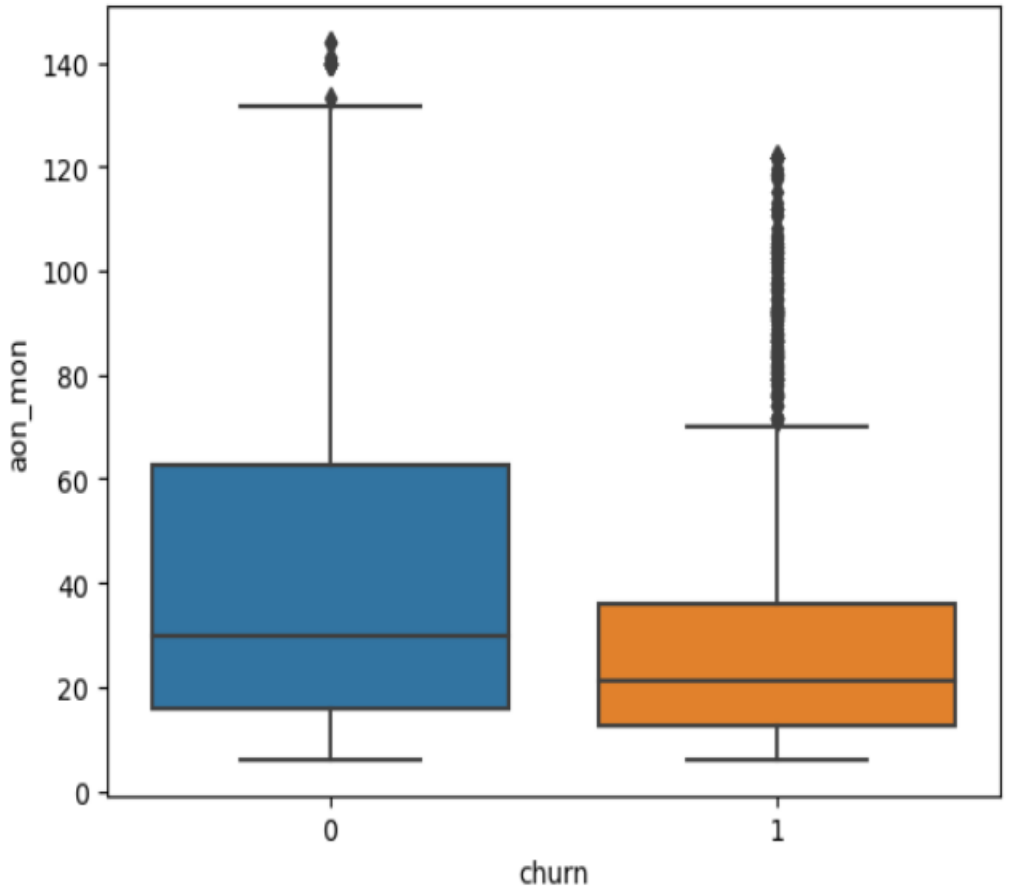
Few Observations:

- Avg std Outgoing Calls for 6th and 7th months & Avg roaming Outgoing calls for 6th and 7th months are positively correlated with churn.
- Minutes of usage for incoming local calls for 8th month) & Minutes of usage for outgoing local calls for 8th month & Average revenue per user for the 8th month has negative correlation with churn



