

House Price Prediction

Submitted by:

VARSHA THAKUR

**ACKNOWLEDGMENT**

REFERENCES 1. Michael J. Ball, “Recent Empirical Work on the Determinants of Relative House Prices,” in Urban Studies, 10, pp. 213-233, 1973.

2. Park, B., & Kwon Bae, J., “Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data,” in Expert Systems with Applications, vol. 42, issue 6, pp. 2928–2934, 2015.

3. Breiman, L., “Random forests,” in Machine Learning, vol. 45, issue 1, pp. 5–32, 2001.

4. Breiman, L et al., “Classification and regression trees,” New York: Chapman & Hall/CRC Press, 1984.

5. Ranadip Pal, “Overview of predictive modeling based on genomic characterizations,” in Predictive Modeling of Drug Sensitivity, 2017.

6. Li et al., “A SVR based forecasting approach for real estate price prediction” in International Conference on Machine Learning and Cybernetics, Hebei, 2009.

Abstract:

For socioeconomic development and the well-being of citizens, developing a precise model for predicting housing prices is always required. So that, a real estate broker or a house seller/buyer can get an intuition in making well-knowledgeable decisions from the model. In this work, a various set of machine learning algorithms such as Linear Regression, Decision Tree, Random Forest are being implemented to predict the housing prices using available datasets. The housing datasets of 506 samples and 13 feature variables from January 2015 to November 2019 were taken from the StatLib library which is maintained at Carnegie Mellon University. Since housing price is emphatically connected to different factors like location, area, the number of rooms; it requires all of this information to predict individual housing prices. This paper will apply both traditional and advanced machine learning approaches to investigate the difference among several advanced models to explore various impacts of features on prediction methods. This paper will also provide an optimistic result for housing price prediction by comprehensively validating multiple techniques in model execution on regression.

**INTRODUCTION**

A house is usually the single largest purchase an individual will make in their lifetime. Such significant purchase warrants being well-informed about what a house’s selling price should be; for the buyer, as well as the seller or real estate broker involved. The power of machine learning provides us with the tools we need to look at a large data set and spit out a predicted value, which was our main goal in this project. Using a dataset containing information on houses in Ames, Iowa, our team leveraged different machine learning techniques to predict sale prices based on both practical intuition and those observed through our exploratory data analysis and model fitting processes.

The general outline of the process was:

* Imputing Missing Values
* Exploratory Data Analysis
* Feature Engineering/Dimension Reduction
* Fixing Skewness and Outliers
* Modeling

**The Data Set:**

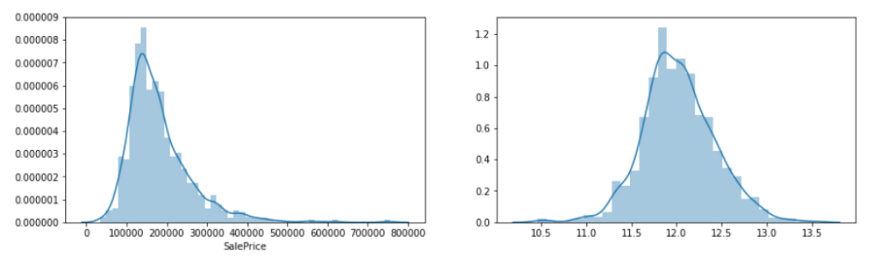
A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them on at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia.

The company is looking at prospective properties to buy to enter the market. You are required to build a regression model using regularisation in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

