PROFESSIONAL TRAINING REPORT

at

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY

(Deemed to be University)

Submitted in partial fulfillment of the requirements for the award of

Bachelor of Engineering Degree

In

Computer Science and Engineering

By

DHEERAJ PULAKHANDAM

(Reg. No. 40110319)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTNG

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade “A” by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

NOVEMBER 2022

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY

(Established under Section 3 of UGC Act, 1956)

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai -600119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **DHEERAJ PULAKHANDAM (Reg.no.40110319)** who carried out the project entitled **"Telecommunication Churn"** under our supervision from Jan 2020 to April 2020.

Internal Guide

Dr. R. Sethuraman, M.E.,Ph.D.,

Head of the Department

Dr. L. Lakshmanan, M.E., Ph.D
Dr. S. Vigneshwari,M.E.,Ph.D

Submitted for Viva voce Examination held on______________________

Internal Examiner **External Examiner**

DECLARATION

I **DHEERAJ PULAKHANDAM (Reg. No. 40110319)** hereby declare that the Project Report entitled **“Telecommunication Churn”** done by me under the guidance of **Dr. R. Sethuraman, M.E., Ph.D.,** Sathyabama Institute of Science and Technology is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE: **SIGNATURE OF THE CANDIDATE**

ACKNOWLEDGEMENT

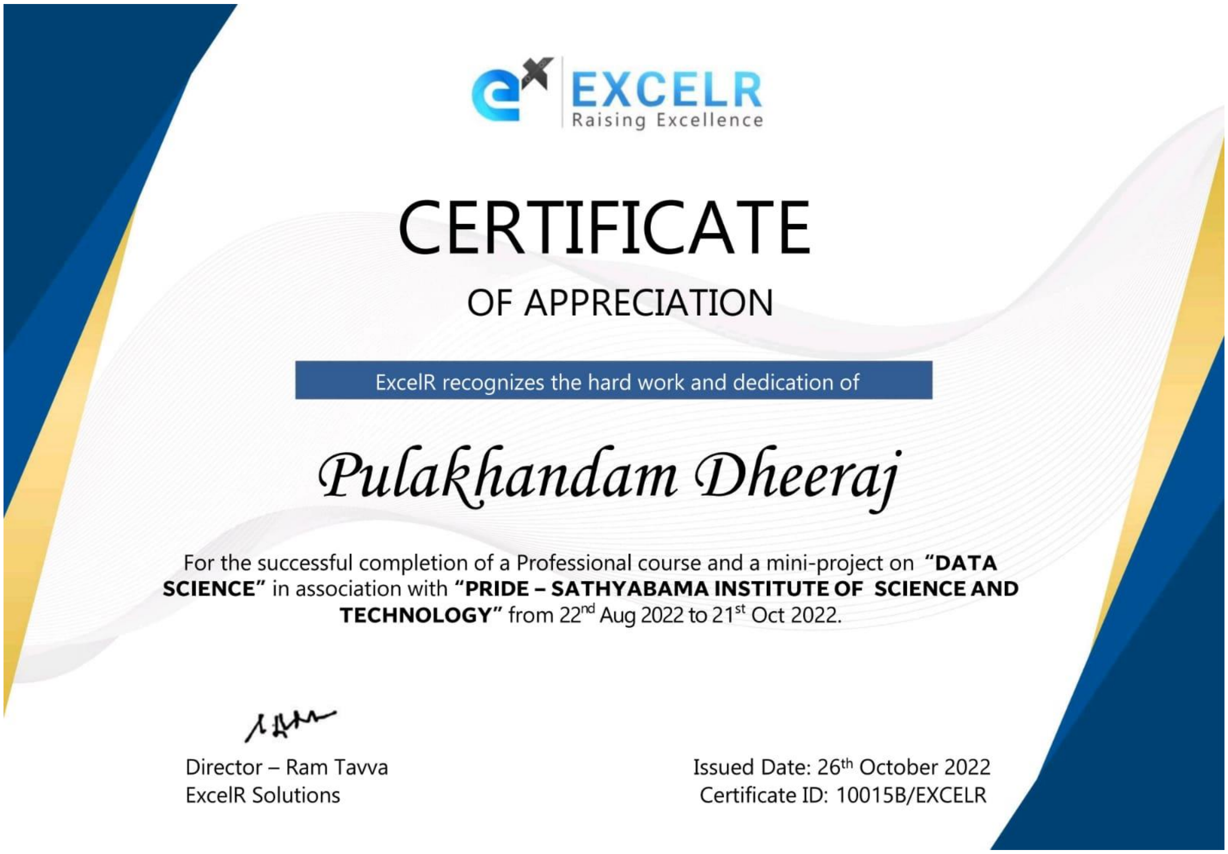
I am pleased to acknowledge my sincere thanks to Board of Management of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. SASIKALA. T, M.E., Ph.D., Dean, School of Computing** and **Dr. L. LAKSHMANAN, M.E., Ph.D., Head of the Department, Dept. of Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. R. Sethuraman, M.E., Ph.D.,** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Computer Science and Engineering who were helpful in many ways for the completion of the project.

TRAINING CERTIFICATE



EXCELR
Raising Excellence

CERTIFICATE

OF APPRECIATION

ExcelR recognizes the hard work and dedication of

Pulakhandam Dheeraj

For the successful completion of a Professional course and a mini-project on **"DATA SCIENCE"** in association with **"PRIDE – SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY"** from 22nd Aug 2022 to 21st Oct 2022.

Director – Ram Tavva
ExcelR Solutions

Issued Date: 26th October 2022
Certificate ID: 10015B/EXCELR

ABSTRACT

In the telecommunication industry that is working with a different number of subscribers or consumers daily, the dividends of the company are mostly dependent on the payments provided by these subscribers. Because it has been observed that the subscribers get frustrated sometimes with the services as well as the response of the company to their queries and based on those situations the subscribers decide to stop using the services of the organization ultimately resulting in the organization's loses. For this problem faced by organizations, our project intends to find out the factors that influence the subscriber's mindset while taking decisions related to the services of a particular telecom organization and use the same to predict whether new subscribers will behave in the same manner or not. Our project initially focuses on using (EDA)Exploratory Data Analysis to gather important factors related to the type of customers who can churn out in the company.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER-1

1. INTRODUCTION

In competitive Telecom market, the customers want competitive pricing, value for money and high quality service. Today's customers won't hesitate to switch providers if they don't find what they are looking for. This phenomenon is called churning. Customer churning is directly related to customer satisfaction. Since the cost of winning a new customer is far greater than cost of retaining an existing one, mobile carriers have now shifted their focus from customer acquisition to customer retention [1]. After substantial research in the field of churn prediction over many years, BigData analytics with Machine Learning was found to be an efficient way for identifying churn. These achieve results more efficiently and receive insights that sets alarm bells ringing before any damage could happen, giving companies an opportunity to take precautionary measures. These techniques are usually applied to predict customer churn by building models and learning from historical data [2]. However, most of these techniques provide a result that customers might churn or not, but only few tell us why they churn. Conducting experiments with end users' perspective, gathering their opinions on network, data normalization, preprocessing data sets [7], employing feature selection [6], eliminating class imbalance and missing values [5], replacing existing variables with derived variables [1] improves the accuracy of churn prediction which assists Telecom industries to retain their customers more efficiently. Comparatively, a smaller study was done on user's perspective, taking into consideration their quality of experience. In fact, no study was done taking into consideration only user's data volumes. Estimation of Quality of Experience by finding relationships between QoE and traffic characteristics could help the service providers to continuously monitor the user satisfaction level, react timely and appropriately to rectify the performance problems and reduce the churn [3] [4].

1.1 MOTIVATION

The Telecom industry is humongous, vibrant and dynamic with extremely large base of customers, making customer acquisition and customer retention imperative concerns for its survival and good profitability. The new entrants focus on customer acquisition, while old and matured one emphasize to focus on customer retention. Globalization enables customers to choose the best available services, which encourages the customers not to stick with a single company, rather opt from a diverse range of products/services. Customer churning is directly related to customer satisfaction [1]. Since the cost of acquiring new customer is much higher than retaining old news, operators lay preeminent significance on various customer related methodologies and analytics to ensure customer retention. There is no clear common consensus on the prediction technique to be used to identify churn. Significant research in the field of churn prediction is being carried out using various statistical and data mining techniques since a decade. BigData analytics with Machine Learning were found to be an efficient way for churn prediction. Several previous works [1] [7] [8] [14] focused on various data mining techniques for churn prediction based on call detail records. The work in [13] focused on service failures and disconnections recorded to identify churn. Study [5] focusses to detect early warnings of churn by assigning “Churn Score” for numerous customer transaction logs.

1.2 METHODOLOGY

This thesis aims to study and analyze customer churn based on data usage volumes with respect to QoE and users’ perspective using BigData Analytics. Three different analysis includes calculation and analysis of Mean, Standard deviation, Autocorrelations and Confidence intervals. Decision tree analysis includes data acquisition, data preparation that includes normalization, data preprocessing, data extraction and finally decision making.

1.3 OBJECTIVE

Churn analytics is the process of measuring the rate at which customers quit the product, site, or service. It answers the questions “Are we losing customers?” and “If so, how?” to allow teams to take action. Lower churn rates lead to happier customers, larger margins, and higher profits

Telecom operators need to be able to accurately predict churn in order to respond in time. The prime aspiration of this thesis is to predict customer churn from monthly and weekly mobile data usage volumes using BigData analytics. This thesis along with two partner theses is collaborated and united to form a lone crucial and dominant project. The main objectives include:

- Conduct survey with different sections of people regarding their data usage and numerous other questions [28].

- Analyze weekly QoE polls and volume measurements by Android-based tool compared [27].

- Study the importance of data preprocessing, data normalization and feature selection.

- Carefully analyze and assess six-month aggregate data usage volumes for active and churned users given by an anonymous Telecom provider.

- Carry out statistical and decision tree analyses for three datasets; one from Telecom provider and others from accompanying theses.

- Correlate and compare the results to know to which extent only data usage volumes could be used to predict churn.

- Finally, affirm the necessary information required for prominent churn prediction

1.4 MAIN CONTRIBUTION

Telecom operators need to be able to accurately predict churn in order to respond in time. The prime aspiration of this thesis is to predict customer churn from monthly and weekly mobile data usage volumes using BigData analytics. This thesis along with two partner theses is collaborated and united to form a lone crucial and dominant project. The main objectives include:

- Conduct survey with different sections of people regarding their data usage and numerous other questions [28].
- Analyze weekly QoE polls and volume measurements by Android-based tool compared [27]. • Study the importance of data preprocessing, data normalization and feature selection.
- Carefully analyze and assess six-month aggregate data usage volumes for active and churned users given by an anonymous Telecom provider.
- Carry out statistical and decision tree analyses for three datasets; one from Telecom provider and others from accompanying theses.
- Correlate and compare the results to know to which extent only data usage volumes could be used to predict churn.
- Finally, affirm the necessary information required for prominent churn prediction.

1.5 RELATED WORK

In today's scenario there already exist several projects detecting customer churn in telecom organizations. Each project consists of some pros and cons.

Ullah et al. [3] propose a churn prediction model that uses classification, as well as clustering techniques to identify the churn customers and provides the factors behind the churning of customers in the telecom sector. Feature selection is performed by using information gain and correlation attribute ranking filter. The proposed model first classifies churn customers data using classification algorithms, in which the algorithm performed well with 88.63% correctly classified instances. Creating effective retention policies is an essential task of the CRM to prevent churners. After classification, the proposed model segments the churning customer's data by categorizing the churn customers in groups using cosine similarity to provide group-based retention offers.

This paper also identified churn factors that are essential in determining the root causes of churn.

Li et al. [4] discuss the development of a model using big data analytics strategy for the dataset provided and using it predictions are made regarding the list of customers with their susceptibility listed in descending order. After getting the list through the techniques mentioned in the previous step, user segmentation and piecewise regression are used to find the highly relevant parameters followed by the division of customers into different categories based on the above-found parameters. Using regression analysis one can estimate the prediction rates for different groups of customers. High computing storage took some time in predicting results and accuracy rates of about 80% were achieved. Based on the results the organization was able to prioritize the customers who shall be given extra attention to influence their mindsets of continuing their relations with thecompany.

K. Sandhya Rani et al. [5] proposed a churn prediction model for telecommunication companies using machine learning techniques namely logistic regression. A comparison is done on the efficiency of the algorithm on the available dataset. They have used R programming language which is not as reliable as Python. Essam Shaaban et al. [6] examined historical activity, and predictive modeling attempts to make predictions about future client behavior. Predictive modeling may be used to identify customers who are at risk of leaving [7]. Customer Relationship Management (CRM) data and DM are used to create client-level models that explain the chance that a certain customer would perform a specific action.Sales, marketing, and client retention-related activities are common. Various models may be used to characterize the distinction between churners and non-churners in an organization in terms of their behavior.

While Lee et al. [8] looked into the relationship between customer happiness and switching cost in the context of French mobile communications, they found that the greater the switching cost the less probable it is for customers to churn when customer satisfaction remains constant. According to Madden et al. [9], the key variables driving customer turnover were monthly ISP use and family income. Win-back tactics were proposed by Amin et al. [10] after analyzing customer attrition from the viewpoints of businesses, rivals, and consumers and concluding. Han et al. [11] investigated the link between consumer attitude, switching obstacles, customer satisfaction, and customer retention, concluding that customer satisfaction was positively connected with customer retention. For this reason, Oghojafor et al. [12] came up with techniques to reduce the rate of client turnover. Effective customer win- back should be traced back to the root cause of churn. Customer churn is an important factor to consider when determining whether or not customers can be won back, according to Tokman and colleagues [13].

While Lee et al. [8] looked into the relationship between customer happiness and switching cost in the context of French mobile communications, they found that the greater the switching cost the less probable it is for customers to churn when customer satisfaction remains constant. According to Madden et al. [9], the key variables driving

customer turnover were monthly ISP use and family income. Win-back tactics were proposed by Amin et al. [10] after analyzing customer attrition from the viewpoints of businesses, rivals, and consumers and concluding. Han et al. [11] investigated the link between consumer attitude, switching obstacles, customer satisfaction, and customer retention, concluding that customer satisfaction was positively connected with customer retention. For this reason, Oghojafor et al. [12] came up with techniques to reduce the rate of client turnover. Effective customer win- back should be traced back to the root cause of churn. Customer churn is an important factor to consider when determining whether or not customers can be won back, according to Tokman and colleagues [13].

Almuqren et.al. [15] addressed this in their research, which studies customer churn, prediction models. Because they depend on previous customer data, the existing churn prediction algorithms have a limited shelf life. The data loses predictive value with time [16], which may not offer telecom firms the optimal churn prediction experience. Structural data framework and realtime analytics need to be integrated to target consumers in real-time [17]. The existing churn prediction algorithms do not take into account regional and linguistic characteristics, which results in geographical and cultural sampling mistakes [18].

CHAPTER-2

2.1 AIM

To determine the reasons why the telecom company losses its customers.

2.2 PROBLEM STATEMENT

In the competitive Telecom industry, public policies and standardization of mobile communication allow customers to easily switch over from one carrier to another, resulting in a strained fluidic market. Churn prediction, or the task of identifying customers who are likely to discontinue use of a service, is an important and lucrative concern of the Telecom industry. The aim of this thesis is to study and analyze customer churn prediction based on mobile data usage volumes with respect to QoE and users’ perspective with the help of BigData analytics.

2.3 CUSTOMER CHURN ANALYSIS

To perform a customer churn analysis, you’ll need a database of customer information and a spreadsheet or other program to dig into the data. You may be able to export stats relevant to customer churn such as churn rate and customer renewal rates directly from various ERP modules to save time and improve accuracy.

While you can and should conduct a high-level churn analysis for all customers, to get meaningful insights, you'll want to break that data down by product, region, client segment or other granular metrics specific to your company. That will give your team insights into where and potentially why you're losing customers.

Say you own a salon chain with seven locations and an average 5% churn. If one site spikes to 10% or more across a few months, it's clear that there is an issue at that shop. Or, for a software-as-a-service (SaaS) provider that sells multiple software solutions, when one has a much higher churn rate than others, you'll want to evaluate whether the application needs improvements, or if it's time to adjust pricing and packaging. When you resolve problems, you should see your churn rate decline.

HOW TO CALCULATE CUSTOMER CHURN

To calculate customer churn, you'll need two stats: the number of customers at the start of the period and the number at the end. The delta is the number of customers that left your service or stopped buying your product.

HOW DO YOU CALCULATE CUSTOMER CHURN

Follow this formula to calculate the total churn rate for a month, quarter or year

Churn Rate = Number of Lost or Cancelled Customers / Ending Total Customer Count

2.4 BENEFITS OF ANALYZING CUSTOMER CHURN

Why devote effort to customer churn analysis when you could spend the time on a long list of other projects? Here's a snapshot of benefits that accrue from analyzing customer churn.

Increase profits

A business sells its products or services to make money. The ultimate goal of a customer churn analysis, then, is to increase profits by lowering customer attrition. If more customers stay around for a longer period, you should see revenues increase and profits follow.

Improve the customer experience

One of the worst reasons to lose a customer is an easily avoidable mistake, like shipping the wrong item. Understanding why customers churn can help you learn about their priorities, identify your own weaknesses and improve the overall customer experience.

Also referred to as "CX," customer experience is your customers' perception or opinion about their interactions with your business. Their view of your brand is shaped throughout the buyer's journey, from their first interactions to post-sale support, and has a lasting impact on your company, including your bottom line.

Optimize your products and services

If customers are leaving because of specific problems with your products or services or delivery methods, you now have opportunities to improve. Not only will acting on these insights decrease customer churn, it leads to a better overall product or service that earns you more future growth.

Customer retention

The opposite of customer churn is customer retention—a business's ability to keep its customers and continue to generate revenue from them. Strong customer retention allows a business to increase the profitability of existing clients and maximize their lifetime value (LTV).

If you sell a service that costs $1,000 per month, keeping a customer for three additional months means you'll bring in an additional $3,000 in revenue per customer without spending on acquisition. The scale and dollar amount vary by business, but the concept is universal: Repeat business is profitable business.

CHAPTER-3

3.1 IMPLEMENTATION

Steps implemented for Data Preparation The “Capture and Analyze” and “Report and Predict” components that handle the collection and aggregation of data into a single, easy to refer to as well as a compatible dataset making it easy to work with the data analysis and model building stages as well as using the dataset to find out about what factors influence the customer churn scenario concerning a particular organization and using the same to build machine learning models that will predict whether a new customer will churn out or not. Before passing the dataset for the further stages, we also focus our efforts on cleaning the dataset (making sure that the data present within it has no abnormalities related to missing values, outdated matches, etc.) The data collection and assimilation process consist of collecting the data from the following sources: label. We also used a correlation matrix and heat map for the depiction of the degree of relationship between features and labels. After the conclusion of our EDA part we jumped on to our model building part in which we tried out the following four algorithms: - 1) KNearest Neighbors 2) Random Forest Classifier 3) Decision Tree Classifier 4) Principal Component Analysis. Once we found KNN to be the best model giving the most accurate results we dumped it using pickle and deployed our project using Flask

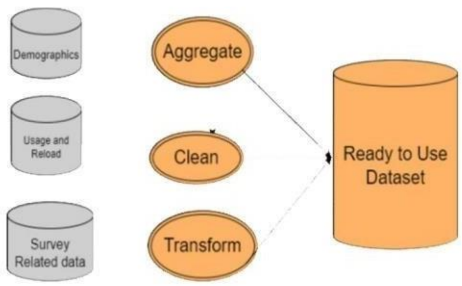


Fig. 1. Capture and Analyze

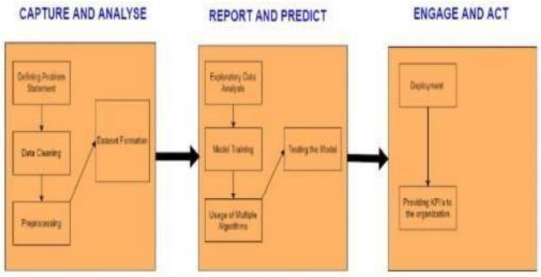


Fig. 2. Architecture Diagram

a) Customer Demographic Data:

A. Dataset This data is concerned with the reports released by telecom organizations which provide us with important factors from the organization’s perspective in a realtime environment. These factors are related to the crucial ones that impact a customer’s decision related to utilizing the services of a particular telecom organization

b) Usage and Reload Data: Several research studies have utilized datasets that contain information about the parameters related to subscribers' utilization of the network. Extensive work has been done on datasets through various other data scientists and data analysts that are shared on public sites such as Google Dataset research and Kaggle are also utilized here.

c) Survey related Data: Lots of online surveys were conducted that collected customer feedback and their expectations in the present as well as future from telecommunication services. So, through these responses, we were able to map it to the real-time attributes that we collected in the usage and reload data module.

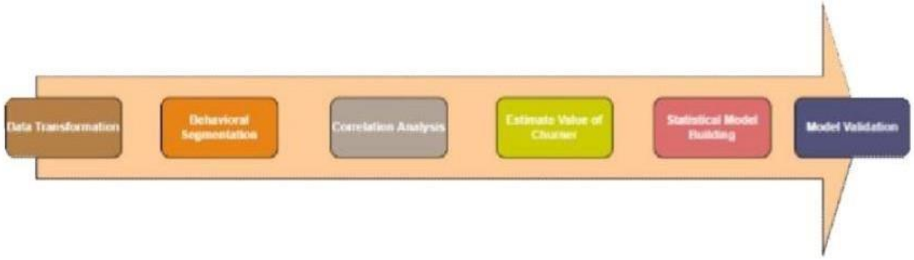


Fig. 3. Report and Predict

Under the report and predict section we carried out Exploratory data analysis in which we performed Data Sourcing, Data Cleaning, Univariate, and Bivariate Analysis. Under Data sourcing we imported our dataset and performed some basic data transformation. We also did the behavioral segmentation by segregating the churn ratio. Under Bivariate and Univariate analysis, we drew various insights and comparisons between various features and also the churn In our dataset, there are 21 columns and 7043 rows which were after the Exploratory Data Analysis part converted to 51 columns and 7043 rows. Our dataset contains the “Churn” column which is the label and the rest of the 50 columns are predictors. As we have imported our dataset from Kaggle it was imbalanced with some missing values. We have dealt with all the problems and have rectified them in the data cleaning and exploration section of ourEDA. B. Exploratory Data Analysis Exploratory Data Analysis is the method of data summarization using various charts and graphs. It’s also used to get familiar with the dataset. The objective of EDA is to increase confidence in our dataset to such a level so that we can build some model out of it.

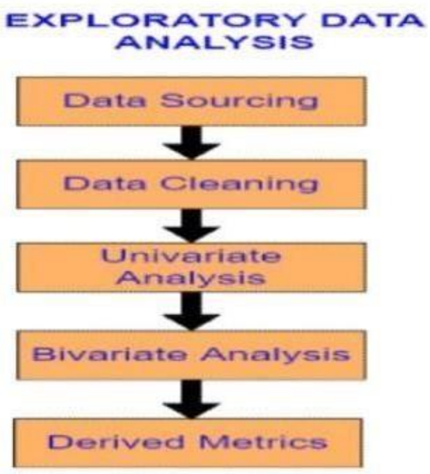
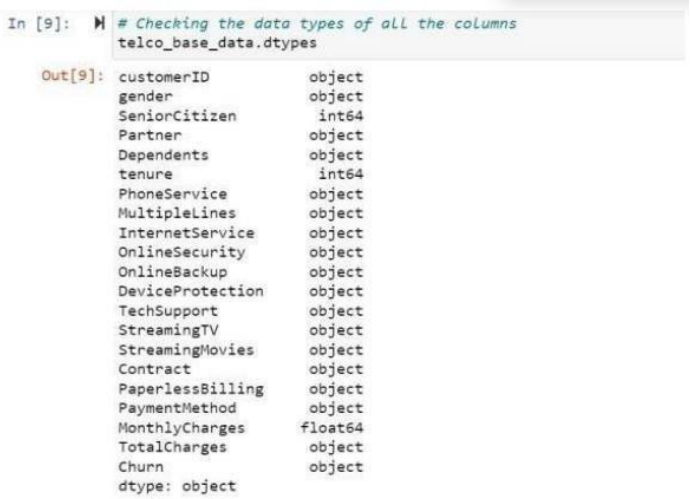


Fig. 4. EDA flow Diagram

In figure 4, we can see that through EDA we have known about the data types of all the features and labels smoothly. So from the above figure, we found out that there are only three numerical type features i.e. Senior Citizen, tenure, Monthly Chargers



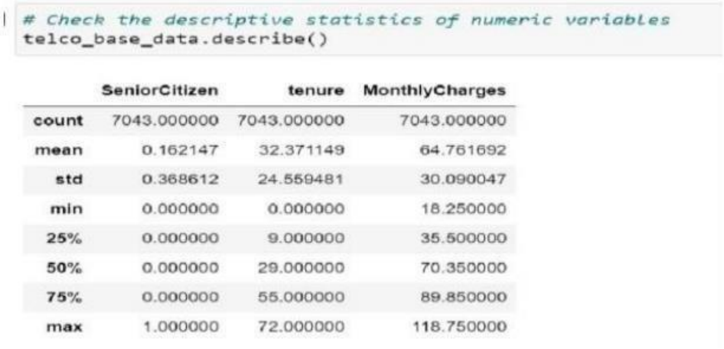
```
In [9]:  # Checking the data types of all the columns
         telco_base_data.dtypes

Out[9]:  customerID           object
         gender               object
         SeniorCitizen         int64
         Partner              object
         Dependents           object
         tenure                int64
         PhoneService         object
         MultipleLines        object
         InternetService      object
         OnlineSecurity       object
         OnlineBackup         object
         DeviceProtection     object
         TechSupport          object
         StreamingTV          object
         StreamingMovies      object
         Contract             object
         PaperlessBilling     object
         PaymentMethod        object
         MonthlyCharges      float64
         TotalCharges         object
         Churn                object
         dtype: object
```

Fig. 5. Datatypes of various feature

In figure 5, we have used describe function to get a good idea of the 3 numerical datasets. After studying and analyzing the above table we have found three insights. 1)Senior Citizen is a categorical feature hence the 25%-50%-75% distribution is not proper.

2) 75% of customers have a tenure of fewer than 55 months.

3) Average Monthly charges are USD 64.76 whereas 25% of customers pay more than USD 89.85 per month.

