Customer Churn Prediction and Analysis

Data Science Track – Round 2

Group (CAI2_AIS4_S1)

Instructor: Ahmed Moustafa

https://github.com/AlkingZiad/-Customer-Churn-Prediction-and-Analysis-project

Ziad Moataz samy (21067779)

Ahmed Mohamed Radwan (21057072)

Abdelrahman Rawhy (21091992)

Ali eldin ahmed (21085612)

Ahmed Nasser (21074056)

Ahmed gamal (21085587)

Omar osama (21079477)

1.Abstract

Customer churn is a significant concern in the telecommunications industry, as retaining existing customers is often more cost effective than acquiring new ones. This project investigates customer churn behaviour using the Telco Customer Churn dataset through a combination of exploratory data analysis (EDA), churn pattern identification, and machine learning-based prediction. Key insights were drawn from analysing customer demographics, service usage, contract types, and payment methods in relation to churn. The dataset was cleaned and enriched with new features such as Customer Lifetime Value (CLV), service counts, and payment digitalization indicators. Visualization techniques were used to identify critical churn patterns revealing, for example, that month-to-month contracts and electronic payments were highly correlated with churn. Finally, the project prepares the data for machine learning pipelines aimed at building predictive models to accurately identify customers at risk of leaving. This work contributes to understanding churn dynamics and provides actionable insights to improve customer retention strategies.

2.Introduction

In today's competitive telecom industry, customer retention is an important driver of business success. With numerous providers offering similar services at comparable prices, maintaining a loyal customer base has become more challenging than ever. As a result, telecom companies are increasingly relying on data-driven approaches to understand and reduce churn. This project focuses on analysing and predicting customer churn using the Telco Customer Churn dataset. By exploring customer demographics, account information, and service usage patterns, the analysis aims to uncover meaningful insights into the factors contributing to churn. These insights can help telecom providers make informed decisions about customer engagement, loyalty programs, and targeted marketing strategies.

The project follows a structured pipeline that includes data cleaning, exploratory data analysis (EDA), feature engineering, and data preprocessing. It also prepares the dataset for machine learning models that can predict whether a customer is likely to leave the service. Through this end-to-end analysis, the project highlights the power of predictive analytics in addressing real-world business problems and offers practical solutions for improving customer retention in the telecom sector.

3.problem statement

Customer churn is a significant challenge for telecom companies, as acquiring new customers often costs more than retaining existing ones. Understanding the underlying reasons why customers leave is essential for improving customer satisfaction and reducing revenue loss. However, identifying churn patterns from large, complex datasets can be difficult without the use of advanced data analysis and predictive modeling techniques.

The main objective of this project is to analyze customer data and build a predictive model that accurately identifies customers who are likely to churn. By leveraging historical data including customer demographics, service subscriptions, usage behavior, and account details the goal is to discover key factors influencing churn and to provide actionable insights for retention strategies.

4.dataset description

The dataset used in this project is sourced from a telecom company uploaded on Kaggle and contains customer-related data that spans various attributes, including demographics, account information, and service usage details. It is designed to provide insights into customer behaviour and churn patterns.

Shape of the Dataset:

- **Rows (Samples):** 7043
- **Columns (Features):** 21

Feature	Description	Data Type	Values
customerID	Unique identifier for each customer.	Categorical	Unique ID
gender	The gender of the customer.	Categorical	Male, Female
SeniorCitizen	Indicates whether the customer is a senior citizen.	Categorical	1 = Yes, 0 = No
Partner	Whether the customer has a partner.	Categorical	Yes, No
Dependents	Whether the customer has dependents.	Categorical	Yes, No
tenure	The number of months the customer has been with the company.	Numerical	Integer (Months)

PhoneService	Whether the customer has phone service.	Categorical	Yes, No
MultipleLines	Whether the customer has multiple lines.	Categorical	Yes, No, No phone service
InternetService	Type of internet service the customer uses.	Categorical	DSL, Fiber optic, No
OnlineSecurity	Whether the customer has online security.	Categorical	Yes, No, No internet service
OnlineBackup	Whether the customer has online backup.	Categorical	Yes, No, No internet service
DeviceProtection	Whether the customer has device protection.	Categorical	Yes, No, No internet service
TechSupport	Whether the customer has tech support.	Categorical	Yes, No, No internet service
StreamingTV	Whether the customer has streaming TV.	Categorical	Yes, No, No internet service

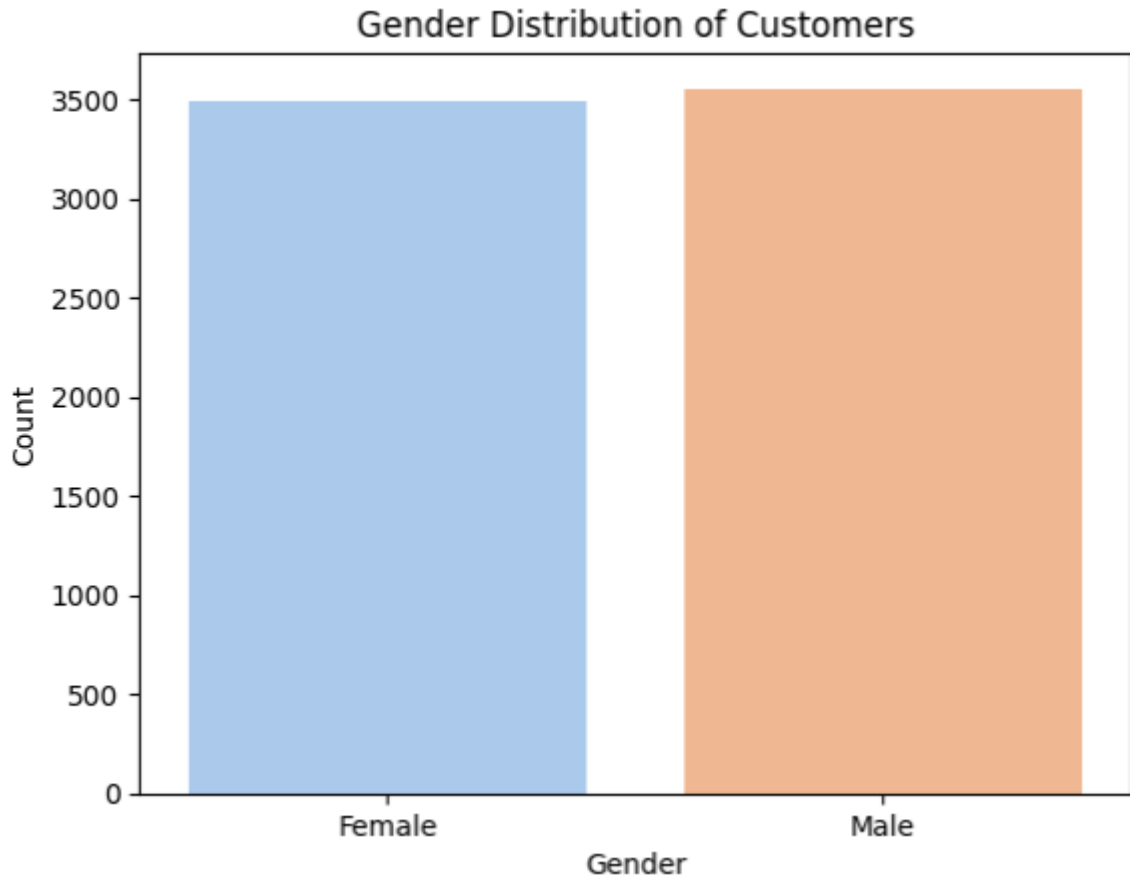
StreamingMovies	Whether the customer has streaming movies.	Categorical	Yes, No, No internet service
Contract	The type of contract the customer has.	Categorical	Month-to-month, One year, Two year
PaperlessBilling	Whether the customer uses paperless billing.	Categorical	Yes, No
PaymentMethod	The payment method the customer uses.	Categorical	Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)
MonthlyCharges	The amount the customer is charged monthly for services.	Numerical	Float (Amount in dollars)
TotalCharges	The total amount the customer has been charged during their tenure with the company.	Numerical	Float (Total amount in dollars)

