House Sale Prices Data Analysis Report

Overview

In this project, we conduct data processing and visualizing on the house prices data with explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa.

In order to have a general understanding of the data, we plot the histogram of the sale prices distribution, the boxplot of sale prices, the boxplot of sale prices in different neighborhoods and the boxplot of sale prices in different house style.

We also apply machine learning techniques to train a linear model that can predict the house given features. First we preprocess the data such as dealing with empty data, then we use cross-validation to choose model parameters, train the model and make the prediction on test data.

We can learn how various features contribute to the final house sale prices from the linear model coefficients. The absolute value of the coefficient indicates the weight of the feature; and the sign indicates it is a positive factor or a negative one.

Data exploration

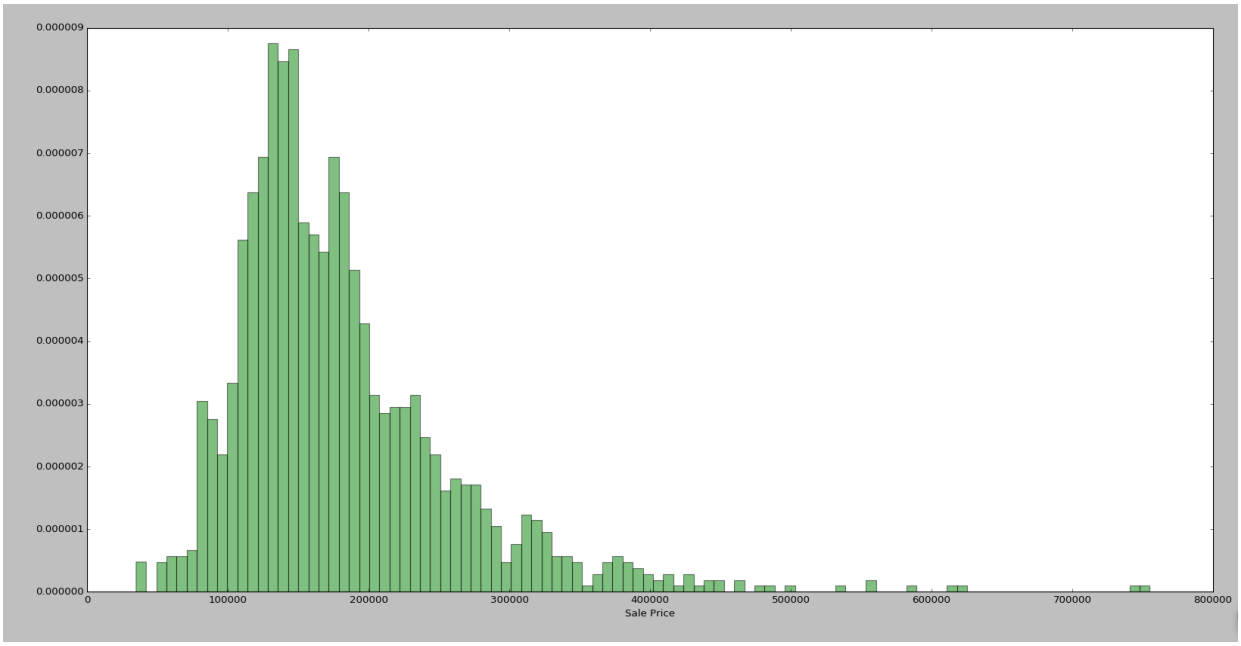

Figure 1. House Sale Prices Histogram

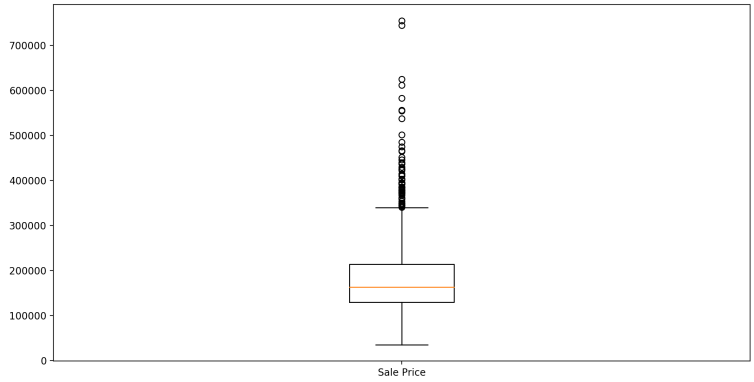

Figure 2. boxplot of sale prices

We can easily tell that the house prices are most in the range between 100000 and 200000.

