|  |
| --- |
| Close-up image showing the leaf-sides of two oversized books side-by-side on a bookshelf, with additional books in soft focus background |
| **TELECOMMUNICATIONS CHURN**  **A Data-Driven Approach to Increase Customer Retention & Reduce Customer Attrition.** |
| |  |  |  | | --- | --- | --- | | Pushkar Mishra | 8/20/23 | BA 723 Capstone | |

**TABLE OF CONTENTS**

**Page Numbers:**

**Executive Summary 3**

**0.1. Executive Introduction 3**

**0.2. Executive Objective 3**

**0.3. Executive Model Description 4**

**0.4. Executive Recommendations 4**

**Introduction 5**

**1.0. Background 5**

**2.0. Problem Statement 6**

**3.0. Objectives & Measurement 6**

**4.0. Assumptions and Limitations 6**

**Data Sources 7**

**5.0. Data Set Introduction 7**

**6.0. Exclusions 11**

**6.1. Initial Data Cleansing or Preparation 12**

**7.0. Data Dictionary 14**

**Data Exploration 22**

**8.0. Data Exploration Techniques 22**

**8.1. Data Manipulation 25**

**9.0. Data Cleansing 46**

**10.0. Summary 47**

**Data Preparation and Feature Engineering 47**

**11.0. Data Preparation Needs 47**

**11.1. Censored Records, Excluded Columns 50**

**11.2. Imputations, Transformations, etc. 51**

**11.3. Up-sampling, Down-sampling, SMOTE 51**

**12.0. Feature Engineering 52**

**12.1. New Variables 53**

**Model Exploration 53 -**

**13.0. Modeling Approach/Introduction 53**

**14.0. Model Technique #01 53**

**14.1. Model Technique #02 55**

**14.0. Model Technique #03 57**

**14.1. Model Technique #04 59**

**14.0. Model Technique #05 62**

**14.0. Model Technique #06 64**

**14.0. Model Technique #07 65**

**14.0. Model Technique #08 68**

**14.0. Model Technique #09 70**

**14.0. Model Technique #10 72**

**14.0. Model Technique #11 74**

**14.0. Model Technique #12 76**

**14.0. Model Technique #13 78**

**14.0. Model Technique #14 80**

**14.0. Model Technique #15 82**

**14.0. Model Technique #16 83**

**14.0. Model Technique #17 84**

**14.0. Model Technique #18 86**

**14.0. Model Technique #19 88**

**15.0. Model Comparison 89**

**Model Recommendation 90**

**16.0 Model Selection 90**

**17.0 Model Theory 90**

**17.1 Model Assumptions and Limitations 91**

**18.0 Model Sensitivity to Key Drivers 92**

**19.0 Additional Models to Address Business Objectives 92**

**Conclusion and Recommendations 92**

**20.0. Impacts on Business Problem 92**

**(Scope of the recommended model)**

**21.0. Recommended Next Steps 92**

**References 93**

**22.0 References 93**

**Executive Summary**

* 1. **Executive Introduction**

The purpose of this report is to provide comprehensive documentation of the data science project undertaken to predict customer churn. Churn prediction is crucial for businesses as it allows them to identify potential customers likely to discontinue using their service or product.

This report documents the comprehensive analysis and modeling process undertaken to predict customer churn for a telecommunication company using a hypothetical dataset from IBM. Churn prediction is pivotal for telecommunication businesses, enabling them to pinpoint customers likely to discontinue their services, thereby formulating strategies to retain them.

The project aims to predict customer churn in the telecommunications sector using a dataset provided by IBM. Churn prediction is vital for businesses as it helps in retaining customers by understanding the reasons for churn and taking preventive measures.

The telecommunications industry is highly competitive, with multiple service providers vying for customer attention. In such an environment, customer churn, which refers to customers leaving one service provider for another, is a significant concern. Understanding the reasons behind churn and predicting which customers are likely to churn can provide a competitive advantage.

* 1. **Executive Objective**

The primary objective of this project was to develop a predictive model that can accurately identify customers at risk of churning. By doing so, businesses can proactively retain these customers, reducing overall churn rates and increasing revenue.

The overarching goal was to devise a model with high predictive accuracy to identify potential churners. By achieving this, the company can implement targeted retention strategies, ultimately reducing churn rates and bolstering revenue.

The primary objective is to build a predictive model that can accurately identify customers at high risk of churning. This will enable the company to take proactive measures to retain these customers, thereby reducing revenue loss and maintaining a strong customer base.

Using the provided dataset, the goal is to predict which customers are likely to churn soon. This prediction can help the company take proactive measures to retain high-risk customers.

* 1. **Executive Model Description**

The project encompassed the exploration of multiple machine learning models. Each model was trained on a dataset, which was a culmination of five distinct source files, merged to form a comprehensive working file. The models' performance was gauged based on their predictive prowess, and a final recommendation was made post a thorough comparison.

Multiple machine learning models were explored, including logistic regression, decision trees, and ensemble methods. Each model's performance was evaluated based on its accuracy, precision, recall, and F1 score. The best-performing model was selected for further analysis and interpretation.

The project encompassed the exploration of multiple machine learning models. Each model was trained on a dataset, which was a culmination of five distinct source files, merged to form a comprehensive working file. The models' performance was gauged based on their predictive prowess, and a final recommendation was made post a thorough comparison.

* 1. **Executive Recommendations**

Based on the findings of this project, it is recommended to deploy the selected model into a production environment where it can be used to score customers in real time. Additionally, businesses should consider implementing retention strategies targeting high-risk customers identified by the model.

Given the project's outcomes, it is advised to integrate the chosen model into a live environment for real-time customer scoring. Concurrently, it is imperative for the business to roll out retention initiatives specifically aimed at high-risk customers as identified by the model.

* Focus on the key features identified as significant predictors of churn.
* Implement targeted marketing strategies for customers identified as high risk.
* Continuously monitor model performance and update as necessary.

Based on the findings of this project, it is recommended to deploy the selected model into a production environment where it can be used to score customers in real time. Additionally, businesses should consider implementing retention strategies targeting high-risk customers identified by the model.

**Introduction**

1. **Background**

Customer churn, often referred to as customer attrition, turnover, or defection, is a significant concern for telecommunication companies. Retaining existing customers is often more cost-effective than acquiring new ones. With the telecommunication industry being highly competitive, companies are constantly seeking ways to reduce churn rates. Predictive analytics, particularly churn prediction models, have emerged as a valuable tool for identifying customers likely to churn soon.

**2.0.** **Problem Statement**

The telecommunication company has been experiencing an increased churn rate over the past few months. This not only leads to a loss of revenue but also increases the costs associated with acquiring new customers. The company needs a solution to identify potential churners in advance, allowing them to take proactive measures.

Predict the likelihood of a customer churning using historical data.

**3.0. Objectives & Measurement**

The primary objective of this project is to develop a predictive model that can identify customers at a high risk of churning. The model's performance will be measured based on its accuracy, precision, recall, and F1-score. A higher accuracy rate will ensure that the company can trust the model's predictions and allocate resources effectively for customer retention strategies.

* **Objective**: Predict customer churn with a high degree of accuracy.
* **Measurement**: Accuracy, Precision, Recall, and F1-Score of the predictive models.

**4.0. Assumptions and Limitations**

1. The data provided is accurate and up-to-date.
2. Historical data patterns related to churn will continue in the future.
3. The model's predictions are based on the features available in the dataset, and external factors not present in the data might influence a customer's decision to churn.
4. The hypothetical dataset from IBM represents a realistic scenario for a telecommunication company, even though it might not capture all possible nuances.
5. The data is representative of the broader customer base.
6. Churn is influenced by the features present in the dataset.
7. The models might not account for external factors influencing churn.
8. The dataset provided is a representative sample of the broader customer base.
9. Features in the dataset are sufficient to predict churn.
10. The model might not capture external factors influencing churn, such as broader market trends or competitor actions.
11. Historical data patterns related to churn will continue in the future.
12. The model's predictions are based on the features available in the dataset, and external factors not present in the data might influence a customer's decision to churn.
13. The hypothetical dataset from IBM represents a realistic scenario for a telecommunication company, even though it might not capture all possible nuances.

**Data Sources**

**5.0. Data Set Introduction**

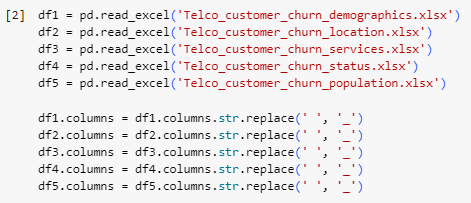
The dataset used for this project is a hypothetical dataset from IBM, tailored for churn prediction in the telecommunication industry. It is a culmination of five distinct source files, which were merged to form a comprehensive working file. These files provide a holistic view of the customers, capturing demographics, location, population density, services availed, and their current churn status.

The dataset is sourced from IBM and contains various customer attributes, including demographic information, service usage, and billing information.

There were 5 source data files downloaded from the IBM website carrying these filenames in MS Excel format:

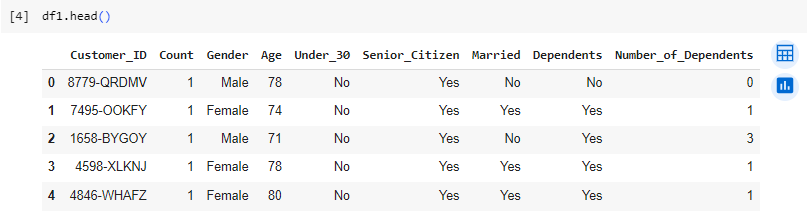
1. **Telco\_customer\_churn\_demographics.xlsx**
2. **Telco\_customer\_churn\_location.xlsx**
3. **Telco\_customer\_churn\_population.xlsx**
4. **Telco\_customer\_churn\_services.xlsx**
5. **Telco\_customer\_churn\_status.xlsx**

Reading the dataset in the Google Collaboratory Python Environment setting up the project (DF stands for Data Frame):



Reading the First 5 rows displaying all the columns of the 5 datasets.

DF1 (Demographics):



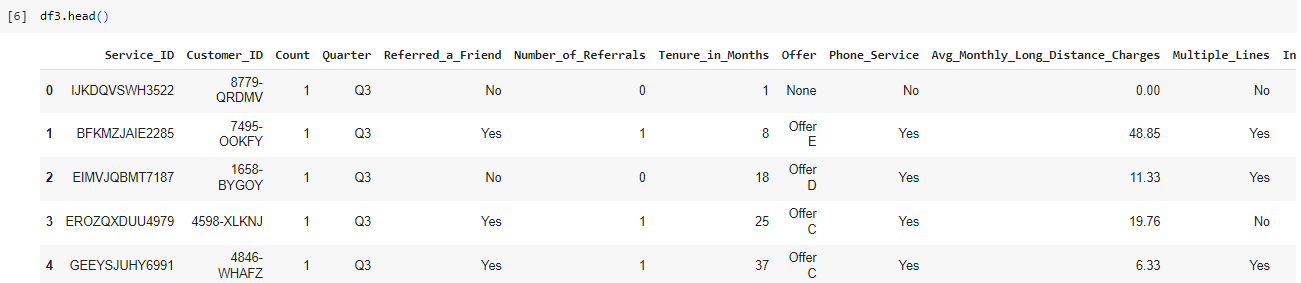
DF2 (Location):



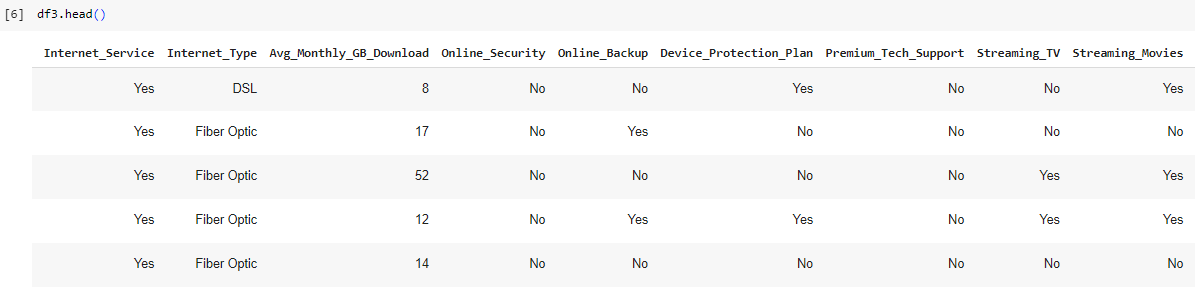
DF3 (Services),

(Multiple snips due to a larger data frame with many columns):

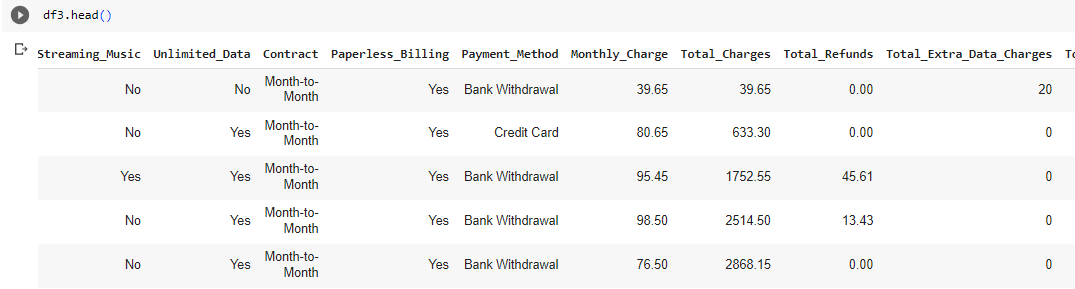
Snip1 :



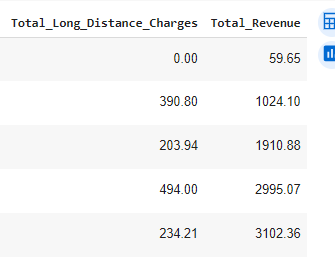
Snip2:



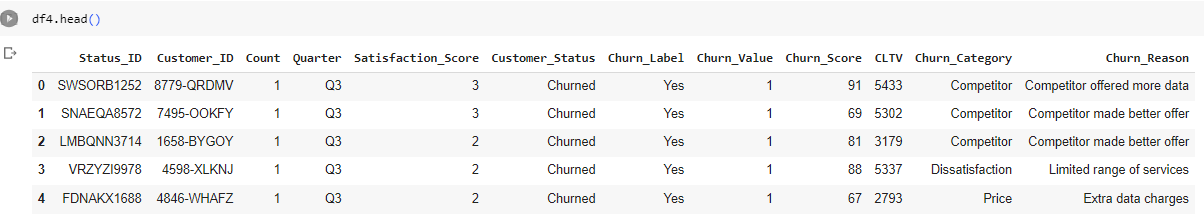
Snip3:



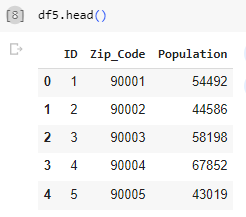
Snip4:



DF4 (Status):



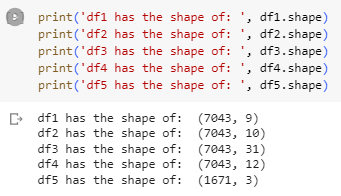
DF5 (Population):



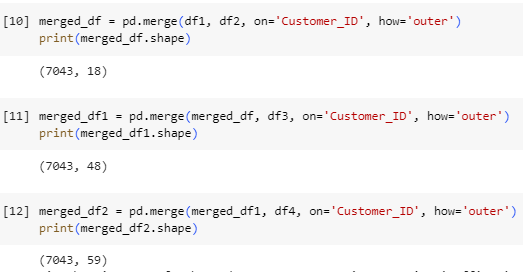
**6.0. Exclusions**

Certain records and columns that were deemed irrelevant or redundant for the modeling process were excluded based on their relevance to the prediction task. Since there were 5 files merged into one working file, the df5 Specific exclusions are detailed here:

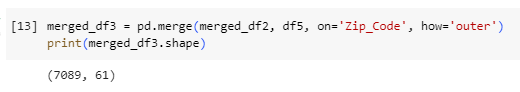
The shape of the 5 files:



Merging the first 4 files using the “Outer Join” because I wanted everything to assess the relevance, and joining on the common field for all 4 datafiles, i.e., “Customer\_ID”, with the new shape displayed:



And finally, merge the last file using the “Zip\_Code” field and once again with an “Outer Join” for the same reason as above.