Churn	The target variable indicating whether the customer has churned.	Categorical	Yes, No

5. Exploratory Data Analysis (EDA)

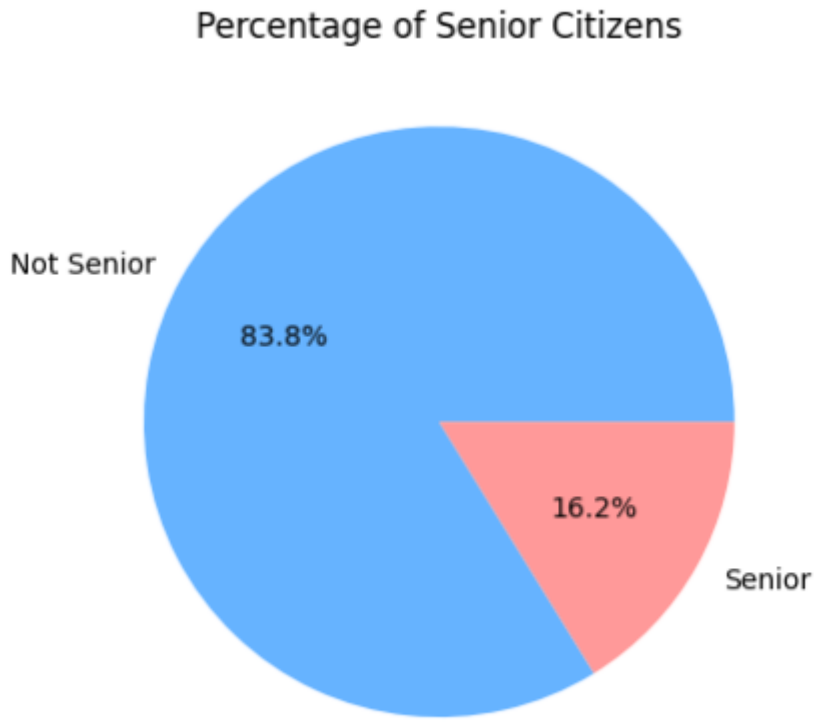
5.1 Gender Distribution

We first analyse the gender distribution of customers to understand whether the dataset is balanced in terms of gender.



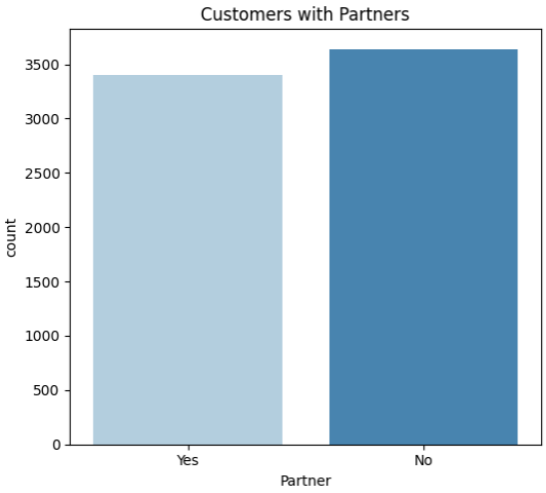
5.2 Senior Citizens

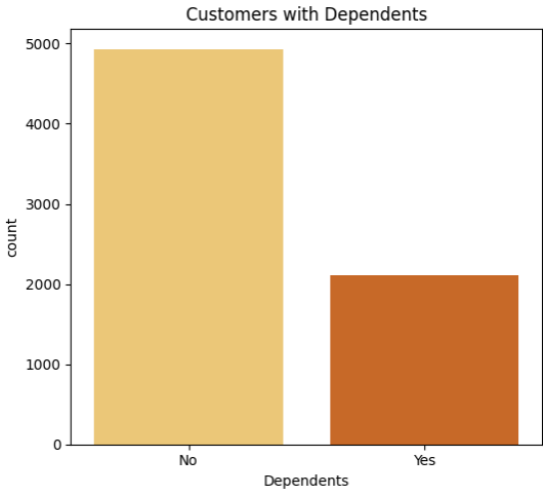
We investigate the percentage of senior citizens in the dataset and compare them to non-senior citizens.



