Department of Industrial and Systems Engineering
UNIVERSITY OF WISCONSIN-MADISON

ISyE 521: Machine Learning in Action for Industrial Engineers

Sales Price Prediction of Houses in Ames, Iowa

Submitted by
Kannan Walter
Prageeshwar Chandran
Varun Srinivasan Kalaiselvan

Introduction

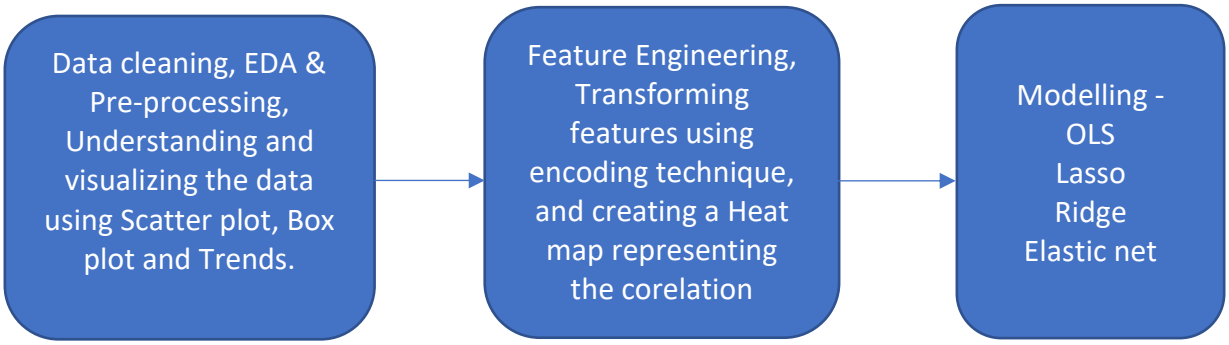
Buying a house is an ambition for many in the United States but it requires a lot of time, effort and most importantly money. Potential buyers approach homes with aspirations for their future but they eventually end up in the tricky situation of finding themselves evaluating the price quoted by the owner of the house. Some might feel stuck, some back out and others go ahead with the purchase without considering whether the price quoted by the owner is optimal or not. For the buyers to find this optimal price, several questions are to be answered. What drives the price of homes? Are the buyers aware of the important factors that influence the price of the homes? Is it the number of bedrooms, the kitchen area or the living area that determines the price? Most buyers don't have answers to all these questions, but they needn't have answers to all, they just need to have an answer for one. All these questions could be resolved with an answer to a simple question, do buyers require a secondary opinion on whether they are paying the right price? If Yes, could it be done?. Secondary opinions are sometimes irrelevant and might go against your beliefs but in this case, with the amount of money involved along the invaluable aspirations that the buyer has, a secondary opinion would do a world of good. The answer to the question if a secondary opinion could be developed is a resounding, yes. This report has addressed this exact question. This has been done by building models using regression techniques. It is to be noted that the price of a real estate property is sophisticatedly linked with our economy. Building a model that predicts the price based on important features can help the buyer to know whether the asking price of a house is higher or lower than the true value of the house. It will also bring transparency and trust between the buyer and seller while improving the real estate sector.

Dataset Information

The dataset is obtained from the Kaggle website under the title "House prices – Advanced Regression". The dataset has records of 1460 houses in Ames, Iowa and covers 79 different attributes.
https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Methods

It is important that a structured methodology is followed in a machine learning process because it helps to ensure that the process is well-defined, repeatable, and effective. By following a structured methodology, it is possible to build high-quality machine learning models that are accurate, reliable, and able to effectively solve real-world problems. As the first step of our structured approach, we performed data cleaning and then conducted Exploratory Data analysis to understand the features. After completion of the EDA, we started to build a baseline model using Ordinary Least squares method. Then we tried to improve the prediction accuracy by using regularization techniques such as Lasso, Ridge, and Elastic net. The images below show our workflow while undertaking this project.



Data Cleaning

Data cleaning is essential because, regardless of how sophisticated our ML algorithm is, we can't obtain good results from data that is not complete. So, the first step was to analyze the data frame for empty cells. This was visualized and the image below represents the same. The white spaces in the image represent cells in each feature that do not have entries.

