CITY UNIVERSITY LONDON

MSc in Data Science

INM430 - Coursework

House Price Prediction by Using Regression techniques

Name: Shengqiang Fan

Student ID:180045953

Content

1. Analysis Domain, Questions, Plan (471 words)

1.1 Analysis Domain and Motivation

House price is a hot topic all over the world, but house price prediction is always a challenge due to the complexity of demand and supply. In other word, several factors would influence the house price, such as location, house type, built date, surrounding transportation, economic cycle, interest rate etc. In this coursework, a house price related data set is chosen from Kaggle, which is the Ames Housing Price data. It collects all the recorded house sale price with related house descriptions in Ames, IA from January 2006 to July 2010 [1]. With this data set, different features of residential houses are given for further evaluation. Therefore, house price prediction in this case will achieved by using regression techniques. Due to complexity of house price prediction, it is necessary to identify the most relevant features for house price and relationship between different features.

1.2 Analytical Problems

The major analytical objective is to make a prediction of house price based on given different properties of a curtain residential house. In order to grasp major factors leads to fluctuation of house price, the relationship between features and sale price needed to be identified, which will help to conduct feature engineering process. In addition, bunches of missing data needed to be treated of this data set.

1.3 Analysis Strategy and Plan

My plan is divided in to 4 parts:

a. Data collection and preparation
b. Exploratory data analysis
c. Feature Engineering and Model Selection
d. Results comparison from different models

After checked several kernels from Kaggle, I found most kernels combined train data and test data into one dataset for dealing with missing data, which I think is not appropriate. In that case, train dataset will obtain extra information from test dataset. Therefore, I tend to deal with missing data with train set and test set separately.

In terms of exploratory data analysis, the first thing needed to be done is to understand each feature's meaning through data description. Then, scatter plots/box plots/heatmap etc are needed to conduct for exploring relationship between features and objectives (SalePrice) as well as inter-relationship between different features. After this stage, some features will be deleted if evidence shows it has less influence of house price and outliers will be pointed out.

In terms of model section, based on scikit-learn algorithm cheat sheet, Lasso or ElasticNet are preferred to conduct this regression task [2]. Thus, Lasso regression and gradient boosting will be used for this supervised regression task. The major reason for using lasso regression and gradient boosting is both algorithms can compute feature importance when make prediction, which will help to understand relationship between

features and sale price. Then some important features from Lasso/gradient boosting will be combined together to create new features. Different combination should be tested to see the actual influence.

2. Analytical Process

The whole process is including:

Data Collection and Preparation → Exploratory Data Analysis

→Feature Engineering → Model Selection and Results

More details can be founded:
https://smcse.city.ac.uk/student/aczd147/INM430/PoDS_Notebook_Shengqiang_Fan.html

3. Findings and Reflections (957 words)

3.1 Findings on feature selection

Firstly, box plots are used especially for discrete data groups. Whole set of plotting can be found in appendix link. In this report, only some representative figures are picked for demonstration. House Sale Price against Overall Quality (rates the overall material and finish of the house) and Overall Condition (rates the overall condition of the house) is shown in figure 1. It is obvious that both Overall Quality and Overall Condition have positive relationship to Sale Price. Similar trend also found in features as LotShape, Alley, HeatingQC, TotRmsAbvGrd, FirePlaces, GarageYrBlt, GarageCars. In terms of other discrete features, low relationship founded based on box plots such as Land Slope and Roof Style as shown in figure 2. In addition, another interesting trend is that most of abnormal sale price value are above upper quartile rather than below lower quartile. Which indicates that the real estate developers always win price game with individual investor/buyer to sale their house above average market value.

