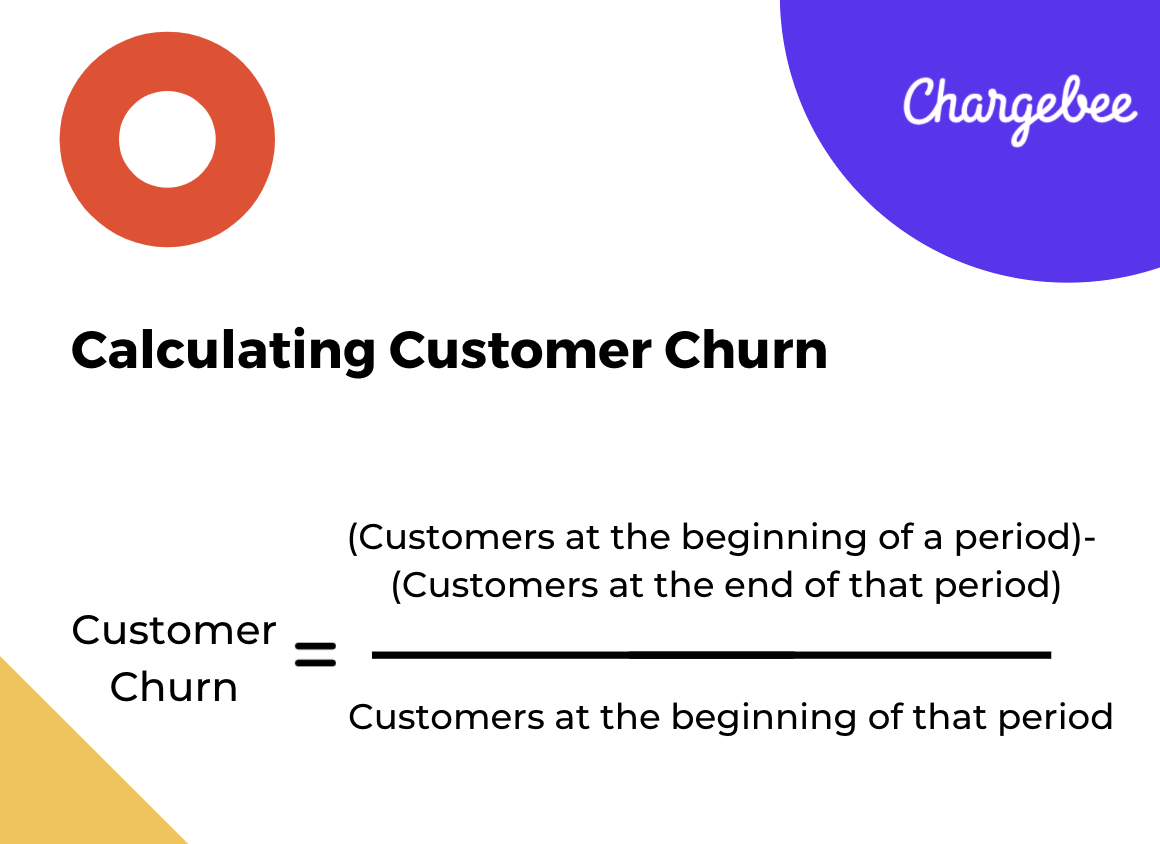
**Blog for**

**Cutomer Churn Analysis**

Submitted by – Chris Chhotai

**Introduction :-**

Churn Analysis is a major problem faced by companies these days. Churn is the nightmare for every business and is faced by almost all companies. Customer Churn is calculated as follow:

