Introduction

Housing prices database was chosen as a starting point for the project. You may find detailed description of the data at Amstat website[1]. Dataset contains housing data for Aimes municipality in Iowa. It includes 79 explanatory variables, of which 23 are nominal, 23 ordinal, 14 discrete and 20 continuous. Data was collected between 2006 and 2010. The dataset will be used to build a prediction model on the housing price.

Our objective is to develop a model with the high predictive power to forecast the housing prices to be used by the real estate agents in pricing and negotiations. Our criteria used for the selection of the model are MSE and MAE scores. Obviously, we want to minimize those.

Our approach to the model development and selection is centered around 5 steps. We appreciate the fact that quality of our recommendation depends on the quality of the dataset that we use. Thus, we invested time in the understanding of our data, cleaning the dataset and selecting the correct predictors. To select the set of predictors we used both exploratory data analysis and random forest importance scores.

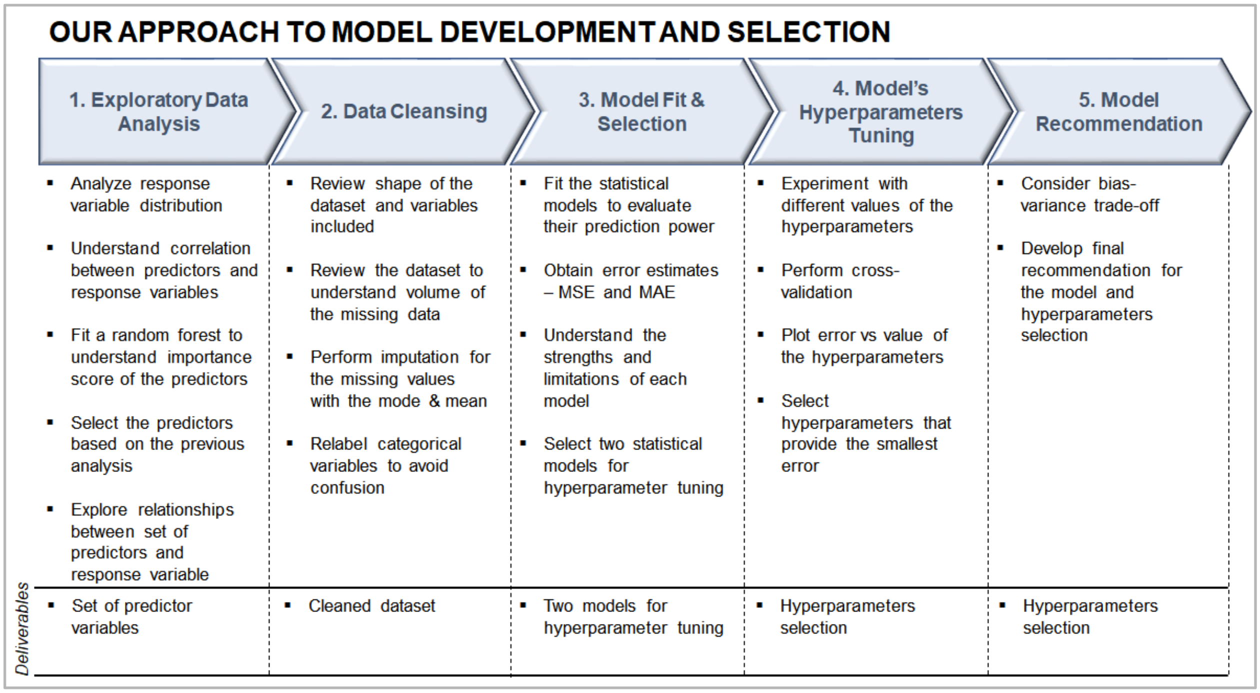
After completing each step in our analysis we recorded a deliverable attributable to the particular step. Detailed description of our analysis and outcomes is presented in the main body of the report.

Background and Motivation

Housing dataset is ideal for perfecting machine learning techniques, especially regression analysis. Our goal is to develop the model with strong performance on the house price prediction. However, the goal runs into several practical issues. The dataset obtained contains significant number of variables. Relying on such extensive list of features is expensive. Going forward it will be increasingly difficult to maintain integrity of the data. The other problem is that our dataset contains large number of subjective variables, such as garage quality or house quality. We want minimize human factor/error in the house pricing. Therefore we will evaluate importance of the predictors. Our objective is to select 10 to 15 predictors, which are technically feasible to collect and maintain in the long-run. They should not display large inter-item

[1] https://ww2.amstat.org/publications/jse/v19n3/decock.pdf

covariance and be a good predictor of the sales price, meaning high correlation with it.



OUR APPROACH TO MODEL DEVELOPMENT AND SELECTION

	1. Exploratory Data Analysis	2. Data Cleansing	3. Model Fit & Selection	4. Model's Hyperparameters Tuning	5. Model Recommendation
	▪ Analyze response variable distribution ▪ Understand correlation between predictors and response variables ▪ Fit a random forest to understand importance score of the predictors ▪ Select the predictors based on the previous analysis ▪ Explore relationships between set of predictors and response variable	▪ Review shape of the dataset and variables included ▪ Review the dataset to understand volume of the missing data ▪ Perform imputation for the missing values with the mode & mean ▪ Relabel categorical variables to avoid confusion	▪ Fit the statistical models to evaluate their prediction power ▪ Obtain error estimates – MSE and MAE ▪ Understand the strengths and limitations of each model ▪ Select two statistical models for hyperparameter tuning	▪ Experiment with different values of the hyperparameters ▪ Perform cross-validation ▪ Plot error vs value of the hyperparameters ▪ Select hyperparameters that provide the smallest error	▪ Consider bias-variance trade-off ▪ Develop final recommendation for the model and hyperparameters selection
Deliverables	▪ Set of predictor variables	▪ Cleaned dataset	▪ Two models for hyperparameter tuning	▪ Hyperparameters selection	▪ Hyperparameters selection

Our predictors can be grouped into 5 categories: real estate, access, lot, house, garage and amenities. Real estate data contains general information on the type of the dwelling - 1, 2 or 3 stories house and year of construction, as well as zoning classification. Access describes the transportation accessibility of the property, access to the road, proximity to highway and major landmarks in the city of Aimes. Lot contains variables on the size of the property's land plot, it's configuration and pavement. House is the largest group with all the variables, describing the actual house, from the building type, year of construction to external finishing materials quality and condition. This group also contains information on the total house area, basement, first and second floors area and number of rooms. Garage contains information on the size and volume, it's quality and condition. Finally, amenities contain all the nice details that distinguish one similar sized property from the other, such as quality of the kitchen finishings, swimming pool, number of bathrooms and fireplaces. This category also employs practical items, for instance type of the heating and electrical systems, access to utilities. Detailed list of all the variables and their descriptions is provided in the appendix section.

Our motivation is to review and apply techniques learnt throughout BAIT 509 Machine Learning class to build predictive model for the housing price. We foremost treat this exercise as a learning experience that gives us the opportunity to collaborate and share knowledge within the team. Secondly, we seek to answer practical statistical question that brings us closer to the real-world application of machine learning techniques.

A discussion about questions

Real estate agents and mutual funds often grapple with the question of pricing the asset - real estate. The market often presents a mismatch between sellers and buyers expectations, leading to the distortions and slow moving properties. Our business objective is to build a housing price predictive model that would provide accurate estimates of the real estate properties value based on the historical housing price data. We recognize that this model would be applied only to houses in municipality of Aimes in Iowa, where the original data set was collected.

Statistical question that we seek to answer - what is the predicted price of the house in Aimes with the given characteristics of the access, lot, house, garage and amenities? Answering this statistical question will provide business insight into the correct market price of the house on sale. Correctly predicting the sales will allow to identify undervalued properties on the market to be taken advantage of. On the other hand, such model will allow fair estimation to minimize market distortion and guide expectations of the players in the market.

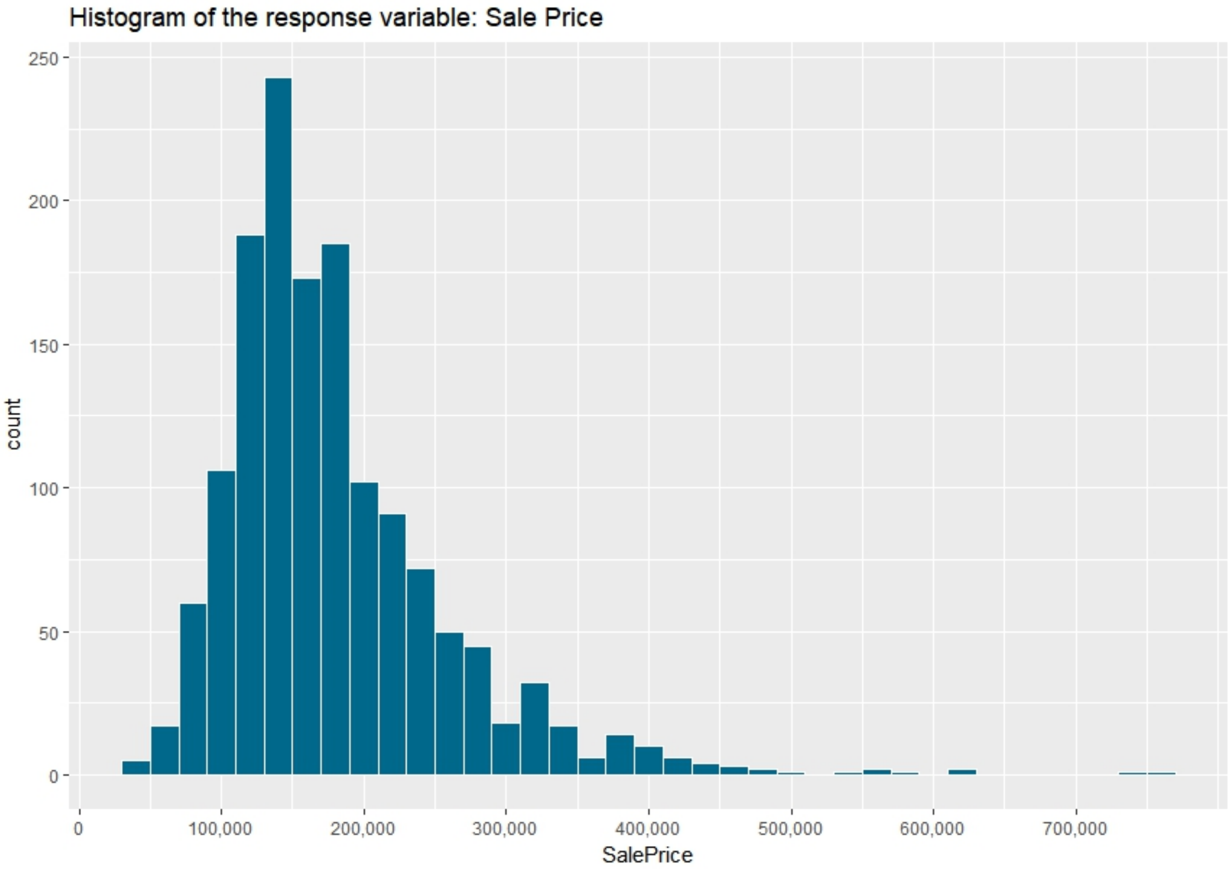
Yet, it’s important to understand the drawbacks of the model. Properties may possess unique characteristics that increase the price that are not reflected in the model. Model doesn’t account for the sudden change in the market, because relies on the historical data for the prediction. Finally, confidence band around point estimate is the best prediction we can do - also recognized as irreducible error. In other words, the model won’t give the one and only true prediction. It’s also important to recognize the reducible error. Our choice of the predictors and models may be suboptimal, thus the variance and the mean of the predicted sales price may be incorrect. To minimize reducible error model should be extensively validated on the test set.

