Data Mining and House Price Prediction: Multiple Linear Regression and Regression Tree Techniques

Registration No:

Word count: 3078

Abstract

INTRODUCTION: In 2016, Kaggle hold a competition of house price prediction and in the provided dataset, there are 79 aspects related to houses in Ames, Iowa from 2006 to 2010. The aim of this research is to use and compare two data mining techniques to build models in order to predict the house price.

DATA MINING HISTORY: Data mining is the process of extracting information and knowledge from data. There are two suitable techniques can be applied in this case: multiple linear regression and regression tree. It is because both of them have achieved similar predictive data mining tasks. In order to evaluate model performance, the author chooses R^2 and RMSE as evaluation measure as well as scatter plot of real values to predicted values.

DATA EXPLORATION AND PREPARATION: Prior to experimental setup, the author briefly describes the sold houses and explores SalePrice with several important attributes. In addition, missing values are handled and three potential attributes sets are selected and will be tested.

EXPERIMENTAL SETUP: In order to find the best model, multiple linear regression tests three different attribute sets and regression tree model uses cross-validation and determine two important parameters: tree level and fold number. All used nodes and settings are summarised in Table 4.3.

RESULTS AND DISCUSSION: Several take-away messages are listed:

(1) Over half of houses' prices are between 100 - 200 thousand dollars and every June and July are peak period for sale

(2) Most sold houses are built during 1940 - 2000 and located in residential low-density zone with one-story and two-story house style

(3) The best model of multiple linear regression is better than that of regression tree and it can explain 84.6% house price and root mean squared error is 28719 dollars.

(4) Top 3 important features are above grade living area, overall material and finish quality and lot size

One limitation of this study is that the intercept and coefficients of attributes are very high. The author suggests that SalePrice can be normalised when building model and then de-normalised as predicted value.

REFLECTION AND CONCLUSION: The benefits of this study case are to fully understand several fundamental data mining principles and greatly develop data processing skills. Some encountered challenges include feature selection and overfitting. According to other studies, further feature engineering should be undertaken to improve model performance.

Table of Contents

1. Introduction

Data mining is the process of exacting information from data and one of data mining categories is prediction: estimate an unknown value that could be in the future (Provost & Fawcett, 2013). In August 2016, Kaggle, an online community of data scientists and machine learners, offered a famous prediction competition of house price. Competitors are offered a dataset which describes the sale of residential property in Ames, Iowa from 2006 to 2010 (Cock, 2011). In addition, there are 79 explanatory variables in this dataset, such as overall condition rating (OverallCond) and original construction date (YearBuilt). The aim of this research is to predict the final price of each home on the basis of 79 attributes. In this report, the author uses the techniques of multiple linear regression and regression tree to achieve this goal.

This paper has been divided into five parts. The first part gives a brief overview and motivation of chosen algorithms and evaluation method for model performance. The second section provides a description of the data (e.g. statistic measures) and explains the exploratory approach by Tableau. The third section describes the experimental process for each data mining method and the results of experiments are presented and discussed in the fourth part. Finally, the author reflects on the chosen data mining methods and summarises the main findings.

2. Data mining theory

According to a definition provided by Witten, Frank and Hall (2011), data mining is "techniques for finding and describing structural patterns in data, as a tool for helping to explain that data and make predictions from it" (p. 8). In addition, data mining techniques can be classified broadly as predictive and descriptive and two of prediction techniques are multiple linear regression and regression tree (Witten et al., 2011). In this section, the author will briefly introduce the theory and motivation of these two data mining methods and provide the evaluation methods for models.

2.1 Multiple Linear regression

Multiple linear regression is a linear approach to model the relationship between a dependent variable and more independent variables:

$$x = w_0 + w_1 a_1 + w_2 a_2 + \ldots + w_k a_k$$

Where x is the dependent variable; a_1, a_2, …, a_k are the independent values; and w_0, w_1, …, w_k are weights which are calculated to minimise the distance of the instances from the line (Freedman, 2009). In addition, there is clear evidence to suggest that linear regression models can be used to predict a numeric outcome. In their study, Kelleher, Namee and D'Arcy (2014) successfully predict rental price with three variables: size, floor and broadband rate and found the following equation:

$$\begin{aligned} \textsc{Rental Price} = \mathbf{w}[0] + \mathbf{w}[1] \times \textsc{Size} + \mathbf{w}[2] \times \textsc{Floor} \\ + \mathbf{w}[3] \times \textsc{Broadband Rate} \end{aligned}$$

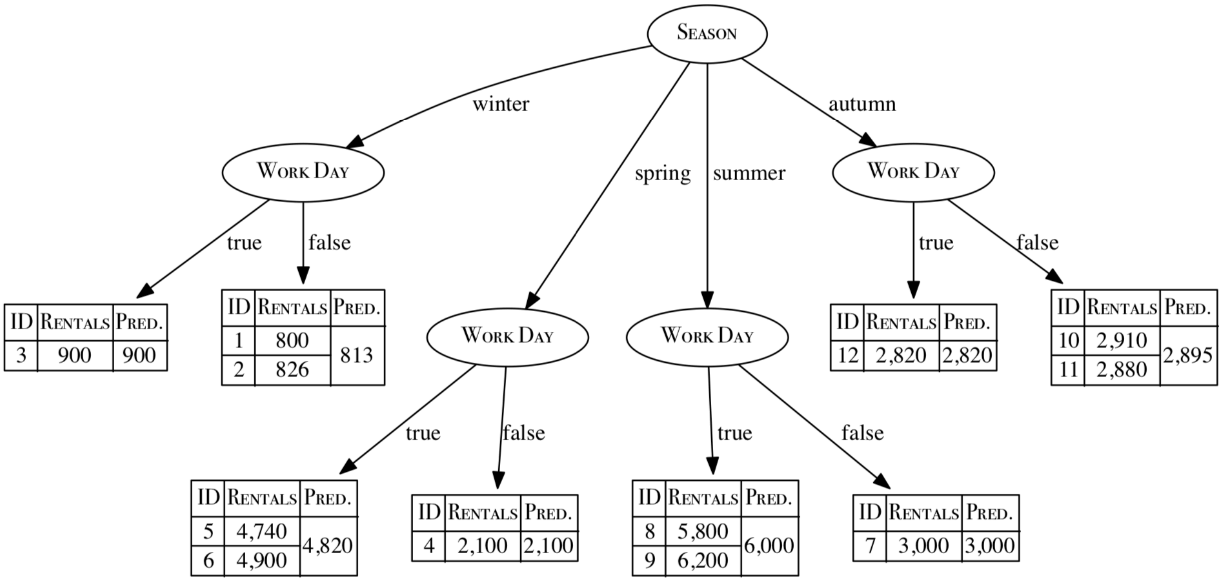
where w[0] = -0.1513, w[1] = 0.6270, w[2] = -0.1781 and w[3] = 0.0714. Another multiple linear regression model for prediction is CPU performance. The regression equation below, given by Witten et al. (2011), is to predict the CPU performance data:

Performance = −55.9 + 0.0489*Cycle time+ 0.0153*Minimum memory + 0.0056*Maximum memory + 0.6410*Cache – 0.2700*Minimum channels + 1.480*Maximum channels

By this regression equation, CPU performance value can be calculated with input data. It is, therefore, argued that multiple linear regression is feasible to numeric value prediction. The motivation of using multiple linear regression as a study method is that house price is a numeric outcome and it can apply multiple linear regression. In addition, Witten et al. (2011) hold the view that one of the advantages of linear regression is that it can give insight into which variables are most important in predicting the output. For instance, according to the regression equation of rental price and CPU performance, size of office and maximum channels are the most important variable because of the highest coefficient.

