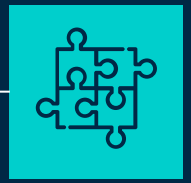
Ames Housing Regression Analysis

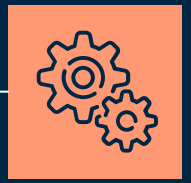
By Marcus Tan

TABLE OF CONTENTS



01

Problem Statement & Exploratory Data Analysis



02

Inferential Visualisations

03

Conclusion & Recommendations

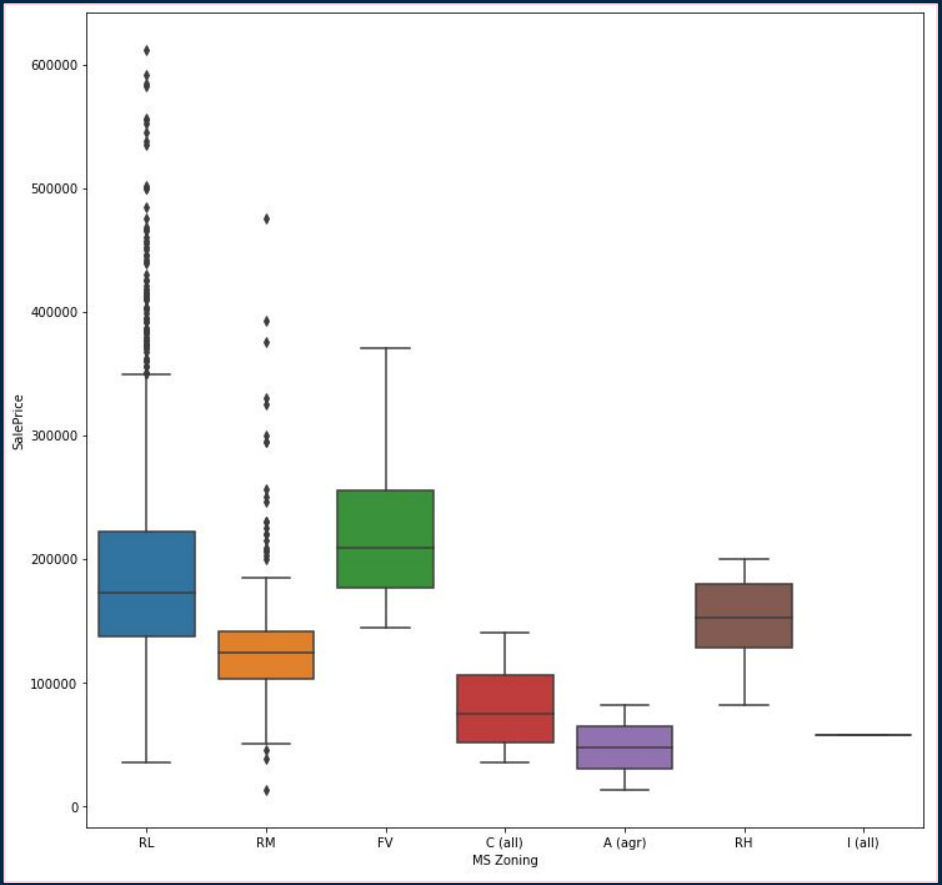
Problem Statement & Exploratory Data Analysis

01

Problem Statement

- Tasked to create a regression model that predicts house Sale Price.
- Provided with Ames Housing Dataset.
- Select the best model and provide some business recommendations.

Data Cleaning



- Dropped columns that have less than 50% data filled (e.g. `Alley', 'Pool Quality', 'Fence', 'Fireplace Quality' & 'Misc Feature')
- Removed rows for properties in the classification of “Commercial, Agriculture and Industrial”.
- Median price below 100k

Getting dummies

- Categorical Features

- Logical imputations:
  - Variable “Street” replacing “1” if “Paved” and “0” if “Gravel”.

  - Variable “Central Air Conditioning” replacing “1” if “Yes” and “0” if “No”.

  - Variable “Kitchen Quality” replacing “1” if “Excellent or Good” and “0” for anything else.