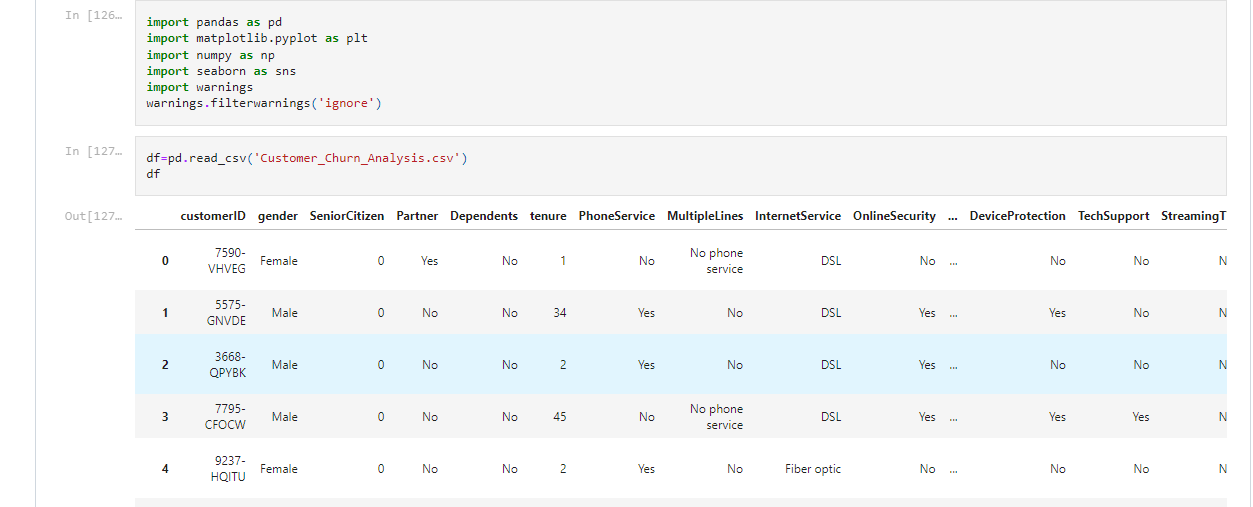
**Problem Definition:**



Customer churn occurs when a company's customers cease to conduct business with it. Churn is closely monitored by businesses since maintaining an existing client is significantly less expensive than obtaining a new one. Working leads through a sales funnel and using marketing and sales expenditures to obtain additional clients are all part of starting a new firm. Existing customers have a larger volume of service usage and are more likely to refer new customers.

Good customer service and merchandise can help you keep your customers. However, the most efficient strategy for a corporation to prevent client attrition is to fully understand them. Churn prediction models can be built using the massive amounts of data collected about customers. Knowing who is most likely to defect allows a corporation to focus its marketing efforts on that segment of its client base.

In the telecommunications industry, preventing client turnover is vital, as the obstacles to entry are high.

**Importing the Data:** ****

Here we have imported the libraries and imported the dataset which was in .csv format in the Jupyter notebook.

This data set contains Independent and Dependent(target) variables.

**Independent Variable**: These are the known as Input variables. These are the input for a process that is being analyzed.

**Dependent Variable**: These are known as Output or Target variables. These are dependent on Indepenndent variables for their outcome.

After importing we will display a sample data. The variables in this dataset are as follows:

* customerID
* gender
* SeniorCitizen
* Partner
* Dependents
* tenure
* PhoneService
* MultipleLines
* InternetService
* OnlineSecurity
* OnlineBackup
* DeviceProtection
* TechSupport
* StreamningTV
* StreamingMovies
* Contract
* PaperlessBilling
* PaymentMethod
* MonthlyCharges
* TotalCharges
* Churn

**DATA ANALYSIS (EDA):**

Now we will understand the dataset by perfoming the Exploratory Data Analysis.

Let’s check the shape of dataset:



As we can see that this dataset contains 7043 rows and 21 columns.

We can't see the null values for the datset here, which will have an impact on our data and, ultimately, the expected result. As a result, we should examine the dataset for null values.

**Checking for null values in the dataset:**

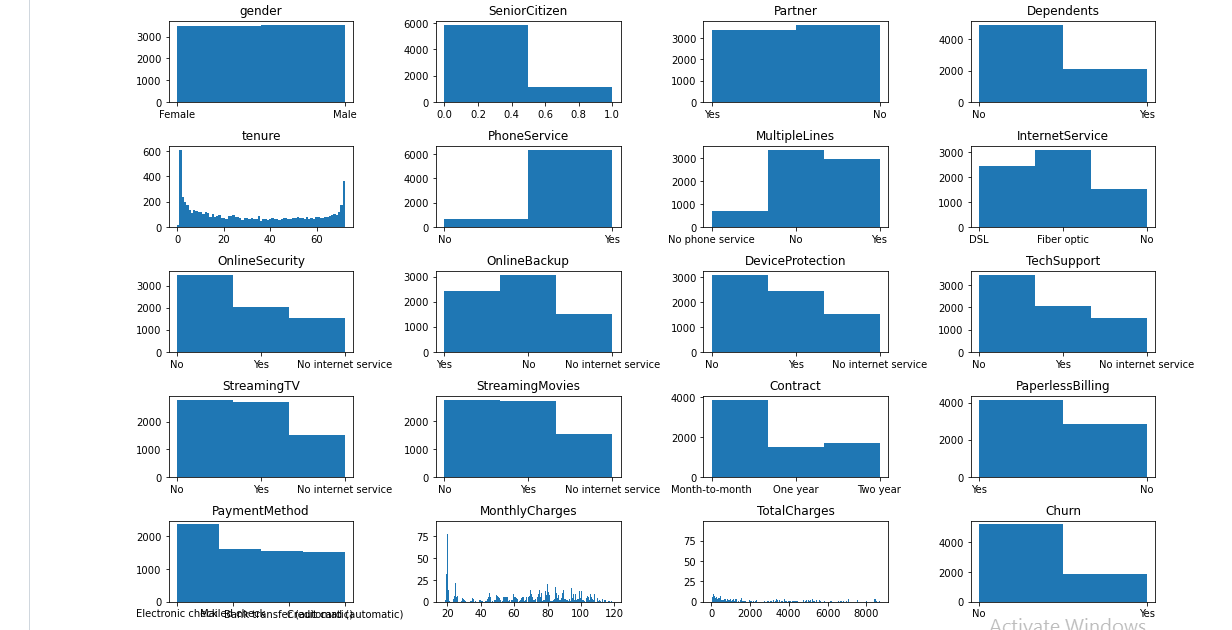


There are no null values in the dataset, as we can see. If the dataset contained null values. If there were any null values in the dataset, the mean, median, or mode would have been used to replace them.

