Group 4

Topic: Telco-Customer-Churn change

Student Name	Student ID
Asma Mohammad	23203269
WONG WING YIN	22231137
Li yueting	22227660

Table of content

Overview

A. Purpose

The primary purpose of this report is to outline analysis strategies to predict whether a telecom contract will be canceled or not. By doing so, telecom can employ proactive tactics, including targeted marketing, to lower the churn rate (27%) and make better decisions with predictive models.

B. Importance

High churn rates will significantly impact the company's revenue and growth.

```
Churn
0    5174
1    1869
Name: count, dtype: int64
```

```
total_monthly_charges: $ 456116.6
mean_monthly_charges: 64.76169246059918
```

The average monthly charge for a contract is $64.7, and the churn is 1869. So, in sum, the company is losing around 1869*64.7 = $120,924.3 per month. If the company can reduce the churn rate by just 20%, it could retain approximately 374 customers (20% of 1,869). The company could earn $24,184 more per month, which is $290,218 per year.

Therefore, the company must utilize predictive analytics to identify at-risk customers, develop targeted retention strategies such as special offers and loyalty rewards, and reduce churn rates through proactive engagement.

These allow the business to comprehend the wants and preferences of its customers. Additionally, leveraging data to tailor client interactions and service offers. and accomplish the ultimate objective of making a sizable profit.

Data analysis

A. Data Exploration Analysis

EDA, or exploratory data analysis, is a crucial initial stage in data science projects. Visualizing and analyzing data to understand its key characteristics, uncover patterns, and identify relationships between variables is the process of studying and examining record sets to identify outliers, find patterns, and understand their predominant traits.EDA is normally carried out as a preliminary step before undertaking extra formal statistical analyses or modeling.

EDA can be classified as univariate, bivariate, or multivariate based on the number of columns being analyzed.

1. Analysis of Univariates

The goal of univariate analysis is to comprehend the internal structure of a single variable. Finding patterns in a single characteristic and characterizing the data are its main concerns. Character variable analysis inside the data set is a specialization of this type of examination. It involves summarizing and visualizing an unmarried variable at a time to understand its distribution, relevant tendency, unfold, and different applicable records.

As show below in graph I, within the dataset, "SeniorCitizen" shows the distribution of senior citizens and non-senior citizens. A binary representation is shown on the x-axis, with 0 being non-senior citizens and 1 denoting senior citizen. The count is shown on the y-axis. With roughly 6000 non-senior citizens and 1000 senior citizens, it is clear from the chart that there are many more non-senior citizens than senior citizens. This discrepancy indicates that the majority of the clientele is younger. Businesses need to understand this demographic dispersion in order to adjust their marketing tactics and

offerings appropriately. In particular, they make sure that the requirements of older persons are met through focused initiatives while concentrating on the greater section of non-senior residents. Making well-informed decisions to improve customer happiness and retention is made easier with this understanding of client demographics.
According to the graph II, With the x-axis representing tenure in months (0 to 75) and the y-axis representing the number of customers, the histogram named "tenure" shows the distribution of customer tenure within the dataset. The chart displays notable peaks at the start of the tenure range (0–5 months) and at the end (70–75 months), signifying a large number of new clients as well as a solid foundation of devoted, long-term clients. With fewer clients and a more even distribution, the middle range (10–60 months) indicates stable but modest retention. These findings underline the significance of retention initiatives for recently acquired clients and point to ways to improve engagement tactics for mid-tenure clients in order to foster enduring loyalty. All things considered, comprehending these tenure trends is essential for creating focused retention plans and raising client delight.

2. Analysis of Bivariate

Examining the relationship between variables is part of bivariate evaluation. It makes it possible to identify dependencies, correlations, and relationships between variable pairs. An essential type of exploratory data analysis that looks at the relationship between two variables is bivariate analysis. Several essential methods for bivariate analysis include.

In the graph III, there are two plots in the graphic that shed light on customer attrition. The scatter plot on the left illustrates the correlation between monthly charges (y-axis) and customer tenure (x-axis). Data points are coloured according to churn status, with orange representing churned customers and blue representing non-churned customers. It is clear from this plot that customers who have shorter tenure and greater monthly expenses are more likely to leave.

