House Price Prediction: Advanced Regression Techniques

Madhuri Konnur
Machine Learning Intern
AI Technology and Systems
magicwordsmadhu@gmail.com
www.ai-techsystems.com

***Abstract* - In today's world, real estate market is one of the most competitive field where pricing tends to vary significantly based on lot of factors. Hence it becomes good candidate for applying the machine learning techniques to optimize and predict the prices with high accuracy. This is an effort to apply important features to use while predicting housing prices with good accuracy. In below section described regression models, using various features to have lower error. In a regression model some feature engineering is required for better prediction. Often a set of features (multiple regressions) is used for making a better model fit. These models are expected to be susceptible towards over fitting ridge regression which is used to reduce it. This use-case is one of the techniques of regression models using Kaggle House Price Dataset.**

Keywords – Machine- Learning, Linear Regression, RandomForest, GradientBoostingRegressor

I. INTRODUCTION

As house is the basic need for human being. But predicting the housing prices is always challenge. Investing in the real estate usually seems to be profitable and return on investment is always appreciable. Machine Learning engineers and researchers are trying to come up with model/s. Which can predict house price more accurately with high accuracy and least error. While creating models, various features and attributes needs to be considered. Such as built up area, total area, ventilation, bedrooms and property age etc. Model to model based on engineers feature weightage method price predictions may vary. Every single feature in our model is given certain weightage and it determines how important is that feature towards our model prediction. This is what defines feature engineering. Most of the companies such as "magicbricks.com", "Zillow.com" majorly deal with real estate business. There might be probability of billions of different features to choose from however one of the drawbacks of having a large number of features involved is that, it requires heavy computations. These real estate companies often tend to have large dataset of houses. whose prices they predict using machine learning techniques. One of the techniques they use is regression to learn the nature of models from the previous results (houses which were sold off previously which are used as training data). In this use case defined linear model, RandomForest, Gradient Boosting using several features as its input and predicting the Sales Price of house.

II. RELATED WORK

From few decades demands of houses are increasing. And house prices are based on several factors. Due to rise in the demand of houses, proper house price prediction model is need. Which needs to be unbiased. Here we are considering Kaggle House price data set. And prediction label is SalesPrice. Here notable work effort is feature extraction. which used visual features to predict the housing prices. To build predictive model real world basic knowledge on real estate is required.

From Kaggle.com obtained housing price scraped data. First pre-processed data by removing outliers and missing values. By using regression techniques Lasso, Ridge etc. On the same data set used various machine learning algorithms such as Linear Regression, Random Forest and Gradient Boosting to predict house sales price.

Based on correlation of features and real estate world basic idea, divided 2 sets of high influence features. And above-mentioned machine learning algorithms compared root-mean-squared errors (RMSE) obtained.

III. METHODOLOGY

Used several models and calculated root mean squared error for each. Visualized by plotting a graph. These techniques are applied on 2 sets of features (Multivariate) and attempted for decent house sale price prediction.

A. Data Collection

Kaggle.com- House price dataset consists of around 3000 records with 80 parameters /variables. These variables influence on property sales price. Considering some feature parameters such as Overall quality which rates the overall condition and finishing of the house, Location, year of built, Area in square meters, Bedrooms, year of built, quality of house, Bath rooms and age of property etc. Predicting SalesPrice using regression(multinomial). They are majorly two types of values for parameters Numerical and Categorical. Listing few parameters:

Parameters	Description	Datatype
SalePrice	The property's Sale Price in dollars.	Numerical
MSSubClass	The building class	Categorical
LotArea	Lot size in square feet	Numerical
Street	Type of road access	Categorical
GrLivArea	Above grade living area square feet	Numerical
GarageCars	Size of Garage in car capacity	Numerical
YrSold	Year Sold	Numerical
BldgType	Type of dwelling	Categorical

RoofStyle	Type of roof	Categorical
PoolArea	Pool area in square feet	Numerical

B. Data Preprocessing

Important step for converting raw data into understandable form, which will help further analysis of data. This process involves finding a missing values and redundant data. In this House price dataset scope of missing value is high rather than r redundant data.

Specific to this dataset NaN values are present in numerical as well categorical columns. Now filling these missing values with some sensible values. The NaN values in the numerical columns were replaced with the median of the values of the particular column whereas the NaN values in the categorical columns were replaced by simply a string 'None' so that each 'None' value can be encoded for further processing.

