

**INTRODUCTION**

In the current market, there is fierce competition for E-commerce companies and DTH providers (you can select any of these two areas). As a result, it has become difficult to hold onto current clients. As a result, the business is trying to create a model that will allow them to predict account churn and send targeted offers to those who might consider canceling. Because a single account can have several customers, account churn is a big concern for this organization. Therefore, the business may lose more than one customer as a result of losing one account.

**ABOUT: DATASET**

The main columns that are commonly present in PG churn prediction datasets are listed below, along with an explanation of their importance:

Each client record in the dataset is uniquely identified by its customer ID (unique identifier).

Details on the demographics:

Age: Can reveal a customer's changing wants and preferences as they get older.

Gender: May have an impact on how a product is used or how responsive it is to promotions.

Location (nation, area, or city): The behavior of customers and pricing policies can be influenced by geographic considerations.

Details of the Subscription:

Start Date of Subscription: The amount of time since the start of the subscription can indicate trends in churn risk.

Subscription Plan (basic, premium, free trial, etc.): The likelihood of cancellation may differ throughout plans.

Payment Method (debit, credit, etc.): The ease of use or complexity of the payment method may have an impact on customer attrition.

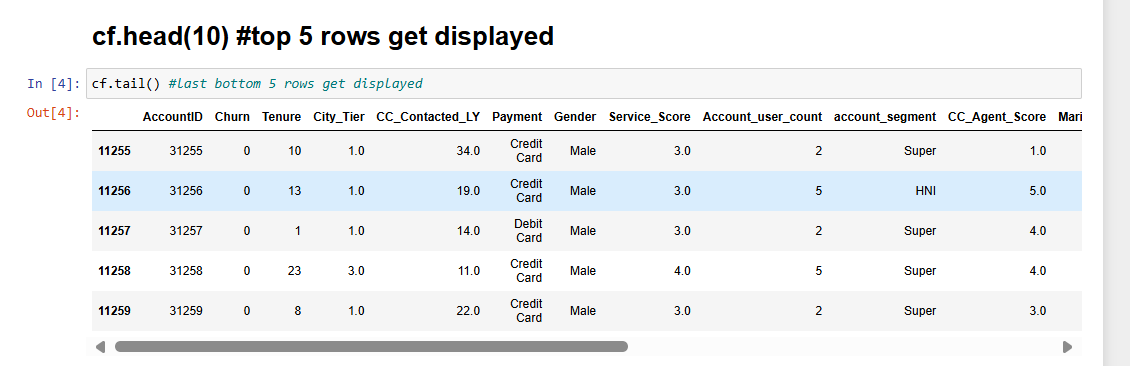
Length of Contract (monthly, yearly, etc.): Durations of contracts have an impact on exit points and churn probability.

