Telecom Churn Analysis

RAM MANOHAR THAKUR,
SATYASURYAPAVAN CHANDURI,
AVINASH ARORA
Data science trainees,
AlmaBetter, Bangalore

ABSTRACT:

The telecoms industry is a highly competitive sector which is constantly challenged by customer churn or attrition. In order to remain steadfast in the consumer business, companies need to have sophisticated churn management strategies that will harness valuable data for business intelligence. Data mining and machine learning are tools which can be used by telecoms companies to monitor the churn behaviour of customers.

This study implemented exploratory data analysis and feature engineering in a public domain Telecoms dataset and this study discussed how these results are essential in reducing customer churn and improving customer service.

The principal contribution of our project is to develop a customer churn analysis using exploratory data analysis which helps telecom operators to analysis Telecom's dataset and how these results are essential in reducing customer churn and improving customer service.

Keywords- Churn prediction, customer churn, exploratory data analysis, feature engineering, Retention, Higher Subscriber Base, Telecommunication, Data mining

INTRODUCTION:

Customer churn is a good indicator of service quality and customer service satisfaction. The telecommunications industry is a dynamic business sector that is primarily composed of companies operating in a subscription-based model. These companies are constantly pressured with higher rates of customers who churned and shifted to rival companies that offer competitive products and services. Thus, some of them employ measures in determining the reasons why their customers churn and seek innovative strategies to improve customer satisfaction and increase the customer base.

Customer Relationship Management is a strategic process of managing customer relations and customer retention. Some companies mine customers' data to better understand the behaviour of their customers and gain actionable insights that help improve customer service. It can also help a company decide and employ proactive retention strategies Customer churn is that a customer ending a subscription to a service provider and choosing the services of another company. Churn rate is defined as the percentage of customers who stop subscribing to a service or percentage of employees leave a job. Churn has affected industries such as banking, insurance, internet streaming and telecommunications etc. There are many reasons for customer churn; some of the major reasons are service dissatisfaction, costly subscription, and better alternatives. Hence, in this paper the problem of churning is addressed and data factors affecting the churn are analysed for their effect on the churn rate.

BUSINESS UNDERSTANDING:

This initial phase of data analysis focuses on understanding the objectives of the project and requirements from a business point of view, and then converting this knowledge into a data analysis problem definition. Customer retention consists of "Identifying which customers are likely to Churn, determining which customers should retain and developing strategies to retain profitable customers". The main thing in retention process is identifying Churn ratio which is a very meaningful and vital determination for many companies. Determination of Churn ratio indicators is also very important. By using those indicators, firms can make prediction on future behaviour of new customers and can develop new strategies much before customers start to think about churn. Thus, it is vital to build a very successful and accurate Churn model during the retention studies.

CUSTOMER CHURN:

If a customer terminates a membership with one company and become a customer of another company, this customer is called as Churn customer. Today's economic trend dictates that price cuts are not the only way to build customer loyalty. Accordingly, adding new value added services to the products has become an industry norm to have loyal customer. The main goal of customer lost study is to figure out a customer who will likely be lost and is to calculate cost of obtaining those customers back again. During the analysis, the most important point is the definition of the churner customer.

Customer's loss is a major problem for companies which are likely to loose their customers easily. Banks, insurance and telecommunication companies can be given as examples. For companies, the cost of acquiring new customers is increasing day by day. Therefore, retaining customer is much more important than anything.

REASONS FOR CUSTOMER CHURNING:

- Price: comparatively high Pricing leads the customers to flee from one carrier to a competitor.
- Service quality: Lack of network coverage may make a customer go to another company with good network coverage.
- Lack of customer service: Slow or no response to customer complaints makes a customer more likely to churn.
- Billing disputes
- New competitors entering the market.
- Competitors introducing new products or technology.

PROBLEM STATEMENT:

Maximize: Company's profit by retaining customer

Minimize: Customer churn by identifying the key cause of the problem

The main goal of the project is to:
Finding factors and cause those influence customers to churn. Retain churn customers by taking appropriate steps providing offers based on affecting factors. Using the data provided, this paper aims to

analyse the data to determine what variables are correlated with customer churn, if any. To identify the people that might churn, will also be analyse.

DATA DESCRIPTION:

The data description phase starts with an initial data collection and proceeds with activities in order to get familiar with the data. Identifying data quality problems, discovering first insights into the data and detecting interesting subsets to form hypotheses from hidden information are activities of this step. Data which is collected from a telecommunication company to get analysed, involves usage details of customers from. The data was taken from Orange Telecom Company. It had 3333 rows and 20 columns. Most columns related to subscriber personal. Other column was indicative of service usage by the subscriber. Based on the business understanding of the data 18 columns was chosen to analyse the data

DATASET PREPARATION:

The customer churn dataset from orange telecom company contains 20 features and 3333 observations. The feature 'Churn' shows customer churn or non-churn based on existing conditions. Approximately 14.5% are churn and 84.5% are no churn. Below Table shows the data features.

Data-set description

Feature Name	Type
State	object
Account length	Int64
Area code	Int64
International plan	object
Voice mail plan	object
Number vmail messages	Int64
Total day minutes	Float64
Total day calls	Int64
Total day charge	Float64
Total eve minutes	Float64
Total eve calls	Int64
Total eve charge	Float64
Total night minutes	Float64
Total night calls	Int64
Total night charge	Float64
Total intl minutes	Float64
Total intl calls	Int64
Total intl charge	Float64
Customer service calls	Int64
Churn	bool

FEATURE BREAKDOWN:

STATE: 51 Unique States

Account Length: Duration of length customer use their The Account

Area Code: There are 3 unique area code present 415, 408, and 510

International Plan: Yes Indicate International Plan is Present and No Indicates no subscription for International Plan

Voice Mail Plan: Yes Indicates Voice Mail Plan is Present and No indicates no subscription for Voice Mail Plan

Number vmail messages: Number of Voice Mail Messages ranging from 0 to 50

Total day minutes Total Number of Minutes Spent by Customers in Morning

Total day calls : Total Number of Calls made by Customer in Morning.

Total day charge: Total Charge to the Customers in Morning.

Total eve minutes: Total Number of Minutes Spent By Customers in Evening

Total eve calls: Total Number of Calls made by Customer in Evening.

Total eve charge: Total Charge to the Customers in Morning.

Total night minutes: Total Number of Minutes Spent By Customers in the Night.

Total night calls: Total Number of Calls made by Customer in Night.

Total night charge: Total Charge to the Customers in Night.

Churn: churning status of the customer

EXPLORATORY DATA ANALYSIS:

If we want to explain EDA in simple terms, it means trying to understand the given data much better, so that we can make some sense out of it. we using univariate frequency analysis was conducted to describe key characteristics of each feature including, minimum and maximum value, average, standard deviation and others. It was also used to produce a value distribution and identify missing values, and outliers.

EDA is a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures. In this chapter, we are going to discuss the steps involved in performing topnotch exploratory data analysis

In statistics, A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task.EDA in Python uses data visualization to draw meaningful patterns and insights

- ### DATA ANALYSIS:

This is one of the most crucial steps that deals with descriptive statistics and analysis of the data. The main tasks involve summarizing the data, finding the hidden correlation and relationships among the data, developing predictive models, evaluating the models, and calculating the accuracies. Some of the techniques used for data summarization are summary tables, graphs, descriptive statistics, inferential statistics, correlation statistics, searching, grouping, and mathematical models.

- ### DATA SOURCING

Data Sourcing is the process of finding and loading the data into our system. Broadly there are two ways in which we can find data.

1. Private Data
2. Public Data

Data collected from several sources must be stored in the correct format and transferred to the right information technology personnel within a company. As mentioned previously, data can be collected from several objects on several events using different types of sensors and storage tools.

- ### DATA PREPROCESSING:

A dataset may contain noise, missing values, and inconsistent data, thus, pre-processing of data is essential to improve the quality of data and time required in the data mining.

- ### DATA CLEANING

After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

Irregularities are of different types of data.

  - Missing Values
  - Incorrect Format
  - Incorrect Headers
  - Anomalies/Outliers

- **DATA TRANSFORMATION:**

Data transformation is the process of normalizing and aggregating the data to further improve the efficiency and accuracy of data mining.

- **DATA DEDUPLICATION:**

It is very likely that your dataset contains duplicate rows. Removing them is essential to enhance the quality of the dataset.

- **DATA REDUCTION:**

Data reduction is the process of reducing data representation while still producing similar results. Discretization (or binning) is applied to numeric attributes by transforming and grouping the continuous values into discrete categories (or interval levels)

- **MISSING VALUES:**

There is a representation of each service and product for each customer. Missing values may occur because not all customers have the same subscription. Some of them may have a number of service and others may have something different. In addition, there are some columns related to system configurations and these columns may have null values but in our orange telecom data set there are no null values present

If there are missing values in the Dataset before doing any statistical analysis, we need to handle those missing values.

There are mainly three types of missing values.

1. MCAR (Missing completely at random): These values do not depend on any other features.
2. MAR (Missing at random): These values may be dependent on some other features.

MNAR (Missing not at random): These missing values have some reason for why they are missing.

- **DROPPING MISSING VALUES:**

One of the ways to handle missing values is to simply remove them from our dataset. We have know that we can use the isnull() and notnull() functions from the pandas library to determine null values

- **HANDLING OUTLIERS:**

Outliers are data points that diverge from other observations for several reasons. During the EDA phase, one of our common tasks is to detect and filter these outliers. The main reason for this detection and filtering of outliers is that the presence of such outliers can cause serious issues in statistical analysis.

There are two types of outliers:

- **UNIVARIATE OUTLIERS:**

Univariate outliers are the data points whose values lie beyond the range of expected values based on one variable.

- **MULTIVARIATE OUTLIERS:**

While plotting data, some values of one variable may not lie beyond the expected range, but when you plot the data with some other variable, these values may lie far from the expected value.

- **MEASURES OF CENTRAL TENDENCY:**

The measure of central tendency tends to describe the average or mean value of datasets that is supposed to provide an optimal summarization of the entire set of measurements. This value is a number that is in some way central to the set. The most common measures for analyzing the distribution frequency of data are the mean, median, and mode.

- **MEASURES OF DISPERSION**:

The second type of descriptive statistics is the measure of dispersion, also known as a measure of variability. If we are analyzing the dataset closely, sometimes, the mean/average might not be the best representation of the data because it will vary when there are large variations between the data. In such a case, a measure of dispersion will represent the variability in a dataset much more accurately.

Multiple techniques provide the measures of dispersion in our dataset. Some commonly used methods are standard deviation (or variance), the minimum and maximum values of the variables, range, kurtosis, and skewness.

- **STANDARDIZING VALUES:**

To perform data analysis on a set of values, we have to make sure the values in the same column should be on the same scale. For example, if the data contains the values of the top speed of different companies' cars, then the whole column should be either in meters/sec scale or miles/sec scale.

- **UNIVARIATE ANALYSIS:**

  If we analyze data over a single variable/column from a dataset, it is known as Univariate Analysis. Univariate analysis looks at one feature at a time. When we analyse a feature independently, we are usually mostly interested in the distribution of its values and ignore other features in the dataset

  Univariate analysis is the simplest form of analyzing data. It means that our data has only one type of variable and that we perform analysis over it. The main purpose of univariate analysis is to take data, summarize that data, and find patterns among the values. It doesn't deal with causes or relationships between the values. Several techniques that describe the patterns found in univariate data include central tendency (that is the mean, mode, and median) and dispersion (that is, the range, variance, maximum and minimum quartiles (including the interquartile range), and standard deviation).

- **BIVARIATE ANALYSIS:**

If we analyze data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis.

- **a)Numeric-Numeric Analysis:**

Analyzing the two numeric variables from a dataset is known as numeric-numeric analysis. We can analyze it in three different ways.

  - Scatter Plot
  - Pair Plot
  - Correlation Matrix

- **b) Numeric - Categorical Analysis:**

Analyzing the one numeric variable and one categorical variable from a dataset is known as numeric-categorical analysis. We analyze those mainly using mean, median, and box plots.

- **MULTIVARIATE ANALYSIS:**

Multivariate analysis is the analysis of three or more variables. This allows us to look at correlations (that is, how one variable changes with respect to another) and attempt to make predictions for future behaviour more accurately than with bivariate analysis.

One common way of plotting multivariate data is to make a matrix scatter plot, known as a pair plot. A matrix plot or pair plot shows each pair of variables plotted against each other. The pair plot allows us to see both the distribution of single variables and the relationships between two variables

- **CORRELATION AMONG VARIABLES**:

In words, the statistical technique that examines the relationship and explains whether, and how strongly, pairs of variables are related to one another is known as correlation. Correlation answers questions such as how one variable changes with respect to another. If it does change, then to what degree or strength? Additionally, if the relation between those variables is strong enough, then we can make predictions for future behaviour

- **GRAPHICAL REPRESENTATION OF THE RESULTS:**

This step involves presenting the dataset to the target audience in the form of graphs, summary tables, maps, and diagrams. This is also an essential step as the result analyzed from the dataset should be interpretable by the business stakeholders, which is one of the major goals of EDA. Most of the graphical analysis techniques include Line chart, Bar chart, Scatter plot, Area plot, and stacked plot Pie chart, Table chart, Polar chart, Histogram, Lollipop chart etc.

CONCLUSIONS:

Telecommunication industry has suffered from high churn rates and immense profit loss due to churning. But we can avoid the customer churn. The importance of this type of research in the telecom market is to help companies make more profit. It has become known that predicting churn is one of the most important sources of income to telecom companies. Hence, this research aimed to build a system that predicts the churn of customers.

REFERENCES:

- Data science for business : what you think about data mining
- Hands-On Exploratory Data Analysis with Python Perform EDA techniques to understand, summarize, and investigate your data by Suresh Kumar Mukhiya, Usman Ahmed (z-lib.org)