SAS® GLOBAL FORUM 2018

USERS PROGRAM

Home Value Prediction Using Datamining

April 8 – 11 | Denver, CO

#SASGF

Zillow’s Home Value Prediction Using Data Mining

Vivek Singh

Oklahoma State University

ABSTRACT

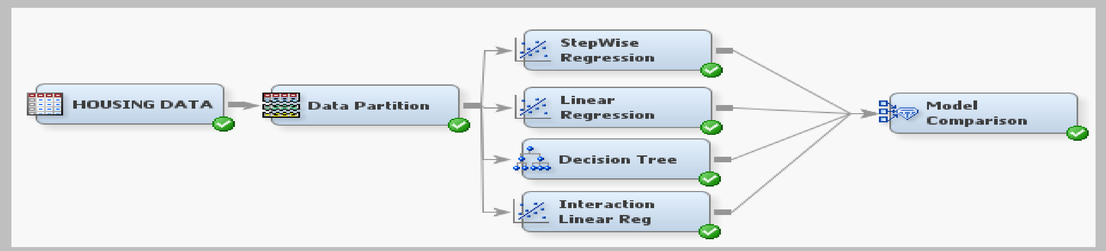
Home valuation has shaken up the U.S. real estate industry since release of online real estate business. A home is often the largest and most expensive purchase a person makes in his or her lifetime. Ensuring homeowners have a trusted way to monitor this asset is incredibly important. The online estimate was created to give consumers as much information as possible about homes and the housing market, marking the first time consumers had access to this type of home value information.

The goal of the paper is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analyzing previous market trends and price ranges, and also upcoming developments future prices will be predicted. The functioning of this paper involves a website which accepts customer’s specifications and then combines the application of multiple linear regression algorithm of data mining. This application will help customers to invest in an estate without approaching an agent. It also decreases the risk involved in the transaction.

METHODS

Data has been obtained from Kaggle. Data Information consists of housing details in King County Seattle, Orange and Ventura, in California. Later for predictive modeling, the data was partitioned into training (70%) and validation (30%) data. Numeric variables were transformed to adjust skewness and kurtosis. Tree based imputation was used to impute missing numeric values.

In this paper, we use the house price data ranging from early 1900 to 2000 to predict the average house price. Linear Regression seems to be the best way to model this dataset. This is because the data follow a highly linear relationship - all we have to do is select features that represent that linear relationship best. We applied several regression method started with a vanilla linear regression, stepwise regression, polynomial regression. The linear regression with two factor interactions and polynomial degree of two turned out to be the winner. This approach resulted with Adjusted R^2 of 78%.



RESULTS

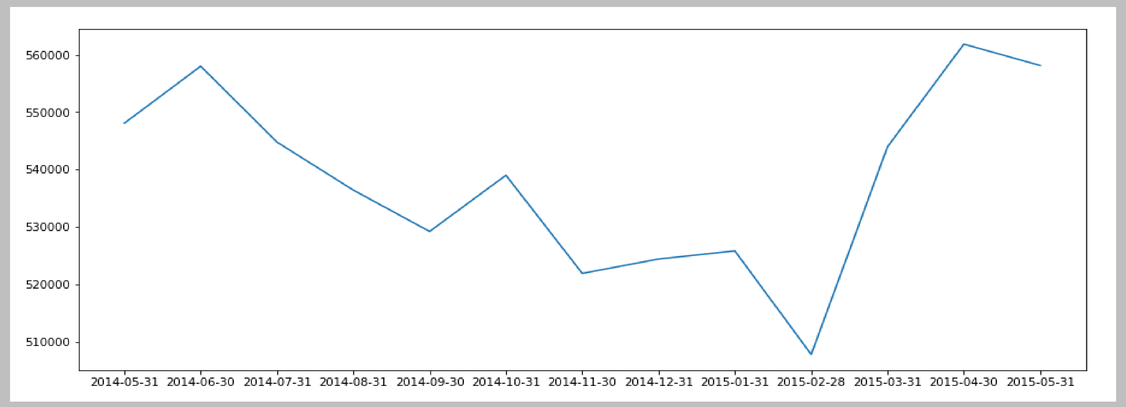
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	127	1.5321762E15	1.206438E13	428.52	<.0001
Error	15001	4.2232834E14	28153345708		
Corrected Total	15128	1.9545046E15			

Model Fit Statistics

R-Square	0.7839	Adj R-Sq	0.7821
AIC	364145.2962	BIC	364149.4805
SBC	365121.2154	C(p)	128.0000

With Time series analysis we found the pattern that there is seasonal effect on house prices, during November through February the house prices go down.



