

CUSTOMER CHURN ANALYSIS: PROJECT REPORT

Submitted by:

UTKARSH VATS

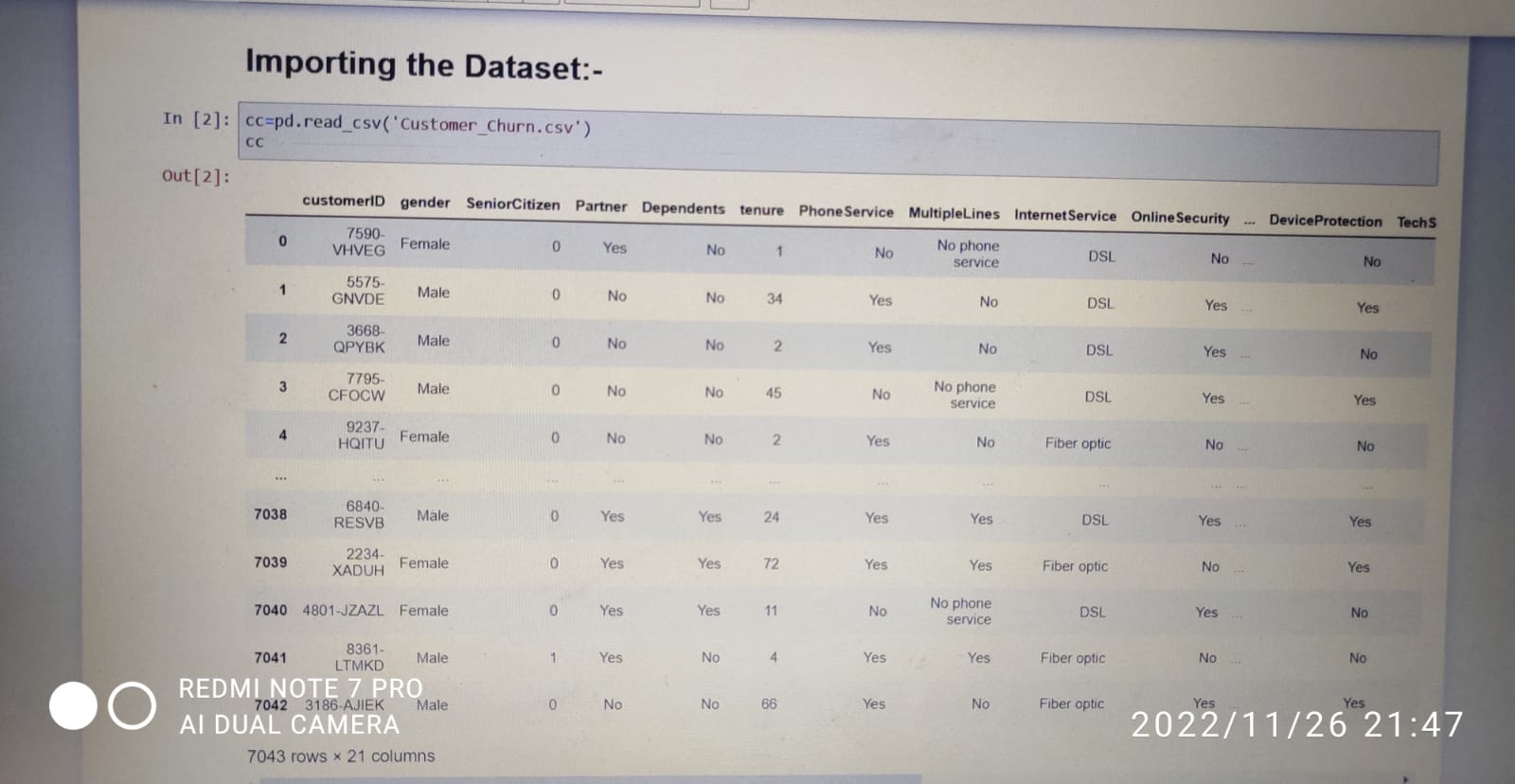
Batch no.: DS0422

* Problem Definition:

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.

