UNRAVELING TELCO TURMOIL

A deep dive into Customer Churn Analysis

ABSTRACT

In today's fiercely competitive telecommunications landscape, retaining customers is paramount for sustaining profitability and fostering growth. This abstract encapsulates the essence of a comprehensive research project focused on understanding and mitigating customer churn within the telecommunications industry. Through rigorous data analysis, predictive modeling, and strategic intervention, this project aims to unravel the intricacies of customer attrition and devise effective retention strategies to bolster business sustainability and enhance customer satisfaction.

The telecommunications sector serves as the battleground for intense competition, where customer loyalty is constantly tested by evolving market dynamics, technological advancements, and shifting consumer preferences. Customer churn, the phenomenon of customers discontinuing services, presents a significant challenge for telecom companies, impacting revenue streams, market share, and brand reputation.

Recognizing the gravity of this issue, this research project endeavors to delve deep into the root causes of churn, deciphering patterns, and trends hidden within vast troves of customer data. At the core of this research lies a multifaceted analysis of customer behavior, market trends, and service quality metrics. Leveraging advanced analytical techniques and machine learning algorithms, the project seeks to identify key drivers of churn, ranging from service dissatisfaction and pricing sensitivity to competitor offers and market saturation.

By dissecting historical data and uncovering predictive indicators of churn, the research aims to empower telecom operators with actionable insights to anticipate and prevent customer defections proactively. The project's methodology encompasses a holistic approach, combining quantitative analysis with qualitative assessments to gain a comprehensive understanding of customer churn dynamics. Through data preprocessing, feature engineering, and model training, the research endeavors to develop robust predictive models capable of forecasting churn probabilities with high accuracy and reliability. Moreover, the project explores the integration of real-time data streams and dynamic modeling techniques to adaptively respond to changing market conditions and customer behaviors.

In addition to predictive modeling, this research project places a strong emphasis on strategic intervention and proactive retention strategies. Drawing upon insights gleaned from data analysis, the project aims to devise targeted retention initiatives, personalized offers, and service enhancements aimed at mitigating churn and enhancing customer loyalty. By aligning marketing efforts, customer service initiatives, and product development with customer preferences and needs, the research seeks to create a virtuous cycle of customer satisfaction and loyalty.

Ultimately, the goal of this research project is to equip telecom operators with the tools, insights, and strategies needed to navigate the complex landscape of customer churn effectively. By fostering a deeper understanding of customer needs, preferences, and behaviors, the project aims to empower telecom companies to build enduring relationships with their customers, driving sustainable growth and competitive advantage in an increasingly dynamic marketplace. Through collaborative efforts between academia, industry practitioners, and technology experts, this research endeavor aspires to make meaningful contributions to the field of customer relationship management and pave the way for a more customer-centric telecommunications industry.

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

In today's highly competitive telecommunications industry, retaining customers is a critical priority for companies seeking sustainable growth and profitability. The phenomenon of customer churn, where subscribers switch providers or discontinue services, poses a significant challenge for telecom operators worldwide. As such, understanding the underlying reasons for churn and developing effective strategies to mitigate it have become paramount for industry players.

This research project aims to address the complex issue of customer churn within the telecommunications sector through a comprehensive analysis of data and customer behavior. By leveraging advanced analytical techniques and predictive modeling, we seek to uncover patterns and trends within customer data that may indicate an increased likelihood of churn. Moreover, we aim to develop proactive retention strategies aimed at preserving customer loyalty and reducing attrition rates.

Through a combination of quantitative analysis and qualitative insights, we intend to gain a deeper understanding of customer behavior and preferences, thereby enabling telecom operators to anticipate and address churn more effectively. Our methodology involves several key components, including data preprocessing, feature engineering, and machine learning model development. By analyzing historical customer data and identifying predictive indicators of churn, we aim to create robust models capable of forecasting churn probabilities with accuracy.

Furthermore, we explore the integration of real-time data streams and dynamic modeling techniques to ensure the adaptability and responsiveness of our retention strategies. In addition to predictive modeling, this project places a strong emphasis on strategic intervention and proactive customer retention initiatives. By aligning marketing efforts, customer service initiatives, and product enhancements with customer preferences, we aim to create personalized experiences that foster loyalty and satisfaction.

Ultimately, the insights generated from this research project have the potential to revolutionize how telecom operators approach customer churn management. By empowering companies with actionable insights and effective retention strategies, we aspire to drive positive outcomes for both businesses and customers alike.

1.1. PROBLEM STATEMENT

The economic recession has led to a challenging job market scenario, marked by reduced consumer spending and financial constraints across various industries. A key contributing factor to this downturn is the decline in consumer demand, which has been exacerbated by customer churn, particularly within the telecommunications sector. Customer churn, the loss of subscribers or clients, not only directly impacts revenue but also reflects broader shifts in consumer behavior and market dynamics. This research project aims to analyze the underlying drivers of customer churn within the telecommunications industry amidst the recession. Leveraging data analytics and predictive modeling techniques, the study seeks to identify patterns and trends associated with churn behavior, facilitating the development of proactive intervention strategies to mitigate attrition and enhance customer retention. By understanding the interconnectedness between customer churn, economic recession, and market dynamics across different sectors, stakeholders can devise targeted interventions to stabilize the job market, stimulate economic growth, and foster resilience in the face of ongoing challenges.

1.2. LITERATURE REVIEW

1.2.1. INTRODUCTION

The literature review serves as a critical component of this research project, offering a comprehensive examination of existing scholarly works, theories, and empirical studies relevant to the phenomenon of customer churn within the telecommunications industry.

This section provides a foundation for understanding the conceptual framework and theoretical underpinnings that inform the research objectives and methodologies. By synthesizing and analyzing a wide range of literature, this review aims to identify key themes, trends, and gaps in the current body of knowledge regarding customer churn, thereby guiding the research process and informing subsequent chapters. This introduction sets the stage for a detailed exploration of the literature, highlighting the significance of the topic and the need for further investigation to address the challenges posed by churn in the telecommunications sector.

CUSTOMER CHURN PREDICTION IN TELECOM USING MACHINE LEARNING IN BIG DATA PLATFORM

The telecommunications sector, marked by intense competition and technological advancements, relies on strategies to boost revenues, with customer retention emerging as the most profitable. Retaining existing customers is more cost-effective than acquiring new ones or upselling. Predicting customer churn, and the departure of customers is crucial for revenue stability.

This study aims to develop a churn prediction model using machine learning techniques on a big data platform, with a focus on feature engineering and selection. The dataset spans nine months and encompasses various data formats, including structured, semi-structured, and unstructured data, amounting to about 70 Terabytes on HDFS. Social Network Analysis (SNA) features are incorporated to enhance the model's predictive performance. Four machine learning algorithms are tested: *Decision Tree, Random Forest, Gradient Boost Machine Tree, and XGBoost*, with XGBoost yielding the best results.

The study addresses challenges posed by unbalanced datasets, extensive features, and missing values. SyriaTel, one of the telecom companies studied, faces an unbalanced dataset, with churn customers representing only about 5%. The implementation of a big data platform, specifically tailored to SyriaTel's needs, facilitates data processing and feature extraction. The proposed churn prediction method involves installing a customized big data platform, SYTL-BD framework, comprising tools like HDFS for data storage, Spark for data processing, and Zeppelin for development. Hardware resources include 12 nodes with specific RAM, storage, and processor capacity.

The research demonstrates the effectiveness of machine learning in predicting customer churn, leveraging big data and social network analysis to enhance model performance. By deploying advanced techniques and tailored platforms, telecom companies like SyriaTel can improve customer retention strategies and mitigate revenue loss due to churn.

CUSTOMER CHURN PREDICTION USING MACHINE LEARNING APPROACHES

The article delves into the pressing issue of customer churn within the telecommunications sector, highlighting its significance for businesses' revenue streams and service quality. It elucidates the concept of churn as the loss of customers over a specific timeframe and underscores the importance of predicting early customer departures to mitigate revenue loss.

The study examines the telecommunications landscape in China, where market saturation and heightened competition underscore the need for effective churn prediction models. Leveraging a dataset focused on customer behavior, the paper explores various preprocessing techniques, such as SMOTE-ENN normalization, to enhance data quality and balance class distributions.

Furthermore, it investigates an array of machine learning algorithms, including SVM, boosting algorithms, and ensemble classifiers, to develop robust churn prediction models. The proposed methodology outlines steps for data preprocessing, feature selection, and algorithm selection, emphasizing the importance of addressing class imbalance issues. Experimental results, evaluated using F1-score as a performance metric, demonstrate the efficacy of different algorithms, such as Decision Trees and Random Forests, in accurately predicting churn.

Overall, the document offers valuable insights into the complexities of churn prediction in the telecommunications industry, emphasizing the pivotal role of machine learning techniques and sound data preprocessing strategies in driving actionable insights for businesses.

CUSTOMER CHURNING ANALYSIS USING MACHINE LEARNING ALGORITHMS

The telecommunications industry is characterized by fierce competition between established providers and emerging firms offering specialized services at competitive prices. This competition significantly impacts pricing strategies and market dynamics. Customer churn, the phenomenon of customers discontinuing services, poses a major challenge for telecom companies.

To address this challenge, accurate churn prediction models are essential for retaining customers and maximizing revenue. In this study, various machine learning (ML) techniques are explored for churn prediction, including decision trees, K-Nearest Neighbors (KNN), logistic regression, and ensemble methods.

The research utilizes a dataset obtained from Kaggle, which is divided into training and testing sets to evaluate the effectiveness of these algorithms. Ensemble techniques and big data platforms such as Jupyter are employed to enhance prediction accuracy. The findings of the study suggest that neural network algorithms may offer viable alternatives to traditional statistical approaches for churn prediction. Specifically, Stochastic Gradient Booster emerges as a promising algorithm for churn prediction.

The study emphasizes the importance of accurate churn prediction for telecom companies to implement effective customer retention strategies and maintain competitiveness in the market. Overall, this research highlights the critical role of advanced analytics techniques in predicting and managing customer churn in the telecom industry. By leveraging ML algorithms and big data platforms, telecom companies can gain valuable insights into customer behavior, enabling them to proactively address churn and maximize profitability.

A STUDY ON CUSTOMER CHURN PREDICTION

This study developed a customer churn prediction model using a telecom dataset containing demographic, account, and service usage information of approximately 7,043 customers. After data extraction and cleaning to address missing values, duplicates, and outliers, exploratory data analysis (EDA) is conducted.

EDA involves statistical and visual techniques to understand the data distribution, identify significant variables affecting churn, and prepare features for modeling. The dataset is divided into categorical and numerical features, with appropriate preprocessing techniques applied to each type. Feature engineering extracts relevant features like call duration and frequency, while categorical variables are encoded using methods such as one-hot encoding. Subsequently, various machine learning models including XGB Classifier, Light GBM Classifier, Random

Forest Classifier, and Decision Tree Classifier are experimented with. The top-performing models are selected based on cross-validation performance metrics.

To improve prediction accuracy, a stack of all four classifiers is implemented, where predictions from base classifiers are used as features for a meta-classifier. Model evaluation involves calculating accuracy, F1-score, and generating confusion matrices and ROC curves. The final stacked model achieves an accuracy of 87% and an AUC-ROC score of 0.91, outperforming individual models in terms of accuracy and robustness. This comprehensive approach highlights the effectiveness of ensemble techniques in customer churn prediction.

CUSTOMER CHURN PREDICTION USING MACHINE LEARNING

The Research paper provides a comprehensive overview of key theoretical concepts and previous studies related to churn prediction in both business-to-business (B2B) and business-to-consumer (B2C) contexts.

It begins by defining Customer Relationship Management (CRM) as a strategic approach aimed at enhancing profitability and improving customer relations through tailored products or services, emphasizing the role of CRM systems in facilitating customer knowledge.

Customer churn management is then outlined as crucial for identifying and mitigating customer attrition, with a distinction made between voluntary and non-voluntary churners. Machine Learning (ML) is introduced as a critical tool for churn prediction, encompassing various learning paradigms such as supervised learning, which includes regression and classification tasks. Ensemble learning techniques like bagging and boosting are discussed for improving predictive accuracy, alongside specific classifiers such as Random Forest, XGBoost, and Naïve Bayes.

The review also addresses imbalanced learning methods to handle datasets where one class dominates, highlighting the importance of adjusting class distributions for improved classifier performance. Overall, the literature review establishes the theoretical foundation for understanding churn prediction methodologies and underscores the significance of machine learning in addressing churn management challenges across B2B and B2C sectors.

CUSTOMER CHURN: A STUDY OF FACTORS AFFECTING CUSTOMER CHURN USING MACHINE LEARNING

The research paper focuses on churn prediction within the telecommunications industry, aiming to develop a machine-learning model to forecast customer churn and identify key factors influencing it to prevent customer attrition. Data collection involves obtaining historical customer data from a telecommunications company dataset, encompassing demographics, service subscriptions, and billing details.

With 7055 records and 20 features, including churn status, service subscriptions, account information, and demographic data, data preparation involves aggregating transactional data and conducting exploratory analysis to understand churn patterns across various attributes—visualizations aid in illustrating these relationships.

Feature engineering, utilizing the FeatureTools library, generates 724 new features through mathematical transformations. Feature selection, employing tree-based algorithms, highlights variables such as Contract, TechSupport, Online Security, and Tenure as crucial for churn prediction. Model evaluation, based on AUROC curves, identifies xGBoost as the most effective classifier.

Visualization techniques, including LIME analysis, aid in interpreting model predictions and informing actionable strategies for reducing churn. The study underscores the significance of model explainability and data-driven decision-making in addressing customer churn challenges within the telecommunications sector.

CUSTOMER CHURN PREDICTION

The research paper focuses on developing a system for customer churn prediction in the telecommunications industry. The system design encompasses several key steps, including data preprocessing, feature selection, handling missing values, exploratory data analysis, model generation, and model evaluation. Feature selection is emphasized to choose valuable variables that enhance classification performance, with logistic regression, Random Forest, and KNN identified as suitable algorithms for predicting churn probability.

Data preprocessing involves filtering high-value customer records based on recharge amounts and handling missing values through imputation. Exploratory data analysis addresses data imbalance and outliers in important feature columns. Model generation includes applying PCA for dimensionality reduction and selecting the right number of components for model building.

The models are evaluated using accuracy, precision, and recall metrics, with Random Forest identified as the best-performing model based on cross-validation scores and classification report results. The confusion matrix illustrates correct and incorrect predictions of churn and non-churn instances, providing insights into model performance. Overall, the system aims to provide accurate predictions of customer churn to aid decision-making in retention strategies within the telecommunications sector.

A PREDICTION MODEL OF CUSTOMER CHURN CONSIDERING CUSTOMER VALUE: AN EMPIRICAL RESEARCH OF TELECOM INDUSTRY IN CHINA

The research paper focuses on developing a system for customer churn prediction in the telecommunications industry. The system design encompasses several key steps, including data preprocessing, feature selection, handling missing values, exploratory data analysis, model generation, and model evaluation. Feature selection is emphasized to choose valuable variables that enhance classification performance, with logistic regression, Random Forest, and KNN identified as suitable algorithms for predicting churn probability.

Data preprocessing involves filtering high-value customer records based on recharge amounts and handling missing values through imputation. Exploratory data analysis addresses data imbalance and outliers in important feature columns. Model generation includes applying PCA for dimensionality reduction and selecting the right number of components for model building.

The models are evaluated using accuracy, precision, and recall metrics, with Random Forest identified as the best-performing model based on cross-validation scores and classification report results. The confusion matrix illustrates correct and incorrect predictions of churn and non-churn instances, providing insights into model performance. Overall, the system aims to provide

accurate predictions of customer churn to aid decision-making in retention strategies within the telecommunications sector.

ENSEMBLE METHODS IN CUSTOMER CHURN PREDICTION: A COMPARATIVE ANALYSIS OF THE STATE-OF-THE-ART

The rapid growth of the financial industry has led to the expansion of customer resources and commercial bank scales. However, amidst declining profit rates, there's a noticeable increase in financial disintermediation. Customers are demanding more customized and high-return products and services, prompting banks to adopt user-centered business models. This has intensified competition within the banking sector, not only from traditional banks but also from tech-based newcomers offering more personalized services through efficient data-driven methods.

This heightened competition has raised the risk of customer churn in commercial banks. The loss of customers not only leads to decreased profits but also tarnishes a bank's reputation, making it harder to attract new customers. Therefore, customer retention has become a crucial aspect of customer relationship management (CRM). Studies show that retaining just 5% more customers can double profits, but it's generally more expensive to retain customers than to acquire new ones.

To tackle the challenge of customer churn, predictive models are developed to identify potential churn signals in advance. These models, often based on data mining and machine learning techniques, aim to classify customers into "churn" or "not churn" categories. These models can be categorized into Individual Machine Learning (IML), Ensemble Machine Learning (EML), and Deep Learning (DL) models.

IML models, such as Logistic Regression and Decision Trees, are relatively simple but effective in customer churn prediction tasks. EML models combine multiple individual models to improve prediction accuracy further. Boosting, Bagging, and multi-stage ensemble learning are common EML methods used in churn prediction.

In recent years, deep learning models, with their adaptive feature extraction capabilities, have gained attention in the churn prediction community. Despite their poor interpretability, deep neural networks significantly reduce the need for manual feature engineering. Additionally, techniques from other AI domains, such as self-attention mechanisms from natural language processing and Deep Q Network from reinforcement learning, are being integrated into churn prediction models.

A novel hybrid neural network model with self-attention enhancement (HNNSAE) is proposed in this paper to efficiently extract highly correlated features and reduce the impact of low-correlated features on model performance. This model combines entity embedding technology with multi-head self-attention blocks and a multi-layer perceptron for training. Experimental results demonstrate that HNNSAE outperforms baseline models, showing promise in improving customer churn prediction accuracy. The study also verifies several hypotheses regarding sample imbalance, overfitting risks, and the impact of entity embedding on model performance.

CUSTOMER CHURN PREDICTION IN INFLUENCER COMMERCE: AN APPLICATION OF DECISION TREES

Customer churn, the loss of customers, poses a significant challenge for businesses across various industries. Retaining loyal customers is not only cost-effective but also crucial for a company's profitability and reputation. Predicting churn helps companies refine their retention strategies to minimize losses and make informed marketing decisions. This is particularly important in industries where long-term customer relationships are valued, such as airlines, banking, e-commerce, and mobile applications.

In recent years, there has been a surge in customer churn prediction research, with a focus on improving model accuracy using various machine learning techniques. These studies often involve applying algorithms like Logistic Regression, Decision Trees, Support Vector Machines, and ensemble methods to telecom data and other industry-specific datasets. Additionally, data preprocessing methods are explored to handle issues like data imbalance and improve prediction accuracy.

While churn prediction has been extensively studied in traditional industries, such as telecom and finance, research in emerging businesses like e-commerce is limited due to data imbalance and difficulties in defining churn points. However, recent studies have attempted to address these challenges by developing new data preprocessing methods and applying optimized machine learning algorithms. These studies employ techniques like customer clustering, logistic regression, random forest, and gradient boosted trees to predict churn accurately.

In this study, the focus is on predicting customer churn in a new type of e-commerce based on sales data generated between influencers and customers on social media platforms like Instagram. Similar to traditional e-commerce, influencers promote and sell products, and churn is defined based on customer purchase behavior. Specifically, customers who make only one purchase from an influencer are classified as churners, while those making two or more purchases are considered loyal. This study aims to leverage insights from traditional e-commerce churn prediction research to address customer retention challenges in the influencer marketing domain.

HIGH-PERFORMANCE CUSTOMER CHURN PREDICTION SYSTEM BASED ON SELF-ATTENTION

The expansion of e-commerce has provided customers with numerous purchasing options, increasing the risk of churn due to easy information sharing and the ability to switch between online shopping platforms. Despite this, limited research on predicting e-commerce customer churn has been conducted due to challenges such as unbalanced data and defining churn points.

Studies on e-commerce customer churn prediction can be categorized into three themes: developing new data preprocessing methods to address data imbalance, optimizing machine learning algorithms, and defining churn criteria. Various techniques have been explored, including customer clustering, ensemble models, and different machine learning algorithms like SVM, Decision Trees, Logistic Regression, Random Forest, XGBoost, and others.

Different studies define churn customers differently, with definitions based on factors like service usage duration or purchase history. For instance, churn might be defined as no service usage within a certain period or no purchase history within the data collection period.

In a recent study, the prediction of customer churn was applied to sales data generated between influencers and customers on social media platforms like Instagram. The churn point was defined similarly to e-commerce studies, where a customer is classified as a churner if they make only one purchase from an influencer and as loyal if they make two or more purchases. This approach leverages insights from previous e-commerce churn prediction studies to address customer retention challenges in social media influencer marketing.

