Customer Churn in Telecom

Akshda Shandilya, Bashaer Alkhattabi, Jay Patel

NJCU School of Business

FINC535: Data Management

Dr. Xiaodi Zhu

December 18, 2022

# Table of Contents

**Table of Contents ……………………………………...……………………………………….2**

**Introduction …………………………………………………………………………………….3**

**Methodology ……………………………………………………………………………………4**

**Data Collection ………………………………….……...………………………………4**

**Variable Description ………………….…………………..……………………………4**

**Designing a relational database …………………….…………………………………5**

**Demo data selection …………………...……………….………………………………6**

**Data Storage …………………………………………..…..……………………………9**

**Potential Data Issues …………………………………...……….………………...……9**

**Missing information ……………………………………………………………9**

**Security ………………………………….………...…....………………………9**

**Safety ………………..………………….………………………….………..…10**

**Results ………………………………………………….…………..…………….……………10**

**Conclusion …………………………………………….…………...……...….……..…………11**

**References …………………………………………….……...………...…….……..…………12**

# **Introduction**

The main focus of the Telcom industry is attracting new customers retaining their current customers and avoiding termination services which are defined as churn. Churning affects companies' revenue and there are many reasons for it, like unprofessional customer service or low service coverage or competitive price offers from other companies. Churn analytics take a valuable place because it can predict the number of customer churn and outline the reasons causing it. The churn matrix is often displayed as the percentage of consumers canceling service within a particular time frame. Moreover, Telcos utilize machine learning algorithms to anticipate churn over an individual client basis and implement solutions like special offers that can help to reduce churn.

## Problem Definition

The major problem is predicting whether a specific consumer would churn or not. This could be obtained using a machine-learning model that uses 80% of the data sample. The rest of the data sample is used to assess that predictive power with regards to “churn / not churn. An important question should be raised, which feature leads clients to churn, the data can be utilized to identify that feature and come up with solutions that satisfy customers and prevent churn.

The accuracy is evaluated to compare models and choose the best one for this task. Other metrics are examined if necessary based on other aspects of the data, such as the balance between classes (number of "churners" vs. "non-churners" in the data set).

# Methodology

## Data Collection:

This study involves collecting data from various telecom industry stakeholders. However, most companies do not have an API, from where the developers can download the data. In such a case, the data can be downloaded from different open-source websites like Github and Kaggel for free to run an analysis on customer churn. Additionally, the IBM Watson Customer Churn dataset, which provides this information is also available for free on the UCI Machine Learning Repository for free. Finally, the data could also be acquired using data crawling for customer information. The dataset for this study was acquired from the Kaggle Customer Churn dataset.

## Variable Description:

The data collected contained demographic information from the customers, including age, gender, income, etc. It also contained other, relevant information, such as usage patterns, including text, call, and internet usage patterns, and churn labels indicating if the customer was churned or converted from one mobile network to another or not. The variables used in this study are described as follows:

|  | **Variable** | **Data Type** |
| --- | --- | --- |
|  | Customer ID | Numeric |
|  | Age | Numeric |
|  | Gender | Categorical |
|  | Income | Numeric |
|  | Number of Calls made | Numeric |
|  | Number of text messages sent | Numeric |
|  | Churn label | Binary |

Table 1

## Designing a relational database:

For this dataset, a potential design could involve designing the following table:

| Primary Heading | Subheadings | Data type |
| --- | --- | --- |
| Customer | Customer ID | Numeric |
| Age | Numeric |
| Gender | Categorical |
| Income | Numeric |
| Usage | Number of calls made | Numeric |
| Number of text messages sent | Numeric |
| Churn | Churn label for customer | Binary |

