Assignment 1 - Classification and Regression Methods

Deepak Raj Mohan Raj Zeming Zhang

2023-02-22

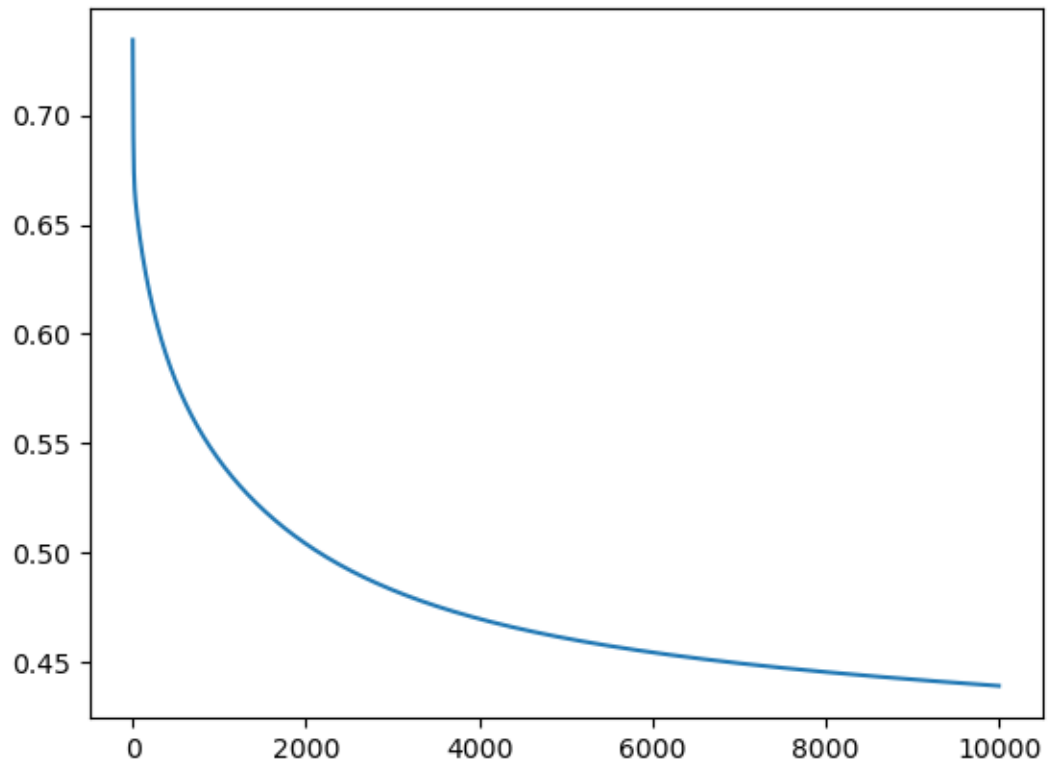
Team Member	Assignment Part	Contribution (%)
dmohanra	Part I, Part II, Part III, Bonus	50%
zemingzh	Part I, Part II, Part III, Bonus	50%

