House Price Prediction: Solutions from Linear Regression Model

M.Sc. in New Media Programme, School of Journalism and Communication

COMM 5962: Topical Studies in Communication: Data Science in New Media

Zeng Shuman 1155160734

April 23, 2022

0. Problem Statement

When considering to buy a house, the house price plays an important role in deciding whether to buy it. Then we may wonder whether the house price is appropriate or not? To answer this question, we are supposed to know what factors will influence the house price. We may consider many conditions such as the area, location, transport convenience, the quality of building and so on. Therefore, this essay would like to discuss the importance of factors which will affect the house price and how can we build a model to predict the house price? And the predicting result could be a reference comparing with the real sale price of the house.

Data source: Kaggle (https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data)

By the way, in order to have a better standing of the whole work, please go to see the coding version in the GitHub.

GitHub (Coding Version): https://github.com/nkdxgxh/COMM-5962_Data-Science-in-New-Media/blob/main/House%20Price%20Prediction%20Solutions%20from%20Linear%20Regression%20Model.ipynb

1. Descriptive statistical

In the first stage, I started from dealing with the problem by having a good understanding of the dataset downloaded from Kaggle, which focused on the descriptive statistical part.

1.1 Basic information

First of all, in order to better understand the dataset, the variable should be well clarified. The dataset is collected from Kaggle. The dependent variable is 'SalePrice', which is the target I should make a prediction. According to the dataset, there are totally 78 independent variables which could be divided into two groups, namely numerical variable and categorical variable. The data types also vary from 'float64(11)', 'int64(26)' to 'object(43)'.

And each variable is clarified in the appendix file.

1.2 Correlation matrix

As we know, in this dataset, 'SalePrice' is the dependent variable for what we want to predict. And the rest of variables are regarded as independent variables. Thus, it's important to learn the correlation between them. Here I utilized a correlation matrix with heatmap visual style to clarify the relationship (Figure.1).

