Experiment 5

Aim: Regression Analysis

a) Perform Logistic regression to find out relation between variables

b) Apply regression model technique to predict the data on the dataset.

About Dataset:

The data set contains customer demographics (age, gender, marital status, citytype, stay in current city), product details (productid and product category) and total purchase amount from last month. From this dataset a company can understand the customer purchase behavior (specifically, purchase amount) against various products of different categories. From this dataset we can build a model to predict the purchase amount of customers against various products which will help them to create personalized offers for customers against different products.

The dataset contains details such as the Airline Company, Origin and destination of flight, Departure and arrival time of the flight. The task is to predict whether a given flight will be delayed. As the number of flights is increasing everyday and people are starting to travel more and more using flights, the delay in flights is inevitable. So predicting the delay in flights is a necessity.

Theory:

Regression is a supervised machine learning technique which is used to predict continuous values. The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data. It helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables. Based on the number of input features and output labels, regression is classified as linear (one input and one output), multiple (many inputs and one output) and multivariate (many outputs). In Regression, we plot a graph between the variables which best fits the given data points, using this plot, the machine learning model can make predictions about the data.

Terminologies Related to the Regression Analysis:

- Dependent Variable: The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called target Variable.
- Independent Variable: The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variables, also called as a predictor.

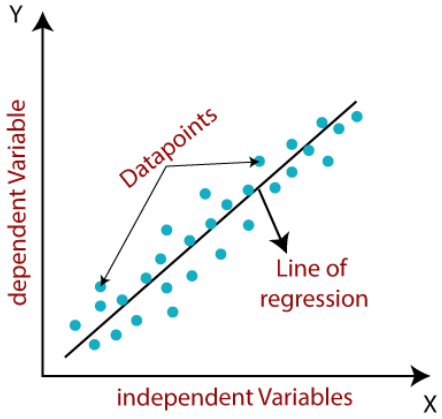
- Outliers: Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.
- Multicollinearity: If the independent variables are highly correlated with each other than other variables, then such a condition is called Multicollinearity. It should not be present in the dataset, because it creates a problem while ranking the most affecting variable.
- Underfitting and Overfitting: If our algorithm works well with the training dataset but not well with the test dataset, then such a problem is called Overfitting. And if our algorithm does not perform well even with a training dataset, then such a problem is called underfitting.

1. Linear regression:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Linear regression is used in many different fields, including finance, economics, and psychology, to understand and predict the behavior of a particular variable. For example, in finance, linear regression might be used to understand the relationship between a company's stock price and its earnings, or to predict the future value of a currency based on its past performance.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**

  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- **Multiple Linear regression:**

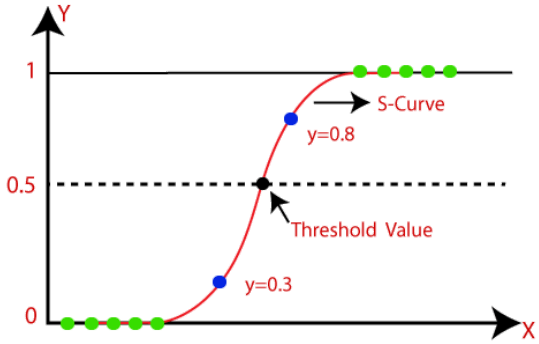
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

2. Logistic Regression:

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



Type of Logistic Regression:

Logistic Regression can be classified into three types:

- **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
- **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
- **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Implementation:-

```
import pandas  as pd #Data manipulation
import numpy as np #Data manipulation
import matplotlib.pyplot as plt # Visualization
import seaborn as sns #Visualization
plt.rcParams['figure.figsize'] = [8,5]
plt.rcParams['font.size'] =14
plt.rcParams['font.weight']= 'bold'
plt.style.use('seaborn-whitegrid')
```

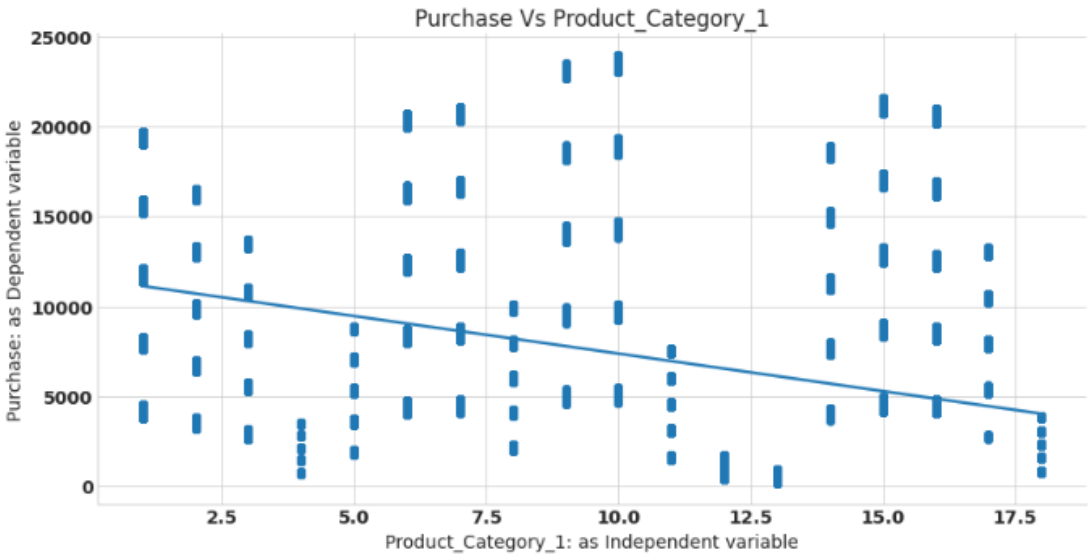
```
print('\nNumber of rows and columns in the data set: ',df.shape)
print('')

#Lets look into top few rows and columns in the dataset
df.head()
```

```
Number of rows and columns in the data set:  (545915, 11)
```

