# Data analysis and machine learning with Python for real-world

# Introduction

## 1. 1 Background Overview

Modern organizations and companies collect massive amounts of information in the course of conducting business. If this data is examined properly, then it can be very useful for modifying future decision making procedures. To this extent, predictive analysis and machine learning are some of the power tools that organizations use to analyze such data in order to be in a position to predict future trends and/or behaviors.

## 1. 2 Objectives

Thus, the main goal of this paper is to highlight the areas of machine learning in business and its application to the case of customers’ churn prediction. More precisely, this report seeks to understand the potential of predictive analytics and analyze the efficiency of different types of machine learning algorithms.

## 1. 3 Scope of Study

In this research, customer churn prediction in the telecommunications industry is of interest and data from the Telco Customer Churn dataset are used. Data cleaning, exploratory data analysis, training of the model and evaluating the same will be part of this analysis.

## 1. 4 Role of the Predictive Analytics and Machine Learning for Data Driven Decision-making

Today business decisions cannot be made without predictive analytics and machine learning. Being able to learn from historical records, these technologies are capable of detecting patterns and thus predicting future occurrences, giving business the ability to act and prevent instead of acting on a projected aftermath. They also assist in enhancing the quality and speed of work such as customer loyalty, sales prediction, and risk control.

## 1. 5 Real world business and social problems solved by machine learning

The usage of machine learning is common in different fields. In business, it assists in achieving efficiency in the delivery of services to customers, identify fraud and cut on the expenses required to run the business. It has its use in health care where it is used in forecasting disease incidences and in the diagnosis of the disease. In societal pertinent issues it supports decision making on resources sharing, prediction of criminal activities, and disease control. It is also a rather universal option capable of providing reasonable approaches to the vast variety of issues facing businesses and societies.

# Exploratory Data Analysis

## 2.1 Dataset Selection and Problem Identification

Subset for this project is the Telco Customer Churn dataset collected from [Kaggle](https://www.kaggle.com/datasets/blastchar/telco-customer-churn). The first and focal business issue that has been addressed is to forecast the ‘churn’, meaning the customers who are likely to leave the telecommunication service. Since the program can accurately predict which customers are likely to defect, it can assist the telecom companies to design ways of protecting their customer base and therefore, contain the possible revenue loss.

## 2. 2 Data Types and the Structure

The dataset consists of the following data types:

* Categorical Variables: Gender, Partner, Dependents, Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Contract, Paperless Billing, Payment Method, Churn
* Numerical Variables: SeniorCitizen, tenure, MonthlyCharges, TotalCharges
* Target Variable: Churn (whether the customer walked away or not).