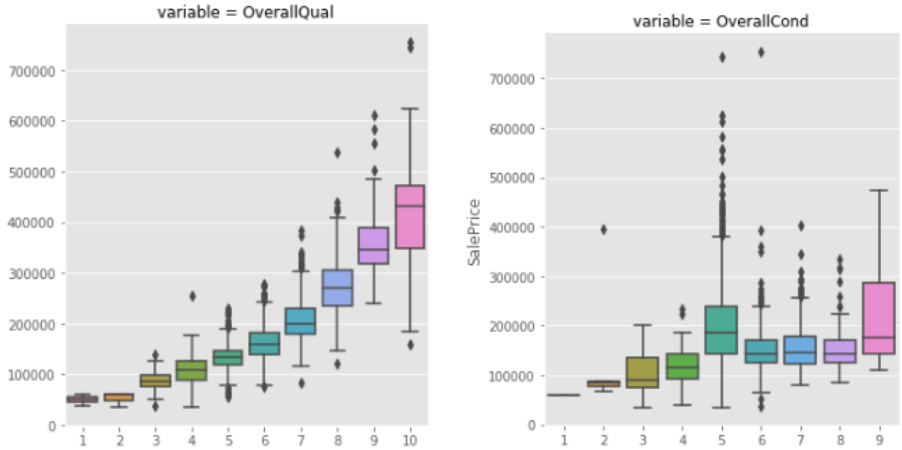

Fig 1: Box plots of Sale Price against Overall Quality (left) and Overall Condition

(Right)

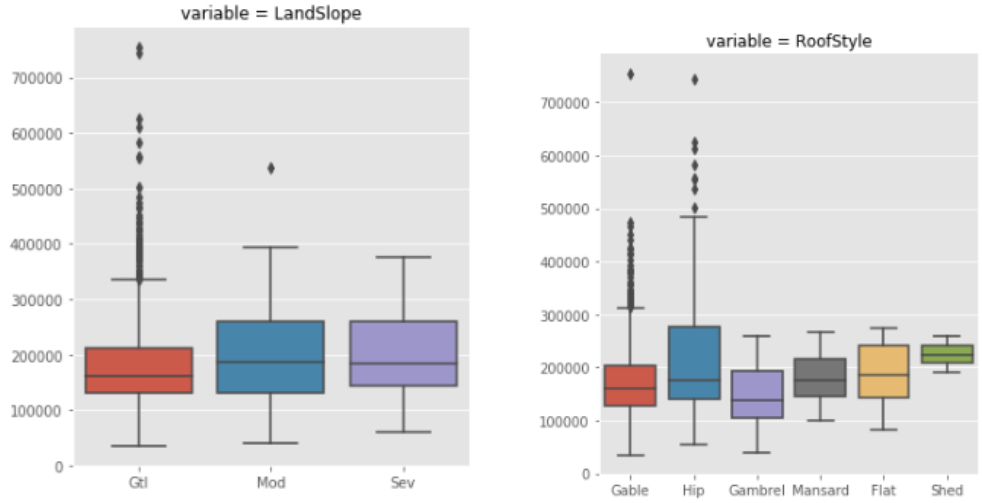

Fig 2: Box plots of Sale Price against Land Slope (left) and Roof Style (right)

Secondly, scatter plots are used for continuous data groups. Due to length limits, only some representative plots are chosen for illustration. For example, as shown in figure 3, it is evidence to say that sale price is proportional to the increase of GrLivArea (above ground living area square feet) and TotalBsmtSF (total square feet of basement area). Similar trend is also found in features as YearBuilt, YearRemodAdd, BsmtFinSF1, 1stFlrSF, 2ndFlrSF and GarageArea. Moreover, for feature which has strong relationship with sale price, outliers can also be pointed easily. The point which is not in the reasonable trend range area is treated as outlier. For instance, two points at right bottom corner in Fig 6 (left) and one point at right bottom corner in Fig 6 (right). Except features discussed above, no evidence can prove their relationship with sale price based on scatter plots, such as MasVnrArea (Masonry veneer area in square feet) and BsmtUnfSF (Unfinished square feet of basement area) in fig4.