5.3 Customers with Partners and Dependents

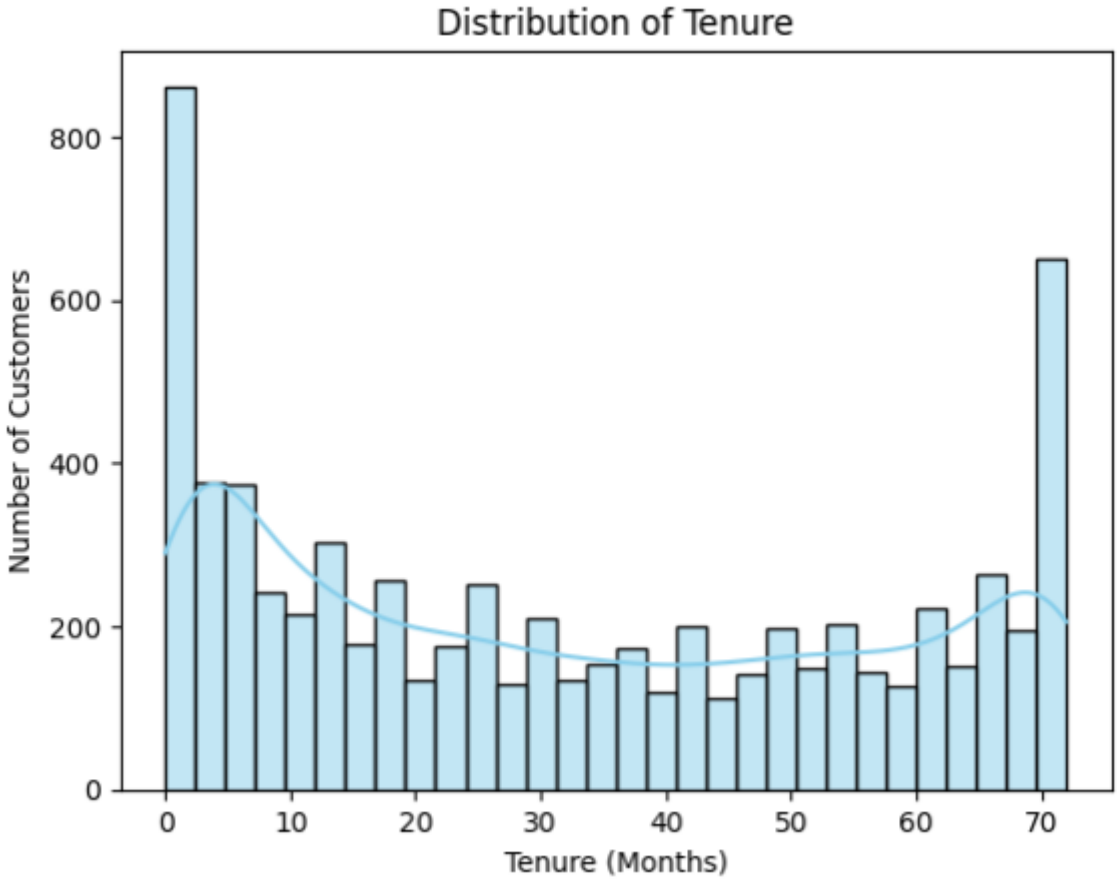
We explore how many customers have partners and dependents, which may influence customer behaviour and churn.





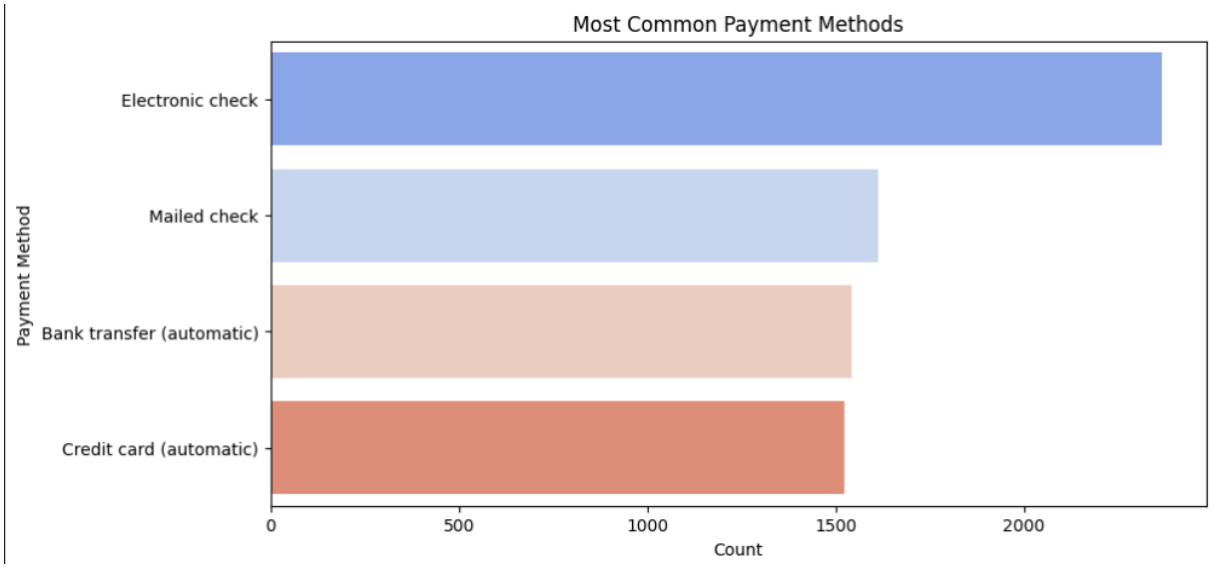
5.4 Distribution of Tenure

The length of time customers have stayed with the company is an important factor in churn prediction. We visualize this by plotting the distribution of the tenure feature.



5.5 Most Common Payment Methods

We analyse the most common payment methods used by customers. Understanding this can help determine if payment method types are associated with churn.

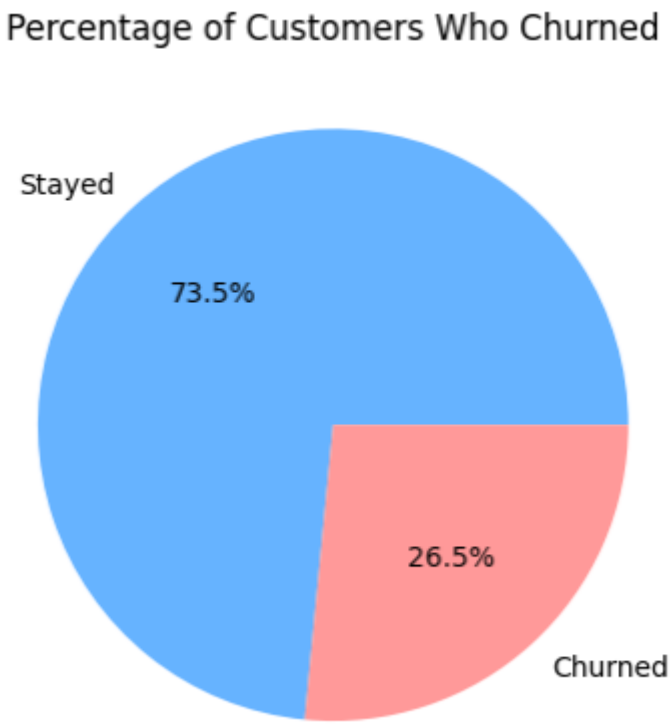


6. Churn Analysis

Customer churn refers to customers who leave the service. In this section, we analyse the churn rate and various factors related to churn.