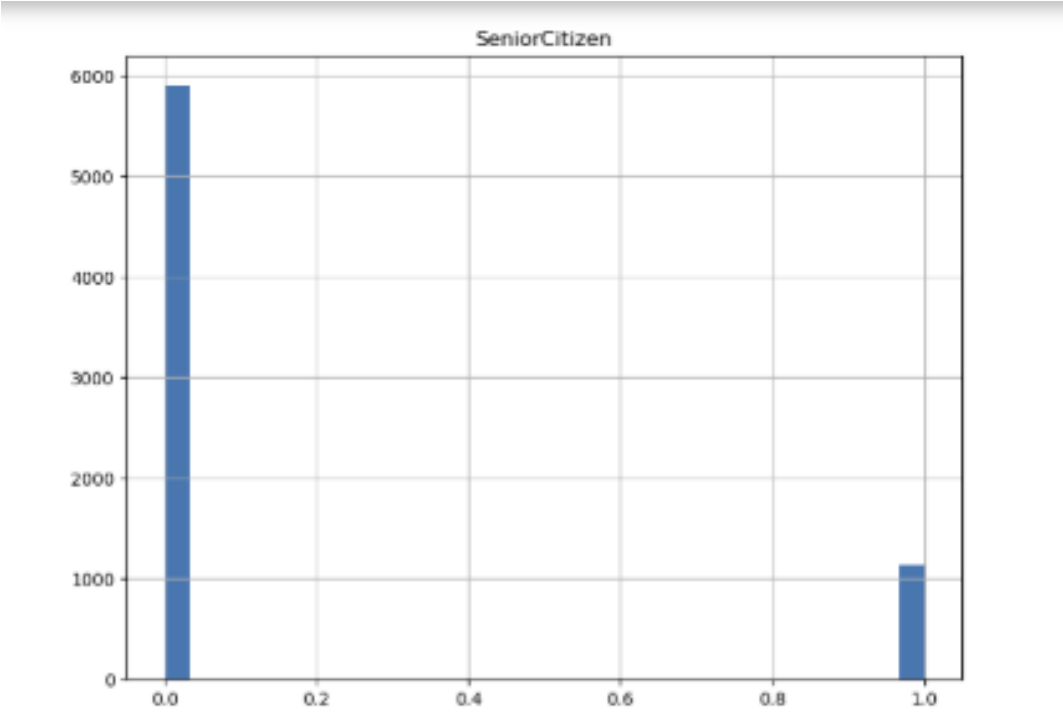
The monthly fee distribution for both churned and non-churned customers is shown in the density figure on the right. Non-churned consumers are represented by the blue line, which peaks at lower monthly rates and shows that most loyal customers typically pay less. The orange line for churned clients, on the other hand, displays a larger density in the mid- to high-range monthly charge. This implies that these clients typically spend more. According to these findings, in order to lower attrition and increase retention, pricing strategies and customer interactions should be modified to meet the demands of clients who pay more.

