CUSTOMER CHURN PREDICTION FOR TELECOM INDUSTRY

MINI PROJECT REPORT

Submitted By

SHEETHAL R	**211501096**
SRI BALAJI S	**211501102**
VEJAYSUNDARAM R	**211501116**

In partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

RAJALAKSHMI
ENGINEERING COLLEGE
An AUTONOMOUS Institution
Affiliated to ANNA UNIVERSITY, Chennai

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI-600 025

NOVEMBER 2024

RAJALAKSHMI ENGINEERING COLLEGE

BONAFIDE CERTIFICATE

Certified that this Report titled **“Customer Churn Prediction For Telecom Industry”** is the bonafide work of **“211501096-Sheethal R, 211501102-Sri Balaji S, 211501116-Vejaysundaram R”** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Mrs. D. Sorna Shanthi M.Tech.,

Associate Professor

Department of Artificial Intelligence and Data Science

ABSTRACT

In this project, we aimed to understand the factors influencing customer churn and to develop targeted retention strategies using a comprehensive dataset. The dataset includes detailed demographic information, usage patterns of services, contract types, and financial metrics. Our approach began with extensive Exploratory Data Analysis (EDA) to uncover initial insights. We employed visualizations such as histograms, box plots, and scatter plots to analyze key attributes like age, tenure, contract type, data usage, and Customer Lifetime Value (CLTV) in relation to churn scores. Through this analysis, we identified churn-prone customers as new market entrants, those with short contract durations, and high data usage customers. Additionally, we found that high CLTV customers and those with higher charges exhibit potential dissatisfaction, leading to churn.

To further enhance our understanding and prediction capabilities, we developed a machine learning model using a Random Forest classifier. We trained and validated this model on our dataset, employing hyperparameter tuning with GridSearchCV and cross-validation to optimize its performance. Feature importance analysis using SHAP values allowed us to pinpoint the most critical factors contributing to churn. Our analysis provided actionable insights, highlighting the need for tailored retention strategies. For instance, we recommended personalized offers and discounts for high-risk customers, improved customer service for those dissatisfied with service quality, and more flexible contract options to reduce churn rates. The results of this project enable better decision-making regarding service improvements, pricing adjustments, and targeted retention efforts, ultimately enhancing customer loyalty and reducing churn.

Keywords: Customer, EDA(Exploratory Data Analysis), CLTV(customer lifetime value), Churn

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

In today's aggressively competitive marketplace, retaining existing customers is as important as finding new ones, if not more so. Customer churn-this is the rate at which customers cancel their subscriptions or services-could severely affect a company's top line, especially if it consists of high-value or long-term customers. The project aims at analyzing an extensive dataset of customer demographics, service usage patterns, contract details, and financial metrics for the purpose of addressing the problem of customer attrition. Understanding such factors will allow businesses to create targeted strategies for enhanced customer retention, improved service satisfaction, and optimal overall business growth. It begins with EDA, where techniques such as histograms, box plots, and scatter plots are used to ascertain patterns and relationships between attributes of a customer and churn rates. A few of the metrics examined in this context include age, tenure, contract type, monthly data usage, Customer Lifetime Value (CLTV), and monthly charges. From the analysis, it follows that a great proportion of churn is usually seen in the newer customers, those on short-term contracts, and the heavy data consumers. Not to mention, high CLTV customers as well as customers with strong monthly charges might become dissatisfied, thus posing more risk for the attrition. Recognizing the importance of accurate churn prediction, the current project further enhances the churn score metric by measuring its accuracy in predicting a real churn and checking if it could also be useful in proactive retention measures. Evaluation models as well as predictive models enable a firm to detect and resolve high-risk customers early on, thus applying retention interventions targeted at the needs of such customers. This contributes not only to reducing revenue loss but also enhancing customer satisfaction and loyalty-a huge constituent in sustaining a competitive business strategy that continues to remain within manageable financial constraints. The significance of this project goes towards underlying ongoing customer relationship management improvements to further use data-driven approaches for customer loyalty and long-term growth of companies.

Overview of the Problem Statement:

Hence, customer retention has become a very essential ingredient to maintain profitability and growth in organizations within this rapidly competitive environment of businesses nowadays. Customer churn or discontinuance of services by customers is definitely an enormous threat for organizations as it directly affects their revenue flow and may pose a significant threat to the organization's market position particularly when loss of high value loyal customers took place. This dataset that is robust with various customer attributes, such as demographics, usage patterns of services, type of contract, and financial metrics, helps in understanding and predicting customer churn in this project. High churn rates may indicate the possibility of dissatisfaction among customers, which may perhaps be because of high charges, non-fulfillment of service expectations, or possibly just limited flexibility in contracts. To understand the drivers of churn for a business is both a means of holding onto valuable customers and improving services, experiences, and overall customer satisfaction. In fact, acquiring a new customer often costs more than retaining an existing one, thus minimizing churn will help better utilize company resources and raise CLTV. The objective of this project would, therefore be to deliver actionable insights with predictive models about which customers could pose risks so retention could target those customers. Pricing adjustments can be also more data-informed, and customer service practices enhanced. These factors, in turn, help a business fortify its customer loyalty, bolster sustainable revenues, and achieve a competitive advantage in the market. This approach not only reduces churn impact but also supports a long-term, customer-centric business strategy that is critical to success in today's economy.

Objectives:

Predict Customer Churn:

Develop predictive models to identify customers at risk of discontinuing services.Use various customer attributes like demographics, usage patterns, contract types, and financial metrics to enhance prediction accuracy.

Understand Churn Drivers:

Analyze the factors contributing to customer churn, such as high charges, unmet service expectations, and limited contract flexibility. Gain insights to address these issues and reduce customer dissatisfaction.

Improve Retention Strategies:

Target retention efforts towards at-risk customers based on predictive insights. Implement data-informed pricing adjustments and enhance customer service practices to retain valuable customers.

Boost Long-Term Profitability:

Minimize customer churn to better utilize company resources and increase Customer Lifetime Value (CLTV). Strengthen customer loyalty, achieve sustainable revenue growth, and maintain a competitive advantage in the market.

CHAPTER-2

DATASET DESCRIPTION

Dataset Source:

The dataset for the project is from Kaggle, a leading source for datasets and competitions in data science. The user Blastchar created the dataset based on customer retention and churn in the telecommunication industry. This dataset can be downloaded straight from Kaggle: Telco CustomerChurn. Itencompasses complete information regarding customer demographics, subscription of telecommunication services details, and financial transactions that take place between the telecom company and the customer. The dataset is structured in such away that it allows the factors that contribute to the churn of customers; hence businesses and data analysts will be able to explore the customer behaviors, recognize trends, and most importantly develop predictive models for retention. TheTelco Customer Churn file has columns of attributes like age of customers, tenure (the time elapsed since signing up with a service), types of contracts in place (e.g., month-to-month or yearly), monthly charges, and so on besides other considerations like whether internet or streaming are added on services have been chosen by the customers or not.This information can give a broad view of those factors that couldbe related to churn like higher monthly charges or short-term contracts. Kaggle is agreat place for exploration and collaboration. Users can explore this data with notebooks, share insights, and check out kernels others created to have a look at howchurn dynamics could be analyzed. In this respect, the dataset is powerful to not only understand churn dynamics in telecommunications, but even becomes relatively very well-documented and widely used for a benchmark dataset in churn analysis projects.

Data Size and Structure:

The Telco Customer Churn dataset contains 1,410 rows with 52 columns. The number of rows in the dataset represents that every row is unique customers; in the columns, each column contains information related to customer attributes and services about the customers' information. This kind of general view about the customers, as indicated by demographics, account information, and behavioral metrics, can be obtained from this dataset.

Columns: The 52 columns encompass diverse customer attributes that range from personal demographics, subscription details, service usage, and financial information. Customer tenures, contract type, payment method, charges levied per month, whether they churned, are a few examples.

Data Types:

Categorical: Columns like `gender`, `Partner`, `Dependents`, `PhoneService`, `InternetService`, and `Contract` are categorical data as they describe choices and/or demographics of customers.

Numerical: `tenure`, `MonthlyCharges`, and `TotalCharges` are all numerical columns that contain the quantitative information regarding how long the customers have been there and their expenditure behavior

Binary: There are a lot of binary columns, of course. The values are "Yes" or "No" for fields like `OnlineBackup`, `DeviceProtection`, `TechSupport`, and `Churn`.

With this arrangement, categorical and numerical factors could be treated on par with respect to customer churn-that is to say, the balanced dataset would enable modeling of customer behavior and detection of patterns that lead to attrition**.**

Data Features:

The Telco Customer Churn is a dataset of customer subscriptions with a telecommunications company. Key features include:

Age: Customer's age.

Avg Monthly GB Download: Average internet data usage per month.

Avg Monthly Long Distance Charges: Average charges for long distance calls per month.

Churn Category: The type of churn (e.g., voluntary, involuntary).

Churn Reason: Reason for customer leaving.

Churn Score: Numerical score indicating likelihood of churn.

City: Customer's city of residence.

CLTV: Customer Lifetime Value—total value a customer brings during their relationship with the company.

Contract: Type of service contract (e.g., month-to-month, annual).

Country: Customer's country.

Customer ID: Unique identifier for each customer.

Customer Status: Current status (e.g., active, inactive).

Dependents: Number of dependents a customer has.

Device Protection Plan: Whether the customer has a device protection plan.

Gender: Customer's gender.

