House pricing Prediction

Using Ames Iowa Housing dataset

Carol Duan

Overview

- Problem Defining
- Data Exploring & Cleaning
- Feature Engineering & Selection
- Modeling & Evaluation
- Conclusion
- Next Step

Problem Defining

Task: Create a regression model based on the Ames Housing Dataset. This model will predict the price of a house at sale.

Target: House price in Ames, IA (sale_price is a continuous variable)

Problem Type: Regression & Predictions

My Workflow:

- **Inferences:** Get knowledge of the most important price-related features of the property that people care about when buying a house in Ames, IA
- **Predictions:** Predict the price of house at sale in Ames, IA based on the important features of the property

Data Exploring & Cleaning

- Check null values
- Check data type for each columns
- Split features and target
- Get dummies for categorical features (for training / testing / predicting data)
- Training / Testing data splitting (.75 / .25)

Data Exploring & Cleaning

Clean up the null values

Investigation:

- Check meaning & data type of each column to understand the meaning of null (real missing?)

Finding:

- **Numerical columns:** the null values should be 0
- **Categorical columns:** the null values are the same as "None" or "doesn't have", which can also be replace by "0" for further analysis

Solution:

- df.fillna(value=0,inplace=True)

Pool QC	2042
Misc Feature	1986
Alley	1911
Fence	1651
Fireplace Qu	1000
Lot Frontage	330
Garage Finish	114
Garage Cond	114
Garage Qual	114
Garage Yr Blt	114
Garage Type	113
Bsmt Exposure	58
BsmtFin Type 2	56
BsmtFin Type 1	55
Bsmt Cond	55
Bsmt Qual	55
Mas Vnr Type	22
Mas Vnr Area	22
Bsmt Half Bath	2
Bsmt Full Bath	2
Garage Cars	1
Garage Area	1
Bsmt Unf SF	1
BsmtFin SF 2	1
Total Bsmt SF	1
BsmtFin SF 1	1

Feature Engineering & Selection

- Manually drop unrelated ‘id’ and ‘pid’ columns
- Get polynomial features (**37400 total features**)
- Standardize the features
- Use Lasso regression model to help feature selections by eliminating some of the features (**368 total features, eliminated 90%**)

Modeling & Evaluation

A. Modeling

- Use Cross-validation to try Linear / Lasso / Ridge / Elastic Net regression model and see which one perform better on the training dataset
- Compared the cross-validation results, Ridge is the best (R^2 score = 0.92), but we still want to test it on the test dataset to make sure the model is not overfitting

