

NC STATE UNIVERSITY

North Carolina State University

Department of Financial Mathematics

FIM 590 003 Machine Learning in Finance

Project 1 Iowa House Price Prediction

Author:

Haozhe Cui

Jinjia Peng

Jinyi Yang

Zhen He

October 9, 2024

Contents

| | |
|-----------------------------------------------------|-----------|
| 1. Introduction..... | 1 |
| 2. Data Preparation..... | 2 |
| 2.1.Feature Encoding..... | 2 |
| 2.2.Handling Missing Values:..... | 2 |
| 2.3.Multicollinearity..... | 2 |
| 2.4.Outlier..... | 3 |
| 2.5.Data Normalization:..... | 5 |
| 3. Model Explanation..... | 6 |
| 3.1.Data Split..... | 6 |
| 3.2.Model..... | 6 |
| 3.3.Validation..... | 7 |
| 3.4.Test..... | 8 |
| 3.5.Performance Summary..... | 8 |
| 4. Report Finding..... | 9 |
| 4.1.Neighborhood Analysis..... | 10 |
| 4.2.Other Features Analysis..... | 11 |
| 5. Predicting the Real-time House Price..... | 15 |
| 6. Summary..... | 17 |
| 7. Reference..... | 18 |

FIM590-003 Project 1 Iowa House Price Prediction

1. Introduction

In this study, we are concentrating on forecasting property prices through the use of machine learning techniques. Our goal is to create a regression model that can effectively predict housing costs by taking into account property characteristics, like neighborhood, area and room numbers. We are looking at data to determine the factors that impact property prices and using this knowledge to enhance the precision of our forecasts.

We began by cleaning and preparing the dataset (from IA_House_Price_Original_Data.xlsx) by addressing any data points and identifying any values while selecting important characteristics, for analysis. In instances where data was missing; we substituted those values with zeros to uphold data integrity and employed graphical representations to identify any outliers that might skew the analysis. We included a mix of numerical and categorical attributes such as the quality of the basement and neighborhood demographics utilizing methods like assigning order-based numbers and placeholder variables, for encoding purposes.

Once the data was ready, for analysis we split it into training, validation, and test groups to make sure our model could work effectively with the data. Next, we. Assessed a variety of regression models to choose the one that performed the best. By following this process our goal was to create a model that could help in estimating real estate prices offering useful information, for individuals looking to buy, sell, or invest in properties.

2. Data Preparation

2.1.Feature Encoding

The dataset includes both categorical variables that represent facets of the property like its size, quality, and location.

In the dataset there are 80 characteristics available; however, for our regression study, we concentrate on 23 specific features comprising 21 numerical and 2 categorical variables that are significant, for forecasting housing costs. This consists of a variable denoting basement quality based on ceiling height and another indicating the locality of the property. (Hull, 2021)

Basement Condition (Type of Feature):

The quality of the basement is determined by the height of the ceiling. Is categorized into six groups ranging from "Excellent " for ceilings over 100 inches to "Poor " for ceilings under 70 inches high or for basements with no ceiling at all. Each category is assigned a value from 5 for "Excellent" to 0 for "No basement " indicating a progression, in the assessment of this characteristic. We can consider how the quality of basements influences housing prices; taller ceiling heights often indicate improved living conditions. Consequently, leads to property values.

Neighborhood (Category) :

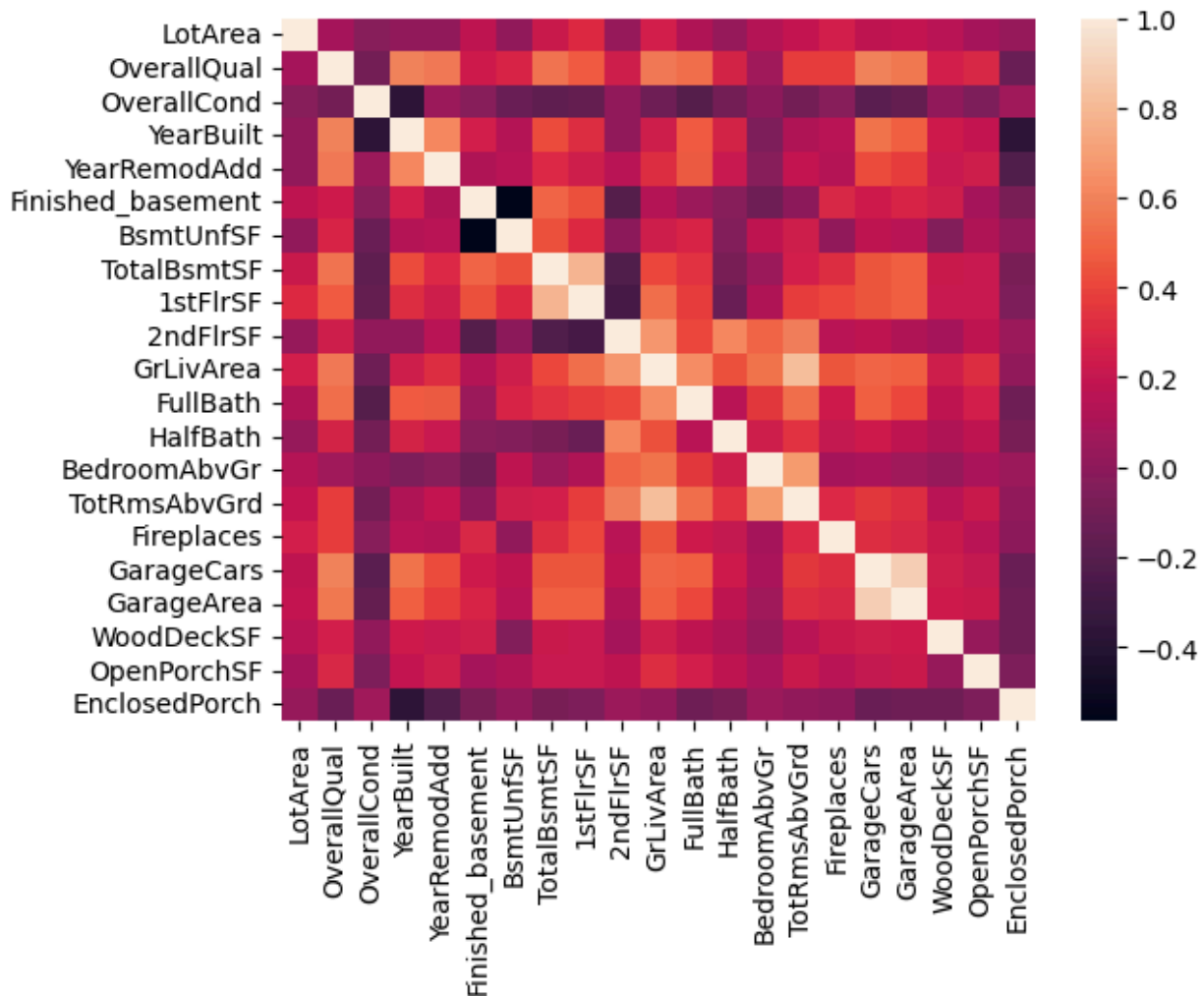
The dataset also encompasses the neighborhood as a factor since the location significantly influences real estate prices. There are 25 neighborhoods considered in the analysis by generating 25 variables to indicate whether a house is situated in a specific neighborhood. This approach guarantees that the impact of neighborhoods is comprehensively accounted for in the model and enables us to explore the correlation between location and property values.