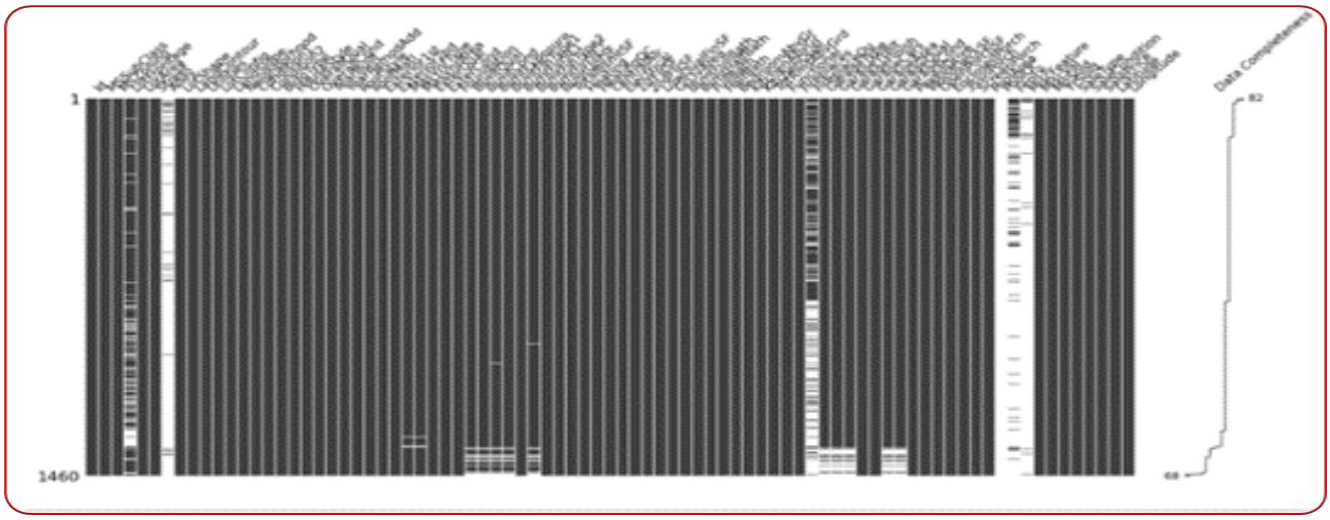


Figure 1: Data Missingness

Upon further analysis it was found that these cells had NaN values. To address this problem, the data description file was referred. It was found that NaN values did not necessarily mean null values but rather the absence of that utility. For example, NaN values in the 'Alley' Feature meant that the house did not have alley access and not an empty cell. This needed to be addressed, so the change was made in the data frame to reflect this. The image below shows the code written to replace the NaN values.

```
df['Alley'].replace(np.nan, 'No alley access', inplace=True)
df['BsmtQual'].replace(np.nan,'No Basement', inplace=True)
df['BsmtCond'].replace(np.nan,'No Basement', inplace=True)
df['BsmtExposure'].replace(np.nan,'No Basement', inplace=True)
df['BsmtFinType1'].replace(np.nan,'No Basement', inplace=True)
df['BsmtFinType2'].replace(np.nan,'No Basement', inplace=True)
df['FireplaceQu'].replace(np.nan,'No Fireplace', inplace=True)
df['GarageType'].replace(np.nan,'No Garage', inplace=True)
df['GarageFinish'].replace(np.nan,'No Garage', inplace=True)
df['GarageQual'].replace(np.nan,'No Garage', inplace=True)
df['GarageCond'].replace(np.nan,'No Garage', inplace=True)
df['PoolQC'].replace(np.nan,'No Pool', inplace=True)
df['Fence'].replace(np.nan,'No Fence', inplace=True)
df['MiscFeature'].replace(np.nan,'None', inplace=True)
```

Figure 2: Conversion of NaN values

Since there were several records, it is possible that there might have been duplication of records. If that were the case, then it would affect the model and our subsequent predictions as well. After checking the data frame for such scenarios, it was found that there were no

duplicates in the data frame (Refer Figure 3 below), and it made sense to conclude the data cleaning and proceed to the next step, exploratory data analysis.

```
There are 0 duplicates in dataset.
```

Figure 3: Result of the check for duplicates

Exploratory Data Analysis

Explanatory Data Analysis is the next step before we build any model. EDA helps one to determine how best to manipulate the data features to get answers that are needed, thus making it easier to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

Figure 4: Distribution of the homes

Generally, the most important factor which influences the house price in the United States is, the neighborhood. To verify this theory and check if neighborhoods close to each other have similar prices, the houses were needed to be plotted against the map of Ames, Iowa. Upon referring to the dataset it was found that though the dataset had the feature of Neighborhood it didn't have the longitude and latitude features, which are essential to plot houses against the backdrop of a map. So, using feature engineering, two new features 'Latitude' and 'Longitude' were created to study the distribution of homes based on neighborhood. Using these features and with the help of the folium library, neighborhoods were plotted in the Ames, Iowa map to understand the relationship. The average sale price for each neighborhood is calculated and it is observed that the neighborhood of NoRidge has the highest Average Sale Price of $335,296.32. Figure 4 depicts the top 5 neighborhoods. Furthermore, an average sale price distribution is created to visualize the average sale price of all neighborhoods, figure 5 depicts the same. This enables us to compare the neighborhoods and test the theory.

