

CUSTOMER CHURN PREDICTION

(REPORT)

Submitted by

MOHAMMED RIZAD IBRAHIM M

MOHAMAD RIZWAN S

ABSTRACT

This report provides a comprehensive analysis of **customer churn** in the **telecom industry**, focusing on predictive **modeling** techniques to anticipate **customer behavior**. The report examines various factors influencing churn and proposes predictive models to help telecom companies mitigate customer attrition. Through extensive data analysis and **machine learning techniques**, this study aims to provide valuable insights for telecom businesses to enhance customer retention strategies.

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.

In the rapidly evolving landscape of the **telecom industry**, understanding and **predicting customer churn** have become pivotal for **sustaining business growth** and enhancing customer satisfaction. This abstract provides a detailed overview of the research conducted in this report, highlighting the **methodologies, findings, and implications**.

The telecom sector, characterized by fierce competition and dynamic customer preferences, faces **substantial challenges** in retaining its customer base. This study delves into the phenomenon of customer churn by analyzing multifaceted data encompassing demographic attributes, service usage patterns, customer service interactions, and pricing strategies. Through **meticulous data collection** and preprocessing techniques, the research explores the intricate factors influencing customer decisions to switch telecom service providers.

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

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.

By accurately **predicting** which customers are likely to churn, the company can proactively take measures to retain them. This includes offering targeted promotions, improving customer service, or customizing communication to address specific needs, **ultimately enhancing customer satisfaction** and **reducing attrition**.

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 modeling** during the retention studies.

Customer retention, as defined in this context, is a multifaceted process. It involves three key components:

- **Identifying Likely Churners:** The first step is to identify which customers are likely to churn. Through data analysis and predictive modeling, the goal is to pinpoint specific patterns and behaviors in the data that indicate a high likelihood of a customer leaving the service provider.
- **Determining Retention Strategies:** Once potential churners are identified, the next step is determining which customers are valuable and should be retained. Not all customers are equal; some are more profitable or have higher potential for future revenue. It's crucial to prioritize these customers in retention efforts.
- **Developing Retention Strategies:** Armed with insights from data analysis, companies can develop targeted strategies to retain these profitable customers. These strategies can range from personalized offers and promotions to improved customer service experiences, all designed to enhance customer satisfaction and loyalty.

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** and **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**.

Objectives:

- **Identifying Key Churn Indicators:** Determine the most significant factors contributing to customer churn by conducting in-depth analysis of historical customer data, including demographic information, service usage patterns, customer service interactions, and billing data.
- **Developing Predictive Models:** Utilize advanced machine learning algorithms, including but not limited to logistic regression, decision trees, random forest, and neural networks, to create predictive models capable of accurately forecasting customer churn.
- **Real-time Churn Prediction:** Develop a real-time churn prediction system that can process incoming data streams and provide timely alerts when a customer is exhibiting signs of potential churn. This system should be scalable and capable of handling a large volume of data in real time.
- **Evaluation and Validation:** Rigorously evaluate the developed models using appropriate metrics such as accuracy, precision, recall, and F1-score. Validate the models on unseen data to ensure their effectiveness in predicting churn for new customers and diverse market segments.

DATA DESCRIPTION & DATASET PREPARATION

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 **4250 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.

The **customer churn dataset** from **orange telecom company** contains **20 features and 4250 observations**. The feature **‘Churn’** shows customer **churn** or **non-churn** based on existing conditions. Approximately **14.1%** are **churn** and **85.9%** are **non-churn**. Below table shows **data-set descriptions, Features and Type**.

state	object
account_length	int64
area_code	object
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
number_customer_service_calls	int64
churn	object

FEATURE DESCRIPTION

- **ACCOUNT LENGTH:** It is the length that the customer used their account.
- **STATE :** There are 51 unique state present.
- **AREA CODE :** There are 3 unique area code present.
- **INTERNATIONAL PLAN & VOICEMAIL PLAN :** Both column are described as a categorical feature . yes means plan taken no means plan not taken.
- **NO: OF VOICEMAIL MESSAGES :** The number of voicemail make by the voicemail plan taken customer.
- **TOTAL(DAY/EVE/NIGHT/INTERNATIONAL)(MINUTES/CALLS /CHARGES) :** These are total 12 columns, and all are numerical data types These contain the data of calls, minutes, charges of the customer with respective to the various time of the day and plan.
- **CUSTOMER SERVICE CALLS :** It is the number of calls made by the customer to operator service centre.
- **CHURN :** It is our target dependent variable having boolean data type of true and false.