2.2.Handling Missing Values:

After checking the data set, there are no missing values. However, after importing files into the data frame, the importing function automatically converts the 'NA' string in basement quality into a null value. Our team filled it with 0 for mapping convenience.

2.3.Multicollinearity

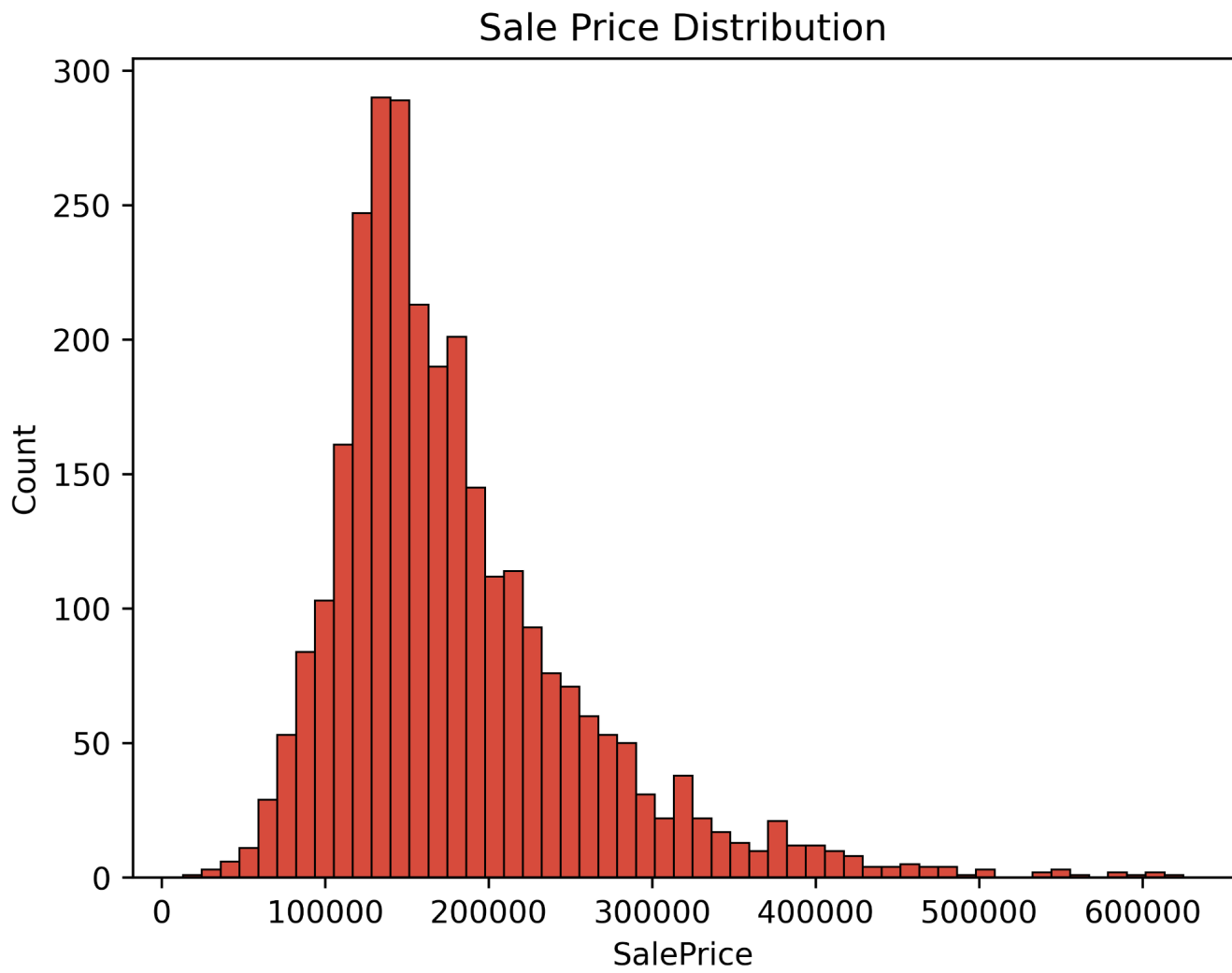
When creating a linear model that uses many features to make predictions, some of those features can be highly correlated with each other. This isn't a problem that's going to break the model; it will still make predictions and it might have good performance metrics. However, it is an issue on the interpretation of the coefficients for your model because it becomes hard to tell which features are truly important.