	Neighborhood	Avg SalePrice
13	NoRidge	335295.317073
15	NridgHt	314313.657895
22	StoneBr	310499.000000
23	Timber	244267.648649
24	Veenker	238772.727273

Figure 5: Top 5 Neighborhood based on Average Sale Price

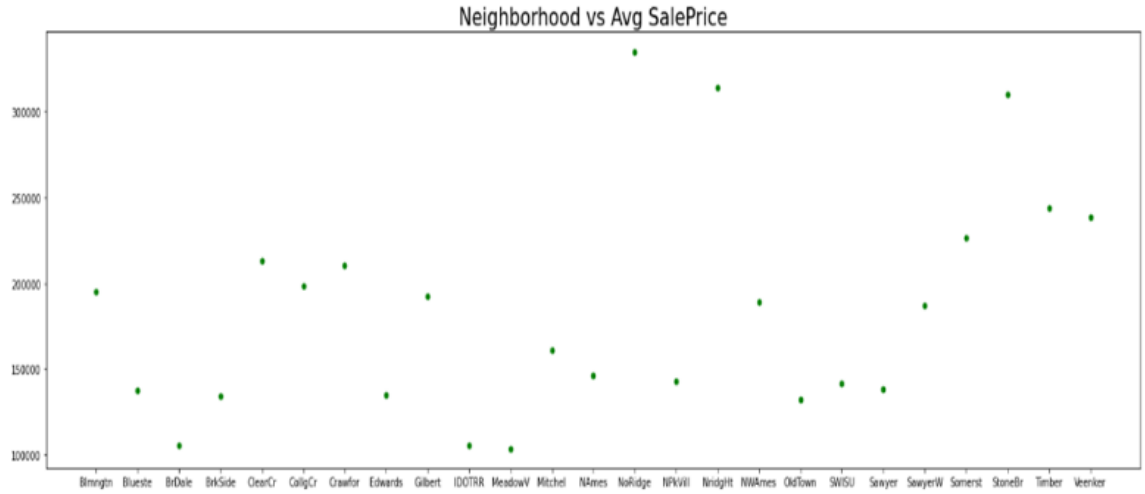


Figure 6: Distribution of the homes

While one can identify neighborhoods which had the highest average sale prices of homes, it wasn't the case for the theory, in fact the difference in the average sale price was quite large among neighborhoods that were close. It can thus be concluded that homes in adjacent neighborhoods don't necessarily have similar sale prices.

Important features affect a model's performance heavily. To determine the correlation between the features in the dataset, a heatmap was created, which included all features. As we are building our model with sale price being our target feature, a list of positively correlated features with sale price is determined and figure 8 depicts the same.

SalePrice	1.000000
OverallQual	0.786304
GrLivArea	0.709783
GarageCars	0.636173
GarageArea	0.607197
TotalBsmtSF	0.603284
1stFlrSF	0.596087
FullBath	0.558902
TotRmsAbvGrd	0.541189
YearBuilt	0.508127
YearRemodAdd	0.506063
GarageYrBlt	0.486658
MasVnrArea	0.468055
Fireplaces	0.450948
BsmtFinSF1	0.373057

Figure 7: Positively correlated features to sale price

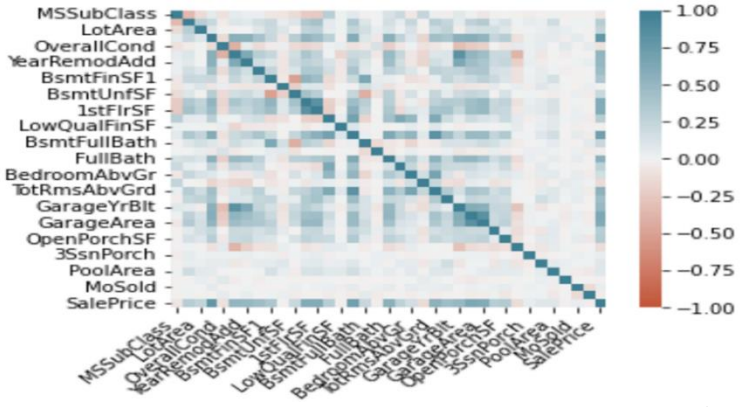


Figure 8: Heatmap depicting the correlation between the features

The feature ‘OverallQual’ is highly correlated with sale price and is our most important feature. The relationship between the two features is plotted against that of the heating system feature in a 3-dimensional scatter plot, which is given in Figure 8. To find the trend between the features and the sales price of houses, creation of a variety of visualizations like scatter plot, box plot, line graph is implemented against the sales price feature.

