**House Prices: Advanced Bayesian Techniques**

***Deepak Agarwal***

**Abstract**

Understanding the customer needs and predicting customer’s purchase intents form the core of success for any business. Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this project dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence. For the vastly diversified realty market, with prices of properties increasing exponentially, it becomes very essential to study the factors which affect directly or indirectly when a customer decides to buy a house and to predict the market trend. In general, for any purchase, a potential customer makes the decision based on the value for the money.

1. **Introduction**
2. Identify the Problem

The abstract was taken from the website Kaggle. I chose this specific problem because it provided me an opportunity to take an existing dataset and implement Bayesian Model Averaging (a Bayesian Approach) and compare it with different Bayesian models and also against the Multiple Regression Models calculated in the dataset.

1. Software & Implementation Details

I used R & RStudio IDE for my analysis and the following packages in R: Mice, Ameila, BAS and ggplot2 for plots and graphs

1. Link to Data, Code, Screenshots

<https://drive.google.com/drive/folders/1Wsy1Oh4v8TWe0iAPGndKi14cLaN6Blza?usp=sharing>

1. **Background**

I had two different data sets namely train dataset and test dataset. Both contained numerous variables in terms of features which were describing a house. The training dataset contained 1460 observations for which sale price of the house was provided. The test dataset contained 1459 observations for which the sale price was predicted.

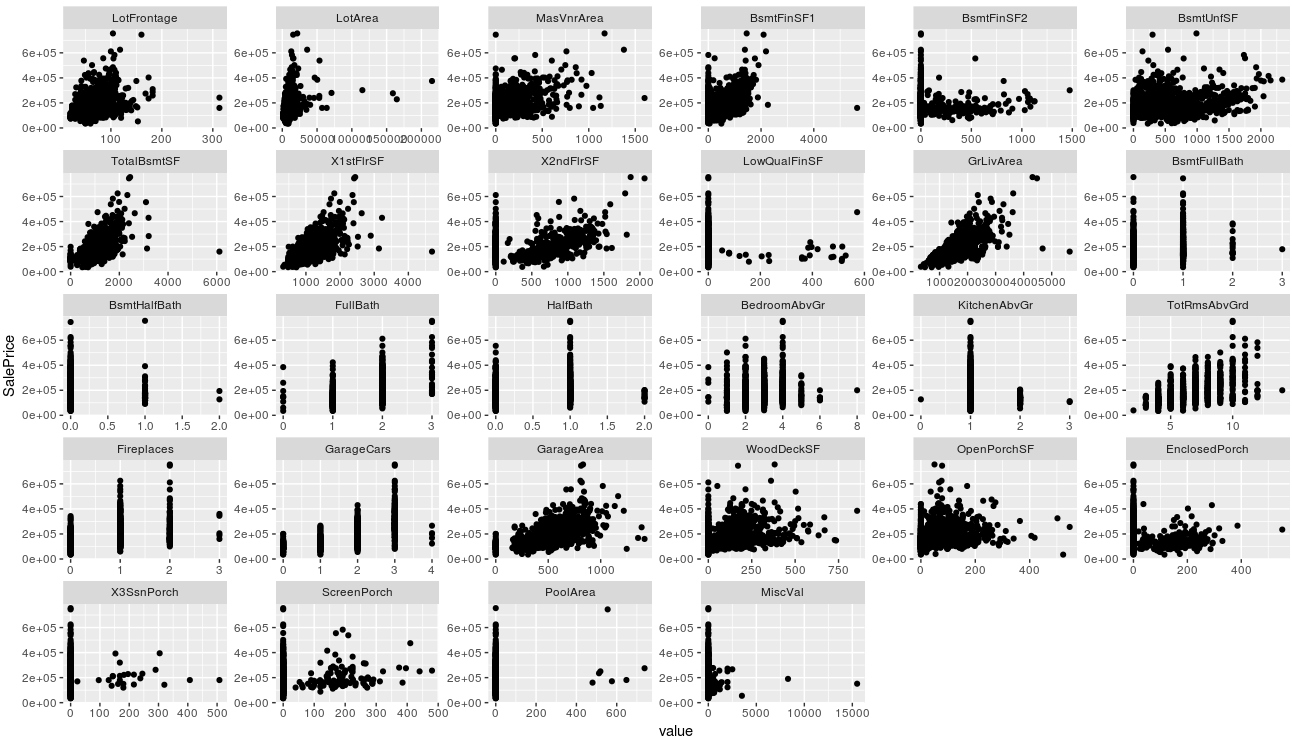
In total 80 variables focus on the quality and quantity of many physical attributes of the property. Most of the variables are exactly the type of information that a typical home buyer would want to know about a potential property (e.g. When was it built? How big is the lot? How many square feet of living space is in the dwelling? Is the basement finished? How many bathrooms are there?).

The dataset contained 23 nominal variables and 23 ordinal variables. Nominal Includes variables like the weather condition and material used for construction. For the nominal and ordinal variables, the levels were in the range of 2 to 28. Total of 14 discrete variables comprise the number of kitchens, washrooms, and bedrooms. This also includes the garage capacity and construction or re-modeling dates. 20 continuous variables describe the area dimension of each observation. Lot size and total dwelling square footage are common home listing available online. Area measurements on the basement, porches and main living area are further classified into respective classes based on quality and type.

1. **Exploratory Data Analysis**

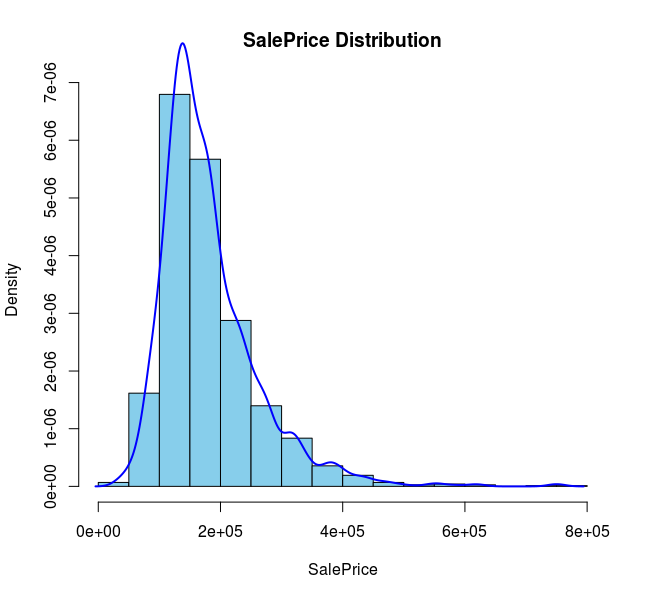
For every data analysis, initial data cleaning, and preprocessing is required. Otherwise outliers in the data set produces wrong prediction models and that can seriously affect the model accuracy. Initially, I decided to analyze data through visualization which can help us understand distribution of data as well as detect outliers. I compared the response variable (sale price) with respect to few important quantitative predictor variables like ‘Garage living area’, ‘Lot frontage’ etc. Scatter plots were plotted and few outliers were observed right away. Also, the box plot helped us to visualize sales price vs categorical values (see screenshots folder).

***Figure 1: Scatter plot of Sales Price vs Numeric Values***



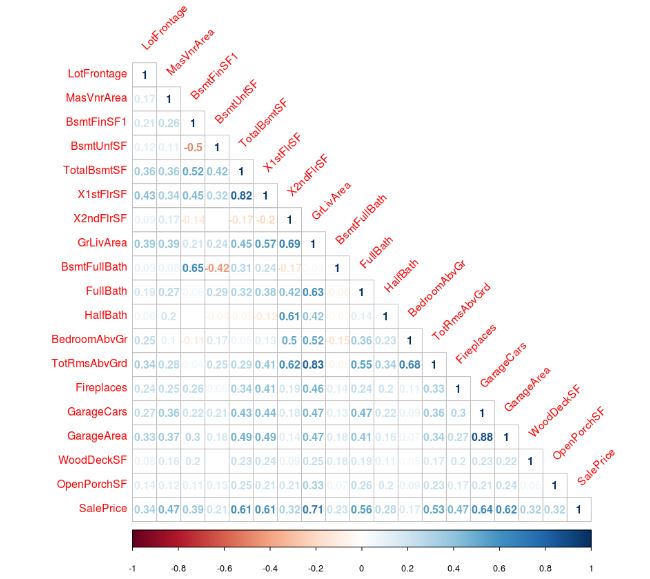
As we know, to apply linear regression the response variable should be normally distributed if not we can apply some transformation. As we can see from below histogram and density plot SalePrice is not normally distributed.