According to this graph, there does not exist any highly correlated features.

2.4.Outlier

We utilized the describe function to examine the data. Verified that there were no outliers as all parameters were, within a range. Furthermore, we graphed the price distribution which did not reveal any obvious outliers.



Housing Prices Stats:

| | |
|--------------------|-----------|
| Number of Houses | 2908 |
| Mean Price | \$180,272 |
| Standard Deviation | \$78,139 |
| Minimum Price | \$12,789 |
| 25% Quantile | \$129,362 |
| 50% Quantile | \$160,000 |
| 75% Quantile | \$213,310 |
| Maximum Price | \$625,000 |

2.5.Data Normalization:

Data standardization is a part of our modeling procedure to guarantee that each feature has an impact on the model's outcome. In our dataset, various features are measured in varying scales. For example, the living area is in square feet while the year built is recorded as a four-digit number. If we want to get a more efficient model and find out the most significant features based on their regression coefficients, we should keep all feature manners at the same level, which is, standardization. In this project, we used z-score:

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}$$

Where Z_i is the z-score vector for the i-th feature, X_i is the i-th feature vector, μ_i is the mean of the i-th feature, and σ_i is the standard deviation of the i-th feature. For unity, we also applied this normalization method to the sale prices, so that we would get a suitable and comparable size of parameters from regression models.

3. Model Explanation

Now we are ready to apply our model to the cleaned data set. To start the algorithm, we need to decide the size of the training set, validation set, and testing set. We have decided to use 1800 data sets as our training size, 600 data sets as the validation size, and 508 data sets as the testing size. We will calculate the mean absolute value (MAE), for each lambda value we choose(0.10, 0.30, and 0.60), on the validation set. By comparing MAE based on the validation set, we can select the comparatively well-performed model as our final model. Then we test this model by testing sets.

3.1.Data Split

We divided the dataset into three sections, for training and assessment of the model; a training set with 1800 observations, a validation set with 600 observations, and a test set with the remaining 508 observations.

- The training data is utilized to teach the model by enabling it to grasp the connections between characteristics and the outcome variable.
- The validation dataset helps refine the model and tweak hyperparameters to avoid overfitting so that the model can perform effectively on data.
- The evaluation phase reserves the test set for assessing the model's performance without bias to predict how well it will handle data in actual use cases.

This way of organizing the data ensures that we have a method for training the model and evaluating its performance accurately throughout the different stages of development.

3.2.Model

Ridge Regression:

Ridge regression is a form of regularization method used in machine learning. It adds a regularization term to the loss function by adding a multiplication of a parameter λ and the sum of the squares of the model coefficients, which penalizes large coefficient values. This constraint prevents the model from fitting the training data too perfectly, improving its generalization to new, unseen data. The formula is given by:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - (y^{(i)}))^2 + \lambda \sum_{j=1}^n \theta_j^2$$

Lasso Regression:

Similar to ridge regression, instead of using the square of each coefficient, LASSO regression adds a penalty to the coefficient values by using the absolute value of the coefficients. Note that due to the attributes of this algorithm, the penalty can push some coefficients to exactly zero when the regularization parameter λ is sufficiently large. This can help us visualize the features that are considered important in our model.

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - (y^{(i)}))^2 + \lambda \sum_{j=1}^n |\theta_j|$$

3.3.Validation

There were a total of 6 models in the validation pool, and we used MAE as a measure of training error.

$$MAE = \frac{1}{m} \sum_{i=1}^m |y^{(i)} - y_{predict}^{(i)}|$$

The result showed us that the ridge regression model with an alpha of 0.1 was the best model in the validation pool.

| Model: | Error (MAE): |
|--------------------------------|---------------------|
| Lasso_0.02': Lasso(alpha=0.02) | 0.27579459343707263 |
| Lasso_0.06': Lasso(alpha=0.06) | 0.28434590939251325 |
| Lasso_0.1': Lasso(alpha=0.1) | 0.29686077635091607 |
| Ridge_0.1': Ridge(alpha=0.1) | 0.23865776911304362 |
| Ridge_0.3': Ridge(alpha=0.3) | 0.238680839744581 |
| Ridge_0.6': Ridge(alpha=0.6) | 0.23875604250280294 |

3.4.Test

Based on the validation result, we tested the ridge regression model with alpha of 0.1 using testing sets, and used R squared and MAE as measures of performance.

