CUSTOMER CHURN PREDICTION

**TEAM LEADER:** GURU VARSHENI A

**REGISTER NO:** 211521244015

**TEAM MEMBER:** DHANUSHREE.S

**REGISTER NO:** 211521244011

**TEAM MEMBER:** ISHA.K

**REGISTER NO:** 211521244018

**TEAM MEMBER:** POOJA.S

**REGISTER NO:** 211521244039

***LOADING AND PREPROCESSING THE DATASET:***

**Pandas**

Pandas is a software library written for Python. It is very famous in the data science community because it offers powerful, expressive, and flexible data structures that make data manipulation, analysis easy AND it is freely available. To use the pandas library, you need to first import it.

import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
import seaborn as sns  
import plotly.express as px  
import plotly.graph\_objects as go  
from plotly.subplots import make\_subplots  
import warnings  
warnings.filterwarnings('ignore')  
from sklearn.preprocessing import LabelEncoder  
from sklearn.preprocessing import StandardScaler  
from sklearn.model\_selection import train\_test\_split  
from sklearn.neighbors import KNeighborsClassifier  
from sklearn.preprocessing import StandardScaler  
from sklearn.preprocessing import LabelEncoder  
from sklearn.tree import DecisionTreeClassifier  
from sklearn.ensemble import RandomForestClassifier  
from sklearn.naive\_bayes import GaussianNB  
from sklearn.neighbors import KNeighborsClassifier  
from sklearn.svm import SVC  
from sklearn.neural\_network import MLPClassifier  
from sklearn.ensemble import AdaBoostClassifier  
from sklearn.ensemble import GradientBoostingClassifier  
from sklearn.ensemble import ExtraTreesClassifier  
from sklearn.linear\_model import LogisticRegression  
from sklearn.model\_selection import train\_test\_split  
from sklearn.metrics import accuracy\_score  
from xgboost import XGBClassifier  
from catboost import CatBoostClassifier  
from sklearn import metrics  
from sklearn.metrics import roc\_curve  
from sklearn.metrics import recall\_score, confusion\_matrix, precision\_score, f1\_score, accuracy\_score, classification\_report  
from sklearn.ensemble import VotingClassifier

from sklearn.metrics import confusion\_matrix, accuracy\_score  
from sklearn.metrics import f1\_score, precision\_score, recall\_score, fbeta\_score  
from statsmodels.stats.outliers\_influence import variance\_inflation\_factorfrom sklearn.model\_selection import cross\_val\_scorefrom sklearn.model\_selection import GridSearchCVfrom sklearn.model\_selection import ShuffleSplitfrom sklearn.model\_selection import KFoldfrom sklearn import feature\_selectionfrom sklearn import model\_selectionfrom sklearn import metricsfrom sklearn.metrics import classification\_report, precision\_recall\_curvefrom sklearn.metrics import auc, roc\_auc\_score, roc\_curvefrom sklearn.metrics import make\_scorer, recall\_score, log\_lossfrom sklearn.metrics import average\_precision\_score

**Loading Data**

The first step for data preparation is to get some data. Since we have a .csv file, we load it up in our system using the .read\_csv() function in pandas.

data = pd.read\_csv("data.csv")  
data.head()

customerID gender SeniorCitizen Partner Dependents tenure PhoneService \  
0 7590-VHVEG Female 0 Yes No 1 No   
1 5575-GNVDE Male 0 No No 34 Yes   
2 3668-QPYBK Male 0 No No 2 Yes   
3 7795-CFOCW Male 0 No No 45 No   
4 9237-HQITU Female 0 No No 2 Yes   
  
 MultipleLines InternetService OnlineSecurity ... DeviceProtection \  
0 No phone service DSL No ... No   
1 No DSL Yes ... Yes   
2 No DSL Yes ... No   
3 No phone service DSL Yes ... Yes   
4 No Fiber optic No ... No   
  
 TechSupport StreamingTV StreamingMovies Contract PaperlessBilling \  
0 No No No Month-to-month Yes   
1 No No No One year No   
2 No No No Month-to-month Yes   
3 Yes No No One year No   
4 No No No Month-to-month Yes   
  
 PaymentMethod MonthlyCharges TotalCharges Churn   
0 Electronic check 29.85 29.85 No   
1 Mailed check 56.95 1889.5 No   
2 Mailed check 53.85 108.15 Yes   
3 Bank transfer (automatic) 42.30 1840.75 No   
4 Electronic check 70.70 151.65 Yes   
  
[5 rows x 21 columns]