Once replacing a missing value is done. next needs convert categorical data into numeric data for easy analysis. This was done using LabelEncoder of scikit-learn package. In the testing dataset, there were some values which were not seen in the training dataset. So, the LabelEncoder was first fitted to combined dataset consisting of all the training and testing categorical samples and then the training and testing datasets were transformed.

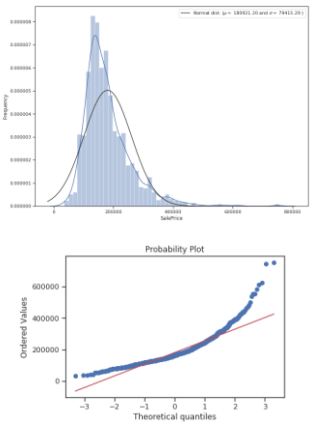
After the transformation. OneHotEncoding was done on the training and testing datasets. A one hot encoding is a representation of categorical variables as binary vectors. This was done to ensure that none of the categorical values that were converted to numerical values have a higher priority.

C. Data Analysis

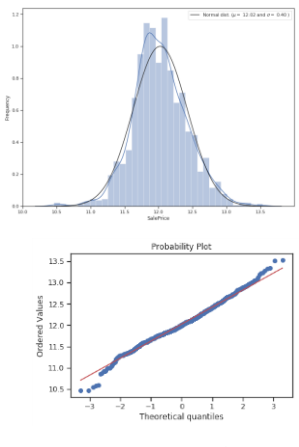
Before applying any model to our dataset, we need to find out characteristics of our dataset. Thus, we need to analyze our dataset and study the different parameters and relationship between these parameters. We can also find out the outliers present in our dataset. Outliers occur due to some kind of experimental errors and they need to be excluded from the dataset.

First, let's take a look at the response variable "Sale Price". It's positively skewed; most houses sold for between $100,000 and $250,000, but some sold for substantially more.

To maximize the performance of our model, we want to normalize our features and response variable. By applying a log transformation, Sale Price now resembles a normal distribution



SalePrice without Log transformation

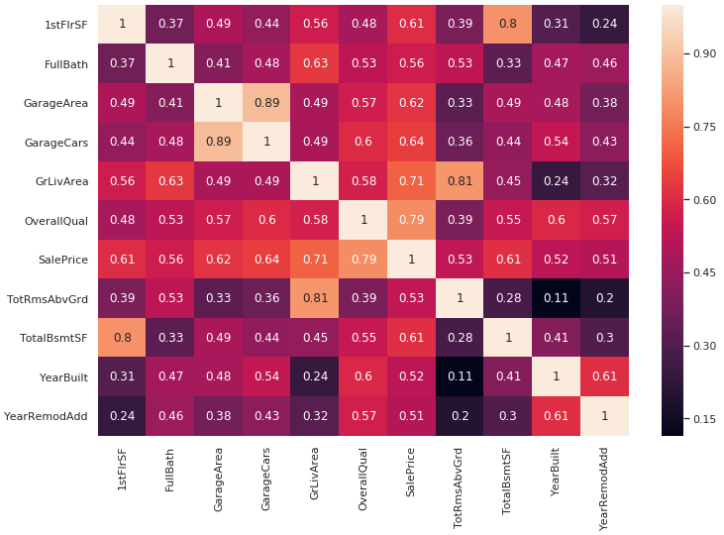


SalePrice with Log Transformation

Around 80 features represent house property, like number of bathrooms, basement square footage, year built, garage square footage, etc. Let's verify using how each is correlated and responses to Sale Price by using Heatmap.

[Correlation exists between +1 to -1. positive number represents a positive correlation between two variables and vice versa.]

.