Part I: Logistic Regression

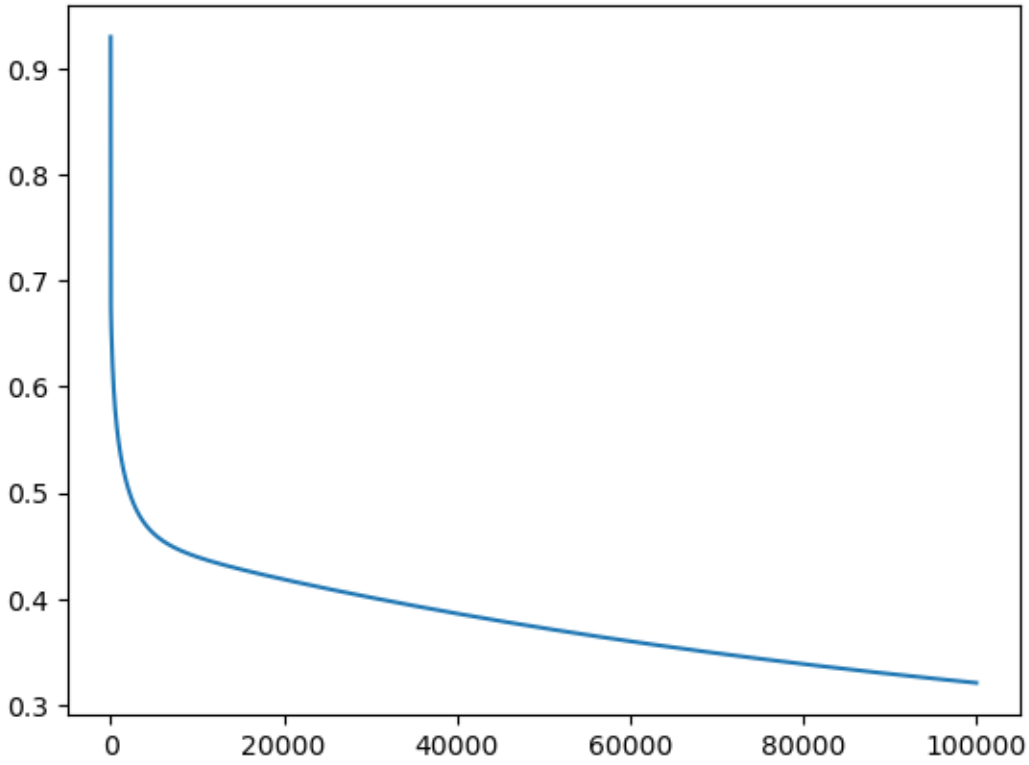
`Best Accuracy` - The best accuracy was found for the hyperparamaters: Learning rate = 0.001 and Iterations = 100000 with the accuracy of around 92.53

`Loss` - The following are the loss graphs for three different set of hyperparameters:

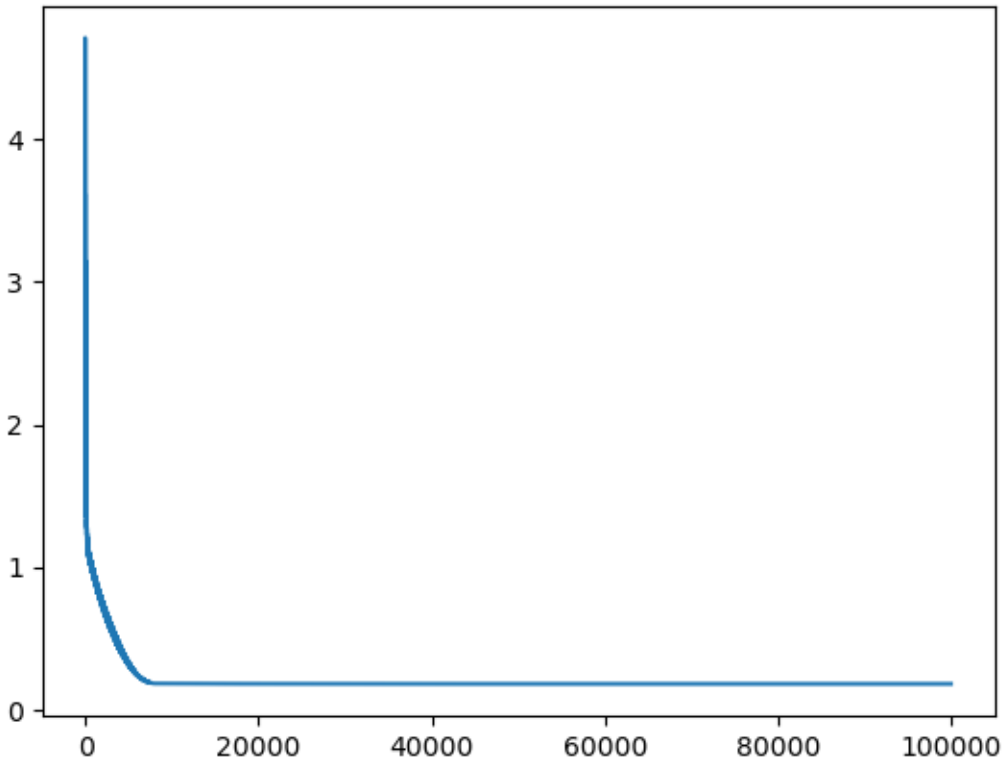
1. `Learning rate = 0.001 and Iterations = 10,000`



