Assignment 2

Group no: 118

Jhanvi Kasundra

Harsh Nisar

**Report for Part-1**

1. Provide brief details about the nature of your dataset. What is it about? What type of data are we encountering? How many entries and variables does the dataset comprise?

Ans: The first datasets we used are

* 1. Diamonds.

It gives us information about the diamond’s carat, price, depth, etc. which id basically its features. The data set consists of 53940 rows and 11 columns as in attributes or features. Here’s the insight of dataset:

A screenshot of a computer

Description automatically generated with low confidence

* 1. Titanic

This dataset gives us information about – how many people survived, age, sex, Fares, people aboard, etc. The dataset consists of 887 rows and 8 columns. Here’s are the insights of the data set:

A screenshot of a computer

Description automatically generated with medium confidence

* 1. Amazon’s top selling books

This dataset gives us information about – authors, books, prices, ratings, reviews, year books were released, etc. The dataset consists of 550 rows and 7 columns. Here’s are the insights of the data set:

Table

Description automatically generated

1. Provide the main statistics about the entries of the dataset (mean, std, number of missing values, etc.

Ans: Dataset1- Diamonds

Table

Description automatically generated

Dataset2- Titanic

Table

Description automatically generated

Dataset3- Amazon’s top selling books

Table

Description automatically generated

1. Provide at least 5 visualization graphs with short description for each graph, e.g. discuss if there any interesting patterns or correlations.

Ans:

Dataset1- Diamonds

Graph1- This is a pair plot graph which helps us see both distribution of single variables and relationships between two variables. We’ve used reg as kind which means regression to plot data and a linear regression model fit. In this graph all column values are plotted against all column values giving us a gist of relationship between each of them.You can see that below:

A picture containing shoji

Description automatically generated

Graph2- This is a pair plot graph which helps us see both distribution of single variables and relationships between two variables. We’ve used kde which is a method for visualizing the distribution of observations in a dataset, like a histogram. It represents the data using a continuous probability density curve in one or more dimensions and forms a less cluttered graph. The first part shows clarity w.r.t. carat, second graph shows clarity, cut and other two carat with depth and cut and the shading part shows the price variation .You can see that below:

A picture containing calendar

Description automatically generated

Graph3- This is a histogram which gives us information about carat with respect to cut of the diamonds. You can see that below:

Chart, bar chart

Description automatically generated

Graph4- This graph displays price based on the cut of diamonds. As we see, price is not necessarily increasing with cut. It increases till midway and then goes low. You can see that below:

Logo, company name

Description automatically generated

Graph5- This is a barR graph. It gives us insights about the prices based on the carat of diamonds. You can see that below:

Chart, histogram

Description automatically generated

Dataset2- Titanic

Graph1- This is a pair plot graph which helps us see both distribution of single variables and relationships between two variables. We’ve used reg as kind which means regression to plot data and a linear regression model fit. You can see that below:

A picture containing shoji, building, window

Description automatically generated

Graph2- This is a pair plot graph which helps us see both distribution of single variables and relationships between two variables. We’ve used kde which represents the data using a continuous probability density curve in one or more dimensions. You can see that below:

A picture containing text, electronics, keyboard

Description automatically generated

Graph3- This is a histogram which gives us information about people who survived with respect to their age. You can see that below:

Chart, histogram

Description automatically generated

Graph4- This graph displays fare with respect to age of people . You can see that below:

Chart

Description automatically generated

Graph5- This is a line graph. It gives us insights about the age of the people who survived. You can see that below:

A picture containing diagram

Description automatically generated

Graph6- This is a bar graph. It gives us insights about the Pclass and age. You can see that below:

Chart, histogram

Description automatically generated

Dataset3-

Graph1- This is a bar graph which gives us insights about how many top selling books were released in each year. You can see that below:

Chart, bar chart

Description automatically generated

Graph2- This is pair plot which gives us insights about Genre, Rating and Author w.r.t., Genre and Price. You can see that below:

Chart, scatter chart

Description automatically generated

Graph3- This is displot which gives us insights about Authors and ratings. You can see that below:

Chart, histogram

Description automatically generated

Graph4- This is factor plot which is used to draw a categorical plot onto a Facet Grid. You can see that below:

Chart, line chart

Description automatically generated

Graph5- This is displot which gives us insights about Genre and price. You can see that below:

Chart, histogram

Description automatically generated

**Report for Part 2- Logistic Regression**

1. Provide your best accuracy and the weight vector

The best accuracy we’ve received is 86.15%Text

Description automatically generated

And the weight vector and bias are:

Table

Description automatically generated

1. Include loss graph and provide a short description

Loss graph is plotting training and testing error over the number of iterations of model

Chart, line chart

Description automatically generated

1. Explain how hyperparameters influence the accuracy of the model. Provide at least 3 different setups with learning rate and #iterations and discuss the results Graphical user interface, chart, line chart

   Description automatically generatedTable

   Description automatically generated

Graphical user interface

Description automatically generated

Text

Description automatically generated

Graphical user interface

Description automatically generated

Text

Description automatically generated

As we see, the graph of 80% accuracy is similar to graph of 84% accuracy . For 78% accuracy the loss graph is more curved comparatively.

1. Discuss the benefits/drawbacks of using a Logistic Regression model.

Logistic regression uses a logistic function to model dependent variable and estimates the parameters of a logistic model. It is used when the target variable is categorical.

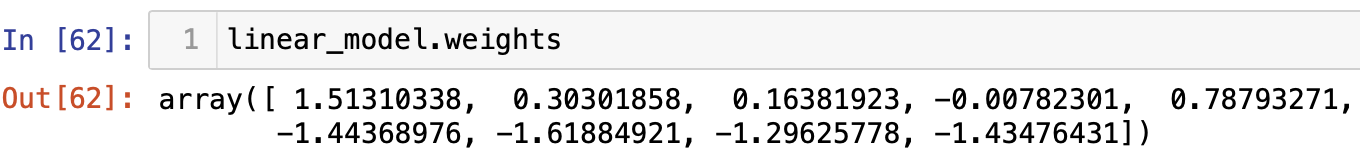
Logistic regression is efficient to train and easy to implement. If the number of observations < number of features, Logistic Regression leads to overfitting.

Logistic models can be updated easily with new data using gradient descent.

When there are multiple or non-linear decision boundaries logistic regression will not perform well. It assumes there is linearity between the target variable and the independent variables. It cannot work properly if the input data has errors or if data is not preprocessed properly.

**Report for Part 3- OLS Linear Regression**

1. Provide your loss value and the weight vector



1. Show the plot comparing the predictions vs the actual test data

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

1. Discuss the benefits/drawbacks of using OLS estimate for computing weights

OLS is used for estimating the unknown parameters in a linear regression model. It minimizes the sum of squared vertical distances between the observed responses in the dataset and the responses predicted by the linear approximation. Benefits of OLS is that it reveals information about cost structures and distinguishes between different features affecting output. The adjustment turns the OLS into a “frontier” approach. The drawback of OLS is that a large data set is necessary to obtain reliable results. OLS estimator has a minimum variance. It is unbiased and has the minimum variance .

Report for PART - 4

1. Provide your loss value and the weight vector
2. Show the plot comparing the predictions vs the actual test data
3. Discuss the difference between Linear and Ridge regressions. What is the main motivation of using l2 regularization?

Linear Regression establishes a relationship between target variable and one or more independent variables using regression line.

The dependent variable is continuous, independent variables can be continuous or discrete, and regression line is linear.

Ridge Regression is used when the independent variables are highly correlated.

Even though the OLS are unbiased, their variances are large which deviates the observed value far from the true value. Adding a degree of bias reduces the standard errors.

The main motivation of using L2 regularization is to reduce the chance of model overfitting.

|  |  |  |
| --- | --- | --- |
| Team Member | Assignment Part | Contribution |
| Jhanvi Kasundra | Part 1 | 60% |
| Part 2 | 50% |
| Part 3 | 100% |
| Report | 80% |
| Harsh Nisar | Part 1 | 40% |
| Part 2 | 50% |
| Part 4 | 100% |
| Report | 20% |