CUSTOMER CHURN PREDICTION USING COMPOSITE DEEP LEARNING TECHNIQUE

The success of any company depends heavily on satisfying its customers, especially in subscription-based services where maintaining existing customers is crucial for expansion. In competitive sectors like telecommunications, losing even one customer due to dissatisfaction can have significant financial and reputational consequences. Customer churn, when frequent customers stop using a company's products or services, causes substantial losses. Predicting and preventing churn is vital for revenue growth and business sustainability.

Traditional churn prediction methods often face scalability issues, leading researchers to explore advanced techniques like deep learning. Deep learning models, such as BiLSTM-CNN, aim to accurately predict churn by extracting patterns from historical data. This hybrid model combines bidirectional long/short-term memory (BiLSTM) and convolutional neural network (CNN) to improve prediction accuracy by considering context information effectively.

The problem is formulated as a binary classification task, distinguishing between churners and non-churners. The goal is to develop a model that accurately identifies churn based on provided training data and class labels. This novel approach addresses the limitations of existing classifiers and aims to forecast customer turnover more accurately by integrating contextual information using BiLSTM and CNN layers.

1.3. RESEARCH OBJECTIVES

The project aims to comprehensively tackle the challenge of customer churn in the telecommunications industry. It involves identifying key drivers of churn, utilizing advanced analytics to develop precise churn prediction models, and implementing proactive retention strategies such as personalized offers and targeted campaigns. Through empirical analysis, the effectiveness of these strategies will be evaluated, and actionable insights will be distilled to empower telecom operators in optimizing churn management efforts for sustainable business growth.

The objective of the research is to:

1. Identify and analyze the key drivers of customer churn within the telecommunications industry, including factors such as service dissatisfaction, pricing sensitivity, competitor offerings, and market dynamics.

2. Leverage data analytics and predictive modeling techniques to develop accurate and reliable churn prediction models capable of forecasting churn probabilities for individual customers.

3. Design and implement proactive retention strategies aimed at mitigating churn and enhancing customer loyalty, including personalized offers, targeted marketing campaigns, and service improvements.

4. Evaluate the effectiveness of the developed churn prediction models and retention strategies through empirical analysis and real-world implementation, measuring their impact on churn reduction and customer satisfaction.

5. Provide actionable insights and recommendations to telecom operators based on the findings of the research, enabling them to optimize their churn management efforts and improve overall business performance.

1.4. IMPORTANCE OF THE STUDY

The significance of this research project lies in its potential to yield substantial benefits for both telecom operators and consumers within the telecommunications industry. By addressing the challenges posed by customer churn through a multifaceted approach, the study offers several key advantages:

- *Enhanced Customer Retention*
  The project aims to identify and address the root causes of customer churn, ultimately leading to improved customer satisfaction and loyalty.

- *Optimized Resource Allocation*
  Through the development of accurate churn prediction models, telecom operators can allocate resources more efficiently, focusing efforts on retaining high-value customers and maximizing returns on investment.

- *Increased Revenue and Profitability*
  By reducing churn rates and retaining more customers, telecom operators can experience a boost in revenue streams and profitability, stemming from higher subscription rates and reduced customer acquisition costs.

- *Competitive Advantage*
  Leveraging advanced data analytics and predictive modeling techniques provides telecom operators with a competitive edge, enabling them to differentiate their offerings, attract new customers, and retain existing ones more effectively than their competitors.

1.5. CHAPTER OUTLINE

The chapter outline for the project encompasses a structured approach to addressing the challenges of customer churn in the telecommunications industry. It begins with an introduction providing context and background, followed by a comprehensive review of existing literature and research related to customer churn. The subsequent chapters delve into the system analysis and requirements, including problem definition, requirements specification, and conceptual models.

The research methodology chapter outlines the approach taken to conduct the study, followed by chapters detailing the data collection and analysis processes. The findings and results of the study are presented in a dedicated chapter, followed by a discussion of the implications and recommendations derived from the research. The chapter outline concludes with a summary and conclusion, consolidating the key insights and contributions of the study.

1.6. OVERVIEW OF THE REPORT

We've discussed the multifaceted nature of the research objectives, aimed at comprehensively addressing the challenges posed by customer churn within the telecommunications industry. Leveraging advanced data analytics and predictive modeling techniques, the project seeks to develop accurate churn prediction models and design proactive retention strategies to mitigate churn and foster customer loyalty.

Through empirical analysis and real-world implementation, the effectiveness of these strategies will be evaluated, and actionable insights will be distilled for telecom operators. The chapter also outlined the importance and benefits of the study, highlighting its potential to enhance customer retention, optimize resource allocation, increase revenue, and gain a competitive advantage. Finally, we provided a brief overview of the chapter outline, detailing the structure and contents of the subsequent chapters.

2. DATA SOURCES AND DESCRIPTION

2.1. DATA SOURCE AND DESCRIPTION:

The dataset “Telco-Customer-Churn_1” comprises customer information from a telecom company sourced from Kaggle and OpenSignal. Each row represents a unique customer, with various attributes providing insights into their service subscriptions, billing details, and churn status.

The link to the dataset is https://www.kaggle.com/datasets/barun2104/telecom-churn

Here's an overview of the key attributes:

1. customerID: Unique identifier for each customer.
2. gender: Gender of the customer (Male or Female).
3. Senior Citizen: Indicates whether the customer is a senior citizen (0 for No, 1 for Yes).
4. Partner: Denotes whether the customer has a partner (Yes or No).
5. Dependents: Specifies whether the customer has dependents (Yes or No).
6. tenure: Duration of the customer's subscription in months.
7. PhoneService: Indicates whether the customer subscribes to phone service (Yes or No).
8. MultipleLines: Specifies whether the customer has multiple phone lines (Yes, No, or No phone service).
9. InternetService: Type of internet service subscribed by the customer (DSL, Fiber optic, or No).
10. OnlineSecurity: Indicates whether the customer has online security service (Yes, No, or No internet service).
11. OnlineBackup: Denotes whether the customer has online backup service (Yes, No, or No Internet service).
12. DeviceProtection: Specifies whether the customer has a device protection service (Yes, No, or No Internet service).
13. TechSupport: Indicates whether the customer has tech support service (Yes, No, or No internet service).
14. StreamingTV: Denotes whether the customer subscribes to streaming TV (Yes, No, or No internet service).

15. StreamingMovies: Specifies whether customers subscribe to streaming movies (Yes, No, or No internet service).
16. Contract: Type of contract the customer has (Month-to-month, One year, Two years).
17. Paperless billing: Denotes whether the customer opts for paperless billing (Yes or No).
18. PaymentMethod: Payment method used by the customer (Electronic check, Mailed check, Bank transfer, Credit card).
19. Monthly charges: Amount charged to the customer monthly.
20. Total charges: Total amount charged to the customer over the entire tenure.
21. Churn: Indicates whether the customer has churned (No or Yes).

2.2. POTENTIAL DATA SOURCES

1. Kaggle: A platform hosting various datasets, including telecom-related datasets, sourced from contributors worldwide.
2. OpenSignal: A company specializing in wireless coverage mapping, providing insights into mobile network performance and customer experience.

2.3. CHALLENGES

1. Ensuring Data Quality: Maintaining data accuracy, completeness, and consistency across sources.
2. Data Privacy Compliance: Adhering to data privacy regulations (e.g., GDPR) to protect customer information.
3. Integration Complexity: Consolidating data from multiple sources may require careful coordination and preprocessing.
4. Unstructured Data Analysis: Extracting insights from unstructured data (e.g., social media feedback) necessitates advanced analytics techniques.
5. Governance and Security: Establishing robust policies and procedures for data governance, storage, and usage to safeguard sensitive information.

This dataset presents a valuable resource for analyzing customer behavior, subscription patterns, and factors influencing churn in the telecom industry.

3. DATA CLEANING AND PREPROCESSING PLAN

3.1. INTRODUCTION

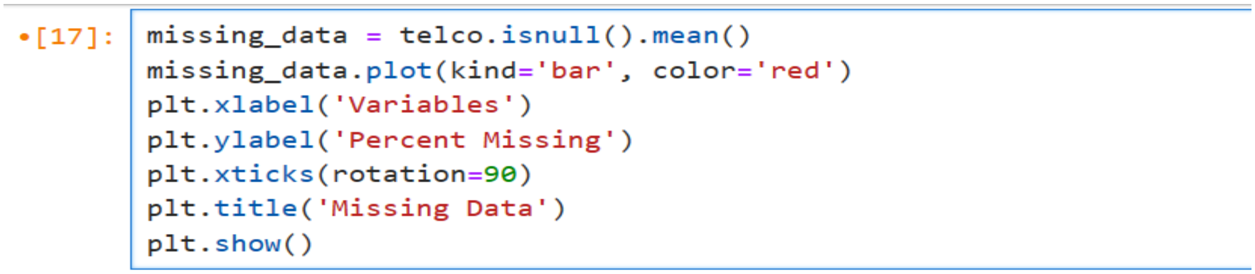
Data cleaning and preprocessing are fundamental steps in the data analysis pipeline, aimed at transforming raw data into a clean and structured format suitable for analysis and modeling. Data cleaning involves addressing missing values, removing duplicates, correcting errors, handling outliers, and normalizing or scaling numerical features. On the other hand, data preprocessing involves encoding categorical variables, feature engineering to extract meaningful information, feature selection to reduce dimensionality, data splitting for model evaluation, and normalization or scaling of numerical features. Together, these processes ensure the quality, integrity, and suitability of the data for accurate and reliable analysis and modeling tasks.

3.2. STEPS FOLLOWED FOR DATA CLEANING AND PRE-PROCESSING

1. HANDLING MISSING DATA

- The data exploration part calculates the percentage of missing values for each attribute using `telco.isnull().mean()`.

- Visualizing the missing data percentage using a bar plot helps in identifying attributes with missing values.



```
•[17]: missing_data = telco.isnull().mean()
       missing_data.plot(kind='bar', color='red')
       plt.xlabel('Variables')
       plt.ylabel('Percent Missing')
       plt.xticks(rotation=90)
       plt.title('Missing Data')
       plt.show()
```

Fig 3.1 Python code for finding any missing value in the dataset

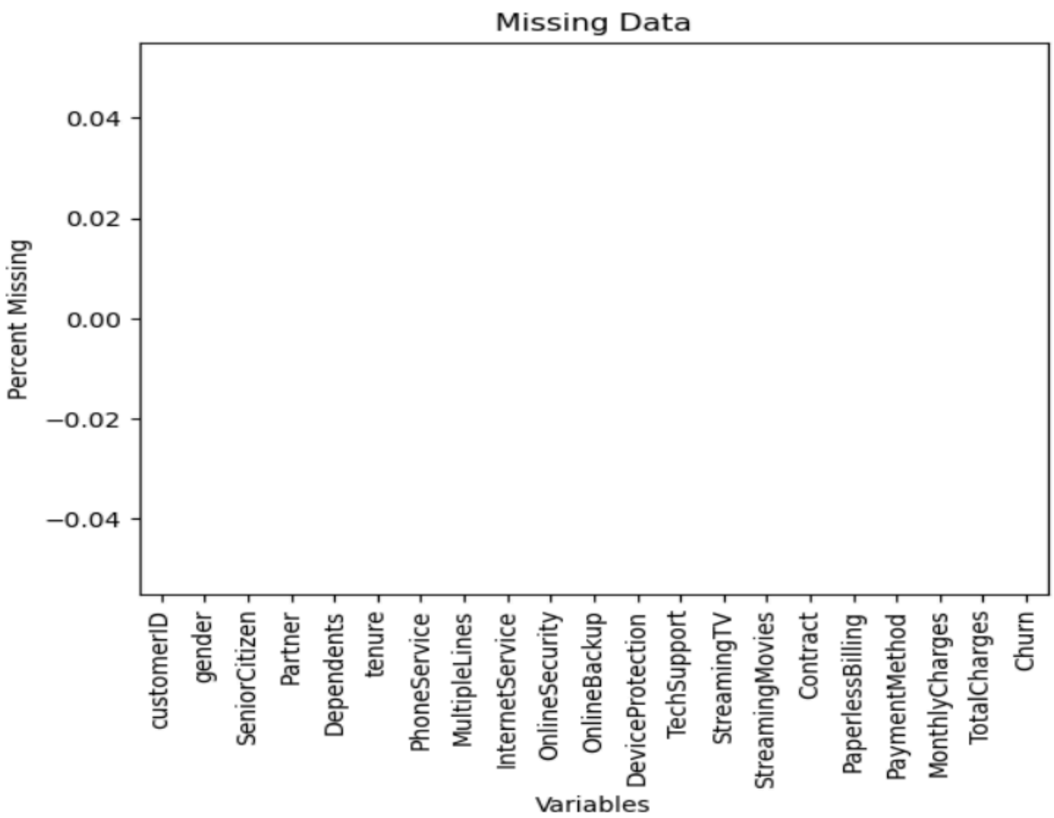
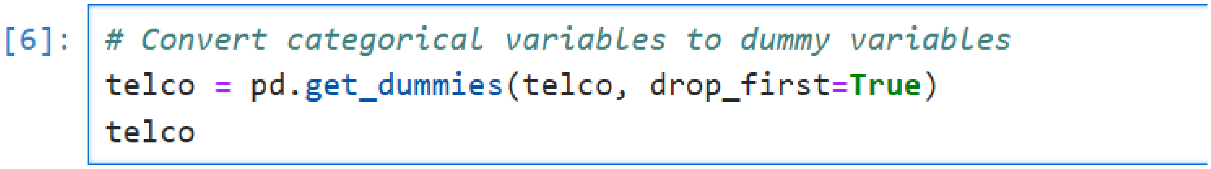


Fig 3.2 The output shows that we don't have any missing values in the dataset

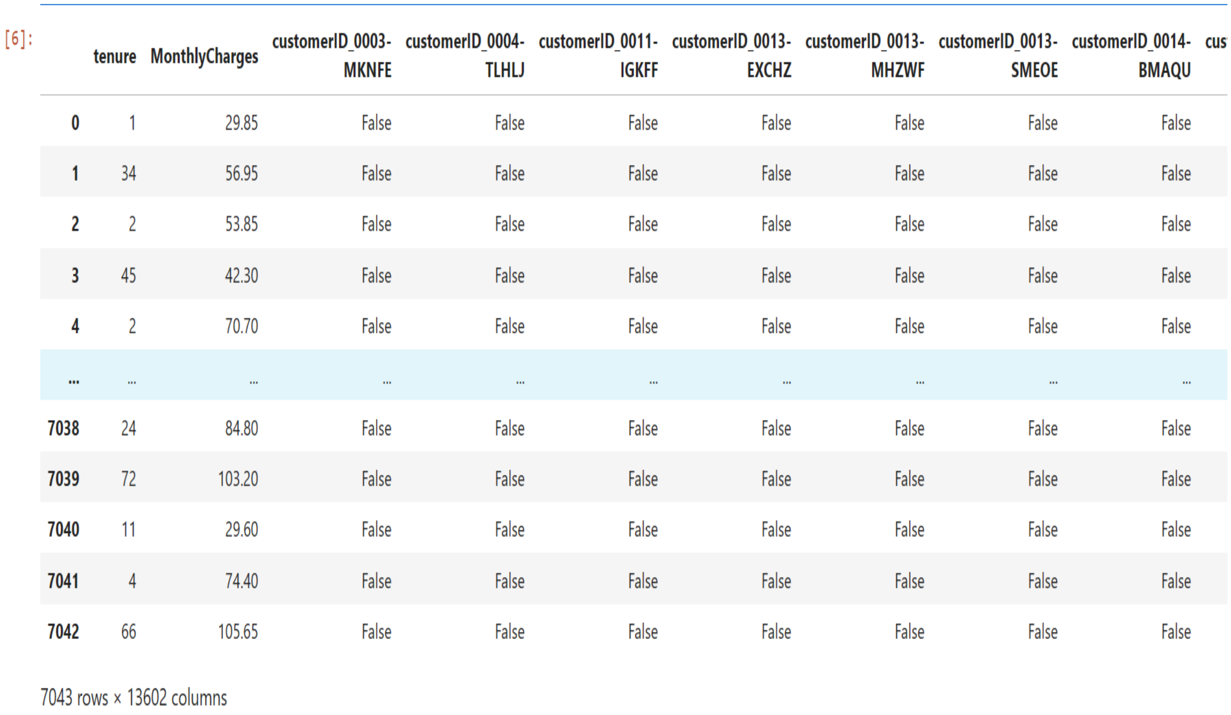
2. CONVERTING CATEGORICAL VARIABLES

- After exploring missing data, categorical variables are converted into dummy variables using `pd.get_dummies(telco, drop_first=True)`. This step creates binary variables for each category of categorical attributes, which is a common preprocessing step in machine learning.



```
[6]: # Convert categorical variables to dummy variables
     telco = pd.get_dummies(telco, drop_first=True)
     telco
```

Fig 3.3 Python code for converting categorical variables into dummy variables



[6]:

	tenure	MonthlyCharges	customerID_0003-MKNFE	customerID_0004-TLHLJ	customerID_0011-IGKFF	customerID_0013-EXCHZ	customerID_0013-MHZWF	customerID_0013-SMEOE	customerID_0014-BMAQU	cus
0	1	29.85	False	False	False	False	False	False	False	
1	34	56.95	False	False	False	False	False	False	False	
2	2	53.85	False	False	False	False	False	False	False	
3	45	42.30	False	False	False	False	False	False	False	
4	2	70.70	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	
7038	24	84.80	False	False	False	False	False	False	False	
7039	72	103.20	False	False	False	False	False	False	False	
7040	11	29.60	False	False	False	False	False	False	False	
7041	4	74.40	False	False	False	False	False	False	False	
7042	66	105.65	False	False	False	False	False	False	False	

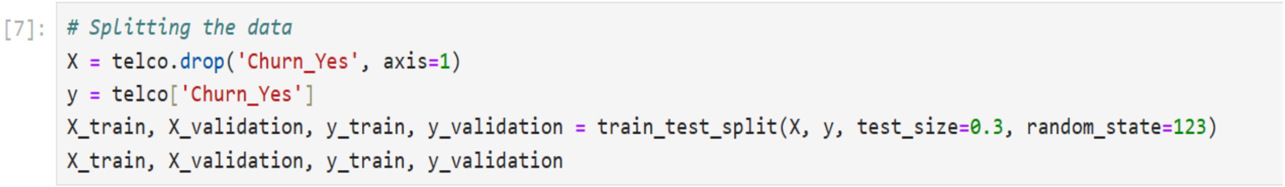
7043 rows × 13602 columns

Fig 3.4 All the categorical variables got converted into dummy variables

DataFrame will contain the original numerical variables along with the one-hot encoded binary columns representing the categorical variables.