2. Learning rate = 0.001 and Iterations = 100,000



3. Learning rate = 0.1 and Iterations = 100,000



On analyzing the above three graphs, we can see that as the learning rate increases the loss drops down drastically. This is also followed by the fact that the number of iterations also increases, thus making the grap loos steeper.

Accuracy - As the number of iterations increases, the accuracy of the model also increases. For the above three trials, we have the following accuracy,

1. 83.58

2. 85.07

3. 92.53

In aaddition to the increasing number of iterations, the accuracy also increases along with increase in learning rate.

Benefits/drawbacks of using a Logistic Regression model: Benefits:

1. Logistic Regression provides easily interpretable coefficients that can help understand the impact of each predictor variable on the outcome.
2. Logistic regression models are computationally efficient and can be trained quickly, making them a good choice for large-scale applications.

Drawbacks:

1. Logistic regression models are only capable of performing binary classification, which means they cannot be used for multi-class classification problems without modifications. For example, inorder to make the species as target we need to make modification.
2. Logistic regression models require carefully selected and engineered input features to perform well. This can be time-consuming and may require domain expertise.

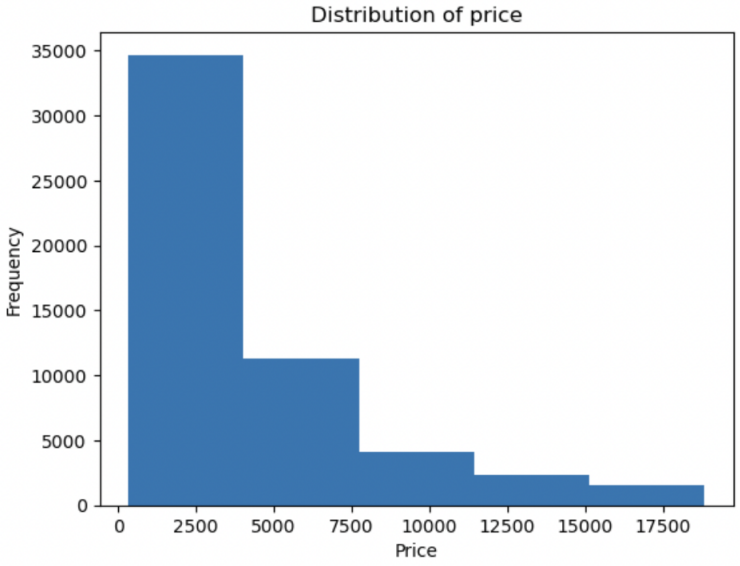
Part II: Linear Regression

1. The diamond dataset contains information about the physical characteristics of diamonds, such as their carat weight, cut, color, and clarity, as well as their dimensions (length, width, and depth) and price. This is a structured dataset with numerical and categorical variables. There are 53,940 entries and 10 variables.
2. The main statistics about the entries of the dataset is as follows:

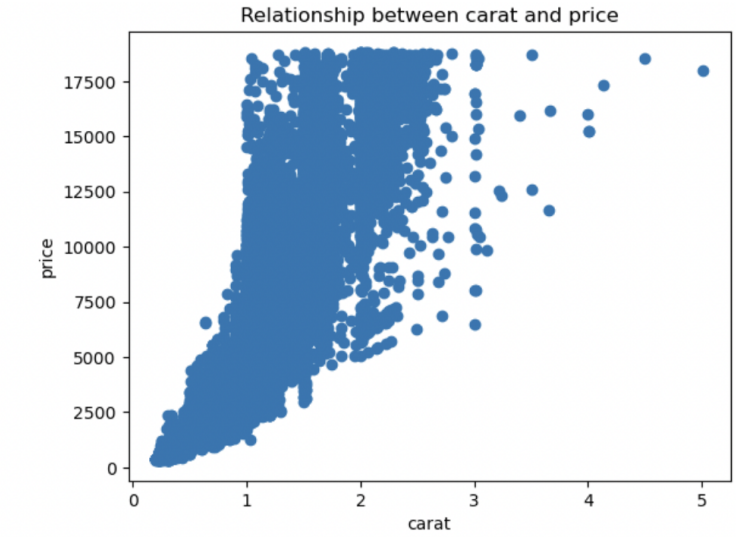
	Unnamed: 0	carat	depth	table	price
count	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000
mean	26970.500000	0.797940	61.749405	57.457184	3932.799722
std	15571.281097	0.474011	1.432621	2.234491	3989.439738
min	1.000000	0.200000	43.000000	43.000000	326.000000
25%	13485.750000	0.400000	61.000000	56.000000	950.000000
50%	26970.500000	0.700000	61.800000	57.000000	2401.000000
75%	40455.250000	1.040000	62.500000	59.000000	5324.250000
max	53940.000000	5.010000	79.000000	95.000000	18823.000000

	x	y	z
count	53940.000000	53940.000000	53940.000000
mean	5.731157	5.734526	3.538734
std	1.121761	1.142135	0.705699
min	0.000000	0.000000	0.000000
25%	4.710000	4.720000	2.910000
50%	5.700000	5.710000	3.530000
75%	6.540000	6.540000	4.040000
max	10.740000	58.900000	31.800000

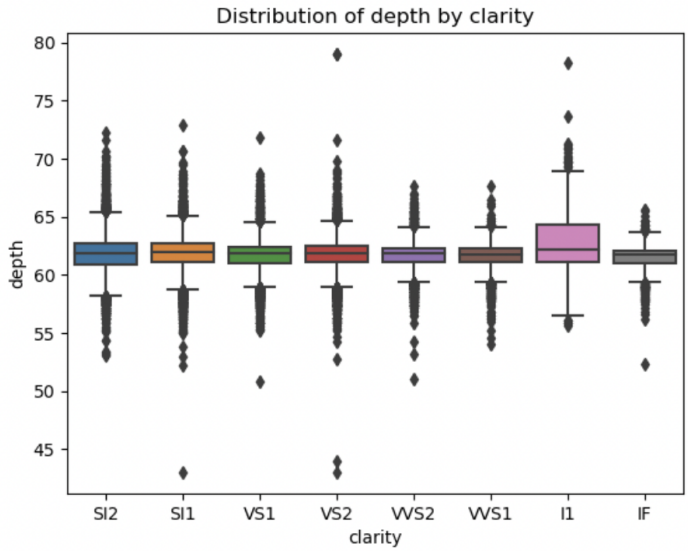
3. Five visualization graphs with a brief description for each graph:



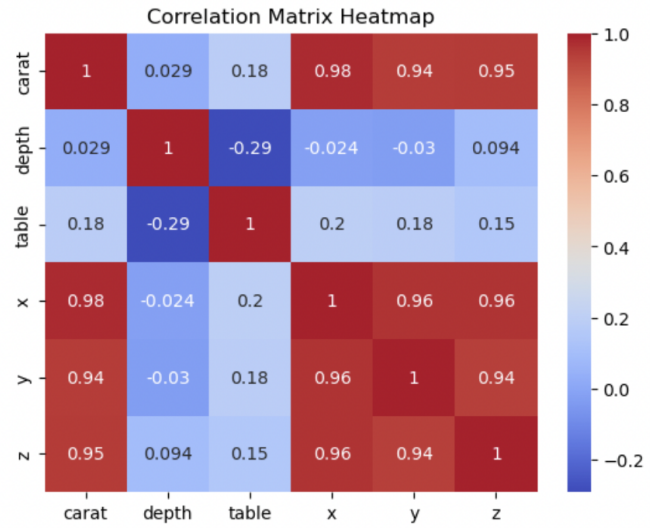
Based on the graph above, it can be observed that the target is primarily focused on values below 2500, and there are only a small number of diamonds with prices as high as 17500.