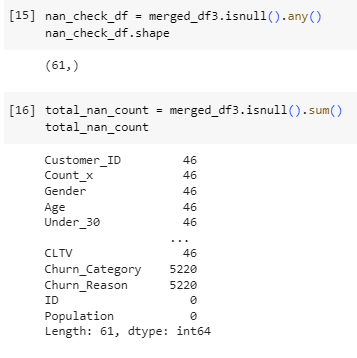


**6.1. Initial Data Cleansing or Preparation**

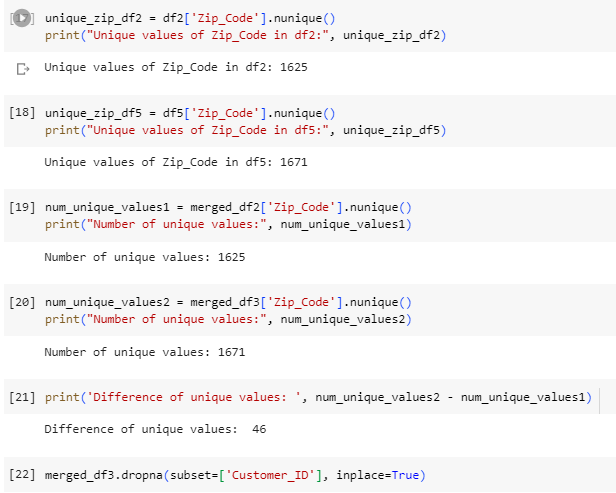
Initial data cleansing involved handling missing values, correcting data formats, and removing any inconsistencies present in the data. A detailed account of these steps will be provided in the data exploration and preparation sections.

Data preprocessing steps include handling missing values, encoding categorical variables, and scaling numerical features.

After the merge, it was noticed that there were extra rows with null values or NaNs (NaN stands for “Not a Number”.).



Further investigation revealed that one file had more zip codes than the other file and had no relevant information, except population. Rest everything was NaN. There were 46 such records after the merge, and were removed:



**7.0. Data Dictionary**

A data dictionary, providing a description of each column, its data type, and potential values, will be detailed in the subsequent sections. This dictionary serves as a reference to understand the nature and structure of the dataset.

A detailed data dictionary would provide descriptions for each feature in the dataset. The final merged dataset consists of 7043 entries and 31 columns. Here is a brief overview of the columns, clubbed under 4 categories by me:

1. **Demographics**: **Gender**, **Age**, **Senior\_Citizen**, **Married**, **Dependents**
2. **Subscription Info**: **Referred\_a\_Friend**, **Tenure\_in\_Months**, **Offer**, **Phone\_Service**, **Multiple\_Lines**, **Internet\_Service**, **Internet\_Type**, **Online\_Security**, **Online\_Backup**, **Device\_Protection\_Plan**, **Premium\_Tech\_Support**, **Streaming\_TV**, **Streaming\_Movies**, **Streaming\_Music**, **Unlimited\_Data**, **Contract**, **Paperless\_Billing**, **Payment\_Method**
3. **Charges and Usage**: **Avg\_Monthly\_Long\_Distance\_Charges**, **Avg\_Monthly\_GB\_Download**, **Monthly\_Charge**, **Total\_Charges**, **Total\_Extra\_Data\_Charges**, **Total\_Long\_Distance\_Charges**, **Total\_Revenue**
4. **Target Variable**: **Churn\_Label**

However, the detailed data dictionary for all 5 source files, as defined by the owners of the dataset, viz. IBM is as follows:

**Demographics**

**CustomerID:**

**Count:** A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

**Gender:** The customer’s gender: Male, Female

**Age:** The customer’s current age, in years, at the time the fiscal quarter ended.

**Senior Citizen:** Indicates if the customer is 65 or older: Yes, No

**Married:** Indicates if the customer is married: Yes, No

**Dependents:** Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.

**Number of Dependents:** Indicates the number of dependents that live with the customer.

**Location**

**CustomerID:** A unique ID that identifies each customer.

**Count:** A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

**Country:** The country of the customer’s primary residence.

**State:** The state of the customer’s primary residence.

**City:** The city of the customer’s primary residence.

**Zip Code:** The zip code of the customer’s primary residence.

**Lat Long:** The combined latitude and longitude of the customer’s primary residence.

**Latitude:** The latitude of the customer’s primary residence.

**Longitude:** The longitude of the customer’s primary residence.

**Population**

**ID:** A unique ID that identifies each row.

**Zip Code:** The zip code of the customer’s primary residence.

**Population:** A current population estimate for the entire Zip Code area.

**Services**

**CustomerID:** A unique ID that identifies each customer.

**Count:** A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

**Quarter:** The fiscal quarter that the data has been derived from (e.g. Q3).

**Referred a Friend:** Indicates if the customer has ever referred a friend or family member to this company: Yes, No

**Number of Referrals:** Indicates the number of referrals to date that the customer has made.

**Tenure in Months:** Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above.

**Offer:** Identifies the last marketing offer that the customer accepted, if applicable. Values include None, Offer A, Offer B, Offer C, Offer D, and Offer E.

**Phone Service:** Indicates if the customer subscribes to home phone service with the company: Yes, No

**Avg Monthly Long-Distance Charges:** Indicates the customer’s average long-distance charges, calculated to the end of the quarter specified above.

**Multiple Lines:** Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No

**Internet Service:** Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable.

**Avg Monthly GB Download:** Indicates the customer’s average download volume in gigabytes, calculated to the end of the quarter specified above.

**Online Security:** Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No

**Online Backup:** Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No

**Device Protection Plan:** Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No

**Premium Tech Support:** Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No

**Streaming TV:** Indicates if the customer uses their Internet service to stream television programming from a third-party provider: Yes, No. The company does not charge an additional fee for this service.

**Streaming Movies:** Indicates if the customer uses their Internet service to stream movies from a third-party provider: Yes, No. The company does not charge an additional fee for this service.

**Streaming Music:** Indicates if the customer uses their Internet service to stream music from a third-party provider: Yes, No. The company does not charge an additional fee for this service.

**Unlimited Data:** Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads: Yes, No

**Contract:** Indicates the customer’s current contract type: Month-to-Month, One Year, Two Year.

**Paperless Billing:** Indicates if the customer has chosen paperless billing: Yes, No

**Payment Method:** Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check

**Monthly Charge:** Indicates the customer’s current total monthly charge for all their services from the company.

**Total Charges:** Indicates the customer’s total charges, calculated to the end of the quarter specified above.

**Total Refunds:** Indicates the customer’s total refunds, calculated to the end of the quarter specified above.

**Total Extra Data Charges:** Indicates the customer’s total charges for extra data downloads above those specified in their plan, by the end of the quarter specified above.

**Total Long-Distance Charges:** Indicates the customer’s total charges for long-distance above those specified in their plan, by the end of the quarter specified above.

**Status**

**CustomerID:** A unique ID that identifies each customer.

**Count:** A value used in reporting/dashboarding, to sum up the number of customers in a filtered set.

**Quarter:** The fiscal quarter that the data has been derived from (e.g., Q3).

**Satisfaction Score:** A customer’s overall satisfaction rating of the company from 1 (Very Unsatisfied) to 5 (Very Satisfied).

**Satisfaction Score Label:** Indicates the text version of the score (1-5) as a text string.

**Customer Status:** Indicates the status of the customer at the end of the quarter: Churned, Stayed, or Joined

**Churn Label:** Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value.

**Churn Value:** 1 = the customer left the company this quarter. 0 = the customer remained with the company. Directly related to Churn Label.

**Churn Score:** A value from 0-100 that is calculated using the predictive tool IBM SPSS Modeler. The model incorporates multiple factors known to cause churn. The higher the score, the more likely the customer will churn.

**Churn Score Category:** A calculation that assigns a Churn Score to one of the following categories: 0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, and 91-100

**CLTV:** Customer Lifetime Value. A predicted CLTV is calculated using corporate formulas and existing data. The higher the value, the more valuable the customer. High value customers should be monitored for churn.

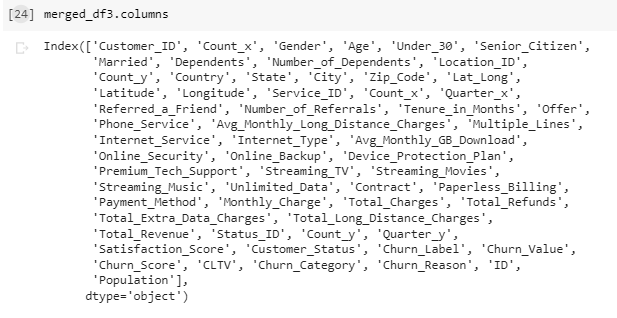
**CLTV Category:** A calculation that assigns a CLTV value to one of the following categories: 2000-2500, 2501-3000, 3001-3500, 3501-4000, 4001-4500, 4501-5000, 5001-5500, 5501-6000, 6001-6500, and 6501-7000.

**Churn Category:** A high-level category for the customer’s reason for churning: Attitude, Competitor, Dissatisfaction, Other, Price. When they leave the company, all customers are asked about their reasons for leaving. Directly related to Churn Reason.

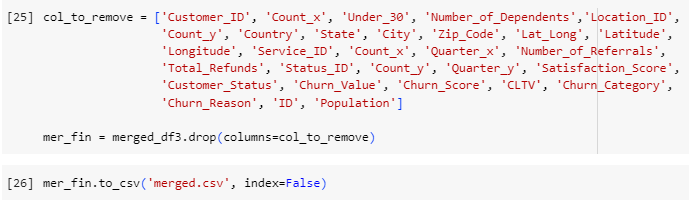
**Churn Reason:** A customer’s specific reason for leaving the company. Directly related to Churn Category.

<https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113>

The final merged dataset had the following 61 columns:



The final file had some calculated columns from the source files and some geospatial information, in addition, there were some more columns that were not deemed necessary for the models and were removed. The initial removal was as follows:



The file is converted into a csv format and saved as “merged.csv”. It had the shape of 7043 rows and 31 columns:



**Data Exploration**

**8.0. Data Exploration Techniques**

Various data exploration techniques were employed to understand the dataset's characteristics and distributions. These techniques include:

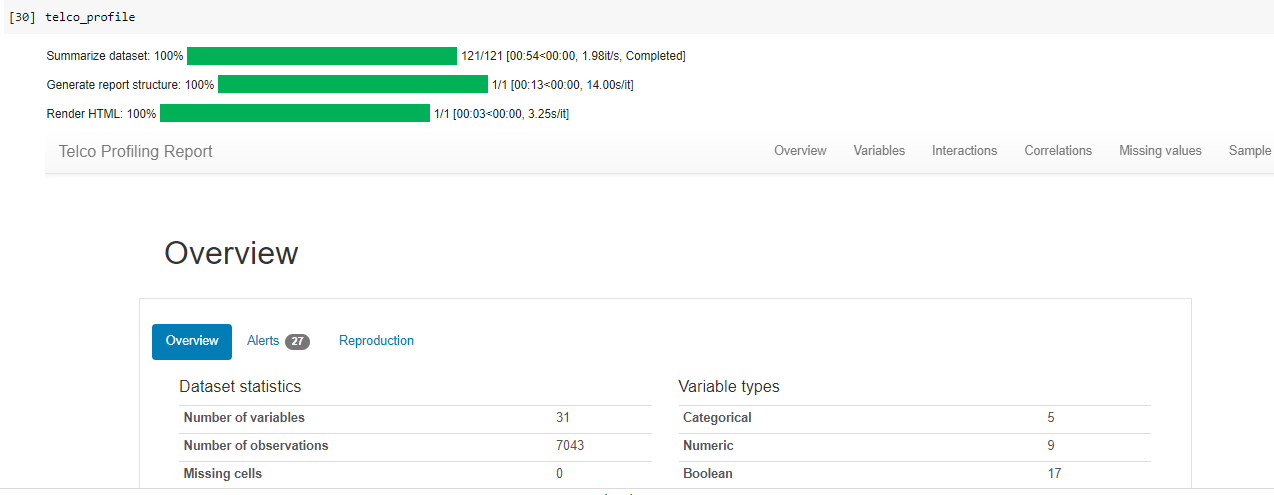
* Descriptive statistics to summarize the central tendency, dispersion, and shape of the dataset's distribution.
* Visualization techniques such as histograms, bar plots, and box plots to visualize the distribution of individual features.
* Correlation matrices to understand the relationships between different features.
* Identification of outliers and anomalies in the data.
* Exploratory Data Analysis (EDA) techniques, including data visualization, were used to understand the distribution and relationships of features.

To get a high-level overview of the dataset, its interaction with the variables, bar graphs, correlations, and other statistical features, and to get a complete profile, a very nice tool was installed and utilized to get the profile:

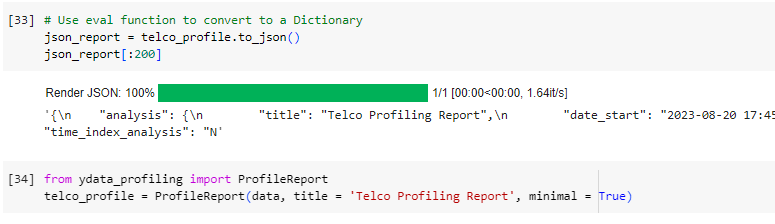




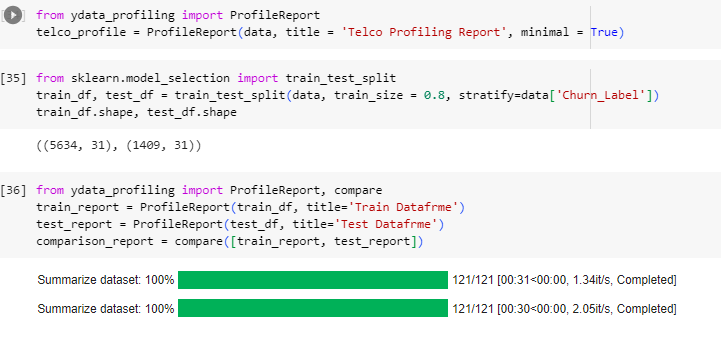
And the profile was created:



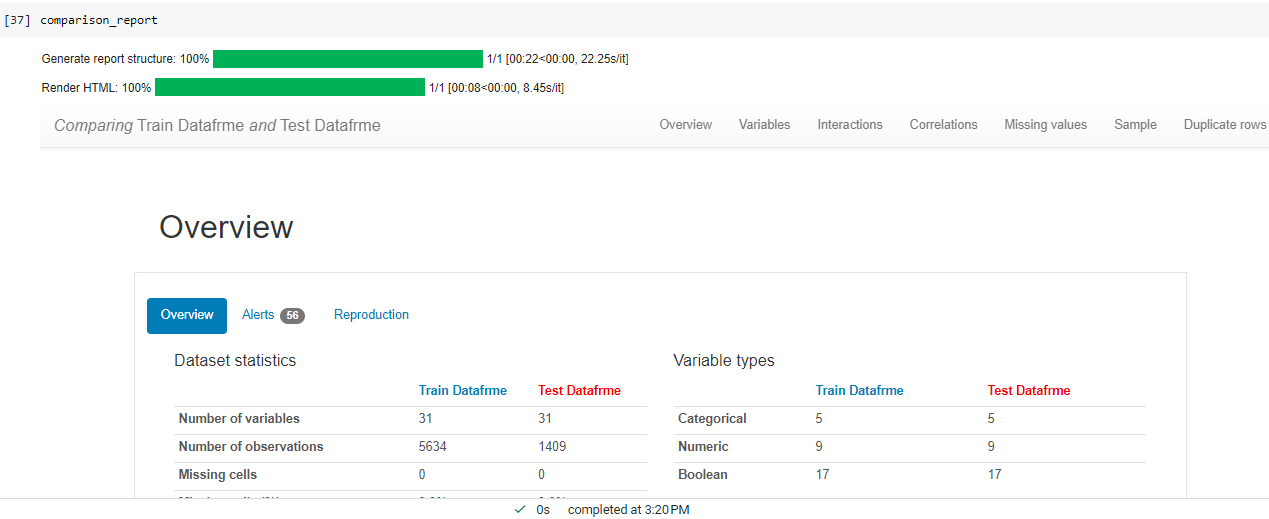
The profile report was then exported as save files with different formats:



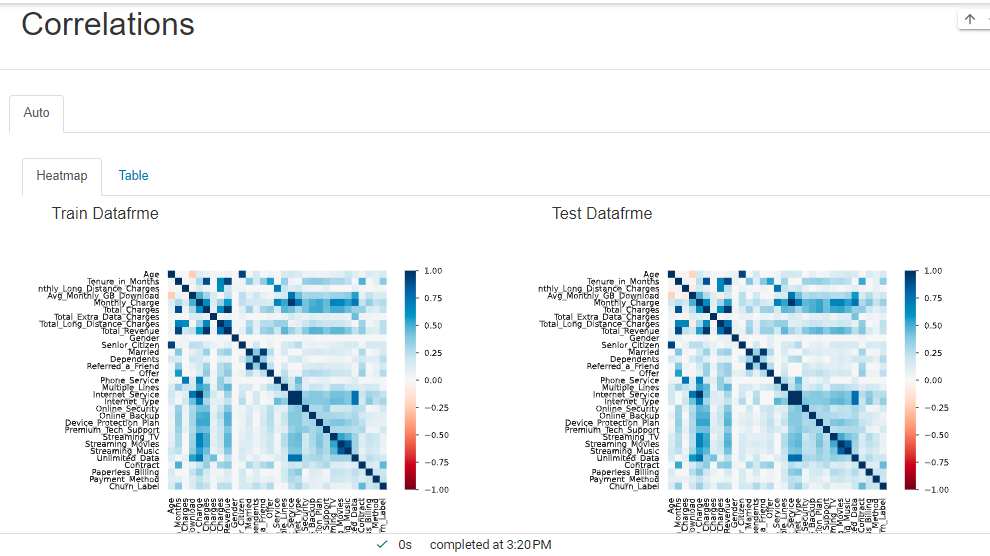
Continuing further with this tool, the dataset was split to visualize and utilize the comparison tool to have a high-level view of the train and test dataset to identify any anomalies.