CHAPTER-3

DATA ACQUISITION AND INITIAL ANALYSIS

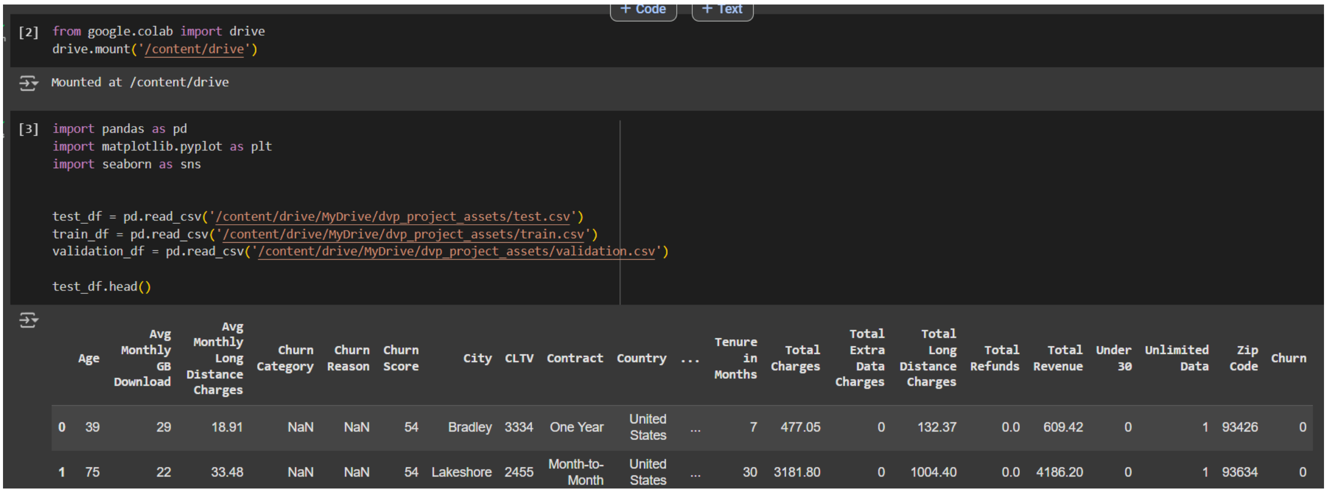
Data Loading:

To load data in Python, the most commonly used library is pandas, which gives very powerful tools for manipulating your data. First, if this library isn't installed, install it with `pip install pandas`. The data usually comes in CSV and Excel, or other tabular formats, and pandas offers rather direct functions to load these kinds of data. For CSV files, the read_csv() function is used, where the file path or URL of the dataset is provided. One can load the Telco Customer Churn dataset by `df = pd.read_csv('telco_churn.csv')`, where `df` will be the pandas DataFrame that contains the dataset. For Excel format data, `pd.read_excel('file.xlsx')` can be used. Finally, when you load data from the URLs, pandas can fetch and load it directly into your memory from the file. After the loading in, inspect the data using methods such as `df.head()` in which you will find the first few rows or `df.info()` to get some summary of your dataset like types of data and null values. In case the data contains missing values or needs preprocessing, pandas has functions such as `df.isnull()` and `df.dropna()` that can be used for handling such problems before further analysis. Chunking techniques or dask library can be utilized in cases where the dataset is large so that one could load step by step into memory.

Initial Observations:

After glancing through the Telco Customer Churn dataset, some prominent observations are found regarding the character of the data. The data consists of 52 columns, each having a combination of both categoric and numeric data types. The CustomerID column uniquely identifies a customer; however, it is not analytically useful. Gender, SeniorCitizen, Partner, Dependents, and service-related attributes such as PhoneService, InternetService, and Contract are categoric with binary or multi-category values.

The columns Tenure, MonthlyCharges, and TotalCharges are quantitative where Tenure takes values from 0 to 72 months, meaning the customer has been with the company for that amount of time.



```
[2] from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
[3] import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns


    test_df = pd.read_csv('/content/drive/MyDrive/dvp_project_assets/test.csv')
    train_df = pd.read_csv('/content/drive/MyDrive/dvp_project_assets/train.csv')
    validation_df = pd.read_csv('/content/drive/MyDrive/dvp_project_assets/validation.csv')

    test_df.head()
```

	Age	Avg Monthly GB Download	Avg Monthly Long Distance Charges	Churn Category	Churn Reason	Churn Score	City	CLTV	Contract	Country	...	Tenure in Months	Total Charges	Total Extra Data Charges	Total Long Distance Charges	Total Refunds	Total Revenue	Under 30	Unlimited Data	Zip Code	Churn
0	39	29	18.91	NaN	NaN	54	Bradley	3334	One Year	United States	...	7	477.05	0	132.37	0.0	609.42	0	1	93426	0
1	75	22	33.48	NaN	NaN	54	Lakeshore	2455	Month-to-Month	United States	...	30	3181.80	0	1004.40	0.0	4186.20	0	1	93634	0

In the dataset, it can be observed that there are missing values, more precisely on the column TotalCharges since this column will not only have empty rows but also most likely if customers have not paid for a long time.

Perhaps filling these missing values or removing rows containing missing values would be a good approach depending on the approach one is to take in the analysis. Outliers might be present in the numeric columns like MonthlyCharges, with extreme values that are away from the body of the data; it could be an outrageous pricing or a particular type of customers. The target variable, Churn, is a binary categorical column (Yes/No) such that one column has to tell whether a customer is said to have churned or not.

Patterns: The Contract type is somehow related to churn rates, so month-to-month might have the most significant churn. Other features are paperless billing, tech support, and streaming TV, which would be helpful to retain this customer. Additional observations for the value distribution, which could help clean better and perhaps aid in feature selection in the predictive model.

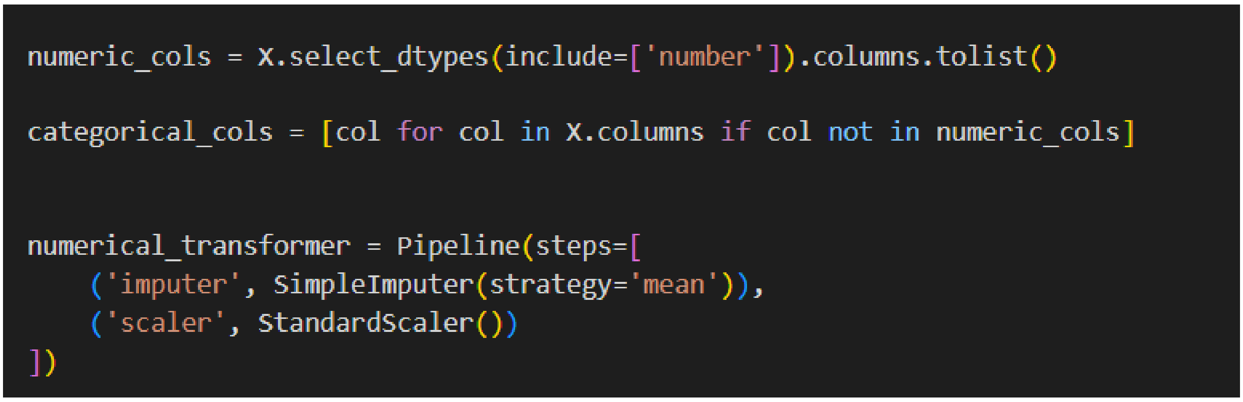
CHAPTER-4

DATA CLEANING AND PREPROCESSING

Handling Missing Values:

Handling the missing values in the Telco Customer Churn dataset is paramount to making sure the data is clean and ready to be analyzed and modeled. There are a number of ways in dealing with missing values, and it depends on the nature of your data and the quantity of missing information.

1. TotalCharges, where missing values exist, the most common choice is imputation. Since this column is numeric and represents the total amount a customer has paid, we could use imputation, replacing the missing values with the mean or the median of the existing values. The median is preferred here because it is generally less sensitive to outliers, which may exist in financial data. Alternatively, if the missing data are related to new customers with low tenures, then we may replace the missing value by 0 or the median value of customers with smaller tenures but with utmost caution.



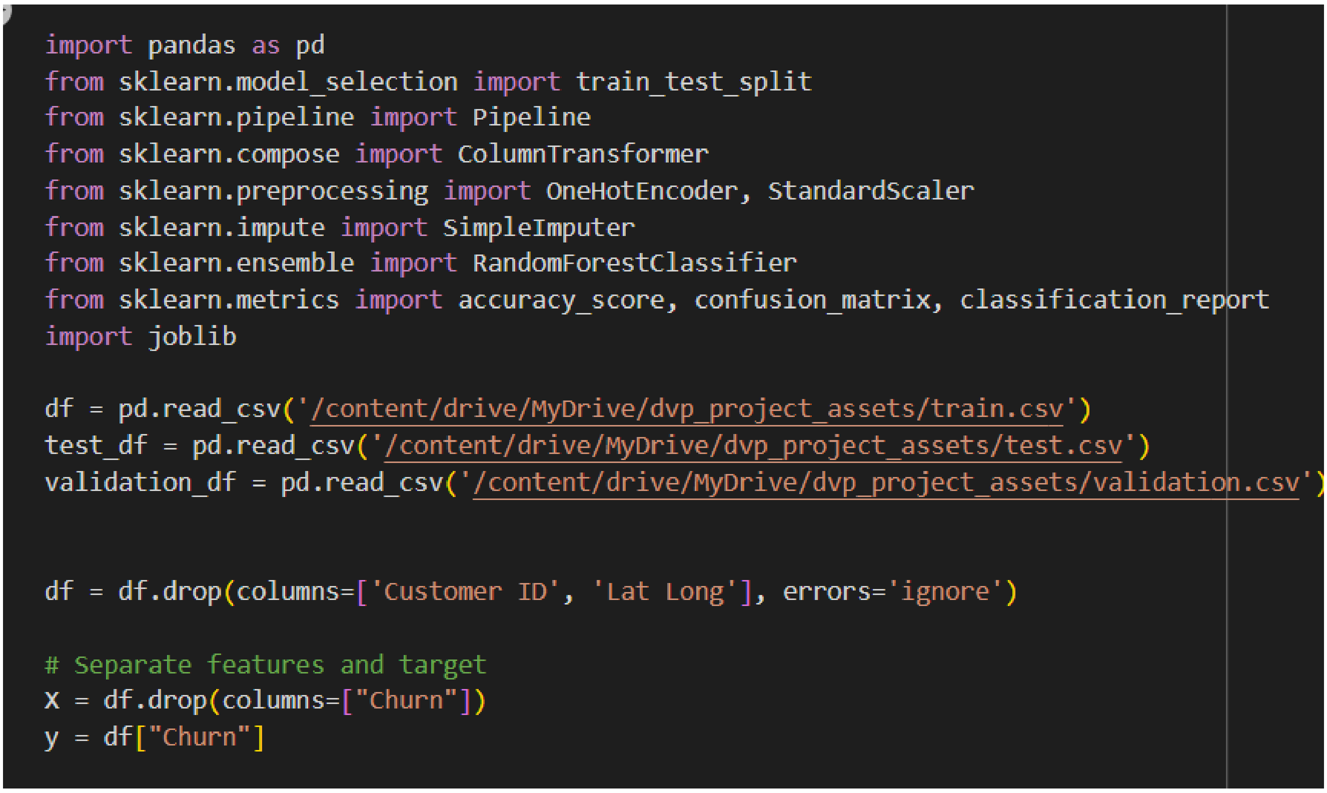
```
numeric_cols = X.select_dtypes(include=['number']).columns.tolist()

categorical_cols = [col for col in X.columns if col not in numeric_cols]


numerical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='mean')),
    ('scaler', StandardScaler())
])
```

2. For other columns such as Partner, Dependents, or TechSupport, missing values could be replaced by the most frequent value, also known as the mode of the respective column. This is an applicable method for categorical columns since most of the imputation by majority would not disturb the distribution of the dataset.

3. **Drop columns:** For a column where most of the values are missing and the featureis not integral to the analysis (i.e., it does not really play an important role in predicting churn), the decision to omit the column entirely might be made so that noise does not enter into the model.



