House Sales Prediction using Bootstrapping as Resampling

Introduction

We have been given house-data to analyses and understand the housing market. We will be drawing sum-maries and plotting relationships between the data to make analysis. We will than be using Machine Learning techniques to predict the Overall Condition and the Sale Price for the houses. The language to be used is R. We will be drawing conclusions about the data from all these steps.

Numerical and Graphical Summaries

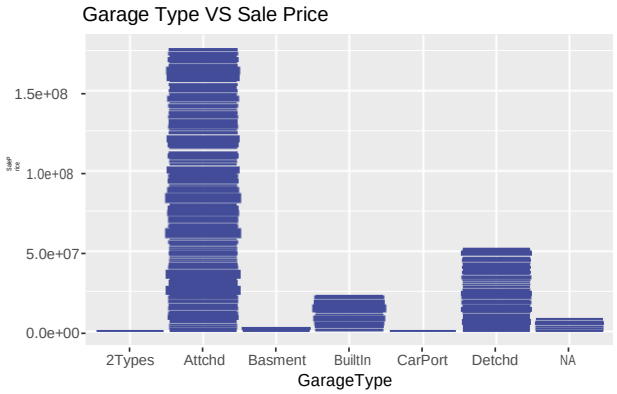
We have been provided data containing information about houses. The data consists of 1461 entries with 51 features. The data contains mostly categorical features with very few numerical ones. We use the Summary(data) function in R to get the summary the data. The function returns the Mean, Median, Min, Max, Quartiles and NA of all numerical columns. We still do not know a lot about the categorical and descriptive data features. We use the function describe from library(psych) to gain that. Describing the data returns a concise statistical summary according to each variable be it numerical or categorical. The function output statistical summaries like variables, count, mean, standard deviation, median, min, max etc. The outputs can be read below, due to limitations of the report we cannot include the detailed summaries here:

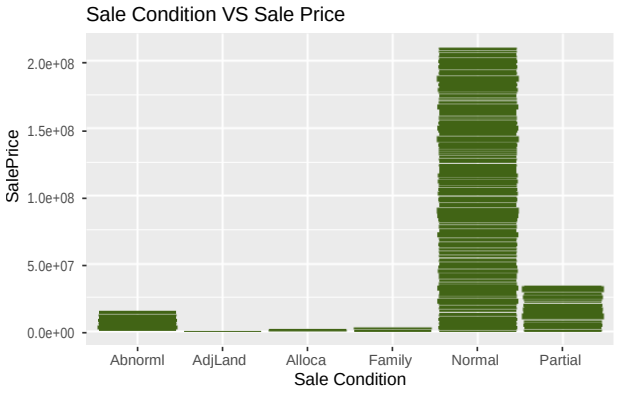
As mentioned above the data is categorical, meaning we cannot generate heatmaps or correlation plots. We will be analysing the impact of variables on the response variable “Sale Price” using bivariate plots. We will be considering the variable “Sale Price” as the major response variable. We have made around 19 plots for all major features. Some will be included directly, and others will be given as comments about the data. Key Takes from the data:

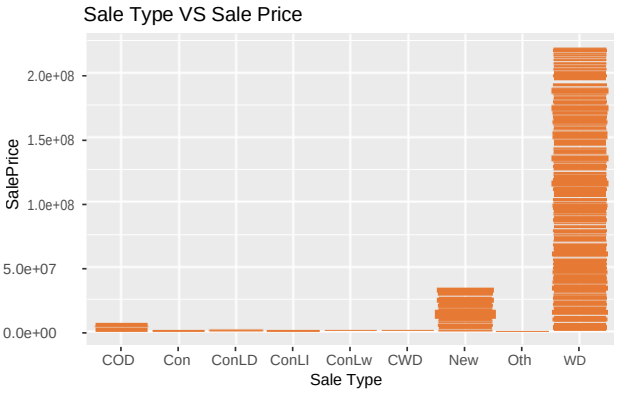
1. It can noted that most house with high prices had average overall condition while better condition sells at better price.
2. We notice that Attached Garages tend to make the houses sell at higher rates.
3. Houses with Normal Sale Condition tends to have higher sale prices followed by Partial condition.
4. Warranty Deed - Conventional Sale Type tend to bring higher sales for houses.
5. We also notice that houses last sold from 2013 - 2015 bring higher prices for sale.
6. Houses with Typical Functionality tend to have sale prices on the higher end of plot.
7. House with 3 bedrooms followed by 2 have higher sale price ranges.
8. Houses equipped with Gas forced warm air furnace heating are mostly sold at higher prices.

