METCS521 Final Project Write-Up

Predicting Housing Price

Introduction:

In this project, we explore a dataset containing dozens of housing-related descriptive variables. Our goal is to select a number of variables that have a large impact on the selling price of a home and implement approachable machine learning algorithms. These algorithms can use the variables we ultimately select to help sellers of homes as well as potential buyers predict the price of a home. Throughout the project, we created two jupyter notebooks and the order of the projects was as follows：

1. Download dataset
2. Exploring & integrating data variables that are subjectively considered relatively desirable
3. Visualize the data variables to make the primary analysis
4. Make a correlation analysis & linear regression
5. Check the coefficients of the variable to the model to optimize the model

Downloading the dataset

In order to reproduce the dataset, the users need to run **Downloadcode.ipynb script.** The script will produce a csv file to your current directory. Next, the users will need to run **Analysis.ipynb** will read this csv file into a data frame and will able to see our work.

Exploring & integrating variable data variables that are subjectively considered relatively desirable

Before we visualize or analyze the data, we briefly glanced at the data. It was found that there were 81 variables in total, each with more than 2000 records of housing data. Since sale price is the main variable of housing prediction, we initially subjectively judge that these missing values may affect the subsequent modeling and analysis.

To facilitate the processing and integration of the data, we imported a module called Panda. Since most of the models need to be analyzed based on the relationship between the sale price and other variables, we used dropna to remove the data records with missing values in the sale price.

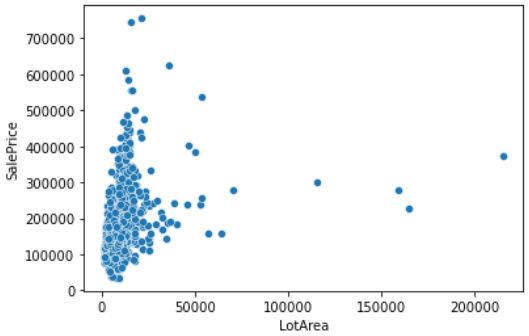
Visualize the data variables to make the primary analysis

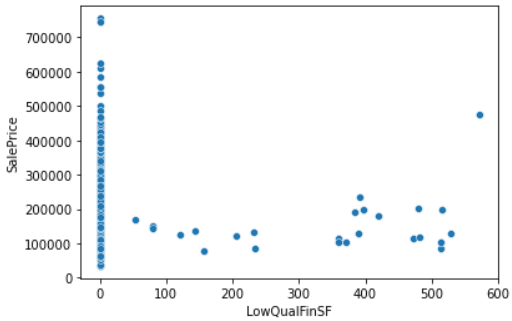
We have studied the variables used in our competitor's product Zillow and selected 15 variables for visualization, including the size of the house, the number of toilets, the size of the garage, and the year of completion and reconstruction.

We made a preliminary determination of what type of visualization to use for different variables. We used bar charts for non-numerical variables like housing type and for variables like number of bathrooms, which require a more visual observation of the difference in the value of the dependent variable when the value of the independent variable varies. On the other hand, for variables such as housing area, where the amount of data is large and the trend of the variable needs to be observed, we use a scatter plot.

We found that among the variables for which we used bar charts, except for the variable related to the number of toilets in the basement, the rest of the variables had larger differences in their sale price when the values were different. However, since we consider that the bar chart is not able to observe the existence of extreme values (because the bar chart takes the average value), we do not make preliminary subjective judgments and analysis of these variables for the time being.

On the other hand, we found that the distribution of Lot Area and LowQualFinSF (Low Quality Finished Square Foot) in the scatter plot was too concentrated and visually irregular, so we initially judged that the data of these two variables would most likely have little impact on the analysis.



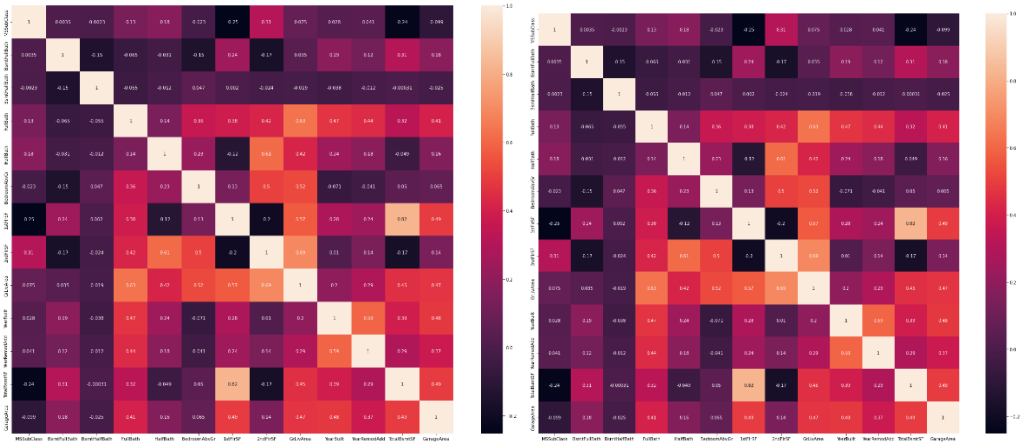


Make a correlation analysis & linear regression(Extension 1 & 2)

We regrouped the selected fifteen variables into a new data frame and dropped the missing values in these variables. Next, in order to prepare for the linear regression analysis, we set up the independent and dependent variables.

In order to way the presence of multicollinearity, we did an analysis of the correlation between predictor Fortunately, there are no cases of strongly correlated predictors among the values of the variables we chose (we define a strong correlation as a correlation between the two that is greater than 0.9)

Next, in order to test the accuracy of the model on the one hand, and to verify whether the model is over-fitting on the other hand, we divided the 15 variables into two groups, 70% of the variable data were assigned to the training group and 30% of the variable data were assigned to the testing group, and then we used the training data to construct the regression analysis model. We examined the accuracy of the training group and the testing group separately and found that, unlike the conventional case, the accuracy of our training group was slightly lower than that of the testing group, but the difference was within an acceptable range. Finally, we used the existing model for housing sale price prediction (where the input values are within the range of the minimum to maximum values of each variable in the data set)



Check the coefficients of the variable to the model to optimize the model(extension 3)ok

Before we finished our first housing prediction, we checked the coefficients of all independent variables and found that the coefficients of Lot Area and LowQualFinSF were very small compared to the coefficients of the other variables, which confirmed our prediction at the beginning of the data visualization that these two variables might not have a significant effect on the sale price.

To further optimize our model, we removed the variables with low coefficient and re-ran the regression analysis and did the accuracy test again. The accuracy difference between the training group and the testing group is still within the acceptable range.

Finally, we made another housing sale price prediction with the new model. With the same input values, the output value produced an error of more than 0.3% with the last prediction, so we believe that the model based on this set of dataset is relatively stable.

Conclusion

We observed and explored the data to find that the number of rooms and restrooms and the ground area were the decisive factors affecting the modeling. In our initial selection of variables, we were fortunate to be able to successfully hypothesize these variables and not find any multicollinearity between them through our own subjective knowledge and experience, as well as by referring to other platforms such as Zillow and apartment.com, which are competing simulations. In order to optimize our model, we finally chose to remove the data variables with small coefficients such as Lot Area, which may have a negative impact on the prediction model, and thus we believe that our prediction model based on this dataset has stabilized and can be used for real prediction.