**Handling Missing Data**

Missing data can arise in the dataset due to multiple reasons: the data for the specific field was not added by the user/data collection application, data was lost while transferring manually, a programming error, etc. It is sometimes essential to understand the cause because this will influence how you deal with such data. We deal it in the following way:

**There are 11 records with missing Total charges**

data[data["tenure"] == 0]

gender SeniorCitizen Partner Dependents tenure PhoneService \488 Female 0 Yes Yes 0 No 753 Male 0 No Yes 0 Yes 936 Female 0 Yes Yes 0 Yes 1082 Male 0 Yes Yes 0 Yes 1340 Female 0 Yes Yes 0 No 3331 Male 0 Yes Yes 0 Yes 3826 Male 0 Yes Yes 0 Yes 4380 Female 0 Yes Yes 0 Yes 5218 Male 0 Yes Yes 0 Yes 6670 Female 0 Yes Yes 0 Yes 6754 Male 0 No Yes 0 Yes  MultipleLines InternetService OnlineSecurity \488 No phone service DSL Yes 753 No No No internet service 936 No DSL Yes 1082 Yes No No internet service 1340 No phone service DSL Yes 3331 No No No internet service 3826 Yes No No internet service 4380 No No No internet service 5218 No No No internet service 6670 Yes DSL No 6754 Yes DSL Yes  OnlineBackup DeviceProtection TechSupport \488 No Yes Yes 753 No internet service No internet service No internet service 936 Yes Yes No 1082 No internet service No internet service No internet service 1340 Yes Yes Yes 3331 No internet service No internet service No internet service 3826 No internet service No internet service No internet service 4380 No internet service No internet service No internet service 5218 No internet service No internet service No internet service 6670 Yes Yes Yes 6754 Yes No Yes  StreamingTV StreamingMovies Contract PaperlessBilling \488 Yes No Two year Yes 753 No internet service No internet service Two year No 936 Yes Yes Two year No 1082 No internet service No internet service Two year No 1340 Yes No Two year No 3331 No internet service No internet service Two year No 3826 No internet service No internet service Two year No 4380 No internet service No internet service Two year No 5218 No internet service No internet service One year Yes 6670 Yes No Two year No 6754 No No Two year Yes  PaymentMethod MonthlyCharges TotalCharges Churn 488 Bank transfer (automatic) 52.55 NaN No 753 Mailed check 20.25 NaN No 936 Mailed check 80.85 NaN No 1082 Mailed check 25.75 NaN No 1340 Credit card (automatic) 56.05 NaN No 3331 Mailed check 19.85 NaN No 3826 Mailed check 25.35 NaN No 4380 Mailed check 20.00 NaN No 5218 Mailed check 19.70 NaN No 6670 Mailed check 73.35 NaN No 6754 Bank transfer (automatic) 61.90 NaN No

**Filling in Missing Data**

To replace or rather "fill in" the null data, you can use the fillna() function. We fill in the null values with the mean.

data.drop(labels=data[data["tenure"] == 0].index, axis = 0, inplace = True)

data.fillna(data["TotalCharges"].mean())

gender SeniorCitizen Partner Dependents tenure PhoneService \  
0 Female 0 Yes No 1 No   
1 Male 0 No No 34 Yes   
2 Male 0 No No 2 Yes   
3 Male 0 No No 45 No   
4 Female 0 No No 2 Yes   
... ... ... ... ... ... ...   
7038 Male 0 Yes Yes 24 Yes   
7039 Female 0 Yes Yes 72 Yes   
7040 Female 0 Yes Yes 11 No   
7041 Male 1 Yes No 4 Yes   
7042 Male 0 No No 66 Yes   
  
 MultipleLines InternetService OnlineSecurity OnlineBackup \  
0 No phone service DSL No Yes   
1 No DSL Yes No   
2 No DSL Yes Yes   
3 No phone service DSL Yes No   
4 No Fiber optic No No   
... ... ... ... ...   
7038 Yes DSL Yes No   
7039 Yes Fiber optic No Yes   
7040 No phone service DSL Yes No   
7041 Yes Fiber optic No No   
7042 No Fiber optic Yes No   
  
 DeviceProtection TechSupport StreamingTV StreamingMovies Contract \  
0 No No No No Month-to-month   
1 Yes No No No One year   
2 No No No No Month-to-month   
3 Yes Yes No No One year   
4 No No No No Month-to-month   
... ... ... ... ... ...   
7038 Yes Yes Yes Yes One year   
7039 Yes No Yes Yes One year   
7040 No No No No Month-to-month   
7041 No No No No Month-to-month   
7042 Yes Yes Yes Yes Two year   
  
 PaperlessBilling PaymentMethod MonthlyCharges \  
0 Yes Electronic check 29.85   
1 No Mailed check 56.95   
2 Yes Mailed check 53.85   
3 No Bank transfer (automatic) 42.30   
4 Yes Electronic check 70.70   
... ... ... ...   
7038 Yes Mailed check 84.80   
7039 Yes Credit card (automatic) 103.20   
7040 Yes Electronic check 29.60   
7041 Yes Mailed check 74.40   
7042 Yes Bank transfer (automatic) 105.65   
  
 TotalCharges Churn   
0 29.85 No   
1 1889.50 No   
2 108.15 Yes   
3 1840.75 No   
4 151.65 Yes   
... ... ...   
7038 1990.50 No   
7039 7362.90 No   
7040 346.45 No   
7041 306.60 Yes   
7042 6844.50 No

**Introduction to IBM Cognos Analytics:**

IBM Cognos Analytics is a powerful business intelligence tool that enables organizations to gain actionable insights from their data, make data-driven decisions, and improve business performance their totality. It allows businesses to visually analyze data, create interactive dashboards, and generate meaningful reports.

With powerful features and capabilities, IBM Cognos Analytics is an ideal solution for analysing and minimizing disruptions. By leveraging advanced analytics and reporting capabilities, businesses can uncover valuable insights into customer behaviour, preferences, and trends to proactively manage churn and improve ability to retain customers.

**Leverage IBM Cognos Analytics to retain customers**

IBM Cognos Analytics offers many features and functionality specifically designed to improve customer retention. Through interactive dashboards, businesses can monitor in real time, visualize trends, and track the effectiveness of loyalty initiatives.

Additionally, IBM Cognos Analytics enables businesses to perform advanced predictive analytics to predict and prevent disruptions. By creating sophisticated models that incorporate many different customer attributes and behaviours, businesses can proactively identify customers at risk of churn and take appropriate action to retain them.

Through integration with other CRM and marketing automation platforms, IBM Cognos Analytics enables businesses to deploy automated campaigns to build customer loyalty.

**Analyze customer behaviour patterns with IBM Cognos Analytics**

Understanding customer behavioural patterns is key to reducing churn. IBM Cognos Analytics allows businesses to analyze customer interactions across different channels, such as website visits, social media engagement, and email communications.

By integrating and analysing these omnichannel touchpoints, businesses can identify customer preferences, concerns, and problems. This customer information can be used to tailor marketing strategies, product offerings and customer service interactions, effectively improving customer satisfaction and reducing churn.

**Leverage Data Visualization in Churn Analytics with IBM Cognos Analytics**

Data visualization is a powerful technique for understanding complex data and communicating information effectively. IBM Cognos Analytics offers many visualization options, including charts, graphs, and interactive heat maps.

By visualizing churn rates and related metrics, businesses can more easily identify trends, patterns, and outliers. This visual representation of data allows businesses to identify areas for improvement and make informed decisions to effectively reduce disruption.

Additionally, by sharing visualizations with stakeholders and decision makers, businesses can foster a data-driven culture and ensure alignment across the organization by reducing costs and improving ability to retain customers.

**Visualization in IBM Cognos:**

To create a visualization in IBM Cognos Analytics, we can change the visualization type or the columns used in the visualization. You can also see thumbnails of your visualizations, called cards, in the navigation pane to the left of the main view.