Comparison Report was generated, stored locally and visualized:



The correlation map from this comparison report:

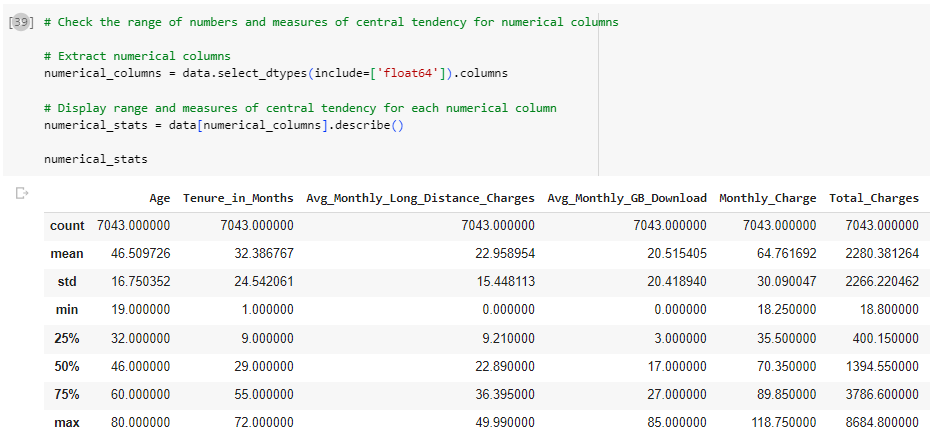


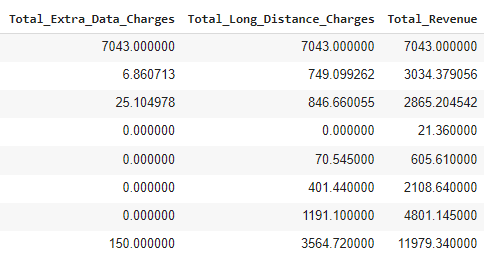
**8.1. Data Manipulation**

This subsection will detail specific insights and findings from the data exploration phase, including patterns, anomalies, and potential areas of interest that could influence the modeling process.

Visualizations such as bar plots and histograms were used to understand the distribution of categorical and numerical features, respectively.

The central tendency of the numerical values in the dataset (split in 2 snips):





The dataset has the following columns and their nature:

Here are the columns and their respective data types:

* **Gender**: object (categorical)
* **Age**: float64 (numerical)
* **Senior\_Citizen**: object (categorical)
* **Married**: object (categorical)
* **Dependents**: object (categorical)
* **Referred\_a\_Friend**: object (categorical)
* **Tenure\_in\_Months**: float64 (numerical)
* **Offer**: object (categorical)
* **Phone\_Service**: object (categorical)
* **Avg\_Monthly\_Long\_Distance\_Charges**: float64 (numerical)
* **Multiple\_Lines**: object (categorical)
* **Internet\_Service**: object (categorical)
* **Internet\_Type**: object (categorical)
* **Avg\_Monthly\_GB\_Download**: float64 (numerical)
* **Online\_Security**: object (categorical)
* **Online\_Backup**: object (categorical)
* **Device\_Protection\_Plan**: object (categorical)
* **Premium\_Tech\_Support**: object (categorical)
* **Streaming\_TV**: object (categorical)
* **Streaming\_Movies**: object (categorical)
* **Streaming\_Music**: object (categorical)
* **Unlimited\_Data**: object (categorical)
* **Contract**: object (categorical)
* **Paperless\_Billing**: object (categorical)
* **Payment\_Method**: object (categorical)
* **Monthly\_Charge**: float64 (numerical)
* **Total\_Charges**: float64 (numerical)
* **Total\_Extra\_Data\_Charges**: float64 (numerical)
* **Total\_Long\_Distance\_Charges**: float64 (numerical)
* **Total\_Revenue**: float64 (numerical)

**Target Variablee to be Predicted: Churn\_Label**: object (categorical)

The dataset was checked for potential issues with the data:

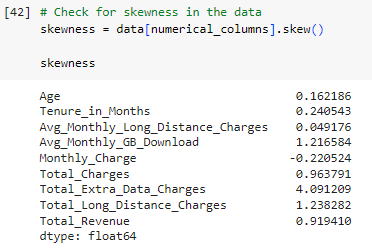
1. Illogical nominal categories.
2. Numbers stored as characters (e.g., "one" is 1).
3. Check the range of numbers.
4. Measures of central tendency (mean, median, mode, standard deviation) for numerical columns.

Here are the unique values for each categorical column:

* **Gender**: ['Male', 'Female']
* **Senior\_Citizen**: ['Yes', 'No']
* **Married**: ['No', 'Yes']
* **Dependents**: ['No', 'Yes']
* **Referred\_a\_Friend**: ['No', 'Yes']
* **Offer**: ['None', 'Offer E', 'Offer D', 'Offer C', 'Offer B', 'Offer A']
* **Phone\_Service**: ['No', 'Yes']
* **Multiple\_Lines**: ['No', 'Yes']
* **Internet\_Service**: ['Yes', 'No']
* **Internet\_Type**: ['DSL', 'Fiber Optic', 'Cable', 'None']
* **Online\_Security**: ['No', 'Yes']
* **Online\_Backup**: ['No', 'Yes']
* **Device\_Protection\_Plan**: ['Yes', 'No']
* **Premium\_Tech\_Support**: ['No', 'Yes']
* **Streaming\_TV**: ['No', 'Yes']
* **Streaming\_Movies**: ['Yes', 'No']
* **Streaming\_Music**: ['No', 'Yes']
* **Unlimited\_Data**: ['No', 'Yes']
* **Contract**: ['Month-to-Month', 'Two Year', 'One Year']
* **Paperless\_Billing**: ['Yes', 'No']
* **Payment\_Method**: ['Bank Withdrawal', 'Credit Card', 'Mailed Check']
* **Churn\_Label**: ['Yes', 'No']

From the above values, there do not seem to be any illogical nominal categories or numbers stored as characters.

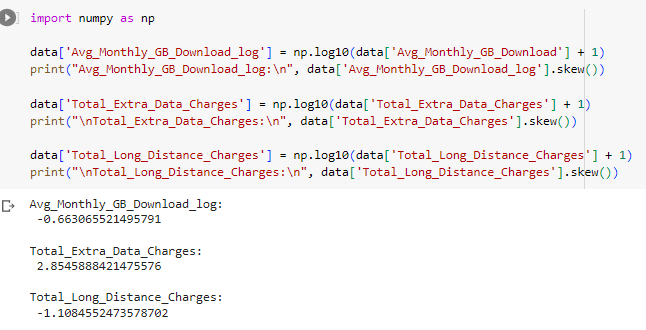
Here is the skewness for the numerical columns:



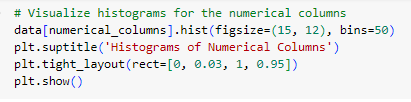
* **Age**: 0.162
* **Tenure\_in\_Months**: 0.241
* **Avg\_Monthly\_Long\_Distance\_Charges**: 0.049
* **Avg\_Monthly\_GB\_Download**: 1.217
* **Monthly\_Charge**: -0.221
* **Total\_Charges**: 0.964
* **Total\_Extra\_Data\_Charges**: 4.091
* **Total\_Long\_Distance\_Charges**: 1.238
* **Total\_Revenue**: 0.919

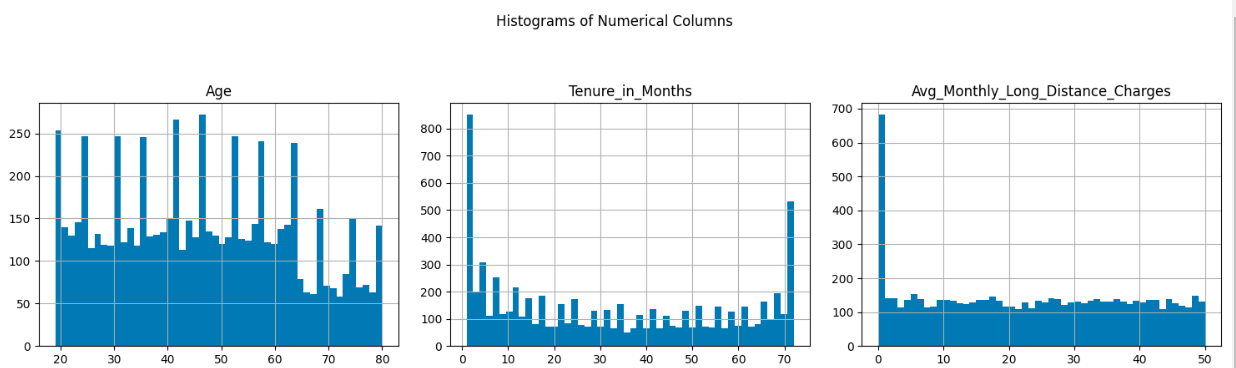
A skewness value close to 0 indicates that the distribution is approximately symmetric. Positive skewness indicates that the distribution is skewed to the right, while negative skewness indicates a left-skewed distribution. Columns like **Avg\_Monthly\_GB\_Download**, **Total\_Extra\_Data\_Charges**, and **Total\_Long\_Distance\_Charges** have a noticeable positive skew.

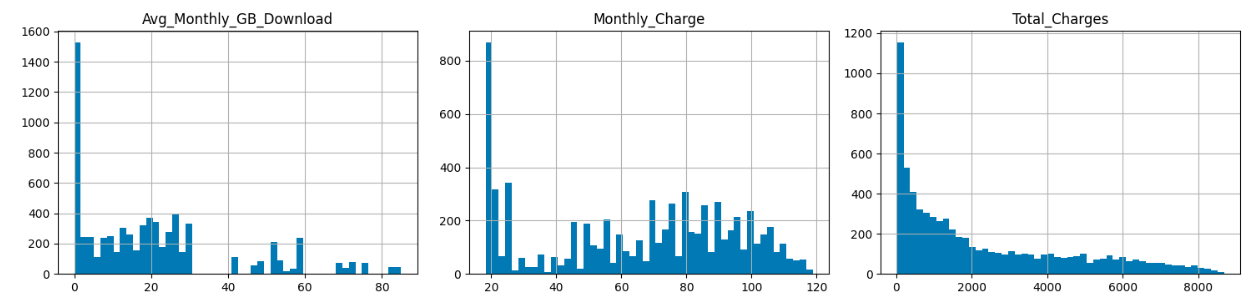
The skewness was handled as follows and the resulting skewness:

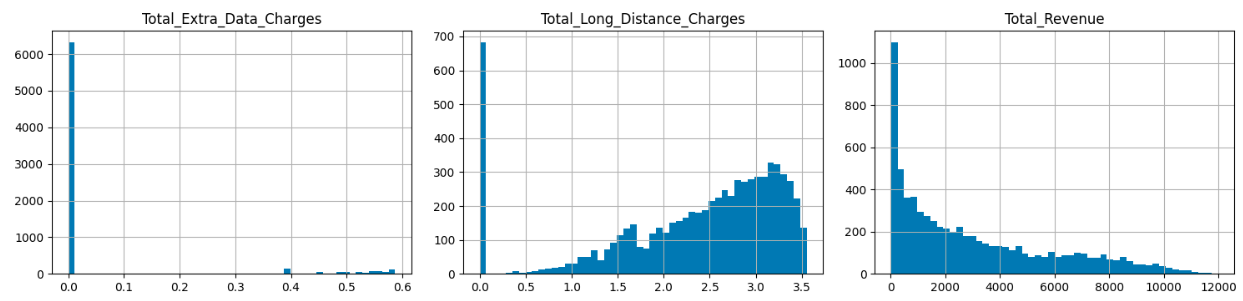
Some visualizations were done to see the distribution of the observations:

(Histograms for numerical variables)



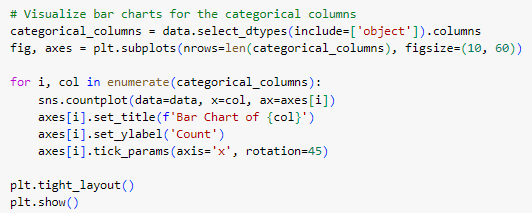




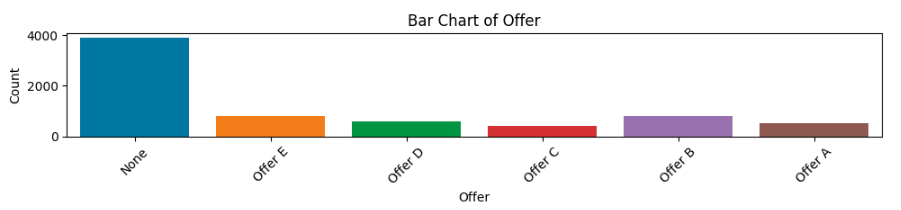


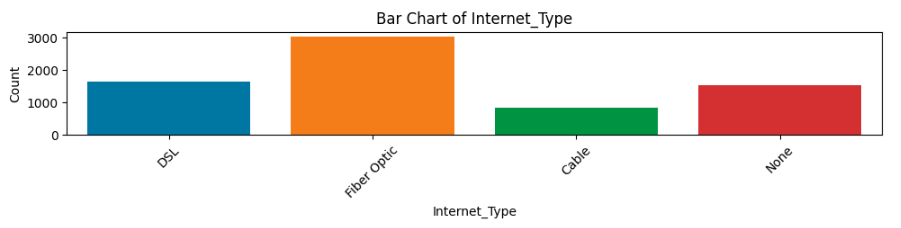
* **Age**: Appears to be uniformly distributed.
* **Tenure\_in\_Months**: Shows a bimodal distribution with peaks at the lower and higher end, indicating many new customers and many long-term customers.
* **Avg\_Monthly\_Long\_Distance\_Charges**: Appears to be normally distributed.
* **Avg\_Monthly\_GB\_Download**: Right-skewed, indicating that most individuals have lower GB downloads, but there are a few with very high downloads.
* **Monthly\_Charge**: Shows a bimodal distribution.
* **Total\_Charges**: Right-skewed, indicating that most individuals have lower total charges, but there are a few with very high charges.
* **Total\_Extra\_Data\_Charges**: Right-skewed with most individuals having lower extra data charges.
* **Total\_Long\_Distance\_Charges**: Appears to be normally distributed.
* **Total\_Revenue**: Right-skewed, similar to Total\_Charges.

(Bar charts for the categorical columns)



Example of some of the barcharts as there are too many of them:





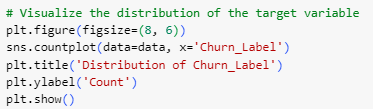
The description of the individual bar charts:

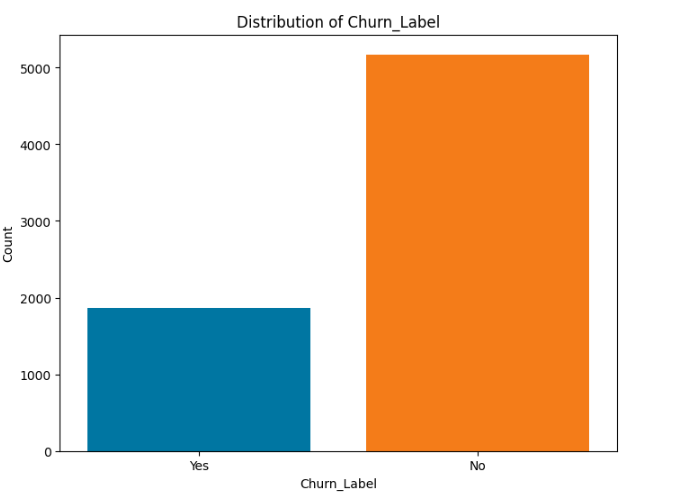
Here are the bar charts for the categorical columns, which provide a visual representation of the distribution of data for each category:

* **Gender**: The distribution between 'Male' and 'Female' is fairly balanced.
* **Senior\_Citizen**: There are fewer senior citizens than non-senior citizens.
* **Married**: The number of unmarried individuals is slightly higher than married ones.
* **Dependents**: A majority of the individuals do not have dependents.
* **Referred\_a\_Friend**: Most individuals have not referred a friend.
* **Offer**: 'None' and 'Offer E' are the most common offers.
* **Phone\_Service**: A majority have phone service.
* **Multiple\_Lines**: The distribution between having multiple lines and not is fairly balanced.
* **Internet\_Service**: Most individuals have internet service.
* **Internet\_Type**: 'Fiber Optic' is the most common internet type.
* **Online\_Security**: Most individuals do not have online security.
* **Online\_Backup**: A majority do not have online backup.
* **Device\_Protection\_Plan**: The distribution is fairly balanced between having a device protection plan and not.
* **Premium\_Tech\_Support**: Most individuals do not have premium tech support.
* **Streaming\_TV**: The distribution is fairly balanced between those who have streaming TV and those who don't.
* **Streaming\_Movies**: Similar to Streaming\_TV, the distribution is balanced.
* **Streaming\_Music**: Most individuals do not have streaming music.
* **Unlimited\_Data**: A majority have unlimited data.
* **Contract**: Most individuals are on a 'Month-to-Month' contract.
* **Paperless\_Billing**: A majority prefer paperless billing.
* **Payment\_Method**: 'Bank Withdrawal' is the most common payment method.