R squared:

The R-squared value of the Ridge regression model, for the test set is 0.9013. This figure shows how effectively the model describes variations in the target variable with a value closer to 1 implying an alignment with the data.

$$R^2 = 1 - \frac{\sum_{i=1}^m (y^{(i)} - y_{predict}^{(i)})^2}{\sum_{i=1}^m (y^{(i)} - \bar{y})^2}$$

It's the fraction of the variability in the dependent variable explained by the regressor.

Mean Absolute Error:

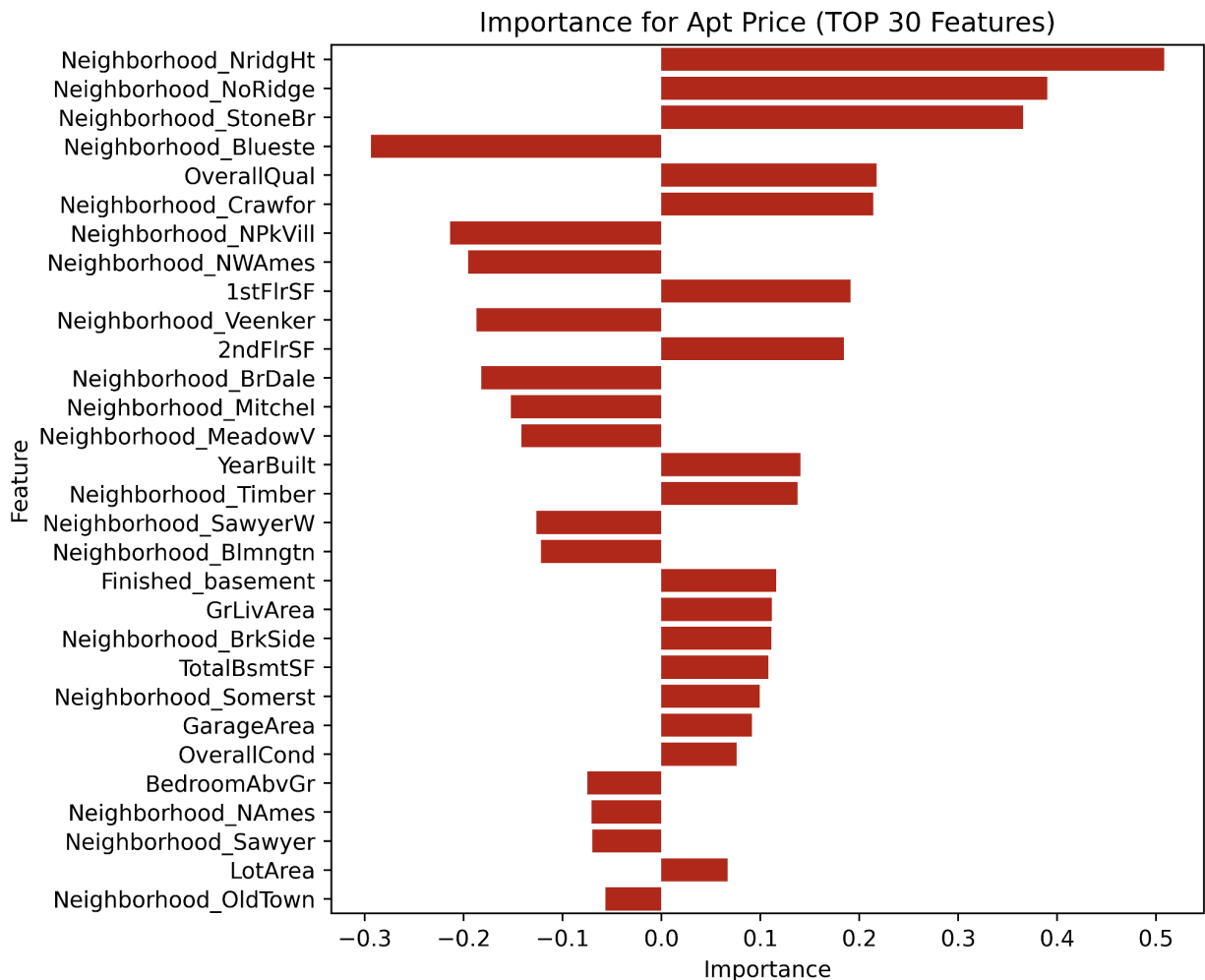
The Ridge regression model's Mean Absolute Error, on the test set is 0.2146. MAE reflects the size of the discrepancies between real values and gives insight into how effectively the model forecasts the target variable.

3.5.Performance Summary

By looking at the R squared and MAE out of the sample, we can conclude that our model, which could explain more than 90% variation and made less than 22% error, did a great job of predicting the sale price. Next, we tried to conduct a feature analysis based on the testing result.

