Predicting House Prices: Feature-Selection-Based Linear Regression Models and Tree-Based Regression Models

1. Introduction

Houses' prices are always increasing year after year, which is proven through the historical data of the average sales of houses sold in the US [1]. The increment of the house's prices might be influenced by some factors such as historical sale prices, neighborhood, size and appeal, age, condition, etc. [2]. Hence, buying or selling house might be challenging since the house prices can differ substantially from the initial evaluation of those factors [2].

Determining house prices involves many techniques in the data science field, especially when there are a lot of variables. One of the studies has implemented the houses price prediction based on the tax object sale value and obtained the minimum prediction error through the combination between regression analysis and the particle swarm optimization [3].

The dataset of this project comes from an ongoing competition on Kaggle website [4]. It is about the sale price prediction for residential homes in Ames, Iowa through 79 predictors, including the data about house decorations, quality evaluations, size, etc. According to the introduction, in buying houses, the buyers' desires will not solely base on the height of the basement ceiling or the houses distance to the railroad [4]. Instead, there are still some other influencers that will take part as the consideration of the buyers decision such as the number of bedroom, fence quality, and others [4].

Therefore, the purpose of this project is to find the most important factors and make accurate regression prediction for house prices, which requires avoiding overfitting, reducing the model complexity, and increasing accuracy.

To this end, multiple feature-selection-based linear regression models and tree-based regression models are applied to the dataset in this project. And then, we will calculate root mean square error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price to evaluate all fitted models. Ultimately, the best prediction model and most important factors will be determined based on the evaluation result.

The major challenges in this project are:

- The misrecognition of data types and a lot of NA values in R
- The transformation of 44 qualitative predictors for regression models
- The determination of parameters for different feature selection methods and regression models

2. Data Preparation

2.1. Dataset Description

The dataset is obtained through Kaggle.com [4], specifically about the predict sales of house price. The provided datasets contain one training-set with observed values for response variable and one validation-set without observed values for response variable. Because the validation-set is only used to submit prediction result, this project will focus on the training-set to train prediction models and call the training-set as the original dataset.

The original dataset consists of sample ID, 79 predictors, and 1response variable (house sale price), totally 1460 samples. Specifically, the 79 predictors contain 35 qualitative predictors and 44 quantitative predictors. In more specific way, the 35 qualitative predictors contain 5 interval, 28 ratio and 2 ordinal predictors. On the other hand, the 44 qualitative predictors comprise 14 ordinal and 30 nominal predictors. The details are listed in **Fig. 1** and the explanations are listed in **Appendix 1**.

The minimum, median, mean, and maximum value for response variable are 34900, 162500, 180624, and 755000, respectively.

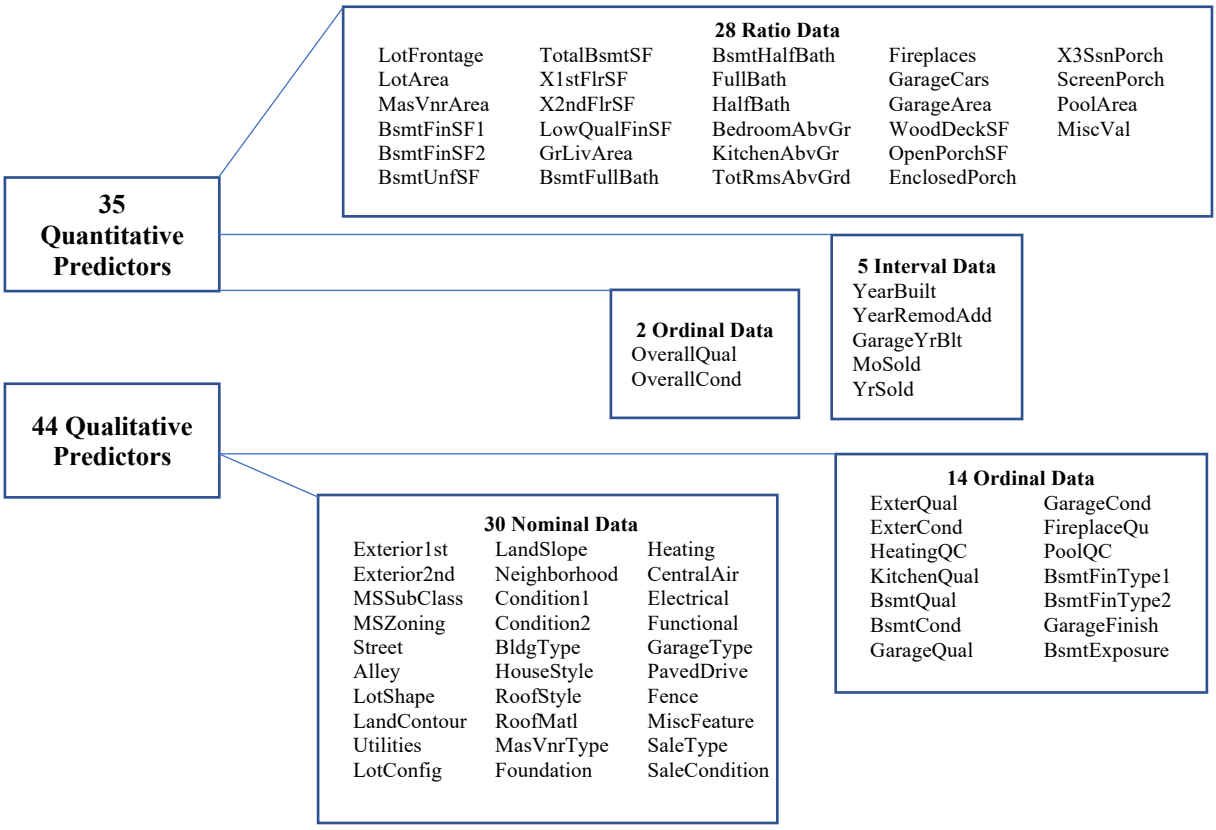

Fig. 1 The division of the original dataset based on different data types

2.2. Data Cleaning

The whole procedures of cleaning data include two main parts which are addressing missing value and dealing with data type problems. The further process explanation of the two aspects will be discussed in the upcoming sections. The graph of procedures for data cleaning is shown in **Appendix 2**. All R-Code and output for data cleaning are listed in **Appendix 3**.

