2. Methodology

2.1 Data Collection

2.1.1 Dataset

This dataset is publicly available at Kaggle under classification datasets for anyone to download. To implement this Customer Travel Churn model a past churn dataset with over 1000 data is being used. This dataset is quite popular on Kaggle as it has hit a usability rate of 10.0. The data is stored as a CSV file.

A Tour & Travels Company Wants To Predict Whether a Customer will Churn or not help build predictive models and save the company's money.

The dataset has a target column named ‘Target’ which depicts whether the customer will churn or not.

|  |  |
| --- | --- |
| Number of Instances | 953 |
| Number of attributes | 07 |
| Number of classes | 02 |
| Number of missing values | Null |
| Related Tasks | Classification |

Table 1: Characteristics of data

2.1.2 Description of Dataset

The following table shows the attributes in the dataset and a description of those attributes showing what they represent.

|  |  |
| --- | --- |
| **Attributes** | **Description** |
| Age | Age of user |
| **FrequentFlyer** | Whether Customer takes frequent flights |
| **AnnualIncomeClass** | Class of annual income of the user |
| **ServicesOpted** | Number of times services opted during recent years |
| **AccountSyncedToSocialMedia** | Whether Company Account Of User Synchronized to Their Social Media |
| **BookedHotelOrNot** | Whether the customer book lodgings/Hotels using company services |
| **Target** | Customer Churns or not |

Table 2: Description of the attributes used

Two main classes of the customer Travel Churn data set are the ‘Churn’ and ‘not Churn’. This represents by a column name called ‘Target’. It has two outputs. One is Yes and the other is No. Yes represents that the customer has left the use of services. Since this should be taken to a machine learning model we have to replace these yes’s and no’s with 1’s and 0’s.

2.2 Data Mapping, Preprocessing, and Dummy Variables

First, need to import relevant libraries to the jupyter Notebook environment. Initially, there are four libraries being imported. They are,

* NumPy – performing statistical and mathematical operations on the Customer Travel Churn data
* Pandas - reading files and creating data frames
* Matplotlib pylot – it is a visualization import that allows visualizing data in MATLAB like plots.
* Seaborn - it is a visualization import that allows the creation of attractive and information-centric statistical graphical representation.

A screenshot of a computer

Description automatically generated

* Using pandas to create a data frame by importing a CSV file.

A screenshot of a computer

Description automatically generated

* Initial snapshot of data table head and information of attributes

A screenshot of a computer

Description automatically generated

* We can use IsNull() function to check whether there are null values.

A screenshot of a computer

Description automatically generated

* Although there are any null values in our dataset, we cannot guarantee the result. So that it would be better if we take another view option to clarify this issue. For that, the best approach is a bird’s eye view. We used the heatmap feature of the seaborn library.

A screenshot of a computer

Description automatically generated

If there are any null values, we could see yellow color bars on the heat map.

* Then we can perform some exploratory data analysis on our dataset.it is a vital thing to get to know about attributes and their relationship before we proceed. Used seaborn count plot. This is to analyze the target class.

A screenshot of a computer

Description automatically generated

* The following graph shows the Churn prediction with the Annual Income of the customer. For this, we have used parameter hue in the countplot.
* Most of the low-income customers have left from services.

A screenshot of a computer

Description automatically generated

* The following dataset is evidence for contains a lot of texts for some columns. Before we put them into any machine learning algorithm we have to convert all of them to numbers.

A picture containing text, screenshot, computer

Description automatically generated

* We have Yes and No’s in three columns. So we have used the map() function to map those data to 1’s and 0’s.

A picture containing text, screenshot, computer

Description automatically generated

* There are multiple lines to map as well. We have only one class called Annual Income class to map they are Middle Income, High Income, and Low income. We made an assumption as M, H, and L accordingly. And also it appears in the dataset after mapping.

A screenshot of a computer

Description automatically generated

* Dummy variables are important when solving real-world problems with machine learning algorithms. The pandas get dummies() allow to convert the categorical variables into indicator variables. We used dummy variables for the Annual Income class.

A screenshot of a computer

Description automatically generated

* Then the task is to concatenate the newly created columns to our existing dataset and drop the columns used for the creation of dummy values.

A screenshot of a computer

Description automatically generated

2.3 Logistic Regression Classifier

2.3.1 Introduction & Selection of the algorithm

The logistic Regression Algorithm is a basic and popular algorithm for solving a classification problem. Logistic regression is the method used to predict a dependent variable by assigning an independent set of variables to classify the dependent variable. A categorical variable is a binary variable holding values like 0 and 1 yes or no. To evaluate the output of the appropriate linear model, we can also use the output of the logistic model and use it for the forecast evaluated by considering a cut-off point. The cut-off point can be taken as 0.5. If it takes 0.5 or a value between 0 and 0.5 then it can be taken as class or 1.