EXPLORATORY DATA ANALYSIS(EDA)

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**.

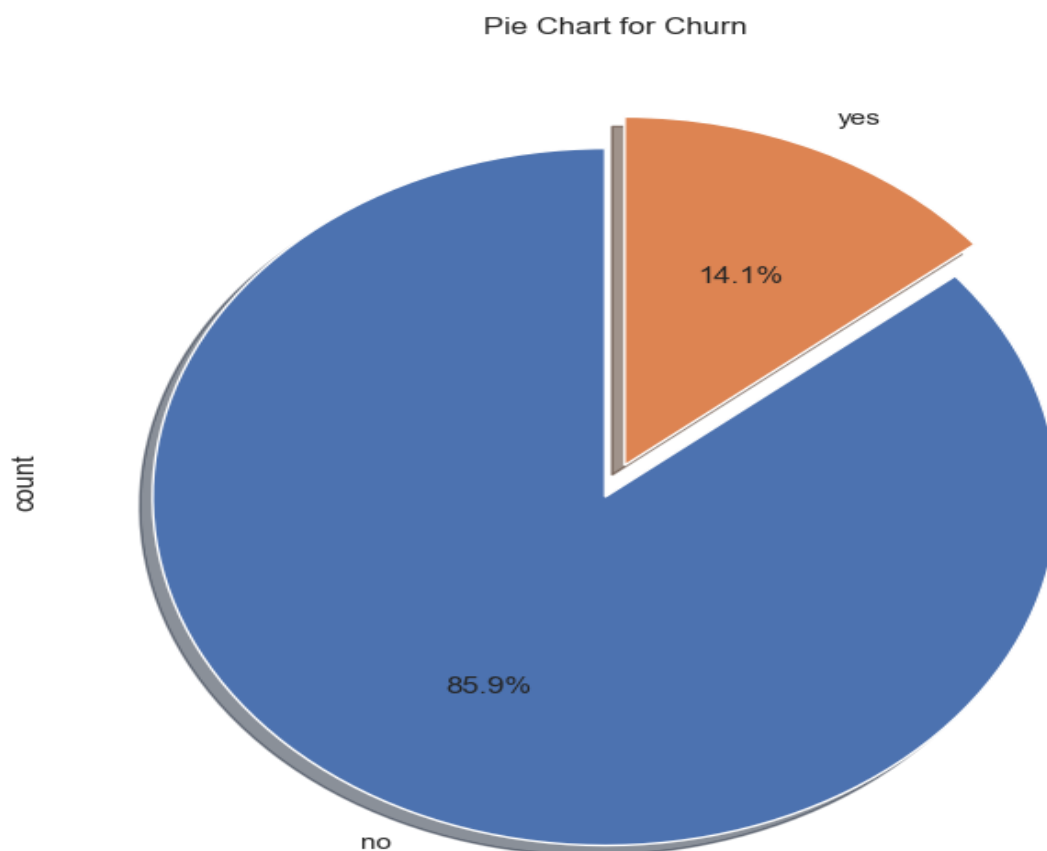
EDA is a process of examining the **available dataset** to discover **patterns**, spot **anomalies**, **test hypotheses**, and check **assumptions** using statistical measures. In this chapter, we are going to discuss the steps involved in performing **topnotch**

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 PREPROCESSING

Data preprocessing is a crucial step in the **customer churn prediction process**. It involves cleaning, transforming, and organizing raw data into a format suitable for **analysis and model training**. In the context of customer churn prediction, data preprocessing plays a **vital role** in ensuring the **accuracy** and effectiveness of **predictive models**.