2.2 Regression tree

According to Rokach and Maimon (2015), regression trees "are decision trees that deal with a continuous which combine decision trees and linear regression to forecast numerical target attribute based on a set of input attributes" (p. 85). In addition, the main idea of this algorithm is to reduce the variance in the target attribute values at each leaf node, but it is difficult to determine the best node spilt. The usual solution is to choose attributes that lead to the lowest weighted variance across all nodes or stop criterion at the appropriate time. Bike rental is another predictive data mining task and can be achieved by the regression tree technique (Nekkanti, 2017). The figure below presents the regression tree view of bike rentals and it can be seen that this task is similar to house price prediction: both of them have different value types such as numeric and nominal attributes.



ID	Rentals	Pred.
3	900	900

ID	Rentals	Pred.
1	800	813
2	826	

ID	Rentals	Pred.
5	4,740	4,820
6	4,900	

ID	Rentals	Pred.
4	2,100	2,100

ID	Rentals	Pred.
8	5,800	6,000
9	6,200	

ID	Rentals	Pred.
7	3,000	3,000

ID	Rentals	Pred.
12	2,820	2,820

ID	Rentals	Pred.
10	2,910	2,895
11	2,880	

Figure 2.1 Regression tree view of rental bikes

2.3 Model building and evaluation

According to James, Witten, Hastie, and Tibshirani (2017), when building a model, supervised learning needs a training dataset and a testing dataset. Ripley (1996) claims that training dataset is used to fit the parameters by comparing prediction values with reference values. As for testing data, Brownlee holds the view that it is independent of training dataset and used to assess the performance of model.

In this report, the author chooses two statistic measures as evaluation measure: coefficient of determination (r-square) and root mean squared error (RMSE). The main purpose of r-square is the prediction of future outcomes (Draper & Smith, 1998; Glantz & Slinker, 2016).

And it is a metric to compare a numeric column's values (r_i) and predicted (p_i) values:

$$R^2=1-SS_{res}/SS_{tot}=1-\Sigma(p_i-r_i)^2/\Sigma(r_i-1/n*\Sigma r_i)^2$$

Regarding to RMSE, it is calculated as $(sqrt(1/n*\Sigma(p_i-r_i)^2))$.Hyndman and Koehler (2006) argue that it is a measure of accuracy and to compare differences between predicted values and the values observed. However, if the r-squared and RMSE of training dataset are good but those of testing dataset are bad, it indicates that the trained model is overfitting and cross-validation should be applied.

Cross-validation, for Provost and Fawcett (2013), is a performance estimates to choose the best parameters and there are three steps to perform. First, data set is split into k partitions and all training and testing dataset are carried out for k times. Second, for every time, one partition is chosen as holdout data to test the performance of model which is learned from other k-1 partitions. Finally, different models are compared in order to determine the best model parameters and all data are deployed again by this best model. The application of cross-validation for house price prediction will be further discussed in the following experimental setup process.

3. Data exploration and preparation

In this section, the author will investigate several important attributes (e.g. house price, sold time and overall quality) in order to describe sold house and provide data cleaning approaches prior to the experimental setup, including handling missing value, normalisation and feature selection.

3.1 Description of attributes

3.1.1 House price

As indicated in the introduction, the house price is the predicted property. Therefore, the author draws the distribution histogram and box plot of sale price. It can be seen from Figure 3.1.1 that about 60% of houses' prices are within the range of 100 - 200 thousand dollars.

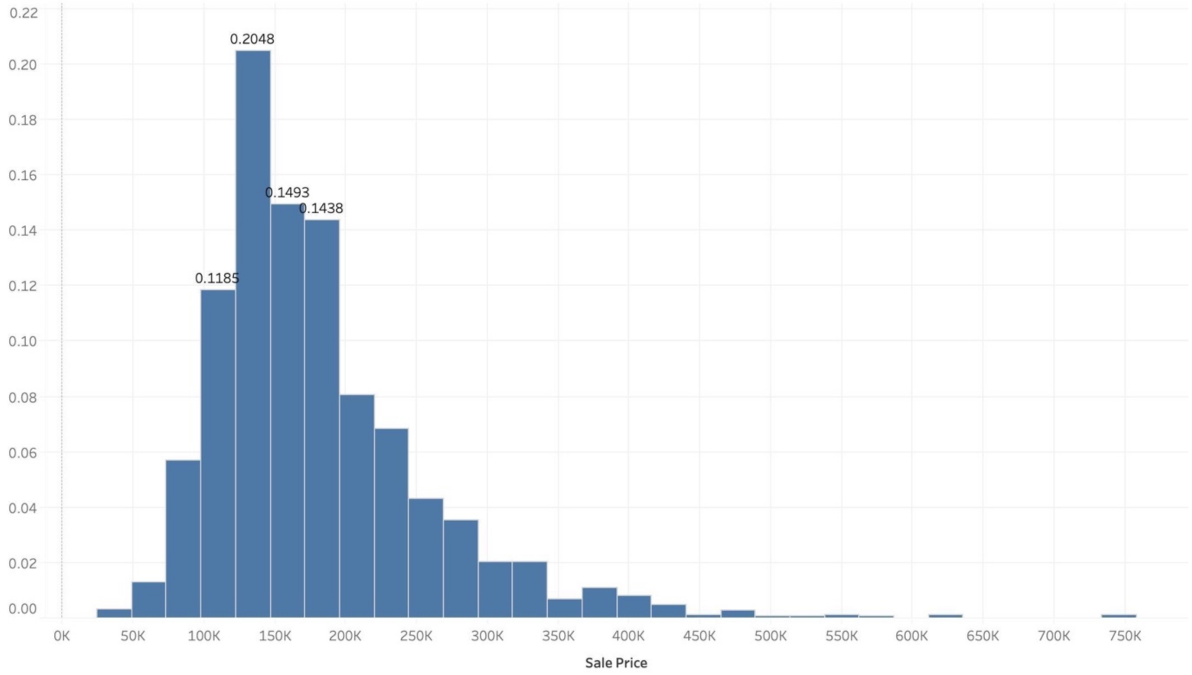


Figure 3.1.1 Distribution of house price

From the boxplot of Figure 3.1.2, we can see following statistic summary: (1) average: $201685 (2) median: $179540 (3) minimum: $34900 (4) maximum: $755000.

