Project – Customer Churn Prediction

# Introduction:

Customer churn prediction is a vital component for customer relationship management and business analysis. It involves the usage of strategies like “Data Analysis” and “Machine Learning” to forecast when and why the customers may stop using a product or service.

The primary goal of customer churn prediction is to identify at risk customers early on and take proactive measures to retain them.By analyzing historical data of the customers and various business factors that may influence businesses can develop many predictive strategies which may help them in maintaining their customers and will probably help in increasing their customer services.

# About Phase 5:

The phase 5 is all about data preprocessing of the given csv file for the purpose of performing various operations such as analysis, exploratory data analysis and visualizing of the dataset.

# Phase 5 of Customer Churn Prediction:

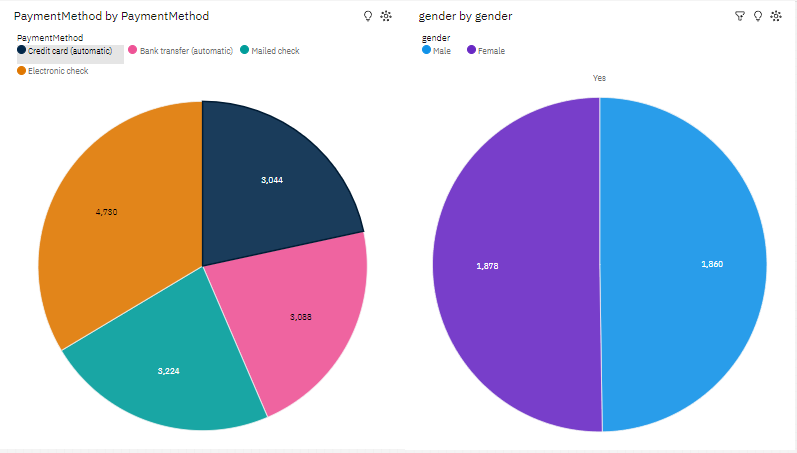
The phase 5 of the project customer churn prediction refers to visualizing of the data using the “IBM Cognos Tool”. The various charts displayed in this document are Bar chart,Pie chart,Line chart and scatter plot.

# About the “IBM Cognos Tool”:

The IBM Cognos tool is used for analyzing the files such as csv files and other files to visualize data from them.

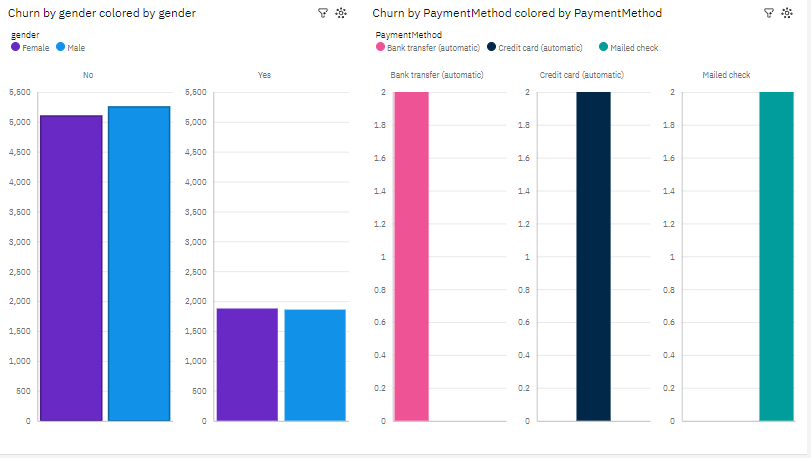
# 1)Pie Chart:

The first chart is about visualizing the data using cognos.The attributes used here are Gender to gender and the comparison of various payment methods.The pie chart is given below.



# 2)Bar Chart:

The second chart is the bar chart which is used for visualizing the contents such as churn by gender to gender as well as the various payments method used in the dataset are compared here.



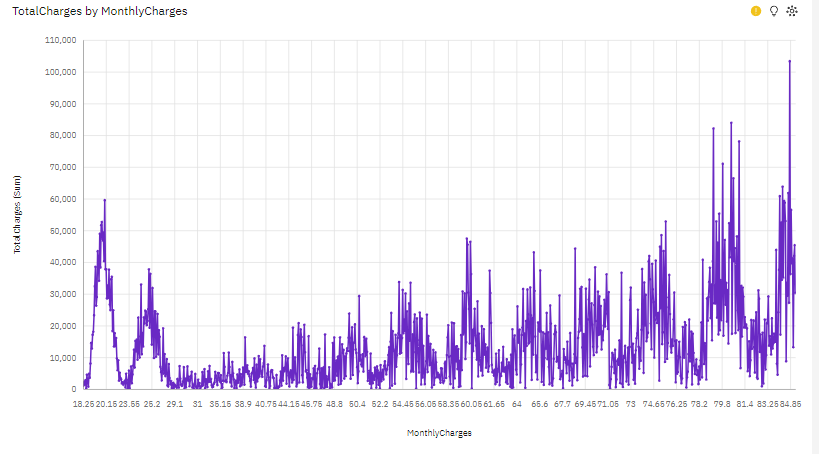
# 3)Scatterplot:

Scatterplot is the most commonly used graph which is used for depicting the relationship of two entities or objects.Here we have used multiple properties which are involved in payment methods.



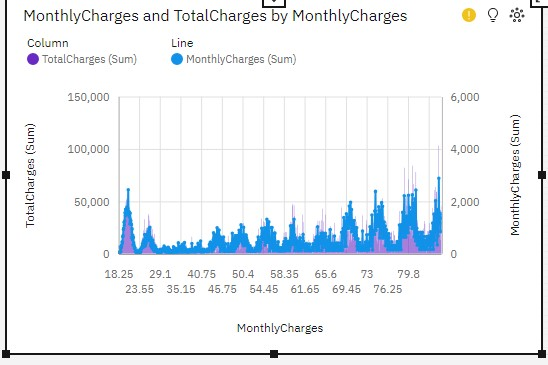
# 4)Line chart:

The line chart is considered to be the most basic chart or graph used for representation.Here the line chart visualizes the different relationships or the properties used here.



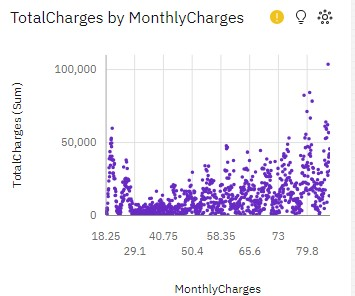
# **5)Line and column chart:**

The line and column chart is the combined version of column chart and the line chart.In the below chart it depicts the relation of Total charges with monthly charges.



# **6)Packed bubble chart:**

The packed bubble chart is a modified version of the scatterplot and point chart.



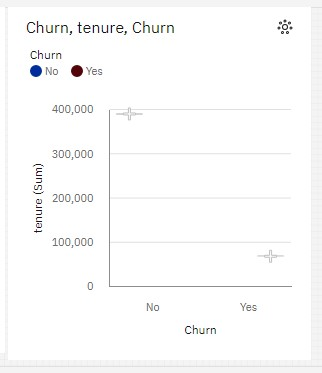
# **Insights:**

For TotalCharges, the most significant values of MonthlyCharges are 99, 84.8, 99.5, and 89.85, whose respective TotalCharges values add up to over 409 thousand, or 1.3 % of the total.

Across all values of MonthlyCharges, the sum of TotalCharges is over 32 million.

# **7)Boxplot:**

A box plot, also known as a whisker plot, is a graphical representation of the distribution of a dataset.



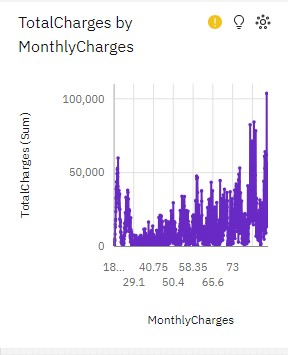
# **Insights:**

No exceeds Yes in tenure by 321,568

Churn No has the highest values of both tenure and TotalCharges.

**8)Bubble chart:**

The bubble chart is a modified version of scatterplot but only uses scatter function.



# **Insights:**

TotalCharges is unusually high when MonthlyCharges is 99, 84.8, 99.5 and 89.85.

Yes MultipleLines accounted for 70% of 99.0 TotalCharges compared to 64% for 84.8.

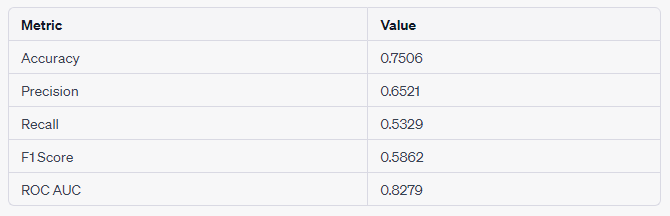
MultipleLines Yes has the highest TotalCharges at almost 21 million, out of which MonthlyCharges 109.55 contributed the most at over 87 thousand.

MonthlyCharges 99.0 has the highest total TotalCharges due to MultipleLines Yes.

Across all values of MonthlyCharges, the sum of TotalCharges is over 32 million.

# **Performance Metrics:**

The performance and evaluation metrics for the dataset and the fields of the csv file has been evaluated. The metrics in the tabular form are,



# **Accuracy:**

Accuracy measures the overall correctness of the model's predictions. In your case, it's the percentage of correct predictions out of all predictions made. An accuracy of 75.06% means that the model correctly predicted customer churn status for approximately 75.06% of the cases in your dataset.

# **Precision:**

Precision is the ratio of true positive predictions to all positive predictions made by the model. It is a measure of how many of the customers predicted as churning by the model actually did churn. A precision of 65.21% means that out of all the customers predicted to churn, 65.21% of them actually did.

# **Recall:**

Recall, also known as sensitivity, is the ratio of true positive predictions to all actual positive cases in the dataset. It measures how many of the actual churning customers were correctly identified by the model. A recall of 53.29% indicates that the model captured approximately 53.29% of the customers who truly churned.

# **F1 Score:**

The F1 score is the harmonic mean of precision and recall. It's a balance between precision and recall. A higher F1 score indicates a better balance between accurately identifying churning customers and minimizing false positives.

# **ROC AUC (Receiver Operating Characteristic Area Under the Curve):**

ROC AUC measures the model's ability to distinguish between the two classes (churn and non-churn) across different probability thresholds. It provides an overall assessment of the model's performance, with a higher value indicating a better-performing model. An ROC AUC of 82.79% suggests that the model is reasonably good at distinguishing between churning and non-churning customers.