Springboard Capstone Project

AMES, IOWA House Sales Price Prediction

Abstract

Data set with different features of houses in AMES, IOWA has been used to demonstrate the learnings from the course. Data has been cleaned, wrangled and different techniques of machine learning applied to predict housing prices.

Rajkumar, Pavithra

Pavithra.rajkumar@gmail.com

Table of Contents

OBJECTIVE

Online real estate marketplaces have become increasingly popular and has become mainstream. In the real estate business, there is an evident shift of power, from the traditional real estate companies along with the traditional model of real estate agents, brokers, to the hands of the sellers and buyers. Different neighborhoods in the US have different factors or variables, that influence the price of houses. All the online real estate firms that exist now, provide an estimate of the housing sales prices. In a very busy housing market, the predicted sales prices, needs constant analysis of the factors that influence the prices. The competitive advantage of the firm will be the best pricing on the sale price of a home so that a buyer can provide the correct offer thus enabling a closure on the deal. On the other hand, the sellers should be able to list the home for the best price to go on the market. Brokers / real estate agents will be able to list more homes, convincing the sellers on the best time to sell and the buyers on the best price to make an offer.

The Objective is to predict prices for a home accurately and have a very low margin of error between the predicted sale price and the actual sales price. The plan is to use different advanced regression techniques and machine learning in Python. We will be use using the statsmodel, scikit, numpy, pandas package to analyze the existing data set to determine prices for houses in AMES, IOWA.

Data set

The AMES Housing Data set from https://www.kaggle.com/c/house-prices-advanced-regression-techniques will be used for the data analysis. Two sets of datasets used are in csv format for training and testing respectively. 'train.csv' has 1460 observations with 80 variables for which the sales price has been provided so the prediction model can be built. 'test.csv' also has 1459 observations and was used to test the model built.

The AMES housing data [1] set has 14 discrete variables, # of kitchens, bedrooms, and bathrooms (full and half) located in the basement and above grade (ground) living areas of the home. Continuous variables are about 20 and are lot size, total dwelling square footage, area of basement, area of living area and so on. There are 23 categorial variables of each of nominal and ordinal data associated with this set. Nominal variables typically identify various types of dwellings, garages, materials and environmental conditions. Ordinal variables will be items with the property like Lot Shape, Utilities.

Data Wrangling

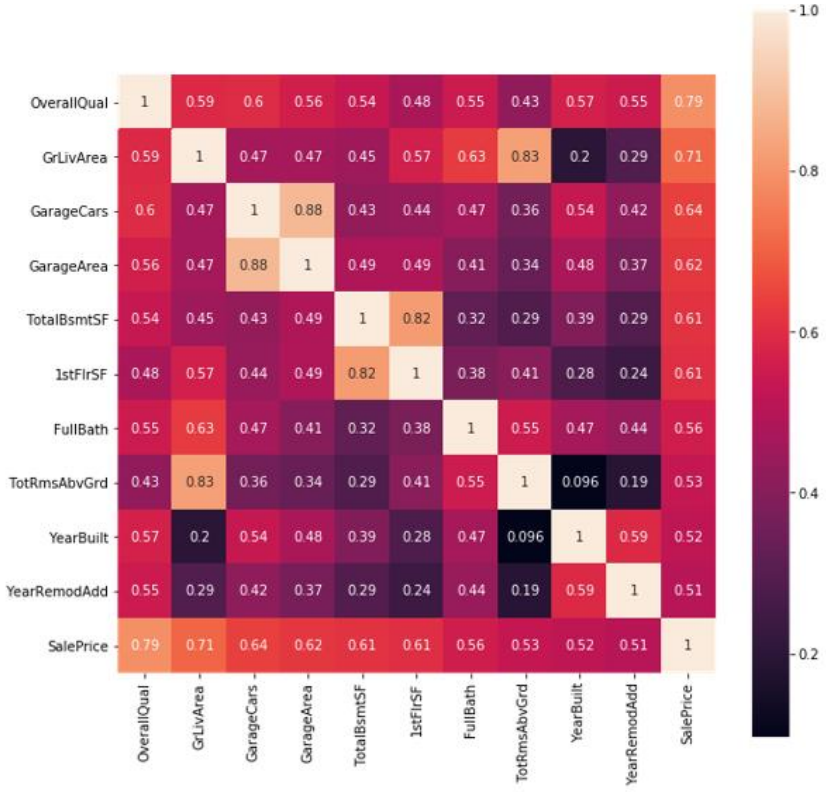
Firstly, the data will be cleaned up and missing data identified along with the outliers. The data will be analyzed for the variables which are exploratory and response variables. The data analysis will also show visualizations of the various regression models, very importantly advanced regression (machine learning), with Exploratory Data Analysis (EDA) techniques. Inference from the data will be made to highlight which variables are highly co-related.