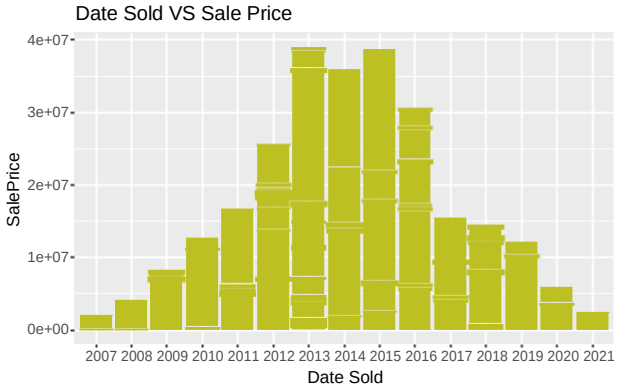
9. Kitchen Quality being Typical/Average and Good is an essential attribute to get a good price for a house.
10. House with foundations made of Poured Contrete followed by Cinder Block have higher sale prices.
11. Houses built post 1980 tend to have higher sale prices.
12. 1 story and 2 story houses are more commonly sold at high prices.
13. Normal Proximity to Conditions tends to make the sale price go to the higher end.
14. House in Neighborhood of College Creek and North Ames have high prices.
15. Availability of All public Utilities tend to get more houses sold.
16. No Alley access tends to get more house sold at high prices.

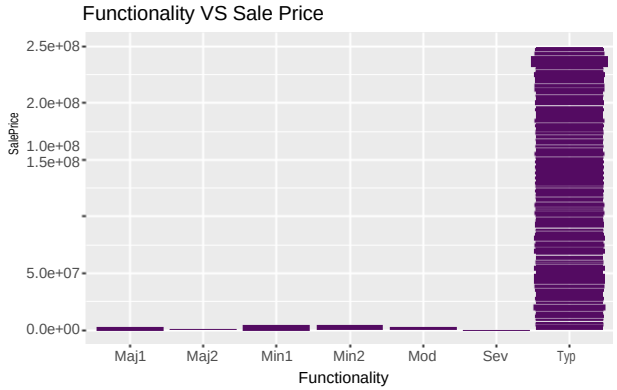

Overall Cond VS Sale Price
Sale Price
1.5e+08
1.0e+08
5.0e+07
0.0e+00
1
2
3
4
5
6
7
8
9
Overall Condition


Garage Type VS Sale Price
1.5e+08
1.0e+08
5.0e+07
0.0e+00
2Types
Attchd
Basment
BuiltIn
CarPort
Detchd
NA
GarageType

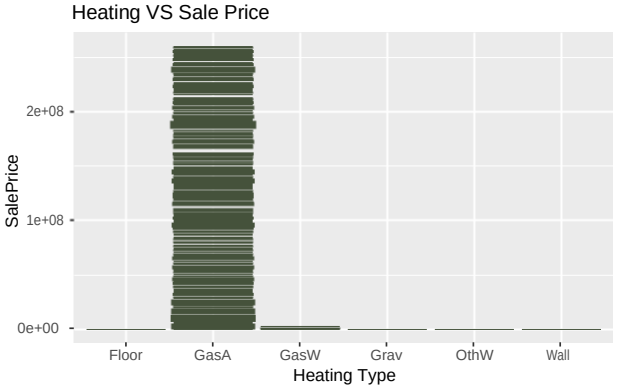

Sale Condition VS Sale Price
SalePrice
2.0e+08
1.5e+08
1.0e+08
5.0e+07
0.0e+00
Abnorml
AdjLand
Alloca
Family
Normal
Partial
Sale Condition


Sale Type VS Sale Price
SalePrice
2.0e+08
1.5e+08
1.0e+08
5.0e+07
0.0e+00
COD
Con
ConLD
ConLI
ConLw
CWD
New
Oth
WD
Sale Type

