Identifying Optimal ML Model For Housing Market Price Prediction

By: Ji Woong Kim, Shashwat Mishra, Otto Goldschmidt

December 2, 2022

An Introduction

- Since Covid-19 the house price increases and, also the increase of the market price affects the house price. Because of this, the house price is very important factor for the people. Predicting of the house price based on the some factors is helpful for the people who aim to have budget. Our project proposes the house price predictors, as there is a famous used car price predictor(carmax).

- Dataset

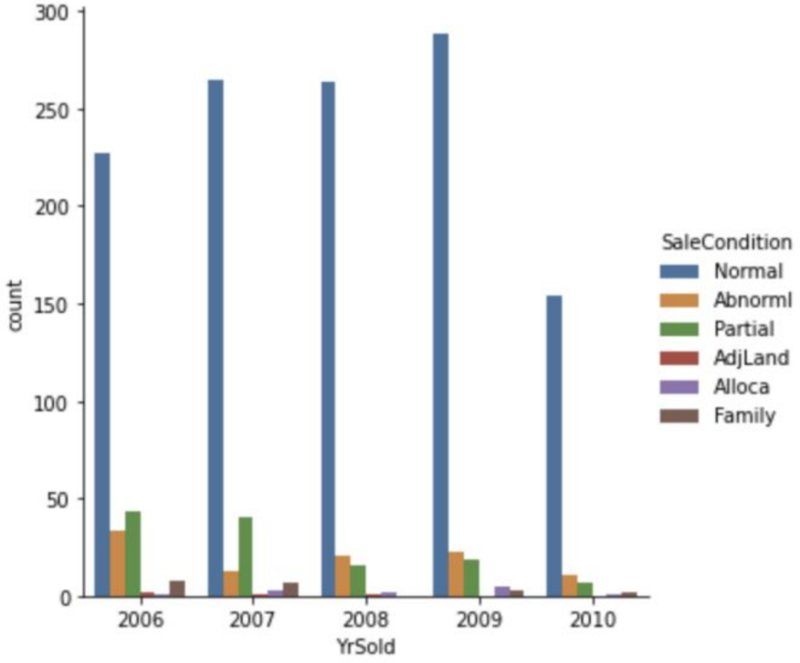
  House Sale Dataset in Iowa from Kaggle

