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

STAT 515 Final Project

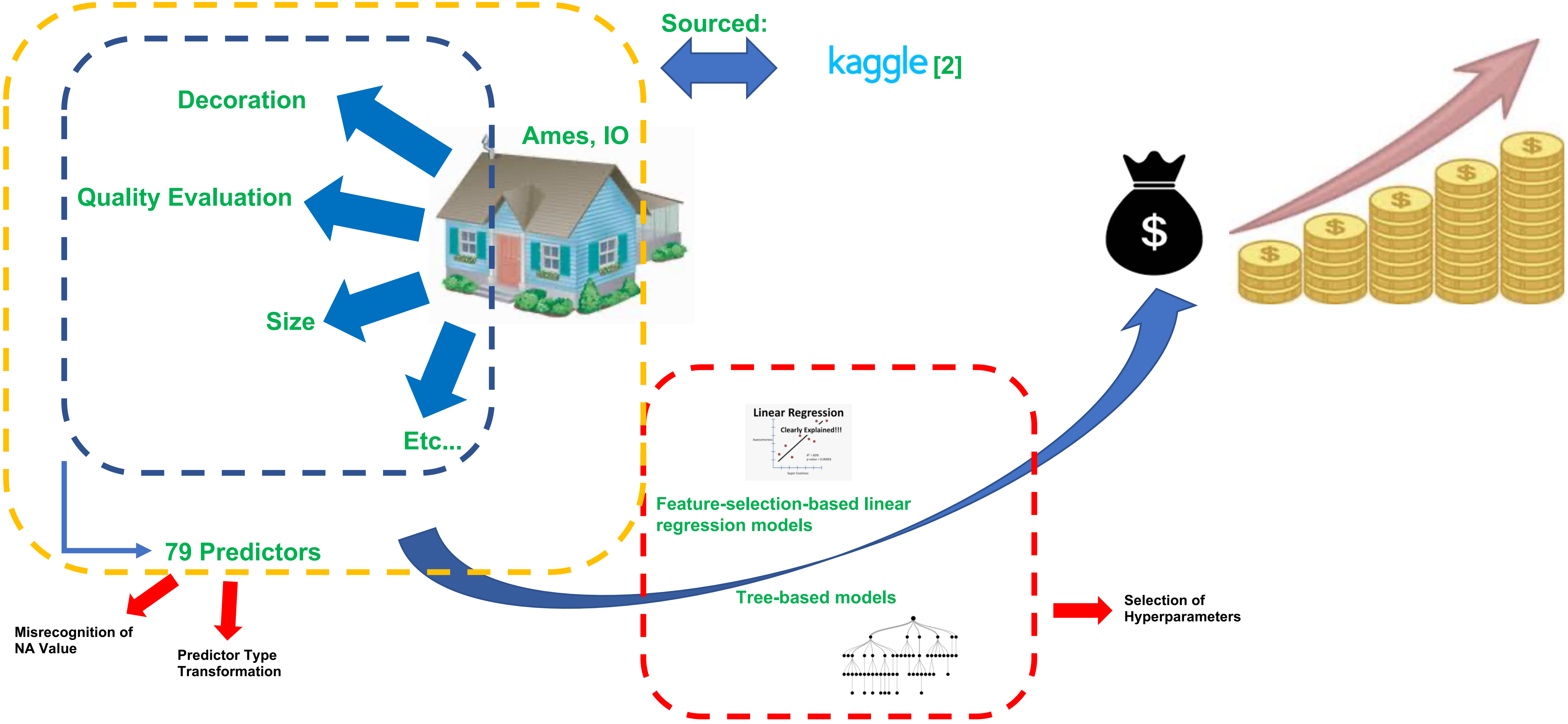
Long Zhang
03.05.2020

Content

1. Introduction
2. Data Preparation
3. Modeling
4. Validation
5. Conclusion

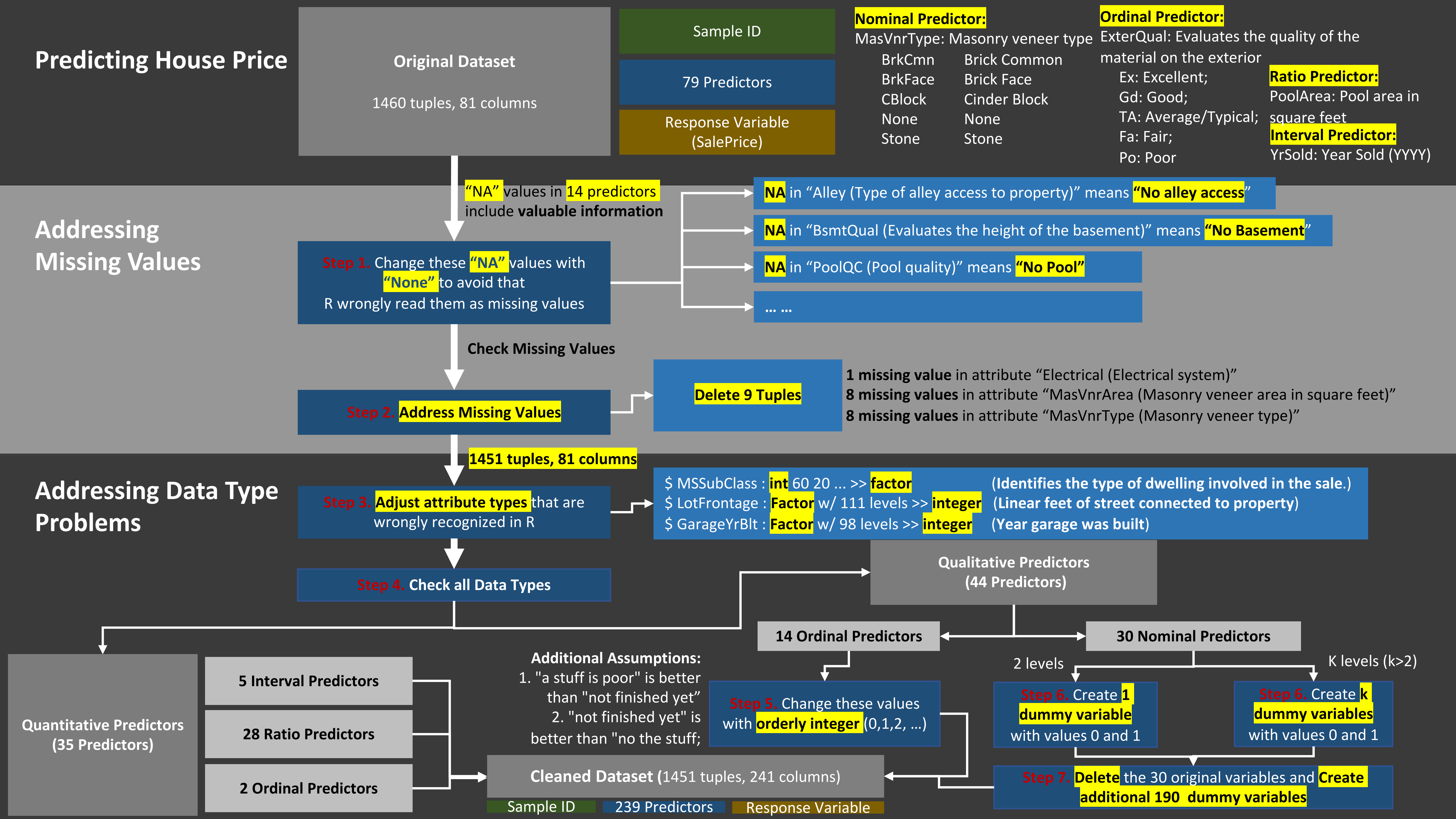
1. Introduction

1. Introduction

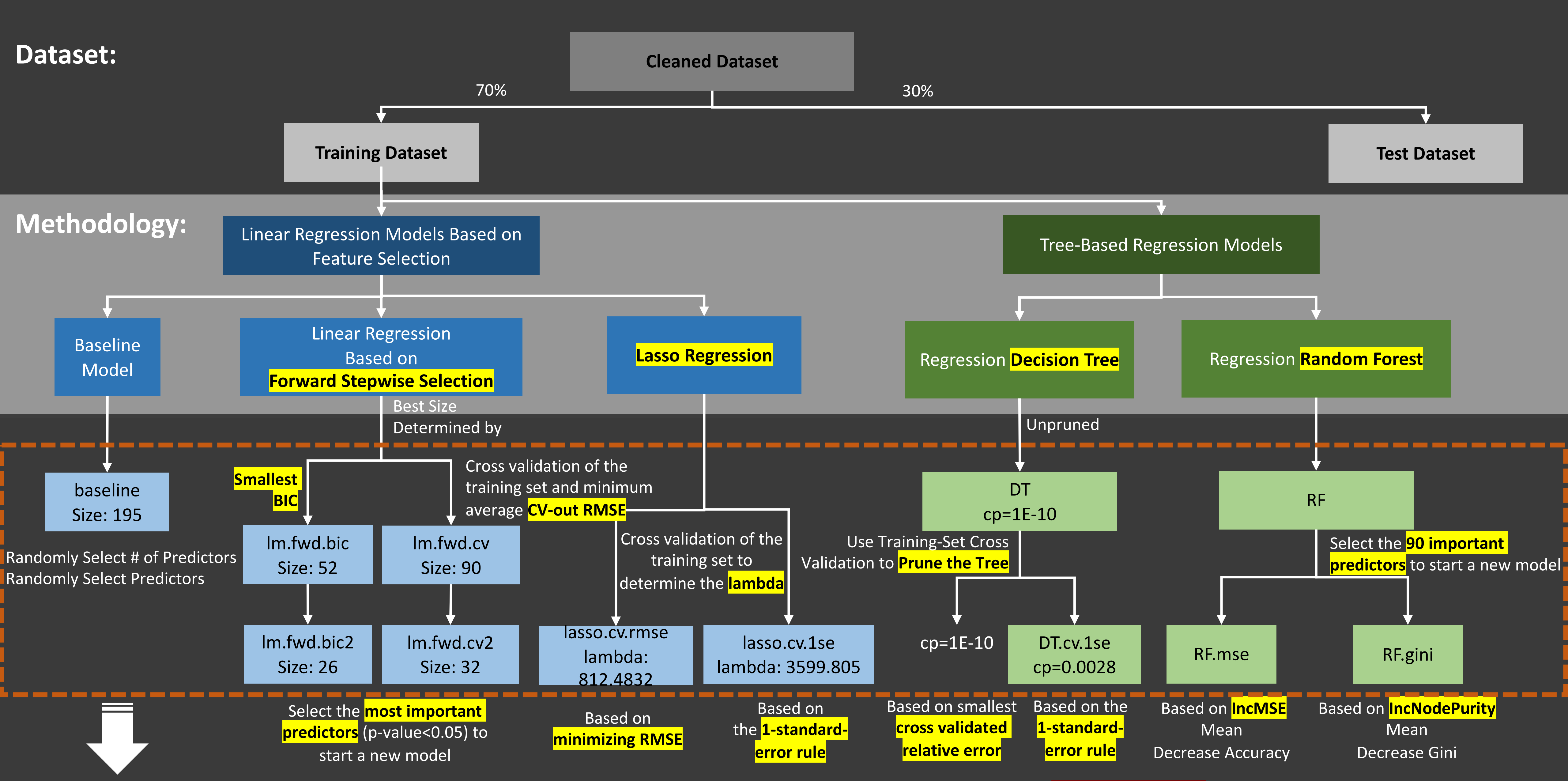


2. Data Preparation

- Dataset Description
- Data Cleaning


Predicting House Price
Original Dataset
1460 tuples, 81 columns
Sample ID
79 Predictors
Response Variable (SalePrice)
Nominal Predictor:
MasVnrType: Masonry veneer type
BrkCmn Brick Common
BrkFace Brick Face
CBlock Cinder Block
None None
Stone Stone
Ordinal Predictor:
ExterQual: Evaluates the quality of the material on the exterior
Ex: Excellent;
Gd: Good;
TA: Average/Typical;
Fa: Fair;
Po: Poor
Ratio Predictor:
PoolArea: Pool area in square feet
Interval Predictor:
YrSold: Year Sold (YYYY)
Addressing Missing Values
"NA" values in 14 predictors include valuable information