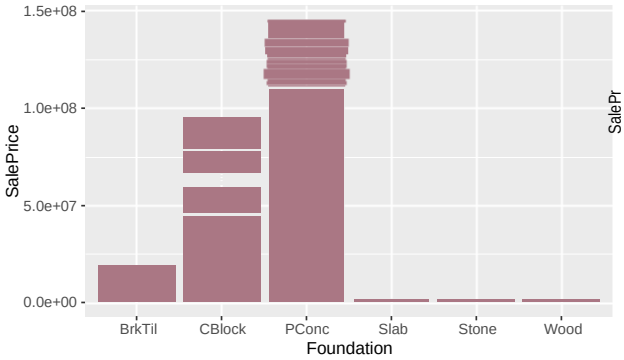

Date Sold VS Sale Price
SalePrice
4e+07
3e+07
2e+07
1e+07
0e+00
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Date Sold


Functionality VS Sale Price
SalePrice
2.5e+08
2.0e+08
1.5e+08
1.0e+08
5.0e+07
0.0e+00
Maj1
Maj2
Min1
Min2
Mod
Sev
Typ
Functionality

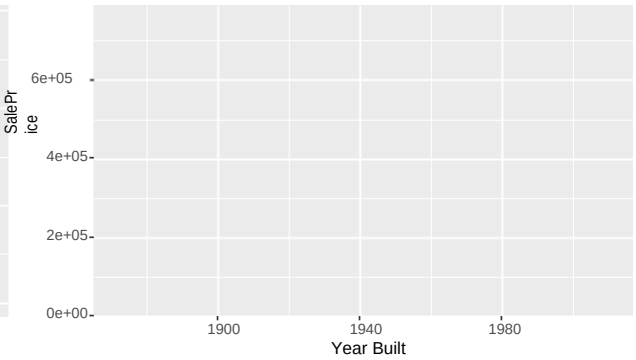

Bedrooms VS Sale Price
SalePrice
1.5e+08
1.0e+08
5.0e+07
0.0e+00
0
1
2
3
4
5
6
8
No. of Bedrooms


Heating VS Sale Price
SalePrice
2e+08
1e+08
0e+00
Floor
GasA
GasW
Grav
OthW
Wall
Heating Type