**Goal**  
You are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns.

**Exploratory Data Analysis:**

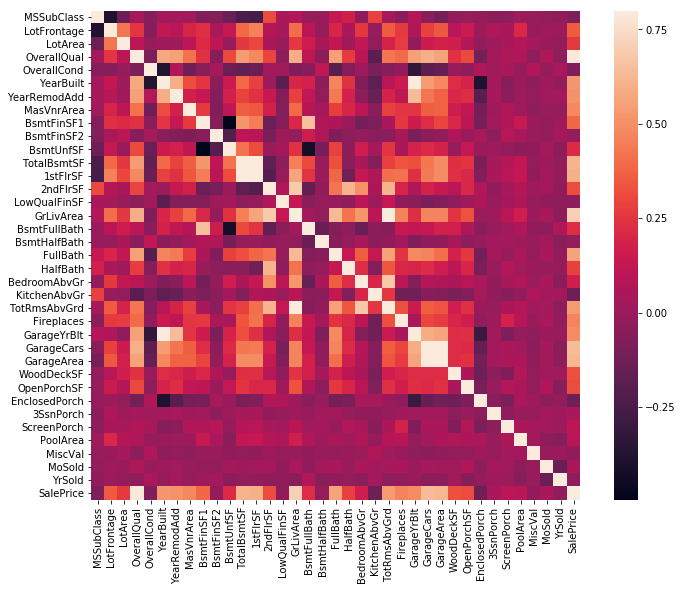
Let's start with the distribution of the target variable, sales price. There is a clear right-skew in the distribution of sale price with a vast majority of property values in the $100,000 — $200,000 range, but with a long tail stretching upward towards $800,000 (bottom left image). Using a Box-Cox transformation, we identified that taking the log of sale price transforms the distribution to approximately normal (bottom right image). This transformation increased the strength of linearity amongst most of our variables against sale price, validating the assumption of linearity in regression.



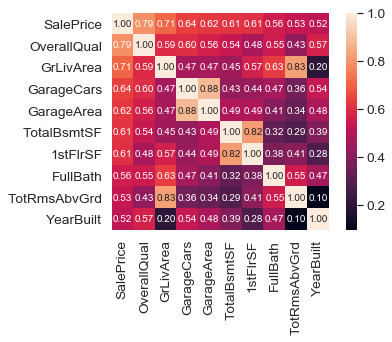
**Multi-Variable Analysis:**

* There are two types of features namely numerical and categorical.
* The categorical data is either ‘Yes’ and ‘No’,’Fare’ and ‘Near’
* The numerical data in numerical form. These features are in a linear relationship with each other. For example, a 2,000 square foot place is 2 times “bigger” than a 1,000 square foot place.

**Correlation Matrix:**



**Top 10 Correlated Variable — Correlation Matrix**

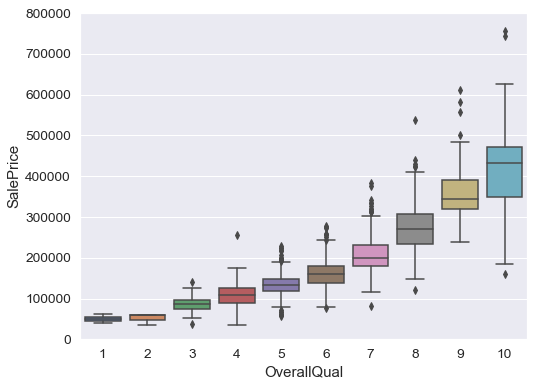


Top 10 Highly Correlated Matrix

The Definition against Top 9 correlated features provided below:

\* OverallQual: Rates the overall material and finish of the house (1 = Very Poor, 10 = Very Excellent)  
\* GrLivArea: Above grade (ground) living area square feet  
\* GarageCars: Size of garage in car capacity  
\* GarageArea: Size of garage in square feet  
\* TotalBsmtSF: Total square feet of the basement area  
\* 1stFlrSF: First Floor square feet  
\* FullBath: Full bathrooms above grade  
\* TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)  
\* YearBuilt: Original construction date