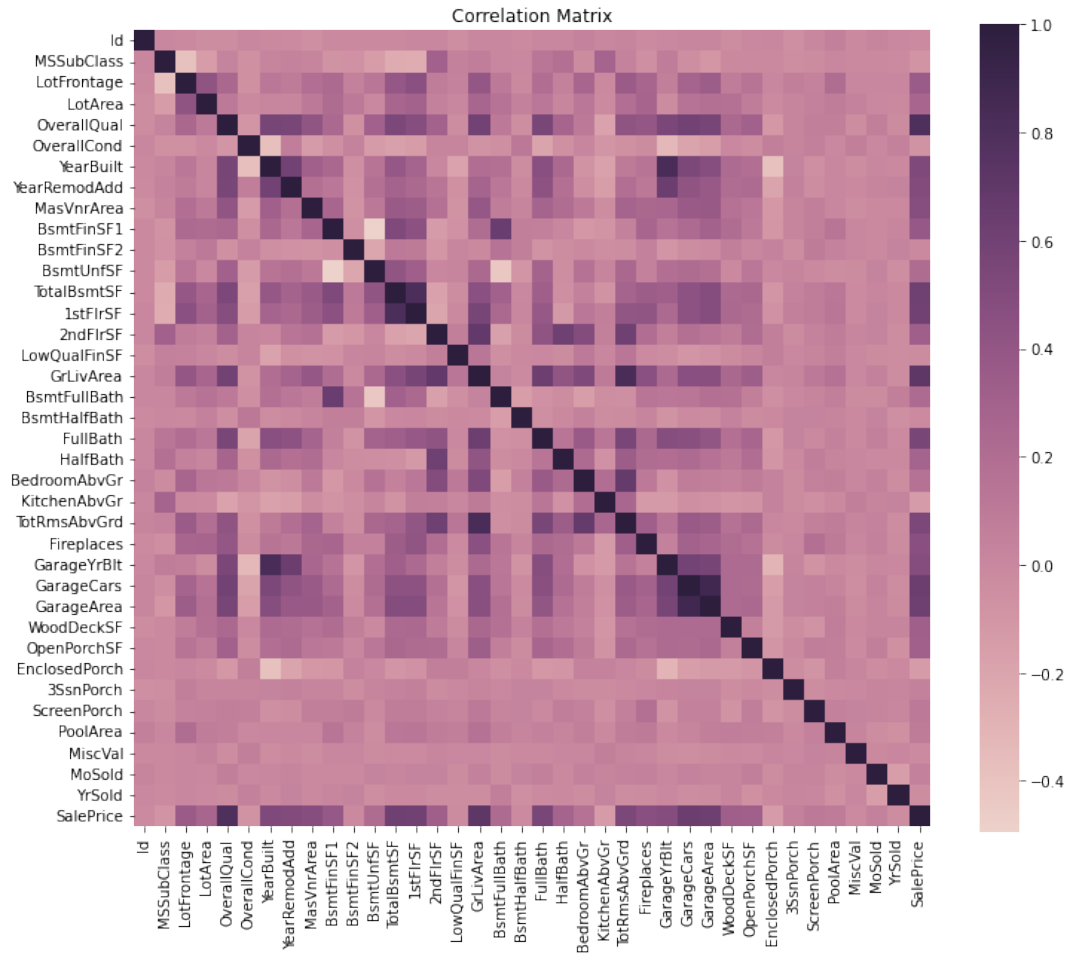

Figure.1 Correlation martix

According to Figure 1, we can learn:

1. OverallQual is highest related to SalePrice. The house price would increase as the quality of the house gets better.

2. GrLivArea is related to location, which illustrates the natural environment plays an important role in deciding the house price.

3. GarageCars and GarageArea also show high correlation with SalePrice. Despite that, these two variables are also highly correlated with each other, which shows a tendency of collinearity. Thus, these two variables should be regarded as one variable and then delete one of them, otherwise, it would affect the accuracy of the prediction.

In summary, there are totally 79 variables in the dataset. As can be seen from the result of correlation matrix (Figure.1), variables such as 'OverallQual', 'YearBuuilt', 'GrLivArea' are highly correlated to 'SalePrice'. However, there are some variables showing low correlation at the same time such as 'KitchenAbvGr', 'EnclosedPorch', 'OverallCond' and so on, which are not supposed to be taken into consideration. Therefore, a more accurate correlation matrix should be created.

2. Data processing

Before we can get further insights from the data, it is supposed be to processed by dealing with the missing values, converting the categorical variables and so on.

2.1 missing value

In order to have a full picture of the missing values in the dataset, a visual bar chart could help to illustrate that (Figure.2).

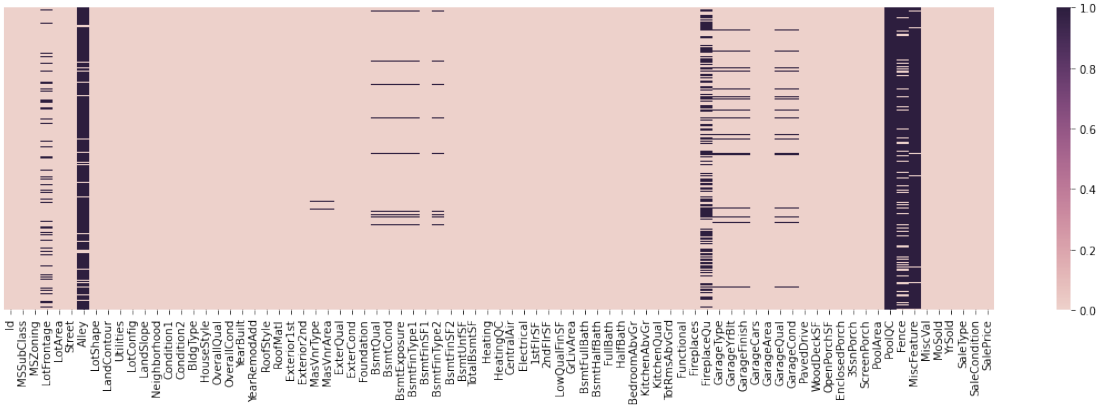

Figure.2 Missing value

From the picture above, we can notice that several variables have a large percentage of missing values such as 'Alley', 'PoolQC', 'Fence', 'MiscFeature' and so on. And the followed table examines the exact ratio of missing values from each variable.