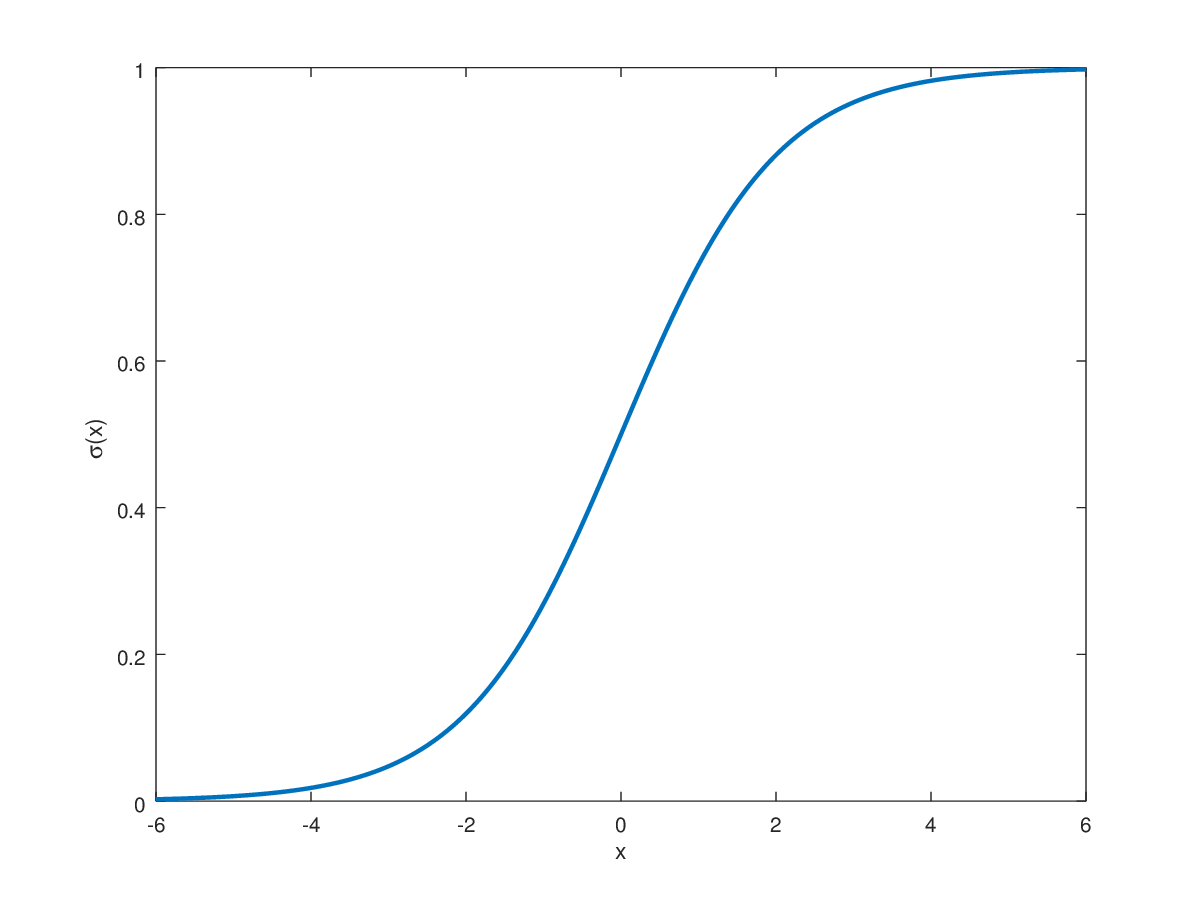
Advantages

1. Can maintain a large sample size.

2. Leads better accuracy

3. Model becomes less overfitting.

Figure 1 - Logistic Regression Sigmoid Curve

 Logistic Regression can also be used to solve problems with more than two classes. It is possible to use predict from a Tour & Travels Company Whether a customer will Churn or not. The logistic function is a sigmoid function that varies between 0 and 1. The function contains the prediction of binary outcomes which corresponds to the probability of the customer churning or not to tours and travels.

2.3.2 Basic Methodology

After activating the logistic regression classifier and attaching our linear model to it, we have divided the dataset into the training data and the test data. We can then evaluate the model performance based on some test data. For this we can use the confusing matrix.

2.3.3 Implementation of Logistic Regression Classifier

1.Seperate the dataset according to features and labels

The attributes we are using to predict the target class are the features and the target predictions are the labels. In here target prediction is the ‘Target’.

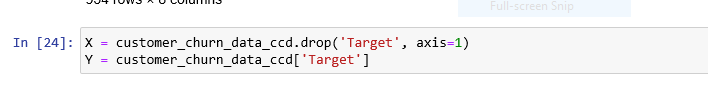


Figure 2 - Split the data

2.Train Test Split

In the Tour & Travels Customer Churn Prediction data has been split to 20% for testing data and 80% to training data.



