DATA ANALYST WITH COGNOS

**TEAM MEMBER**

**411421104022:JOY SLESSOR I**

Phase -4: DEVELOPMENT PART-2

Project: Customer Churn Prediction In Telecommunication



Introduction:

* Customer attrition is one of the biggest expenditures of any organization. Customer churn otherwise known as customer attrition or customer turnover is the percentage of customers that stopped using your company's product or service within a specified period.
* Telecommunication Churn Prediction is a machine learning project which focuses on predict customer churn in the telecommunication industry. Churn prediction is important for telecommunication companies because it helps them to retain their customers and reduce customer acquisition cost figure below depicts the description of the data set importation and screenshot of same.

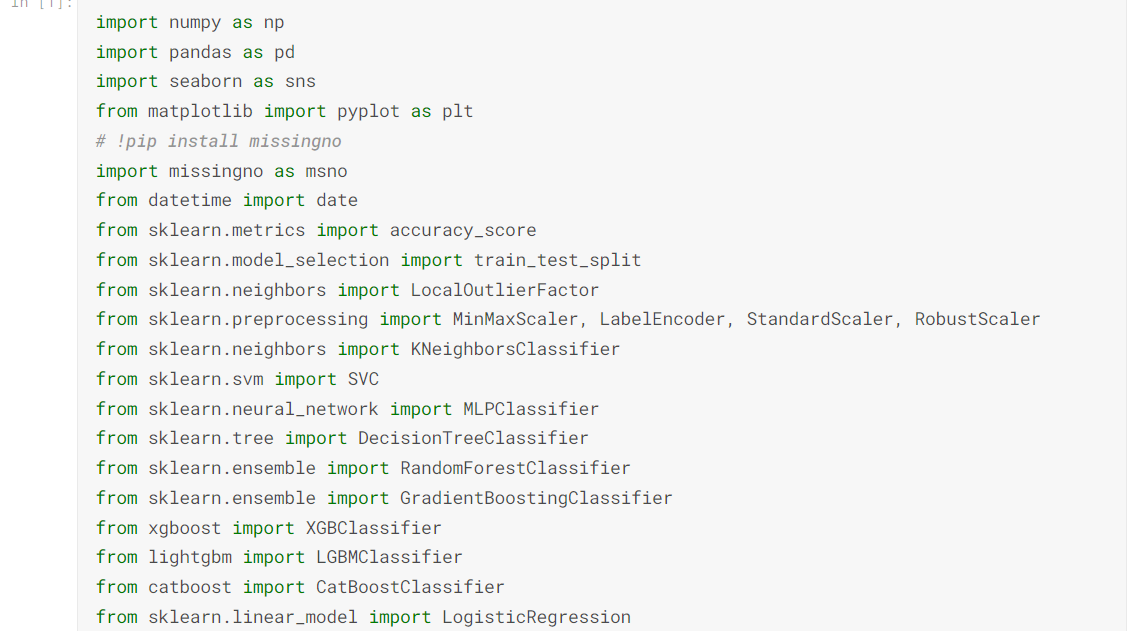
Problem statement:

Telecommunication companies (telco) have a significant problem with customer churn, which is the loss of customers who stop using their services. To solve this problem, telco need to identify the customers who are likely to churn and take preemptive measures to retain them. Machine learning models can help telco in predicting the customers who are most likely to churn, based on various factors such as customer usage patterns, payment history, and demographics.

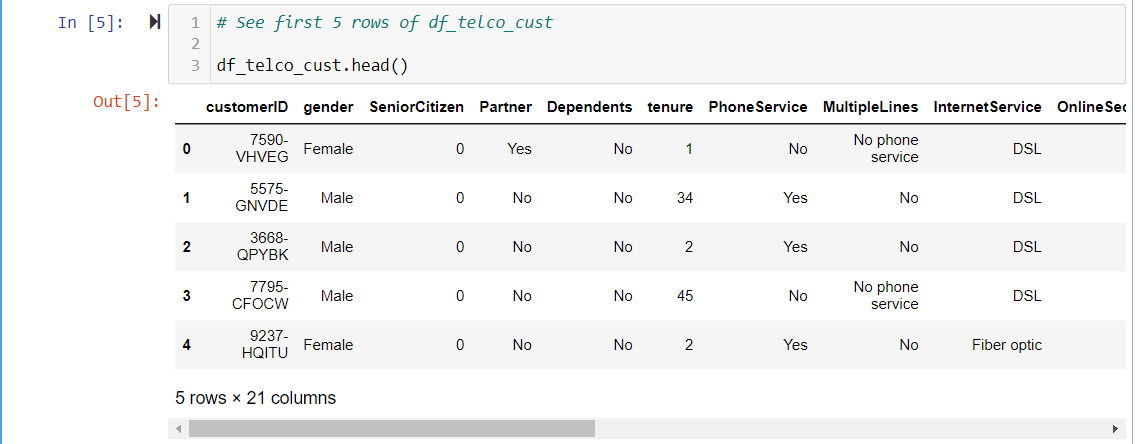
Data Cleaning:

The data set used in this project contains 21 columns and 7,043 rows. The first step in the data cleaning process was to remove any duplicate rows in the data set. After removing duplicates, there were 21 columns and 7,032 rows left in the data set.

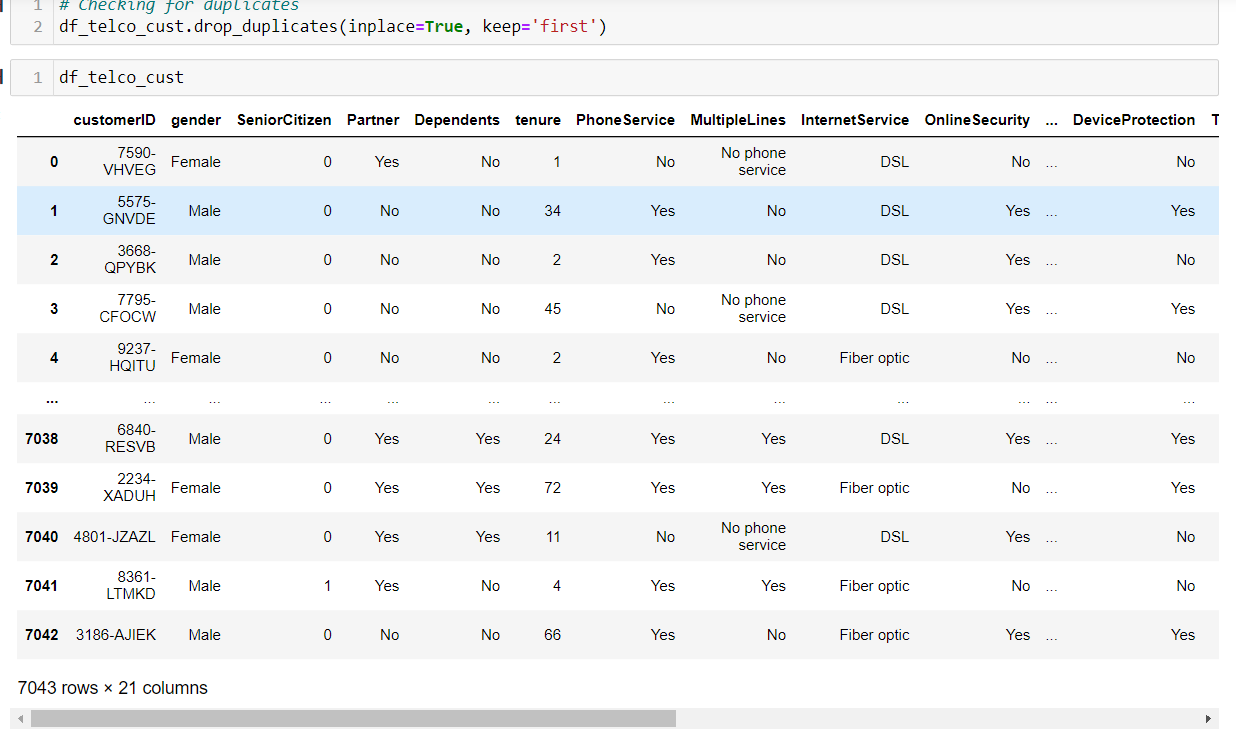
**DATA PREPROCESSING**







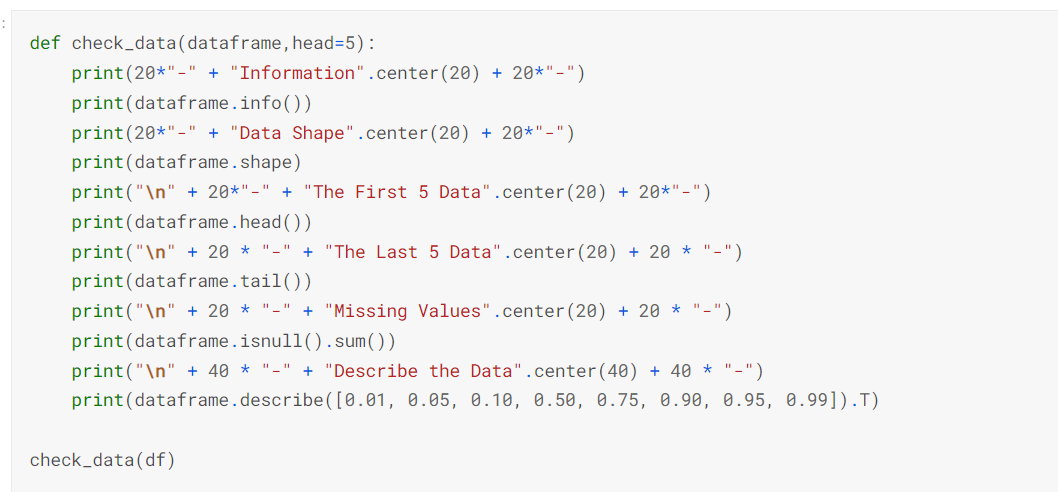
The next step was to check for missing values. There were no missing values in the columns. Interestingly there were no duplicates in the dataset when checked. Below represents the action.



**Data Analysis**

* After cleaning the dataset, the next step was to perform exploratory data analysis (EDA). EDA is important because it helps to understand the distribution of the data and identify any trends or patterns in the data.
* The first step in EDA was to analyze the distribution of gender of the target variable ‘Churn’. it was revealed the distribution is almost equally with a difference of 67 count of male than female.
* The ratio between the Senior Citizens to Non-Senior Citizens was about 1:5. With Non-Senior Citizens constituting about 84% of the