**Data Visualization and EDA Concluding Remarks:**

The 'Churn' feature or variable is the Target feature or variable in the given data. This characteristic has only two distinct values, Y and N (Yes and No), implying that it has only two classifications. As there are only two distinct values, this is referred to as a 'Classification Problem.'

We have plotted all the feature columns againt the count which are shown below.



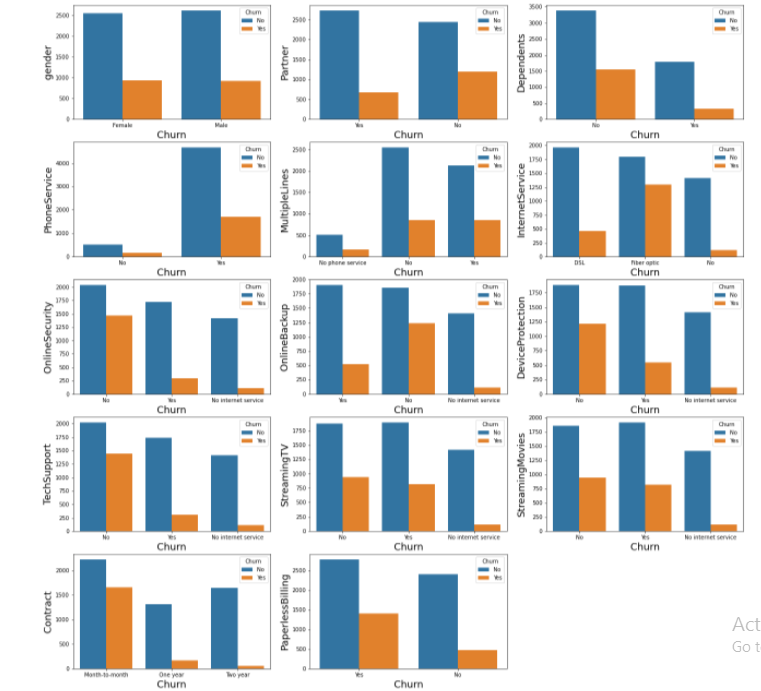
As we can see in this that majority of the count is towards ‘NO’.

The other observations which can be concluded from this are as follows:

* Majority of the customer are Male. There is no much difference in gender count as well.
* Majority of the customers not Senior Citizens.
* Majority of the customers have partner and also there is not larger difference under this column data types.
* Majority of the customers doesn’t have dependents.
* Majority of the customers are with PhoneService.
* Majority of the customers has Paperless Billing.
* Majority of the customers has no Multiple Lines followed by having Multiple Lines and least are with no phone service.
* Majority of the customers has Fiber optic Internet Service followed by DSL and least are with no Internet Service.
* Majority of the customers doesn’t have Online Security followed by customers with online security and least are with no Internet Service.
* Majority of the customers doesn’t have/ use Online Backup service followed by customers with online Backup service and least are with no Internet Service.
* Majority of the customers doesn’t have/ use Device Protection service followed by customers with Device Protection service and least are with no Internet Service.
* Majority of the customers doesn’t have/ use Technical Support service followed by customers using Tech Support service and least are with no Internet Service.
* Majority of the customers doesn’t have/ use Streaming TV service followed by customers using Streaming TV service and least are with no Internet Service.
* Majority of the customers are under Month-to-month contract followed by two year contract and least are under one year contract.
* Majority of the customers use Electronic check as PaymentMethod followed by Mailed check and Bank transfer (automatic) payment methods and least are through Credit card (automatic).

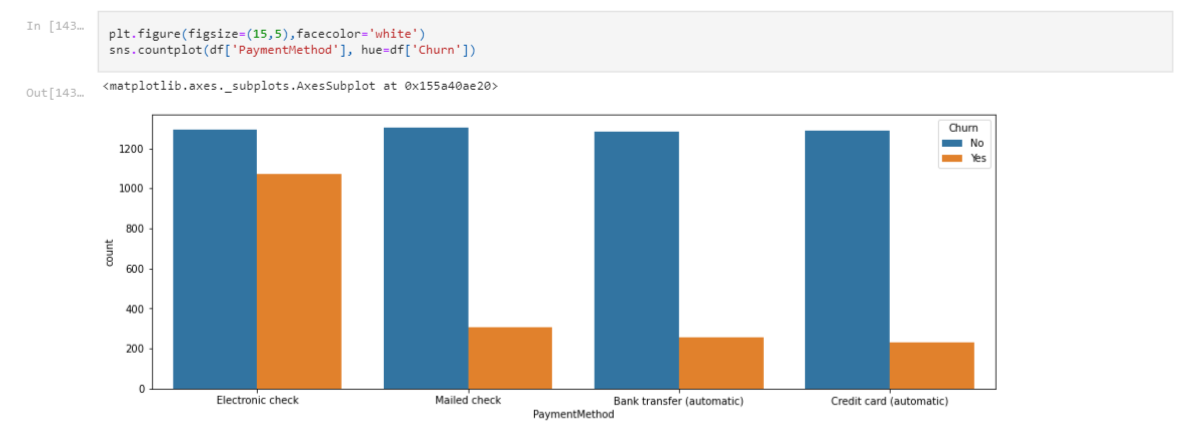
**The plot of every feature column against target variable ‘Churn’ is as follows:**

* PaperlessBilling - Customers who use paperless billing has higher Churn rate.
* gender - There is no much difference in the Churn rate with respect to Gender.
* Partner - Customers who doesn’t have partner are having higher Churn rate.
* Dependents - Customers who doesnt have any dependents are having higher Churn rate.
* PhoneService - Customers who use Phone services has higher Churn rate.
* MultipleLines - Churn rate is almost same for the customers with and without Multiple Lines.
* InternetService - Customers who use Fiber optics based internet service has higher Churn rate.
* OnlineSecurity - Customers who doesn’t have/ use online security service are having higher Churn rate.
* OnlineBackup - Customers who doesn’t have/ use online backup service are having higher Churn rate.
* DeviceProtection - Customers who doesn’t have/ use device protection service are having higher Churn rate.
* TechSupport - Customers who doesn’t have/ use Tech Support service are having higher Churn rate.
* StreamingTV - Customers who doesn’t have/ use streaming TV service are having higher Churn rate.
* StreamingMovies - Customers who doesn’t have/ use streaming movies service are having higher Churn rate.
* Contract - Customers who are under the contract of Month to month are having higher Churn rate.



* Customers who does payment through Electronic check has higher attrition rate / Churn rate and rest 3 payment modes customers have almost same Churn rate.

We can see the above point in below graph.



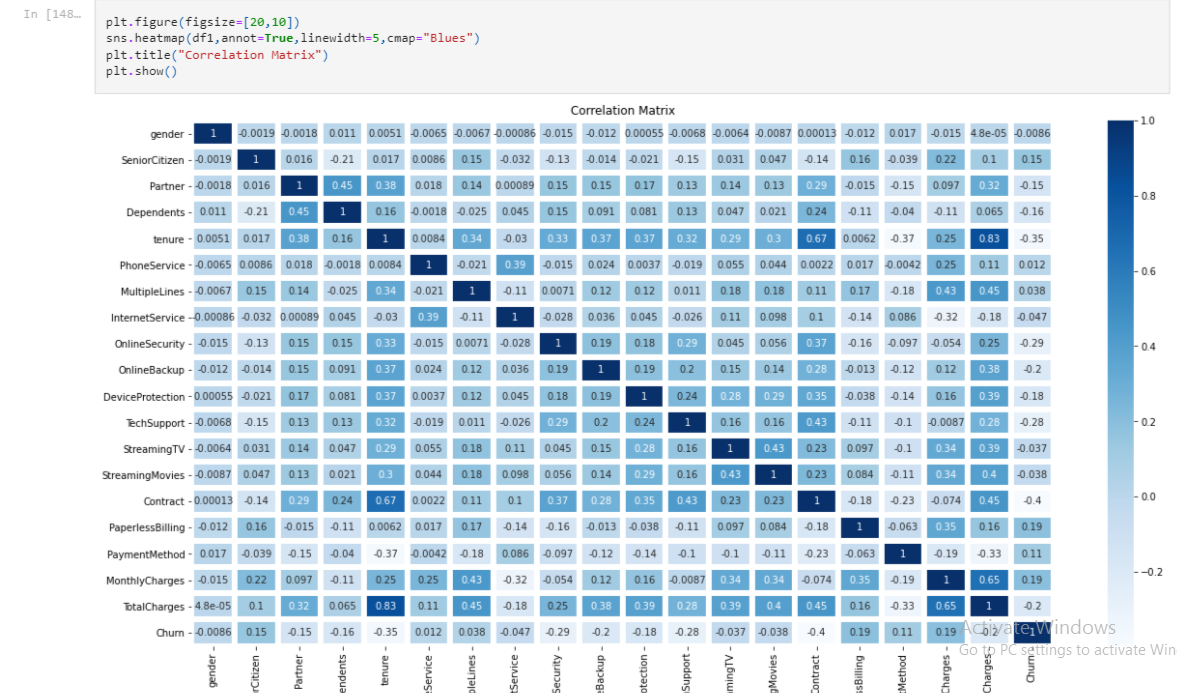
**Pre-processing Pipeline:**

There are also object-type variables, as we can see. They contain strings that the machine learning model won't be able to recognise because it doesn't understand string data types. It is only capable of recognising numerical data.

