**PREDICTING CUSTOMER CHURN IN THE TELECOM SECTOR**

****

**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 prioritise 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.

To examine customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.

**What is customer churn?**

Churn is a metric that shows**customers who stop doing business** with a company or a particular service, also known as customer attrition. By following this metric, what most businesses could do was try to understand the reason behind churn numbers and tackle those factors, with **reactive action plans.**

But what if you could know in advance that a specific customer **is likely to leave your business**, and have a chance to **take proper actions** in time to prevent it from happening?

The reasons that lead customers to the **cancellation decision** can be numerous, coming from poor service quality, delay on customer support, prices, and new competitors entering the market, and so on. Usually, there is no single reason, but a **combination of events** that somehow culminated in customer displeasure.

If your company were not capable to **identify these signals** and take actions prior to the cancel button click, there is no turning back, your customer is already gone. But you still have something valuable: **the data**. Your customer left very good clues about where you left to be desired. It can be a valuable source for **meaningful insights** and to **train customer churn models**. Learn from the past, and **have strategic information at hand** to **improve future experiences**, it’s all about **machine learning.**

When it comes to the telecommunications segment, there is great room for opportunities. The **wealth** and the **amount of customer data** that carriers collect can contribute a lot to shift from a **reactive** to a **proactive position**. The emergence of sophisticated artificial intelligence and data analytics techniques further help **leverage this rich data** to address churn in a much more **effective manner.**

In this article, I’m going to use a customer base [dataset](https://raw.githubusercontent.com/mmcuri/ds_handson/master/data/telecom/WA_Fn-UseC_-Telco-Customer-Churn.csv) from given by data trained, made available by the platform [IBM Developer](https://developer.ibm.com/technologies/data-science/patterns/predict-customer-churn-using-watson-studio-and-jupyter-notebooks/). This dataset contains a total of 7,043 customers and 21 attributes, coming from personal characteristics, services signatures, and contract details. Out of the entries, 5,174 are active customers and 1,869 are churned, which demonstrates that the dataset is highly unbalanced. The target variable for this assessment is going to be the feature Churn.

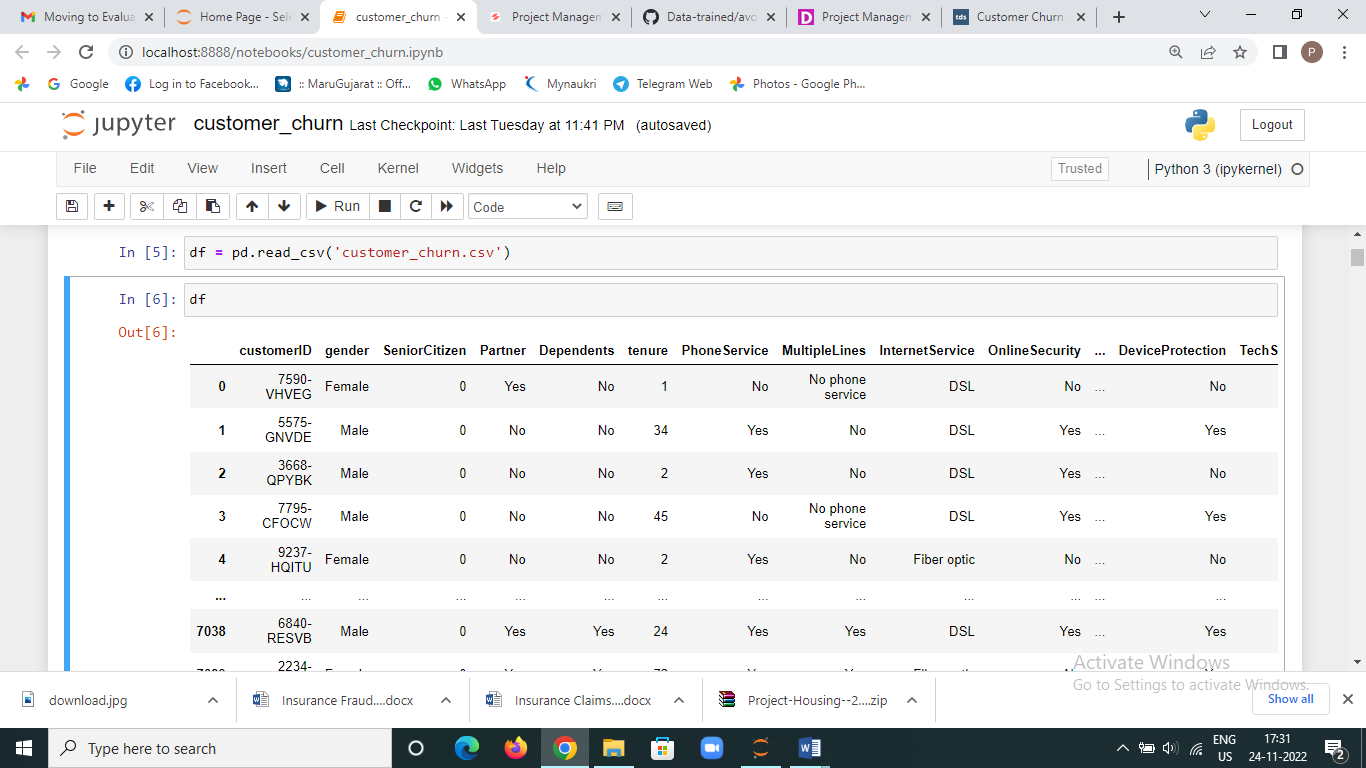
The main goal is to develop a machine learning model capable to predict customer churn based on the customer’s data available. I will use mainly Python, [Pandas](https://pandas.pydata.org/), and [Sk-Learn](https://scikit-learn.org/stable/" \t "_blank) libraries for this implementation. The [complete code](https://github.com/mmcuri/ds_handson/blob/master/Telecom_Churn_Prediction.ipynb) you can find on my [GitHub](https://github.com/mmcuri/ds_handson" \t "_blank).