4. Report Finding

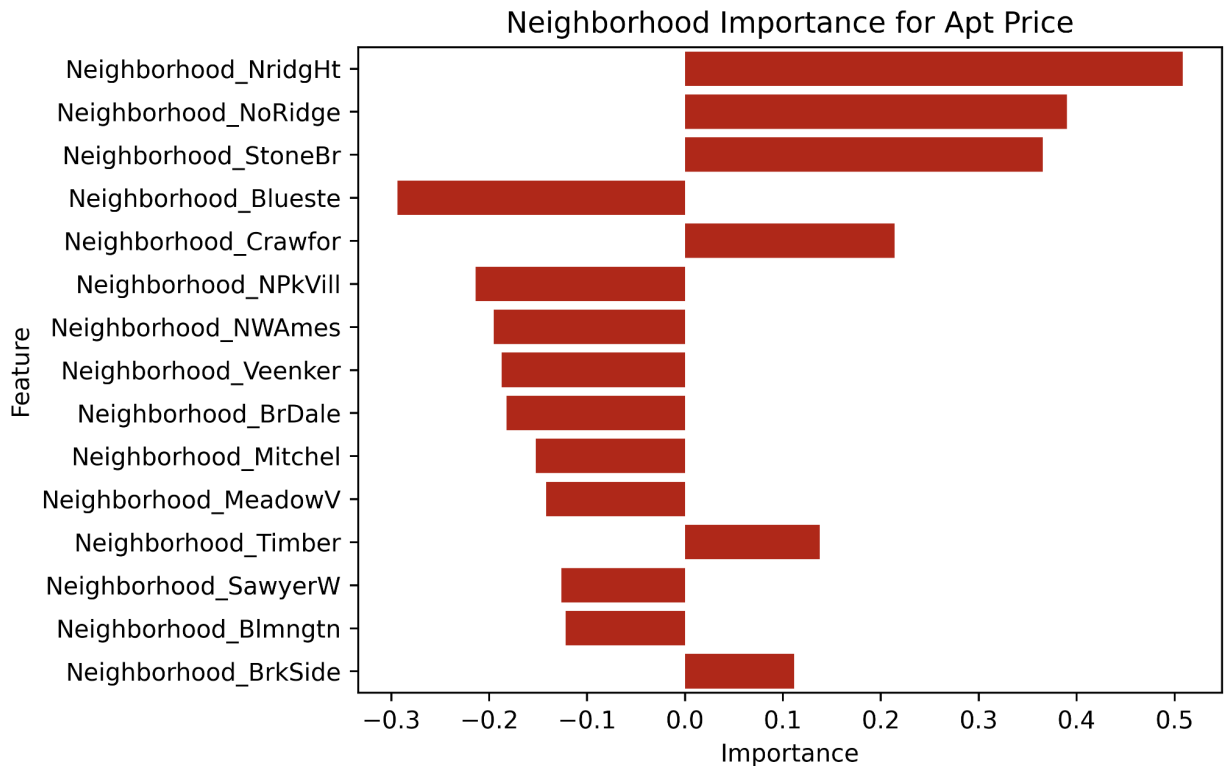
Now we want to dive deeper into our regression result, to see which feature contributes most to the whole model. We visualized the importance, which is the coefficients of features. Note that the absolute value of each coefficient is between 0 and 1, which is indeed the result after the standardization process. Thus we sorted the features by the importance (absolute value) of coefficients, and we ranked them in order from the top to the bottom and generated a graph. However, the actual values of these coefficients are preserved, as shown in the graph that there is a negative value (example: BedroomAbvGr) presented.



From the result, we can see that the feature Neighborhood is the most significant, and there are also other features such as OverallQual.

4.1. Neighborhood Analysis

First, we will look at the impact of different neighborhoods.



Positive and Negative Influences on Price:

In places, like Northridge Heights (abbreviated as NighT) and Northridge (No. Ridge) there are factors at play in terms of pricing indicating that homes in these neighborhoods tend to command higher prices than those in other areas.

Here are some potential explanations; improved infrastructure and top-notch schools attracting buyers, due to crime rates. Appealing amenities, in well-located neighborhoods lead to increased property values.

In contrast, neighborhoods, like Bluestem and Northpark Villa (NPKVill) have price-based strides showing a significant drop in the value of houses, in these regions.