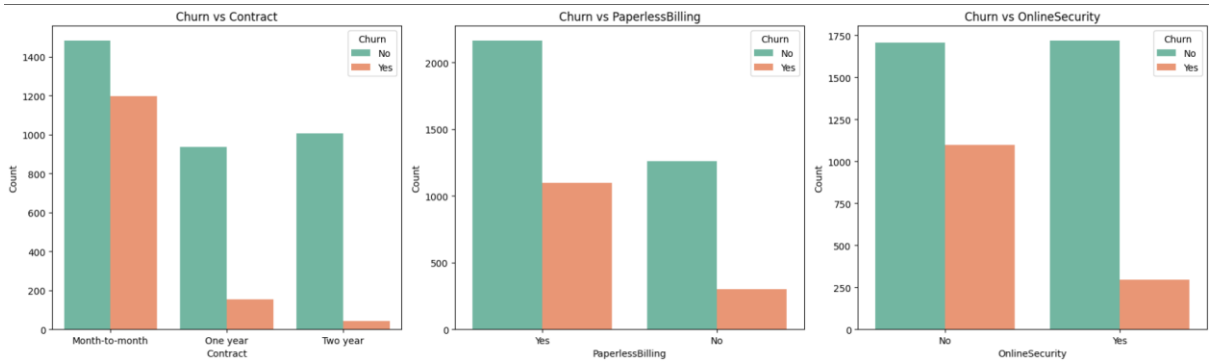
6.1 What Percentage of Customers Have Churned?

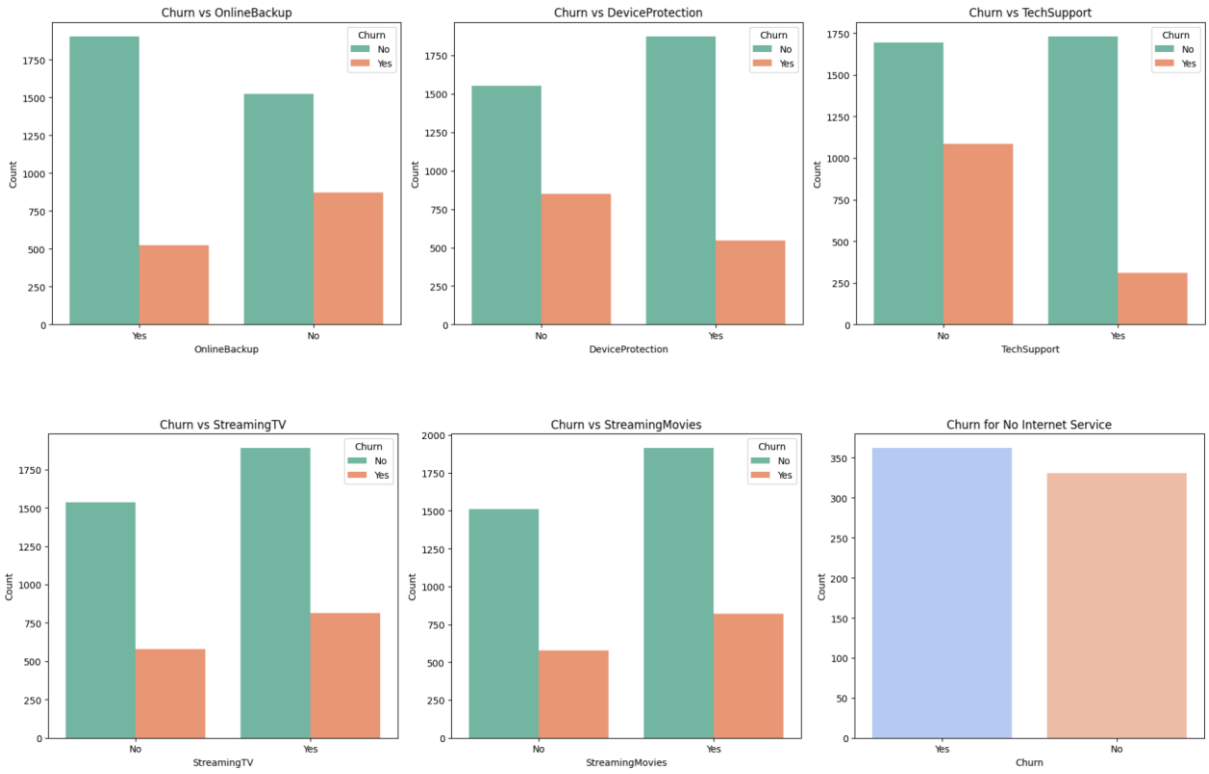
We begin by calculating the percentage of customers who have churned.



6.2 Services Used by Churned Customers

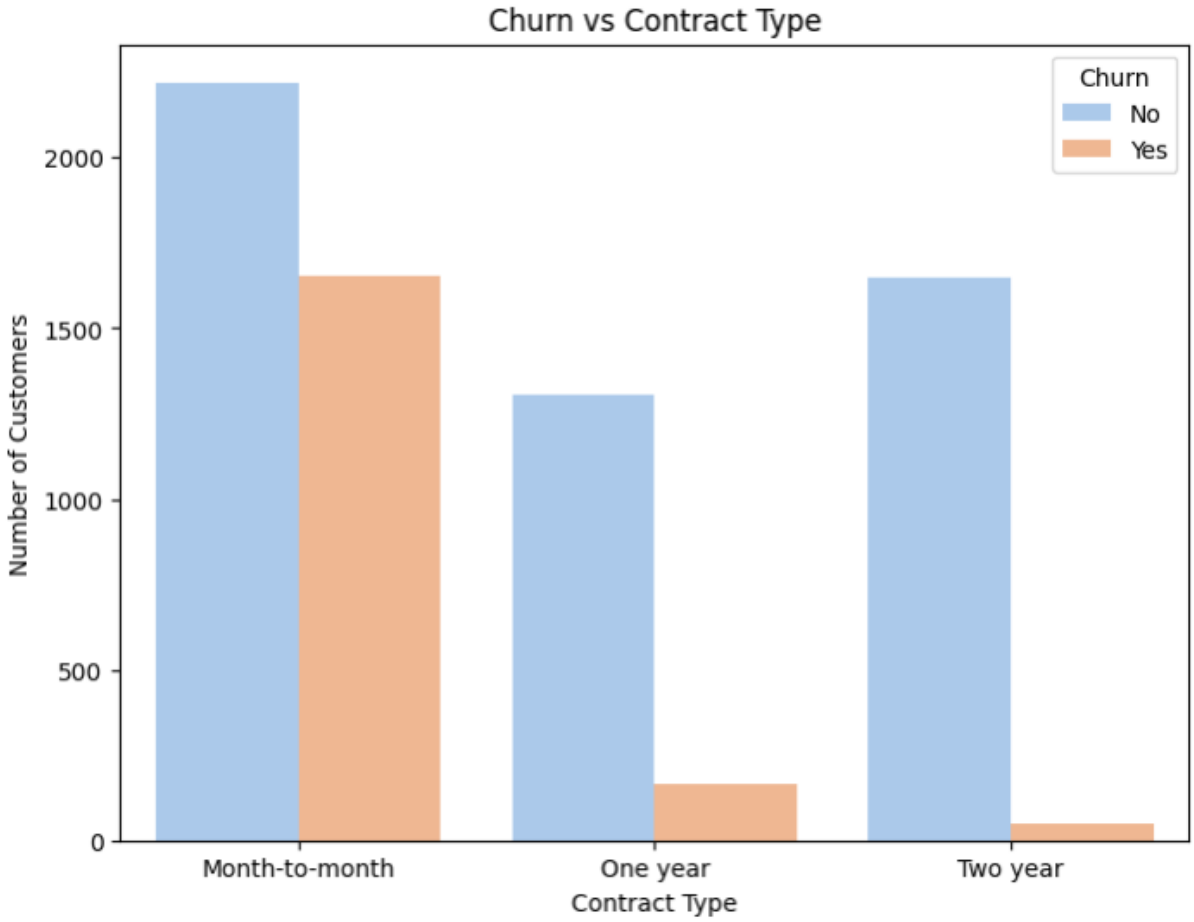
Next, we examine the most common services used by customers who have churned. This helps identify whether certain services are associated with higher churn rates.