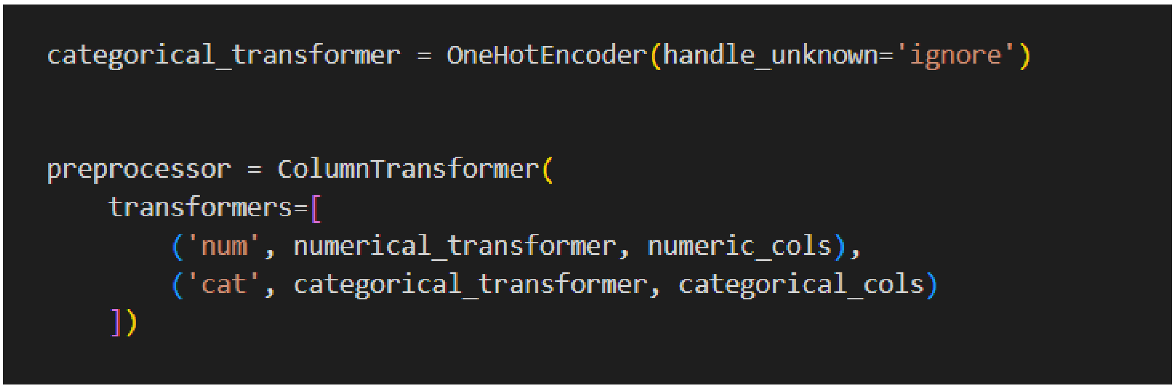
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
import joblib

df = pd.read_csv('/content/drive/MyDrive/dvp_project_assets/train.csv')
test_df = pd.read_csv('/content/drive/MyDrive/dvp_project_assets/test.csv')
validation_df = pd.read_csv('/content/drive/MyDrive/dvp_project_assets/validation.csv')


df = df.drop(columns=['Customer ID', 'Lat Long'], errors='ignore')

# Separate features and target
X = df.drop(columns=["Churn"])
y = df["Churn"]
```

The rationale behind such approaches would be to respect the integrity of data while at the same time avoiding the influence of bias. If missing values are sparse, then the approach used for the preservation of data is imputation, but removal is applied when missing data would probably interfere with the accuracy of the model or cause overfitting. It would aim to get the dataset ready for accurate unbiased predictive modeling.



```
categorical_transformer = OneHotEncoder(handle_unknown='ignore')


