Project 2: Ames Housing Data and Kaggle Challenge

Sophia Joseph

Problem Statement

Given the ongoing inflationary trends affecting our current economy, it becomes increasingly important to seek out discounted prices, particularly when it comes to housing. Therefore, the primary goal of this project is to assess and compare the r2 scores of various models designed to predict house sale prices using different sets of variables.

What is the r² Score?

- Statistical measure
- Finds the proportion of variance
- Can be read by humans
- **Shows how well your data fits your regression model**

Steps

01

Data Cleaning

How each dataset was cleaned

02

Create Model 1

First trial using LinearRegression

03

Create Model 2

Second trial using LinearRegression with PolynomialFeatures

04

Create Model 3

Third trial using LinearRegression and LassoCV

05

Compare r^2

Comparing all the r^2 scores

06

Conclusion

Conclusions and recommendations

Model 1

- Simple data cleaning was performed
  - Filling all the missing values with 0
- Simple LinearRegression model was performed
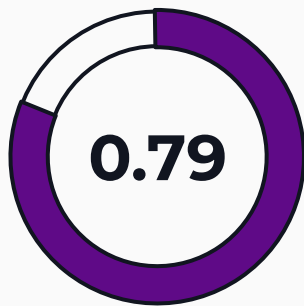- An r^2 score of 0.79 was produced

Model 2

- Data cleaning was completed
  - Custom function to apply the mode and mean
- Simple LinearRegression model was performed
- The data was scaled
- PolynomialFeatures was added on
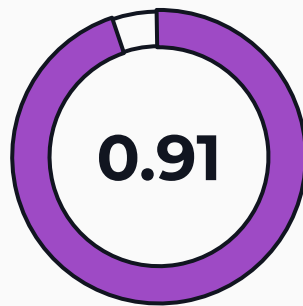- An r^2 score of 0.91 was produced

Model 3

- Data cleaning was completed
  - Custom function to apply the mode and mean
- Simple LinearRegression model was performed
- The data was scaled
- LassoCV was added on
- An r^2 score of 0.89 was produced

Comparing r² Scores

Model 1

Using
LinearRegression

Model 2

Using
LinearRegression with
PolynomialFeatures

Model 3

Using
LinearRegression with
LassoCV

Conclusion

- The model that gave me the best r^2 score was **Model 2**
- Train model score of 0.91
- Test model score of 0.90

Recommendations

- For future analysis:
    - Look at different features of X
    - Use different models
    - Different cleaning techniques for the model

Sources

- r^2 score:
  - https://corporatefinanceinstitute.com/resources/data-science/r-squared/

Thanks!