Predicting House Prices Using Linear Regression

PRESENTED BY

EDDIE REED – DATA SCIENTIST

Problem Statement

Using historical sales data of houses in Ames, Iowa, I will attempt to create a machine learning model to accurately predict the price of future houses. I will be using the linear regression model in this project.

Data Collection and EDA

The dataset came from historical sales of houses in Ames, Iowa. The following key observations were made during EDA:

1. Null values in various columns
2. Redundant features that could be combined
3. Many features to consider (over 80)

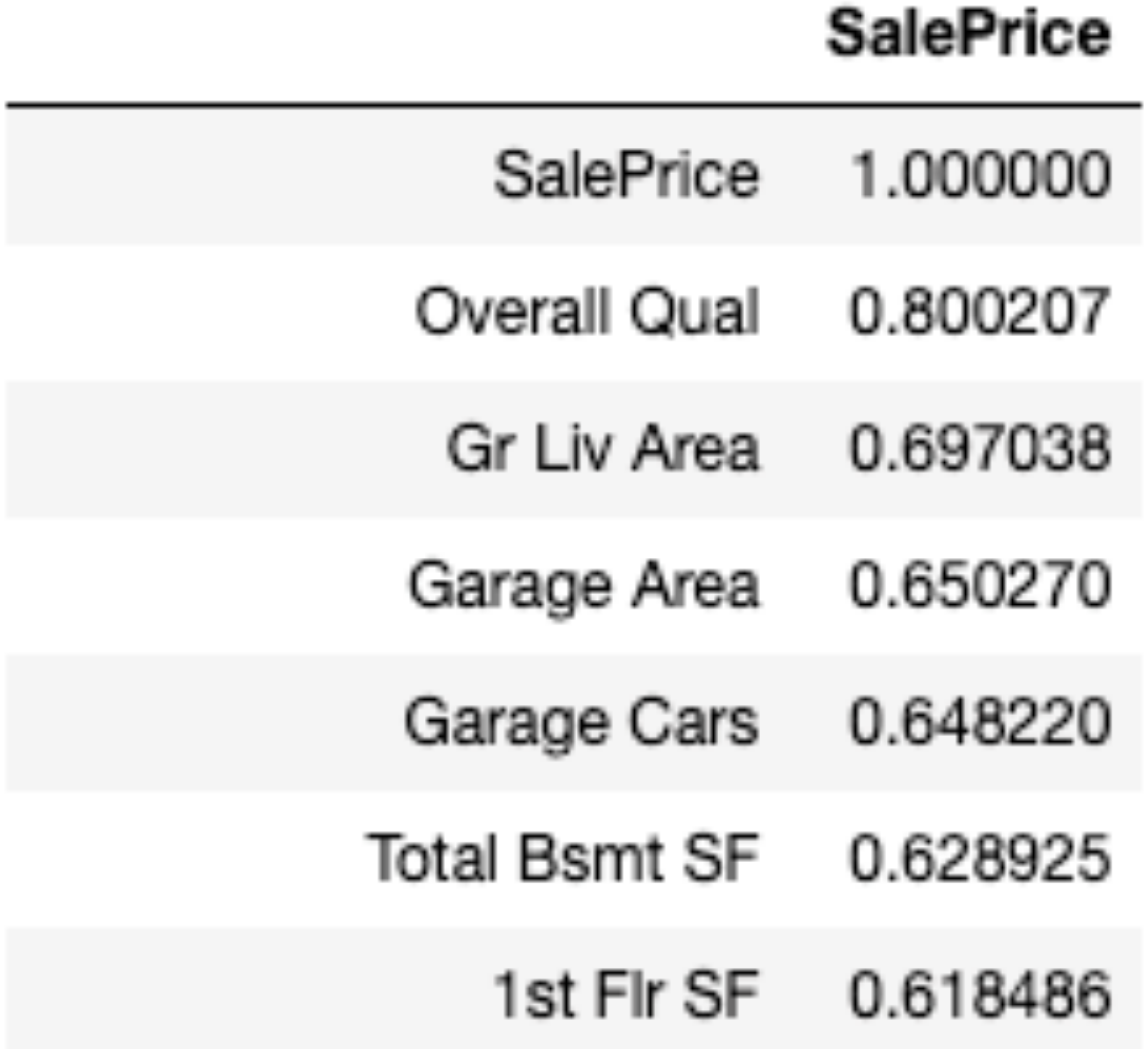
Preprocessing and Modeling

The following steps were taken to prepare the data for modeling:

1. All null values were identified as that feature not being included in the property. Null values were changed to reflect the non feature.
2. Several redundant features were dropped
3. All categorical features were converted to binary values
4. The number of features were reduced by constructing a correlation matrix and selecting the top 6 features that showed the greatest impact on sale prices.

Correlation Matrix of Features to Sale Price (top 6)

Correlation – This value ranges between -1 and 1. The closer the value to 1, the stronger the correlation that feature has on positively impacting the outcome of the sale price.



	SalePrice
SalePrice	1.000000
Overall Qual	0.800207
Gr Liv Area	0.697038
Garage Area	0.650270
Garage Cars	0.648220
Total Bsmt SF	0.628925
1st Flr SF	0.618486

Evaluation of the Model

After fitting the model to the training data and running a set of predictions on the testing dataset, the following observations were noted:

1. The model was 79% accurate on predicting sales prices on the training dataset.
2. The model was 72% accurate on predicting sales prices on the testing dataset (data that the model has not seen before)

Conclusion

Based on the 6 features selected, the model would predict the sales price of a house 72% of the time. Below are some key takeaways:

1. The model underperforms compared to data it has already seen. (e.g. The training dataset. This is called “overfitting” of a model.)
2. More features could be added some feature engineering could be performed to see if it the model performs better.
3. Tuning parameters on the model could also be helpful.

Overall, the model does show that there is a good correlation with some of the top features in houses like overall quality, square footage, and whether the property has a garage, in predictng the sales price of a house.