MINISTRY OF EDUCATION AND TRAINING

EASTERN INTERNATIONAL UNIVERSITY

MIS 395

Artificial Intelligence for Business

Final Exam

Lecturers: ***Mr. Dang Thai Doan***

Name	IRN
Lê Nguyễn Tâm Như	2132300065

Quarter 4/2024-2025

Date Submission: 21/8/2025

Table of content

Question 1: 3
Question 2: 3
Question 3: 8
Question 4: 13
Question 5: 14
Question 6: 15

Github Link: https://github.com/Le-Nguyen-Tam-Nhu/MIS-395/tree/main/Final%20Exam_MIS%20395_2132300065_Le%20Nguyen%20Tam%20Nhu

Question 1:

Business Problem:

The business problem that can be addressed using this dataset is predicting customer churn. Churn refers to customers who cancel or do not renew their subscriptions to the telecom service. The company wants to understand the factors that lead to customer churn and predict which customers are likely to churn in the future.

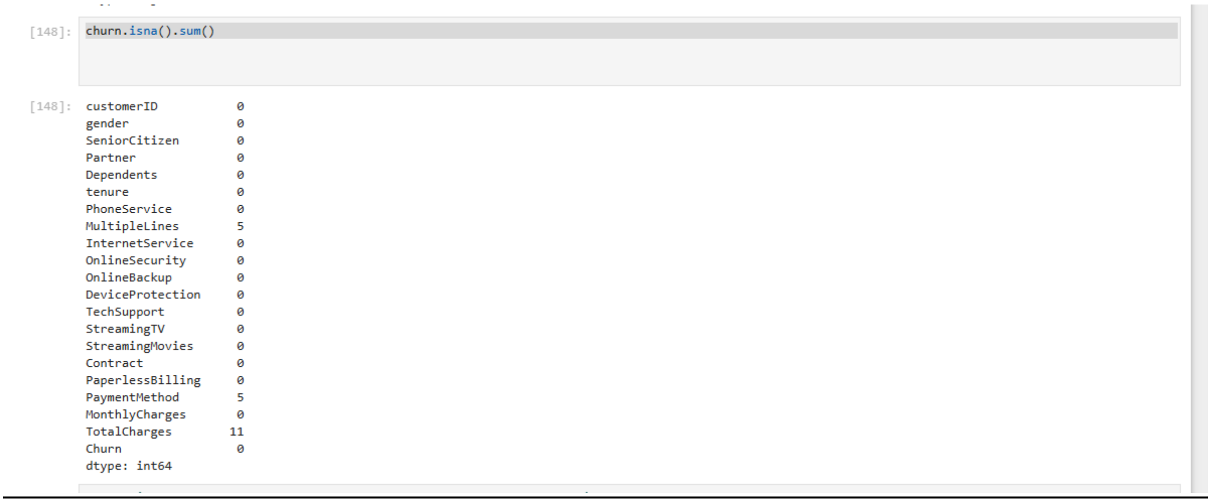
Model Choice:

I would choose a Classification model because the target variable, Churn, is categorical (with values Yes or No). We are trying to predict whether a customer will churn or not, which is a classification problem.

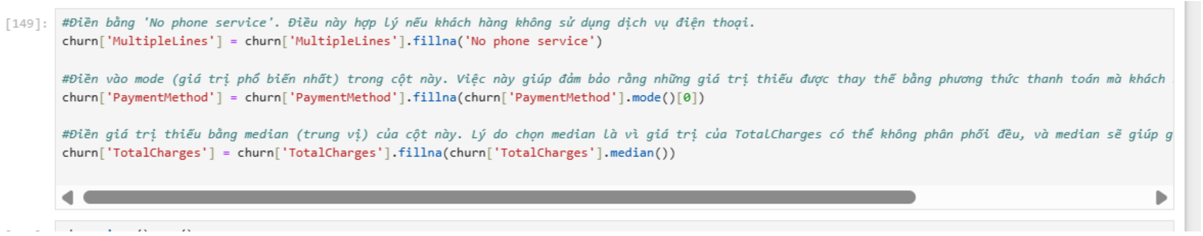
The goal is to identify patterns in customer behavior such as contract type, monthly charges, tenure, payment method, and other features to classify customers into two categories: those who will churn and those who will stay.

Question 2:

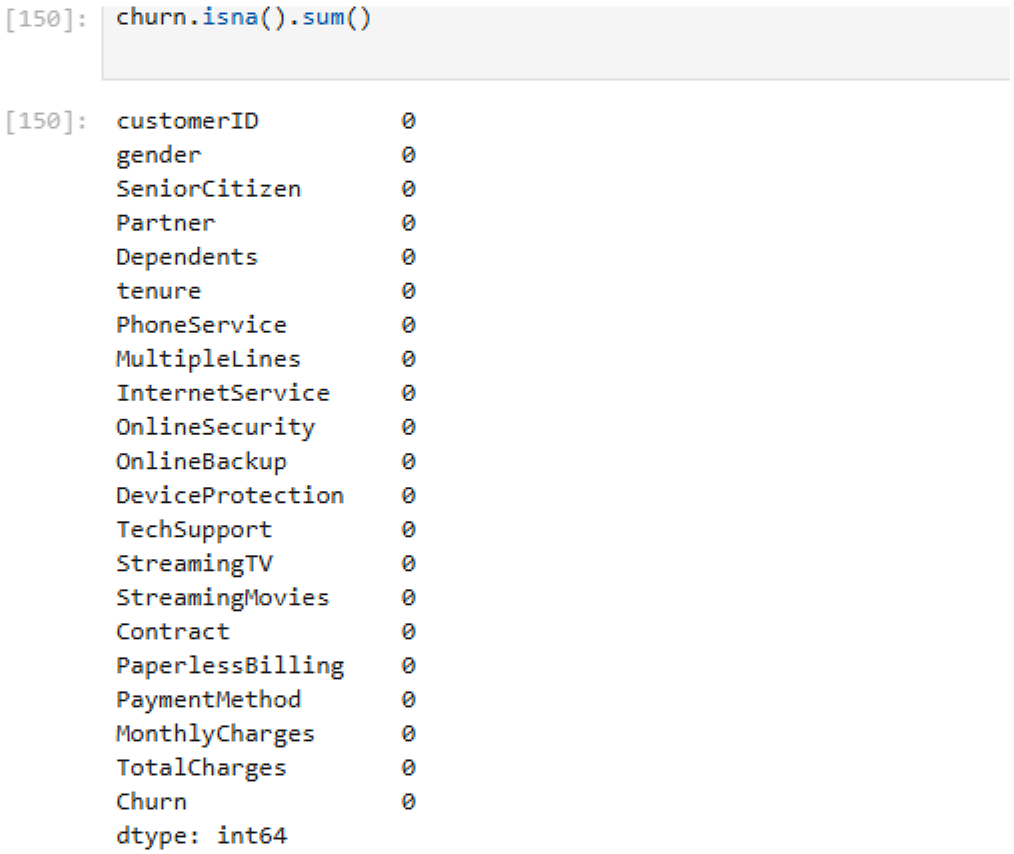
1. Handling Missing Values



```
[148]: churn.isna().sum()

[148]: customerID           0
       gender               0
       SeniorCitizen        0
       Partner              0
       Dependents           0
       tenure               0
       PhoneService         0
       MultipleLines        5
       InternetService      0
       OnlineSecurity       0
       OnlineBackup         0
       DeviceProtection     0
       TechSupport          0
       StreamingTV          0
       StreamingMovies      0
       Contract             0
       PaperlessBilling     0
       PaymentMethod        5
       MonthlyCharges       0
       TotalCharges        11
       Churn                0
       dtype: int64
```



```
[149]: #Điền bằng 'No phone service'. Điều này hợp lý nếu khách hàng không sử dụng dịch vụ điện thoại.
       churn['MultipleLines'] = churn['MultipleLines'].fillna('No phone service')

       #Điền vào mode (giá trị phổ biến nhất) trong cột này. Việc này giúp đảm bảo rằng những giá trị thiếu được thay thế bằng phương thức thanh toán mà khách
       churn['PaymentMethod'] = churn['PaymentMethod'].fillna(churn['PaymentMethod'].mode()[0])

       #Điền giá trị thiếu bằng median (trung vị) của cột này. Lý do chọn median là vì giá trị của TotalCharges có thể không phân phối đều, và median sẽ giúp g
       churn['TotalCharges'] = churn['TotalCharges'].fillna(churn['TotalCharges'].median())
```



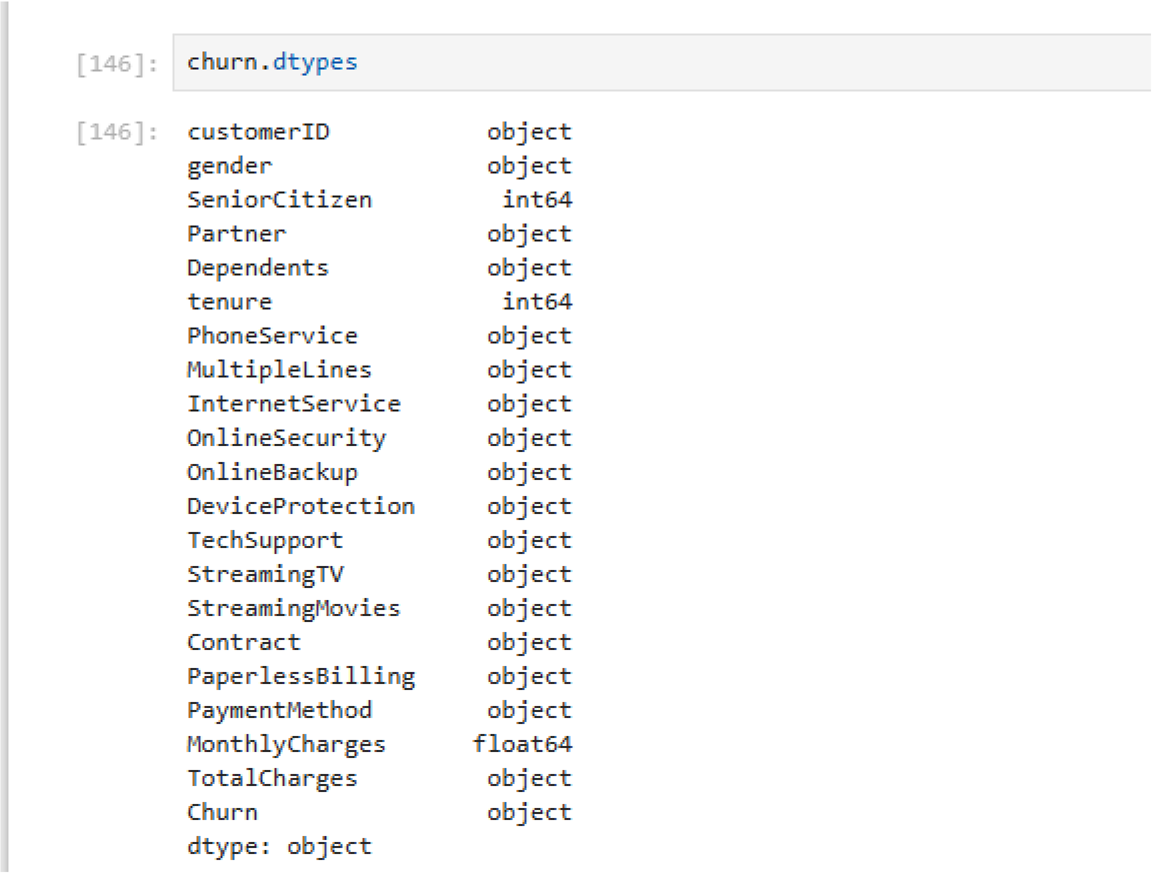
```
[150]: churn.isna().sum()

[150]: customerID          0
       gender              0
       SeniorCitizen       0
       Partner             0
       Dependents          0
       tenure              0
       PhoneService        0
       MultipleLines       0
       InternetService     0
       OnlineSecurity      0
       OnlineBackup        0
       DeviceProtection    0
       TechSupport         0
       StreamingTV         0
       StreamingMovies     0
       Contract            0
       PaperlessBilling    0
       PaymentMethod       0
       MonthlyCharges      0
       TotalCharges        0
       Churn               0
       dtype: int64
```

How to do:

In this step, missing data is addressed by filling in missing values with appropriate default values. For example, in the **MultipleLines** column, missing values are replaced with "No phone service", assuming that customers without multiple lines likely don't have phone service. Similarly, in the **PaymentMethod** column, missing values are replaced with the most common payment method (mode). For **TotalCharges**, missing values are filled with the median value, which is useful when the distribution is skewed or when there are outliers.