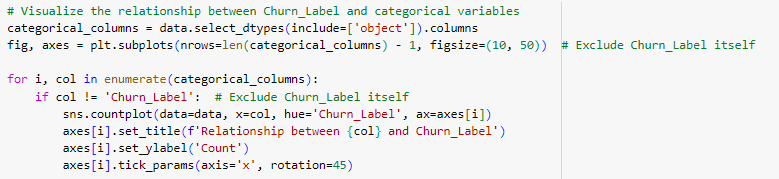
**Churn\_Label**: A majority of individuals have not churned.

The distribution of the target variable was visualized next:



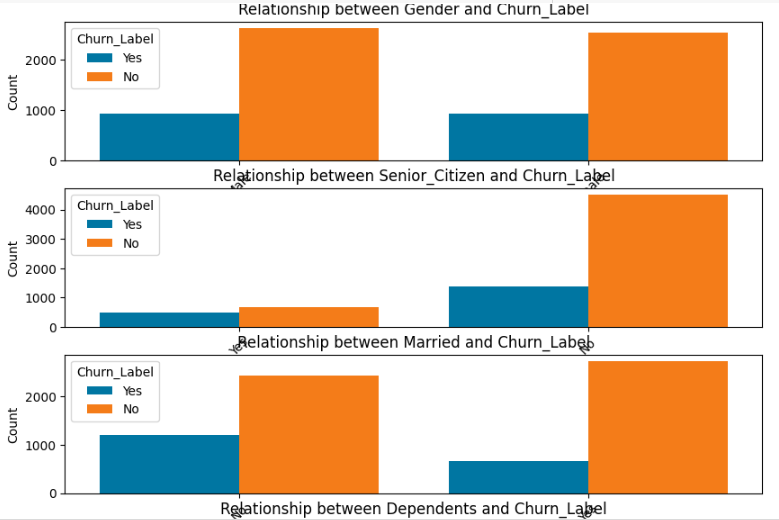


Next, the target variable, **Churn\_Label**, and its relationship with other variables was visualized:

. 

Approximately 73.46% of the customers in the dataset did not churn, while 26.54% did. This indicates that the majority of customers remained with the service, but there's still a significant proportion that churned.

Some of the visualized relationships with churn label:

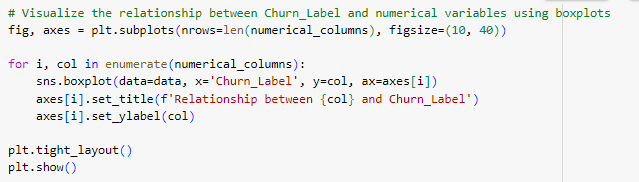


The bar charts provide insights into the relationship between each categorical variable and the **Churn\_Label**:

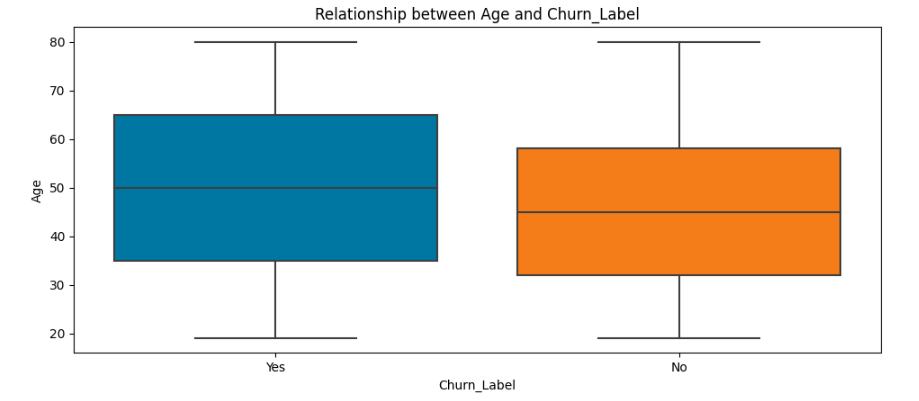
1. **Gender**: Churn rate seems fairly balanced between males and females.
2. **Senior\_Citizen**: Senior citizens have a higher churn rate compared to non-senior citizens.
3. **Married**: Unmarried individuals have a slightly higher churn rate.
4. **Dependents**: Individuals without dependents have a higher churn rate.
5. **Referred\_a\_Friend**: Those who haven't referred a friend have a higher churn rate.
6. **Offer**: Different offers have varying churn rates, with 'Offer A' and 'Offer D' showing higher churn.
7. **Phone\_Service**: Churn rate is higher for those with phone service.
8. **Multiple\_Lines**: Churn rate is slightly higher for those without multiple lines.
9. **Internet\_Service**: Those with internet service, especially 'Fiber Optic', have a higher churn rate.
10. **Internet\_Type**: 'Fiber Optic' users have the highest churn rate.
11. **Online\_Security**: Those without online security have a higher churn rate.
12. **Online\_Backup**: Similar to online security, those without online backup have a higher churn rate.
13. **Device\_Protection\_Plan**: Individuals without a device protection plan have a higher churn rate.
14. **Premium\_Tech\_Support**: Those without premium tech support have a higher churn rate.
15. **Streaming\_TV**: Churn rate is fairly balanced between those who have streaming TV and those who don't.
16. **Streaming\_Movies**: Similar to Streaming\_TV, the churn rate is balanced.
17. **Streaming\_Music**: Those without streaming music have a higher churn rate.
18. **Unlimited\_Data**: Churn rate is higher for those with unlimited data.
19. **Contract**: Month-to-Month contracts have the highest churn rate.
20. **Paperless\_Billing**: Those with paperless billing have a higher churn rate.
21. **Payment\_Method**: 'Credit Card' and 'Bank Withdrawal' methods have higher churn rates.

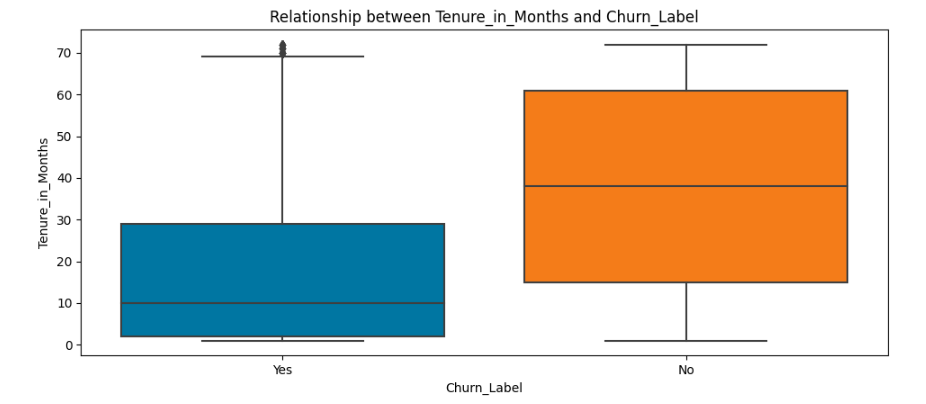
These insights can be valuable for understanding the factors influencing customer churn and developing strategies to retain customers.

Continuing with the data exploration, Boxplots were created for churn against the numerical columns:



Sampling a couple of them:



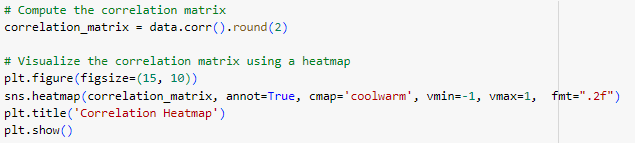


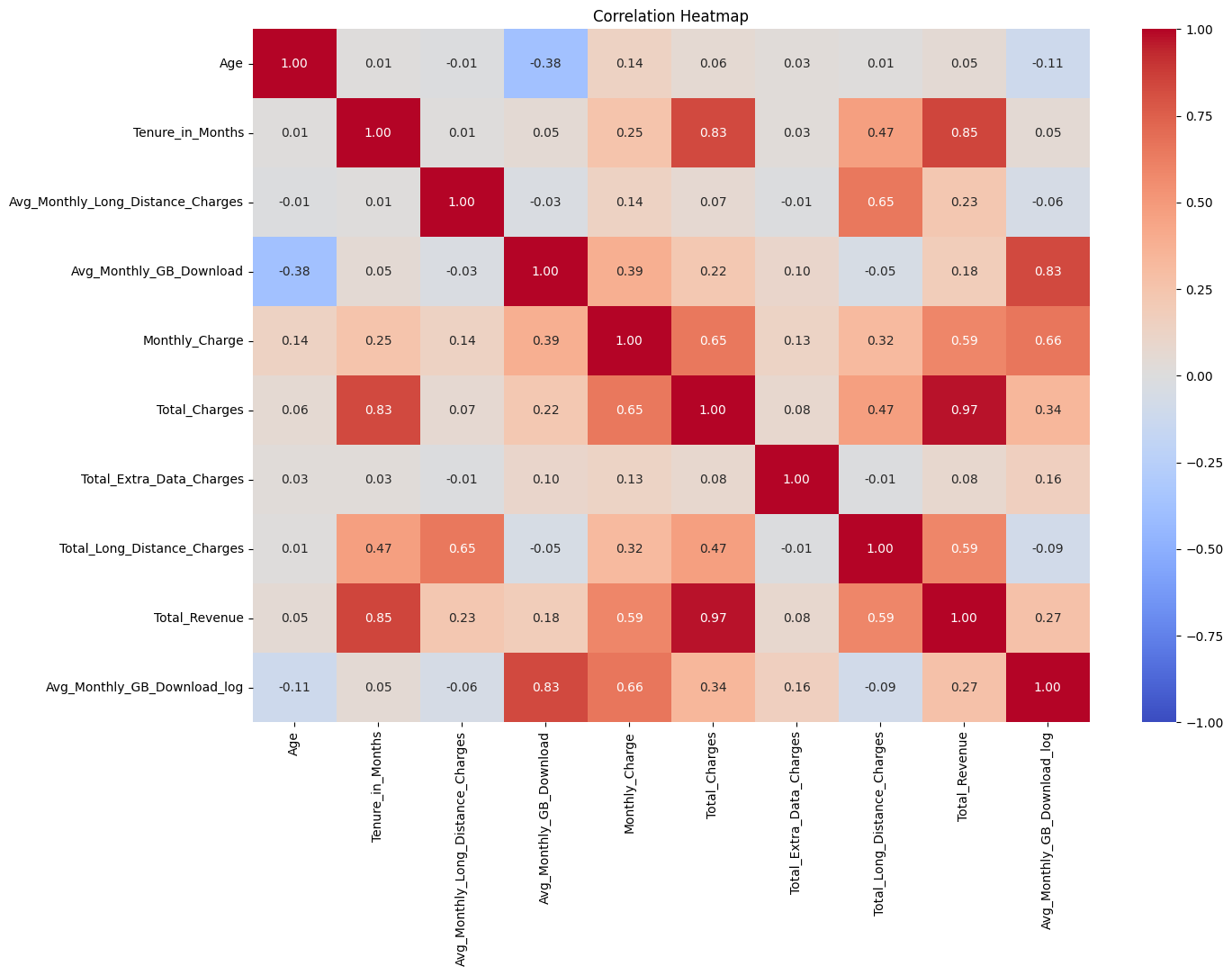
The boxplots provide insights into the relationship between each numerical variable and the **Churn\_Label**:

1. **Age**: The median age of those who churned is slightly lower than those who didn't.
2. **Tenure\_in\_Months**: Customers with lower tenure (newer customers) have a higher tendency to churn.
3. **Avg\_Monthly\_Long\_Distance\_Charges**: The distribution of charges is fairly similar between those who churned and those who didn't.
4. **Avg\_Monthly\_GB\_Download**: The median GB download for those who churned is slightly higher.
5. **Monthly\_Charge**: Customers who churned tend to have higher monthly charges.
6. **Total\_Charges**: The total charges for customers who churned is lower, which aligns with their lower tenure.
7. **Total\_Extra\_Data\_Charges**: The distribution of extra data charges is fairly similar between the two groups.
8. **Total\_Long\_Distance\_Charges**: The distribution of long-distance charges is also fairly similar between those who churned and those who didn't.
9. **Total\_Revenue**: The total revenue from customers who churned is lower, aligning with their lower tenure and total charges.

These insights can help in understanding the behavior of customers who churn and those who stay, especially in terms of their spending patterns, tenure, and usage.

Next a correlation heatmap was created for numerical variables:





The correlation heatmap provides insights into the relationships between numerical variables:

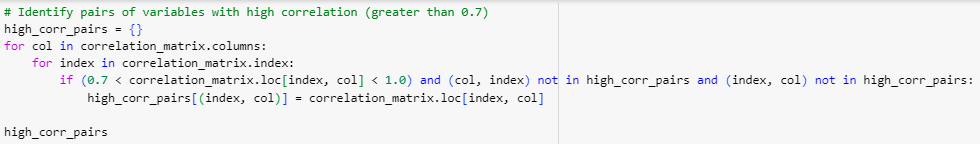
* Values closer to **1** indicate a strong positive correlation.
* Values closer to **-1** indicate a strong negative correlation.
* Values around **0** indicate little to no correlation.

From the heatmap, we can observe:

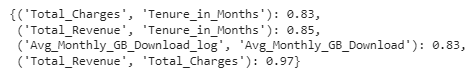
1. **Tenure\_in\_Months** has a strong positive correlation with **Total\_Charges** and **Total\_Revenue**. This is expected as the longer a customer stays, the more they are likely to be charged over time.
2. **Monthly\_Charge** also has a positive correlation with **Total\_Charges** and **Total\_Revenue**.
3. **Total\_Charges** and **Total\_Revenue** are highly correlated, which is expected since revenue is derived from charges.

**Avg\_Monthly\_GB\_Download** has a positive correlation with **Monthly\_Charge**, indicating that customers who download more data tend to have higher monthly charges.

Continuing with the EDA, multicollinearity, which refers to a situation where two or more variables are highly correlated, was addressed. This can be problematic in predictive modeling as it can make it hard to determine the effect of individual variables. l identifed pairs of variables with a correlation greater than 0.7.



Which resulted in:

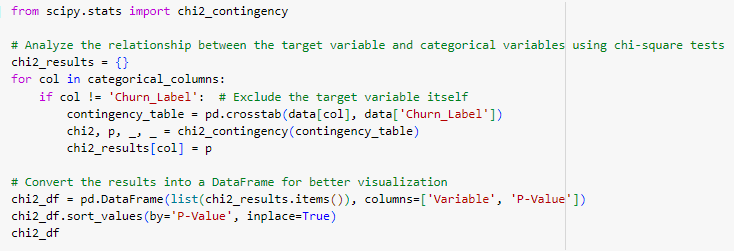


The pairs of variables with a correlation greater than 0.7 are:

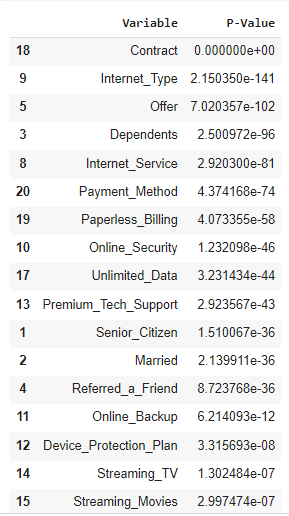
1. **Total\_Charges** and **Tenure\_in\_Months** with a correlation of 0.826.
2. **Total\_Revenue** and **Tenure\_in\_Months** with a correlation of 0.853.
3. **Avg\_Monthly\_GB\_Download\_log** and **Avg\_Monthly\_GB\_Download** with a correlation of 0.832.
4. **Total\_Revenue** and **Total\_Charges** with a very high correlation of 0.972.

These high correlations indicate multicollinearity. In predictive modeling, it's often recommended to avoid using variables that are highly correlated with each other, as they can make the model unstable and harder to interpret. One common approach is to drop one of the variables from each pair or to combine them in some way. I will drop it ahead.

