Customer Churn Prediction

Problem Definition:

Customer churn prediction is a critical challenge in the realm of customer relationship management. The problem revolves around identifying and forecasting the likelihood of customers discontinuing their relationship with a business. By accurately defining this problem, organizations can proactively address the issue and take measures to retain valuable customers. The primary goal is to develop predictive models that can analyze historical customer data, such as purchase history, engagement metrics, and customer feedback, to predict when and why a customer might churn. This involves a multi-faceted approach that includes data preprocessing, feature engineering, and the application of machine learning algorithms to create a robust churn prediction system. An effective problem definition for customer churn prediction serves as the foundation for businesses to implement retention strategies, thereby reducing customer attrition and bolstering long-term profitability.

Customer churn prediction in IBM Cognos involves the critical task of identifying and addressing the factors leading to customer attrition within a business. The problem definition revolves around leveraging the advanced analytics capabilities of IBM Cognos to create predictive models that can accurately forecast which customers are likely to leave a company. By analyzing historical customer data, including their interactions, transactions, and behaviors, the goal is to proactively identify at-risk customers and take strategic actions to retain them. This problem definition also involves selecting the most relevant features and applying machine learning algorithms to develop models that can provide actionable insights for businesses to reduce customer churn, ultimately enhancing customer satisfaction and the bottom line. IBM Cognos serves as a powerful tool in this endeavor, enabling organizations to make data-driven decisions and design effective retention strategies.

Step 1: Define Analysis Objectives

The analysis objectives of customer churn prediction are multifaceted and critical for businesses aiming to retain their customer base. First and foremost, this analysis seeks to identify potential churners by assessing historical data, behavioral patterns, and key indicators. By understanding why customers leave, companies can formulate effective retention strategies.

Step 2: Data Collection:

Data collection is a crucial step in customer churn prediction, as the quality and quantity of data significantly impact the accuracy of predictive models. Several tools and methods are commonly used to collect data for customer churn prediction:

1. Customer Relationship Management (CRM) Systems: CRM systems like Salesforce, HubSpot, or Microsoft Dynamics store a wealth of customer data, including interactions, purchase history, and customer profiles. These platforms often have built-in reporting and data export features.

2. Data Warehousing and ETL (Extract, Transform, Load) Tools: Data warehouses like Amazon Redshift or ETL tools like Apache Nifi can be used to consolidate data from various sources, transform it into a consistent format, and load it into a database for analysis.

3. Database Management Systems (DBMS): DBMS such as MySQL, PostgreSQL, or Microsoft SQL Server are used to store and manage customer data. Queries can be run to extract relevant information for churn prediction.

4. Web Scraping Tools: When dealing with online businesses, web scraping tools like Scrapy or Beautiful Soup can be employed to gather data from websites, social media, and other online sources.

5. Surveys and Feedback Forms: Collecting direct feedback from customers through surveys, questionnaires, and feedback forms can provide valuable insights into their satisfaction levels and reasons for potential churn.

6. Social Media Analytics Tools: Tools like Hootsuite, Brandwatch, or Socialbakers allow you to monitor and collect data from social media platforms to gauge customer sentiment and reactions.

7. Email and Text Analytics: Analyzing email correspondence and text messages can reveal patterns or keywords indicative of churn intent. Tools like Python NLTK or commercial solutions like RapidMiner can be helpful.

8. Call Center Logs: Data from call center interactions, including call recordings and transcripts, can be used to understand customer issues, complaints, or dissatisfaction.

9. IoT Devices and Sensors: In industries like telecommunications or utilities, data from IoT devices and sensors can provide insights into network performance or equipment health, which may indirectly affect customer churn.

10. Third-party Data Providers: Companies may subscribe to data services that provide external data, such as economic indicators, weather data, or industry-specific information, which can be integrated into churn prediction models.

11. Mobile App and Website Analytics: Analyzing user behavior through tools like Google Analytics or Mixpanel can reveal user engagement and navigation patterns that might correlate with churn.

12. Data APIs: Accessing data through application programming interfaces (APIs) provided by third-party services, such as weather data, financial indicators, or social media APIs, can enrich your customer dataset.