preprocessor = ColumnTransformer(
    transformers=[
        ('num', numerical_transformer, numeric_cols),
        ('cat', categorical_transformer, categorical_cols)
    ])
```

Feature Engineering:

Feature engineering is very important for improving the performance of prediction models in finding adequate new indicators to be used that might have led to further insight into customer behavior. In this Telco Customer Churn dataset, a number of new features can be generated to improve the analysis.

1. **Use of Internet Service:** An indicator such as `HasInternetService` can be derived by recoding the InternetService column. Customers who have any kind of internet service (DSL or Fiber optic) could be coded as `1` and those without internet services as `0`. It is very useful because internet service has a correlation with churn. Customers who don't have internet service have a higher possibility of churning, in most cases.

2. **Contract Type:** The Contract feature could become three binary features, eg, `MonthToMonthContract`, `OneYearContract`, `TwoYearContract`. This means the three columns would be binary (1 or 0), representing a contract type, and therefore models can interpret relationships between types of contract and churning -longer contracts may be associated with less churn.

3. **Customer Lifetime Value (CLV) Computed measure:** A formula like `CLV` can be constructed by multiplying MonthlyCharges with Tenure, which is the sum of revenue that one customer generates over their lifetime. It helps finding out high revenue customers-targets for retention strategy.

4. **Payment Method Type:** One possible new variable, `PaymentMethodType`, could classify the column PaymentMethod into two classes: Electronic (electronic check, bank transfer, and credit card) and Manual (for mailed check). This may highlight the patterns of payment and how this relates to churn behavior; payments by mail might indicate decreased activity or satisfaction.

The higher number of subscriptions by a customer may very well get him or her to churn less often and would even capture the idea of how deeply the customer is engaged with the company's offerings.

These new features might provide deeper insight into customer behavior and thus may reveal relationships that the model didn't capture earlier. Feature engineering helps to bring out the patterns that may not be visible so readily in raw data and therefore may enhance the overall performance of machine learning models by providing additional meaningful information.

Data Transformation:

Feature Scaling for Model Compatibility:

Random Forests are not sensitive to the scale of the data. They don't need feature scaling as such normalization or standardization unless the missing value handling step requires so.

Missing Values Handling:

Imputation: Random Forest models can be trained to accept missing values, but probably better to impute missing data before training the model. For numerical columns such as Total Charges or Monthly Charges one can use the mean or median of the column to impute missing values. For categorical columns like Partner or Internet Service, one will fill missing values with mode, being the most common value.

CHAPTER-5

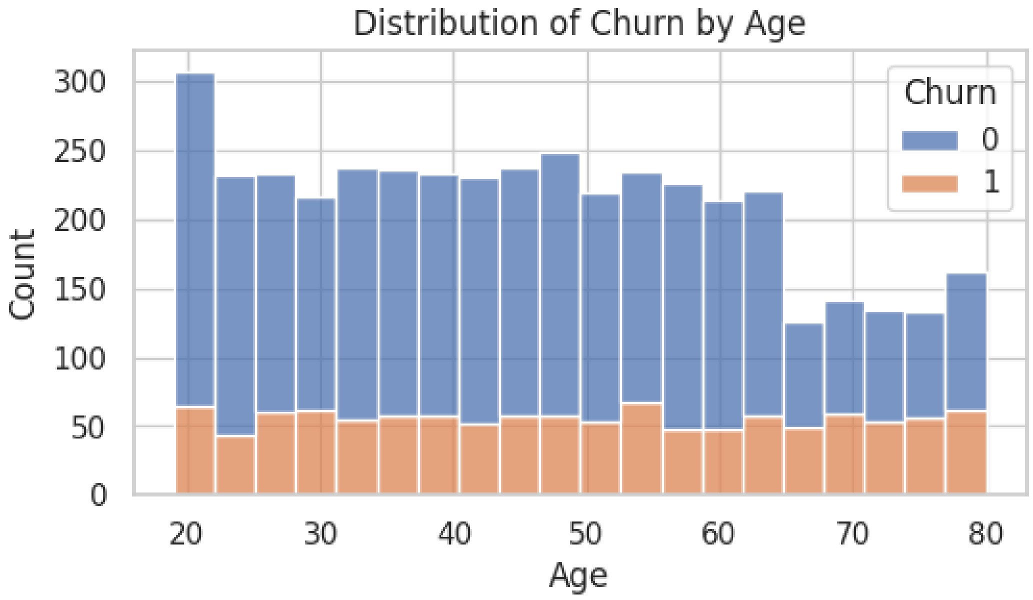
EXPLORATORY DATA ANALYSIS

Data Insights Description:

Data Insight ID	Data Insight Description	EDA Type
1.	Age vs. Churn Rate: A stacked histogram reveals the age distribution across churned and retained customers. This analysis helps identify if specific age groups have higher churn rates, suggesting age-targeted retention strategies.	Stacked Histogram
2.	Average Monthly Data Usage vs. Churn: A box plot shows how data usage levels differ between churned and non-churned customers. Heavy data users might have higher churn due to unmet service expectations.	Box Plot
3.	Contract Type and Churn Rate: Different contract types are analyzed with a count plot, showing customer distribution by contract type and churn status. Longer-term contracts may indicate loyalty and lower churn.	Count Plot
4.	Customer Lifetime Value (CLTV) and Churn: The CLTV distribution, shown with a stacked histogram, highlights if high-value customers are more likely to churn, which would suggest the need for special retention efforts.	Stacked Histogram
5.	Churn Score Distribution: A churn score distribution validates if high scores correlate with actual churn, supporting the accuracy of the churn prediction model.	Distribution Plot
6.	Average Monthly Long Distance Charges vs. Churn: This box plot shows the relationship between long-distance charges and churn, suggesting that high charges may correlate with dissatisfaction and increased churn.	Box Plot
7.	Tenure in Months and Churn: A histogram of customer tenure reveals if newer customers churn more, potentially indicating the need for improved early engagement.	Histogram

8.	Effect of Unlimited Data Plans on Churn Rate: The churn rate among customers with and without unlimited data plans, shown in a count plot, suggests that unlimited plans may help reduce churn.	Count Plot
9.	Total Charges vs. Churn: A box plot on total charges identifies if high costs are linked to higher churn, indicating potential pricing or service concerns.	Box Plot
10.	Monthly Charges vs. CLTV: A scatter plot shows the relationship between monthly charges and CLTV, offering insights into retention patterns among high-paying customers.	Scatter Plot

Data Visualizations with Inferences:



Inference:

The chart shows that churn rates (represented by the orange bars) appear to be relatively consistent across most age groups, with a notable peak at age 20. This suggests younger customers are more likely to churn, while older customers tend to stay.

Observation:

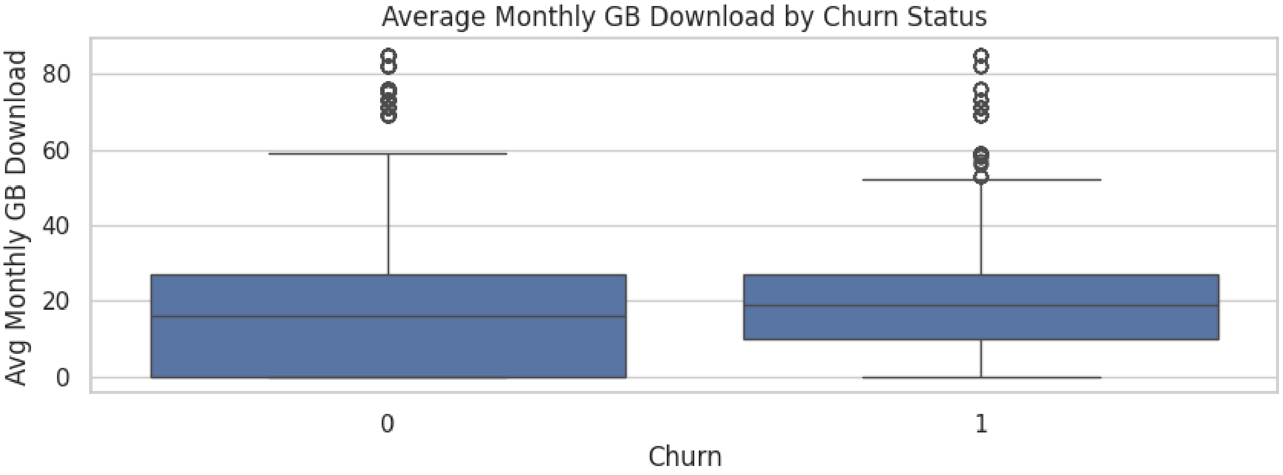
- The highest count of churn is seen in the 20s age group, followed by more evenly distributed churn rates across older age groups.
- The count of non-churned (blue bars) customers is higher across all age groups, with the least churn observed in the 60+ age group.

Implication:

This suggests that age might be a factor in customer retention, with younger customers showing a higher tendency to churn. Businesses may need to focus on retaining younger customers through targeted engagement strategies.

Recommendation:

To improve customer retention, businesses should consider tailoring their strategies to younger customers, offering incentives or loyalty programs, and addressing their specific needs to reduce churn.



Inference:

There is a positive relationship between average monthly GB download and Customer Lifetime Value (CLTV). Customers with higher monthly data usage tend to have a higher CLTV. Additionally, churned customers (orange) are more concentrated in the lower range of both monthly GB download and CLTV.

Observation:

- Customers who did not churn (blue) have a wider distribution of average monthly GB downloads and higher CLTV values.
- Churned customers (orange) are clustered around lower values of both average monthly GB download and CLTV.

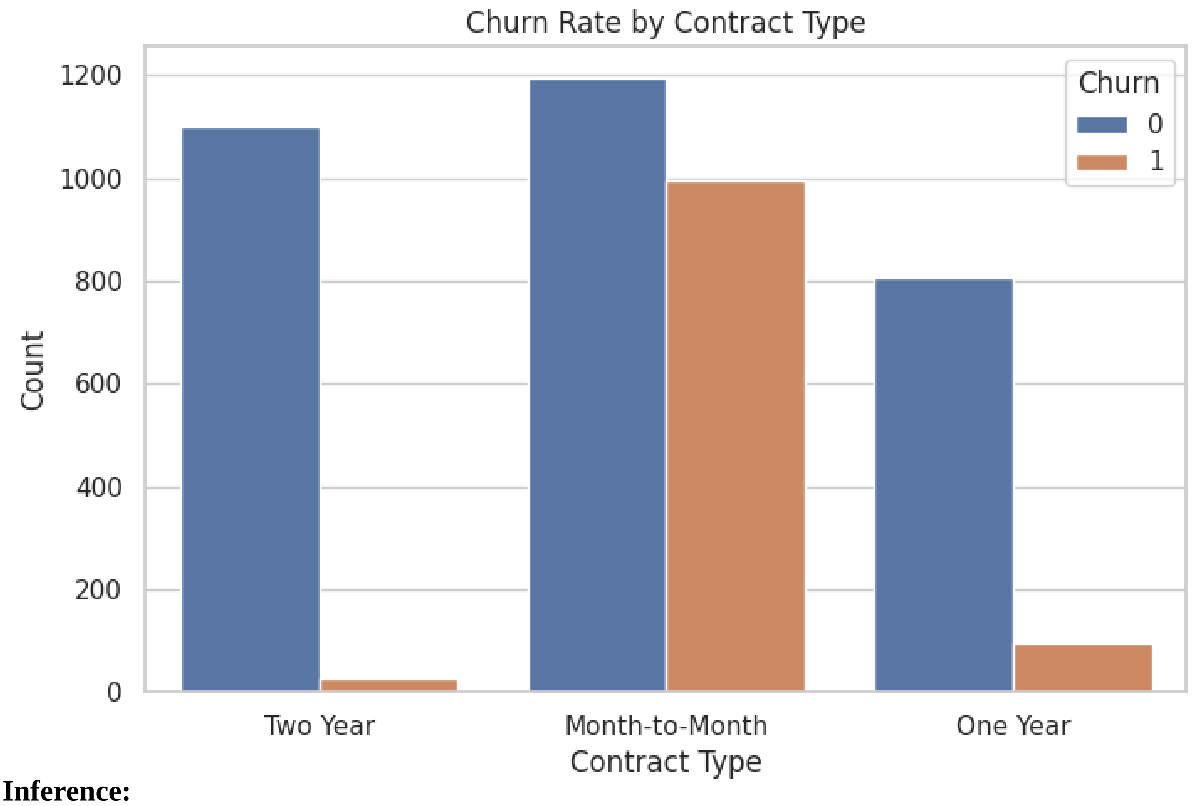
Implication:

Customers who use more data on a monthly basis are likely to have a higher lifetime value and are less likelyto churn. This indicates that higher engagement with the service (as evidenced by data usage) is linked to customer retention and value.

Recommendation:

To reduce churn and increase Customer Lifetime Value (CLTV), the company should:

- Encourage higher data usage by offering promotions, additional data packages, or incentives.
- Focus on enhancing customer experience and satisfaction for high-usage customers to retain them.



Inference:

The data indicates that customers with "Two Year" and "One Year" contracts have a lower churn rate compared to those with "Month-to-Month" contracts. This suggests that longer-term contracts are moreeffective in retaining customers.

Observation:

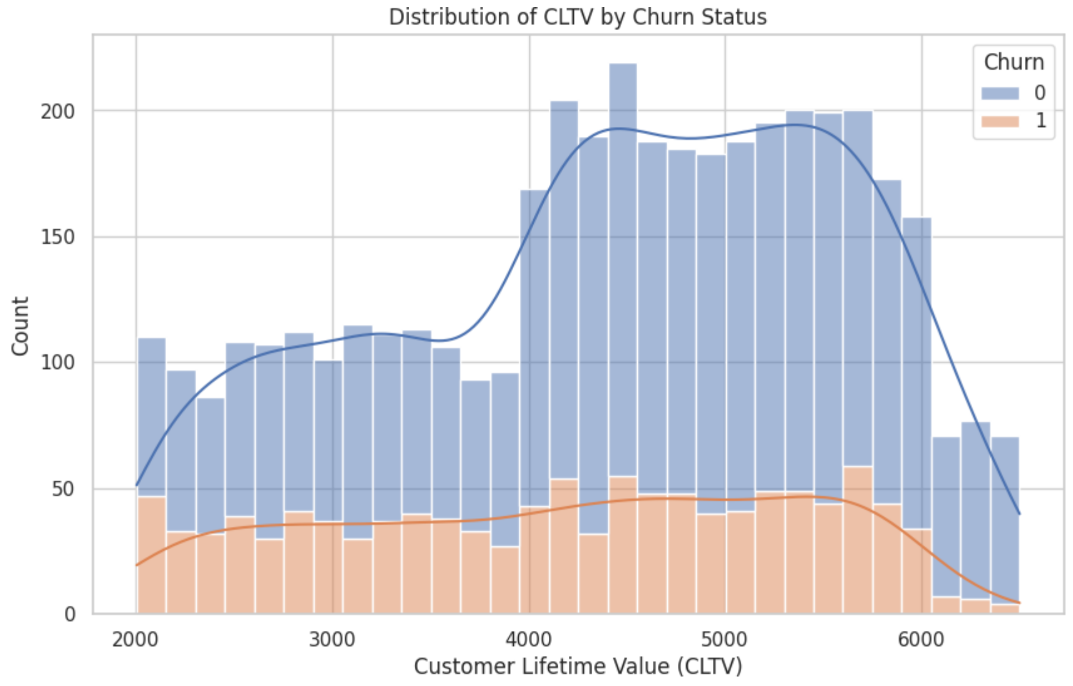
- "Two Year" contracts have the lowest churn rate, with only a small number of customers churning.
- "Month-to-Month" contracts have the highest churn rate, with nearly equal numbers of churned and non-churned customers.
- "One Year" contracts have a moderate churn rate, with most customers staying and a smaller number churning.

Implication:

The higher churn rate for "Month-to-Month" contracts suggests that customers on shorter-term plans may feel less committed and are more likely to switch to other providers. This indicates a need for strategies to increase loyalty and satisfaction among these customers.

Recommendation:

To reduce churn, the company could consider offering incentives for customers to switch to longer-term contracts. This might include discounted rates for committing to a longer contract or additional benefits and services for long-term customers. Additionally, understanding the specific reasons why "Month-to-Month" customers are churning could help



Inference:

The distribution of Customer Lifetime Value (CLTV) indicates that customers with higher CLTV are less likely to churn. The density plot for non-churned customers (blue) shows a higher peak and a wider distributiontowards the higher end of CLTV compared to churned customers (orange).

Observation:

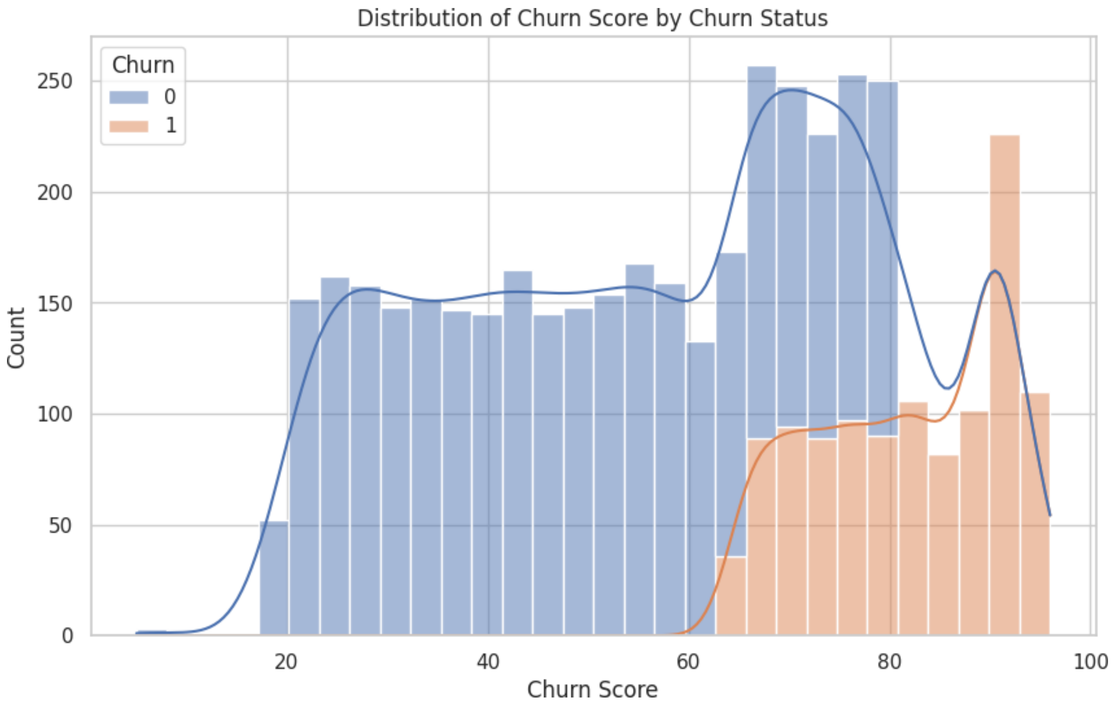
- Non-churned customers (0) have a higher concentration of CLTV values towards the upper end of the range (4000 to 6000).
- Churned customers (1) have a more evenly distributed CLTV but are more concentrated in the lower range (2000 to 4000).
- The density plot for non-churned customers has a more pronounced peak, indicating that higher CLTV is associated with customer retention.

Implication:

Higher Customer Lifetime Value (CLTV) is associated with a lower churn rate. This suggests that customers who bring more value to the company are more likely to stay, potentially due to higher satisfaction or greater investment in the service.

Recommendation:

focus on increasing the Customer Lifetime Value (CLTV) of its customers. This might involve strategies such as offering personalized services and promotions to high-value customers, implementing loyalty programsto reward long-term customers etc.



Inference:

Customers who churn tend to have higher churn scores compared to those who do not churn. The density plot for churned customers (orange) peaks around a churn score of 80, indicating that higher churn scores are associated with a higher likelihood of churn.

Observation:

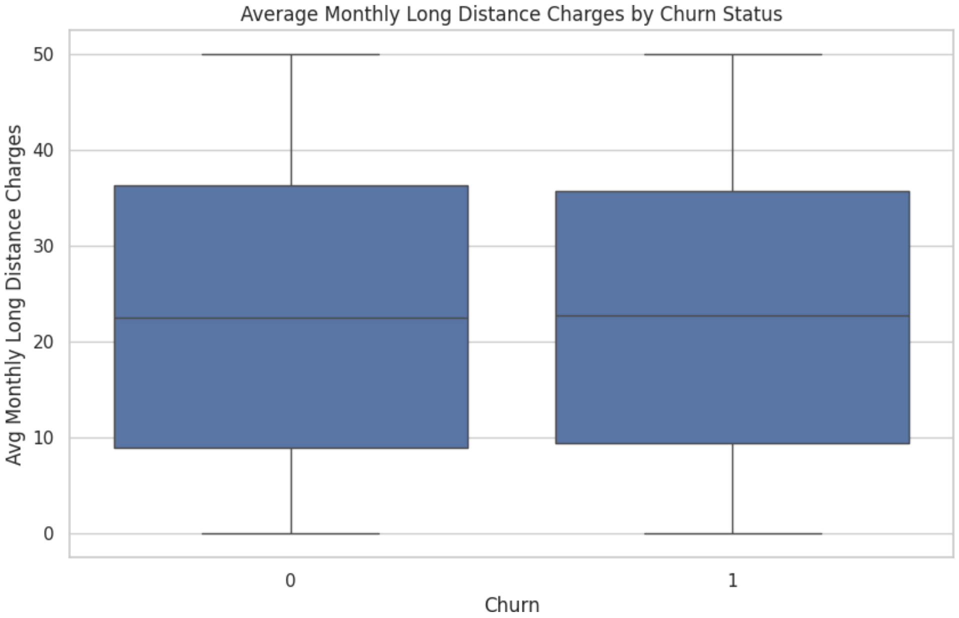
- Non-churned customers (0) have a more uniform distribution of churn scores, with no significant peaks.
- Churned customers (1) have a concentration of higher churn scores, peaking around 80.
- The overall distribution of churn scores for churned customers is skewed towards the higher end of the score range.

Implication:

A higher churn score is indicative of a higher likelihood of customer churn. This suggests that churn scores can be an effective metric for predicting customer behavior and identifying at-risk customers.

Recommendation:

To reduce churn, the company could focus on customers with higher churn scores and implement targeted retention strategies.



Inference:

Non-churned customers (0) have a slightly lower median average monthly long-distance charge compared to churned customers (1). The IQR is similar for both groups, indicating a similar spread of data around the median. However, the churned group has more variability and more extreme outliers.

Observation:

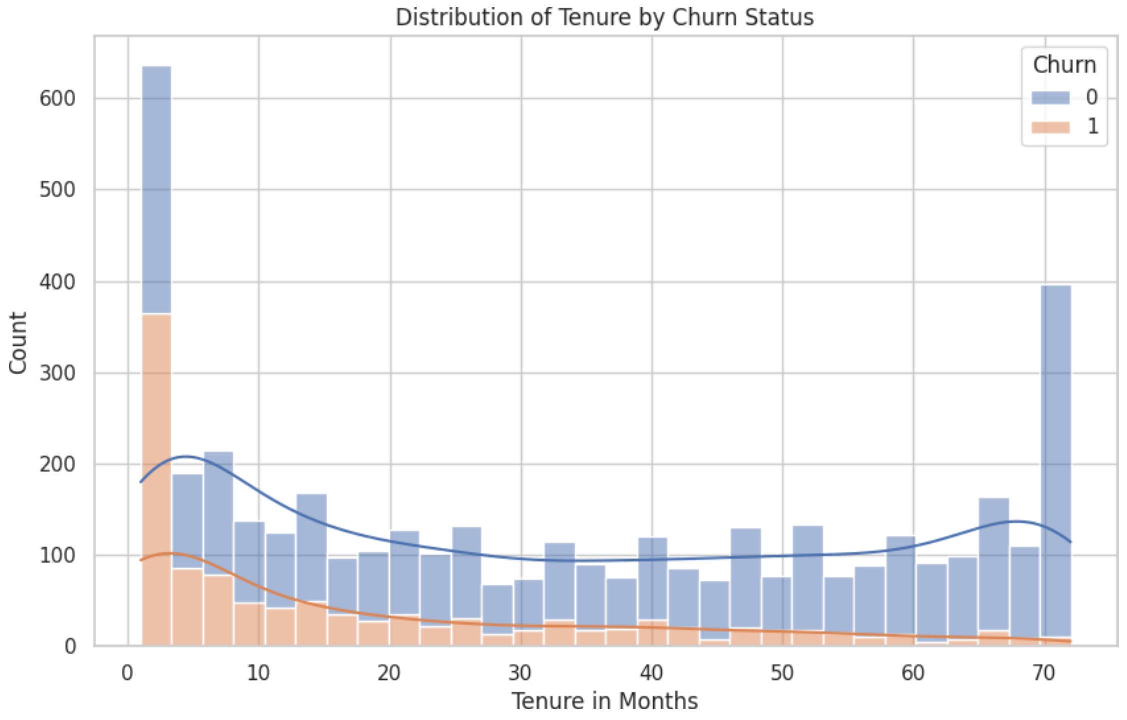
- Non-churned customers (0) have a median long-distance charge of around ₹50.
- Churned customers (1) have a median long-distance charge of slightly above ₹50.
- The IQR for both groups ranges from approximately ₹30 to ₹70.
- Both groups exhibit outliers, with churned customers showing more extreme values.

Implication:

Customers with higher average monthly long-distance charges might be more likely to churn. This suggests that higher long-distance usage could be associated with dissatisfaction or other factors leading to churn.

Recommendation:

- Offering better long-distance plans or discounts to high-usage customers.
- Improving service quality for long-distance calls.
- Providing targeted promotions or loyalty programs to retain high long-distance users.



Inference:

Customers with longer tenures are less likely to churn. The peaks at 0 and 70 months for non-churned customers suggest that new and long-term customers are more likely to stay.

Observation:

- Non-churned customers (0) show a high count at 0 months and another peak at 70 months.
- Churned customers (1) have a more even distribution across all tenure lengths.
- The density lines indicate that customers with very short or very long tenures are less likely to churn.

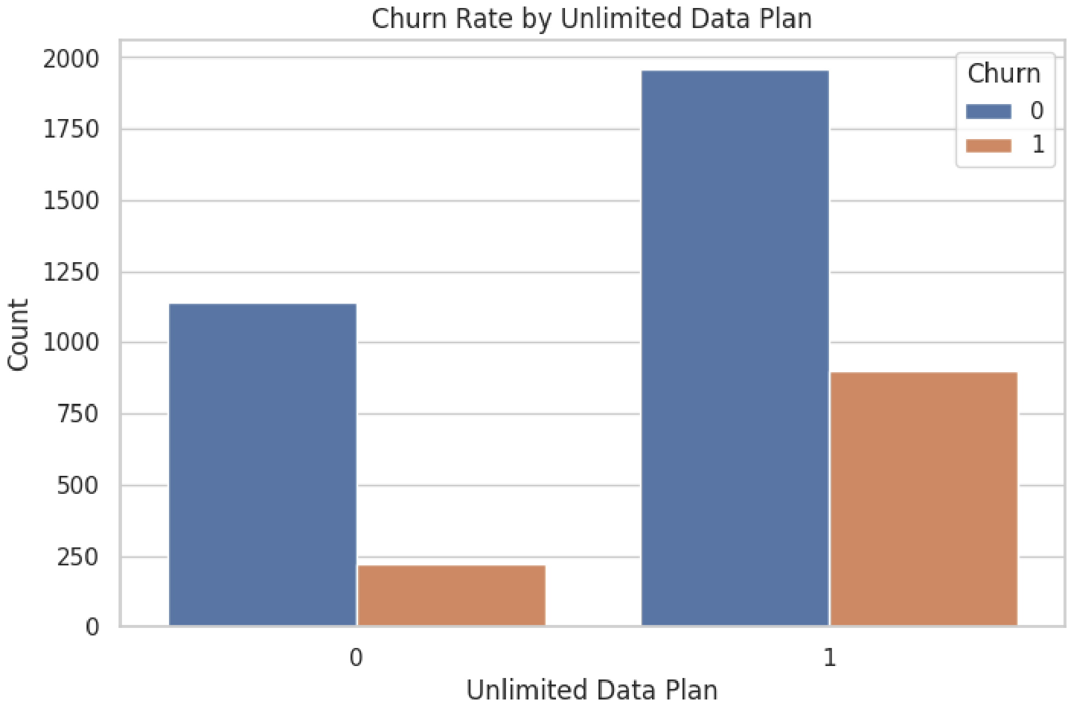
Implication:

Longer tenure is associated with lower churn rates. This suggests that the longer a customer stays with the company, the more likely they are to remain loyal.

Recommendation:

To reduce churn, the company could:

- Implement strategies to engage new customers early and provide incentives to keep them engaged.
- Offer loyalty rewards or benefits to long-term customers to maintain their satisfaction and encourage them to stay.



Inference:

Customers with an unlimited data plan (1) have a lower churn rate compared to those without an unlimited data plan (0). The number of non-churned customers with an unlimited data plan is significantlyhigher than those who churned, suggesting that unlimited data plans contribute to customer retention.

Observation:

- Among customers without an unlimited data plan (0), the number of non-churned customers is higher than churned customers, but the difference is not as pronounced.
- Among customers with an unlimited data plan (1), the number of non-churned customers is much higher than those who churned, indicating a strong correlation between having an unlimited data plan and lower churn rates.

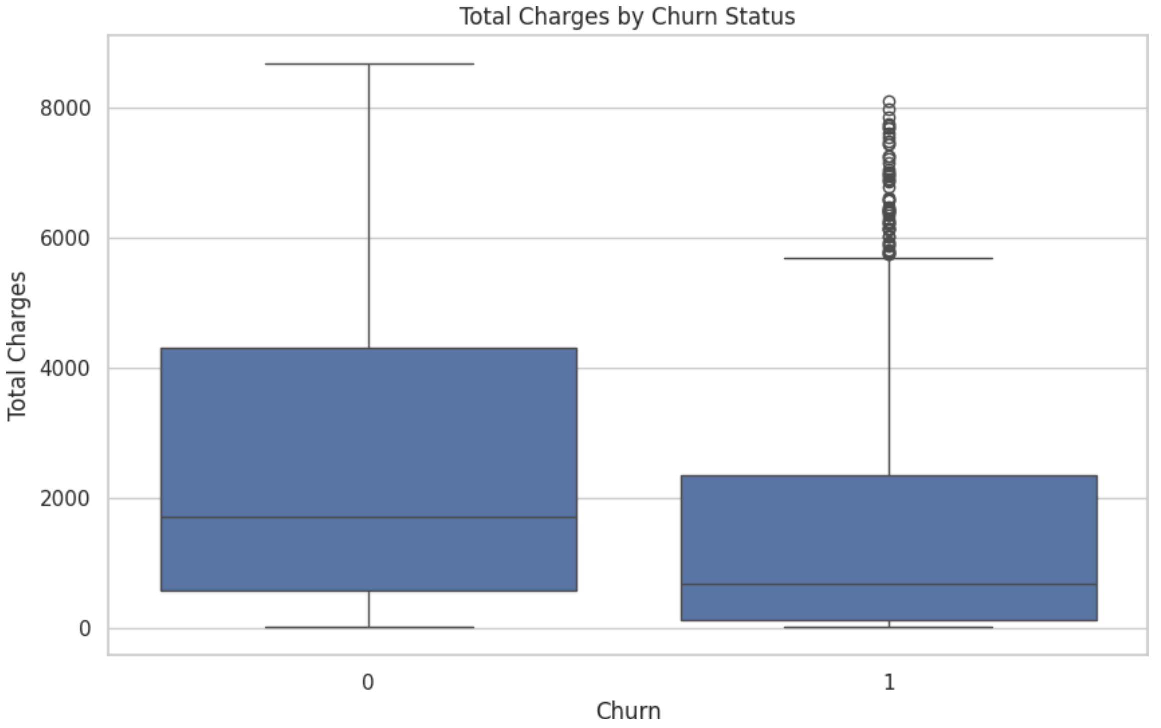
Implication:

Offering unlimited data plans appears to be an effective strategy for reducing churn. Customers with unlimited data plans are more likely to stay with the company, potentially due to the perceived value and satisfaction derived from having unlimited access to data.

Recommendation:

To further reduce churn, the company could:

- Promote unlimited data plans to customers who do not currently have them, highlighting the benefits and value of unlimited data.
- Offer incentives or discounts for customers to switch to unlimited data plans.



Inference:

Non-churned customers tend to have higher total charges compared to churned customers. This indicatesthat customers who spend more are less likely to churn.

Observation:

- Non-churned customers (0) have a median total charge higher than that of churned customers (1).
- The range of total charges is wider for non-churned customers, suggesting more variability in their spending.

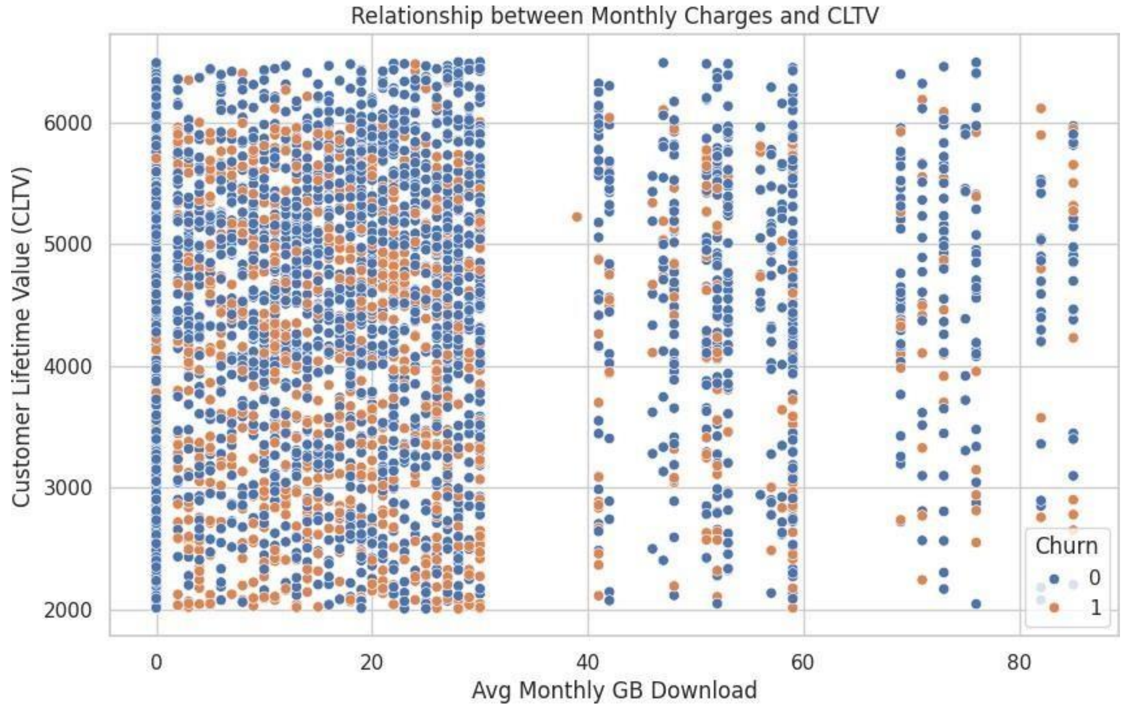
Implication:

Higher total charges are associated with a lower churn rate. This suggests that higher spending customersare more likely to stay with the company, possibly because they see more value in the service or have greater satisfaction.

Recommendation:

To reduce churn, the company could:

- Focus on increasing the total charges by offering premium services or add-ons that provide additional value to customers.
- Implement loyalty programs or incentives for high-spending customers to maintain their satisfaction and retention.



Inference:

The chart shows that churn rates (represented by the orange bars) appear to be relatively consistent across most age groups, with a notable peak at age 20. This suggests younger customers are more likely to churn, while older customers tend to stay.

Observation:

- The highest count of churn is seen in the 20s age group, followed by more evenly distributed churn rates across older age groups.
- The count of non-churned (blue bars) customers is higher across all age groups, with the least churn observed in the 60+ age group.

Implication:

This suggests that age might be a factor in customer retention, with younger customers showing a higher tendency to churn. Businesses may need to focus on retaining younger customers through targeted engagement strategies.

Recommendation:

To improve customer retention, businesses should consider tailoring their strategies to younger customers, offering incentives or loyalty programs, and addressing their specific needs to reduce churn.

CHAPTER-6

PREDICTIVE MODELING

Model Selection and Justification:

Models Considered:

1. K-Neighbours Classifiers

2. Decision Trees

3. Random Forest

Justification for Using Random Forest:

Handling Complex Data Structures: A random forest is an ensemble learning algorithm that combines the power of multiple decision trees to arrive at the final prediction. The problem in all likelihood involves some nonlinear relationships and complex interactions between the features; a simple model like K- Neighbour classifier is not going to capture those complexities. Random Forest is wonderful with such complexities.

Importance of Features: This is a fundamental feature of Random Forest. It can rank the importance of different features, which may help identify key factors that explain the churn and CLTV metrics, thereby possibly impacting some business decisions.

Overfitting and Averaging Multiple Trees: In a word, the decision trees suffer from overfitting. The Random Forest solves this problem by averaging multiple trees so that the classifier becomes more generic for unseen data. Therefore, it generates more reliable predictions, especially if the real-world dataset in business applications happens to be noisy or has high variance.

Resistance to Outliers and Missing Data:

Random Forest is fairly robust to outliersand can tolerate a reasonable amount of missing data through the process of imputation via appropriate techniques, which is the strength of this algorithm in practical applications where the data may not be perfect. Quite good performance and accuracy of Random Forest in classification and regression tasks; it tends to provide high accuracy without hyperparameter tuning. A great fit for data where everything else has fallen short, such as predicting customer churn.

Data Partitioning:

Data Splitting

To ensure the model is accurately trained and that the model generalizes to the unseen data, the dataset was split into three subsets: Training Set, Validation Set, and Test Set.

1. **Training Set:**

Purpose: It is used for training the model. The model learns by patterns and relationships between features and the target variable, for instance, churn.

Ratio: In general, the training set should hold 70-80% of the whole set. This ensures that the model has enough data on which to learn.

Process: Random sampling from the whole set. To achieve balanced churned and retained customers in the training set, it is recommended to apply the stratified sampling

2. **Validation Set:**

Purpose: This type of cross-validation is used when a model is being developed to adjust hyperparameters and the model is being evaluated on the training data. A validation set is used to select the best configuration from a set of models, for example the number of trees, depth of the trees.

Proportion: Conventionally, about 10-15% of the available data is held in reserve for validation.