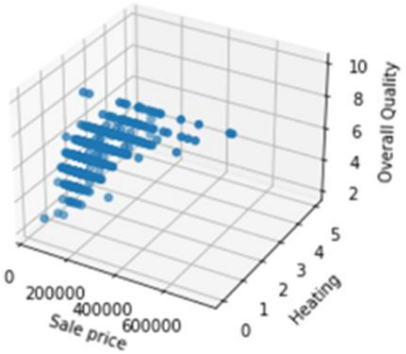


Figure 9: Heating System Vs Sales Price Vs Overall Quality

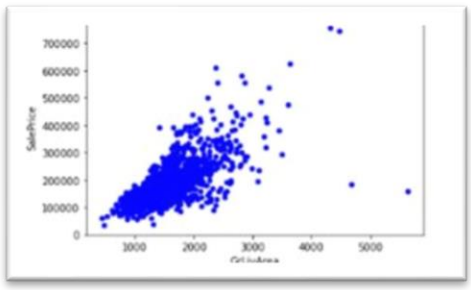


Figure 10: Living Area Vs Sales Price

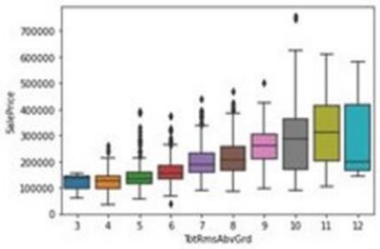


Figure 11: No. Of Rooms Vs Sales Price Price

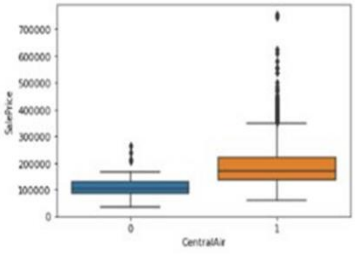


Figure 12: Central Heat Vs Sales Price

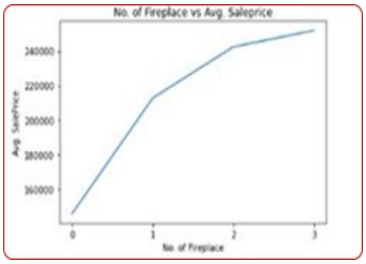


Figure 13: No. of fireplace Vs Sales Price

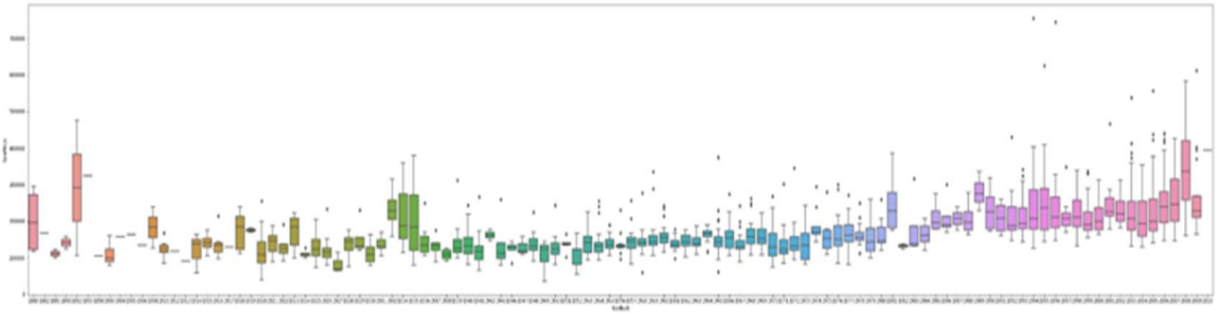


Figure 14: Year Built Vs Sales Price

Regression Techniques

Linear regression techniques are one of the simplest and most efficient methods to predict values of the target variables based on the values of other variables provided. Regression makes predictions based on studying the strength of the relationship between the variables. Our dataset has 79 features and regression techniques are easy in handling large number of features. Our prime objective is to include as many features as possible in our prediction, so the model will be more reliable and versatile.