Adding new features

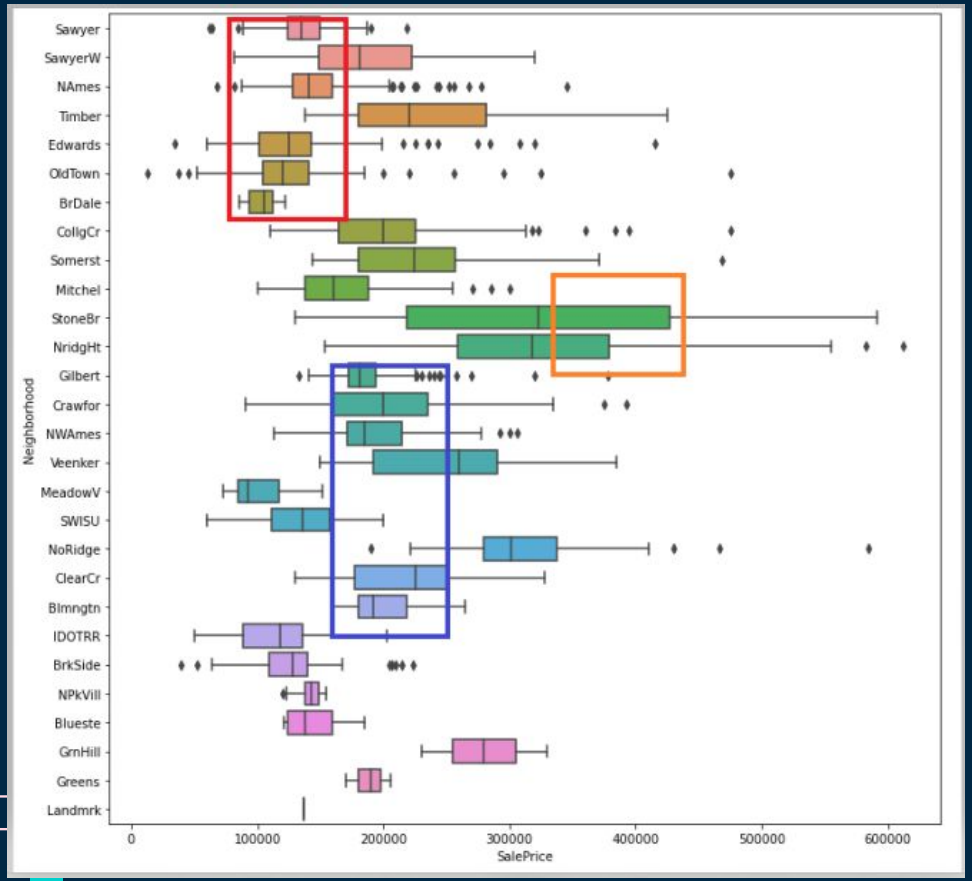
- Feature Engineering

- Logical imputations:
  - Getting age of property by taking the year sold minus the year remodeled.

  - Getting the total finished basement (sqf) of property by taking the total basement (sqf) minus the basement unfinished (sqf).

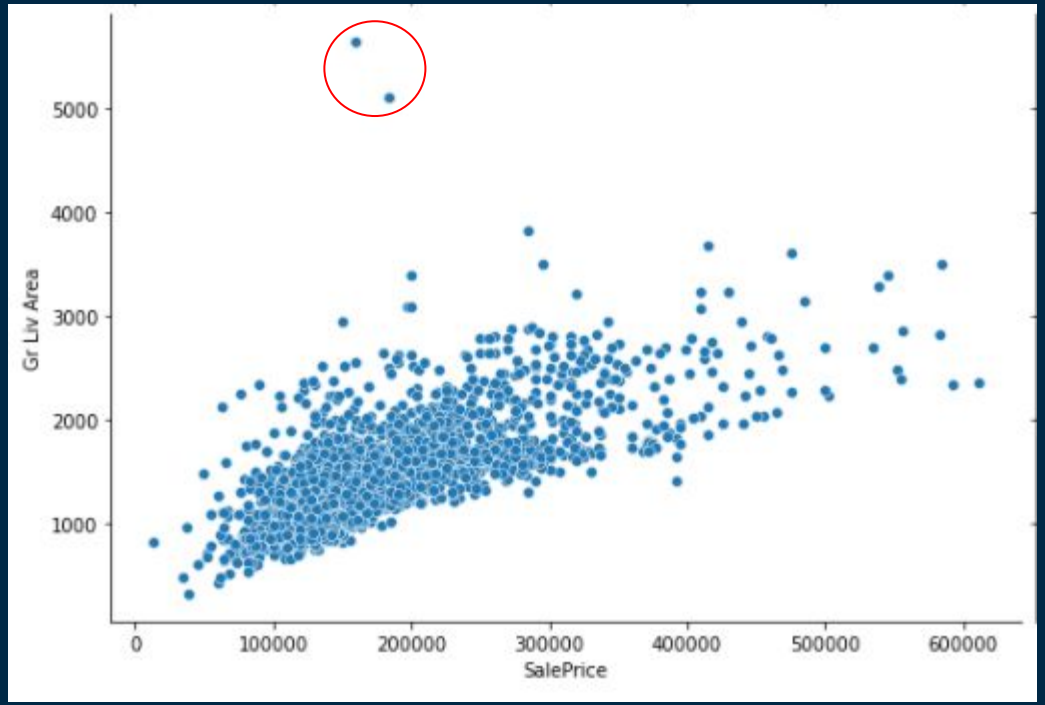
  - Getting the total porch area (sqf) of property by taking the total of open porch, enclosed porch, 3 season porch and screen porch.