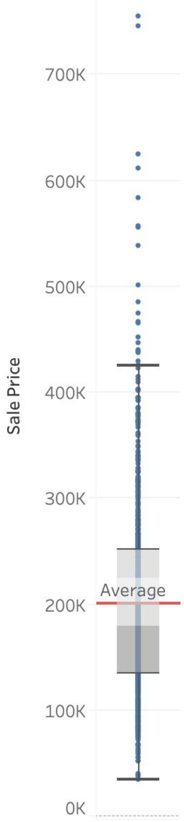
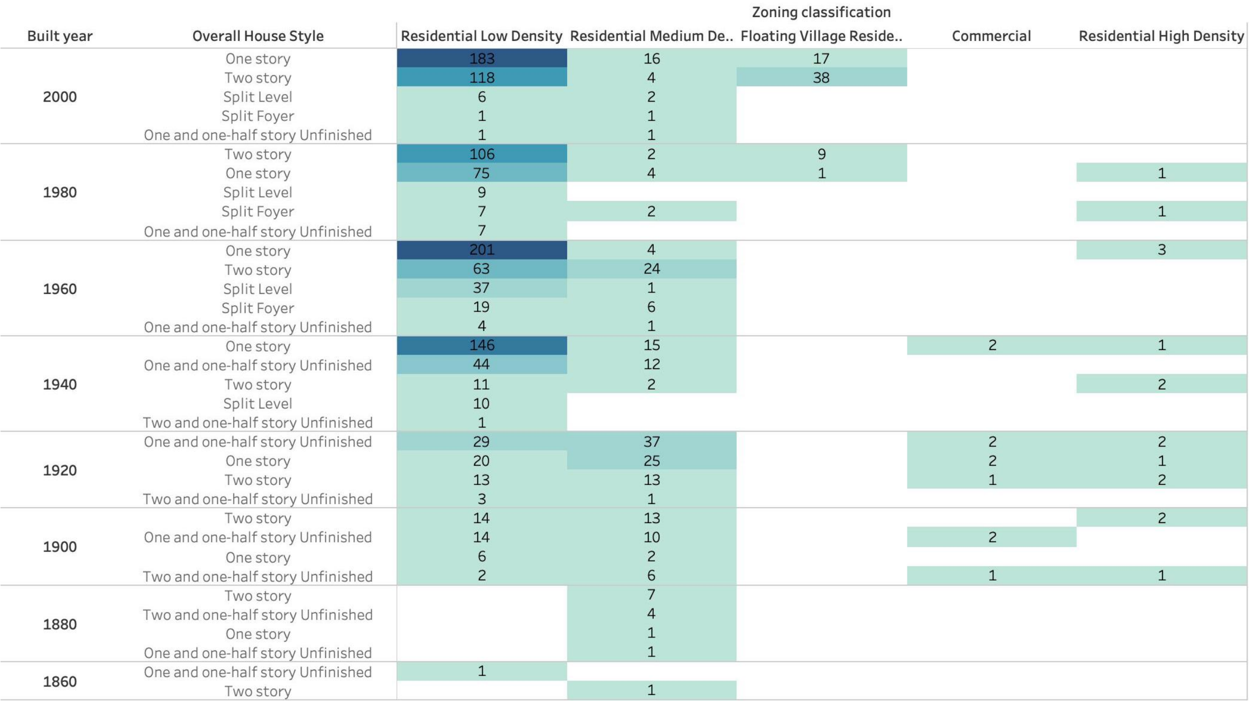


Figure 3.1.2 Box plot of house price

3.1.2 Sold house and sale time

Before exploring attributes, the author would like to investigate background information about the sold house. Three important attributes of the sold house are summarised in Figure 3.1.3: built year (YearBuilt), house style (HouseStyle) and zoning classification (MSZoning). From the figure we can see that most sold house are built during 1940 - 2000 and located in residential low-density zone. In addition, one-story and two-story is main sold house style.



Built year	Overall House Style	Zoning classification Residential Low Density	Residential Medium De..	Floating Village Reside..	Commercial	Residential High Density
2000	One story	183	16	17		
	Two story	118	4	38		
	Split Level	6	2			
	Split Foyer	1	1			
	One and one-half story Unfinished	1	1			
1980	Two story	106	2	9		
	One story	75	4	1		1
	Split Level	9				
	Split Foyer	7	2			1
	One and one-half story Unfinished	7				
1960	One story	201	4			3
	Two story	63	24			
	Split Level	37	1			
	Split Foyer	19	6			
	One and one-half story Unfinished	4	1			
1940	One story	146	15		2	1
	One and one-half story Unfinished	44	12			
	Two story	11	2			2
	Split Level	10				
	Two and one-half story Unfinished	1				
1920	One and one-half story Unfinished	29	37		2	2
	One story	20	25		2	1
	Two story	13	13		1	2
	Two and one-half story Unfinished	3	1			
1900	Two story	14	13			2
	One and one-half story Unfinished	14	10		2	
	One story	6	2			
	Two and one-half story Unfinished	2	6		1	1
1880	Two story		7			
	Two and one-half story Unfinished		4			
	One story		1			
	One and one-half story Unfinished		1			
1860	One and one-half story Unfinished	1				
	Two story		1			

Figure 3.1.3 Classification of sold house

As for sold time, Figure 3.1.4 presents the house sale over sale time which aggregated from sold year (YrSold) and sold month (MoSold). It is argued that every June and July are peak period for sale, apparently higher than others.