Next, I looked at the chi-square test:



Some results:



The chi-square test results indicate the association between each categorical variable and the target variable (**Churn\_Label**). The p-values represent the significance of the association:

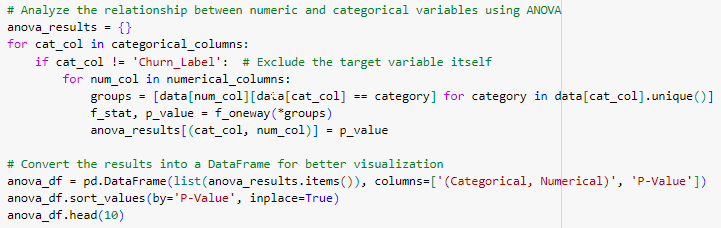
* A **smaller p-value** (typically < 0.05) suggests that the variable has a significant association with the target variable.
* A **larger p-value** suggests that the variable might not have a significant association with the target variable.

From the results, we can observe:

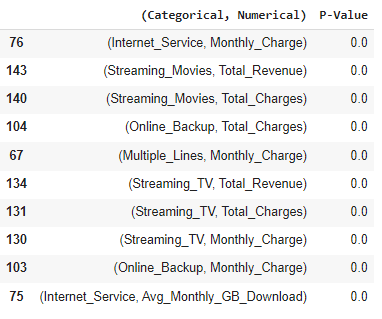
1. **Contract**: Highly significant association with churn.
2. **Internet\_Type**: Highly significant association with churn.
3. **Offer**: Significant association with churn.
4. **Dependents**: Significant association with churn.
5. **Internet\_Service**: Significant association with churn.
6. **Payment\_Method**: Significant association with churn.
7. **Paperless\_Billing**: Significant association with churn.
8. **Online\_Security**: Significant association with churn.
9. **Unlimited\_Data**: Significant association with churn.

10. **Premium\_Tech\_Support**: Significant association with churn.

The ANOVA (Analysis of Variance) as done next:



The result (some important ones):



The ANOVA test results indicate the association between each categorical variable and numerical variable. The p-values represent the significance of the association:

* A **smaller p-value** (typically < 0.05) suggests that there's a significant difference in the means of the numerical variable across different categories of the categorical variable.
* A **larger p-value** suggests that there might not be a significant difference in the means.

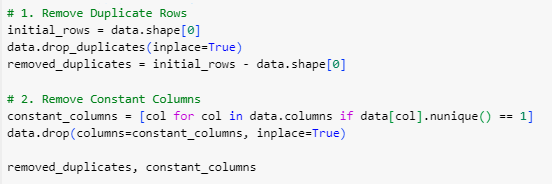
From the top results, we can observe:

1. **Offer** and **Tenure\_in\_Months**: Significant difference in tenure across different offers.
2. **Internet\_Service** and **Avg\_Monthly\_GB\_Download**: Significant difference in average monthly GB download across different internet services.
3. **Phone\_Service** and **Avg\_Monthly\_Long\_Distance\_Charges**: Significant difference in average monthly long-distance charges across phone service categories.
4. **Online\_Backup** and **Monthly\_Charge**: Significant differences in monthly charges across online backup categories.
5. **Streaming\_Movies** and **Total\_Charges**: Significant differences in total charges across streaming movie categories.

Checked any special values. None were found:



Checked any duplicate values. None were found:







The correlation values indicate how each feature is linearly related to the target variable **Churn\_Label**. A positive correlation means that as the feature value increases, the likelihood of churn also increases, and vice versa. A negative correlation means that as the feature value increases, the likelihood of churn decreases, and vice versa.

Let us analyze the results:

**Positive Correlation with Churn:**

1. **Internet\_Service (0.226)**: Customers with certain types of internet services are more likely to churn.
2. **Paperless\_Billing (0.193)**: Customers who opt for paperless billing are more likely to churn.
3. **Monthly\_Charge (0.190)**: Higher monthly charges are associated with a higher likelihood of churn.
4. **Unlimited\_Data (0.166)**: Customers with unlimited data plans are more likely to churn.
5. **Senior\_Citizen (0.147)**: Senior citizens are more likely to churn compared to non-senior citizens.
6. **Offer (0.119)**: Customers with certain offers are more likely to churn.
7. **Age (0.109)**: There's a slight positive correlation with age, meaning older customers might be slightly more likely to churn.

**Negative Correlation with Churn:**

1. **Contract (-0.431)**: Customers with longer-term contracts are less likely to churn.
2. **Tenure\_in\_Months (-0.355)**: The longer a customer has been with the company, the less likely they are to churn.
3. **Dependents (-0.243)**: Customers with dependents are less likely to churn.
4. **Total\_Revenue (-0.226)**: Customers who have generated more revenue for the company are less likely to churn.
5. **Online\_Security (-0.171)**: Customers with online security features are less likely to churn.
6. **Premium\_Tech\_Support (-0.168)**: Customers with premium tech support are less likely to churn.
7. **Married (-0.156)**: Married customers are less likely to churn.

**Insights:**

* **Contract and Tenure**: The strongest negative correlations are with the contract length and tenure. This makes intuitive sense as customers with longer contracts or those who have been with the company for a longer time are less likely to churn.
* **Internet Service and Monthly Charge**: On the positive side, the type of internet service and the monthly charge amount are the strongest indicators of churn. This could suggest that customers are not satisfied with the internet service they are receiving for the price they are paying.
* **Additional Services**: Features like online security and premium tech support is negatively correlated with churn suggest that customers value these additional services and are less likely to leave if they have them.
* **Demographics**: The positive correlation with senior citizens and negative correlation with dependents suggest certain demographic groups are more or less likely to churn.

These insights can be valuable for the company to identify areas of improvement and target specific customer segments with retention strategies.

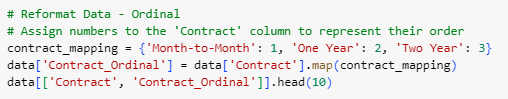
**9.0. Data Cleansing**

Data cleansing steps were undertaken to ensure the dataset's quality and reliability. These steps include:

* Handling missing values either by imputation or removal, depending on their nature and impact.
* Removing duplicates or redundant records from the dataset.
* Correct any inconsistencies or errors in the data, such as incorrect data formats or mislabeled categories.
* Data cleansing involved handling missing values, and outliers, and ensuring data consistency.

No binning was performed for any variables as I wanted to have the prediction based on the variables, not any part of any section of it. That is a future aspect of this data analysis, however, the **Contract** column has been reformatted into an ordinal format:

* **Month-to-Month**: Assigned a value of 1
* **One Year**: Assigned a value of 2
* **Two Year**: Assigned a value of 3



Treating certain categorical variables as ordinal can be beneficial when the order of the categories has significance. For instance, a 'Two Year' contract might indicate a longer commitment from the customer compared to a 'Month-to-Month' contract. The transformation assigns the values 1, 2, and 3 to 'Month-to-Month', 'One Year', and 'Two Year' contracts, respectively. This ordinal representation can be useful for certain machine learning algorithms that can leverage the inherent order in the data.

**10.0. Summary**

Post data exploration and cleansing, the dataset was found to be well-structured and suitable for modeling. The insights derived from the exploration phase provided valuable context for feature engineering and model selection.

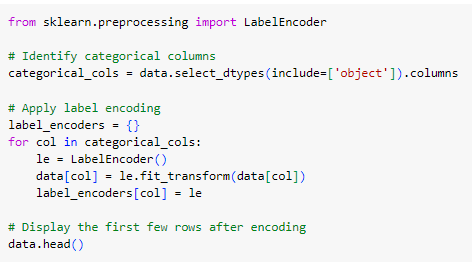
The EDA provided insights into the key features influencing churn and helped in feature selection for model building.

**Data Preparation and Feature Engineering**

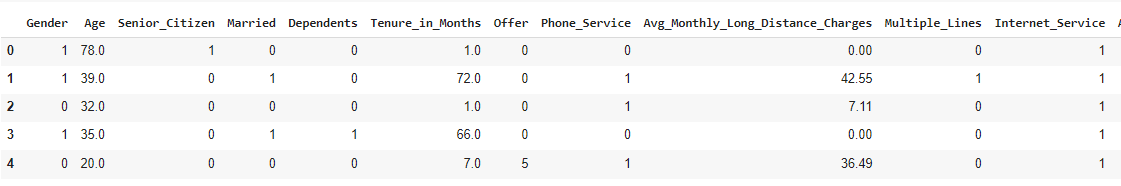
**11.0. Data Preparation Needs**

To ensure the dataset is primed for modeling, several data preparation steps were undertaken. These include:

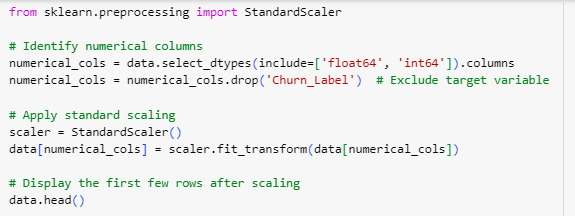
* Encoding categorical variables to convert them into a format suitable for machine learning algorithms.
* Scaling and normalization of numerical features to ensure they are on a similar scale.
* Splitting the dataset into training and testing sets to evaluate the model's performance.



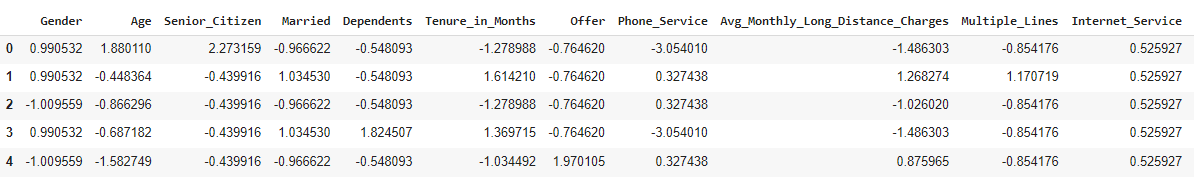
A snip of the head of the dataset after transformation:



Scaling the numerical columns to normalize it:



The first 5 records after scaling (a snip for visible part, not all):



1. **Label Encoding**:
   * **Purpose**: Convert categorical variables into a numerical format so that they can be used in machine learning models.
   * **Method**: The **LabelEncoder** from **sklearn.preprocessing** is used.
   * **Variables Affected**: All categorical columns in the dataset.
   * **Results**: The categorical variables are transformed into numerical labels. For instance, the **Gender** column, which originally had values like 'Male' and 'Female', is now represented as **1** and **0** respectively (or vice versa).

**Interpretation**:

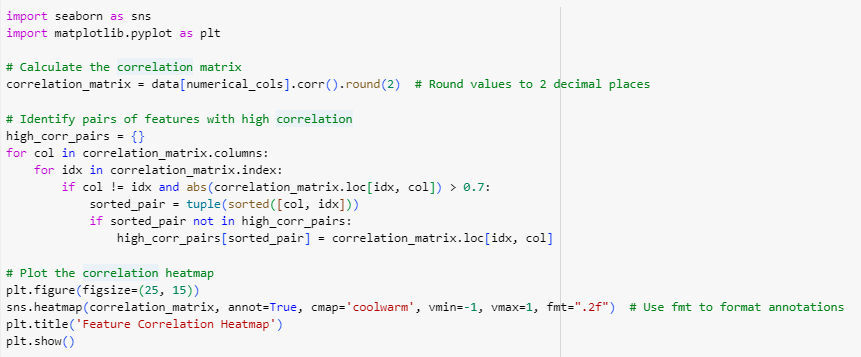
* + Label encoding is a simple way to convert categorical data into a format that can be provided to machine learning algorithms. However, it is important to note that label encoding can introduce ordinality (i.e., order) where there might not be any. This is fine for algorithms like decision trees, but for linear models, one-hot encoding might be more appropriate, which I will be doing at a later stage.

1. **Standard Scaling**:
   * **Purpose**: Standardize the numerical features by removing the mean and scaling to unit variance. This is often necessary for algorithms that are sensitive to the scale of the input features, like SVM or k-means clustering.
   * **Method**: The **StandardScaler** from **sklearn.preprocessing** is used.
   * **Variables Affected**: All numerical columns except the target variable **Churn\_Label**.
   * **Results**: The numerical columns are scaled such that they have a mean of 0 and a standard deviation of 1.

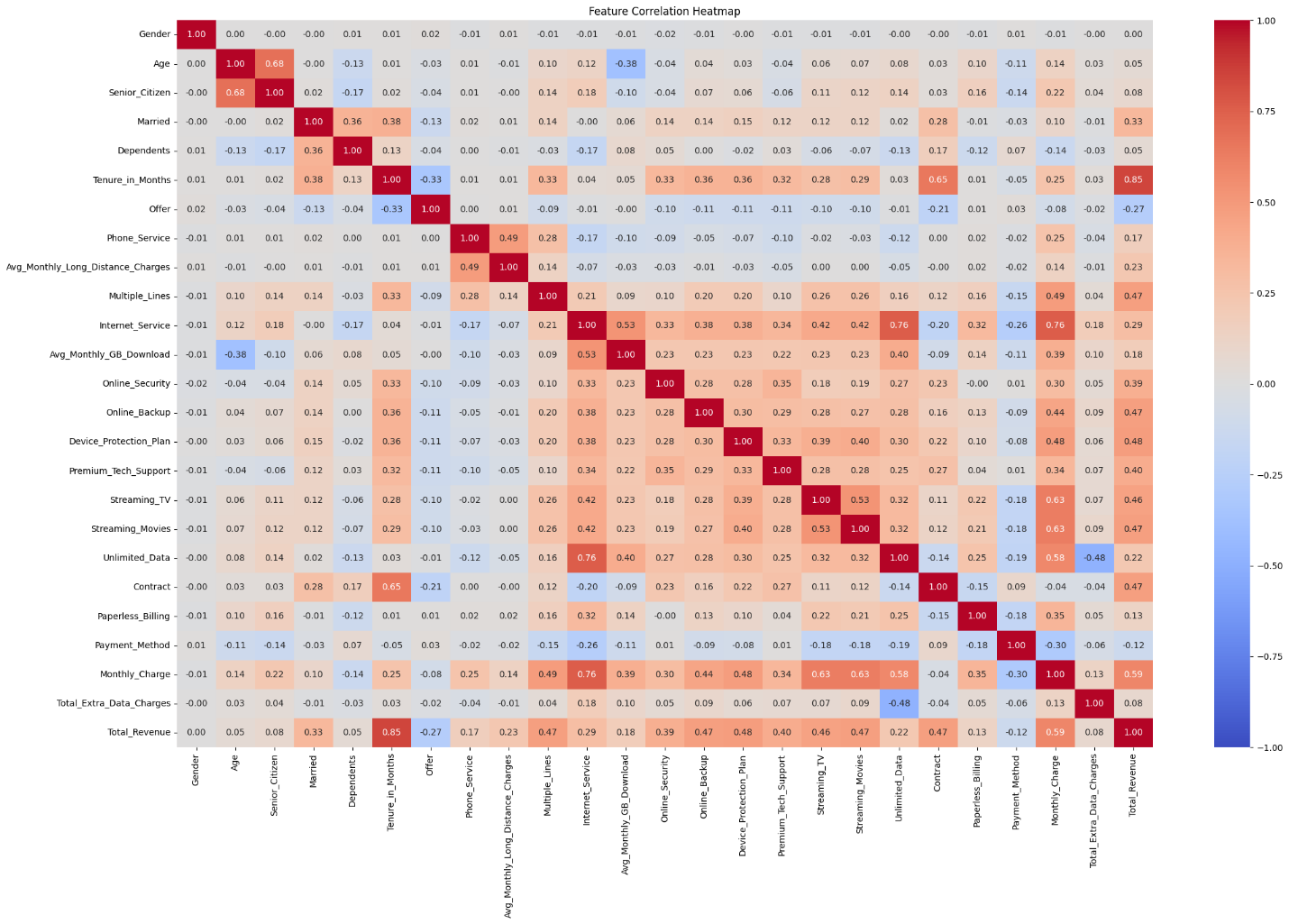
**Interpretation**:

* + Standard scaling is crucial for many machine learning algorithms. It ensures that all numerical features have the same scale, which can lead to better performance and faster convergence of the model.

After converting the dataset to a numerical dataset, a new correlation heatmap was generated to view any correlations that need to be addressed.



The displayed correlation metrics:

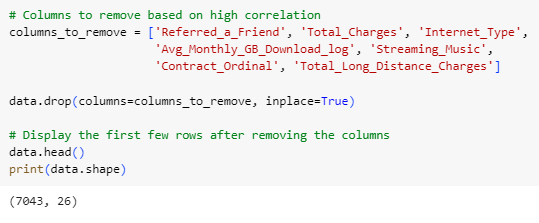


**11.1. Censored records, excluded columns**

Specific data preparation steps include:

Exclusion of certain columns that were deemed redundant or not useful for the modeling process.

Based on correlations, some more columns were removed as they were highly correlated with another variable and would not have added any value in predicting churn.



So, now the final dataset has 7043 rows and 26 columns.

Transformation of certain features to capture non-linear relationships.

