House Price Prediction

Aditya Sai Nallapaneni

06/28/2022

## Abstract:

The main goal of this project is to perform data analysis and prediction of house using a set of features. The data set is taken from Kaggle which is a public website for machine learning data sets. For data analysis, scatter plots, correlation tests using pearson method, ANOVA, box plots and linear regression are used to identify the effects of different set of features on the target variable. The algorithm being used is linear regression. We have obtained an R square value of 79% which means our model was able to explain 79% variability in our target variable. It seeks to answer the question which variable is the most important to predict the house prices. This model will help the real estate agents and the general public to get an estimate about the prices of a house. We could deploy this model to production in this case.

## Introduction:

When you ask a home buyer to describe their ideal home, they are unlikely to start with the basement ceiling height or the closeness to an east-west railroad. However, the data from this playground competition shows that there are much more factors that impact price negotiations than the number of beds or the presence of a white-picket fence. This research focuses on the forecasting the ultimate price of each property using 79 explanatory factors that describe (nearly) every characteristic of residential dwellings in Ames, Iowa.

## Literature Review:

In a recent study by Ravikumar (S, 2018)[1], they implemented machine learning algorithms and data mining techniques on predicting house prices. By using R studio and WEKA graphical interface, ML algorithms like Support vector machine, random Forest, Multiple linear regression, and neural networks are addressed. In his study, they found that among these algorithms random forest and gradient boosted trees performed better with high accuracy and with fewer error values. Moreover, they found few attributes that contributed most to the prediction analysis. Evaluation of the accuracy is done by using mean absolute error which tells the difference between actual and predicted prices. The author suggested that it would be more helpful if there are more attributes like surroundings, marketplaces, and many related variables to the houses. In addition, I’ll perform linear regression and decision tree models and find the most common attributes that helps to predict the model.

According to the study of (Thamarai D. M., 2020)[2], he addressed two machine learning algorithms to predict house prices. As he said these models are implemented to help customers to purchase a house suitable for their needs. The proposed theory tells us the most common attributes or features of the house are likely to be the Age of the house, traveling facility from XYZ location, school facility, and shopping malls availability nearby the house. He implemented these algorithms using sci-kit- learn. As a result, by comparing two algorithms Multiple linear regression is found to perform better in predicting the house prices. However, the author suggested including more features and performing more advanced techniques for predicting house prices. In addition, I do data visualization using scatter plots and histograms, and comparison of linear regression algorithm and decision tree regression is applied which helps real estate agents to predict the prices based on important attributes.

## Research Question:

The main goal of this project is to build a prediction model to predict the prices of a house based on the input features that are present in the data set. Secondly we will perform statistical tests to see which factors affects the prices of a house.

## Theory:

H1: Regression analysis can be used to highlight the relationship between year built and sale price of a house. It will tell us whether the house become old, and its value reduced or not.

H2: An evaluation of the correlation between garage area, lot area, overall condition of the house, and year built can be helpful for identifying highly correlated pairs.

## Dataset:

This project is based on the dataset collected about the real estate prices of a house from publically available dataset <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data>. The dataset consists of around 79 columns but we will subset the data to contain a few columns of interest in this case. Following variables will be used during this project instead of using all available variables: • SalePrice - the property’s sale price in dollars. This is the target variable that you’re trying to predict. • MSSubClass: The building class • LotArea: Lot size in square feet • LandContour: Flatness of the property • OverallQual: Overall material and finish quality • OverallCond: Overall condition rating • YearBuilt: Original construction date • RoofStyle: Type of roof • ExterCond: Present condition of the material on the exterior • TotalBsmtSF: Total square feet of basement area • Heating: Type of heating • CentralAir: Central air conditioning • Bedroom: Number of bedrooms above basement level • Kitchen: Number of kitchens • KitchenQual: Kitchen quality • TotRmsAbvGrd: Total rooms above grade (does not include bathrooms) • GarageArea: Size of garage in square feet • GarageQual: Garage quality • PoolArea: Pool area in square feet • PoolQC: Pool quality • SaleType: Type of sale • SaleCondition: Condition of sale