2.2.1. Addressing Missing Values

- **Step 1: Replace "NA" values by "None" for those predictors which include level "NA"**

In the beginning, almost all samples are recognized as containing missing values, which is extraordinary. By further checking the original dataset, we found that the given values for 14 qualitative predictors include level "NA," which means that value "NA" includes valuable information and it is important for the modelling process. For example:

- "NA" in "Alley (Type of alley access to property)" means "No Alley Access"
- "NA" in "BsmtQual (Evaluates the height of the basement)" means "No Basement"
- "NA" in "PoolQC (Pool quality)" means "No Pool"
- …

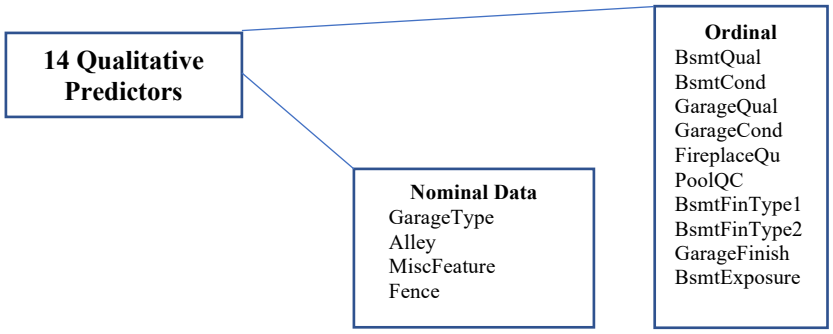

Fig. 2 The 14 predictors which contain level "NA"

Therefore, we need to replace these "NA" values by "None" to avoid that R wrongly reads them as missing values.

- **Step 2: Address real missing values**

  After **Step1**, we further checked the real missing values that are shown in **Fig. 3**.

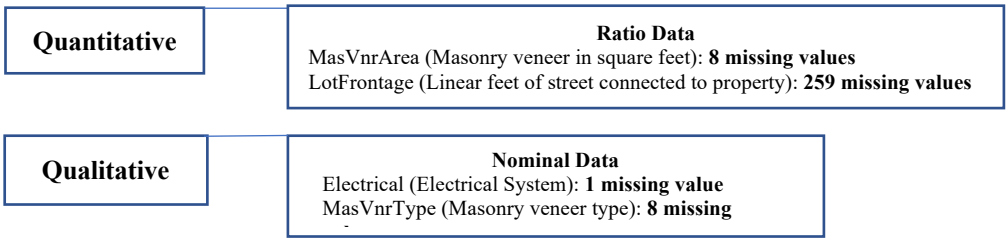

Fig. 3 The predictors containing real missing values.

For predictor "LotFrontage," we replaced its missing values by the average value 70. For predictors "MasVnrType," "MasVnrArea," and "Electrical," we deleted them. And then the total sample size became 1451 from 1460.

2.2.2. Addressing Data Type Problems

In this section, there are two steps that will be done. First, fix the attribute types that cannot be handled in R. Second, check all the data type that to ensure all the data type can be used for the further analysis.

- **Step 3: Adjust attribute types for predictors that are misrecognized in R**

  By checking the dataset description file, we found that levels of nominal data "MSSubClass" are integer, which would result in the misrecognized in R. Therefore, we change its datatype to factor.

MSSubClass: Identifies the type of dwelling involved in the sale.	
20	1-STORY 1946 & NEWER ALL STYLES
30	1-STORY 1945 & OLDER
40	1-STORY W/FINISHED ATTIC ALL AGES
…	……

By checking the structure of the dataset, predictors "LotFrontage (Linear feet of street connected to property)" and "GarageYrBlt (Year garage was built)" have almost 100 levels, which is weird. In fact, "Linear feet of street connected to property" and "Year garage was built" should be quantitative data. Therefore, we changed them to integer in R.

- **Step 4: Check data types and transfer qualitative data**

As the dataset description in **Section 2.1.,** there are 44 qualitative predictors in the original dataset. They cannot be directly used in the linear regression model. Therefore, we need to transform them to quantitative data based on some criterions.

As for the 14 ordinal predictors, they are changed to orderly integer, like 0, 1, 2, …, with the assumptions that "a stuff is poor" is better than "not finished yet" and "not finished yet" is better than "no the stuff." For example, change the levels of predictor "ExterQual (Evaluates the quality of the material on the exterior)" from Ex (Excellent), Gd (Good), TA (Average/Typical), Fa (Fair), and Po (Poor) to 5, 4, 3, 2, and 1, respectively.

As for the 30 nominal predictors, if the number of their levels is 2, we will create one dummy variable with values 0 and 1 to represent the two level; if the number of their levels (k) is greater than 2, we will create k dummy variables with values 0 and 1 to represent whether one specific level the sample is or not. Totally, we created 190 dummy variables. And then, we deleted the original 30 nominal predictors.

Finally, the cleaned dataset contains **239 predictors, 1 housing ID, and 1 response variable**.

3. Modeling

The goal of this project is to use regression models to predict the sale price for residential homes in Ames, Iowa through 79 predictors, including the data about house decorations, quality evaluations, size, and time. All data has been cleaned up in section 2, consisting of 239 integer predictors.

However, so many predictors may result in overfitting, which will reduce the accuracy of the prediction for validation samples. Moreover, irrelevant predictors will lead to unnecessary complexity, which makes it harder to see the effect of important predictors. This project will mainly focus on the two key problems to train the most reliable model that satisfies high accuracy, avoids overfitting, reduces the complexity, and further find the most important predictors.

To this end, **feature selection methods** and **tree-based models** are applied to the dataset in this project. Specifically, based on the linear regression model, **forward stepwise selection** and **lasso regression** are performed to conduct the feature selection and reduce the dimension. Besides this, tree-based models, like **decision tree** and **bagged trees (Random Forest),** can recursively select and partition features through the greedy criteria. Due to the computational reasons for high dimensions, best subset selection method **is not** selected here.

The whole modeling procedures are shown in **Fig. 4** The first step is to divide the dataset to a training-set and a test-set (**Section 3.1.**). And then we will build regression models on the training-set through feature selection method (**Section 3.2.**) and tree-based models (**Section 3.3.**). Finally, we will show all model properties and evaluate built models to select the best model (**Section 3.4.**). All R-Code and output for modeling are listed in **Appendix 4**.