whole distribution. Thus, most of the customers in the dataset are younger people.



* About 50% of the customers have a partner, while only 30% of the total customers have dependents.
* Interestingly, among the customers who have a partner, only about half of them also have a dependent, while other half do not have any dependents.
* Additionally, as expected, among the customers who do not have any partner, a majority (90%) of them do not have any dependents.

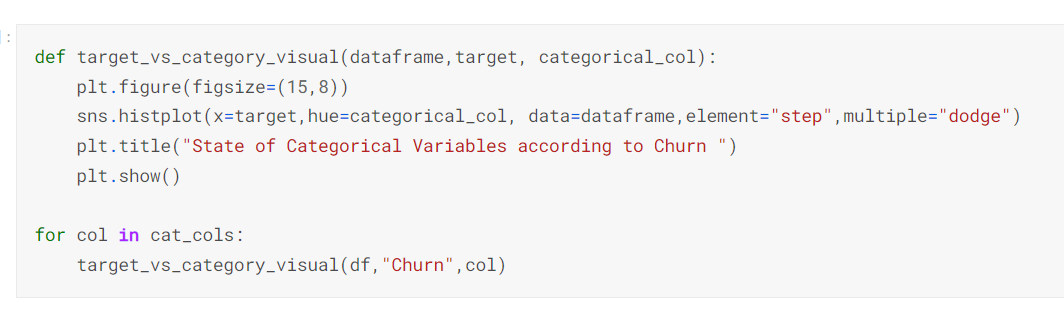
**Overall Churn Rate**

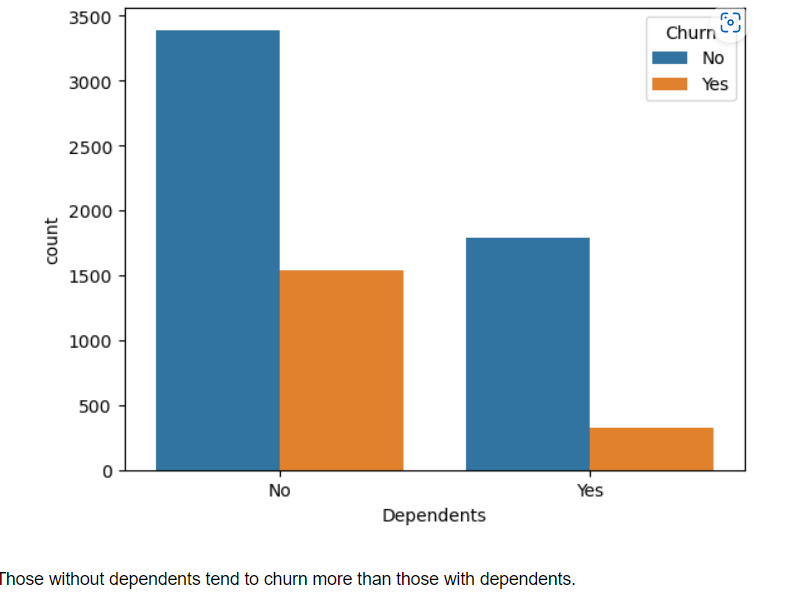
In our data, 74% of the customers do not churn. Clearly the data is skewed as we would expect a large majority of the customers to not churn. This is important to keep in mind for our modelling as skewed-sens could lead to a lot of false negatives. We will see in the modelling section on how to avoid newness in the data.