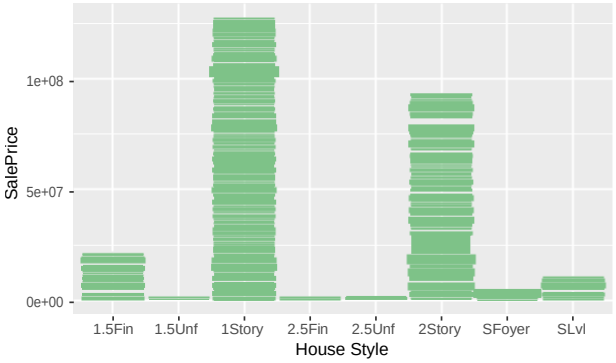
Foundation VS Sale Price



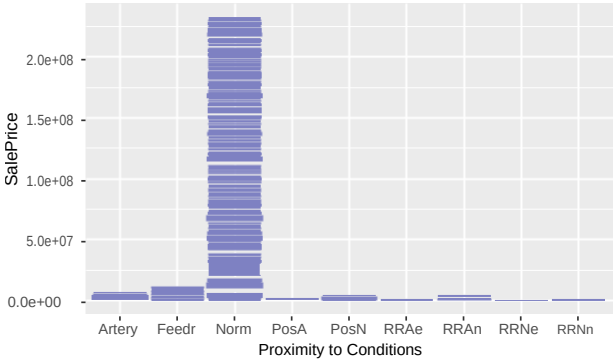
Year Built VS Sale Price



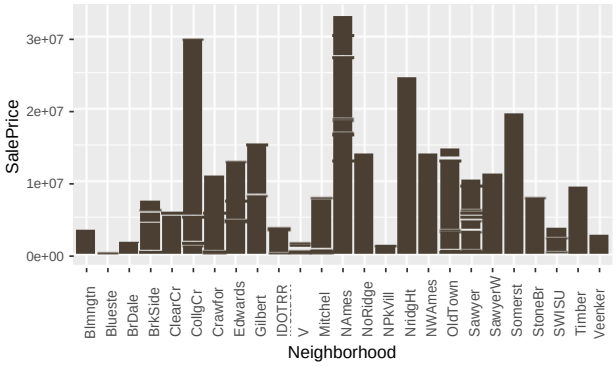
House Style VS Sale Price



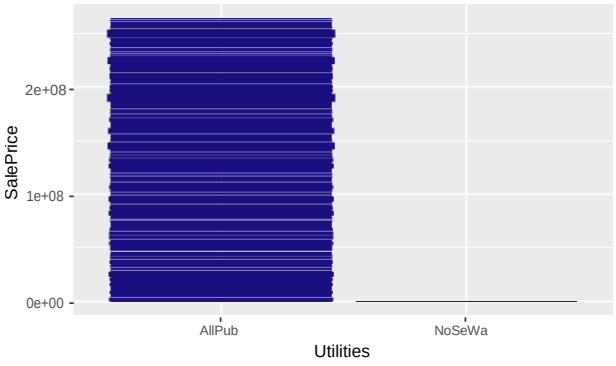
Condition Proximity VS Sale Price



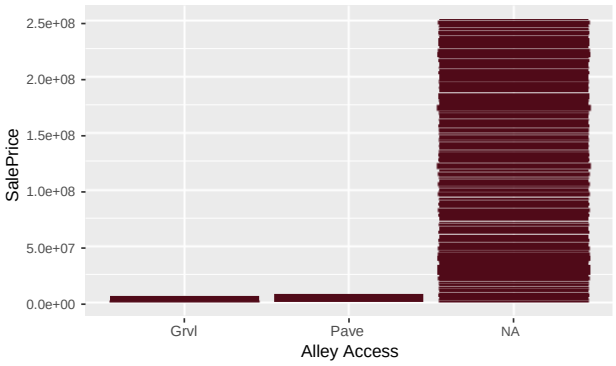
Neighborhood VS Sale Price



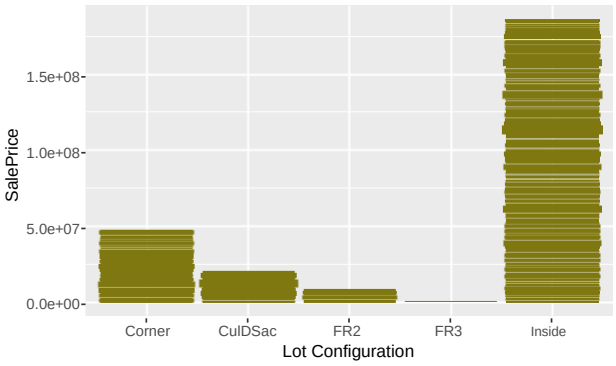
Utilities VS Sale Price



Alley Access VS Sale Price



Lot Configuration VS Sale Price



Logistic Regression

This is a multinomial classification problem due to the fact that there are 3 classes in the output data which are “Poor”, “Average” and “Good”. Before the application of multinomial logistic regression algorithm, some pre-processing to data has been done. First of all, “PoolQC” , “Fence”, “MiscFeature”, “LotFrontage” and “Alley” variables have been removed from dataset. Because %99.5 of “PoolQC”, %80 of “Fence”, %96 of “MiscFeature” ,%17 of “LotFrontage” and %93 of “Alley” is missing. Moreover, “RoofMatl” has also been removed since 1434 value out of 1460 is the same. Therefore, there is no meaning to keep this feature in the model. There are around %2.5 missing values in “BsmtQual” and “BsmtCond” variables. But after inspecting them, It was realized that these values are missing because there is no basement on those houses and NA values have been replaced with “No_Basement” value. The same situation also appears in “GarageType” and “GarageCond”. The missing values in these variables have been replaced with “No_Garage”. Lastly, missing values in “MasVnrArea” have been replaced with 0 because the majority of values in this variable is 0. Then, the dataset has been split into training and test sets with 0.25 test set proportion. The last step in pre-preprocessing was the scaling and centering (Standardization)the numeric values in the dataset. To prevent data leakage, standardization was trained on training set and transformed on the test set by using training mean and standard deviation. R does not support multinomial logistic regression by default. Hence, the “multinom” function from “nnet” package has been used as the algorithm. This function automatically deals with dummy encoding and label encoding automatically. After fitting the model to train data, model are tested with test data. According to confusion matrix created with test data, the model achieves %76.8 accuracy. The model is doing good job on detecting “Average” condition houses. This can be expected as “Average” class is the majority in dataset. The model achieves %58 accuracy for the class “Good” and only 15% accuracy for “Poor” class. The performance on the “Poor” class is not satisfying but it can be expected since only 31 example of 1460 belongs to “Poor” class.