	Product_ID	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Purchase	F	M
0	P00069042	0-17	10	A	2	0	3	9.842329	8370	1	0
1	P00248942	0-17	10	A	2	0	1	6.000000	15200	1	0
2	P00087842	0-17	10	A	2	0	12	9.842329	1422	1	0
3	P00085442	0-17	10	A	2	0	12	14.000000	1057	1	0
4	P00285442	55+	16	C	4+	0	8	9.842329	7969	0	1

```
sns.lmplot(x='Product_Category_1',y='Purchase',data=df,aspect=2,height=6)
plt.xlabel('Product_Category_1: as Independent variable')
plt.ylabel('Purchase: as Dependent variable')
plt.title('Purchase Vs Product_Category_1');
```




```
from scipy import stats

x = df.Product_Category_1
y = df.Purchase

slope, intercept, r, p, std_err = stats.linregress(x, y)

def myfunc(x):
  return slope * x + intercept

purchase = myfunc(10)

print(purchase)
```

```
7248.255767660475
```

Logistic regression:

In [1]:
```
import pandas as pd
import numpy as np
```

Data Description

Flight : ID of Flight

Time : Departure Time

Length : Length of Flight

Airline : Airline ID

AirportFrom : Which airport the flight flew from

AirportTo : Which airport the flight flew to

DayOfWeek : Day of the week of the flight

Class : Delayed(1) or Not(0)

In [2]:
```
df = pd.read_csv("airlines_delay.csv")
df
```

Out[2]:

	Flight	Time	Length	Airline	AirportFrom	AirportTo	DayOfWeek	Class
0	2313.0	1296.0	141.0	DL	ATL	HOU	1	0
1	6948.0	360.0	146.0	OO	COS	ORD	4	0
2	1247.0	1170.0	143.0	B6	BOS	CLT	3	0
3	31.0	1410.0	344.0	US	OGG	PHX	6	0
4	563.0	692.0	98.0	FL	BMI	ATL	4	0
...	...	...	...	...	...	...	...	...
539377	6973.0	530.0	72.0	OO	GEG	SEA	5	1
539378	1264.0	560.0	115.0	WN	LAS	DEN	4	1
539379	5209.0	827.0	74.0	EV	CAE	ATL	2	1
539380	607.0	715.0	65.0	WN	BWI	BUF	4	1
539381	6377.0	770.0	55.0	OO	CPR	DEN	2	1

539382 rows × 8 columns

Model with ouliers and without normalization

In [19]:
```
from sklearn.model_selection import train_test_split
df = pd.read_csv("airlines_delay.csv")
```

In [20]:
```
X = df_model.drop(['Class'], axis=1)
y = df_model.Class
```

In [21]:
```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.30)
```

In [22]:
```
from sklearn.linear_model import LogisticRegression
```

In [23]:
```
model = LogisticRegression()
```

In [24]:
```
clf = model.fit(X_train,y_train)
```

In [25]:
```
pred = clf.predict(X_test)
```

In [26]:
```
clf.score(X_test, y_test)
```

Out[26]:
```
0.6383153601334858
```

Model with ouliers and with normalization

In [27]:
```
def normalize(data, column):
    max = data[column].max()
    min = data[column].min()
    norm_df = (data[column] - min)/(max-min)
    return norm_df
```

In [28]:
```
df = pd.read_csv("airlines_delay.csv")
df['Length'] = normalize(df, 'Length')
df['Time'] = normalize(df, 'Time')
```

In [29]:
```
df_model = pd.get_dummies(data=df, columns=['Airline','AirportFrom','AirportTo'])
```

In [30]:
```
X = df_model.drop(['Class'], axis=1)
y = df_model.Class
```

In [31]:
```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.30)
```

In [32]:
```
from sklearn.linear_model import LogisticRegression
```

In [33]:
```
model = LogisticRegression(max_iter=1000)
clf = model.fit(X_train,y_train)
pred = clf.predict(X_test)
clf.score(X_test, y_test)
```

Conclusion:

In this experiment we have successfully applied linear regression and logistic regression. In the process we came to know about the accuracy of the model and the prediction it gave based on various inputs.