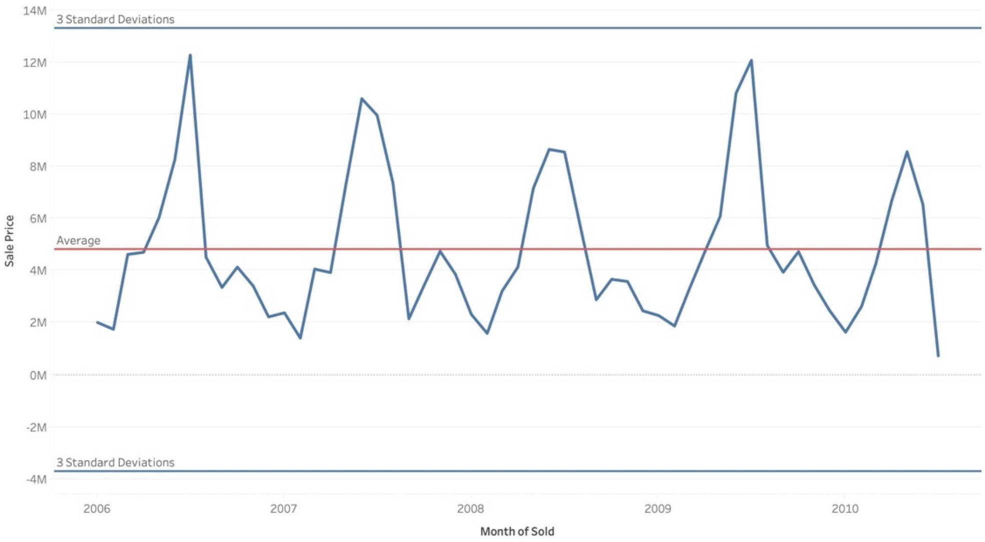


Figure 3.1.4 Sale price over time

3.1.3 Attribute exploration

As introduced above, there are 79 variables in this dataset for house price prediction. In this step, the author chooses four attributes, which are most likely to be related to sale price. The first is ground living area (GrLivArea). It is general knowledge that the bigger ground living area is, the higher price house is. Ground living area - Sale price scatter plot with tread line is presented in Figure 3.1.5 and it indicates that there could be a linear relationship (R-square = 0.5021, P-value < 0.0001).

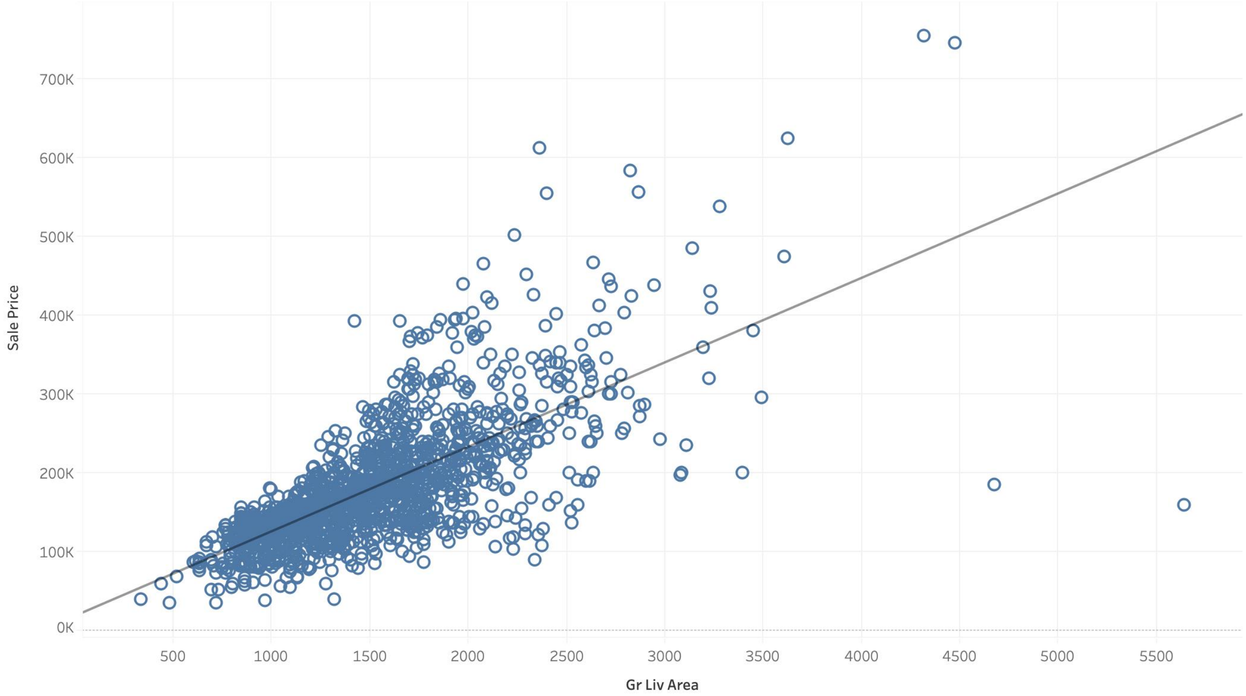


Figure 3.1.5 Sale price to ground living area

The second significant aspect is house quality (OverallQual). From Figure 3.1.6 below, we can see that high-quality house is more expensive, which is reasonable.

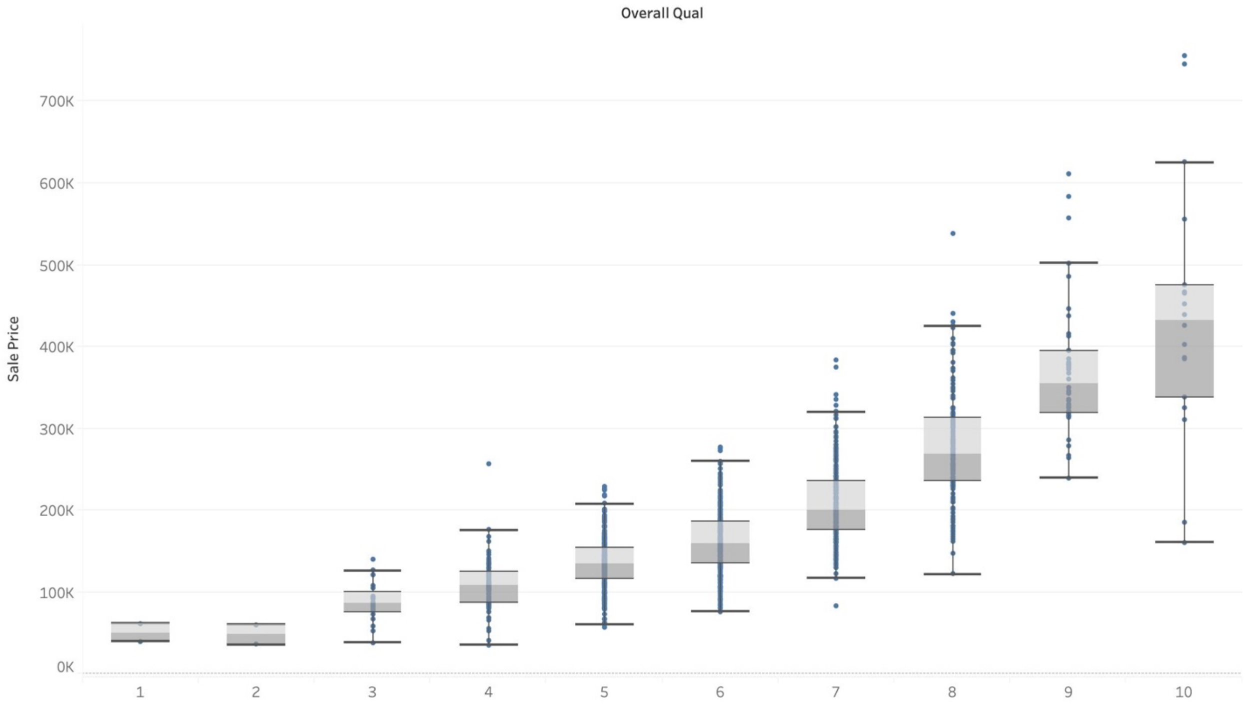


Figure 3.1.6 House price boxplot to house quality

Similarly, it can be seen from Figure 3.1.7 and 3.1.8 that neighbourhood (Neighbourhood) and built year (YearBuilt) also seem to influence sale price. To summarise, ground living area (GrLivArea), house quality (OverallQual), neighbourhood (Neighbourhood) and built year (YearBuilt) are potential attributes.

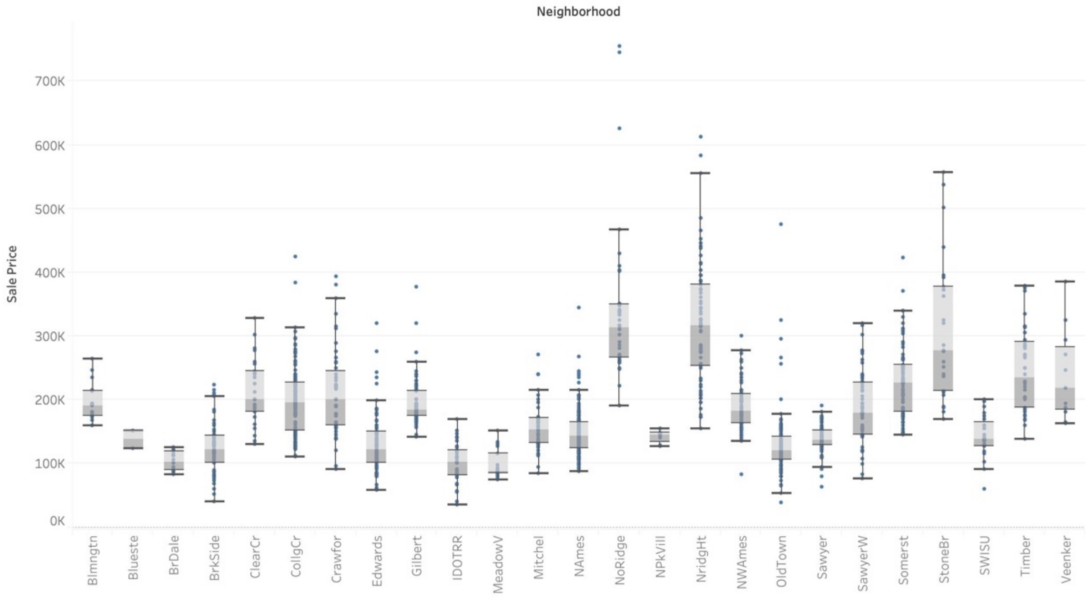


Figure 3.1.7 Box plot of house price to neighbourhood

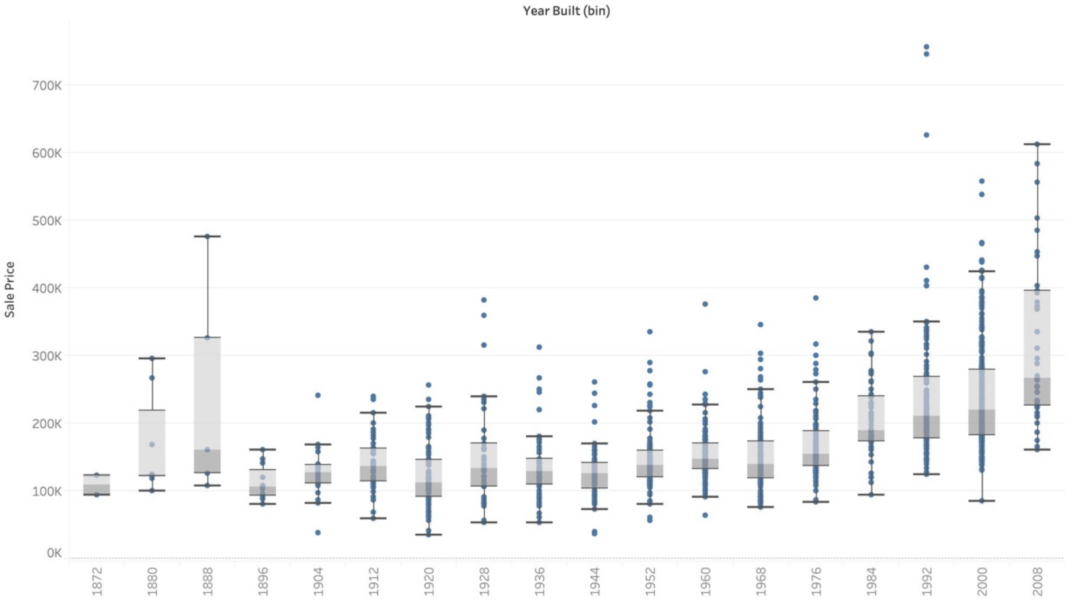
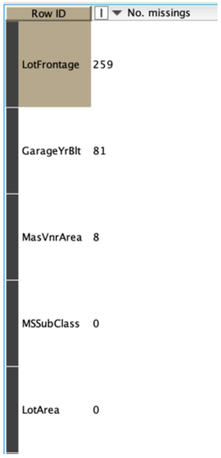


Figure 3.1.8 Box plot of house price to build year

3.2 Data cleaning

3.2.1 Missing data and normalisation

After reading data from CSV file, the first step is to check missing data. In Knime software, "Statistic" node can be used to search the number of missing values of each attribute. It can be seen from Figure 3.2.1 that there are three attributes have missing values: linear feet of street connected to property (LotFrontage), garage-built year (GarageYrBlt) and Masonry veneer area (MasVnrArea). First, for LotFrontage attribute, the author replaces missing values with the average value. Second, missing values in MasVnrArea are replaced with zero. Because all corresponding MasVnrType value is NA which indicates there is no masonry veneer for these sold houses. Finally, GatageYrBlt attribute is dropped because it is closed to build year (YearBuilt) and they are highly correlated. The second step is normalisation and the author adopt min-max normalisation method for all attributes, excluding YearBuilt, YearRemodAdd, MoSold, YrSold, MSSubClass and SalePrice.



Row ID	No. missings
LotFrontage	259
GarageYrBlt	81
MasVnrArea	8
MSSubClass	0
LotArea	0

Figure 3.2.1 The number of missing values

3.2.2 Feature selection

When selecting features, the author chooses two sets of attributes by correlation matrix and personal experience independently and finally combines them together. First, linear correlation node can be used to calculates for each pair of selected columns a correlation coefficient and correlation matrix is shown in Figure 3.2.2.