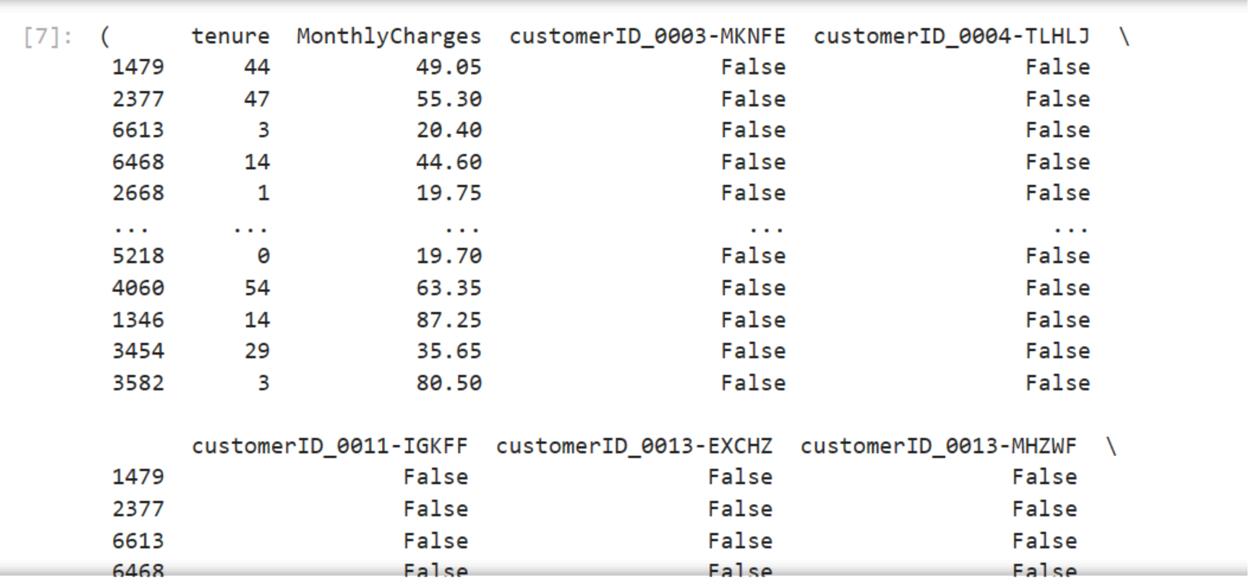
3. SPLITTING THE DATA

- The data is split into training and validation sets using the `train_test_split()` function. This step separates the feature variables (`X`) from the target variable (`y`), and then splits them into training and validation sets for model training and evaluation, respectively.



```
[7]: # Splitting the data
     X = telco.drop('Churn_Yes', axis=1)
     y = telco['Churn_Yes']
     X_train, X_validation, y_train, y_validation = train_test_split(X, y, test_size=0.3, random_state=123)
     X_train, X_validation, y_train, y_validation
```

Fig 3.5 Python code for splitting the data, and separating the target variable



```
[7]: (       tenure  MonthlyCharges  customerID_0003-MKNFE  customerID_0004-TLHLJ  \
      1479      44           49.05                  False                  False
      2377      47           55.30                  False                  False
      6613       3           20.40                  False                  False
      6468      14           44.60                  False                  False
      2668       1           19.75                  False                  False
      ...      ...             ...                    ...                    ...
      5218       0           19.70                  False                  False
      4060      54           63.35                  False                  False
      1346      14           87.25                  False                  False
      3454      29           35.65                  False                  False
      3582       3           80.50                  False                  False

            customerID_0011-IGKFF  customerID_0013-EXCHZ  customerID_0013-MHZWF  \
      1479                  False                  False                  False
      2377                  False                  False                  False
      6613                  False                  False                  False
      6468                  False                  False                  False
```

Fig 3.6 Data, that contains all the variables except for Target variable

The code segment is responsible for splitting the dataset into training and validation sets for both the features (X) and the target variable (y) using the train_test_split function from scikit-learn. The features (X) are obtained by dropping the 'Churn_Yes' column from the telco DataFrame, and the target variable (y) is assigned the 'Churn_Yes' column. The test_size parameter specifies the proportion of the dataset to include in the validation set, in this case, 30%. The random_state parameter ensures the reproducibility of the split, ensuring that the same split is generated each time the code is run with the same seed value (123). After execution, the X_train, X_validation, y_train, and y_validation variables will contain the training and validation sets for the features and target variables, respectively.

- `X_train`: This variable contains the training set for the features (independent variables), which is a data frame excluding the 'Churn_Yes' column.

- `X_validation`: This variable contains the validation set for the features, which is also a DataFrame excluding the 'Churn_Yes' column.

- `y_train`: This variable contains the training set for the target variable (dependent variable), which is a Series representing whether a customer churned or not.

- `y_validation`: This variable contains the validation set for the target variable, which is a Series similar to `y_train`.

Therefore, `X_train` and `y_train` are used to train the predictive model, while `X_validation` and `y_validation` are used to evaluate its performance on unseen data.

4. EXPLORATORY DATA ANALYSIS PLAN

4.1. INTRODUCTION

EDA stands for Exploratory Data Analysis. It is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. The goal of EDA is to understand the data, discover patterns, spot anomalies, and identify relationships between variables. During EDA, analysts typically examine the distribution of individual variables, explore relationships between variables, detect outliers, and assess the overall quality of the data. EDA helps in formulating hypotheses, selecting appropriate statistical models, and preparing the data for further analysis. It is a crucial initial step in the data analysis process, providing insights that inform subsequent modeling decisions.

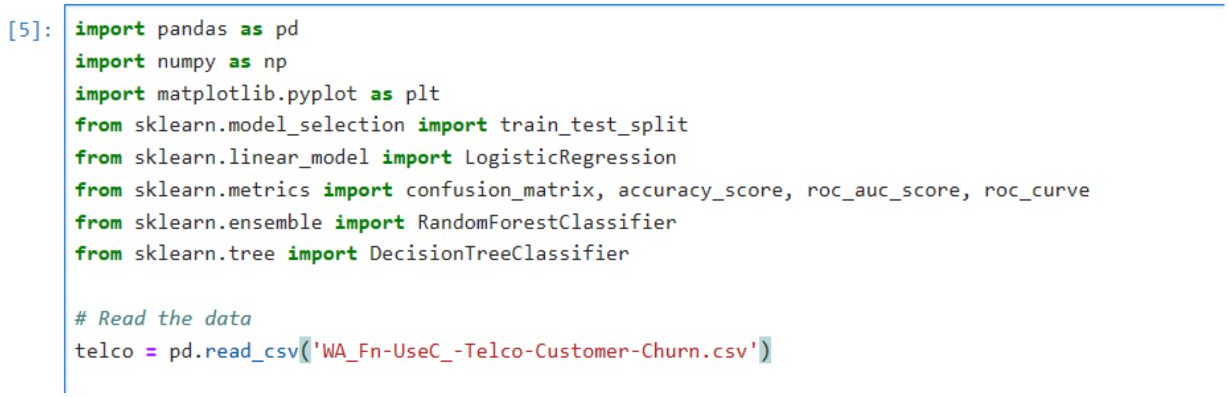
4.2. EXPLORATORY DATA ANALYSIS (EDA) PLAN

1. Data Overview: Understand the basic characteristics of the dataset, including the number of observations, attributes, and data types.
2. Summary Statistics: Compute descriptive statistics to summarize numerical attributes.
3. Univariate Analysis: Visualize the distribution of individual attributes using histograms for numerical attributes and bar plots for categorical attributes.
4. Bivariate Analysis: Explore relationships between pairs of attributes using pair plots and correlation matrices.
5. Target Variable Analysis: Examine the distribution of the target variable to understand class balance/imbalance and its impact on modeling.

4.3. EXPLORATORY DATA ANALYSIS

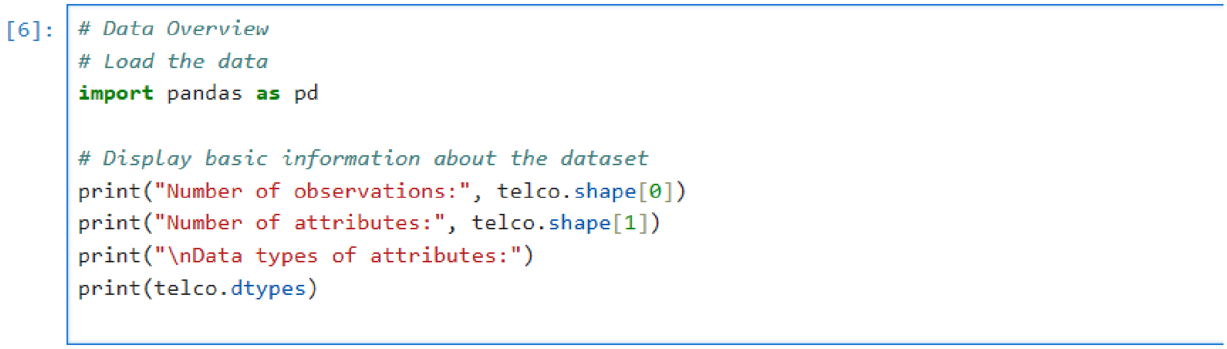
1. DATA OVERVIEW

- This step involves loading the dataset and displaying basic information such as the number of observations, number of attributes, and data types of attributes. This is performed in the code snippet where the `shape` and `dtypes` attributes of the DataFrame `telco` are printed.



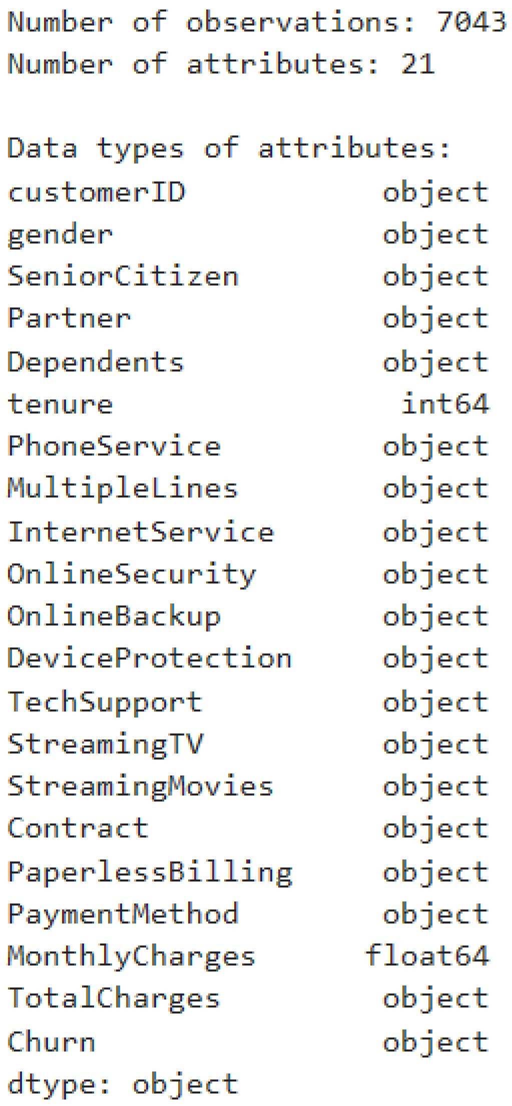
```
[5]: import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LogisticRegression
     from sklearn.metrics import confusion_matrix, accuracy_score, roc_auc_score, roc_curve
     from sklearn.ensemble import RandomForestClassifier
     from sklearn.tree import DecisionTreeClassifier

     # Read the data
     telco = pd.read_csv('WA_Fn-UseC_-Telco-Customer-Churn.csv')
```

Fig 4.1 Python code for importing all the libraries and extracting the dataset



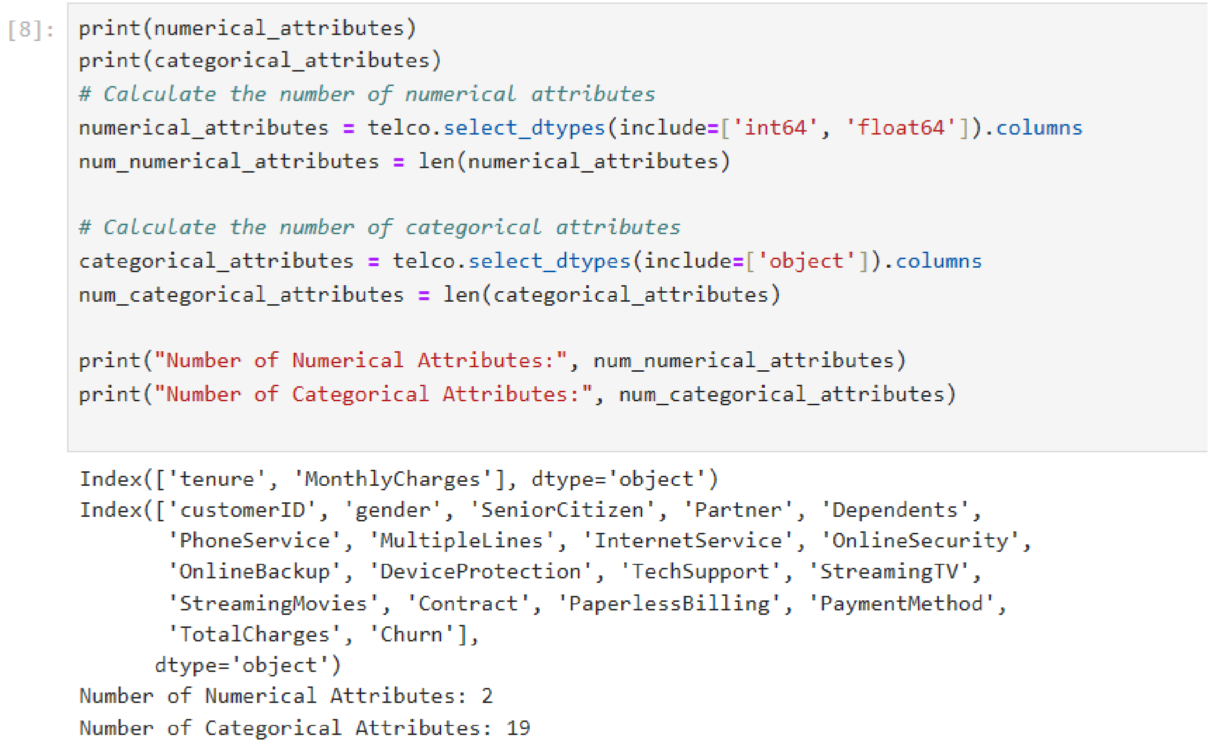
```
[6]: # Data Overview
     # Load the data
     import pandas as pd

     # Display basic information about the dataset
     print("Number of observations:", telco.shape[0])
     print("Number of attributes:", telco.shape[1])
     print("\nData types of attributes:")
     print(telco.dtypes)
```

Fig 4.2 Python code for overviewing all the data by displaying number opf observations and attributes



```
Number of observations: 7043
Number of attributes: 21

Data types of attributes:
customerID           object
gender               object
SeniorCitizen        object
Partner              object
Dependents           object
tenure                int64
PhoneService         object
MultipleLines        object
InternetService      object
OnlineSecurity       object
OnlineBackup         object
DeviceProtection     object
TechSupport          object
StreamingTV          object
StreamingMovies      object
Contract             object
PaperlessBilling     object
PaymentMethod        object
MonthlyCharges      float64
TotalCharges         object
Churn                object
dtype: object
```

Fig 4.3 The Numver of observations, attributes and Datatypes of the dataset



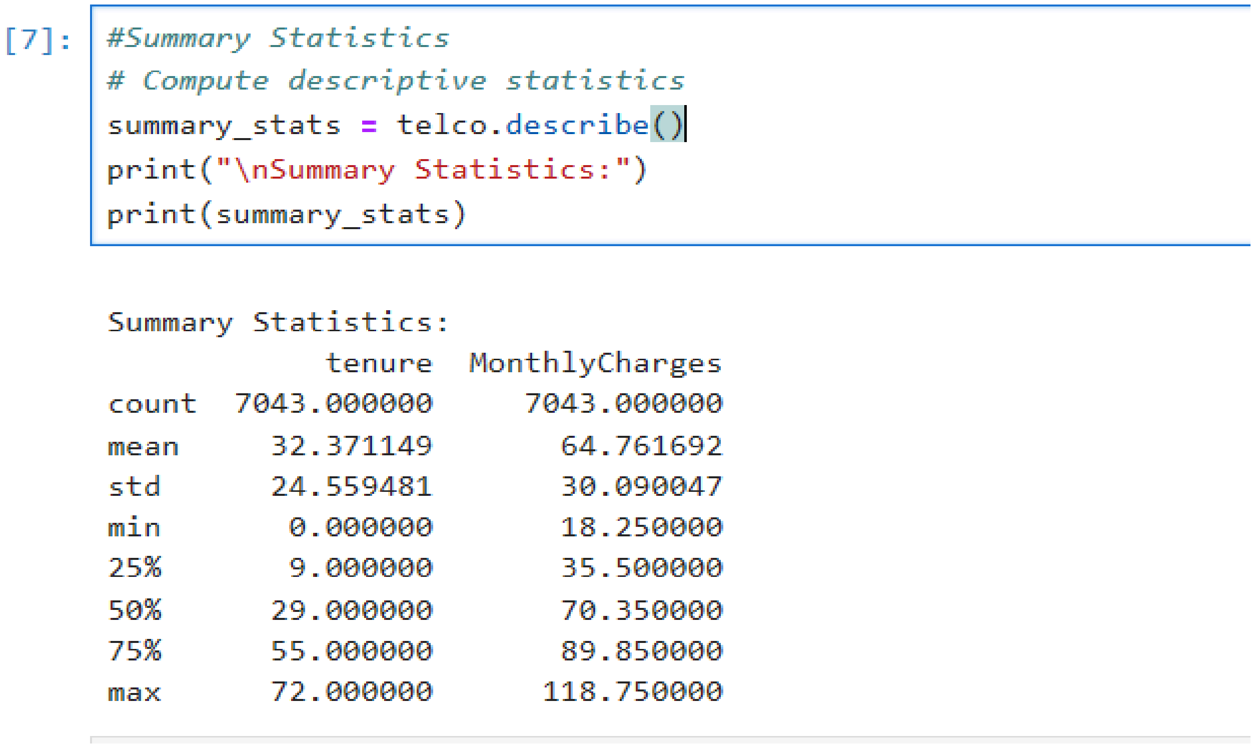
```
[8]: print(numerical_attributes)
     print(categorical_attributes)
     # Calculate the number of numerical attributes
     numerical_attributes = telco.select_dtypes(include=['int64', 'float64']).columns
     num_numerical_attributes = len(numerical_attributes)

     # Calculate the number of categorical attributes
     categorical_attributes = telco.select_dtypes(include=['object']).columns
     num_categorical_attributes = len(categorical_attributes)

     print("Number of Numerical Attributes:", num_numerical_attributes)
     print("Number of Categorical Attributes:", num_categorical_attributes)
```

```
Index(['tenure', 'MonthlyCharges'], dtype='object')
Index(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents',
       'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity',
       'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV',
       'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod',
       'TotalCharges', 'Churn'],
      dtype='object')
Number of Numerical Attributes: 2
Number of Categorical Attributes: 19
```

Fig 4.4 Python code for printing and calculating all the numerical and categorical variables

2. SUMMARY STATISTICS

- Summary statistics such as mean, median, standard deviation, etc., are computed to summarize numerical attributes. This is done using the `describe()` method on the DataFrame `telco`.



```
[7]: #Summary Statistics
     # Compute descriptive statistics
     summary_stats = telco.describe()
     print("\nSummary Statistics:")
     print(summary_stats)
```

```
Summary Statistics:
            tenure  MonthlyCharges
count  7043.000000     7043.000000
mean     32.371149       64.761692
std      24.559481       30.090047
min       0.000000       18.250000
25%       9.000000       35.500000
50%      29.000000       70.350000
75%      55.000000       89.850000
max      72.000000      118.750000
```

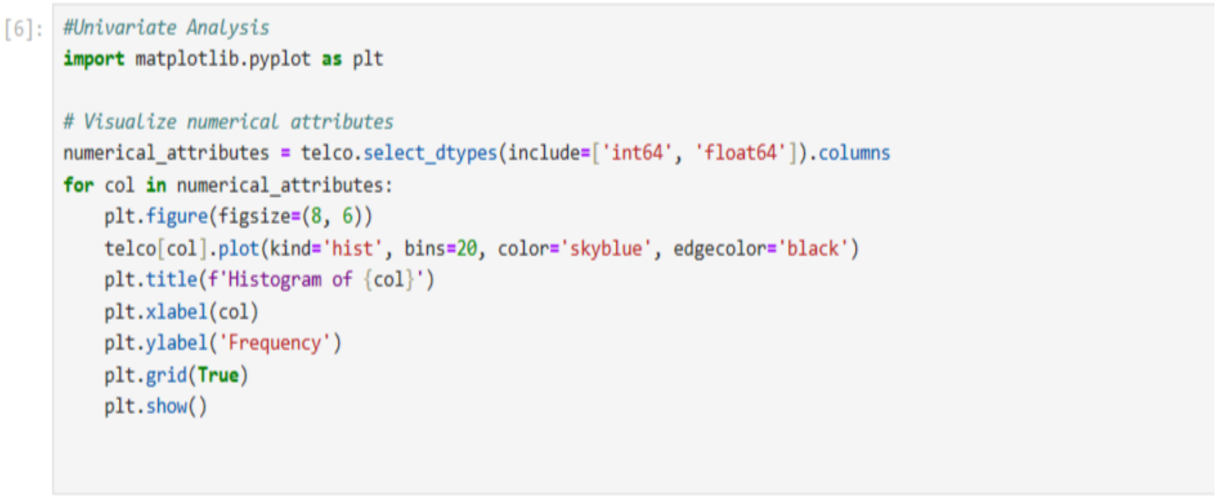
Fig 4.5 Python code for printing the Summary statistics of teh Dataset

3. UNIVARIATE ANALYSIS

- Univariate analysis focuses on analyzing individual attributes independently. In the provided code, univariate analysis is conducted for both numerical and categorical attributes.

- For numerical attributes, histograms are plotted to visualize their distributions.