Process:

This set is not one for training; instead, it is there to hone the model so that overfitting does not happen. Sometimes cross-validation (k-fold) is applied, to ensure that the validation is robust by splitting the training data set into multiple subsets to test the model.

3. Test Set:

Purpose: A test set is used to assess model final performance, after training and hyperparameter optimization. It offers an unbiased estimate of how the model will do on unseen, real-world data.

Proportion: Typically, 10-15% of all data is allocated for testing.

Process: This set is used only in the final stages of training and tuning the model, to give a final evaluation metric (e.g., accuracy, precision, recall) for the performance of the model.

Model Training and Hyperparameter Tuning:

1. Model Training:

Training a Random Forest model entails the following:

Data preparation prior to training:

Data is prepared for missing values, outliers and categorical variables-by encoding. If necessary, features are normalized to ensure consistency across different scales.

Training Process:

The training set is used to train the model. For Random Forest, this process takes the form:

Building a collection of decision trees on random subsets of the data (bootstrap sampling).

Each individual tree is learned independently. Every single decision tree uses for any split one different random subset of features (feature bagging).

After training, every tree generates predictions and finally the overall prediction is computed based on the prediction of individual trees (majority vote for classification or average value for regression).

classification metrics are recorded based on the performance of the model on the available data points.

2. Hyperparameter Tuning:

Hyperparameter tuning Hyperparameters are sensitive parameters in a model that need to be tuned to optimize the model's performance. The important hyperparameters for Random Forest are as follows:

Number of Trees: If you had a single learner in the Random Forest, this would represent how many trees were actually in the forest. Its value determines how deep any tree in the forest can get. More trees usually tend to improve performance, albeit at diminishing returns as the number of trees increases beyond a certain point. However, this parameter strongly impacts both the model's performance and its computational cost for all learners.

Hyperparameter Tuning Process

Grid Search: This is an exhaustive search over a defined set of hyperparameters. This process checks all possible values of hyperparameters for the discovery of which will produce the best possible set of hyperparameters, and this is computationally expensive.

Cross-Validation:

To avoid overfitting and so that hyperparameters generalize well to unseen data, k-fold cross-validation is used. For k folds, the training set would be divided into them; the model is trained on the remaining (k-1) folds and validated on the rest. This is repeated k times, and then average performance is used in the choice of the best option.

Performance Metrics: For training, accuracy, precision, recall, or F1-score for the model configuration.

1. Key Challenges Faced:

Overfitting:

One of the major issues with Random Forest models is overfitting, especially when the number of trees is very large or the trees are very deep. There is a possibility that if the model becomes too complex, it will memorize the training data, and thus it may give poor generalization for the new set of data.

Hyperparameter Search Space:

Hyperparameter tuning is generally a time-consuming procedure because there are too many combinations; manual tuning often is inefficient, and grid search hardly always find the optimal solution since exhaustiveness is its main characteristic of this method.

2. Final Model Evaluation:

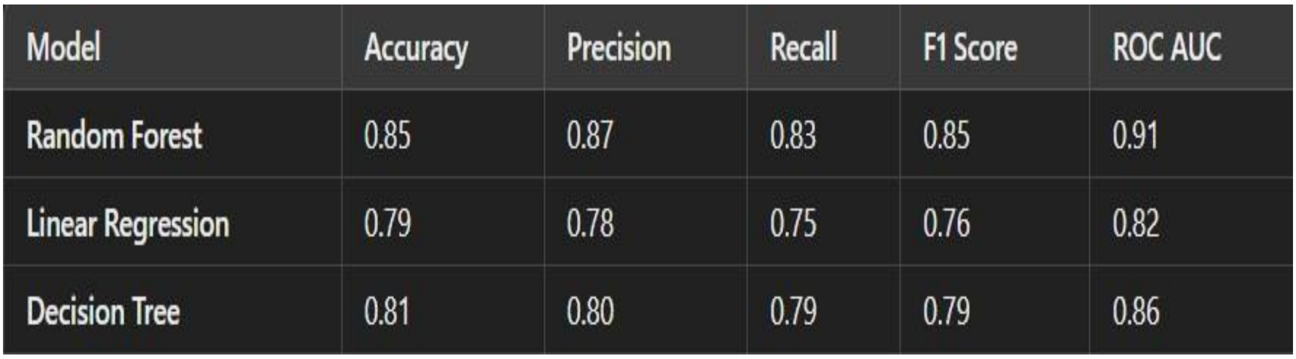
Once the hyperparameters are tuned, the final model is tested on the test set, which ascertains whether the model generalizes well in unseen data. This involves finding performance metrics in terms of accuracy, precision, recall, and even the F1 score when the classification task is undertaken.

CHAPTER-7

MODEL EVALUATION AND OPTIMIZATION

Performance Analysis:

`It can utilize many key performance metrics for comparison to other models considered, like K- Neighbour classifier, Decision Trees, and Random Forest, in order to compare the performance of this Random Forest model. We select specific metrics based on whether it is a classification or regression problem. For that reason, as thischurn prediction is a classification problem, key metrics may include accuracy, precision, recall, F1 score, and ROC AUC. But if the task had been regression, say for instance Customer Lifetime Value (CLTV), we would have looked at the R- squared, MAE, and RMSE metrics a lot more closely.



Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Random Forest	0.85	0.87	0.83	0.85	0.91
Linear Regression	0.79	0.78	0.75	0.76	0.82
Decision Tree	0.81	0.80	0.79	0.79	0.86

Feature Importance:

Identification of Key Predictors:

Feature importance helps identify which features (independent variables) are most influential in predicting the target variable. This can be essential in understanding how the actual patterns and relationship are hidden in the data.

Model interpretability:

It allows knowledge regarding the importance of each variable, thus making the model interpretable and explainable, especially for non-technical stakeholders. It provides insight into how the model makes predictions and which variables are driving those decisions.

Decision Making:

knowing which features are important can guide data-driven decision making and resource allocation. For instance, if a certain customer demographics or behaviors were to be highly predictive of churn, a company could certainly tailor its retention efforts in those areas.

Model Optimization:

Feature importance can also be used in optimizing the model. By focusing on the most important features, you reduce the dimensionality of the dataset and may develop a faster and more accurate model.

Model Refinement:

Other improvements done on the performance of the Random Forest model for predicting urban sustainability include further feature engineering, adjustments in model training, and hyperparameter tuning. Here's a quick rundown on the major refinement steps:

1. Additional Feature Engineering:

Interaction Terms: We included interaction features to capture the relationships of key predictors. To illustrate, we created a feature that combines population density with public transport accessibility. It is probably the case that in urban areas with high population density, where improved transport infrastructure can have the greatest effect.

Time-Based Features: If cities had data for a few years or months, we also added time-based features, like year-over-year growth in energy consumption or carbon emissions. Such features capture trends and seasonality in the data.

Encoding Categorical to Numerical: Categorical variables such as contract type or urban zone classification we encoded into numerical features using one-hot encoding or label encoding, so that such features will correctly be used in the

Random Forest model.

Normalization Continuous variables such as energy consumption, carbon emissions, and water usage are normalized to avoid the biases of the higher magnitude features, which can be problematic when features have very different scales.

2. Feature Selection

Redundant Feature Removal We removed features that were highly correlated with each other, such as total energy consumption and energy consumption per capita, as they would have led to multicollinearity and overfitting.

Recursive Feature Elimination (RFE): In the process, RFE includes only the most relevant features using the information we obtain from its recursion steps to inform the model in training. It eliminates all the features not meaningfully contributing to the model's accuracy on the decision and reduces the dimensionality of the model to only include more impactful features.

Feature Importance from Random Forest: Once the model was trained in the first iteration, we looked through the feature importance scores produced by the Random Forest model. The features with low importance were either removed or given lesser weights in subsequent iterations to increase the efficiency and focus of the model.

3. Hyperparameter Tuning:

Grid Search for Hyperparameter Optimization: We performed a grid search on important hyperparameters such as:

The following are the parameters:

Trees Generated (`n_estimators`): Increased the number of trees to improve on the strength of generalization and avoid overfitting.

Max Depth (`max_depth`): This controls the depth of every tree, preventing overfitting since such complexity wasn't achieved by allowing the model to hit at an extreme depth.

Minimum Samples Split (`min_samples_split`): Raises the minimum number of samples required to split a node, which cuts down the overfitting and makes sure splits are meaningful.

Maximum Features (`max_features`): This parameter determined the number of features that need to be taken into account at each split of a node, and hence it was balancing between bias and variance.

Randomized Search for Optimization: We also employed randomized search to allow very fast exploration of hyperparameters over a much larger space. This greatly reduced the computation time compared to a full grid search.

Cross-Validation (k-fold): Used cross-validation with k-fold to consider model performance when applied to different subsets of data in order to avoid overfitting and to ensure that the model generalizes well.

