

Predictive Modelling

Name: Aditya Gupte

Institute: IIT Gandhinagar

Date: May 10, 2024

Objective

The objective of this project is to develop a predictive model using linear regression to predict the target variable based on multiple features.

Data Collection and Preprocessing

The project started with data collection. The dataset used in this project contains information about various housing features, including numerical and categorical variables. The target variable in this dataset is the pricing of the house.

Data preprocessing was performed to handle missing values, encode categorical variables (one-hot encoding), and split the dataset into features and target variable.

Feature Engineering

Feature engineering was performed to transform the features and make them suitable for modeling. This included creating polynomial features and one-hot encoding categorical variables. Polynomial featuring was carried out to determine the accuracy of the prediction of a polynomial regression model.

Model Training

The dataset was split into training and testing sets using the `train_test_split` function from `scikit-learn`. A linear regression model was created using `scikit-learn`'s `LinearRegression` class. The model was trained on the training data using the `fit` method. Similarly, after creating the polynomial features, a polynomial regression model was also trained on various degrees.

Model Evaluation

Predictions were made on the test data using the trained model.

The performance of the model was evaluated using metrics such as mean squared error (MSE) and R^2 score. These metrics provide insights into how well the model is performing in terms of prediction accuracy and variance explained.

Results Visualization

The results of the model were visualized using scatter plots. One plot showed the actual values against the predicted values, providing a visual representation of the model's performance.

Reasoning: Linear regression was chosen as the modeling technique due to its simplicity, interpretability, and ease of implementation. It is well-suited for this project as it allows us to predict a continuous target variable based on multiple features. Additionally, the use of one-hot encoding ensures that categorical variables are properly handled in the modeling process.

Conclusion

The predictive modeling project using linear regression was successfully implemented. The model was trained and evaluated on the dataset, and its performance was assessed using appropriate metrics. The results indicate that the model is able to make reasonably accurate predictions, as evidenced by the low mean squared error and high R^2 score. Overall, the project demonstrates the effectiveness of linear regression for predictive modeling tasks.