The telecommunications sector has displayed one of the central industries in developed countries. Service companies like these suffer, particularly from the loss of valuable customers due to competitors known as customer churn. The scientific progress and the growing number of operators increased the level of opposition. Companies are pulling hard to survive in this aggressive market, depending on complicated strategies. The customer churn causes a considerable loss of telecom services and becomes a severe problem. Three main approaches have been introduced to generate more profits to get new customers, upsell the current customers, and increase the holding period of customers. However, comparing these strategies using the value of return on investment (ROI) of each into account has shown that the third approach is the most successful strategy, proves that maintaining an existing customer costs much lower than getting a new one, in extension to being held much easier than the upselling tactics. To implement the third strategy, companies have to reduce the potential of customer’s churn, known as “the customer movement from one provider to another.” Customers’ churn is a significant concern in service sectors with great aggressive services. On the other hand, foretelling the customers who are expected to leave the company will serve a potentially big-hearted extra revenue source if it is given in the early phase. Many types of research confirmed that machine learning technology is highly efficient in predicting this situation.

* Data Analysis:
* Software and Libraries Used:
* Jupyter Notebook (whole project analysis).
* Numpy and Pandas (data manipulation and analysis).
* Scipy (Statistics).
* Matplotlib and Seaborn (data visualization).
* Scikit-Learn (Machine Learning Algorithms and Model Building).
* Dataset: As we know, the data set is the starting point for everything; it should have full-fledged data to make the machine learn about the problem. Datasets can be generated or developed from the scrap information available on the internet. Some issues we have to create a dataset that makes sense that tells how to respond based on real-time inputs for the problem datasets can be gathered from the internet every day. A dataset is a collection of data. Most commonly, a data set has contents of a single database table, or a single statistical data matrix, where every column of the table describes a particular variable, and each row matches a given member of the data set in question. The data set lists the values of the variables, such as height, the weight of an object, for each member of the data set. Each value is recognized as a datum. Total numbers of rows: 7043 and columns: 21.



* Data Cleaning: It is very crucial to make the data useful because unwanted or null values can cause unsatisfactory results or may lead to producing less accurate results. In the data set, there are a lot of incorrect values and missing values. We analyzed the whole dataset and listed out only the useful features. The listing of features can result in better accuracy and contains only valuable features.

Observation 1(Null Values): Observations shows that there are no null values present in this dataset.

Observation 2(Data types): Observations shows that some columns are integer and float type and rest are object type and after encoding all have to be converted into integer and float type.

Whereas, Column ‘Total Charges’ contains blank values at some points in dataset that has replaced with np.nan, so that our data work properly and also imputed those null values with mean of this column as it is of float type.

* Exploratory Data Analysis:

* Data Visualization:

There are various kind of plots such as pie chart, box-plot, bar chart and scatter plot to display pictorial analysis of data for better understand.

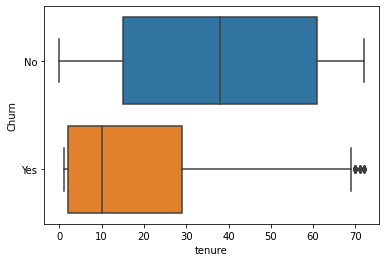
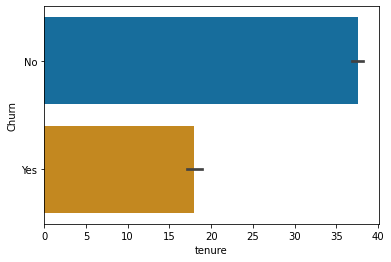
1. Churn Rate plot



Observation:

The count-plot shows that there are three times more records in churn-no class than in churn-yes class, means the dataset is imbalanced and We have a data with 27% churn rate.

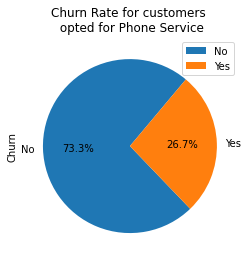
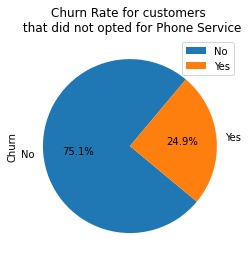
1. Tenure with respect to Churn Rate plot

Observation:

Customer with more tenure tends to churn less.