Figure 3 - Train Test Split

3.Create the Logistic Regression Classifier

First, we have to import the Logistic Regression from the linear model. Then create the object and assign it to the training features and labels.

Figure 4 - Logistic Regression Creation

4.Do the predictions on the test data set

The model will predict whether Tour & Travels Customer Churn or not using testing data.

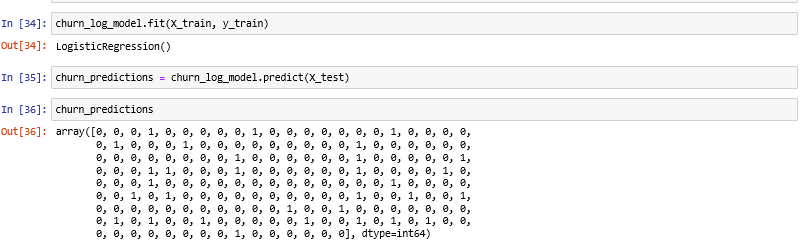


Figure 5 - Prediction

2.4 Testing

After the prediction we need to analyze the accuracy of the Logistic Regression Classifier.

1.Accuracy of the predicted data

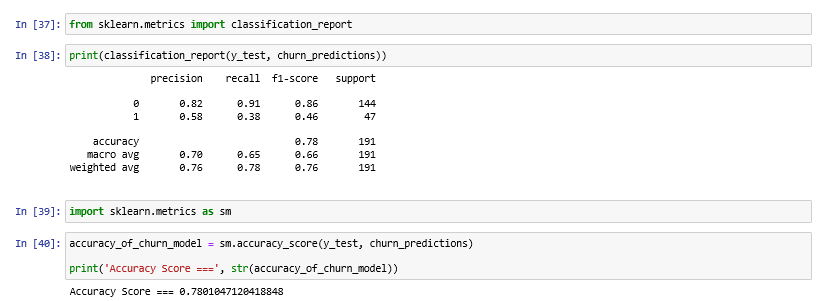
Tour & Travels Customer Churn Prediction dataset has been split into 20% for testing data and 80% to training data. For the prediction use 20% for testing data. The accuracy score of the classifier is 0.7801047120418848 (78%).

Figure 6 - Accuracy Score

2.Evaluate Model Using Confusion Matrix

Confusion Matrix gives a graphical representation of true positives, true negatives, false positives, and false negatives and is used to evaluate the result of the Logistic Regression Classification Algorithm. The model has predicted that 131 people have not churned the tour and travels and falsely about 13 people. The other part predicted that 29 people churn the tour and travel correctly and falsely predicted about 18 people.



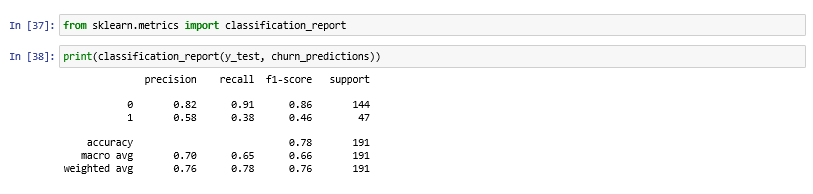
Figure 7 - Confusion Matrix

3.Classification Report Generation

We can use the classification report to list the most important parameters of the classification for each category. This shows a very deep understanding of the classifier. The main parameters set out in this report are:

* Precision - The capability of the Logistic Regression Classifier indicates that examples cannot be labeled positively.
* Recall - Shows the ability of the machine learning classifier to find all positive opportunities
* F1 score - Shows class accuracy for classification of samples belonging to this class compared to other class
* Support - How many actual class events are in the Churn

Figure 8 - Classification Report Generation



4. Feature Importance

This is important to identify which feature affects most the travel and tour customer churn prediction.

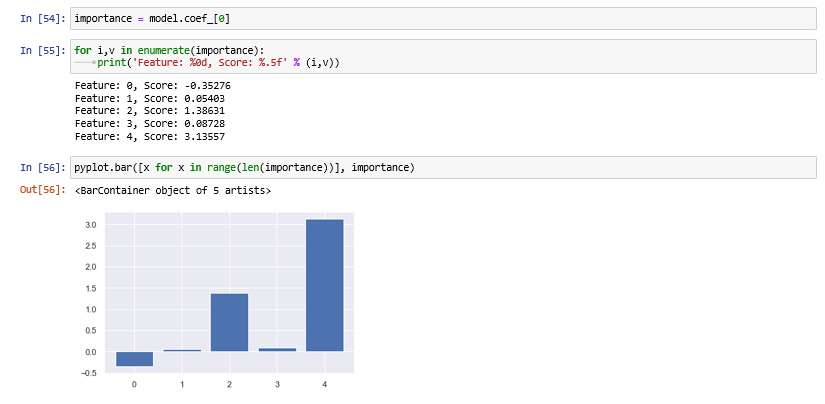


Figure 9 - Feature Importance