A discussion about the model(s) & model selection

1. Exploratory data analysis

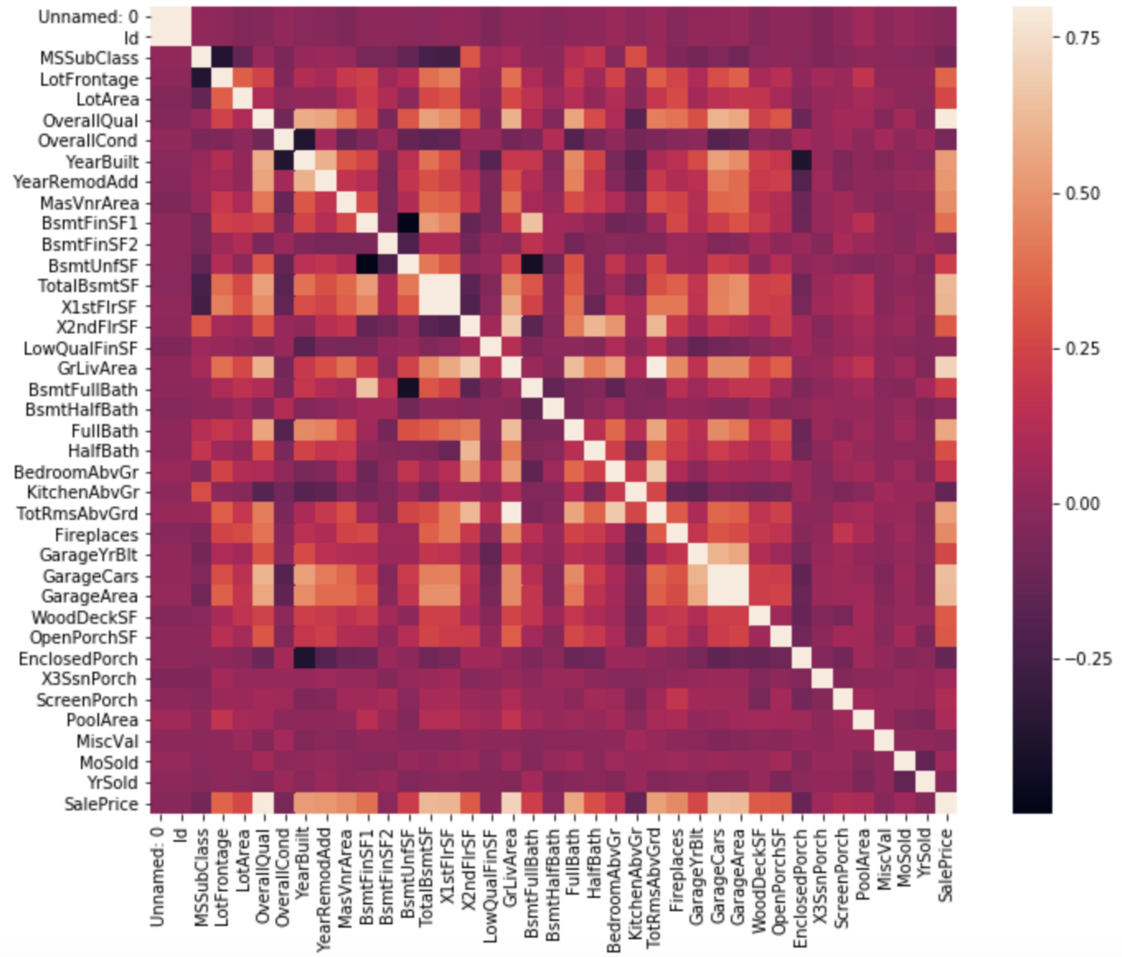
1.1 Response variable

Sale price is the response variable. Based on the histogram sale price values are right skewed. This can be explained by the fact that majority of the expensive houses are less in demand with the public. Summary statistics indicate that minimum price for a house is 34900, maximum is 755000, mean is at 180843 and median is at 16300.



1.2 Correlations between response and predictor response variables

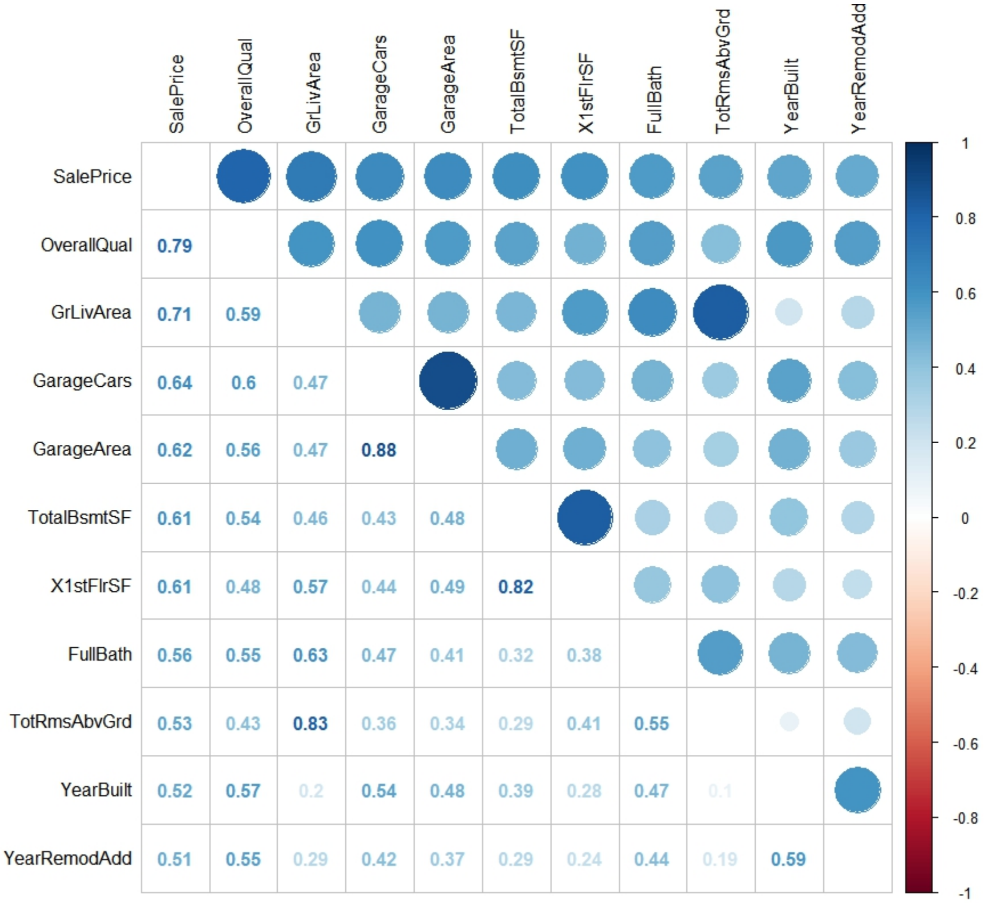
Next, we examine the associations between numeric variables to identify the best predictors of the sale price. We can take advantage of the correlation matrix to select predictors that are highly correlated with SalePrice and check intercorrelation between the predictors.



[2]

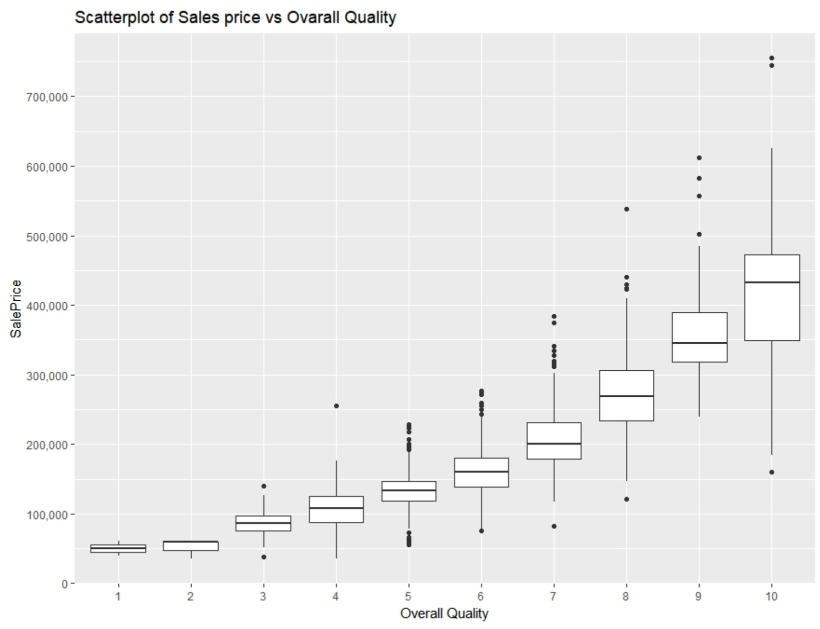
For further analysis we employ only variables that show correlation with the SalePrice above 0.5. It’s important to note that we would have to further examine the relationship between selected predictors and response variables to ensure their relationship is approximately linear, otherwise the value of the correlation coefficient is meaningless.