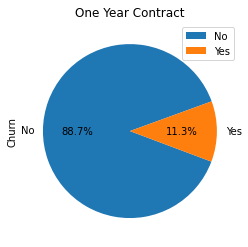
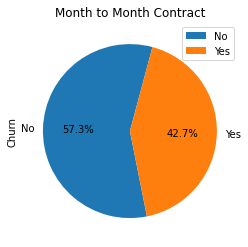
1. Phone Service with respect to Churn Rate plot

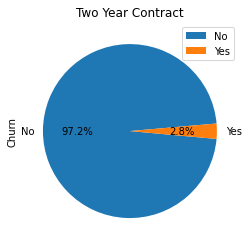
 

Observation:

Opting for phone service does not have a significant impact on churn rate.

1. Contract Period with respect to Churn Rate plot

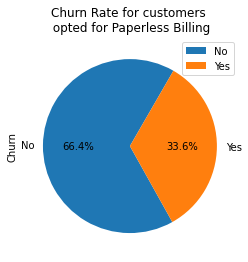
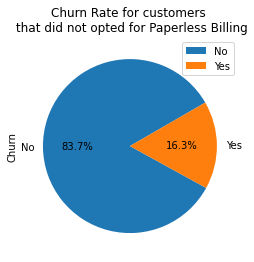




Observation:

Customers with One Year and Two Year contract tends to Churn a lot lesser than month to month contract.

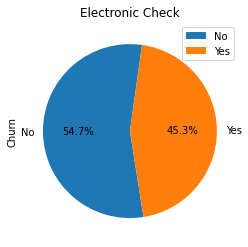
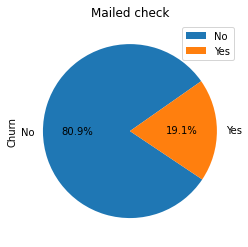
1. Paperless Billing with respect to Churn Rate plot

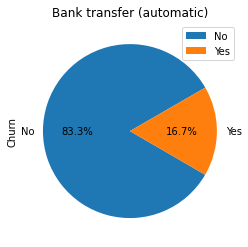
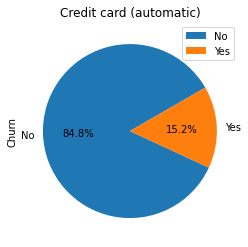
 

Observation:

Churn Rate is higher for the customers who opted for paperless billing.

1. Payment Method with respect to Churn Rate plot

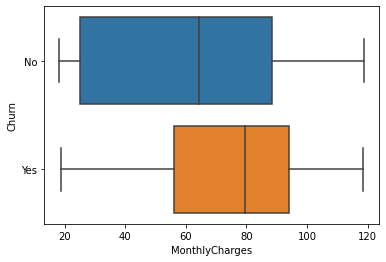
 

Observation:

Customers with Electronic Check tends to churn more than other payment methods.

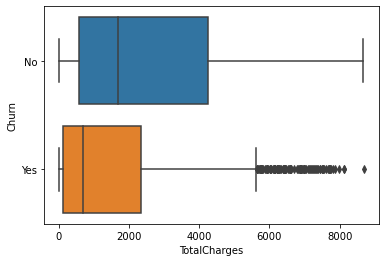
1. Monthly Charges with respect to Churn Rate plot



Observation:

Median Monthly charges are higher for customers who have churned.

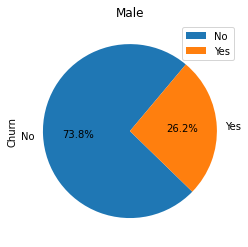
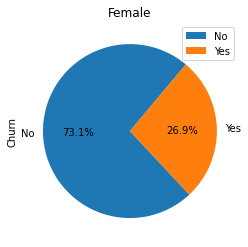
1. Total Charges with respect to Churn Rate plot



Observation:

Median Total charges are low for customers who have churned.

