

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, you will 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).

In the postpaid model, when customers want to switch to another operator, they usually inform the existing operator to terminate the services, and you directly know that this is an instance of churn.

However, in the prepaid model, customers who want to switch to another network can simply stop using the services without any notice, and it is hard to know whether someone has actually churned or is simply not using the services temporarily (e.g. someone may be on a trip abroad for a month or two and then intend to resume using the services again).

Thus, churn prediction is usually more critical (and non-trivial) for prepaid customers, and the term 'churn' should be defined carefully. Also, prepaid is the most common model in India and southeast Asia, while postpaid is more common in Europe in North America.

This project is based on the Indian and Southeast Asian market.

Definitions of Churn

There are various ways to define churn, such as:

Revenue-based churn: Customers who have not utilized 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 you 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, you will use the **usage-based** definition to define churn.

High-value Churn

In the Indian and the southeast Asian market, approximately 80% of revenue comes from the top 20% customers (called high-value customers). Thus, if we can reduce churn of the high-value customers, we will be able to reduce significant revenue leakage.

In this project, you will define **high-value customers** based on a certain metric (mentioned later below) and **predict churn only on high-value customers**.

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 behavior during churn will be helpful.

Understanding Customer Behaviour During Churn

Customers usually do not decide to switch to another competitor instantly, but rather over a period of time (this is especially applicable to high-value customers). In churn prediction, we assume that there are three phases of customer lifecycle :

The 'good' phase: In this phase, the customer is happy with the service and behaves as usual.

The 'action' phase: The customer experience starts to sore in this phase, for e.g. he/she gets a compelling offer from a competitor, faces unjust charges, becomes unhappy with service quality etc. In this phase, the customer usually shows different behavior than the 'good' months. Also, it is crucial to identify high-churn-risk customers in this phase, since some corrective actions can be taken at this point (such as matching the competitor's offer/improving the service quality etc.)

The 'churn' phase: In this phase, the customer is said to have churned. You define churn based on this phase. Also, it is important to note that at the time of prediction (i.e. the action months), this data is not available to you for prediction. Thus, after tagging churn as 1/0 based on this phase, you discard all data corresponding to this phase.

In this case, since you are working over a four-month window, the first two months are the 'good' phase, the third month is the 'action' phase, while the fourth month is the 'churn' phase.

Data Dictionary

The data dictionary contains 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 containing 6, 7, 8, 9 as suffixes imply that those correspond to the months 6, 7, 8, 9 respectively.

Data Preparation

The following data preparation steps are crucial for this problem:

1. Derive 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. Use your business understanding to derive features you think could be important indicators of churn.

2. Filter high-value customers

As mentioned above, you need to predict churn only for the high-value customers. Define high-value customers as follows: Those 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).

After filtering the high-value customers, you should get about 29.9k rows.

3. Tag churners and remove attributes of the churn phase

Now tag the churned customers (churn=1, else 0) based on the fourth month as follows: Those who have not made any calls (either incoming or outgoing) AND have not used mobile internet even once in the churn phase. The attributes you need to use to tag churners are:

total_ic_mou_9

total_og_mou_9

vol_2g_mb_9

vol_3g_mb_9

After tagging churners, remove all the attributes corresponding to the churn phase (all attributes having ‘_9’, etc. in their names).

Modeling

Build models to predict churn. The predictive model that you’re going to build will serve two purposes:

It will be used to predict whether a high-value customer will churn or not, in near future (i.e. churn phase). By knowing this, the company can take action steps such as providing special plans, discounts on recharge etc.

It will be used to identify important variables that are strong predictors of churn. These variables may also indicate why customers choose to switch to other networks.

In some cases, both of the above-stated goals can be achieved by a single machine learning model. But here, you have a large number of attributes, and thus you should try using a dimensionality reduction technique such as PCA and then build a predictive model. After PCA, you can use any classification model.

Also, since the rate of churn is typically low (about 5-10%, this is called class-imbalance) - try using techniques to handle class imbalance.

You can take the following suggestive steps to build the model:

Preprocess data (convert columns to appropriate formats, handle missing values, etc.)

Conduct appropriate exploratory analysis to extract useful insights (whether directly useful for business or for eventual modeling/feature engineering).

Derive new features.

Reduce the number of variables using PCA.

Train a variety of models, tune model hyperparameters, etc. (handle class imbalance using appropriate techniques).

Evaluate the models using appropriate evaluation metrics. Note that it is more important to identify churners than the non-churners accurately - choose an appropriate evaluation metric which reflects this business goal.

Finally, choose a model based on some evaluation metric.

The above model will only be able to achieve one of the two goals - to predict customers who will churn. You can't use the above model to identify the important features for churn. That's because PCA usually creates components which are not easy to interpret.

Therefore, build another model with the main objective of identifying important predictor attributes which help the business understand indicators of churn. A good choice to identify important variables is a logistic regression model or a model from the tree family. In case of logistic regression, make sure to handle multicollinearity.

After identifying important predictors, display them visually - you can use plots, summary tables etc. - whatever you think best conveys the importance of features.

Finally, recommend strategies to manage customer churn based on your observations.

Note: Everything has to be submitted in one Jupyter notebook.

Evaluation Rubrics

Your solution will be evaluated based on the following rubrics

Stage	Meets expectations	Does not meet expectations

<p>Data understanding, preparation, and feature engineering (35%)</p>	<p>All important data quality checks are performed and inconsistent/missing data is handled appropriately.</p> <p>Relevant EDA is done using plots and summaries. The insights from EDA are clearly derived and explained.</p> <p>Filtering high-value customers and tagging churned customers is done correctly.</p> <p>Feature engineering is conducted rigorously and correctly. An appropriate set of features is used to build the model.</p>	<p>Data quality checks are not performed/missing data is not handled correctly.</p> <p>Exploratory analysis is not conducted/useful observations are either not extracted or mentioned clearly.</p> <p>Filtering high-value customers or tagging is done incorrectly.</p> <p>Feature engineering is not conducted or is conducted on an inappropriate set of features.</p>
---	--	--

<p>Modelling (churn prediction) (35%)</p>	<p>Dimensionality reduction is conducted correctly, including data preparation required for it.</p> <p>Class imbalance is handled using at least one of the techniques.</p> <p>Model hyperparameters are tuned using correct principles and the approach is explained clearly.</p> <p>A reasonable number and variety of different models are attempted and the best one is chosen based on key performance metrics.</p> <p>Model evaluation is conducted using an appropriate metric.</p>	<p>Dimensionality reduction is not conducted correctly/data is not preprocessed.</p> <p>Class imbalance is not handled.</p> <p>Model hyperparameters are not tuned correctly or the approach is not explained clearly.</p> <p>Few models are experimented with resulting in suboptimal results.</p> <p>Model evaluation is not conducted using an appropriate metric.</p>
---	--	---

	<p>Model evaluation results are at par with the best possible models on this data set.</p>	<p>The results are suboptimal compared to what is possible on this dataset.</p>
--	--	---

Identifying important churn indicators and business recommendation (20%)	<p>Important churn indicators are identified correctly.</p> <p>Clear actionable recommendations are provided based on supporting evidence.</p>	<p>Important indicators are not identified correctly.</p> <p>Recommendations are unclear, unactionable or not backed with supporting evidence.</p>
Code readability and conciseness (10%)	<p>The code is well commented and text is written in detail to explain the thought process.</p> <p>Efficient, concise code is written.</p>	<p>The code is not commented well / text is not written in detail.</p> <p>Inefficient/verbose code is written.</p>