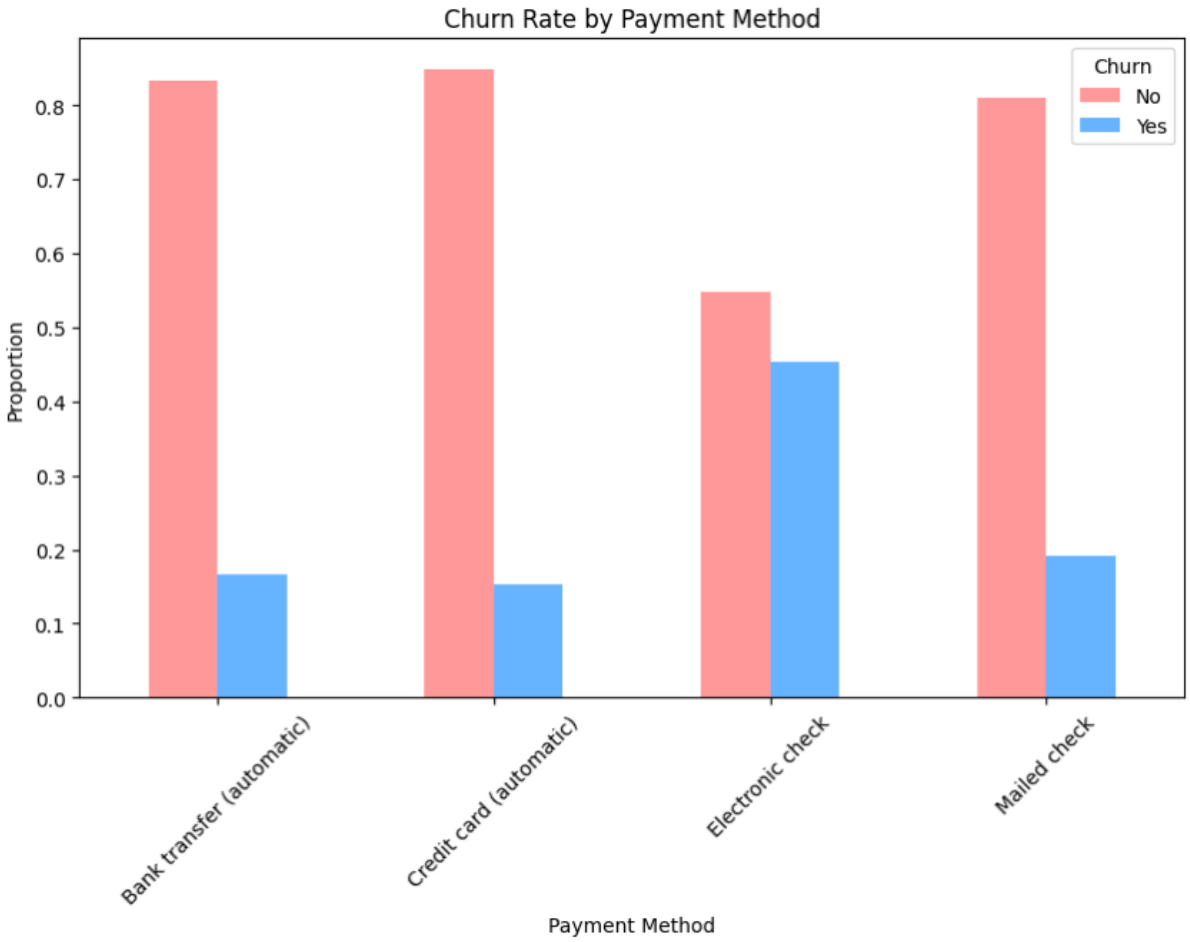
6.3 Churn by Contract Type

We analyse how churn varies based on the type of contract customers have with the company.