Adding new features



- Sorted by mean Sale Price of neighborhoods
- 3 ordinal values
- 3 is for the top 9 area
- 2 for the middle 9
- 1 for the rest

Removing outliers



- Strong positive correlation
- Heteroskedasticity forms after 400k
- 2 outliers

Feature selection and preprocessing

- Features selected based on strong correlation with Sale Price
- Dropped features with strong correlation among the variables
- Scaled only the numerical features with StandardScaler

Inferential Visualisations

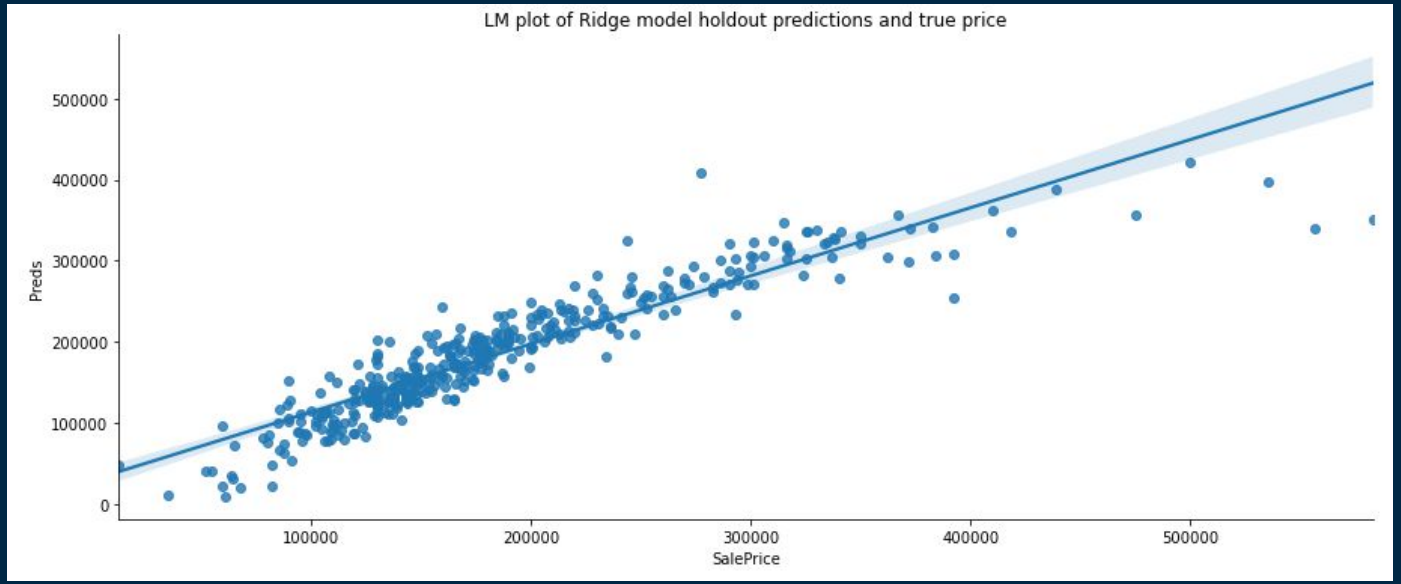
02

Inferential Visualization

Model	Train (RMSE)	Holdout (RMSE)
Linear	29420.91	32021.95
Ridge	29359.60	31957.06
Lasso	29370.87	31995.82
Elastic Net	29370.87	31995.86

