Analysis of Ames Housing Data

Building lowest possible error regression model

Background

- Sale price of residential property in Ames town, Iowa
  - 2006 to 2010 period
  - Compiled by Prof. Dean de Cock
  - 2051 training data
  - 879 test data

- 81 attributes
  - 79 quantity and quality attributes
  - 2 identification columns

Workflow

Data and problem statement identification	Exploration	Cleaning and analysis	Model building and evaluation	Findings
• Build regression model with least possible error • MSE as metric • Training and test data given	• Data size and formats • Understanding attributes • Distribution and correlation of attributes	• Various imputation methods for different missing data • Feature Engineering and removal	• Linear Regression models • Regularised Regression model (LASSO/Ridge)	• Identify best model • Identify key attributes affecting price

Problem statement

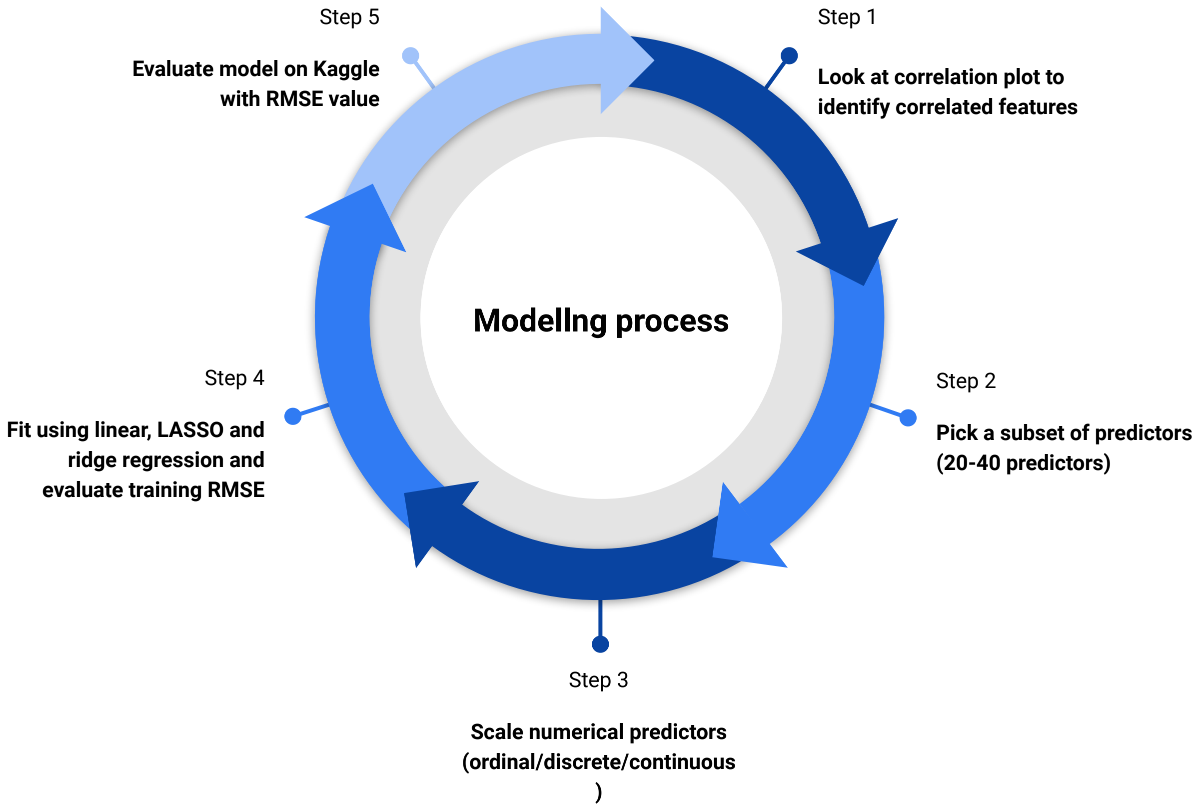
- Develop regression model to predict house value in Ames city with the lowest possible error

- Identifying key predictors that can highly influence housing prices.

Challenges

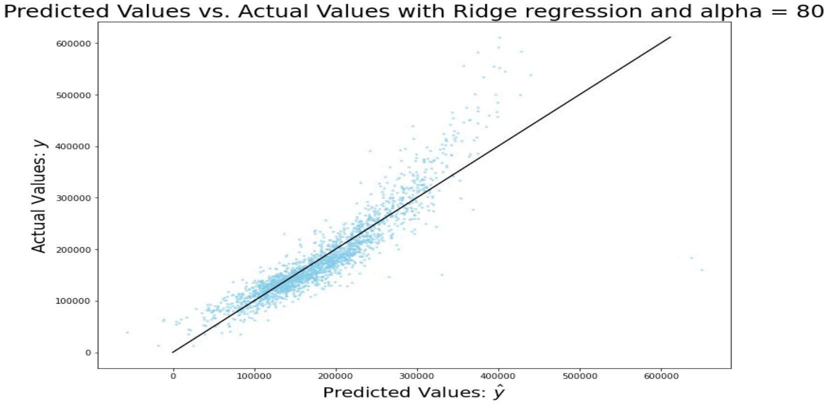
- Mismatch in column names with data dictionary
  - Sale Condition not included
- Interpretation of columns meaning
  - E.g. Lot Frontage
- Missing data and selecting imputation methods
  - Nominal vs Ordinal vs Discrete vs Continuous
  - E.g. missing data for lot frontage