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"
… …
Step 1. Change these "NA" values with "None" to avoid that R wrongly read them as missing values
Check Missing Values
Step 2. Address Missing Values
Delete 9 Tuples
1 missing value in attribute "Electrical (Electrical system)"
8 missing values in attribute "MasVnrArea (Masonry veneer area in square feet)"
8 missing values in attribute "MasVnrType (Masonry veneer type)"
1451 tuples, 81 columns
Addressing Data Type Problems
Step 3. Adjust attribute types that are wrongly recognized in R
$ MSSubClass : int 60 20 ... >> factor (Identifies the type of dwelling involved in the sale.)
$ LotFrontage : Factor w/ 111 levels >> integer (Linear feet of street connected to property)
$ GarageYrBlt : Factor w/ 98 levels >> integer (Year garage was built)
Step 4. Check all Data Types
Qualitative Predictors (44 Predictors)
14 Ordinal Predictors
30 Nominal Predictors
2 levels
K levels (k>2)
Quantitative Predictors (35 Predictors)
5 Interval Predictors
28 Ratio Predictors
2 Ordinal Predictors
Additional Assumptions:
1. "a stuff is poor" is better than "not finished yet"
2. "not finished yet" is better than "no the stuff;
Step 5. Change these values with orderly integer (0,1,2, …)
Step 6. Create 1 dummy variable with values 0 and 1
Step 6. Create k dummy variables with values 0 and 1
Step 7. Delete the 30 original variables and Create additional 190 dummy variables
Cleaned Dataset (1451 tuples, 241 columns)
Sample ID
239 Predictors
Response Variable

3. Modeling



Apply the 12 models to both the training-set and the test-set and then compare their RMSE values

RMSE is calculated between <u>**the logarithm of the predicted value**</u> and <u>**the logarithm of the observed sales price**</u> to ensure that errors in predicting expensive houses and cheap houses affect the result equally.

Dataset:

Cleaned Dataset

70% 30%

Training Dataset

Test Dataset

Methodology:

Linear Regression Models Based on Feature Selection

Baseline Model

Linear Regression Based on **Forward Stepwise Selection**

Best Size Determined by

Smallest BIC

Cross validation of training set and mi… average **CV-out RM…**

baseline Size: 195

Randomly Select # of Predictors
Randomly Select Predictors

lm.fwd.bic Size: 52

lm.fwd.cv Size: 90

lm.fwd.bic2 Size: 26

lm.fwd.cv2 Size: 32

lasso… lam… 812…

Forest

…t the **90 important** …**ctors** to start a new model

RF.gini

Select the **most important predictors** (p-value<0.05) to start a new model

Based on **minimizing RMSE**

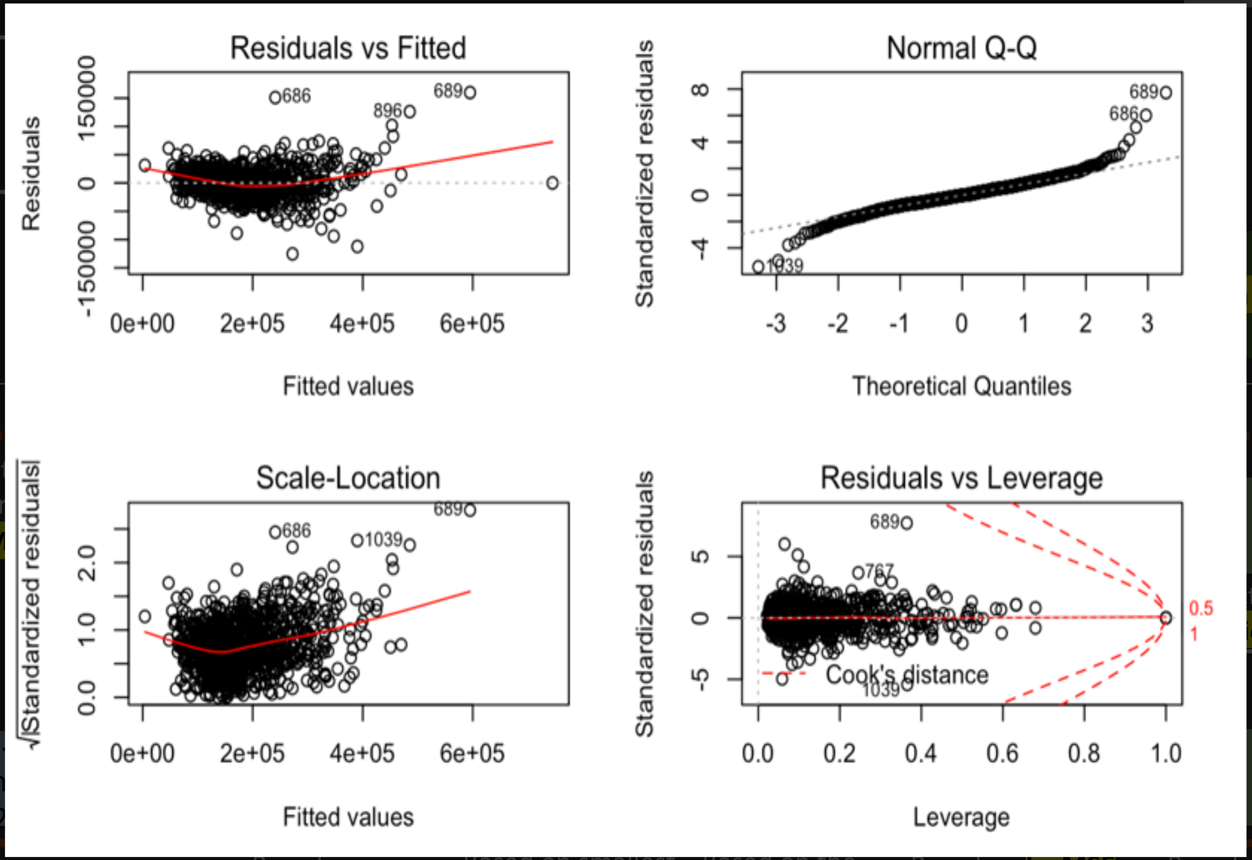
Based on the **1-standard-**…

Based on smallest **cross validated** …

Based on the **1-standard-**…

Based on **IncMSE** Mean …se Accuracy

Based on **IncNodePurity** Mean Decrease Gini



The diagnostic plots for baseline model

Adjusted $R^2 = \mathbf{0.8968}$

Apply the 12 models to both the training-set and the test-set and then compare their RMSE values

RMSE is calculated between the logarithm of the predicted value and the logarithm of the observed sales price to ensure that errors in predicting expensive houses and cheap houses affect the result equally.

Dataset:

Cleaned Dataset

70% — Training Dataset

30% — Test Dataset

Methodology:

Linear Regression Models Based on Feature Selection

Baseline Model

Linear Regression Based on **Forward Stepwise Selection**

Lasso Regres

Best Size Determined by

Smallest BIC

Cross validation of the training set and minimum average **CV-out RMSE**

baseline Size: 195

Randomly Select # of Predictors
Randomly Select Predictors

lm.fwd.bic Size: 52

lm.fwd.cv Size: 90

lm.fwd.bic2 Size: 26

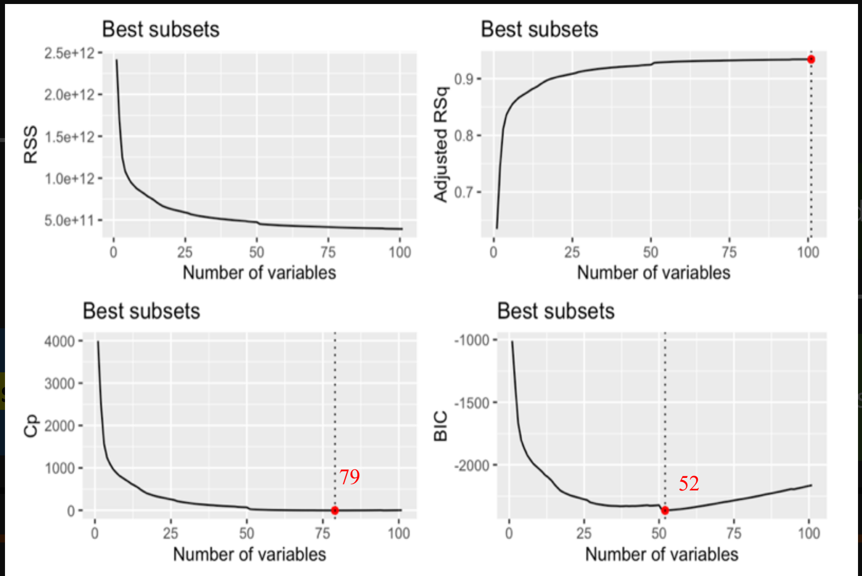
lm.fwd.cv2 Size: 32

Select the **most important predictors** (p-value<0.05) to start a new model

Cross validation training set determine the

lasso.cv.rmse lambda: 812.4832

Based on **minimizing RMSE**



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

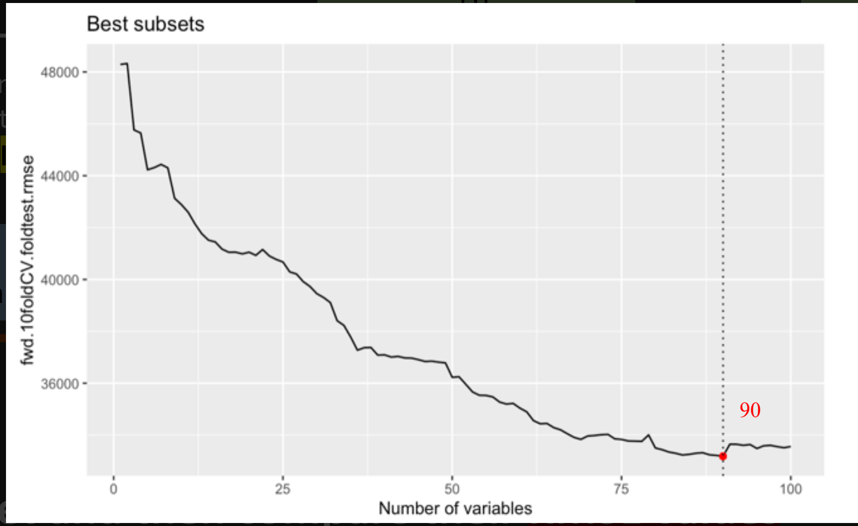
dels

ssion **Random Forest**

DT

RF

Select the **90 important predictors** to start a new model



The plot of average RMSE for fold-out sets for each step

RF.gini

MSE

Based on **IncNodePurity**

Mean Decrease Gini

uracy

Apply the 12 models to both the training-set and the test-se

RMSE is calculated between the logarithm of the predicted value and the logarithm of the observed sales price to ensure that errors in predicting expensive houses and cheap houses affect the result equally.

Dataset:

Cleaned Dataset

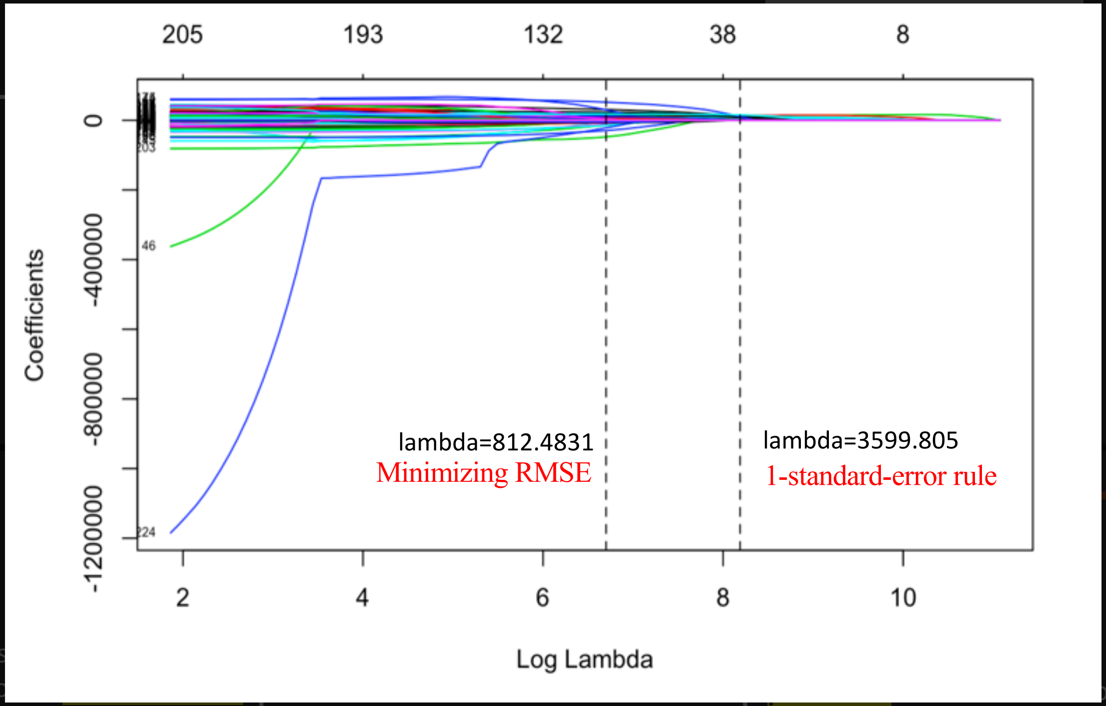
70% — Training Dataset

30%

Methodology:

Linear Regression Models Based on Feature Selection

- Baseline Model → baseline Size: 195
  - Randomly Select # of Predictors
  - Randomly Select Predictors
- Linear Regression Based on **Forward Stepwise Selection** — Best Size Determined by
  - **Smallest BIC** → lm.fwd.bic Size: 52 → lm.fwd.bic2 Size: 26
  - Cross validation of the training set and minimum average **CV-out RMSE** → lm.fwd.cv Size: 90 → lm.fwd.cv2 Size: 32
  - Select the **most important predictors** (p-value<0.05) to start a new model
- **Lasso Regression** — Cross validation of the training set to determine the **lambda**
  - lasso.cv.rmse lambda: 812.4832 — Based on **minimizing RMSE**
  - lasso.cv.1se lambda: 3599.805 — Based on the **1-standard-error rule**



The plot of coefficient paths for lasso regression

cp=1E-10

DT.cv.1se cp=0.0028 — Based on the **1-standard-error rule**

Based on smallest **cross validated relative error**

RF.mse — Based on **IncMSE** Mean Decrease Accuracy

RF.gini — Based on **IncNodePurity** Mean Decrease Gini

Apply the 12 models to both the training-set and the test-set and then compare their RMSE values

RMSE is calculated between the logarithm of the predicted value and the logarithm of the observed sales price to ensure that errors in predicting expensive houses and cheap houses affect the result equally.