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**.

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

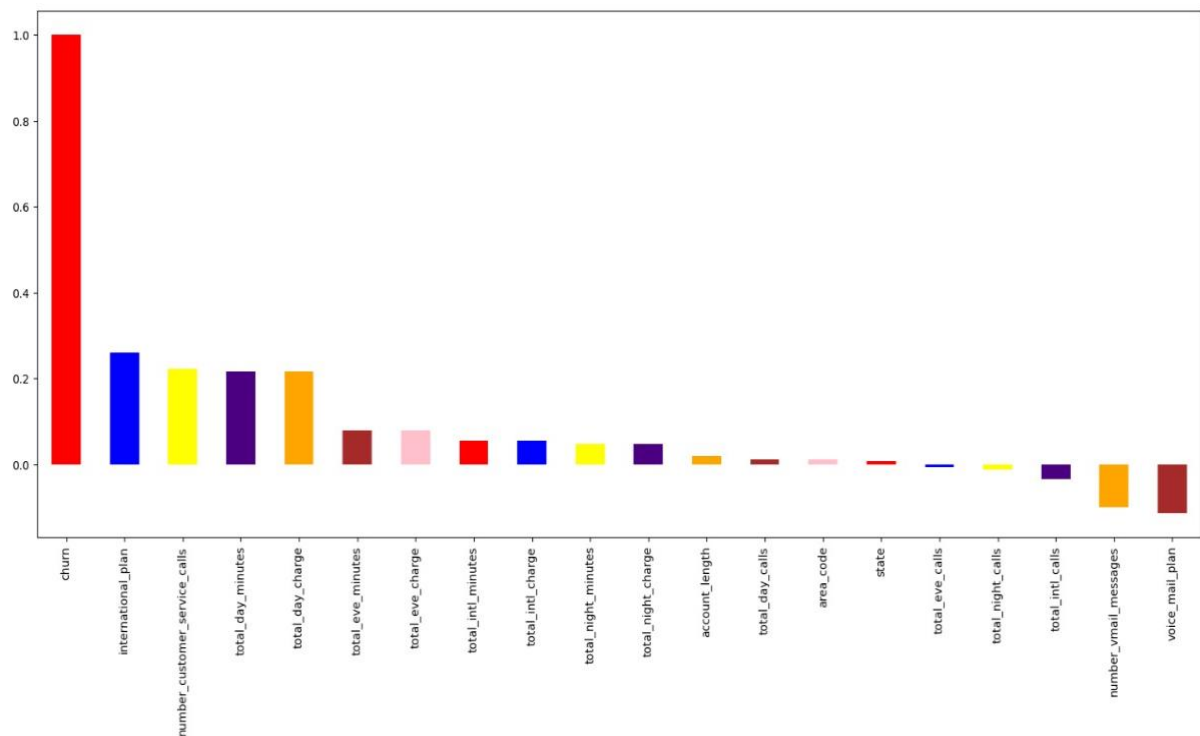
b) Numeric - Categorical Analysis:

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**.