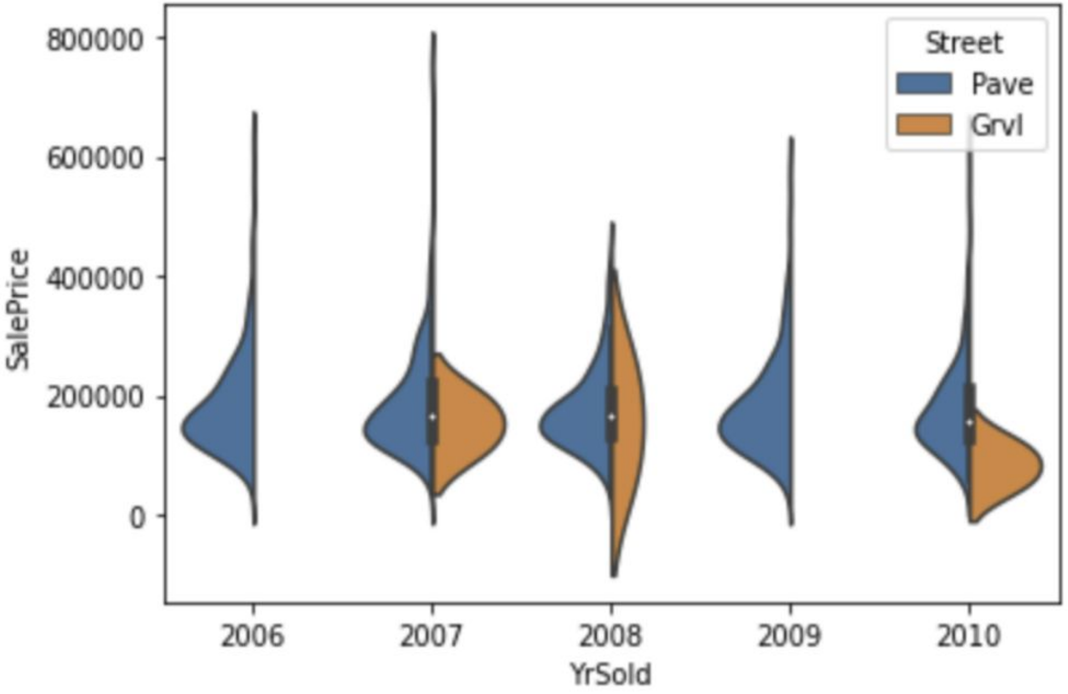
  Predictors : **79**

  Response : House Sale Price

  43 Categorical Variables, **38** Numerical Variables

General Dataset EDA



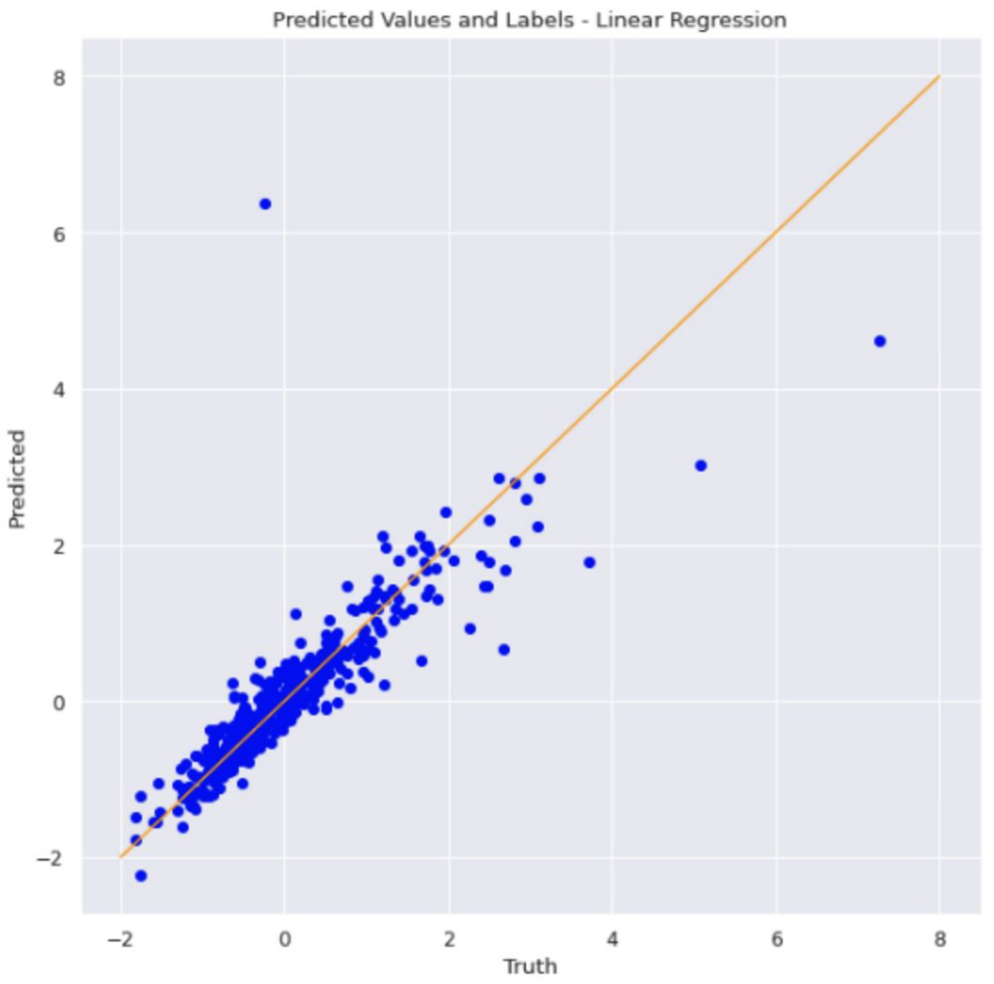


Information about type of Houses and their Roads.

Initial Test with Linear Regression

- After all of the data exploration and cleaning, we decided to perform a few beginner level tests to see how the models would look. We split the data, scaled it, and created predictions. Here is a scatterplot for the predicted Y and the Y test.

- The diagonal axis helps to visualize the difference between predicted values and truth. Points below orange diagonal line are the ones that truth points are greater than predicted points. Otherwise, the points that are located left from the orange axis means predicted values that are estimated lower than labels



Key Issues

- Since Covid-19 the house price increases and, also the increase of the market price affects the house price. For the reason of the house price is very important factor for the people, the prediction of the house price based on the some factors is helpful for the people who aim to have budget.

