BDA601-Big Data and Analytics

Assessment: Visualisation and Model development

Assessment 2 Report and Source Code

Student Name: Aadesh Bohora

Lecture Name: Farshid Keivanian

**Academic Integrity**

I declare that except where I have referenced, the work I am submitting for this assessment task is my work. I have read and am aware of Torrens University Australia Academic Integrity Policy and Procedure viewable online at https://www.torrens.edu.au/policies-and-forms I am aware that I need to keep a copy of all submitted material and their drafts, and I will do so accordingly.

Name: Aadesh Bohora

Table of Contents

[1.Introduction​ 3](#_Toc164021209)

[2.Data Preparation​ 3](#_Toc164021210)

[3 Model Development 5](#_Toc164021211)

[a) Problem statement 5](#_Toc164021212)

[b) Exploratory data analysis 5](#_Toc164021213)

[Visualization 5](#_Toc164021214)

[Tenure Distribution: 6](#_Toc164021215)

[TotalCharge Distribution: 6](#_Toc164021216)

[Churn Distribution: 6](#_Toc164021217)

[SeniorCitizen Distribution: 6](#_Toc164021218)

[Correlation Matrix of Numerical Features: 7](#_Toc164021219)

[Tenure vs. Churn Box Plot: 8](#_Toc164021220)

[MonthlyCharges Vs Churn Scatter Plot: 8](#_Toc164021221)

[c) Data Cleaning and Feature Selection 8](#_Toc164021222)

[Install the required library 8](#_Toc164021223)

[Library Import 9](#_Toc164021224)

[Data loading 9](#_Toc164021225)

[Tenure Histogram and total charges Histogram 9](#_Toc164021226)

[Tenure histogram: 10](#_Toc164021227)

[Churn count plot and SeniorCitizen Count Plot 10](#_Toc164021228)

[Correlation Matrix and Box plot for Tenure vs. churn 10](#_Toc164021229)

[Scatter plot for MonthlyCharges vs Churn 11](#_Toc164021230)

[d) Model Building​ 11](#_Toc164021231)

[Development of a decision Tree model 11](#_Toc164021232)

[Import ML libraries: 12](#_Toc164021233)

[Define Features and target 12](#_Toc164021234)

[Define Preprocessing for Numeric and categorical Features 13](#_Toc164021235)

[Handling Missing Value: 14](#_Toc164021236)

[Interpretation of Churn Analysis: 15](#_Toc164021237)

[Conclusion: 15](#_Toc164021238)

[References 16](#_Toc164021239)

# 

# Introduction​

According to the research (XM Experience Management, n.d.), we can determine that Customer Churn is also known as customer attrition when existing customers stop choosing products or services from their provider. That means they do not want to be customer of their provider. The precise result of this point is that businesses have to go into loss. However, in simple terms, a high customer churn rate is very critical for business for a different reason. In this report, we are building a machine learning (ML) model to predict customer churn by using the principles of ML and Big data tools.

In this report, we are explaining the predictive model from a given data set that follows data mining principles and techniques. And discussing how to handle missing values from a dataset and expanding how to interpret the outcomes of the customer churn analysis.

# Data Preparation​

The modified Telco Customer Churn dataset is downloaded from Torrens Blackboard, or it can be downloaded from Kaggle datasets blastchar which is a dataset originally derived from IBM Telco customer churn sample data. In a modified dataset information contains approximately 7043 telecommunication customers and comprises 16 attributes. this data set is format in a CSV file. and can be manipulated in different applications like Microsoft Excel or Open Office Calc. The sample data of the modified Telco customer Churn dataset is below:

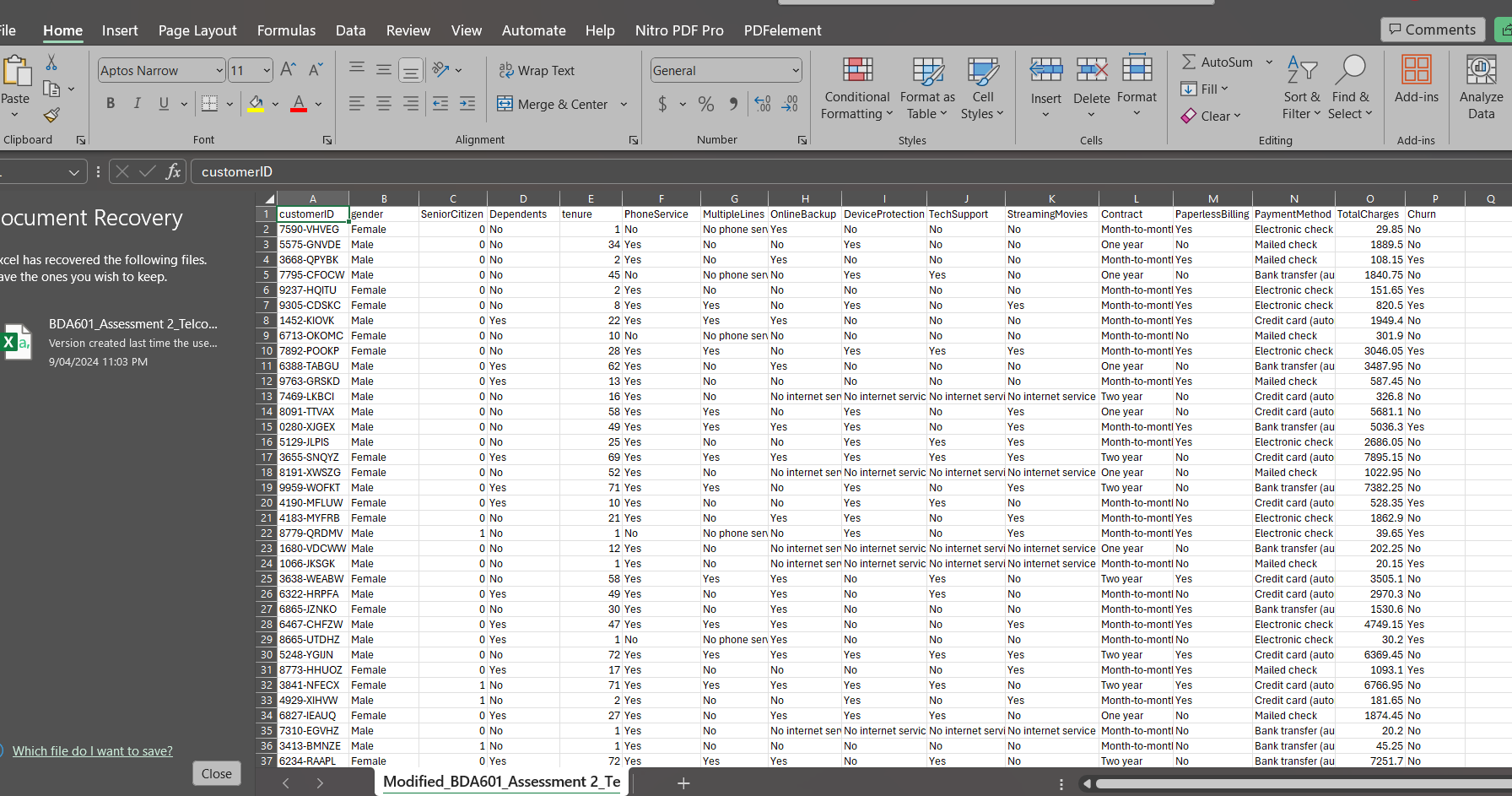
#### Figure 1

A screenshot of a computer program

Description automatically generated

Note: This is a code of modified data .

#### Figure 2



Note: This Excel file is downloaded from Kaggle.

In this modified Kaggle telco churn dataset there are some of the attributes that are removed from the original dataset are as follows:

* MonthlyCharges
* OnlineSecurity
* StreamingTV
* InternetService
* Partner

These attributes are removed from this dataset because it could be deemed less relevant to predicting customer churn or were likely to develop complexity while analyzing the data for valuable results.