**1.Explore the churn rate by dependent, seniority, payment method, monthly**

**charges and total charges to see how it varies by these variables.**

**A.Churn rate by Dependent**

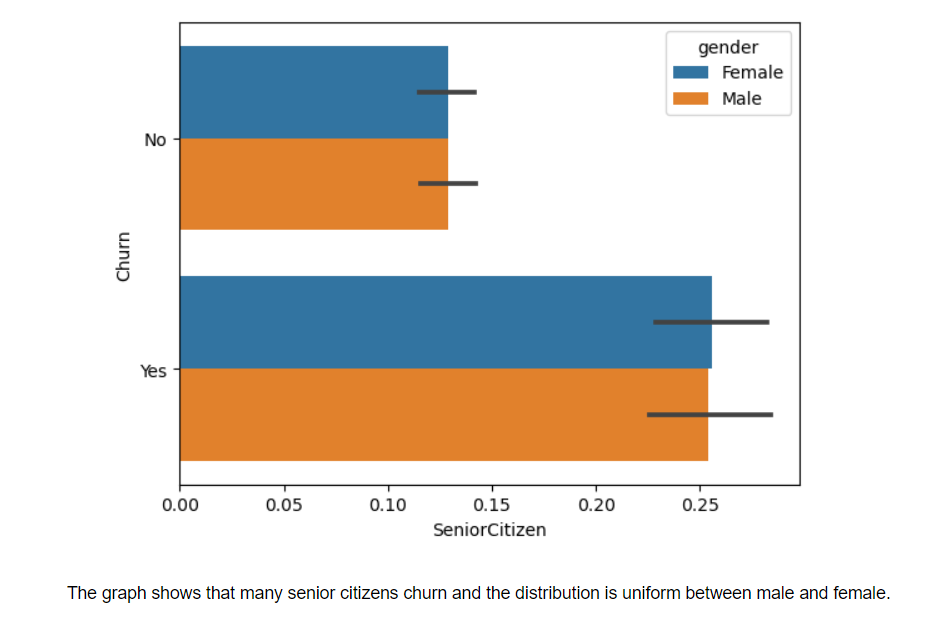




**Churn Rate By Dependent**

As we can see from the above plot, the customers who do not have dependent tend to churn more than those with dependents.

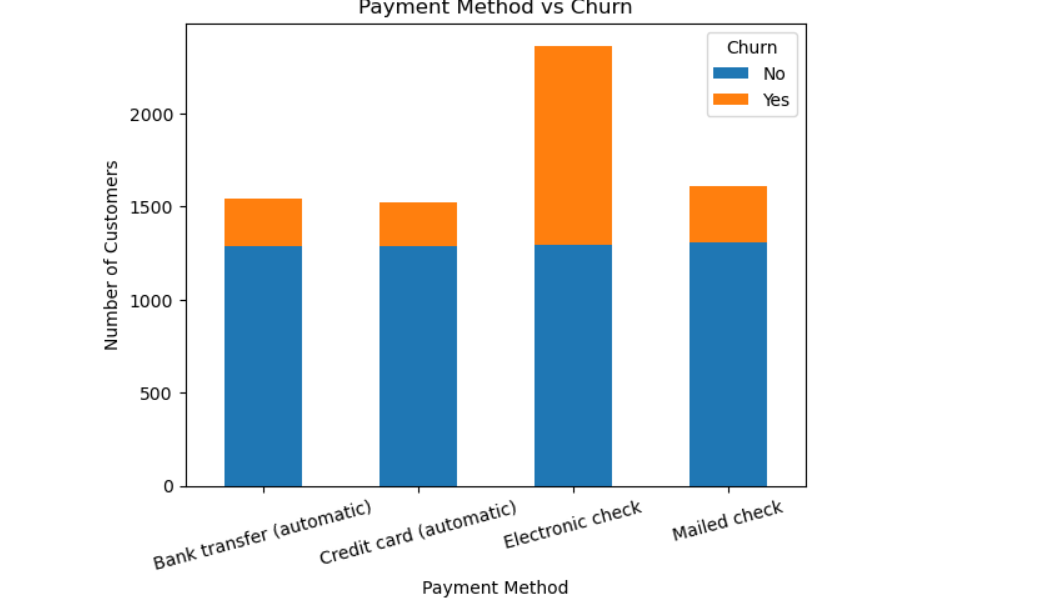
**b. Churn by Seniority**



**Churn by Seniority**

The graph shows that many senior citizens churn and the distribution is uniform between male and female.