1. Housing has always been a big issue for people
2. House prices help experts predict recessions
3. Improvements in this area are always relevant
4. Housing Corporation are spending billions to analyze housing market data to capture the market.

Alternative Approaches to These Issues

- Different machine learning algorithms
- Housing Prices - Government Intervention
- Recession Prediction - Other factors such as income, inflation, political stability, etc.
- Pattern Detection- Detecting patterns

  of price change according to housing policies.

Related Work

- C. Chee Kin, Z. Arabee Bin Abdul Salam and K. Batcha Nowshath, "Machine Learning based House Price Prediction Model," 2022 International Conference on Edge Computing and Applications (ICECAA), 2022, pp. 1423-1426, doi: 10.1109/ICECAA55415.2022.9936336.

Prices here are estimated using a variety of technologies, including chatbots, artificial neural networks, and machine learning (ML). Here, the most efficient housing price was ascertained from the gathered dataset using all of the above mentioned methods.

- C. Zhan, Z. Wu, Y. Liu, Z. Xie and W. Chen, "Housing prices prediction with deep learning: an application for the real estate market in Taiwan," 2020 IEEE 18th International Conference on Industrial Informatics (INDIN), 2020, pp. 719-724, doi: 10.1109/INDIN45582.2020.9442244.

Most econometric or statistical models cannot capture non-linear relationships yet. Therefore, the paper proposed housing price prediction models based on deep learning methods, which can capture non-linear relationships.

Our Approach

- Algorithms which will be used to make predictions are:
  Linear Regression, Support Vector Regression, Decision Tree Regression, Ridge Regression and **LASSO**.

- Feature Engineering: Finding the top most correlated predictors with House Sale Price.

- Evaluation Metrics: **Root Mean Square Error** (RMSE). It is used to prevent the drastic increase of the MSE.

- Fine tuning the hyperparameter (K-Value) : Lasso Regression is the core model _ **Alpha** is fine tuned.

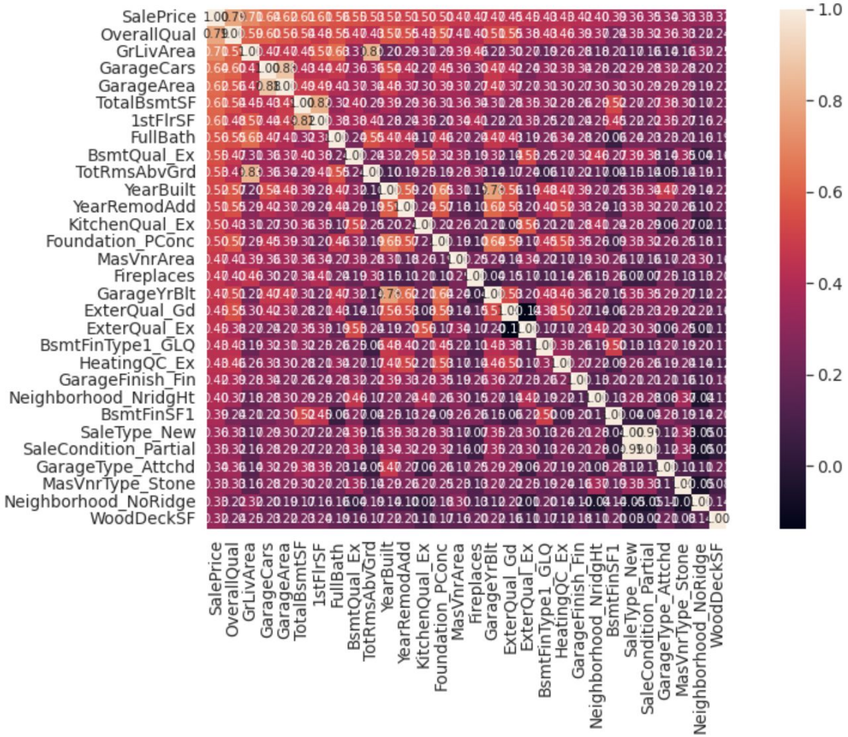
- Summarizing the findings and identifying the best machine learning model.