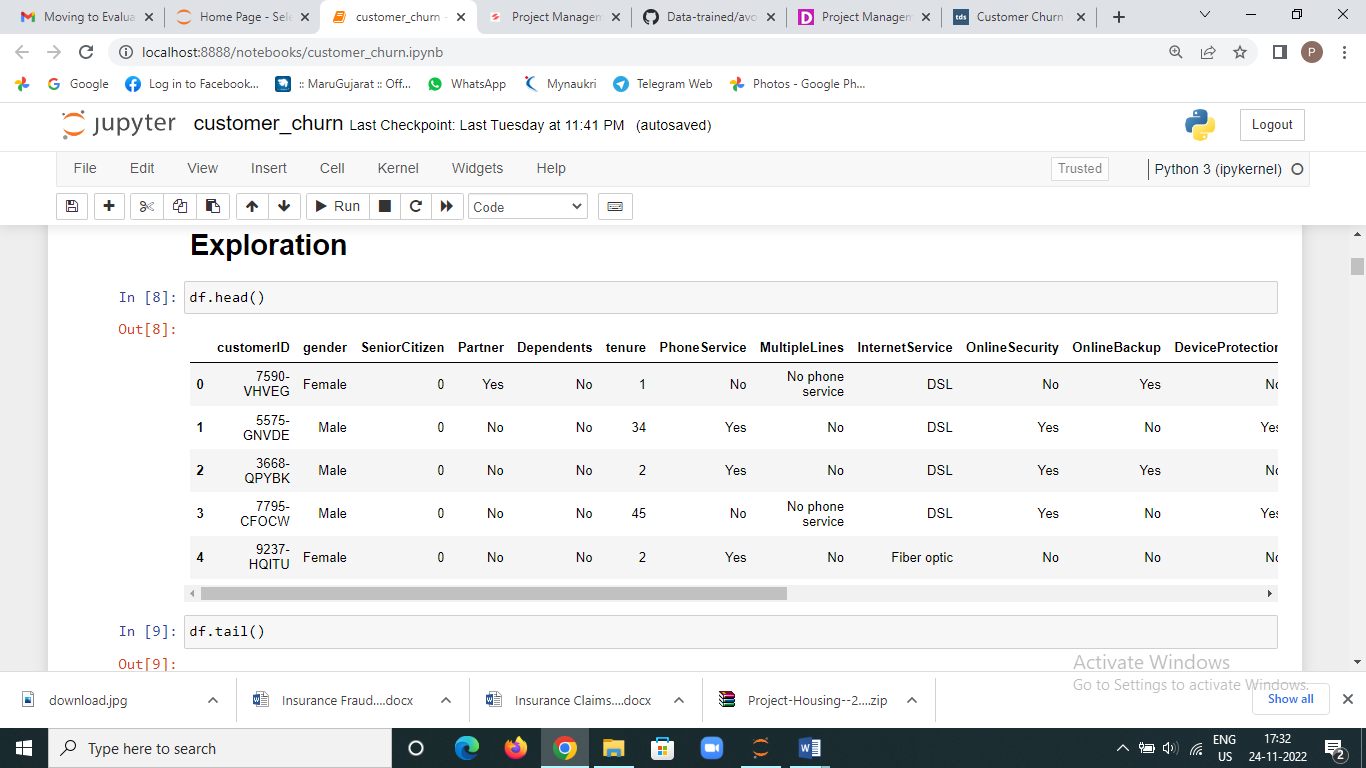
To accomplish that, I will go through the below steps:

* Exploratory analysis (Exploration)
* Data analysis and visualization
* Data pre-processing
* Checking outliers and skewness and VIF calculation for multicollinearity
* Train, tune and model building
* Hyper parameter tuning
* AUC ROC Curve
* Model saving and prediction with original data

**Data Exploration:**

**Reading the csv file by pandas method:**

****

****

Checking the head and tail code to see the data.

As the purpose of this experiment is to **identify patterns** that can yield to customers churn, I will be focusing mainly on the churn portion of the dataset for the exploratory analysis.

Data Description:

Data is having 7043 entries, 0 to 7042

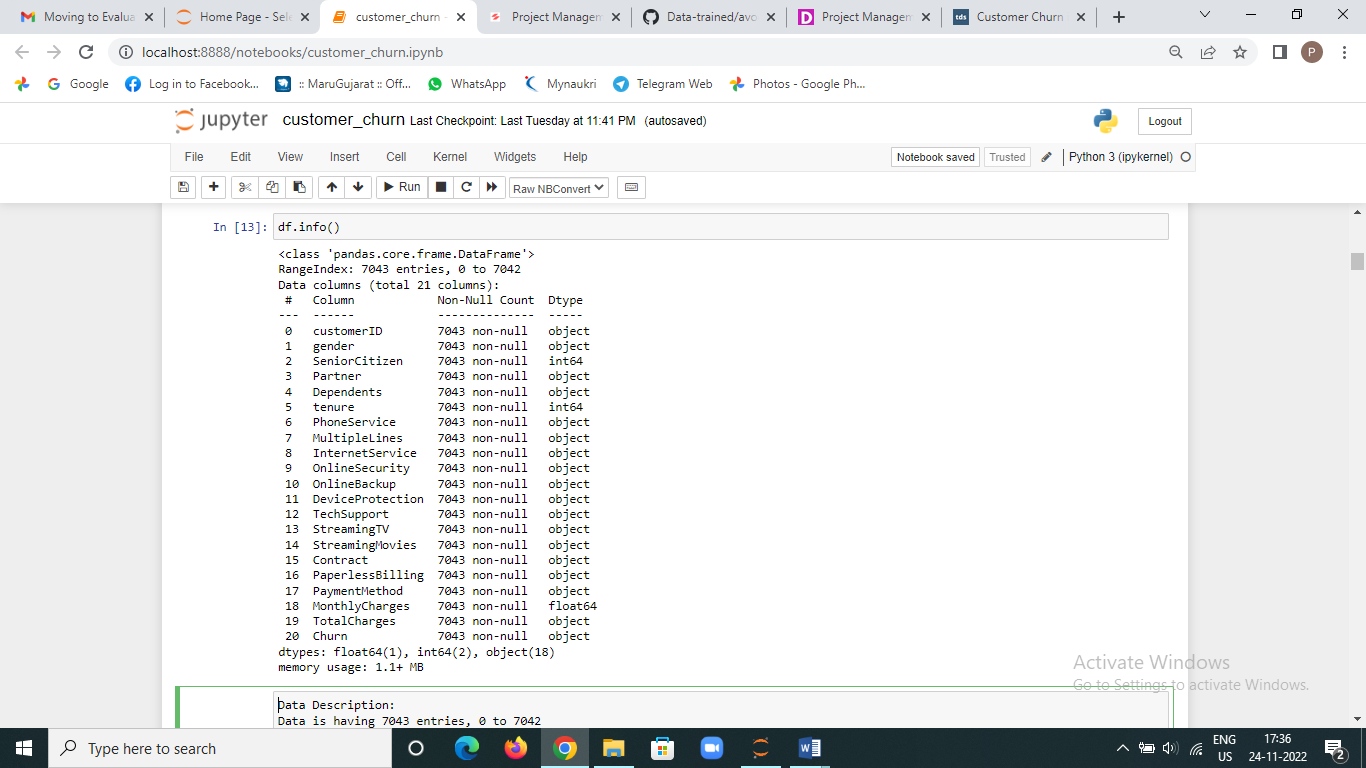
Data having 21 columns and 7043 Rows

All the columns are having 7043 Non-Null value, means data is not having any null values.

Out of 21 columns we have 1 float, 2 int64 and 18 Object type values.