***Figure 2. Distribution for Sale price.***



The histogram is skewed right (not normally distributed) as the outliers were present at the higher price range so we can apply square root, cube root or log transformation. We chose log as data is highly skewed. For missing data, we identified columns which were having missing values.

***Figure 3. Correlation Plot***



By looking at data dictionary it was clear that for some columns values were not missing but they were labeled incorrectly. For example, I found that NA means = no basement, no garage, etc.; not that answer was unavailable. After labeling correctly, I still had missing values which can be viewed in our Missing Map. For it we used mice package to impute data on basis of decision tree methods and checked whether data has been imputed in a sensible way (see screenshots folder) There should not be multi-collinearity in the linear models because it causes more noise in the data. To keep our model working correctly, I removed the variables like LotFrontage, FullBath, Garage Area which had strong correlation (>0.6 or <-0.6) based on our correlation plot also (see above figure 3). Also, I have considered only numeric columns for making predictions to avoid dimensionality issues that comes with categorical columns due to time/scale constraints.

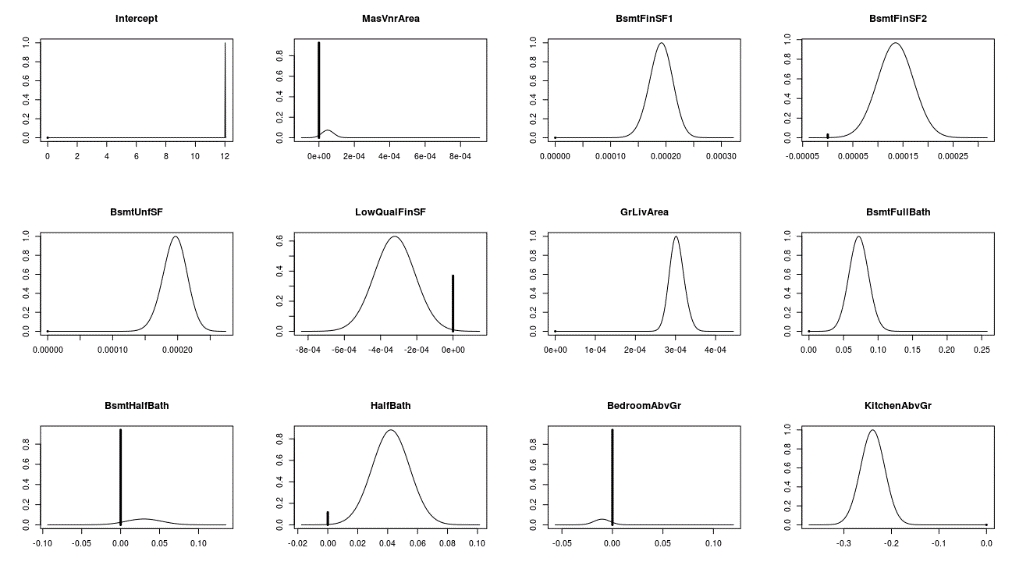
1. **Bayesian Approach**

Linear regression is the basis of most statistical modeling. Bayesian linear regression is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian Inference. When the regression model has errors that have a normal distribution, and if a form of prior distribution is assumed, explicit results are available for the posterior probability distributions of the model's parameters.

When there are multiple models that are equally plausible, a way to choose the best model instead of mistakenly ignoring another is to implement Bayesian model averaging (BMA). I can also use different estimators like ’HPM’ the highest probability model, ’BPM’ the model that is closest to BMA predictions under squared error loss or ’MPM’ the median probability model. Here, multiple models are averaged to obtain posteriors of coefficients and predictions from new data. (See folder of screenshots for Summary Linear Regression on BMA Sales Price, Marginal Posterior Summaries of Coefficients, and 95% confidence interval.)

