Ames Housing Data and Kaggle Challenge

GA SG DSI 26: Project 2

Introduction

- The dataset provided is based on housing information in Ames, which is a city in Story County, Iowa, United States.
- It is the home of Iowa State University, with leading agriculture, design, engineering, and veterinary medicine colleges.
- The Ames Housing Dataset is an exceptionally detailed and robust dataset with over 70 columns of different features relating to houses.

Problem Statement

- A home owner in Ames exploring to sell my current house and want to find out which features I should improve or renovate to increase the house value/ sale price before selling my house.
- This project aims to answer the problem statement and attempt to predict the features that will have a large impact on the sale price of the house.

Executive Summary

- The best model to use is Ridge model.
- Positive features to sale price
  - Area features like ground living area, total basement square feet, garage area
  - Quality features like exterior material quality, kitchen quality and overall quality
- Negative features to sale price
  - Age of the house when sold, exterior materials of wood and vinyl and a rough garage finish

Methodology

1. Data Cleaning
2. EDA
3. Data Visualisation
4. Feature Engineering
5. Pre-processing
6. Model Preparation
7. Model Evaluation
8. Production Model
9. Kaggle Submission

Managing and preparing the data

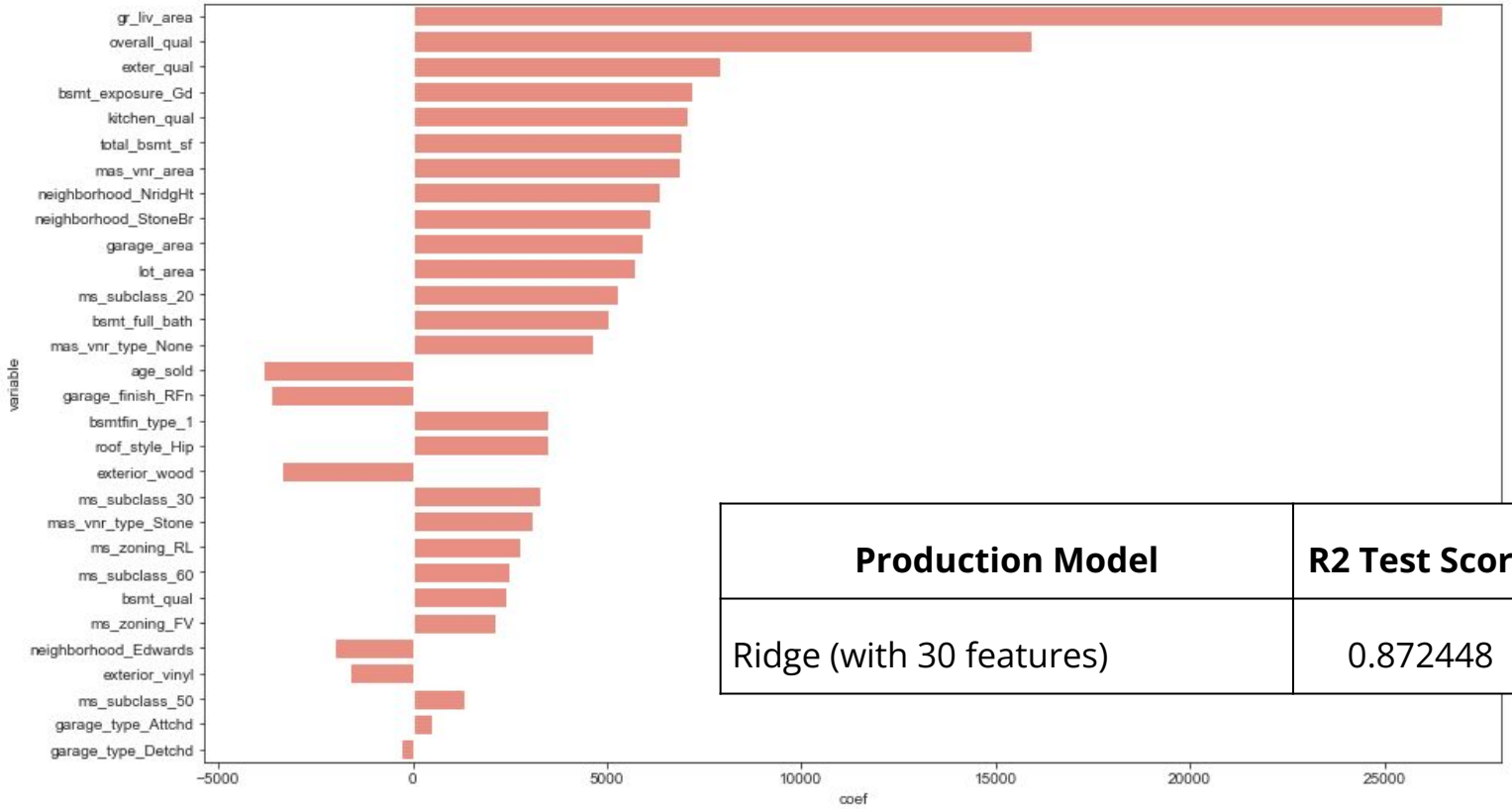
1. Deal with null values
2. Plot graphs (boxplot, histograms, scatterplots) for numerical and categorical data to identify trends and patterns. Set several criterias to assist in dropping of features
3. Explore correlation between the features to identify collinear features
4. Feature engineering to create, combine and change several features (porch, pool, age sold and remodelled columns)

Model Evaluation

Model Evaluation	R2 Train Score	R2 Test Score	RMSE
Dummy Regressor (Strategy: mean)	0.0	-0.00127	76412.97
Linear Regression	0.919402	0.886587	25717.03
Ridge (alpha = 19.1164)	0.918581	0.887734	25586.67
Lasso (alpha = 125.6524)	0.917511	0.887586	25603.47

- Tested with 4 models for model evaluation.
- Ridge is chosen as the best model as it has the lowest RMSE score
- Ridge only reduces the coefficients close to zero but does not carry out any feature selection.
- Allow for sorting of the Ridge coefficients and pick the top 30 coefficients to ensure an easy to interpret production model.

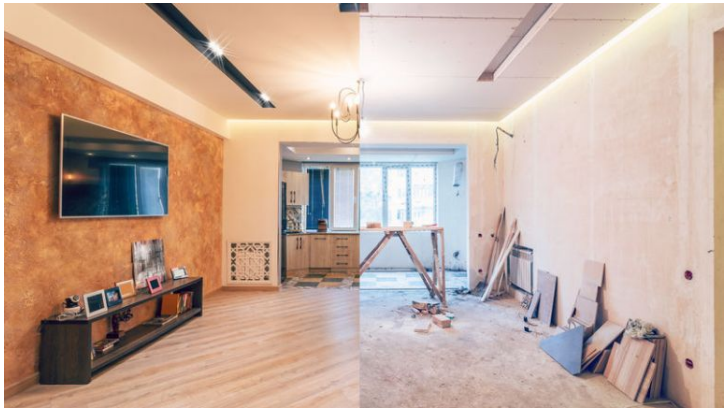
Production Model (Ridge Model Coefficients)



Production Model	R2 Test Score	RMSE
Ridge (with 30 features)	0.872448	27272.96

Recommendations

- Increase the liveable area of my house in the ground living, basement and garage space
- Improve the kitchen, exterior and overall quality of the house before selling the house.
- In addition, if the house contains any of the features that could affect the price of the house like a rough finish in the garage, these can be removed or improved upon to bring up the sale price.



The End

Prepared by: Kwek Jun Hong