	Total	Ratio
PoolQC	1453	99.520548
MiscFeature	1406	96.301370
Alley	1369	93.767123
Fence	1179	80.753425
FireplaceQu	690	47.260274
LotFrontage	259	17.739726
GarageYrBlt	81	5.547945
GarageCond	81	5.547945
GarageType	81	5.547945
GarageFinish	81	5.547945
GarageQual	81	5.547945
BsmtFinType2	38	2.602740
BsmtExposure	38	2.602740
BsmtQual	37	2.534247
BsmtCond	37	2.534247
BsmtFinType1	37	2.534247
MasVnrArea	8	0.547945
MasVnrType	8	0.547945
Electrical	1	0.068493
Id	0	0.000000

Figure.3 Percentage of missing value

In terms of the figure 3, more than 80 percent of data from variables ‘PoolQC’, ’MiscFeature’, ‘Alley’, ’Fence’ had lost, which could highly affect the accuracy of results if these variables are included into the predicting model. Moreover, the valid data from ‘FireplaceQu’ and ‘LotplaceQu’ are less than 85 percent, which should be eliminated from the dataset too. It’s proved that data with a missing value of more than 15% are considered invalid. Thus, according to the result, 'PoolQC', 'MiscFeature', 'Alley', 'Fence', ‘LotplaceQu’ and 'FireplaceQu' are supposed to be removed. Additionally, variable of ‘Electrical’ just has one column of missing value according to figure 3, which could be simply deleted.

2.2 Data transformation

Figure 4 presents the histogram of 'SalePrice'. As can be seen, the value of 'SalePrice' shows a right skewed tendency (Figure.4).

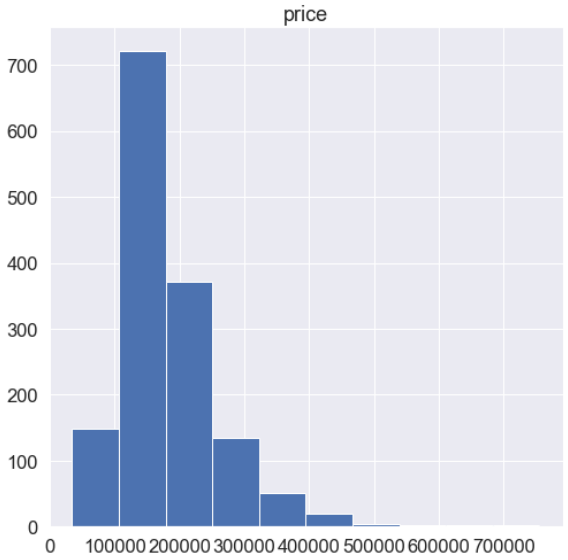

Figure.4 Histogram of SalePrice

In order to make the feature more normal to better fit in the prediction model, here I deal with the data by taking log transformation (Figure 5). After the transformation, the data below shows a normal distribution, which will reduce errors in the prediction model.

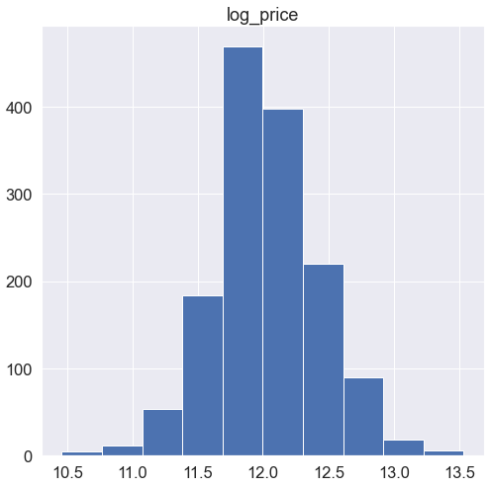

Figure.5 Histogram of SalePrice after transformation

In addition, the rest of data belongs to ‘object’ also would be processed by ‘log transformation’.