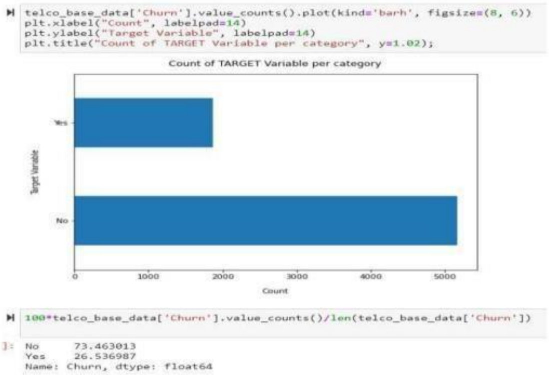


```
# Check the descriptive statistics of numeric variables
telco_base_data.describe()
```

	SeniorCitizen	tenure	MonthlyCharges
count	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692
std	0.368612	24.559481	30.090047
min	0.000000	0.000000	18.250000
25%	0.000000	9.000000	35.500000
50%	0.000000	29.000000	70.350000
75%	0.000000	55.000000	89.850000
max	1.000000	72.000000	118.750000

Fig 6. Description of numerical column

As we already know that our target variable is churn. Hence, we have plotted a bar graph between a count of customers vs churn and no-churn. The above figure illustrates the graph and also the ratio between the churn and non-churn which accounts for 73:27. This is an imbalanced dataset. As our data set is not balanced, we analyzed our data with other features while taking the target values separately to get some insights.



```
telco_base_data['Churn'].value_counts().plot(kind='barh', figsize=(8, 6))
plt.xlabel("Count", labelpad=14)
plt.ylabel("Target Variable", labelpad=14)
plt.title("Count of TARGET Variable per category", y=1.02);
```

```
100*telco_base_data['Churn'].value_counts()/len(telco_base_data['Churn'])
```

```
]:  No      73.463013
    Yes     26.536987
    Name: Churn, dtype: float64
```

Fig 7. Churn Vs Non-Churn ratio

3.2 DATA CLEANING:

This is the 2nd step of our EDA after the data sourcing that we have already seen above. Data cleaning is the process of cleaning data and improving the quality of the dataset for further data analysis and model building. The sole benefit of data cleaning is that we get a good quality dataset that is without any outliers, missing data, and useless columns. This enhances the model's accuracy.

There are plenty of histograms which is not possible to show in this paper. If we talk about the tenure group vs Churn (Figure 14) count the 1-12- month tenure group as having the highest churn ratio as the tenure is less. Whereas 61-72 months are having the lowest churn rate due to the longer duration of tenure. Moreover, in the gender vs churn plot (Figure 12), the churn ratio of both male and female are almost the same. So, we can eliminate these columns as they are not giving us many insights about churn rate. Likewise, In Figure 13 we can be confused for a moment and can think that non-senior citizens are the higher churners than senior citizens. But if we see the churn to a non-churn ratio in both senior citizens and non-senior citizens, we can see that ratio is higher in senior citizens. Hence, we can conclude that the senior citizens are high churners.

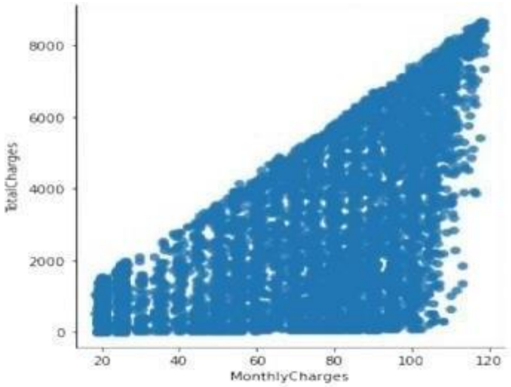


Fig 8. Monthly Charges Vs Total Charges

As we can see in the above figure.15 monthly charges is having a linear relationship with the Total charges as expected.

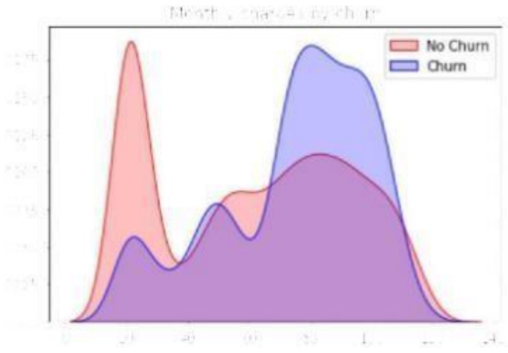


Fig 9. Monthly Charges Vs Density Churn

n Figure [16] it is quite evident that churn density is higher when the monthly charges are high and lower when the monthly charges are low. Bivariate Analysis: This is our 4th and final step in our EDA. Unlike, univariate analysis, Bivariate analysis deals withtwo variables instead of one.

The above shown in Fig depicts the correlation of different features concerning the churn label. In the above matrix, we can see some features having a positive, negative, and neutral correlation with the churn. The features which are neutral i.e. having a correlation score of 0, are not important features. Below figureillustratesthe correlation between different features along with the churn.

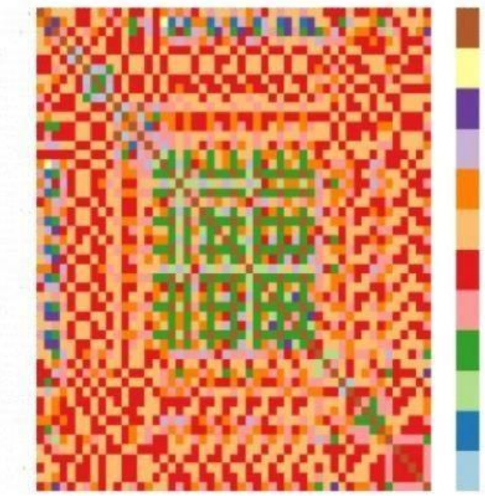


Fig 10. Heat Map

From the below figure it’s evident that the customers irrespective of their gender are high churners during the month-tomonth contract type. Conversely, As the duration of the contract type increased to two years, the churn rate also decreased drastically for both male and female customers.

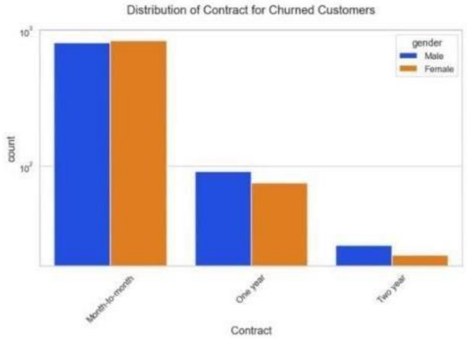
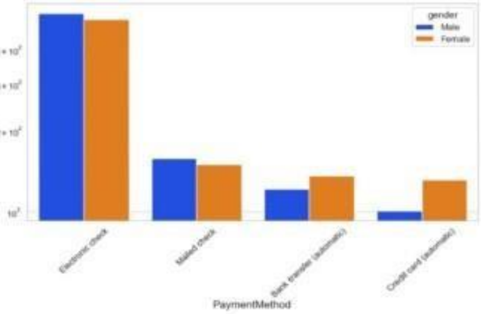


Fig 11 Contract Type Vs Churn Count

In this figure we can visualize an interesting insight that the customers using an electronic check as their payment method are the highest churners. Moreover, male credit card users are higher churners than their female counterparts.



3.3 DECISION TREE

After a substantial research in the field of churn prediction over many years, Big Data analytics were found to be an efficient way for identifying customer churn. Big Data analytics achieve better results more efficiently and receive insights that sets alarm bells ringing before any damage could happen, giving companies an opportunity to take precautionary measures. Decision trees are one of the predictive modeling approaches extensively used in data mining, where in a tree is used to explicitly represent decisions and decision making. These are the structured regression models. The goal of decision tree is to create a model that predicts the value of a target based on several input variables. Each node of a decision tree represents one of the traffic usage attributes of the customer. Leaves represent class labels and branches represent conjunctions of features that lead to class labels. The decision tree used in this thesis is J48, which is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. J48 is specifically chosen for its pruning ability and exceptional handling of missing classes, which no other tree could perform. Algorithm C4.5, often referred to as a statistical classifier is an extension to ID3 algorithm which builds decision trees from a dataset using the concept of information entropy [25]. C4.5 algorithm has been quite successful in achieving the discrepancies of ID3 algorithm as it could handle both continuous and discontinuous attributes by creating thresholds, then branching the values based on these thresholds. This algorithm could propitiously handle the missing attributes and helps in pruning the trees after creation, where in the size of decision tree is reduced by removing branches/leaves that provide very little power to classify instances. 10 Figure 4. Decision tree

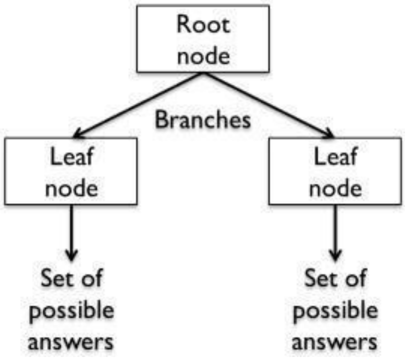


Fig 12. Decision tree

The trees are constructed in a top-down recursive divide-and-conquer manner. In a J48 decision tree, after the decisions are made, leaf nodes have two values. The first value demonstrates the total number of instances reaching the leaf and second value indicates the number of misclassified instances. In the case of missing values, fractional values are exhibited at the leaves. This tree by default uses 10-fold cross validation which states that 90% of the data is used for training and 10% for testing. As 90% is not too far from 100%, it gives affair estimate of the value. These cross validation folds can be switched at the “Test options” from Classify in WEKA. The accuracy achieved with a decision tree is far much higher than other data mining techniques which clearly states that decision tree is an efficient technique to predict churn

3.4 EXPLORATORY DATA ANALYSIS

The data was analytically explored to gain preliminary insights and uncover hidden patterns if they exist in the data. to start with, a correlation analysis was performed to understand the relationships between the variables. From the correlation matrix in Figure 3 below, a perfect positive linear correlation is observed between i q and torque while u d is observed to be highly negatively correlated with torque and i q. The former can be explained by electric drive theory, where either higher torque is exclusively dependent on i q in case of similar sized inductances in d-axis and q-axis or increasing with higher i q and slightly decreasing i d elsewise (more common in practice). The high correlation of torque further corroborates the decision to drop it. There also exists very strong positive correlations between the stator variables i.e., stator yoke, stator tooth and stator winding. This is expected as these are the temperatures of components in close proximity and sometimes in contact in an electric motor. This is however inconsequential as each of these was considered a response variable to be estimated from other variables in the absence of the other temperature variables.

3.5.1 UNIVARIATE ANALYSIS

Uni means one and variate means variable, so in univariate analysis, there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately. It is possible for two kinds of variables- Categorical and Numerical.

Countplot in Python

In this article, we will discuss how we can create a **countplot** using the seaborn library and how the different parameters can be used to infer results from the features of our dataset.

Seaborn library

The seaborn library is widely used among data analysts, the galaxy of plots it contains provides the best possible representation of our data.

The seaborn library can be imported into our working environment using

```
import seaborn as sns
```

Let us now discuss why do we use countplot and what is the significance of its parameters

The countplot is used to represent the occurrence(counts) of the observation present in the categorical variable.

It uses the concept of a bar chart for the visual depiction.

Parameters**-**

The following parameters are specified when we create a countplot, let us get a brief idea of them-

1. x and y- This parameter specifies the data we refer to for representation and then observes the highlighted patterns.
2. color- This parameter specifies the color that can give a good appearance to our plot.
3. palette- It takes the value of the palette. It is mostly used to show the hue variable.
4. hue- This parameter specifies the column name.
5. data- This parameter specifies the data frame we would like to take for the representation. For instance, data can be an array.
6. dodge- This parameter is an optional one and it accepts a Boolean value as input.
7. saturation- This parameter accepts a float value. A variation in the intensity of colors can be observed when we specify this.
8. hue_order- The parameter hue_order takes strings as an input.
9. kwargs- The parameter kwargs specifies the key and value mappings.
10. ax- The parameter ax is an optional one and is used to take axes on which plots are created.
11. orient- The parameter orient is an optional one and tells the orientation of the plot that we need, horizontal or vertical.

Lmplot **:**

lmplot() method is used to draw a scatter plot onto a FacetGrid.Parameters : This method is accepting the following parameters that are described below: x, y: (optional) This parameters are column names in data. data : This parameter is

DataFrame

KDEPLOT **:**

Kdeplot is a Kernel Distribution Estimation Plot which depicts the probability density function of the continuous or non-parametric data variables i.e. we can plot for the univariate or multiple variables altogether. Using the Python Seaborn module, we can build the Kdeplot with various functionality added to it.