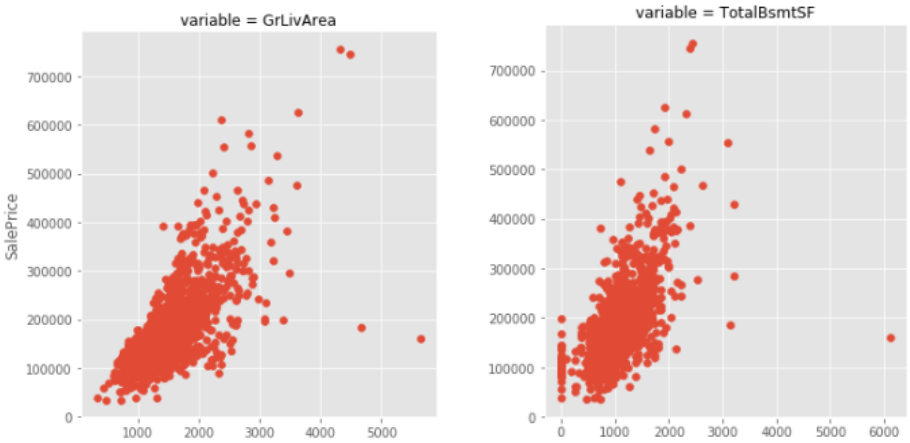

Fig 3: Scatter plots of Sale Price against GrLivArea (left) and TotalBsmtSF (Right)

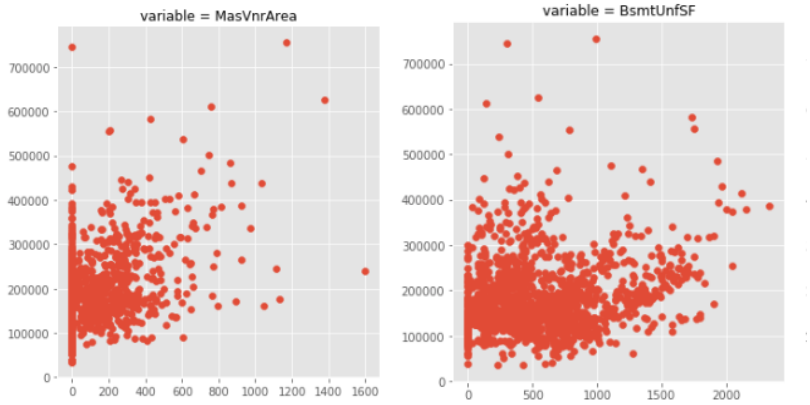

Fig 4: Scatter plots of Sale Price against MasVnrArea (left) and BsmtUnfSF (Right)

Two heatmaps are generated to show the relationship between sale price and features and correlations among features as shown in fig 5 and 6, which is based on pearson method and spearman method respectively. Same color scheme is used for both heatmap, lighter color indicates stronger relationship and darker color shows weak correlation. In terms of relationship between sale price and each feature, both heatmaps highlight OverallQual, YearBuilt, ToatlBsmtSF, 1stFlrSF, GrLiveArea, FullBath, TotRmsAbvGrd, GarageCars and GarageArea. Thus, these highlighted features are very important for further prediction. With respect to correlation among features, there two areas with very light color, which are TotalBsmtSF against1stFlrSF and GarageCars against GarageArea. In other words, TotalBsmtSF (Total square feet of basement area) has a high correlation with 1stFlrSF (First Floor square feet), therefore, one of these two features can be dropped in order to avoid multicollinearity problem.