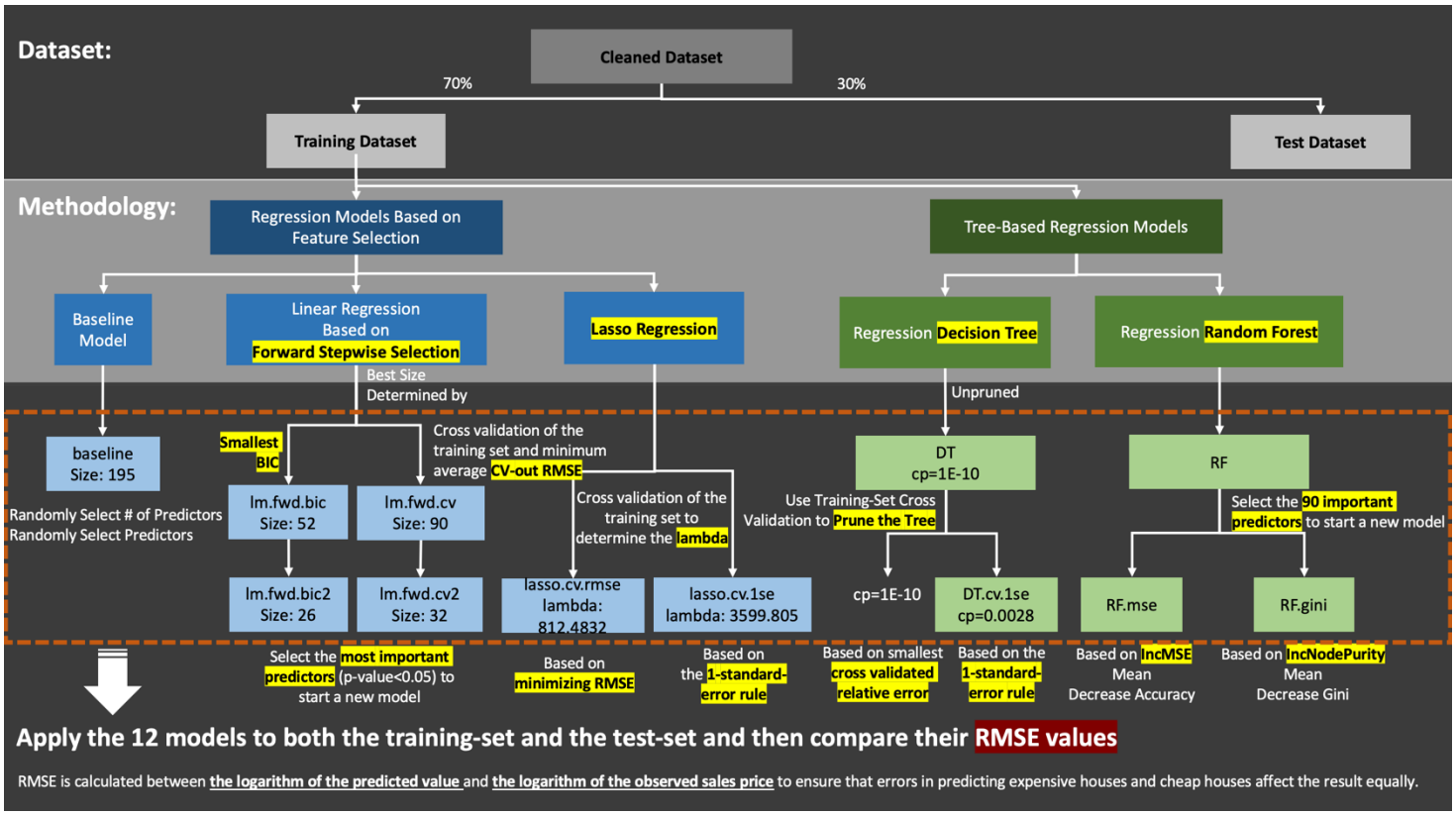

Fig.4 The Procedures of Modeling

3.1. Dataset Division

To evaluate models efficiently and avoid overfitting, the cleaned dataset is randomly divided into two parts: a training set and a test set, accounting for 70% and 30%, respectively. After the division, there are 1015 training samples and 436 test samples.

The 1015 training samples will be applied to train different regression models. And then, based on the fitted models, we will obtain the estimated values for the 436 test samples. Ultimately, the model with the highest accuracy will be determined by the validations of different models based on estimated values and observed values for response variable in test-set.

3.2. Regression Models Based on Feature Selection

3.2.1. Baseline Model: Multiple Linear Regression with Random Feature Selection

1) Methodology

First of all, the multiple linear regression model is specified as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \quad (1)$$

where Y is the response variable ("Sale Price" in this project), $\{\beta_0, \beta_1, \ldots, \beta_p\}$ are p predictors or features, $\{\beta_0, \beta_1, \ldots, \beta_p\}$ are p unknown constants that represent the intercept and slope, also known as coefficients, and ϵ is the error term. The p coefficients are determined by minimum residual sum of square $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where n ($n = 1015$ in this project) is the sample size of training-set, y_i is the observed value for the response variable, $\hat{y}_i$ is the predicted value for the response variable.

2) Experiment and Discussion

This project is to select most important and relevant p predictors from total 239 predictors to predict the sale price as accurate as possible. Hence, to check whether the proposed feature selection models perform well, we **randomly determine the number of predictors, 154**, and select the predictors to build a multiple linear regression baseline model, called "**baseline**." The diagnostic plots are shown in **Fig. 5**.

