1. Introduction

1.1 Objective:-

Develop a machine learning model to predict customer churn based on historical customer data. You will follow a typical machine learning project pipeline, from data preprocessing to model deployment.

What is the problem?

The main problem is to predict if a customer would be credit defaulter or not depending upon the previous data of the customer.

Why is it important?

It is important from a bank’s perspective in order to maintain business and customer relationship/ Apart from that if someone could be predicted as a defaulter then primitive measures can be taken in order to ensure that such violations do not happen.

What is your basic approach?

The basic approach of solving this problem was first studying the data , then bringing out insights from the dataset and after that I have followed a machine learning pipeline in order to solve the problem.

The ML Pipeline that I have followed is :

* Importing the necessary libraries and the dataset
* Performing Data Preprocessing (Exploratory Data Analysis and Data Manipulation)
* Modelling using Logistic Regression, KNN and Random Forest
* Performing Prediction
* Visualization in between Actual and predicted Values

The environment used was python 3.7 and the libraries such as numpy, pandas, matplotlib , Standard Scaler and Scikit Learn module were used for Scientific computations.

2. Problem Definition and Algorithm

The Problem Statement is as follows :-

Customer Churn Prediction:

The ML Pipeline that I have followed is :

Importing the necessary libraries and the dataset:

Here, the libraries such as numpy, pandas and matplotlib were called. Numpy is known to be Numerical Python which is responsible for performing all the Numerical tasks in this project whereas Pandas would make the data frame and Matplotlib was used for visualization..

Performing Data Preprocessing (Exploratory Data Analysis and Data Manipulation:-

In this step the data was thoroughly analysed. Univariate Analysis would be the analysis of a data if we are studying/ analysing a particular data majorly depending upon its type (for e.g Continuous or Categorical Type )

If the data is said to be a continuous data then in order to do a univariate analysis we check the distribution of the data i.e, we check if the data is normally distributed or not and if the data is said to be a categorical data then we would check the bar plot of that data..In order to perform a Bivariate Analysis (The study of two data at an instance) we would have to plot the scatter plot in order to check the relationship in between them.

Modelling using Logistic Regression, KNN and Random Forest

The modelling was done by using Logistic Regression , Random Forest and KNN and at the end it was ensemble by using VotingClassifier.

The Following are the algorithms that were used for performing the predictions

2.2.1 Logistic Regression:-

Classification techniques are an essential part of machine learning and data mining applications. Approximately 70% of problems in Data Science are classification problems. There are lots of classification problems that are available, but the logistics regression is common and is a useful regression method for solving the binary classification problem. Another category of classification is Multinomial classification, which handles the issues where multiple classes are present in the target variable. For example, IRIS dataset a very famous example of multi-class classification. Other examples are classifying article/blog/document category.

Logistic Regression can be used for various classification problems such as spam detection. Diabetes prediction, if a given customer will purchase a particular product or will they churn another competitor, whether the user will click on a given advertisement link or not, and many more examples are in the bucket.

Logistic Regression is one of the most simple and commonly used Machine Learning algorithms for two-class classification. It is easy to implement and can be used as the baseline for any binary classification problem. Its basic fundamental concepts are also constructive in deep learning. Logistic regression describes and estimates the relationship between one dependent binary variable and independent variables.

In this tutorial, you will learn the following things in Logistic Regression:

Introduction to Logistic Regression

Linear Regression Vs. Logistic Regression

Maximum Likelihood Estimation Vs. Ordinary Least Square Method

How do Logistic Regression works?

Model building in Scikit-learn

Model Evaluation using Confusion Matrix.

Advantages and Disadvantages of Logistic Regression

Logistic Regression

Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. For example, it can be used for cancer detection problems. It computes the probability of an event occurrence.

It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.

Linear Regression Equation:

Where, y is dependent variable and x1, x2 ... and Xn are explanatory variables.

Properties of Logistic Regression:

The dependent variable in logistic regression follows Bernoulli Distribution.

Estimation is done through maximum likelihood.

No R Square, Model fitness is calculated through Concordance, KS-Statistics.

Linear Regression Vs. Logistic Regression

Linear regression gives you a continuous output, but logistic regression provides a constant output. An example of the continuous output is house price and stock price. Example's of the discrete output is predicting whether a patient has cancer or not, predicting whether the customer will churn. Linear regression is estimated using Ordinary Least Squares (OLS) while logistic regression is estimated using Maximum Likelihood Estimation (MLE) approach.

Sigmoid Function

The sigmoid function, also called logistic function gives an ‘S’ shaped curve that can take any real-valued number and map it into a value between 0 and 1. If the curve goes to positive infinity, y predicted will become 1, and if the curve goes to negative infinity, y predicted will become 0. If the output of the sigmoid function is more than 0.5, we can classify the outcome as 1 or YES, and if it is less than 0.5, we can classify it as 0 or NO. The outputcannotFor example: If the output is 0.75, we can say in terms of probability as: There is a 75 percent chance that patient will suffer from cancer.

Types of Logistic Regression

Binary Logistic Regression: The target variable has only two possible outcomes such as Spam or Not Spam, Cancer or No Cancer.

Multinomial Logistic Regression: The target variable has three or more nominal categories such as predicting the type of Wine.

Ordinal Logistic Regression: the target variable has three or more ordinal categories such as restaurant or product rating from 1 to 5.

Model building in Scikit-learn