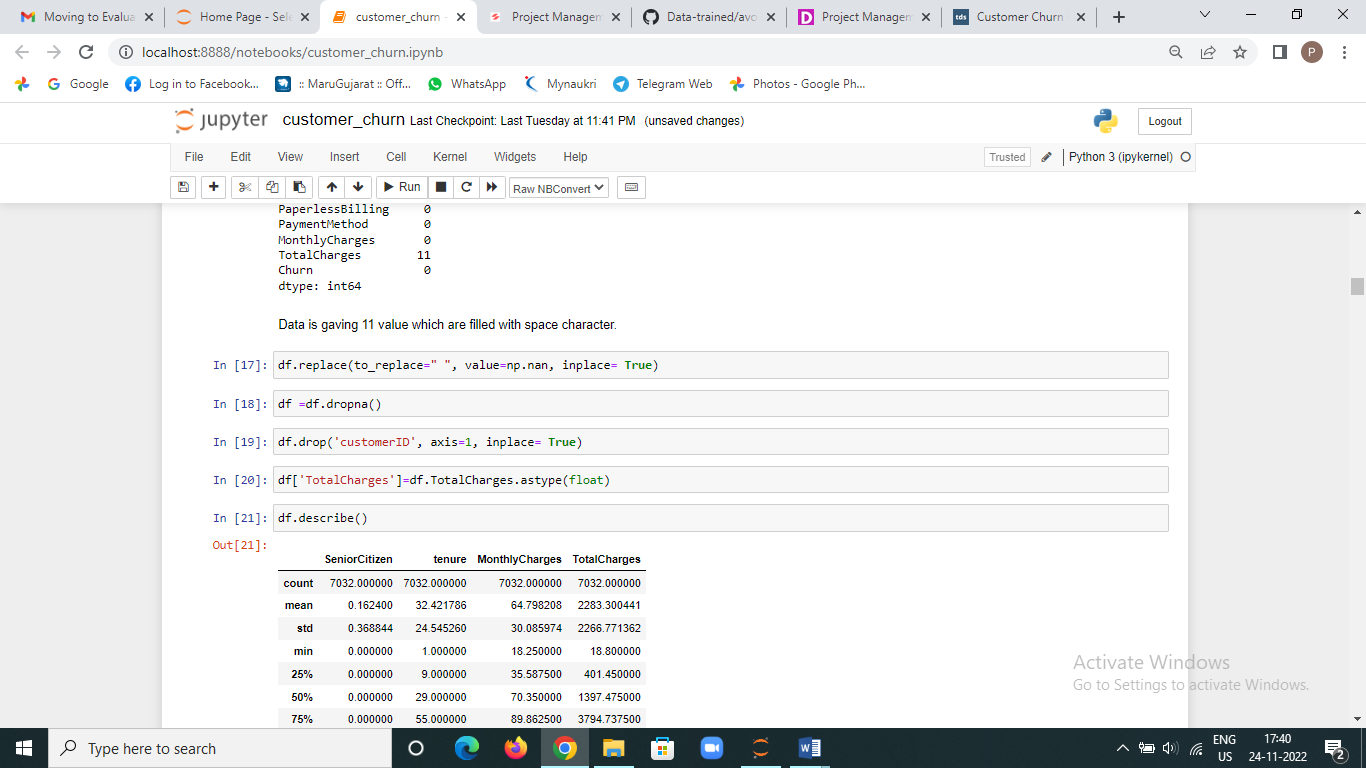
Here we can see Total charges is object type, but the values are float. It might be possible we have space in instead of value we will check where data having space or not.

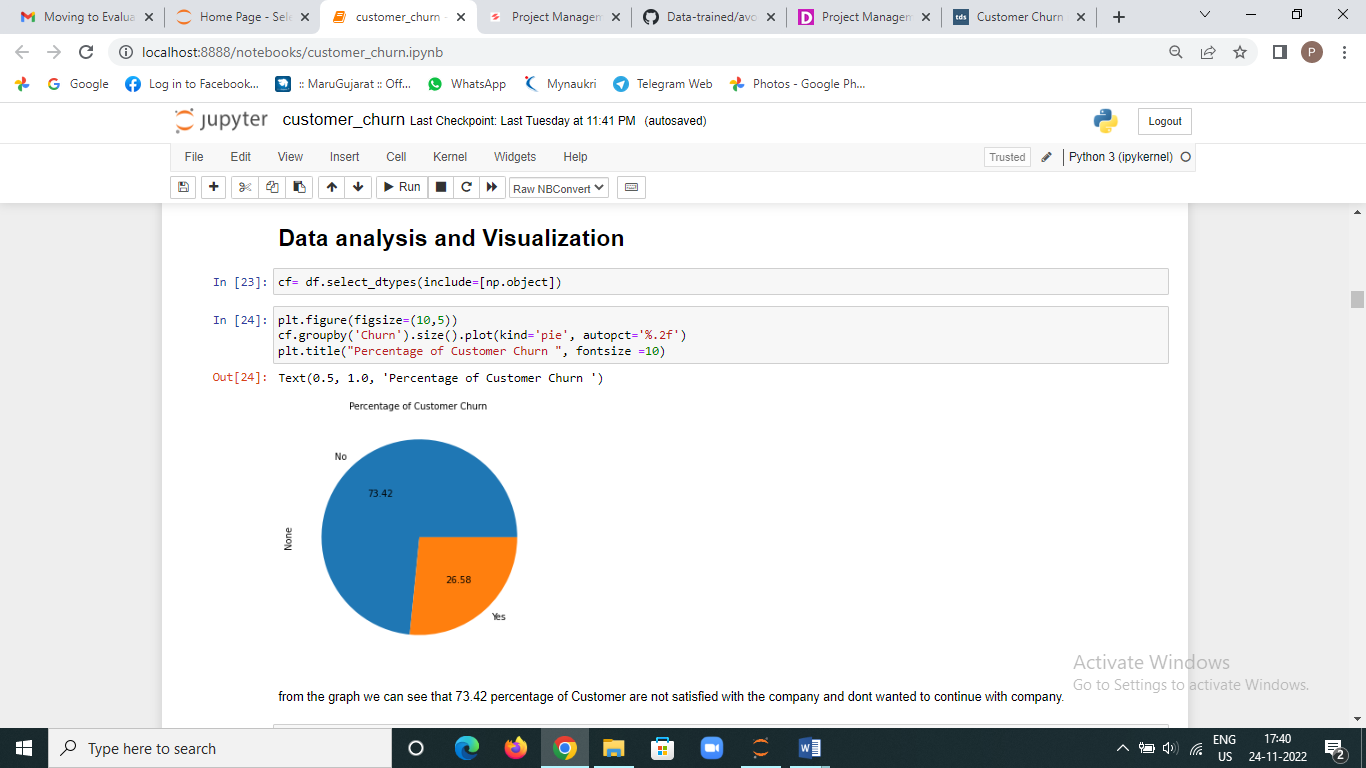
Please see below image for above details:



After that check for missing values:

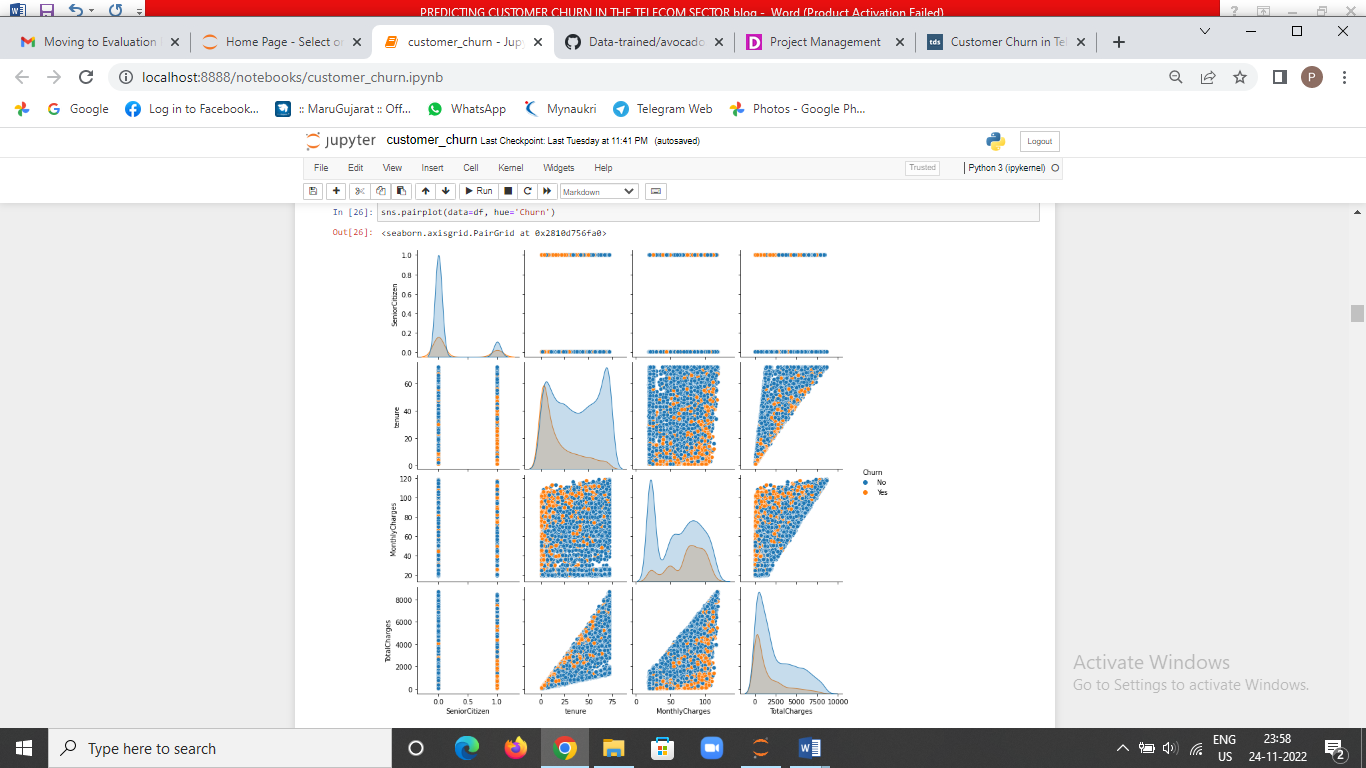
Dataset have 11 missing values in Total charges column which is filled with Nan values and changing the type of total charges from object to float type.