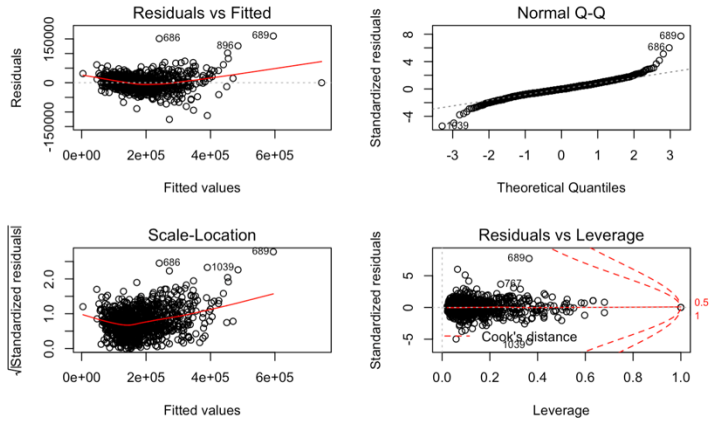

Fig. 5 The diagnostic plots for baseline model

- In the plot of **residuals vs. fitted values**, the red smooth line of residuals around shows a little curvature; and the shape of residuals is cone shape. Therefore, **the linearity assumption** and **independent-errors assumption** are violated.
- In **the normal Q-Q plot**, the points on the left and right are far away the reference line, which means **the normality assumption** is violated.
- In the **scale-location plot**, though the red smooth line is roughly near line y=1, it still shows significant curvature, which means **the equal-variance assumption** is violated.

Therefore, though the adjusted R^2 of "**baseline**" is **0.8968**, it would overfit the training-set and mislead the prediction results. The further evaluation for test-set will be shown in **Section 3.4.**

3.2.2. Linear Regression Based on Forward Stepwise Selection

1) Methodology

First of all, **forward stepwise selection** begins with a null linear regression model containing no predictors, and then adds one predictor at a time that improves the model the most until no improvement is possible. And then, we need to set criterions to choose which size is best. Here, we will apply two kinds of methods, which are:

- **Adjusted R^2, BIC (Bayesian Information Criterion), and Mallow's C_p:** The three methods can adjust the training error for the model size and can be used to select among a set of models with different numbers of variable. Specifically, the smaller value of C_p and BIC indicates a better model, and the larger value for Adjusted R^2 indicates a better model.
- **Cross Validation of the training-set**: For each size, we conduct the cross validation of training-set and calculate the average RMSE for fold-out sets. Best size can be determined by the minimum average RMSE of fold-out sets.

2) Experiment and Discussion

In the beginning, in R, we set **the maximum size of subsets to examine as 100**. For all 100 steps, we obtained the plots of Adjusted R^2, BIC, and C_p for fitted models and the plot of average RMSE for fold-out sets in **Fig. 6** and **Fig. 7**, respectively.

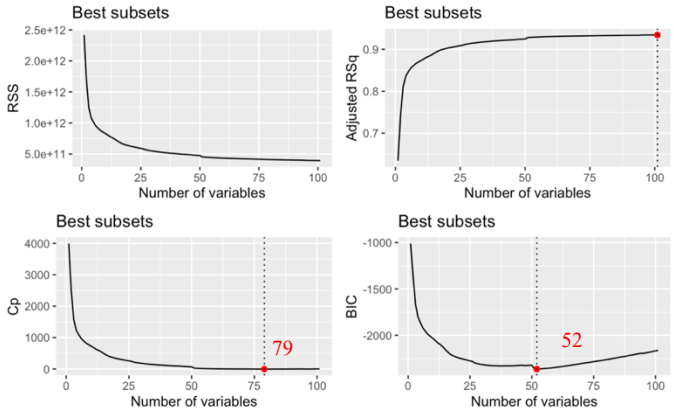

Fig. 6 The plots of Adjusted R^2, BIC, and C_p for each step

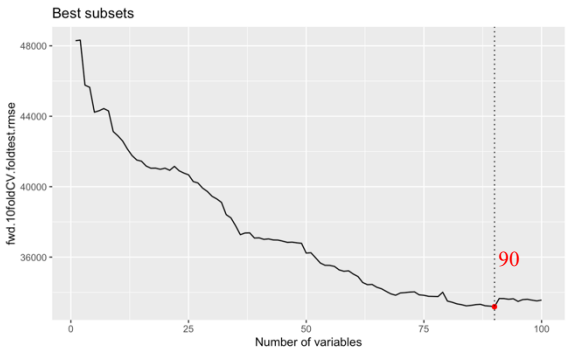

Fig. 7 the plot of average RMSE for fold-out sets for each step

Here, we obtained the best size are **79**, **52**, and **90** for criterions **of C_p**, **BIC**, and **Cross Validation**, respectively. Therefore, we will train the models **lm.fwd.bic** and **lm.fwd.cv** based on the model size obtained from BIC and Cross Validation. Furthermore, we will improve the two models to **lm.fwd.bic2** and **lm.fwd.cv2** through selecting the most important and relevant predictors whose p-value is less than 0.05.

3.2.3. Lasso Regression

1) Methodology

Based on the multiple linear regression formula **(1)** in **Section 3.2.1.**, the Lasso regression improves the method of determining p coefficients by adding a shrinkage penalty and minimize the quantity:$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$, where n ($n = 1015$ in this project) is the sample size of training-set, y_i is the observed value for the response variable, $\hat{y}_i$ is the predicted value for the response variable, λ is a tuning parameter, $\beta_j$is the coefficient for predictor j, and p is the total number of predictors. As λ increases, the Lasso regression coefficients shrinks towards zero. And then, we will achieve the goal of feature selection.

2) Experiment and Discussion

Use training-set to perform lasso regression for 100 lambda values, and then use cross validation of the training set to determines the lambda minimizing RMSE. Also, we can determine lambda for the 1-standard-error rule. Finally, the plot of coefficient paths is shown in **Fig. 8**. And the **lambda minimizing RMSE** is **812.4831**; the **lambda for the 1-standard-error rule** is **3599.805**. We call the two models are **lasso.cv.rmse** and **lasso.cv.1se**, respectively.

