Jerome Veix
Intro to Machine Learning
Assignment 3
2/28/2024

The model I will investigate consists of x-inputs of fare, tip, and surge applied, while the outcome variable is the distance traveled column, which are all elements from the 'taxi_fare' file. After initially loading the data on Google Collab, I assigned the x-inputs from the data frame into a variable and did the same for the y-output. Then, I promptly fitted the data into a linear regression model for the slope and intercept of the model. The slope of -0.27261272 represents the negative value of distance traveled to the input variables of fare, tip, and surge applied when they are zero. Additionally, the slope of the regression line for fare and tip, resulting in scores of 0.05131761 and 0.01819076, respectively, indicates the minimal positive change of distance traveled when these x-inputs change. For instance, if there were no change in all other independent variables, a 1-unit increase in tip would lead to an insignificant increase in distance traveled by 0.01819076 units. As for the surge applied variable, the negative slope of -0.07996727 shows that when one of the variables increases, the other variable will decrease by that measly measure.

Moreover, with the same three input variables and a single output, we split the original data into training and testing data to find how well the model is forming a prediction. After fitting the training datasets into the linear regression model, the model based on the training data produced the coefficient of determination, also known as the R-squared score, of 0.52. The outcome indicates that the model does an average job of predicting the outcome with the result of 0.52. Also, the coefficient of determination for the testing data was a whopping 0.0004, a score that implies the model's complete lack of ability to determine the variability of the variables explained by the regression model.

For an analysis of the model, we used the evaluation metrics of regression, which are Mean Absolute Error (MAE), Mean Squared Error (MSE), Square Root of Mean Squared Error (RMSE), and R-squared. The MAE measures the absolute difference between the predicted outcome and the actual results, and the RMSE evaluates the average variance of errors in the model and how far the predicted values are from the produced outcome. RMSE and MAE are both metrics that are negatively oriented scores, meaning the lower the score, the better the model. Our MAE score is 2.072, and our RMSE score is 228.38. Although the RMSE score is considerably high, representing a pronounced average difference between predicted and actual values, the low MAE score signifies the low difference between the prediction and truth. The MSE assesses the amount of error in a statistical model, hence why the lower the score, the better it is. Our model yields an immensely high result of 52.159, which shows the substantial magnitude of error in the model. The measure of the variability of the model, known as the R-squared square, produced an exceedingly low score of 0.004 that shows the model predicts about zero percent of the relationship between fare, tip, and surge applied with distance traveled.

Ultimately, the various slopes, coefficient of determinant, and multiple evaluation metrics depicting the linear regression model between fare, tip, and surge applied with distance traveled show how remarkably poor the model performs. Despite the favorable MAE score, no multicollinearity, and an okay R-squared score predicting the outcome, the overwhelming analysis of the model designates an underperforming model with a high likelihood of error and weak relationships between the independent and dependent variables. The model does not effectively predict values close enough to the expected outcomes that warrant a positive diagnosis of its performance.