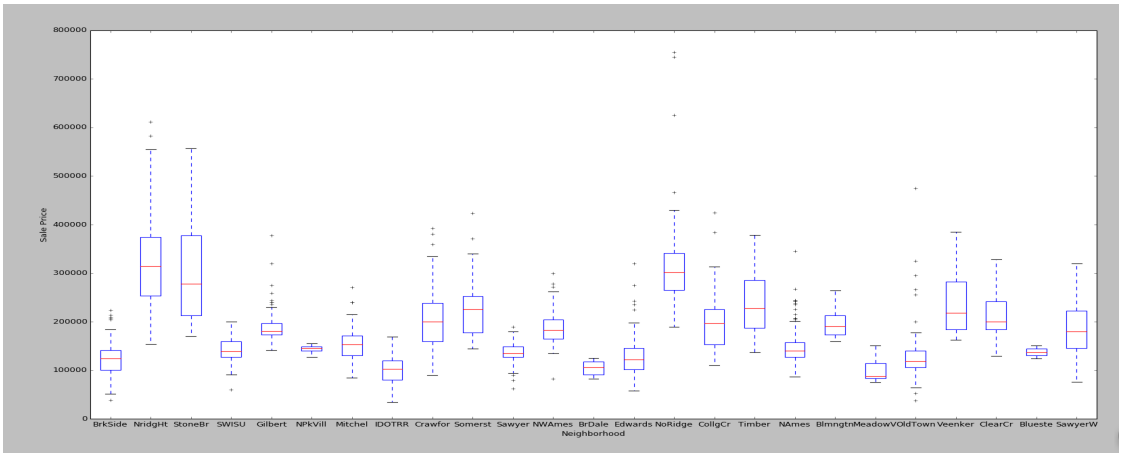

Figure 3. Sale Prices Boxplot of Different Neighborhoods

We can find that the most expensive houses are in NridgHt and NoRidge neighborhoods, as much cheaper houses are in Edwards and Meadow mostly.

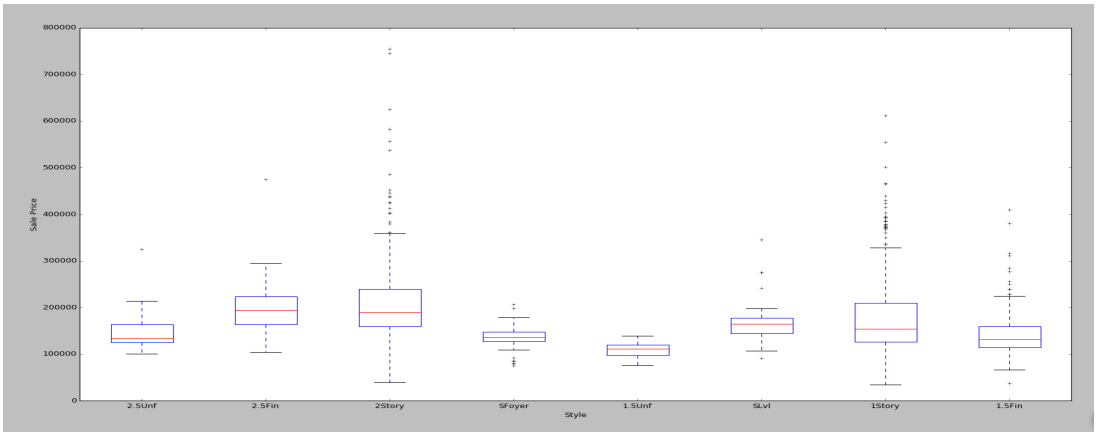

Figure 4. Sale Prices Boxplot for Different House Style

From the result, there is no significant difference among the medians of sale prices of different house styles. However, we can see the big gap between maximum and minimum prices of 2Story and 1Story houses.

Linear Model

- **Data Preprocessing**

Linear regression model will have better performance when applied to the data that have similar distribution with normal distribution. But some features of the house prices data are very skew, so we use log transformation to let them more approximate to the normal distribution. We make the formula as: new_x = log (x + 1).
The LotArea feature data after transformation is as followed, and we can see it becomes more “normal”.