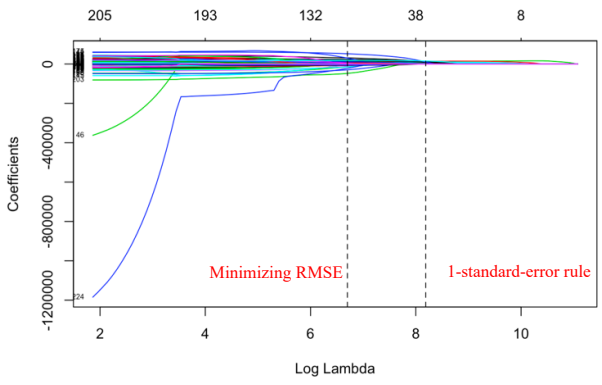

Fig. 8 The plot of coefficient paths for lasso regression

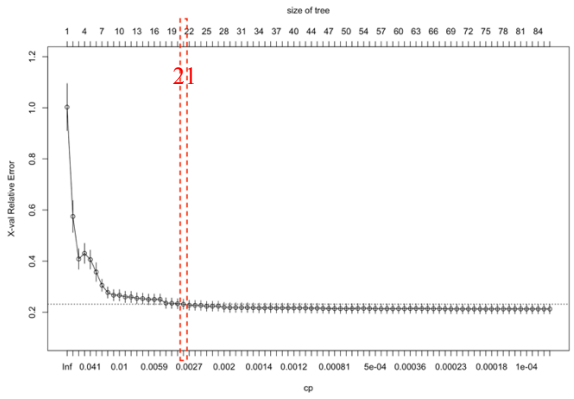

Fig. 9 Cross-Validation-Error plot for unpruned tree (cp=1E-10)

3.3. Tree-Based Regression Models

3.3.1. Regression Decision Tree

The regression decision tree is a top-down, greedy, recursive partitioning algorithm in form of a tree structure [5]. It reduces the samples' overall impurity by breaking down a dataset into smaller and smaller subsets [5]. And the complexity parameter (cp) is used to determine the size of the tree [5]. In the beginning, we built an unpruned tree **DT** with a pretty small cp value, **1E-10**.

And then, we use **cross-validated method** to prune tree to address overfitting. The cross-validation error plot is shown in **Fig.9**. It indicates that the tree with smallest cp value has **the smallest cross validated relative error**, which means the unpruned tree performs well. The plot also indicates that a tree of **size 21** satisfies the **1-standard-error rule** and we can prune to this size of the tree using **cp=0.0028**. The pruned tree plot is shown in **Appendix 4 (2.1.2)** and we call it as **DT.cv.1se**.

3.3.2. Regression Random Forest

Regression random forest consists of many regression decision trees. As for each tree, the predictors are determined by random selection from the whole predictors; and the training samples are selected by bootstrap sampling from the original training-set. The two properties greatly overcome the problem of overfitting. And the final prediction result can be obtained from the average outputs of all individual trees.

In the beginning, we train the model **RF** by setting the number of variables randomly sampled as candidates at each split as 80 ($^{239}/_{3} \approx 80$) and the number of trees to grow as 500. The relative error plot for different size of random forest model is shown in **Appendix 4 (2.2.1)**. We can see that the accuracy of model tends to be stable at size of 250. And then, based on mean decrease accuracy (IncMSE) and mean decrease gini (IncNodePurity), we find the most important 90 predictors to train two new models **RF.mse** and **RF.gini**, respectively.

3.4. Validation and Selection of Models

Above all, we obtained totally 12 models for the problem. To further evaluate their performance, we use them to predict both the training-set and the test-set, and then compare the RMSE values for different models. To ensure that errors in predicting expensive houses and cheap houses affect the result equally, RMSE is calculated between the logarithm of the predicted value and the logarithm of the observed sales price.

Ultimately, the validation results, methodology, properties of all built models are shown in **Table 1**. And the details of all models, including selected predictors, coefficients, plots, and codes, are shown in **Appendix 4**. According to **Table 1**, We can see that model **RF.gini** has best performance on training-set with 30 candidates at each split and the smallest RMSE value, 0.1551. Therefore**, the final model** will be obtained by fitting all training-set and test-set data on **RF.gini.**

According to **Table 1,** all built models perform better than the **baseline** model. Also, by selecting most important predictors, **lm.fwd.bic2, lm.fwd.cv2, RF.gini,** and **RF.mse** all improve the performance of original models. Among all regression models based on feature selection, **lasso.cv.1se** has the best performance with the smallest test-set RMSE value, 0.18. As for the four forward stepwise selection models, **lm.fwd.cv2** performs best. As for the pruned tree **DT.cv.1se** of model **DT**, it does not improve the performance of the original model. However, it performs similar with the forward stepwise selection model **lm.fwd.bic**. and it has the smallest number of predictors, **11**.

	Names	TestSet RMSE	TrainingSet RMSE	Number of Predictors	Parameters	Methods	Property
Regression Models Based on Feature Selection	baseline	0.2918	0.1547	154/239 selected	predictors	RandomSelected	adj.r2=0.8968
	lm.fwd.bic	0.2132	0.1974	52/239 selected	size=52	BIC.min	adj.r2=0.8524
	lm.fwd.bic2	0.204	0.2076	26/239 selected	predictors.in.bic.model	most.important.p<0.05	adj.r2=0.8528
	lm.fwd.cv	0.1941	0.1695	90/239 selected	size=90	CVout.RMSE.min	adj.r2=0.8867
	lm.fwd.cv2	0.1909	0.1692	32/239 selected	predictors.in.cv.model	most.important.p<0.05	adj.r2=0.8837
	lasso.cv.rmse	0.1859	0.1296	98/239 selected	lambda=812.4831	CVout.RMSE.min	-
	lasso.cv.1se	0.18	0.1578	98/239 selected	lambda=3599.805	CVout.1se.rule	-
Tree-Based Models	DT	0.1999	0.1195	29/239 used	cp=0.0000000001	-	nsplit=85
	DT.cv.1se	0.2164	0.1737	11/239 used	cp=0.0028	Pruned-CV-Error-Plot.1se.rule	nsplit=20
	RF	0.1561	0.0621	80/239 candidates	mtry=80,ntree=500	mtry=number.of.variables/3	-
	RF.gini	0.1551	0.0619	30/239 candidates	candidates	most.important.90atts.IncNodePurity	-
	RF.mse	0.1554	0.0614	30/239 candidates	candidates	most.important.90atts.IncMSE	-
The Final Model	Final.Model	**0.1452**	0.0627	30/239 candidates	-	Use **RF.gini** to fit the entire dataset Test-set score obtained from Kaggle	-

Methodology:* lm**: *multiple linear regression;* ***fwd: *forward stepwise selection;* ***cv:*** *cross validation;* ***lasso:*** *lasso regression;* ***rmse***: *root mean square error;* ***1se:*** *one-standard-error rule;* ***DT***: *decision tree;* ***RF:*** *randomforest;* ***IncMSE:*** *mean decrease accurace;* ***IncNodePurity:*** *mean decrease gini.*

Table 1 Final Results

At last, we submitted the result for unobserved data on Kaggle website. The RMSE score is 0.1452. It performs better than all test-set RMSE in this project.