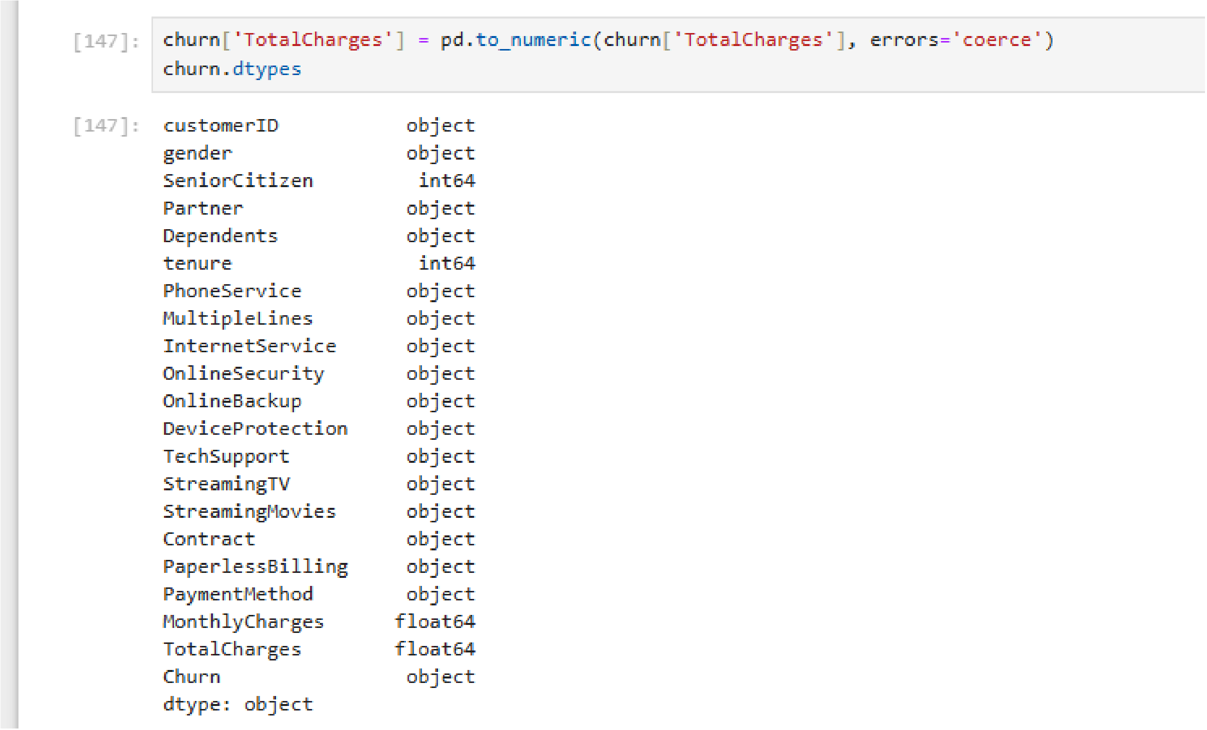
.

2. Data Type Conversion



```
[146]: churn.dtypes
```

```
[146]: customerID           object
       gender               object
       SeniorCitizen         int64
       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
```

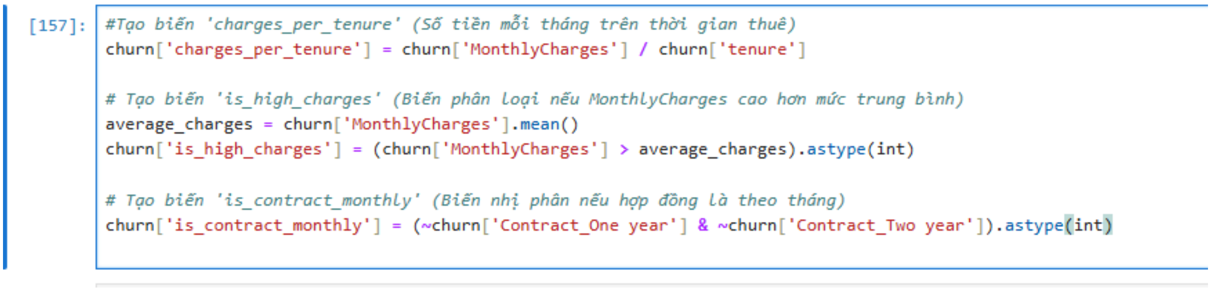


```
[147]: churn['TotalCharges'] = pd.to_numeric(churn['TotalCharges'], errors='coerce')
       churn.dtypes
```

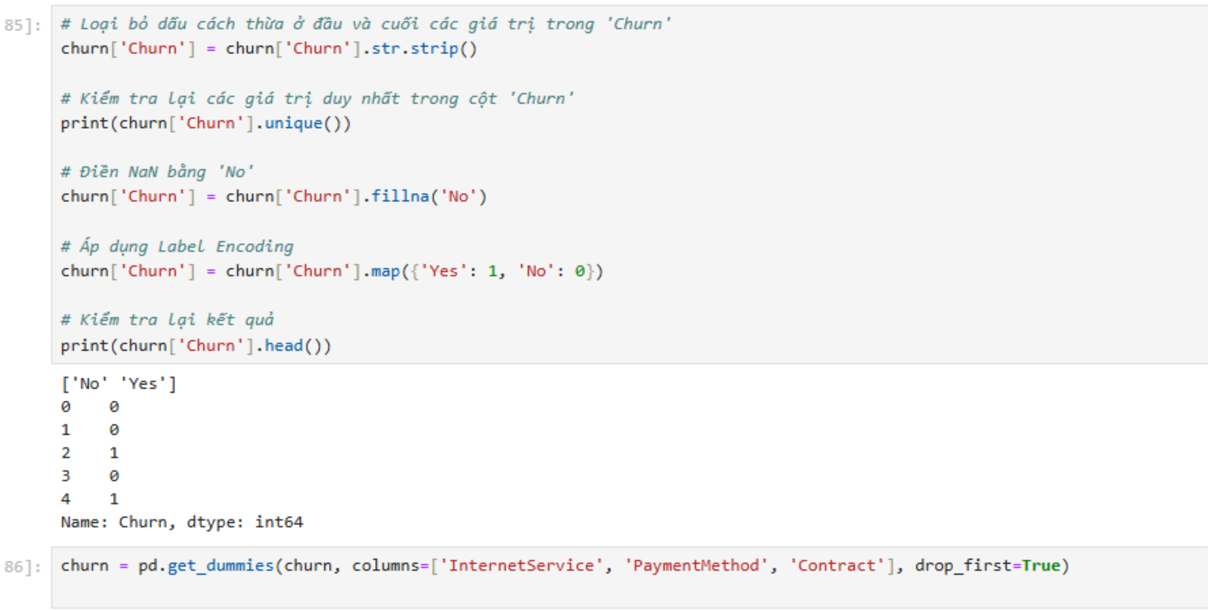
```
[147]: customerID           object
       gender               object
       SeniorCitizen         int64
       Partner              object
       Dependents           object
       tenure                int64
       PhoneService         object
       MultipleLines        object
       InternetService      object
       OnlineSecurity       object
       OnlineBackup         object
       DeviceProtection     object
       TechSupport          object
       StreamingTV          object
       StreamingMovies      object
       Contract             object
       PaperlessBilling     object
       PaymentMethod        object
       MonthlyCharges      float64
       TotalCharges        float64
       Churn                object
       dtype: object
```

Data type conversion ensures that each column is in the correct format for analysis. For example, the **TotalCharges** column, which may initially be treated as an object type, is converted to a numeric type (float64). This is important because machine learning algorithms require numerical values to process the data correctly. Converting the data types ensures that all features are in the appropriate format for the model to interpret.

3. Encoding Categorical Data & Feature Creation



```
[157]: #Tạo biến 'charges_per_tenure' (Số tiền mỗi tháng trên thời gian thuê)
       churn['charges_per_tenure'] = churn['MonthlyCharges'] / churn['tenure']

       # Tạo biến 'is_high_charges' (Biến phân loại nếu MonthlyCharges cao hơn mức trung bình)
       average_charges = churn['MonthlyCharges'].mean()
       churn['is_high_charges'] = (churn['MonthlyCharges'] > average_charges).astype(int)

       # Tạo biến 'is_contract_monthly' (Biến nhị phân nếu hợp đồng là theo tháng)
       churn['is_contract_monthly'] = (~churn['Contract_One year'] & ~churn['Contract_Two year']).astype(int)
```



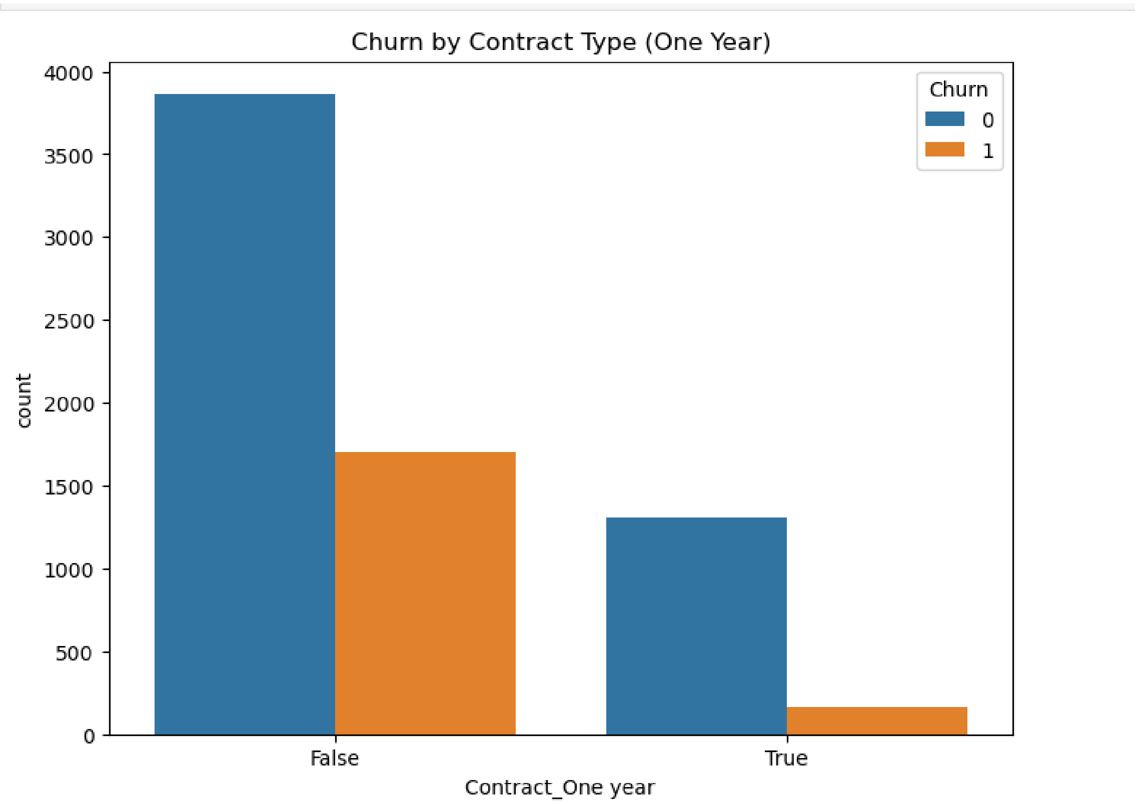
```
85]: # Loại bỏ dấu cách thừa ở đầu và cuối các giá trị trong 'Churn'
     churn['Churn'] = churn['Churn'].str.strip()

     # Kiểm tra lại các giá trị duy nhất trong cột 'Churn'
     print(churn['Churn'].unique())

     # Điền NaN bằng 'No'
     churn['Churn'] = churn['Churn'].fillna('No')

     # Áp dụng Label Encoding
     churn['Churn'] = churn['Churn'].map({'Yes': 1, 'No': 0})

     # Kiểm tra lại kết quả
     print(churn['Churn'].head())
```

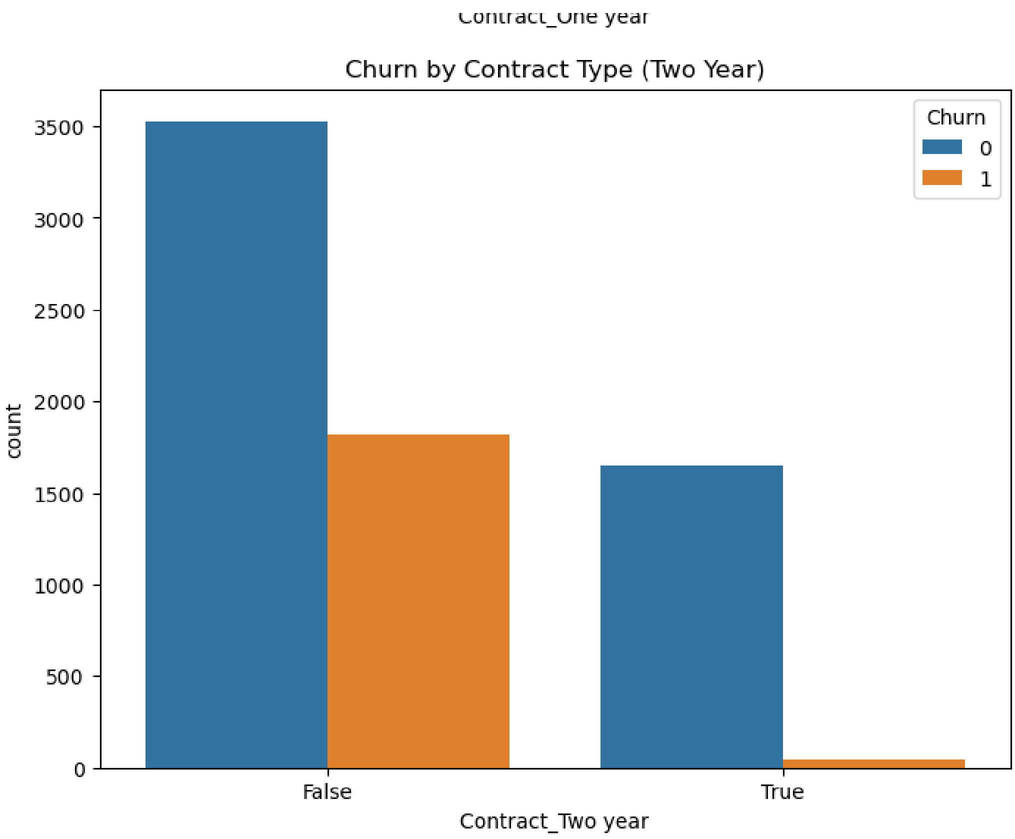
```
['No' 'Yes']
0    0
1    0
2    1
3    0
4    1
Name: Churn, dtype: int64
```

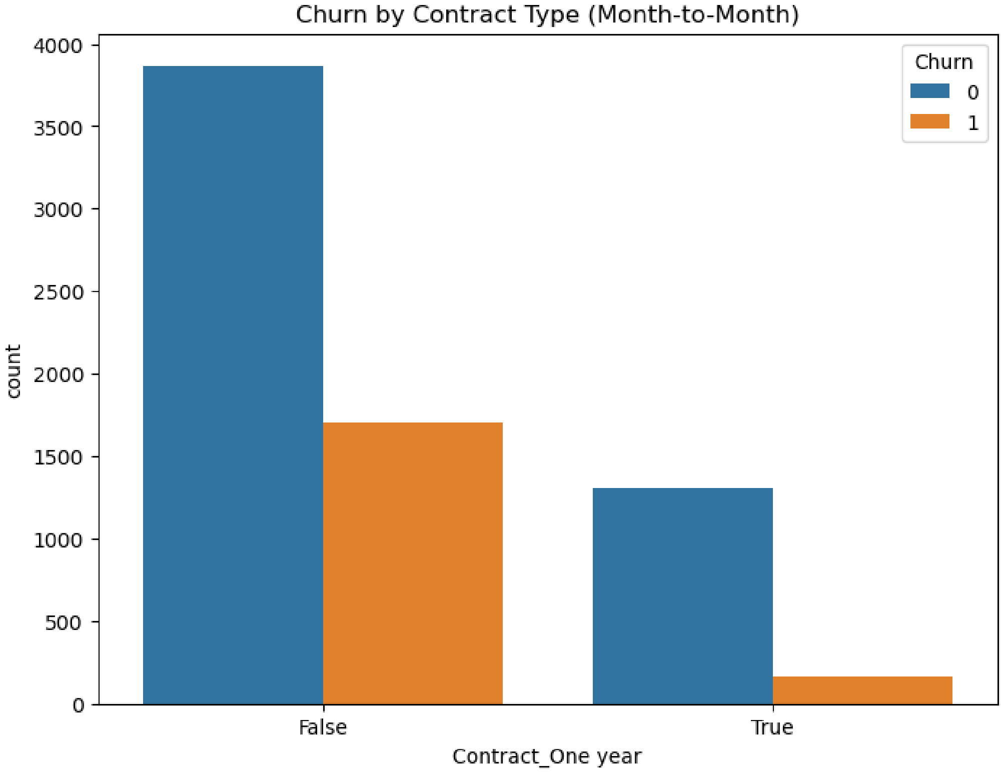
```
86]: churn = pd.get_dummies(churn, columns=['InternetService', 'PaymentMethod', 'Contract'], drop_first=True)
```

Categorical data is transformed into a format that machine learning algorithms can work with. **Label encoding** is used for the **Churn** column, where categorical values ("Yes" and "No") are mapped to numeric values (1 and 0). **One-hot encoding** is applied to columns like **InternetService**, **PaymentMethod**, and **Contract**, creating binary columns for each category. Additionally, **feature creation** is applied to generate new variables that could improve model performance. For example, features like **charges_per_tenure** and **is_high_charges** are created to capture relationships between customer charges, tenure, and churn, which could provide valuable insights for the predictive model

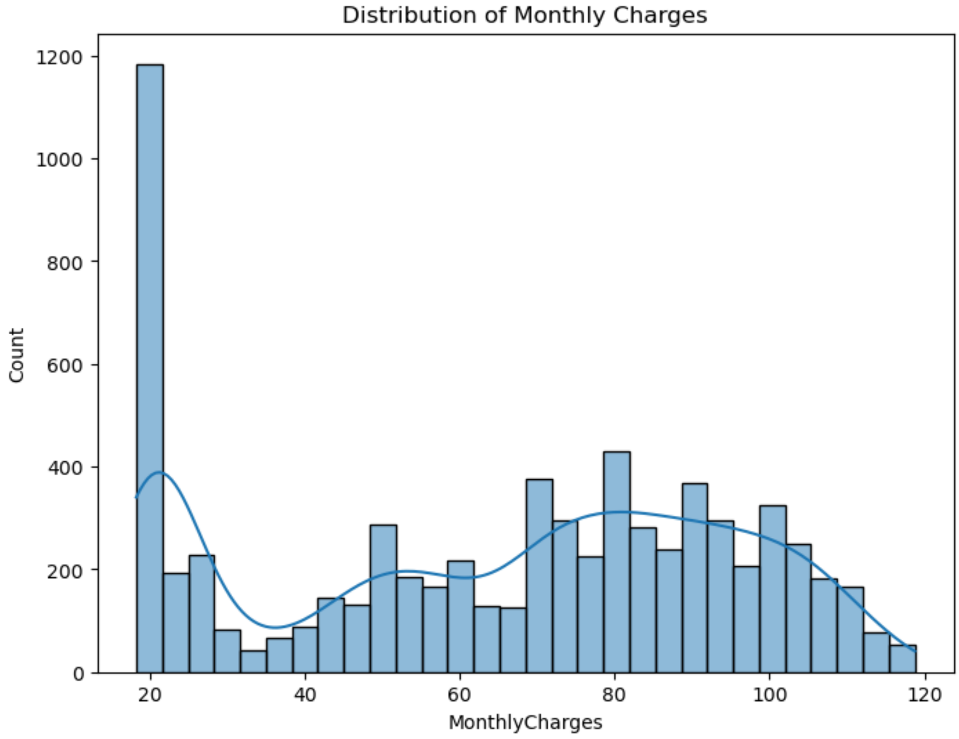
Question 3:


Churn by Contract Type (One Year)
Churn
0
1
4000
3500
3000
2500
2000
1500
1000
500
0
count
False
True
Contract_One year

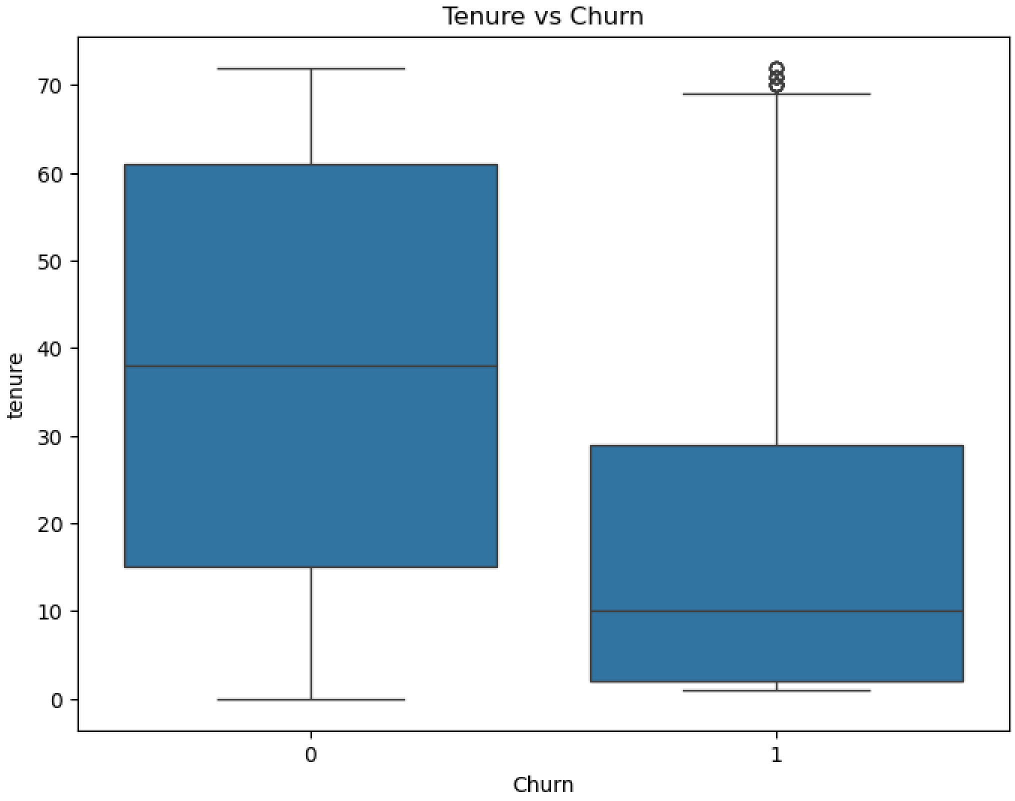

Contract_One year
Churn by Contract Type (Two Year)
Churn
0
1
3500
3000
2500
2000
1500
1000
500
0
count
False
True
Contract_Two year

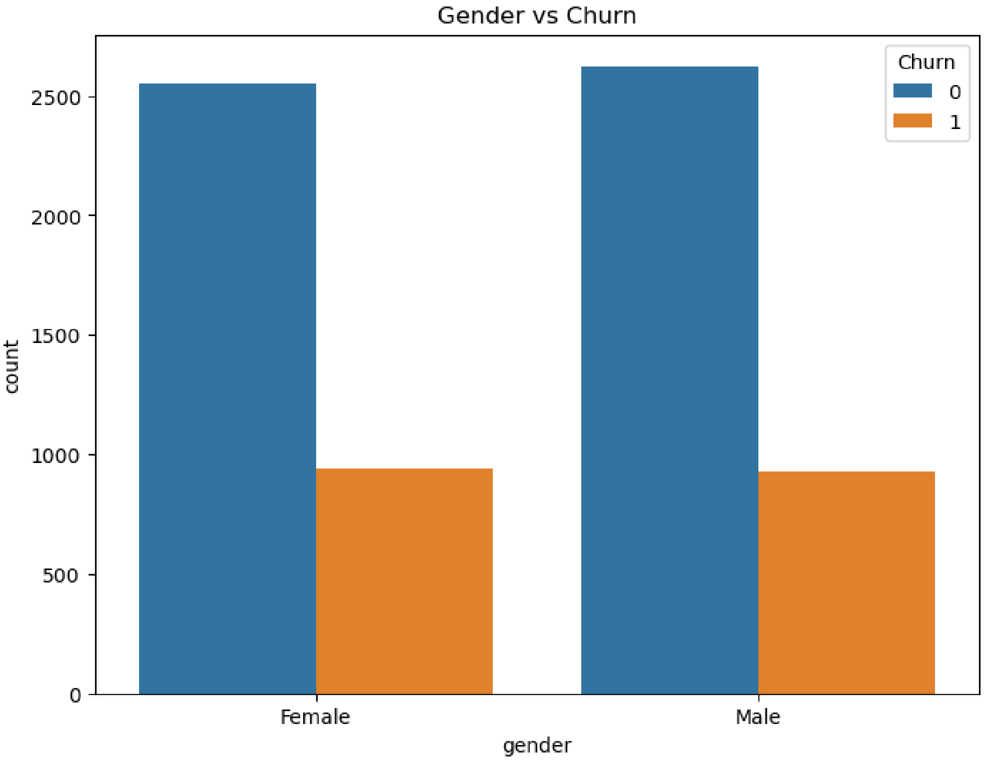

Churn by Contract Type (Month-to-Month)
Churn
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
False
True
Contract_One year

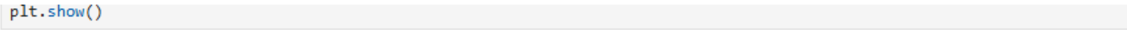
```
plt.show()
```


Distribution of Monthly Charges
Count
1200
1000
800
600
400
200
0
20
40
60
80
100
120
MonthlyCharges

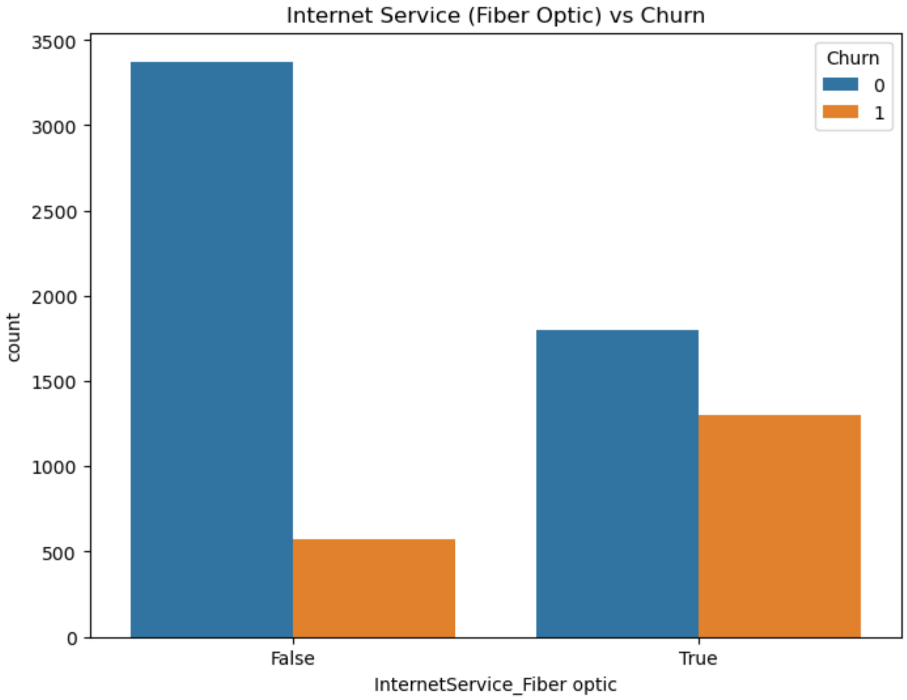
```
plt.show()
```


Tenure vs Churn
tenure
70
60
50
40
30
20
10
0
0
1
Churn


Gender vs Churn
Churn
0
1
count
2500
2000
1500
1000
500
0
Female
Male
gender



```
plt.show()
```



1. Churn by Contract Type (One Year)

Interpretation: Customers with a one-year contract are more likely to stay (indicated by the high count of churn = 0 for the "False" category). However, a smaller proportion of customers with one-year contracts have churned (churn = 1, indicated by the orange bars). This suggests that customers who opt for longer-term contracts tend to stay longer.

2. Churn by Contract Type (Two Year)

Interpretation: Similar to the one-year contract, customers with a two-year contract show a high retention rate (most are in the "False" category, meaning they did not churn). However, the proportion of customers who churned remains small, indicating a strong correlation between longer contracts and customer retention.

3. Churn by Contract Type (Month-to-Month)

Interpretation: A high percentage of customers with month-to-month contracts churn (represented by the orange bars). This supports the idea that customers on month-to-month plans are more likely to leave the service. The lack of a long-term commitment seems to be a major factor in the decision to churn.

4. Distribution of Monthly Charges

Interpretation: There is a peak in the count of customers who have low monthly charges, with a sharp drop-off at higher charge levels. The distribution suggests that most customers are on the lower end of the pricing scale. The relationship between higher charges and churn could be explored further.

5. Tenure vs. Churn

Interpretation: Customers who have been with the service for a longer period (indicated by higher tenure values) are less likely to churn, as seen from the box plot. Customers who churn typically have a lower tenure, highlighting that customer retention tends to improve with longer service periods.

