Ames Housing Sale Price Modeling

By
Dr Monday Oshoikpor

Overview

- Problem Statement
- Exploratory Data Analysis
- Correlation Analysis
- Model for Prediction
- Conclusion and Recommendations

Problem Statement

- I am a member of a newly-hired data science team at Ames Realty Company.
- I am tasked to use housing data collected between 2006 and 2010 to create a model that can take in the housing data and return a predicted true sale price for houses.
- Model success will be evaluated on the Root Mean Squared Error (RMSE) score.

Exploratory Data Analysis

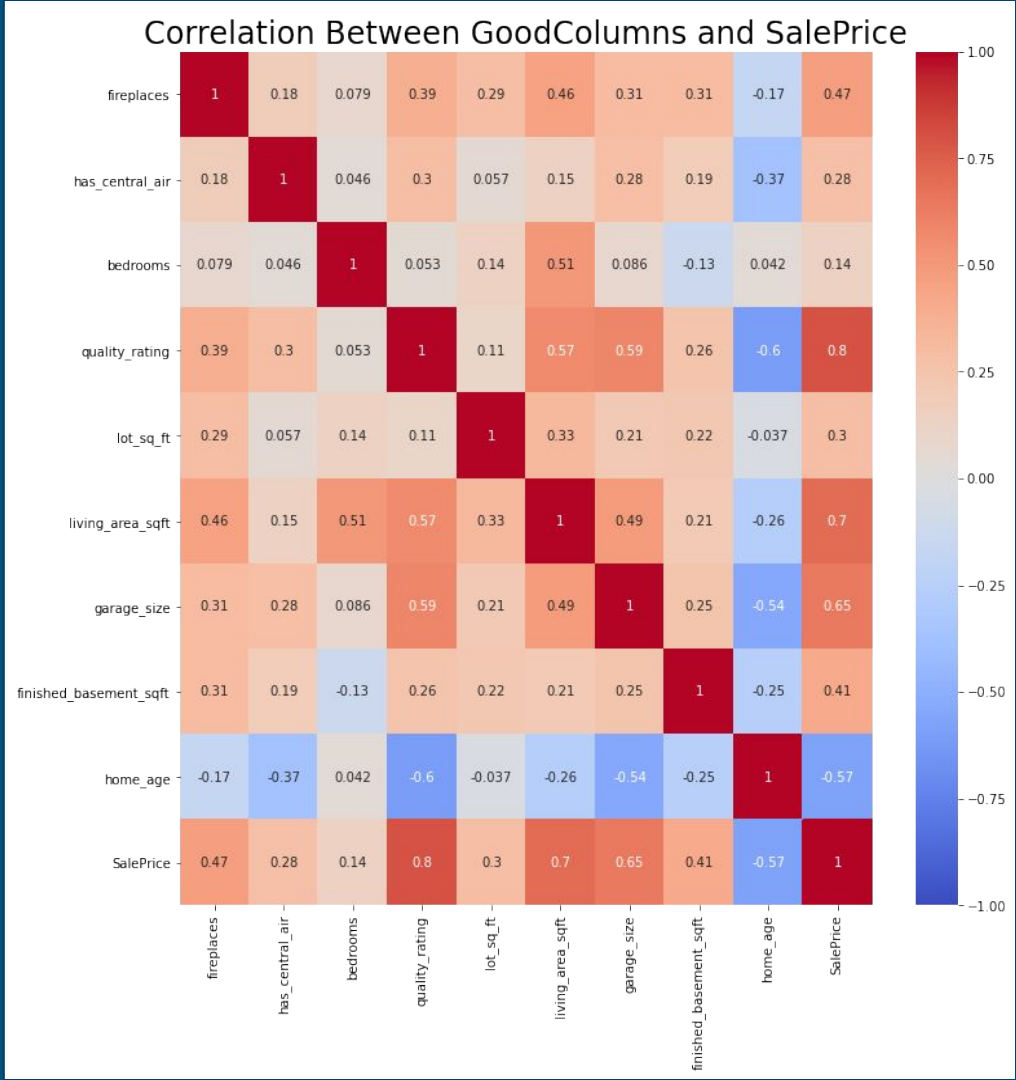
	Id	parcel_id	ms_subclass	zoning	lot_frontage	lot_sq_ft	paved_street	alley	lot_shape	flatness	...	misc_value	month
0	109	533352170	60	RL	0.0	13517	1	0	IR1	Lvl	...	0	
1	544	531379050	60	RL	43.0	11492	1	0	IR1	Lvl	...	0	
2	153	535304180	20	RL	68.0	7922	1	0	Reg	Lvl	...	0	
3	318	916386060	60	RL	73.0	9802	1	0	Reg	Lvl	...	0	
4	255	906425045	50	RL	82.0	14235	1	0	IR1	Lvl	...	0	

5 rows × 78 columns

- Created a function to rename columns
- Filled all null values with a *for* loop
- Created and ran EDA function to clean up dataframe

Correlation Analysis

- The heatmap shows Quality Rating with the highest positive correlation to SalePrice at 80%