Data cleaning steps are applied on the dataset which includes creating a subset of data to have selected columns as proposed in proposal and replacing the null values in categorical columns with mode of that column. After cleaning below is the head of our data:

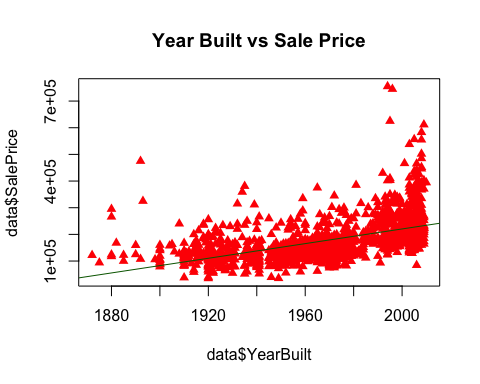
## MSSubClass LotArea LandContour OverallQual OverallCond YearBuilt RoofStyle  
## 1 60 8450 Lvl 7 5 2003 Gable  
## 2 20 9600 Lvl 6 8 1976 Gable  
## 3 60 11250 Lvl 7 5 2001 Gable  
## 4 70 9550 Lvl 7 5 1915 Gable  
## 5 60 14260 Lvl 8 5 2000 Gable  
## 6 50 14115 Lvl 5 5 1993 Gable  
## ExterCond TotalBsmtSF Heating CentralAir KitchenQual TotRmsAbvGrd GarageArea  
## 1 TA 856 GasA Y Gd 8 548  
## 2 TA 1262 GasA Y TA 6 460  
## 3 TA 920 GasA Y Gd 6 608  
## 4 TA 756 GasA Y Gd 7 642  
## 5 TA 1145 GasA Y Gd 9 836  
## 6 TA 796 GasA Y TA 5 480  
## GarageQual PoolArea SaleType SaleCondition SalePrice  
## 1 TA 0 WD Normal 208500  
## 2 TA 0 WD Normal 181500  
## 3 TA 0 WD Normal 223500  
## 4 TA 0 WD Abnorml 140000  
## 5 TA 0 WD Normal 250000  
## 6 TA 0 WD Normal 143000

## Methodology:

For statistical analysis and prediction, we have used ANOVA, visualizations such as box plots, correlation tests and regression model. Apart from this for the prediction of house prices, we will apply multiple linear regression and decision tree regression to compare the performances of both models based on evaluation metrics such as root mean squared error, mean absolute error and so on. Extensive data cleaning is performed on the data set which includes subseting the data to have columns of interest, removing unnecessary columns, replacing null values in categorical columns with mode and null values in numerical columns with mean value of that particular column.

## Results:

Starting with the scatter plot created below. It can be seen that the trend line fitted is a linear trend line however the relationship between both variables is not perfectly linear. There do exist some positive relationship between both variables which tells us that as the year built increases which means as the age of house reduces, the prices of house increases and as the age of house increases, the prices of house decreases.



Below are a couple of correlation tests applied on different variables with respect to target variable sales price. Pearson method [3] is used. The null and alternate hypothesis are attached below:

Null Hypothesis: Each pair for which correlation test is applied, they are not correlated.

Alternate Hypothesis: Each pair for which correlation test is applied, they are correlated.

As the p values for all 4 correlation tests is below the significance level alpha = 0.05, we can reject the null hypothesis and conclude that each pair for which correlation test is applied, they are correlated. The highest positive correlation exists between year built and sales price with a correlation coefficient of 0.5 approximately and garage area and sales price with a correlation coefficient of 0.6 approximately.