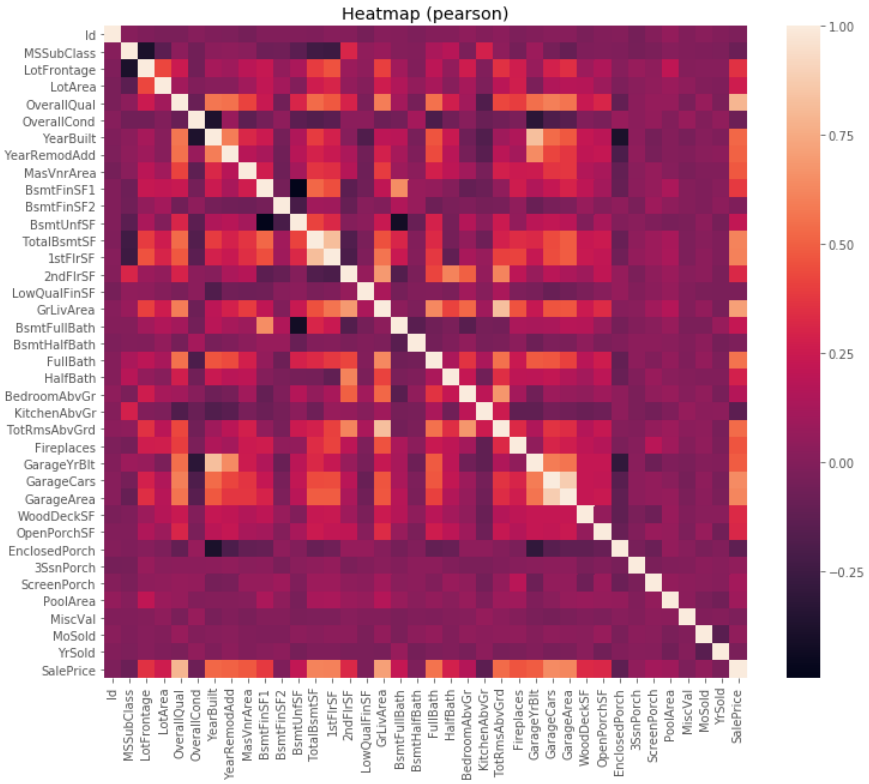

Fig 5: Heatmap based on pearson method

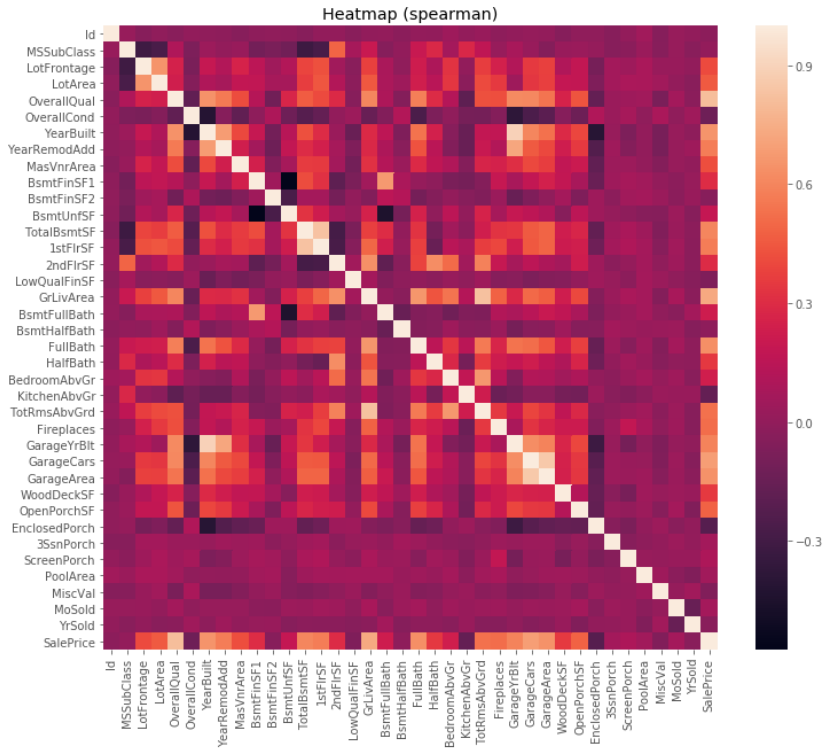

Fig 6: Heatmap based on spearman method

Though results are not totally the same with previous results, feature weight in Lasso and feature importance from gradient boosting algorithm also highlight similar important features as shown in fig 7 and fig 8 (result from lasso and gradient boosting respectively). In fig7, lasso gives higher weight on Overall quality, GreLivArea (above ground living area square feet), Neighborhood, Overall condition etc. As for feature importance from gradient boosting in fig 8, features such as LotArea (Lot size in square feet), GreLivArea, BsmtUnfSF (Unfinished square feet of basement area), GarageArea are pointed out. Therefore, 10 more features are added based on the highlighted features from above plotting. For example, one new feature is overall quality * GreLivArea, another is Overall Quality + Overall Condition. And then based on 10 more added features, lasso and gradient boosting regression are conducted again to compare performance of sale price prediction.