[2] https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

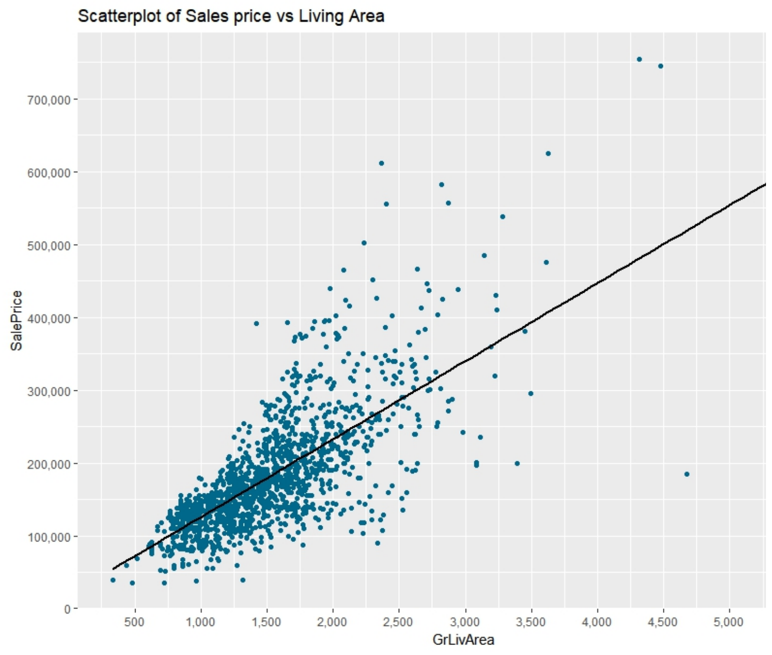


Examining pairwise correlations, we notice that there are 10 predictors that have relatively high correlation with SalePrice, therefore are candidates for the inclusion into the model. However, GarageCars & GarageArea have high intra-correlation, as well as GrLivArea and TotRmsAbvGrd; X1stFlrSF and TotalBsmSF; YearBuilt and YearRemodAdd. For further exploration we will use only one of the variables that exhibit high intra-correlation. We will explore relationship between SalePrice and OverallQual, GrLivArea, GarageCars, X1stFlrSF, FullBath and YearBuilt.

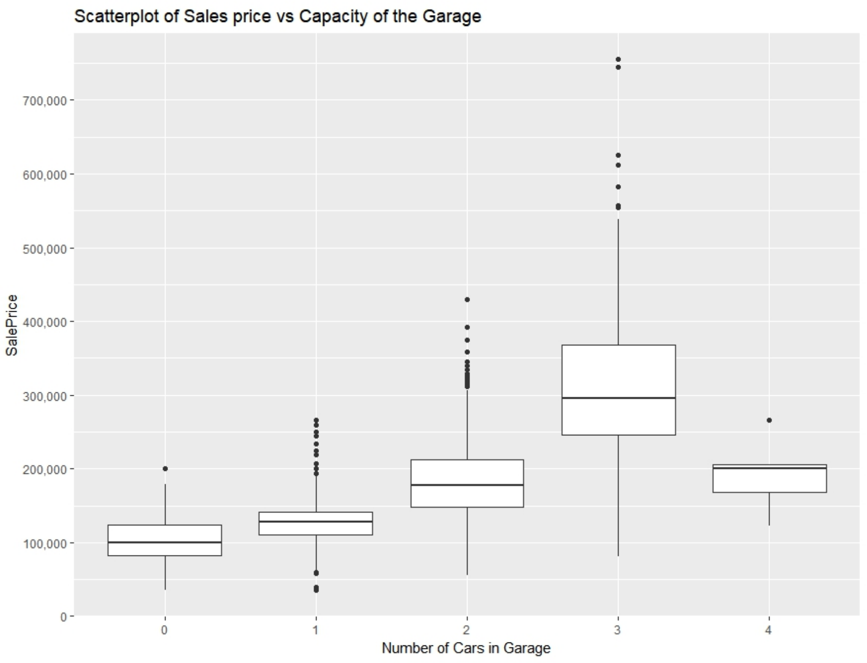
1.3 Exploring relationships between response and predictor variables



- ➔ Overall quality takes on the values from 1 to 10, correlation with SalePrice is 0.79
- ➔ Positive correlation between overall quality and sales price should be noted
- ➔ Relationship is best described by the quadratic function due to the upward curve at the tail



- ➔ Living area is a continuous variable, correlation with SalePrice is 0.71
- ➔ Positive linear association between living area and sales price should be noted
- ➔ Relationship is strong, but variance increases with the living area
- ➔ Few outliers beyond GrLivArea > 3500 and SalePrice > 600,000

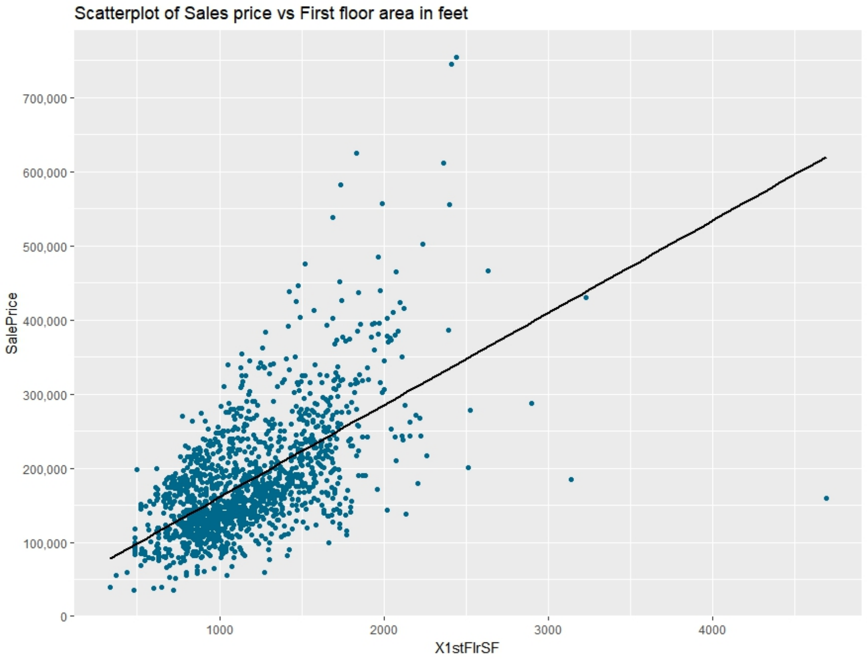


➔ Capacity of garage ranges from 0 to 4 with the most of observations in 0 & 1

➔ Positive correlation with SalePrice is attributable to garages with 0 to 3 cars capacity

➔ Garages for 4 cars are underpriced by the market

➔ Correlation with SalePrice is 0.64 houses with 4 cars

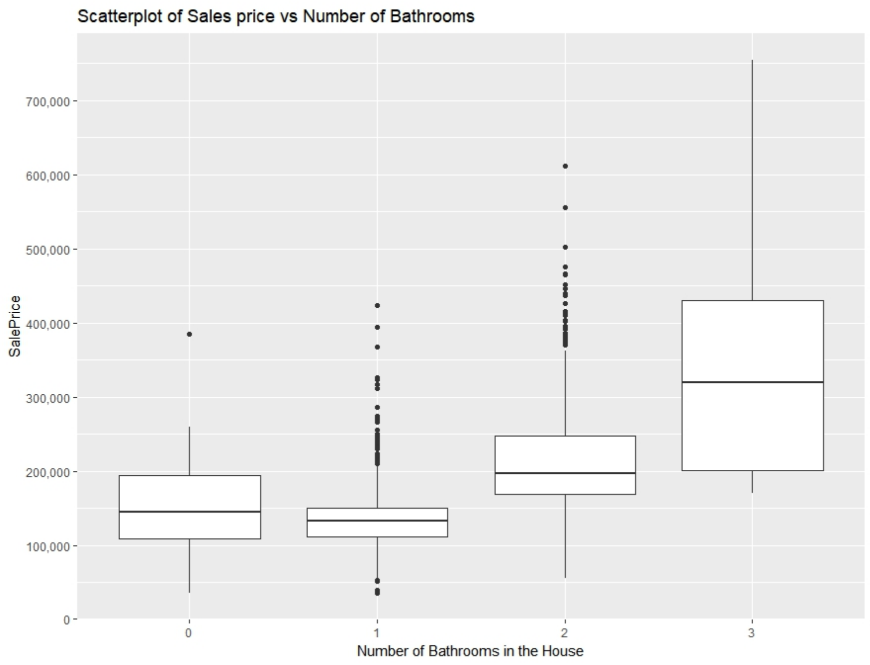


➔ First floor area is shows strong positive linear association with SalePrice

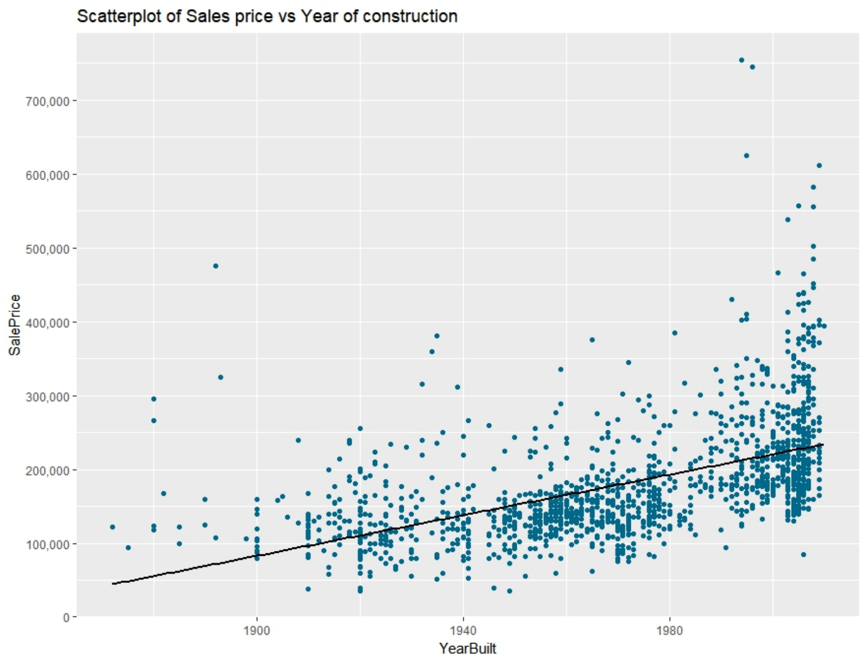
➔ Correlation is 0.61

➔ Beyond 2000 square feet there is large variance. Outliers are beyond 2500 feet and $650 thousand

➔ We must proceed with caution when interpreting results with X1stFlrSF due to large variation



- ➔ Number of bathrooms has values ranging from 0 to 3
- ➔ Houses with 0 bathrooms are outliers, possibly they are not residential properties
- ➔ Houses with 1, 2 and 3 bathrooms demonstrate positive correlation with Sales price