Correlation Analysis

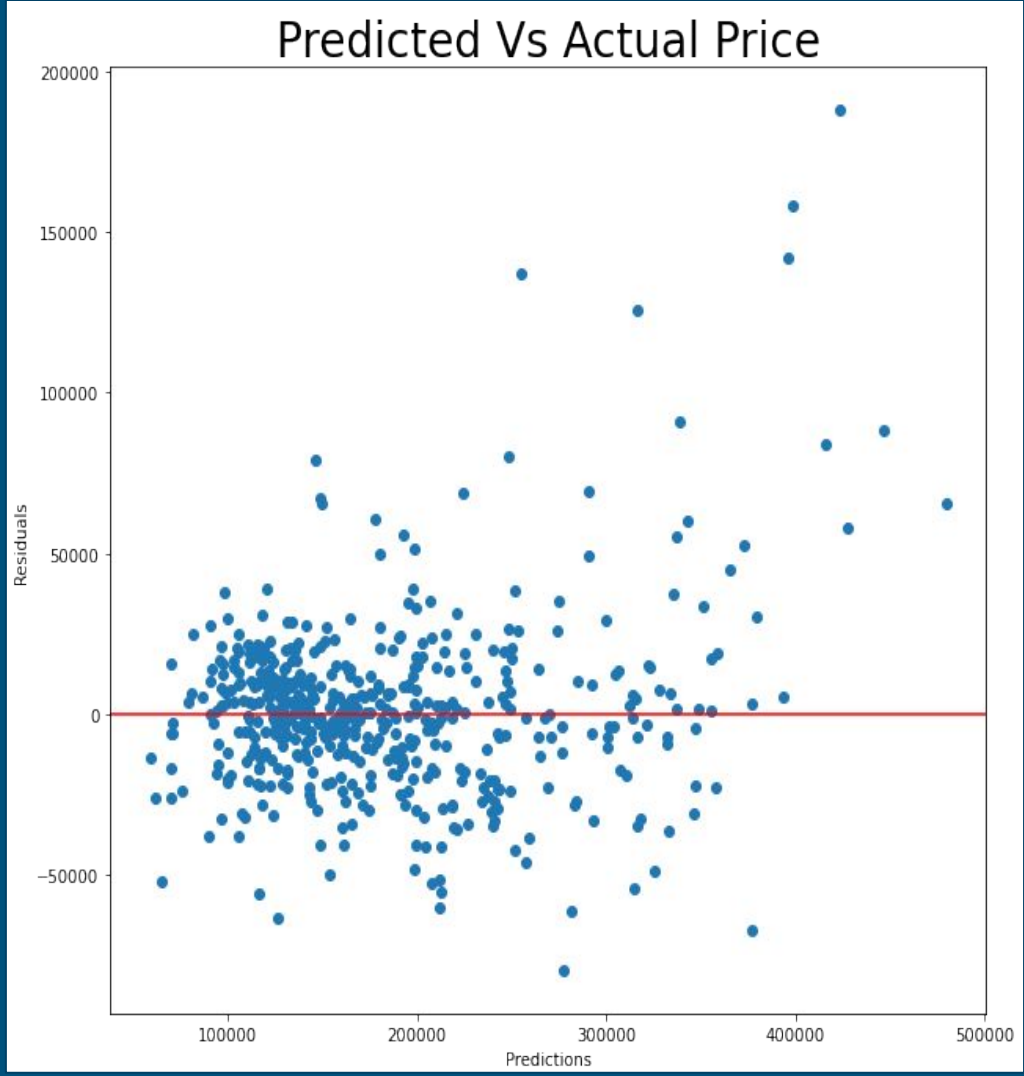
Fireplaces	Has paved driveway
Has central air	Functionality typical
Bedrooms	Building: Single Family, Townhouse End or Inside Unit, Two Family Conversion, Duplex
Quality Rating	
Kitchens	
Lot Sq Feet Living Area Sq Feet	Porch Type: Open, Enclosed, Seasonal or Screen
Baths	Garage Size
Finished Basement Sq Feet	Year Built

- I identified 22 “good” features and created histograms to see the correlation with SalePrice
- Quality Rating scored the highest positive correlation

Model for Prediction

Ridge Model

- High R2 score similar for both the train and test data
  - Train = 0.8885
  - Test = 0.8941
- Low variance and low bias
- Mostly homoskedastic
- A few outliers with a few high price homes which means there's room for improvement
- RMSE Score = 27570



Conclusion

- My model give credence to generalization.
- With a high accuracy score that is similar for both the train and test data, I am really happy my organization will have less problem predicting future sale price
- There are a few outliers that tell me the model is not scoring well on a few high price homes so there is still room for improvement.

Recommendation

- Dive deeper into the categorical features especially in neighborhoods because I found that houses over $500,000 were all in only two places.
- Improve on existing model by increasing features

References

- https://datausa.io/profile/geo/ames-ia/
- https://www.cityofames.org/about-ames/about-ames
- https://unionstreetmedia.com/the-rise-of-machine-learning-in-real-estate/#:~:text=Personalized%20Marketing%20Automation%20%E2%80%93%20machine%20learning,neighborhood%20and%20property%20is%20best
- https://pandas.pydata.org/docs/pandas.pdf
- http://jse.amstat.org/v19n3/decock/DataDocumentation.txt