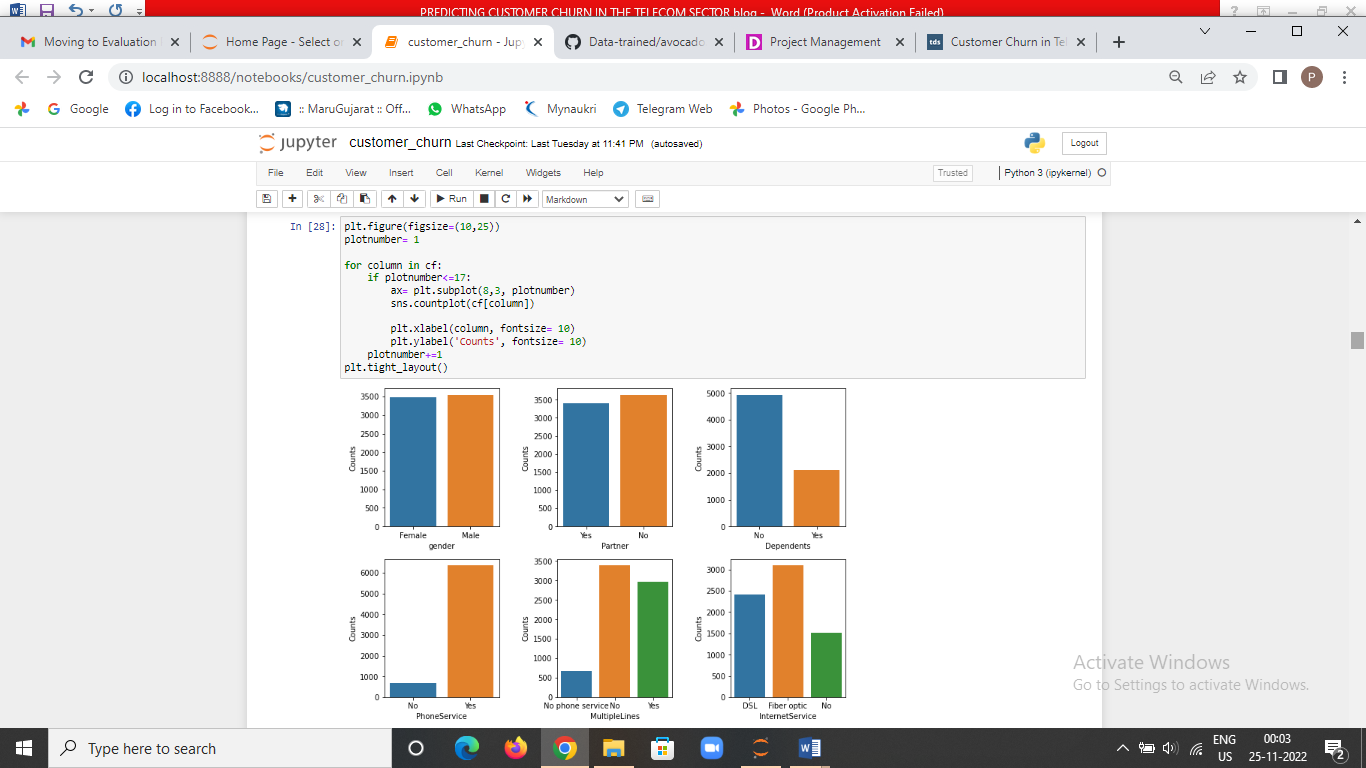


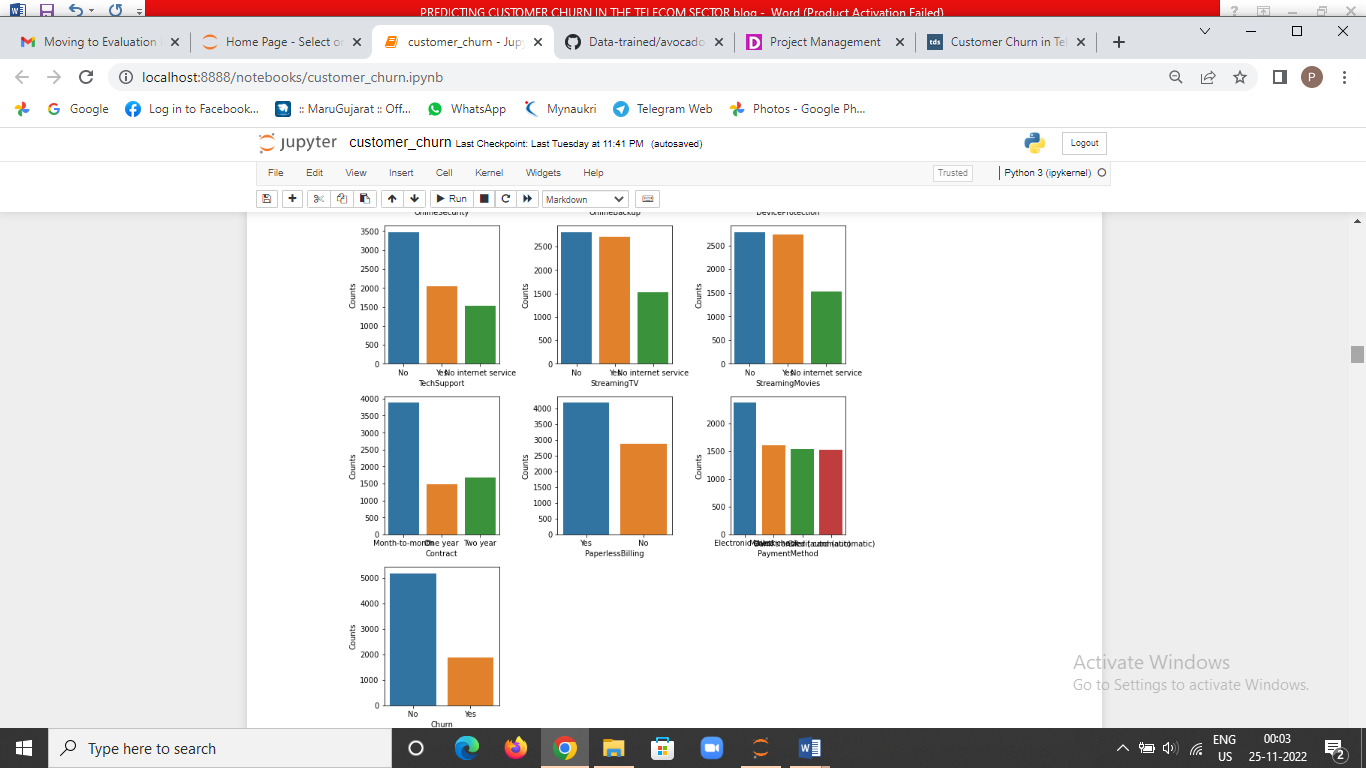
From the graph we can see that 73.42 percentage of Customer are not satisfied with the company and don’t wanted to continue with company.

Then checked pair plot using sea born library with target variable- churn vs. other variables, check below image:



We can see with increase in monthly charge and at low values of total charges churn is more likely to happen. The tenure and total charge increase, chance of churn is less. We can predict Senior Citizen are more likely to churn.





**EDA Concluding Remarks:**

Gender: No of males and females are almost equal

Partner: Count for having partner is less as compare to without partner

Dependents: Near about 70 % customers are not having any Dependent

Phone Service: More than 90 % customers have phone Service

Multiple Lines: More than 48 % customer not having multiple lines

Internet Service: For More than 44 % customers’ internet service provider is Fibre Optics

Online Security: 49 % customers are not having Online Security

Online Backup: More than 43 customers not having online backup

Device Protection: More than 43 customers are not having Device Protection

TechSupport: 49 % customers are not having any technical support

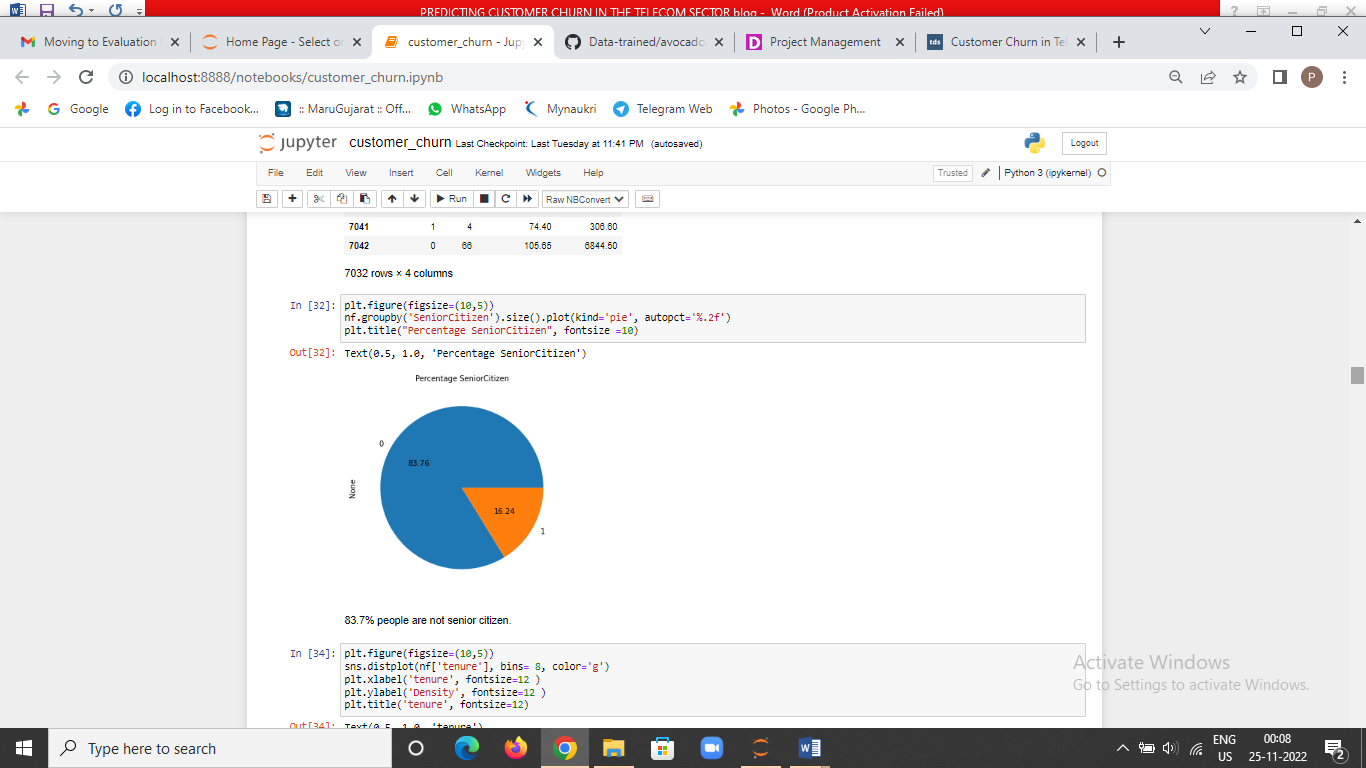
Streaming TV: The count for Streaming TV and Not Streaming TV are almost same

Streaming Movies: The count for Streaming Movies and Not Streaming Movies are almost same

Contract: More than 55 % of the customers are having month-to-month contract

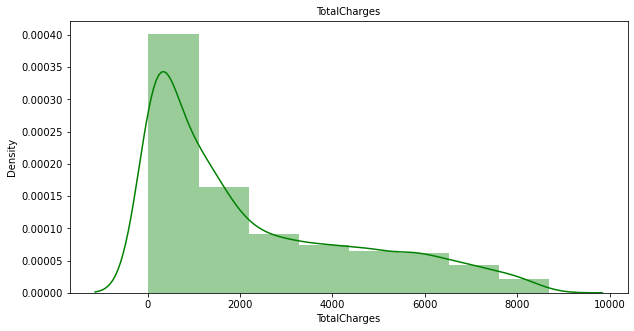
Paperless Billing: we can see more than 59 % of the customers are using Paperless Billing, Most of the customers are using Electronic check