13. Customer Feedback and Reviews: Online customer reviews, forums, and feedback platforms can be a valuable source of unstructured data for sentiment analysis.

14. Payment and Subscription Data: For subscription-based businesses, transaction data, payment history, and churn records are critical sources of information.

15. Market Research and Industry Reports: Industry-specific reports and market research data can offer insights into broader market trends and competitive analysis, helping in understanding the context of churn.

Effective data collection tools and methods should be selected based on the nature of the business, the availability of data sources, and the specific objectives of customer churn prediction. Combining data from multiple sources can provide a more comprehensive view of customer behavior and enhance the accuracy of predictive models.:

Step 3: Data Preprocessing

Data preprocessing is a crucial step in customer churn prediction that involves cleaning, transforming, and organizing the collected data to make it suitable for analysis and modeling. Here are key data preprocessing steps in customer churn prediction:

1. Data Cleaning:

- Handling missing values: Deal with missing data by imputing values or removing incomplete records.

- Outlier detection and treatment: Identify and address data points that significantly deviate from the norm.

2. Data Transformation:

- Feature scaling: Normalize or standardize numerical features to ensure they have a consistent scale.

- Encoding categorical variables: Convert categorical data (e.g., gender, product type) into numerical format using techniques like one-hot encoding.

- Feature engineering: Create new features or transform existing ones to extract more meaningful information (e.g., calculating customer tenure from signup date).

3. Feature Selection:

- Identify and select the most relevant features that contribute to predicting churn. This can reduce model complexity and improve performance.

4. Data Splitting:

- Split the data into training, validation, and test sets to evaluate the model's performance accurately.

5. Handling Class Imbalance:

- In many cases, the number of customers who churn is significantly smaller than those who don't. Techniques like oversampling, undersampling, or using synthetic data can address class imbalance issues.

6. Time Series Handling:

- If the data includes timestamps, consider time-related features and ensure that the dataset is sorted chronologically.

7. Normalization and Scaling:

- Standardize or normalize numerical features to ensure they have similar scales, preventing certain features from dominating the model.

9. Handling Text Data:

- If text data, such as customer feedback, is included, text preprocessing techniques like tokenization, stemming, or sentiment analysis may be applied.

10. Data Balancing:

- To address class imbalance, techniques like oversampling (creating more instances of the minority class), undersampling (reducing instances of the majority class), or using synthetic data (SMOTE) can be employed.

11. Data Splitting:

- Split the preprocessed data into training, validation, and test sets. This allows for model training, hyperparameter tuning, and final evaluation.

12**. Normalization and Scaling:**

- Standardize or normalize numerical features to ensure they have similar scales, preventing certain features from dominating the model.