Based on the hypothesis test on coefficients of best multiple linear regression model **lm.fwd.cv2** in **Table 2**, we can find that the most important 3 predictors are "OverallQual," "MSSubClass_60," and "SaleType_New." Based on the IncNodePurity value of the best tree-based model **RF.gini** in **Table 3**, we can find that the most important 3 predictors are "OverallQual," "GrLivArea," and "ExterQual." These conclusions can help the house developers make decisions for designing and selling strategies, like focusing more on the quality of the material on the exterior (ExterQual).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.340308e+05	1.096256e+04	-12.226228	4.250624e-32
OverallQual	1.329406e+04	1.175376e+03	11.310471	5.739132e-28
MSSubClass_60	2.863308e+04	3.496764e+03	8.188452	8.173150e-16
SaleType_New	2.696264e+04	3.635164e+03	7.417173	2.583909e-13
OverallCond	6.522930e+03	9.148235e+02	7.130261	1.940566e-12
X1stFlrSF	3.743010e+01	5.605400e+00	6.677507	4.060827e-11
LotArea	7.206438e-01	1.087386e-01	6.627306	5.628586e-11
TotRmsAbvGrd	4.967481e+03	7.901341e+02	6.286883	4.865376e-10
GarageType_BuiltIn	2.581007e+04	4.167805e+03	6.192725	8.680606e-10
MasVnrArea	3.679381e+01	6.055975e+00	6.075621	1.764411e-09

Table 2 Most important 10 predictors in model **lm.fwd.cv2**

	%IncMSE	IncNodePurity
OverallQual	25.687378	1.579919e+12
GrLivArea	36.102186	8.068590e+11
ExterQual	14.533457	6.892731e+11
GarageCars	11.835467	4.420522e+11
TotalBsmtSF	16.761648	3.145091e+11
YearBuilt	11.460444	2.983114e+11
BsmtFinSF1	18.456440	2.780989e+11
X1stFlrSF	13.191865	2.554032e+11
KitchenQual	8.973273	1.821043e+11
GarageArea	10.772787	1.797083e+11

Table 3 Most important 10 predictors in model **RF.gini**

4. Conclusion and Future Works

This project established multiple kinds of feature selection and tree-based regression models to predict house price. Ultimately, we built the final model through fitting the entire dataset on the best model **RF.gini**. Therefore, we can use it to predict unobserved samples to evaluate the house price. Moreover, the importance evaluation of predictors can guide house developers to make decisions for designing and selling strategies. In the future, to further improve the accuracy of feature selection models and overcome overfitting, we can apply ensemble learning methods on those models.

References:

[1] J. FOLGER, "The Truth About Real Estate Prices," Investopedia, 22 April 2020. [Online]. Available: https://www.investopedia.com/articles/mortages-real-estate/11/the-truth-about-the-real-estate-market.asp. [Accessed 11 May 2020].

[2] A. JOHANSSON, "6 factors that influence a home's value," inman, 7 August 2017. [Online]. Available: https://www.inman.com/2017/08/07/6-factors-that-influence-a-homes-value/. [Accessed 11 May 2020].

[3] A. N. Alfiyatin, H. Taufiq, R. E. Febrita and W. F. Mahmudy, "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization," *International Journal of Advanced Computer Science and Applications,* vol. 8, no. 10, pp. 323-326, 2017.

[4] kaggle, "House Prices: Advanced Regression Techniques," Kaggle, [Online]. Available: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview. [Accessed 11 May 2020].

[5] S. Sayad, "Decision tree - regression," [Online]. Available: https://saedsayad.com/decision_tree_reg.htm.

[6] M. Becker, "Sorry, But Your Home Isn't an 'Investment'," The Simple Dollar, 29 October 2019. [Online]. Available: https://www.thesimpledollar.com/investing/real-estate/sorry-but-your-home-isnt-a-good-investment/. [Accessed 11 May 2020].

Appendix 1: Dataset Description

Names/Values	Description	Names/Values	Description
MSSubClass	**: Identifies the type of dwelling involved in the sale.**	**MasVnrType**	**: Masonry veneer type**
20	1-STORY 1946 & NEWER ALL STYLES	BrkCmn	Brick Common
30	1-STORY 1945 & OLDER	BrkFace	Brick Face
40	1-STORY W/FINISHED ATTIC ALL AGES	CBlock	Cinder Block
45	1-1/2 STORY - UNFINISHED ALL AGES	None	None
50	1-1/2 STORY FINISHED ALL AGES	Stone	Stone
60	2-STORY 1946 & NEWER	**MasVnrArea**	**: Masonry veneer area in square feet**
70	2-STORY 1945 & OLDER	**ExterQual**	**: Evaluates the quality of the material on the exterior**
75	2-1/2 STORY ALL AGES	Ex	Excellent
80	SPLIT OR MULTI-LEVEL	Gd	Good
85	SPLIT FOYER	TA	Average/Typical
90	DUPLEX - ALL STYLES AND AGES	Fa	Fair
120	1-STORY PUD (Planned Unit Development) - 1946 & NEWER	Po	Poor