3. Outliers

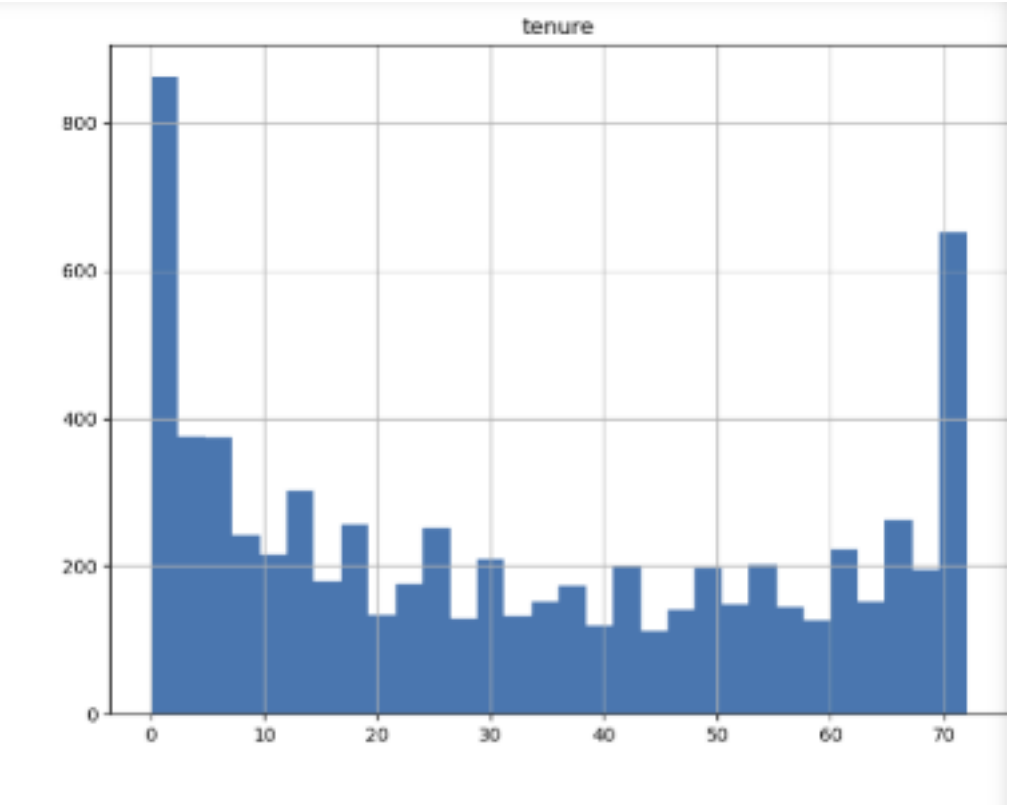
The Graph IV shows a box plot called "Box plot of tenure," which shows how tenure data is distributed. With the label "tenure," the x-axis most likely shows how long in months clients have been with the business. In a box plot, the median is at about 30 is represented by a vertical line inside a rectangular box that runs from the first quartile which is roughly 10 to the third quartile which is roughly 50. From the box's edges to the dataset's least about 0 month and highest of roughly 70 months

The median tenure and the range that the majority of data points fall within may be quickly identified thanks to this visualization, which offers a succinct assessment of the central tendency, spread, and skewness of the tenure data. With a median tenure of almost 30 months, the plot shows that the majority of customers have tenures ranging from 10 to 50 months. Given that some clients have extremely short or very lengthy tenures, the whiskers indicate a wider distribution of the data.

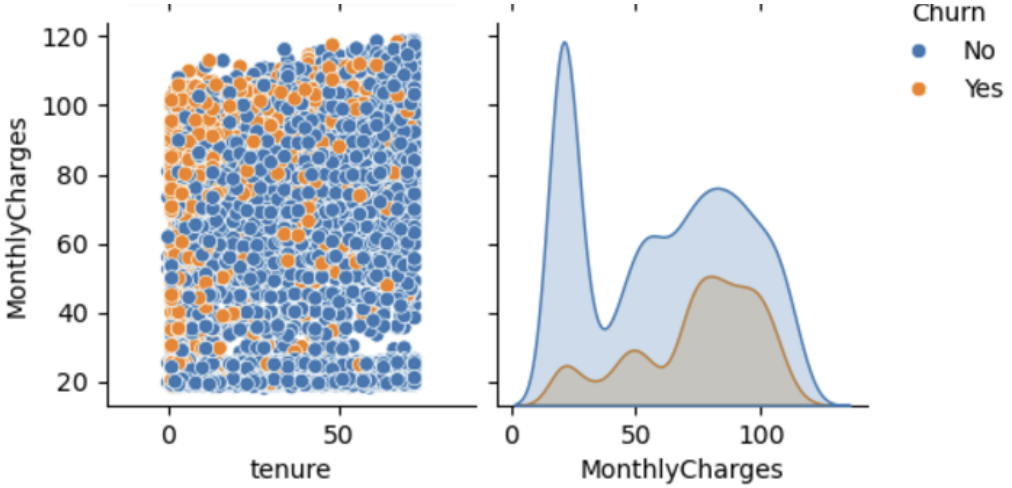
Graph I:



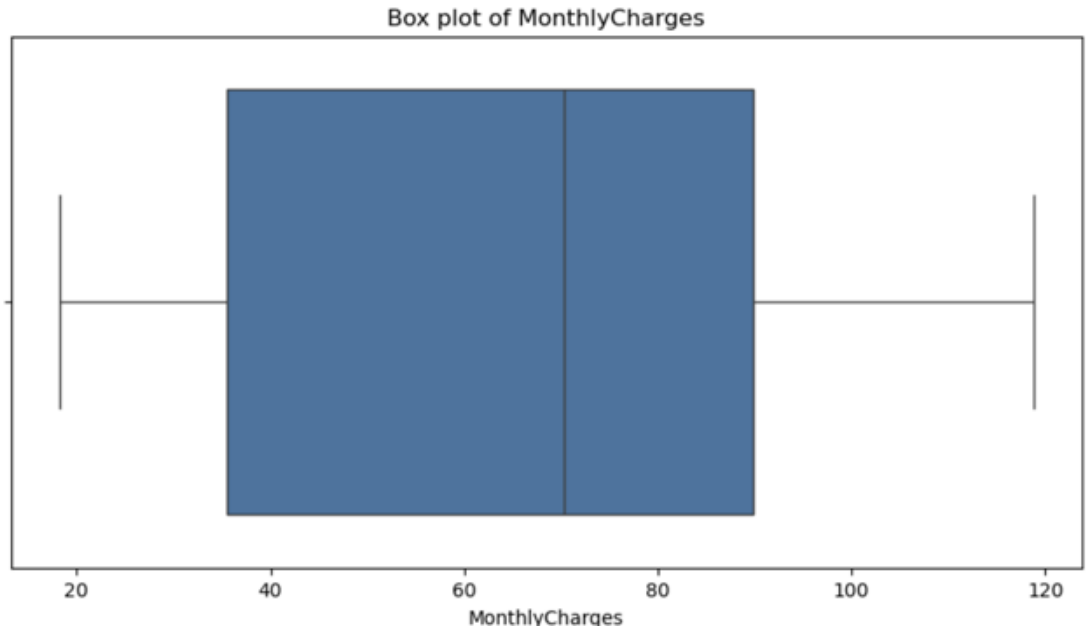
Graph II:



Graph III & Graph IV:



Graph V:



B. Data Pre-processing

Data preparation is an important step in the data analysis process, which makes sure the dataset is consistent, clean, and ready for analysis. In this process, we handled the missing values, categorical variables are encoded, numerical characteristics are scaled, and any data inconsistencies are addressed. This builds strong and accurate prediction models requires effective preprocessing since it helps to remove potential biases and mistakes that could affect the study.

Number of rows and Number columns make up our dataset, which includes both categorical and numerical variables. While the categories columns include elements like gender, partner status, and internet service type, the numerical columns contain attributes like tenure, monthly charges, and total charges. The first step in getting ready for preprocessing is to comprehend the composition and structure of the dataset.
In data analysis, missing values can provide serious difficulties that, if left unchecked, might result in biases and errors. To maintain the data's central tendency, the mean was used to fill in the missing values for numerical columns. The most prevalent categories

in the dataset were preserved by using the mode to fill in the missing values for categorical columns. This method guarantees that all missing data is dealt with properly, preserving the integrity of the dataset.

The required data type conversions were carried out to guarantee accuracy and consistency in the analysis. For example, the Total Charges column was changed from an object to a numeric data type. In order to perform statistical analysis and mathematical calculations on the column, this conversion is essential. During analysis, errors and inconsistencies can be avoided by making sure that each column contains the correct data type.

Because duplicate rows give repeated observations too much weight, they can skew the outcomes of data analysis. In order to fix this, duplicate entries were eliminated from the dataset once it was examined for them. By ensuring that every observation in the dataset is distinct, this phase contributes to a more accurate representation of the data.

For machine learning algorithms to employ categorical variables, they must be converted into numerical values. This was accomplished by converting categorical data into an analysis-ready format using one-hot encoding or label encoding. For example, multi-category variables like internet service type were converted into several binary columns, whereas binary variables like gender were encoded with values 0 and 1.

For algorithms that are sensitive to the size of the data, like gradient descent-based techniques, scaling numerical features guarantees that all variables are on a similar scale. The numerical columns were standardized using Standards Caler to a common scale with a standard deviation of one and a mean of zero. The accuracy and rate of convergence of machine learning models are enhanced by this step.

The dataset was separated into training and testing sets in order to assess the predictive model's performance. The model was trained on the training set, and its performance was assessed on the testing set. This method offers a more accurate evaluation of the model's prediction power and helps avoid overfitting. To ensure there was enough data for both training and evaluation, the data was divided in an 80-20 ratio.

```
# Identify numerical and categorical columns
numeric_cols = data.select_dtypes(include=['int64', 'float64']).columns.tolist()
categorical_cols = ['customerID', 'gender', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines',
                    'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport',
                    'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'Churn']

# Convert 'TotalCharges' to numeric, coercing errors
data['TotalCharges'] = pd.to_numeric(data['TotalCharges'], errors='coerce')

# Handle missing values for numerical columns
data[numeric_cols] = data[numeric_cols].fillna(data[numeric_cols].mean())

# Handle missing values for categorical columns
for col in categorical_cols:
    data[col] = data[col].fillna(data[col].mode()[0])
```

```
data['TotalCharges'] = pd.to_numeric(data['TotalCharges'], errors='coerce')
```

```
# Checking for duplicates
duplicates = data.duplicated().sum()
print(f'Number of duplicate rows: {duplicates}')

# Dropping duplicates
data = data.drop_duplicates()
```

```
Number of duplicate rows: 0
```

```
# Convert categorical columns to 'category' data type
for col in categorical_cols:
    data[col] = data[col].astype('category')

# Example of one-hot encoding
data = pd.get_dummies(data, columns=categorical_cols, drop_first=True)
```