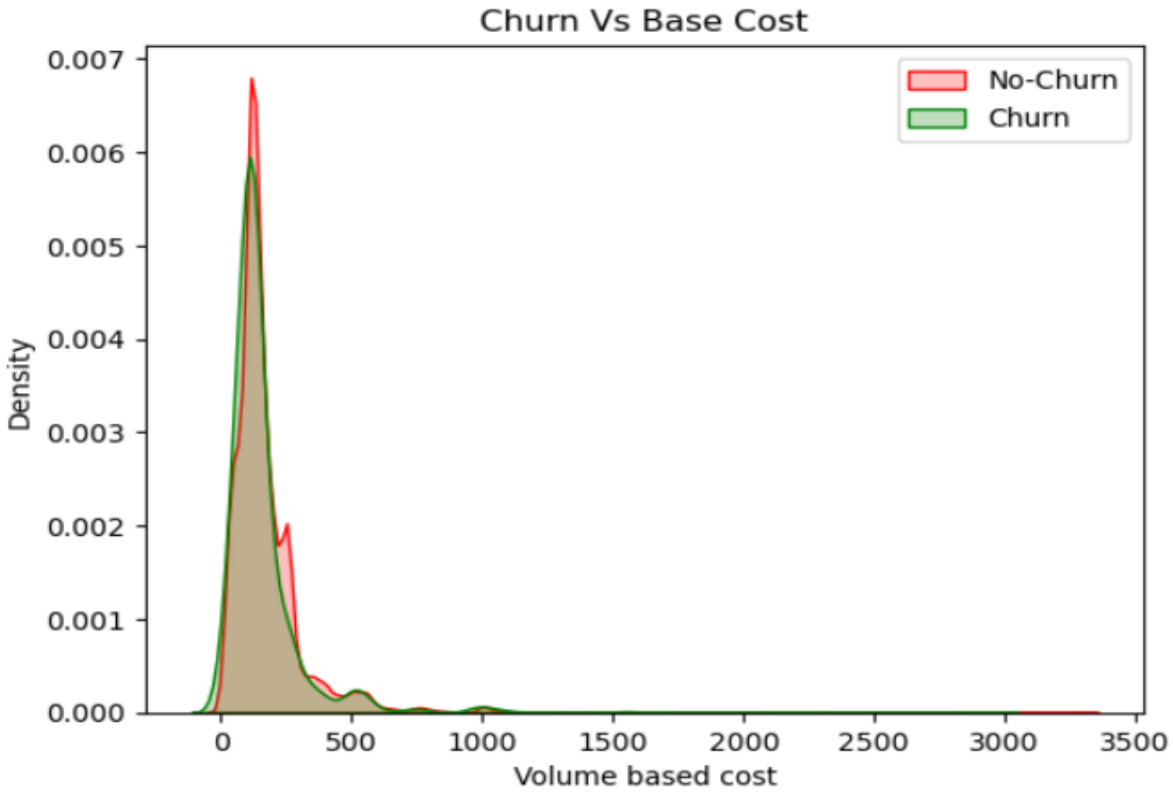
Data Preparation

- Few Observations:
  - Tenured customers tend not to churn much and continue to use telecom services comparing to non-tenured customers, it suggests a positive correlation between tenure and customer retention



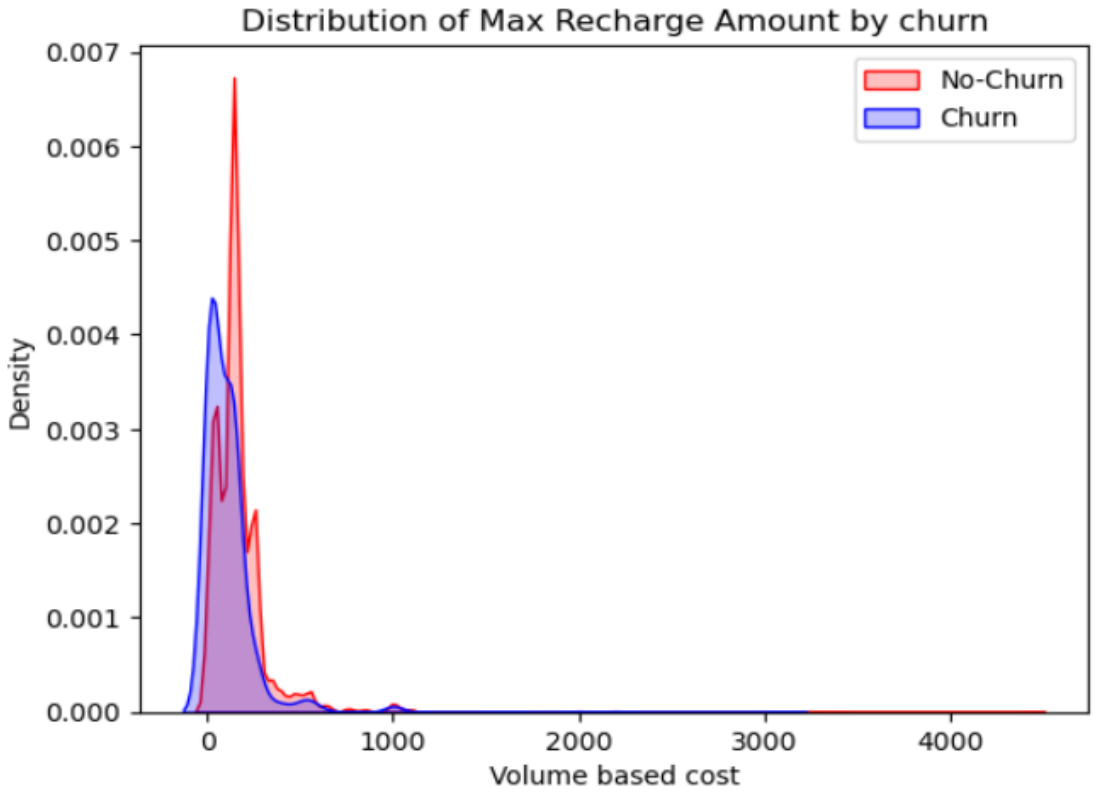
Data Preparation

- Few Observations:
  - The presence of churn results in lower and more concentrated costs, while the absence of churn leads to higher and more dispersed costs across different values.



