**INFO 5505 – APPLIED MACHINE LEARNING FOR DATA SCIETISTS**

**Assignment -1**

**Data:**

This data set is a monet painting data set which has 86 observations with Price of the painting with some other features of the painting. The data is cleaned data set.

The objective of this assignment is to develop predictive models that can predict the price of paintings using different independent variables. There are two kinds of Linear regression models are developed, simple linear regression and multiple linear regression.

1. **Data Loading**

The data is loaded into a data frame and verified using pandas library as shown below.

Table

Description automatically generated with medium confidence

As we can see from the above result there are 6 variables in the data set.

Graphical user interface, application

Description automatically generated

This data set has 430 observations.

1. **Checking for missing values:**

While building any predictive model it is very important to check whether the give data set has any missing values in it so that we can replace or remove those missing values.

Graphical user interface

Description automatically generated with medium confidence

From the above msno matrix diagram it can be inferred that the data set has no missing values in any feature.

1. **Creating size variable**

A new variable size is created in the data set which is the multiplication of height and width variables and Height

Table

Description automatically generated

1. **Skewness and transforming variables.**

Skewness value shows whether a variable is normally distributed or not. Below are the values of skewness of all the numeric variables.

Graphical user interface, text, application, chat or text message

Description automatically generated

From the above values it can be inferred that Price, Height, Width and size are positively skewed whereas Signed is negatively skewed.

These variables are log transformed to make them normally distributed.

Graphical user interface, text, application

Description automatically generated

1. **Finding Correlation between variables.**

It is important to find out the correlation between all the variables in the data set. This is checked by using the heatmap as shown below.

A screenshot of a computer

Description automatically generated with medium confidence

From the above correlation heatmap it can be inferred that painting price is highly related (38%) to size of the painting followed by width and height. It is very less related to picture, signed and house features.

1. **Developing models:**

Model 1 – Simple Linear Regression:

Independent variable: log\_size

Dependent variable: log\_price

As show below X data set is taken with the column log\_size and y data set log\_price. These data sets are split into train and test data sets with 80% and 20% volume of original data set respectively.

Graphical user interface, text, application, email

Description automatically generated

Regression plot for the model:

Chart, scatter chart

Description automatically generated

From the above plot it can be inferred that log\_size and log\_price is linearly related. The line represents the predicted values of the price with the given size values. The dots on the plot represents actual values of the price.

The distance between the line and actual values (The average difference between the predicted and actual prices) is calculated by using Root Mean square error.

**Root Mean square error of the model:**

RMS error represents the error of the prediction with the test data set. It is calculated as shown below.

Graphical user interface, text

Description automatically generated

The error of the prediction with the test data set in this data set is 0.69

**Model 2- Simple Linear Regression:**

Independent variable: log\_width

Dependent variable: log\_price

As show below X data set is taken with the column log\_size and y data set log\_price. These data sets are split into train and test data sets with 80% and 20% volume of original data set respectively.

Graphical user interface, text, application, email

Description automatically generated

Regression plot:

Chart, scatter chart

Description automatically generated

From the above plot it can be inferred that the line represents the predicted values of the price with the given width values. The dots on the plot represents actual values of the price. In this model there are lot points are away from the line which says that there is lot of difference between actual price and predicted price.

**Root mean square error of the model:**



The RMS error of this model is 2.32 which is way higher than the previous model.

**Model 3: Multivariate Linear Regression:**

Independent variable: log\_width, log\_Height, log\_size, House, signed, picture

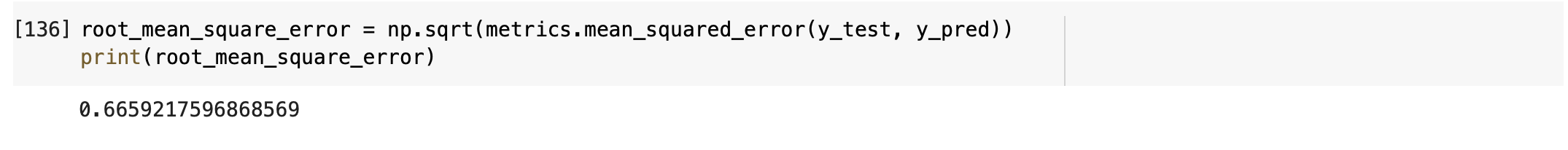
Dependent variable: log\_price

As show below X data set is taken with the column log\_size and y data set log\_price. These data sets are split into train and test data sets with 80% and 20% volume of original data set respectively.

Graphical user interface, text, application

Description automatically generated

**Root Mean square error of the model:**



The rms error of this model is 0.67.

**Conclusion:**

From the all the rms error values of all the three models it can be inferred that model 3 has less error. This means the painting price can be predicted with all the given features of the painting with more accuracy rather than a single feature.