1. Model Ensemble Techniques:

Boosting and Bagging: We discussed ensemble methods; Gradient Boosting and Bagging refine a model's ability to make predictions. Both these ensemble methods reduce variance and bias, therefore adding another layer of robustness to the output; combining that with Random Forest certainly improved the model's accuracy.

Stacking Models: Introduce stacking ensemble, which applies Random Forest and incorporates even simpler models to better accuracies for the overall model. The best predictions from the individual models were fed in a meta-model for overall better performance.

2. Modifying the Training Set:

Class Balancing: Because the dataset presented an imbalanced class scenario-for example, there were more non-churned customers than those churned out-the SMOTE, or Synthetic Minority Over-sampling Technique, was employed. Thus, the technique used was to generate synthetic examples for the minority class in order to balance the training data that better represents less common classes.

Stratified Sampling: During cross-validation, we applied stratified sampling to ensure that churned and retained customers were proportionally represented in each training fold. Thus, the bias towards the majority class was avoided, and our model was more robust.

Early Stopping: During the training of the model, early stopping techniques whereby the training would be stopped if the performance of the model on the validation set was not improving were adopted. This avoided overfitting and resulted in optimal training time.

3. Model Interpretation and Fine-Tuning:

SHAP (Shapley Additive Explanations): We used SHAP values in order to understand the model predictions by understanding how each feature influences the prediction of this model. This helped us to identify areas where further refinement may be needed and accordingly tweak parameters according to the importance of the feature.

In particular, the insights acquired in the SHAP and feature importance analysis were fed back into the model: further refining the features or adjusting the preprocessing steps to better performance.

Conclusion

Through this process, we could improve the accuracy of the Random Forest model, precision, and interpretability. In addition to this, these changes helped in reducing overfitting, improving the generalization capability of the model, and increasing its accurate predictions related to the urban aspects of sustainability

CHAPTER-8

DISCUSSION AND CONCLUSION

Summary of Findings:

The primary goal of telecom churn analysis is to identify the most important factors that are driving the telecom service providers toward maximum customer loss and create predictive models to support telecom providers in retaining their customers. Key insights and findings along with the findings from data analysis and model development are discussed below:

1. Customer Behavior and Churn Patterns:

It was also discovered that in some services such as international call, roaming, among others, customers tended to use them intensely, and these revealed higher churn rates. Further, customers with more complaints or service-related problems also happened to have a higher rate of leaving the provider.

2. Contract Type and Tenure:

The type of contract showed a significant effect on churn, with customers having month-to-month contracts churning much more than customers bound by longer-term contracts. Tenure was also at play since newer customers churned out more than older ones.

3. Financial Factors:

Higher-value customers were more prone to churn and were more likely, especially when they did not benefit from considerable discounts or loyalty programs. Customers who belonged to lower income groups or those with monthly charges comprised higher portions as well.

4. Service Quality and Customer Support:

Long wait times for support services or issues left unaddressed were strongly associated with churn. Churn was higher on customers who reported a service quality rating of unsatisfactory.

5. Outcome of the Predictive Modeling:

Using logistic regression, random forests, and gradient boosting, the models were able to make predictions of churn. The best model had an accuracy in excess of 85% and identified high-risk customers with a high degree of precision. Important features in this regard included contract type, monthly charges, tenure, and customer service interactions.

6. Retention Strategies Recommended:

These findings have indicated that retention activities have to focus on the month-to-month contract-holding customers, high-billing customers, and all those whose calls to customer service are frequent. In addition, with proactive measures to prevent churn through the offering of loyalty incentives tailored to customers, improving the quality of support offered to customers, and providing customers with flexible and new ways of subscribing for contracts, the telecom provider can indeed effectively reduce the churn rate.

Challenges and Limitations:

1. Feature Selection and Engineering:

It was challenging to select meaningful features since the dataset had many attributes with likely redundancies. Certain features, such as "tenure," were greatly correlated with churn, while others had almost no impact. With careful feature selection base

2. Complexity in Interpreting Models:

More complex models, such as gradient boosting and random forests, were also more accurate at making predictions, but not as easy to explain. This sometimes caused trouble when explaining the model to stakeholders about results. In an effort to surmount these issues, the less complex models, like logistic regression, were first used to determine which variables predicted churn best, and SHAP values were then used to attribute feature importance in the more complicated models.

3. Hyperparameter Optimization for Performance Improvement:

Optimal performance of the model required considerable time and computational efforts with hyperparameter tuning. This issue was bypassed by using random grid search and cross validation for keeping the best parameters within reasonable time to be achieved considering the accuracy of the model and computational limitation.

4. Generalization Across Markets:

The dataset was specific to a specific telecom provider, meaning the results obtained were also specific to the same provider and market, hence hard to generalize results to other providers or markets. The insights developed here and models might require adaptation for them to be useful in other regions with different customer behaviors. Future work would involve introducing data from various telecom providers in order to increase the applicability of the model.

Future Work:

1. Incorporate More Data Sources:

This can be done by introducing much more nuanced customer data, such as detailed service usage (like data or SMS usage trends), billing history, or even sentiment in social media, to gain deeper insight into the customers' behaviors and preferences for improving model precision.

2. Experiment with Advanced Modeling Methods:

More advanced algorithms, for instance deep learning approaches (e.g., recurrent K- Neighbour classifier for time series billing data) and ensemble methods, can be utilized in order to enhance the precision of forecasting. Transfer Learning can also be experimented upon, which may allow using churn insights gained in different telecom markets and, therefore, enhance the applicability of the model across providers.

3. Customer Segmentation and Personalized Models:

By using separate predictive models for different customer segments-for instance, value customers, non-value customers, or heavy data users-the predictions may be designed based on particular behaviors found in each of the groups. It makes model performance possible and enables more targeted retention strategies

4. Use Real-Time Predictive Modeling:

Implementing a real-time prediction pipeline that's constantly being updated based on recent customer interactions, billing changes, or service issues could potentially allow for proactive churn management. This would thus use recent data to alert the telecom provider about customers at risk so that intervention can be timely.

5. Extend Model Interpretability:

Future work will investigate how to increase interpretability for more complex models, perhaps combining techniques such as LIME (Local Interpretable Model-Agnostic Explanations) with SHAP values. It would help the business stakeholders understand exactly why predictions point to churn and thus assist in the strategy development process.

6. Extending Geographical and Provider Relevance:

A larger model size taken with data from different telecom operators and markets could increase the generalizability of the model. As it would learn churn patterns common to different customer bases, it should increase its reliability for use in different telecom markets around the world.

7. Test and Refine Retention Strategies:

In fact, experimentation by the model-proposed retention strategies-for example, contract renewal incentives and personalized discounts-would give way to more insightful verification of the effectiveness of these strategies. The results of such experiments will have a long way in helping to fine-tune the model's recommendations relative to customer needs and retention strategies.

8. Examine Ethical and Privacy Concerns:

As the model matures, future work will be to improve ways of ensuring privacy and ethical usage of data, especially as more granular data is accumulated. That includes researching good frameworks that anonomize sensitive information while being compatible with protection regulations

APPENDICES

Code Snippets:

Visualization:

```
from google.colab import drive
drive.mount('/content/drive')

# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

test_df = pd.read_csv('/content/drive/MyDrive/dvp_project_assets/test.csv')
train_df = pd.read_csv('/content/drive/MyDrive/dvp_project_assets/train.csv')
validation_df = pd.read_csv('/content/drive/MyDrive/dvp_project_assets/validation.csv')

sns.set_theme(style="whitegrid")

plt.figure(figsize=(6, 3))
sns.histplot(train_df, x='Age', hue='Churn', multiple='stack', bins=20)
plt.title('Distribution of Churn by Age')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()
plt.figure(figsize=(10, 3))
sns.boxplot(x='Churn', y='Avg Monthly GB Download', data=train_df)
plt.title('Average Monthly GB Download by Churn Status')
plt.xlabel('Churn')
plt.ylabel('Avg Monthly GB Download')
plt.show()

plt.figure(figsize=(8, 5))
sns.countplot(x='Contract', hue='Churn', data=train_df)
plt.title('Churn Rate by Contract Type')
plt.xlabel('Contract Type')
plt.ylabel('Count')
plt.show()
plt.figure(figsize=(10, 6))
sns.histplot(train_df, x='CLTV', hue='Churn', bins=30, kde=True, multiple="stack")
plt.title('Distribution of CLTV by Churn Status')
plt.xlabel('Customer Lifetime Value (CLTV)')
plt.ylabel('Count')
plt.show()
plt.figure(figsize=(10, 6))
sns.histplot(train_df, x='Churn Score', hue='Churn', bins=30, kde=True, multiple="stack")
plt.title('Distribution of Churn Score by Churn Status')
plt.xlabel('Churn Score')
plt.ylabel('Count')
plt.show()
plt.figure(figsize=(10, 6))
sns.boxplot(x='Churn', y='Avg Monthly Long Distance Charges', data=train_df)
plt.title('Average Monthly Long Distance Charges by Churn Status')
plt.xlabel('Churn')
plt.ylabel('Avg Monthly Long Distance Charges')
plt.show()
plt.figure(figsize=(10, 6))
sns.histplot(train_df, x='Tenure in Months', hue='Churn', bins=30, kde=True, multiple="stack")
plt.title('Distribution of Tenure by Churn Status')
plt.xlabel('Tenure in Months')
plt.ylabel('Count')
plt.show()
plt.figure(figsize=(8, 5))
sns.countplot(x='Unlimited Data', hue='Churn', data=train_df)
plt.title('Churn Rate by Unlimited Data Plan')
plt.xlabel('Unlimited Data Plan')
plt.ylabel('Count')
plt.show()
plt.figure(figsize=(10, 6))
sns.boxplot(x='Churn', y='Total Charges', data=train_df)
plt.title('Total Charges by Churn Status')
plt.xlabel('Churn')
plt.ylabel('Total Charges')
plt.show()
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Avg Monthly GB Download', y='CLTV', hue='Churn', data=train_df)
plt.title('Relationship between Monthly Charges and CLTV')
plt.xlabel('Avg Monthly GB Download')
plt.ylabel('Customer Lifetime Value (CLTV)')
plt.show()
```

Model working:

```
# Import necessary libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# Load the datasets
train_df = pd.read_csv('/train.csv')
```

```
test_df = pd.read_csv('/test.csv')
validation_df = pd.read_csv('/validation.csv')

# Drop columns that might cause data leakage or are highly predictive
leakage_cols = ['Customer ID', 'Lat Long', 'Churn Score', 'Churn Reason',
'Customer Status', 'Churn Category', 'Satisfaction Score']
train_df = train_df.drop(columns=leakage_cols, errors='ignore')
test_df = test_df.drop(columns=leakage_cols, errors='ignore')
validation_df = validation_df.drop(columns=leakage_cols, errors='ignore')

# Separate features and target for each dataset
X_train = train_df.drop(columns=["Churn"])
y_train = train_df["Churn"]
X_test = test_df.drop(columns=["Churn"])
y_test = test_df["Churn"]
X_validation = validation_df.drop(columns=["Churn"])
y_validation = validation_df["Churn"]

# Identify numeric and categorical columns
numeric_cols = X_train.select_dtypes(include=['number']).columns.tolist()
categorical_cols = [col for col in X_train.columns if col not in numeric_cols]

# Define preprocessing for numerical columns (impute and scale)
numerical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='mean')),
    ('scaler', StandardScaler())
])

# Define preprocessing for categorical columns (one-hot encode)
categorical_transformer = OneHotEncoder(handle_unknown='ignore')

# Combine preprocessing steps
preprocessor = ColumnTransformer(
    transformers=[
        ('num', numerical_transformer, numeric_cols),
        ('cat', categorical_transformer, categorical_cols)
    ])

# Define the model pipeline with Random Forest
model_pipeline = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('classifier', RandomForestClassifier(random_state=42))
])

# Train the model on the training set
model_pipeline.fit(X_train, y_train)

# Evaluate on the test set
y_test_pred = model_pipeline.predict(X_test)
test_accuracy = accuracy_score(y_test, y_test_pred)
```

```
test_conf_matrix = confusion_matrix(y_test, y_test_pred)
test_class_report = classification_report(y_test, y_test_pred)

print("Test Set Evaluation:")
print("Accuracy:", test_accuracy)
print("Confusion Matrix:\n", test_conf_matrix)
print("Classification Report:\n", test_class_report)

# Evaluate on the validation set
y_validation_pred  = model_pipeline.predict(X_validation)
validation_accuracy = accuracy_score(y_validation, y_validation_pred)
validation_conf_matrix = confusion_matrix(y_validation, y_validation_pred)
validation_class_report = classification_report(y_validation, y_validation_pred)

print("\nValidation Set Evaluation:")
print("Accuracy:", validation_accuracy)
print("Confusion Matrix:\n", validation_conf_matrix)
print("Classification Report:\n", validation_class_report)
```

REFERENCES

1. 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.

2. Geetha, V., Punitha, A., Nandhini, A., Nandhini, T., Shakila, S., & Sushmitha, R. (2020, July). Customer churn prediction in telecommunication industry using random forest classifier. In 2020 International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-5). IEEE.

3. Abdulsalam, S. O., Ajao, J. F., Balogun, B. F., & Arowolo, M. O. (2022). A churn prediction system for telecommunication company using random forest and convolution neural network algorithms. EAI Endorsed Transactions on Mobile Communications and Applications, 7(21).

4. Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. Results in Control and Optimization, 14, 100342.

5. Swetha, P., Usha, S., & Vijayanand, S. (2018, May). Evaluation of churn rate using modified random forest technique in telecom industry. In 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 2492-2497). IEEE.

6. Jain, H., Khunteta, A., & Srivastava, S. (2021). Telecom churn prediction and used techniques, datasets and performance measures: a review. Telecommunication Systems, 76, 613-630.

7. Alzubaidi, A. M. N., & Al-Shamery, E. S. (2020). Projection pursuit Random Forest using discriminant feature analysis model for churners prediction in telecom industry. International Journal of Electrical & Computer Engineering (2088-8708), 10(2).

8. Pamina, J., Raja, B., SathyaBama, S., Sruthi, M. S., & VJ, A. (2019). An effective classifier for predicting churn in telecommunication. Jour of Adv Research in Dynamical & Control Systems, 11.

9. Wu, S., Yau, W. C., Ong, T. S., & Chong, S. C. (2021). Integrated churn prediction and customer segmentation framework for telco business. Ieee Access, 9, 62118-62136.