3. Feature engineering

After processing the data roughly, followed step is to observe the correlation and obtain the important feature variables via correlation matrix.

3.1 Correlation again: Top10 correlation matrix

As mentioned above, a more detailed correlation should be examined in this stage. Thus, the correlation matrix with top 10 correlation coefficient is shown as follows (Figure 6).

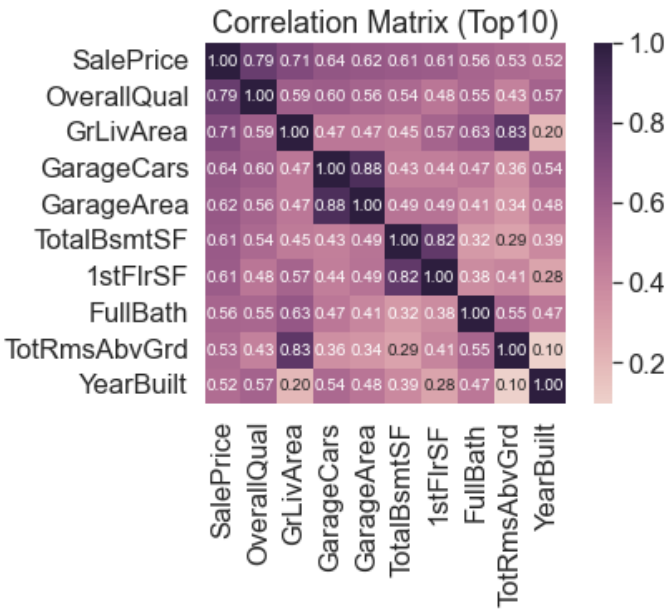

Figure.6 Correlation matrix (Top 10)

According to the figure of correlation matrix (Top10), we can learn the variables which are most related to 'SalePrice':

1. 'OverallQual (0.79)' is highest related to 'SalePrice'.
2. Followed is 'GrLivArea (0.71)', which illustrates the significance of natural environment in deciding the house price.

3. 'GarageCars' and 'GarageArea' both show high correlation with 'SalePrice (0.64/0.62)', but the correlation between these two variables is also high (0.88). In order to avoid the collinearity, ‘GarageArea’ would be removed.
4. 'TotalBsmtSF' and '1stFlrSF' seem have the same value of the correlation coefficient (0.61) with ‘SalePrice’. Meanwhile, ‘TotalBsmtSF’ and ‘1stFlrSF’ are highly correlated with each other, and here ‘1stFlrSF’ is removed.
5. It's interesting that 'FullBath (0.56)' is slightly correlated with 'SalePrice' here. Because 'FullBath' means full bathrooms above grade in the house and I don’t figure out why it is important.
6. 'TotRmsAbvGrd' is highly correlated to 'GrLivArea'(0.83), which should also be removed in order to avoid the collinearity.
7. 'YearBuild (0.52)' also shows slight correlation with 'SalePrice'. In reality, people would consider the construction date of the house, which could affect the house price.

3.2 Dealing with Outliers

First, I am dealing with the outliers from feature variable ‘GrLivArea’. As can be seen from the chart (Figure 7), there are two outlier points in the bottom-right corner, which should be deleted.