We started to build a baseline model using Ordinary Linear Squares (OLS) regression technique as this type of regression is the natural choice for baseline for any regression problems. OLS regression is a simple and widely used method for estimating the coefficients

of a linear regression model. It is a powerful tool for understanding the relationship between variables and for making predictions based on that relationship.

To further improve and compare our results, regularization techniques such as Lasso, Ridge, and Elastic net are introduced. Regularization is a technique used to prevent overfitting in regression models. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the amount of data. This can lead to a model that performs well on the training data but does not generalize well to new data.

Lasso, which stands for Least Absolute Shrinkage and Selection Operator, is a regularization method that applies a penalty equal to the absolute value of the magnitude of the coefficients. This has the effect of forcing some of the coefficients to be exactly equal to zero, which can be used to perform feature selection by eliminating irrelevant or redundant features.

Ridge regression, on the other hand, applies a penalty equal to the square of the magnitude of the coefficients. This has the effect of shrinking the coefficients, but it does not set any of them to exactly zero. As a result, ridge regression is not capable of performing feature selection, but it can still be useful for preventing overfitting.

Elastic net is a combination of lasso and ridge regression. It applies to both an L1 penalty, which is the same as the penalty used in lasso, and an L2 penalty, which is the same as the penalty used in ridge regression. This allows elastic net to achieve the benefits of both lasso and ridge regression, including feature selection and shrinkage of the coefficients.

In summary, the main difference between lasso, ridge, and elastic net is the way that they apply regularization penalties to the coefficients of the regression model. Lasso uses an L1 penalty, ridge uses an L2 penalty, and elastic net uses a combination of both.

Results

We achieved an R-squared value of 0.845 in our baseline model using Ordinary Least Squares (OLS) method. We were then able to improvise our model using regularization techniques. By using Lasso, we were able to get an R-squared value of 0.8590 and Ridge and Elasticnet models generated the R-squared value of 0.8586 and 0.855 respectively.

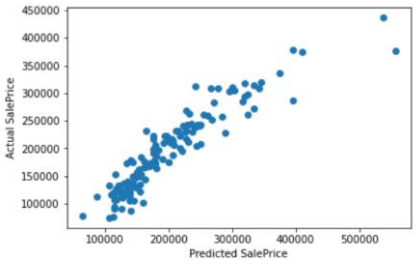


Figure 15: Predicted Sale Price Vs Actual Sale Price using Lasso

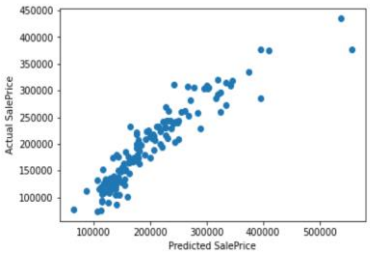


Figure 16: Predicted Sale Price Vs Actual Sale Price using Ridge

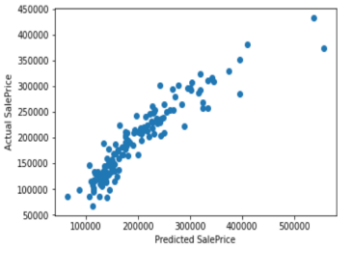


Figure 17: Predicted Sale Price Vs Actual Sale Price using Elastic Net

\

As we can see by using regularization, we were able to enhance our results and the Lasso model gave us the best results among the four models.

Conclusions

Using the model's predictions, prospective home buyers located in Ames, Iowa can use the predictions as a secondary reference to evaluate the price quoted by the seller. Generally, the buyers are not usually aware of the factors which influence the price of the home, but using this model helps the buyers to eliminate the middleman e.g., real estate agents, cost of transfer fees and finally help to make an informed decision. Also, predicting house prices can help to support the development of new financial products and services. For example, banks and other financial institutions can use house price predictions to develop more sophisticated mortgage products and services that are better tailored to the needs of their customers.

Future Directions

We have built a model that predicts the house price of houses in Ames, Iowa. In future the method can be extended to predict the price of houses for a bigger region. We could also enhance the model by adding few more attributes such as school district, public transport facilities, etc. By adding these attributes, we may be able to enhance the results by being more specific and will be effective. This approach will also make the model more versatile. In the perspective of improving the model performance, using Random Forest method could give us better solutions, because we have considerable number of features which would give us the chance to train many trees for the model.