Deep Salunkhe
21102A0014

Assignment : 3

D D M M Y Y Y Y

Q1] You are a data Scientist at a real estate company your task is to predict house prices based on various features such as square footage, number of bedrooms and location. Using linear regression, build a model to make these prediction. Disuss the steps you would address them.
Objective : Apply linear regression to a real-world problem and discuss data preparation and validation techniques

⇒

Data Preparation.

1. Data Collection : Gather a comprehensive dataset containing house prices squar footage, number of bedrooms, and location information

2. Data cleaning :

- Handle missing values: Impute missing values using techniques like mean, median mod.
- Outlier detection : identify and handle outlier using statistical method or visualization
- Feature Engineering : Create new feature if necessary

3. Data exploration:

- Descriptive statistics : Calculate summary statistics for numerical featur to understand distribution.

Visualization: Create histogram, scatter plots, and correlation matrices to explore relationships.

3. Feature Selection

1. Feature Importance: Use techniques like correlation analysis, feature importance or statistical tests to identify the most relevant features

2. Dimensionality reduction: if dealing with a large number of features, consider techniques like Principal component analysis to reduce dimensionality

Model Building.

linear regression: Create a linear regression model using selected features.

Model Training: Fit the model to the training dataset

Model Evaluation

Splitting Data: Divide the dataset into training and testing set to evaluate model performance on unseen data.

Model Evaluation Metrics: Use metrics like Mean squared Error (MSE), Root mean squared Error (RMSE), Mean absolute Error (MAE), and R-squared.

Cross-validation: Employ cross-validation to get a more robust estimate of model performance.

Potential issues and solutions

- Non-linearity: If the relationship between features and the target variable is non linear, consider transformation, using non-linear models

- Overfitting: If the model performs well on the training data but poorly on the testing data, consider regularization, feature selection or increasing the dataset size.

Underfitting: If the model performs poorly on both training and testing data, consider adding more features, increasing model complexity or improving data quality