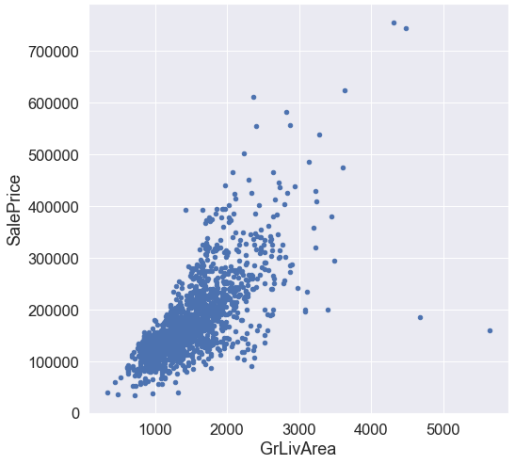

Figure.7 Outliers of GrLivArea

3.3 Categorical variables transformation

As mentioned before, there are several categorical variables among the feature variables I picked. Thus, in this step, I will transform these categorical data into numerical format by using the 'get_dummies()' method of python.

Finally, I get a prepared dataset with 1092 rows and 221 columns.

4. Model

Before put the data to train the prediction model, here I employ 'Sklearn' package from python to spilt the dataset into training dataset (70%) and testing dataset (30%). Since it is the first time for me to complete a data science project individually, I would like to try something simple. Following are the prediction models which I can handle with. Firstly, I am going to use linear regression models from the scikit learn module. Then, I am planning to try both Lasso and Ridge regression models which belong to regularized linear regression model.

4.1 Linear regression model

Model	Alpha	Train score	Test score	Features selected
Linear Regression	/	0.95	0.90	220

Table 1. Results of Linear regression model

According to Table1, we can learn:

The train score of linear regression model is 0.95, 0.05 higher than the test score (0.90), with 220 features selected. Overall, the linear regression model performs well in the dataset.

4.2 Lasso regression model

Model	Alpha	Train score	Test score	Features selected
Lasso_01	0.1	/	/	/
Lasso_0005	0.0005	0.41	0.40	2
Lasso_0001	0.0001	0.85	0.84	12

Table2. Results of Lasso model

According to Table2, we can learn:

When Alpha is equal to 0.1, the model performs worse since the scores are negative, which means there is an underfitting situation.

When Alpha is equal to 0.005, the train score is 0.41 and the test score is 0.40, which contains 2 features.

When Alpha is equal to 0.001, the model performs best, with 12 feature variables included. Meanwhile, the gap between train score and test score is small, which means there is no overfitting situation.

4.3 Ridge regression model

Model	Alpha	Train score	Test score	Features selected
Ridge_01	1	0.95	0.91	214

Table3. Results of Ridge model

According to Table3, we can learn:

When Alpha is equal to 1, the train score of ridge model is 0.95, which is sightly higher than the test score (0.91), with 214 features included.

In summary, as can be seen from the table 4, the linear regression model and the Ridge model (Alpha=1) seems perform better compared to the remains.

Model	Alpha	Train score	Test score	Features selected
Linear Regression	/	0.95	0.90	220
Lasso_01	0.1	/	/	/
Lasso_0005	0.005	0.41	0.40	2
Lasso_0001	0.001	0.85	0.84	12
Ridge_01	1	0.95	0.91	214

Table4. Results of all models

5. Conclusion

1. When deal with the missing values, I focused on them without further optimization, such as making the dataset more complete by adding the mean or the mode. There is still room for more progress.
2. As for this project, the Linear regression model is used, and although the results are acceptable, the model is too simple to be realistic. In the future, I need to try to use other models for validation.
3. Although I didn't handle with the feature engineering well in the project, I have learned something from the process: When a variable is found to be useful for the goal to be accomplished, it can be employed as a feature for the model. Thus, as many features as possible should be generated at first, and then the prediction model can pick out the most useful ones, which is a continuous process.
4. Last but not least, in the process of analyzing data and visualizing data, my python skills have been improved a lot. In addition, I also realize that python is a useful tool with great power. But only when I learn the basic knowledge such as data processing and data visualization well, can I utilize it better.

Reference

基于 sklearn 的 Lasso 实现与效果分析. (2018, May 23). Blog.csdn.net.

https://blog.csdn.net/weixin_41500849/article/details/80425110

Appendix

Data fields

Here's a brief version of the data description.

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality

- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Exterior material quality
- ExterCond: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating

- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning
- Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity

- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale