**Customer Churn Analysis in Telecom Industry**

**Objective:**

In this article, I’m going to apply data analytics skills on the telecom customer churn dataset which contains customer level information and try to find insights about the key attributes which are causative for customer analysis, we try to find the hidden trends from customer behavior through visualization techniques, build several predictive classification models. This article concludes by providing successfully predict potential customer churn, Companies can use such ML pipelines to initiate retention strategies on those customers who are classified as likely targets of churn.

**Abstract:**

With rapid development of telecommunication industry, the service providers are more inclined towards expansion of the customer base. Customer acquisition and retention has become a key concern for several industries and is particularly acute in fiercely competitive and growing business. Finding the Key factors which triggers the customer churn plays important role in early initiation of customer retention policies and cut back the churn. We will focus on analyzing the customer data, perform exploratory data analysis, to get insight about which variables are contributing to customer churn, implementing the machine learning algorithms to identify potential churn customers and label them based on usage patterns and visualize the results.

**Problem Statement:**

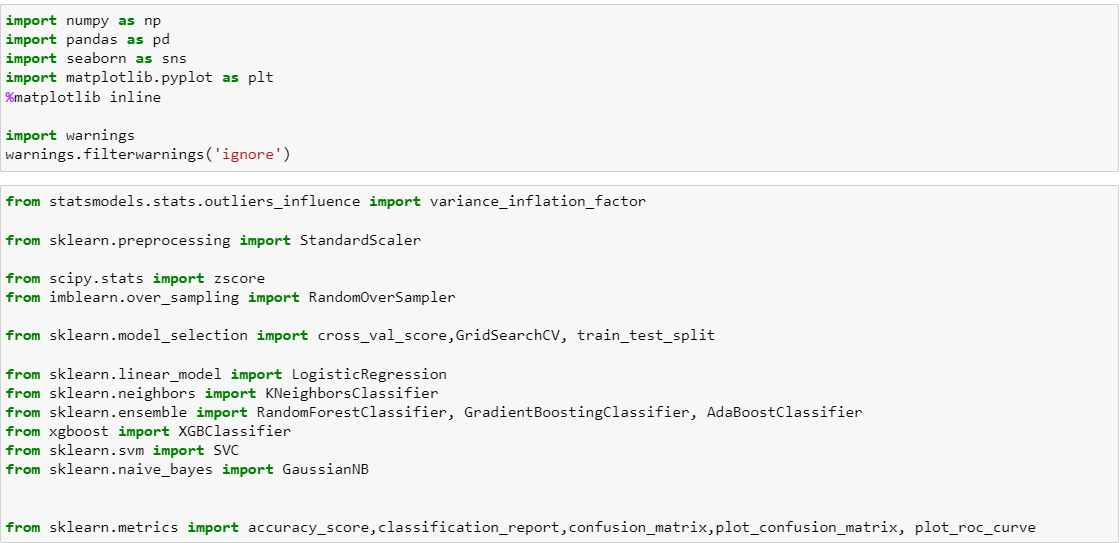
Customer churn is when a company’s customers stop doing business with that company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers. Existing customers will often have a higher volume of service consumption and can generate additional customer referrals.

Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to defect means that a company can prioritize focused marketing efforts on that subset of their customer base.

Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low.

**Data Analysis:**

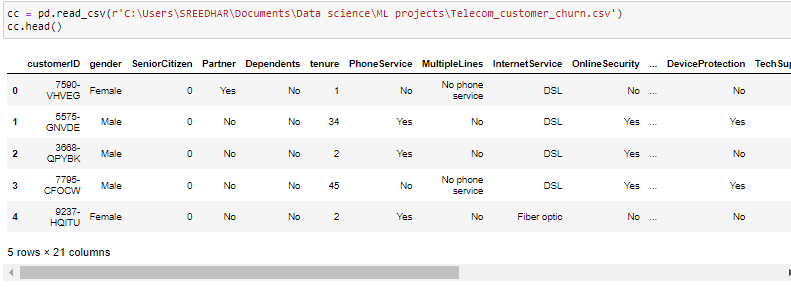
Let us import all the necessary libraries

****

Overview of dataset

Let us import the dataset, understand the data and what each record in row and column is about. Here each column consists of details of customer information which is recorded by company during the service. There are 4 types of information,

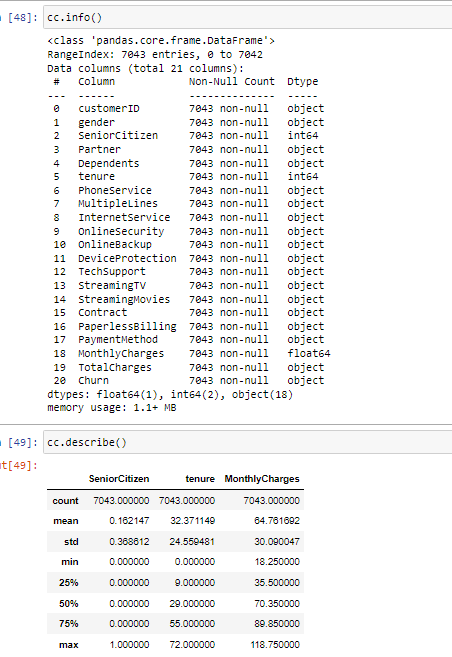
*i. Services used by customer. ii. Customer Demographic information (basic details) iii. Customer churn Details. Iv. Customer Account information.*

**

Following are the attributes available from the datasets which contains all features and (Churn) target variable.

Index(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents','tenure', 'PhoneService', MultipleLines', 'InternetService','OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport',

'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'Churn'], dtype='object')



Observations:

1. The dataset consists of 7043 rows and 21 columns,
2. The dataset contains 18 categorical and 3 numerical values. Total charges should be numerical data but it is showing categorical data, let us check what is missing.
3. There are no null values in the dataset.
4. The descriptive statistical overview of the dataset is available.
5. The below table provides brief understanding of all the columns and types of data present,

|  |  |  |
| --- | --- | --- |
| Column | Definition | Comments |
| Customer ID | Unique ID provided by company |  |
| gender | Gender of customer | Male, Female |
| Senior Citizen | If customer is Senior Citizen | 1= Yes, 0=No |
| Partner | If customer have partner | Yes, No |
| Dependents | If customer have any dependents | Yes, No |
| tenure | Since how many years customer is using service of company |  |
| Phone Service | If customer has phone service | Yes, No |
| Multiple Lines | If customer uses multiple line service | Yes, No, No phone service |
| Internet Service | Does customer have internet service | Fiber Optic, DSL, No |
| Online Security | Does customer use online security service | Yes, No, No internet service |
| Online Backup | Does customer use online backup service | Yes, No, No internet service |
| Device Protection | Does customer use device protection service | Yes, No, No internet service |
| Tech Support | Does customer use tech support of company | Yes, No, No internet service |
| Streaming TV | Does customer streams TV? | Yes, No, No internet service |
| Streaming Movies | Does customer Streams movies ? | Yes, No, No internet service |
| Contract | What type of contract does customer use? | Month-to-Month, One year, Two year |
| Paperless Billing | Does customer prefer paperless billing? | Yes, No |
| Payment Method | What is mode of payment customer opt for? | Payment Method |
| Monthly Charges | What is monthly charge of customer |  |
| Total Charges | Total charges since using the service |  |
| 'Churn' | Does customer leaves company or continues with the company service? | Yes, no |

**Exploratory Data Analysis**

Let us proceed with the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test, check assumptions with the help of summary statistics and graphical representations. And summarize the main characteristics for model building.

Univariate Analysis:

Gender Distribution and Senior Citizens:

* There are equal number of customers in our data set that male while the other half are female.
* There are 1142 Senior Citizens which comprise of only 16% and 5901 Non-Senior Citizens using the service.

****

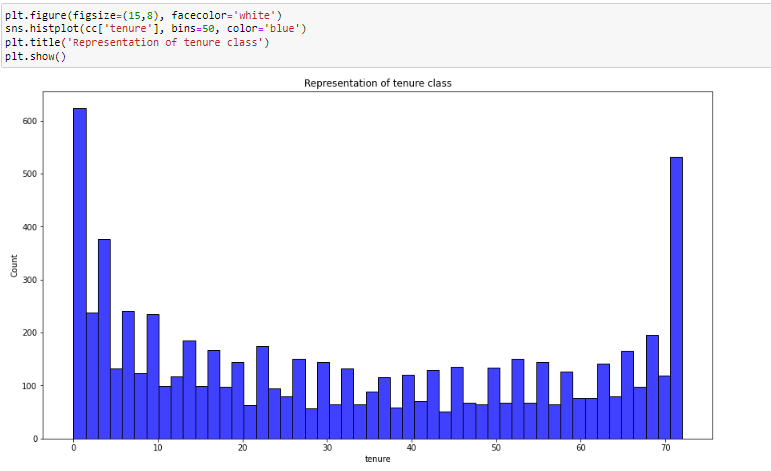
Partner and Dependents Distribution:

* There 3402, nearly 49-50% of customers having partners, and 70% of customers having dependents



Tenure:

* The average number of people staying in the service is nearly 150, Most number of people churn is seen after 1-2 months gradually, we can see more number of people staying in service after 70months.