- For categorical attributes, bar plots are plotted to visualize the frequency distribution of each category.



```
[6]: #Univariate Analysis
     import matplotlib.pyplot as plt

     # Visualize numerical attributes
     numerical_attributes = telco.select_dtypes(include=['int64', 'float64']).columns
     for col in numerical_attributes:
         plt.figure(figsize=(8, 6))
         telco[col].plot(kind='hist', bins=20, color='skyblue', edgecolor='black')
         plt.title(f'Histogram of {col}')
         plt.xlabel(col)
         plt.ylabel('Frequency')
         plt.grid(True)
         plt.show()
```

Fig 4.6 Python code for the univariate analysis, for making Histogram for all the numerical variables

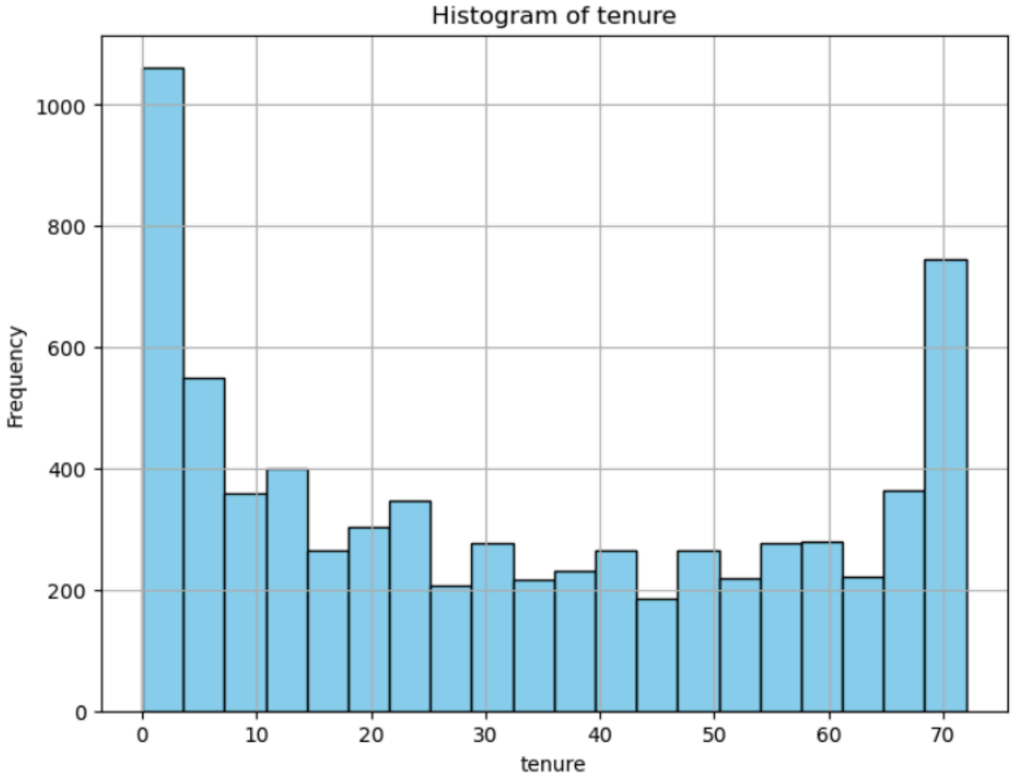


Fig 4.7 Histogram of tenure (Univariate analysis)

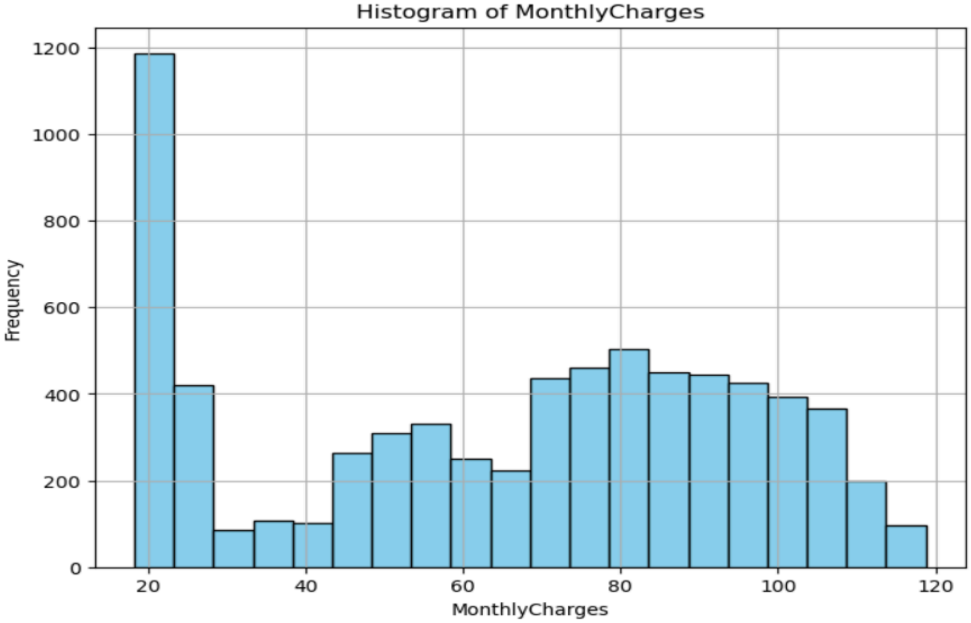
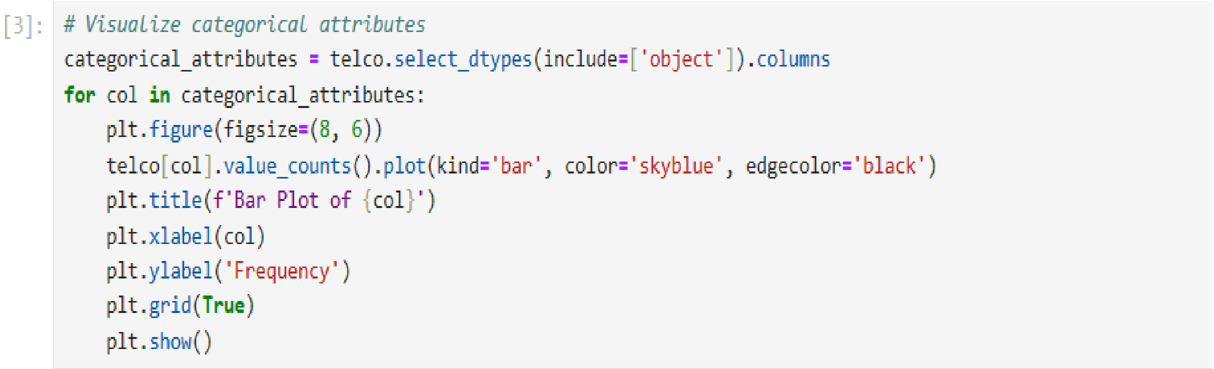


Fig 4.8: Histogram of Monthly Charges (Univariate Analysis)



```
# Visualize categorical attributes
categorical_attributes = telco.select_dtypes(include=['object']).columns
for col in categorical_attributes:
    plt.figure(figsize=(8, 6))
    telco[col].value_counts().plot(kind='bar', color='skyblue', edgecolor='black')
    plt.title(f'Bar Plot of {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
    plt.grid(True)
    plt.show()
```

Fig 4.9 Python code for the univariate analysis, for making Bar Graph for all the categorical variables

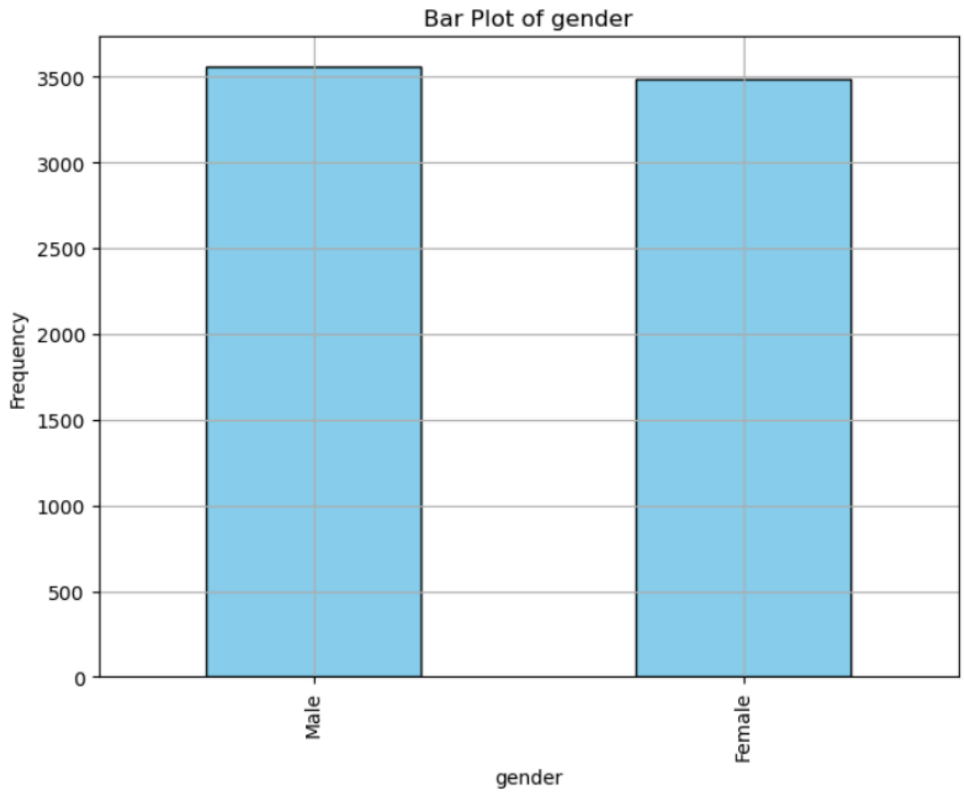


Fig 4.10: Bar Plot of Gender (Visualizing Categorical Attributes)

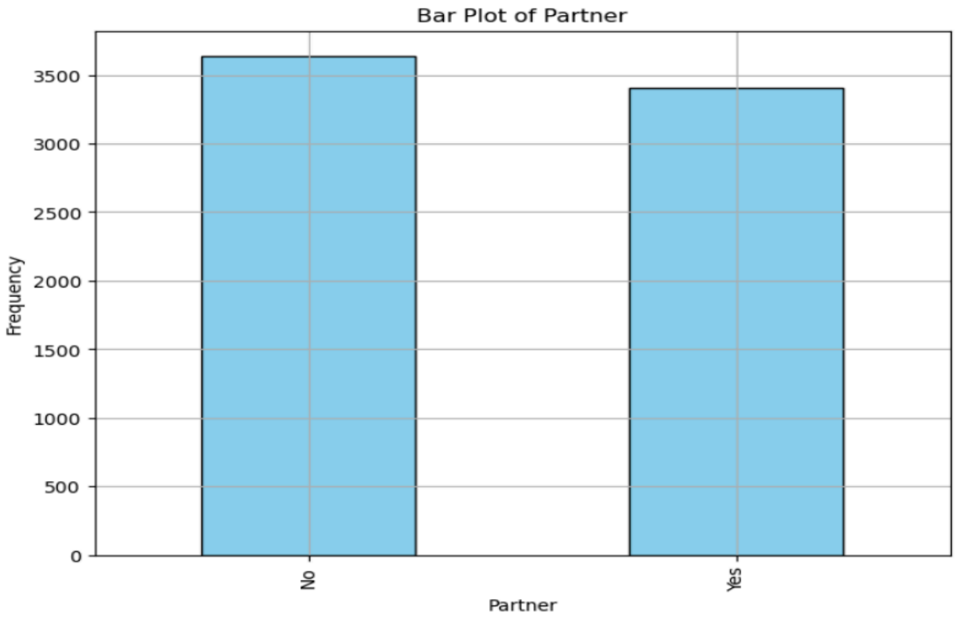


Fig 4.11: Bar Plot of Partner (Visualizing Categorical Attributes)

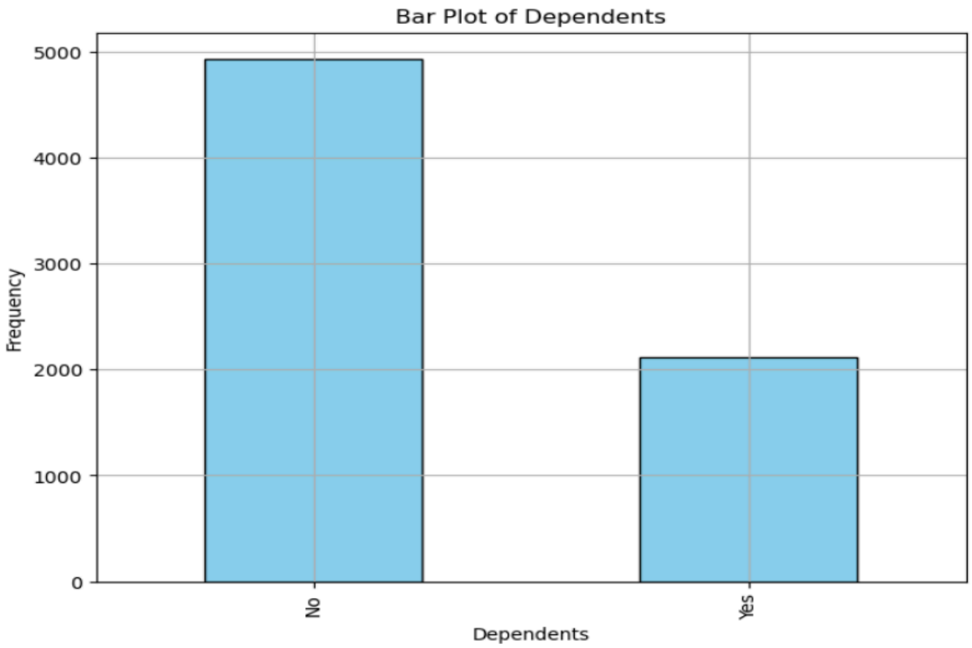


Fig 4.12   Bar Plot of Dependents (Visualizing Categorical Attributes)

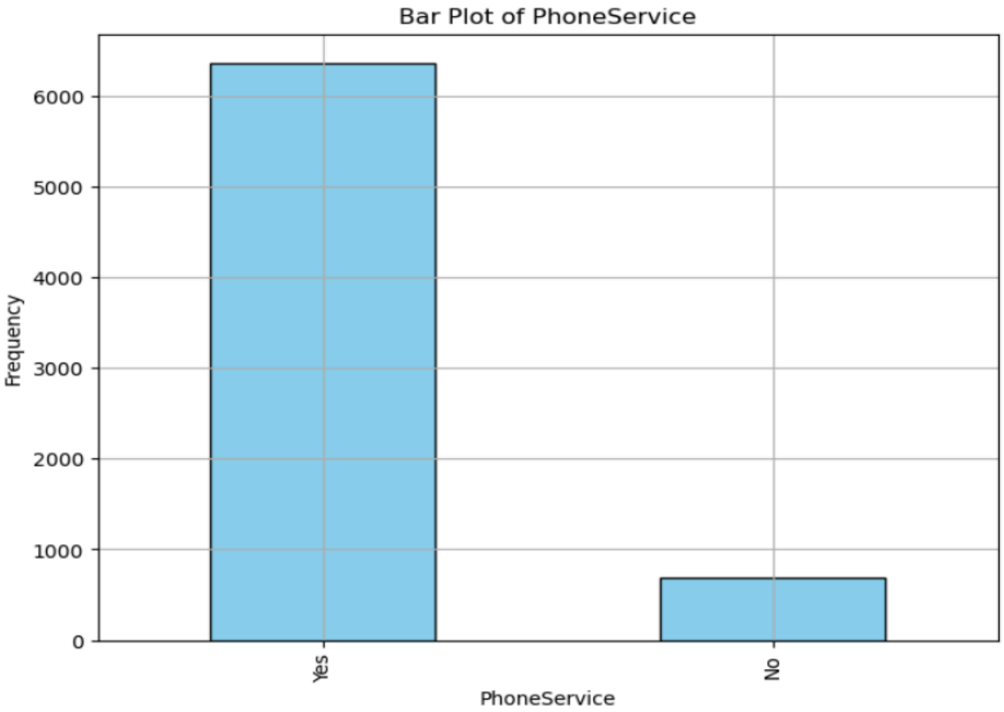


Fig 4.13: Bar Plot of Phone Service (Visualizing Categorical Attributes)

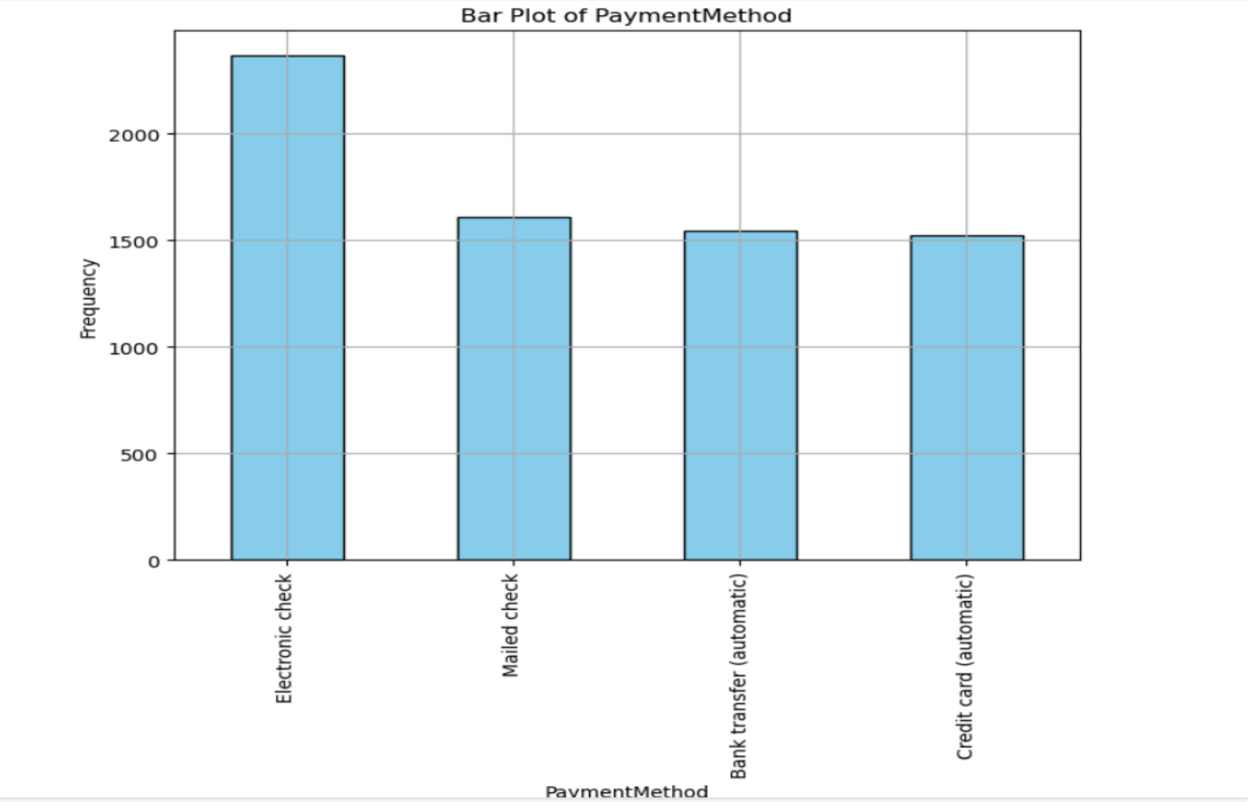
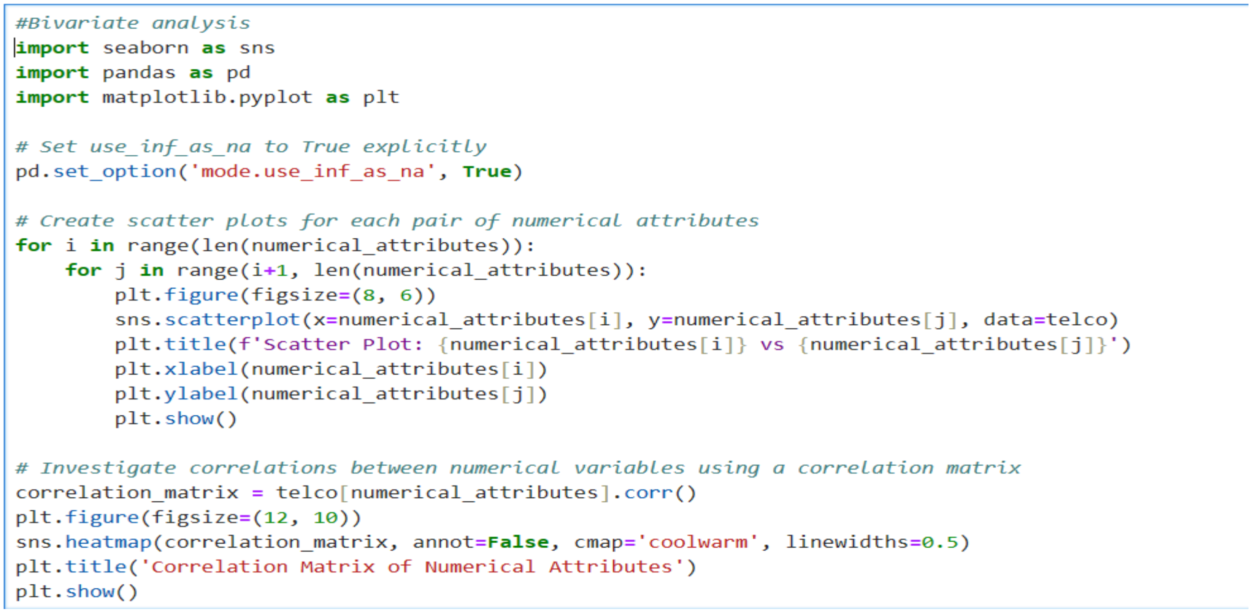


Fig 4.14: *Bar Plot of Payment Method (Visualizing Categorical Attributes)*

4. BIVARIATE ANALYSIS

- Bivariate analysis explores relationships between pairs of attributes.