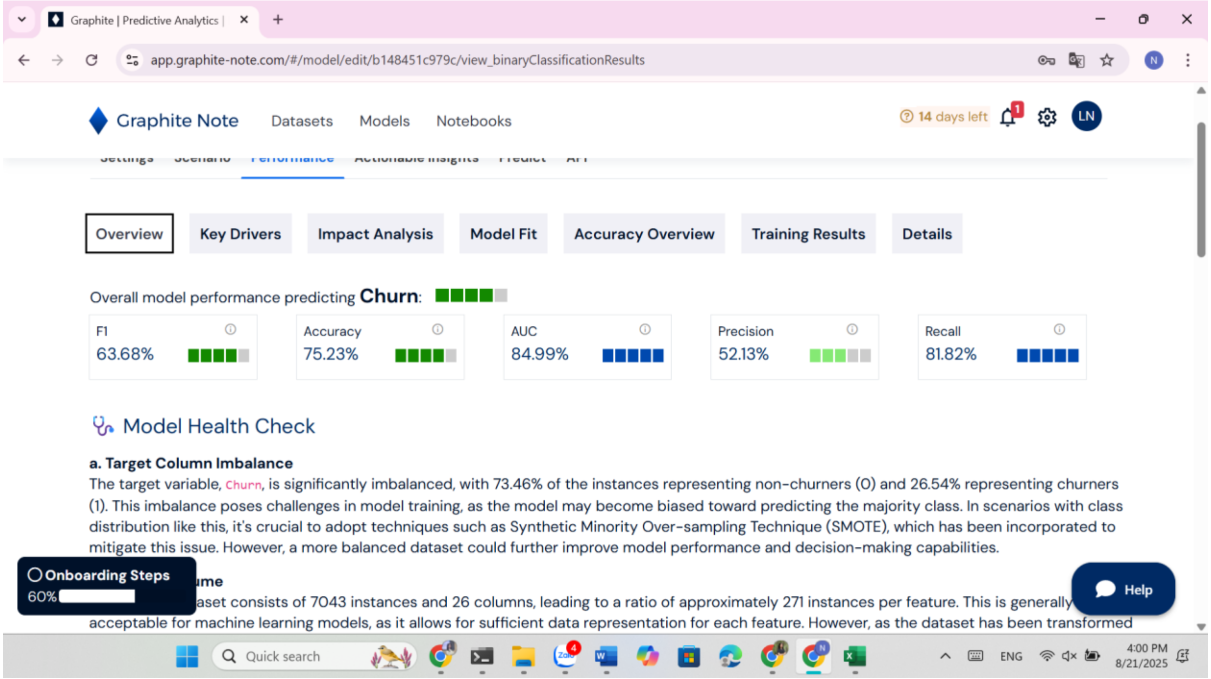
6. Gender vs. Churn

Interpretation: The number of male and female customers who churn is relatively balanced, with women showing a slightly higher rate of churn. However, the overall churn rate is still quite low compared to the number of customers who stay, regardless of gender.

7. Internet Service (Fiber Optic) vs. Churn

Interpretation: Customers with fiber optic internet service are more likely to stay (high number of churn = 0 for the "True" category). However, there is a noticeable number of churns in this group, suggesting that despite the perceived high-quality service, some fiber optic customers do leave.

Question 4:



Overall model performance predicting **Churn**: 🟩🟩🟩🟩

F1	Accuracy	AUC	Precision	Recall
63.68%	75.23%	84.99%	52.13%	81.82%

Model Health Check

a. Target Column Imbalance

The target variable, Churn, is significantly imbalanced, with 73.46% of the instances representing non-churners (0) and 26.54% representing churners (1). This imbalance poses challenges in model training, as the model may become biased toward predicting the majority class. In scenarios with class distribution like this, it's crucial to adopt techniques such as Synthetic Minority Over-sampling Technique (SMOTE), which has been incorporated to mitigate this issue. However, a more balanced dataset could further improve model performance and decision-making capabilities.

me

aset consists of 7043 instances and 26 columns, leading to a ratio of approximately 271 instances per feature. This is generally acceptable for machine learning models, as it allows for sufficient data representation for each feature. However, as the dataset has been transformed

1. What type of model did you build?

I built a **Binary Classification** model to predict customer churn, aiming to classify customers into two categories: those who will churn ("Yes") and those who will not churn ("No").

2. What algorithm did you use?

I used Graphite Note's machine learning tools, which likely employ algorithms such as **Logistic Regression**, **Random Forest**, or **Gradient Boosting** for binary classification, optimizing performance on imbalanced data.

3. Which features did you include in the model, and why?

I included the following features:

- **Contract Type**: Month-to-month contracts are linked to higher churn rates.
- **Monthly Charges**: Higher charges may correlate with increased churn.
- **Tenure**: Longer tenure typically reduces the likelihood of churn.
- **Internet Service**: Service type can impact customer satisfaction and churn.
- **Payment Method**: Certain payment methods could indicate higher or lower churn risk.

These features were chosen based on their relevance to customer retention.

4. What are the key predictions or patterns your model reveals?

The model shows:

- **Accuracy**: 75.23%, meaning the model is fairly effective in predicting churn.
- **Recall**: 81.82%, indicating good identification of churners.
- **Precision**: 52.13%, showing room for improvement in reducing false positives.
- **AUC**: 84.99%, indicating strong discriminatory power between churners and non-churners.

The model could be improved by addressing data imbalance using techniques like **SMOTE**.

Question 5:

To evaluate the model's performance, we used several key metrics:

- **Accuracy**: 75.23% - This metric indicates that the model correctly predicted whether a customer would churn or not in about 75% of the cases. While this is a solid result, there is still room for improvement.
- **Recall**: 81.82% - This shows that the model is good at identifying customers who will churn. The higher recall suggests fewer churners are missed, which is important for retaining at-risk customers.
- **Precision**: 52.13% - Precision indicates that when the model predicts a customer will churn, it is correct 52.13% of the time. This relatively low precision indicates the need for improvement in reducing false positives (predicting churn when the customer will actually stay).
- **AUC (Area Under the Curve)**: 84.99% - AUC measures the model's ability to distinguish between churners and non-churners. A high AUC score of 84.99% means that the model performs well in separating the two classes.

Suggestions for Improvement:

- **Address Data Imbalance**: Since the dataset is imbalanced (73.46% non-churners vs. 26.54% churners), applying **SMOTE** (Synthetic Minority Over-sampling Technique) to balance the classes could improve precision.
- **Feature Engineering**: Further refining features (e.g., adding interaction terms or using advanced feature selection) could enhance the model's predictive power, especially by reducing false positives.

Question 6:

Based on the analysis and predictive modeling, we can summarize the following key findings:

- **Key Findings**: Customers on month-to-month contracts and those with higher monthly charges are more likely to churn. Additionally, customers with longer tenures tend to have lower churn rates. By analyzing these patterns, we can identify at-risk customers and take proactive measures to retain them.
- **Trends and Risks**: The model shows that a significant portion of customers with shorter contracts (month-to-month) leave the service, which presents a potential risk. Additionally, the relationship between high charges and churn suggests that pricing strategies could influence retention.
- **Actionable Recommendations**:
    1. **Offer Incentives for Long-Term Contracts**: Encourage month-to-month customers to sign longer-term contracts by offering discounts or additional perks. This will likely reduce churn and improve customer retention.
    2. **Review Pricing Structure**: Consider adjusting monthly charges or providing special offers for high-paying customers who may be more price-sensitive. Lowering costs could reduce churn rates.

3. **Target Retention Campaigns**: Use the model to identify at-risk customers based on their contract type and charges, and offer personalized retention offers such as discounts or loyalty rewards.

Proposed AI/Analytics Application:
The company could explore **predictive analytics for customer segmentation**. By analyzing different customer segments, such as high-value vs. low-value customers, the company can tailor marketing strategies and improve service offerings to maximize retention and revenue.