[1] http://www.amstat.org/publications/jse/v19n3/decock/DataDocumentation.txt

Data Wrangling: Visualization /Correlation of variables

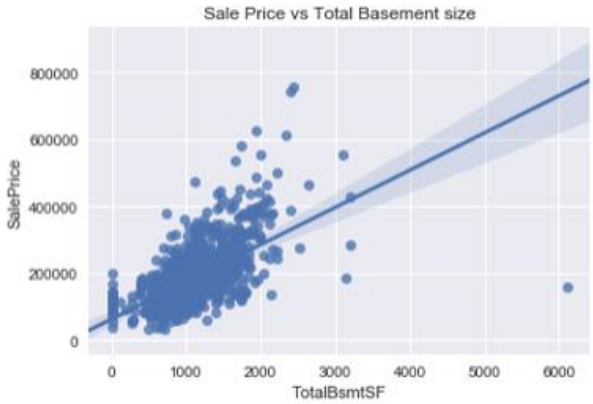
We see that the OverallQual and GroundFloor Living Area has a good correlation to the Sales Price. So we can make below a heatmap to show the correlation matrix for all the variables which are above 50%

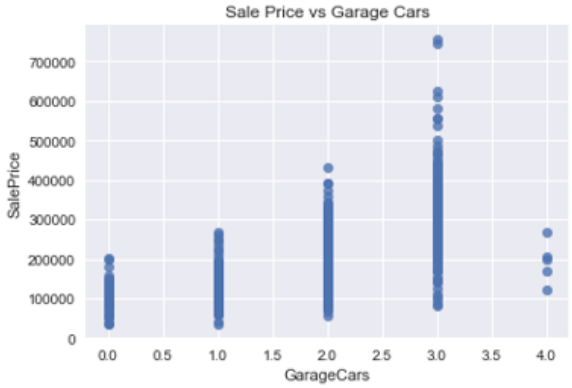
Other variables which showed > 50% correlation with sales price were: OverallQual GrLivArea GarageCars GarageArea TotalBsmtSF 1stFlrSF FullBath TotRmsAbvGrd YearBuilt



Data Wrangling: Outliers

We now check into dropping the Outliers by analyzing the scatter plots since they are used in outlier detection. For outliers, we consider anything more than -1.5 and 1.5 Interquartile Range, three or more standard deviations from the mean as outliers. Also, any data that are isolated points from the trend line can be dropped as well.





Four outliers in the graphs above will be dropped. These corresponding to -Ground living area > 4500 -Year Built graph sales price between 7000 and 8000 -The isolated TotalBsmtSF point > 6000 is also included as part of these outliers which will be removed

Data Wrangling: Missing Values

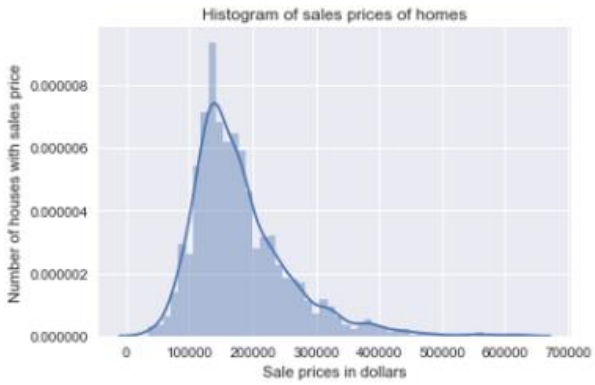
From checking the training data set, we can see that it missing more than 90 % of data the following columns 'PoolQC', 'MiscFeature', 'Alley', 'FirePlaceQC' are being excluded from our analysis. Other variables with Basement and Garage seem to have less missing values, we need to check if these variables have NaN since the houses have that feature missing. In that case we need to have that value NaN as 'None'.