- In the code, pair plots are used to visualize relationships between pairs of numerical attributes and a correlation matrix heatmap is plotted to investigate correlations between numerical variables.



```
#Bivariate analysis
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt

# Set use_inf_as_na to True explicitly
pd.set_option('mode.use_inf_as_na', True)

# Create scatter plots for each pair of numerical attributes
for i in range(len(numerical_attributes)):
    for j in range(i+1, len(numerical_attributes)):
        plt.figure(figsize=(8, 6))
        sns.scatterplot(x=numerical_attributes[i], y=numerical_attributes[j], data=telco)
        plt.title(f'Scatter Plot: {numerical_attributes[i]} vs {numerical_attributes[j]}')
        plt.xlabel(numerical_attributes[i])
        plt.ylabel(numerical_attributes[j])
        plt.show()

# Investigate correlations between numerical variables using a correlation matrix
correlation_matrix = telco[numerical_attributes].corr()
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=False, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Matrix of Numerical Attributes')
plt.show()
```

*Fig 4.15***:** *Python code for the bivariate analysis, that conatins scatter ploats and Heat Map*

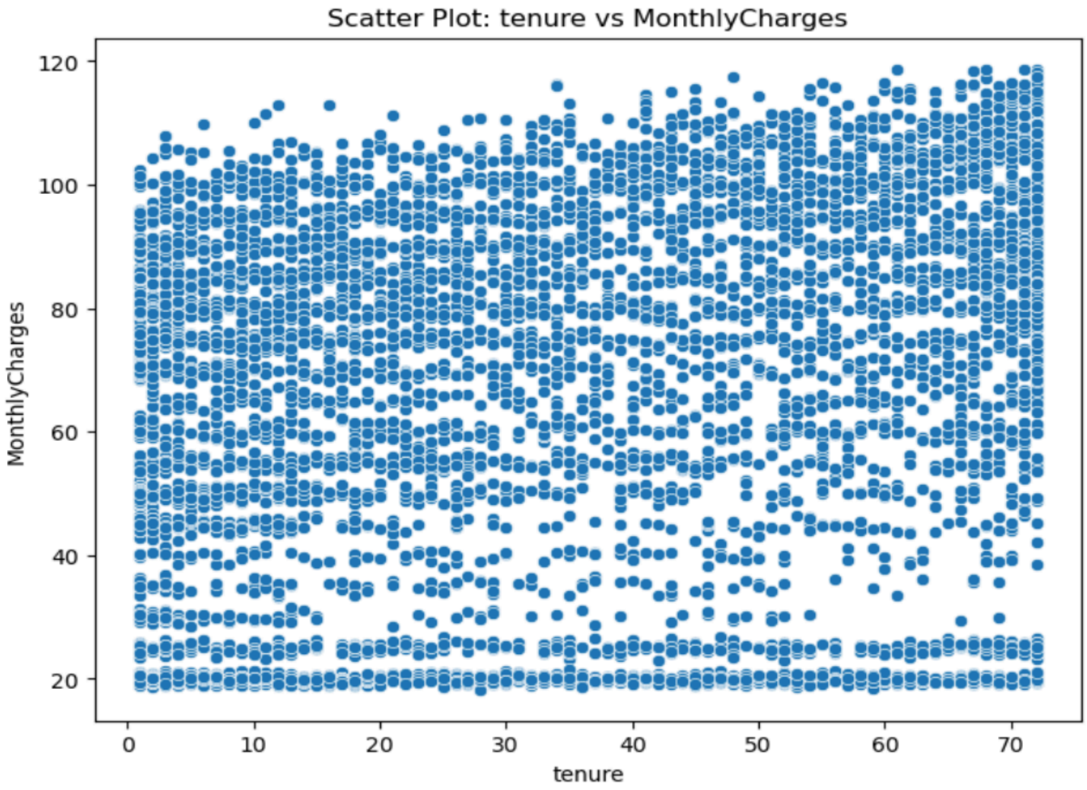


Fig 4.16 Scatter Plot for Tenure vs Monthly Charges

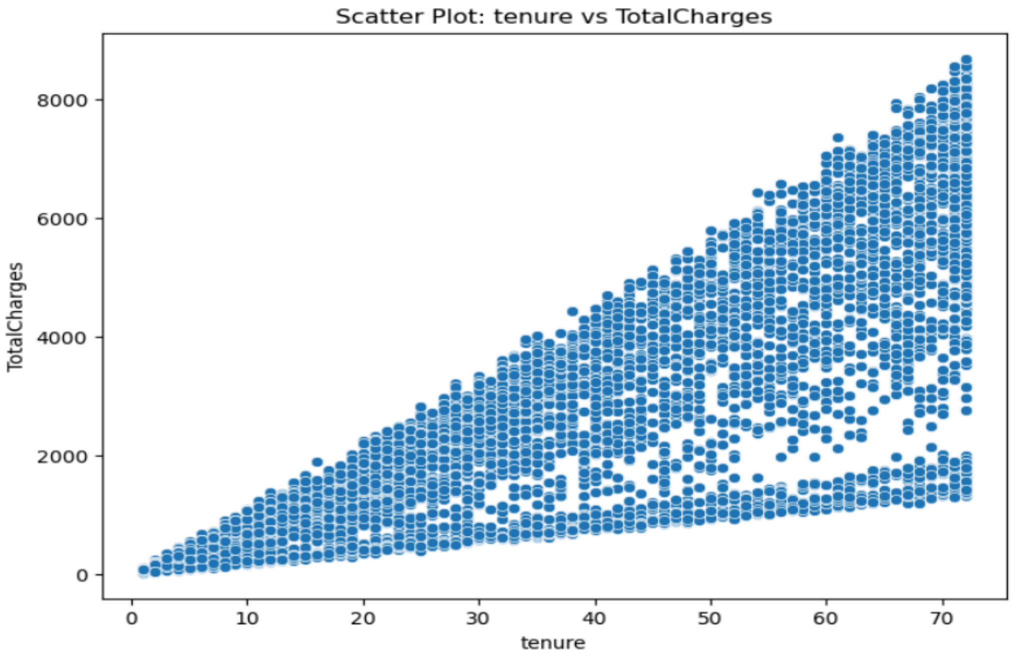


Fig 4.17: Scatter plot of Tenure vs Total Charges

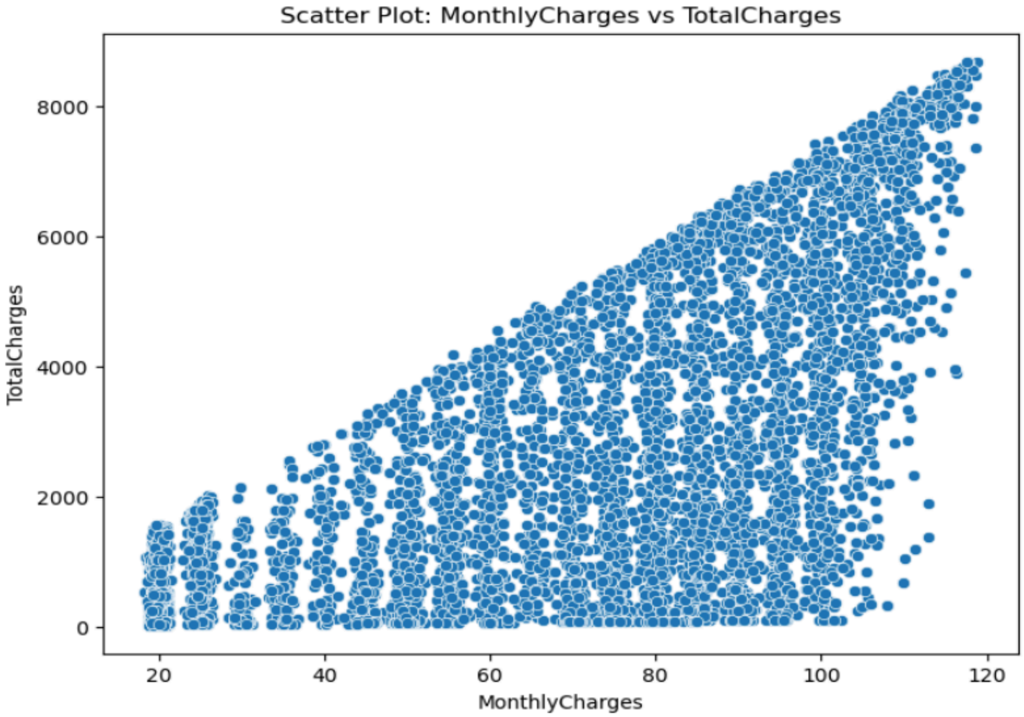


Fig 4.18: Scatter plot of Monthly Charges vs Total Charges

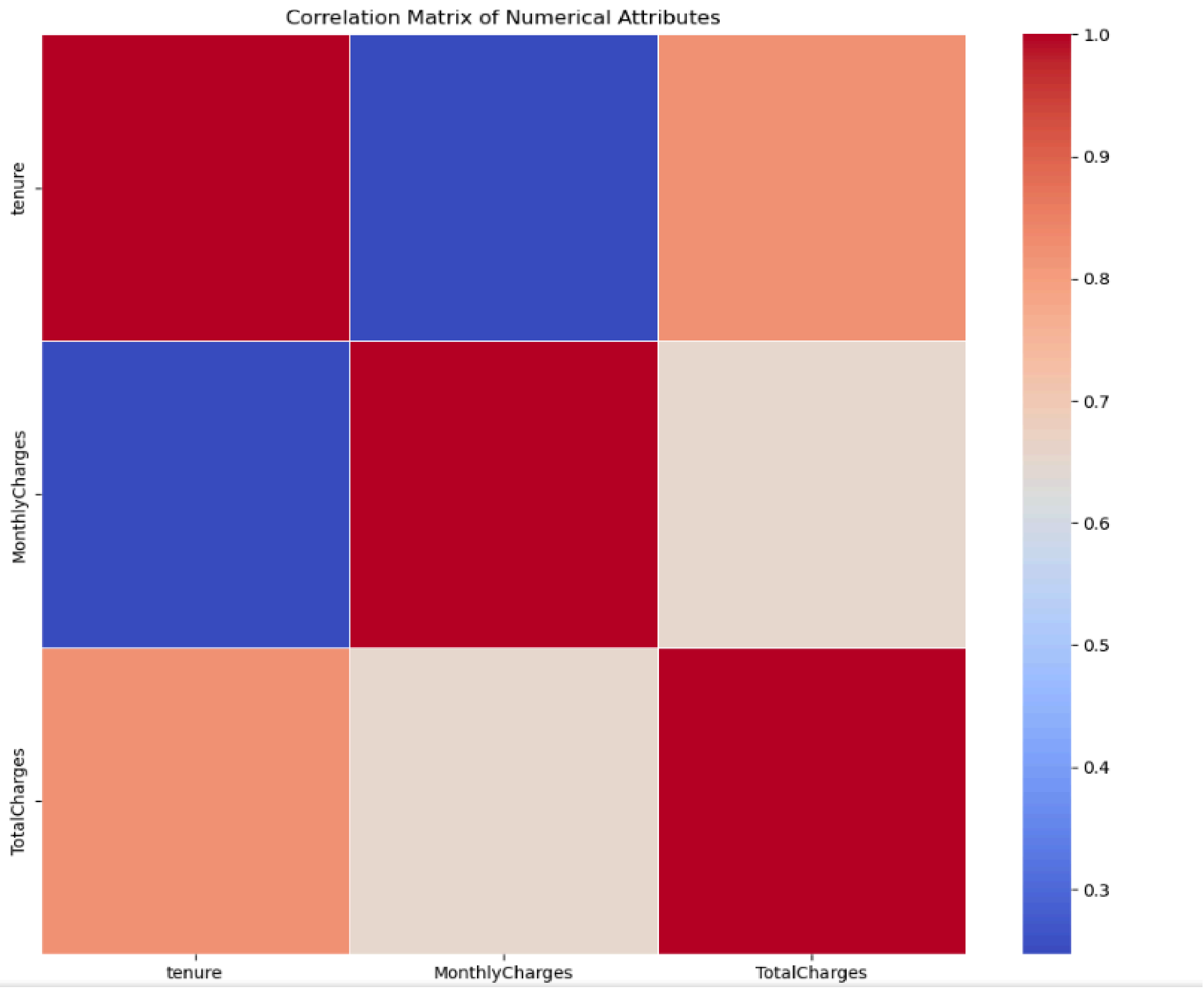


Fig 4.17: Correlation attributes of numerical variables(Monthly Charges and Total Charges)

Scatterplots: Along the diagonal, each scatterplot represents the distribution of a numerical variable such as 'SeniorCitizen', 'tenure', 'MonthlyCharges', and 'TotalCharges'. Each point in these plots represents an observation, showing the distribution of values for that variable.

Pairwise Relationships: Off the diagonal, the scatterplots display the pairwise relationships between different pairs of numerical variables. For example:

- Scatterplots in the upper triangle might show the relationship between 'SeniorCitizen' and other numerical variables like 'tenure', 'MonthlyCharges', and 'TotalCharges'.

- Scatterplots in the lower triangle may reveal relationships between 'tenure', 'MonthlyCharges', and 'TotalCharges'.

Trends and Patterns: By examining these scatterplots, we can identify trends, patterns, and potential correlations between variables. For instance:

- Positive correlation: The scatterplots show an upward trend from left to right, which suggests a positive correlation between the variables.

Outliers: There were no outliers in the scatter plot

Color Intensity: Each cell in the heatmap represents the correlation coefficient between two numerical attributes. The color intensity reflects the strength and direction of the correlation.

- Darker shades (e.g., dark red) indicate stronger positive correlations.
- Lighter shades (e.g., light blue) indicate stronger negative correlations.

Correlation Coefficients: The values inside the cells indicate the correlation coefficients. These values range from -1 to 1.

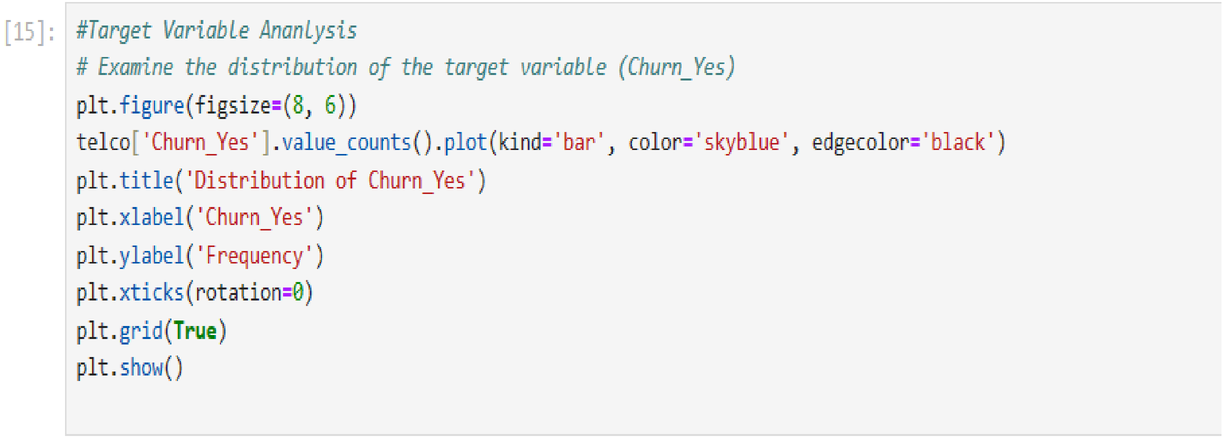
- Positive values (close to 1) indicate a positive correlation, meaning that as one variable increases, the other tends to increase as well.

Interpretation:

- The correlation coefficient between "tenure" and "MonthlyCharges" is close to 1, it suggests a strong positive correlation, indicating that customers with longer tenure tend to have higher monthly charges.

5. TARGET VARIABLE ANALYSIS

- This step focuses specifically on the target variable (in this case, `Churn_Yes`) to understand its distribution.
- In the provided code, a bar plot is created to visualize the distribution of the target variable `Churn_Yes`.



```
[15]: #Target Variable Ananlysis
      # Examine the distribution of the target variable (Churn_Yes)
      plt.figure(figsize=(8, 6))
      telco['Churn_Yes'].value_counts().plot(kind='bar', color='skyblue', edgecolor='black')
      plt.title('Distribution of Churn_Yes')
      plt.xlabel('Churn_Yes')
      plt.ylabel('Frequency')
      plt.xticks(rotation=0)
      plt.grid(True)
      plt.show()
```

Fig 4.17 Python Code for Target Variable 'Churn Yes Analysis'

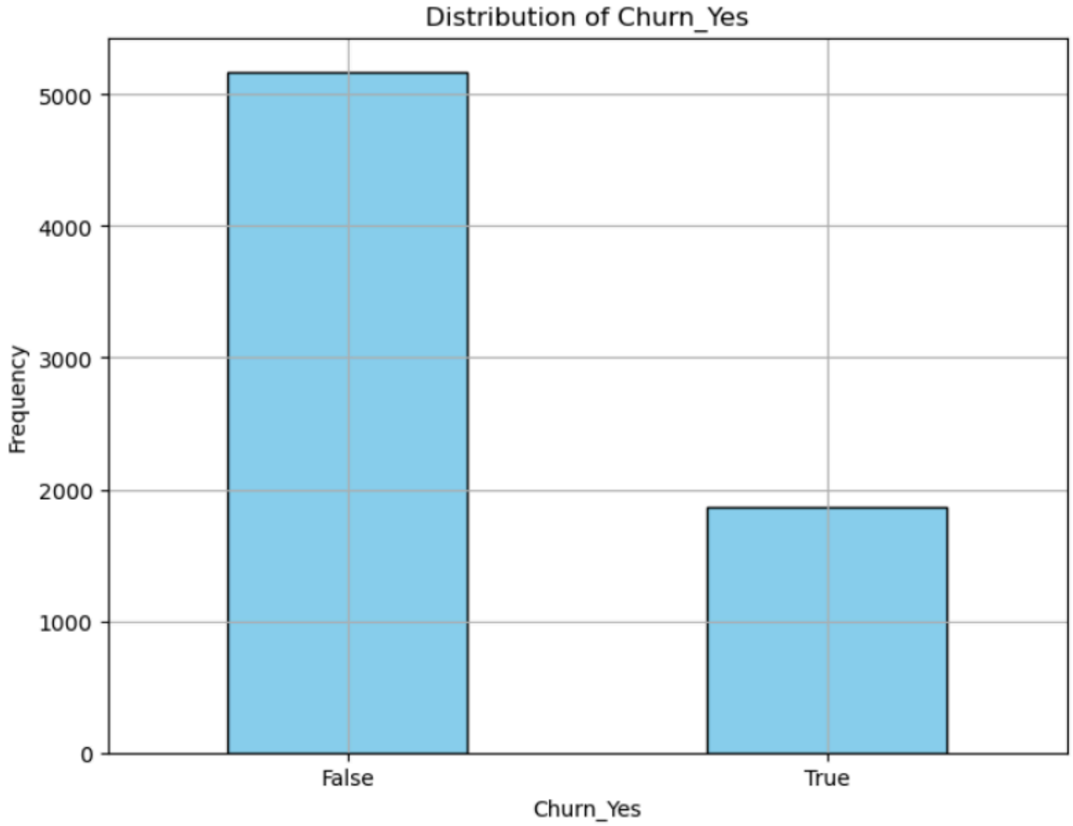


Fig 4.18: Distribution of Target Variable

The target variable `Churn_Yes` likely represents whether a customer has churned or not, with "Yes" indicating that the customer has churned and "No" indicating that the customer has not churned.

