Predicting House Prices in Ames, Iowa

GA DSI Project 2 by Suma Karanam, 30 August 2021

Background

- Available data:
  - Extensive data[1] about various features of 2051 houses sold in Ames, Iowa between 2006 and 2010
  - Examples of features: Square footage, Lot Area, Neighborhood, Overall Quality, # Bedrooms, # Bathrooms etc.
- Goals:
  - Build a Linear Regression model to predict the home price for homes in this region
  - Tune the model to minimize RMSE
  - Determine if predicting log(SalePrice) instead of SalePrice improves model performance

[1] *Ames housing data description*

Contents

- Part 1

  Exploratory Data Analysis

- Part 2

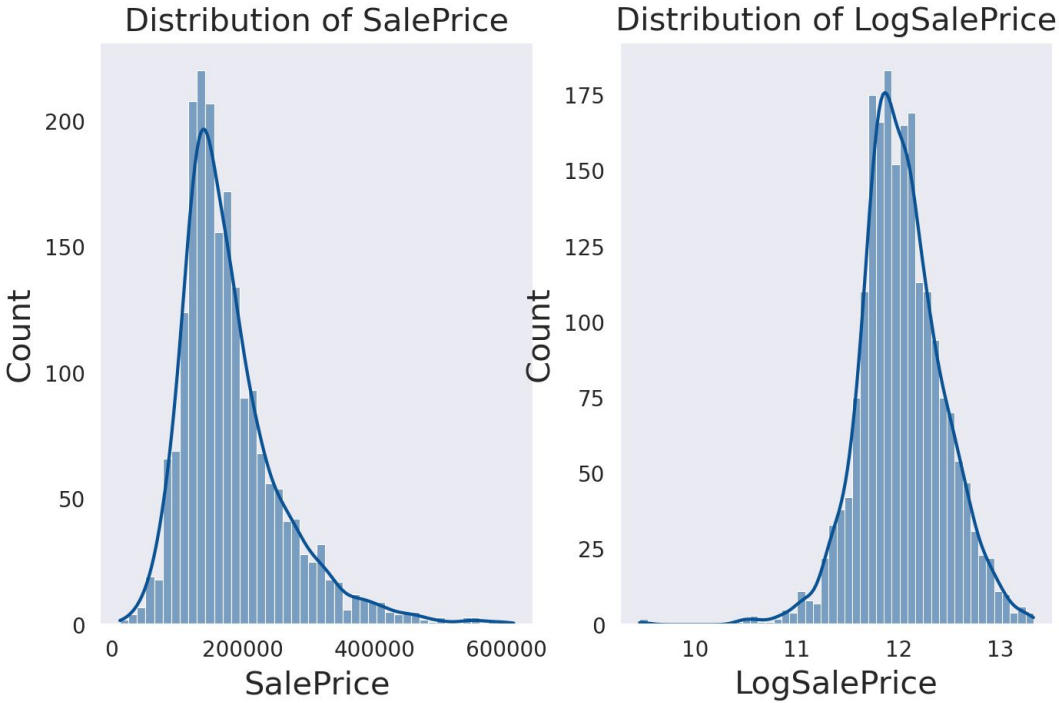
  Overview of the model

- Part 3

  Model Performance

Part 1

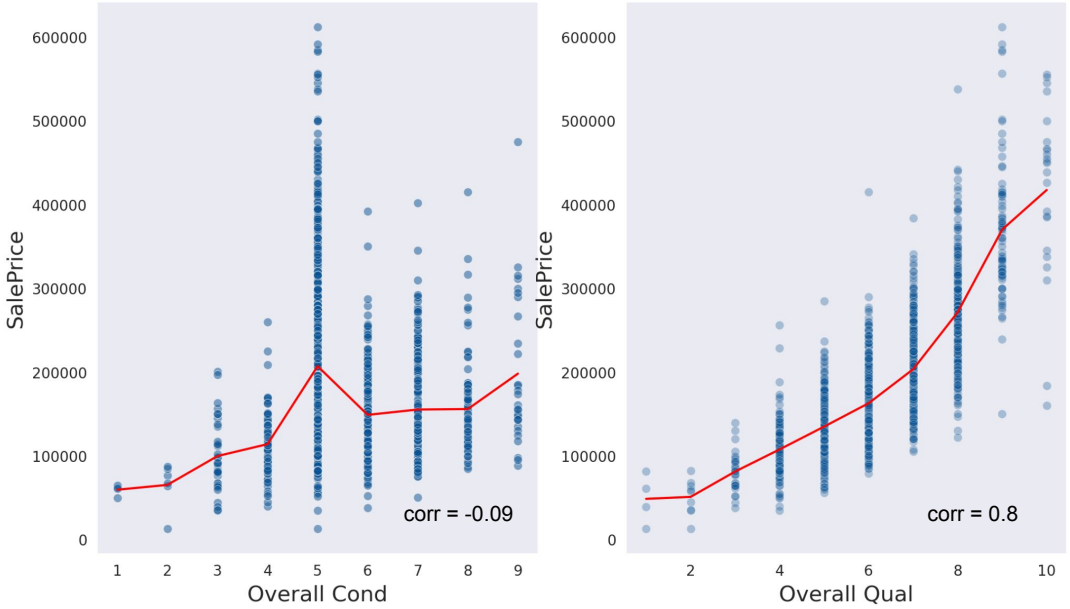