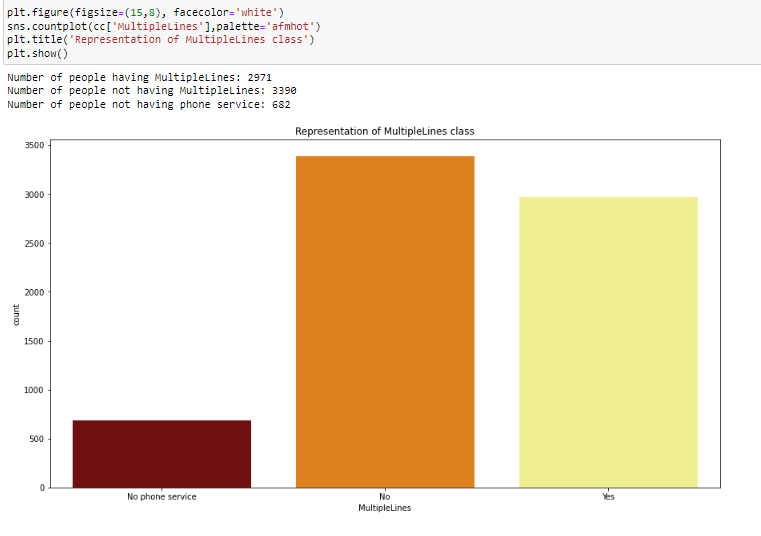
Phone Service and Internet Service:

* 21% of the customer data, do not have Phone Service, they are not contribution for any other service as well.
* 2. 79% of customer of overall has Internet Service, and 44% of them choose for Fiber Optics, and 35% choose for DSL Internet Service.



Multiple line Services:

* Only 40% of total customer has multiple lines, and 48% do not opt for multiple lines, 22% do not have phone service.



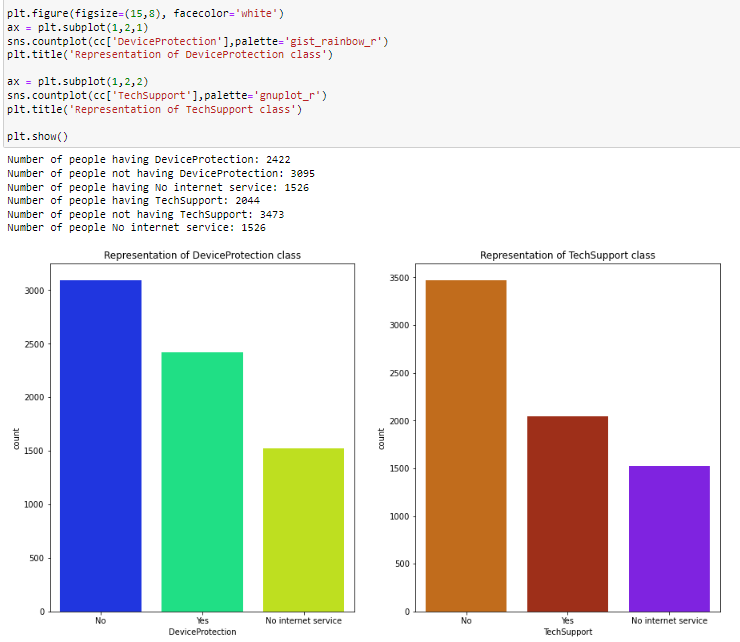
Online Security and Online Backup:

* Only 28% of total customer choose Online Security service, and 49% do not choose online Security service.
* And 34% of customer choose online backup service, 44% do not choose, and 22% do not have Internet Service.



Device Protection and Tech Support:

* 29% of customer Choose Tech Support service from the company,
* 34% customer choose Device protection service.



Streaming TV class and Streaming Movies class:

1. Nearly 50% of customers using Internet stream TV.
2. 48% of customers using internet will go for streaming Movies
3. 21% of customers do not have Internet services.



Contract class and Paperless Billing:

* 54% of customer choose Month-to-month contract, and 22% choose for 1 year contract,24% choose 2-year contract.
* 42% customer go for paper billing and 58% go for paperless billing.