Applying a similar model studied in MA321 SVM linear classifier

We also carried out a SVM linear classifier as studied in MA321 to classify the Overall condition of the house. As we were comparing the two different classification methods. We decided to use the same training and testing split as was applied for the logistic regression to ensure a fair comparison. Comparing the two classifiers. The accuracy of the model determines the overall correctly classified observations. The following are the Accuracy scores. • Logistic Regression 77.41% Accuracy • SVM Linear Classifier 75.48% Accuracy

Comparing the Sensitivity and Specificity

Sensitivity is the rate of True Positives predicted by the model which is to say that the rate of each class being predicted correctly in this multi-nomial classification. The specificity is the rate of True Negatives which is to say how many observations were correctly not classified in the specific class (observations not belonging to that class) were predicted to not belong to that class correctly.

Comparing the two Model's Sensitivity we can see that the SVM model is better at predicting whether an Overall condition is actually poor. The Logistic Regression is better at predicting the Average and Good Class. The same goes for the Specificity of the three classes.

Sale Price Prediction (Regression Model training and evaluation)

Prediction of House Prices:

In this part of assignment, we've to choose two models that can fit and predict the house prices at their best. Now normally, Regression techniques are the most prevalent and proven methods for this kind of problem, but firstly we were supposed to use ensemble methods and secondly, the final two techniques that I'll be discussing ahead performed very well overall.

First model or ensemble technique used is Random Forest Regressor. A RF can build multiple decision trees to be combined together for precise predictions. The best part about assembling different individual decision trees is that they are uncorrelated and stochastic at the same time, giving us way better results than single trees. This process of multitude decision trees is known as Bagging.

Hyperparameter tuning that was optimal for our models was done using trial and error-based methods. Results obtained by the method gave R2 value of around 0.88 with validation R2 of 0.85. It was trained on the pre-processed data I got from the previous part, that was performed by my colleague.

Second technique used, despite the fact that it is a greedy algorithm, and they tend to overfit the model, was Gradient Boosting. The working behind the model is that it depends on weak learner i.e., Decision Trees and lean on the intuition that the next model will be the best one. That allows the subsequent models to combine and reduce the residual effect of the predictions as a whole making it a better fit. This comparison of predictions and assignment of weights to the best learner is known as Boosting Technique. Results obtained by the method gave R2 value of 0.88 with validation R2 of 0.83.

Resampling Methods

Now that we've modelled and predicted the house prices in the previous part, there is still a very vital step that can enhance the performance of the model by leaps and bounds. This step is multiple techniques that can be followed to improve the estimate of the population and it is known as Resampling of the data. Now statistically, sampling in a nutshell is the preference of data that suits our requirement from a set of huge population, carrying a hypothesis with it that it represents the whole population in some aspect. Now with this can come problem of its own. One can be of nature that a bias could've been introduced while specifying a subset of the population that can guide our predictions to different direction thus giving modelling results. There can also be random variations associations with the data that can affect our results ultimately. Dealing with these kinds of issues can be hefty and time consuming if dealt humanly. Here comes the resampling methods to save the day.

Now that we have only modelled our data in one way i.e., the way it was chosen from a population, with no idea of uncertainty it carries along; we can sample the same data into chunks or folds that can be run multiple times to see the variation and uncertainty. This will be eventually help us to choose the best fold of them all. This technique is known as Resampling. There are different methods that can be used to resample are to be known as Resampling Methods.

First method practiced in the assignment was K-Folds Cross Validation. It depends on the Train-Test split of the data that can give us a fair idea about the distribution of the data and which fold of the data is least skewed. As the name suggests data is randomly split into the given number of folds. Then with the given number of folds, model is fitted for each of them, and we get scores and associated error for each fold. This process is repeated for the given number of folds. K-Folds was applied to both of the previous provided models i.e., Random Forest and Gradient Boost and results were recorded.