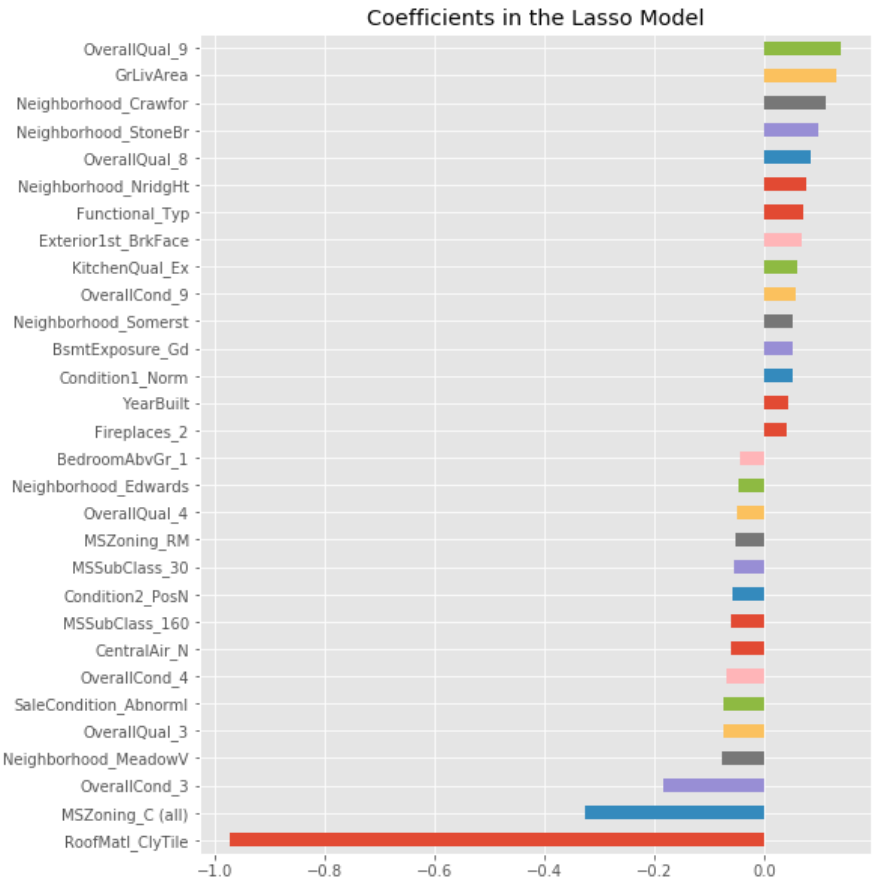

Fig 7: Feature weight in Lasso

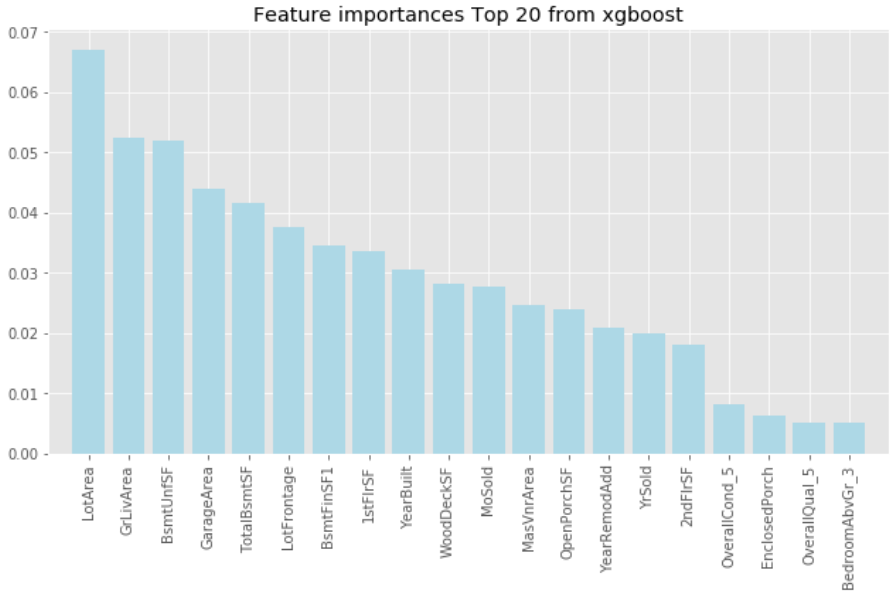

Fig 8: Feature importance Top20 from gradient boosting

3.2 Findings on sale price prediction

In terms of model selection and prediction, RMSE (Root-Mean-Squared-Error) should be used as defined in Kaggle specification. Thus, lower RMSE score indicates better prediction performance. Grid search is used for tuning hyperparameter in both Lasso and Gradient Boosting model. Table 1 shows the comparison of prediction precision before and after new features added. Cross validation score is computed from train set for 5-fold CV, while prediction score for test set is provided by Kaggle automatically. Result gives a great support for feature engineering process since performance of both algorithms are improved. Hence, highlighted features in above section such as GrLivArea (above ground living area square feet), TotalBsmtSF (total square feet of basement area) YearBuilt, Overall Quality and Overall Condition are highly correlated to sale price. In terms of comparison of Lasso and Gradient Boosting, lasso has better performance than gradient boosting in this case. In addition, cross validation score is always better than prediction score, which indicates overfitting/local optimal problem occurs.

Table 1: Comparison of before and after new features added

Model	No new added features		New features added	
Score method	Cross Validation Score	Prediction Score (Kaggle)	Cross Validation Score	Prediction Score (Kaggle)
Lasso	0.13892	0.13872	0.13432	0.12368
Gradient Boosting	0.12829	0.14577	0.12295	0.14101

In the meanwhile, ensemble two algorithm is considered to overcome overfitting/local optimal problem although lasso gives better prediction compared to gradient boosting. Ensemble method used in this report is using different weighting for two algorithms as shown in table 2. The best combination is shown to be 85% weight of lasso and 15% gradient boosting, which gives the better score by Kaggle 0.12318 (top25%).

Table 2: Prediction score of different combination