Predictions are better for newer houses.

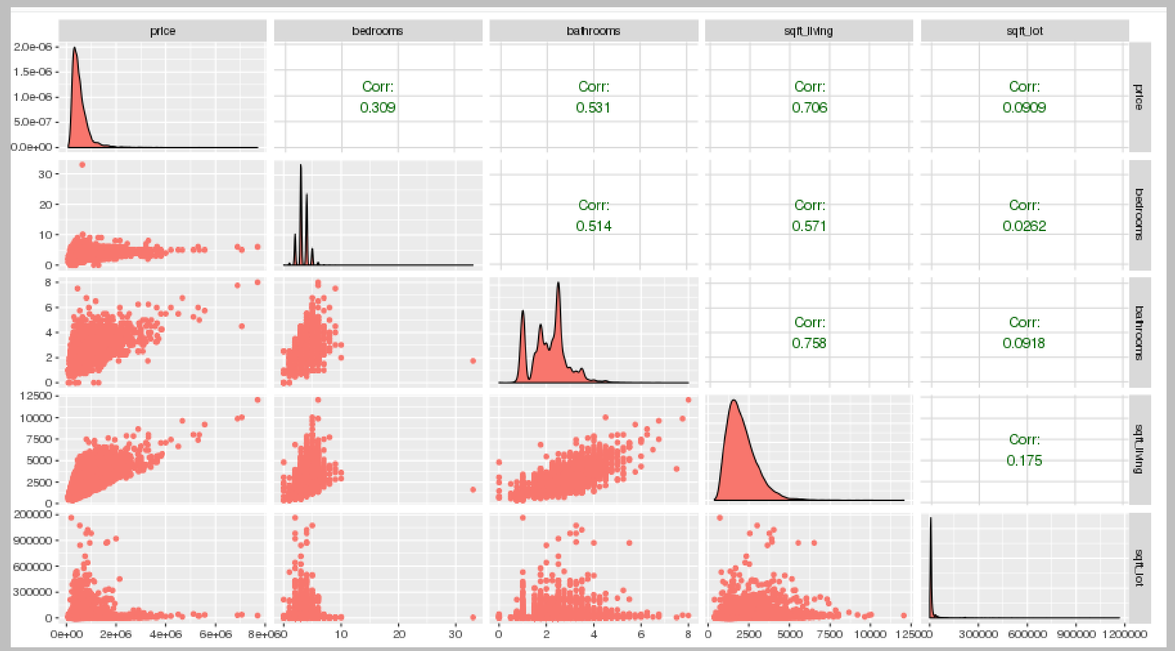
Prices for house is highly correlated with number of bathrooms, square feet living area and grade whereas weak correlation with the number of bedrooms, Lot size, floor, view, and year built.

Zillow’s Home Value Prediction Using Data Mining

Vivek Singh

Oklahoma State University

RESULTS CONTINUED



CONCLUSIONS

This data has many limitations. The largest one is that we have no information about potential buyers. An auction can be extremely unpredictability. Ego, biding wars, buyer’s susceptibility to a seller or realtors pressure, and even the weather during the time the house was on the market, are all great examples of missing information. This is why even human realtors can only predict the selling price of a home within 10-15% at best - and they have access to a lot of that missing information! Furthermore, the value of a house can be heavily influenced by features that are extremely subjective. Artistic quality of the residence, architecture, and a specific style might appeal strongly to one buyer but not another.

However, with a combination of various approaches we were able to come up with a method: by combining regression, clustered regression, random forest, and clustered random forest we actually came very close to the oracle with an average error of only about 20%. Throughout this process we gained a far deeper understanding of the housing market as well as features that were most closely correlated with price, which could assist us in developing a more general framework for predicting housing prices in different parts of the world in the future. Finally, we expanded upon many basic methods that were introduced in lecture, gaining a very deep understanding for how they work and how we might use them for future projects.

REFERENCES

https://web.stanford.edu/class/cs221/2017/restricted/p-final/ianjones/final.pdf

http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6073000

http://thesai.org/Downloads/Volume8No10/Paper_42-Modeling_House_Price_Prediction_using_Linear_Regression.pdf

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF