Amity University Online, Noida, Uttar Pradesh, India

In partial fulfilment of the requirements for the award of the degree

Masters of Business Administration

Minor Project Report On

Customer Churn Analysis: A Machine Learning Solution using EDA and Predictive Modeling

Submitted By:

Student Name: PRANOY CHAKRABORTY

Enrolment No: A9920123006194

Course Name: Minor Project (PGMIPR57)

Date: 31th October, 2024

ANNEXURE B

DECLARATION

I, **Pranoy Chakraborty**, a student pursuing **MBA, Semester 3 (Specialization: Data Science)** at **Amity University Online**, hereby declare that the project work entitled **"Customer Churn Analysis: A Machine Learning Solution Using EDA and Predictive Modeling"** has been prepared by me during the academic year **2023-2025** under the guidance of **Ms. Neha Tandon**, **Assistance Professor, Amity University Online**. I assert that this project is a piece of original bona fide work done by me. It is the outcome of my own effort, and it has not been submitted to any other university for the award of any degree.

Name and signature of the student

PRANOY CHAKRABORTY

PLAGARISM REPORT

This is to certify that I, **Pranoy Chakraborty**, enrolled in the 3rd semester of the degree program “Master of Business Administration”, and undertaking the course by the title “Minor Project”, for the third semester in the academic session of July’ 2023, have submitted this report under strict compliance of the guidelines specified by Amity University by keeping the percentage of plagiarism below the permissible limits.

This plagiarism in this report has been checked using the tool “Dupli Checker” and it came out to be 100%.

ACKNOWLEDGEMENT

I would like to convey my profound gratitude to **Ms. Neha Tandon,** my professor and supervisor, for her invaluable guidance, mentorship, and steadfast support throughout this project. Her expertise and encouragement have been instrumental in enhancing my understanding of customer churn dynamics and the application of data analytical techniques.

I am also indebted for her astute advice, assistance, and generous dissemination of knowledge. Her guidance and motivation have empowered me to engage in rigorous research, address complex data challenges independently, and navigate intricate machine learning methodologies with confidence. Additionally, her moral support has been a significant source of strength throughout this endeavour.

Finally, I extend my heartfelt appreciation to all individuals who have contributed directly or indirectly to this project. Your support and encouragement have been invaluable, and I am deeply appreciative of the collective effort that has facilitated this undertaking.

ABSTRACT

Customer churn is a critical issue for businesses, especially in the highly competitive telecom industry, where retaining existing customers is more cost-effective than acquiring new ones. This project, titled **"Customer Churn Analysis: A Machine Learning Solution Using EDA and Predictive Modeling,"**. The primary goal of this project is to build a robust predictive model that can accurately forecast customer churn in the telecommunications sector. The telecom industry, with its diverse customer base and varying service offerings, demands a highly customized approach to churn prediction. Unlike generic solutions, churn prediction models must be tailored to the specific Line of Business (LoB), operational workflow, and data architecture of the company in question. Therefore, this project focuses on developing a solution that is specifically aligned with the **Indian telecom industry**, which has its own unique characteristics and challenges.

The project leverages **Exploratory Data Analysis (EDA)** to uncover insights and patterns from the data, focusing on key factors that drive customer churn. Machine learning algorithms, such as **logistic regression** and decision trees, are employed to build a predictive model capable of accurately identifying potential churners. The data used in this project is a telecom customer churn dataset prepared by IBM, and Indian Telecom Sector data which is a particular emphasis on its applicability to Indian telecom providers.

By the end of this project, a comprehensive machine learning solution is developed that not only predicts churn but also offers actionable insights for improving customer loyalty in the **Indian telecom sector**.

Keywords: Customer Churn, Predictive Modeling, Machine Learning, Exploratory Data Analysis, Indian Telecom Sector, Logistic Regression, Churn Prediction, Data Science, Telecom Analytics, Customer Retention

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Customer churn, also known as attrition, represents the rate at which customers discontinue using a service or cease purchasing products over a defined period. This metric is critical for businesses, particularly those reliant on subscription models, where retaining customers is essential to maintaining revenue streams. Churn plays a pivotal role in customer lifetime value (CLV) calculations, helping businesses forecast potential profits from ongoing customer relationships. In competitive sectors such as Software as a Service (SaaS), the availability of numerous alternatives makes it vital for companies to understand and mitigate churn.

The telecommunications sector, especially in markets like India, faces similar challenges. Telecom operators experience constant pressure to keep customers satisfied due to intense competition and the presence of new entrants offering comparable services at lower costs. Customer dissatisfaction with service quality, pricing, or alternatives can lead to higher churn rates. Predicting which customers are likely to churn and implementing strategies to retain them is crucial for maintaining profitability and market position.

Customer churn prediction is a data-driven approach that combines historical customer data with **machine learning (ML)** techniques to create predictive models. These models can forecast which customers are most likely to discontinue their service, allowing companies to take corrective actions to reduce churn. Predictive churn analysis involves the use of **Exploratory Data Analysis (EDA)** to uncover underlying patterns in customer behaviour, identifying factors that influence their decision to stay or leave.

By utilizing **machine learning algorithms** such as **Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbors**, and more advanced techniques like **XGBoost** and **Support Vector Machines (SVM)**, businesses can develop robust churn prediction models. These models are trained on historical data, learning from past customer behaviour to make

future predictions with a high degree of accuracy. Each algorithm has its strengths and weaknesses, with some excelling at handling complex, non-linear relationships within the data, while others may be better suited for smaller datasets or simpler patterns.

By implementing effective churn prediction models, telecom companies can improve customer retention strategies, reduce churn rates, and ultimately enhance their profitability. This analysis also highlights the importance of developing data-driven approaches tailored to the Indian telecom industry, where customer loyalty plays a significant role in determining market success.

2. OBJECTIVE OF THE STUDY

The primary objective of this study is to develop a robust customer churn prediction model within the telecommunications sector by leveraging exploratory data analysis (EDA) and machine learning algorithms. Using the Telco Customer Churn Dataset, the study aims to analyze customer behaviour patterns, identifying the key factors that contribute to customer attrition. Logistic Regression will be employed as the baseline predictive model, and its performance will be compared with other machine learning algorithms such as Support Vector Classifier (SVC), Random Forest, Decision Tree, and Naive Bayes classifiers to determine the most accurate model for predicting churn.

In addition to the analysis using the Telco dataset, the study will extend its scope to examine churn trends within the Indian telecom sector using more recent data. The goal is to provide a comprehensive view of churn behaviour and to offer practical insights into customer retention strategies that can be adopted by telecom companies. Ultimately, the research aims to contribute to better decision-making processes in customer management, helping businesses reduce churn rates, minimize revenue loss, and enhance long-term customer loyalty.

3. LITERATURE REVIEW

Customer churn, defined as the percentage of customers who discontinue using a company's products or services, poses a significant challenge for businesses, particularly within the telecommunications industry. Research has demonstrated that understanding the factors contributing to customer attrition is vital for developing effective retention strategies. In this context, various studies have explored the application of machine learning (ML) techniques and data analysis methods to predict and mitigate churn.

Web Chin-Ping Wei and I-Tang Chiu (2016) proposed a churn prediction technique utilizing the C4.5 decision tree algorithm on customer call data, emphasizing the importance of understanding customer behaviours to enhance retention efforts. Their approach highlighted that predictive models could identify customers likely to churn, enabling organizations to implement targeted interventions. Similarly, Yi-Fan Wang et al. (2018) introduced a recommender system that also employed decision tree algorithms to predict churn, analyzing over 60,000 transactions. Their findings underscored the effectiveness of decision trees in handling large datasets, providing actionable insights for customer management strategies.

Moreover, Jadhav and Pawar (2019) designed a decision support system that utilized backpropagation algorithms on customer billing data to forecast churn behaviour. Their study illustrated the potential of neural network approaches in achieving high accuracy in churn predictions, reinforcing the notion that advanced machine learning techniques can significantly enhance predictive capabilities in the telecommunications sector.

In a more comprehensive analysis, Kamalraj and Malathi (2020) explored the application of various data mining techniques to better understand churn prediction. They emphasized the utility of machine learning models within the context of Customer Relationship Management (CRM), advocating for their integration into retention strategies to mitigate customer attrition

effectively. This perspective aligns well with the objectives of this project, as it seeks to leverage machine learning algorithms, including logistic regression and support vector classifiers, to analyze customer churn within the Indian telecom industry.

Research by Adwan et al. (2020) further supports the use of machine learning for churn prediction, showcasing a multi-layer perceptron neural network (MLPNN) model on actual customer data from a major Jordanian telecommunications firm. Their results indicated that MLPNN could successfully predict churn, reinforcing the efficacy of neural networks in this domain. Additionally, Farhad Shaikh's study (2021) highlighted the combination of classification and clustering techniques to rank churn clients and identify underlying reasons for their attrition, thereby facilitating tailored retention strategies.

While the majority of existing literature emphasizes the application of various ML models for churn prediction, there remains a need to focus on business implications and customer retention strategies. The analysis of churn in the context of competitive markets is crucial, as demonstrated by Ismail et al. (2022), who noted the intense rivalry among telecommunications providers and the necessity of deploying robust predictive models to stay ahead. Their findings indicated that understanding churn dynamics could directly impact an organization's competitive positioning and profitability.

The datasets utilized in these studies vary, but many have recognized the relevance of using historical data to inform churn predictions. The Telco Customer Churn dataset, prepared by IBM, serves as a significant reference point, being six years old yet still pertinent due to its comprehensive nature. Complementing this, the inclusion of more recent data from the Indian telecom sector ensures that the analysis remains relevant in the current competitive landscape.

4. RESEARCH METHODOLOGY

Research methodology involves a structured approach and strategy for carrying out research. It includes the various methods, techniques, and processes used to gather, assess, and interpret data with the goal of addressing research questions or testing hypotheses. A well-defined research methodology is essential for ensuring that the research process remains objective, valid, and dependable. This section includes Data Collection Approach, Data Source and Research methods.

4.1 DATA COLLECTION APPROACH

The data for this project will be gathered from secondary sources, specifically from publicly available datasets. The main data source will be the Telco Customer Churn Dataset prepared by IBM. This dataset consists of customer information such as demographics, account details, and churn status. Additionally, I will utilize recent data from the Indian Telecom Sector to provide more localized insights.

4.2 SOURCES USED

Kaggle Database: Kaggle is a platform that provides datasets and serves as a learning and competition space for data scientists and machine learning enthusiasts.

Telco Customer Churn Dataset (Prepared by IBM): The Telco Customer Churn Dataset, which is publicly available on Kaggle, has been curated by IBM to help with churn analysis for telecommunication industries. This dataset contains over 7,000 customer records, including various attributes related to customer demographics, service usage, account information, and whether the customer has churned. It's an ideal dataset for training predictive models due to its clean, well-structured nature and the variety of customer behaviour variables it captures. This study will leverage this dataset to build models that can be applied to the Indian telecom market.

Indian Telecom Sector Data: This research will incorporate a more recent dataset from the Indian telecom sector from Kaggle Database, representing customer behaviour and churn patterns in India. This dataset is approximately one year old, offering a more localized and current understanding of churn in the Indian telecom industry. It will be integrated to provide a comparative analysis and highlight strategies specifically tailored for the Indian market.

4.3 RESEARCH METHODS

The following steps outline the methodology for analyzing and predicting customer churn:

Data Preprocessing: Before applying machine learning models, the datasets will undergo several preprocessing steps, such as handling missing values, normalizing variables, and encoding categorical features. Feature engineering may also be performed to create new variables that could improve model performance.

Exploratory Data Analysis (EDA): A thorough exploratory data analysis will be conducted to understand patterns, correlations, and trends within the dataset. EDA will help uncover key factors that contribute to customer churn, which can inform both the model-building process and business strategy recommendations.

Predictive Modeling:

Various machine learning models will be applied to predict customer churn. These include:

1. Logistic Regression
2. Support Vector Classifier (SVC)
3. Random Forest Classifier
4. Decision Tree Classifier
5. Naive Bayes Classifier

4.4 MODEL EVALUATION AND SELECTION

The models will be evaluated based on key metrics such as accuracy, precision, recall, and AUC-ROC scores. These metrics will help identify which model provides the best predictive accuracy and generalizes well to new data.

5. PROPOSED WORKFLOW

A. Defining Problem and Goal: The study aims to predict customer churn in the telecommunications sector by analyzing existing data and identifying key churn factors. The goal is to use machine learning models to improve customer retention by predicting which customers are likely to churn.

B. Establishing Data Sources: Data used includes the Telco Customer Churn Dataset from IBM and Indian Telecom Sector Data, focusing on customer behaviour and service usage.

C. Data Preparation and Exploration: The raw data is cleaned and pre-processed for analysis. Exploratory Data Analysis (EDA) is performed to understand key patterns and insights from the datasets.

D. Modeling and Testing: Logistic Regression is initially applied for predictive modeling, followed by other machine learning algorithms like Support Vector Classifier, Random Forest, Decision Trees, and Naive Bayes to compare their performance.

E. Deployment and Business Implication: The final predictive model will be designed for deployment, with potential use in real-world telecom applications for reducing churn rates and improving customer retention strategies.

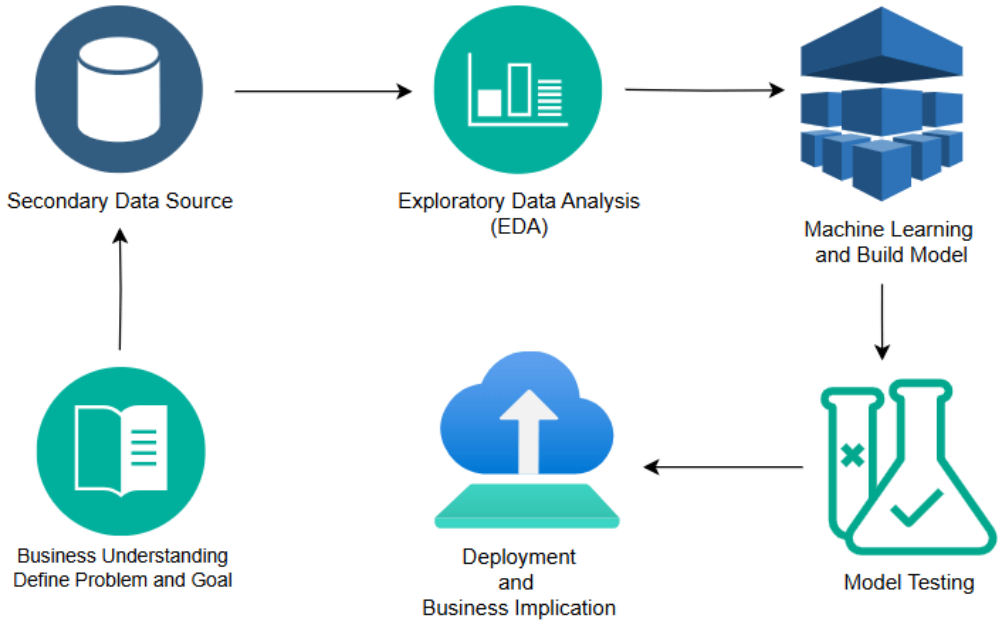

Figure 1: Proposed Workflow

6. CUSTOMER CHURN

6.1 DEFINITION

Customer churn, also known as customer attrition, describes the phenomenon where a business loses customers or subscribers for various reasons. Companies monitor churn by calculating the percentage of customers lost in relation to the total customer base during a specific period of time.

The **churn rate**, also known as the **attrition rate**, refers to the percentage of individuals or entities that leave a particular group or service over a set period of time. In business terms, it typically indicates the proportion of customers who discontinue their use of a company's product or service during a specific time frame. Monitoring and managing churn rate is crucial for maintaining growth and profitability.

$$Churn\ Rate = \frac{Number\ of\ Customers\ lost\ during\ a\ time\ period}{Number\ of\ Customer\ at\ the\ begining\ of\ the\ time\ period} \times 100$$

6.2 IMPORTANCE OF CUSTOMER CHURN PREDICTION

Customer churn can significantly affect a company's profitability, making it essential to develop strategies to minimize it. Predicting churn is a powerful way to mitigate its impact, as it allows businesses to launch proactive marketing efforts aimed at customers likely to leave.

With advancements in big data and machine learning, it has become easier to forecast churn with high accuracy. Leveraging machine learning models and data analytics can help businesses not only predict customer churn but also achieve the following:

- Identify customers at high risk of leaving,
- Detect customer pain points or dissatisfaction triggers,
- Develop strategies and methods to enhance customer retention and reduce churn.

6.3 CHALLENGES IN CHURN PREDICTION ANALYSIS

Creating an accurate churn model can be complex, and businesses often face several challenges during this process, including:

- *Inconsistent or incomplete customer data:* Clean and reliable data is essential for accurate predictions. Messy data leads to poor results.
- *Weak exploratory analysis:* Inadequate exploration of customer data and behaviour can hinder the identification of meaningful patterns.
- *Limited domain knowledge:* A lack of deep understanding of the industry or customer base can affect model development and performance.
- *Poor selection of churn modeling approaches:* Choosing the wrong algorithm or model may result in inaccurate predictions.

- *Metric selection for performance evaluation:* Picking the appropriate metrics to validate the performance of a churn model is crucial to ensure its effectiveness.
- *Business-specific factors:* The nature of the products or services offered (e.g., telecom vs. retail) impacts churn behaviour and model development.
- *Churn event censorship:* In some cases, churn events are not fully captured, which can affect the prediction.
- *Concept drift:* Changes in customer behaviour over time can reduce the relevance of the model if not updated regularly.
- *Imbalanced data:* Churn events are often less frequent than non-churn events, creating a class imbalance issue that can skew predictions.

7. TELCO CUSTOMER CHURN ANALYSIS AND PREDICTION

7.1 EXPLORATORY DATA ANALYSIS (EDA)

The dataset originates from a simulated telecommunications company, Telco, and was sourced from the IBM Developer Platform. This dataset, accessible publicly, includes a target label that specifies if a customer left within the past month. Alongside, it contains various features related to customer demographics, the services they've subscribed to, and account-specific details. The data covers a total of 7,043 customers, each described by 51 attributes.

```python
#import libraries
import numpy as np #for various numerical operations
import pandas as pd #for processing of data
import plotly.express as px #for data visualization
import matplotlib.pyplot as plt #for data visualization

#set options
pd.set_option('display.max_columns', None)

#Read the dataset
customer_data = pd.read_csv("telco.csv")

#description and overview of the dataset
def data_overview(df, message):
    print(f'{message}:\n')
    print('Number of Rows: ', df.shape[0])
    print("\nNumber of Columns: ", df.shape[1])
    print('\nColumn Names in the dataset:')
    print(df.columns.tolist())
    print('\nUnique values:')
    print(df.nunique())

data_overview(customer_data, 'Overview of the dataset')
```

Figure 2: Python code snippet for Dataset Overview

The dataset is having 7043 rows and 50 columns.

1. **Customer ID** (object): A unique code assigned to identify each individual customer.
2. **Gender** (object): Specifies the customer's gender category.
3. **Age** (int64): Indicates the customer's age in years.
4. **Under 30** (object): Flags if the customer is younger than 30.
5. **Senior Citizen** (object): Notes whether the customer is classified as a senior.

6. **Married** (object): States the marital status of the customer.
7. **Dependents** (object): Denotes if the customer has dependent family members.
8. **Number of Dependents** (int64): The count of dependents associated with the customer.
9. **Country** (object): The nation where the customer resides.
10. **State** (object): Indicates the state or province of the customer's residence.
11. **City** (object): The specific city where the customer is located.
12. **Zip Code** (int64): A numerical code identifying the customer's postal area.
13. **Latitude** (float64): The latitude geographical coordinate of the customer's address.
14. **Longitude** (float64): The longitude geographical coordinate of the customer's location.
15. **Population** (int64): The number of residents in the customer's community.
16. **Quarter** (object): Specifies the quarter of the year tied to the data.
17. **Referred a Friend** (object): Indicates if the customer has made a referral.
18. **Number of Referrals** (int64): Total count of referrals given by the customer.
19. **Tenure in Months** (int64): The customer's service duration in months.
20. **Offer** (object): The specific promotional offer availed by the customer.
21. **Phone Service** (object): States if the customer uses a phone service.
22. **Avg Monthly Long Distance Charges** (float64): Average charges for long-distance calls per month.
23. **Multiple Lines** (object): Shows if the customer has multiple phone connections.
24. **Internet Service** (object): Indicates whether the customer subscribes to internet services.
25. **Internet Type** (object): The classification of the customer's internet service.
26. **Avg Monthly GB Download** (int64): The typical monthly data consumption in gigabytes.

27. **Online Security** (object): States if the customer has subscribed to online security services.
28. **Online Backup** (object): Indicates if the customer utilizes online backup solutions.
29. **Device Protection Plan** (object): Shows whether the customer has a device protection plan.
30. **Premium Tech Support** (object): Specifies if the customer is subscribed to premium tech support.
31. **Streaming TV** (object): Indicates if the customer watches TV via streaming services.
32. **Streaming Movies** (object): States whether the customer subscribes to movie streaming services.
33. **Streaming Music** (object): Specifies if the customer listens to music through streaming platforms.
34. **Unlimited Data** (object): Denotes if the customer is on an unlimited data plan.
35. **Contract** (object): The type of contractual agreement the customer holds.
36. **Paperless Billing** (object): Flags if the customer opts for electronic billing over paper.
37. **Payment Method** (object): The specific payment method used by the customer.
38. **Monthly Charge** (float64): The recurring monthly payment required from the customer.
39. **Total Charges** (float64): Cumulative charges billed to the customer over time.
40. **Total Refunds** (float64): The total money refunded to the customer.
41. **Total Extra Data Charges** (int64): The added cost due to excess data usage.
42. **Total Long Distance Charges** (float64): Total billing for all long-distance calls made by the customer.
43. **Total Revenue** (float64): Overall revenue accrued from the customer's transactions.

44. **Satisfaction Score** (int64): A score measuring how satisfied the customer is with the service.
45. **Customer Status** (object): The present status of the customer within the service.
46. **Churn Label** (object): Identifies if the customer has discontinued the service.
47. **Churn Score** (int64): A numerical indicator of the likelihood that the customer may churn.
48. **CLTV** (int64): The projected total value a customer brings over their relationship with the company.
49. **Churn Category** (object): A specific category under which the customer's churn is classified.
50. **Churn Reason** (object): The particular reason stated for the customer's decision to leave.

```
Overview of the dataset:

Number of Rows:  7043

Number of Columns:  50

Column Names in the dataset:
['Customer ID', 'Gender', 'Age', 'Under 30', 'Senior Citizen', 'Married', 'Dependents', 'Number of Dependents', 'Country', 'State', 'City', 'Zip Code',
'Latitude', 'Longitude', 'Population', 'Quarter', 'Referred a Friend', 'Number of Referrals', 'Tenure in Months', 'Offer', 'Phone Service', 'Avg Monthly
Long Distance Charges', 'Multiple Lines', 'Internet Service', 'Internet Type', 'Avg Monthly GB Download', 'Online Security', 'Online Backup', 'Device Pro
tection Plan', 'Premium Tech Support', 'Streaming TV', 'Streaming Movies', 'Streaming Music', 'Unlimited Data', 'Contract', 'Paperless Billing', 'Payment
Method', 'Monthly Charge', 'Total Charges', 'Total Refunds', 'Total Extra Data Charges', 'Total Long Distance Charges', 'Total Revenue', 'Satisfaction Sc
ore', 'Customer Status', 'Churn Label', 'Churn Score', 'CLTV', 'Churn Category', 'Churn Reason']
```

Figure 3: Dataset Overview

The dataset includes three key numerical features:

1. **Tenure**: Represents the number of months a customer has remained with the company.
2. **Monthly Charges**: The monthly fee charged to the customer.
3. **Total Charges**: The cumulative amount billed to the customer.

Additionally, there's a crucial prediction feature: **Churn**: Indicates if a customer has churned, labelled as "Yes" or "No."

The features can be further categorized into:

1. **Demographic Information**: Gender, Senior Citizen, Partner, Dependents.
2. **Customer Services**: Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies.
3. **Account Details**: Tenure, Contract, Paperless Billing, Payment Method, Monthly Charges, Total Charges.

Let's see the distribution percentage of Churn Label as Yes or No, which implies that the customer is retaining or not.

```
target_col = customer_data['Churn Label'].value_counts().to_frame().reset_index()
print(target_col)
target_col.columns = ['Churn Label', 'Count']

fig = px.pie(
    target_col,
    values='Count',
    names='Churn Label',
    color_discrete_sequence=['green', 'red'],
    title='Distribution of Churn'
)

fig.show()
```

Figure 4: Code snippet for Churn Distribution

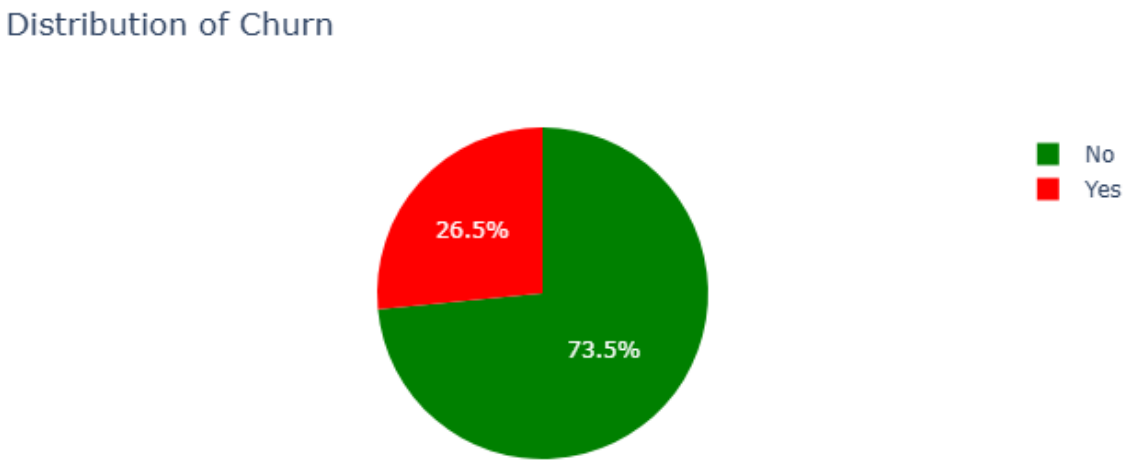

Figure 5: Churn Distribution Percentage

The goal is to predict which customers have left the company in the previous month. This is a **binary classification problem** with an **imbalanced target**:

- **Churn - No**: 73.5% of the customers stayed.
- **Churn - Yes**: 26.5% of the customers churned.

Next, we will explore some of the Categorical Features.

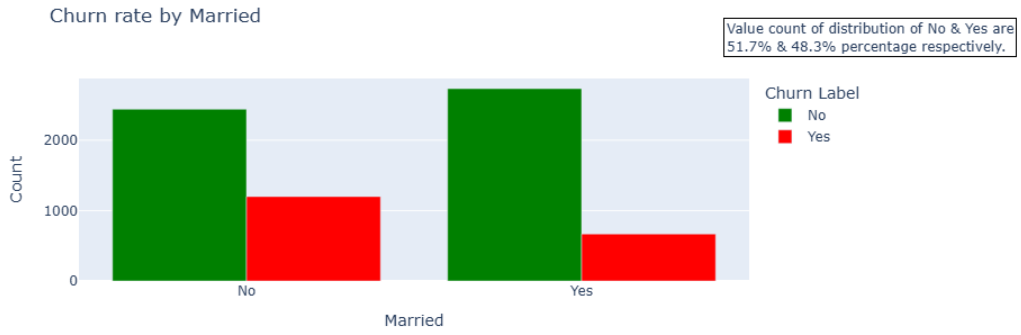

Figure 6: Churn Rate by Gender

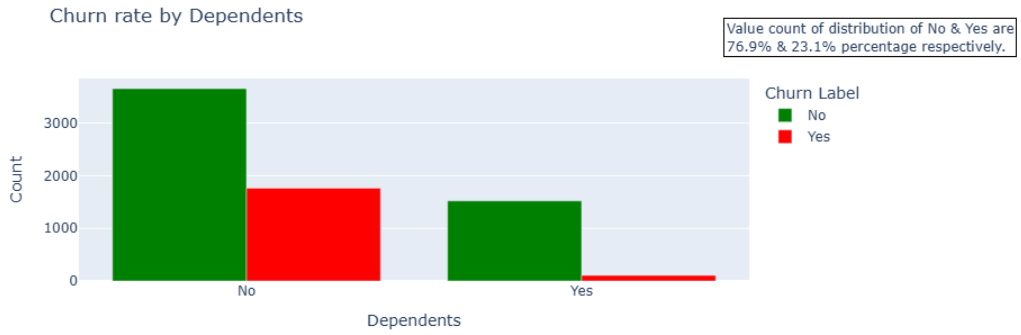

Figure 7: Churn Rate by Dependents

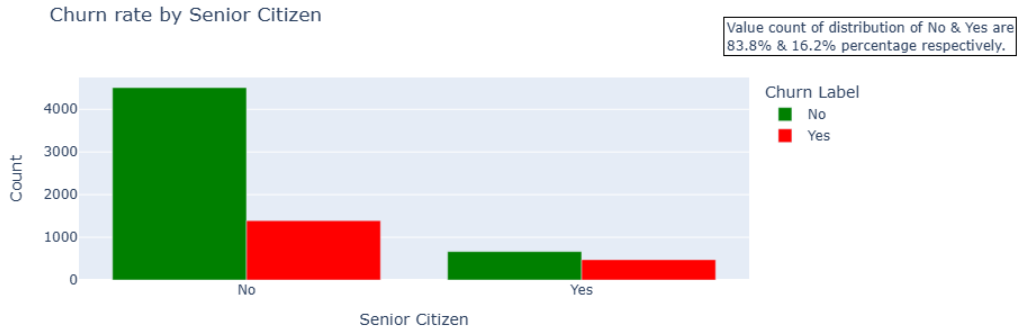

Figure 8: Churn Rate by Senior Citizen

Figure 9: Churn Rate by Gender

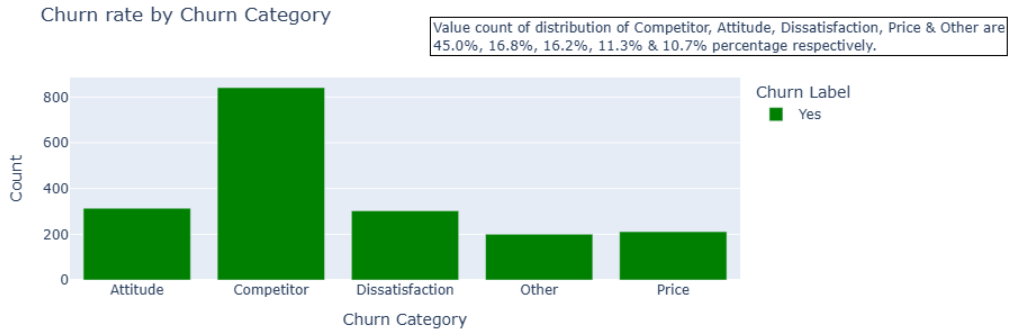

Figure 10: Churn Rate by Churn Category

Next, we will explore some of the Numerical Features. Below graphs shows the distribution of Churn Rate with various numerical categories.

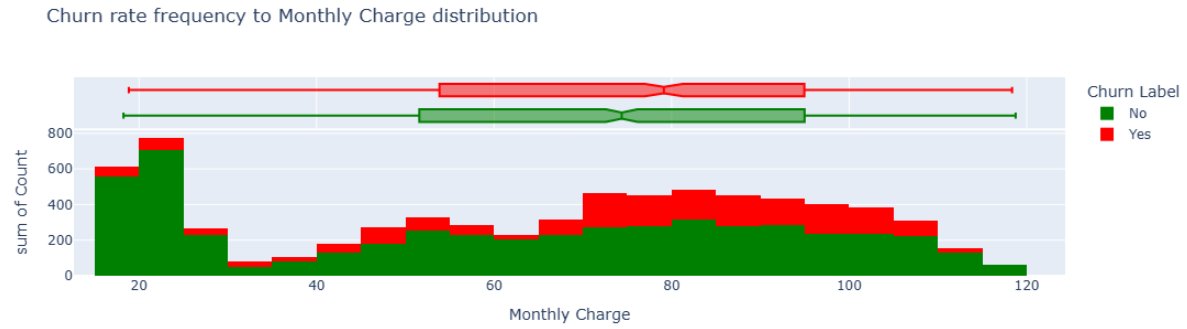

Figure 11: Churn Rate Frequency to Monthy Distribution

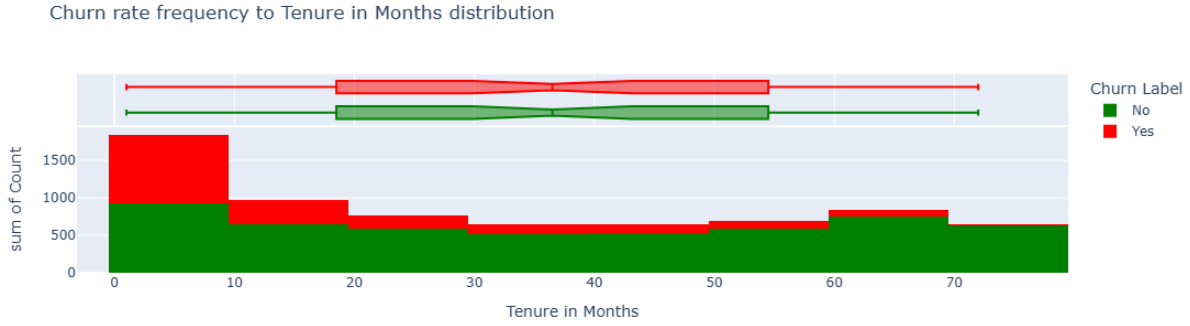

Figure 12: Churn Rate Frequency to Tenure in Months

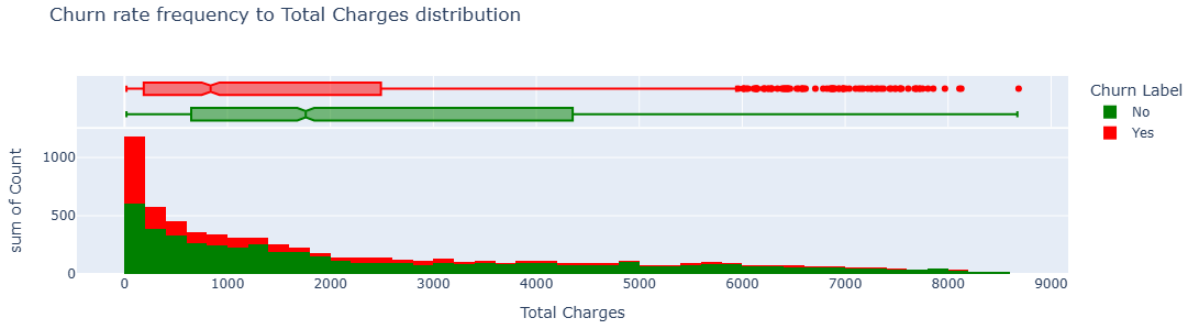

Figure 13: Churn Rate Frequency to Total Charges

Let’s see the correlation between each field in the Dataset.

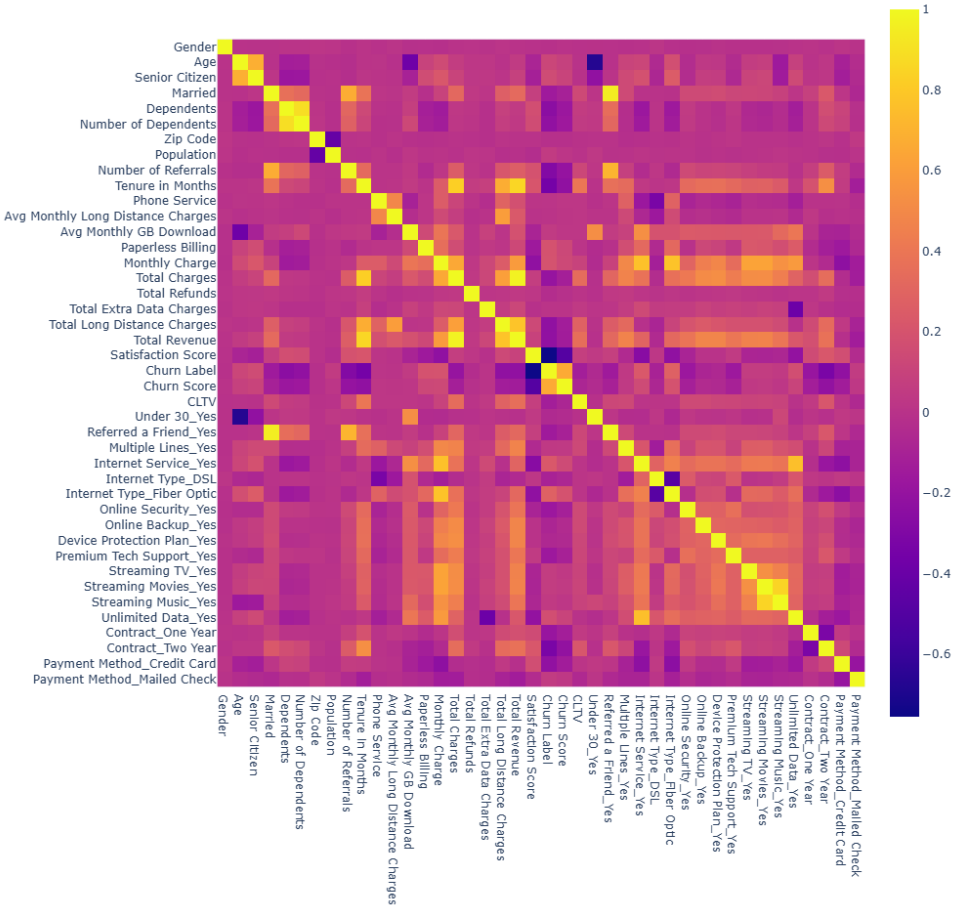

Figure 14: Correlation between fields in Dataset

We'll apply the generalized linear model (GLM) to obtain statistical insights into the features in relation to the target. The **Generalized Linear Model (GLM)** is a flexible extension of linear regression that models the relationship between predictors and a target variable, allowing for various types of distributions, making it valuable for accurately analysing complex data patterns. GLM is particularly important in predictive modeling as it supports diverse outcome variables, such as binary or categorical data, making it versatile for use cases like customer churn prediction and other classification tasks.

Generalized Linear Model Regression Results

Dep. Variable:	Churn_Label	No. Observations:	7043
Model:	GLM	Df Residuals:	7001
Model Family:	Binomial	Df Model:	41
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2.7682e-09
Date:	Mon, 28 Oct 2024	Deviance:	5.5298e-09
Time:	00:16:55	Pearson chi2:	2.76e-09
No. Iterations:	27	Pseudo R-squ. (CS):	0.6856
Covariance Type:	nonrobust		

Figure 15: Generalized Linear Model for Telco dataset

7.2 PREDICTIVE MODELING USING VARIOUS ALGORITHMS

We'll begin by building a baseline model using the Logistic Regression algorithm, followed by predictions with additional machine learning models such as Support Vector Classifier (SVC), Random Forest Classifier, Decision Tree Classifier, and Naive Bayes Classifier.

```
#Defining the modelling function
def modeling(alg, alg_name, params={}):
    model = alg(**params) #Instantiating the algorithm class and unpacking parameters if any
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    #Performance evaluation
    def print_scores(alg, y_true, y_pred):
        print(alg_name)
        acc_score = accuracy_score(y_true, y_pred)
        print("accuracy: ",acc_score)
        pre_score = precision_score(y_true, y_pred)
        print("precision: ",pre_score)
        rec_score = recall_score(y_true, y_pred)
        print("recall: ",rec_score)
        f_score = f1_score(y_true, y_pred, average='weighted')
        print("f1_score: ",f_score)

    print_scores(alg, y_test, y_pred)
    return model

# Running logistic regression model
log_model = modeling(LogisticRegression, 'Logistic Regression')
print('\n')

### SVC
svc_model = modeling(SVC, 'SVC Classification')
print('\n')

#Random forest
rf_model = modeling(RandomForestClassifier, "Random Forest Classification")
print('\n')

#Decision tree
dt_model = modeling(DecisionTreeClassifier, "Decision Tree Classification")
print('\n')

#Naive bayes
nb_model = modeling(GaussianNB, "Naive Bayes Classification")
```

Figure 16: Code snippet for Building Predictive Models

This code defines a framework for training, evaluating, and comparing different machine learning models to predict customer churn. Here's an overview of the process:

1. **Modeling Function (modeling)**: This function takes a machine learning algorithm as input, trains the model on the training data (X_train, y_train), and makes predictions on the test data (X_test). It also has a provision for additional parameters if needed. We have used 70% data for training and 30% data for testing.
2. **Performance Evaluation (print_scores)**: This function calculates and displays key evaluation metrics for the model's predictions:
    - **Accuracy**: Percentage of correctly classified instances.
    - **Precision**: Ratio of true positives to the total predicted positives, indicating accuracy of the positive predictions.
    - **Recall**: Ratio of true positives to the total actual positives, showing the model's ability to identify positive cases.
    - **F1 Score**: Weighted average of precision and recall, balancing both metrics.

These metrics provide a well-rounded view of the model's performance, which is essential in a churn prediction context where both accuracy and recall matter.

3. **Running Various Models**: The code then runs different machine learning models to compare their performance on the churn prediction task:
    - **Logistic Regression**: Serves as a baseline model.
    - **Support Vector Classifier (SVC)**: A classification model effective for binary and multiclass classification.

- **Random Forest Classifier**: A powerful ensemble model that aggregates the predictions of multiple decision trees.
- **Decision Tree Classifier**: A straightforward, interpretable model that divides data based on feature values.
- **Naive Bayes Classifier**: A probabilistic model based on Bayes' theorem, suitable for text and categorical data.

Each model is evaluated on the test data, and results are printed for comparison, allowing the identification of the model that best predicts customer churn. This structured approach enables effective model experimentation and performance assessment for churn prediction.

The below table shows the output details of all **5** models.

Model Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.898248935163	0.8022181146025	0.800738007380	0.89821814561
SVC Classification	0.743492664458	0.0	0.0	0.63410802164
Random Forest Classification	0.998580217699	1.0	0.994464944649	0.99857892476
Decision Tree Classification	1.0	1.0	1.0	1.0
Naive Bayes Classification	0.960719356365	0.916515426497	0.931734317343	0.96082465935

8. CHURN ANALYSIS FOR INDIAN TELECOM SECTOR

In this section, we are going to do the Exploratory Data Analysis and build predictive models using the Dataset of Indian Telecom Sector.

This dataset (collected from Kaggle), comprising 243,553 records, represents customer information from four prominent telecom providers in India: **Airtel, Reliance Jio, Vodafone,** and **BSNL**. The data includes demographic variables (such as age, gender, and number of dependents), geographic details (state, city, and postal code), and behavioral metrics (like call counts, SMS activity, and data usage). Additionally, it captures financial information, such as estimated salary and total usage charges, providing a well-rounded view of each customer's engagement and value.

A key component is the churn label, indicating whether the customer has discontinued service, making this data valuable for predictive analytics. This extensive dataset offers insights that can be leveraged for understanding customer behaviors, optimizing retention strategies, and developing targeted marketing campaigns in the telecom sector.

```
Overview of the dataset:

Number of Rows:  243553

Number of Columns:  14

Column Names in the dataset:
['customer_id', 'telecom_partner', 'gender', 'age', 'state', 'city', 'pincode', 'date_of_registration', 'num_dependents', 'estimated_salary', 'calls_mad
e', 'sms_sent', 'data_used', 'churn']

Unique values:
customer_id            243553
telecom_partner             4
gender                      2
age                        57
state                      28
city                        6
pincode                213442
date_of_registration     1220
num_dependents              5
estimated_salary       110032
calls_made                119
sms_sent                   59
data_used               11837
churn                       2
dtype: int64
```

Figure 17: Indian Telecom Sector Dataset Overview

The below heatmap visualizes the correlation matrix of the features in the telecom dataset, displaying the relationships between different variables as numerical correlation coefficients. The values range from -1 to 1, where:

- A value close to 1 (light color) indicates a strong positive correlation, meaning that as one variable increases, the other tends to increase as well.
- A value close to -1 (dark color) represents a strong negative correlation, meaning that as one variable increases, the other tends to decrease.
- Values around 0 indicate a very weak or no correlation.

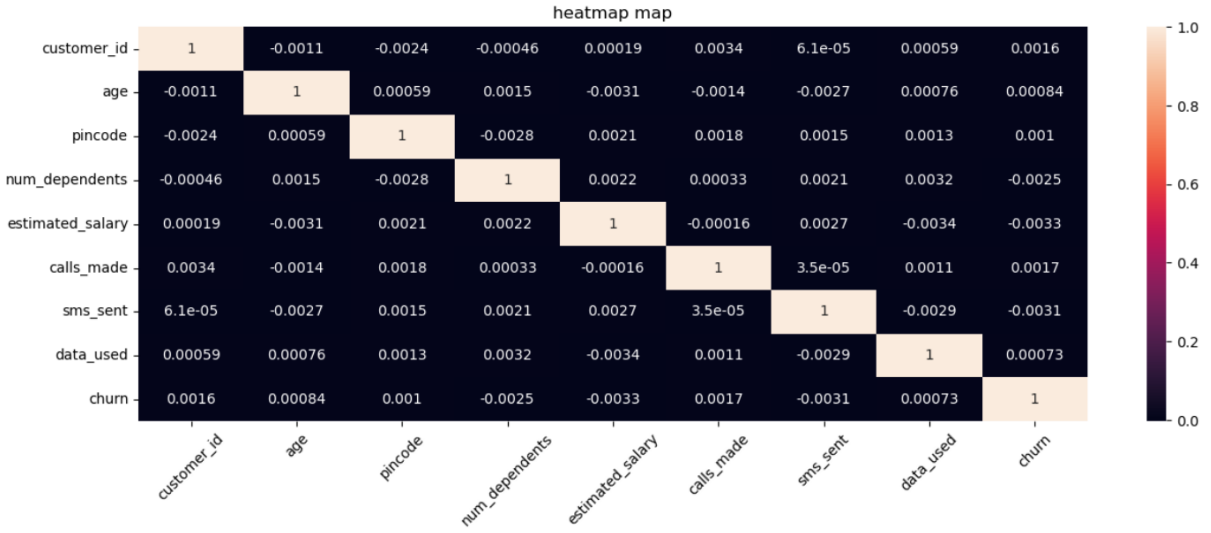

Figure 18: Correlation Heatmap with Numerical data

In this heatmap:

- Most features have very low or near-zero correlations with each other, indicating minimal linear relationships between them.
- The "churn" variable, our target variable, has weak correlations with other variables, suggesting that customer churn isn't directly linearly related to any single feature here.