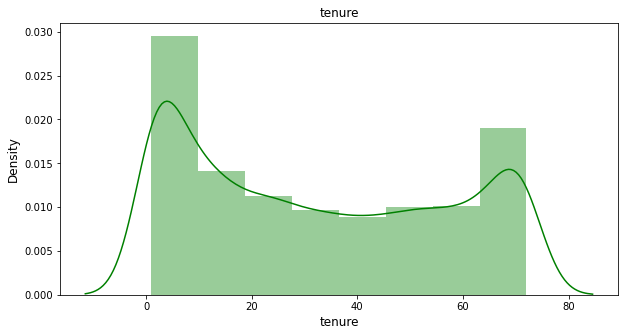
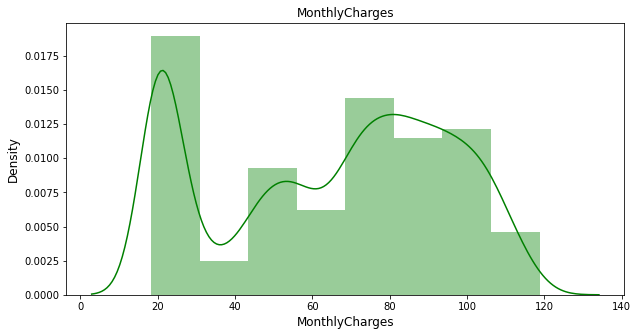
Churn: 26 % customer out 100 are churned.



From above image the graph shows 83.7% people are not senior citizen.

After that I have check tenure, monthly charges and total charges distplot and also check whether the data skewed or not in below images.

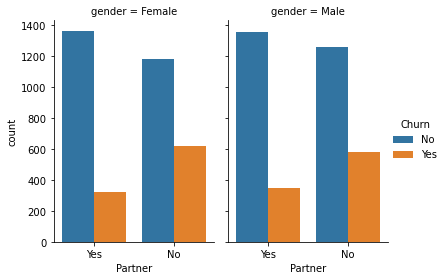




Total charges: we can see that customers in the range 0 to 2000 are more and data are skewed.

Monthly charges: a chance of customers having monthly charge in the range of 70 to 105, is more as compare to other charges.

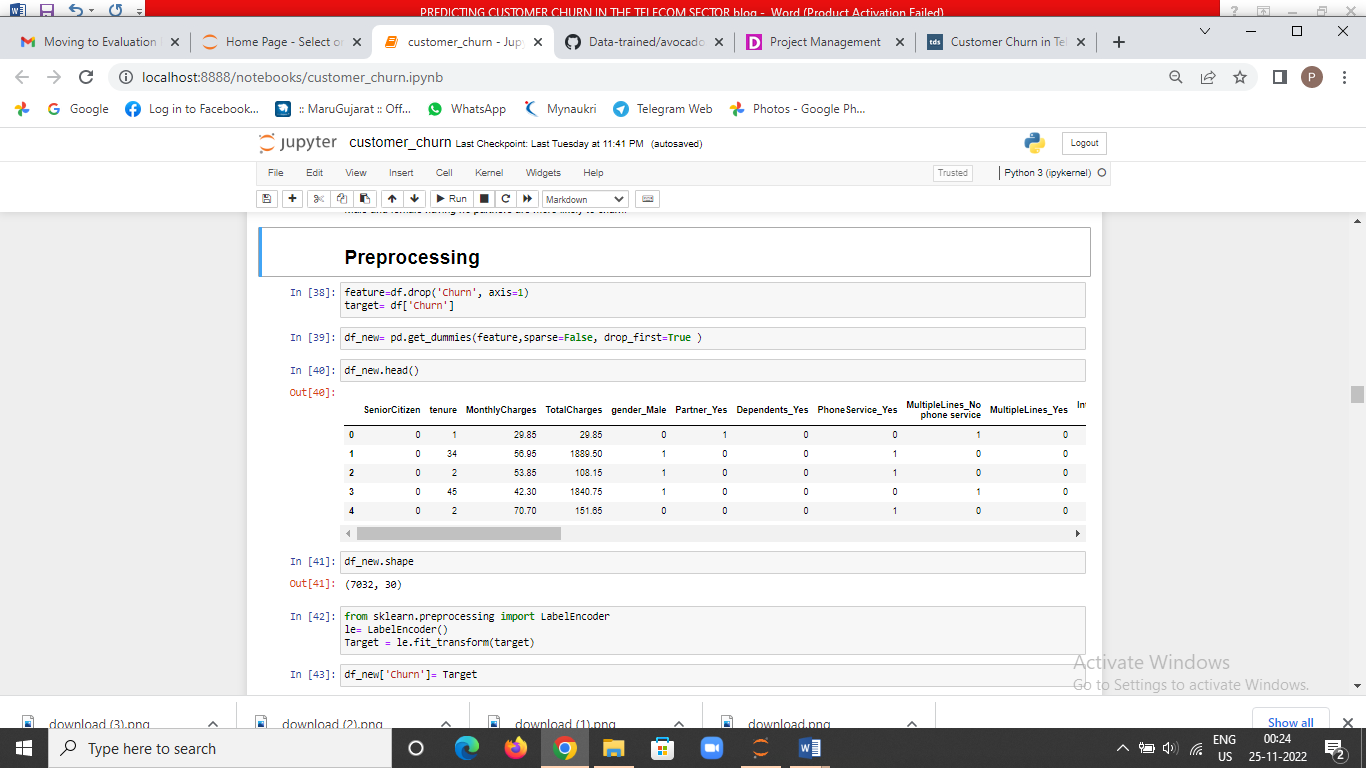
Tenure: We can see that the probability to have tenure 0-10 and more than 65-70 is more.



Male and female having no partners are more likely to churn.

**Pre-processing Pipeline:**

Separate data in feature and target.



**Feature Encoding**

Machine learning algorithms can only read **numerical values**. Hence, one essential part of data preparation is to apply the encoding of categorical features into numerical values, this process is called feature encoding.

For categorical features with only two classes (binary), such as gender, SeniorCitizen, Partner, Dependents, PhoneService and PaperlessBilling I used [Label Encoder](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html) from sk-Learn, a utility class to help normalize labels such that they contain only values between 0 and n\_classes-1.

When it comes to categorical features with more than two classes, such as MultipleLines , InternetService , OnlineSecurity , OnlineBackup , DeviceProtection , TechSupport, StreamingTV , StreamingMovies , Contract and PaymentMethod I used method [get dummies](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html) from Pandas Data frame, which converts categorical variables into dummy/indicator variables.

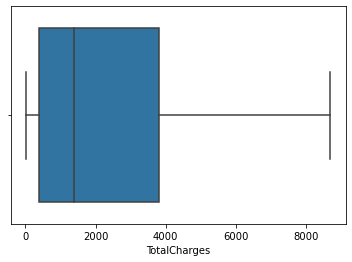
**Correlation matrix:**

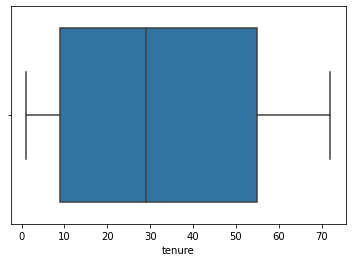
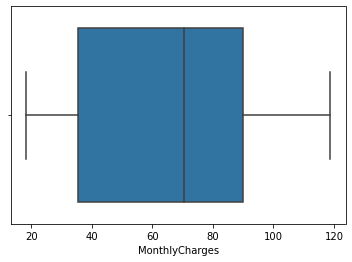
A correlation matrix is a table containing correlation coefficients between variables. Each cell in the table represents the correlation between two variables. The value lies between -1 and 1. A correlation matrix is used to summarize data, as a diagnostic for advanced analyses and as an input into a more advanced analysis. The two key components of the correlation are:

* **Magnitude:** larger the magnitude, stronger the correlation.
* **Sign:** if positive, there is a regular correlation. If negative, there is an inverse correlation.

