DSC412-project_Tharun Mandadi

Title: Iowa House Price Prediction

Summary: The objective of this project is to develop and evaluate predictive models for estimating house prices in Iowa, using a dataset that encompasses a wide range of property attributes.

1. Data Collection

Gather hosing price data for a large number of assets. Data set is uploaded in the current repository

2. Data Preprocessing

The dataset consists of multiple columns, each representing different attributes related to house prices, with close to 3,000 data points in total. To assess the uniqueness of the data, the number of unique values was calculated for each column, and each column will be checked for missing (NaN) values and duplicates.

3. Feature Selection

Using the text Machine Learning in Business by John Hull as a reference, features will be selected from the dataset to build the model. All features will be transformed to numerical form, to ensure the data could be used to develop our regression model.

4. Splitting Data and Scaling

To build and find the best model, the dataset will be split into three subsets: training, validation, and test sets. This allowed for proper evaluation of model performance during training and final testing. After splitting the data, feature scaling will be applied to the features and the target variable Sale Price.

5. Model Building

To begin we first fit a linear regression model to capture the relationship between Sale Price and our selected features. However, a linear regression model without regularization may yield suboptimal results due to multicollinearity among features. To simplify the model we will employ the regularization techniques Ridge and Lasso regression. By evaluating these methods with different values of λ, we aim to identify the best-performing model.

6. Model Evaluation

We further evaluate the model with the test data set.