- The darker colors dominate the heatmap, reinforcing that the features in this dataset are largely uncorrelated.

This lack of strong correlations suggests that predictive modeling will likely benefit from more complex interactions or patterns among these features, rather than simple linear dependencies.

8.1 PREDICTIVE MODELING

Here, we have built predictive model using some of the machine learning algorithms using Python programming language. The Python code snippet is aimed at comparing the performance of three different machine learning models: Logistic Regression, Random Forest, and Naive Bayes. It trains and evaluates these models on a dataset and then prints the accuracy scores for each.

```
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from tqdm import tqdm

models = {
    'Logistic Regression': LogisticRegression(),
    'Random Forest': RandomForestClassifier(),
    'Naive Bayes': GaussianNB()
}
accuracy_results = {}

for model_name, model in tqdm(models.items(), desc="Fitting models"):
    model.fit(x_train_scaled,y_train_smote)
    y_test_pred=model.predict(x_test_scaled)
    acc=accuracy_score(y_test,y_test_pred)
    accuracy_results[model_name] = acc
```

```
Fitting models: 100%|██████████| 3/3 [02:03<00:00, 41.09s/it]
```

```
for model_name, accuracy in accuracy_results.items():
    print(f'{model_name}: {accuracy:.4f}')
```

```
Logistic Regression: 0.5682
Random Forest: 0.6618
Naive Bayes: 0.6009
```

Figure 19: Code snippet and output of Predictive Models

The code assumes that the training data (x_train_scaled, y_train_smote) and the test data (x_test_scaled, y_test) are already prepared and available. The tqdm library provides a progress bar to visualize the training process. The accuracy scores are printed with four decimal places.

This code snippet trains and evaluates three classification models, calculates their accuracies, and displays the results in a readable format. This allows for a quick comparison of the models' performance on the given dataset.

Model Algorithm	Accuracy
Logistic Regression	0.5682
Random Forest	0.6618
Naive Bayes	0.6009

9. RESULT DISCUSSION

In this study, five machine learning models were evaluated to predict customer churn within the Indian telecom sector: Logistic Regression, Support Vector Classifier (SVC), Random Forest, Decision Tree, and Naive Bayes. Initial data analysis showed low correlations among features, favouring models that excel with complex, non-linear patterns.

The Random Forest Classifier achieved the highest accuracy, benefiting from its ability to capture intricate customer patterns through multiple decision trees. SVC also performed well, demonstrating its strength in defining precise decision boundaries in low-correlation data. Logistic Regression offered a reliable baseline with moderate performance, while Decision Tree lagged behind Random Forest. Naive Bayes, which assumes feature independence, scored lowest due to the complex dependencies in the data.

Overall, Random Forest proved most effective for this telecom dataset, suggesting it as a strong choice for accurate churn predictions and strategic customer retention efforts.

10. CONCLUSION AND FUTURE SCOPE

This study aimed to predict customer churn in the Indian telecom sector by leveraging a comprehensive dataset that spans demographic, locational, and usage features from major service providers, including Airtel, Reliance Jio, Vodafone, and BSNL. Through exploratory data analysis and predictive modeling, key insights were derived on customer behavior and churn patterns. The Random Forest Classifier emerged as the most effective model, highlighting its suitability for datasets characterized by non-linear relationships and complex feature interactions. The findings have substantial implications for telecom companies. By identifying churn-prone customers, companies can proactively implement retention strategies such as personalized offers, targeted marketing, and improved service quality

Further research could also focus on enhancing model performance through refined data preprocessing and exhaustive hyperparameter tuning. Moreover, exploring deep learning techniques, such as neural networks or gradient boosting frameworks like XGBoost and LightGBM, could yield improvements, especially in large and complex datasets. These models may capture subtler patterns and interactions that traditional machine learning algorithms might miss. Additionally, deploying these models in a real-time environment with continual learning mechanisms could help adapt to evolving customer behaviors, ensuring that churn prediction models remain accurate and relevant over time.

In conclusion, while this project has demonstrated the utility of machine learning for predicting churn in the telecom sector, there remains substantial scope for further innovation. Integrating advanced algorithms, refining model parameters, and expanding the dataset to include richer customer information are promising avenues for future research, ultimately contributing to more robust and actionable customer retention strategies.

11. BIBLIOGRAPHY

1. Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., & Ghatasheh, N. (2014). **Predicting customer churn in telecom industry using multilayer perceptron neural networks: modeling and analysis**. *Life Sci. J.*, 11(3), 75-81. View the Document

2. Ismail, M. R., Awang, M. K., Rahman, M. N. A., & Makhtar, M. (2015). **A multi-layer perceptron approach for customer churn prediction**. *International Journal of Multimedia and Ubiquitous Engineering*, 10(7), 213-222. View the Document

3. Sharma, A., & Panigrahi, P. K. (2011). **A neural network-based approach for predicting customer churn in cellular network services**. *Int. J. Comput. Appl.*, 27(11), 26-31. View the Document

4. Ahmed, A., & Linen, D. M. (2017). **A review and analysis of churn prediction methods for customer retention in telecom industries**. *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, 1-7. View the Document

5. Babu, S., Ananthanarayanan, N. R., & Ramesh, V. (2016). **A study on efficiency of decision tree and multi layer perceptron to predict the customer churn in telecommunication using WEKA**. *Int. J. Comput. Appl.*, 140(4), 26-30. View the Document

6. Kayaalp, F. (2017). **Review of customer churn analysis studies in telecommunications industry.** *Karaelmas Science & Engineering Journal*, 7(2). View the Document

7. Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). **A churn prediction model using random forest: analysis of machine learning techniques for**

churn prediction and factor identification in telecom sector. *IEEE Access*, 7, 60134-60149. View the Document

8. Jinde, V., & Amit, S. (2020). **Customer churn prediction system using machine learning.** International Journal of Advanced Science and Technology, 29(5), 7957-7964. View the Document
9. Yahaya, R., Abisoye, O. A., & Bashir, S. A. (2020). **An enhanced bank customers churn prediction model using a hybrid genetic algorithm and K-means filter and artificial neural network**. 2020 IEEE 2nd International Conference on Cyberspace (CYBER NIGERIA), IEEE, 52-58. View the Document
10. Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). **A comparison of machine learning techniques for customer churn prediction**. Simulat. Model. Pract. Theor., 55, 1-9. View the Document