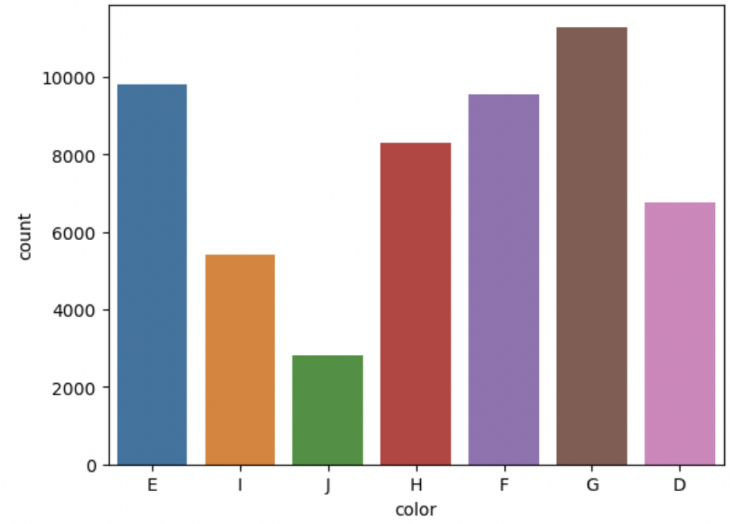
Based on the graph above, it can be observed that carat weight is predominantly concentrated below 2, and only a limited number of diamonds have a weight greater than 3, with prices that reflect their rarity. Furthermore, the graph indicates that diamond prices experience a sudden increase before reaching 1 carat in weight.



The graph above demonstrates that L1 clarity diamonds are distributed across a wide range of depths, and outliers in terms of depth are present for diamonds with VS2 and LSI clarity. Interestingly, the median depth for all clarity types is nearly identical, and there is no evidence of skewness in the clarity data.