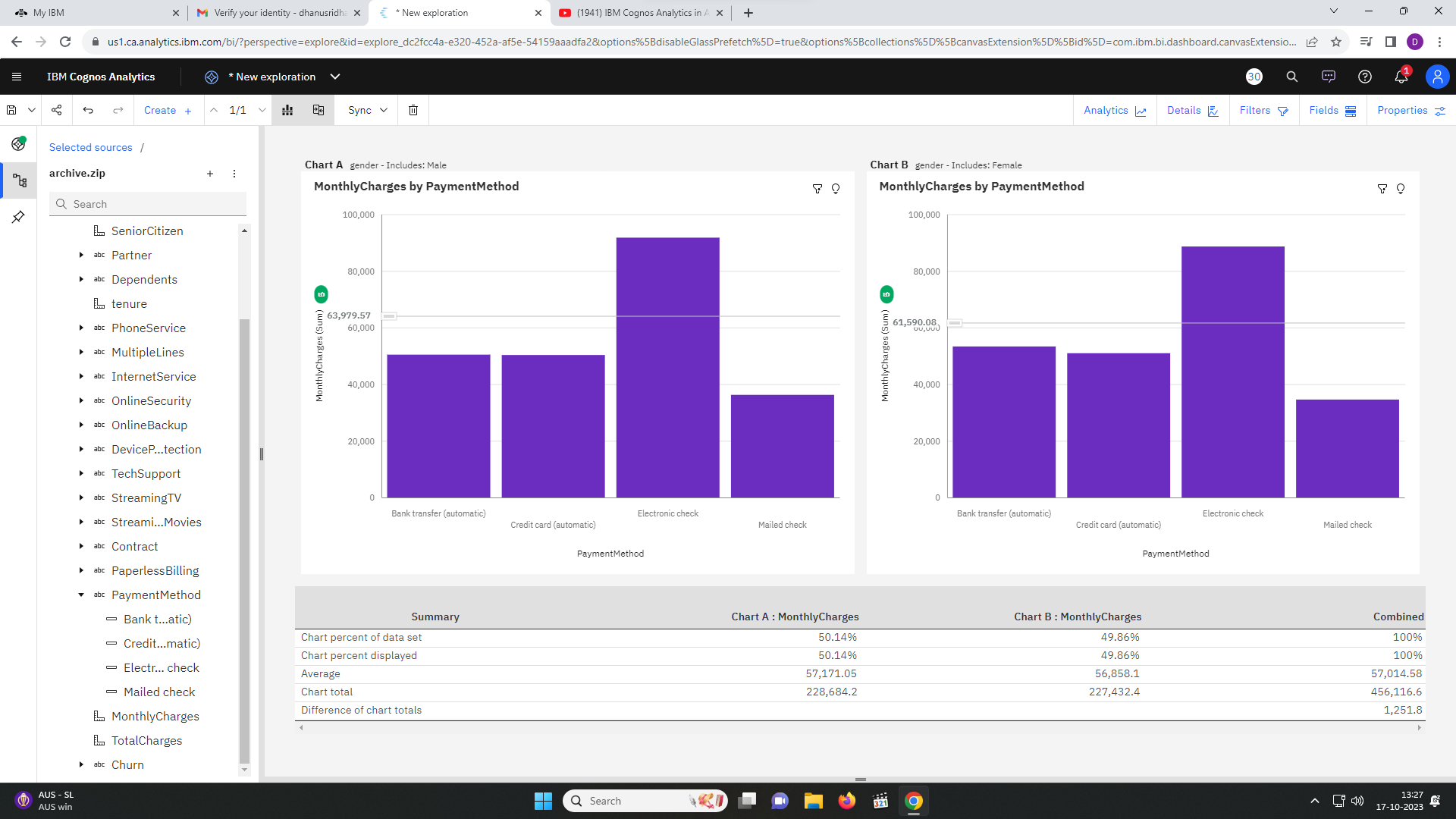
IBM Cognos Analytics provides analytical insights that help you discover and validate important relationships and meaningful differences based on data presented visually. You can choose from different visualization types, such as bar charts, line charts, scatter plots, and more. to convey comparisons, relationships, and trends.

We can also create your own comparisons to analyze data between two visualizations. You can also start with a suggested comparison. In both cases, a summary of key information and differences between the two visualizations is created.

Given below is the representation of MonthlyCharges by PaymentMethod in the form of column charts. We take PaymentMethod as the X-axis and the MonthlyCharges as the Y-axis.

In IBM Cognos we can create your own comparison to analyze the data between two visualizations or start with a recommended comparison. In either case, a summary of key information and differences between the two visualizations is generated.

We compare two charts with the filter of gender where the right side includes only the male gender and the left one includes only the female gender.