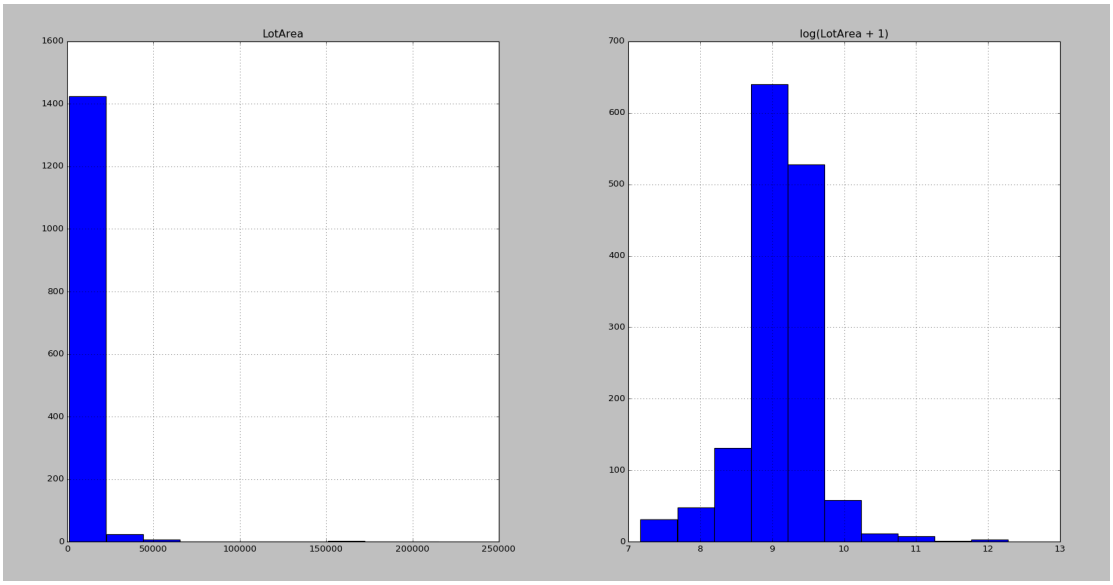

Figure 4. LotArea before and after transformation

Next we convert categorical variables to indicator variables to be used in the linear model. For example, the CentralAir has two categories “Y” or “N”. This feature is converted to features CentralAir_Y and CentralAir_N. The CentralAir_Y value of a sample is either 1 or 0, and it is 1 if and only if the CentralAir is “Y”. The CentralAir_N value of a sample is either 1 or 0, and it is 1 if and only if the CentralAir is “N”.

Then, we will deal with an empty value. We can simply discard features or samples that have an empty value, but that will make the data much smaller. Another strategy is to replace the empty value with a default value. We choose to replace the empty value with a mean value. Since we don’t know its value, we can just assume it is an average one.

- **Model Training**

We choose Lasso linear model in the scikit-learn. Because the Lasso model can automatically choose useful features for us and remove features that are not so important.

It has a regularization parameter alpha to avoid overfitting. Different alpha may have very different performance. So we use cross-validation to select the best parameter. From the candidate values [0.00001, 0.0001, 0.001, 0.01, 0.1], the cross-validation choose alpha = 0.001 which has the mean squared error [0.01243842 0.0186352 0.01735224] for 3-fold cross-validation. Then we can train a Lasso linear model on the training data. We just use the fit method with max iteration set to 10000 which means the algorithm will run at most 10000 iterations to optimize the result.

After training, each feature of the model has a coefficient, this coefficient directly determined how this feature will contribute to the final sale prices. If the coefficient is 0, then the model doesn't consider this feature. The result shows that our trained model selects 80 variables and removes 208 variables.

Then we extract the 10 features that have the largest coefficient which means they have the most significantly positive influence on sale prices and the 10 features that have the smallest coefficient which means they have the biggest negative influence.

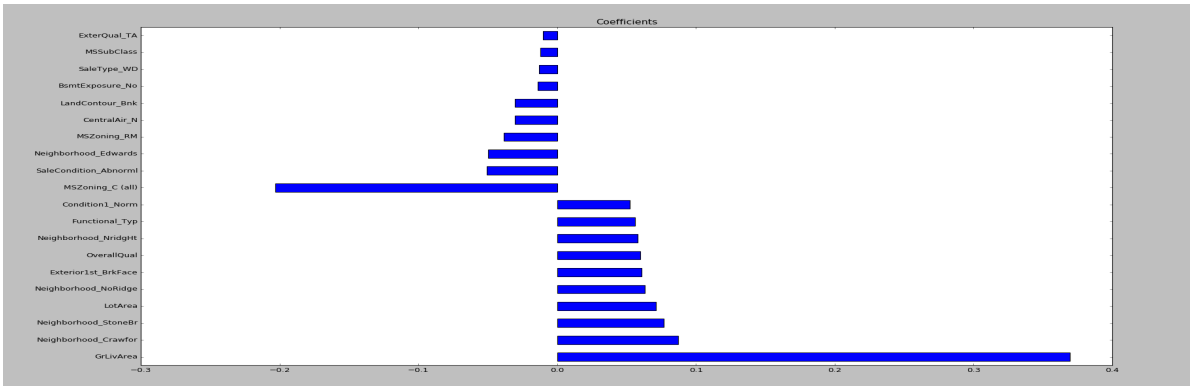

Figure 5. Largest 10 and smallest 10 feature coefficients

The largest positive coefficient feature is GrLivArea which is reasonable, the above grade (ground) living area has the greatest impact on prices. There are 4 Neighborhood indicator variables in the largest 10 which indicate that the Neighborhood has big influence on house prices. OverallQual among the largest 10 also have common sense, since overall material and finish quality can obviously influence the prices.

10 smallest coefficients are also fully justified by the reason why they have big negative impact on sale prices. The general zoning classification of type C influences prices mostly in the negative way. Abnormal sale condition, bad neighborhood, no central air and so on are all the factors would discount the value of houses.

In summary, we can see how various features related to house prices through their coefficients, and we can easily tell several biggest influence features the model found are quite reasonable.

At last, we use our trained model to predict on the test data and save to file.