# Model Development

## Problem statement

In this analysis, the priority is to develop a predictive model to decrease customer churn in the telecommunication industry with the help of machine learning techniques, more importantly with a decision-tree model implemented using Pysark MLlib, the main goal of this analysis is to find the customer accurately who are trying to churn based on their historical usage patterns and demographic information. According to the research (Ahmad et al., 2019)

we can determine that there are three main strategies to generate more revenues which are.

1. acquire new customers.
2. upsell the existing customers.
3. increase the retention period of customers.

the last one is the most effective way to profitable strategy. therefore, it is important to identify potential churners so that companies can implement targeted retention strategies to decrease customer attrition and improve outcomes of business goals.

## Exploratory data analysis

### Visualization

#### Figure 3

A screenshot of a computer

Description automatically generated

### Tenure Distribution:

This diagram shows the distribution of customer tenure. It can be seen that the large amount of tenure at the beginning time sharply decreases after some time and stabilizes after that. However, there is significant growth at the end of the 70-month mark This could mean that there are quite a few long-term customers.

### TotalCharge Distribution:

This histogram for TotalCharges shows a right-skewed distribution which means that the most of customers have lower total charges and only some of the customers have very high total charges. Thus, there are significantly fewer spend customers than high-value customers.

### Churn Distribution:

The Churn Distribution is depicted in the bar plot, suggesting that the number of customers who have not churned significantly outnumber those who have.

### SeniorCitizen Distribution:

This bar plot shows the distribution of customers who is classified into senior citizens, in this bar plot we can see that most of the customers are senior citizens.

#### Figure 4

A screenshot of a computer

Description automatically generated

### Correlation Matrix of Numerical Features:

The heatmap displays that the tenure and “TotalCharge” have a higher positive value than the senior citizen by 1.0 which means that the tenure and “TotalCharges” can increase as well. In the long term, customers generally contribute more revenue over time than the others as shown in this correlation matrix. According to the correlation Matrix of Numerical features, we can conclude that the “SeniorCitizen” and other numerical features are quite low, suggesting no strong linear relationship (Dr. Farshid , 2024).

#### Figure 5

A screenshot of a computer screen

Description automatically generated

### Tenure vs. Churn Box Plot:

This box plot demonstrates that customers who have not churned often have a longer stay. This also means that, if customers stay on the company for a longer time they will be less likely to leave the company.

### MonthlyCharges Vs Churn Scatter Plot:

This Scatter plot gives information about the customers who have not churned yet and there is a vast amount of monthly charges to them .thus the result also shows that there is a strong chance of higher monthly charges to customers . This could be a reason to customer’s decision to churn.

### Data Cleaning and Feature Selection

Install the required library

#### Figure 6

A screen shot of a computer screen

Description automatically generated

### Library Import

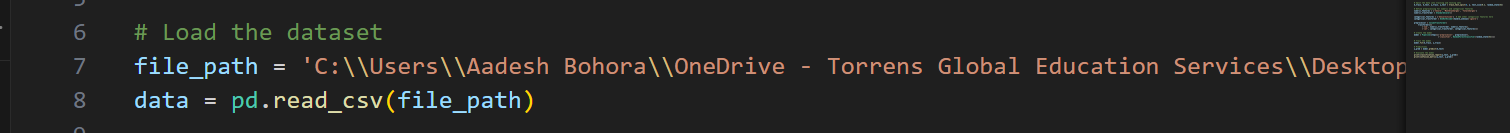
Figure 7A black screen with text

Description automatically generated

This section shows the necessary Python libraries. Which is explaining about the “Panda” that is used for data manipulation and analysis. Moreover, it shows “NumPy” which is also for numerical operations. “matplotlib.pyplot” and “seaborn are mainly for plotting and data visualization (Dr. Farshid, 2024).