Then I checked correlation matrix using heat map.

1. 'PhoneService\_Yes','SeniorCitizen','MultipleLines\_Yes','StreamingMovies\_Yes','StreamingTV\_Yes', 'PaperlessBilling\_Yes', 'Monthly Charges', 'PaymentMethod\_Electronic check','InternetService\_Fiber optic' are negatively correlates with churn.

**Checking outliers and skewness:** 

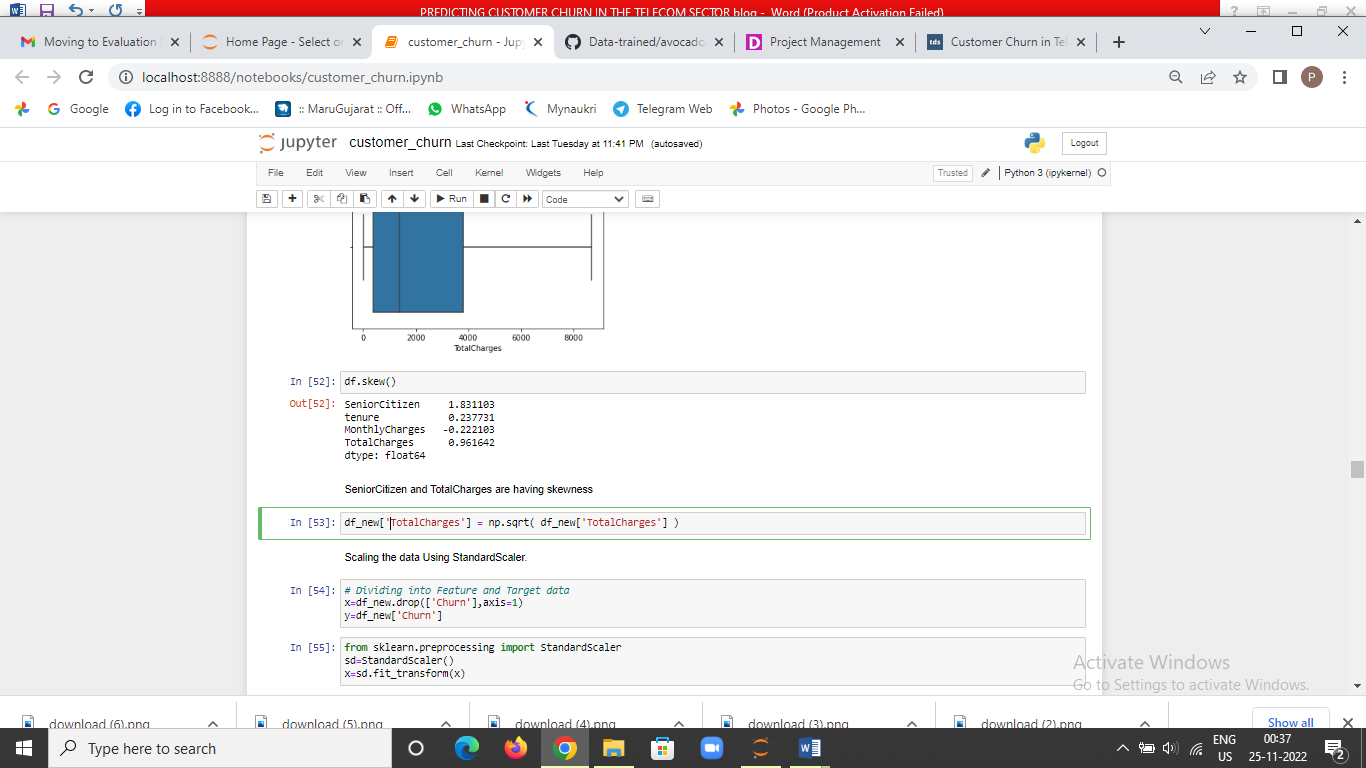
****

**No outliers present in these columns.**

Senior Citizen and Total Charges are having skewness and by using square root method skewness has removed.

Also scaled the data using standard scaler **to handle highly varying magnitudes or values or units**. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

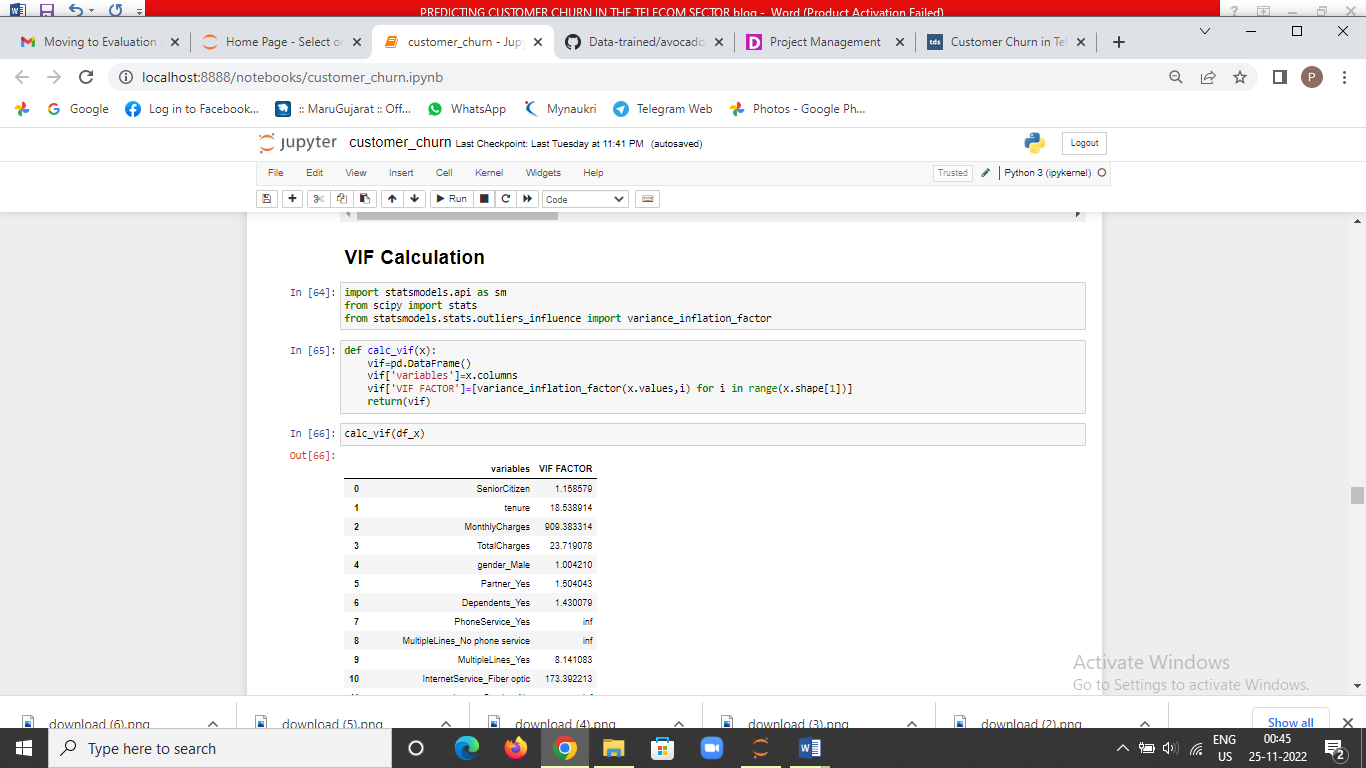
See below image:

****

**Imbalanced Data**

The target variable which will be used for ML model training will be Churn. We need to make sure that both Churn classes ('Yes' and 'No') have equal distribution otherwise, it can introduce bias into the model.