Dataset:

Cleaned Dataset

70%

30%

Training Dataset

Test Dataset

Methodology:

Regression Models Based on Feature Selection

Tree-Based Regression Models

Regression Decision Tree

Regression Random Forest

Unpruned

DT
cp=1E-10

RF

Use Training-Set Cross Validation to Prune the Tree

Select the 90 important predictors to start a new model

cv.1se
3599.805

cp=1E-10

DT.cv.1se
cp=0.0028

RF.mse

RF.gini

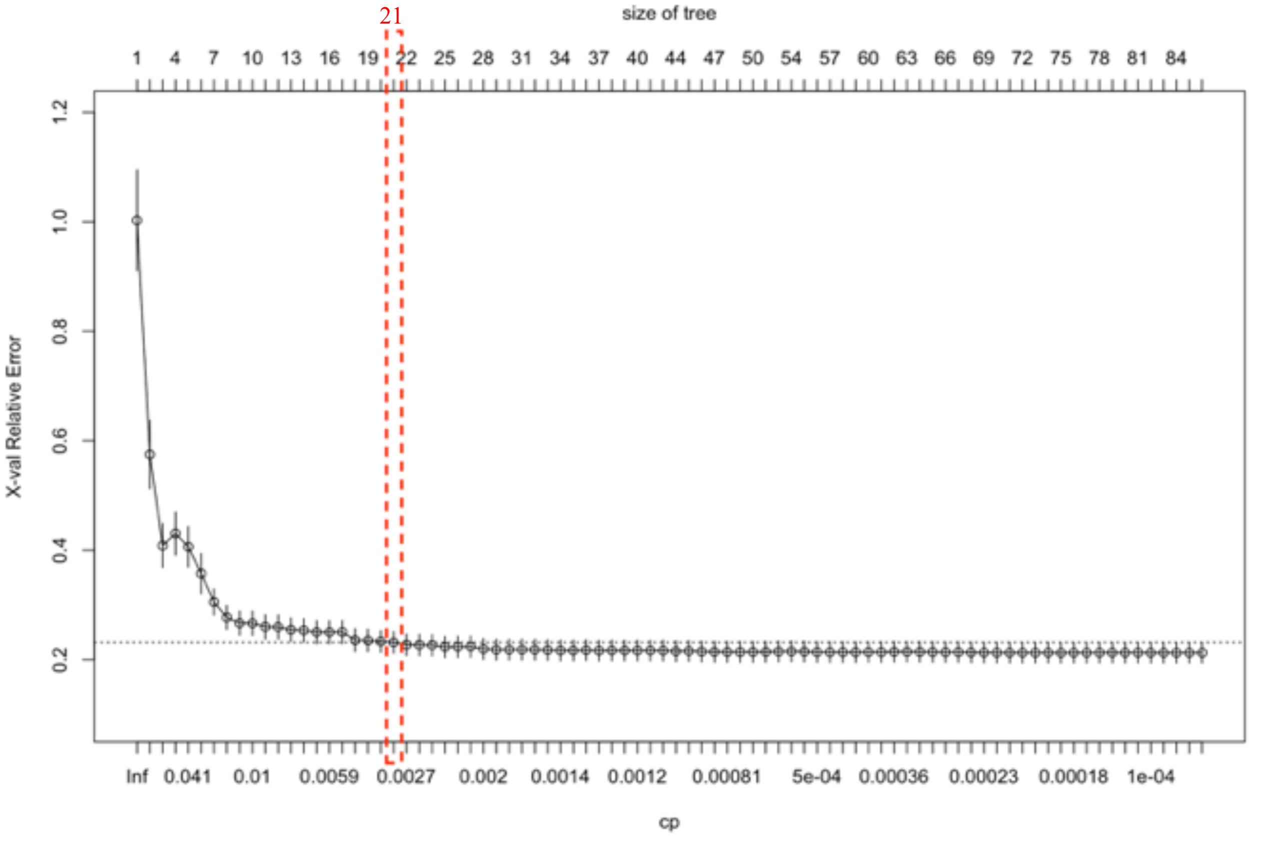
d on andard- rule

Based on smallest cross validated relative error

Based on the 1-standard-error rule

Based on IncMSE Mean Decrease Accuracy

Based on IncNodePurity Mean Decrease Gini



Apply the 12 models to both the training-set and the test-set and then compare their RMSE values

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

RMSE is calculated between the logarithm of the predicted value and the logarithm of the observed sales price to ensure that errors in predicting expensive houses and cheap houses affect the result equally.

Dataset:

Cleaned Dataset

70% 30%

Training Dataset

Test Dataset

Methodology:

Tree-Based Regression Models

Regression Decision Tree

Regression Random Forest

Unpruned

DT
cp=1E-10

RF

Select the 90 important predictors to start a new model

RF.mse

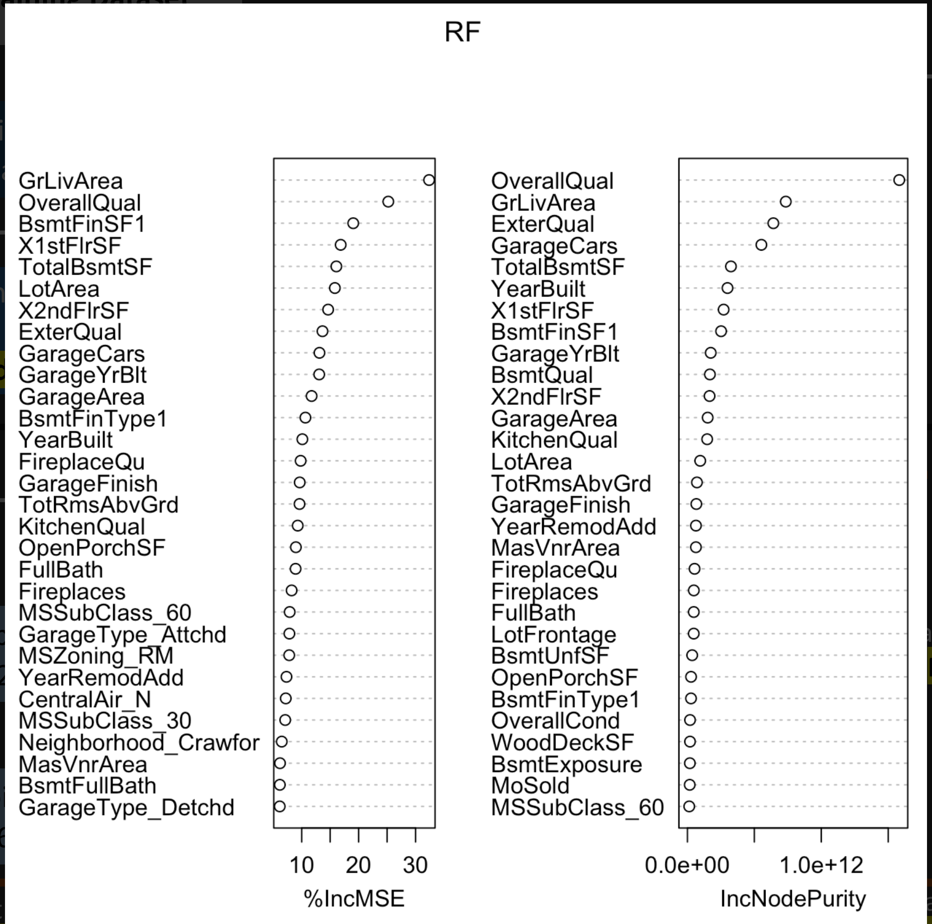
RF.gini

Based on IncMSE
Mean
Decrease Accuracy

Based on IncNodePurity
Mean
Decrease Gini

baseline
Size: 195

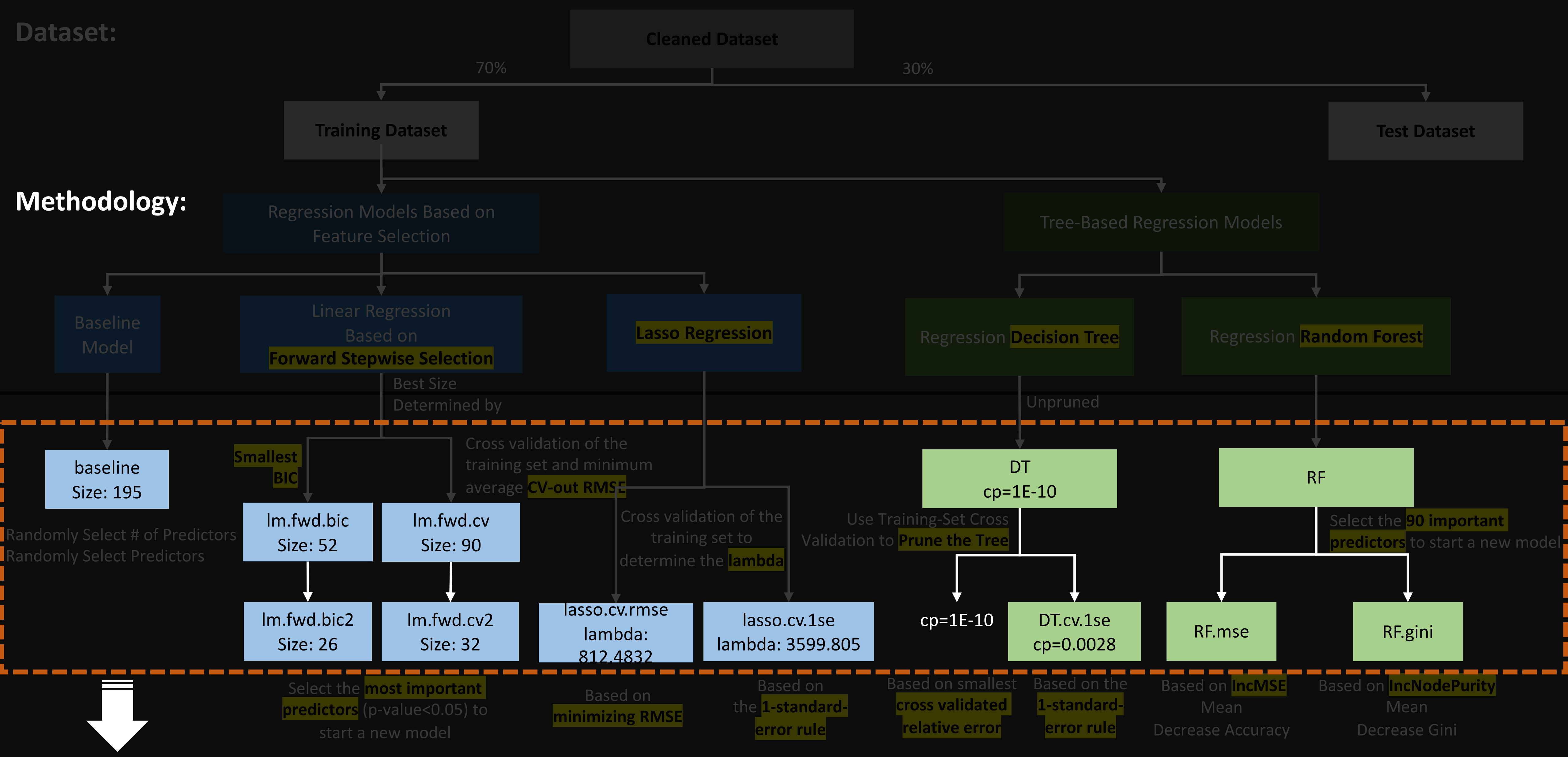
Randomly Select # of Predictors
Randomly Select Predictors



Dotchart of variable importance for model RF

Apply the 12 models to both the training-set and the test-set and then compare their RMSE values

RMSE is calculated between the logarithm of the predicted value and the logarithm of the observed sales price to ensure that errors in predicting expensive houses and cheap houses affect the result equally.