**Overall Quality Vs Sale Price**

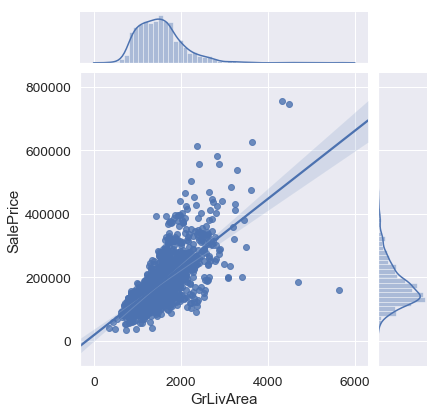


Overall Quality Vs SalePrice

**Observation**-

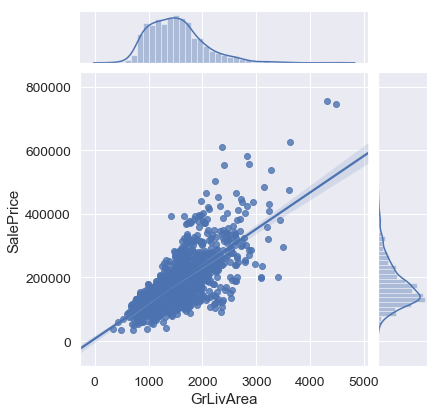
* The plot defines we can state that as OverallQual increases, the SalePrice also increases.
* The above assumption is true on normal assumption also, As OverallQual is more the SalePrice is also more.

**GrLivArea Vs SalePrice:**

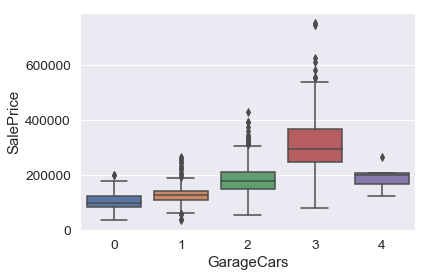


**Observation -**

* People pay more for more living area.
* In the above plot, there is a value which has the least cost for more living area. It is better to remove it
* GrLivArea seems outside the range of 4000 to 200000 appears to be an outlier. In the below plot we can see the variable after removing outlier.



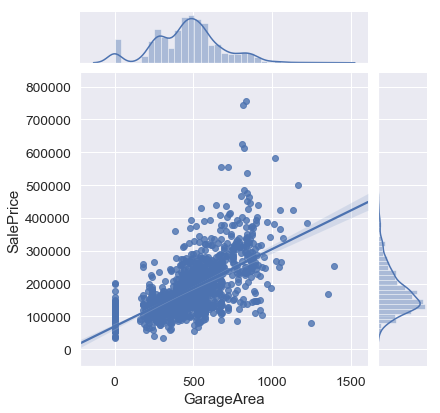
**GarageCars Vs SalePrice**



**Observation**

* From the above plot we can say that for GarageCars of 3, the SalePrice is more.
* But strange is that the GarageCars==4 having less cost, As it is an outlier we can remove this outlier

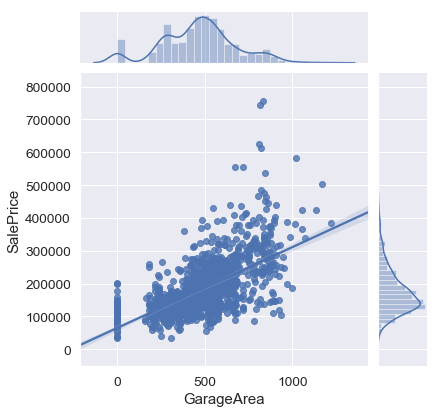
**GarageArea Vs SalePrice**



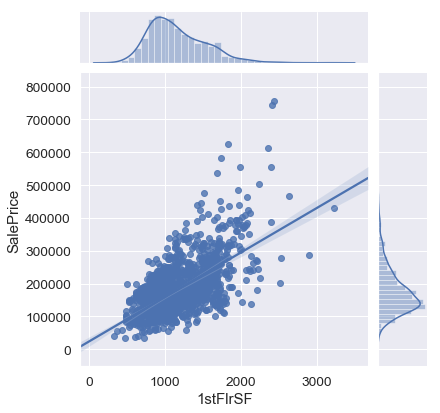
**Observation**

* From the above plot, we can say that for GarageArea of 200–1000 has most of the SalePrice.
* But strange is that the GarageArea>=1000 having less cost, As it is an outlier we can remove this outlier