Combination	Lasso %	Gradient Boosting%	Prediction Score (Kaggle)
1	65%	35%	0.12430
2	70%	30%	0.12383
3	75%	25%	0.12349
4	80%	20%	0.12327
5	85%	15%	0.12318
6	90%	10%	0.12322

3.3 Short conclusion

In a conclusion, lasso and gradient boosting can give a good prediction for sale price although accuracy is not extremely high. When ensemble two models, best combination is tested to be 85% lasso and 15% gradient boosting. In addition, important features are

highlighted in this report. Although GrLivArea (above ground living area square feet), Overall Quality and Overall Condition are already accepted as importance features by the public, some features such as GarageArea and TotalBsmtSF (total square feet of basement area) may not considered as important feature in daily life but play important roles in sale price prediction. Hence, these highlighted features should be considered when evaluate a house price whether for individual buyers or real estate investors.

4. Limitations and Future Works

House Price Index (HPI) is used in many countries for evaluating house price change [3]. If more related features such as interest rate, local salary/income, population and property tax are available, it will be very useful for local house price prediction. In terms of regression model, there are still lots of techniques can be conducted for regression task. Ensemble several regression model would give better prediction results.

Reference

[1] https://www.kaggle.com/c/house-prices-advanced-regression-techniques
[2] https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
[3] Eli Beracha, Ben T Gilbert, Tyler Kjorstad, Kiplan womack, "On the Relation between Local Amenities and House Price Dynamics", Journal of Real estate Economics, Aug. 2016.

Appendix

Codes of analysis process:
HTML:
https://smcse.city.ac.uk/student/aczd147/INM430/PoDS_Notebook_Shengqiang_Fan.html

Ipynb:
https://smcse.city.ac.uk/student/aczd147/INM430/PoDS_Notebook_Shengqiang_Fan.ipynb

File directory of PoDS:
https://smcse.city.ac.uk/student/aczd147/INM430/