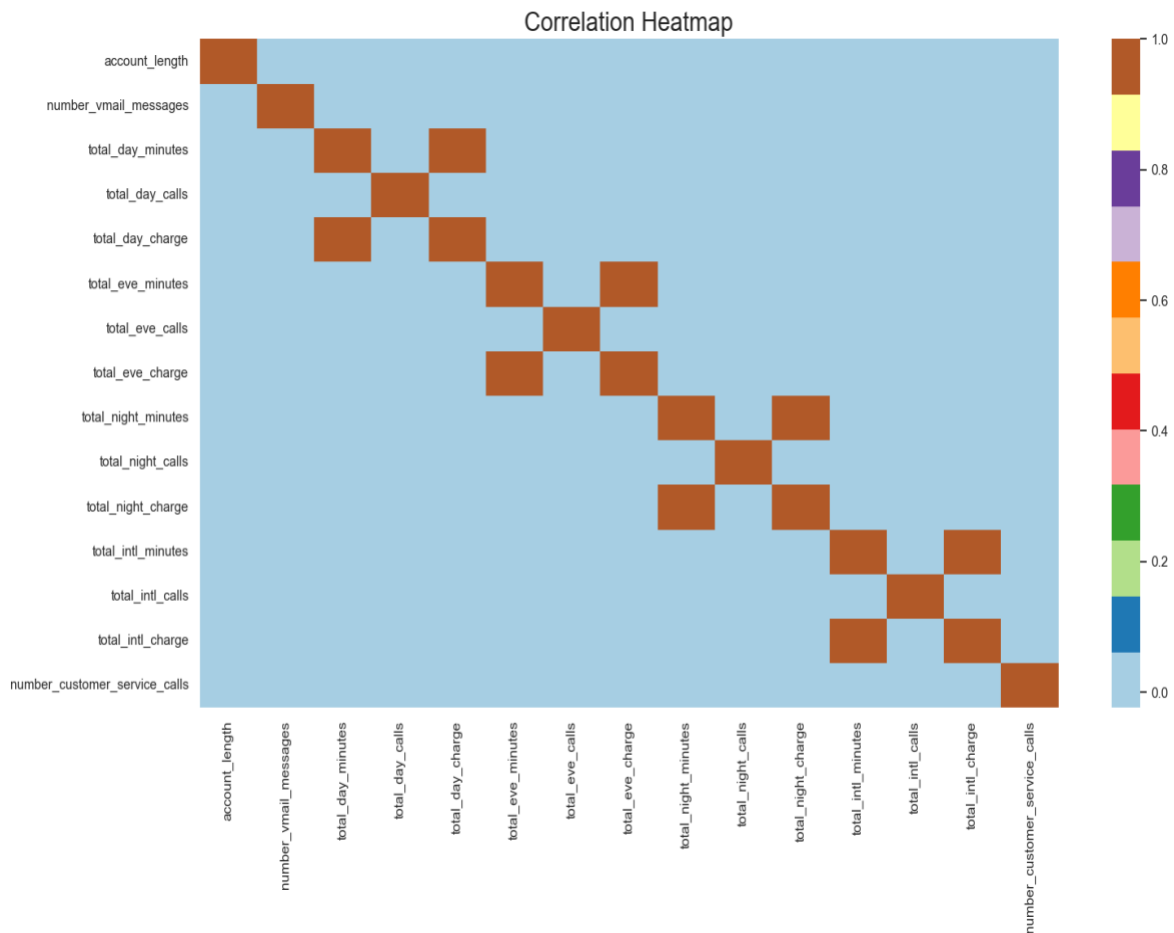
One common way of **plotting multivariate data** is to make a **matrix scatter plot**, known as a **pair plot**. A matrix plot or pair plot shows each pair of variables plotted against each other. The **pair plot** or **Correlation plot** and **Correlation Heatmap** allows us to see both the **distribution**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