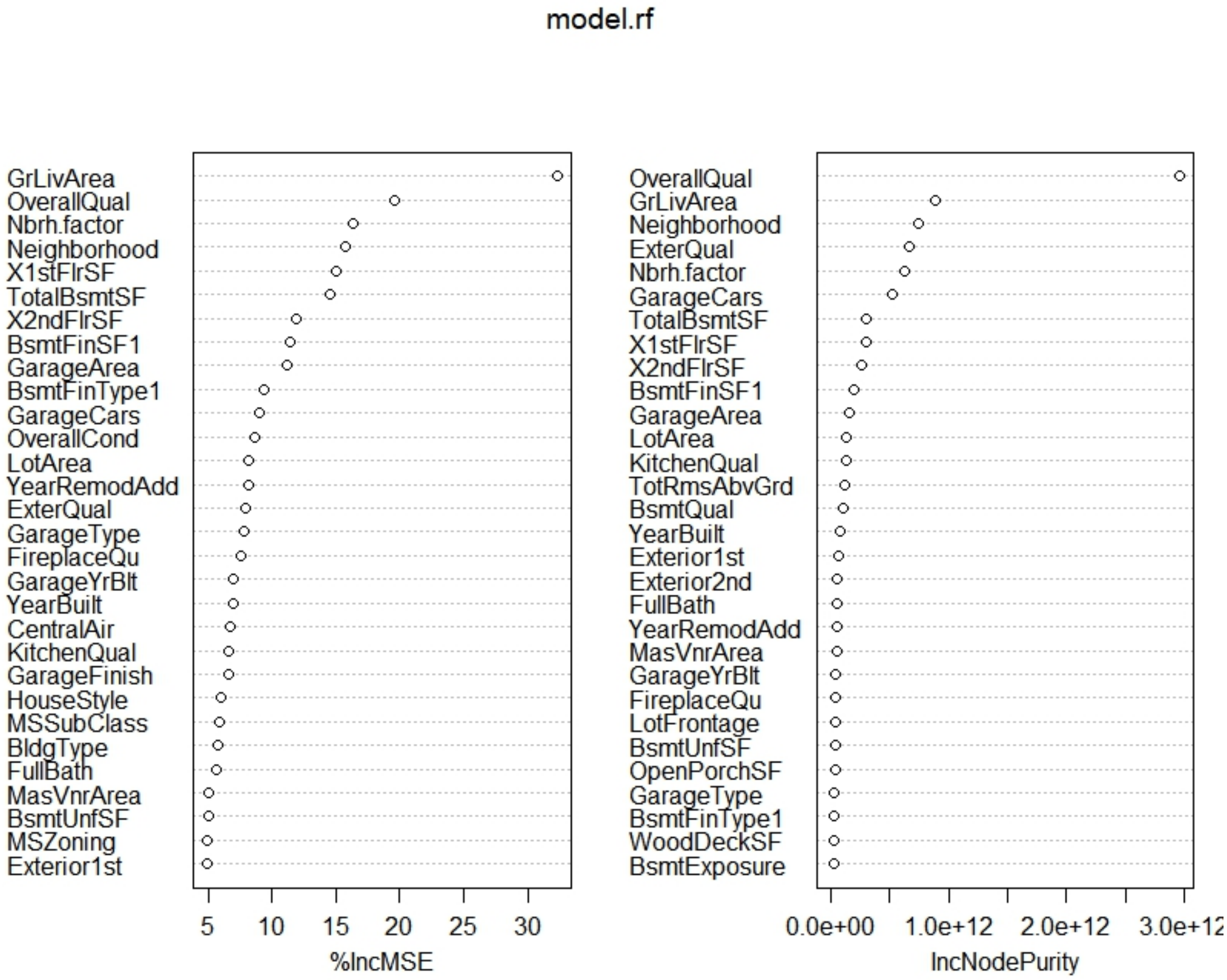


- ➔ Relationship between YearBuilt and Sales price is approximately linear
- ➔ It’s easily noticeable that higher priced properties pull the regression line upwards
- ➔ Meaning there is significant variance within the prices of houses built in different years
- ➔ We must use this variable with caution

YearBuilt, FullBath and X1stFlrSF have relationships with SalePrice that can’t be well described by the linear function. At this point in our analysis we don’t apply any transformations to our predictor variables, rather we cautiously proceed with the data analysis, taking into the account drawback of the variables.

1.4 Random Forest importance score

We employ random forest to extract importance scores for our predictors. This will ensure that we don’t forget important categorical variables in the model. And for the quantitative variables additional information obtained will ensure that we stay on the right path. Below is the output from random forest importance scores.



It’s easily noticeable, there are similarities between IncNodePurity and the intra-correlation analysis pertaining to the selection of the predictor variables. For instance, it’s clear that the OverallQual is the most influential predictor both on random forest importance and the correlation coefficient. Yet, there are additional predictors that were not accounted by the correlation analysis: Neighborhood, ExterQual, BsmtFinSF1, KitchenQual, X2ndFlrSF, LotArea.

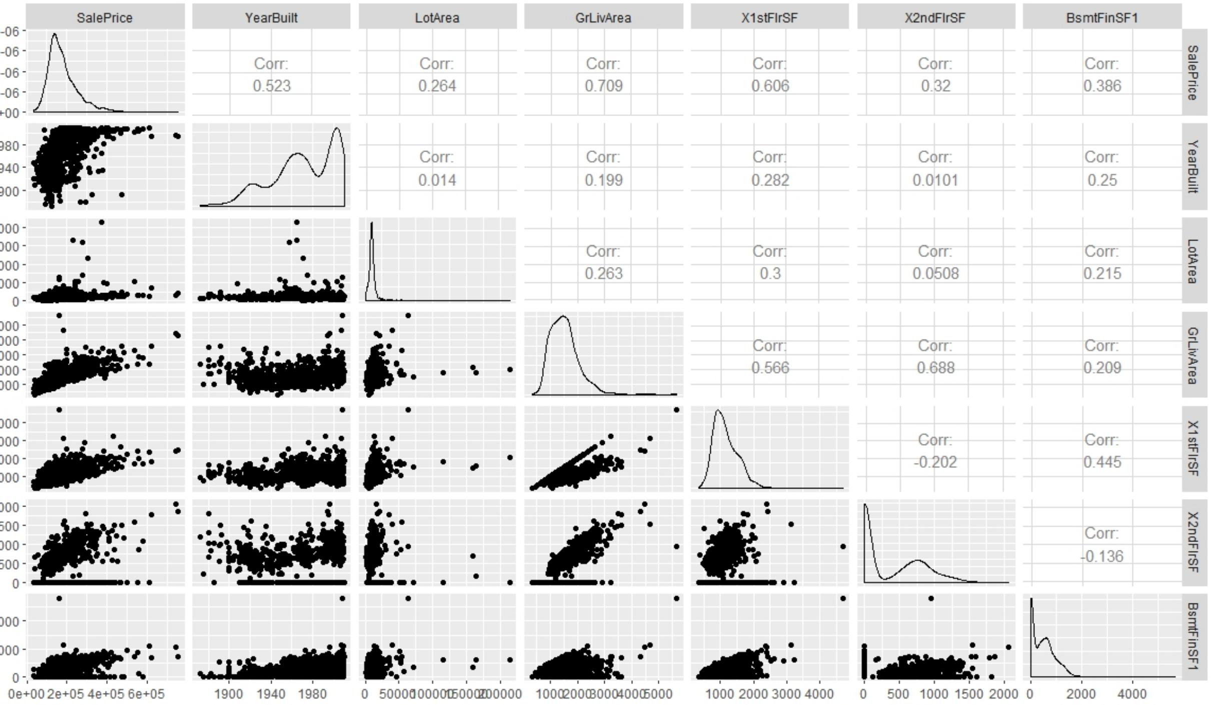
The first 4 are categorical variables and the last 2 are quantitative. We would like to incorporate those into our analysis. Since we use tree based regression methods for our model development, it will be beneficial to employ those variables in the analysis.

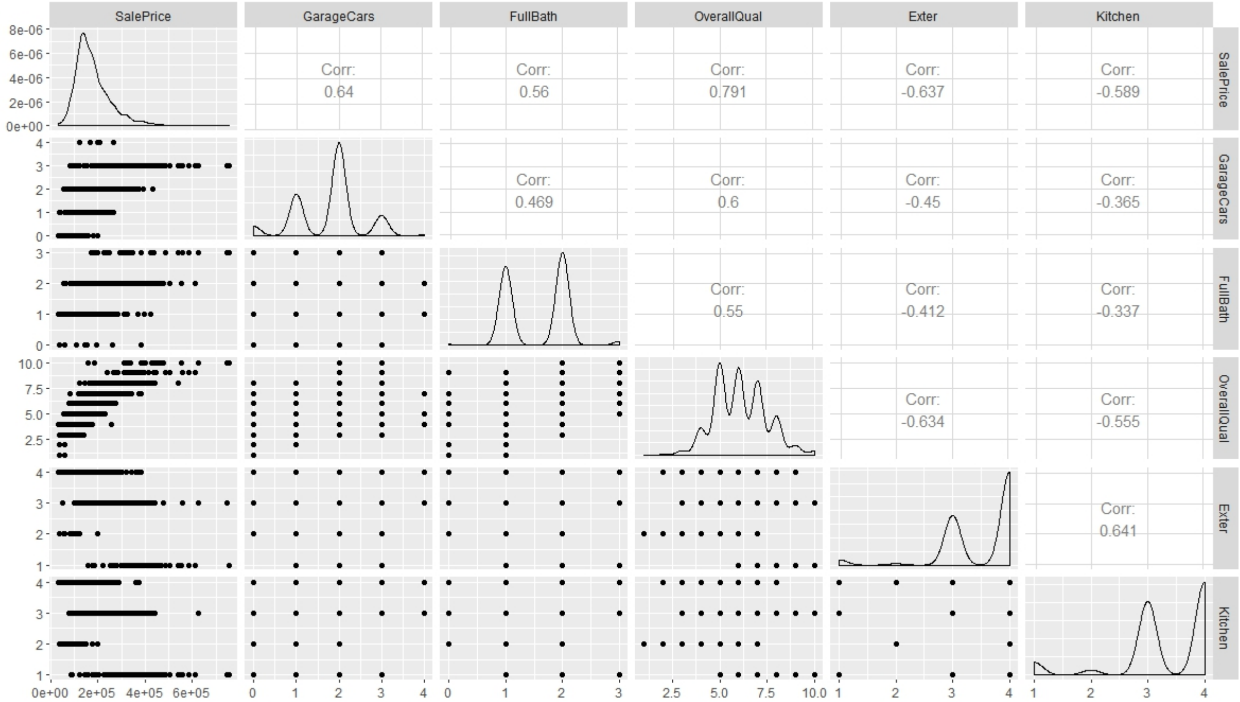
1.5 Our Choice of the Predictors

We select the following variables for the model development:

1. *Neighborhood* - 25 neighborhoods in Aimes;
2. *YearBuilt* - year of the house construction;
3. *LotArea* - area of the land plot in square feet;
4. *OverallQual* - overall quality of the building from very poor (1) to excellent (10);
5. *KitchenQual* - quality of the kitchen from poor (1) to excellent(5);
6. *GrLivArea* - above grade living area in feet;
7. *GarageCars* - size of the garage in car capacity;
8. *ExterQual* - quality of the external materials from poor (1) to excellent (5);
9. *X1stFlrSF* - first floor square feet;
10. *X2ndFlrSF* - second floor square feet;
11. *BsmtFinSF1* - finished basement in square feet;
12. *FullBath* - number of full baths in the house.