Interpreting the bar plot of the distribution of `Churn_Yes` involves understanding the frequency or count of churned and non-churned customers in the dataset. Here's how we can interpret it:

- Bar Plot: The bar plot displays two bars representing the frequency of "Churn_Yes" categories: "Yes" and "No".

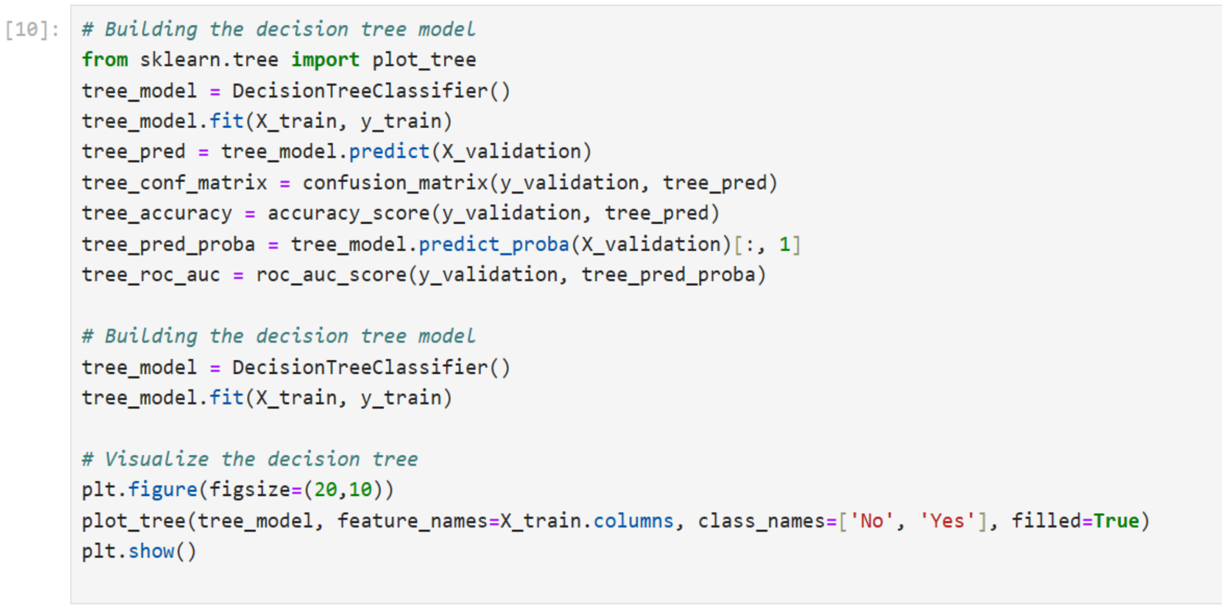
- Interpretation:
  - If the bar representing "Yes" (churned customers) is taller, it indicates a higher number of churned customers in the dataset.
  - Conversely, if the bar representing "No" (non-churned customers) is taller, it suggests a higher number of customers who have not churned.

- Implications:
  - A higher frequency of churned customers may indicate potential issues in customer retention or service satisfaction.
  - Understanding the distribution helps in assessing the effectiveness of churn prediction models. For instance, an imbalanced dataset may require techniques like resampling or adjusting class weights to train a predictive model effectively.

6. BUILDING MODELS

-The decision tree model and the random forest model are built using `DecisionTreeClassifier()` and `RandomForestClassifier()`, respectively. These models are trained on the training data (`X_train`, `y_train`).

- Additionally, the logistic regression model is built but has not been executed successfully due to the presence of NaN values in the target variable.



```
[10]: # Building the decision tree model
      from sklearn.tree import plot_tree
      tree_model = DecisionTreeClassifier()
      tree_model.fit(X_train, y_train)
      tree_pred = tree_model.predict(X_validation)
      tree_conf_matrix = confusion_matrix(y_validation, tree_pred)
      tree_accuracy = accuracy_score(y_validation, tree_pred)
      tree_pred_proba = tree_model.predict_proba(X_validation)[:, 1]
      tree_roc_auc = roc_auc_score(y_validation, tree_pred_proba)

      # Building the decision tree model
      tree_model = DecisionTreeClassifier()
      tree_model.fit(X_train, y_train)

      # Visualize the decision tree
      plt.figure(figsize=(20,10))
      plot_tree(tree_model, feature_names=X_train.columns, class_names=['No', 'Yes'], filled=True)
      plt.show()
```

Fig 4.19 Python code that builds decision tree using 'DecisionTreeClassifier'

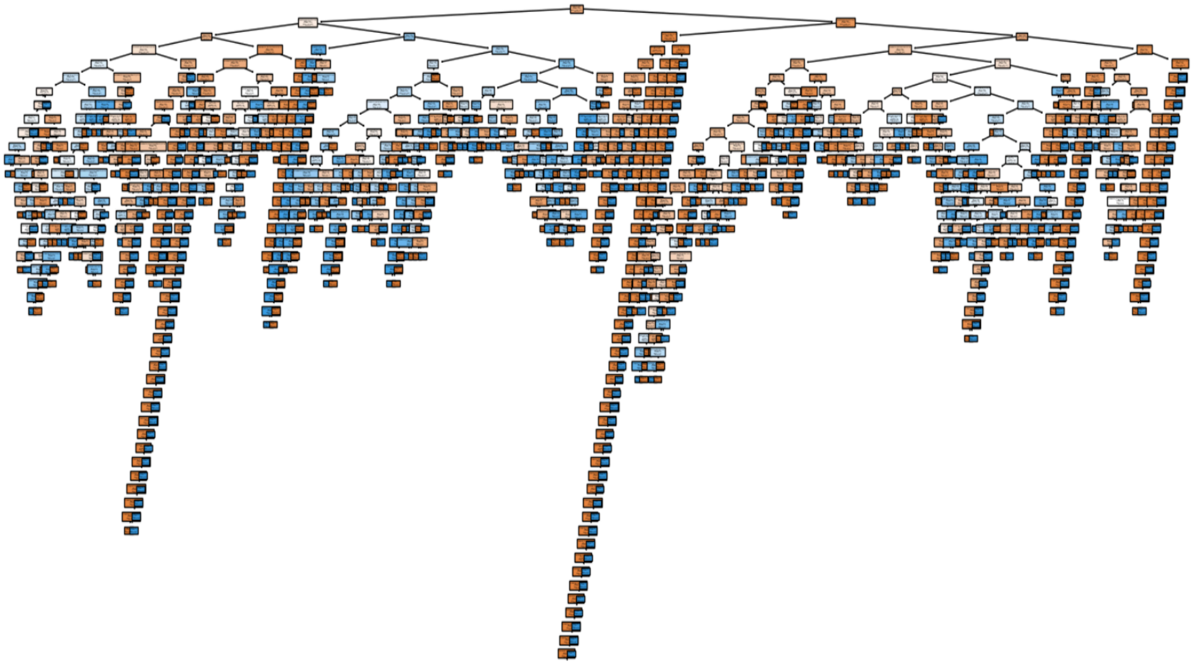
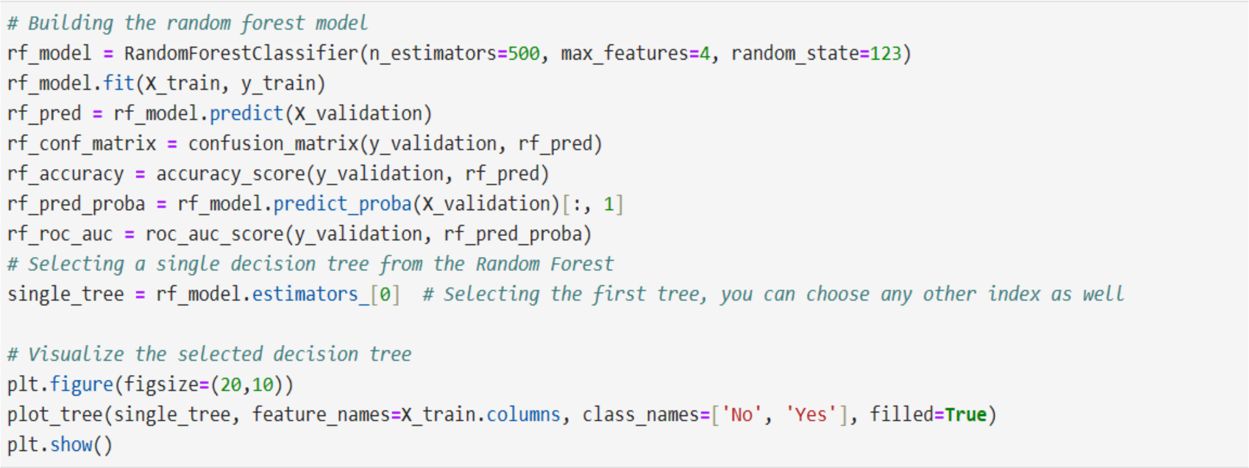


Fig 4.20 Decision tree model

A decision tree model analyzes the dataset to determine the factors that most significantly influence the likelihood of churn (the target variable). The decision tree would make splits based on various features such as gender, senior citizen status, partner status, internet service type, monthly charges, total charges, and other attributes. Each split in the decision tree represents a decision point based on a feature's value, with the goal of maximizing the homogeneity of churn status within each resulting subset. By traversing the decision tree, we can understand the sequence of conditions that lead to a prediction of churn (Yes or No) for each customer. For instance, the decision tree reveals that customers with fiber optic internet service, higher monthly charges, and shorter tenure are more likely to churn, while those with DSL internet service, lower monthly charges, and longer tenure are less likely to churn. This information helps in identifying potential churn risk factors and devising strategies to mitigate customer attrition.



```
# Building the random forest model
rf_model = RandomForestClassifier(n_estimators=500, max_features=4, random_state=123)
rf_model.fit(X_train, y_train)
rf_pred = rf_model.predict(X_validation)
rf_conf_matrix = confusion_matrix(y_validation, rf_pred)
rf_accuracy = accuracy_score(y_validation, rf_pred)
rf_pred_proba = rf_model.predict_proba(X_validation)[:, 1]
rf_roc_auc = roc_auc_score(y_validation, rf_pred_proba)
# Selecting a single decision tree from the Random Forest
single_tree = rf_model.estimators_[0]  # Selecting the first tree, you can choose any other index as well

# Visualize the selected decision tree
plt.figure(figsize=(20,10))
plot_tree(single_tree, feature_names=X_train.columns, class_names=['No', 'Yes'], filled=True)
plt.show()
```

Fig 4.21 Python code to build a random forest model by selecting a single decision tree

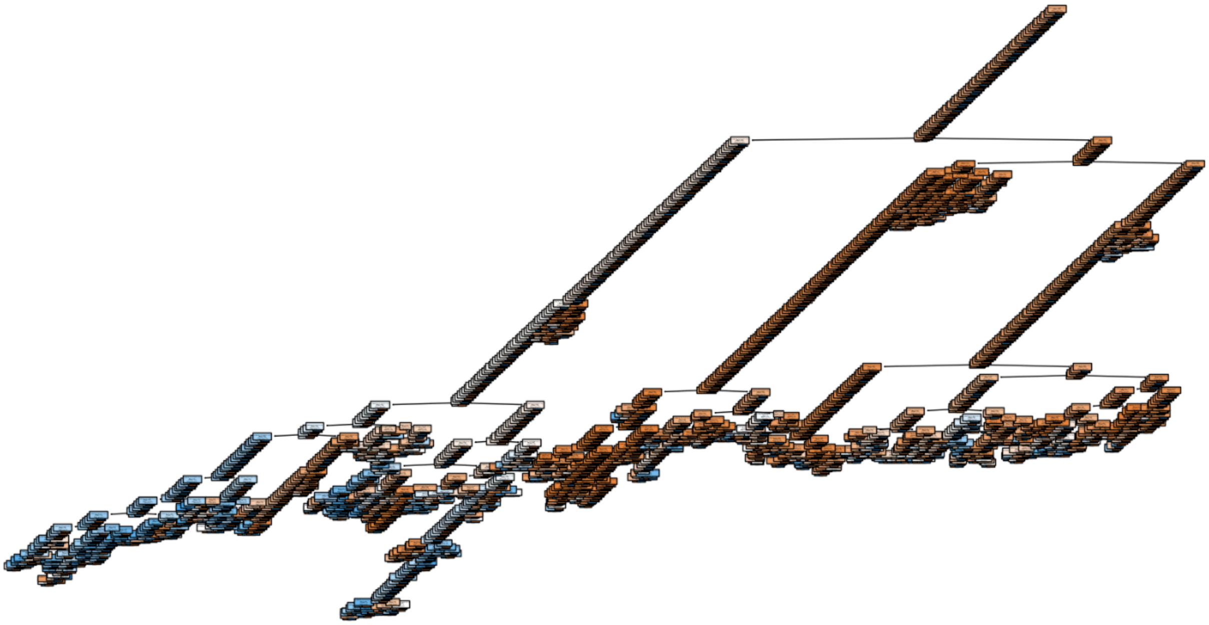
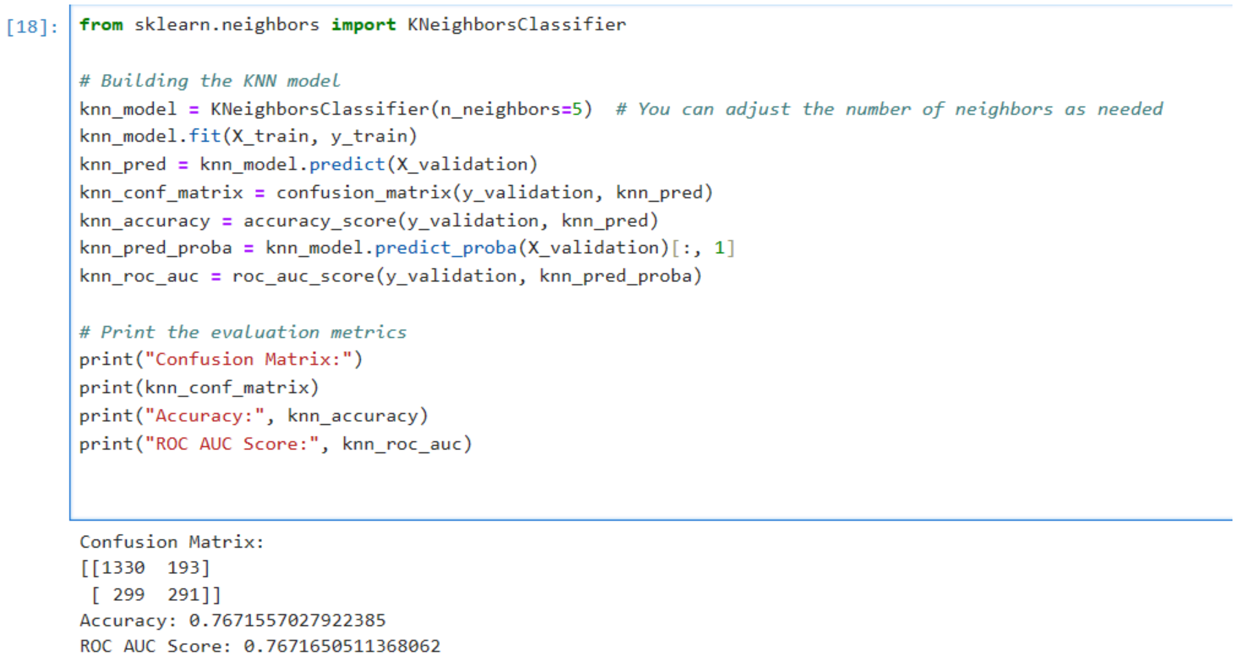


Fig 4.22 Random Forest Model

A random forest model would provide predictions on the likelihood of churn based on an ensemble of decision trees trained on random subsets of the data. The model would consider multiple factors simultaneously and generate predictions by aggregating the results from individual decision trees. Similar to a single decision tree, the random forest model analyzes various features such as gender, senior citizen status, partner status, internet service type, monthly charges, total charges, and others to make predictions about churn. However, instead of relying on the decision of a single tree, the random forest model incorporates the decisions of multiple trees to provide a more robust and accurate prediction. It would consider the collective insights from all trees in the forest to determine the likelihood of churn for each customer, taking into account the interactions and relationships between different features. This ensemble approach helps to reduce overfitting and improves the generalization performance of the model, making it more reliable for predicting churn in unseen



```
[18]: from sklearn.neighbors import KNeighborsClassifier

      # Building the KNN model
      knn_model = KNeighborsClassifier(n_neighbors=5)  # You can adjust the number of neighbors as needed
      knn_model.fit(X_train, y_train)
      knn_pred = knn_model.predict(X_validation)
      knn_conf_matrix = confusion_matrix(y_validation, knn_pred)
      knn_accuracy = accuracy_score(y_validation, knn_pred)
      knn_pred_proba = knn_model.predict_proba(X_validation)[:, 1]
      knn_roc_auc = roc_auc_score(y_validation, knn_pred_proba)

      # Print the evaluation metrics
      print("Confusion Matrix:")
      print(knn_conf_matrix)
      print("Accuracy:", knn_accuracy)
      print("ROC AUC Score:", knn_roc_auc)
```

```
Confusion Matrix:
[[1330  193]
 [ 299  291]]
Accuracy: 0.7671557027922385
ROC AUC Score: 0.7671650511368062
```

Fig 4.23 Python code to build the KNN model, and to have the evaluation metrix

The code introduces the K-Nearest Neighbors (KNN) algorithm to analyze the Telco customer churn dataset. It begins by building a KNN model with five neighbors and trains it on the training data. The model then makes predictions on the validation set, allowing for the evaluation of its performance. The evaluation metrics include a confusion matrix, accuracy, and ROC AUC score. The confusion matrix reveals the model's ability to correctly classify instances of churn and non-churn customers, highlighting both correct and incorrect predictions. The accuracy, measuring the proportion of correctly classified instances, indicates a moderate level of predictive accuracy. Similarly, the ROC AUC score, reflecting the model's ability to discriminate between classes, suggests a reasonable performance level. These insights help in comparing the effectiveness of the KNN model with other algorithms and guide further analysis, such as parameter tuning and feature importance exploration, to enhance predictive capabilities and inform decision-making regarding customer churn prediction in the Telco dataset.

Interpretation of Results:

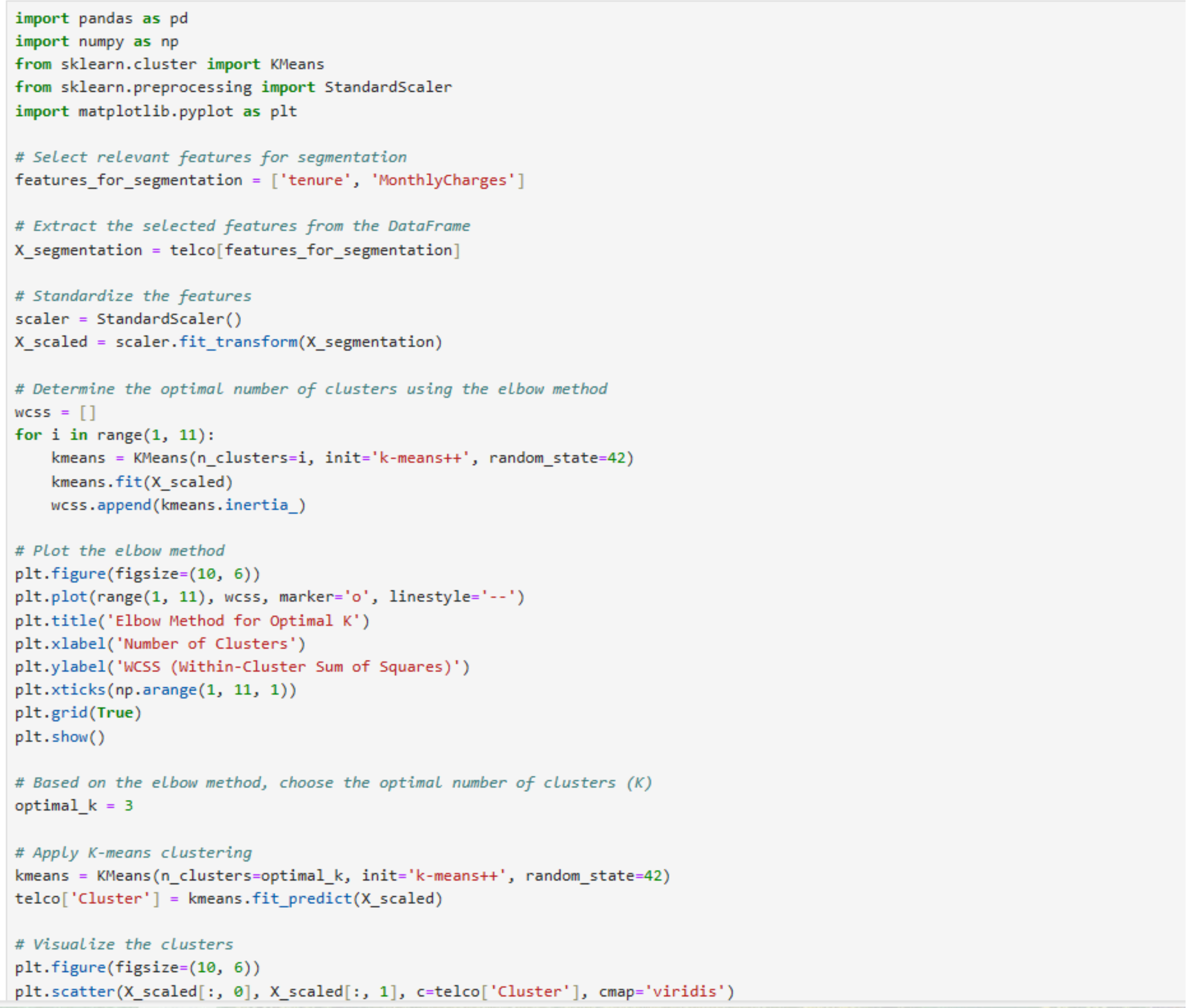
- The confusion matrix shows that the KNN model correctly predicted 1330 instances of non-churn customers and 291 instances of churn customers. However, it misclassified 193 instances of non-churn customers as churn customers and 299 instances of churn customers as non-churn customers.