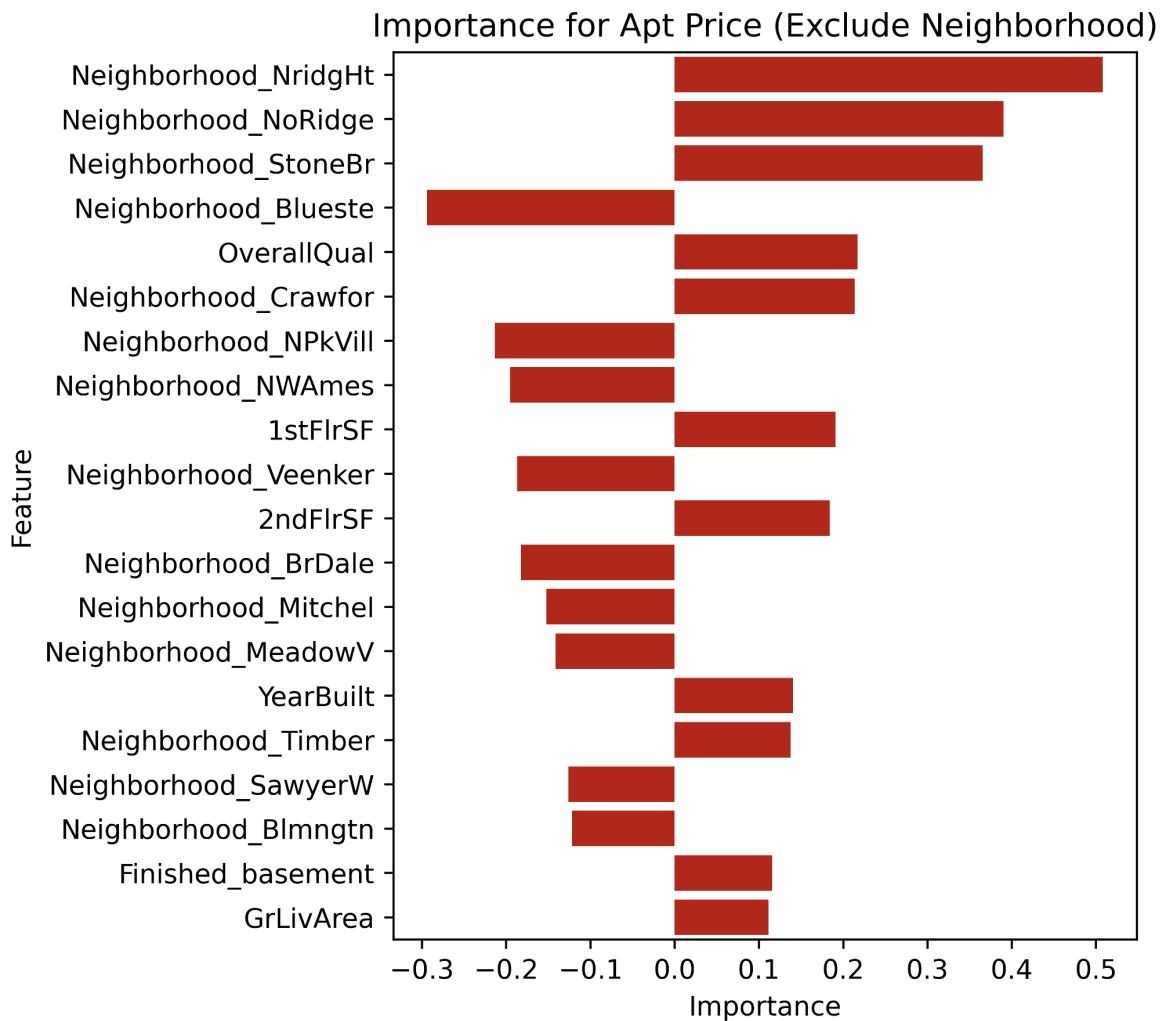
This might be because of reasons, like increased crime rates or lower quality amenities in the area and limited access to services and commercial places as well as recent economic declines or

lack of appealing facilities in those neighborhoods which could lead to a drop in their appeal and consequently lower property values.

In Summary, the area where a house is situated greatly influences its price tag according to the chart that displays the differences, in real estate values, among various neighborhoods.

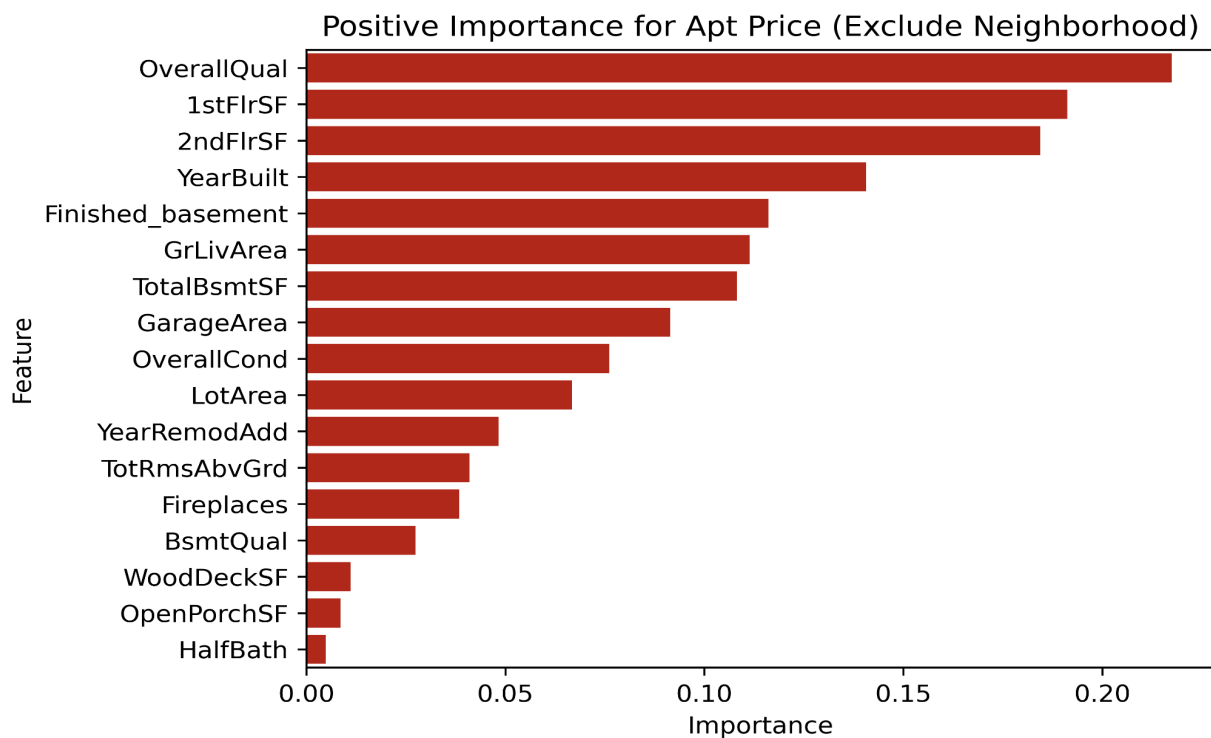
4.2.Other Features Analysis

Now we will take a look at the features other than the neighborhoods. Notice that after feature normalization to some degree, we can see each feature's significance from the absolute value of their regression coefficients. Thus we sorted the features by the importance (absolute value) of coefficients, and we ranked them in order from the top to the bottom and generated a graph. However, the actual values of these coefficients are preserved, as shown in the graph that there is a negative value (example: BedroomAbvGr) presented.