****

Payment Methods and Churn:

* 27% of overall customer have churned and 73% of customer have not churned. The target data is unbalanced, and it need to be treated.
* Nearly 32% of payments have been done through electronic check, around 22-25% by other methods(Mailed check, Bank transfer, Credit Card)

****

Bivariate Analysis:

Let us visualize the relationship between the attributes and target variables.

Gender vs Churn and Senior Citizen vs Churn:

* There is equal number of male and female who are leaving company
* Out of 5901, 29% of non-senior customers and 30 % of senior customer have churned. i.e., most of young age people are leaving the services, hence need to attract young people by providing attractive offers.



Partner vs Churn, Dependents vs Churn:

* 22% of people having parters and 36% of people not having partners have churned.
* 16% of people having dependents have churned.



Phone Service vs Churn, Multiple lines vs Churn:

* People subscriber for phone service seems to have more churn, then that of not having service, as they might have discontinued using the service and stopped subscribing as well.
* Less percentage of customers who took multiple lines connection seems to have high churn ratio.



Internet Service vs Churn, Tenure vs Churn:

* The people <30 seems to move out at higher rate than that >30 tenure.
* People who opted for Fiber Optics Internet service have highest churn percentage, need to check if there is issue with this service.



Online Security vs Churn, Online Backup vs Churn:

* We can infer that people who not opted above 2 services seem to have churned more, we can think of developing more ideas in order to attract and make people to choose this service.

****

Payment Method vs Churn Paperless Billing vs Churn:

* People who choose credit card method seems to stay more than those of electronic check method.
* People who choose TV, Movies streaming services having higher ratio of churn. Need to focus on improvising in this area.



Tenure vs Monthly charges with Churn as hue:

* The monthly charges have been increase all over the tenures except at 2 spikes, but overall growth is positive.

****

Observations:

1. We have Gender, Partner, Dependents, Phone Service, Paperless Billing, Churn with 2 values(Yes, No). Let us replace them by 1 and 0 respectively.

2. Multiple Lines, Internet Service, Online Backup, Online Security, Device Protection, Tech Support, Streaming TV, Streaming Movies, Contract have 3 values which are to be One hot Encoded.

3. Payment Method has 4 values to be encoded.

4. As the customer ID column is having all unique values indicating ID, no use for analysis.

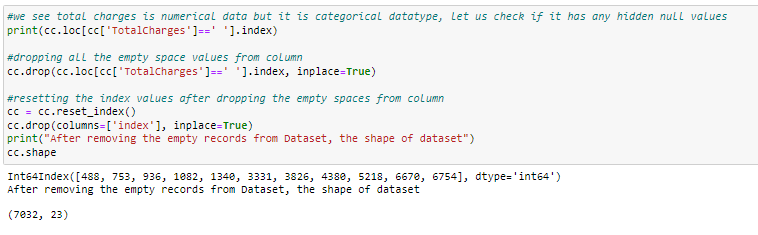
**Data Preprocessing**

Data preprocessing is the most important phase in prediction models as the data consists of ambiguities, errors, redundancy which needs to be cleaned beforehand.

The following represents the columns which are being encoded.



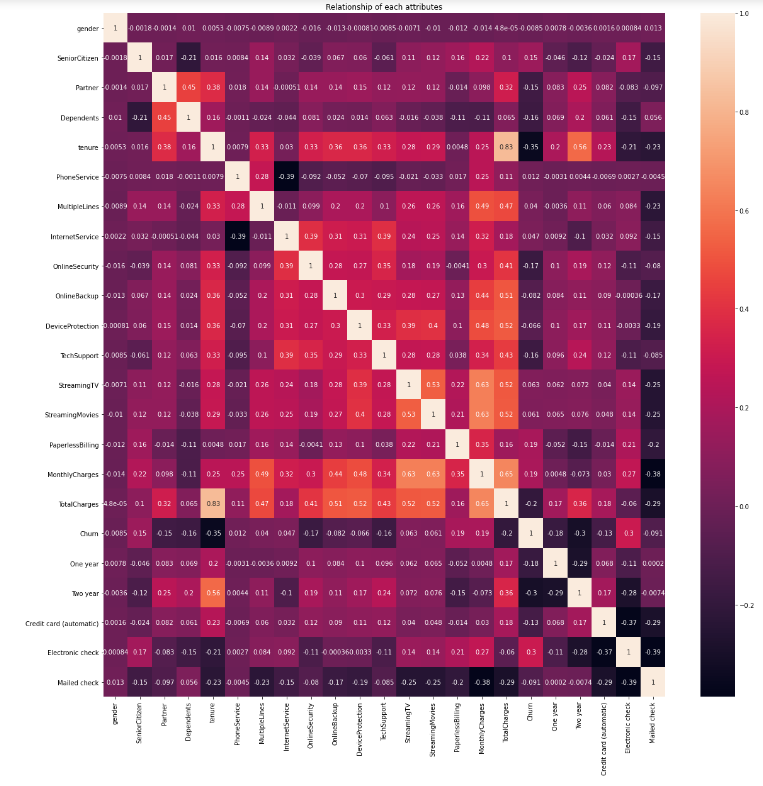
Let us dig deep into attributes and check if Total Charges column is having any hidden missing values or anomalies other than numbers.