Feature Engineering

- Selecting optimal predictors from the dataset based on the correlation
  - Based on the response (Sale Price)
  - The correlation matrix describes the most correlated predictors with Response (Sale Price)

- Top 10 Most Correlated Predictors and their correlation scores

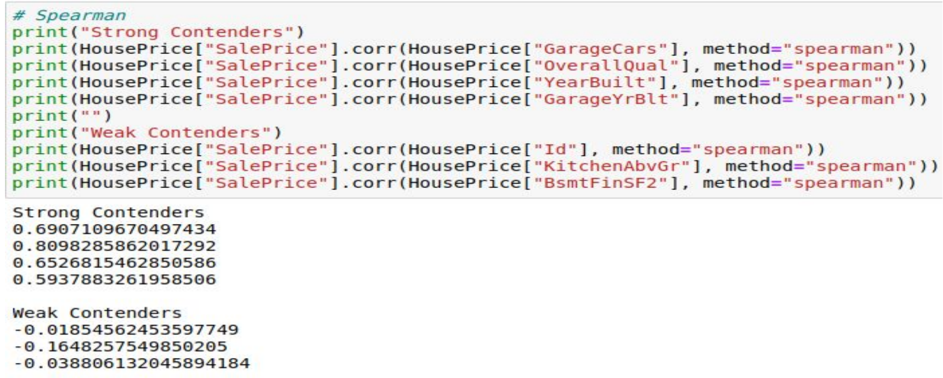
OverallQual	0.790982
GrLivArea	0.708624
GarageCars	0.640409
GarageArea	0.623431
TotalBsmtSF	0.613581
1stFlrSF	0.605852
FullBath	0.560664
BsmtQual_Ex	0.553105
TotRmsAbvGrd	0.533723
YearBuilt	0.522897



Feature Engineering

Kendall Tau & Spearman

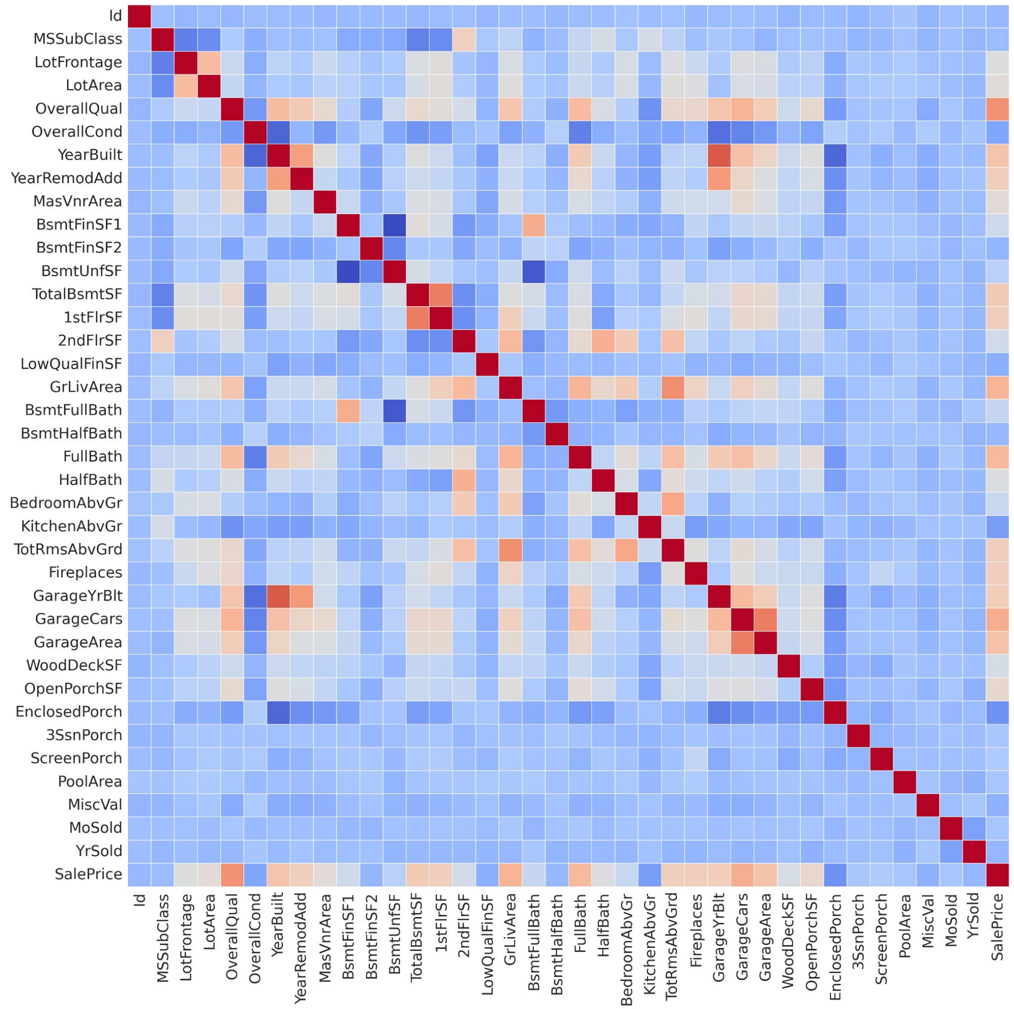
- Kendall’s tau is a measure of the correspondence between two rankings. Values close to 1 indicate strong agreement.
- Kendall Tau from scipy.stats library
- Additional feature selection
- Verifying feature importance
- Additional insights



```
# Spearman
print("Strong Contenders")
print(HousePrice["SalePrice"].corr(HousePrice["GarageCars"], method="spearman"))
print(HousePrice["SalePrice"].corr(HousePrice["OverallQual"], method="spearman"))
print(HousePrice["SalePrice"].corr(HousePrice["YearBuilt"], method="spearman"))
print(HousePrice["SalePrice"].corr(HousePrice["GarageYrBlt"], method="spearman"))
print("")
print("Weak Contenders")
print(HousePrice["SalePrice"].corr(HousePrice["Id"], method="spearman"))
print(HousePrice["SalePrice"].corr(HousePrice["KitchenAbvGr"], method="spearman"))
print(HousePrice["SalePrice"].corr(HousePrice["BsmtFinSF2"], method="spearman"))
```

```
Strong Contenders
0.6907109670497434
0.8098285862017292
0.6526815462850586
0.5937883261958506

Weak Contenders
-0.018545624535977749
-0.1648257549850205
-0.038806132045894184
```