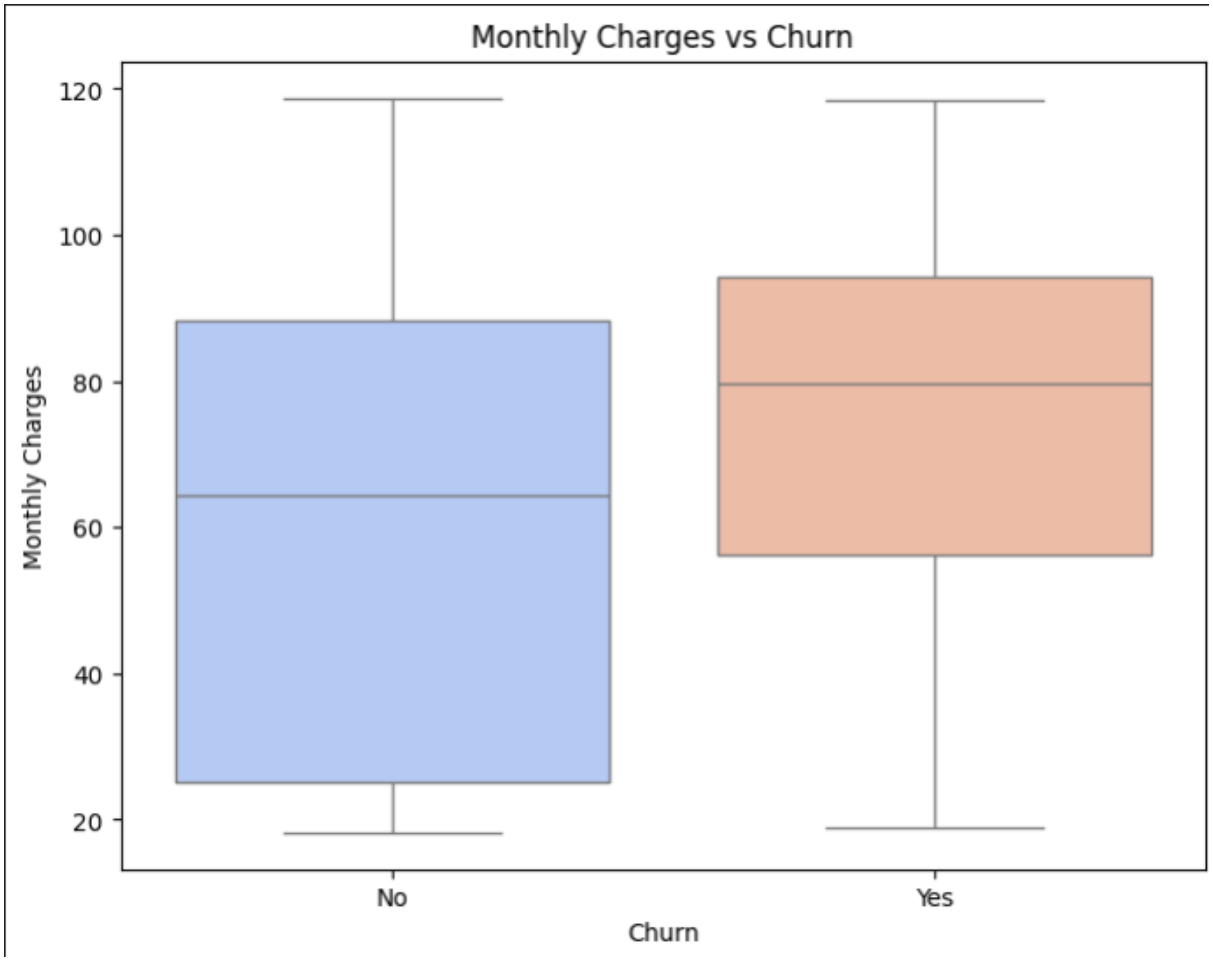
6.4 Churn Rate by Payment Method

We investigate whether there is any relationship between payment methods and churn rate.



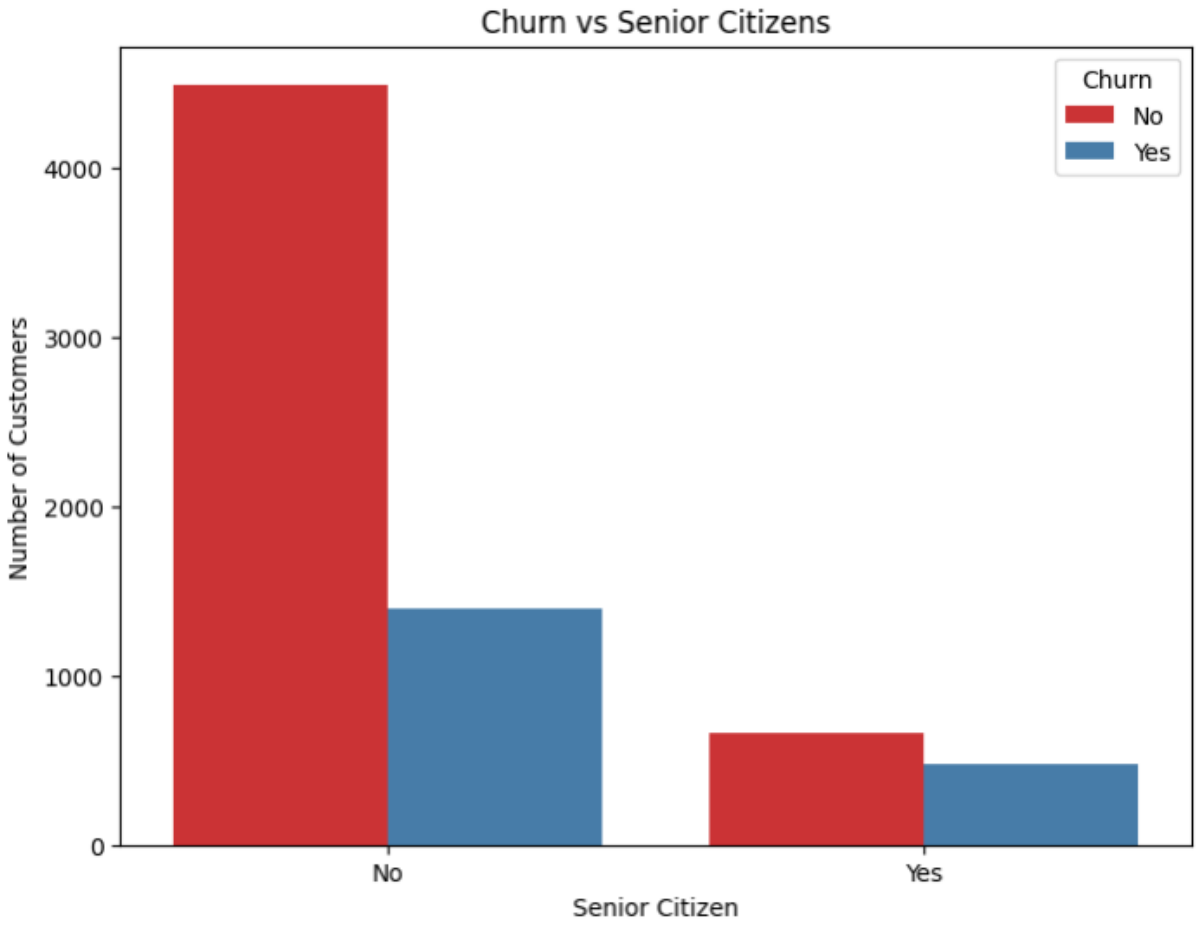
6.5 Relationship Between Monthly Charges and Churn

We explore whether there is a significant difference in monthly charges between customers who churned and those who stayed.