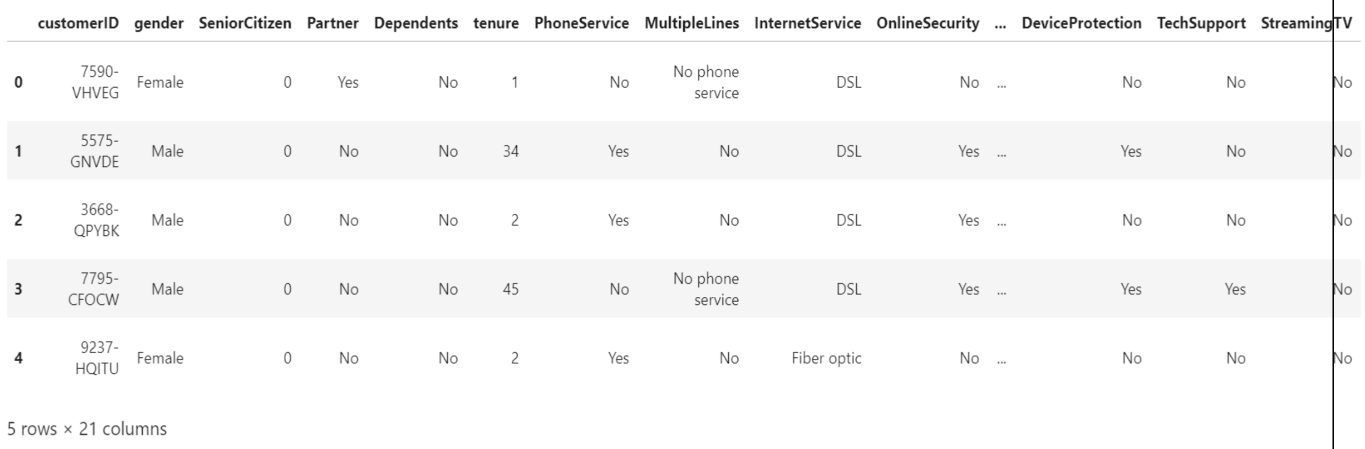
13**. Correlation Analysis:**

- Examine correlations between features to identify redundant or highly correlated variables that can be removed.

Data preprocessing ensures that the data used for customer churn prediction is in a suitable format for training and testing machine learning models. It can significantly impact the performance and accuracy of the predictive models.

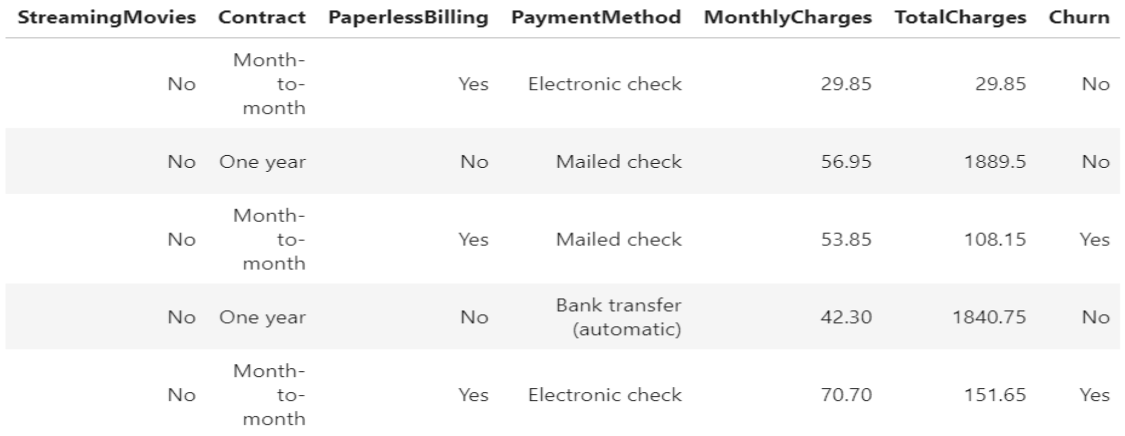
Code:

```
import numpy as np # linear algebra

import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

import seaborn as sns # For creating plots

import matplotlib.ticker as mtick # For specifying the axes tick format

import matplotlib.pyplot as plt

from sklearn.preprocessing import LabelEncoder

from sklearn.preprocessing import StandardScaler

sns.set(style = 'white')

# Input data files are available in the "../input/" directory.

import os

print(os.listdir("../input"))

# Any results you write to the current directory are saved as output.
```



	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	StreamingTV
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	No	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	No	No
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	No	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	Yes	No
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	No	No

5 rows × 21 columns



StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
No	Month-to-month	Yes	Electronic check	29.85	29.85	No
No	One year	No	Mailed check	56.95	1889.5	No
No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes

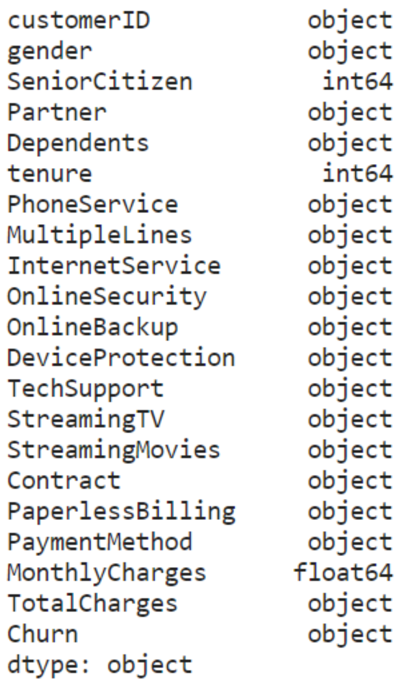
```
telecom_cust = pd.read_csv('/kaggle/input/customer-churn-prediction/WA_Fn-UseC_-Telco-Customer-Churn.csv')
```

```
telecom_cust.head()
```

```
telecom_cust.columns.values
```