According to the chart illustrating how different factors affect apartment prices, it's clear that Overall Quality, along with First Floor Area (1stFlrSF) and Second Floor Area (2ndFlrSF) significantly boost house prices positively. Conversely, the Number of Bedrooms (Bedroom bvGr) and Unfinished Basement Area (Bsmt sfSF) seem to exert even slightly adverse impacts, on the price levels.

Positive Coefficient Analysis:



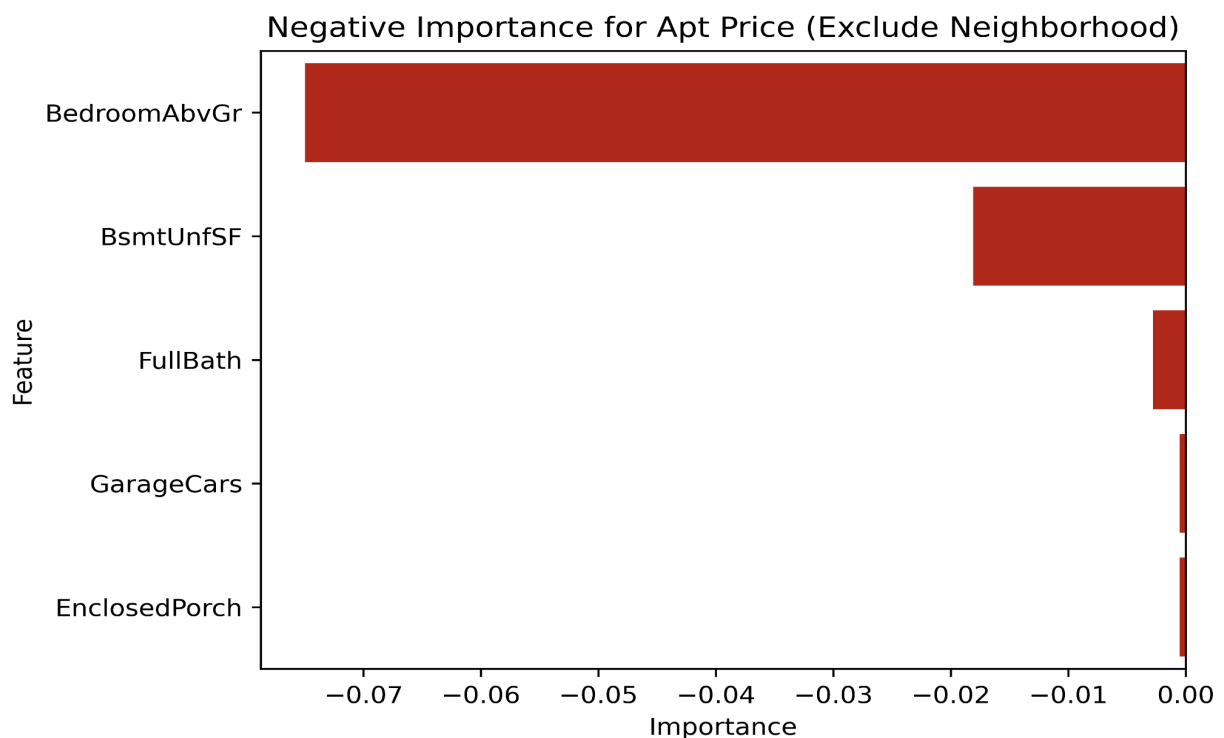
The summary chart showing the effects reveals that each feature mentioned contributes to the value of homes to different extents. Quality of construction has the impact of boosting prices significantly by adding value to the property. Likewise, both the size of the floor and the second floor also have influences indicating that buyers highly esteem the larger living spaces they provide. Newer homes hold appeal due to their designs and reduced maintenance needs—considering the Year Built is a factor in the decision-making process.

Having a completed basement and a generous total basement area can boost property prices well; although not significantly as the main living spaces do. Suggesting that well-designed and roomy basements can enhance the overall value of a home too. Additionally, the size of the living area and garage are factors that impact the value of a property since living areas and adequate garage space are highly sought after by buyers looking for convenience and comfort. Maintained homes

with lots of recent renovations tend to be more appealing, in the market due to factors, like Overall Condition and Year Remodeled having a positive influence.

Minor positive effects are seen from aspects like the number of bedrooms, above ground level the presence of fireplaces, and the quality of the basement in a home. All contributing both functionally and aesthetically to its appeal. Even elements such as the size of a wood deck or open porch and the presence of bathrooms exhibit beneficial impacts on property desirability. Indicating that added living conveniences and outdoor spaces play a role in enhancing overall attractiveness, to some extent. Ultimately a built residence, with living areas and outdoor spaces along with extra facilities can significantly boost the property's attractiveness, in the market as potential buyers are often willing to invest more for such appealing attributes.