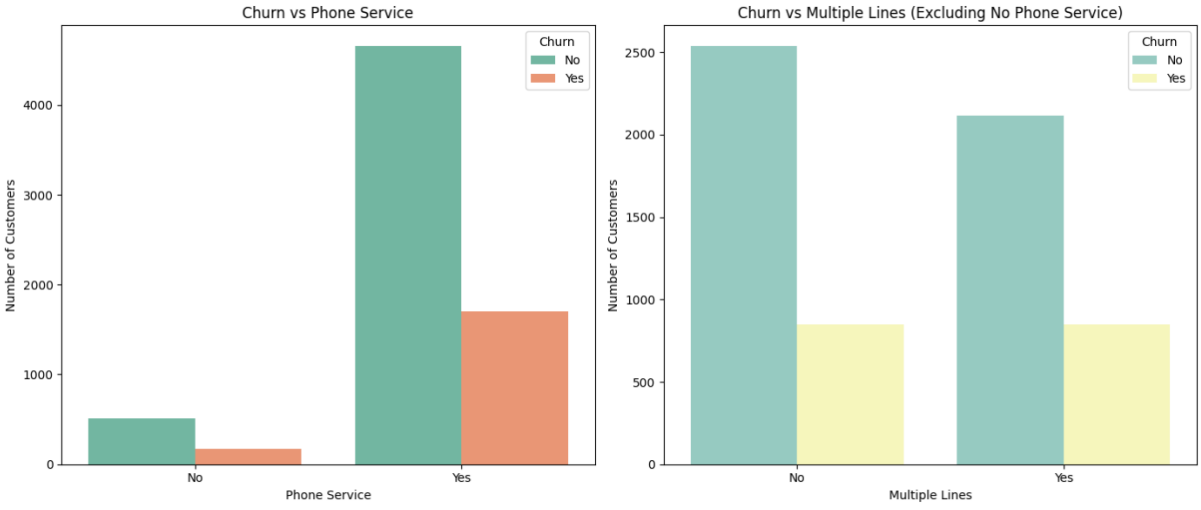


6.6 Churn by Senior Citizens

We analyze whether senior citizens have a higher churn rate compared to younger customers.

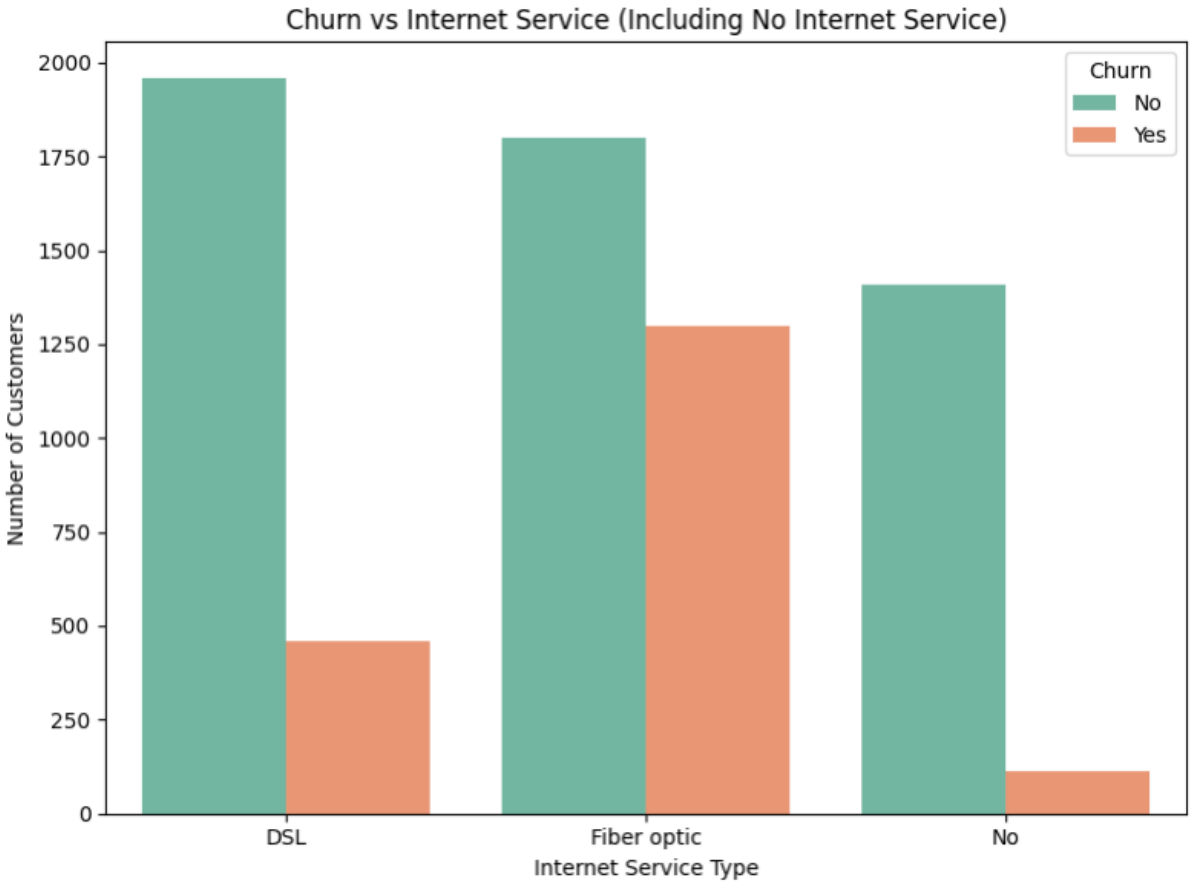


6.7 Relationship Between Phone Service, Multiple Lines, and Churn



6.8 Relationship Between Internet Services and Churn

We investigate how the type of internet service is related to churn.



7.Data Cleaning

It was essential to clean the dataset to ensure data integrity and consistency. The raw Telco Customer Churn dataset included various types of inconsistencies such as missing values, incorrect data types, and redundant information. Below is a summary of the key data cleaning steps that were undertaken:

1. Data Type Conversion

Several columns were incorrectly interpreted as object types, especially numeric columns like TotalCharges. This was resolved by:

- Converting TotalCharges from object to float using pd.to_numeric().
- Ensuring categorical features remained as object or were later converted to category as needed for modeling.

2. Handling Missing Values

After conversion, it was observed that the TotalCharges column had some missing values. This occurred because some customers had no total charges (likely due to very short tenure).

- Rows with missing TotalCharges were removed since they represented a very small portion of the dataset and were not significant for churn prediction.

3. Duplicate Data Check

The dataset was checked for duplicate records using df.duplicated(), and no duplicates were found.

4. Whitespace and Formatting Issues

Whitespace and inconsistencies were found in some categorical columns such as SeniorCitizen and MultipleLines.

- Stripped extra spaces using .str.strip().
- Ensured binary columns were consistently labeled (Yes/No, Male/Female, etc.).

5. Standardizing Values

Some categorical features had values such as 'No internet service', which were semantically similar to 'No'. These were standardized for consistency.

This step helped simplify the model and reduced the number of categories for encoding.