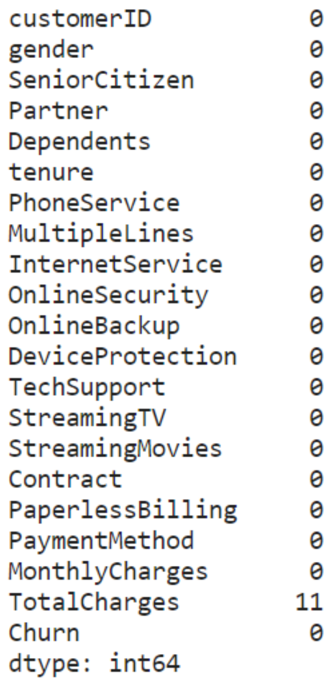
```
array(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents',
       'tenure', 'PhoneService', 'MultipleLines', 'InternetService',
       'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
       'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',
       'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges',
       'TotalCharges', 'Churn'], dtype=object)
```

```
telecom_cust.dtypes
```



```
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
```

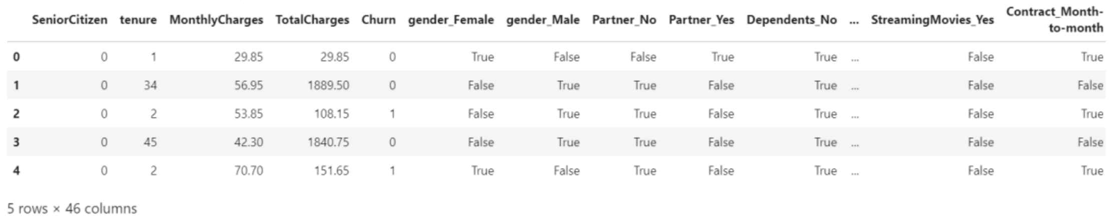
telecom_cust.isnull().sum()



```
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        11
Churn                0
dtype: int64
```

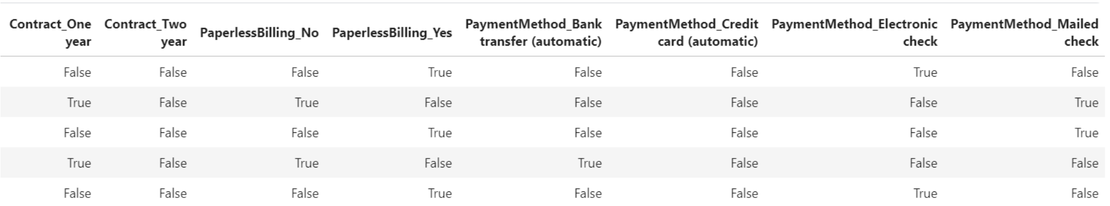
```
#Removing missing values
telecom_cust.dropna(inplace = True)


#Remove customer IDs from the data set
df2 = telecom_cust.iloc[:,1:]
#Convertin the predictor variable in a binary numeric variable
df2['Churn'].replace(to_replace='Yes', value=1, inplace=True)
df2['Churn'].replace(to_replace='No',  value=0, inplace=True)

#Let's convert all the categorical variables into dummy variables
```

```
df_dummies = pd.get_dummies(df2)
df_dummies.head()
```