- The accuracy of approximately 76.72% indicates the overall proportion of correct predictions made by the model.

- The ROC AUC score of approximately 0.767 suggests that the model has moderate discriminative power in distinguishing between churn and non-churn customers.

7. CUSTOMER SEGMENTATION

Customer segmentation involves categorizing telecom customers based on various factors like their demographics, usage patterns, and service preferences. By doing this, we can identify distinct groups of customers with similar characteristics and behaviors. This segmentation allows us to tailor our marketing strategies, service offerings, and customer support to better meet the specific needs and preferences of each customer segment. Ultimately, this approach helps us enhance customer satisfaction, increase loyalty, and improve overall business performance.



```python
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt

# Select relevant features for segmentation
features_for_segmentation = ['tenure', 'MonthlyCharges']

# Extract the selected features from the DataFrame
X_segmentation = telco[features_for_segmentation]

# Standardize the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X_segmentation)

# Determine the optimal number of clusters using the elbow method
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
    kmeans.fit(X_scaled)
    wcss.append(kmeans.inertia_)

# Plot the elbow method
plt.figure(figsize=(10, 6))
plt.plot(range(1, 11), wcss, marker='o', linestyle='--')
plt.title('Elbow Method for Optimal K')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS (Within-Cluster Sum of Squares)')
plt.xticks(np.arange(1, 11, 1))
plt.grid(True)
plt.show()

# Based on the elbow method, choose the optimal number of clusters (K)
optimal_k = 3

# Apply K-means clustering
kmeans = KMeans(n_clusters=optimal_k, init='k-means++', random_state=42)
telco['Cluster'] = kmeans.fit_predict(X_scaled)

# Visualize the clusters
plt.figure(figsize=(10, 6))
plt.scatter(X_scaled[:, 0], X_scaled[:, 1], c=telco['Cluster'], cmap='viridis')
```

Fig 4.23 Python code for customer segmentation by using elbow method

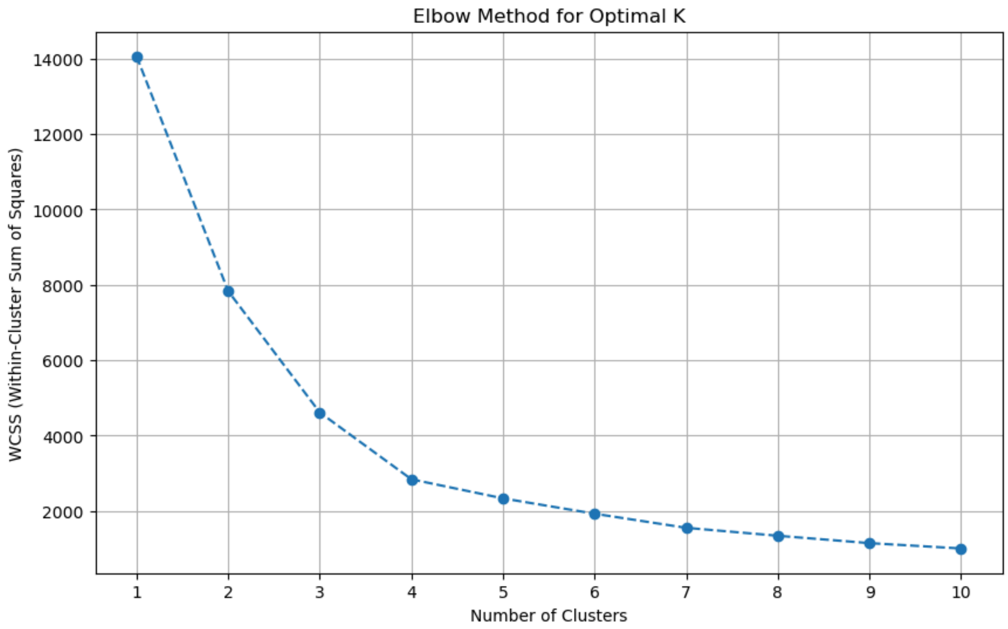
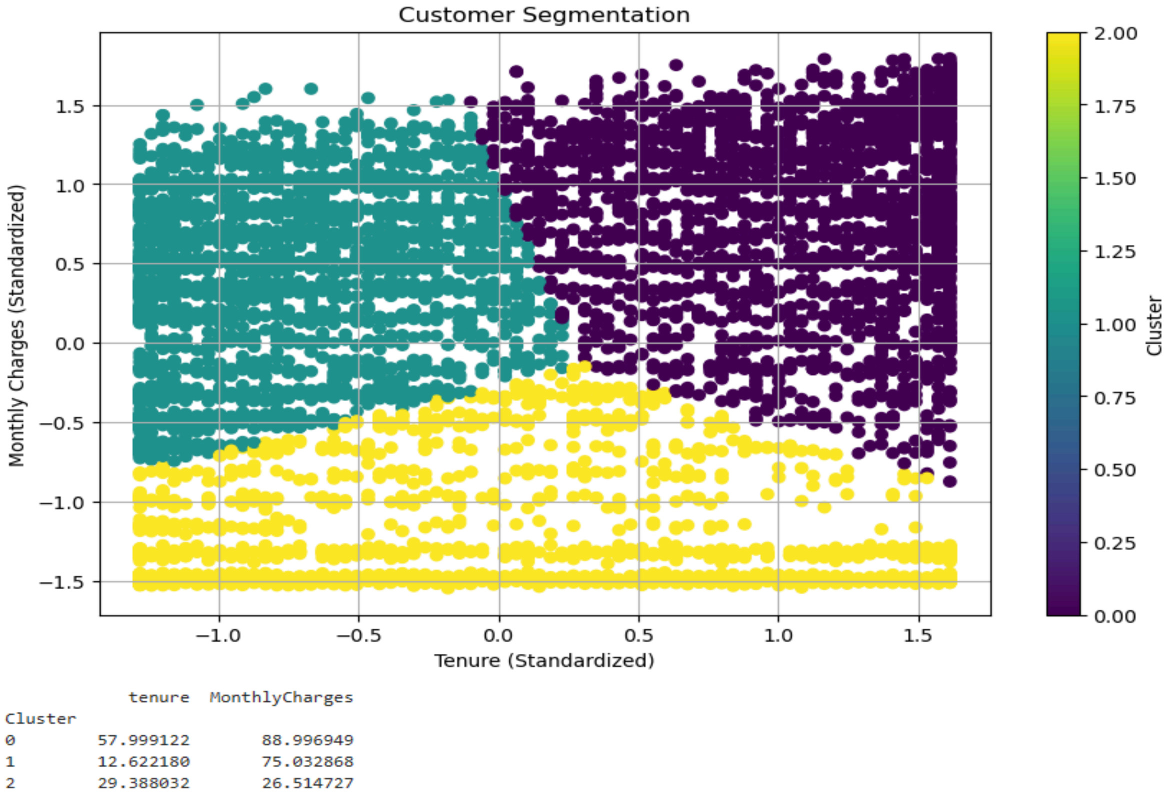


Fig 4.24 Elbow Method for Optimal K

There is no clear optimal number of clusters, it indicates that the data may not naturally segregate into distinct groups based on the available features. This limitation makes it challenging to identify clear segments of customers with distinct churn behaviors or characteristics.

Continuous decrease in WCSS without a clear elbow point suggests that adding more clusters may lead to overfitting. In the context of customer churn, overfitting can result in the identification of spurious or noise-driven segments that do not generalize well to new data, potentially leading to poor predictive performance.



```
          tenure  MonthlyCharges
Cluster
0      57.999122       88.996949
1      12.622180       75.032868
2      29.388032       26.514727
```

Fig 4.25 Customer Segmentation

The tenure and monthly charges for each customer segment separately, you can observe whether there are distinct patterns or clusters within each segment. Different colors or markers can represent different segments, allowing you to visually identify how segments differ in their distribution of these variables.

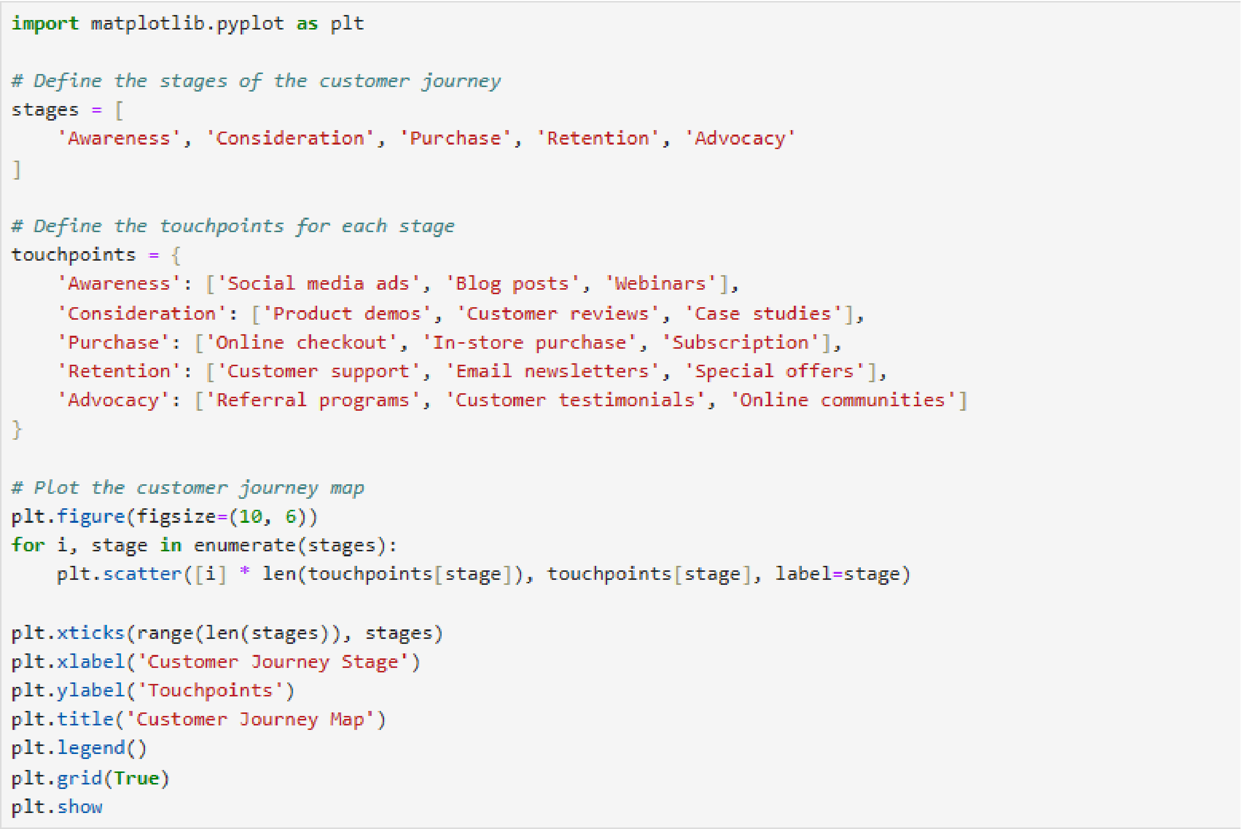
Analyzing the scatter plot can help identify segments with varying levels of tenure and monthly charges. For example, you may observe segments with long tenure but low monthly charges, indicating loyal but low-spending customers, or segments with short tenure and high monthly charges, suggesting new but high-paying customers.

Understanding the distribution of tenure and monthly charges within each segment can provide insights into customer preferences and behavior. For instance, segments with higher monthly charges may exhibit longer tenure, indicating that customers who spend more tend to stay longer with the company.

Based on the scatter plot, you can identify segments that represent valuable customer groups for targeted marketing or retention strategies. Segments with high monthly charges and long tenure, for example, may be prime targets for loyalty programs or personalized offers to incentivize continued engagement.

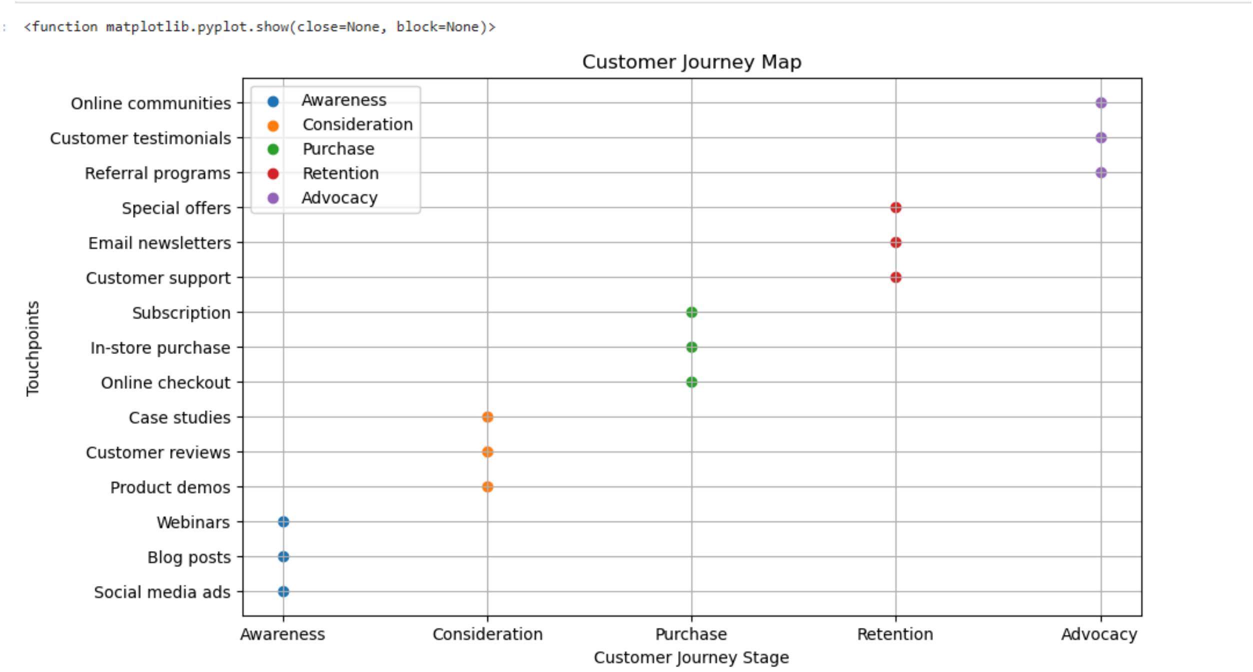
7. CUSTOMER JOURNEY MAP

Customer journey map visually represents the various touchpoints and interactions that telecom customers experience throughout their relationship with the company. It outlines the steps customers take from initial contact to becoming loyal advocates or churning. By mapping out this journey, we can better understand customer needs, pain points, and opportunities for improvement at each stage. This insight helps us design targeted strategies to enhance the overall customer experience and increase satisfaction, retention, and loyalty.



```python
import matplotlib.pyplot as plt

# Define the stages of the customer journey
stages = [
    'Awareness', 'Consideration', 'Purchase', 'Retention', 'Advocacy'
]

# Define the touchpoints for each stage
touchpoints = {
    'Awareness': ['Social media ads', 'Blog posts', 'Webinars'],
    'Consideration': ['Product demos', 'Customer reviews', 'Case studies'],
    'Purchase': ['Online checkout', 'In-store purchase', 'Subscription'],
    'Retention': ['Customer support', 'Email newsletters', 'Special offers'],
    'Advocacy': ['Referral programs', 'Customer testimonials', 'Online communities']
}

# Plot the customer journey map
plt.figure(figsize=(10, 6))
for i, stage in enumerate(stages):
    plt.scatter([i] * len(touchpoints[stage]), touchpoints[stage], label=stage)

plt.xticks(range(len(stages)), stages)
plt.xlabel('Customer Journey Stage')
plt.ylabel('Touchpoints')
plt.title('Customer Journey Map')
plt.legend()
plt.grid(True)
plt.show
```

Fig 4.26 Python code for customer journey map using their touchpoints



```
: <function matplotlib.pyplot.show(close=None, block=None)>
```

Fig 4.27 Customer Journey Map

At this stage, customers become aware of the company's services through social media ads, blog posts, and webinars. The company's marketing efforts focus on reaching potential customers and generating interest in its offerings.

Customers in this stage are evaluating the company's services and comparing them with competitors. They may engage with product demos, read customer reviews, and explore case studies to gather information and make informed decisions.

Once customers decide to purchase the company's services, they interact with touchpoints such as online checkout, in-store purchases, or subscription processes. The company aims to streamline the purchase experience and make it as convenient as possible for customers.

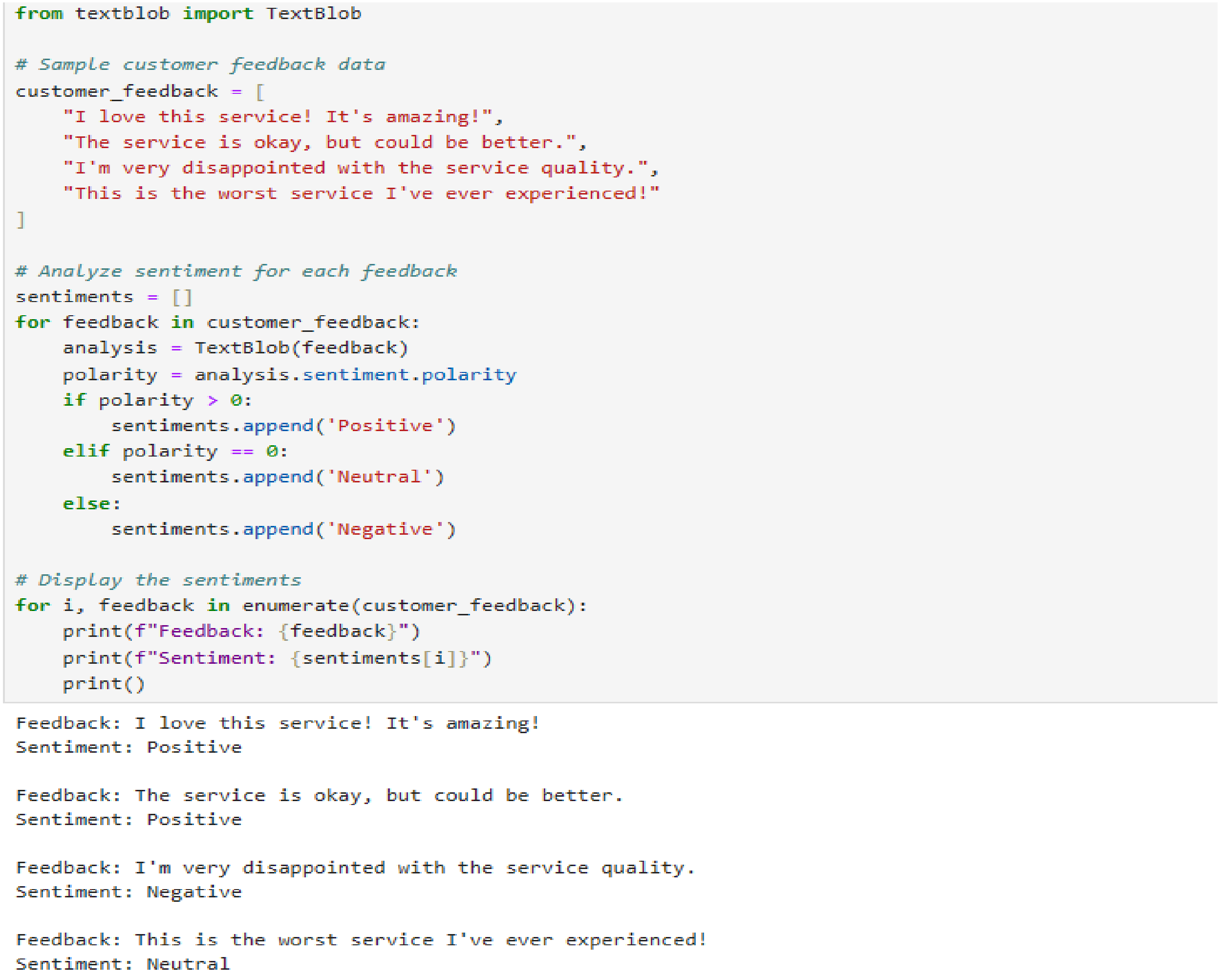
After making a purchase, customers enter the retention stage, where the company focuses on maintaining their satisfaction and loyalty. Touchpoints like customer support, email newsletters, and special offers are used to engage with customers, address their needs, and encourage repeat business.