Negative Coefficient Analysis:



On the other hand, it appears that certain property attributes like the Number of Bedrooms Above Grade (BedroomAbvGr) hurt house prices – suggesting that having too many bedrooms may not necessarily increase property value and could even lower its overall appeal due to cramped room sizes or inefficient layout utilization. Another significant factor contributing negatively is the Unfinished Basement Area (BsmtUnfSF) indicating that buyers generally do not perceive spaces as valuable compared to finished areas, in a property.

Having bathrooms (Full Bathrooms abbreviation; Full Bath) and garage spaces (Garage Cars abbreviation; GarageCars) seems to have minimal negative effects, on property value, suggesting that exceeding a certain threshold in adding these features may not substantially raise property value but might indicate excessive or underutilized space instead.

Additionally, Enclosed Porches (Enclosed Porch abbreviation; Enclosed Porch) although having an impact on value as per the data analysis results may not be particularly appealing to potential homebuyers. In terms and, from a negative perspective. These drawbacks underscore the principle that having more isn't necessarily better all the time; certain features that appear beneficial could diminish a property's worth if they're too abundant or not put to good use.

5. Predicting the Real-time House Price

We have successfully developed a high-performing model, and now we are eager to test its effectiveness in real-world conditions. We discovered a property for sale at 3206 Aspen Rd, Ames, IA 50014, listed on Zillow.com, and we applied our model to predict its price. (Zillow, 2024)

Here is the information about this house:

| Characteristics | Value | Note |
|-------------------|------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Sale Price | \$ 457,000 | - |
| LotArea | 11761.2 | sqft |
| OverallQual | 10 | A subjective assessment of the quality of the material and finish of the house, not directly stated but inferred to be high due to premium features (Granite countertops, stainless steel appliances, etc.) |
| OverallCond | 9 | The overall condition of the house is inferred to be good based on the well-maintained state and recent updates |
| YearBuilt | 1994 | 1994 |
| YearRemodAdd | 1994 | No specific remodel year was mentioned; possibly 1994 or later due to the modern updates |
| Finished_basement | 1041 | sqft |
| BsmtUnfSF | 0 | Not specified in the document but possibly little or no unfinished space due to the mention of a fully finished basement |
| TotalBsmtSF | 1041 | sqft |
| 1stFlrSF | 1134 | sqft |
| 2ndFlrSF | 2034 | Not directly provided, but the house is a two-story structure, so the implied value is 2034 sqft |
| GrLivArea | 2034 | sqft |
| FullBath | 1 | - |
| HalfBath | 3 | 3/4 bathrooms: 2 1/2 bathrooms: 1 See them as half-bathrooms |

| | | |
|---------------|--------------------|--------------------------------------------------------------------------------------------------------------------|
| BedroomAbvGr | 3 | - |
| TotRmsAbvGrd | 12 | Total rooms not directly specified, but includes living room, dining room, kitchen, and bedrooms, totally 12 |
| Fireplaces | 2 | 2 fireplaces (one on the main level and one in the basement) |
| GarageCars | 2 | - |
| GarageArea | 600 | Not specified, but the property includes a garage so implying the value by observing the floor plan |
| WoodDeckSF | 600 | Not specified, but the property includes a new multi-tiered deck so implying the value by observing the floor plan |
| OpenPorchSF | 200 | Not specified, but the property includes an open front porch so implying the value by observing the floor plan |
| EnclosedPorch | 0 | Not mentioned in the listing |
| Neighborhood | Northridge Heights | - |
| BsmtQual | 5 | Likely high, inferred from the fully finished basement with living quarters |

Now we have defined its features and price, then we used the same scaler to normalize these features and price and inputted the standardized features to get the predicted value, which is \$ 396,849.1, then we computed the predicting error:

$$\text{Predicting Error} = \frac{|\text{Actual Price} - \text{Predicting Price}|}{\text{Actual Price}}$$

The predicted Error is 13.2%. Our model still made a good prediction in the current house market!

6. Summary

In essence, the analysis of housing prices in the study highlights how different property features greatly influence real estate prices. The model developed through rigorous data preparation, feature encoding, and regression analysis demonstrates a strong predictive capability, achieving an R-squared value of 0.9013 and a mean absolute error of 0.2146. This suggests that the model is capable of accounting for over 90% of the variability in housing prices while maintaining a low error margin.

The analysis of neighborhoods indicates that the location significantly impacts property values. For instance, Northridge Heights fetch prices due to amenities and infrastructure while areas like Bluestem and Northpark Villa see a decrease in property values possibly due to factors such as crime rates and limited service accessibility.

Moreover the analysis of features pinpoints characteristics that boost housing values such as the quality of the property and the size of the living space along with having completed basements available for residents to use effectively. Conversely, some aspects like the number of bedrooms and unfinished basement spaces are shown to have an effect on the worth of a property.

The model's successful use, in forecasting the value of a property at 3206 Aspen Rd in Ames demonstrates its application in real-life situations well. The predicted price of \$ 396,849.1, with a prediction error of 13.2%, underscores the model's effectiveness in estimating housing costs in the current market.

In terms, this project doesn't just offer perspectives on the elements that impact housing costs but also sets up a strong structure, for upcoming real estate price forecasts assisting buyers sellers, and investors, in making well-informed choices.

7. Reference

Hull, J. (2021). *In Machine Learning in Business: An introduction to the world of Data Science*. Independently published.

Zillow. (2024). *Northridge 3206 Aspen Rd*. Zillow.
https://www.zillow.com/homedetails/3206-Aspen-Rd-Ames-IA-50014/93953617_zpid/