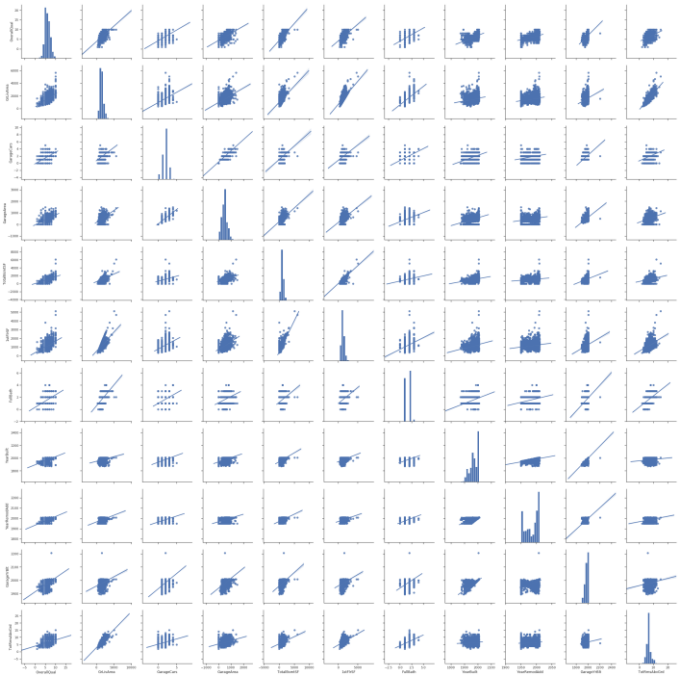


Heatmap for maximum correlating parameters with respect to 'SalePrice'

This gives us information about the feature importance in predicting the Sale Price and indicates where there may be multicolinearity. The overall quality of the home "OverallQual" is highly correlated with Sale Price, not surprisingly. In contrast, the year the home was sold "YrSold" has little correlation with the Sale Price.

From the heatmap, we can deduce that 'GrLivArea', 'TotalBsmtSF' and 'OverallQual' are the most correlated with 'SalePrice'.

We can analyze the columns that are most correlated to 'SalePrice' by using a Pair plot. A Pairplot allows us to see both distribution of single variables and relationships between two variables.'

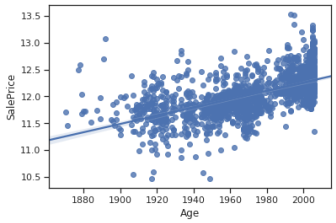


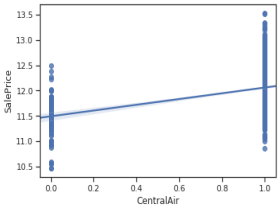
Pair plot

For Multinomial /Multivariance coming up with two sets of features

Referring to heatmap
SET 1 feature - ['OverallQual', 'GrLivArea', 'GarageCars', 'GarageArea', 'TotalBsmtSF', '1stFlrSF', 'FullBath', 'YearBuilt','YearRemodAdd', 'GarageYrBlt', 'TotRmsAbvGrd']

Referring to real estate world let's check 2 more variables which influences SalePrice, property age and CentralAir
.





Age increases, sales price decreases

Presence of CentralAir increases SalesPrice

(note: here higher no . of yr means newer home)

SET 2 feature - ['OverallQual', 'OverallCond', 'YearBuilt', 'CentralAir', '1stFlrSF', '2ndFlrSF', 'BedroomAbvGr' ,'YrSold']

Age = YearBuilt - YrSold

D. Training

Once we gained the insight of our dataset, we are ready to build our model. Since our label data is continuous in nature need to fit regression model. For prediction of house sales price.

For example: Simple Linear Regression - works on 1 feature vs label.

Equation looks like this:

$$F(x) = w0+w1x$$

Where: x=feature (1), F(x) =price w0=intercept term, w1=coefficient

Since we are building Multivariate *regression models* on SET 1 and SET 2
Dividing training data into train and test (validation) sets in order to test model and minimize root mean square error.

1. Multinomial /Multivariate Regression

In multivariate models instead of 1 feature, we use several features as based on many correlative factors.

$$F(x) = w0 + w1_1*X_1 + W_2*X_2 + \ldots + w_p*X_p$$

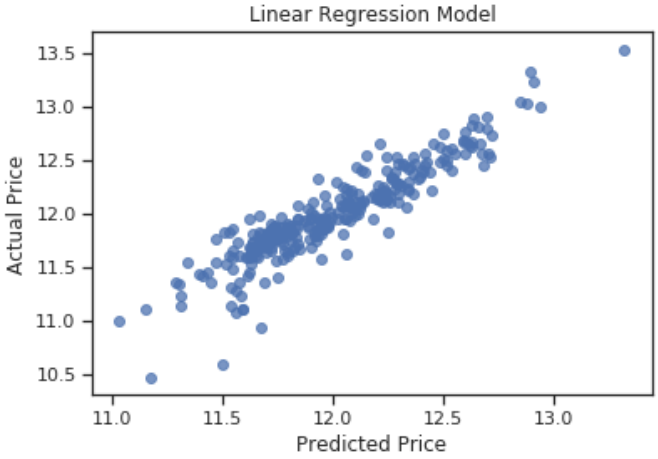
Where: X1,X2,…= features, F(x) =price w0=intercept term, W1,W2=coefficients of respective features