### Data loading

#### Figure 8



In this code, we can see that the dataset is loaded from a CSV file into pandas, DataFrame, “data” by using “pd.read\_csv()” function. The location of the dataset is held by the “file\_path”. (Dr. Farshid, 2024)

### Tenure Histogram and total charges Histogram

#### Figure 9

A computer screen with colorful text

Description automatically generated

### Tenure histogram:

Builds a histogram with a kernel Density Estimate(KDE)overlay for the “tenure” column, and the given plot is accordingly colored and sized through specified parameters moreover, the “TotalCharges” histogram is also created like the Tenure histogram. (Dr. Farshid, 2024)

### Churn count plot and SeniorCitizen Count Plot

#### Figure 10

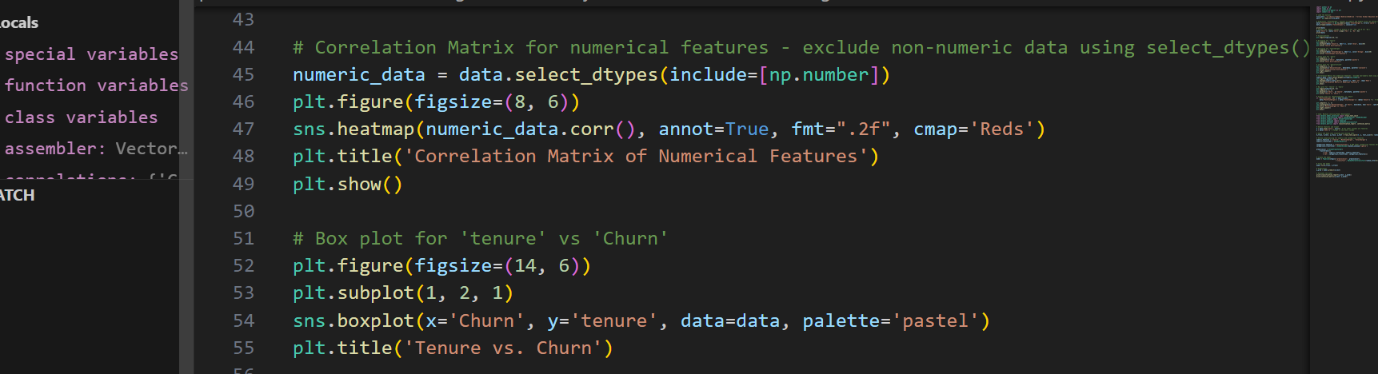
A screen shot of a computer code

Description automatically generated

In this figure, we can see the Churn Count plot which it is developed to visualize the distribution of churned and non-churned customers. SeniorCitizen Count Plot produces a bar lot for providing a different color palette (Dr. Farshid, 2024)

### Correlation Matrix and Box plot for Tenure vs. churn

#### Figure 11



This code demonstrates a correlation matrix that is for numerical features and only excludes non-numeric data using select\_dtypes(). Box Plot for tenure vs Churn: this code is for using box plot for tenure vs churn and for distribution of tenure between customers who have churned and those who have not (Dr. Farshid , 2024).

### Scatter plot for MonthlyCharges vs Churn

#### Figure 12

A computer screen with text on it

Description automatically generated

This data set for MonthlyCharges vs churn and it shows that if “MonthlyCharges” is not in data columns then the “MonthlyCharges” is calculated by dividing “TotalCharges by “tenure”. After that it adds 1 to avoid division by zero.

## Model Building​

Before Building the model, It is very important to split the dataset into training and testing sets using “train\_test\_split from “sklearn.model\_selection”. After that build an instance of “decisionTreeclassifier” from “sklearn.tree”. Generally, a common split is 80% for training and 20% for testing (Dr. Farshid , 2024).

However, Fit the model to the training data using the “fit “method. When the model is trained, they can make predictions on the test set and evaluate the model using the “classification\_report and “confusion matrix from “sklearn.metrics” (Dr. Farshid , 2024). “the plot\_tree” function from “sklearn.tree” can be implemented to visualization the decision tree.

### Development of a Decision Tree model

This process consists of data preparation, splitting it into training and test sets, building a Decision Tree Classifier, fitting the model , making predictions, evaluating the model, and visualizing the decision tree.

#### Figure 13

A screen shot of a computer code

Description automatically generated

And moreover, if “churn consist values like “yes” or “no , that means it eeds to be encoded as a binary variable before it can be implemented in the machine learning model.