As a result, we'll convert it to numerical data using Label Encoding. The target variable 'Churn' has two distinct values, Y and N, which will be changed to 0 and 1 after encoding. Similarly, if there were three distinct values, they were changed to 0, 1, and 2.

**Checking the correlation between Independent and Target Variable:**

The heat map of correlation between all independent variable and target column churn is shown below.

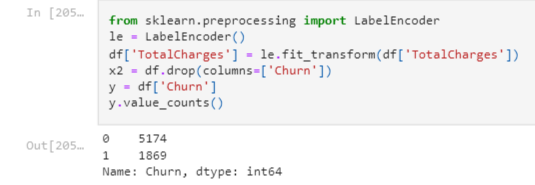


We can see that the datasets are linearly related, thus we don't need to remove any columns. We must remove any strongly linked columns from the dataset since they alter the dataset and bias the model towards it.

Now that we've done the preprocessing, we can move on to data modelling and prediction.

**Building Machine Learning Models:**

We will split the data into independent and target variables.

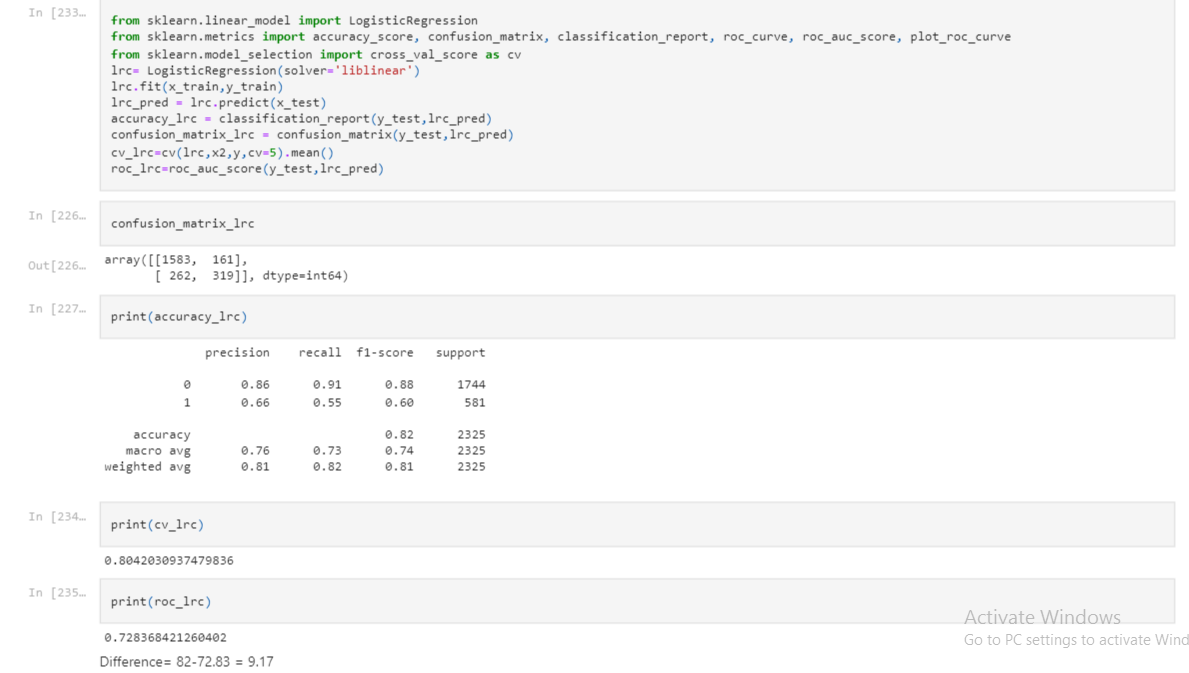


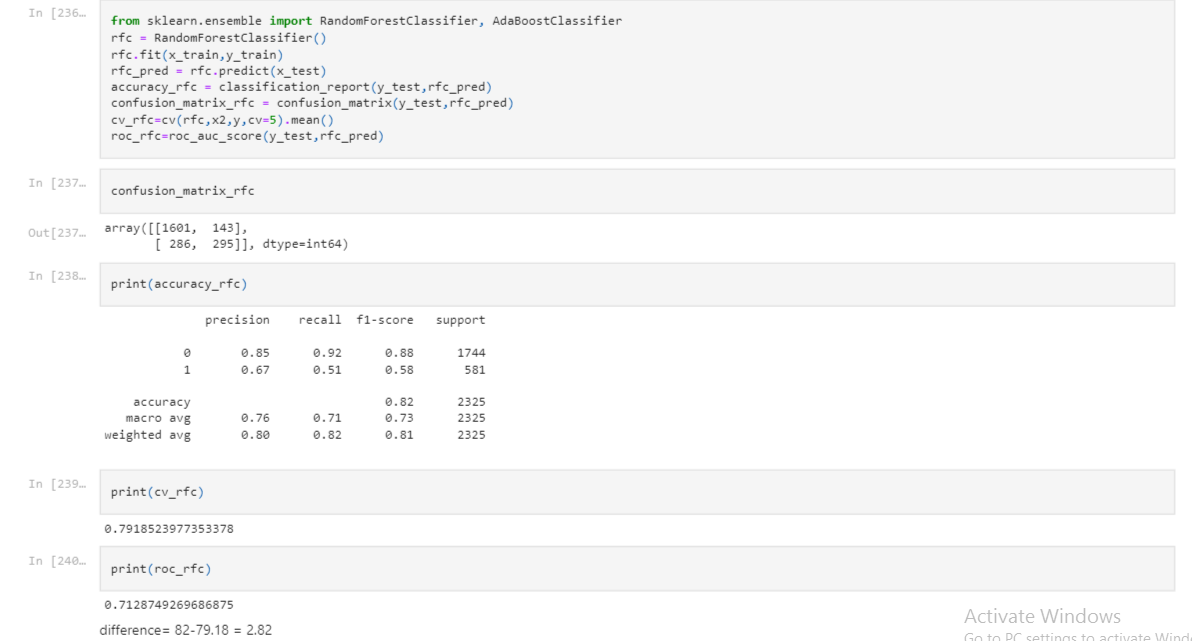
As can be seen, the goal variable is 'Chrun,' whereas the others are independent variables.

We'll now divide the data into training and testing datasets and determine which random state is the best.

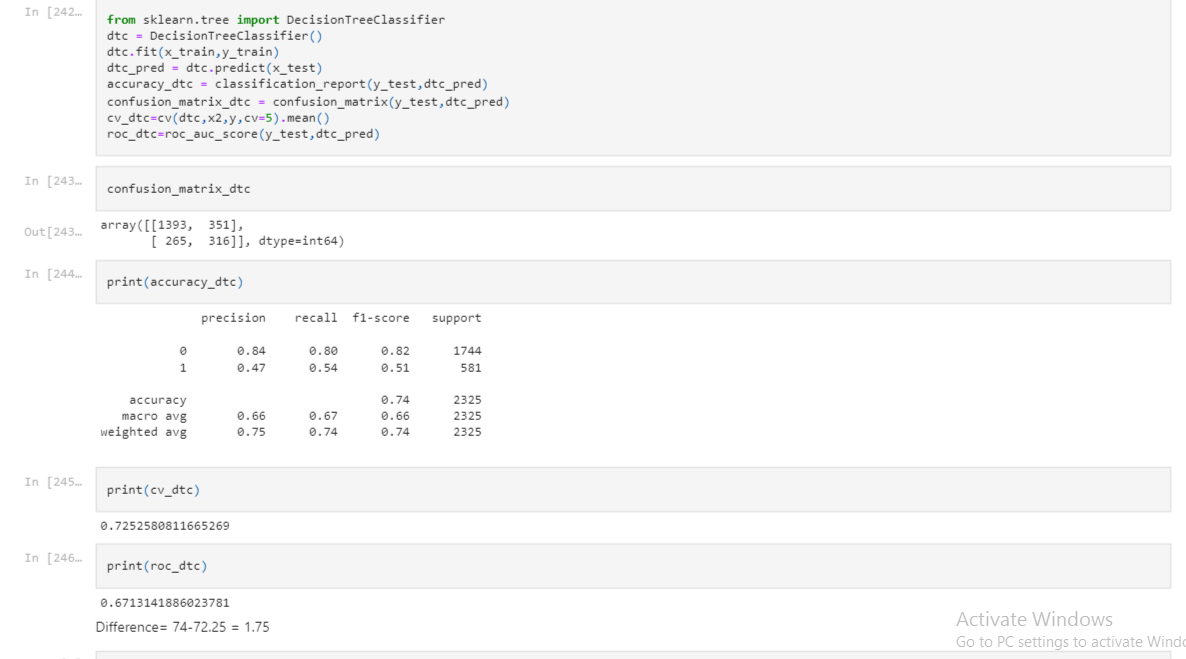
We found that 24 is the greatest random state for this, so we'll use that for our model.We'll now apply a Machine Learning model to learn from the training dataset and forecast the testing set.Because the target variable 'Churn' contains categorical data, it is a categorical problem.To anticipate the data, we'll apply five different Machine Learning models and pick the best one.