**How to import data set**

****

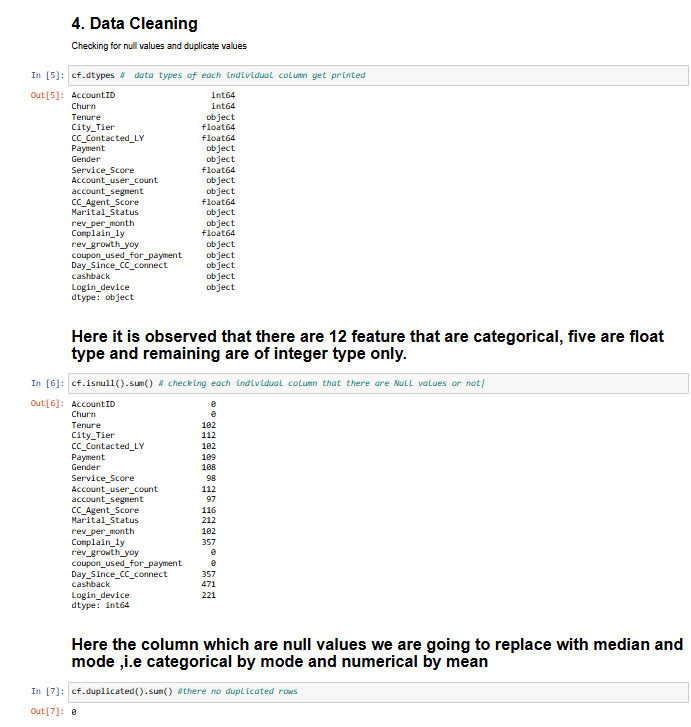
**HOW TO FETCH TOP AND LAST ROWS**

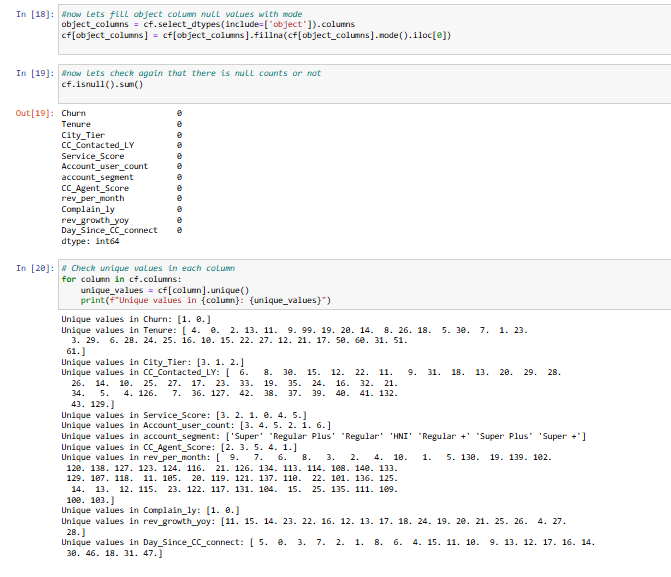
By the use of head() and tail() function



**Data cleaning**

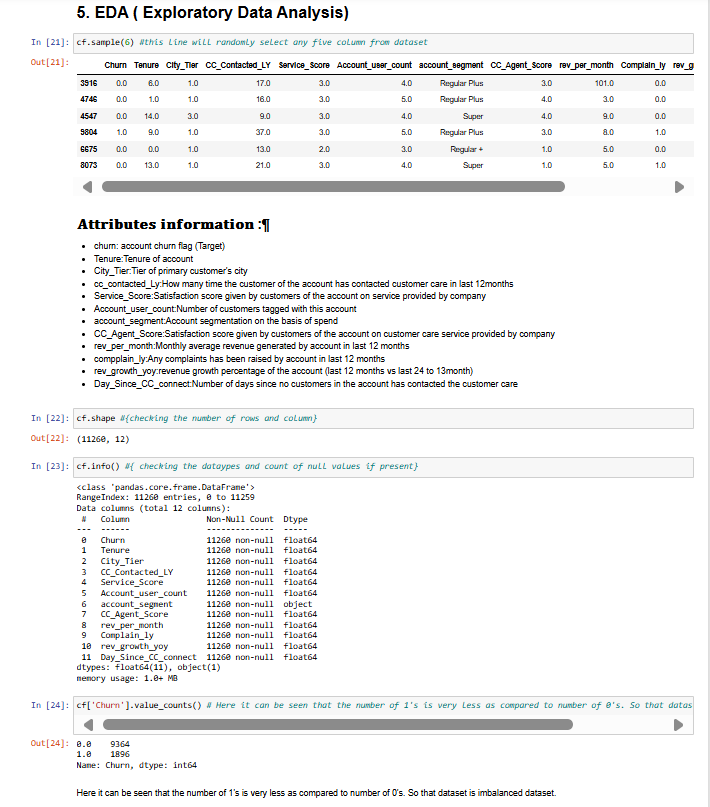
* Common Data Quality Issues
* Missing Values: Instances where data points are absent.
* Duplicate Entries: Repeated data entries that can skew analysis.
* Inconsistent Data: Data that is not uniform, such as different formats for dates.
* Outliers: Data points that deviate significantly from the rest of the dataset.
* Incorrect Data: Errors in data entry that lead to invalid data points.
* Irrelevant Data: Data that does not contribute to the analysis or model.

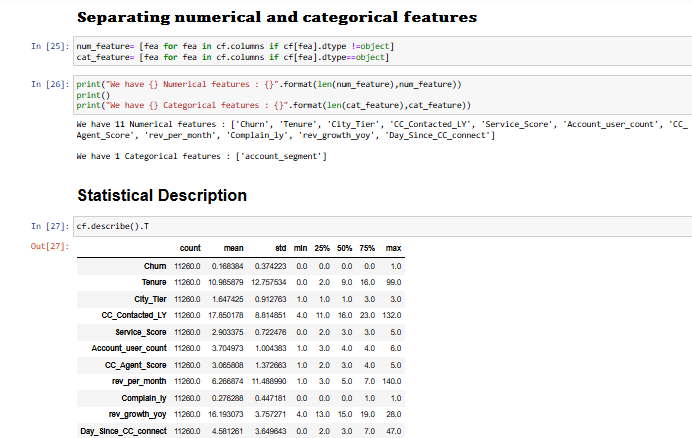
****

****

**Exploratory Data Analysis**

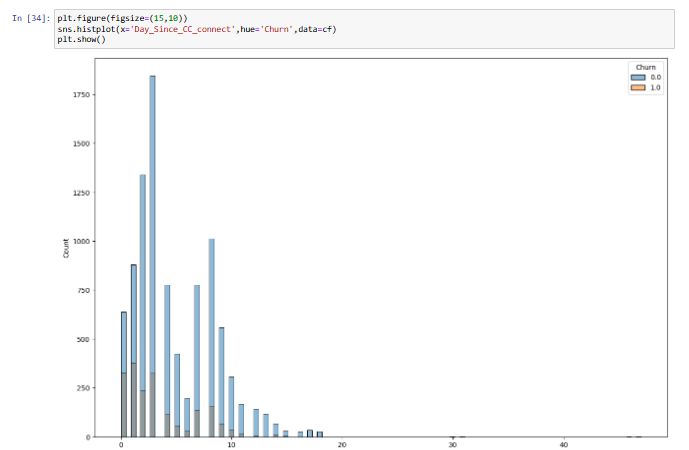
* The significance of EDA
* Data Understanding: Offers a thorough understanding of the dataset, covering its relationships, primary features, and structure.
* Finding trends, patterns, and outliers that might not be immediately obvious is made easier with the aid of insight generation.
* Assumption Validation: Verifies the validity of the underlying assumptions in statistical models.
* Selecting features: Assists in determining the most pertinent variables to include in a model.
* Data cleaning: Makes it easier to find errors and abnormalities that should be fixed before doing more analysis.



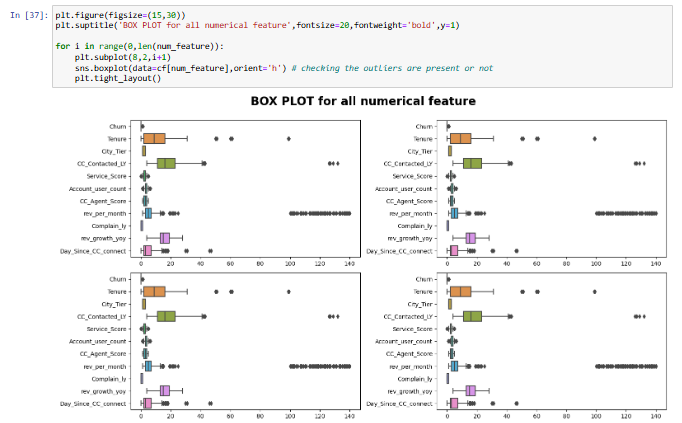


**ANALYSIS with the help of graph**

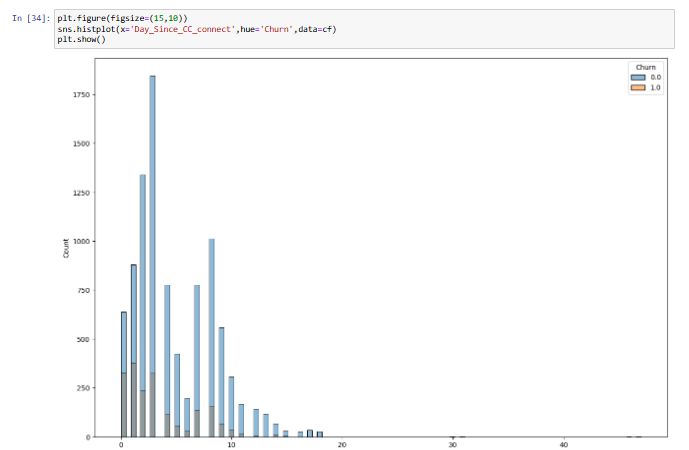
* Histograms:  
    
  A single numerical variable's frequency distribution should be displayed.helpful in comprehending the distribution and central tendency of the data.



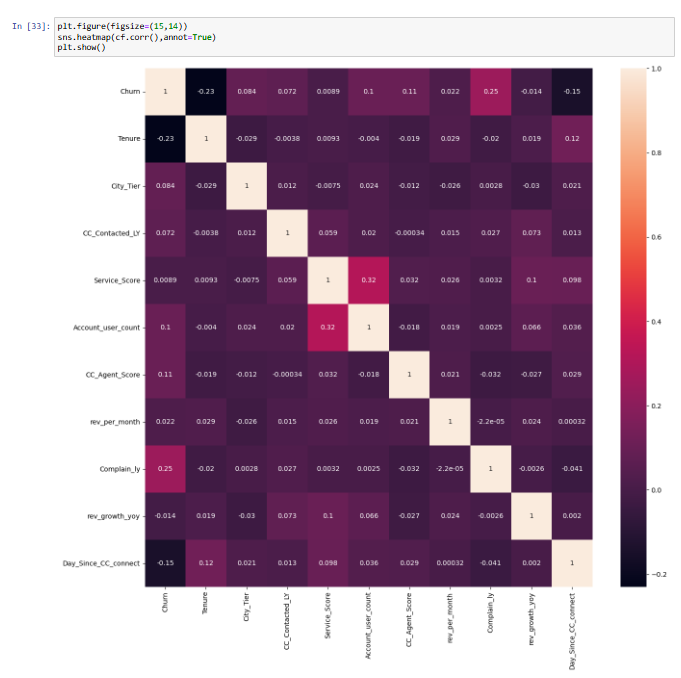
* Plots in boxes:  
    
  Show a dataset's variability, central value, and distribution.  
  efficient in comparing distributions between groups and identifying outliers.



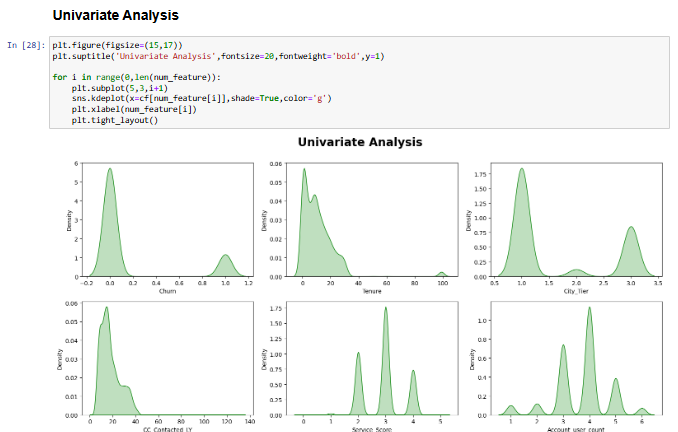
* Bar Diagrams:  
    
  Use rectangular bars to show categorical data.  
  helpful in comparing the number or frequency of groups.



* Heat maps:  
    
  Make use of color gradients to visualize the variables' correlation matrix.  
  Excellent for identifying robust correlations and patterns among several variables.