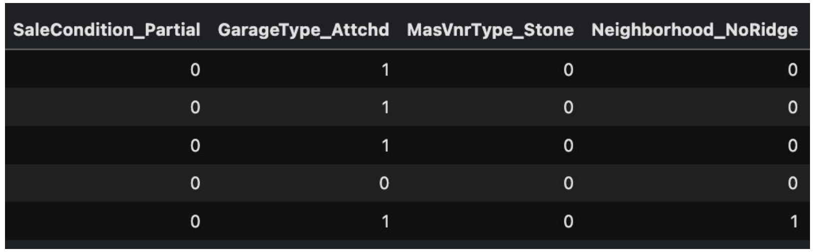


Feature Engineering

- Handling Null Values
    - Columns that have too many NaN values don’t have a significant impact on the training phrases. For this reason, the columns that have more than 100 rows with null values are truncated.

- Converting Categorical Values to Dummy variables using One-Hot Encoding
    - categorical values to the dummy variables using One-Hot Encoding. Converted columns have values 1 and 0 filled in it for true or false case but now can be used as a numeric variable.
        - ex) SaleCondition -> SaleCondition_Normal, SaleCondition_Abnormal, SaleCondition_Partial

Categorical Variables →



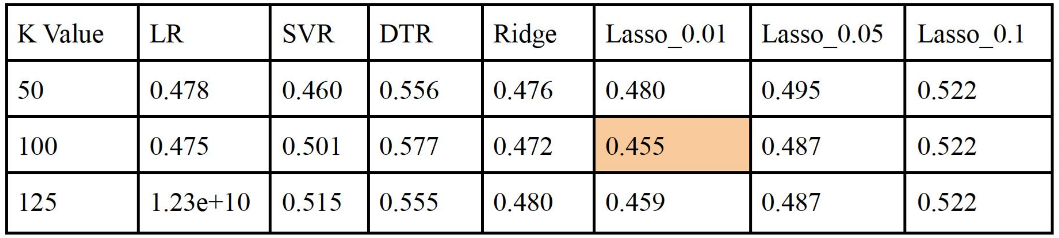
SaleCondition_Partial	GarageType_Attchd	MasVnrType_Stone	Neighborhood_NoRidge
0	1	0	0
0	1	0	0
0	1	0	0
0	0	0	0
0	1	0	1

- Filling Median Values to the Rows with Null Values
    - ○ Null values from the 'GarageYrBlt' and 'MasVnrArea' predictors are replaced with median values.