Apply the 12 models to both the training-set and the test-set and then compare their RMSE values

RMSE is calculated between **the logarithm of the predicted value** and **the logarithm of the observed sales price** to ensure that errors in predicting expensive houses and cheap houses affect the result equally.

4. Validation

4. Validation

	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	0.14522	0.0627	30/239 candidates	-	Use **RF.gini** to fit the entire dataset	-

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

4. Validation

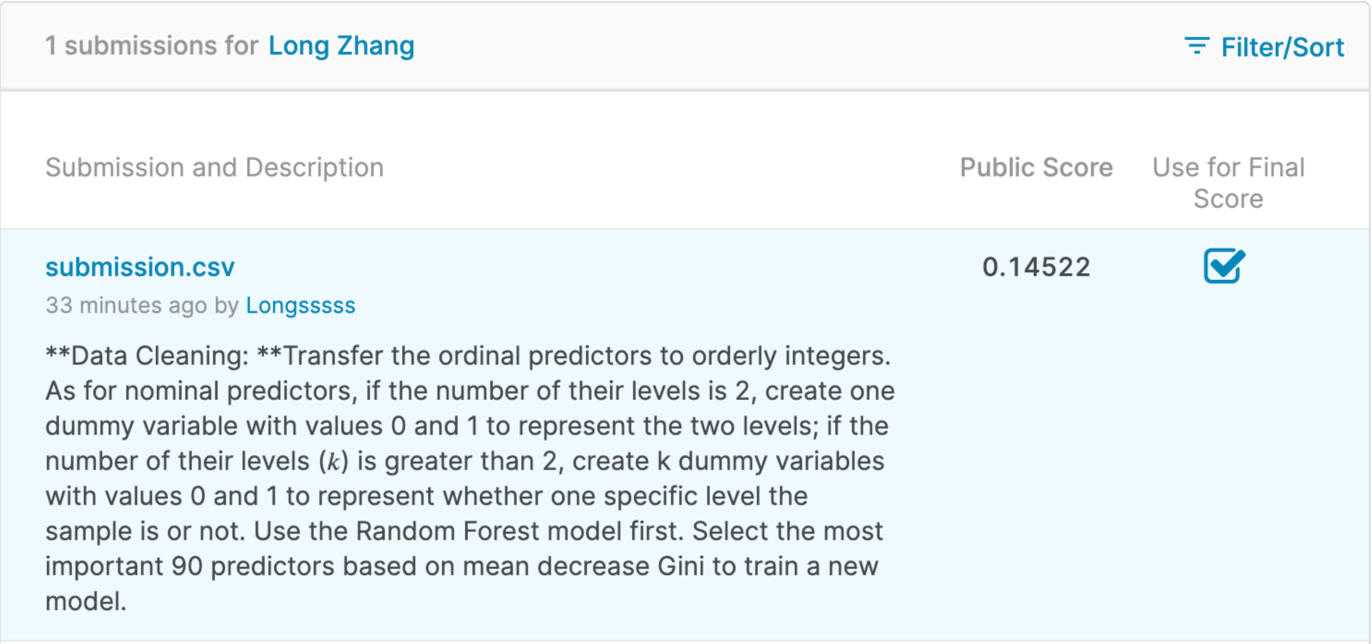
	Names	TestSet RMSE	TrainingSet RMSE	Number of Predictors	Parameters	Methods	Property
Regression Models Based on Feature Selection	baseline	0.2918	0.1547	154/239 selected	predictors	RandomSelected	adj.r2=0.8968
	lm.fwd.bic	0.2132	0.1974	52/239 selected	size=52	BIC.min	adj.r2=0.8524
	lm.fwd.bic2	0.204	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	0.14522	0.0627	30/239 candidates	-	Use **RF.gini** to fit the entire dataset	-

***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

4. Validation

	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	0.20				
	lm.fwd.cv	0.1941	0.16				
	lm.fwd.cv2	0.1909	0.16				
	lasso.cv.rmse	0.1859	0.12				
	lasso.cv.1se	0.18	0.15				
Tree-Based Models	DT	0.1999	0.11				
	DT.cv.1se	0.2164	0.17				
	RF	0.1561	0.06				
	RF.gini	$\boxed{0.1551}$	0.06				
	RF.mse	0.1554	0.06				
The Final Model	Final.Model	0.14522	0.06				



***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

4. Validation

	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**

5. Conclusion

5. Conclusion

- This project established multiple kinds of **feature-selection-based linear regression models** and **tree-based regression models** to predict house price.

- Ultimately, we built the model through fitting the entire dataset on the best model **RF.gini**. Therefore, we can use it to predict unobserved instancess 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.

THANK YOU

References:
[1] 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.
[2] kaggle, "House Prices: Advanced Regression Techniques," Kaggle, [Online]. Available: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview. [Accessed 11 May 2020].