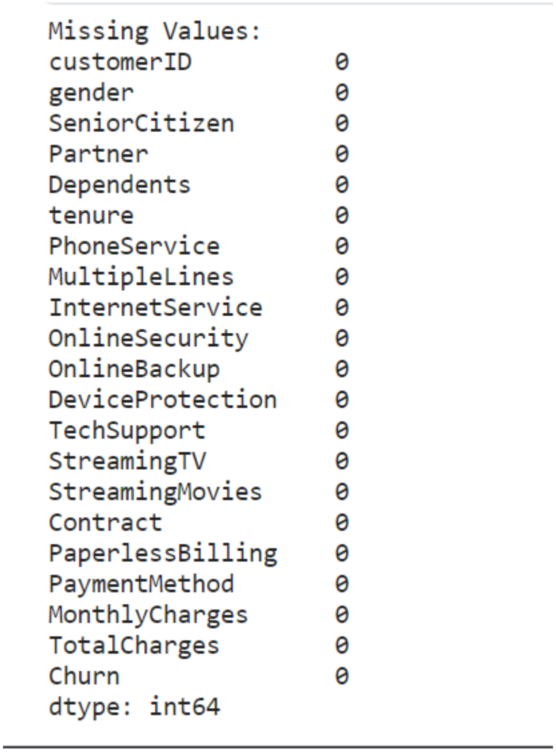
	SeniorCitizen	tenure	MonthlyCharges	TotalCharges	Churn	gender_Female	gender_Male	Partner_No	Partner_Yes	Dependents_No	...	StreamingMovies_Yes	Contract_Month-to-month
0	0	1	29.85	29.85	0	True	False	False	True	True	...	False	True
1	0	34	56.95	1889.50	0	False	True	True	False	True	...	False	False
2	0	2	53.85	108.15	1	False	True	True	False	True	...	False	True
3	0	45	42.30	1840.75	0	False	True	True	False	True	...	False	False
4	0	2	70.70	151.65	1	True	False	True	False	True	...	False	True

5 rows × 46 columns



Contract_One year	Contract_Two year	PaperlessBilling_No	PaperlessBilling_Yes	PaymentMethod_Bank transfer (automatic)	PaymentMethod_Credit card (automatic)	PaymentMethod_Electronic check	PaymentMethod_Mailed check
False	False	False	True	False	False	True	False
True	False	True	False	False	False	False	True
False	False	False	True	False	False	False	True
True	False	True	False	True	False	False	False
False	False	False	True	False	False	True	False

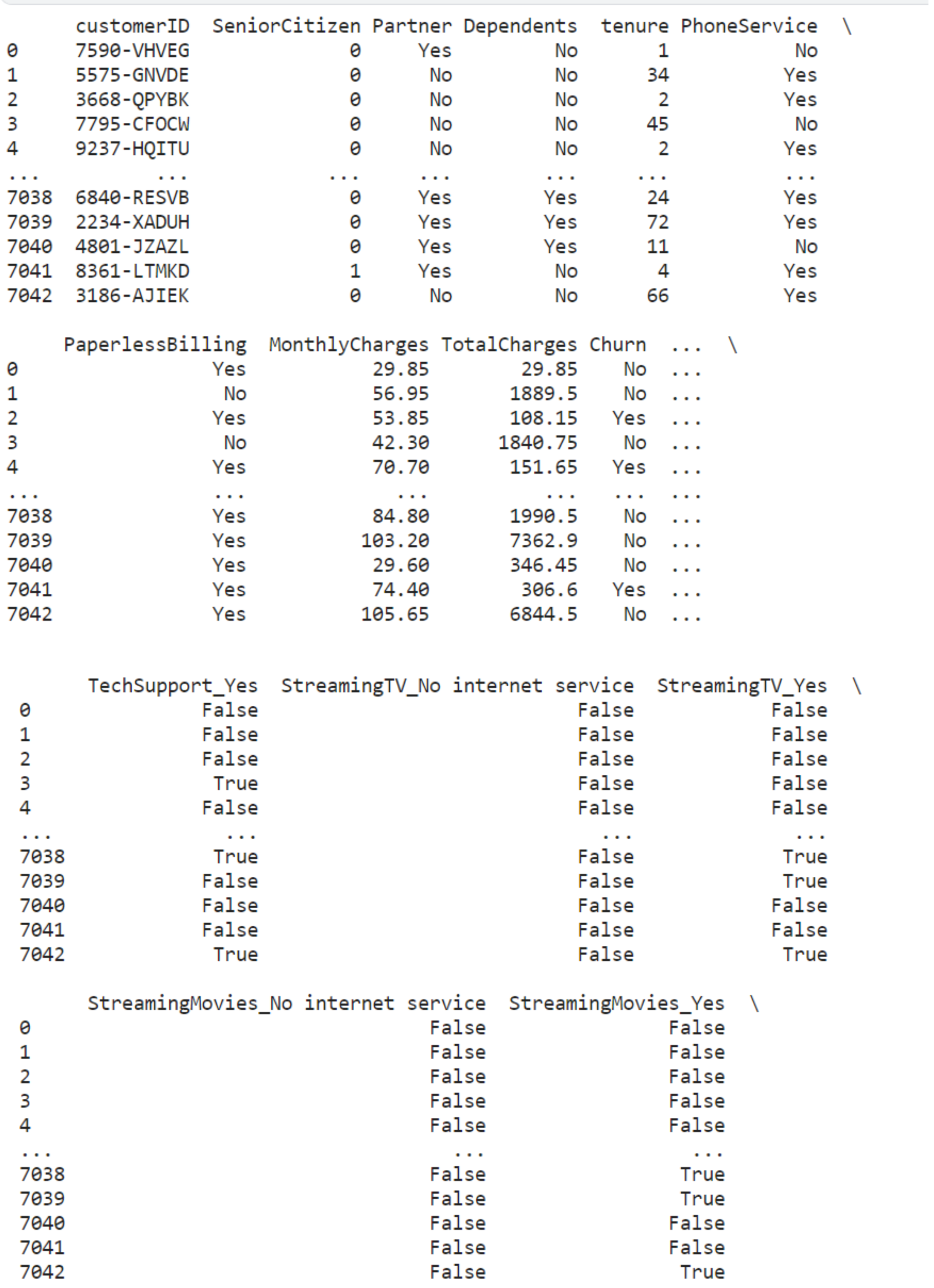
```
missing_values = telecom_cust.isnull().sum()
print("Missing Values:")
print(missing_values)
```