Table 2

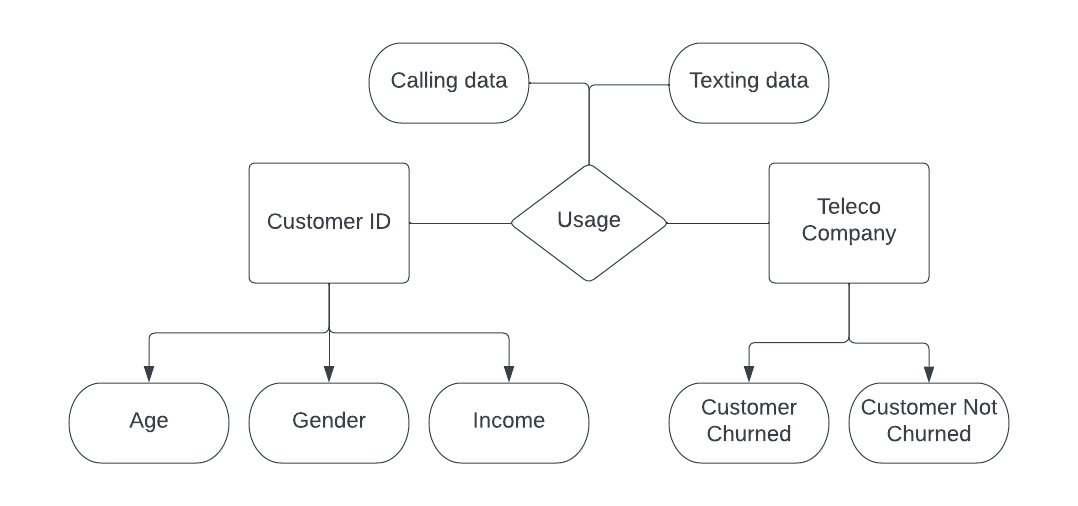
This database can use the customer ID as a key to access the data. The following is a flowchart representing the dataset as stored in the database.

Image 1

## Demo data selection:

Queries that can be used for this dataset:

| **Code** | **Usage** |
| --- | --- |
| CREATE DATABASE Projects; | Create the database for this project |
| CREATE TABLE Churn (  RowNumber INT NOT NULL AUTO\_INCREMENT,  CustomerId DECIMAL(10) NULL,  Gender VARCHAR(255) NOT NULL,  Age DECIMAL(3) NULL,  Income DECIMAL(10 , 2 ) NULL,  NumOfCalls DECIMAL(3) NULL,  NumOfTexts DECIMAL(3) NULL,  Exited DECIMAL(3) NULL,  PRIMARY KEY (RowNumber)  ); | Create the table. Here, “Exited” refers to the customers who churned. |
| LOAD DATA INFILE '<files>/Churn\_Dataset.csv'  INTO TABLE Churn  FIELDS TERMINATED BY ','  ENCLOSED BY '"'  LINES TERMINATED BY '\n'  IGNORE 1 ROWS; | Add data to the table |
| SELECT \* FROM Churn; | View the table |
| Use Projects;  ALTER TABLE Churn DROP RowNumber;  ALTER TABLE Churn DROP CustomerId; | Dropping irrelevant columns |
| Use Projects;  SELECT COUNT(\*) AS null\_values FROM Churn WHERE Churn IS NULL; | Finding null values |
| DELETE FROM Churn WHERE Exited = '';  DELETE FROM Churn WHERE Exited IS NULL; | Deleting null values |

Table 3

After this data has been prepared, it can be moved to a Python code running platform like Google Collab or Jupyter Notebook to run further analysis.