##   
## Pearson's product-moment correlation  
##   
## data: data$YearBuilt and data$SalePrice  
## t = 23.424, df = 1458, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.4845947 0.5591987  
## sample estimates:  
## cor   
## 0.5228973

##   
## Pearson's product-moment correlation  
##   
## data: data$GarageArea and data$SalePrice  
## t = 30.446, df = 1458, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.5910324 0.6538222  
## sample estimates:  
## cor   
## 0.6234314

##   
## Pearson's product-moment correlation  
##   
## data: data$OverallCond and data$SalePrice  
## t = -2.9819, df = 1458, p-value = 0.002912  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.12864437 -0.02666008  
## sample estimates:  
## cor   
## -0.07785589

##   
## Pearson's product-moment correlation  
##   
## data: data$LotArea and data$SalePrice  
## t = 10.445, df = 1458, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2154574 0.3109369  
## sample estimates:  
## cor   
## 0.2638434

Moving forward, in order to see whether sales type affects the sales price, one way ANOVA test [4] is applied. The null and alternate hypothesis are attached below:

Null Hypothesis: Sales type has no effect on sales price.

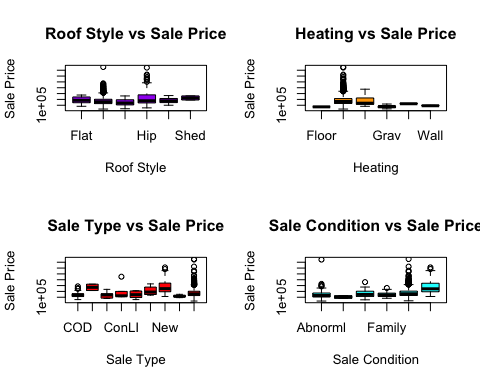
Alternate Hypothesis: Sales type affects the sales price.

As the p value of one way anova [4] is less than significance level 0.05, we can reject the null hypothesis and conclude that sales type affects the sales price.

## Df Sum Sq Mean Sq F value Pr(>F)   
## data$SaleType 8 1.264e+12 1.580e+11 28.86 <2e-16 \*\*\*  
## Residuals 1451 7.944e+12 5.475e+09   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

To further explore the data, 4 different box plots have been created with respect to the target variable. Following conclusions could be drawn from these box plots:

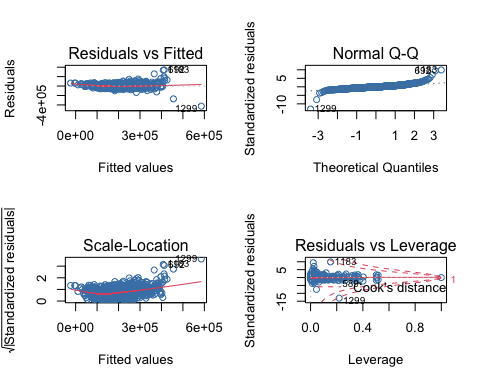
* The average sales price for houses with roof type gambrel is lowest and highest for roof type Shed.
* The average sales price for house with GasA heating has highest while it is lowest for Grav.
* The average sales price for sale type Con is highest while it is lowest for Oth.
* The average sales price for sale condition partial is highest and lowest for AdjLand.



Finally, a regression model [5] is built to predict the house prices and the summary of regression model is attached below. It can be seen that we have several variables where the p value is less than significance level alpha = 0.05 and they are significant. The highest positive effect amongst significant variables is of Overall condition and sales type set to new. The R square of the model is 79% which means the model was able to explain 79% variability in the data set which is good in this case. The p value of overall model is also less than significance level alpha = 0.05, which makes our model significant.