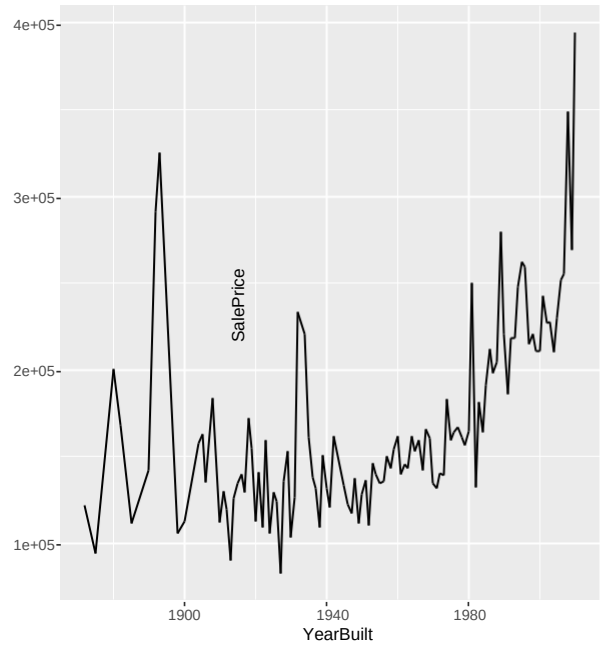
Second method was Bootstrapping. It is widely used to draw statistical inferences from a given population. It follows in by extracting a sample of size n from the population and creating another random sample by replacing the original sample. These replications are done for given B times, giving a total of B Bootstrap samples. With the results obtained, evaluation is carried out using the population parameters i.e., population mean, population standard deviation, population variances and confidence intervals. Calculations were done on 95% CI to get BCa (Bias Corrected and Accelerated) bootstrap intervals. Results showed that the data was positively skewed and is to be adjusted towards right.

Exploring Further the Dataset

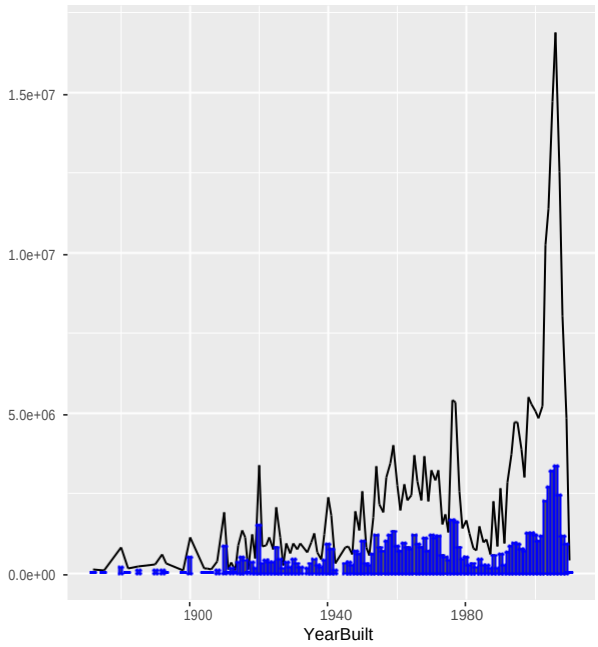
In this section we aim to explore the effect of house prices regarding the year they were build. Normally, in the Real State industry a house looses value as older it is. This happens due lots of factors but the most relevant is that the house is just old.The constructions techniques and materials used in the house are out-dated and they cannot compete with a similar house in characteristics but that has been built more recently. Also, investors are more cautious when buying old houses since they are more promt to have structural issues and other things affecting the house habitability.

We have grouped our data by the year it was built and computed the mean of sale price.

1 : Mean Sale Price by House Built Year



2 : Aggregated Sale Price by built Year and Sales Count



Older houses (Figure 1) are prompt to be more cheap in a general basis. Nevertheless, this also depends on the house characteristics as it's not the same a 100 ft. sq. house than a 1000 ft. sq. house. To visualize the following we have aggregated the sale price instead of computing the mean and produced the following plot.

We can also see (Figure 2) that the aggregated sale price for every built year houses increases exponentially as we approach the current year.Also, in blue we show the number of houses that were built that year. This shows us that most of the houses sold were built from 1980 and above. As we decrease in time, less houses are being sold and for less money.