Below we plot the final choice of our predictor and response variables:





2. Data cleansing and wrangling

The raw dataset contains the training dataset and the test dataset. After discovering the whole dataset, we found that the test set did not contain the response variable "SalePrice". So in order to build supervised machine learning model, we will use the training dataset to analyze and build models based on it.

After a quick review of the training dataset, we found that there are 81 variables including the response variable. And 19 of total variables have missing values. The variables are shown below:

PoolQC	*MiscFeature*	*Alley*	*Fence*	*FireplaceQu*	*LotFrontage*	*GarageType*
1453	1406	1369	1179	690	259	81
GarageYrBlt	*GarageFinish*	*GarageQual*	*GarageCond*	*BsmtExposure*	*BsmtFinType2*	*BsmtQual*
81	81	81	81	38	38	37
BsmtCond	*BsmtFinType1*	*MasVnrType*	*MasVnrArea*	*Electrical*		
37	37	8	8	1		

PoolQC:

The PoolQC is the variable with most NAs. The description is as follows:

The house's Pool Quality:

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
NA	No pool

So, it is obvious that we need to just assign 'No Pool' to the NAs. Also, the high number of NAs makes sense as normally only a small proportion of houses have a pool.

MiscFeature:

Within Miscellaneous Feature, there are 1406 NAs. The description is as follows:

Values:

Elev	Elevator
Gar2	2nd Garage (if not described in
Othr	Other
Shed	Shed (over 100 SF)
TenC	Tennis Court
NA	None

So, it is obvious that we need to just assign 'No Pool' to the NAs.

Alley:

Within Alley, there are 1369 NAs. The description is as follows:

Grvl	Gravel
Pave	Paved
NA	No alley access

So, it is obvious that we need to just assign ‘No alley access’ to the NAs.

Fence:

Within Fence, there are 1179 NAs. The values seem to be ordinal.

Values:

GdPrv	Good Privacy
MnPrv	Minimun Privacy
GdWo	Good Wood
MnWm	Minimun Wood/Wire
NA	No fence

So, it is obvious that we need to just assign ‘No fence’ to the NAs.

FireplaceQu:

The number of NAs in FireplaceQu matches the number of houses with 0 fireplaces. This means that I can safely replace the NAs in FireplaceQu with ‘no fireplace’. The values are ordinal, and I can use the Qualities vector that I have already created for the Pool Quality.

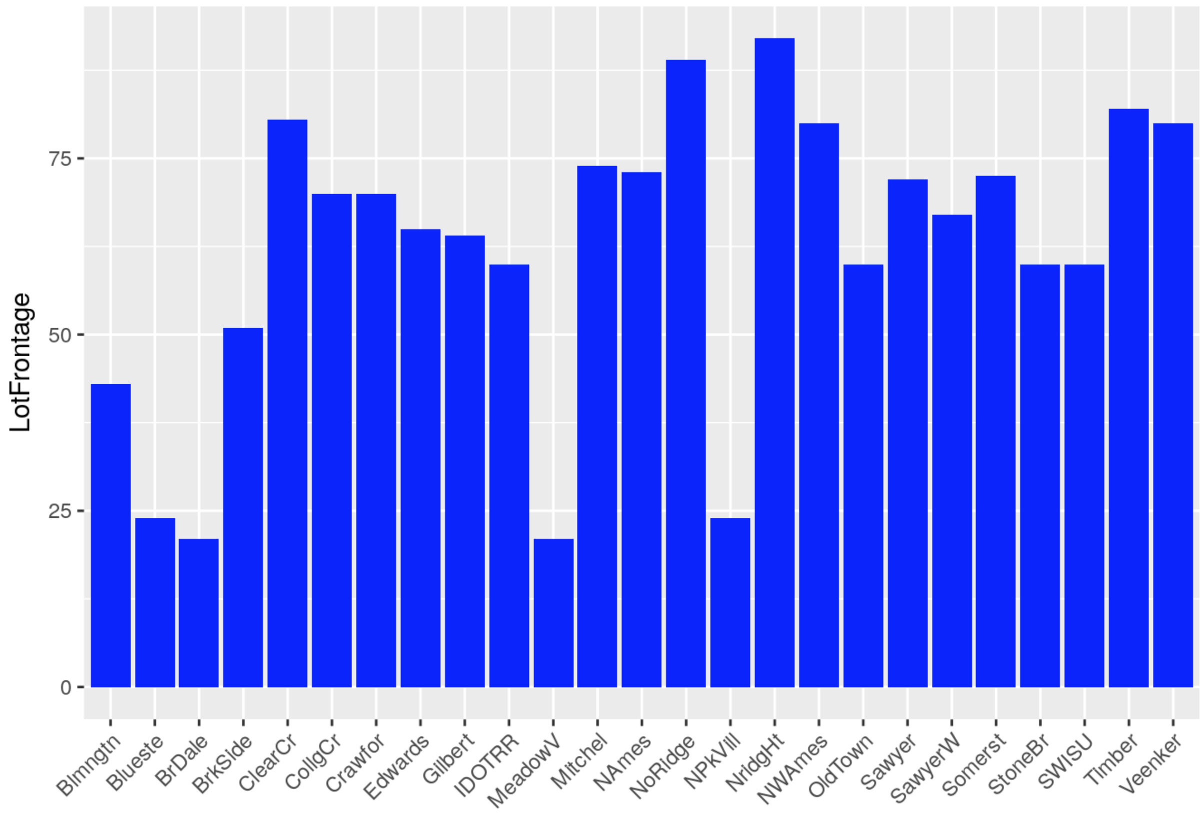
Values:

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

So, it is obvious that we need to just assign ‘No fireplace’ to the NAs.

LotFrontage:

This variable contains 259 NAs. The most reasonable imputation seems to take the median per neighborhood since every neighborhood has similar lot frontage length. The mean of lot frontage is shown below per neighborhood:



GarageType & GarageYrBlt & GarageFinish & GarageQual & GarageCond:

These five variable should be dealt together because they are all related to the garage and should be consistent with each other.

The NA in these five variable means that there is no garage in the house.

Also, we found that the number of missing value in these five variable is the same. So we can replace all missing value in GarageType & GarageFinish & GarageQual & GarageCond with "No Garage" and replace all missing value in GarageYrBlt with 0.

BsmtExposure & BsmtFinType2 & BsmtQual & BsmtCond & BsmtFinType1:

These five variable should be dealt together because they are all related to the basement and should be consistent with each other.

The NA in these five variable means that there is no basement in the house.

But the missing number in the five variables are not the same.

BsmtExposure	BsmtFinType2	BsmtQual	BsmtCond	BsmtFinType1
38	38	37	37	37

So there might be something wrong in few observation about the basement variables.
After exploring the the dataset a bit, we found two observations that are problematic.

	BsmtQual	BsmtCond	BsmtExposure	BsmtFinType1	BsmtFinSF1	BsmtFinType2	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF	BsmtFullBath	BsmtHalfBath
949	Gd	TA	<NA>	Unf	0	Unf	0	936	936	0	0

The BsmtExposure in No.949 observation indicates that the house did not have basement. But according to other variable value, it indicates that the house had a basement. So it is not consistent with other variables. We decide to delete this observation instead.

	BsmtQual	BsmtCond	BsmtExposure	BsmtFinType1	BsmtFinSF1	BsmtFinType2	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF	BsmtFullBath	BsmtHalfBath
333	Gd	TA	No	GLQ	1124	<NA>	479	1603	3206	1	0

The BsmtFinType2 in No.333 observation indicates that the house did not have basement. But according to other variable value, it indicates that the house had a basement. So it is not consistent with other variables. We also decide to delete this observation instead.

MasVnrType & MasVnrArea:

These two variable should be dealt together because they are all related to the masonry veneer and should be consistent with each other.

The description of the values in the two variables is shown below:

MasVnrType: Masonry veneer type

BrkCmn	*Brick Common*
BrkFace	*Brick Face*
CBlock	*Cinder Block*
None	*None*
Stone	*Stone*

MasVnrArea: Masonry veneer area in square feet

So the NA means that there is no veneer in the house. Also, we found that the number of missing value in these two variable is the same. So we can replace all missing value in MasVnrType with "None" and replace all missing value in MasVnrArea with 0.

Electrical:
There are only 1 NA. Values descriptions are shown below:

Electrical: Electrical system

SBrkr	*Standard Circuit Breakers & Romex*
FuseA	*Fuse Box over 60 AMP and all Romex wiring (Average)*
FuseF	*60 AMP Fuse Box and mostly Romex wiring (Fair)*
FuseP	*60 AMP Fuse Box and mostly knob & tube wiring (poor)*
Mix	*Mixed*

After we check out the distribution of the values of electrical system, we decide to use the mode

of it to replace this missing value.
The frequency table of Electrical is shown below:

FuseA	FuseF	FuseP	Mix	SBrkr
94	27	3	1	1334

As we can see from the table, 91% of the electrical system are “SBrkr”. So replacing the missing value with “SBrkr” would be reasonable.

3. Model fit and selection

We divided our cleaned data. 70% are training set and 30% of that are the validation set. And we use the the training set to build models and the validation set to assess the model performance.

3.1 Naive model

The naive model is to use the mean of SalePrice in the training set to as our SalePrice prediction. The mean of SalePrice in the training set = 162000, so the prediction of sale price is 162000 in the naive model. And the **Mean Absolute Error (MAE)** in the test set turns out to be **59054.0936073.** We will use this value as the baseline to evaluate other machine learning models we use.