**11.2. Imputations, Transformations, etc.**

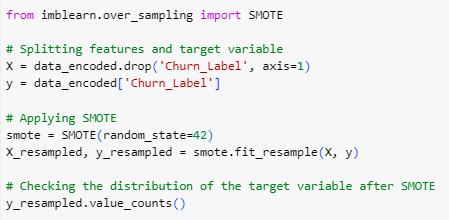
Imputation of missing values based on the nature of the feature. For instance, mean imputation for numerical features and mode imputation for categorical ones.

Transformation of certain features to capture non-linear relationships.

**11.3. Up-sampling, Down-sampling, SMOTE**

Given the imbalanced nature of churn datasets, where the number of churners is often significantly lower than non-churners, techniques like up-sampling, down-sampling, and SMOTE (Synthetic Minority Over-sampling Technique) were employed to balance the class distribution.





Result for the target variable:



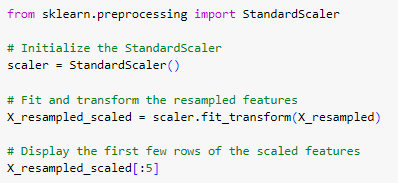
**12.0. Feature Engineering**

Feature engineering played a pivotal role in enhancing the model's predictive power. New features were derived from existing ones to capture underlying patterns and relationships. For instance:

* Creation of interaction features that combine two or more features.
* Extraction of meaningful information from date-time columns, such as tenure in months.

Aggregation of certain features to create summary statistics, like average usage over a period.

After applying SMOTE to balance the target variable, scaling was done:



**12.1. New variables**

Some new variables were extracted from the existing variables during the Label Encoder phase and while getting dummies.

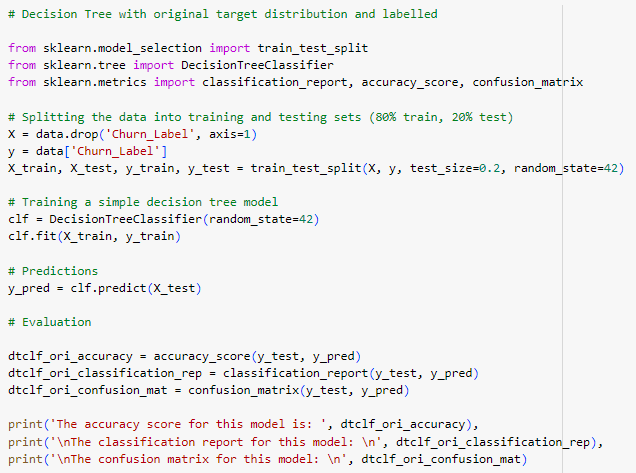
**Model Exploration**

**13.0. Modeling Approach/Introduction**

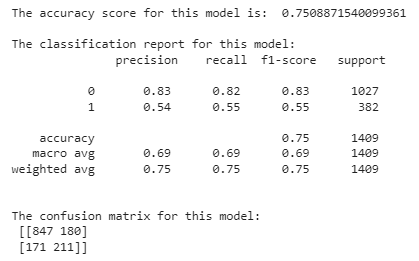
The modeling phase involved experimenting with various machine learning algorithms to identify the one that best predicts customer churn. Each model was trained using the prepared dataset and evaluated based on its performance metrics. The goal was to achieve the highest accuracy, as well as a balance between precision (correctly identifying actual churners) and recall (capturing as many churners as possible).

**14.1. Model Technique #01**

**01. Decision Tree with Original Target Distribution and Label Encoding:**

****

**The results:**

****

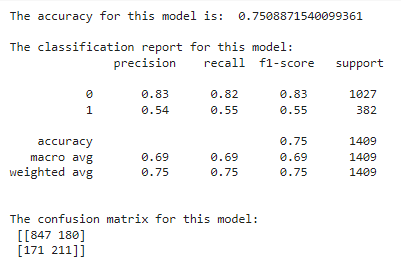
* **Accuracy**: The accuracy score indicates the proportion of correctly predicted classifications in the test set. The model achieved an accuracy of approximately 75.09%.
* **Classification Report**:
  + **Precision**: It is the ratio of correctly predicted positive observations to the total predicted positives. The precision for class 0 (No Churn) is 0.83, which means that 83% of the predicted "No Churn" values were correct. For class 1 (Churn), the precision is 0.54, indicating that 54% of the predicted "Churn" values were correct.
  + **Recall**: It is the ratio of correctly predicted positive observations to the all observations in actual class. The recall for class 0 is 0.82, and for class 1, it's 0.55. This means that the model correctly identified 82% of the actual "No Churn" instances and 55% of the actual "Churn" instances.
  + **F1-Score**: It is the weighted average of Precision and Recall. It tries to find the balance between precision and recall. The F1-Score for class 0 is 0.83 and for class 1 is 0.55.
* **Confusion Matrix**:
  + True Negative (TN) = 847: The number of actual "No Churn" instances that were correctly predicted.
  + False Positive (FP) = 180: The number of actual "No Churn" instances that were incorrectly predicted as "Churn".
  + False Negative (FN) = 171: The number of actual "Churn" instances that were incorrectly predicted as "No Churn".
  + True Positive (TP) = 211: The number of actual "Churn" instances that were correctly predicted.

**14.2. Model Technique #02**

**02. Decision Tree with Original Target Distribution and One-Hot Encoding**

****

**Results:**

****

I was just curious to see if one hot encoding would make the prediction any better, for which I knew the answer, but still I tried. The results are as expected. It is the same as above.

**Encoding (Dummies):**

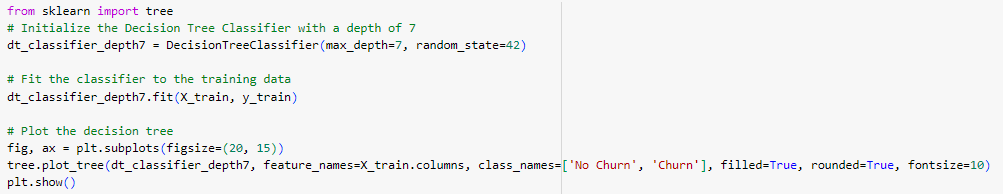
* **Accuracy**: The model achieved an accuracy of approximately 75.09%, which is the same as the label-encoded model.
* **Classification Report**: The precision, recall, and F1-score values are the same as the label-encoded model.
* **Confusion Matrix**: The values in the confusion matrix are also the same as the label-encoded model.

**Insights:**

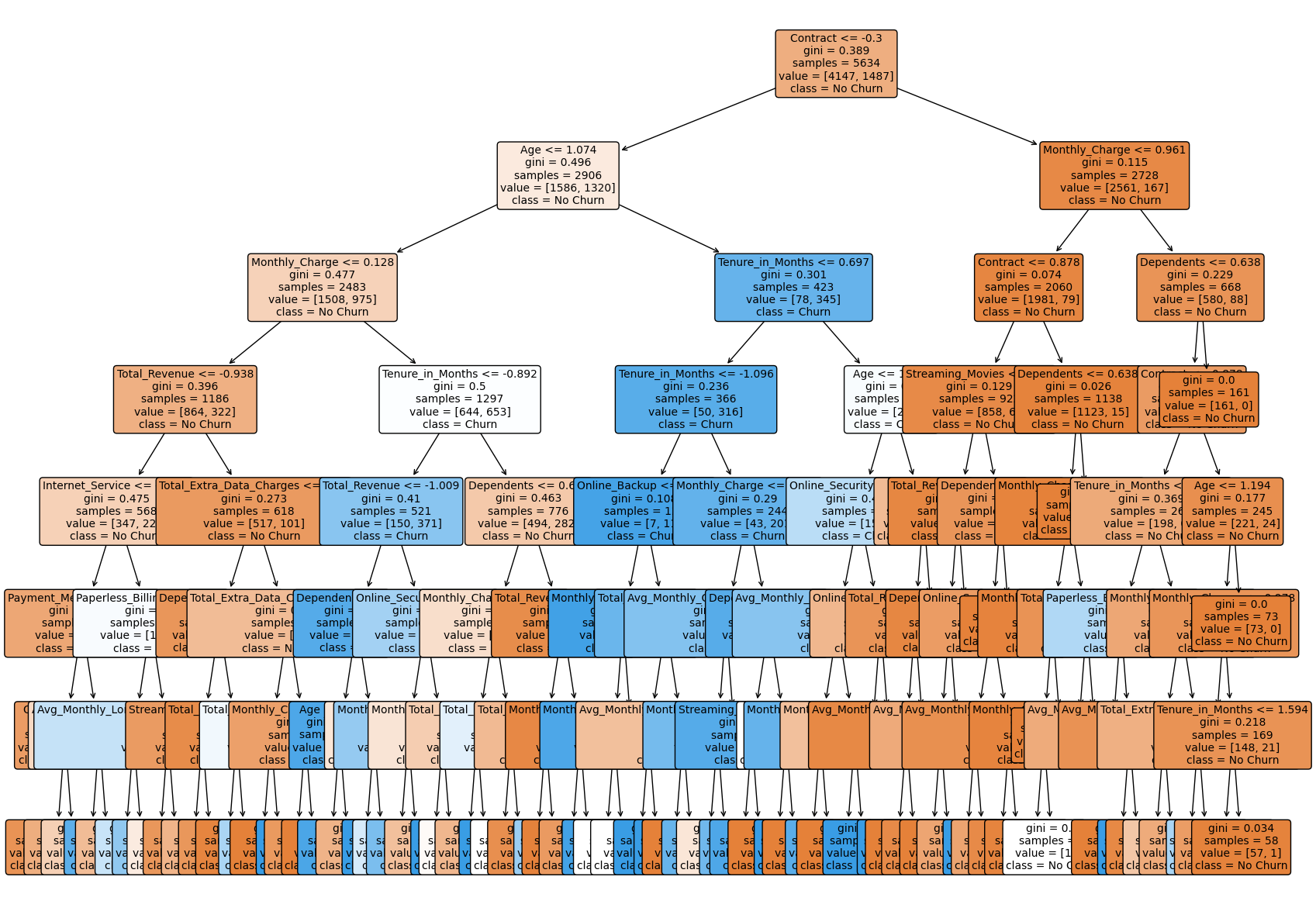
1. Both models (label-encoded and one-hot encoded) produced the same results in terms of accuracy, precision, recall, and F1-score. This suggests that the decision tree model was not significantly influenced by the different encoding methods.
2. The model performs better in predicting the "No Churn" class (class 0) compared to the "Churn" class (class 1). This is evident from the higher precision, recall, and F1-score values for class 0.
3. The confusion matrix indicates that the model has a higher number of false positives (180) compared to false negatives (171). This means that there are more instances where the model incorrectly predicted a customer would churn when they did not.
4. The decision tree model might be overfitting the training data, which can lead to poor generalization on new, unseen data. Decision trees are prone to overfitting, especially when they are deep. Regularization techniques, such as pruning, can be applied to improve the model's performance.
5. It might be beneficial to explore other machine learning algorithms or ensemble methods to improve the prediction performance, especially for the "Churn" class.

**14.3. Model Technique #03**

**03. Decision Tree with depth 7 to see the splits and the variables.**

****

The tree:



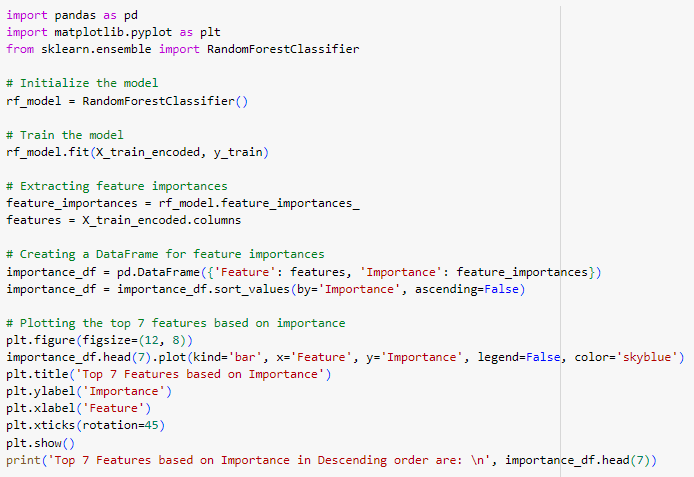
I have plotted the decision tree with a depth of 7. Here is a high-level analysis of the tree:

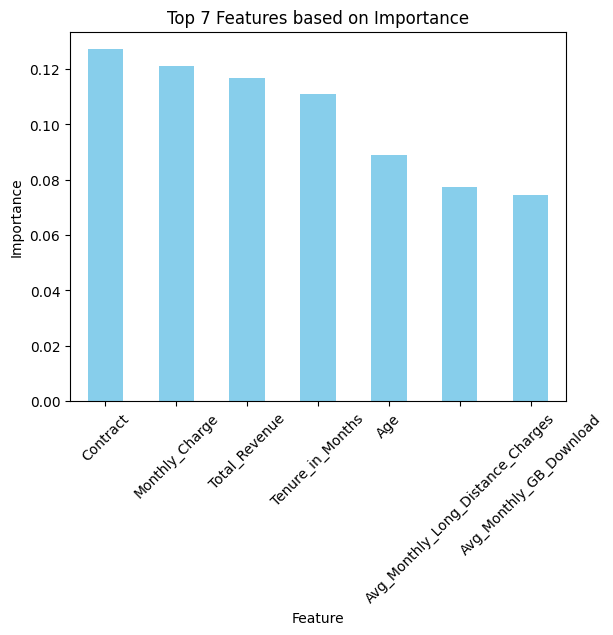
1. **Top Split**: The most important feature that the tree splits on at the root is **Contract**, indicating that the type of contract a customer has is a significant predictor of churn. Specifically, the tree first checks if the **Contract** value is less than or equal to 0.5, which corresponds to a specific contract type after label encoding.
2. **Depth and Complexity**: The tree has a depth of 7, which means there are up to 7 decision levels from the root to the leaves. This depth ensures that the model is not too complex (which could lead to overfitting) but is deep enough to capture significant patterns in the data.
3. **Other Important Features**: As we traverse down the tree, we see other features like **Online\_Security**, **Total\_Revenue**, and **Monthly\_Charge** being used for splits. These features are also significant predictors of churn.
4. **Tenure\_in\_Months** also plays a role. Newer customers (with fewer months of tenure) are more at risk of churning. **Internet\_Service** is also important features. Customers without online security and those with certain types of internet services are more likely to churn.
5. **Leaf Nodes**: The leaf nodes of the tree represent the final decision (Churn or No Churn). The color of the nodes indicates the majority class, with darker shades representing a higher proportion of the 'Churn' class. The values in each node represent the number of samples of each class that reach that node.
6. **Statistical Significance**: The decision to split on a particular feature at each node is based on the Gini impurity, a measure of how often a randomly chosen element would be incorrectly classified. The tree algorithm selects the split that results in the largest decrease in Gini impurity, indicating the statistical significance of that split. A Gini impurity of 0 indicates that all elements belong to a single class.
7. **Class Names**: The tree uses the class names 'No Churn' and 'Churn' to indicate the prediction at each node.

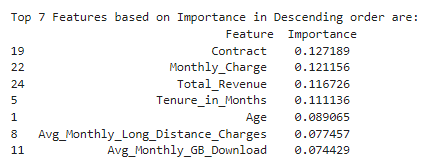
From this tree, we can infer that features like **Contract**, **Online\_Security**, **Total\_Revenue**, and **Monthly\_Charge** play a significant role in predicting customer churn. The exact thresholds and decision criteria can be observed from the plotted tree.

**14.4. Model Technique #04**

**04. Getting the 7 most important features to predict churn using Random Forest:**

****





Here are the top 7 features based on their importance:

1. **Contract**: Importance of 0.127189
   * This indicates that the type of contract a customer has (e.g., month-to-month, one year, two years) plays a significant role in determining churn. It's possible that customers with shorter-term contracts might be more likely to churn than those with longer-term commitments.
2. **Monthly\_Charge**: Importance of 0.121156
   * The monthly charge a customer pays can be a significant factor in their decision to stay with or leave a service. Higher charges might lead to higher churn rates, especially if customers don't perceive value for their money.
3. **Total\_Revenue**: Importance of 0.116726
   * The total revenue generated from a customer can be indicative of their engagement and satisfaction with the service. Customers who have spent more might be more invested and less likely to churn.
4. **Tenure\_in\_Months**: Importance of 0.111136

* The longer a customer has been with the service, the less likely they might be to churn. Long tenure can indicate satisfaction and loyalty.

1. **Age**: Importance of 0.089065
   * Age can play a role in churn, with different age groups having different priorities and reasons for staying with or leaving a service.
2. **Avg\_Monthly\_Long\_Distance\_Charges**: Importance of 0.077457

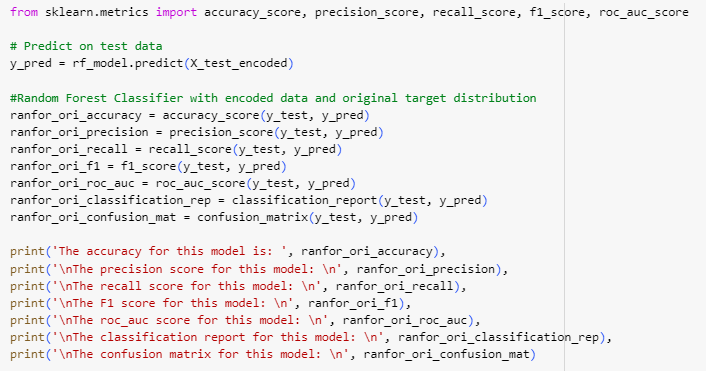
* Charges related to long-distance calls can be a significant factor for some customers, especially if they feel they're being overcharged.