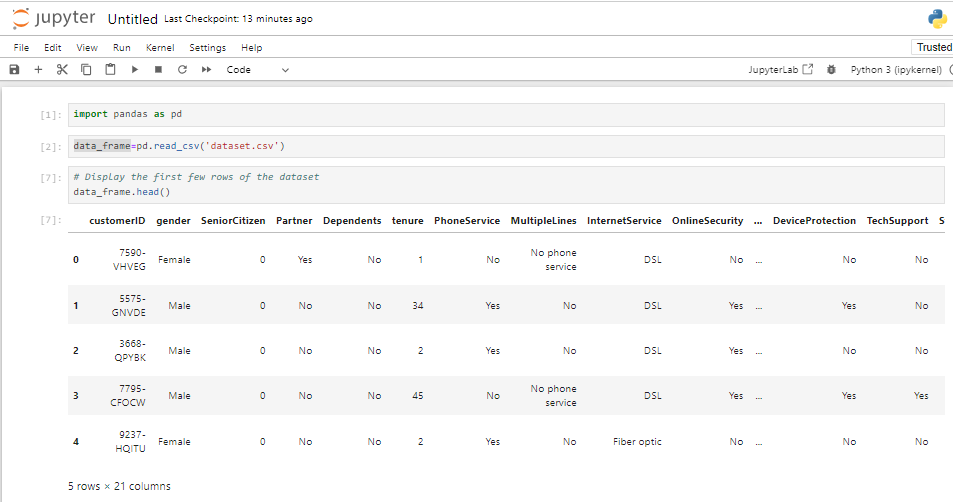


Figure : Jupyter Notebook: Data Loading and Initial Inspection

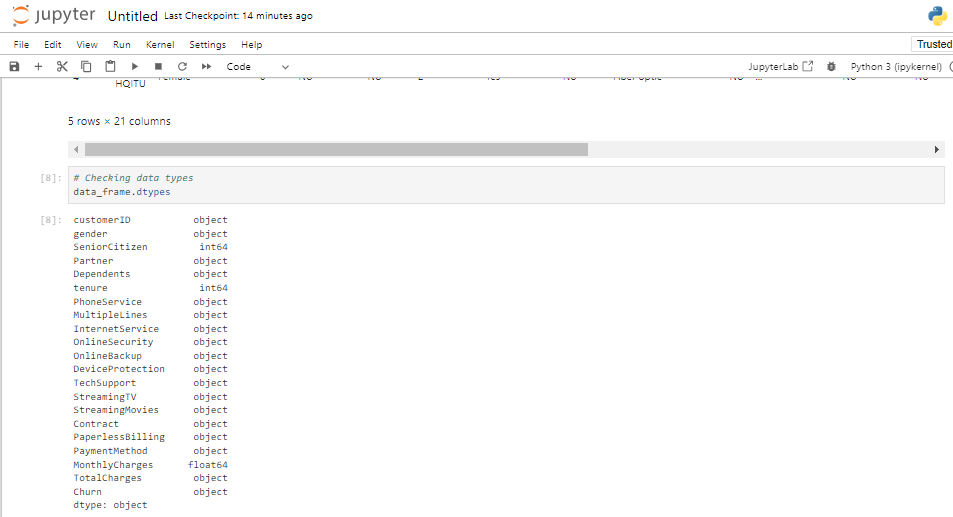


Figure : Jupyter Notebook Code and Output: Data Types Exploration

The used data set is composed of different customer characteristics including the taken services, explicit and implicit profile and account data. Thus, the aim is to maximize the accuracy in estimating the Churn column, which is binary in nature.

## 2.3 Data Cleaning and Encoding

It is important to preprocess the data before constructing the models of machine learning. This includes feature preprocessing to handle missing values, and also encoding of categorical variables for compatibility with the models.

* **Handling Missing Values**: In the dataset there may be some columns with missing or incorrect data, for example, the TotalCharges column. These will either be filled or dispensed, which are available in the table below;

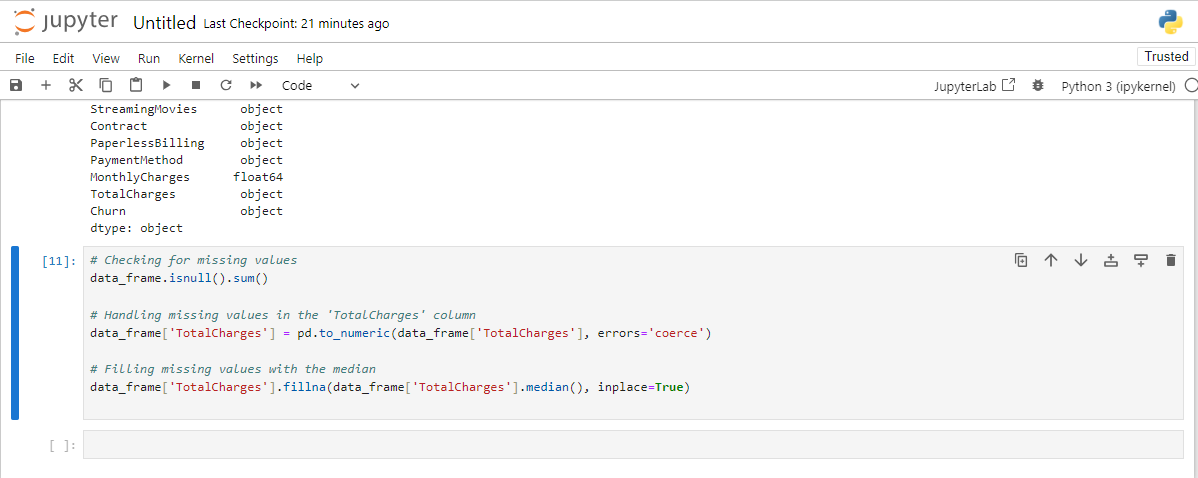


Figure : Jupyter Notebook Code and Output: Data Type Exploration and Handling Missing Values

* **Encoding Categorical Variables**: Gender, Contract and PaymentMethod would be categorical inputs that would have to be converted by the algorithm to numerical form for analysis.