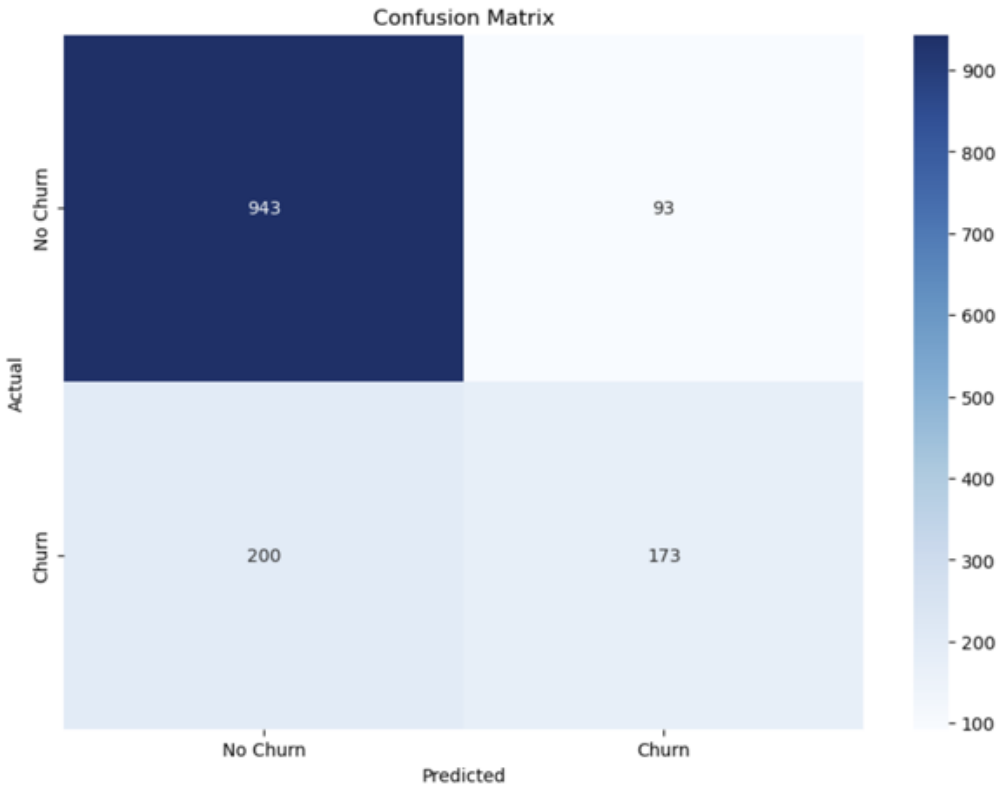
```
from sklearn.model_selection import train_test_split

X = data.drop('Churn_Yes', axis=1)
y = data['Churn_Yes']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Predictive Analysis

Confusion Matrix

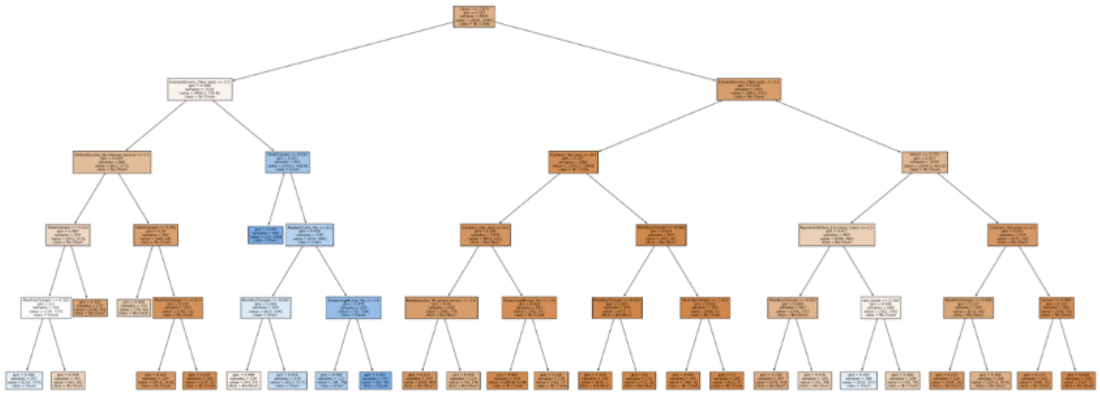
The graph below shows a confusion matrix for a classification model predicting customer churn. The matrix on the vertical axis (No Churn and Churn) and predicted classes on the horizontal axis (No Churn and Churn). The colour intensity of the cells represents the count of predictions, with darker colours indicating higher counts. The confusion matrix provides a detailed breakdown of the model's performance, showing how many instances were correctly or incorrectly classified. This information is crucial for understanding the model's accuracy, precision, recall, and overall effectiveness in predicting customer churn.

The model evaluation results indicate an accuracy of 79.2%, showing that the model correctly predicts customer churn. The classification report highlights precision, recall. For non-churned customers (0), the model achieves a high precision (0.83), recall (0.91). For churned customers (1), the precision is 0.65, recall is 0.46 indicating moderate performance. The confusion matrix further shows the model's performance with 943 true negatives, 93 false positives, 200 false negatives, and 173 true positives. These metrics suggest that even if the model performs well overall, it still has room for improvement specially when predicting churned customers.



Decision Tree

The classification report for the decision tree model reveals an accuracy of 77.6%, indicating that the model correctly predicts customer churn in approximately 77.6% of cases. For non-churned customers, the model shows high precision (0.82) and recall (0.88). However, the model's performance in predicting churned customers is moderate, with a precision of 0.61, recall of 0.48. The macro average and weighted average metrics also reflect this discrepancy, with overall precision, recall. These results suggest that while the decision tree effectively predicts non-churned customers, it has room for improvement in accurately identifying those at risk of churn. This insight is crucial for refining the model and enhancing its predictive capabilities.