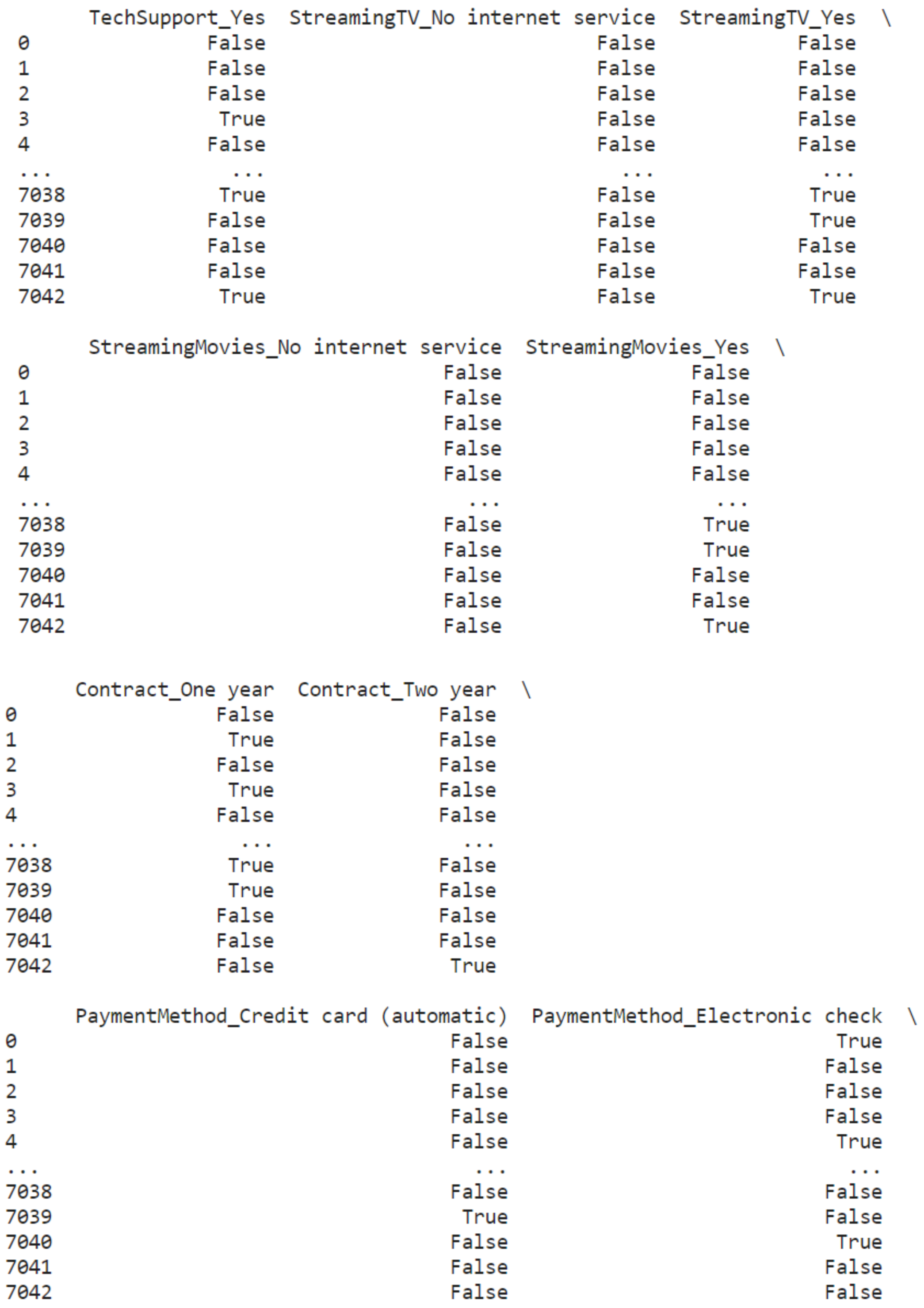
```
Missing Values:
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport     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