##   
## Call:  
## lm(formula = SalePrice ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -425747 -18504 -2168 14573 342070   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.681e+05 1.148e+05 -5.819 7.33e-09 \*\*\*  
## MSSubClass -7.060e+01 2.481e+01 -2.845 0.004499 \*\*   
## LotArea 6.988e-01 1.124e-01 6.219 6.58e-10 \*\*\*  
## LandContourHLS 2.938e+04 7.229e+03 4.064 5.09e-05 \*\*\*  
## LandContourLow 3.090e+04 8.507e+03 3.633 0.000291 \*\*\*  
## LandContourLvl 1.480e+04 5.076e+03 2.915 0.003607 \*\*   
## OverallQual 2.059e+04 1.206e+03 17.077 < 2e-16 \*\*\*  
## OverallCond 5.073e+03 1.113e+03 4.558 5.62e-06 \*\*\*  
## YearBuilt 3.183e+02 5.389e+01 5.906 4.39e-09 \*\*\*  
## RoofStyleGable -1.308e+04 1.100e+04 -1.189 0.234676   
## RoofStyleGambrel -7.913e+03 1.588e+04 -0.498 0.618450   
## RoofStyleHip -6.633e+03 1.119e+04 -0.593 0.553451   
## RoofStyleMansard -8.930e+03 1.773e+04 -0.504 0.614641   
## RoofStyleShed -7.151e+03 2.828e+04 -0.253 0.800445   
## ExterCondFa 5.372e+03 2.328e+04 0.231 0.817567   
## ExterCondGd 4.221e+03 2.205e+04 0.191 0.848202   
## ExterCondPo 2.540e+04 4.402e+04 0.577 0.564109   
## ExterCondTA 2.521e+03 2.200e+04 0.115 0.908794   
## TotalBsmtSF 2.452e+01 3.054e+00 8.030 2.03e-15 \*\*\*  
## HeatingGasA -2.798e+04 3.744e+04 -0.747 0.454944   
## HeatingGasW -2.774e+04 3.840e+04 -0.723 0.470058   
## HeatingGrav -2.348e+04 3.998e+04 -0.587 0.557108   
## HeatingOthW -5.417e+04 4.582e+04 -1.182 0.237336   
## HeatingWall -3.730e+03 4.144e+04 -0.090 0.928301   
## CentralAirY 1.734e+03 5.048e+03 0.343 0.731316   
## KitchenQualFa -4.569e+04 8.552e+03 -5.343 1.06e-07 \*\*\*  
## KitchenQualGd -4.624e+04 4.451e+03 -10.389 < 2e-16 \*\*\*  
## KitchenQualTA -5.804e+04 5.088e+03 -11.408 < 2e-16 \*\*\*  
## TotRmsAbvGrd 1.072e+04 7.129e+02 15.032 < 2e-16 \*\*\*  
## GarageArea 4.550e+01 6.111e+00 7.446 1.67e-13 \*\*\*  
## PoolArea 1.036e+01 2.491e+01 0.416 0.677521   
## SaleTypeCon 5.804e+04 2.689e+04 2.158 0.031074 \*   
## SaleTypeConLD 1.033e+04 1.416e+04 0.730 0.465662   
## SaleTypeConLI 4.075e+03 1.761e+04 0.231 0.817000   
## SaleTypeConLw 2.144e+04 1.795e+04 1.195 0.232405   
## SaleTypeCWD 3.352e+04 1.972e+04 1.700 0.089383 .   
## SaleTypeNew 6.230e+04 2.321e+04 2.684 0.007352 \*\*   
## SaleTypeOth 2.807e+04 2.231e+04 1.258 0.208488   
## SaleTypeWD 1.023e+04 6.160e+03 1.660 0.097084 .   
## SaleConditionAdjLand 3.641e+03 1.950e+04 0.187 0.851950   
## SaleConditionAlloca 1.482e+04 1.177e+04 1.260 0.207908   
## SaleConditionFamily -1.642e+04 9.297e+03 -1.767 0.077500 .   
## SaleConditionNormal 6.059e+03 4.216e+03 1.437 0.150823   
## SaleConditionPartial -3.679e+04 2.245e+04 -1.639 0.101535   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 36900 on 1416 degrees of freedom  
## Multiple R-squared: 0.7906, Adjusted R-squared: 0.7843   
## F-statistic: 124.3 on 43 and 1416 DF, p-value: < 2.2e-16

The diagnostic plots are attached below. The residuals vs fitted plot shows a linear trend and pattern which tells us that our model is a better fitted model. The residuals are not normalized as the values deviates at tails of 45 degree line drawn.

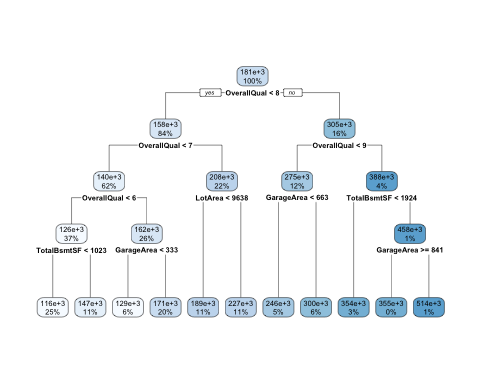


Finally the RMSE of regression on the data set is attached below and it is 36338.3.

Next a regression tree is trained and the summary of tree is attached. We can see that the value of n in this case is 1460 and 31 nodes are used to construct the regression tree in this case.

## n= 1460   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 1460 9.207911e+12 180921.2   
## 2) OverallQual< 7.5 1231 2.987549e+12 157832.4   
## 4) OverallQual< 6.5 912 1.287323e+12 140384.0   
## 8) OverallQual< 5.5 538 5.160363e+11 125633.2   
## 16) TotalBsmtSF< 1023 372 2.889255e+11 116053.0 \*  
## 17) TotalBsmtSF>=1023 166 1.164577e+11 147102.0 \*  
## 9) OverallQual>=5.5 374 4.858330e+11 161603.0   
## 18) GarageArea< 333 87 7.308988e+10 129076.5 \*  
## 19) GarageArea>=333 287 2.927973e+11 171463.0 \*  
## 5) OverallQual>=6.5 319 6.287649e+11 207716.4   
## 10) LotArea< 9637.5 161 1.593479e+11 189083.7 \*  
## 11) LotArea>=9637.5 158 3.565642e+11 226702.9 \*  
## 3) OverallQual>=7.5 229 2.036506e+12 305035.9   
## 6) OverallQual< 8.5 168 6.818726e+11 274735.5   
## 12) GarageArea< 662.5 78 2.438476e+11 245578.7 \*  
## 13) GarageArea>=662.5 90 3.142478e+11 300004.8 \*  
## 7) OverallQual>=8.5 61 7.755905e+11 388486.1   
## 14) TotalBsmtSF< 1924 41 1.702665e+11 354347.0 \*  
## 15) TotalBsmtSF>=1924 20 4.595807e+11 458471.3   
## 30) GarageArea>=841 7 1.048383e+11 355099.7 \*  
## 31) GarageArea< 841 13 2.396657e+11 514132.9 \*

A plot of regression tree created in this case is attached below. The RMSE of the regression tree on the data set is 40205.37.



Based on the RMSE value, we can see that the root mean square error of the regression model is lower as compared to decision tree regression. Hence we can say that the accurate model is the regression model.

## Implications:

More and more findings could be explored as we have used a subset of data for this project. The original data set has 79 different data points which could be explored and other factors which highly contribute to the sales price of a house. The study could further be taken one step forward by enhancing this machine learning model and deploying it to a website where the general public can input certain parameters and the model gives an estimation of price of house.

## Conclusion:

Concluding this project, several key points have been identified. A multiple regression model is built with an R square value of 79%. We saw that age of house, condition, roofing, heating matters in determining the price of a house when it comes to sales. Even sales type has an affect on the price, the garage area, the lot area and the basement square foot affects the prices positively which means increase in those values increases the house prices as well.

## References:

1. Ravikumar, A. S. (n.d.).S, R. K. (2018). Real Estate Price Prediction Using Machine Learning. School of Computing National College of Ireland.

2.hamarai, D. M. (2020). House Price Prediction Modeling Using Machine Learning. I.J. Information Engineering and Electronic Business, 2020, , 15-20.

1. Coefficient, I., (Hand), H., Excel, F., SPSS, P., & Minitab, H. (2022). Correlation Coefficient: Simple Definition, Formula, Easy Steps. Statistics How To. Retrieved 7 June 2022, from <https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/>.
2. One-Way ANOVA Test in R - Easy Guides - Wiki - STHDA. Sthda.com. (2022). Retrieved 7 June 2022, from <http://www.sthda.com/english/wiki/one-way-anova-test-in-r>.
3. Base, K., & (Examples), M. (2022). Multiple Linear Regression | A Quick Guide (Examples). Scribbr. Retrieved 7 June 2022, from <https://www.scribbr.com/statistics/multiple-linear-regression/>.