****

We find there are empty space values other than numbers, hence it is showing as a categorical column, we are dropping the empty space values present in the attributes present in column.

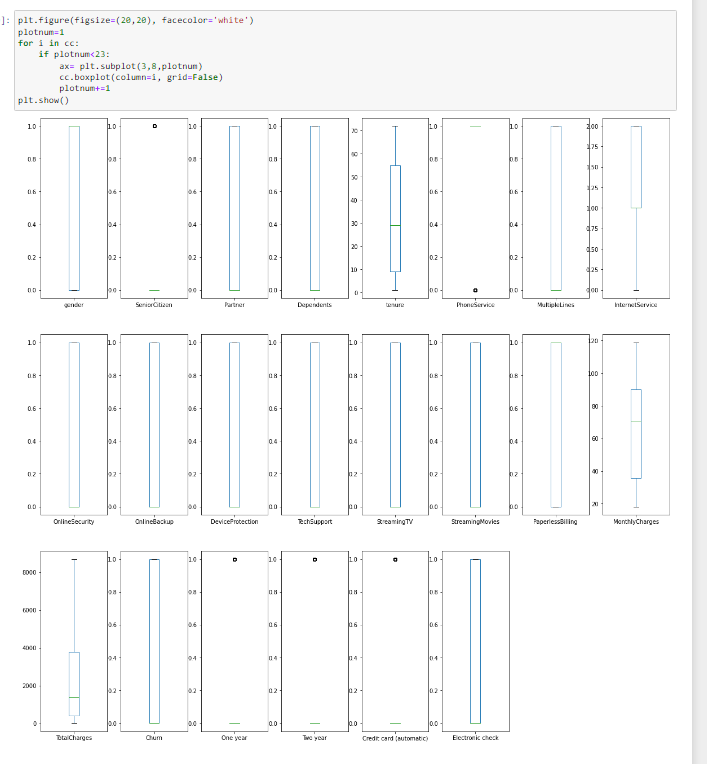
We are good to proceed with further part of data analysis, Let us check for the correlation of the attributes.

Bright color indicates correlation values as 1 which indicates that there is stronger correlation, but dark color indicates there is negative correlation between the attributes.

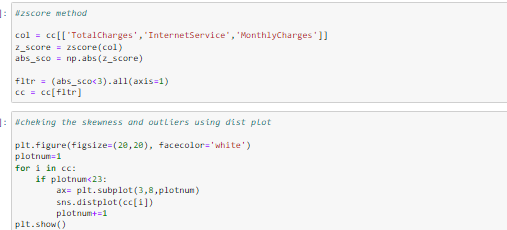
****

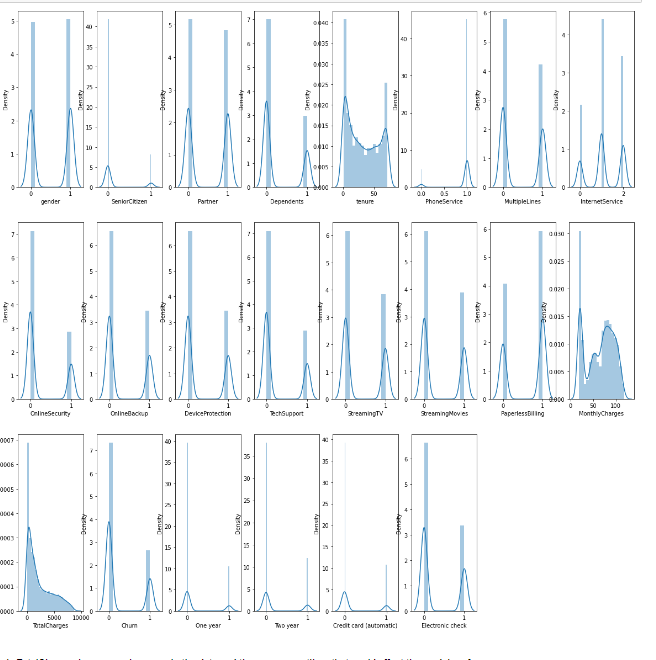
Let us check for Outliers:

There are two graphical techniques for identifying outliers, scatter plots and box plots. The box plot uses the median and the lower and upper quartiles (defined as the 25th and 75th percentiles). If the lower quartile is Q1 and the upper quartile is Q3, then the difference (Q3 - Q1) is called the inter quartile range or IQ.



We have used box plots to examine the outliers. The above graph shows the outliers present in the dataset. There are few negative skewed attributes and very few outliers are present in the dataset.

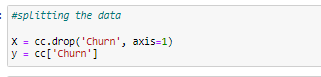




* Only Total Charges has some skewness in the data, and there are no outliers that could affect the model performance.
* Let us analyze the correlation of the attributes and check how are each column contributing to model performance.

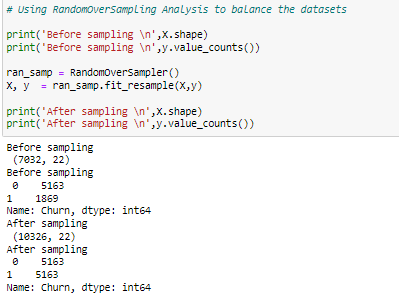
Feature Selection:

Let us split the Features and Label separately and analyze which are the important features that are impacting more, and which are having less importance using VIF technique.



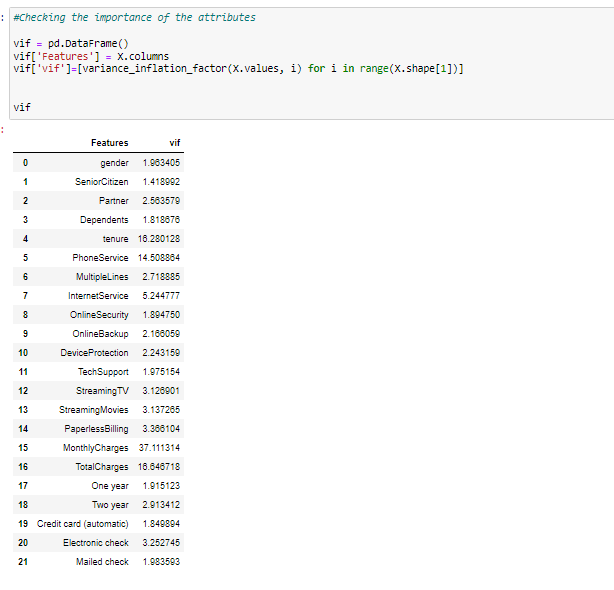
We see that target variable is unbalanced; unbalanced data refers to classification problems where we have unequal instances for different classes. Most machine learning classification algorithms are sensitive to unbalance in the predictor classes.

We are balancing the data sets using oversampling technique before proceeding with model building.



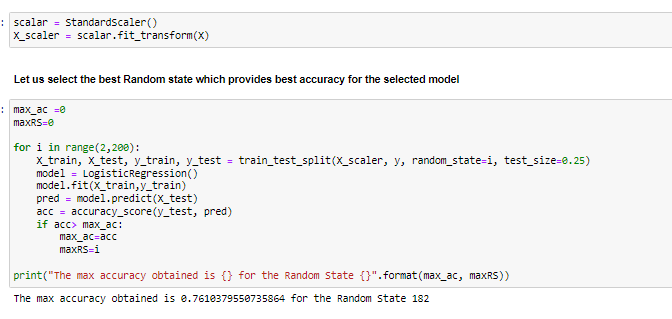
VIF score less than 5 indicates that there is no multi co linearity in between the attributes, VIF score higher than 5.

('tenure', 'Phone service', 'Monthly Charges', 'Total Charges') will need to be treated by removing single column from the data.



**Building Machine Learning Models:**

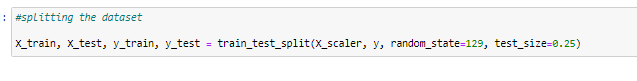
Let us standardize our data, before splitting the datasets.

****

We find the best random state to split the data into training set and test set, so that models give the best score.