```
Accuracy: 0.7756743965925225
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.88      0.85      1539
           1       0.61      0.48      0.54       574

    accuracy                           0.78      2113
   macro avg       0.72      0.68      0.70      2113
weighted avg       0.76      0.78      0.77      2113
```

Logistic Regression

Logistic regression is a data analysis technique that uses mathematical methods to uncover relationships between two factors and predict the value of one factor based on the other. This makes logistic regression particularly suitable for binary classification problems, such as the Telco customer churn case.

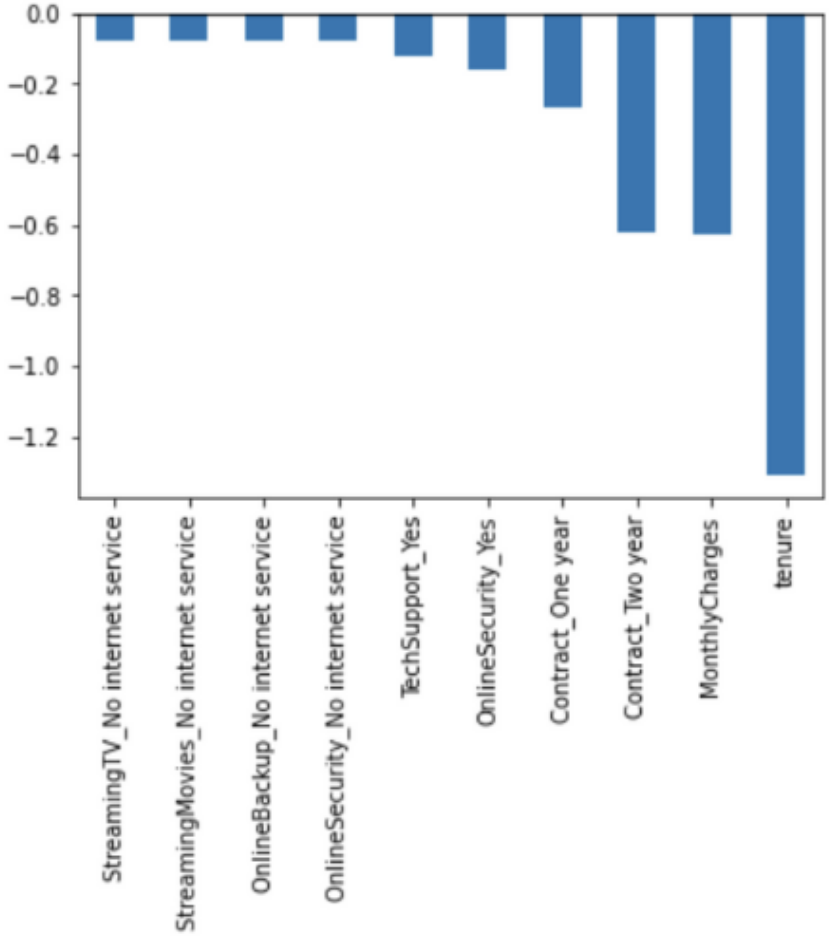
In this context, the target variable is "churn." Before applying the logistic regression model, several preprocessing steps are necessary. First, we identify categorical data and encode these categorical variables. Next, we address any null values and drop columns that are less relevant to the analysis.

```
Accuracy: 0.8204400283889283
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.90      0.88      1036
           1       0.69      0.60      0.64       373