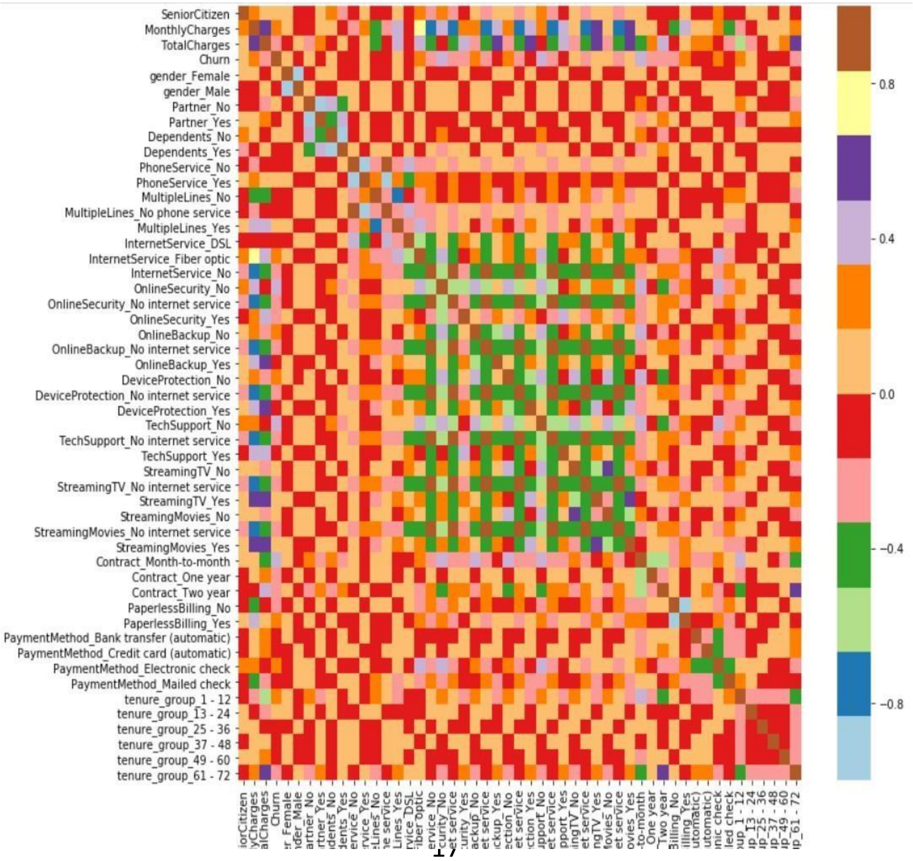
Derived Insight**:**

HIGH Churn seen in case of Month to month contracts, No online security, No Tech support, First year of subscription and Fibre Optics Internet

LOW Churn is seens in case of Long term contracts, Subscriptions without internet service and The customers engaged for 5+ years

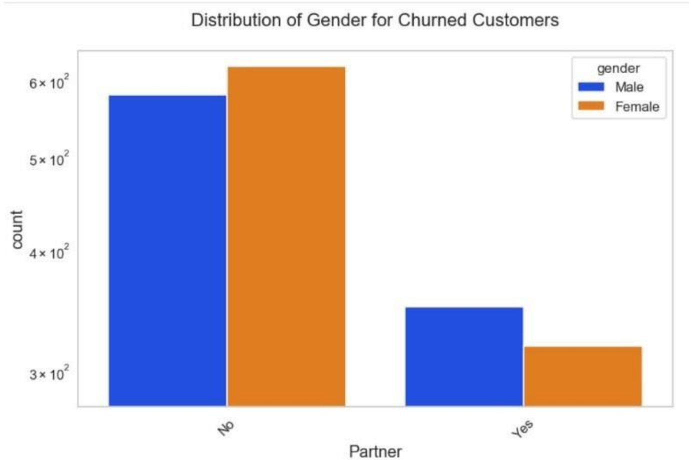
Factors like Gender, Availability of PhoneService and # of multiple lines have alomost NO impact on Churn

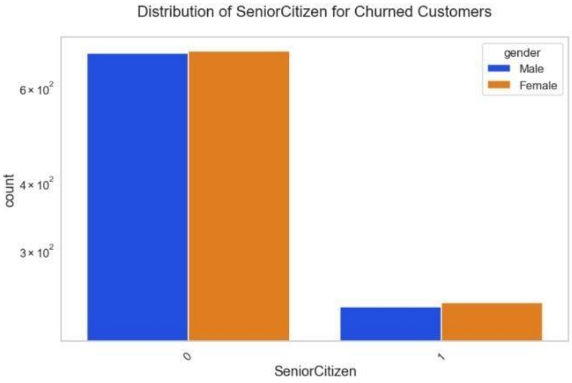
This is also evident from the Heatmap below



3.5.2 BIVARIATE ANALYSIS :

target variable of interest (or) using 2 variables and finding the relationship between Box plot, Violin plot. is performed to find the relationship between each variable in the dataset and the





3.6 RANDOM FOREST CLASSIFIER:

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5**:** For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

4. RESULTS AND DISCUSSIONS

4.1 IMPORTANT EDA INSIGHTS:

After the completion of EDA, we found some extremely useful and interesting insights and results. Firstly, Senior Citizen is categorical hence the 25%-50%75% distribution is not proper. Secondly, Higher Monthly charges, Lower tenure, and Lower Total charges are linked to High Churn. Thirdly, HIGH Churn has seen in case of Month-tomonth contracts, No online security, No Tech support, First year of subscription, and Fiber Optics Internet. Fourthly, LOW Churn is seen in the case of Long-term contracts, Subscriptions without internet service, and the customers engaged for 5+ years. Fifth, Factors like Gender, Availability of phone service and the number of multiple lines have almost NO impact on Churn. These are some more quick insights: - 1. Electronic check medium is the highest churner.

4.2 MODEL ACCURACY:

A propensity score is used to train machine learning algorithms to manage client attrition. The propensity score is vital since it indicates the likelihood of a client leaving soon. CCP models' forecast accuracy relies on this. The next section will outline the different stages of the churn prediction process and the associated machine learning algorithms. We tried out many algorithms. Out of which three of the algorithms gave us the best accuracy. Random Forest Classifier with SMOTEENN gave us a whopping 9393.5% accuracy. 0 and 1 stand for non-churners and churners respectively. The precision determines the accuracy of the prediction. The recall is the ratio of actual positives and the total positive cases. F-1 score is the mean of precision and recall

TABLE I. RANDOM FOREST CLASSIFICATION REPORT

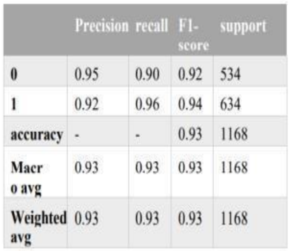
	Precision	recall	F1-score	support
0	0.94	0.90	0.92	530
1	0.92	0.96	0.94	649
accuracy	-	-	0.93	1179
Macro avg	0.93	0.93	0.93	1179
Weighted avg	0.93	0.93	0.93	1179

TABLE II. RANDOM FOREST CONFUSION MATRIX

	Positive	Negative
Positive	TP = 479	FN = 51
Negativ e	FP = 28	TN = 621

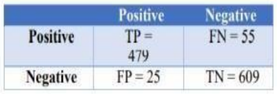
Decision Tree Classifier with SMOTEEN gave us around the same 93.1% accuracy.

TABLE III. DECISION TREE CLASSIFICATION REPORT



	Precision	recall	F1-score	support
0	0.95	0.90	0.92	534
1	0.92	0.96	0.94	634
accuracy	-	-	0.93	1168
Macr o avg	0.93	0.93	0.93	1168
Weighted avg	0.93	0.93	0.93	1168

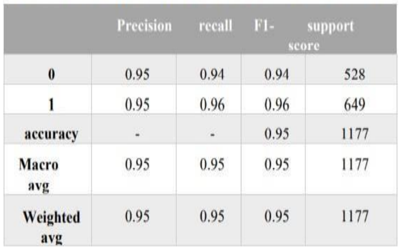
TABLE IV. DECISION TREE CONFUSION MATRIX



	Positive	Negative
Positive	TP = 479	FN = 55
Negative	FP = 25	TN = 609

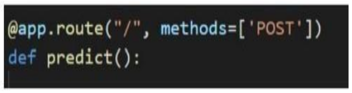
KNN with SMOTEEN yielded an impressive 94-95% accuracy. This is the model which gave us the best accuracy.

TABLE V. K-NEAREST NEIGHBOR CLASSIFICATION REPORT



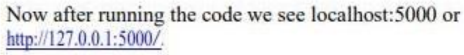
	Precision	recall	F1-score	support
0	0.95	0.94	0.94	528
1	0.95	0.96	0.96	649
accuracy	-	-	0.95	1177
Macro avg	0.95	0.95	0.95	1177
Weighted avg	0.95	0.95	0.95	1177

The predict method is our POST method, which is called when we pass all the inputs from our front end and click SUBMIT.



```
@app.route("/", methods=['POST'])
def predict():
```

The run() method of the Flask class runs the application on the local development server.



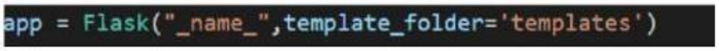
Now after running the code we see localhost:5000 or http://127.0.0.1:5000/.

```
app.run()
```

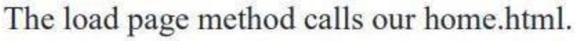
```
* Running on http://127.0.0.1:5000 (Press CTRL+C to quit)
```

5. SUMMARY AND CONCLUSION

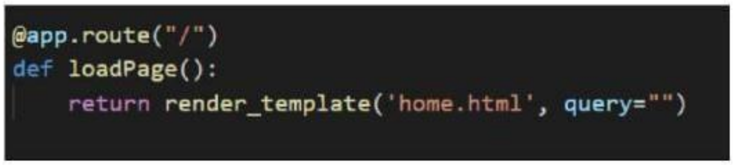
The importance of this kind of project is very necessary for today's competitive environment. Especially in the telecom sector where customer quite often churns and results in big losses to the telecom company. From churning losses, we should learn a big lesson to be more aware and interactive with the customers to retain them in case they are willing to churn. One possible way could be through hiring and training a Customer Relationship Manager (CRM) to such an extent that customer never thinks of churning and become an Unconditionally loyal subscriber. The old and traditional ways of working in telecommunication organizations should be amended. Many companies are still not aware of the benefits of Machine Learning and Data Science in predicting the likelihood of churn. As a result, many companies tend to make mistakes by not hiring ML engineers and then losing customers. Our model has been prepared by using a variety of Algorithms like k-means, k-means++, Random Forest Classifier, Decision Tree Classifier, and K-Nearest- Neighbors. Out of many only three algorithms were selected and finally a single algorithm i.e. KNN was selected for our model building giving 94-95% accuracy. Apart from accuracy, we have also improved that our F1 score, Precision, and Recall are good



```
app = Flask("_name_",template_folder='templates')
```



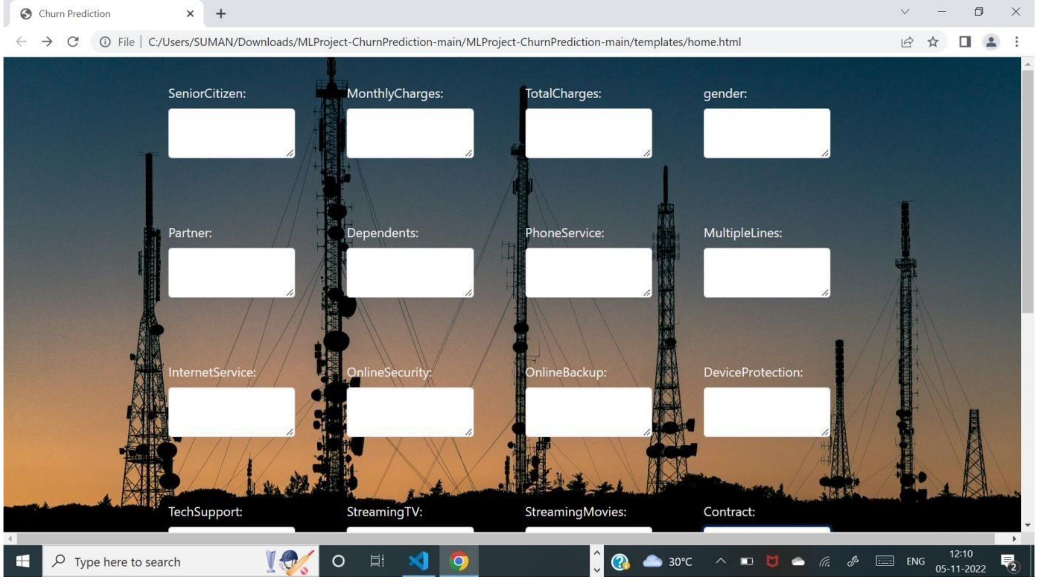
The load page method calls our home.html.

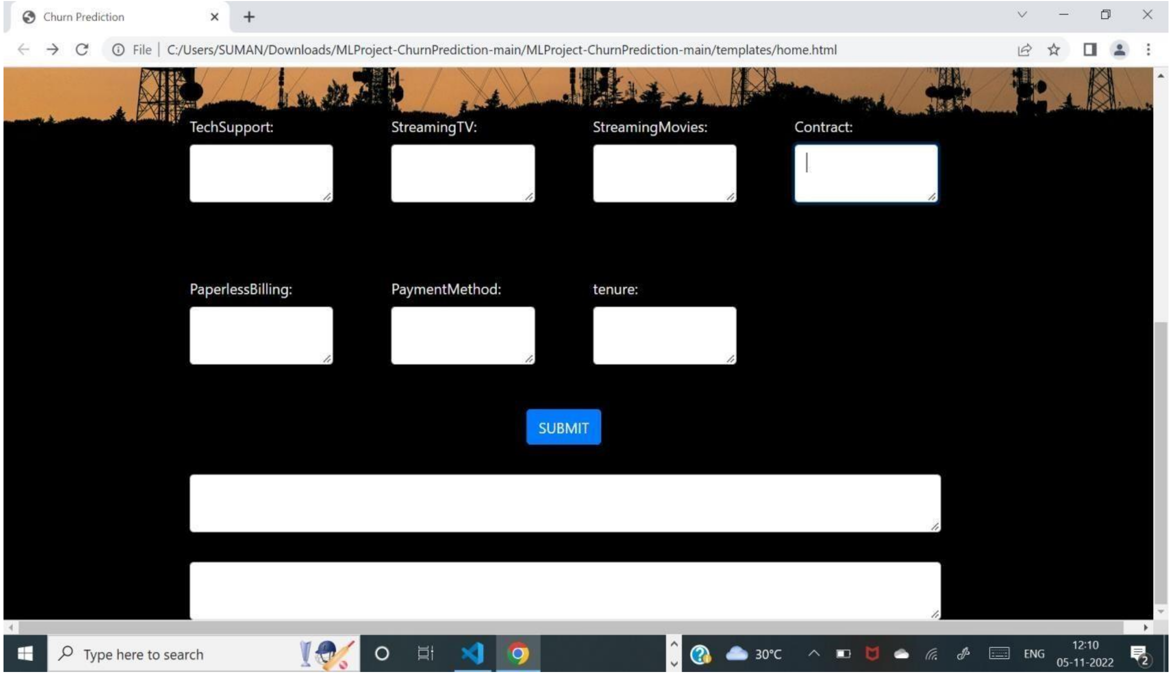


```
@app.route("/")
def loadPage():
    return render_template('home.html', query="")
```

APPENDIX :

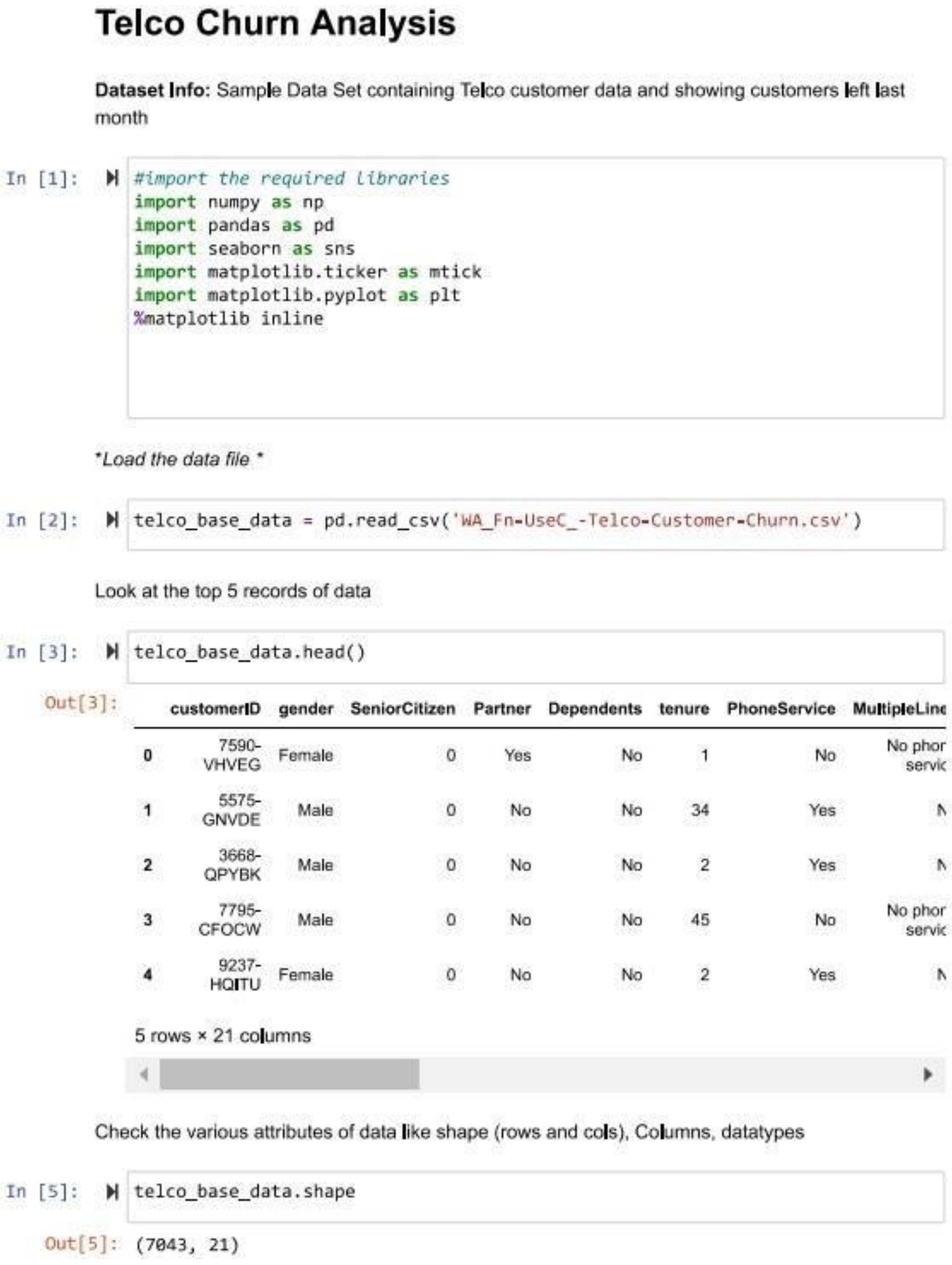
(A). SCREENSHOTS :





(B).SOURCE CODE :

EDA



Telco Churn Analysis

Dataset Info: Sample Data Set containing Telco customer data and showing customers left last month

```
In [1]:  #import the required libraries
         import numpy as np
         import pandas as pd
         import seaborn as sns
         import matplotlib.ticker as mtick
         import matplotlib.pyplot as plt
         %matplotlib inline
```

*Load the data file *

```
In [2]:  telco_base_data = pd.read_csv('WA_Fn-UseC_-Telco-Customer-Churn.csv')
```

Look at the top 5 records of data

```
In [3]:  telco_base_data.head()
```

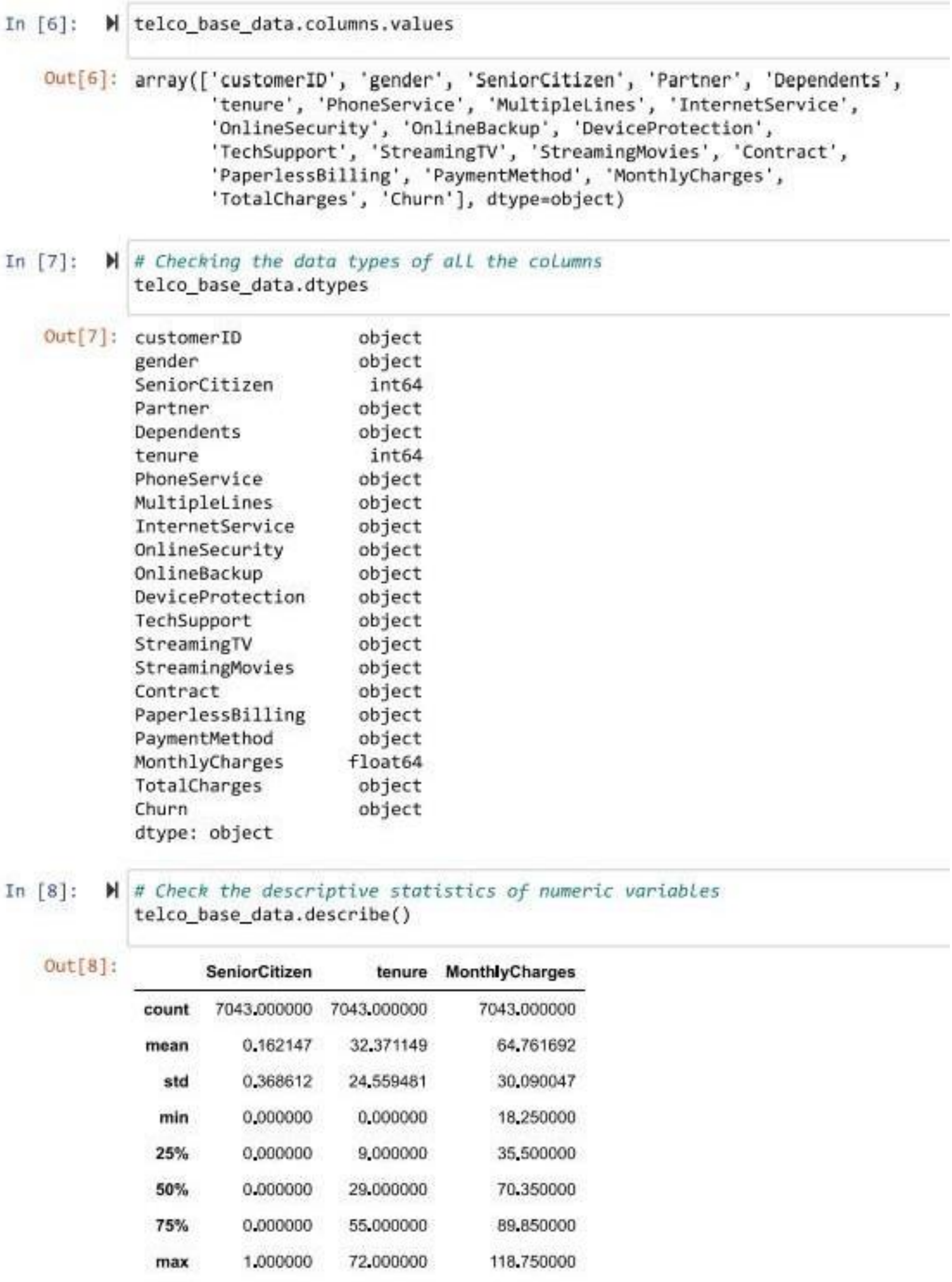
Out[3]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLine
0	7590-VHVEG	Female	0	Yes	No	1	No	No phor servic
1	5575-GNVDE	Male	0	No	No	34	Yes	N
2	3668-QPYBK	Male	0	No	No	2	Yes	N
3	7795-CFOCW	Male	0	No	No	45	No	No phor servic
4	9237-HQITU	Female	0	No	No	2	Yes	N

5 rows × 21 columns

Check the various attributes of data like shape (rows and cols), Columns, datatypes

```
In [5]:  telco_base_data.shape

Out[5]:  (7043, 21)
```



```
In [6]:  telco_base_data.columns.values

Out[6]:  array(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents',
                'tenure', 'PhoneService', 'MultipleLines', 'InternetService',
                'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
                'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',
                'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges',
                'TotalCharges', 'Churn'], dtype=object)
```

```
In [7]:  # Checking the data types of all the columns
         telco_base_data.dtypes

Out[7]:  customerID           object
         gender               object
         SeniorCitizen         int64
         Partner              object
         Dependents           object
         tenure                int64
         PhoneService         object
         MultipleLines        object
         InternetService      object
         OnlineSecurity       object
         OnlineBackup         object
         DeviceProtection     object
         TechSupport          object
         StreamingTV          object
         StreamingMovies      object
         Contract             object
         PaperlessBilling     object
         PaymentMethod        object
         MonthlyCharges      float64
         TotalCharges         object
         Churn                object
         dtype: object
```

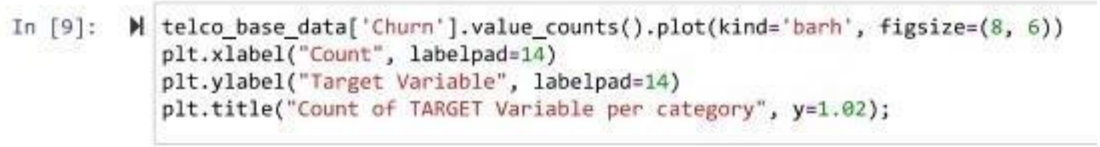
```
In [8]:  # Check the descriptive statistics of numeric variables
         telco_base_data.describe()

Out[8]:
```

	SeniorCitizen	tenure	MonthlyCharges
count	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692
std	0.368612	24.559481	30.090047
min	0.000000	0.000000	18.250000
25%	0.000000	9.000000	35.500000
50%	0.000000	29.000000	70.350000
75%	0.000000	55.000000	89.850000
max	1.000000	72.000000	118.750000

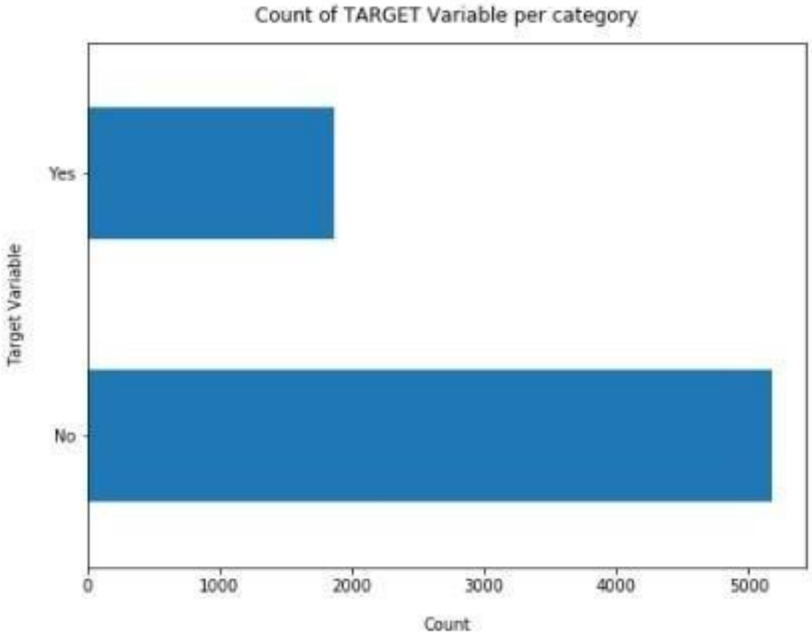
SeniorCitizen is actually a categorical hence the 25%-50%-75% distribution is not propoer

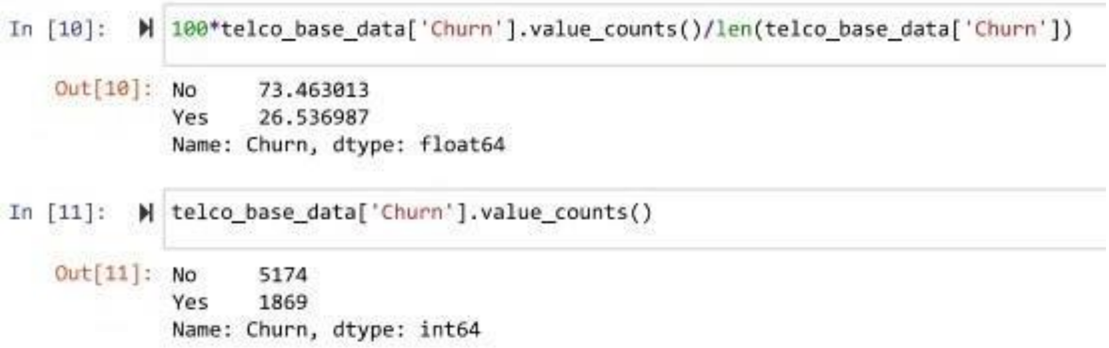
75% customers have tenure less than 55 months