```
telecom_cust = pd.read_csv('/kaggle/input/customer-churn-prediction/WA_Fn-UseC_-Telco-Customer-Churn.csv')
```

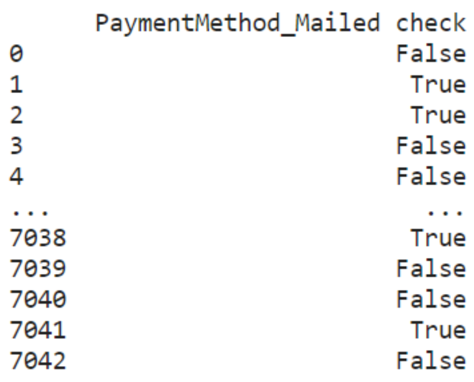
```
# Define the list of columns that need one-hot encoding
categorical_cols = ['gender', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup',
                    'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',
                 'PaymentMethod']

# Use get_dummies to one-hot encode the specified columns
telecom_cust = pd.get_dummies(telecom_cust, columns=categorical_cols, drop_first=True)

# Display the updated DataFrame
print(telecom_cust)
```



```
      customerID  SeniorCitizen Partner Dependents  tenure PhoneService  \
0     7590-VHVEG              0     Yes         No       1           No
1     5575-GNVDE              0      No         No      34          Yes
2     3668-QPYBK              0      No         No       2          Yes
3     7795-CFOCW              0      No         No      45           No
4     9237-HQITU              0      No         No       2          Yes
...          ...            ...     ...        ...     ...          ...
7038  6840-RESVB              0     Yes        Yes      24          Yes
7039  2234-XADUH              0     Yes        Yes      72          Yes
7040  4801-JZAZL              0     Yes        Yes      11           No
7041  8361-LTMKD              1     Yes         No       4          Yes
7042  3186-AJIEK              0      No         No      66          Yes

     PaperlessBilling  MonthlyCharges TotalCharges Churn  ...  \
0                 Yes           29.85        29.85    No  ...
1                  No           56.95       1889.5    No  ...
2                 Yes           53.85       108.15   Yes  ...
3                  No           42.30      1840.75    No  ...
4                 Yes           70.70       151.65   Yes  ...
...               ...             ...          ...   ...  ...
7038              Yes           84.80       1990.5    No  ...
7039              Yes          103.20       7362.9    No  ...
7040              Yes           29.60       346.45    No  ...
7041              Yes           74.40        306.6   Yes  ...
7042              Yes          105.65       6844.5    No  ...