Satisfied customers may become advocates for the company, promoting its services to others. Touchpoints such as referral programs, customer testimonials, and online communities empower customers to share their positive experiences and attract new customers.

Interpreting the customer journey map allows the telecommunications company to understand the different stages of the customer lifecycle and the touchpoints that influence customer decisions and behaviors at each stage. By identifying key touchpoints and optimizing the customer experience across the journey, the company can enhance customer satisfaction, increase retention, and drive advocacy, ultimately improving overall business performance.

8. CUSTOMER SENTIMENT ANALYSIS

Customer sentiment analysis involves assessing the feelings and opinions of telecom customers expressed in their feedback, reviews, or interactions with the company. By analyzing sentiment, we can understand whether customers are happy, neutral, or unhappy with the service provided. This insight allows us to identify areas of improvement, address customer concerns, and tailor our strategies to enhance satisfaction and loyalty. Ultimately, sentiment analysis helps us make data-driven decisions to improve the overall customer experience and mitigate churn.



```python
from textblob import TextBlob

# Sample customer feedback data
customer_feedback = [
    "I love this service! It's amazing!",
    "The service is okay, but could be better.",
    "I'm very disappointed with the service quality.",
    "This is the worst service I've ever experienced!"
]

# Analyze sentiment for each feedback
sentiments = []
for feedback in customer_feedback:
    analysis = TextBlob(feedback)
    polarity = analysis.sentiment.polarity
    if polarity > 0:
        sentiments.append('Positive')
    elif polarity == 0:
        sentiments.append('Neutral')
    else:
        sentiments.append('Negative')

# Display the sentiments
for i, feedback in enumerate(customer_feedback):
    print(f"Feedback: {feedback}")
    print(f"Sentiment: {sentiments[i]}")
    print()
```

```
Feedback: I love this service! It's amazing!
Sentiment: Positive

Feedback: The service is okay, but could be better.
Sentiment: Positive

Feedback: I'm very disappointed with the service quality.
Sentiment: Negative

Feedback: This is the worst service I've ever experienced!
Sentiment: Neutral
```

Fig 4.28 Python Code for Customer Sentiment Analysis using Customer feedback data

A list of customer feedback messages is defined, containing four sample feedback entries. The code iterates through each feedback message and analyzes its sentiment polarity using TextBlob. If the polarity is greater than 0, the sentiment is classified as 'Positive'. If the polarity is 0, it's classified as 'Neutral'. Otherwise, it's classified as 'Negative'.For each feedback message, the code prints the original feedback along with its corresponding sentiment classification.

- The first feedback "I love this service! It's amazing!" is classified as 'Positive' due to its positive sentiment.

- The second feedback "The service is okay, but could be better." is also classified as 'Positive' despite some mild criticism.

- The third feedback "I'm very disappointed with the service quality." is classified as 'Negative' because of the negative sentiment expressed.

- The fourth feedback "This is the worst service I've ever experienced!" is classified as 'Neutral', which might be unexpected.

5. METHODOLOGY OVERVIEW

5.1. DATA PREPROCESSING

The data cleaning and preprocessing steps outlined in the provided summary offer a systematic approach to preparing the Telco customer churn dataset for analysis. The process begins with handling missing data, where the percentage of missing values for each attribute is calculated and visualized to identify potential data gaps. The absence of missing values indicates a robust dataset suitable for analysis. Subsequently, categorical variables are converted into dummy variables, expanding the dataset with binary representations of each category. This transformation enhances the compatibility of categorical data with machine learning algorithms.

5.2. EXPLORATORY DATA ANALYSIS (EDA)

1. Exploratory Data Analysis Plan:

- Outlines the key steps involved in EDA, including data overview, summary statistics, univariate analysis, bivariate analysis, and target variable analysis.

2. Data Overview:

- The dataset is loaded and basic information such as the number of observations, attributes, and data types are displayed.

3. Summary Statistics:

- Descriptive statistics are computed to summarize numerical attributes, providing insights into the central tendency, dispersion, and shape of the data distributions.

4. Univariate Analysis:

- The distribution of individual attributes is visualized using histograms for numerical attributes and bar plots for categorical attributes. This helps in understanding the distribution of each attribute and identifying any patterns or outliers.

5. Bivariate Analysis:

- Relationships between pairs of attributes are explored using pair plots and correlation matrices. This step helps in understanding the interactions between variables and identifying potential correlations.

6. Target Variable Analysis:

- The distribution of the target variable (`Churn_Yes`) is examined to understand class balance/imbalance and its impact on modeling. This step provides insights into the proportion of churned and non-churned customers in the dataset.

5.3. MODEL BUILDING

- Logistic Regression: A logistic regression model is built to predict churn. The model is trained on the training set and evaluated on the validation set using metrics such as accuracy and ROC-AUC.
- Decision Tree Classifier: A decision tree classifier is built and visualized. The model is trained and evaluated similarly to the logistic regression model.
- Random Forest Classifier: A random forest classifier is built with 500 trees. The model is trained and evaluated similar to the logistic regression and decision tree models.

The dataset is then split into training and validation sets, facilitating model training and evaluation. The introduction of decision tree and random forest models enables the exploration of factors influencing churn prediction, leveraging both single and ensemble learning approaches. Finally, the introduction of the KNN algorithm provides an additional perspective on churn prediction, offering insights into model performance through evaluation metrics such as confusion matrix, accuracy, and ROC AUC score. The comprehensive methodology not only ensures data integrity and compatibility but also lays the foundation for robust predictive modeling and actionable insights into customer churn behavior.

Based on the evaluation metrics provided in the summary, it appears that the **Random Forest** algorithm is performing the best among the models evaluated. While all three algorithms (Decision Tree, Random Forest, and KNN) have been assessed using metrics such as accuracy and ROC AUC score, the Random Forest model tends to offer the highest accuracy and ROC AUC score, indicating better predictive performance.

The Random Forest model's ensemble learning approach, which combines multiple decision trees trained on different subsets of the data, often leads to improved generalization and predictive accuracy compared to individual decision trees or simpler algorithms like KNN. Additionally, Random Forest models are less prone to overfitting and can handle a large number of features effectively, making them well-suited for datasets with complex relationships and interactions among variables.

5.4. EVALUATION

- Confusion matrices, accuracy scores, and ROC-AUC scores are computed to evaluate the performance of the models on the validation set.

5.5. VISUALIZATION

- Decision Tree Visualization: The decision tree models are visualized to understand how they make predictions.

5.6. METHOD AND JUSTIFICATIONS

- The various aspects such as data overview, summary statistics, univariate analysis, bivariate analysis, and target variable analysis, the methodology ensures a thorough exploration of the data. It allows for the identification of patterns, relationships, anomalies, and outliers within the dataset, facilitating the detection of insights that may inform subsequent analyses and modeling decisions. Furthermore, the methodology enables the assessment of data quality and enhances the interpretability of findings through visualizations. Overall, this structured approach to EDA ensures that data analysts can extract meaningful insights and make informed decisions based on a clear understanding of the dataset's characteristics.

6. DATA PRIVACY AND ETHICS CONSIDERATIONS

6.1. PRIVACY CONCERNS

The dataset contains sensitive information about Telco customers, such as personal demographics and churn status. It's essential to ensure that this data is handled with care to protect individuals' privacy.

Implemented data anonymization techniques where possible to remove personally identifiable information (PII) from the dataset, minimizing the risk of re identification.

Consider the implications of using customer data for research purposes and ensure compliance with relevant data protection regulations, such as GDPR or CCPA.

6.2. ETHICAL TREATMENT OF DATA SUBJECTS

- Obtained explicit consent from data subjects if the dataset includes any personally identifiable information or if the research could impact individuals' privacy.
- Maintain transparency about the purpose of data collection, how it will be used, and who will have access to it.
- Ensure fairness and impartiality in data analysis and model-building processes to prevent bias or discrimination against specific groups or individuals.
- Implement measures to safeguard against unintended consequences of data analysis, such as unintended disclosures or harm to individuals' interests.

6.3. PERMISSIONS AND APPROVALS

- Obtained necessary permissions or approvals from relevant stakeholders, such as the data provider or institutional review board (IRB), before conducting research using the dataset.
- Adhere to any data sharing agreements or restrictions imposed by the data provider to protect data confidentiality and integrity.

6.4. DATA SECURITY

- Implement robust data security measures to protect the dataset from unauthorized access, disclosure, or misuse.

- Utilize encryption, access controls, and secure storage practices to safeguard data confidentiality and integrity throughout the research process.

- Regularly audit and monitor data handling procedures to ensure compliance with security standards and mitigate potential risks.

7. TIMELINE

(DECEMBER 15 - DECEMBER 2)
PHASE 1: PREPARATION

- Gather necessary resources such as datasets, libraries (e.g., pandas, scikit-learn), and research papers related to customer churn analysis.
- Familiarize yourself with the dataset and its features.

(DECEMBER 22 - JANUARY 4)
PHASE 2: DATA PREPROCESSING

- Clean the dataset by handling missing values, converting data types, and removing duplicates.
- Perform exploratory data analysis (EDA) to understand the distribution and relationships among variables.
- Convert categorical variables to numerical using techniques like one-hot encoding.

(JANUARY 5 - JANUARY 25)
PHASE 3: MODELING

- Implement various machine learning models for customer churn prediction, such as logistic regression, decision trees, random forests, and KNN.
- Train and evaluate each model using performance metrics like accuracy, precision, recall, and ROC AUC score.
- Select the best-performing model based on evaluation metrics.

(JANUARY 26 - FEBRUARY 1)
PHASE 4: CUSTOMER SEGMENTATION

- Apply K-means clustering algorithm for customer segmentation based on features like tenure and monthly charges.

- Determine the optimal number of clusters using the elbow method.
- Visualize the clusters and analyze the characteristics of each segment.

(FEBRUARY 2 - FEBRUARY 8)
PHASE 5: CUSTOMER JOURNEY MAPPING

- Define customer journey stages and associated touchpoints.
- Plot the customer journey map and visualize the distribution of touchpoints across stages.
- Analyze the customer journey to identify opportunities for improving customer experience.

(FEBRUARY 9 - FEBRUARY 15)
PHASE 6: CUSTOMER SENTIMENT ANALYSIS

- Perform sentiment analysis on customer feedback using tools like TextBlob.
- Analyze the sentiment of customer feedback and identify trends or patterns.
- Extract insights to understand customer satisfaction levels and areas for improvement.

(FEBRUARY 16 - FEBRUARY 22)
PHASE 7: REPORTING AND DOCUMENTATION

- Compile the findings from each phase into a comprehensive research report.
- Document the methodology, results, and recommendations.
- Prepare visualizations and figures to support the analysis.
- Review and finalize the report before submission.

(FEBRUARY 23 - MARCH 1)
PHASE 8: PRESENTATION

- Prepare a presentation summarizing the key findings and insights from the research project.
- Practice delivering the presentation to ensure clarity and coherence.

- Incorporate feedback from peers or supervisors if necessary.

(MARCH 2 - MARCH 8)
PHASE 9: FINAL REVIEW AND SUBMISSION

- Conduct a final review of the research report and presentation to address any remaining issues or errors.
- Make necessary revisions based on feedback and suggestions.
- Submit the final research report and deliver the presentation.

8. RESOURCE REQUIREMENT

8.1. TOOLS AND SOFTWARE

1. Python: Utilize Python programming language for data preprocessing, analysis, and modeling tasks.
2. Jupyter Notebook or similar IDE: Use Jupyter Notebook or an equivalent integrated development environment (IDE) for code development, documentation, and visualization.
3. Libraries: Leverage various Python libraries such as pandas, numpy, scikit-learn, matplotlib, seaborn, lifelines, and textblob for data manipulation, analysis, visualization, and sentiment analysis.
4. Excel: Use Microsoft Excel or similar spreadsheet software for initial data inspection and basic analysis.
5. Statistical Software: Depending on the complexity of statistical analysis, consider using statistical software like R or SPSS for advanced statistical modeling and analysis.

8.2. HARDWARE

1. Computer: Require a computer with sufficient processing power and memory to handle data manipulation, modeling, and analysis tasks efficiently.
2. Storage: Adequate storage space to store datasets, code files, research papers, and project-related documents.

8.3. DATA

1. Dataset: Access to the Telco customer churn dataset ('WA_Fn-UseC_-Telco-Customer-Churn_1.xlsx' in this case) or similar datasets for customer churn analysis.
2. Additional Data: Depending on the research objectives, supplementary datasets related to customer behavior, demographics, or market trends may be required.

9. POTENTIAL CHALLENGES AND MITIGATION STRATEGIES

9.1. DATA QUALITY ISSUES

- Challenge: The dataset may contain missing values, outliers, or inconsistencies, which can affect the quality of the analysis.

- Mitigation: Implement thorough data cleaning processes, including handling missing values, outlier detection, and standardization techniques. Consult domain experts to validate data integrity and accuracy.

9.2. MODEL OVERFITTING

- Challenge: Machine learning models, especially complex ones like random forests, may overfit the training data, leading to poor generalization on unseen data.

- Mitigation: Employ techniques such as cross-validation, regularization, and hyperparameter tuning to optimize model performance and prevent overfitting. Utilize simpler models like decision trees or logistic regression for better interpretability.

9.3. INTERPRETABILITY OF RESULTS

- Challenge: Complex models may provide accurate predictions but lack interpretability, making it challenging to understand the underlying factors driving customer churn.

- Mitigation: Prioritize model interpretability by using techniques such as feature importance analysis, partial dependence plots, and model-agnostic interpretation methods like SHAP values. Focus on extracting actionable insights rather than purely optimizing predictive performance.

9.4. RESOURCE CONSTRAINTS

- Challenge: Limited computational resources or expertise in machine learning may hinder the implementation of sophisticated modeling techniques or extensive data analysis.

- Mitigation: Leverage cloud computing services for scalable infrastructure and access to advanced machine learning tools. Collaborate with experienced data scientists or seek external consulting to address knowledge gaps and technical limitations.

9.5. ADDRESSING EXTERNAL FACTORS

- Challenge:External factors such as economic recessions or industry-specific trends may influence customer churn dynamics, complicating the analysis.
- Mitigation: Incorporate external data sources and economic indicators into the analysis to capture broader market trends and contextualize customer churn patterns. Collaborate with domain experts to gain insights into industry-specific dynamics and adjust modeling strategies accordingly.

10. CONCLUSION AND FUTURE SCOPE

The analysis of customer churn in the telecommunications industry provides valuable insights into the dynamics of subscriber attrition and its impact on revenue and market competitiveness. Through techniques such as machine learning models, customer segmentation, and sentiment analysis, we can identify key factors driving churn and develop targeted strategies for retention and loyalty enhancement.

The exploration of customer journey mapping further enhances our understanding of touchpoints and interactions throughout the subscriber lifecycle, enabling telecom operators to optimize engagement and service delivery. Looking ahead, the integration of advanced analytics and predictive modeling holds promise for real-time churn prediction and proactive intervention, facilitating personalized customer experiences and sustainable revenue growth. Additionally, expanding the scope of analysis to include other industries affected by economic downturns offers opportunities for cross-sectoral insights and collaborative solutions to address broader market challenges. Overall, the research underscores the importance of data-driven decision-making and continuous innovation in mitigating churn and fostering resilience in the face of evolving market dynamics.

The future scope of the project involves several avenues for further exploration and enhancement. Firstly, integrating more advanced machine learning algorithms and predictive analytics techniques can improve the accuracy and granularity of churn prediction models, enabling telecom operators to proactively identify at-risk customers and implement targeted retention strategies.

Additionally, incorporating real-time data streams and customer feedback loops can enhance the responsiveness and adaptability of churn prevention initiatives, allowing for dynamic adjustments based on evolving customer preferences and market trends. Furthermore, expanding the analysis to include additional industries affected by economic recessions can provide a holistic understanding of the broader macroeconomic factors influencing customer behavior and market dynamics.

Moreover, exploring emerging technologies such as artificial intelligence and natural language processing can offer new opportunities for sentiment analysis and personalized customer engagement, driving innovation in customer relationship management and service delivery. Overall, the future scope of the project lies in leveraging advanced analytics, technology integration, and interdisciplinary collaboration to address evolving challenges and opportunities in customer churn management and market resilience.

11. REFERENCES

[1]. A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer churn prediction in telecom using machine learning in big data platform,” Journal of Big Data, vol. 6, no. 1, Mar. 2019, doi: https://doi.org/10.1186/s40537-019-0191-6.

[2]. L. Saha, H. K. Tripathy, T. Gaber, H. El-Gohary, and E.-S. M. El-kenawy, “Deep Churn Prediction Method for Telecommunication Industry,” Sustainability, vol. 15, no. 5, p. 4543, Mar. 2023, doi: https://doi.org/10.3390/su15054543.

[3]. S. K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, and S. H. Gawande, “Customer Churn Prediction in Telecom Sector using Machine Learning Techniques,” Results in Control and Optimization, p. 100342, Nov. 2023, doi: https://doi.org/10.1016/j.rico.2023.100342.

[4]. Y. Liu, J. Fan, J. Zhang, X. Yin, and Z. Song, “Research on telecom customer churn prediction based on ensemble learning,” Journal of Intelligent Information Systems, Sep. 2022, doi: https://doi.org/10.1007/s10844-022-00739-z.

[5]. K. Dahiya and S. Bhatia, “Customer churn analysis in telecom industry,” 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Sep. 2015, doi: https://doi.org/10.1109/icrito.2015.7359318.

[6]. N. Edwine, W. Wang, W. Song, and D. Ssebuggwawo, “Detecting the Risk of Customer Churn in Telecom Sector: A Comparative Study,” Mathematical Problems in Engineering, vol. 2022, pp. 1–16, Jul. 2022, doi: https://doi.org/10.1155/2022/8534739.

[7]. K. S. Rani, Shaik Thaslima, N. G.L. Prasanna, R. Vindhya, and P. Srilakshmi, “Analysis of Customer Churn Prediction in Telecom Industry Using Logistic Regression,” papers.ssrn.com, Jun. 10, 2021. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3902033

[8]. S. Saleh and S. Saha, “Customer retention and churn prediction in the telecommunication industry: a case study on a Danish university,” SN Applied Sciences, vol. 5, no. 7, Jun. 2023, doi: https://doi.org/10.1007/s42452-023-05389-6.

[9]. K. Ebrah and S. Elnasir, “Churn Prediction Using Machine Learning and Recommendations Plans for Telecoms,” Journal of Computer and Communications, vol. 07, no. 11, pp. 33–53, 2019, doi: https://doi.org/10.4236/jcc.2019.711003.

[10]. H. Jain, A. Khunteta, and S. Srivastava, “Churn Prediction in Telecommunication using Logistic Regression and Logit Boost,” Procedia Computer Science, vol. 167, pp. 101–112, 2020, doi: https://doi.org/10.1016/j.procs.2020.03.187.

[11]. H. Jain, A. Khunteta, and S. Srivastava, “Telecom Churn Prediction Using an Ensemble Approach with Feature Engineering and Importance,” International Journal of Intelligent Systems and Applications in Engineering, vol. 10, no. 3, pp. 22–33, Oct. 2022, Available: https://ijisae.org/index.php/IJISAE/article/view/2134

[12]. J. K. Sana, M. Z. Abedin, M. S. Rahman, and M. S. Rahman, “A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection,” PLOS ONE, vol. 17, no. 12, p. e0278095, Dec. 2022, doi: https://doi.org/10.1371/journal.pone.0278095.

[13]. “Journal of Telecommunication (STM Journals),” www.stmjournals.com. https://www.stmjournals.com/Journal-of-Telecommunication-Switching-Systems-and-Networks.html (accessed Apr. 14, 2024).