**I have used SMOTE from imb-learn to balance the data.**

**Check multicollinearity and remove by VIF calculation: **

**Split train and test data**

The subscribers' dataset was split into 80% for training and 20% for testing. The training set will be used to generate the model the chosen algorithms will use when **exposed to new data.** The test set is the final dataset I will touch to measure**model performance** based on some metrics.

**Train, tune and evaluate machine learning models**

**Metrics**

For the churn problem, the ideal metric to be used is [Recall](https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall). Recall answers the following question: what proportion of **actual positives** was identified correctly? In that case, Recall measures the percentage of **churns** that were correctly classified out of the total churns, which is what we will be looking for to analyse the performance of ML classifiers.

As an example, consider a re-engagement campaign which gives 1 GB of data usage for free. You’d likely want to ensure that the precision for this bonus gift is high. In other words, you would want to minimize the number of happy users who will receive the bonus, and instead have this bonus hitting almost exclusively users in danger of churning.

**Building Machine Learning Models:**

In this experiment, I applied four different ML algorithms to analyse and compare the Recall score obtained by each of them. Those are listed below:

* 1. Logistic regression
  2. Support vector machine classifier
  3. KNeighbors classifier
  4. Decision tree classifier
  5. Random forest classifier
  6. Adaboost classifier
  7. GaussianNB

To compare their performances, as a first step, I applied [cross-validation](https://scikit-learn.org/stable/modules/cross_validation.html) method which is a technique that partitions the data into subsets, training the data on a subset and use the other subset to evaluate the model’s performance. The top-performers were Random Forest Classifier accuracy score: 84.22071636011617 Cross Val Score: 84.8934078835655, Adaboost classifier accuracy score: 81.26815101645693

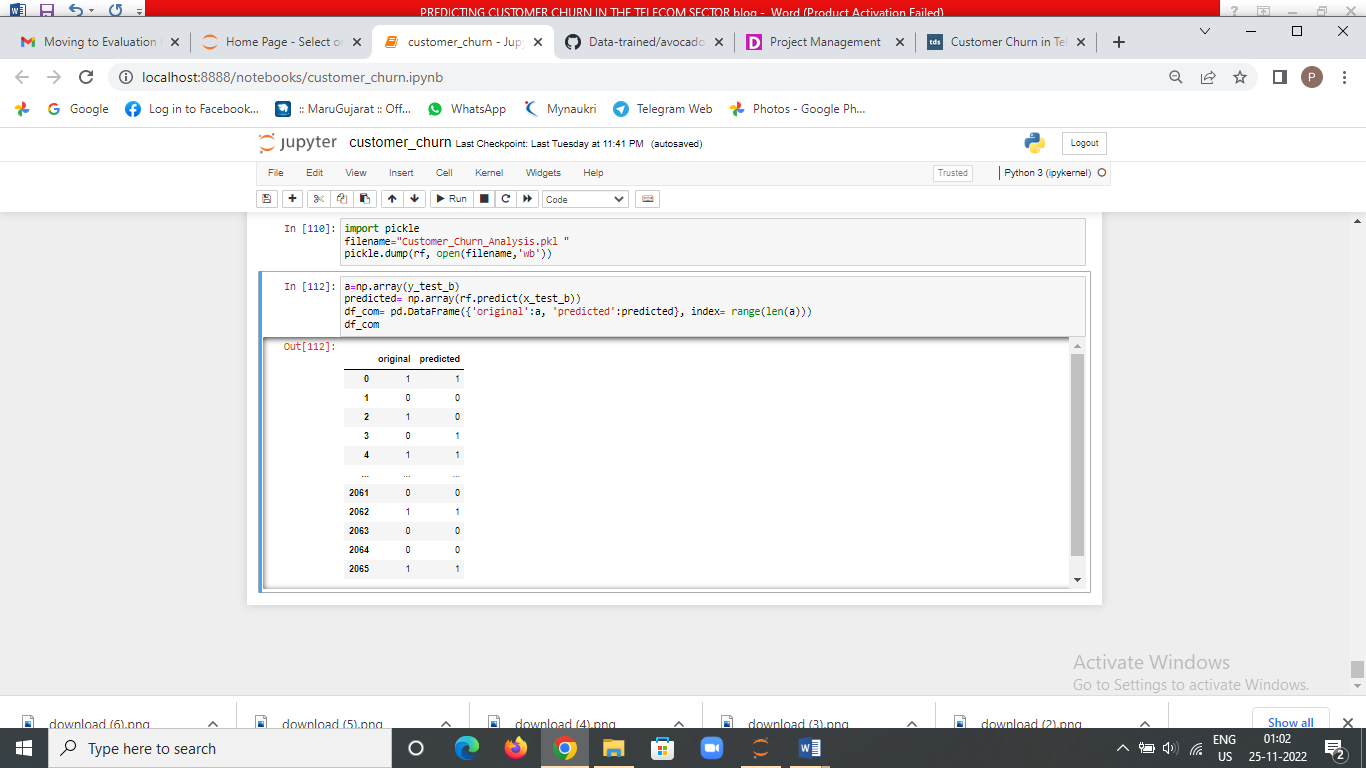
Cross Val Score: 80.54463725412815 and SVC Accuracy core: 80.63891577928364 Cross Val Score: 81.24184628718896.But there is still room for optimization.

**Tuning via Hyper parameters**

To improve the overall performance when it comes to [recall](https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall) metric, I tuned classifiers hyper parameters using [grid search](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html?highlight=grid%20search#sklearn.model_selection.GridSearchCV) CV for Random Forest Classifier, Adaboost Classifier and SVC.

The **most expressive** improvement came from Random Forest classifier model, accuracy score: 84.17231364956437 Cross Val Score: 85.14520323846641. This means the implemented ML model based on Random forest **delivers 84% of precision while predicting customer churn.** These results can be seen in the below correlation matrix, where 1 means Churn and 0 means not Churn.

I saved the model using pickle as ‘Customer\_Churn\_Analysis.pkl’.



**Concluding remarks:**

No algorithm will predict churn with 100% accuracy. There will always be a trade-off between precision and recall. That's why it's important to test and understand the **strengths and weaknesses**of each classifier and get the best out of each.

If the goal is to engage and reach out to the customers to prevent them from churning, it's acceptable to engage with those who are mistakenly tagged as ‘not churned,’ as it does not cause any **negative impact**. It could potentially make them even happier with the service. This is the kind of model that can **add value** from day one if proper action is taken out of meaningful information it produces.

**Thank you for your time!**