Names/Values	Description	Names/Values	Description
150	1-1/2 STORY PUD - ALL AGES	**ExterCond**	**: Evaluates the present condition of the material on the exterior**
160	2-STORY PUD - 1946 & NEWER	Ex	Excellent
180	PUD - MULTILEVEL - INCL SPLIT LEV/FOYER	Gd	Good
190	2 FAMILY CONVERSION - ALL STYLES AND AGES	TA	Average/Typical
MSZoning	**: Identifies the general zoning classification of the sale.**	Fa	Fair
A	Agriculture	Po	Poor
C	Commercial	**Foundation**	**: Type of foundation**
FV	Floating Village Residential	BrkTil	Brick & Tile
I	Industrial	CBlock	Cinder Block
RH	Residential High Density	PConc	Poured Contrete
RL	Residential Low Density	Slab	Slab
RP	Residential Low-Density Park	Stone	Stone
RM	Residential Medium Density	Wood	Wood
LotFrontage	**: Linear feet of street connected to property**	**BsmtQual**	**: Evaluates the height of the basement**
LotArea	**: Lot size in square feet**	Ex	Excellent (100+ inches)
Street	**: Type of road access to property**	Gd	Good (90-99 inches)
Grvl	Gravel	TA	Typical (80-89 inches)
Pave	Paved	Fa	Fair (70-79 inches)
Alley	**: Type of alley access to property**	Po	Poor (<70 inches
Grvl	Gravel	NA	No Basement
Pave	Paved	**BsmtCond**	**: Evaluates the general condition of the basement**
NA	No alley access	Ex	Excellent
LotShape	**: General shape of property**	Gd	Good
Reg	Regular	TA	Typical - slight dampness allowed
IR1	Slightly irregular	Fa	Fair - dampness or some cracking or settling
IR2	Moderately Irregular	Po	Poor - Severe cracking, settling, or wetness
IR3	Irregular	NA	No Basement
LandContour	**: Flatness of the property**	**BsmtExposure**	**: Refers to walkout or garden level walls**
Lvl	Near Flat/Level	Gd	Good Exposure
Bnk	Banked - Quick and significant rise from street grade to building	Av	Average Exposure (split levels or foyers typically score average or above)
HLS	Hillside - Significant slope from side to side	Mn	Mimimum Exposure
Low	Depression	No	No Exposure
Utilities	**: Type of utilities available**	NA	No Basement
AllPub	All public Utilities (E, G, W,& S)	**BsmtFinType1**	**: Rating of basement finished area**

Names/Values	Description	Names/Values	Description
NoSewr	Electricity, Gas, and Water (Septic Tank)	GLQ	Good Living Quarters
NoSeWa	Electricity and Gas Only	ALQ	Average Living Quarters
ELO	Electricity only	BLQ	Below Average Living Quarters
LotConfig	**: Lot configuration**	Rec	Average Rec Room
Inside	Inside lot	LwQ	Low Quality
Corner	Corner lot	Unf	Unfinshed
CulDSac	Cul-de-sac	NA	No Basement
FR2	Frontage on 2 sides of property	**BsmtFinSF1**	**: Type 1 finished square feet**
FR3	Frontage on 3 sides of property	**BsmtFinType2**	**: Rating of basement finished area (if multiple types)**
LandSlope	**: Slope of property**	GLQ	Good Living Quarters
Gtl	Gentle slope	ALQ	Average Living Quarters
Mod	Moderate Slope	BLQ	Below Average Living Quarters
Sev	Severe Slope	Rec	Average Rec Room
Neighborhood	**: Physical locations within Ames city limits**	LwQ	Low Quality
Blmngtn	Bloomington Heights	Unf	Unfinshed
Blueste	Bluestem	NA	No Basement
BrDale	Briardale	**BsmtFinSF2**	**: Type 2 finished square feet**
BrkSide	Brookside	**BsmtUnfSF**	**: Unfinished square feet of basement area**
ClearCr	Clear Creek	**TotalBsmtSF**	**: Total square feet of basement area**
CollgCr	College Creek	**Heating**	**: Type of heating**
Crawfor	Crawford	Floor	Floor Furnace
Edwards	Edwards	GasA	Gas forced warm air furnace
Gilbert	Gilbert	GasW	Gas hot water or steam heat
IDOTRR	Iowa DOT and Railroad	Grav	Gravity furnace
MeadowV	Meadow Village	OthW	Hot water or steam heat other than gas
Mitchel	Mitchell	Wall	Wall furnace
Names	North Ames	**HeatingQC**	**: Heating quality and condition**
NoRidge	Northridge	Ex	Excellent
NPkVill	Northpark Villa	Gd	Good
NridgHt	Northridge Heights	TA	Average/Typical
NWAmes	Northwest Ames	Fa	Fair
OldTown	Old Town	Po	Poor
SWISU	South & West of Iowa State University	**CentralAir**	**: Central air conditioning**
Sawyer	Sawyer	N	No
SawyerW	Sawyer West	Y	Yes
Somerst	Somerset	**Electrical**	**: Electrical system**
StoneBr	Stone Brook	SBrkr	Standard Circuit Breakers & Romex

Names/Values	Description	Names/Values	Description
Timber	Timberland	FuseA	Fuse Box over 60 AMP and all Romex wiring (Average)
Veenker	Veenker	FuseF	60 AMP Fuse Box and mostly Romex wiring (Fair)
Condition1	**: Proximity to various conditions**	FuseP	60 AMP Fuse Box and mostly knob & tube wiring (poor)
Artery	Adjacent to arterial street	Mix	Mixed
Feedr	Adjacent to feeder street	**1stFlrSF**	**: First Floor square feet**
Norm	Normal	**2ndFlrSF**	**: Second floor square feet**
RRNn	Within 200' of North-South Railroad	**LowQualFinSF**	**: Low quality finished square feet (all floors)**
RRAn	Adjacent to North-South Railroad	**GrLivArea**	**: Above grade (ground) living area square feet**
PosN	Near positive off-site feature--park, greenbelt, etc.	**BsmtFullBath**	**: Basement full bathrooms**
PosA	Adjacent to positive off-site feature	**BsmtHalfBath**	**: Basement half bathrooms**
RRNe	Within 200' of East-West Railroad	**FullBath**	**: Full bathrooms above grade**
RRAe	Adjacent to East-West Railroad	**HalfBath**	**: Half baths above grade**
Condition2	**: Proximity to various conditions (if more than one is present)**	**Bedroom**	**: Bedrooms above grade (does NOT include basement bedrooms)**
Artery	Adjacent to arterial street	**Kitchen**	**: Kitchens above grade**
Feedr	Adjacent to feeder street	**KitchenQual**	**: Kitchen quality**
Norm	Normal	Ex	Excellent
RRNn	Within 200' of North-South Railroad	Gd	Good
RRAn	Adjacent to North-South Railroad	TA	Typical/Average
PosN	Near positive off-site feature--park, greenbelt, etc.	Fa	Fair
PosA	Adjacent to positive off-site feature	Po	Poor
RRNe	Within 200' of East-West Railroad	**TotRmsAbvGrd**	**: Total rooms above grade (does not include bathrooms)**
RRAe	Adjacent to East-West Railroad	**Functional**	**: Home functionality (Assume typical unless deductions are warranted)**
BldgType	**: Type of dwelling**	Typ	Typical Functionality
1Fam	Single-family Detached	Min1	Minor Deductions 1
2FmCon	Two-family Conversion; originally built as one-family dwelling	Min2	Minor Deductions 2
Duplx	Duplex	Mod	Moderate Deductions
TwnhsE	Townhouse End Unit	Maj1	Major Deductions 1
TwnhsI	Townhouse Inside Unit	Maj2	Major Deductions 2
HouseStyle	**: Style of dwelling**	Sev	Severely Damaged
1Story	One story	Sal	Salvage only
1.5Fin	One and one-half story: 2nd level finished	**Fireplaces**	**: Number of fireplaces**
1.5Unf	One and one-half story: 2nd level unfinished	**FireplaceQu**	**: Fireplace quality**