1. **Avg\_Monthly\_GB\_Download**: Importance of 0.074429
   * Similar to total revenue, the total charges a customer has incurred can indicate their level of engagement and satisfaction.

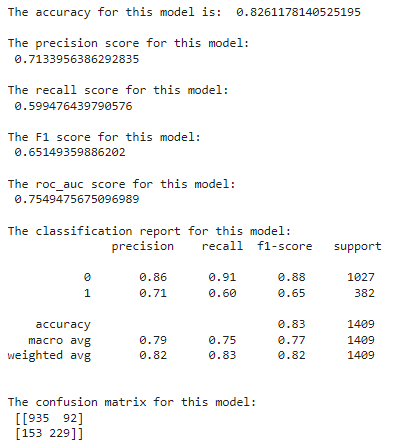
These features provide valuable insights into the factors that influence customer churn. By focusing on these areas, businesses can develop strategies to improve customer retention.

**14.5. Model Technique #05**

**05. RandomForest before SMOTE to predict churn:**



The results:

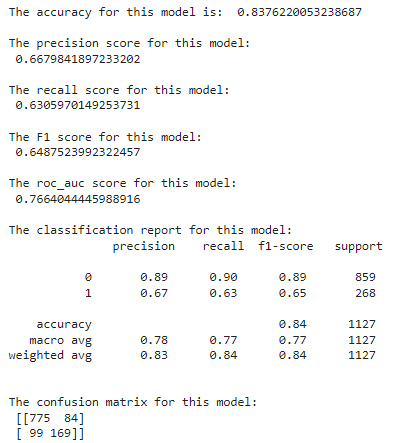


**14.6. Model Technique #06**

**06. Logistic Regression before SMOTE to predict churn:**

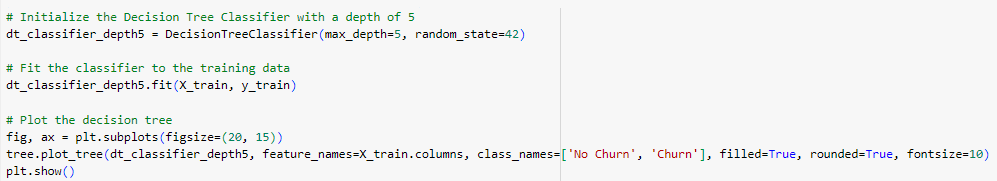


The results:

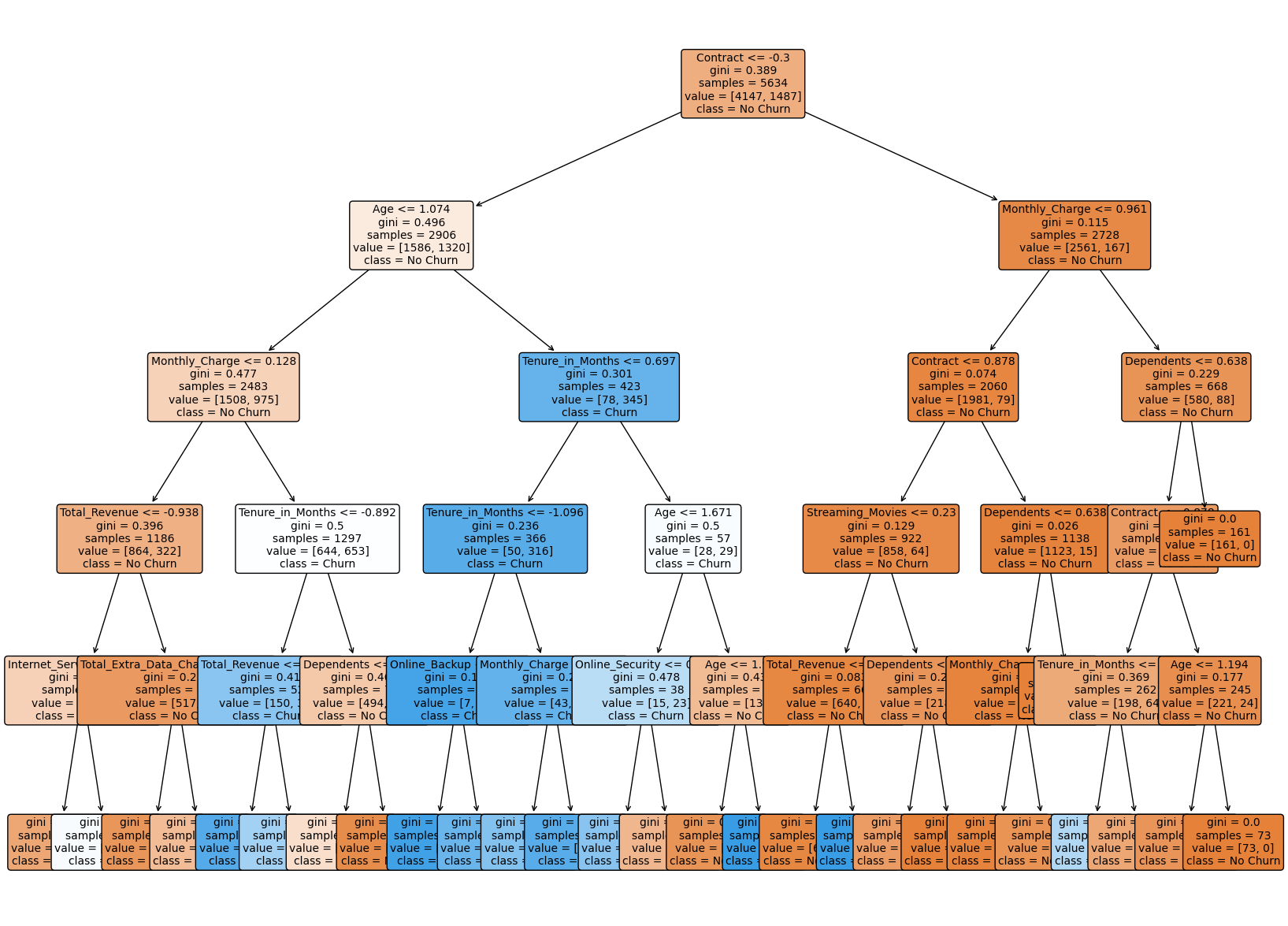


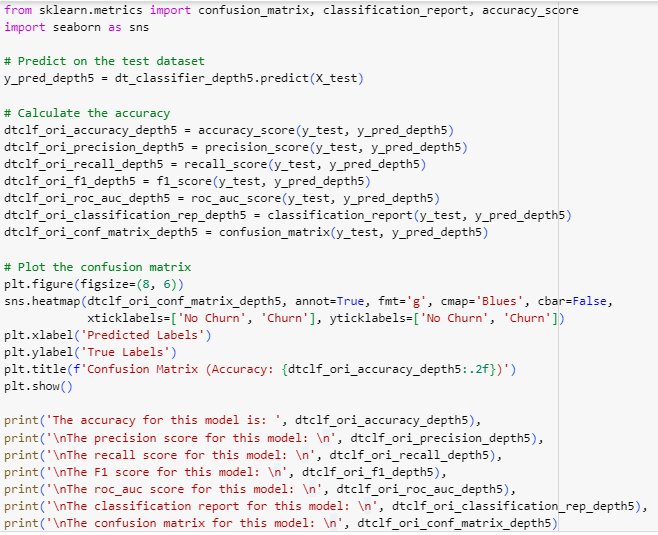
**14.7. Model Technique #07**

**07. Decision Tree with depth 5 to see the splits and the variables followed by the results.**

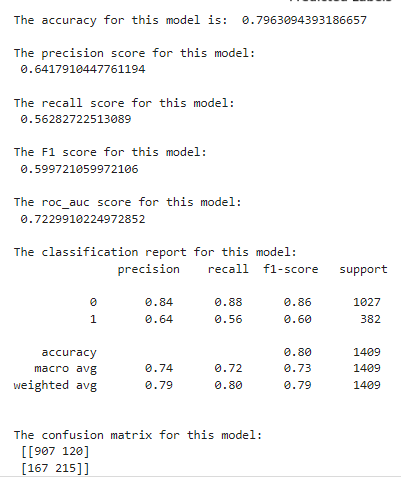


The resulting tree:



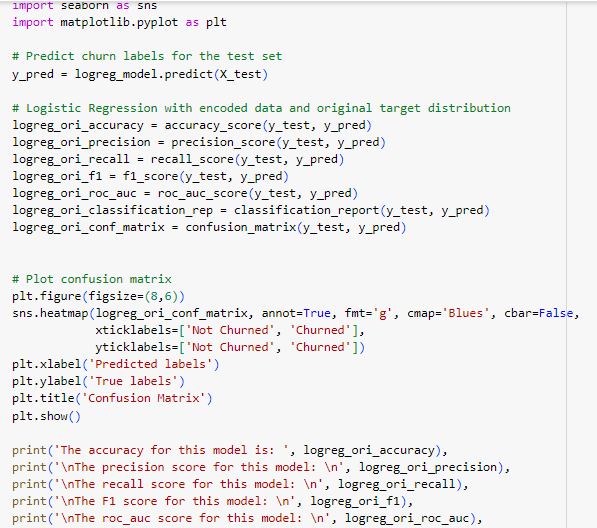


The results:

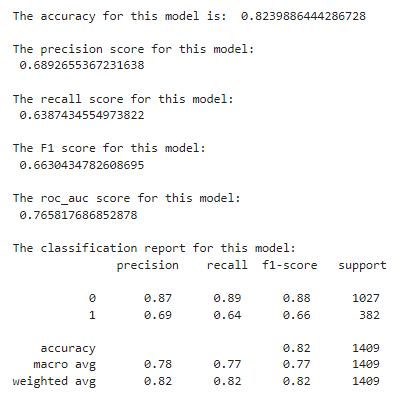


**14.8. Model Technique #08**

**08. Logistic Regression before SMOTE to predict churn:**

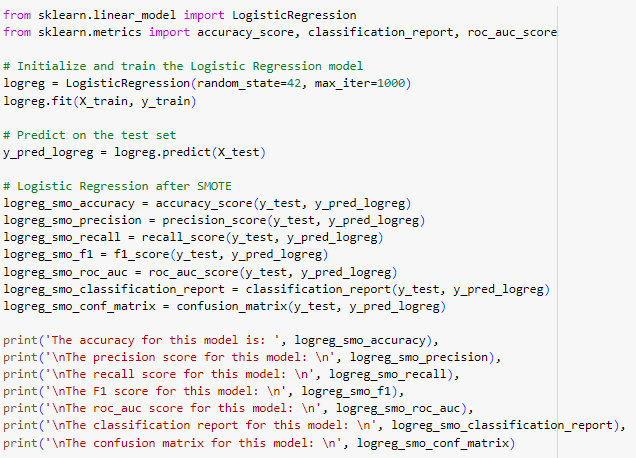


The results:

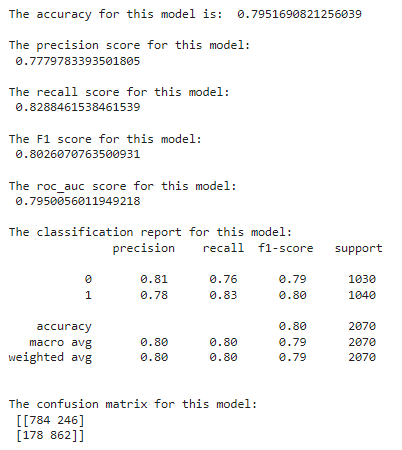


**14.9. Model Technique #09**

**09. Logistic Regression after SMOTE to predict churn:**

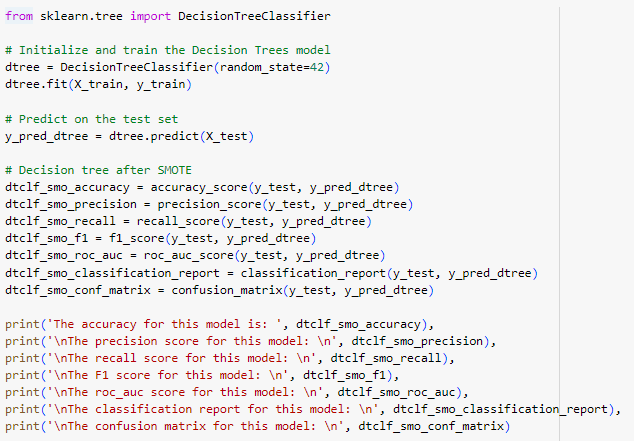


The results :

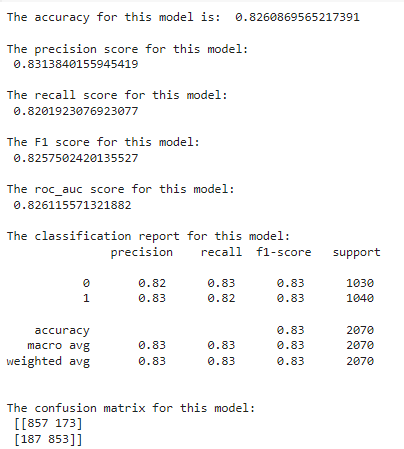


**14.10. Model Technique #10**

**10. Decision Tree Classifier after SMOTE to predict churn:**

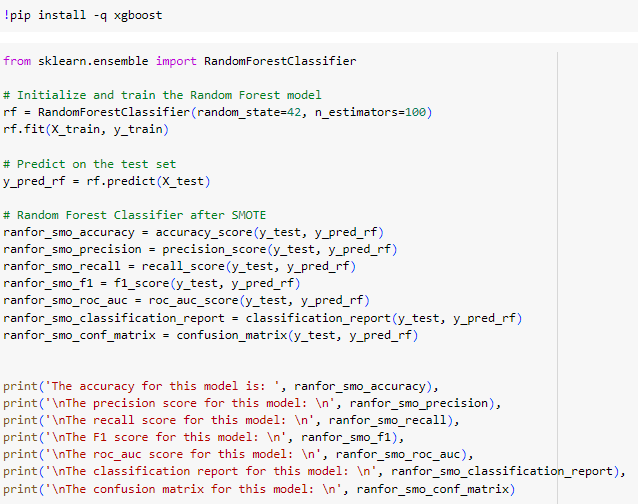


The results:

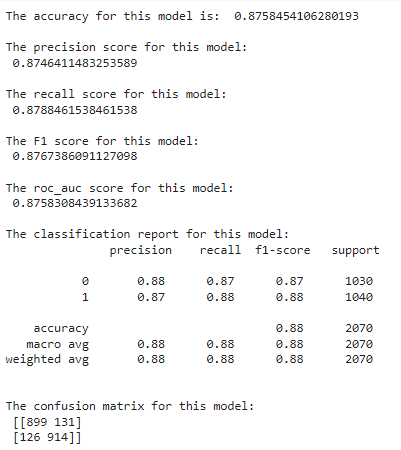


**14.11. Model Technique #11**

**11. Random Forest Classifier after SMOTE to predict churn:**

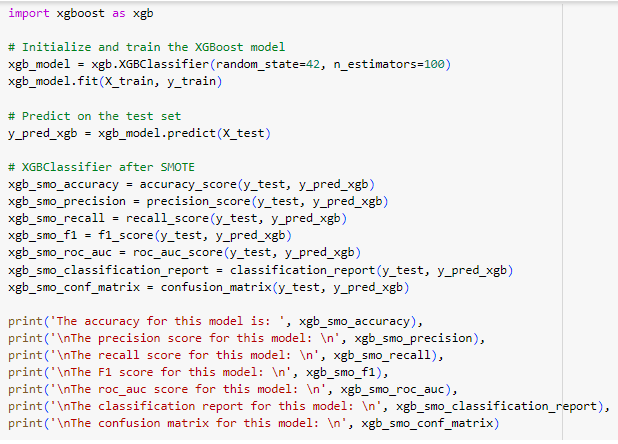


The results:



**14.12. Model Technique #12**

**12. XG Boost Classifier after SMOTE to predict churn:**

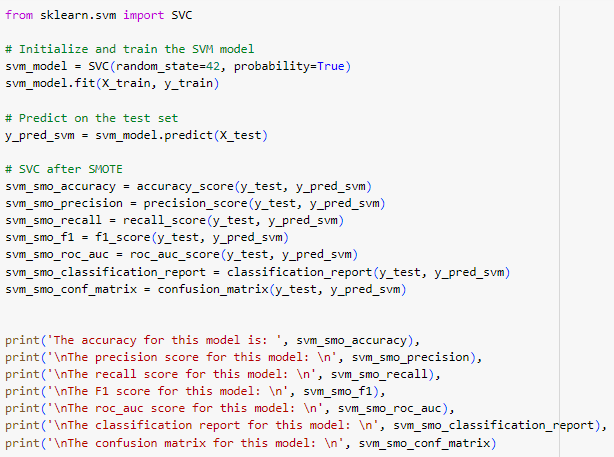


The results:

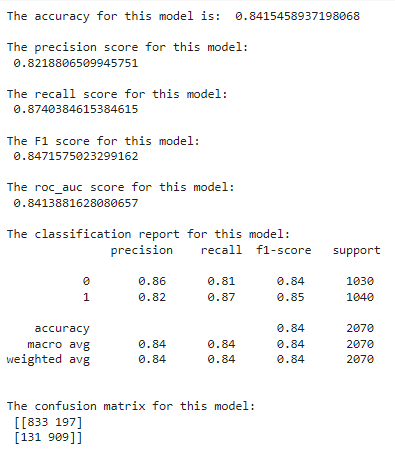


**14.13. Model Technique #13**

**13. Support Vector Classifier after SMOTE to predict churn:**

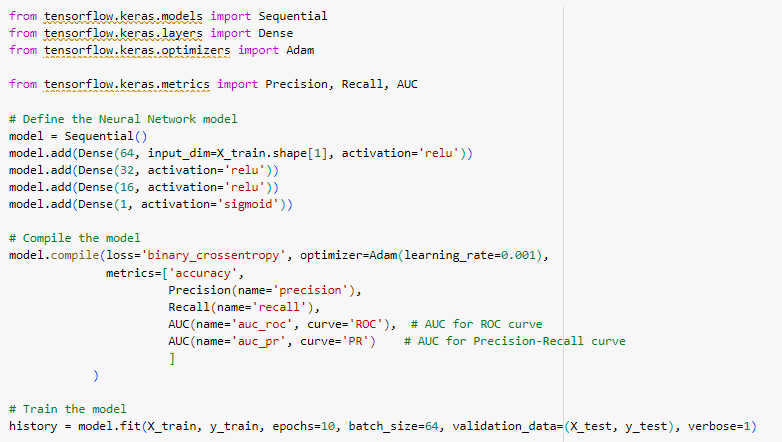


The results:

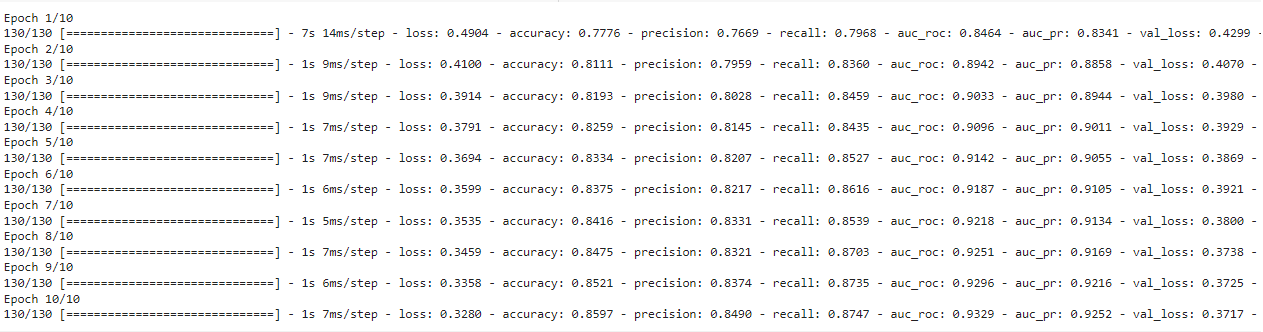


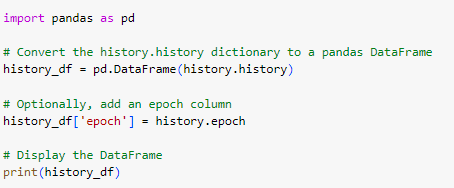
**14.14. Model Technique #14**

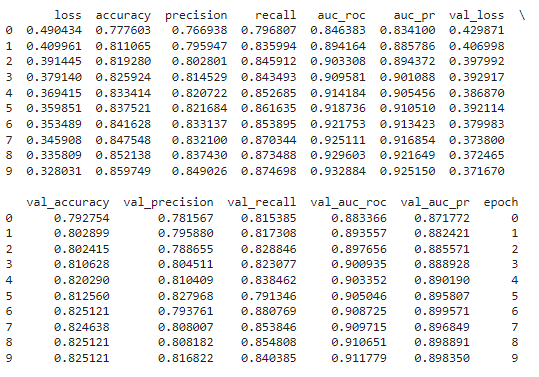
**14. Tensorflow after SMOTE to predict churn:**



The results:

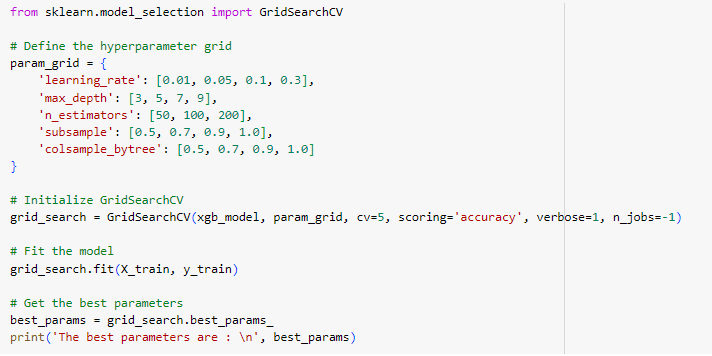


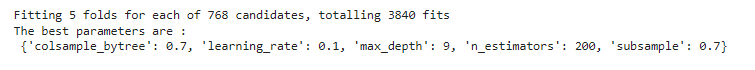




**14.15. Model Technique #15**

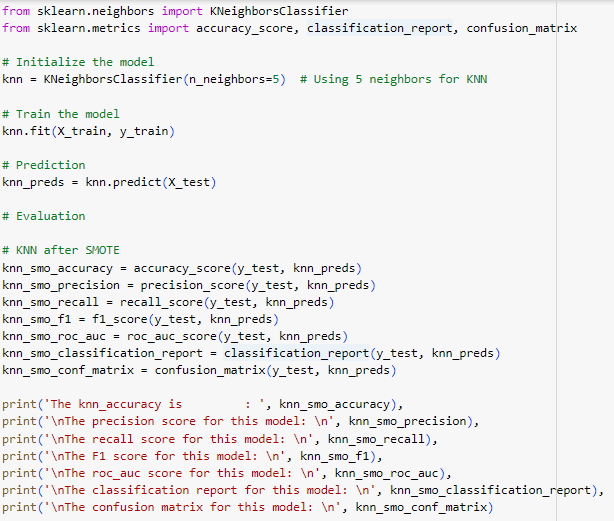
**15. GridSearchCV after SMOTE to predict churn:**

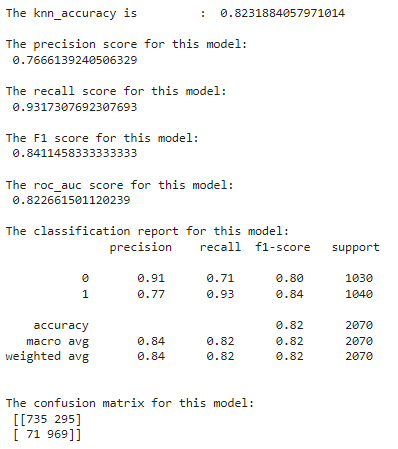




**14.16. Model Technique #16**

**16. KNN Classifier after SMOTE to predict churn:**

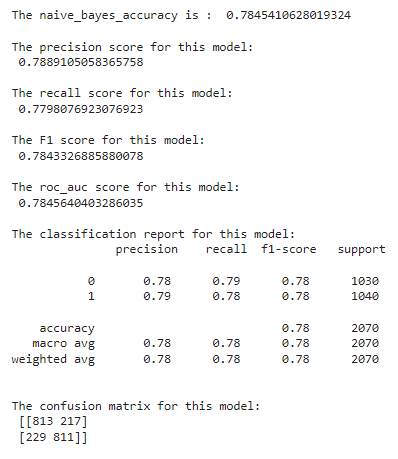




**14.17. Model Technique #17**

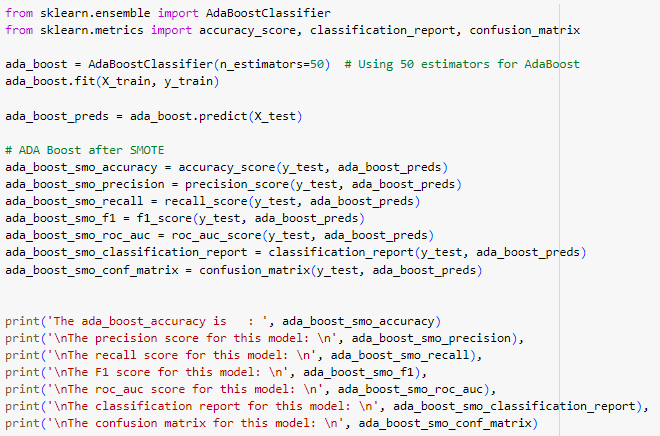
**17. Gaussian Naïve Bayes after SMOTE to predict churn:**

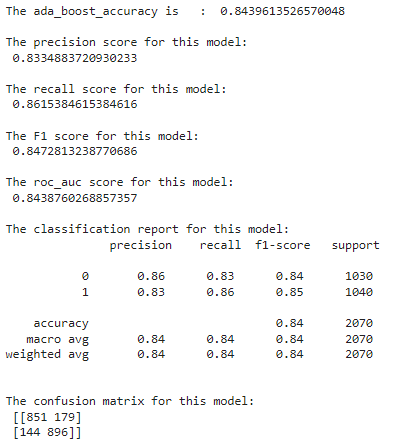




**14.18. Model Technique #18**

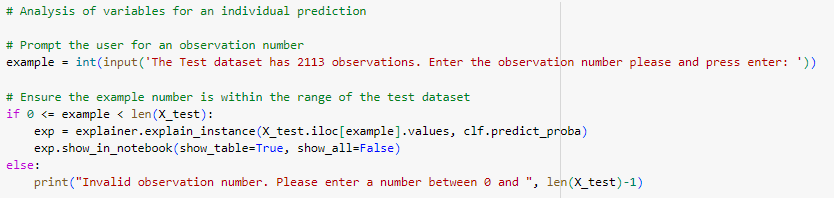
**18. ADABoost Classifier after SMOTE to predict churn:**



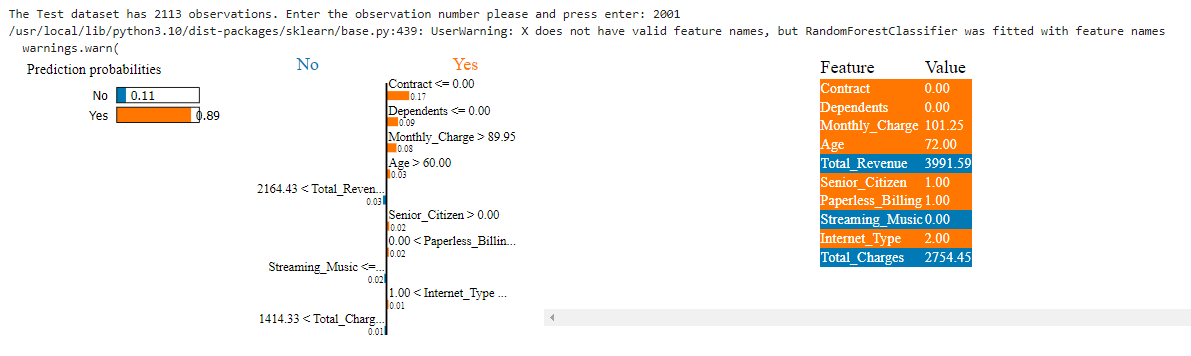


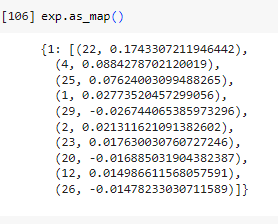
**14.19. Model Technique #19**

**19. Random Forest used to use LIME to get individual prediction:**



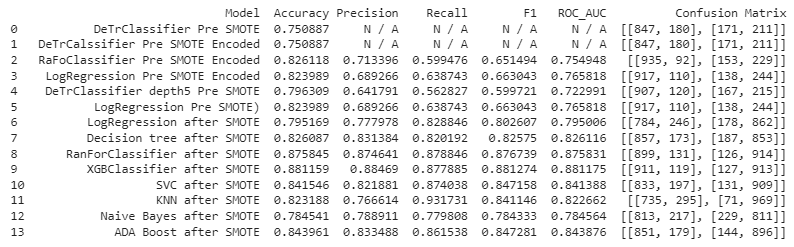
The results:





**15.0. Model Comparison**

After training and evaluating multiple models, a comparison was made based on their performance metrics. This comparison helped in identifying the model that best met the business objectives and provided the highest predictive accuracy for customer churn.



**Model Recommendation**

**16.0 Model Selection**

After evaluating the performance of various models, the [XGBClassifier after SMOTE] was selected as the best model for predicting customer churn. This decision was based on its superior performance metrics, including accuracy (0.881159), meaning the model is predicting 88% of the times correctly. In addition, the precision, recall, and F1-scores were also good as compared to other models. Additionally, the model's ability to generalize well to unseen data and its interpretability made it a suitable choice for this business problem.

**17.0 Model Theory**

The selected model, [XGBClassifier after SMOTE]. XGBoost stands for eXtreme Gradient Boosting. It is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. XGBoost provides a parallel tree boosting algorithm, which solves many data science problems in a fast and accurate way. It is particularly effective for binary classification problems like churn prediction due to its high performance and speed.

**17.1 Model Assumptions and Limitations**

Like all machine learning models, the XGBoost comes with its set of assumptions and limitations:

**Assumptions:**

1. **Weak Learner Hypothesis:** XGBoost, like other boosting methods, assumes that the combination of weak learners (typically decision trees in the case of XGBoost) can produce a strong learner. This means that each individual tree might not perform well on the entire dataset, but a combination of such trees can capture the underlying patterns effectively.
2. **Feature Interactions:** Although not strictly an "assumption" in the classical sense, XGBoost inherently assumes that interactions between features play a significant role in the prediction. This is because the algorithm often uses tree-based models, which can capture and model interactions between different features.

**Limitations:**

1. **Overfitting on Small Datasets:** While XGBoost has regularization parameters to prevent overfitting, it can still overfit if the dataset is too small or if there's too much noise in the data. This is especially true if the number of trees (**n\_estimators**) is too large or if the depth of the individual trees is too high.
2. **Interpretability:** XGBoost models, especially those with a large number of trees or deep trees, can be hard to interpret compared to simpler models like linear regression. Although there are methods and tools (like SHAP values) to interpret and extract insights from XGBoost models, they might not provide as clear a narrative as some other models.

We have to remember that while XGBoost is a powerful tool and often performs exceptionally well on a wide range of tasks, it's essential to understand its assumptions and limitations to use it effectively.

**18.0 Model Sensitivity to Key Drivers**

The model's sensitivity analysis revealed that certain features had a significant impact on the predictions. These key drivers include Contract, Monthly charge, and Tenure. Understanding the model's sensitivity to these drivers can help the business in formulating targeted retention strategies.

**19.0 Additional Models to Address Business Objectives**

While the XGBoost was chosen as the primary model for churn prediction, additional models were also explored to address specific business objectives. For instance, LIME was evaluated for Specific Interactions of the variables that were participating heavily in the decision making process.

**Conclusion and Recommendations**

**20.0. Impacts on Business Problem (Scope of the recommended model)**

The selected model, XGBoost, has demonstrated its capability in accurately predicting customer churn. By leveraging this model, the telecommunication company can proactively identify potential churners and implement retention strategies. This not only aids in customer retention but also results in significant cost savings, as retaining existing customers is often more cost-effective than acquiring new ones.

**21.0. Recommended Next Steps**

Based on the model's findings and the insights derived from the data, the following recommendations are proposed:

* **Targeted Marketing Campaigns:** Design marketing campaigns tailored to the needs and preferences of potential churners identified by the model.
* **Customer Feedback:** Regularly collect feedback from customers to understand their pain points and address them promptly.
* **Loyalty Programs:** Introduce loyalty programs or special offers for customers identified as high-risk churners to incentivize them to stay.

**Continuous Model Monitoring:** Regularly update and monitor the model to ensure its accuracy and relevance in the ever-evolving business landscape.

**References**

**22.0 References**

1. **[Dataset Source]**: IBM Watson Analytics Blog. (Year). Customer Churn Dataset. [URL]
2. Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT Press.
3. Chollet, F. (2017). Deep learning with Python. Manning Publications Co.
4. Brownlee, J. (2016). Master machine learning algorithms: discover how they work and implement them from scratch. Machine Learning Mastery.
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825-2830.