To detect a machine learning model behavior, we need to use observations that aren’t used in the training process. Otherwise, the evaluation of the model would be biased.

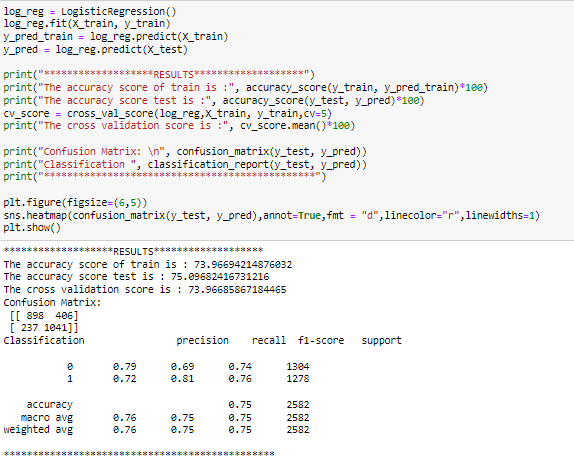
The simplest method is to divide the whole dataset into two sets. Then use one for training and the other for model evaluation. This is called the holdout method.



We are using Classification machine learning algorithms like , Logistic Regression, KNeighbors Classifier, Random Forest Classifier, Boosting Algorithms, Support Vector Machine Classifier, Gaussian Naïve Bays algorithms for predictive analysis.

1. Logistic Regression

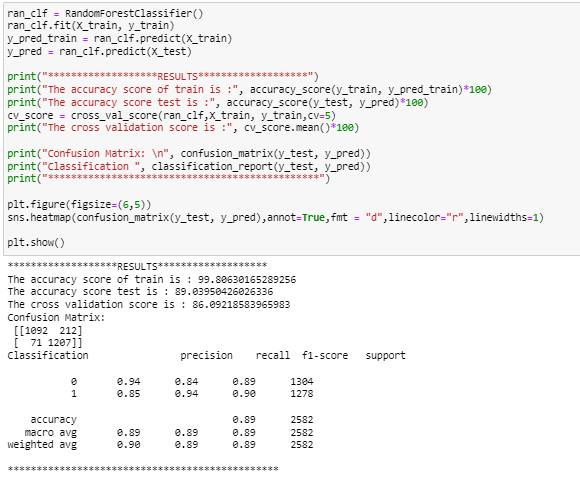
A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.



The accuracy score: 75.29 and CV score: 73.92, we can try for some more better performing models.

2. Random Forest Classifier

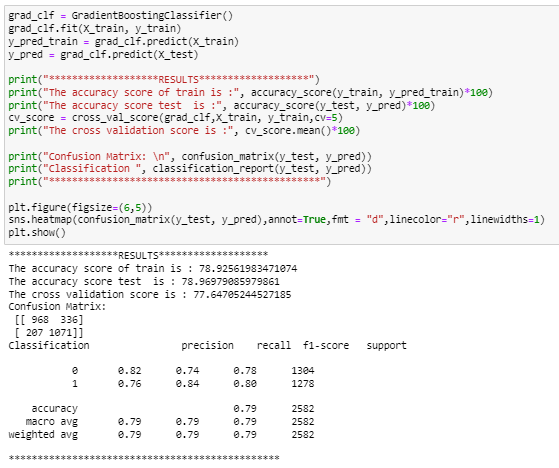
Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.



The accuracy score: 89.07 and CV score: 86.37, the accuracy and CV are too good, but high difference in values.

1. Gradient Boosting Algorithm

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model.