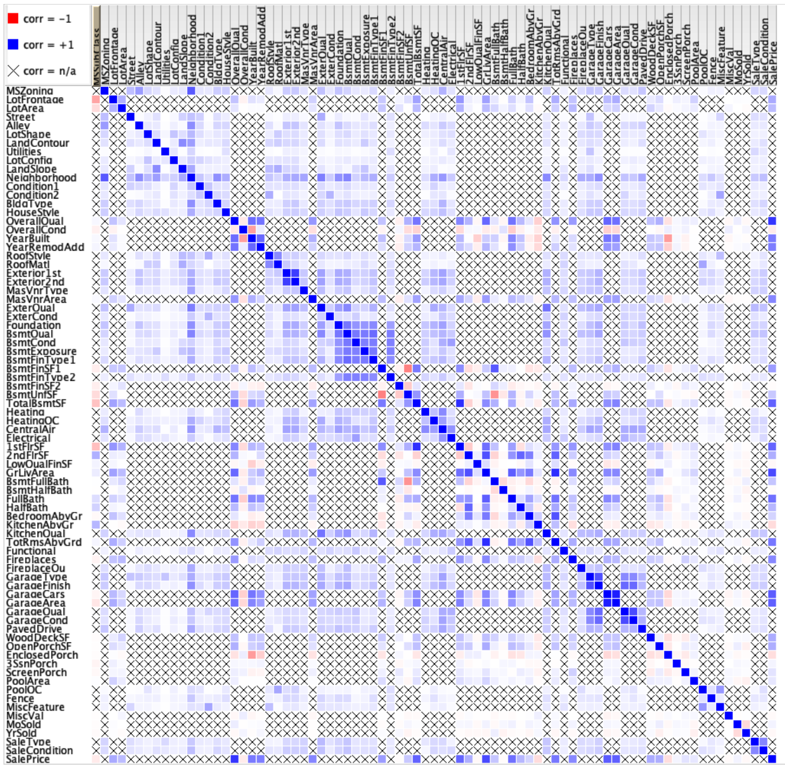


Figure 3.2.2 Correlation matrix of all attributes

The author select 10 attributes whose correlation coefficient with sale price is more than 0.5: OverallQual (0.791), GrLivArea (0.7086), GarageCars (0.6404), GarageAreas (0.6234), TotalBsmtSF (0.6136), 1stFlrSF (0.6059), FullBath (0.5607), TotRmsAbvGrd (0.5337), YearBuilt (0.5229) and YearRemodAdd (0.5071). In addition, this result includes three of four attributes explored before. These attributes, however, could also be correlated and need to be removed. The figure below shows that there are four pairs of strong correlated attributes: GarageAreas - GarageCars, YearBuilt - YearRemodAdd, TotalBsmtSF - 1stFlrSF and GrLivArea - TotalBsmtSF. The author, thus, filters GarageCars and YearRemodAdd. Finally, there are six attributes in the first set: OverallQual, GrLivArea, GarageAreas, 1stFlrSF, FullBath and YearBuilt.

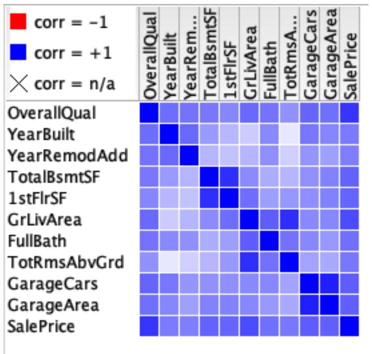


Figure 3.2.3 Correlation matrix of 10 attributes

Second, according to personal experience, the author assumes following attributes as the second set: LotArea, Neighborhood, Condition1 Condition2, BldgType, HouseStyle, YearBuilt, YearRemodAdd, OverallQual, MoSold, YrSold, ExterQual and ExterCond. Fortunately, as shown in the correlation matrix, there is no strong relationship between these attributes. In the next section, these two attribute sets will be tested and further selected.

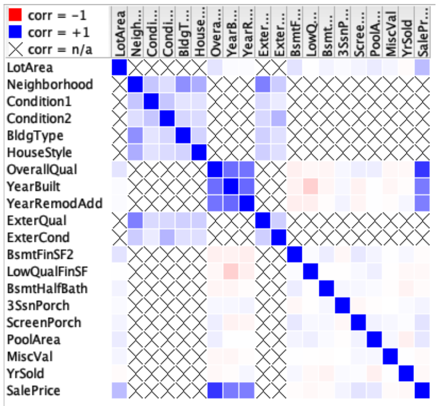


Figure 3.2.4 Correlation matrix of 13 attributes

4. Experimental setup

In this part, the author will describe the experimental design and validation process for two data mining methods and provide a table including all used nodes and parameter settings in this process.

4.1 Multiple linear regression

As discussed above, the author provides two sets of attributes and test them in following multiple linear regression models.

There are six attributes in the first set: OverallQual, GrLivArea, GarageAreas, 1stFlrSF, FullBath and YearBuilt. The first step is to use Partitioning node to split data into two partitions: choose 80% data randomly as train data and the rest as test data. The second step is to perform multiple linear regression: import train data to Linear Regression Learner node and connect the output model to Regression Predictor with test data. The final step is to evaluate performance by scatter plot and statistics measures: use Scatter Plot node to create scatter plot of SalePrice and Prediction (SalePrice) and compute R-Square and root mean square error by Numeric Scorer node.

Similarly, the second attribute set is also split into train data (80%) and test data (20%). After performing multiple linear regression, several outliers are deleted: Row Id = 899, 826, 804, 186 and 113.

Finally, the author attempt to combine these two sets together and form the third set including OverallQual, GrLivArea, GarageAreas, 1stFlrSF, FullBath and YearBuilt, LotArea, Neighborhood, Condition1 Condition2, BldgType, HouseStyle, MoSold, YrSold, ExterQual and ExterCond. Then, data is split and outliers (Row ID = 1183, 1047 and 899) are removed to perform regression.

4.2 Regression tree

In the regression tree model, the author uses the third attribute set above as dependent attributes. In addition, after splitting data, cross validation is achieved using X-Partitioner and X-Aggregator nodes. The aim of cross-validation is to estimate the performance of a model and choose the best parameter. There are, however, another two hyper-parameters need to be optimised: the number of cross-validation (fold) and the number of tree levels. The author configures fold value with five or ten and tree levels with five or ten, of which results are compared by statistics measures in the next section.

4.3 Summarise

The aim of this section is to describe the experiment process of multiple linear regression and regression tree techniques in detail. When performing multiple linear regression, three attribute sets are chosen and compared and outliers are removed for better results. As for the regression tree, cross-validation is applied and configured to avoid over-trained. All used nodes with parameter setting are summarised in the table below.

Table 4.3 Parameter settings

Data processing			Node name	Configure content
Import data			CSV Reader	Browse: dataset file path
Handle missing data			Missing Value	Column Setting: LotFrontage – Mean GarageYrBlt – Do nothing MasVnrArea – Fix Value with 0
			Column Filter	Exclude: GarageYrBlt
Normalize data			Normalizer	Exclude: MSSubClass, YearBuilt, YearRemodAdd, MoSold, YrSold, SalePrice; Min-Max Normalization: Min: 0.0; Max: 1.0
Multiple linear regression	Attribute set 1	Select features	Column Filter	Include: LotArea, Neighbourhood, Condition1, Condition2, BldgType, HouseStyle, OverallQual, YearBuilt, YearRemodAdd, ExterQual, ExterCond, BsmtFinSF2, LowQualFinSF, BsmtHalfBath, 3SsnPorch, ScreenPorch, PoolArea, MiscVal, YrSold, SalePrice;

		Remove outliers	Row Filter	Column to set: SalePrice;
				Exclude rows by attribute value:
				61157, 475000, 385000, 383970;

	Attribute set 2	Select features	Column Filter	Include: OverallQual, YearBuilt, 1stFlrSF, GrLivArea, FullBath, GarageArea, SalePrice;
		Remove outliers	Row Filter	Exclude rows by row ID; Regular expression: 1183, 1047, 899
	Attribute set 3	Select features	Column Filter	Include: 1stFlrSF, GrLivArea, FullBath, GarageArea, LotArea, Neighbourhood, Condition1, Condition2, BldgType, HouseStyle, OverallQual, YearBuilt, YearRemodAdd, ExterQual, ExterCond, BsmtFinSF2, LowQualFinSF, BsmtHalfBath, 3SsnPorch, ScreenPorch, PoolArea, MiscVal, YrSold, SalePrice
		Remove outliers	Row Filter	Exclude rows by row ID; Regular expression: 1183, 1047, 899
	Split Data		Partitioning	Relative [%]: 80;

			Draw randomly
	Perform regression	Linear Regression Learner	Target: SalePrice
		Regression Predictor	Default setting
	Produce Results	Scatter Plot	
		Numeric Scorer	
Regression tree	Select features	Column Filter	Include: 1stFlrSF, GrLivArea, FullBath, GarageArea, LotArea, Neighbourhood, Condition1, Condition2, BldgType, HouseStyle, OverallQual, YearBuilt, YearRemodAdd, ExterQual, ExterCond, BsmtFinSF2, LowQualFinSF, BsmtHalfBath, 3SsnPorch, ScreenPorch, PoolArea, MiscVal, YrSold, SalePrice;
	Cross-validation	X-Partitioner	Number of validations: 5; Random sampling
		X-Aggregator	Target column: SalePrice; Prediction column: Prediction (SalePrice)
	Perform model	Simple Regression Tree Learner	Use binary splits for nominal attributes; Limit number of levels: 5
		Simple Regression Tree Predictor	Default setting
	Produce Results	Scatter Plot	
		Numeric Scorer	

5. Results and discussion

To evaluate the performance of model, r-squared and RMSE is used and the results of two data mining techniques are presented. Table 5.1 provides the results obtained from different attribute sets. It can be seen that, compare to attribute set 2, attribute set 1 has higher r-squared (0.817 to 0.795) but RMSE is also higher, which indicates that model 1 can explain more data but error is bigger. As described in the previous section, the author combines attribute set 1 with attribute set 2 to form the third attribute set. As shown in Table 5.1, attribute set 3 is the best model of linear regression: r-square is the highest and RMSE is the lowest than other models (r-square = 0.846, RMSE = 28719). In regard to regression tree, there are two parameters need to figure out: the number of fold and tree level.

Table 5.1 results of different attribute sets

	attribute set 1	attribute set 2	attribute set 3
Adjusted R^2 (training)	0.7797	0.7625	0.8383
R^2 (testing)	0.817	0.795	0.846
RMSE	34114	33130	28719

Table 5.2 provides the results of combination of different values of fold and tree level. It is apparent that when fold = 5 and tree level = 5, the performance of model is the best: r-square = 0.745 and RMSE = 40066.

Table 5.2 results of different parameters

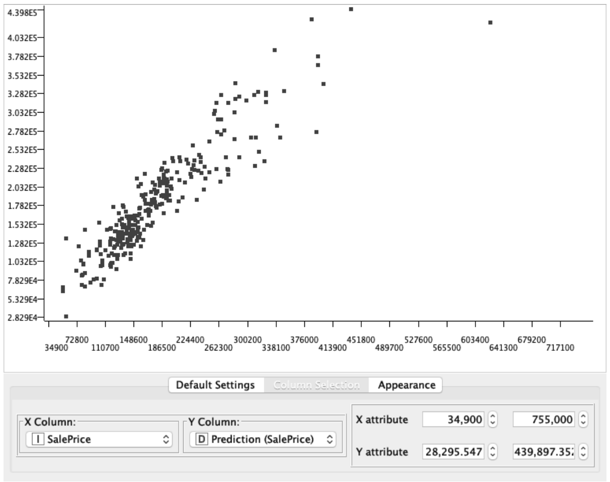
	fold = 5 tree level = 5	fold = 10 tree level = 5	fold = 10 tree level = 10	fold = 5 tree level = 10
R^2	0.745	0.702	0.696	0.71
RMSE	40066	43329	43762	42748

After determining the best model for two methods, it is necessary to identify the best method. Table 5.3 contracts the best model of two methods and illustrates that multiple linear regression is better: r-square is higher (0.846 to 0.745) and RMSE is lower (28719 to 40066).

Table 5.3 Comparison of two data mining methods

	linear regression	regression tree
R^2	0.846	0.745
RMSE	28719	40066

In addition, the scatter plots of real house price to predicted house price for two data mining methods are set out in Figure 5.4 for multiple linear regression and Figure 5.5 for regression tree. It can be seen that the performance of multiple linear regression is better.



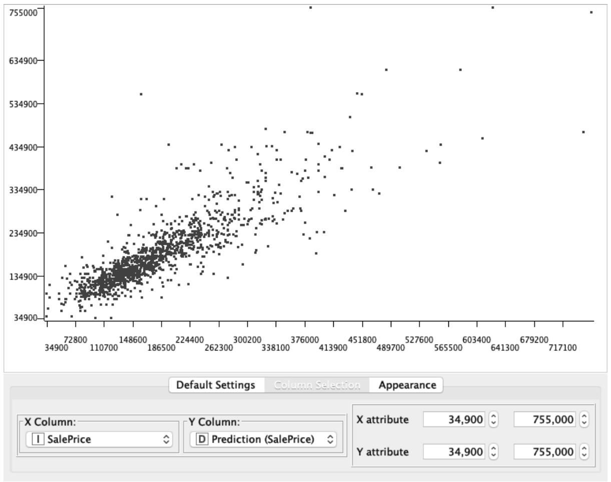
X Column: I SalePrice

Y Column: D Prediction (SalePrice)

X attribute 34,900 755,000

Y attribute 28,295.547 439,897.35:

Figure 5.4 Scatter of linear regression



X Column: I SalePrice

Y Column: D Prediction (SalePrice)

X attribute 34,900 755,000

Y attribute 34,900 755,000

Figure 5.5 Scatter plot of regression tree

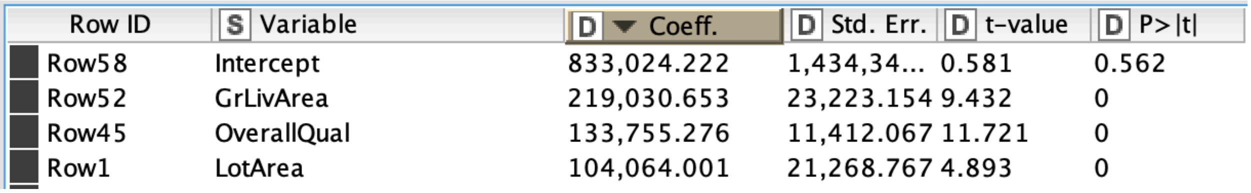
Moreover, Table 5.6 summarised the most accurate and inaccurate cases of two techniques. From the table, we can see that the percent error of the most inaccurate case of regression tree is two times more than that of multiple linear regression, although the most accurate case of both of them has almost the same percent error (0.0 and 0.00496 individually). It can thus be argued that multiple linear regression is the best method and the result is: r-square = 0.846, RMSE = 28719.

Table 5.6 The results of the most accurate and inaccurate cases

	multiple linear regression		regression tree		
real sale price	60,000	325,300	160,000	145,000	290,000
predicted sale price	28,295.54	325,316.15	555,000	145,000	290,000
percent error	112.047	0.00496	246.875	0.0	0.0

Percent error = |(real sale price – predicted sale price)| / real sale price * 100

As noted above, another aim of this case is to determine the important attributes that could influence the house price. The figure below shows the coefficient values of dependent attributes and we can see that top 3 important features are above grade living area (GrLivArea), overall material and finish quality (OverallQual) and lot size (LotArea).



Row ID	S Variable	D Coeff.	D Std. Err.	D t-value	D P>|t|
Row58	Intercept	833,024.222	1,434,34...	0.581	0.562
Row52	GrLivArea	219,030.653	23,223.154	9.432	0
Row45	OverallQual	133,755.276	11,412.067	11.721	0
Row1	LotArea	104,064.001	21,268.767	4.893	0

Figure 5.7 Coefficients of top 3 features

The best multiple linear regression model can explain 84.6% house price and root mean squared error is 28719 dollars. In order to compare with other models, the author downloads and performs the test data, provided by Kaggle and upload the submission. The rank is top 75%, although the evaluation metrics is different: it is only ranked by RMSE.

One limitation of this study is that the intercept and coefficients of attributes are very high. It is because attribute values are normalized between 0 – 1, but SalePrice is the real value, over 100 thousand. The author, thus, suggest that SalePrice can be normalized when building the model and should be de-normalized as predicted values.

6. Reflection and conclusion

During this study, the author obtains several benefits of data mining. First, several fundamental principles of data mining are fully understood, such as multiple linear regression, regression tree and cross-validation. Second, data processing skills are greatly developed, including feature selection, missing value handle and so forth. In regard to challenges, the first one is to determine the attribute set because there are 79 attributes in this dataset and we need to select significant features to build the model. The second challenge is overfitting. As discussed before, overfitting becomes more likely as the model becomes more complex. Therefore, the author adopts cross-validation in order to estimate model performance and choose the best parameters.

Compared to linear regression models of other research, the performance of this model is not good and further improvement can be done in feature engineering. Marcelino (2017) has confirmed that it is possible to combine several related attributes into one feature, such as basement feature including the height of basement (BsmtQual), the general condition of basement (BsmtCond) and etc. In addition, the report of Cohen-Solal (2016) has confirmed that simplifying existing features can also improve performance. For instance, transforming the value of house quality 1-3 into 1 (bad); 4-6 into 2 (average); 7-9 into 3 (good).

To conclude, the aim of this study is to predict the final price of each home on the basis of 79 attributes. The result of this model shows that above grade living area (GrLivArea), overall material and finish quality (OverallQual) and lot size (LotArea) are top 3 important features. In addition, the performance of multiple linear regression model is better than that of regression tree: R^2 = 87.6% and RMSE =

28719. Further feature engineering work should to conducted, such as feature simplification and combination.

References

Brownlee, J. (2017, July 14). What is the Difference Between Test and Validation Datasets [Blog post]. Retrieved from https://machinelearningmastery.com/difference-test-validation-datasets/

Cock, D. D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. Journal of Statistics Education, 19(3). https://doi.org/10.1080/10691898.2011.11889627

Cohen-Solal, J. (2016). A study on Regression applied to the Ames dataset [Blog post]. Retrieved from https://www.kaggle.com/juliencs/a-study-on-regression-applied-to-the-ames-dataset

Draper, N. R., & Smith, H. (1998). Applied Regression Analysis (3rd ed.). New Jersey: Wiley-Blackwell.

Freedman, D. A. (2009). Statistical Models: Theory and Practice (2nd ed.). Cambridge: Cambridge University Press.

Glantz, S. A., Slinker, B. K., & Neilands, T. B. (2016). Primer Of Applied Regression & Analysis Of Variance (3rd ed.). New York: McGraw-Hill Education.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. Internal Journey of Forecasting, 22(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An Introduction to Statistical Learning with Applications in R (8th ed.). New York: Springer. https://doi.org/10.1007/978-1-4614-7138-7

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2014). Fundamentals of Machine Learning for Predictive Data Analytics: algorithms, worked examples, and case studies. Cambridge: MIT Press.

Marcelino, P. (2017, February). COMPREHENSIVE DATA EXPLORATION WITH PYTHON [Blog Post]. Retrieved from https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

Nekkanti, O. (2017). PREDICTION OF 2016 RENTAL DEMAND IN GREAT RIDES BIKE SHARE PROGRAM (Unpublished master's dissertation). North Dakota State University, Fargo, USA.

Provost, F., & Fawcett, T. (2013). Data Science for Business. California: O'Reilly Media.

Ripley, B. D. (1996). Pattern recognition and neural networks. Cambridge: Cambridge University Press.

Rokach, L., & Maimon, O. (2014). Data Mining with Decision Trees: theory and applications (2nd ed.). Singapore: World Scientific.

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). Massachusetts: Elsevier.