Import ML libraries:

Figure 14A screen shot of a computer program

Description automatically generated

### Import ML libraries:

This Blocks imports the important components from scikit-learn: “train\_test\_split’ to split data into training and testing sets, StandardScaler” and “OneHotencoder” for feature scaling and encoding. To transform different coulmns, “columnTransformer “can be applied for it.

The “pipeline” can be used for transforming sequentially applied and for a final estimator. “RandomForestClassifier” as the machine learning model and evaluation functions ”classification\_report” and “confusion\_matrix”.

Define Features and target.

In this code, it shows that the x stores the features which is independent variables, and “y” stores the target which is the dependent variable which could be a store that is churned or not in this case.

#### Figure 15

A screen shot of a computer code

Description automatically generated

Split the data: This split the data into a training set”9X\_train, y\_train)” and a testing set “(X\_test,y\_test)” with 20% of the data reserved for testing.

Define Preprocessing for Numeric and categorical Features.

This numerical feature will be scaled by using “StandardScaler” which is to standardise them by erasing the means and increasing to unit variance. Use 'OneHotEncoder' to encode categorical features.​

* Create a column Transformer: “ColumnTransformer” permits distinct columns of the input to be converted separately. The features which are developed by each transformer are integrated to form a single feature space.
* Train the model: This stage is the train of the model which Trains the RandomForest.
* Predictions: This Trained model can be used to make predictions on the test set.
* Evaluate the model:This coding will print the classification report and confusion\_matrix result which gives meaningful and accurate types of results and types of errors made by the classifier.

Output:

#### Figure 16

A screenshot of a computer program

Description automatically generated

# Handling Missing Value:

According to the report, we can determine that the given dataset has the missing value In this dataset there are many categorical features and numerical ones. And according to the given dataset, there are no missing values, however, we can find the “TotalCharges” column which is the object type of the dataset that may contain non-numeric values that need to be converted to numeric values. As seen in the dataset if there are any non-numerical values or any errors in this conversion, they will be seen as NaN values after the conversion attempt in the following data set.

To proceed we need to follow the step (handling missing Values) that the approach detailed in the code: by converting “TotalCharges” to a numeric type and, after that handling any resulting missing values However, demonstrate a proper strategy for handling these.

For handling missing values there are many strategies which are:

* Imputation
* Deletion
* Prediction model

The decision tree classifier shows that “TotalCharges” is the most essential attribute for predicting churn with approximately 67.6%. On the other hand, The “tenure” attribute is the second most important at around 29.2%. However, “SeniorCitizen” has less important than others and it has about 3.2%.

# Interpretation of Churn Analysis:

To the interpretation of churn Analysis, we have to do further analysis on the given dataset. There are some involvements to the analysis of the model’s prediction on churn and discuss the implications of business strategy, customer segments churning, and potential reasons. To improve the model’s accuracy, we could do more sophisticated imputation methods, by including more features.

# Conclusion:

In conclusion, we can conclude that the churn analysis offers valuable insights into customer behavior and churn prediction, however, there are so many areas to improve. From this analysis, we can describe the importance of handling missing values, building predictive models, and interpreting o churn analysis to inform business strategies significantly. More importantly, the accuracy of churn analysis was developed through critical data preprocessing, model development, and interpretation of results. From this report, we find the missing value which is the “TotalCharges” attribute.

# References

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1). <https://doi.org/10.1186/s40537-019-0191-6>

Dr. Farshid, K. (2024). *BDA601 (455): Big Data & Analytics​*. Torrens University Australia.

Farshid Keivanian. (2024). *Sessions Python/Algorithm Decision Tree Classifier\_for\_Customer Churn Prediction.py at main · FarshidKeivanian/Sessions\_Python*. GitHub. <https://github.com/FarshidKeivanian/Sessions_Python/blob/main/Algorithm_Decision%20Tree%20Classifier_for_Customer%20Churn%20Prediction.py>

XM Experience Management. (n.d.). *What is Customer Churn? Learn To Measure & Prevent It.* Qualtrics AU. Retrieved April 10, 2024, from <https://www.qualtrics.com/au/experience-management/customer/customer-churn/>