| **Code** | **Explanation** |
| --- | --- |
| import mysql.connector  import pymysql  import pandas as pd  mydb = mysql.connector.connect(  host="localhost",  user="root",  password="<insert password here>",  database="Projects",  auth\_plugin="mysql\_native\_password"  )  df = pd.read\_sql('SELECT \* FROM Churn', con=mydb) | Importing the data to the Jupyter notebook |
| df.head(5) | Viewing the top five rows of the data |
| df.shape | Viewing the shape(Details) of the dataset |
| df["Exited"] = df["Exited"].astype(int)  Y = df["Exited"].values  X = df.drop(labels = ["Exited"],axis = 1)  from sklearn.model\_selection import train\_test\_split  X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=101) | Splitting data for analysis |
| import matplotlib.pyplot as plt  # Data for plotting  sizes = df['Exited'].value\_counts(sort = True)  colors = ["lightblue","red"]  my\_labels = 'Not Churned','Churned'  # Plotting  plt.pie(sizes, labels=my\_labels, colors=colors,  autopct='%1.1f%%', shadow=True, startangle=270,)  plt.title('Percentage of Churn in Dataset')  plt.show() | Displaying data using different visualization tools. |
| from sklearn.linear\_model import LogisticRegression  model = LogisticRegression()  result = model.fit(X\_train, y\_train) | Logistic Regresssion model Training |
| from sklearn import metrics  prediction\_test = model.predict(X\_test)  # Print the prediction accuracy  print (metrics.accuracy\_score(y\_test, prediction\_test)) | Logistic regression model testing |
| # get the weights of all variables  weights = pd.Series(model.coef\_[0], index=X.columns.values)  weights.sort\_values(ascending = False) | To get the best prediction model |

## Data Storage:

In this study, using a relational database is the most appropriate method to store the information.

*Why not a NoSQL database?*

This is because the dataset, in this case, is structured and well-defined. A NoSQL database is most effective in datasets of large volume, that are unstructured or semi-structured. Such databases can scale easily and do not require a strict schema for storage. However, the NoSQL databases are not very efficient to run queries and analyses on the databases. Therefore, it is most appropriate to store the database in a Relational Database.

## **Potential Data Issues**:

When attempting to predict customer churn for a telecommunications company, several potential data issues may arise. Some of the most common issues include

**Missing information**: If some customer data is missing or incomplete, it can affect the accuracy of the churn prediction model. This can make it challenging to accurately predict churn risk, as there may be gaps in important information.

**Incorrect entries**: The reliability of the estimation method can also be affected by data entry errors, which can be brought on by either human or computer errors.

**Security**: A critical problem is preventing unauthorized access to sensitive client information. This contains any personal details that may be gathered as well as information on invoicing and payment history. To protect user privacy and avoid potential legal problems, it is crucial to maintain the security of this data.

**Safety**: Ensuring the protection of data collected and stored on customers is also essential. This involves storing data securely and protecting it from security breaches and threats. When examining and utilizing client data, experts must also follow moral and ethical data handling principles.

# Results

Predicting customer churn is important for telecommunications companies because it enables them to identify and keep valuable customers and maintain a strong customer base. It is important for a telecoms company to understand which customers are at risk of leaving and take steps to prevent it, as predicting customer churn is a typical business issue.

In order to address the problem of forecasting customer attrition, it would be necessary to gather data on the company's customers and their interactions with the company. This data may include details about how the company's services have been used, billing and payment records, and any customer service interactions.

After the data has been collected, experts could use machine learning methods to create a model that predicts the probability of a customer leaving. This model could be trained using historical data and then used to make predictions about new customers as well as current customers whose risk of churn has possibly altered.

To determine the efficiency of the model, the data scientist could use various evaluation metrics, such as accuracy, precision, and recall. They could also employ strategies like cross-validation to verify that the model is applicable and not overfitted to the training data.

# Conclusion

In conclusion, predicting customer churn is a critical business problem for telecommunications companies, as it allows them to identify and retain valuable customers and maintain a healthy customer base. By utilizing machine learning techniques, data scientists can build models that can accurately predict churn risk based on data collected on customers and their interactions with the company. However, there may be various data issues that arise, such as missing or incorrect data and security and safety concerns. To effectively deal with these data issues, data scientists must make sure they are collecting and storing data accurately and securely, and using ethical and responsible data practices when analyzing and using customer data Overall, predicting customer churn for a telecommunications company is a complex business problem that requires careful consideration of data issues in order to accurately predict and prevent churn.

# References

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*, *94*, 290–301. https://doi.org/10.1016/j.jbusres.2018.03.003

Kohli, C., & Suri, R. F. (2018). Customer centricity: A review and future directions. Journal of Business Research, 84, 1-10. https://doi.org/10.1016/j.jbusres.2018.03.003