CORRELATION AMONG VARIABLES

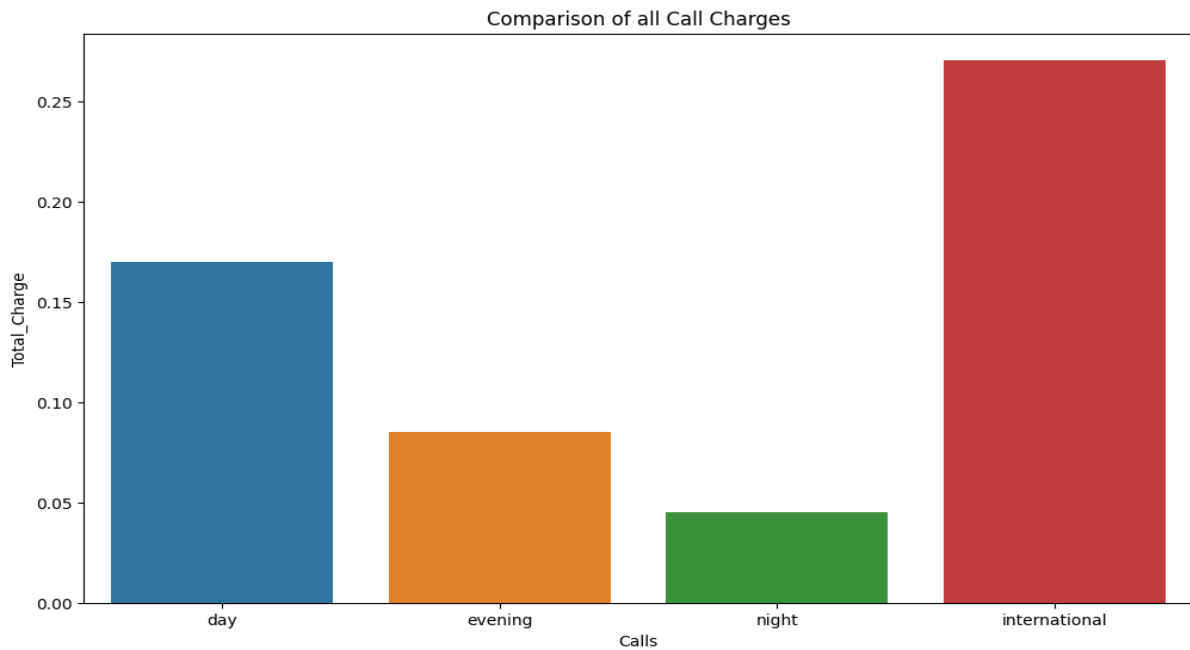
The **statistical technique** that examines the relationship and explains whether, and how strongly, pairs of variables are related to one another is known as correlation. **Correlation HeatMap** answers questions such as how one variable changes with respect to another. Also if the relation between those variables is strong enough, then we can make predictions for future behaviour.

CORRELATION HEATMAP



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.



MODEL SELECTION AND TRAINING

Throughout the project, we employed a range of predictive models, each designed to capture different aspects of customer churn:

- 1) Linear Regression
- 2) K Nearest Neighbors (KNN)
- 3) Random Forest
- 4) Support Vector Machines (SVM)
- 5) Decision Trees
- 6) Neural Networks .

RANDOM FOREST ALGORITHM

Random Forest is a powerful ensemble learning algorithm commonly used for customer churn prediction due to its ability to handle complex relationships in the data and mitigate overfitting.

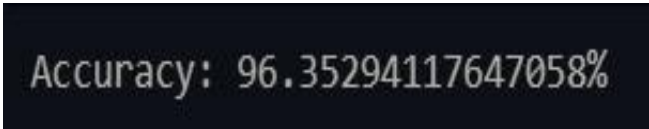
We used **Random Forest Algorithm** , this is because Random Forest achieves **higher accuracy** in **customer churn prediction** compared to other algorithms due to its ability to handle **non-linear** relationships, feature importance analysis, and the ensemble effect.

However, it's important to note that the performance of any **machine learning algorithm**, including **Random Forest**, depends on the specific dataset and problem at hand.

Random Forest might **outperform** other algorithms in a particular dataset, but this is not universally true for all **datasets**.

Therefore, it's essential to perform rigorous experimentation and validation to determine the most suitable algorithm for a specific **customer churn prediction** task.

Finally we have attained **96.35%** of accuracy by using Random Forest Algorithm.



Accuracy: 96.35294117647058%

CONCLUSION

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

- https://book.akij.net/eBooks/2018/May/5aef50939a868/Data_Science_for_Bus.pdf
- **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)**
- <https://bunker2.zlibcdn.com/dtoken/01c5fc197a94283bfb0c0943bd5b2d0c>