Average Monthly charges are USD 64.76 whereas 25% customers pay more than USD 89.85 per month



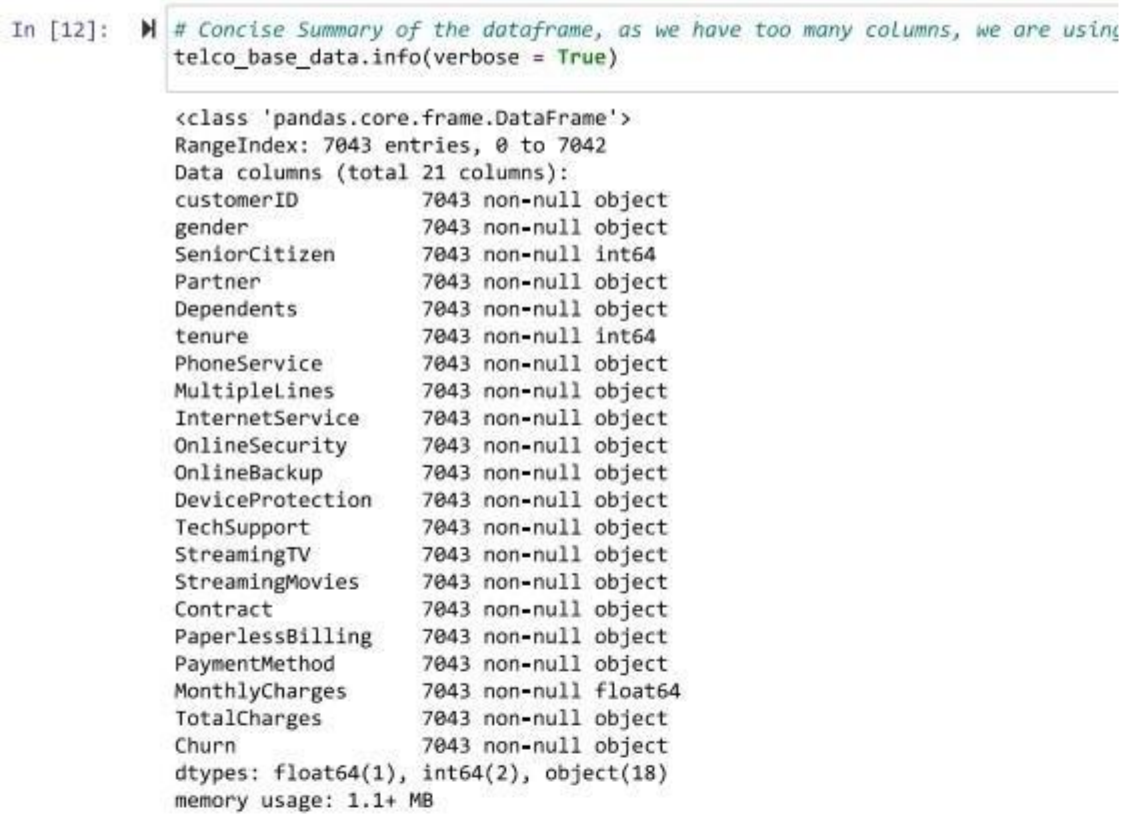
```
In [9]:  telco_base_data['Churn'].value_counts().plot(kind='barh', figsize=(8, 6))
         plt.xlabel("Count", labelpad=14)
         plt.ylabel("Target Variable", labelpad=14)
         plt.title("Count of TARGET Variable per category", y=1.02);
```



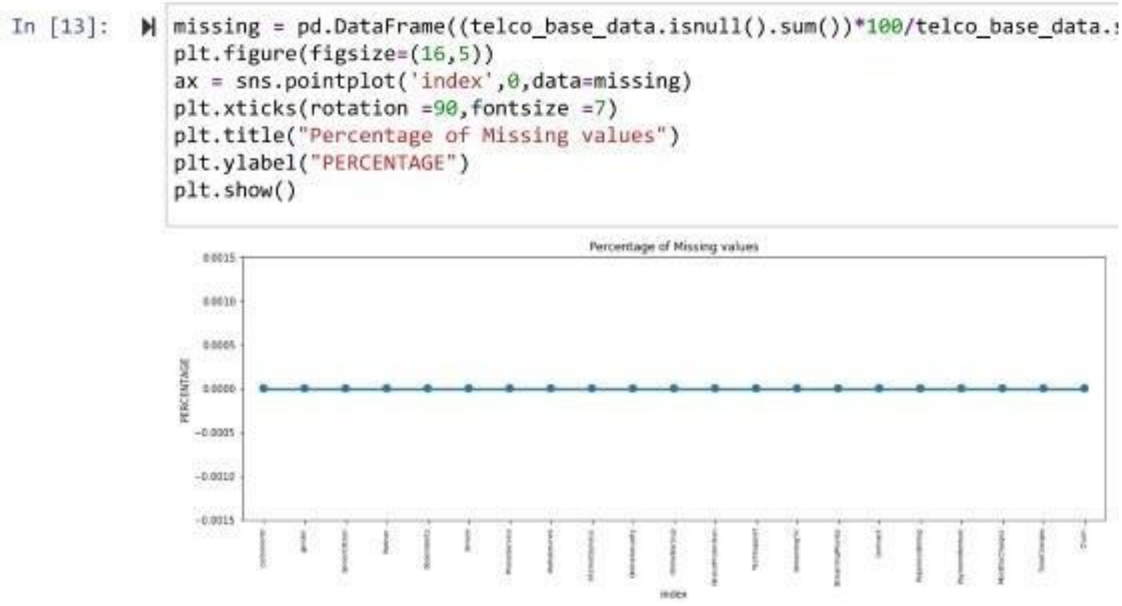


```
In [10]:  100*telco_base_data['Churn'].value_counts()/len(telco_base_data['Churn'])

Out[10]: No     73.463013
         Yes    26.536987
         Name: Churn, dtype: float64

In [11]:  telco_base_data['Churn'].value_counts()

Out[11]: No     5174
         Yes    1869
         Name: Churn, dtype: int64
```

- Data is highly imbalanced, ratio = 73:27
- So we analyse the data with other features while taking the target values separately to get

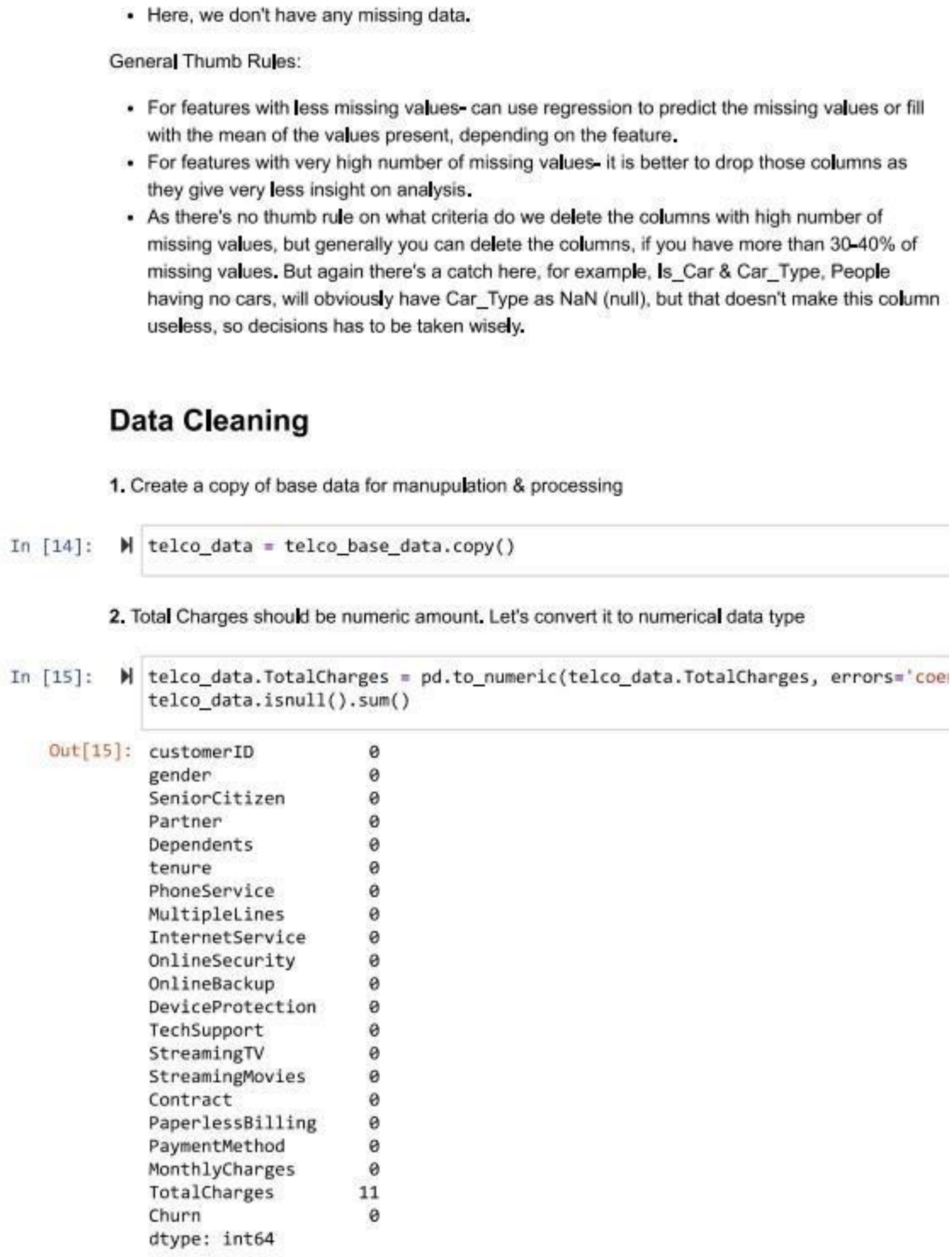


```
In [12]:  # Concise Summary of the dataframe, as we have too many columns, we are using
           telco_base_data.info(verbose = True)

           <class 'pandas.core.frame.DataFrame'>
           RangeIndex: 7043 entries, 0 to 7042
           Data columns (total 21 columns):
           customerID          7043 non-null object
           gender              7043 non-null object
           SeniorCitizen       7043 non-null int64
           Partner             7043 non-null object
           Dependents          7043 non-null object
           tenure              7043 non-null int64
           PhoneService        7043 non-null object
           MultipleLines       7043 non-null object
           InternetService     7043 non-null object
           OnlineSecurity      7043 non-null object
           OnlineBackup        7043 non-null object
           DeviceProtection    7043 non-null object
           TechSupport         7043 non-null object
           StreamingTV         7043 non-null object
           StreamingMovies     7043 non-null object
           Contract            7043 non-null object
           PaperlessBilling    7043 non-null object
           PaymentMethod       7043 non-null object
           MonthlyCharges      7043 non-null float64
           TotalCharges        7043 non-null object
           Churn               7043 non-null object
           dtypes: float64(1), int64(2), object(18)
           memory usage: 1.1+ MB
```



```
In [13]:  missing = pd.DataFrame((telco_base_data.isnull().sum())*100/telco_base_data.s
           plt.figure(figsize=(16,5))
           ax = sns.pointplot('index',0,data=missing)
           plt.xticks(rotation =90,fontsize =7)
           plt.title("Percentage of Missing values")
           plt.ylabel("PERCENTAGE")
           plt.show()
```

Missing Data - Initial Intuition



- Here, we don't have any missing data.

General Thumb Rules:

- For features with less missing values- can use regression to predict the missing values or fill with the mean of the values present, depending on the feature.
- For features with very high number of missing values- it is better to drop those columns as they give very less insight on analysis.
- As there's no thumb rule on what criteria do we delete the columns with high number of missing values, but generally you can delete the columns, if you have more than 30-40% of missing values. But again there's a catch here, for example, Is_Car & Car_Type, People having no cars, will obviously have Car_Type as NaN (null), but that doesn't make this column useless, so decisions has to be taken wisely.

Data Cleaning

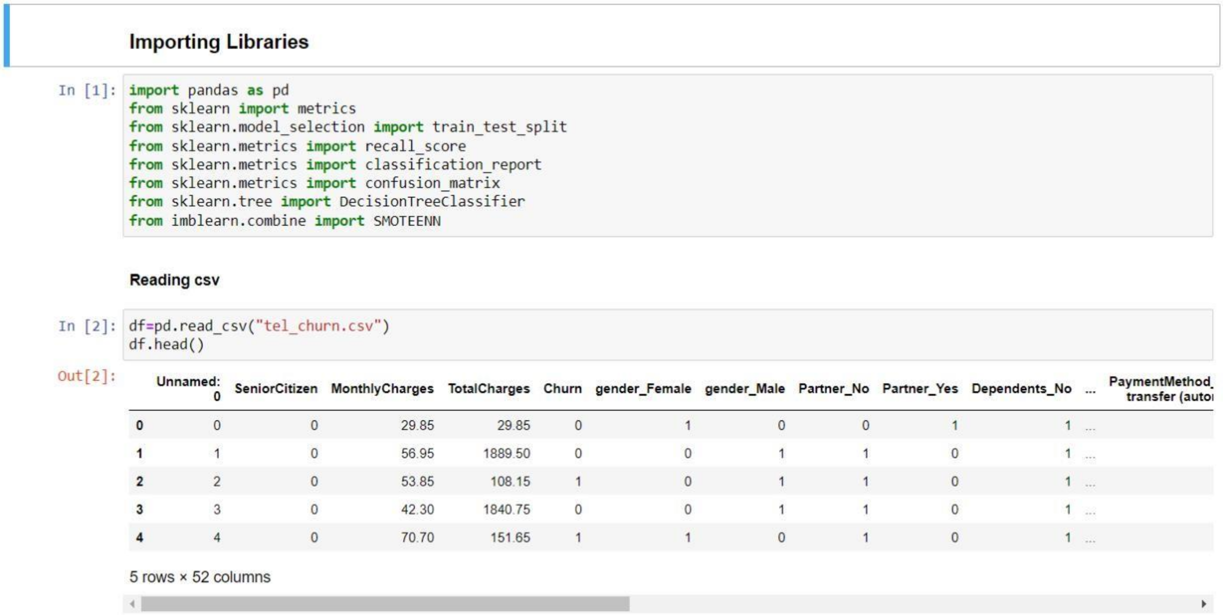
1. Create a copy of base data for manupulation & processing

```
In [14]:  telco_data = telco_base_data.copy()
```

2. Total Charges should be numeric amount. Let's convert it to numerical data type

```
In [15]:  telco_data.TotalCharges = pd.to_numeric(telco_data.TotalCharges, errors='coe
          telco_data.isnull().sum()

Out[15]:  customerID           0
          gender               0
          SeniorCitizen        0
          Partner              0
          Dependents           0
          tenure               0
          PhoneService         0
          MultipleLines        0
          InternetService      0
          OnlineSecurity       0
          OnlineBackup         0
          DeviceProtection     0
          TechSupport          0
          StreamingTV          0
          StreamingMovies      0
          Contract             0
          PaperlessBilling     0
          PaymentMethod        0
          MonthlyCharges       0
          TotalCharges        11
          Churn                0
          dtype: int64
```

MODEL BUILDING :



Importing Libraries

```
In [1]: import pandas as pd
        from sklearn import metrics
        from sklearn.model_selection import train_test_split
        from sklearn.metrics import recall_score
        from sklearn.metrics import classification_report
        from sklearn.metrics import confusion_matrix
        from sklearn.tree import DecisionTreeClassifier
        from imblearn.combine import SMOTEENN
```

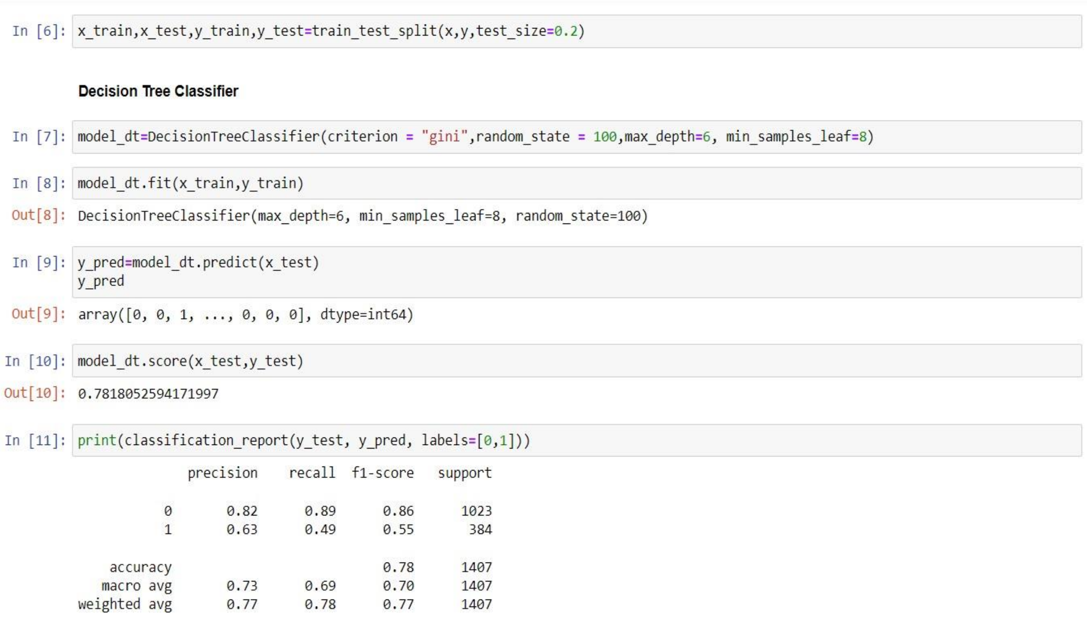
Reading csv

```
In [2]: df=pd.read_csv("tel_churn.csv")
        df.head()
```

Out[2]:

	Unnamed: 0	SeniorCitizen	MonthlyCharges	TotalCharges	Churn	gender_Female	gender_Male	Partner_No	Partner_Yes	Dependents_No	...	PaymentMethod_ transfer (autoı
0	0	0	29.85	29.85	0	1	0	0	1	1	...	
1	1	0	56.95	1889.50	0	0	1	1	0	1	...	
2	2	0	53.85	108.15	1	0	1	1	0	1	...	
3	3	0	42.30	1840.75	0	0	1	1	0	1	...	
4	4	0	70.70	151.65	1	1	0	1	0	1	...	

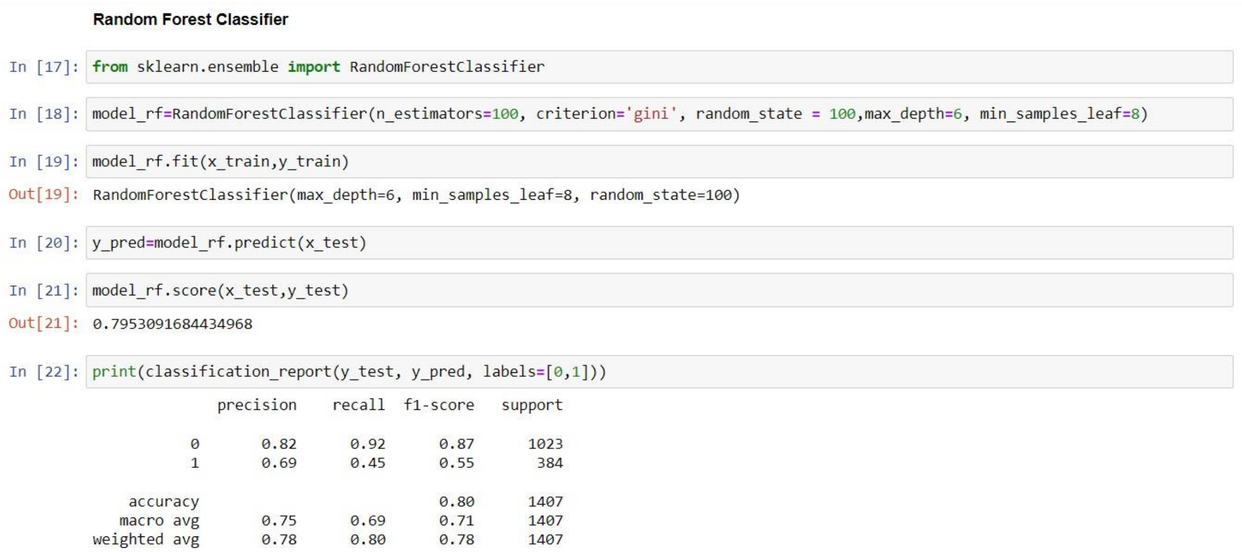
5 rows × 52 columns



```
In [6]: x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2)
```

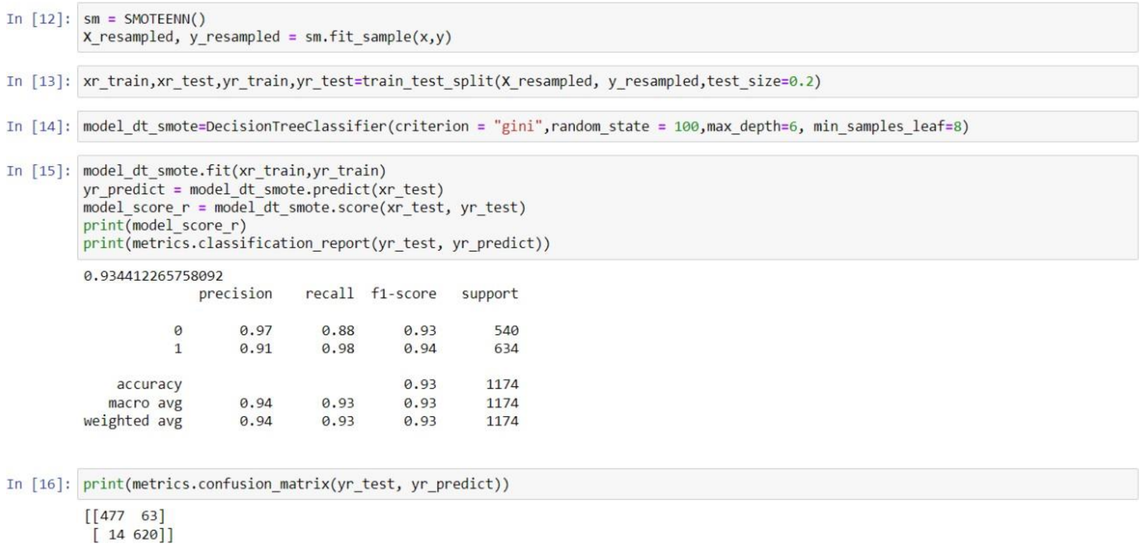
Decision Tree Classifier

```
In [7]: model_dt=DecisionTreeClassifier(criterion = "gini",random_state = 100,max_depth=6, min_samples_leaf=8)

In [8]: model_dt.fit(x_train,y_train)

Out[8]: DecisionTreeClassifier(max_depth=6, min_samples_leaf=8, random_state=100)

In [9]: y_pred=model_dt.predict(x_test)
        y_pred

Out[9]: array([0, 0, 1, ..., 0, 0, 0], dtype=int64)

In [10]: model_dt.score(x_test,y_test)

Out[10]: 0.7818052594171997

In [11]: print(classification_report(y_test, y_pred, labels=[0,1]))

                  precision    recall  f1-score   support

               0       0.82      0.89      0.86      1023
               1       0.63      0.49      0.55       384

        accuracy                           0.78      1407
       macro avg       0.73      0.69      0.70      1407
    weighted avg       0.77      0.78      0.77      1407
```

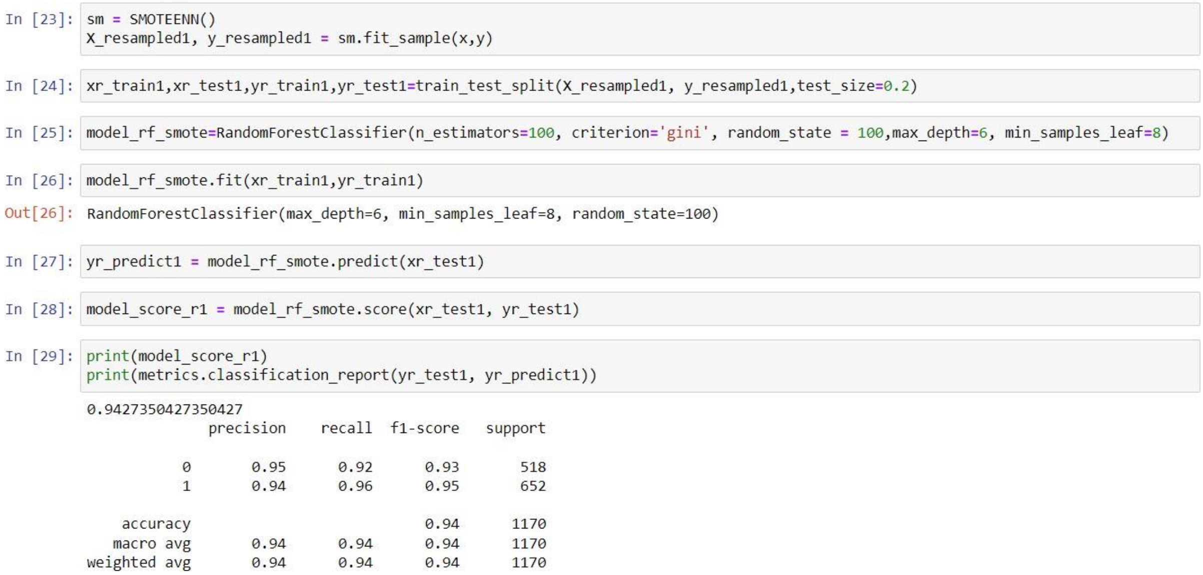


Random Forest Classifier

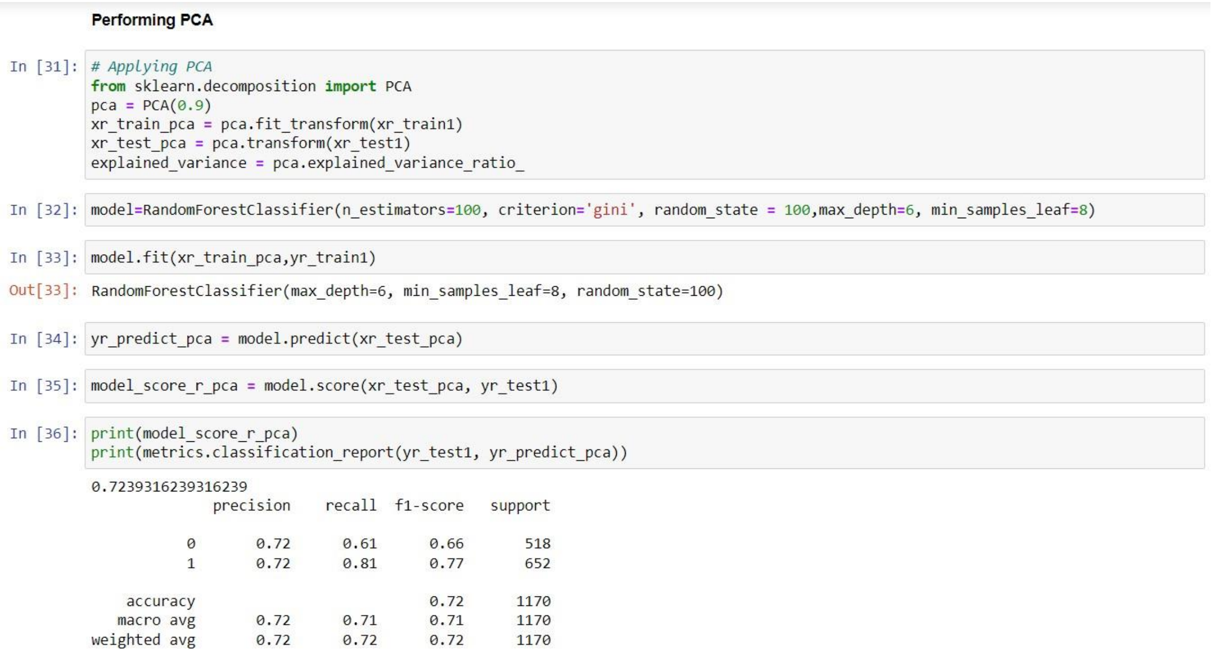
```
In [17]: from sklearn.ensemble import RandomForestClassifier

In [18]: model_rf=RandomForestClassifier(n_estimators=100, criterion='gini', random_state = 100,max_depth=6, min_samples_leaf=8)

In [19]: model_rf.fit(x_train,y_train)

Out[19]: RandomForestClassifier(max_depth=6, min_samples_leaf=8, random_state=100)

In [20]: y_pred=model_rf.predict(x_test)

In [21]: model_rf.score(x_test,y_test)

Out[21]: 0.7953091684434968

In [22]: print(classification_report(y_test, y_pred, labels=[0,1]))

              precision    recall  f1-score   support

           0       0.82      0.92      0.87      1023
           1       0.69      0.45      0.55       384

    accuracy                           0.80      1407
   macro avg       0.75      0.69      0.71      1407
weighted avg       0.78      0.80      0.78      1407
```



```
In [12]: sm = SMOTEENN()
         X_resampled, y_resampled = sm.fit_sample(x,y)

In [13]: xr_train,xr_test,yr_train,yr_test=train_test_split(X_resampled, y_resampled,test_size=0.2)

In [14]: model_dt_smote=DecisionTreeClassifier(criterion = "gini",random_state = 100,max_depth=6, min_samples_leaf=8)

In [15]: model_dt_smote.fit(xr_train,yr_train)
         yr_predict = model_dt_smote.predict(xr_test)
         model_score_r = model_dt_smote.score(xr_test, yr_test)
         print(model_score_r)
         print(metrics.classification_report(yr_test, yr_predict))

         0.934412265758092
                       precision    recall  f1-score   support

                    0       0.97      0.88      0.93       540
                    1       0.91      0.98      0.94       634

             accuracy                           0.93      1174
            macro avg       0.94      0.93      0.93      1174
         weighted avg       0.94      0.93      0.93      1174

In [16]: print(metrics.confusion_matrix(yr_test, yr_predict))

         [[477  63]
          [ 14 620]]
```

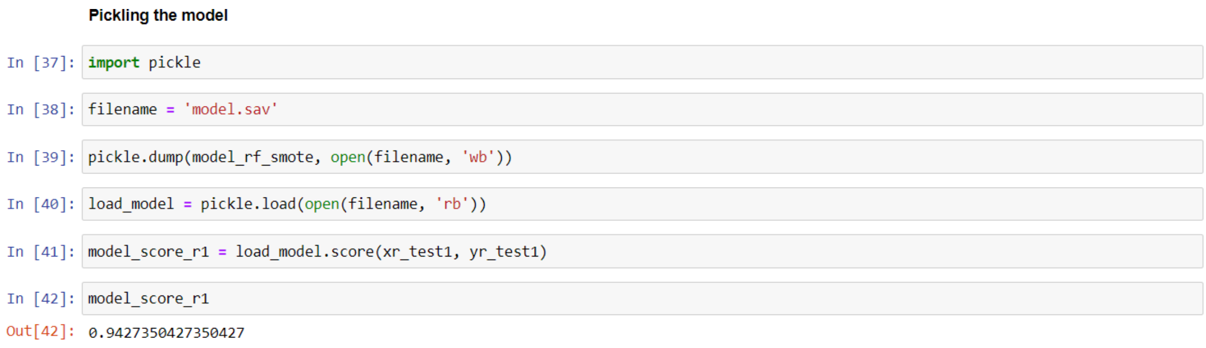


```
In [23]: sm = SMOTEENN()
         X_resampled1, y_resampled1 = sm.fit_sample(x,y)

In [24]: xr_train1,xr_test1,yr_train1,yr_test1=train_test_split(X_resampled1, y_resampled1,test_size=0.2)

In [25]: model_rf_smote=RandomForestClassifier(n_estimators=100, criterion='gini', random_state = 100,max_depth=6, min_samples_leaf=8)

In [26]: model_rf_smote.fit(xr_train1,yr_train1)

Out[26]: RandomForestClassifier(max_depth=6, min_samples_leaf=8, random_state=100)

In [27]: yr_predict1 = model_rf_smote.predict(xr_test1)

In [28]: model_score_r1 = model_rf_smote.score(xr_test1, yr_test1)

In [29]: print(model_score_r1)
         print(metrics.classification_report(yr_test1, yr_predict1))

         0.9427350427350427
                       precision    recall  f1-score   support

                    0       0.95      0.92      0.93       518
                    1       0.94      0.96      0.95       652

             accuracy                           0.94      1170
            macro avg       0.94      0.94      0.94      1170
         weighted avg       0.94      0.94      0.94      1170
```



Performing PCA

```
In [31]: # Applying PCA
         from sklearn.decomposition import PCA
         pca = PCA(0.9)
         xr_train_pca = pca.fit_transform(xr_train1)
         xr_test_pca = pca.transform(xr_test1)
         explained_variance = pca.explained_variance_ratio_

In [32]: model=RandomForestClassifier(n_estimators=100, criterion='gini', random_state = 100,max_depth=6, min_samples_leaf=8)

In [33]: model.fit(xr_train_pca,yr_train1)

Out[33]: RandomForestClassifier(max_depth=6, min_samples_leaf=8, random_state=100)

In [34]: yr_predict_pca = model.predict(xr_test_pca)

In [35]: model_score_r_pca = model.score(xr_test_pca, yr_test1)

In [36]: print(model_score_r_pca)
         print(metrics.classification_report(yr_test1, yr_predict_pca))

         0.7239316239316239
                       precision    recall  f1-score   support

                    0       0.72      0.61      0.66       518
                    1       0.72      0.81      0.77       652

             accuracy                           0.72      1170
            macro avg       0.72      0.71      0.71      1170
         weighted avg       0.72      0.72      0.72      1170
```



Pickling the model

```
In [37]: import pickle

In [38]: filename = 'model.sav'

In [39]: pickle.dump(model_rf_smote, open(filename, 'wb'))

In [40]: load_model = pickle.load(open(filename, 'rb'))

In [41]: model_score_r1 = load_model.score(xr_test1, yr_test1)

In [42]: model_score_r1

Out[42]: 0.9427350427350427
```

HTML CODE :

```
<html>
<head>
  <link rel="stylesheet"
href="https://maxcdn.bootstrapcdn.com/bootstrap/4.3.1/css/bootstrap.min.css">
  <link rel="stylesheet" href="./style.css">

</head>
  <body>
     <img src="./mario-caruso-0C9VmZUqcT8-unsplash.jpg" alt="">
      <title>Churn Prediction</title>
```

```
<div class="container">
<div class="row">

<form action="http://localhost:5000/" method="POST">
  <div class="col-sm-9 container1">
  <div class="form-group  purple-border">
    <label for="comment">SeniorCitizen:</label>
    <textarea class="form-control" rows="2" id="query1" name="query1" rows="2"
cols="5" autofocus></textarea>
</div>

<div class="form-group  purple-border">
  <label for="comment">MonthlyCharges:</label>
  <textarea class="form-control" rows="2" id="query2" name="query2" rows="2"
cols="5" autofocus></textarea>
</div>

<div class="form-group  purple-border">
  <label for="comment">TotalCharges:</label>

<textarea class="form-control" rows="2" id="query3" name="query3" rows="2"
cols="5" autofocus></textarea>
</div>

<div class="form-group  purple-border">
   <label for="comment">gender:</label>
  <textarea class="form-control" rows="2" id="query4" name="query4" rows="2"
cols="5" autofocus></textarea>
</div>

<div class="form-group  purple-border">

<label for="comment">Partner:</label>

 <textarea class="form-control" rows="2" id="query5" name="query5" rows="2"
cols="5" autofocus></textarea>
</div>

<div class="form-group  purple-border">

 <label for="comment">Dependents:</label>

<textarea class="form-control" rows="2" id="query6" name="query6" rows="2"
cols="5" autofocus></textarea></div>
```

```
<div class="form-group  purple-border">

<label for="comment">PhoneService:</label>

<textarea class="form-control" rows="2" id="query7" name="query7" rows="2"
cols="5" autofocus></textarea>

</div>

<div class="form-group  purple-border">

 <label for="comment">MultipleLines:</label>

<textarea class="form-control" rows="2" id="query8" name="query8" rows="2"
cols="5" autofocus></textarea>

</div>

<div class="form-group  purple-border">
 <label for="comment">InternetService:</label>
<textarea class="form-control" rows="2" id="query9" name="query9" rows="2"
cols="5" autofocus></textarea>
</div>

<div class="form-group  purple-border">

<label for="comment">OnlineSecurity:</label>

 <textarea class="form-control" rows="2" id="query10" name="query10" rows="2"
cols="5" autofocus></textarea>

</div>

<div class="form-group  purple-border">

 <label for="comment">OnlineBackup:</label>

 <textarea class="form-control" rows="2" id="query11" name="query11" rows="2"
cols="5" autofocus></textarea>

</div>

<div class="form-group  purple-border">
```

```
<label for="comment">DeviceProtection:</label>

 <textarea class="form-control" rows="2" id="query12" name="query12" rows="2"
cols="5" autofocus></textarea>

</div>

<div class="form-group  purple-border">

 <label for="comment">TechSupport:</label>

<textarea class="form-control" rows="2" id="query13" name="query13" rows="2"
cols="5" autofocus></textarea>

</div>

<div class="form-group  purple-border">

<label for="comment">StreamingTV:</label>

 <textarea class="form-control" rows="2" id="query14" name="query14" rows="2"
cols="5" autofocus></textarea>

</div>

<div class="form-group  purple-border">

<label for="comment">StreamingMovies:</label>

<textarea class="form-control" rows="2" id="query15" name="query15" rows="2"
cols="5" autofocus></textarea>

</div>

<div class="form-group  purple-border">

<label for="comment">Contract:</label>

<textarea class="form-control" rows="2" id="query16" name="query16" rows="2"
cols="5" autofocus></textarea>

</div>
```

```
<div class="form-group  purple-border">

<label for="comment">PaperlessBilling:</label>

<textarea class="form-control" rows="2" id="query17" name="query17" rows="2"
cols="5" autofocus></textarea>

</div>

<div class="form-group  purple-border">

<label for="comment">PaymentMethod:</label>

<textarea class="form-control" rows="2" id="query18" name="query18" rows="2"
cols="5" autofocus></textarea>

</div>

<div class="form-group  purple-border">

<label for="comment">tenure:</label>

 <textarea class="form-control" rows="2" id="query19" name="query19" rows="2"
cols="5" autofocus></textarea>

</div>
</div>

<div class="col-sm-3 submit">

<center><button type="submit" class="btn btn-primary"
name="submit">SUBMIT</button></center>

</div>
    </form>

 </div>


<div class="row">

<div class="col-sm-9 container2">
```

```
<textarea class="form-control" rows="2" id="comment" name="query6" rows="2"
cols="5" autofocus>output1</textarea>

<textarea class="form-control" rows="2" id="comment" name="query7" rows="2"
cols="5" autofocus>output2</textarea>
</div>
</div>
</div>

  <script
src="https://ajax.googleapis.com/ajax/libs/jquery/3.4.1/jquery.min.js"></script>

  <script
src="https://cdnjs.cloudflare.com/ajax/libs/popper.js/1.14.7/umd/popper.min.js"></script>

<script
src="https://maxcdn.bootstrapcdn.com/bootstrap/4.3.1/js/bootstrap.min.js"></script>

</body>
</html>
```

REFERENCES :

[1] Lalwani, P., Mishra, M.K., Chadha, J.S. et al. Customer churn prediction system: a machine learning approach. Computing 104, 271– 294 (2022). https://doi.org/10.1007/s00607-021- 00908-y.

[2] Chen, H., Chiang, R.H., Storey, V.C.: Business intelligence and analytics: From big data to big impact. MIS quarterly pp. 1165– 1188 (2012).

[3] Ullah, Irfan & Raza, Basit & Malik, Ahmad & Imran, Muhammad & Islam, Saif & Kim, Sung Won. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2914999.

[4] L. Li, T. Chi, T. Hao, T. Yu Customer demand analysis of the electronic commerce supply chain using big data Ann. Oper. Res., 268 (2018), pp. 113-128.

[5] Rani, K. Sandhya and ., Shaik Thaslima and ., N.G.L. Prasanna and ., R.Vindhya and ., P. Srilakshmi, Analysis of Customer Churn Prediction in Telecom Industry Using Logistic Regression (JUNE 10, 2021). International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347- 5552, Volume9, Issue- 4, July 2021, https://doi.org/10.21276/ijircst.2021.9.4.6

[6] Shaaban, Essam & Helmy, Yehia & Khedr, Ayman & Nasr, Mona. (2012). A Proposed Churn Prediction Model. International Journal of Engineering Research and Applications (IJERA. 2. 693-697