IA Construction

Project Proposal

Kaitlyn Ellsweig
SmartConsulting

TABLE OF CONTENTS

01 Problem Statement

02 Ames Dataset

03 EDA & Data Cleaning

04 Model Development

05 Initial Findings

06 Next Steps

IA Construction

IA Construction is a developer in the state of Iowa. They are looking to build development(s) in Ames, Iowa or possibly the surrounding area.

Ames Dataset Kaggle

- 80 variables
- 2,051 observations
- Ames Iowa

EDA & Data Cleaning Goals

Goal 1

Identify & Fix Missing Variables

Goal 2

Identify and fix data type issues

Goal 3

Determine which variables are most correlated with Sale Price

Goal 4

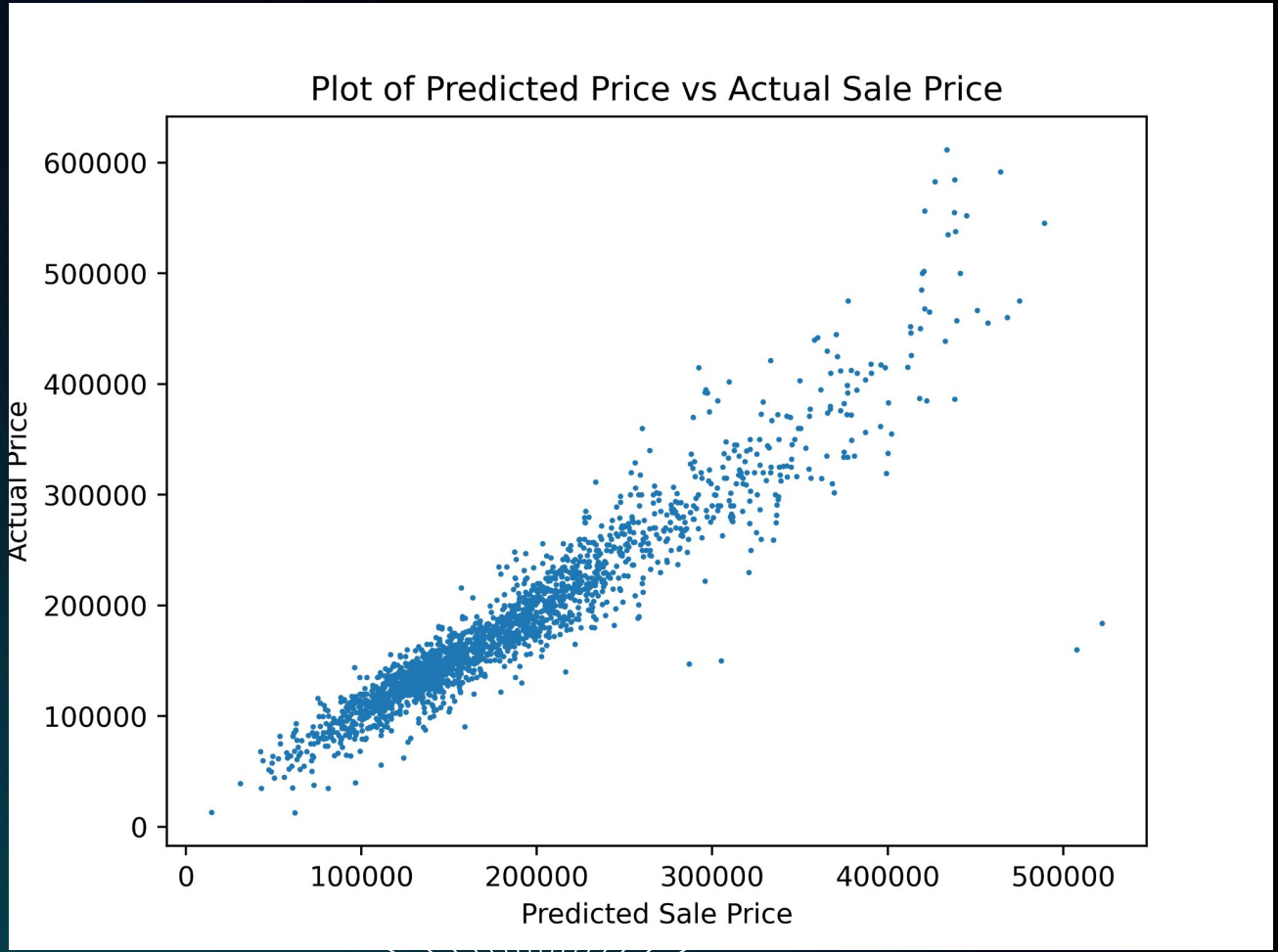
Identify any multicollinearity or other issues