Names/Values	Description
2Story	Two story
2.5Fin	Two and one-half story: 2nd level finished
2.5Unf	Two and one-half story: 2nd level unfinished
SFoyer	Split Foyer
SLvl	Split Level
OverallQual	**: Rates the overall material and finish of the house**
10	Very Excellent
9	Excellent
8	Very Good
7	Good
6	Above Average
5	Average
4	Below Average
3	Fair
2	Poor
1	Very Poor
OverallCond	**: Rates the overall condition of the house**
10	Very Excellent
9	Excellent
8	Very Good
7	Good
6	Above Average
5	Average
4	Below Average
3	Fair
2	Poor
1	Very Poor
YearBuilt	**: Original construction date**
YearRemodAdd	**: Remodel date (same as construction date if no remodeling or additions)**
RoofStyle	**: Type of roof**
Flat	Flat
Gable	Gable
Gambrel	Gabrel (Barn)
Hip	Hip
Mansard	Mansard

Names/Values	Description
Ex	Excellent - Exceptional Masonry Fireplace
Gd	Good - Masonry Fireplace in main level
TA	Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement
Fa	Fair - Prefabricated Fireplace in basement
Po	Poor - Ben Franklin Stove
NA	No Fireplace
GarageType	**: Garage location**
2Types	More than one type of garage
Attchd	Attached to home
Basment	Basement Garage
BuiltIn	Built-In (Garage part of house - typically has room above garage)
CarPort	Car Port
Detchd	Detached from home
NA	No Garage
GarageYrBlt	**: Year garage was built**
GarageFinish	**: Interior finish of the garage**
Fin	Finished
RFn	Rough Finished
Unf	Unfinished
NA	No Garage
GarageCars	**: Size of garage in car capacity**
GarageArea	**: Size of garage in square feet**
GarageQual	**: Garage quality**
Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor
NA	No Garage
GarageCond	**: Garage condition**
Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor

Names/Values	Description	Names/Values	Description
Shed	Shed	NA	No Garage
RoofMatl	**: Roof material**	**PavedDrive**	**: Paved driveway**
ClyTile	Clay or Tile	Y	Paved
CompShg	Standard (Composite) Shingle	P	Partial Pavement
Membran	Membrane	N	Dirt/Gravel
Metal	Metal	**WoodDeckSF**	**: Wood deck area in square feet**
Roll	Roll	**OpenPorchSF**	**: Open porch area in square feet**
Tar&Grv	Gravel & Tar	**EnclosedPorch**	**: Enclosed porch area in square feet**
WdShake	Wood Shakes	**3SsnPorch**	**: Three season porch area in square feet**
WdShngl	Wood Shingles	**ScreenPorch**	**: Screen porch area in square feet**
Exterior1st	**: Exterior covering on house**	**PoolArea**	**: Pool area in square feet**
AsbShng	Asbestos Shingles	**PoolQC**	**: Pool quality**
AsphShn	Asphalt Shingles	Ex	Excellent
BrkComm	Brick Common	Gd	Good
BrkFace	Brick Face	TA	Average/Typical
CBlock	Cinder Block	Fa	Fair
CemntBd	Cement Board	NA	No Pool
HdBoard	Hard Board	**Fence**	**: Fence quality**
ImStucc	Imitation Stucco	GdPrv	Good Privacy
MetalSd	Metal Siding	MnPrv	Minimum Privacy
Other	Other	GdWo	Good Wood
Plywood	Plywood	MnWw	Minimum Wood/Wire
PreCast	PreCast	NA	No Fence
Stone	Stone	**MiscFeature**	**: Miscellaneous feature not covered in other categories**
Stucco	Stucco	Elev	Elevator
VinylSd	Vinyl Siding	Gar2	2nd Garage (if not described in garage section)
Wd Sdng	Wood Siding	Othr	Other
WdShing	Wood Shingles	Shed	Shed (over 100 SF)
Exterior2nd	**: Exterior covering on house (if more than one material)**	TenC	Tennis Court
AsbShng	Asbestos Shingles	NA	None
AsphShn	Asphalt Shingles	**MiscVal**	**: $Value of miscellaneous feature**
BrkComm	Brick Common	**MoSold**	**: Month Sold (MM)**
BrkFace	Brick Face	**YrSold**	**: Year Sold (YYYY)**
CBlock	Cinder Block	**SaleType**	**: Type of sale**
CemntBd	Cement Board	WD	Warranty Deed - Conventional
HdBoard	Hard Board	CWD	Warranty Deed - Cash
ImStucc	Imitation Stucco	VWD	Warranty Deed - VA Loan

Names/Values	Description	Names/Values	Description
MetalSd	Metal Siding	New	Home just constructed and sold
Other	Other	COD	Court Officer Deed/Estate
Plywood	Plywood	Con	Contract 15% Down payment regular terms
PreCast	PreCast	ConLw	Contract Low Down payment and low interest
Stone	Stone	ConLI	Contract Low Interest
Stucco	Stucco	ConLD	Contract Low Down
VinylSd	Vinyl Siding	Oth	Other
Wd Sdng	Wood Siding	**SaleCondition**	**: Condition of sale**
WdShing	Wood Shingles	Normal	Normal Sale
		Abnorml	Abnormal Sale - trade, foreclosure, short sale
		AdjLand	Adjoining Land Purchase
		Alloca	Allocation - two linked properties with separate deeds, typically condo with a garage unit
		Family	Sale between family members
		Partial	Home was not completed when last assessed (associated with New Homes)

Appendix 2: The procedures of data cleaning