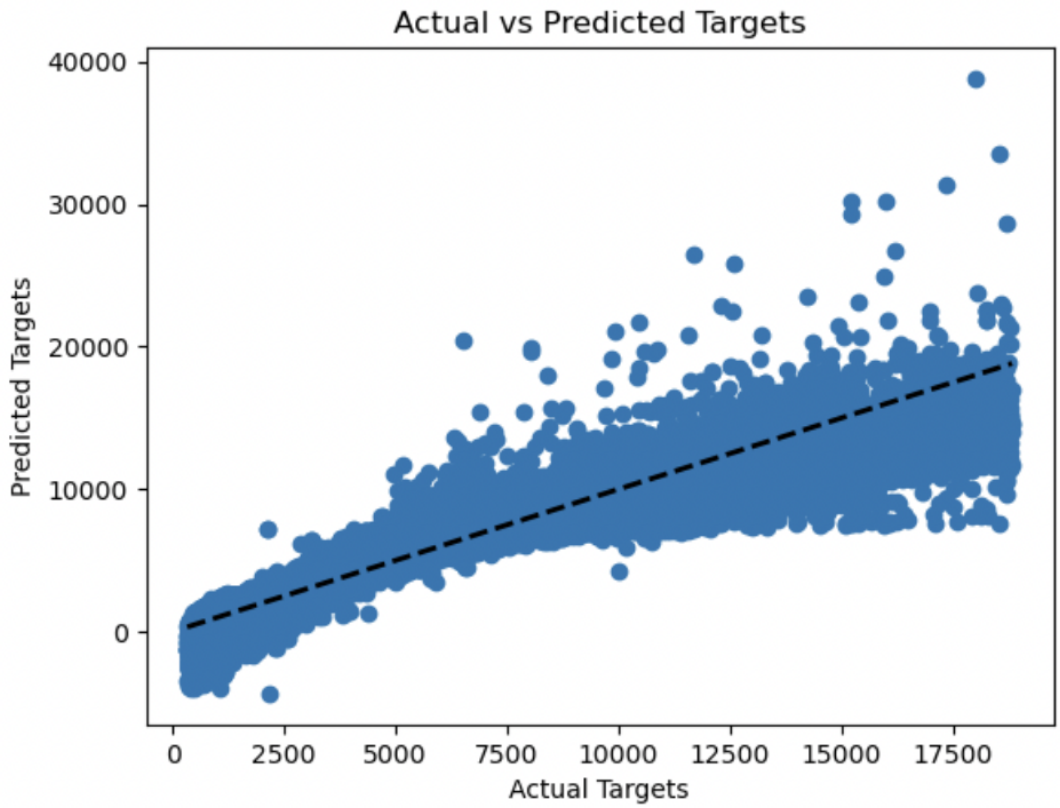


The graph above clearly indicates a strong correlation between the X, Y, and Z dimensions of the diamond. Moreover, it is noteworthy that these dimensions display a significant correlation with the carat weight, while exhibiting only weak correlations with depth and table measurements.



Upon observing the graph above, it is evident that the brown color of the diamond is the most frequently occurring color in the dataset. In contrast, the green color is the least commonly found color. This implies that brown diamonds are more prevalent and widely available, whereas green diamonds are much rarer and less commonly found

4. The loss value is 1318453.2574061984
5. The plot comparing the predictions vs the actual test data



6. Benefits and drawbacks of using OLS estimate for computing the weights:

- Benefits:
  - The estimated results are simple to understand and interpret as they provide coefficients that represent the relationship between the predictor variables and the response variable.
  - Computationally efficient for small to medium-sized datasets because it has closed-form solutions and can be easily computed using matrix algebra.
  - Solutions can be directly calculated without the need for iterative algorithms.
- Drawbacks:
  - OLS assumes that the errors (the differences between the predicted and actual values) have a normal distribution with a constant variance. This assumption may not hold true in some cases, which can affect the accuracy of the model's predictions.
  - Assumes that there are no influential outliers in the data, which can significantly affect the model's predictions. If influential outliers do exist, the model may not accurately represent the relationship between the predictor variables and the response variable.

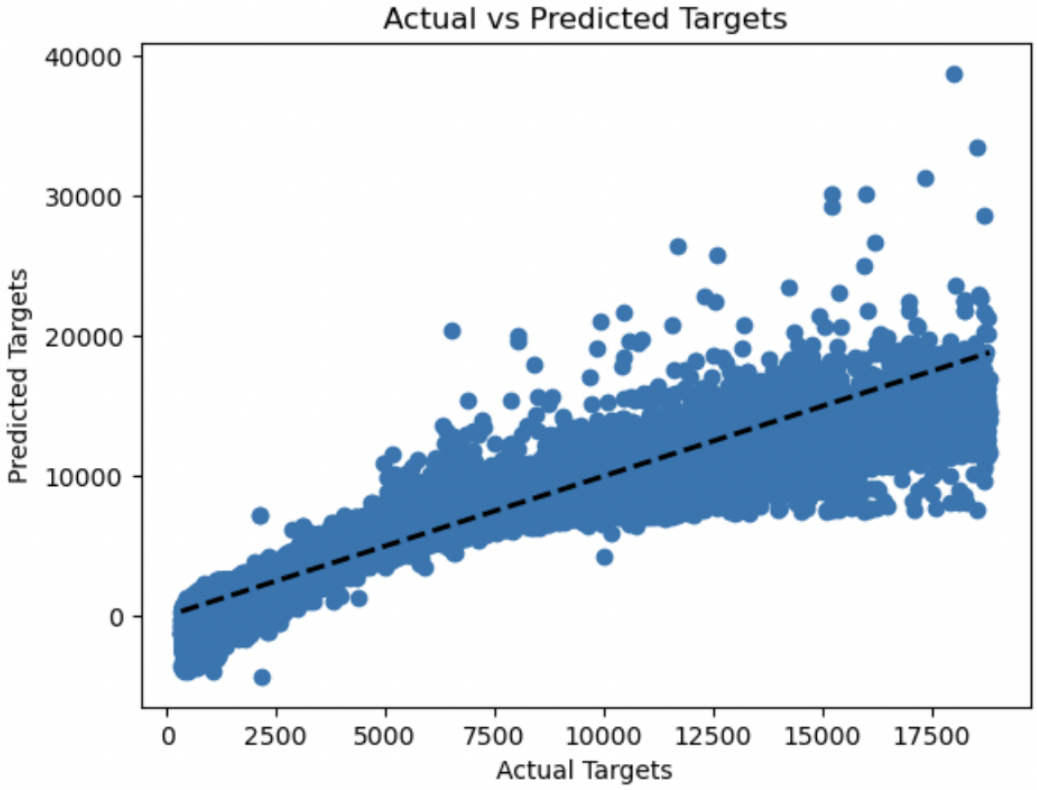
7. Benefits and drawbacks of using a Linear Regression model:

- Benefits:
  - Easy to comprehend: Linear regression is a simple and intuitive model that is easy to understand and interpret.
  - Can capture linear relationships between predictor variables and the response variable
- Drawbacks:
  - Linear regression assumes that the relationship between the predictor variables and the response variable is linear. This assumption may not hold true in some cases, which can affect the accuracy of the model's predictions.
  - It may not perform well when there are non-linear relationships or interactions between predictor variables, as it cannot capture these complex relationships.
  - It assumes that the errors (the differences between the predicted and actual values) have a normal

distribution with a constant variance. This assumption may not hold true in some cases, which can affect the accuracy of the model's predictions.
  - Also assumes that there are no influential outliers in the data, which can significantly affect the model's predictions. If influential outliers do exist, the model may not accurately represent the relationship between the predictor variables and the response variable.

Part III: Ridge Regression

1. Loss value is 35863807174.385956
2. The plot comparing the predictions vs the actual test data



3. The primary difference between linear regression and ridge regression is that linear regression minimizes the sum of squared errors between the predicted and actual values, while ridge regression minimizes the sum of squared errors plus a penalty term proportional to the square of the coefficients' magnitude. The regularization term is a squared L2 norm of the coefficients, multiplied by a regularization parameter alpha. The main purpose of adding L2 regularization is to prevent overfitting and improve the model's generalization performance.
4. The benefits/drawbacks of using a Ridge Regression model.

- Benefits:
  - Ridge regression is a regularization method that can improve the generalization performance of the model by preventing overfitting.
  - It can handle multicollinearity, a situation where the independent variables are highly correlated with each other.
  - Ridge regression can be used to select the most important features in the dataset by setting some coefficients to zero.
- Drawbacks:

  - The choice of the regularization parameter alpha can be challenging and may require cross-validation to find the optimal value.
  - The interpretation of the coefficients in ridge regression is less straightforward than in linear regression, as the coefficients are shrunk towards zero.
  - Ridge regression assumes that the relationship between the independent and dependent variables is linear, which may not always be the case.

References

- Linear Regression:
  - Andrew Ng's Machine Learning course on Coursera:
    - https://www.coursera.org/learn/machine-learning
  - Khan Academy's Linear Regression tutorial:
    - https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/more-on-regression/v/regression-line-example
- Ridge Regression:
  - Ridge Regression explained by StatQuest:
    - https://www.youtube.com/watch?v=Q81RR3yKn30
  - Ridge Regression tutorial by towards data science:
    - https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db
- Elastic Net Regression:
  - Elastic Net Regression explained by StatQuest:
    - https://www.youtube.com/watch?v=1dKRdX9bfIo