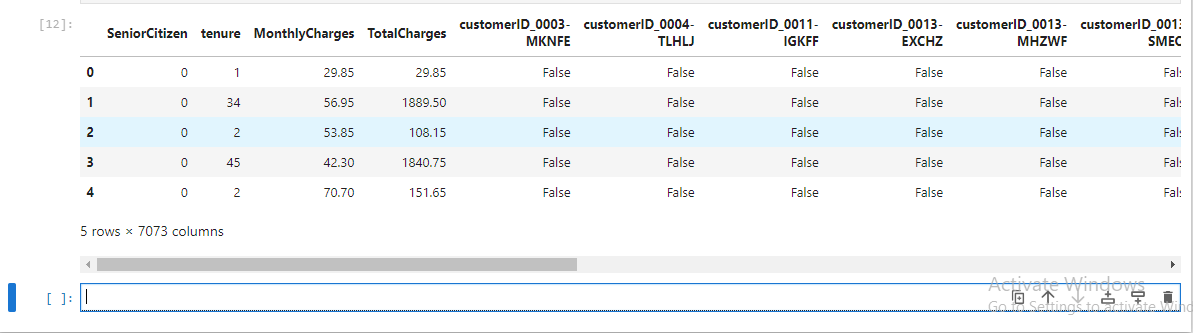


Figure : Jupyter Notebook Code and Output: Data Exploration and Handling Missing Values

## 2.4 Exploratory Data Analysis (EDA)

### 2.4.1 Descriptive Statistics

Also known as summarizing statistics, they let you understand the density and centrality of the data in terms of its features.

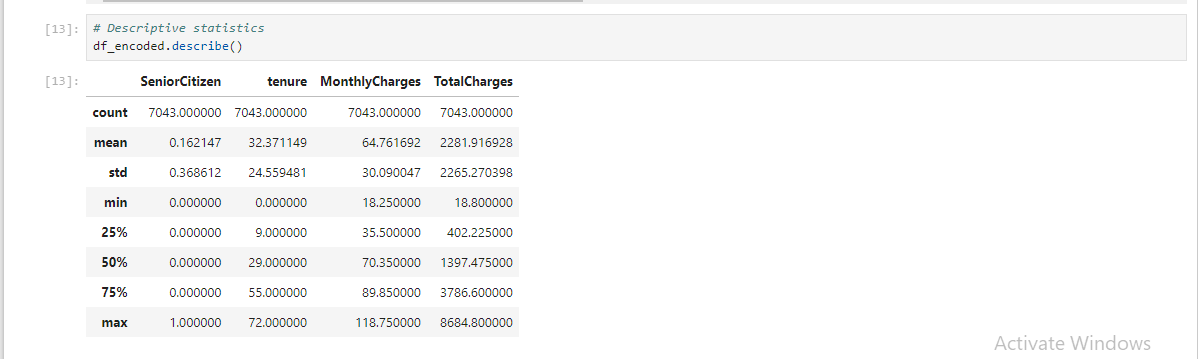


Figure : Jupyter Notebook Output: Descriptive Statistics for Numerical Columns

It is possible to derive certain qualitative conclusions from such statistical generalization; average tenure, MonthlyCharges, the percentage of SeniorCitizen customers, etc.

### 2.4.2 Data Visualizations

This means that, by presenting a view of the data, some insights can be made as to the interaction between different features and the target variable known as Churn.