      TechSupport_Yes  StreamingTV_No internet service  StreamingTV_Yes  \
0               False                            False            False
1               False                            False            False
2               False                            False            False
3                True                            False            False
4               False                            False            False
...               ...                              ...              ...
7038             True                            False             True
7039            False                            False             True
7040            False                            False            False
7041            False                            False            False
7042             True                            False             True

      StreamingMovies_No internet service  StreamingMovies_Yes  \
0                                   False                False
1                                   False                False
2                                   False                False
3                                   False                False
4                                   False                False
...                                   ...                  ...
7038                                False                 True
7039                                False                 True
7040                                False                False
7041                                False                False
7042                                False                 True
```



```
      TechSupport_Yes  StreamingTV_No internet service  StreamingTV_Yes  \
0               False                            False            False   
1               False                            False            False   
2               False                            False            False   
3                True                            False            False   
4               False                            False            False   
...               ...                              ...              ...   
7038             True                            False             True   
7039            False                            False             True   
7040            False                            False            False   
7041            False                            False            False   
7042             True                            False             True   

      StreamingMovies_No internet service  StreamingMovies_Yes  \
0                                   False                False   
1                                   False                False   
2                                   False                False   
3                                   False                False   
4                                   False                False   
...                                   ...                  ...   
7038                                False                 True   
7039                                False                 True   
7040                                False                False   
7041                                False                False   
7042                                False                 True   

      Contract_One year  Contract_Two year  \
0                 False              False   
1                  True              False   
2                 False              False   
3                  True              False   
4                 False              False   
...                 ...                ...   
7038               True              False   
7039               True              False   
7040              False              False   
7041              False              False   
7042              False               True   

      PaymentMethod_Credit card (automatic)  PaymentMethod_Electronic check  \
0                                     False                            True   
1                                     False                           False   
2                                     False                           False   
3                                     False                           False   
4                                     False                            True   
...                                     ...                             ...   
7038                                  False                           False   
7039                                   True                           False   
7040                                  False                            True   
7041                                  False                           False   
7042                                  False                           False   
```



```
      PaymentMethod_Mailed check
0                         False
1                          True
2                          True
3                         False
4                         False
...                         ...
7038                       True
7039                      False
7040                      False
7041                       True
7042                      False
```

Step 4: Data Analysis with IBM

Now that your data is clean and ready, you can use IBM Cognos for data analysis. Here's a simplified process for analysis:

a.Data Connection

Load your preprocessed dataset into IBM Cognos by connecting to the data source. IBM Cognos provides various connectors to different data sources.

b.Data Exploration

Explore the dataset to understand its structure and contents.

c. Create Visualizations

Create various visualizations (e.g., bar charts, pie charts, histograms, heatmaps) to present the data effectively

d. Perform Descriptive Statistics

Calculate and visualize descriptive statistics

e. Hypothesis Testing

If applicable, conduct statistical tests to validate hypotheses

Step 5: Interpret and communicate results

Interpreting and effectively communicating the results of customer churn prediction is a critical aspect of leveraging this analysis for business decision-making. Once the predictive model is trained and tested, the findings need to be translated into actionable insights. This involves identifying the key drivers of churn, which could be factors such as decreased activity, overdue payments, or specific customer behaviors.

Step 6: Continuous monitoring and development

Monitor the campaign's effectiveness over time and make data-driven decisions to improve future campaigns. Collect and analyze additional data as needed to refine your analysis and recommendations. This process provides a high-level overview of how to build a public health awareness campaign analysis using IBM Cognos. Remember that the specific steps and techniques you use may vary depending on your analysis objectives and the nature of the data.

Conclusion:

data preprocessing is a foundational and indispensable phase in the customer churn prediction process. It plays a pivotal role in ensuring the accuracy and effectiveness of predictive models. By meticulously cleaning and transforming the data, handling missing values, and engineering relevant features, businesses can enhance the quality of the dataset, thereby improving the performance of the subsequent predictive algorithms. Data preprocessing also involves the critical task of standardizing or scaling features to ensure that variables with different scales do not unduly influence the modeling process. Furthermore, the process may include encoding categorical variables and splitting the data into training and testing sets. This preliminary phase sets the stage for more advanced predictive modeling and allows organizations to extract actionable insights from their data. In essence, data preprocessing serves as the cornerstone upon which robust customer churn prediction models are built, ultimately helping businesses make informed decisions, reduce churn rates, and foster customer retention.