Fine Tuning

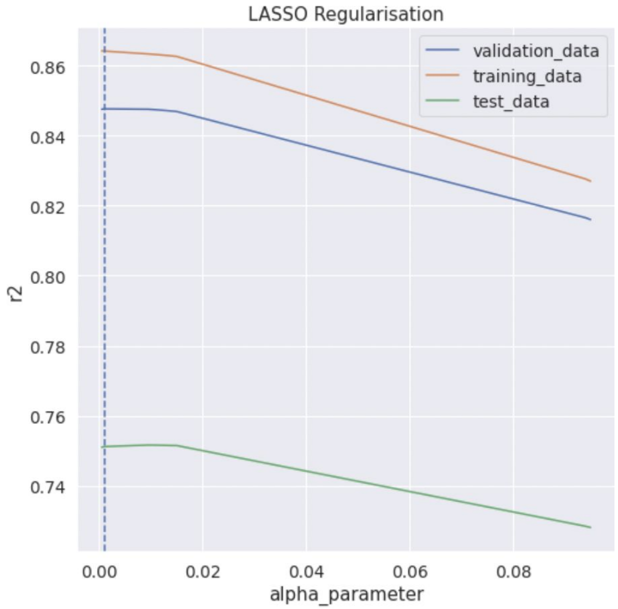
- Detecting the optimal Alpha value for the Lasso Regression
  - Three Alpha values were used to find the optimal alpha.
    0.01 / 0.05 / 0.1

RMSE



K Value	LR	SVR	DTR	Ridge	Lasso_0.01	Lasso_0.05	Lasso_0.1
50	0.478	0.460	0.556	0.476	0.480	0.495	0.522
100	0.475	0.501	0.577	0.472	0.455	0.487	0.522
125	1.23e+10	0.515	0.555	0.480	0.459	0.487	0.522

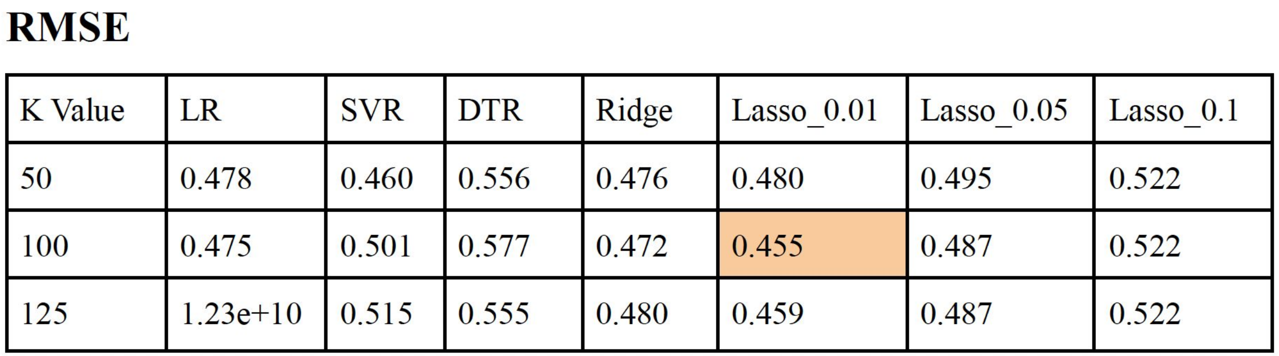
- Grid Search Cross Validation
  - Using the gridsearchcv provided by scikit-learn, even better fine-tuned alpha hyper-parameter is proposed.
    Alpha : 0.01 –> 0.096



Evaluation

- Comparison Algorithms
  - (Linear Regression, Support Vector Regression, Ridge Regression, Decision Tree Regression)

- K-Values: Top K-th most correlated predictors.
  - Three K-Values are used to find the optimal number of predictors.