After removing Outlier



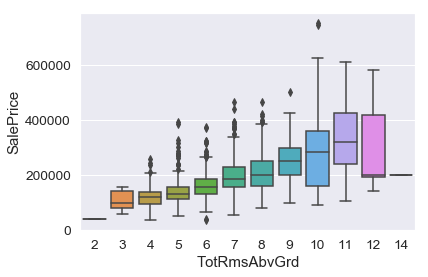
**1stFlrSF Vs SalePrice**



**Observation**

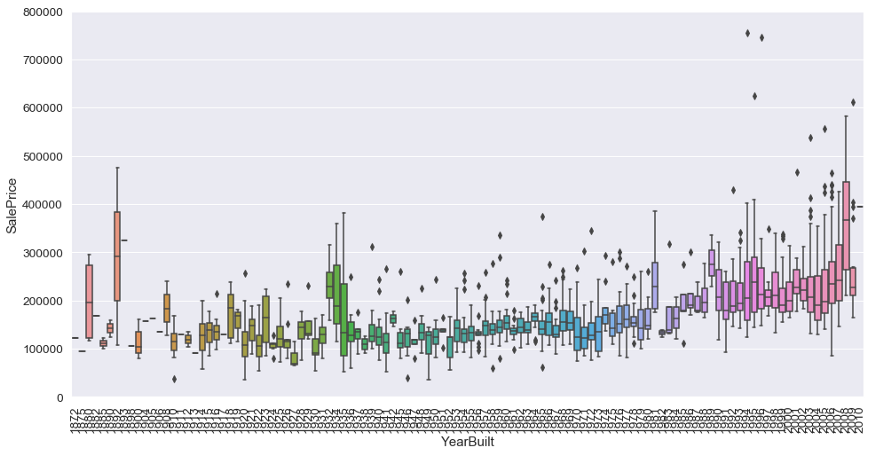
* From the above plot, we can say that for 1stFirSF of 500–2000 has most of the SalePrice.
* We can see as 1st-floor square feet increases, the SalePrice also increases

**TotRmsAbvGrd Vs SalePrice**



**Observation**

* From the above plot, we can say that for TotRmsAbvGrd having more than 11 rows has less weight.
* May be those are old enough due to which they cost less, But it is just an assumption.

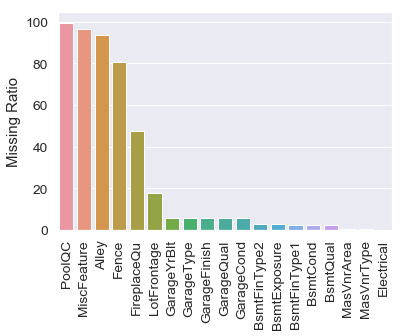


**Observation**-

* Even we see as the building ages the cost is both increasing and decreasing. Our assumption was if the building cost is less for a big house, it might be because it was old. But now even if the building is old, The cost seems to be high
* So the reason may be because of stock market crashes.

**Missing value treatment**

* We have to know the reason behind the missing values, instead of dropping the rows blindly less us proceed in convicting the column so that the missing data process is not biased and hiding an inconvenient truth.



**Imputing Missing Values**

* PoolQC : data description says NA means “No Pool”
* MiscFeature: data description says NA means “no misc feature”
* Alley: data description says NA means “no alley access”
* Fence : data description says NA means “no fence”
* FireplaceQu: data description says NA means “no fireplace”
* lot frontage: Since the area of each street connected to the house property most likely have a similar area to other houses in its neighborhood, we can fill in missing values by the median LotFrontage of the neighborhood.
* GarageType, GarageFinish, GarageQual, and GarageCond: Replacing missing data with “None”.
* GarageYrBlt, GarageArea, and GarageCars: Replacing missing data with 0.  
  - BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, BsmtFullBath and BsmtHalfBath: Replacing missing data with 0.
* BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, and BsmtFinType2: For all these categorical basement-related features, NaN means that there isn’t a basement.
* MasVnrArea and MasVnrType: NA most likely means no masonry veneer for these houses. We can fill 0 for the area and None for the type.
* MSZoning (The general zoning classification): ‘RL’ is by far the most common value. So we can fill in missing values with ‘RL’.
* Utilities: For this categorical feature all records are “AllPub”, except for one “NoSeWa” and 2 NA. Since the house with ‘NoSewa’ is in the training set, this feature won’t help in predictive modeling. We can then safely remove it.
* Functional: data description says NA means typical.
* Electrical: It has one NA value. Since this feature has mostly ‘SBrkr’, we can set that for the missing value.
* KitchenQual: Only one NA value, and the same as Electrical, we set ‘TA’ (which is the most frequent) for the missing value in KitchenQual.
* Exterior1st and Exterior2nd: Both Exterior 1 & 2 have only one missing value. We will just substitute in the most  
  common string
* SaleType: Fill in again with most frequent which is “WD”
* MSSubClass: Na most likely means No building class. We can replace missing values with None