Data Preparation

- Few Observations:
  - Churned customers have lower and more dispersed maximum recharge amounts/costs
  - Non-churned customers have higher and more concentrated maximum recharge amounts/costs clustered around a specific value.



Model Preparation

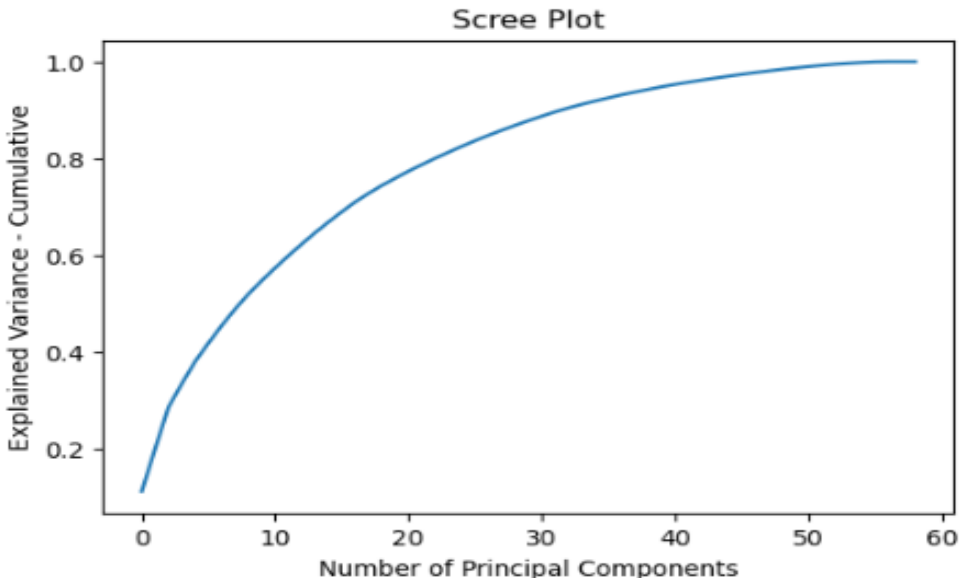
- The data was used to predict whether a high-value customer would churn or not in the near future (i.e., churn phase). By having this information, the company could take action steps such as providing special plans or discounts on recharges to retain customers.
- Additionally, the data was utilized to identify significant variables that served as strong predictors of churn. These variables could shed light on the reasons why customers chose to switch to other networks.
- We used following models for our purpose
  - Logistic Regression
  - Principal Component Analysis (PCA)
  - Decision Tree Model
  - Random Forest Model

Model Preparation

- Logistic Regression
  - Following are observations from logistics Regression Model
    - Precision: High for class 0 (non-churning customers) and low for class 1 (churning customers)
    - Recall (Sensitivity): Good for both classes but slightly higher for class 1
    - F1-Score: Higher for class 0, indicating better balance between precision and recall
    - Support: More instances of class 0 compared to class 1
    - Accuracy: Overall accuracy is 78%, indicating correct predictions for 78% of instances
    - In summary, the model performs well in identifying non-churning customers but needs improvement in predicting churning customers.
- Decision Tree Model
  - Following are observations from Decision Tree Model
    - 89% accuracy on the test dataset
    - lots of false positives in the confusion matrix

Model Preparation

- Principal Component Analysis (PCA)
  - Following are observations from PCA
    - Model has 87% Accuracy
    - 33 features can explain 90% variance in the dataset
    - Most important features:
      - arpu_8,
      - onnet_mou_8
      - offnet_mou_8
      - roam_ic_mou_8
      - roam_og_mou_8



Model Preparation

- Random Forest Model
  - Following are observations from Random Forest Model
    - Local Incoming for Month 8, Average Revenue Per Customer for Month 8 and Max Recharge Amount for Month 8 are the most important predictor variables to predict churn.

Conclusion

- Std Outgoing Calls and Revenue Per Customer emerged as robust indicators of churn.
- Local Incoming and Outgoing Calls in the 8th month, along with average revenue in the 8th month, proved to be the most influential columns for predicting churn.
- Customers with a tenure of fewer than 4 years are predisposed to churn.
- Max Recharge Amount stands out as a significant feature for predicting churn.
- Random Forest yielded the most accurate prediction results, with SVM following closely behind.