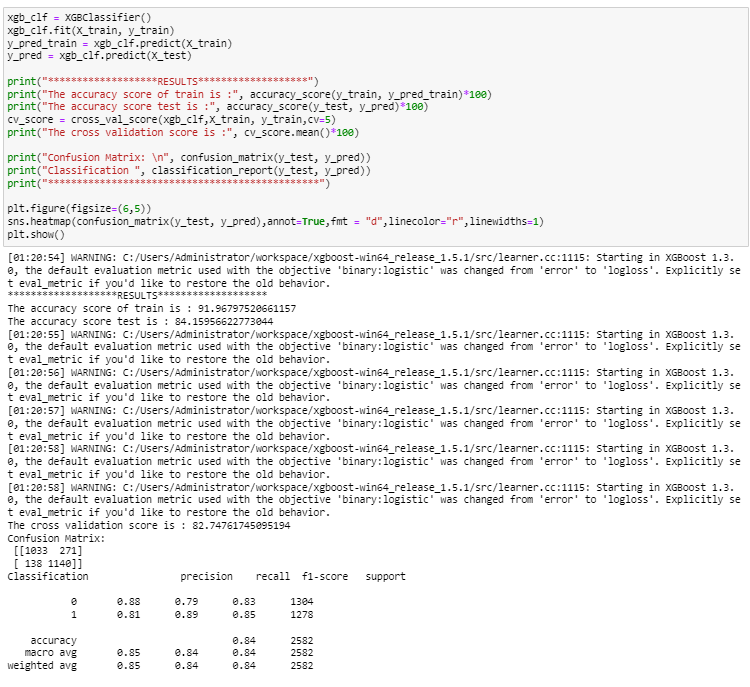


The accuracy score: 78.96 and CV score: 77.64, the accuracy and CV are descent and values have less difference.

1. XG Boosting Classifier

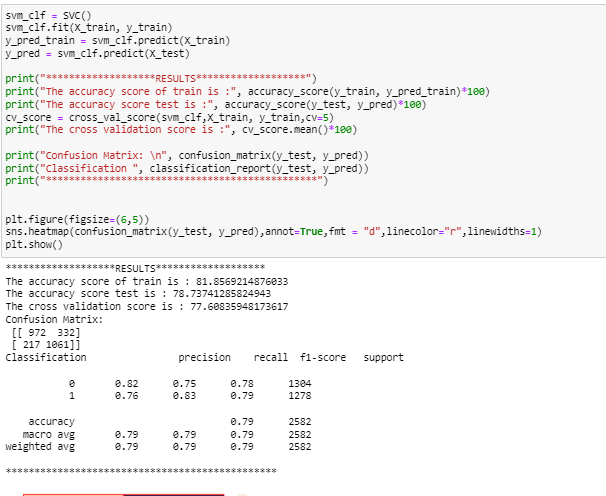
XG Boost is a popular and efficient open-source implementation of the gradient boosted trees algorithm. XG Boost is a more regularized form of Gradient Boosting. XG Boost uses advanced regularization (L1 & L2), which improves model generalization capabilities. XG Boost delivers high performance as compared to Gradient Boosting.

The accuracy score: 84.15 and CV score: 82.74, the accuracy and CV are good and values have much difference.



1. SVM Classifier

SVM is a supervised machine learning algorithm which can be used for classification or regression problems.



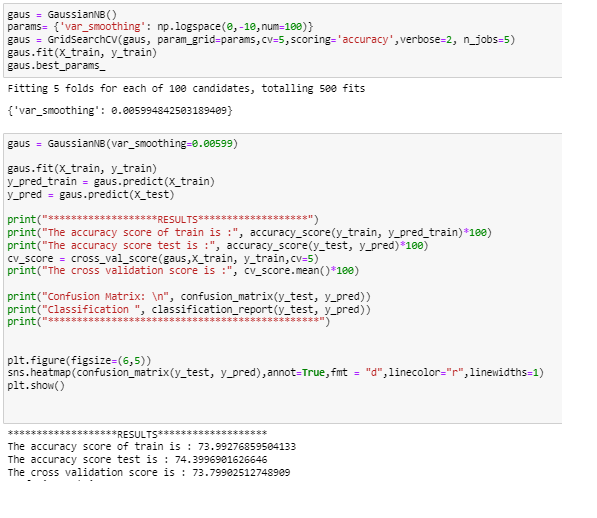
The accuracy score: 78.73 and CV score: 77.60, the accuracy and CV are descent and values have lesser difference.

From the above models we see that Gaussian NB and SVM algorithms have performed well, that is CV Score and Accuracy score values have least difference. Let us tune the parameters of these models to obtain higher accuracy.

1. SVC



2 Gaussian NB



Graphical user interface, application

Description automatically generated

After the Hyper parameter tuning, we got the accuracy nearly to be equal to that of previous one, From the graph we can see that SVC and XGB Classifier are giving the same area under the curve. Hence Let us save SVC as best model

Saving the Best Model:

Graphical user interface, text, application

Description automatically generated

**Conclusion:**

Telecommunication industry has suffered from high churn rates and enormous churning loss. Good methods need to be developed and existing methods must be enhanced to prevent the telecommunication industry to face challenges. In this article we discussed the various prediction models and compared the quality measures of prediction models. We found that the accuracy achieved with SVM Classifier is far much higher than the logistic regression technique which clearly states that decision tree is an efficient technique.