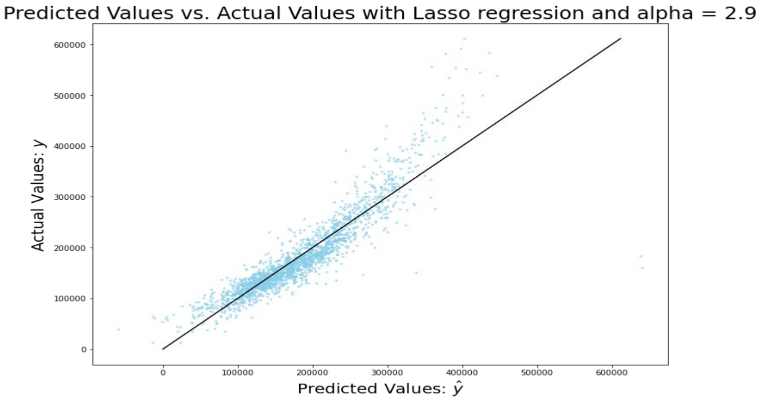
Modelling process

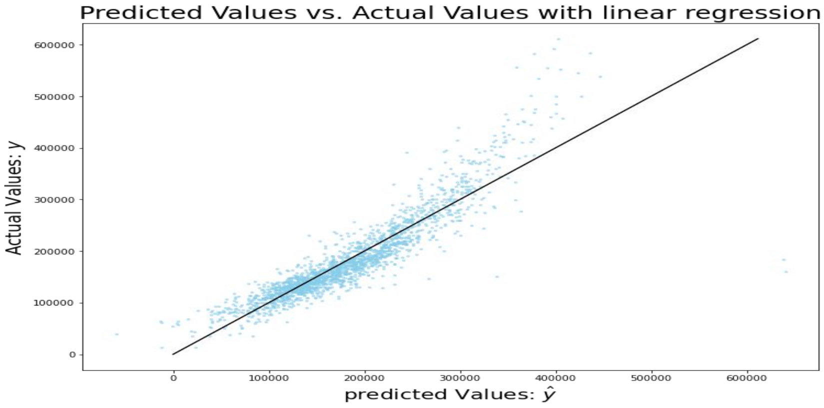


Model Performance

- Training RMSE comparison
  - Ridge: 34615
  - Lasso: 35059
  - Linear Regression : 35070