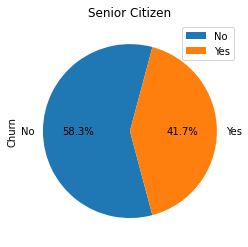
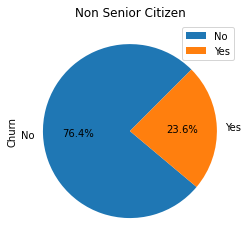
1. Gender with respect to Churn Rate plot



Observation:

There is no effect of Gender on Churn.

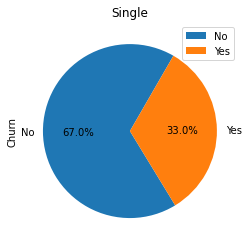
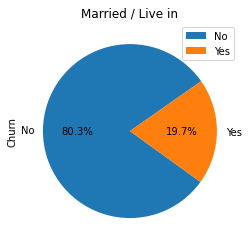
1. Senior Citizen with respect to Churn Rate plot

Observation:

Senior Citizens tends to churn more.

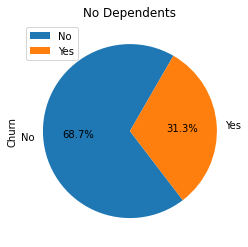
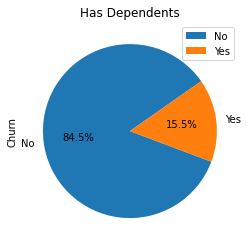
1. Partner with respect to Churn Rate plot



Observation:

Singles have higher churn rate.

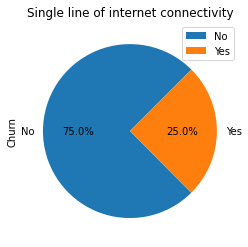
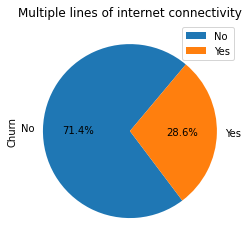
1. Dependents with respect to Churn Rate plot



Observation:

Customers with no dependents tends to churn more.

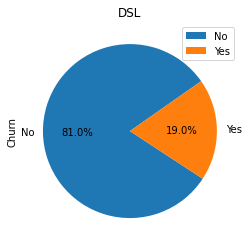
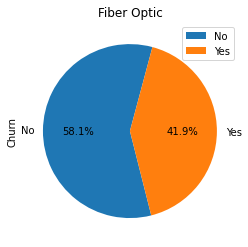
1. Multiple Lines with respect to Churn Rate plot

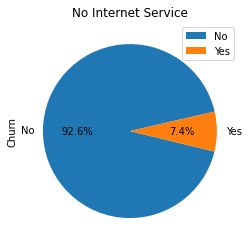


Observation:

Multiple lines of internet connectivity does not affects churn that much.

1. Internet Service with respect to Churn Rate plot

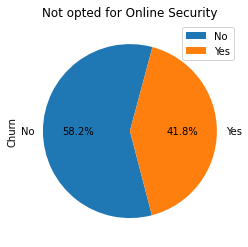
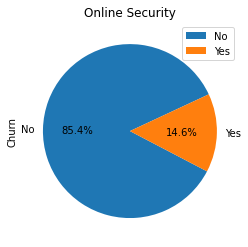




Observation:

Customers with Fiber Optic Connection churn the most whereas least probability of churning for those with no internet service.

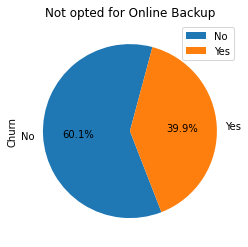
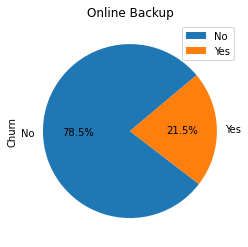
1. Online Security with respect to Churn Rate plot

****

Observation:

Customers opted for Online Security churn less than who have not opted.

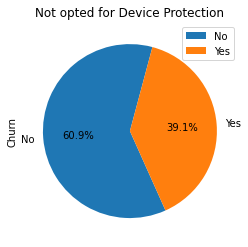
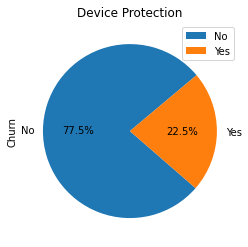
1. Online Backup with respect to Churn Rate plot

****

Observation:

Customers opted for Online Backup churn less than who have not opted.

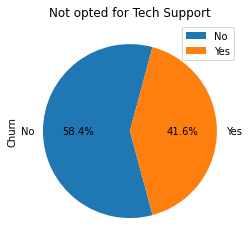
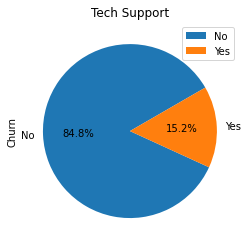
1. Device Protection with respect to Churn Rate plot

****

Observation:

Customers opted for Device Protection churn less than who have not opted.

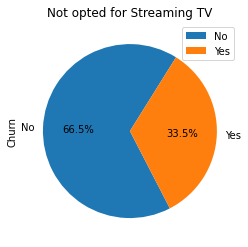
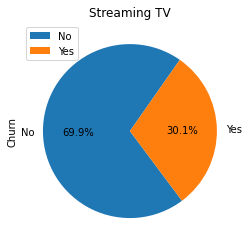
1. Tech Support with respect to Churn Rate plot



Observation:

Customers opted for Tech Support churn less than who have not opted.

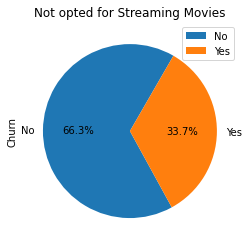
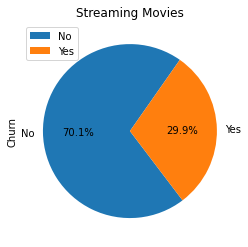
1. Streaming TV with respect to Churn Rate plot



Observation:

Streaming TV doesn't make such impact on churning.

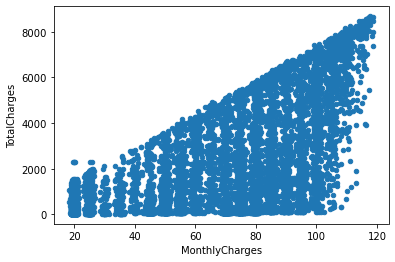
1. Streaming Movies with respect to Churn Rate plot



Observation:

Streaming Movies doesn't make such impact on churning.

1. Monthly Charges with respect to Total Charges plot



Observation:

We will observe that the total charges increase as the monthly bill for a customer increase.

* Pre-Processing Pipeline:
* Transforming the data type:

In previous observation, it has shown that some columns are in object data type that need to convert into numeric form for further process. So, with the help of Label Encoder, I have converted the object type data into integer and float type which is completely in numeric form.

* Correlation:

After data type transform, then comes correlation part. In correlation, I have checked how much input variables are correlated with the output/target variable “Churn”. For that I have developed a heatmap for better understanding. The Heatmap shows that

'InternetService','StreamingMovies','StreamingTV','customerID','gender','PhoneService' and 'MultipleLines' are less correlated with the target variable 'Churn' and rest are highly correlated.

* Feature Selection:

After the correlation process, I have dropped those columns which were less correlated ['InternetService','StreamingMovies','StreamingTV','customerID','gender','PhoneService' and 'MultipleLines'] to find the best features for further processing and results.

* Checking for the Skewness:

The features/columns (14 columns) that I got after dropping irrelevant and less correlated columns were used to check if there is any skewness present in the data or not. For that, I have plotted a distribution plots with those columns which shows that the data is not normalized and the building blocks are out of normalized curve.

* Outliers Checking and Removal:

For checking the outliers in the data, I have plotted the boxplots and also performing mathematical calculation with the Z-Score formula which shows that the data is becoming biased as it not considering the case of Senior Citizen. So, I will not consider/remove it as outlier. That’s why after the removal of outliers still the number of rows: 7043 and columns: 14 and also there is no data loss.

* Transforming the data to remove Skewness:

In order to remove the skewness from the data, I have used power transform ‘yeo-johnson’ method. Then checking the standard deviation which comes “1.0”.