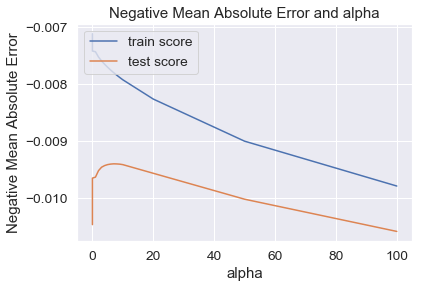
**Model Building**

**Ridge Regression :**

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable.

For ridge regression, we introduce GridSearchCV. This will allow us to automatically perform 5-fold cross-validation with a range of different regularization parameters in order to find the optimal value of alpha.

Plot decide on the optimum value of alpha:



**Observation**

* The above plot defines the way to decide the optimum value of alpha.
* The point in which train and test score has less gap between them is the value which we take as an optimum value of alpha
* From the above plot, we came to know that the value with alpha = 10 has a minimum gap between the test and the training score.
* The R2 value for optimum alpha value:

The r2 value for train data using Rigde Regression 0.9368658742480027  
The r2 value for train data using Rigde Regression 0.9011752415932732

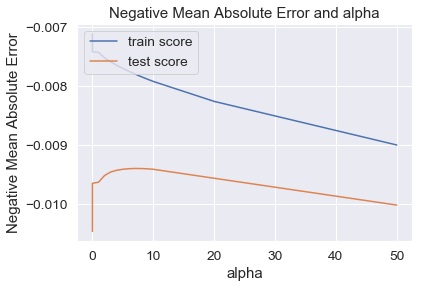
* From the above result, we can see the train data has 0.94 as its R2 value, on test data we have 0.90 as R2 value. So it is pretty much predicting well.
* We can say it hasn't overfitted because the test data(90% r2 value) comparable value when compared to train data(94% r2 value)

**Lasso Regression:**

Lasso regression analysis is a shrinkage and variable selection method for linear regression models. The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that cause regression coefficients for some variables to shrink toward zero. Variables with a regression coefficient equal to zero after the shrinkage process are excluded from the model. Variables with non-zero regression coefficients variables are most strongly associated with the response variable. Explanatory variables can be either quantitative, categorical, or both.

For lasso, we follow a very similar process to ridge regression:

Plot decide on the optimum value of alpha



**Observation**

* The above plot defines the way to decide the optimum value of alpha.
* The point in which train and test score has less gap between them is the value which we take as an optimum value of alpha
* From the above plot, we came to know that the value with alpha = 0.0001 has a minimum gap between the test and the training score.
* The R2 value for optimum alpha value:

The r2 value for train data using Rigde Regression 0.929902822301988  
The r2 value for train data using Rigde Regression 0.8982673232022554

* From the above result, we can see the train data has 0.93 as its R2 value, on test data we have 0.90 as R2 value. So it is pretty much predicting well.
* We can say it hasn't overfitted because the test data(90% r2 value) comparable value when compared to train data(94% r2 value)

**Conclusion**

* From the above two techniques of Lasso and Ridge Regression, we can say that both almost having the same r2 value.
* When comparing the complexity, it is better to use Lasso because as we have 221 variables, Lasso will make the feature selection among the present variables, but Ridge will not reduce columns, it will keep all 221 variables with the reducing the coefficient of variables.