6. Conversion of SeniorCitizen Column

The SeniorCitizen column was originally an integer (0 or 1), which was transformed into a categorical variable for clarity

8.Feature engineering

Feature engineering was an important step in transforming raw data into meaningful input for the machine learning model. This process involved creating new features, simplifying existing ones, and converting data into formats better suited for analysis and modeling. Below are the key feature engineering steps applied:

1. Customer Lifetime Value (CLV)

A new feature, **Customer Lifetime Value (CLV)**, was created to estimate the potential value of a customer to the company

This metric helped quantify customer profitability over their tenure and proved valuable in churn prediction.

2. Average Monthly Charges

To better understand customer spending behavior, the **average monthly charge** was calculated

This helped account for variations in billing over time and mitigated the influence of one-time charges.

3. Service Count

To represent how many services a customer is subscribed to, a **NumServices** feature was created by counting the number of 'Yes' values in multiple service-related columns

This provided a compact and informative way to capture service usage behavior

4. Tenure Grouping

To reduce the granularity of the tenure feature and capture customer loyalty trends, tenure was binned into groups:

This made the data more interpretable and helped identify churn patterns over different stages of customer lifecycle.

5. Binary Mapping

Binary categorical columns such as Partner, Dependents, and PaperlessBilling were mapped from 'Yes'/'No' to 1/0 for compatibility with machine learning algorithms.

The target column Churn was also transformed to a binary label: 1 for churned customers, 0 otherwise.

6. Dropping Redundant Columns

After feature creation, the following columns were dropped:

- customerID – Not useful for modeling.
- Columns fully encoded by new features (e.g., individual service columns could be dropped after creating NumServices if desired).

9. Data Preprocessing

Before building machine learning models, the data needed to be transformed into a clean and model-ready format. The preprocessing steps ensured all variables were appropriately encoded, scaled, and free of inconsistencies or null values.

1. Encoding Categorical Variables

Categorical variables were processed based on the number of unique categories:

- **Binary categories (Yes/No)** were label encoded as 1/0 (e.g., Partner, Dependents, PaperlessBilling, Churn).

- **Multiclass categorical columns** (e.g., InternetService, Contract, PaymentMethod, TenureGroup) were **one-hot encoded** using pd.get_dummies():

2. Feature Scaling

Numerical columns such as MonthlyCharges, TotalCharges, tenure, CLV, and AvgMonthlyCharge were scaled using **StandardScaler** to normalize their distribution and ensure equal contribution to the model

This step improved model convergence and performance, especially for distance-based algorithms.

3. Train-Test Split

To evaluate model performance, the data was split into training and testing sets using **stratified sampling** to maintain class balance:

- train_size: 80% of data for training.
- test_size: 20% of data for testing.
- stratify=y: Ensures churn class distribution is preserved.

4. Final Check

After preprocessing, the final dataset contained:

- Fully numeric and scaled features.
- No missing values.
- Balanced class distribution across training and test sets.

10. Model Building and Evaluation

After completing data cleaning, feature engineering, and preprocessing, we proceeded to build and evaluate several machine learning models for predicting customer churn. The target variable, Churn, is a binary indicator where 1 denotes a customer who churned and 0 denotes a customer who remained.

The dataset was split into training and testing sets using stratified sampling to preserve the original distribution of the Churn variable. 80% of the data was used for training and 20% for testing. A full pipeline was implemented to handle both categorical and numerical features, as well as to automate the feature engineering process.

10.1 Preprocessing Pipeline

The pipeline included:

- **Feature Engineering**:
    - Customer Lifetime Value (CLV)
    - Average Monthly Charges
    - Binary indicators for services (Phone, Internet)
    - Count of subscribed services
    - New customer flag (tenure ≤ 6 months)
    - Contract type encoded as an ordinal feature
    - Digital payment method flag
- **Numerical Features**:
    - Imputed using the median strategy
    - Standardized using StandardScaler

- **Categorical Features**:
    - Imputed with a constant placeholder "missing"
    - One-hot encoded using OneHotEncoder with handle_unknown='ignore'

10.2 Model Training

The preprocessed dataset was used to train a variety of supervised classification models. All models were evaluated using accuracy as primary matrix

Below are the models that were trained:

10.3.2 Logistic Regression

- **Description**: A baseline linear model for binary classification.
- **Strengths**: Fast, interpretable, good for linearly separable data.
- **Performance**:
    - Accuracy: 80.4%
- **Observation**: Logistic Regression served as a strong baseline model, but struggled with nonlinear patterns.

10.2.2 Random Forest Classifier

- **Description**: An ensemble method that builds multiple decision trees and averages their predictions.
- **Strengths**: Handles both numerical and categorical features well, robust to outliers, captures nonlinearity.
- **Performance**:
    - Accuracy: 80.3%

- **Observation**: Delivered strong performance, particularly in handling complex interactions between features like service combinations and contract types.

10.3 Final Model Selection

After comparing all the models, the **Logistic Regression** was selected as the final model due to its superior performance in accuracy. It was also effective in handling the moderately imbalanced target variable and the nonlinear relationships introduced through feature engineering.

11.Results and Discussion

In this project, we investigated customer churn in a telecommunications company using exploratory data analysis (EDA) and machine learning techniques. The dataset, consisting of 7,032 customer records, revealed valuable insights such as a higher churn rate among customers without internet service, those with month-to-month contracts, and those lacking services like "TechSupport" and "OnlineSecurity." Additionally, the analysis highlighted that customers with higher monthly charges or those using "Electronic check" payment methods were more likely to churn. These findings suggest that pricing, service offerings, and contract flexibility significantly influence customer retention.

The machine learning models built for churn prediction, including Logistic Regression, Random Forest showed strong performance, with the Logistic Regression model achieving an accuracy of 80.%. The model's feature importance analysis highlighted key churn predictors such as "Contract type," "Internet service," "MonthlyCharges," and "TechSupport," confirming insights from the EDA. The Random Forest model demonstrated excellent precision and recall, and hyperparameter tuning slightly improved its performance. Overall, this project demonstrates how machine learning can provide valuable insights into customer

behaviour, helping businesses develop targeted strategies to reduce churn and improve customer retention.

12.conclusion

In conclusion, this Telco Customer Churn analysis and prediction project offers valuable insights into the factors influencing customer retention and churn in the telecommunications industry. Through thorough exploratory data analysis (EDA) and feature engineering, we explored customer demographics, service usage patterns, and payment methods, identifying key variables related to churn. We observed notable relationships between factors such as contract type, internet service, and payment methods, providing actionable insights for improving customer retention strategies. The application of advanced feature engineering techniques helped enhance the model's predictive power by adding customer lifetime value, average monthly charges, and other relevant features. By building a comprehensive machine learning pipeline for churn prediction, we were able to assess the effectiveness of various models and identify the most critical factors contributing to churn. Ultimately, this project demonstrates the power of data-driven analysis in understanding customer behaviour and informs strategies that could potentially reduce churn rates and improve customer satisfaction.