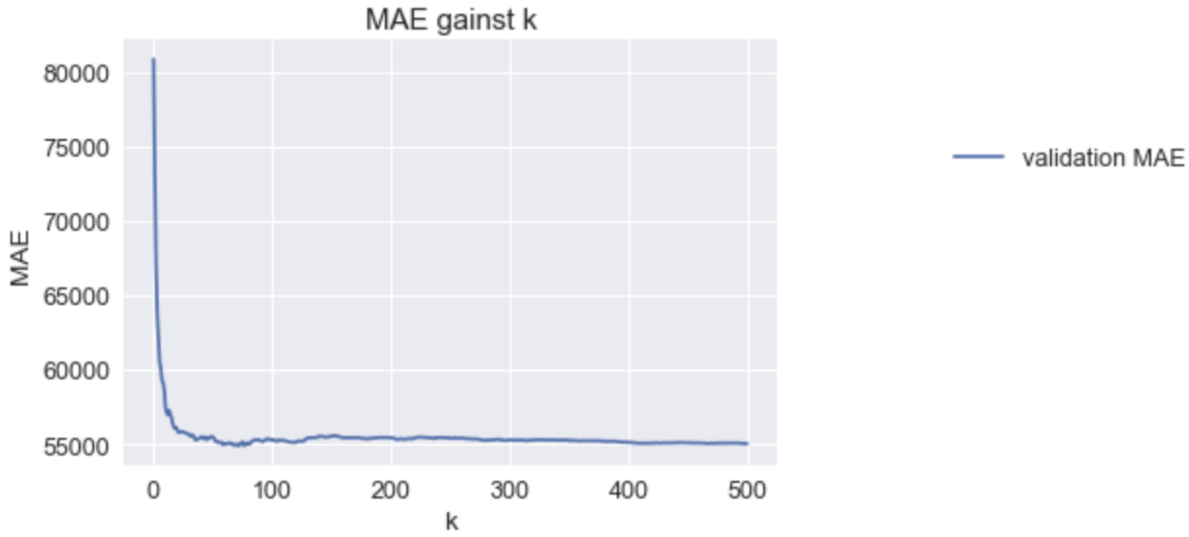
3.2 k-Nearest Neighbors(KNN)

The k-nearest neighbors algorithm can be used for regression and give us the average value of its k nearest neighbors. There is no assumptions in terms of the relationship between the predictors and response variable. Therefore, we think KNN might be a good method to SalePrice prediction and it is easy to use.

We build a KNN model to see whether it will be better than the naive method. We tried different number of nearest observation (k) and found different MAE.

After some trial and errors, we realized that there can be an optimal k which can give us the lowest MAE of k-NN in the validation set.

```
When K is 76, the MAE is 54907.95949712387.
```



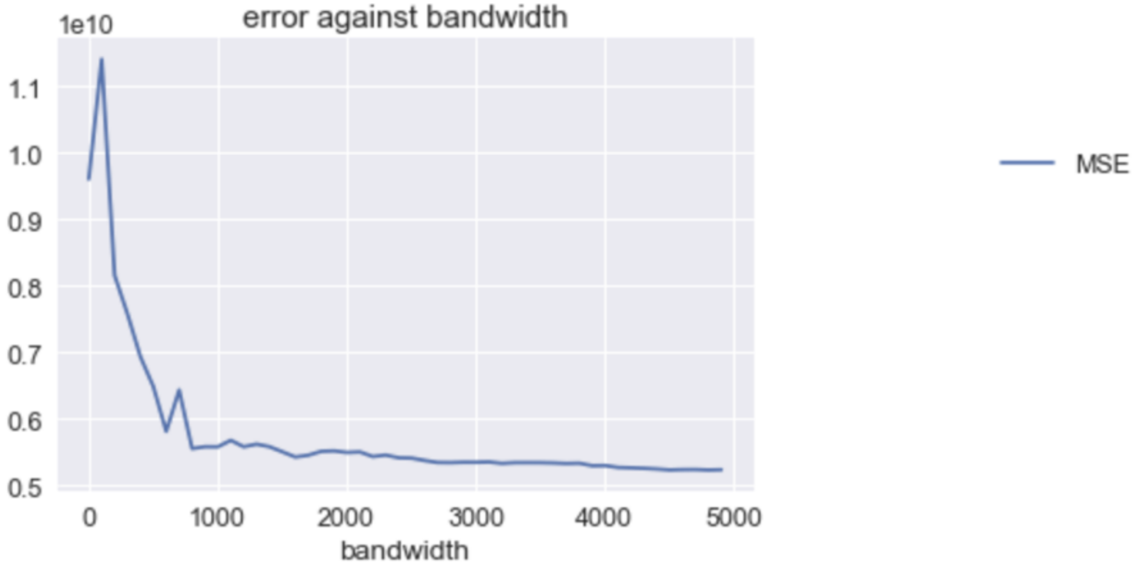
The MAE is slightly lower than the naive method(59054.0936073), and 76 nearest observation is acceptable because we are predicting a house price based on that of the 76 nearest houses. In addition, it is computationally expensive to find an optimal k nearest neighbors. Besides, KNN is not interpretable because we cannot see how does the change of dependent variables(12 predictors) affect the response variable (SalePrice). Hence, we thought KNN in an option to forecast house price, but there won't be a lot value added by this model.

3.3 Local Regression (loess)

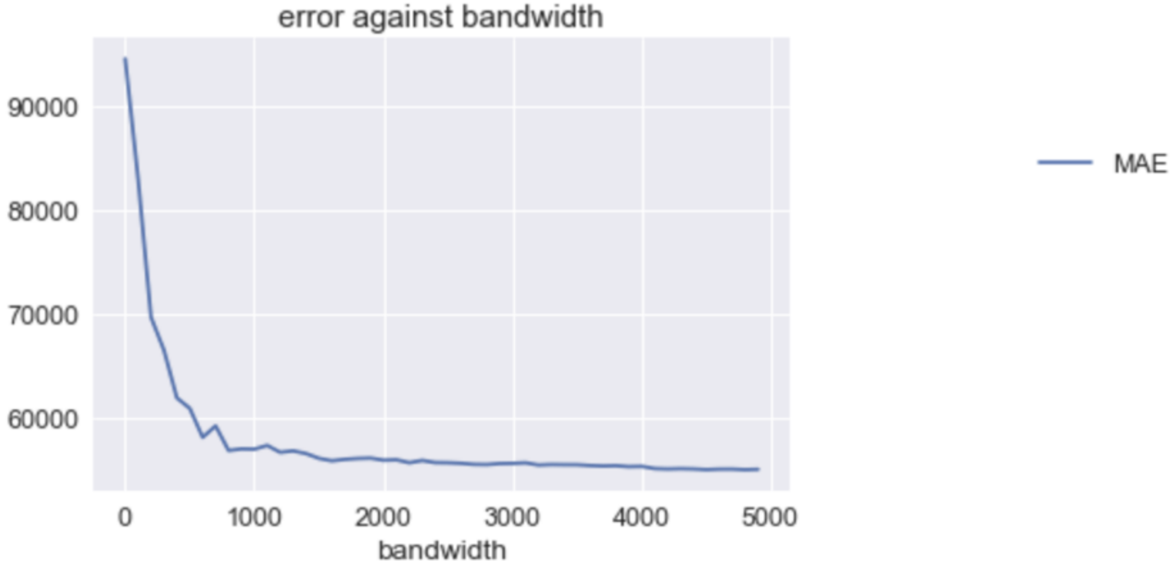
We further decide to try a Loess model because loess is a non parametric regression method that combines multiple regression models in a *k*-nearest-neighbor-based meta-model. [3] Loess is easy to use and flexible to adjust parameters.

Next, we tried Loess model. We shrink the range of the radius between 1 and 5000, and we choose R=48 because the MAE and MSE are both smallest. In addition, MAE is acceptable compared to that of the naive model (59054.0936073).

[3] https://en.wikipedia.org/wiki/Local_regression



```
When R is 48, the MSE is 5234453127.18507.
```

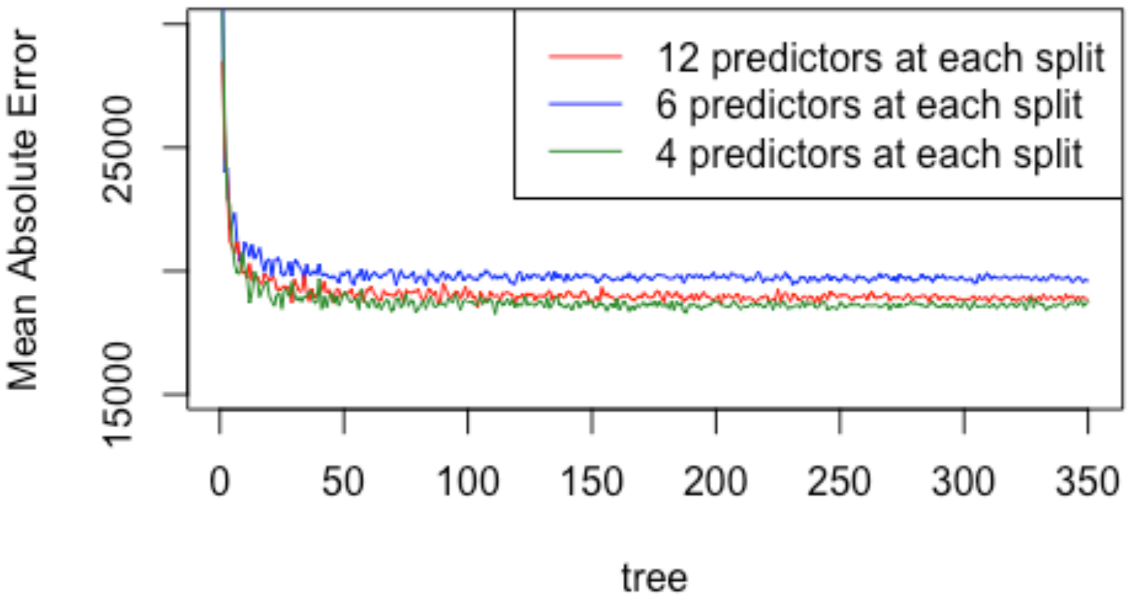


```
When R is 48, the MAE is 54932.68997727367.
```

However, the model cannot be interpreted because we cannot see how does the change of dependent variables(12 predictors) affect the response variable (SalePrice).Hence, we thought loess is an option to forecast house price, but there won't be a lot value added by this model.