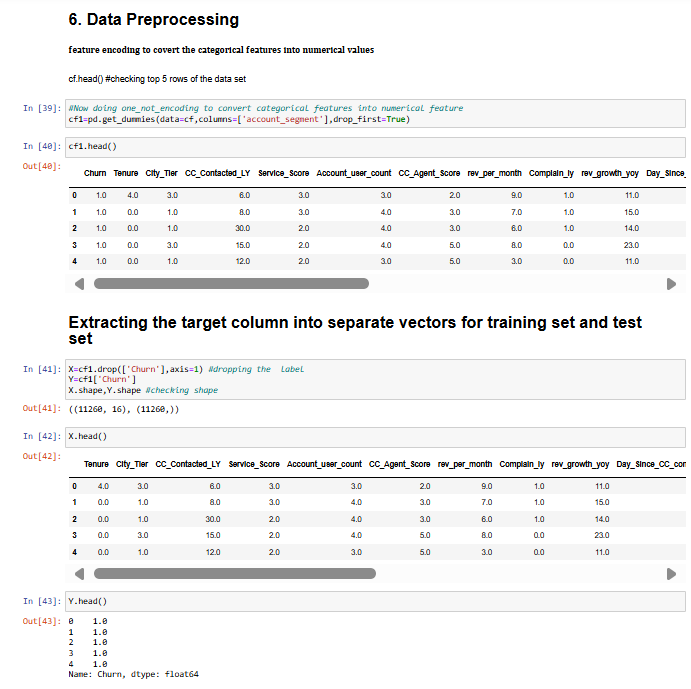


* Line Diagrams:  
    
  Monitor a variable's changes over time.  
  beneficial for identifying cycles, trends, and seasonal patterns in time-series data



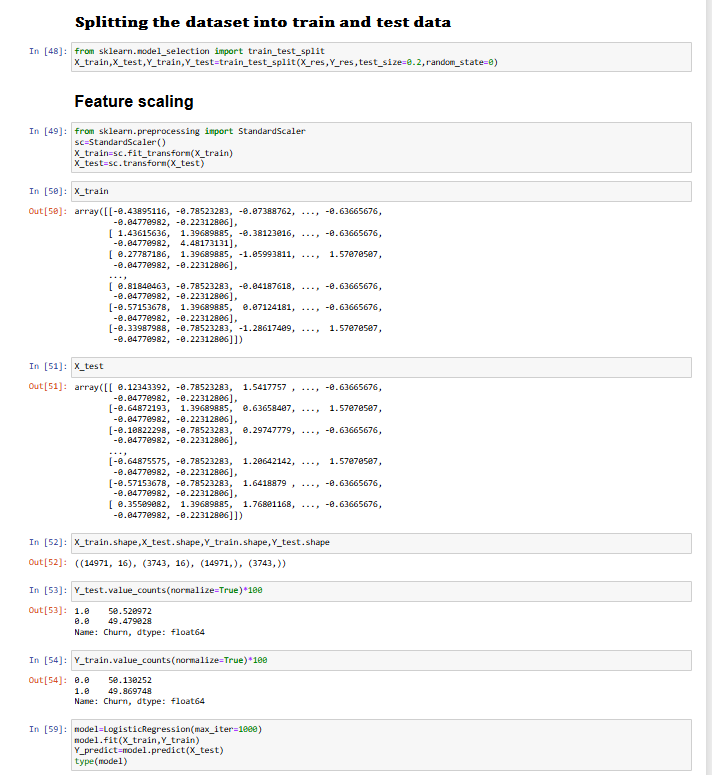
**DATA PRE-PROCESSING**

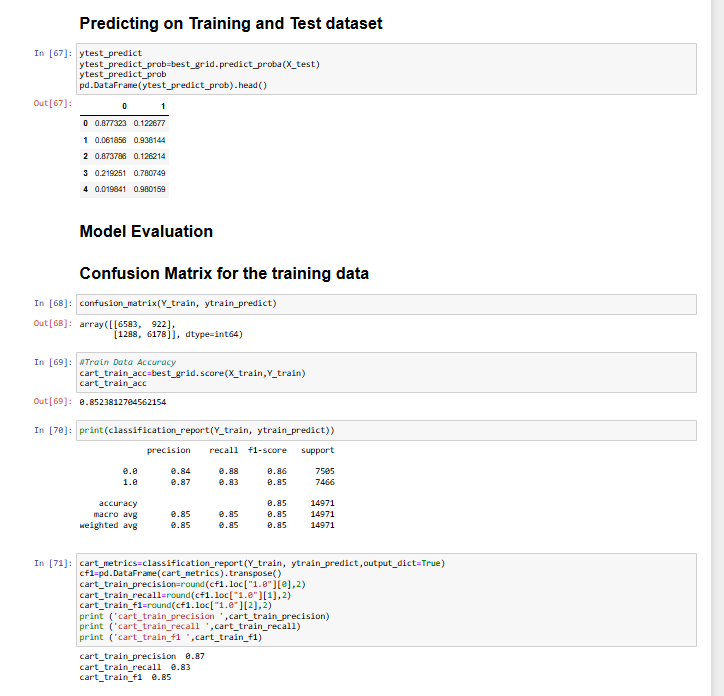
* The Value of Preparing Data
* Enhanced Data Quality: Assures that the information is accurate, dependable, and clean.
* Improved Model Performance: Results in improved machine learning model training.
* Effective Analysis: Reduces complexity of the data, facilitating manipulation and examination.
* Decreased Variance and Bias: Aids in dataset balancing and error reduction.



**NOW ABOUT TRAIN AND TEST OF MODEL**

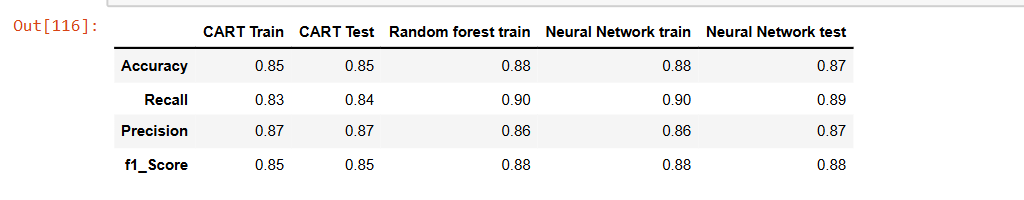
* Train-Test Split Model Validation is important because it offers a strong framework for evaluating the model's capacity to generalize to new data.  
  Performance Metrics: Provides the ability to compute metrics on unseen data, including accuracy, precision, recall, and F1-score.  
  Preventing Overfitting: Assesses the model on an independent test set to assist in identifying and reducing overfitting.
* Steps for Data Splitting in Train-Test Split: splitting the dataset, usually in an 80-20 or 70-30 ratio, into training and testing sets.  
  Model Training: Using the training set, which includes most of the data, the model is constructed.  
  Model testing is the process of assessing a model's performance using data from the test set that was not observed by the model during training.



****

**Precision, Recall, and F1 Score in Brief**

* Precision:  
    
  The ratio of accurately predicted positive observations to the total number of predicted positives is known as precision.  
  It gauges how accurate positive predictions are, which is crucial when the expense of false positives is substantial.
* Recall:  
    
  The ratio of accurately predicted positive observations to all actual positive observations is known as recall.  
  It gauges how well the model can find every pertinent instance—a critical function when it comes to avoiding costly false positives.
* F1 Points:  
    
  The F1 score strikes a balance between recall and precision by taking the harmonic mean of the two.  
  It offers a solitary statistic for assessing the model in situations where recall and precision are equally significant.

****

**CONCLUSION**

Based on the provided output from the dataset, here is an interpretation of the accuracy scores for different models:

1. Decision Tree (CART) Model:

- Train Data Accuracy: The decision tree model achieved a training accuracy of approximately 86.96% (0.8696). This indicates that the model correctly predicted the churn outcome for about 86.96% of the training data.

2. Random Forest Model:

# - Train Data Accuracy: The random forest model achieved a training accuracy of around 85.76% (0.8576). This suggests that the model correctly predicted the churn outcome for approximately 85.76% of the training data.

3. Neural Network Model:

- Train Data Accuracy: The neural network model achieved a training accuracy of approximately 87.73% (0.8773). This indicates that the model correctly predicted the churn outcome for around 87.73% of the training data.

- Test Data Accuracy: The neural network model achieved a test accuracy of approximately 87.54% (0.8754). This suggests that the model correctly predicted the churn outcome for approximately 87.54% of the test data.

Based on above observation our data set is good fit and good model

Based on these results, the neural network model seems to have the highest accuracy both on the training and test data, indicating its better performance compared to the decision tree and random forest models. However, it is important to consider other factors such as model complexity, interpretability, and potential overfitting when selecting the best model for your specific churn prediction task.