* LogisticRegression
* RandomForest Classifier
* DecisionTree Classifier
* SuppoortVector Classifier
* KNeighbors Classifier

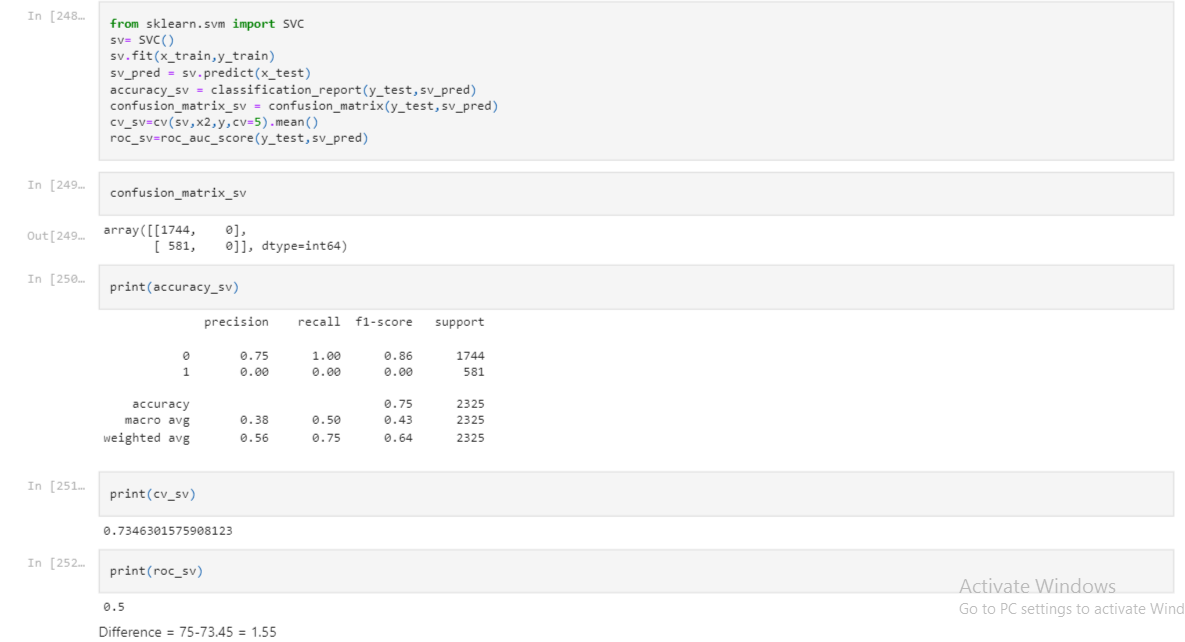
1. **LogisticRegression:**
2. **RandomForest Classifier:**

****

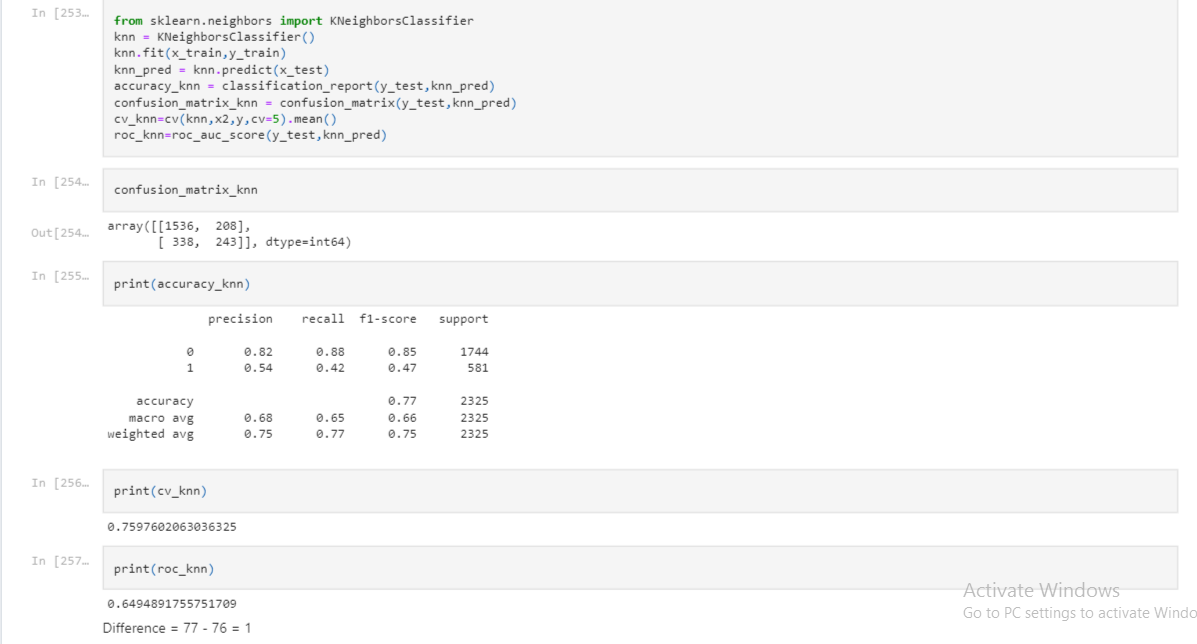
1. **DecisionTree Classifier:**

****

1. **Support Vector Classifier(SVC):**

****

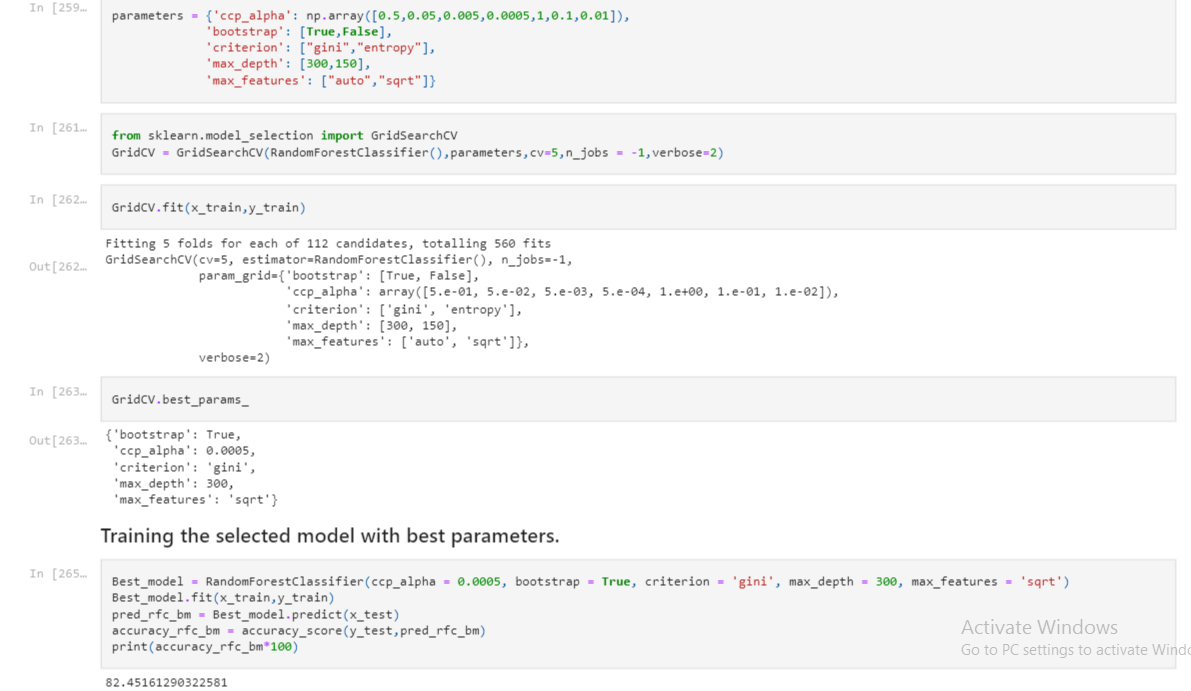
1. **KNeighbors Classifier:**

****

|  |  |  |  |
| --- | --- | --- | --- |
| **Machine Learning Model** | **Accuracy** | **Cross\_Validation score** | **Standard Deviation** |
| LogisticRegression | 0.82 | 0.80 | 0.026 |
| RandomForstClassifier | 0.82 | 0.8068 | 0.013 |
| DecisionTreeClassifier | 0.74 | 0.72 | 0.02 |
| Supportvectorclassifier | 0.75 | 0.73 | 0.02 |
| KNeighborsClassifier | 0.77 | 0.75 | 0.02 |

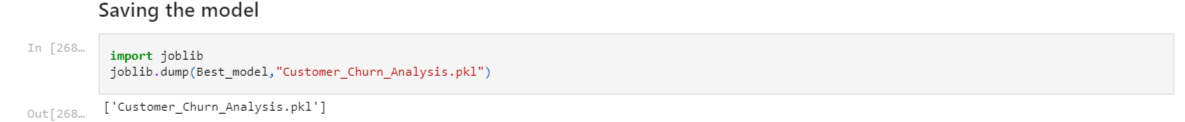
We can see from the Cross Val Score and accuracy score that the RandomForestClassifier has the least difference and has a high accuracy, hence we will choose the RandomForestClassifier model.

After finalizing the model we will hyper tune the model with GridSearchCV and select the best parameters for the model.



As we can see here that the accuracy score of the model has been increased.

After this we will save the model.



**Concluding Remarks:**

In this type of situation, The most crucial step is pre-processing and data cleaning. We must properly manage both categorical and numerical data, as well as check by generating multiple ML models on the same dataset. We must examine each model's accuracy and cross-val score and select the one that has the best of both.

The results show an accuracy of 82 percent, indicating that our algorithm correctly forecasts client retention 82% of the time. Customer churn forecast is critical to a company's long-term financial survival. This concludes our procedure. With an accuracy of 82 percent, we have successfully trained our model to predict customer data from Sample Data Sets with the goal of constructing and evaluating different customer churn prediction models.