The plot below shows the density of Posterior distributions of each Predictor variable (which shows which variables are important or significance for predicting the response variable). The higher the bar in each plot shows that variable has little importance on sales price.

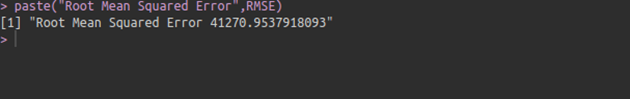
***Figure 4. Posterior BMA***



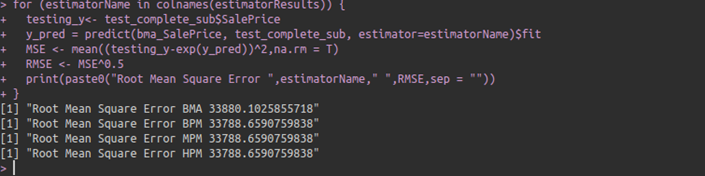
1. **Results/Conclusion**

Even though the Bayesian Approach had less Multiple R-Squared than the Linear Regression, the Bayes predictions were better as RMSE (Bayes), figure 5, was less than RMSE (Linear Regression), figure 6. The values were superimposed on each to show a positive linear relationship.

***Figure 5. Root Mean Squared Error (Linear Regression)***

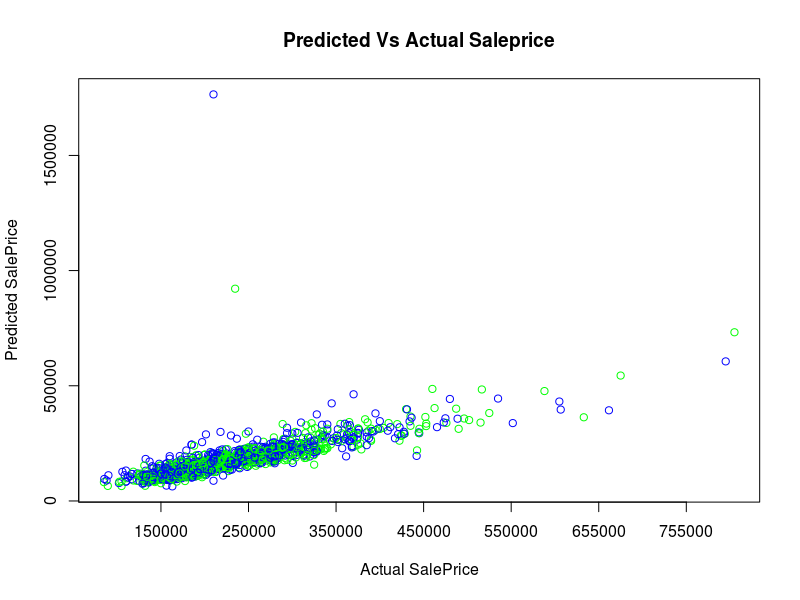


***Figure 6. Root Mean Squared Error (Bayesian Linear Regression)***



The figure below shows predicted versus actual sale price. (Based on Bayesian approach)

***Figure 7. Predicted Vs Actual Sales Price***



***References:***

1. Bayesian Model Comparison <https://ned.ipac.caltech.edu/level5/Sept13/Trotta/Trotta4.html>
2. BIC Prior in Bayesian Linear Regression

<https://stats.stackexchange.com/questions/234096/what-is-the-bic-prior-for-bayesian-linear-regression>

1. Understanding Bayesian Posterior Distributions

<https://www.stat.washington.edu/~raftery/Research/PDF/rmh1997.pdf>

1. Sample Bayesian Linear Regression

<https://galeascience.wordpress.com/2016/09/11/bayesian-linear-regression-on-the-swiss-dataset/>

1. AIC vs BIC

<https://stats.stackexchange.com/questions/577/is-there-any-reason-to-prefer-the-aic-or-bic-over-the-other>