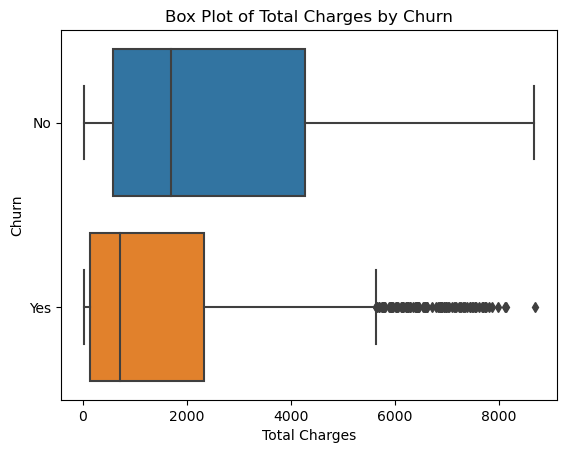
**c. Churn by Payment Methods**



**Churn by Payment Methods**

From the stack-bar chart above, it appears that those who churn most are those with electronic check as a payment method.

**d. Churn by Total Charges**

****

It seems that customers who tend to churn have their median Total Charges of about 500 but with several outliers beyond 5000.

**Hypothesis Testing**

Customers who have been with the company for a longer tenure are less likely to churn compared to customers with shorter tenure.Three set of hypothesis testing were used but one is selected for this article.

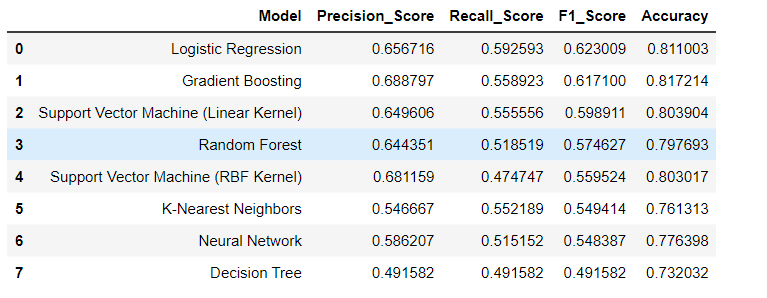
Null: Tenure of a customer does not affect churn rate.

**Alternative:**

* Customers have been with the company for a longer tenure are less likely to churn compared to customers with shorter tenure.
* If the churn rate for this subset is significantly higher than the overall churn rate, then it indicates that customers in that subset are more likely to churn.
* To test the above hypothesis, we used code to create a subset of the data based on the conditions mentioned in the hypothesis and calculate the churn rate for this subset.
* We ignore the Null hypothesis since the churn rate for the subset is significantly lower than the overall churn rate.

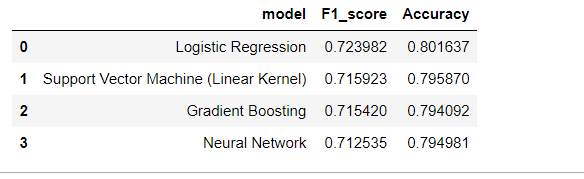
**Machine learning Modeling:**

After performing EDA, the next step was to build a machine learning model to predict customer churn. The following models were used for this project; Logistic Regression, Random Forest, Gradient Boosting, Neural Network, Support Vector Machine (Linear Kernel), Support Vector Machine (RBF Kernel), K-Nearest Neighbors, Decision Tree.



The 8 models were trained on the training data and evaluated on the test data using accuracy, Precision, Recall, and F1-score metrics. The Logistic Regression model had the highest performance with an F1 score of 0.62. The top four performing models were selected and tuned.

**Summary of Results:**



* Based on the Accuracy and F1-Score metrics, the best model for the classification problem appears to be Logistic Regression model. Although all the other models had similar accuracy scores, Logistic Regression model had the highest F1-Score and Accuracy.
* While Logistic Regression model had a good accuracy score of 0.80, its F1-Score, is slightly higher than the other three models. Therefore, based on the metrics evaluated Logistic Regression model appears to be the best model for this classification problem.
* However, it is important to note that other factors such as model complexity, computerization, efficiency, and interpret-ability such also be considered when choosing a model for deployment.

**Conclusion:**

**My interpretation of the categorical variables in line with the graphs is as follows:**

* **There is no significant difference in the churn numbers of male and female customers.**
* **It seems that clients who pay by electronic check have a high churn.**
* **It seems that clients with monthly contracts have a high churn.**
* **It seems that clients using fiber optic internet have a high churn.**
* **These comments were made according to the graphics. A/B testing can be done to verify these.**