3.4 Random Forest

We also tried to use the random forest to predict the sale price. After fitting the model and assess the goodness of fit, we found that when the number of tree in the ensemble is greater than 200 the mean absolute errors are stabilized. And we tuning the parameters of the “mtry” to select how many of the predictors we choose in each split. At last, we found that when mtry=4, the mean absolute error is the smallest. The figure below shows the changes of the MAE as tuning parameters.



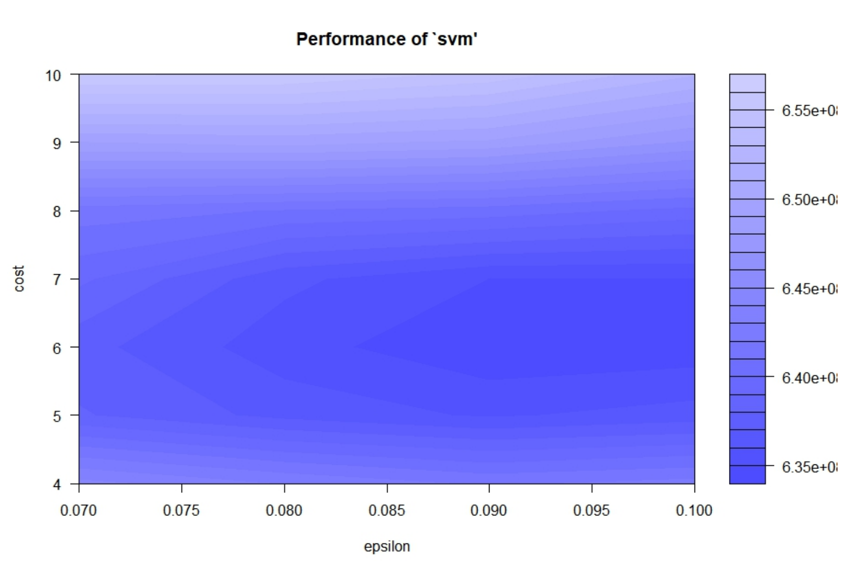
So the MAE is 18306.41 in our best RF model, which is largely decreased compared to the local method such as loess and the k-NN. So based on the error value, the RF model is much better than the local methods.

3.5 Multiple Linear Regression (MLR)

Linear Regression remains a strong method for the prediction purposes, especially when the model can be trained on large number of features. We were interested to see the performance of MLR in comparison to other methods. We fitted a model using our standard set of the predictors. For this model MAE is equal to 20604. But FullBath and GrLivArea haven’t passed t-test, meaning they don’t explain variation in our response variable. After removing these two variables from our MLR model, we obtained MAE equal to 20525. The error for MLR model is higher than Random Forest, which means that regression isn’t as efficient in the extraction of information from predictor variables to explain variation in our response.

3.6 Support Vector Regression (SVR)

We employ SVR based on SVM classifier learned during classroom sessions to predict sales prices. In the case of SVR response is a quantitative variable. By fitting a hyperplane to our data, we hope to improve results obtained by MLR. First, we fit a model with our selected predictors. MAE obtained is equal to 18132.56. Next, we seek to tune our hyperparameters, so we employ "tune" function in R.

	Best performance is obtained with the hyperparameters: ● epsilon = 0.09 ● cost = 6 After fitting the model with the recommended hyperparameters, we obtained MAE equal to 18460. Therefore, we couldn't improve the model's performance.

Support Vector Regression based model proved to be robust for our purposes, outperforming other statistical methods. But before committing to the recommendation, we want to fit a model, using gradient boosting method.

3.7 Gradient Boosting Method(GBM)

At last, we tried to fit the dataset with GBM model. We choose the number of tree in the ensemble as 4000 because the MAE is stabilized at there. And we choose the depth of tree as 4 because we do not want to overfit the data there. In GBM, the boosting will reduce the errors gradiently thus decrease the biased steadily. So we don't fit a high-depth tree to the dataset. And after tuning the model, the MAE is minimized when the depth is 4. So 4 is a reasonable number for our GBM model.

The MAE is 17915.71 in our GBM model, which is less than SVR model.

3.8 Model selection

The table below shows the comparison of each of four models.

Model	Advantage	Disadvantage
k Nearest Neighbours	● Easy to understand. ● Simple to implement. ● No assumptions required about the relationship between the response and predictors.	● **Large error** in validation set: the **Mean absolute error (MAE)** is 55000. ● It is computationally expensive to find the k nearest neighbours. ● The model can not be interpreted.
Local Regression	● It does not require the specification of a function to fit a model to all of the data. ● The model is easy and flexible to adjust the parameters.	● The model can not be interpreted. ● **Large error:** The **MAE** is 54932.
Random Forest	● **Small error**: The **MAE** is only 18306; ● **Less variance**: By using multiple trees, RF reduces the chance of stumbling across a classifier that doesn't perform well because of the relationship between the train and test data; ● **No assumptions required** about the relationship between the response and predictors.	● They are much harder and time-consuming to construct a bunch of tree to averaging the results than local methods. ● The model can not be interpreted.
Multiple Linear Model	● **Small error** in validation set: The **MAE** is 20525. ● MLR performs better than many other "proven" methods. ● MLR provides summary statistics that is easily interpretable.	● MLR is not the most efficient method for extracting information through multiple predictors. ● **Not all of our predictors** demonstrated **linear relationship with SalesPrice**, diminishing MLR effectiveness.
Support Vector Regression	● **Small error**: The **MAE** is 18132. ● SVR enables strong prediction performance. ● Good at learning from the data, rather fast. Model allows further tuning and possibly better performance.	● Tuning the model is computationally expensive, since it requires iterating over all the possible hyperparameters. ● It's difficult to find global optimum with this method.
Gradient	● **Lowest error**: The MAE of	● **Computational expensive**: GBM

Boosting Method	validation set is only 17915.71 in our models. ● **No assumptions required** about the relationship between the response and predictors. ● **Less bias**: Boosting gradually improves predictions by learning on residuals. Using multiple error prediction trees, GBM reduces the bias in the weak model and captures patterns in the data that the initial tree missed.	training process took a longer time than random forest. ● The model cannot be interpreted.

Aftering weighing the pros and cons of each, we decide to choose **Gradient Boosting Method.**

- From quantitative perspective, it gives us the lowest error value. So it has the best performance of prediction.

- From the qualitative angle, the relationship between the response variable and the features is complex, which is not linear or quadratic that we can identify intuitively. So using GBM can avoid to predefine the relationship and get a inaccurate prediction based on incorrect assumption we made.

- Although the GBM model cannot be interpreted, it identifies the internal pattern better than the interpretable model (e.g. Linear regression/decision tree) between the sale price and other factors in the dataset. The trade-off between biased and variance is well balanced as well in the GBM.

Therefore, we recommend to use gradient boosting method to make a prediction.

Communication of results, and advice to a non-expert

After exploring and cleaning the dataset, we found that overall quality of a house is the most influential feature of the sale price throughout the 80 features. And we also found that the above ground living area, the located neighborhood and the garage size of a house are important features that influence the sale price of a house. And since some of the features are highly correlated, it is unnecessary and inappropriate to use all features to predict the sale prices. Choosing a number of important features that are not highly intercorrelated is a good choice to forecast.

After fitting and evaluating different kinds of models, we recommend to use gradient boosting method which is an algorithm that produces a prediction model in the form of a collection of weak prediction models, typically decision trees with low variance and high bias.[4] It can help us forecast the house price with higher accurateness through the elimination of non-optimal predictions, measured by the errors. The high bias is cut down by adding error adjustments in the decision tree sequentially. Finally, the adjusted decision tree will be with low bias and ideal for forecasting.

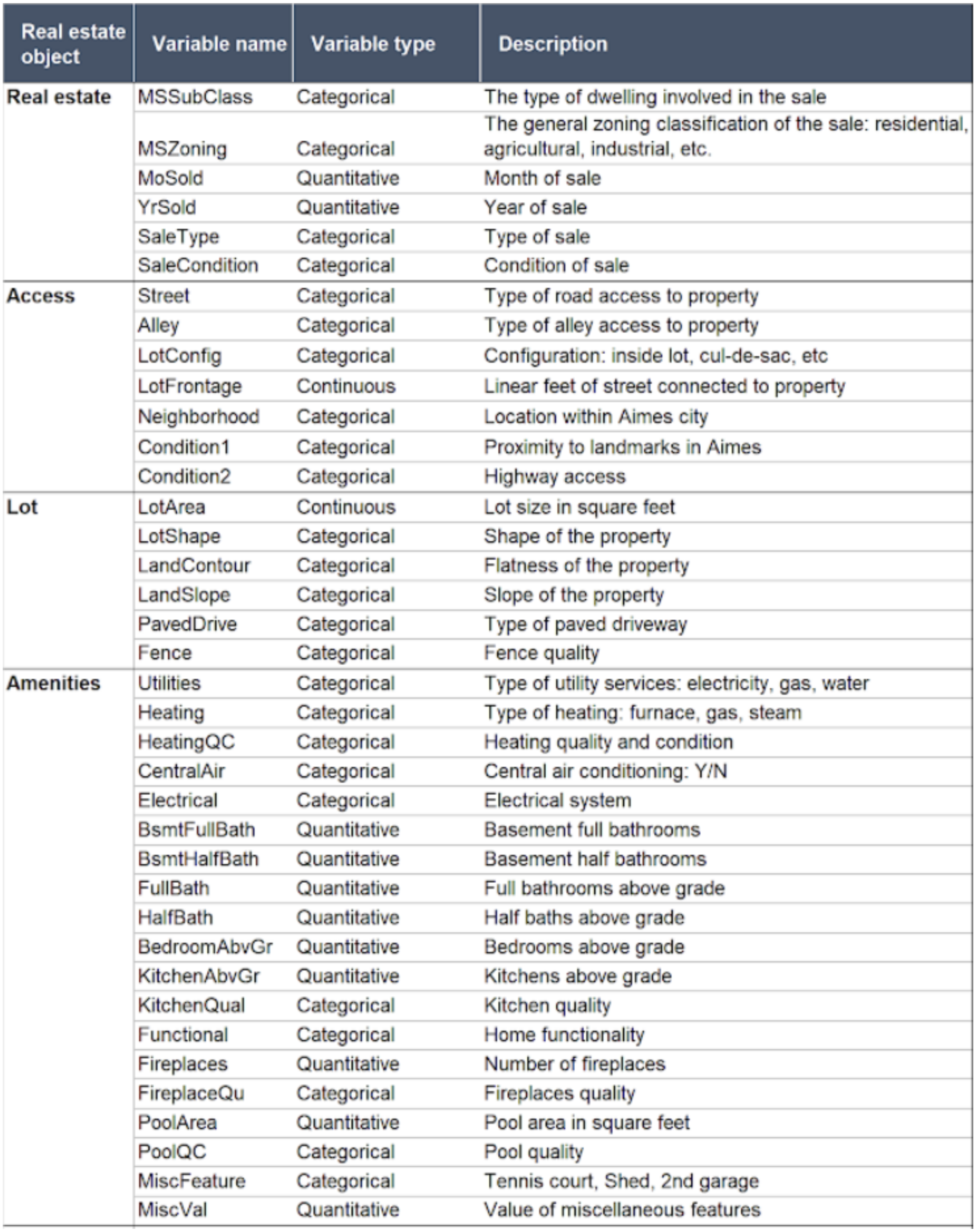
During our research, we found that machine learning techniques are widely used in the real estate industry. For example, Real Estate Wire (REW) is a real estate marketplace and information hub in BC and Ontario, which provide buyer, seller, and third parties to search house price based on city, neighborhood, address, and school, etc. This is a platform for people who are buying and selling houses to search easily and access information based on their preferences.[5] We thought it will be better to include more predictors that we used to develop our prediction model such as price, area, garage and bathroom in ‘advanced search’. In this way, customers can search houses that are more fitted with their preferences.