* Building Machine Learning Models:

In this process, I have variable ‘x’ for input variables and ‘y’ for the target variable. For machine learning model building, all the features or I can say the complete should be first transformed into numeric form.

Then checking whether the data is balanced or not. It was unbalanced because the value counts for target variable namely 0(No) and 1(Yes) were not equal. Then, after applying the Oversampling Technique (SMOTE) our data becomes balanced.

* Libraries Imported for Model Building:

1. Train test split - For model selection.
2. Logistic Regression – For Linear Model.
3. Accuracy Score, Confusion Matrix and Classification Report - For Metrics.
4. Grid Search CV – For Hyperparameter Tuning.

* Choosing The Best Random State:

For the Further Model testing, we used for loop with a range of (1,200) to predict the best random state with the training size/phase of 80% and testing size/phase of 20%. Then, we got the best random state “56” out of “200” with maximum accuracy of “81.3%”.

* Algorithms Used and Their Cross Validation:

1. Logistic Regression:

By using logistic regression, we can predict the probability of a churn i.e., the likelihood of a customer to cancel the subscription. Logistic regression is a supervised learning algorithm used for classification. In Logistic regression, we set a threshold; based on the limit, and only the classification is made using logistic regression. The threshold value is variable, and it is dependent on the classification problem itself.

I got “81.3%” of accuracy score with cross validation value upto9 in Logistic Regression.

1. K-Nearest Neighbors:

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

I got “78.9%” of accuracy score with cross validation value upto9 in KNN.

1. Decision Tree Classifier:

The decision tree classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute. I got “73%” of accuracy score with cross validation value upto9 in Decision Tree Classifier.

1. Random Forest Classifier:

Random Forest uses Decision trees for classifying whether the customer is going to cancel his subscription. The random forest consists of a large number of decision trees. A decision tree points to a specific class. A class with more number of votes will be the classifier for a particular customer.

I got “80.2%” of accuracy score with cross validation value upto9 in Random Forest Classifier.

1. Ada Boost Classifier:

AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem. The most suited and therefore most common algorithm used with AdaBoost are decision trees with one level. I got “81.1%” of accuracy score with cross validation value upto9 in Ada Boost Classifier.

1. Gradient Boosting Classifier:

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. I got “80.8%” of accuracy score with cross validation value upto9 in Gradient Boosting Classifier.

Assumption:

Therefore, the difference between cv score and r2 score in very less in case of Logistic Regression that is [0.8126-0.8120=0.0006], So we will choose this model for further process.

* Hyper-Parameter Tuning:

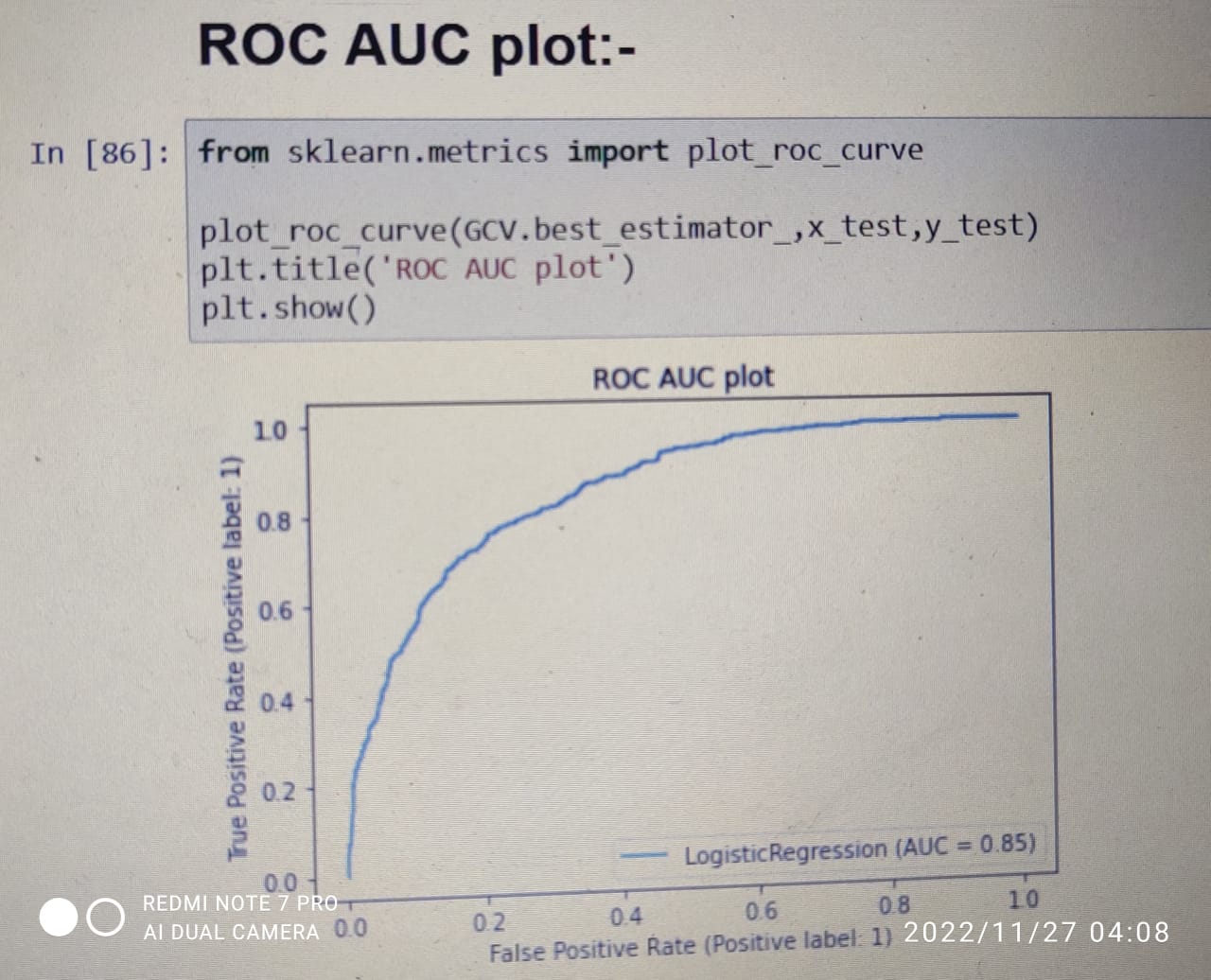
After that, I have performed hyperparameter tuning with logistic regression with its respective parameters for model prediction.

The best parameters/estimators were:

1. Max\_iter: 50
2. Penalty: l1
3. Multi\_class: ‘auto’
4. Solver: liblinear
5. CV\_Score: 5

This Hyper-Parameter yields model accuracy of “81%” with almost same cross validation score which shows that our model is working well enough.

* ROC-AUC Plot:

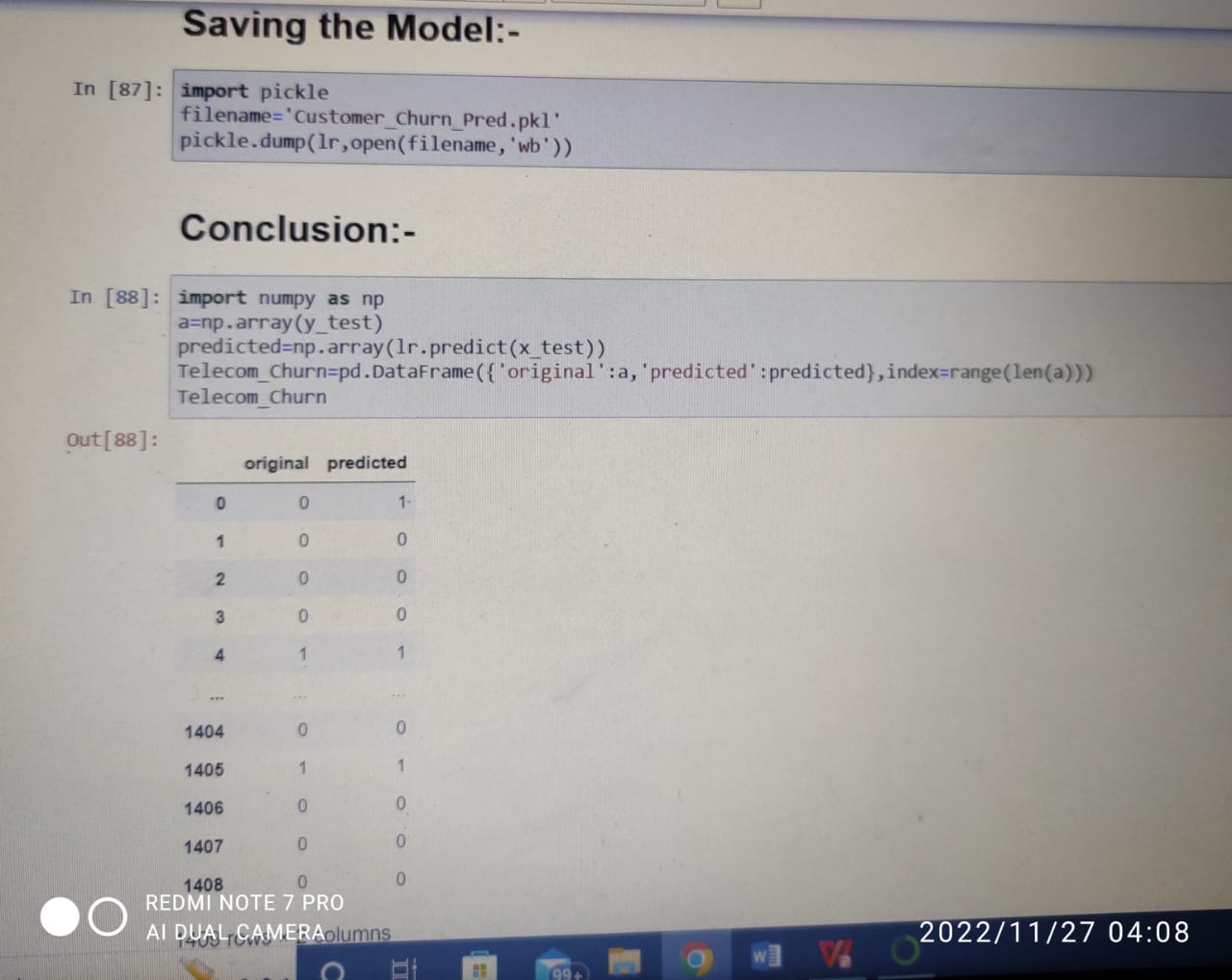


This is the AUC-ROC Curve for Logistic Regression mentioned on a snapshot above.

* Conclusion:

In this article, we walked through several steps of the Customer Churning rate prediction and model evaluation:

* Data Cleaning and Formatting.
* Exploratory Data Analysis.
* Feature Engineering and Selection.
* Compare Machine Learning models on a performance metrics.
* Perform Hyperparameter Tuning on the best model.
* Evaluate the best model on the testing set.
* Interpret the model results.
* Draw conclusion and document work.



The results of this study suggest following outputs which might be useful for

Telecom industry to reduce the churn rate and extend their business:

1. The importance of churn prediction will help many companies, mainly in telecom industries, to have a profitable income and achieve good revenue. Customer churn prediction is the major issue in the Telecom Industry, and due to this, companies are trying to keep the existing ones from leaving rather than acquiring a new customer. Three tree-based algorithms were chosen because of their applicability and diversity in this type of application.
2. Data Analysis shows that total churn rate that we have observed is 27% which is good but needs to down a little more. ‘Phone Service’, ‘Electronic Check’, ‘Senior Citizen’, ‘No dependents’ and ‘Singles’ tends to churn more means they are impact more on the churn rate. Whereas, ‘Tenure’, ‘Contract’, ‘Online Security and Backup’, ‘Tech Support’ and ‘Device Protection’ have less impact on churn rate. Some factors namely ‘Gender’, ‘Multiple Lines of Internet’ and ‘Streaming TV and Movies’ have no impact on churn.