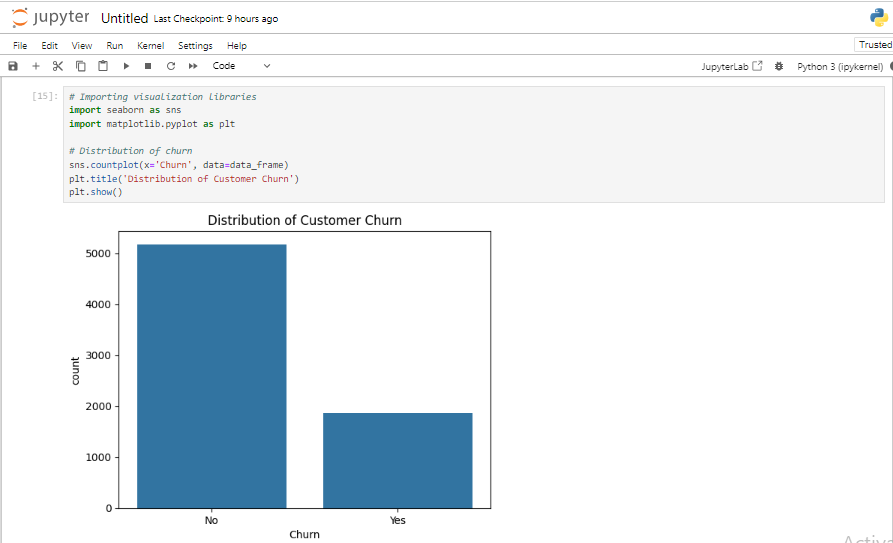


Figure : Jupyter Notebook Code and Output: Distribution of Customer Churn

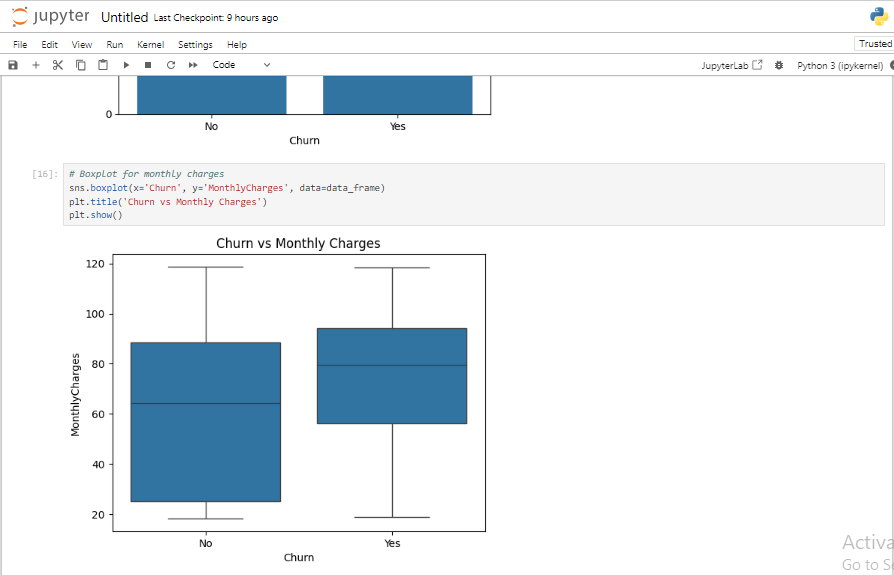


Figure : Jupyter Notebook Code and Output: Boxplot of Monthly Charges by Churn

* **Churn Distribution**: The count plot is a measure of the customers who have churned against those who have not.
* **Monthly Charges vs. Churn**: The boxplot gives a view as to how the monthly charges are spread out between the churning and the non- churning customers.
* **Correlation Heatmap**: This helps to visualize past correlations between different variables and shows how ‘strong’ each feature is with Churn.

### 2. 4. 3 Insights from EDA

1. **Churn Rate**: Currently, customer churn has been experienced within a significant percentage which underlines the need to come up with efficient methods of utilizing customer churn.
2. **Monthly Charges**: Mobile customers who are being charged a relatively high amount in a month are more likely to churn thus implying that improper pricing strategy might be a reason for customer churn.
3. **Service-Related Features**: Metrics like Contract, TechSupport and OnlineSecurity have a high degree of correlation with Churn implying that customer with shorter contractual period or who do not avail technical support are likely to churn. It would be expected that some of these variables to be one of the independent variables when developing common machine learning models.

These insights culminated from the EDA phase are very important especially when model building and evaluation is ongoing.

# 3. Task 2: Machine Learning Models

## 3. 1 Model Selection

Therefore, two ML models have been selected for performance of the task related to customer churn. They are ideal for classification problems, and they are prevalently applied in churn prediction.

### 3. 1. 1 Logistic Regression

Logistic Regression is linear model for binary classification that is applied for prediction of whether a customer is going to churn or not. It quantifies the propensity of a customer to churn by employing the input features and does a linear regression between the target and input features where the target variable is the log-odds of churning.