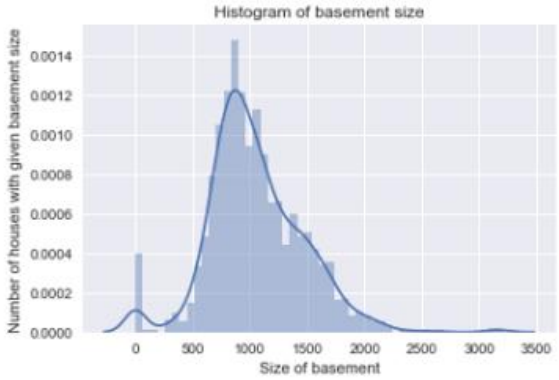
We can group all variables with the NaN values. Variables with similar categories have been grouped for comparison e.g. Basement, MasVnr, Garage etc. Non-categorical variables have also been included (e.g Pool Area, MasVnrArea) to help decide if the NaN in categorical variable of the same category need to be set to None or dropped. If PoolQC is NaN and PoolArea is 0 we can see that NaN will be that there is no pool and should be set to None.

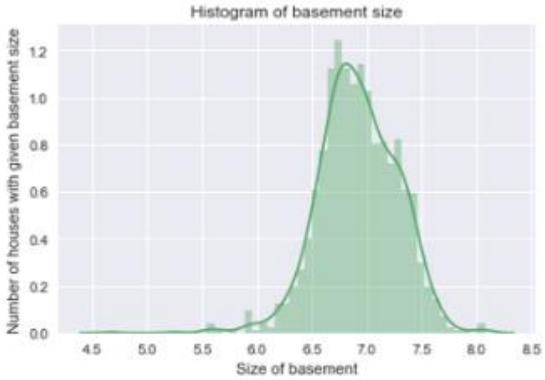
So for the categorical variables, it shows we can check for the NaN values. For Garage columns: where there are NaN values, they are in all garage variables suggesting the absence of garages for those houses. So, for Pool - PoolQC is set to NaN corresponding to PoolArea 0, which means the absence of Pool. This should be set to 'None' Basement, Alley, Fence, Miscfeatures has NaN values and means absence of this feature in the house and should be set to 'None'. In case of the Electrical variable, when we find NaN value setting it to 'None' does not make sense as it is a feature that all houses should have. Since it is a small portion say 0.1% we can remove the rows with that value.

Data Wrangling: Normality & Linearity

We need to check if the data set has data that follows a normal distribution and is linear so that we can decide on how the model can be built.







We see that the distributions above are skewed (on the left), we would need to have the data standardized so individual features look like standard normally distributed data: Gaussian with zero mean and unit variance. One way of doing this is by removing the mean value of each feature, then scale it by dividing non-constant features by their standard deviation. Another way is to do a log transformation can be applied although it can't be applied to zero or negative values. Our second histogram (on the right) above shows some zero values for basement size which would not be suitable for log transformation unless they are removed.

Scatter plots after the log transformation

We can see that the dense clutter in the scatter plots are now shifted towards the center following log transformation. As a result, we will have the data exhibit less heteroskedasticity (absence of the conical shape like in the previous plots).

Model building: Advanced Regression analysis

Model was built using the following regression machine learning algorithm

1. Ridge Regression
2. Lasso Regression
3. Support Vector Machines
4. Gradient Boosting Regression

The independent variables that are used for the model to predict the dependent variable SalesPrice of a home are the following

OverallQual GrLivArea GarageCars TotalBsmtSF FullBath YearBuilt

Result Summary for R2 Score			
Ridge Regression	Lasso Regression	Support Vector Machine	Gradient Boosting Regression
0.825067	0.826482	0.826482	0.834250

We see that the Gradient Boosting Regression has a good R2 value and for the test set, we will use that Regression model to predict the Sale Price of homes. All the home prices are in the attached spreadsheet

https://github.com/pavithraraj/AMESHousePricePrediction/blob/master/submission_predict.csv

Github Project code

- https://github.com/pavithraraj/AMESHousePricePrediction/blob/master/A_Capstone_test.ipynb

Potential Data set to use

All the firms that compete in the market, provide home prices, but do not provide any custom option for a buyer. By analyzing various factors that influence home prices, we can provide more options for a buyer t

References

http://ww2.amstat.org/publications/jse/v19n3/decock.pdf

http://www.amstat.org/publications/jse/v19n3/decock/DataDocumentation.txt