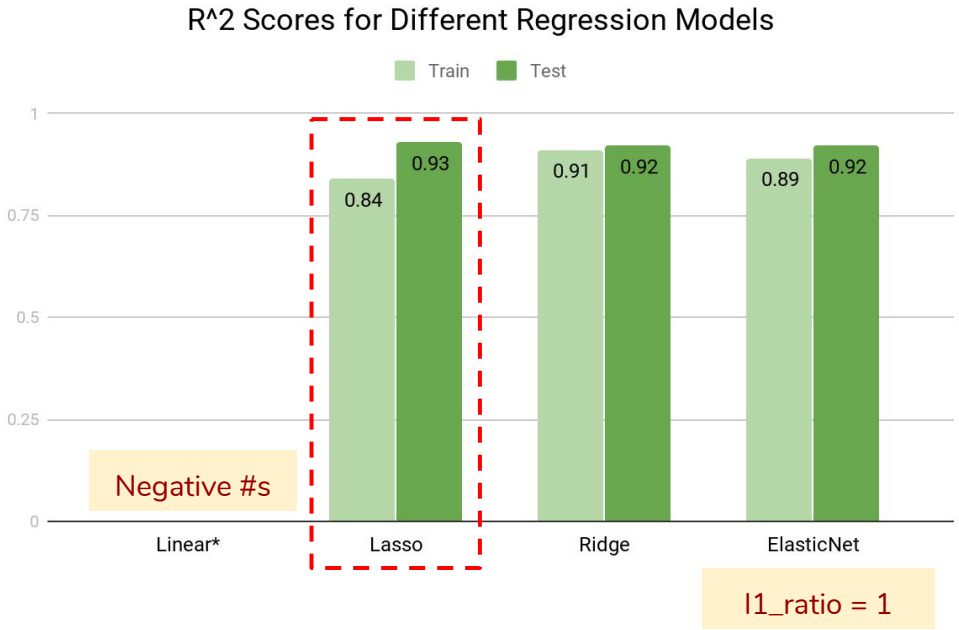
B. Evaluation

- Clean up & feature engineering on testing dataset - Same process to training data
- Fit the models using training data and use the testing dataset to score each models (Linear / Lasso / Ridge / Elastic Net regression) and see which one has the highest score
- Lasso got the highest score on testing dataset (R^2 score = 0.93), which is pretty good. So I decided to use this model to do prediction

C. Optimization

- Build the real model using the entire training + testing dataset I have

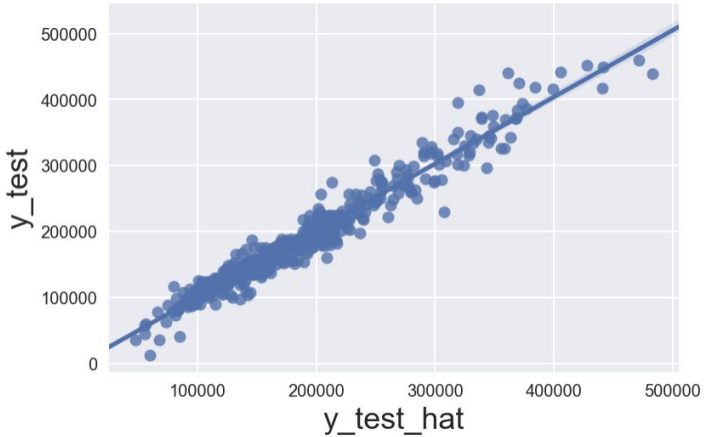
Modeling & Evaluation



Optimization:

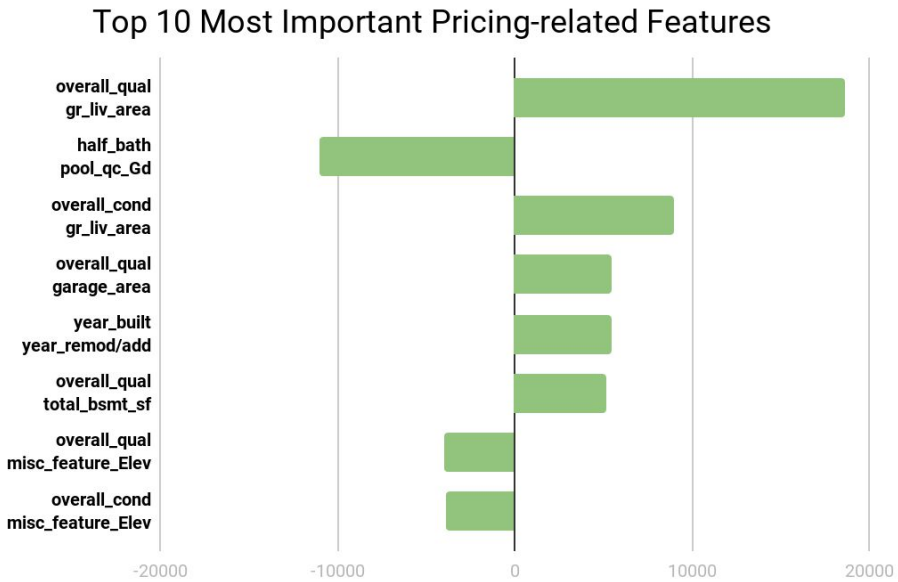
Fit Lasso regression model using entire dataset (Train + Test)

R^2 score: 0.95



Conclusion

Inferences:



Prediction:

Predict the price of house at sale in Ames, IA based on the important features of the property.

Model Type: Lasso Regression

RMSE: 36,258 (Prediction)

Model R^2: 0.95 (Test)

Kaggle Rank: 6

Next Step

- Do research and better understand the study (features)
- Find a better way to eliminate the unrelated features (still a big amount now)
- Continually tune and improve my model for prediction purpose
- Collect more data for testing
- Thoroughly Interpret my model and visualize it
- Give recommendation to support the house industry in Ames, IA

Thank You for Watching!