Name: Akshata Mulye

Subjective Questions [20 marks]

Answer the following questions only in the notebook. Include the visualisations/methodologies/insights/outcomes from all the above steps in your report.

Subjective Questions based on Assignment

Question 1. [2 marks]

Are there any categorical variables in the data? From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

Question 2. [1 marks]

What does test_size = 0.2 refer to during splitting the data into training and test sets?

Answer:

Question 3. [1 marks]

Looking at the heatmap, which one has the highest correlation with the target variable?

Answer:

from Heat map we see that there is no relation between delivey time and other featurs

Question 4. [2 marks]

What was your approach to detect the outliers? How did you address them?

Answer:

Question 5. [2 marks]

Based on the final model, which are the top 3 features significantly affecting the delivery time?

Answer:

General Subjective Questions

Question 6. [3 marks]

Explain the linear regression algorithm in detail

Answer:

Linear regression is one of the simplest and most widely used algorithms in statistics and machine learning. It models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. Equation of a Line: In simple linear regression (one independent variable), the relationship is represented by the equation: Y = b0 + b1X where: Y is the dependent variable (predicted output). X is the independent variable (input feature). b0 is the y-intercept (the value of Y when X is 0). b1 is the slope of the line (how much Y changes for each unit change in X). For multiple linear regression (multiple independent variables), the equation extends to: Y = b0 + b1X1 + b2X2 + ... + bnXn

Question 7. [2 marks]

Explain the difference between simple linear regression and multiple linear regression

Answer:

Simple linear regression and multiple linear regression are both statistical methods used to model the relationship between variables. The key difference lies in the number of independent (predictor) variables used. Simple linear regression uses only one independent variable, while multiple linear regression uses two or more.

Question 8. [2 marks]

What is the role of the cost function in linear regression, and how is it minimized?

Answer:

The cost function measures how well the model's predictions match the actual data and guides the optimization of parameters to minimize errors and find the best fit.

Question 9. [2 marks]

Explain the difference between overfitting and underfitting.

Answer:

Overfitting: Training error is low, but testing error is significantly higher.overfit models experience high variance—they give accurate results for the training set but not for the test set. More model training results in less bias but variance can increase

Underfitting: Errors are consistently high across training and testing data sets. Underfit models experience high bias—they give inaccurate results for both the training data and test set.

Question 10. [3 marks]

How do residual plots help in diagnosing a linear regression model?

Answer:

Residual plots are crucial for diagnosing linear regression models as they visually assess the model's assumptions about the data. By plotting residuals (the difference between observed and predicted values) against predicted values or independent variables, these plots help identify if the model is a good fit and if its assumptions are met.