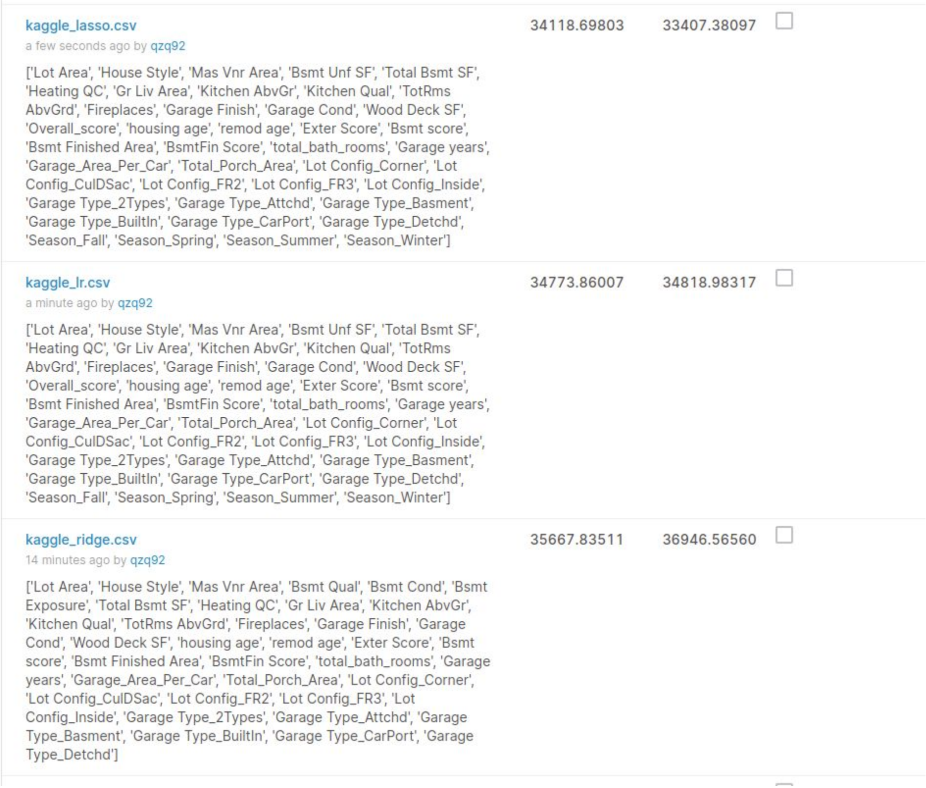






Model Performance

- Lasso regression has the best Kaggle RMSE
    - RMSE: 33407, alpha = 2.9
- Number of attributes: 40
    - Numerical: 24
    - Dummy encoded categories: 16
        - 3 Categories (Season, Lot Config and Garage Type)



kaggle_lasso.csv a few seconds ago by qzq92 ['Lot Area', 'House Style', 'Mas Vnr Area', 'Bsmt Unf SF', 'Total Bsmt SF', 'Heating QC', 'Gr Liv Area', 'Kitchen AbvGr', 'Kitchen Qual', 'TotRms AbvGrd', 'Fireplaces', 'Garage Finish', 'Garage Cond', 'Wood Deck SF', 'Overall_score', 'housing age', 'remod age', 'Exter Score', 'Bsmt score', 'Bsmt Finished Area', 'BsmtFin Score', 'total_bath_rooms', 'Garage years', 'Garage_Area_Per_Car', 'Total_Porch_Area', 'Lot Config_Corner', 'Lot Config_CulDSac', 'Lot Config_FR2', 'Lot Config_FR3', 'Lot Config_Inside', 'Garage Type_2Types', 'Garage Type_Attchd', 'Garage Type_Basment', 'Garage Type_BuiltIn', 'Garage Type_CarPort', 'Garage Type_Detchd', 'Season_Fall', 'Season_Spring', 'Season_Summer', 'Season_Winter']	34118.69803	33407.38097	☐
kaggle_lr.csv a minute ago by qzq92 ['Lot Area', 'House Style', 'Mas Vnr Area', 'Bsmt Unf SF', 'Total Bsmt SF', 'Heating QC', 'Gr Liv Area', 'Kitchen AbvGr', 'Kitchen Qual', 'TotRms AbvGrd', 'Fireplaces', 'Garage Finish', 'Garage Cond', 'Wood Deck SF', 'Overall_score', 'housing age', 'remod age', 'Exter Score', 'Bsmt score', 'Bsmt Finished Area', 'BsmtFin Score', 'total_bath_rooms', 'Garage years', 'Garage_Area_Per_Car', 'Total_Porch_Area', 'Lot Config_Corner', 'Lot Config_CulDSac', 'Lot Config_FR2', 'Lot Config_FR3', 'Lot Config_Inside', 'Garage Type_2Types', 'Garage Type_Attchd', 'Garage Type_Basment', 'Garage Type_BuiltIn', 'Garage Type_CarPort', 'Garage Type_Detchd', 'Season_Fall', 'Season_Spring', 'Season_Summer', 'Season_Winter']	34773.86007	34818.98317	☐
kaggle_ridge.csv 14 minutes ago by qzq92 ['Lot Area', 'House Style', 'Mas Vnr Area', 'Bsmt Qual', 'Bsmt Cond', 'Bsmt Exposure', 'Total Bsmt SF', 'Heating QC', 'Gr Liv Area', 'Kitchen AbvGr', 'Kitchen Qual', 'TotRms AbvGrd', 'Fireplaces', 'Garage Finish', 'Garage Cond', 'Wood Deck SF', 'housing age', 'remod age', 'Exter Score', 'Bsmt score', 'Bsmt Finished Area', 'BsmtFin Score', 'total_bath_rooms', 'Garage years', 'Garage_Area_Per_Car', 'Total_Porch_Area', 'Lot Config_Corner', 'Lot Config_CulDSac', 'Lot Config_FR2', 'Lot Config_FR3', 'Lot Config_Inside', 'Garage Type_2Types', 'Garage Type_Attchd', 'Garage Type_Basment', 'Garage Type_BuiltIn', 'Garage Type_CarPort', 'Garage Type_Detchd']	35667.83511	36946.56560	☐

Findings and conclusion

- Lasso Regression is the best compared with Ridge and Linear regression models

- Three main attributes affecting price in absolute terms
  - Heating quality and condition (positive)
  - Total porch area (negative)
  - Wood deck area (positive)

- Possible explanations
  - Natural disasters (floods, rainstorms)
  - Damage and destruction by natural disasters
  - Wood deck as insulator of heat and comfort

Analysis constraints and recommendations

- Analysis bounded by given data
- More granular data required
  - E.g buyer data, seller financial background and coordinates of home sales
- Subsequent studies with granular data
  - Buyer and seller behaviours
  - Neighborhood

End