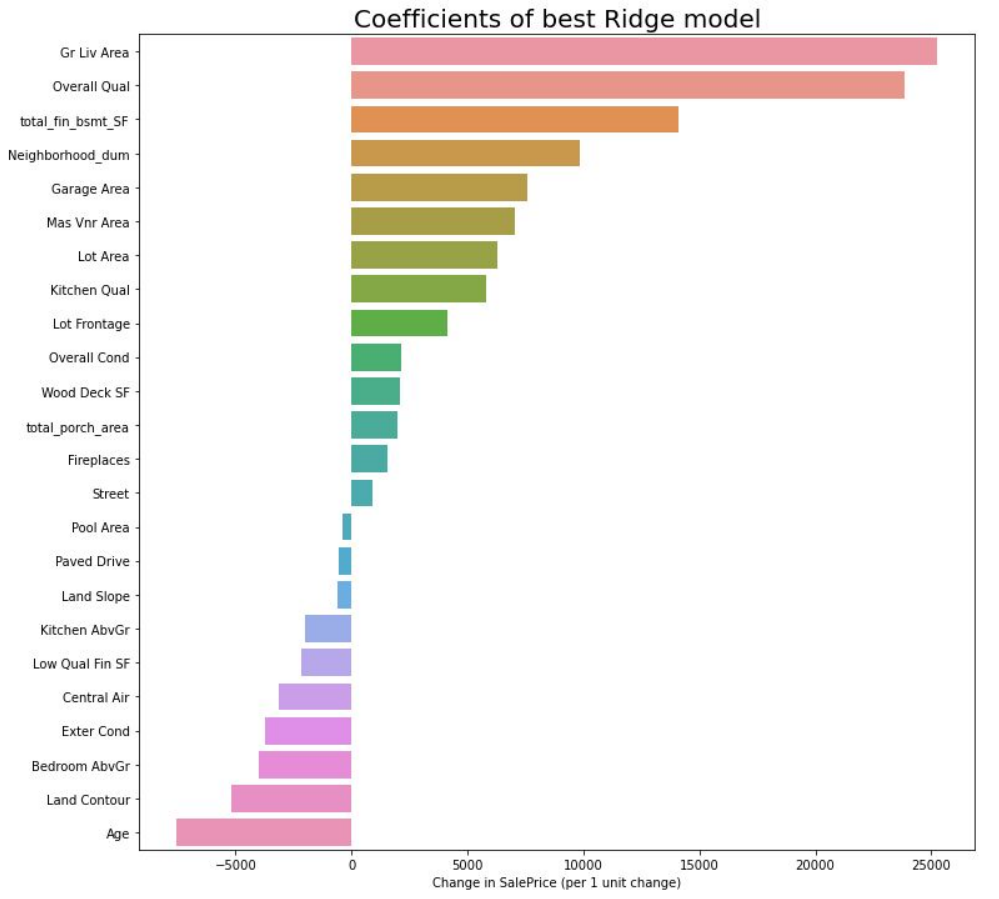
- 4 regression models were used.
- Ridge model has the best score.
- Elastic Net score is similar to Lasso (L1 ratio of 1).
- Shrinks the variable features.
- Reduces multicollinearity.

Inferential Visualization



- Performed well in the range of 100k to 300k
- Heteroskedasticity forms after 300k

Inferential Visualization



- Ground Living Area adds the most value to a home price
- Age hurts the value of a home the most
- Anomalies:
  - Pool area
  - Kitchen above ground
  - Bedrooms above ground

Conclusion & Recommendations

03

Conclusion and Recommendations

Conclusion:

- Ridge model performs the best.
- For housing in the range of 100k to 300k.
- Will not generalize to housing prices in other cities.

Ways to increase home value:

- Increase the Ground Living Area / overall material and finish of their homes.
- Construction/renovation cost may outweigh the increase in value.
- Request mayor to build more infrastructures and amenities around the Neighborhood.

Neighborhoods to consider:

- Stone Brook and Northridge Heights.
- High median sale price (300k and above).
- Range concentrated from 200k to 400k.

Do you have any questions?

THANKS