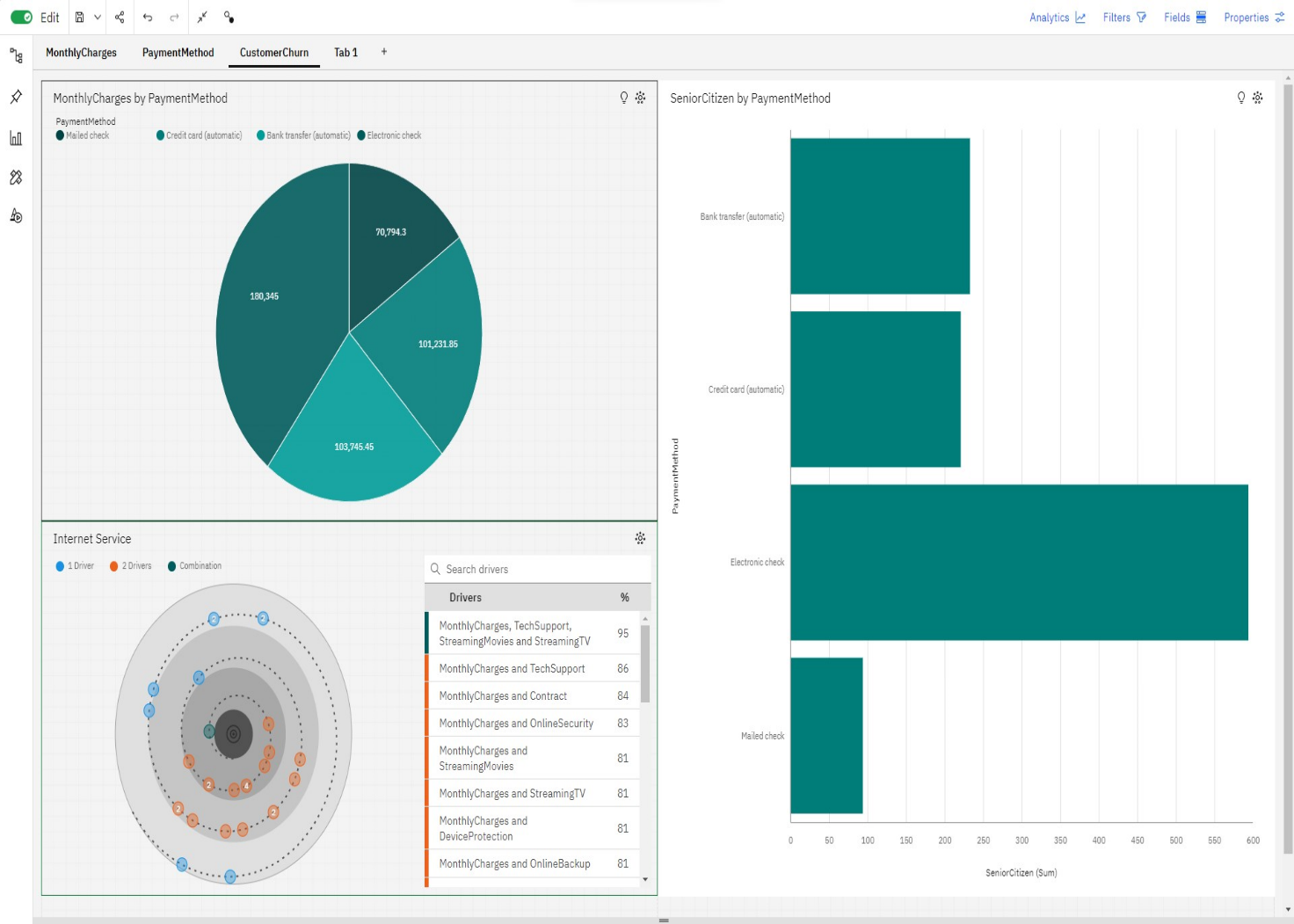


IBM Cognos Analytics is a business intelligence tool that offers many features, including dashboards. A dashboard is a collection of visualizations that provide insight into data. IBM Cognos Analytics allows you to create dashboards with powerful visualization capabilities that help you discover patterns and relationships in your data. You can also share these dashboards with others.

Given below is a dashboard which has three visualisations. The first one is a pie chart. Pie charts are useful for highlighting proportions. They use segments of a circle to show the relationship of parts to the whole. This pie chart represents MonthlyCharges by PaymentMethod.

Next, we have the spiral visualisation which shows the Internet Services. A spiral visualization shows you the key drivers, or predictors, for a given target. The closer the driver is to the centre, the stronger that driver is.

Lastly, we have SeniorCitizen by PaymentMethod in the Y-axis and the X-axis respectively. It shows that SeniorCitzen mostly use the Electronic check as the payment method.



Given below is another dashboard which shows the count of the distinct customerID which is estimated to be 7.04K. Also, we have three other visualisations, one of which is the line chart which represents SeniorCitizen by PaymentMethod in the Y-axis and the X-axis respectively. It shows that SeniorCitzen mostly use the Electronic check as the payment method.

Next, we have a bar chart that shows tenure by Partner in the Y-axis and the X-axis respectively. It shows that people who have partners have more tenure than those who don’t.

Finally, we have a decision tree which shows the top 5 nodes and the the category is set to yes.