Linear regression

Model = Regression trained using SET 1 [['OverallQual', 'GrLivArea', 'GarageCars', 'GarageArea', 'TotalBsmtSF', '1stFlrSF', 'FullBath', 'YearBuilt','YearRemodAdd', 'GarageYrBlt', 'TotRmsAbvGrd']]

Coefficients for SET 1 **[** 9.39571055e-02,
2.12427198e-04,
6.47515137e-02,
2.84348701e-05,
6.59757013e-05,
6.60260293e-05,
-2.34754808e-02,
1.90012719e-03,
2.35768700e-03,
4.87134155e-05,
.03744290e-03**]**

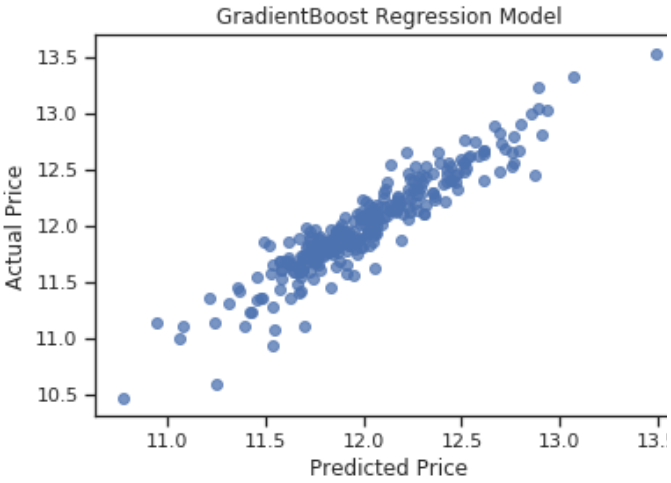
Intercept for SET 1 **[**2.323**]**



Relationship between predictions and actual_values

Performed predicting House Sale price on 2 sets of feature variables. Using Linear Regression (LR), Random Forest (RF) and Gradient Boost (GBR).

On Set 2 GBR model, predictions accuracy is 87.5% better than other models and RMSE is 0.02



GBR predictions and actual prices are comparatively closer

Gradient Boosting Regressors are trees that function on boosting. Boosting is a mechanism in which samples which were not fit well in a tree are given higher probability to be utilized in the next tree. In this way, the algorithm focuses on increasing accuracy of prediction on all samples sequentially. Boosting takes advantage of weak learners and perfects them one by one. Using Gradient Boosting Regressors, the cross-validation accuracy jumped up to 87.50% and the RMSE was down to 0.02.

IV. CONCLUSION

,

I have built model using 3 regression methods on 2 sets of features to predict the house prices. Mentioned step by step analysis on dataset. Key steps include assigning appropriate values for NAs, normalizing variables, optimizing hyperparameters for candidate models, and choosing the best model

Gradient Boosting Regression method gave 87.5% accuracy and low RMSE 0.02

For future work, other preprocessing techniques can be used and the training dataset can be clustered by taking into account features such as Neighborhood, Class to observe predictions.

V. REFERENCES

[1] Mansural Bhuiyan and Mohammad Al Hasan (2016) "Waiting to be Sold: Prediction of Time- Dependent House Selling Probability" IEEE *International Conference on Data Science and Advanced Analytics* pp468-477

[2] Sean Arietta, Alexa, A Efros, Ravi Ramamoorthi, and Maneesh Agarwal City Forensics: Using attributes to predict non visual Attributes visual elements for *prediction IEEE Transactions on Visualization and Computer Graphics* vol. 20(12) pp 2624-2633

[3] Neelam Shinde, Kiran Gawande, "Valuation Of House Prices using Predictive Techniques", Volume-5, Issue-6, Jun.2018

[4] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh, "A Hybrid Regression Technique for House Prices Prediction", December-2017

[5] Limsombunchai, Visit. "House price prediction: hedonic price model vs. artificial neural network."New Zealand Agricultural and Resource Economics Society Conference. 2004.