[4] https://en.wikipedia.org/wiki/Gradient_boosting

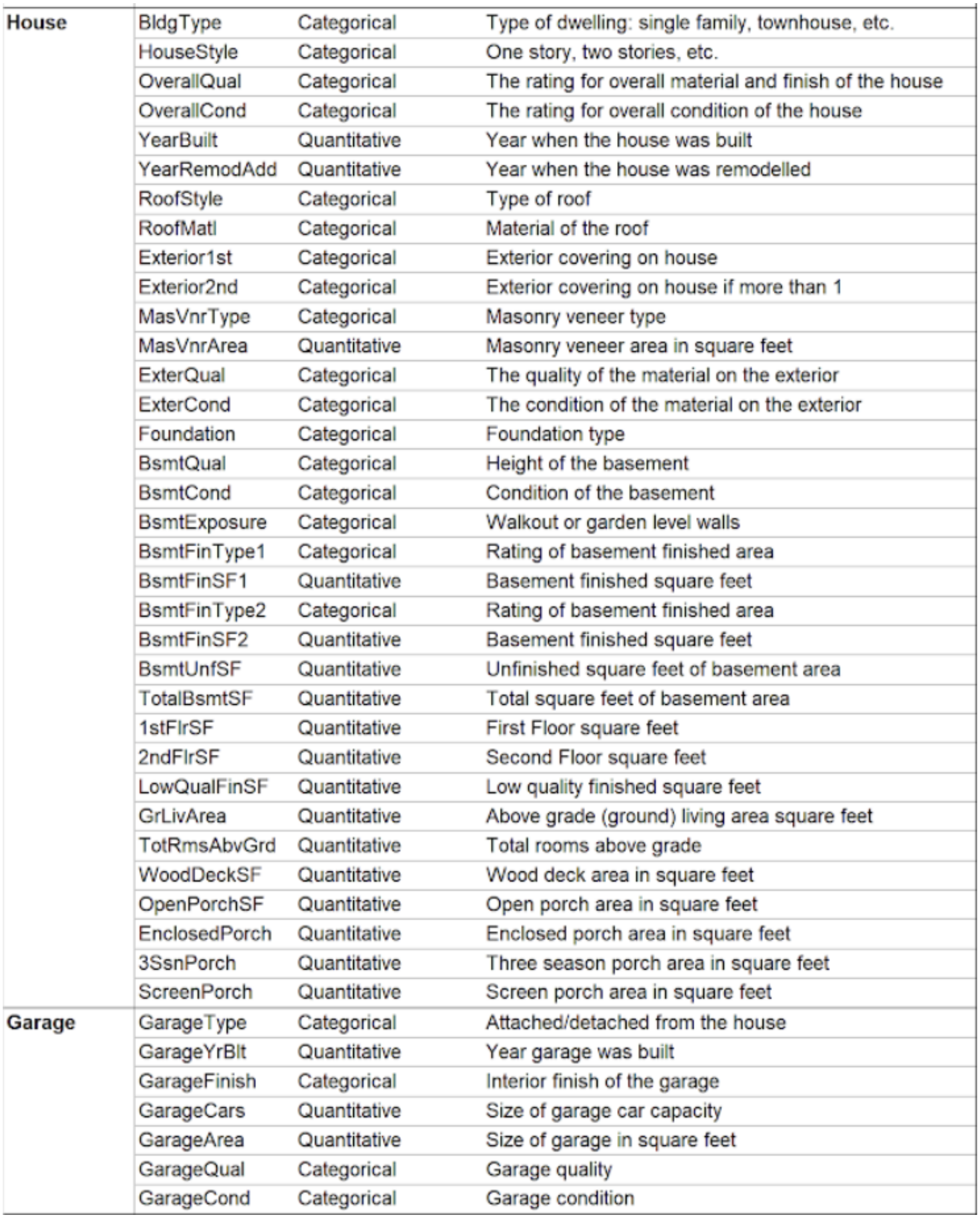
[5] https://www.rew.ca/about-us

Appendix

1. Index of the predictor variables



Real estate object	Variable name	Variable type	Description
Real estate	MSSubClass	Categorical	The type of dwelling involved in the sale
	MSZoning	Categorical	The general zoning classification of the sale: residential, agricultural, industrial, etc.
	MoSold	Quantitative	Month of sale
	YrSold	Quantitative	Year of sale
	SaleType	Categorical	Type of sale
	SaleCondition	Categorical	Condition of sale
Access	Street	Categorical	Type of road access to property
	Alley	Categorical	Type of alley access to property
	LotConfig	Categorical	Configuration: inside lot, cul-de-sac, etc
	LotFrontage	Continuous	Linear feet of street connected to property
	Neighborhood	Categorical	Location within Aimes city
	Condition1	Categorical	Proximity to landmarks in Aimes
	Condition2	Categorical	Highway access
Lot	LotArea	Continuous	Lot size in square feet
	LotShape	Categorical	Shape of the property
	LandContour	Categorical	Flatness of the property
	LandSlope	Categorical	Slope of the property
	PavedDrive	Categorical	Type of paved driveway
	Fence	Categorical	Fence quality
Amenities	Utilities	Categorical	Type of utility services: electricity, gas, water
	Heating	Categorical	Type of heating: furnace, gas, steam
	HeatingQC	Categorical	Heating quality and condition
	CentralAir	Categorical	Central air conditioning: Y/N
	Electrical	Categorical	Electrical system
	BsmtFullBath	Quantitative	Basement full bathrooms
	BsmtHalfBath	Quantitative	Basement half bathrooms
	FullBath	Quantitative	Full bathrooms above grade
	HalfBath	Quantitative	Half baths above grade
	BedroomAbvGr	Quantitative	Bedrooms above grade
	KitchenAbvGr	Quantitative	Kitchens above grade
	KitchenQual	Categorical	Kitchen quality
	Functional	Categorical	Home functionality
	Fireplaces	Quantitative	Number of fireplaces
	FireplaceQu	Categorical	Fireplaces quality
	PoolArea	Quantitative	Pool area in square feet
	PoolQC	Categorical	Pool quality
	MiscFeature	Categorical	Tennis court, Shed, 2nd garage
	MiscVal	Quantitative	Value of miscellaneous features



House	BldgType	Categorical	Type of dwelling: single family, townhouse, etc.
	HouseStyle	Categorical	One story, two stories, etc.
	OverallQual	Categorical	The rating for overall material and finish of the house
	OverallCond	Categorical	The rating for overall condition of the house
	YearBuilt	Quantitative	Year when the house was built
	YearRemodAdd	Quantitative	Year when the house was remodelled
	RoofStyle	Categorical	Type of roof
	RoofMatl	Categorical	Material of the roof
	Exterior1st	Categorical	Exterior covering on house
	Exterior2nd	Categorical	Exterior covering on house if more than 1
	MasVnrType	Categorical	Masonry veneer type
	MasVnrArea	Quantitative	Masonry veneer area in square feet
	ExterQual	Categorical	The quality of the material on the exterior
	ExterCond	Categorical	The condition of the material on the exterior
	Foundation	Categorical	Foundation type
	BsmtQual	Categorical	Height of the basement
	BsmtCond	Categorical	Condition of the basement
	BsmtExposure	Categorical	Walkout or garden level walls
	BsmtFinType1	Categorical	Rating of basement finished area
	BsmtFinSF1	Quantitative	Basement finished square feet
	BsmtFinType2	Categorical	Rating of basement finished area
	BsmtFinSF2	Quantitative	Basement finished square feet
	BsmtUnfSF	Quantitative	Unfinished square feet of basement area
	TotalBsmtSF	Quantitative	Total square feet of basement area
	1stFlrSF	Quantitative	First Floor square feet
	2ndFlrSF	Quantitative	Second Floor square feet
	LowQualFinSF	Quantitative	Low quality finished square feet
	GrLivArea	Quantitative	Above grade (ground) living area square feet
	TotRmsAbvGrd	Quantitative	Total rooms above grade
	WoodDeckSF	Quantitative	Wood deck area in square feet
	OpenPorchSF	Quantitative	Open porch area in square feet
	EnclosedPorch	Quantitative	Enclosed porch area in square feet
	3SsnPorch	Quantitative	Three season porch area in square feet
	ScreenPorch	Quantitative	Screen porch area in square feet
Garage	GarageType	Categorical	Attached/detached from the house
	GarageYrBlt	Quantitative	Year garage was built
	GarageFinish	Categorical	Interior finish of the garage
	GarageCars	Quantitative	Size of garage car capacity
	GarageArea	Quantitative	Size of garage in square feet
	GarageQual	Categorical	Garage quality
	GarageCond	Categorical	Garage condition

2. Random forest scores

We built a random forest includes all variables to assess the importance of the each variable. The most important predictors are shown below:

OverallQual	1.23E+12
GrLivArea	6.88E+11
GarageCars	6.45E+11
Neighborhood	6.34E+11
ExterQual	4.20E+11
TotalBsmtSF	2.63E+11
GarageArea	2.47E+11
X1stFlrSF	2.46E+11
KitchenQual	2.29E+11
YearBuilt	2.25E+11
BsmtFinSF1	1.50E+11
LotArea	1.03E+11
X2ndFlrSF	1.01E+11
TotRmsAbvGrd	9.83E+10
FullBath	9.61E+10