    accuracy                           0.82      1409
   macro avg       0.77      0.75      0.76      1409
weighted avg       0.81      0.82      0.82      1409
```

The results of the logistic regression model show an accuracy of 0.8204, indicating that 82% of the samples were correctly classified. In the Classification Report, the precision for class 0 (non-churn) is 0.86, while the precision for class 1 (churn) is only 0.69. This indicates that only 69% of the instances classified as positive (churn) are true positives. Regarding recall, class 0 has a recall of 0.90, but class 1 shows that only 60% of positive instances were correctly identified.

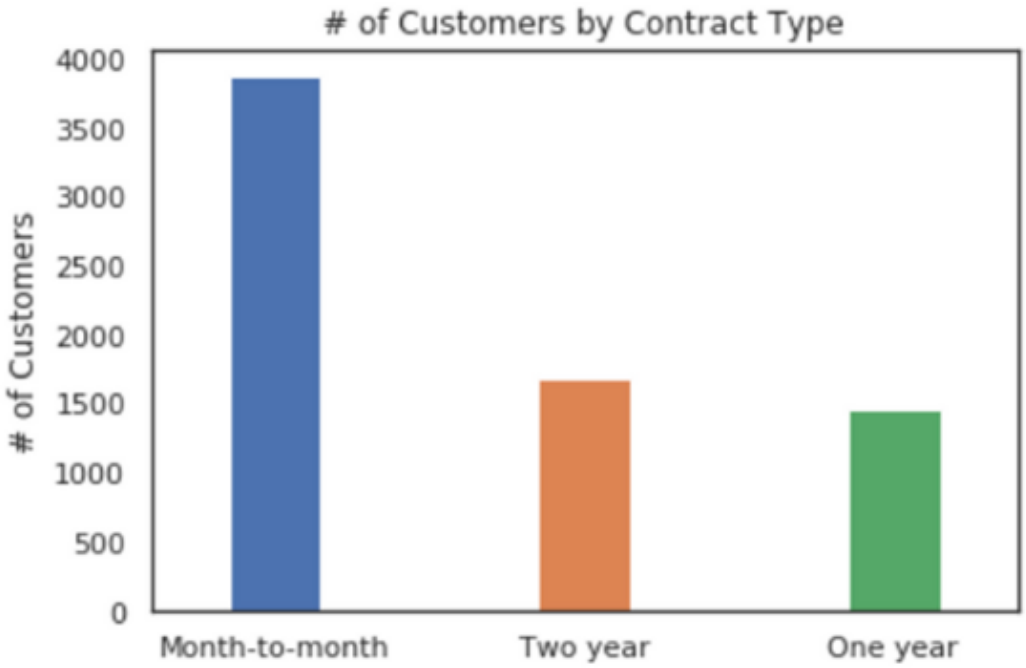
Overall, the data suggests that the logistic regression model demonstrates reasonably good performance, but there is still a need for improvement in predicting the positive class (churn).

Finding and recommendation



From these two figures, we can find that higher fibre optic internet services and total charges can lead to higher churn rates. At the same time, monthly charges, longer contracts, and tenure can lead to a lower churn rate.
Therefore, we have proposed two suggestions to the company.

1. Encourage Longer Contracts



Reason: Offering longer contract terms leads to lower churn rates.

Importance: The figure shows that more than 40% of the current customers are on a month-to-month contract, which means that there is a significant opportunity for the telecom company to convert these customers into long-term contract holders, potentially reducing the overall churn rate.

Correlation: Customers who commit for an extended period may feel a sense of obligation to stay with the service and loyalty. Also, longer contracts can provide customers with predictable pricing, reducing anxiety over potential rate increases.

2. Lower the Total Charges

Reason: High total fees lead to higher churn rates.

Correlation: Many consumers are sensitive to price, especially in competitive markets. When faced with high costs, they are more likely to explore other service providers that may offer more favorable pricing. Also, in terms of economic uncertainty, customers may become more cautious about their spending and not be willing to allocate a large portion of their budget to telecom services as this may limit the available funds they could use in emergencies.

V. Conclusion

To sum up, we have tested three different models, including Confusion Matrix, decision tree, and logistic Regression. Below is their performance.

Confusion Matrix	Decision tree	Logistic Regression
Precision (1):0.69	Precision (1):0.58	Precision (1):0.69
Recall (1):0.58	Recall (1):0.47	Recall (1):0.60
Accuracy:0.79	Accuracy:0.77	Accuracy:0.82

As we can see, the logistic regression has got the highest in all three criteria, therefore we would choose it as the best prediction model for churn. Also, we would recommend the company to encourage the customer to sign longer Contracts and lower the total charges in the contract to lower the churn rate.

Reference:

1. Data Preprocessing in Data Mining (Sep 2024)

Data Preprocessing in Data Mining - GeeksforGeeks

2. Sze Zhong LIM(Apr 06, 2024) Mastering Exploratory Data Analysis (EDA)

Mastering Exploratory Data Analysis (EDA): Everything You Need To Know | by Sze Zhong LIM | Data And Beyond | Medium