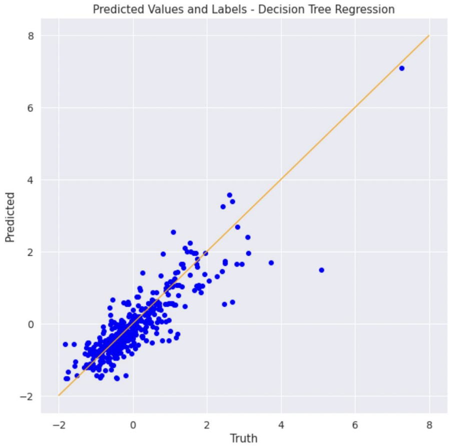
RMSE

K Value	LR	SVR	DTR	Ridge	Lasso_0.01	Lasso_0.05	Lasso_0.1
50	0.478	0.460	0.556	0.476	0.480	0.495	0.522
100	0.475	0.501	0.577	0.472	0.455	0.487	0.522
125	1.23e+10	0.515	0.555	0.480	0.459	0.487	0.522

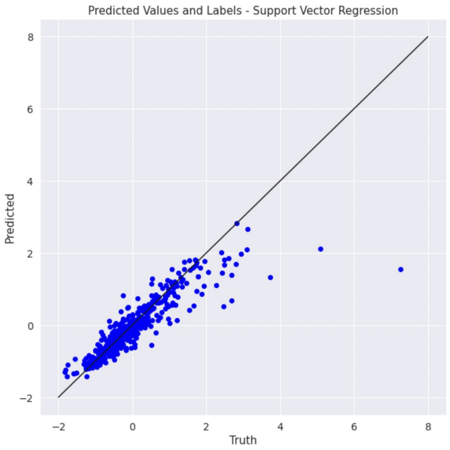
As we can see above, Lasso with alpha = 0.01 with 100 most correlated features gives the best model.

Evaluation

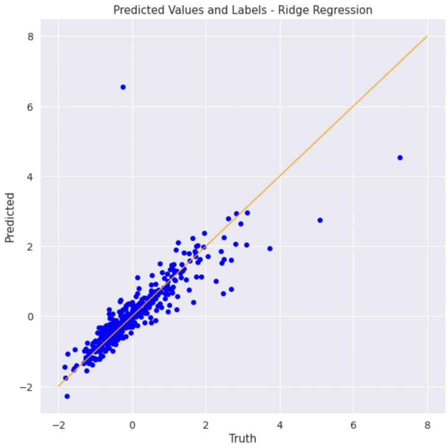
- Comparison Algorithms
  - Linear Regression, Support Vector Regression, Ridge Regression, Decision Tree



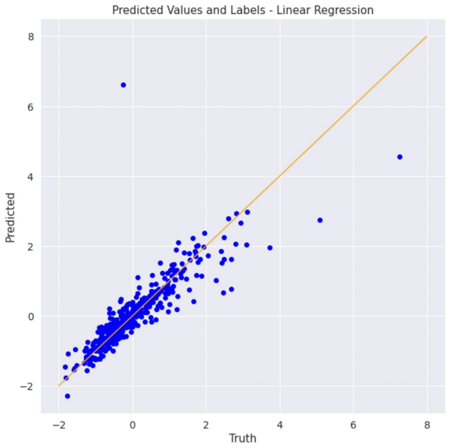
Decision Tree Regression



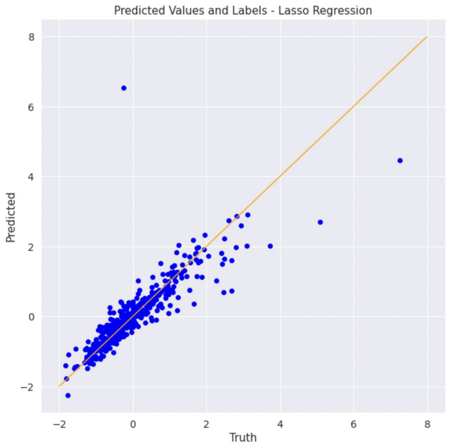
Support Vector Regression



Ridge Regression



Linear Regression



Lasso Regression

Key Contributions

- Identified the core predictors that can most affect the house price. And they could be potential significant predictors that can be applied other house price dataset.
- Developed the combination of the feature engineering that can optimize the house price prediction model.
- Identified the best model (of the tested ones) for the Iowa house price dataset - LASSO Regression.
- Explored how different models and hyper parameters create different prediction.

Future Work

- Only deals with one city. So, we can apply our models for multiple cities and compare their results for a better outcome.
- We can use training models from different cities against each other and test out the accuracy.
- We can analyze housing prices, model and accuracies with respect to the housing market rules and regulations in the particular city to find correlation between them.

Thank You