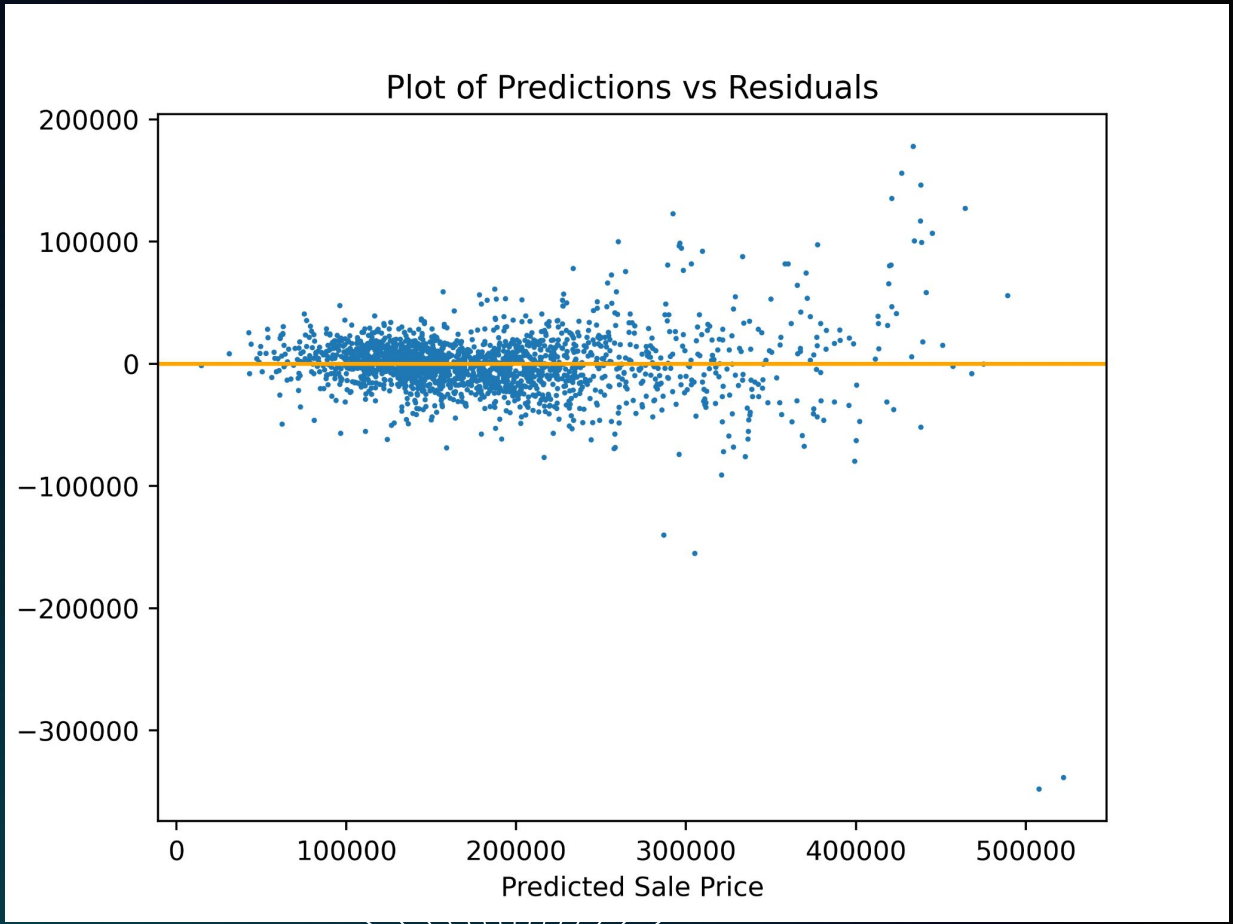
Model Development

- LASSO, Elastic Net
- Polynomial Features
- ***One-hot Encoding***
- ***Imputation using median***
- ***Variable Selection***

Model Metrics

	Train r2	Test r2	MSE
Final Model	.8911	.9079	548,835,999
Elastic Net	0.9379	.8183	1,020,246,568
Polynomial Features	.9419	.8999	679,339,772


Plot of Predicted Price vs Actual Sale Price
Actual Price
Predicted Sale Price
0
100000
200000
300000
400000
500000
600000
0
100000
200000
300000
400000
500000


Plot of Predictions vs Residuals
200000
100000
0
−100000
−200000
−300000
0
100000
200000
300000
400000
500000
Predicted Sale Price

Initial Findings

- The model was not overfit
- Bias may be high despite a relatively strong r2 value
- While the validation r2 was high compared with other models, since the MSE and r2 of the validation data after the train-test split were higher than the metrics of the training data, it seems likely that there are still opportunities to reduce bias and improve the model without increasing variance and overfitting the model.

Specific Variables:

- Positive Coefficients:
  - Garage (higher ~$10,00/car space)
  - Basement (only $4 per sq ft)
  - Bathrooms ($5,550 per each additional bathroom)
- Categorical Variables:
  - House Type (stories, development)
  - Neighborhood

Next Steps

Resources:

- Pandas Library https://pandas.pydata.org/pandas-docs/version/1.3/index.html
- MatPlot Lib Library https://matplotlib.org/
- SciKit Learn Library https://scikit-learn.org/stable/index.html
- Numpy
- Seaborn
- StatsModels

Special Thanks

***Image credit: individuals' slack profiles*

Susan Hopper

Help with simplifying code OLS Summary table

Qingxin Wei

Help with One-Hot-Encoder syntax to ignore new info

Any Questions?