EDA - Distribution of the sale prices



- Distribution of house prices seem typical - more homes selling closer to mean price and a few selling at high prices
- Log of sale price looks closer to normal distribution

Part 1

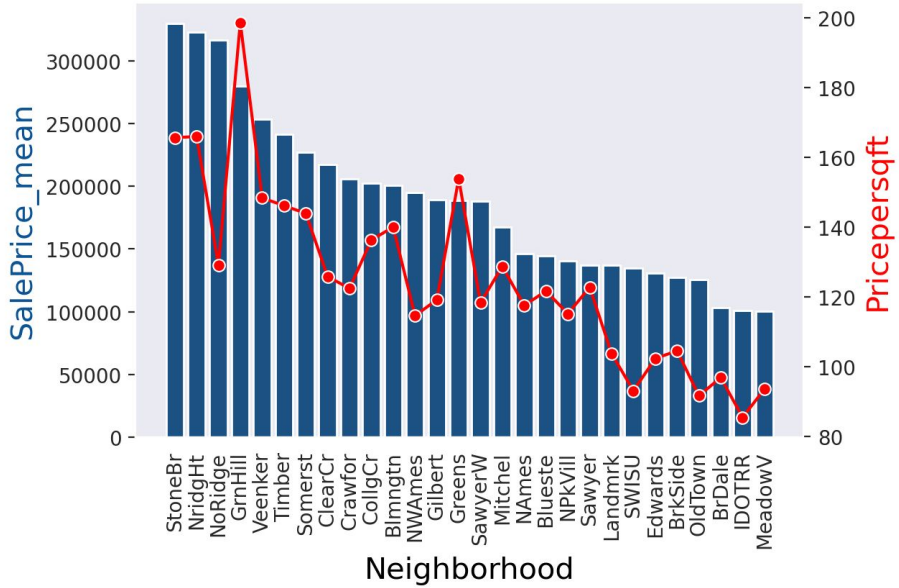
EDA - ‘Overall Cond’ and ‘Overall Qual’



- Prices increase with 'Overall Qual' as expected
- There is a small negative correlation with 'Overall Cond'
- 'Overall Cond' not used in the model

Part 1

EDA - Neighborhood



- The mean sale price of a home varies considerably by neighborhood
- The average price per sqft also varies considerably with neighborhood
- Two attempts to capture the effect of neighborhood on home prices:
  - Mean Encode the neighborhood feature
  - Find an ‘implied neighborhood price’ for each home and use that as feature

Part 2

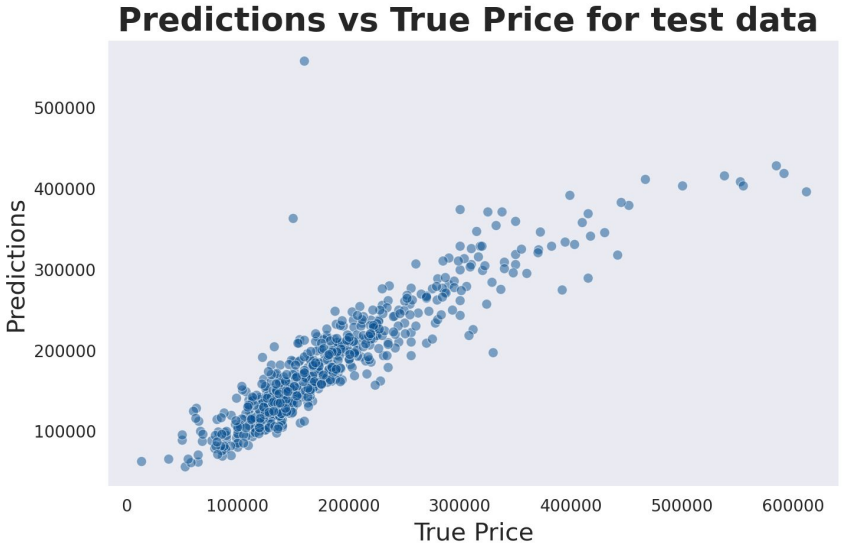
Overview of the model

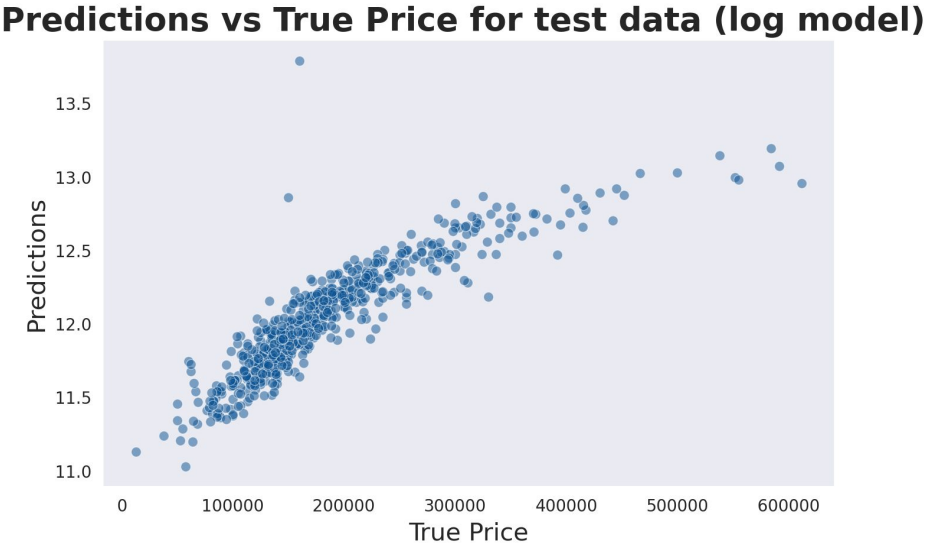
Features used in the model

- Mean Encoded: 'Overall Qual' , 'MS SubClass', 'Kitchen Qual'
- Binarized: 'MS Zoning', 'Lot Config', 'Condition 1', 'Bldg Type', 'House Style', 'Roof Style', 'Exterior 2nd', 'Exter Qual', 'Foundation', 'Heating QC', 'Sale Type'
- Neighborhood: assigning an implied ‘neighborhood sale price’
- Other numerical columns: 'Gr Liv Area', 'Year Built', 'Year Remod/Add', 'Garage Area', 'TotRms AbvGrd', 'Open Porch SF', 'Wood Deck SF', 'Lot Area'
- One hot encoded: 'Lot Shape’

Part 3

Model Performance



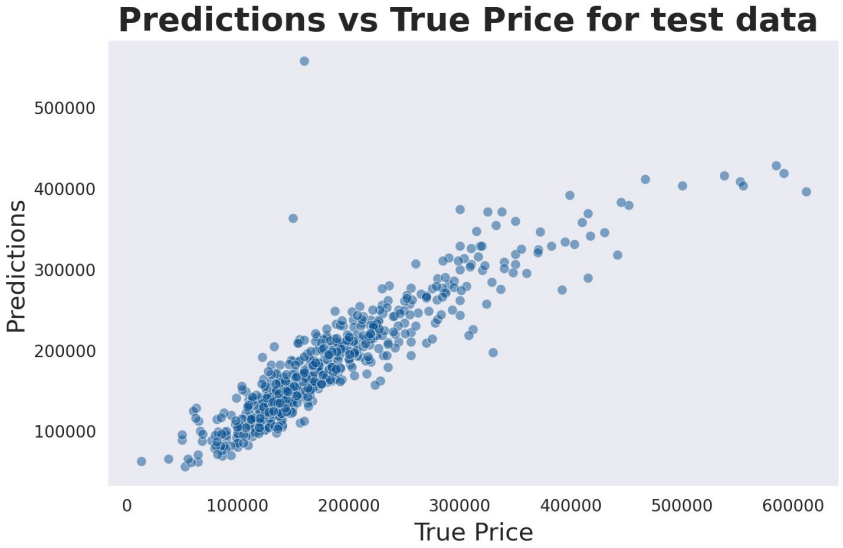


	Predicting SalePrice		Predicting Log SalePrice	
	Training data	Test Data	Training Data	Test Data
R2	0.88	0.81	0.88	0.82
RMSE	26412	36203	29136	44008

Top coefficients of Predicting SalesPrice model	
Feature	Coefficient
Overall Qual	26260.64245
Gr Liv Area	23520.66593
Neighborhood	12391.02904

Part 3

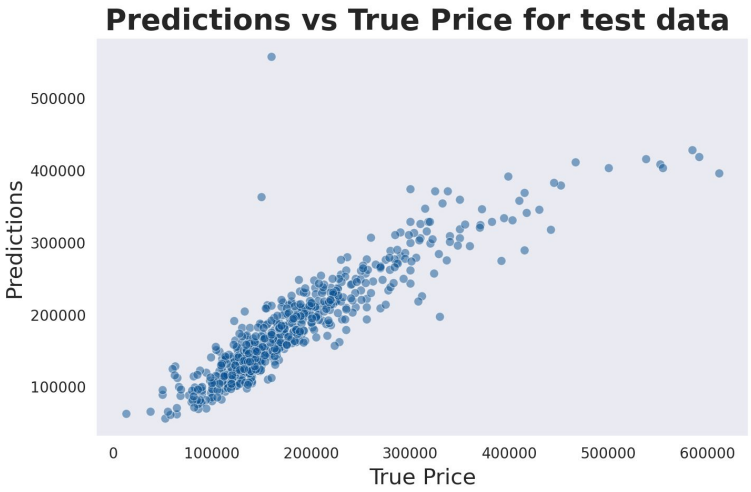
Model Performance

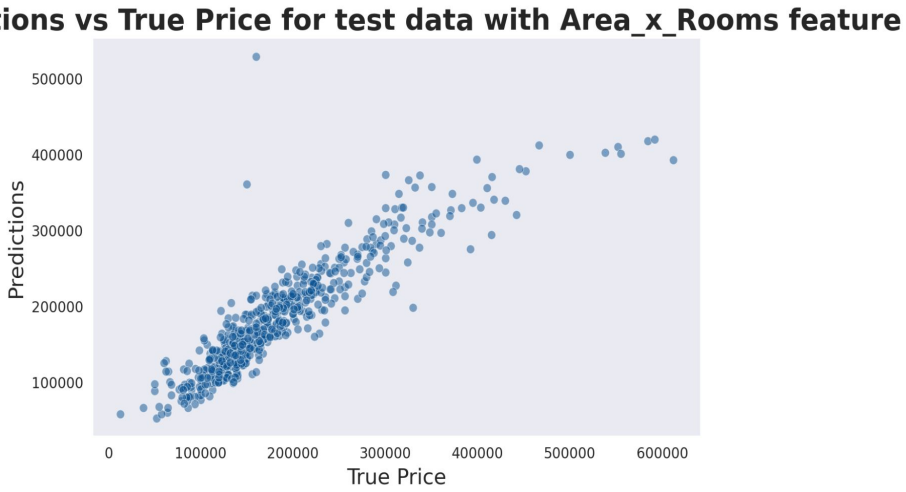


	With average price per neighborhood		With sq.footage * average price per sqft for each neighborhood	
	Training data	Test Data	Training Data	Test Data
R2	0.88	0.81	0.89	0.82
RMSE	26412	36203	25480	35073

Part 3

Model Performance





	With average price per neighborhood		With Area_X_#Rooms	
	Training data	Test Data	Training Data	Test Data
R2	0.88	0.81	0.88	0.82
RMSE	26412	36203	26308	35925

TODO

- Building a model with a different set of features
  - Example: 'Garage Cars' instead of ‘Garage Area’
- Including other interaction terms:
  - Products of the two highly correlated similar variables
  - Example: Area * 1st Floor Area
- Building different models for different sets of Neighborhoods
- Investigate other ways to address multicollinearity - examples: reduce the number of features, creating meaningful interaction features