Statistical Analysis of Random Forest on Real Estate Prediction

Joylin Zeffora.A, R. Shobarani

Abstract: ***Prediction models in real estate have a significant role to play in telling the future of the real estate industry. They have a role in forecasting that is essential to investors who use the information to come up with effective decisions. Random Forest model's accuracy in estimating residential property prices are much better when compared to other models as the marginal error is comparatively less.***

Keywords: Random Forest, Real Estate model, Statistics, Predictive analysis, Price prediction

I. INTRODUCTION

There are various predictive models in real estate and they include Linear Regression, Stepwise Regression, Decision Tree, Random Forest and Boosted model. Real estate industry has been experiencing various challenges which have been limiting its ability to realize full potential. Therefore, it is essential to focus on various real estate models while comparing different types of models.

Random forests model was originally developed as a method of combining several CART individual decision trees using bagging. Mousa and Saadeh further educate us that the development was influenced by the novice random subspace model. The proponents of this approach of random split selection from Dietterich (2010) on feature selection. Notably, several of the key ideas used in random forests model is also found in the early work of on ensembles of the decision tree. Since their introduction, random forests models have developed from a single algorithm to complex framework of models, and have been applied to in many fields.

Some argue that even although, many studies on mass appraisal have used this technique, the mathematical forces underlying the whole process are not well understood. The justification for this is the work of Breiman (2010), who based on mathematical and intuition heuristics, was quite explicit and difficult to comprehend (Biau, 2012). There are two core theoretical issues that are associated with random forest model. The first is the consistency of estimators that are randomly generated by the algorithm (Zurada, 2011). This is paramount since it guarantees convergence to the most optimal solution as the data set grows considerably large. The other issue is the rate of node convergence. However, in this study, the researcher focuses on the consistency which is established in the Breiman's original algorithm. Most studies involving this technique typically focus on this algorithm.

An example of this is the research conducted which aimed at studying random forests models in one dimension with random splitting decision trees. Notably, this technique ensures easy tractability (Biau 2012).

II. RANDOM FOREST

Random Forest model approximates the sale prices of apartments whilst comparing it against the traditional multiple regression analysis (MRA). It is imperative to note that Mousa, (2010) notes that RF model resulted in having almost accurate prices of twice the predicted values within five-percent of the apartment sales price than their MRA model had forecast on a test sample of 100 houses.

Let *prf* represent the accuracy of predicted price of the RF model and *pra* represent the accurate predicted price of MRA.

prf=2(pra) within the first 5% in a test sample of 100 houses.

The study concluded that the RF model performs better than other regression models for estimating the value of residential properties in the US. The independent variables included; the number of bathrooms, number of garages, age, total square footage, number of fireplaces, size, and the number of floors. Notably, the dependent variable was the selling price.

Random Forest model's accuracy in estimating residential property prices have been used in various parts of the world. They also investigated the impacts on the average forecast error when outliers were both removed in the test data and the training dataset. The research established that when outliers are omitted from the data set, the RF model works well to predict the value property. Notably, the absolute error was averaged. Nonetheless, not all researches have reported positive or favorable results from the use of this technique. Various authors reviewed a number of these project researches and presented a paradigm of what can happen when different RF models are used for predicting the value of the real estate properties. (McCluskey, 2014). The studies concluded that that optimal random forest models solely depends upon the time period involved and the specific data sets. Besides, they also established that if the same data sets are combined with varying model settings can produce differing results. For this reason, it is recommended that scholars be cautious during the development and use of the random forests networks in real estate property appraisal.

Revised Manuscript Received on May 22, 2019.

Joylin Zeffora.A, Research Scholar, Dr. M.G.R. Educational and Research Institute

R. Shobarani, Professor, Dr. M.G.R. Educational and Research Institute

III. REAL ESTATE MODEL

The raw data in the Real Estate dataset originally contains 266 observations made from 23 variables. The task is to develop a model for predicting what a house should sell for using the Real Estate data set. To accomplish this task, I started by the first step of determining the appropriate sample to use in developing the model. I have determined the sample by carrying out data cleaning. I have taken several steps in carrying out the cleaning. These steps are as outlined below:

Firstly, I have started by dropping the variables that I have considered to have no effect on the price a house should. In that category, I have dropped MLS - Multiple Listing Service - serves as a house ID in the real estate system. I have also dropped the variable COUNTY since all the observations are from Chester County. This implies that county would have no influence on the model I intend to develop.

Similarly, I have also dropped the variables CONDITION - the condition of the house - condition does not have a standard definition and Ltd. - listing date - the day the house was put on the market. I dropped these two variables since they did not have any observation recorded. In summary, I have dropped 4 out of the 23 variables leaving a total of 19 variables.

Secondly, I have dropped all the observations without any records i.e. those with N/A and blanks. I did this for each and every variable with such entries. I did this since the entries could end up influencing the final outcome of the model so desired by influencing the sample size. There were numerous cases of blank and/or N/A entries in various variables such as HIGH, MIDDLE, and SQ FOOT just to mention a few.

Thirdly, I have also dropped all the entries with 0 in the ACREAGE variable. This is important since a house cannot have a zero acreage coverage of land. I have also dropped all the entries of 0 in the RETAXES variable. This is essential since 0 entry means that there is no tax imposed in the building, which is not true.

Consequently, I have grouped related entries in the variables. This is evidenced in the DESIGN variable where there are observations recorded as "1-story and 1 story". I have grouped these together and coded them as 1. Similarly, there are observations recorded as "split/multi and Split/multi". I have again grouped them and coded them as 5.

Finally, I have identified and removed outliers and potentially inaccurate data. An outlier is an entry or observation that lies in an abnormal position from the rest of the other observations (Lind, et al. 2008).is an In the DOM variable, the days on the market, I have considered days less than 30 (one month). This is not a significant time frame to influence the price of a building. Similarly, in the GARAGE variable, I have considered 0 as an outlier/inaccurate entry since it would have a very insignificant influence on the price of a building. Instead, it can lead to erroneous model since it will influence the sample size.

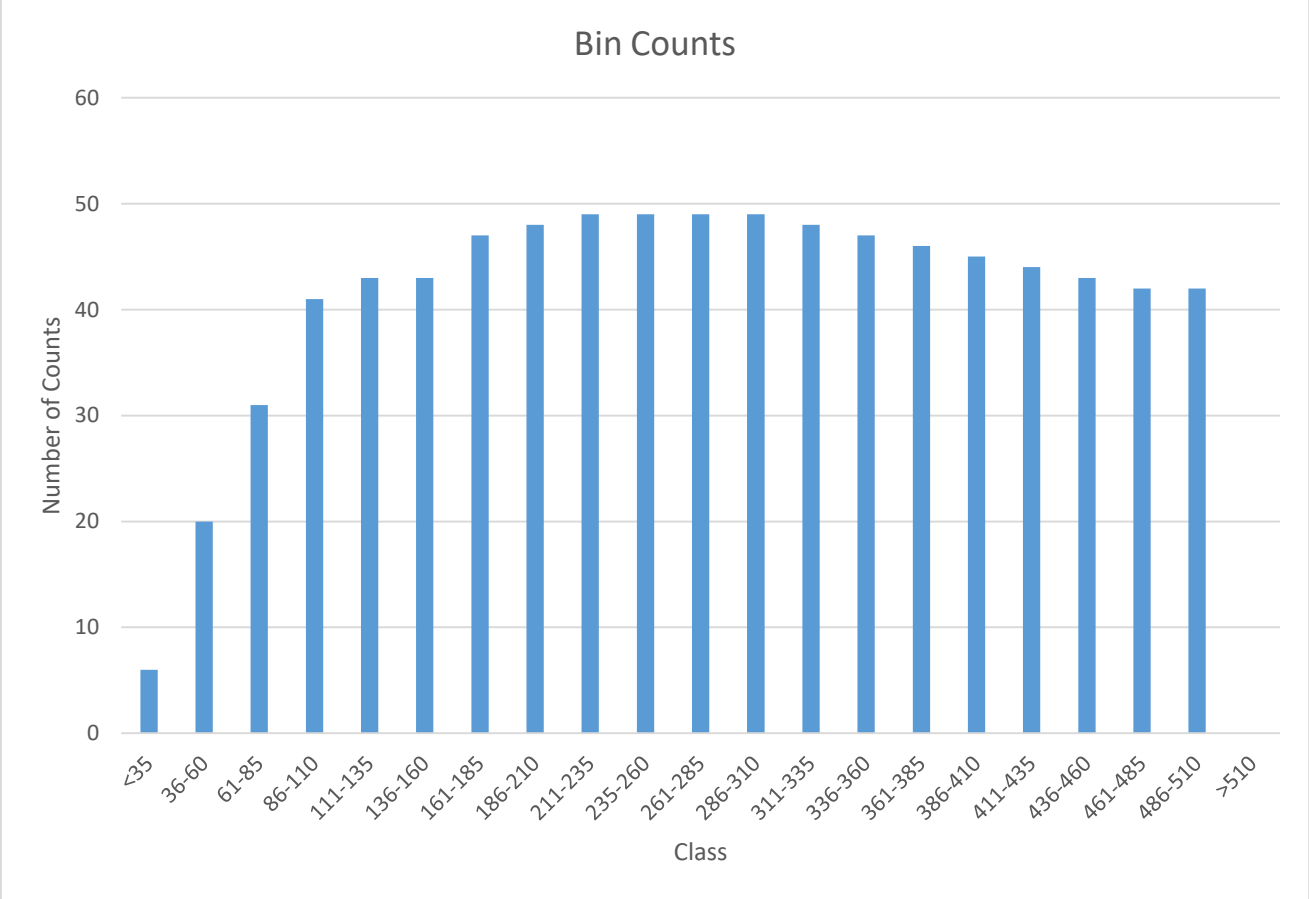
In conclusion, after the cleaning, I ended up with 19 variables and 61 observations only. Out of the 19 variables, there are both categorical and continuous variables. We know that variables can either numerical or categorical. A numerical data can either be discrete or continuous (Lind, et al. 2008). A discrete variable can only take a set of particular values while continuous variables can take any value within a specified range (Lind, et al. 2008). Similarly, categorical variables can either be nominal, attribute or ordinal (Frankfort, et al. 2015). An attribute (also called dichotomous or dummy) variable has got only two categories (Jackson, et al. 2005). A nominal variable, on the other hand, has got several unordered categories (Kaye, et al. 2001). An ordinal variable has got several categories that can take a given order (Stuart, et al. 1999).

From the above descriptions, I easily categorized the 19 variables as follows; The continuous variables in the sample include GARAGE, ACRE, ASSESS, RETAXES, SQ FOOT, AGE, ½ BATH, BATH, BED, PRICE, and DOM. Similarly, the categorical variables include SUBDIV / NE, SCHOOL DISTRICT, HIGH, MIDDLE, ELEM, TYPE, DESIGN, and STYLE.

Out of the categorical variables, there were nominal, ordinal or attribute variables. On the other hand, the nominal variables include; SUB DISTRICT, HIGH, MIDDLE, ELEM, STYLE, and TYPE. The Ordinal variable is the DESIGN while the Attribute variable is DISTRICT.The categorical variables are coded appropriately; for the variable STYLE, the codes are as follows: *carriagehouse=1,colonial=2,colonial,Endunit/Row=3,colonial,,Traditional=4,contemporary=5, Farmhouse=6, French=7 and Ranch=8.* For the variable TYPE; *Row/townhouse=0, single=1, single/detac=2.*
For the variable ELEM; *chads=0, E Bradford=1, Exton=2, Hilendale=3, Hillsdale=4, Mhouse=5, Penn wood=6, pocopson=7, strkwether=8, Unionville=9 and wstwn=10.*
For the variable DESIGN; *1.5 story=0, 1 story=1, 2story=2, 3 story=3, bi-level= 4 and split/multi=5.* For the variable SCHOOL DISTRICT; *Unionville=0 and West Chester=1.* For the variable HIGH; *East=0, Henderson=1, Ruston=2 and Unionville=3.* For the variable MIDDLE; *C F patton=0, E N peirce=1, J R Rugett=2 and Stetson = 3.* Finally, for the variable SUB; *Appligate=1, Bala=2, Beversrede=3, Birmingham=4, Blue Rock Meadows=5, Bradford pointe=6, Brandywine overloo=7, Brandywine River E=8, Brandywine E=9, Brandywine Thorn, chads ford knol=10, Deerfield greene=11, dilwortown oak E=12, est at chads=13, fair hill=14, field point=15, ffolkstone=16, folke monor=17, Grandview=18, Hamilton place=19, harmony hill=20, highland farm=21, knol of birgham=22, marshaltown=23, new south hill=24, newin greene=25, oadbourne=26, old oak=27, olstead=28, pleasant=29, Radley run iv=30, red bridge farm=31, revolutionary=32, sconneltown=33, silverwood=34, spring mdws=35, squires lea=42, thornby estate=43 and whiteland crest=44.*

I have also decided to bin the DOM variable. Binning is the process of grouping data in a graphical manner like the histogram (Stuart, et al. 1999).I have decided to bin this variable since it has a wide range of entries from as low as 31 to as high as 497.

Furthermore, the observations are quite numerous. I have used a lower limit of 35- a value that is just slightly above the lowest entry which is 31. I have also used an upper limit of 510- a value that is slightly above the highest value which is 497. Consequently, I have used an interval of 25. The histogram produced after binning is as shown in the figure below.



Descriptive statistics is a way of analysing data in which the quantitative and qualitative characteristics of the variables are identified (Stuart, et al. 1999). The table below outlines the descriptive characteristics of the continuous data.

From the table, for example, we can see that the mean price of a house in Chester county is 575, 554.10, mean date on market is 104.56 and the mean tax on houses is 7, 483.74. The full descriptive statistics is outlined in the table.

	DOM	PRICE	AGE	SQFOOT	RETAXES	ASSESS	ACRE
Mean	104.56	575,554.10	18.84	3,763.16	7,483.74	321,372.61	0.98
Standard Error	10.49	27,537.41	1.78	201.76	462.08	17,863.36	0.12
Median	78.00	542,000.00	15.00	3,418.00	6,580.00	288,810.00	0.75
Mode	43.00	725,000.00	10.00	2,129.00	#N/A	154,800.00	1.00
Standard Deviation	81.95	215,074.03	13.94	1,575.80	3,608.93	139,517.33	0.92
Sample Variance	6,716.22	46,256,839,857.92	194.31	2,483,151.04	13,024,355.23	19,465,084,166.44	0.85

Kurtosis	8.07	-0.17	0.27	1.53	1.66	1.03	7.29
Skegness	2.41	0.59	1.08	1.23	1.25	1.03	2.44
Range	466.00	916,000.00	55.00	7,608.00	17,614.00	664,660.00	4.87
Minimum	31.00	269,000.00	2.00	1,392.00	2,413.00	119,510.00	0.03
Maximum	497.00	1,185,000.00	57.00	9,000.00	20,027.00	784,170.00	4.90
Sum	6,378.00	35,108,800.00	1,149.00	229,553.00	456,508.00	19,603,729.00	59.73
Count	61.00	61.00	61.00	61.00	61.00	61.00	61.00
Largest(1)	497.00	1,185,000.00	57.00	9,000.00	20,027.00	784,170.00	4.90
Smallest(1)	31.00	269,000.00	2.00	1,392.00	2,413.00	119,510.00	0.03
Confidence Level (95.0%)	20.99	55,083.02	3.57	403.58	924.29	35,732.05	0.24

Correlation is the measure of the degree of association between variables (Stuart, et al. 1999). A correlation of 1 indicates a perfect correlation. Correlation can be negative or positive. For example, from the output below, it is clear that there is a week negative correlation between the price and DOM. This suggests that as the date on the market increases, the price of a building tends to reduce. Similarly, the correlation between the AGE and DOM is 0.20, which is a week positive correlation. This suggests that as age increases, DOM increases as well. Again, we can see that the correlation between RETAXES and PRICE is 0.82 which is a strong correlation. This suggests that as the price of a house increases, the tax increases as well.

	DOM	PRICE	AGE	SQ FOOT	RETAXES	ASSESS	ACRE
DOM	1.00						
PRICE	-0.03	1.00					
AGE	0.20	-0.62	1.00				
SQ FOOT	0.16	0.86	-0.55	1.00			
RETAXES	0.03	0.82	-0.57	0.81	1.00		
ASSESS	0.03	0.93	-0.65	0.87	0.92	1.00	
ACRE	0.20	0.59	-0.04	0.65	0.49	0.53	1.00

IV. CONCLUSION

Concisely, prediction models play a significant role in the real estate industry. There are various prediction models and they include decision Tree, Random Forest, linear regression and bankruptcy prediction model. All these models have a common goal which is to provide a clear picture on how the real state can reach its full potential. Random Forest model's accuracy in estimating residential property prices are much better when compared to other models.

REFERENCES

1. Altman, EL. & Sabato G (2007): Modeling Credit Risk for SMEs: Evidence from US Market. A JournalApproach. Abacus, 43(3):303–324.
2. Moon, T. & Sohn, S. (2010): Technology credit scoring model considering both SME characteristics andof Accounting, Finance and Business Studies (ABACUS), 43(3):332-357
3. Peat, Maurice. "Factors affecting the probability of bankruptcy: A managerial decision based approach." Abacus 43.3 (2007): 303-324.
4. Řezáč, Martin, and František Řezáč. "How to measure the quality of credit scoring models." Finance a úvěr: Czech Journal of Economics and Finance 61.5 (2011): 486-507.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

5. Frankfort-Nachias, C., & Leon-Guerrero, A. (2015). Social Statistics for a diverse society,(7thed.). Thousand Oaks, CA: Sage Publications.
6. Jackson, S., Marcus, K., McDonald, C., Wehner,T., & Palmquist, M. (2005).
Methods. Writing@CSU Colorado State University Department of English. Retrieved December 15, 2009
from http://writing.colostate.edu/guides/research/stats/
7. Kaye, David H., and David A. Freedman. "Reference guide on statistics." Reference manual on scientific evidence, (2011): 211-302.
8. Lind, Douglas A., William G. Marchal, and Samuel Adam Wathen. Statistical techniques in business & economics. New York, NY: McGraw-Hill/Irwin,, 2012.
9. Stuart A., Ord K., Arnold S. (1999), Kendall's Advanced Theory of Statistics: Volume 2A- Classical Inference & the linear Model.
10. Mousa, A.A. & Saadeh, M., 2010.Automatic valuation of Jordanian estates using a genetically-optimised artificial neural network approach. WSEAS TRANSACTIONS on SYSTEMS (ISSN 1109-2777), 9(8), pp.905-916.
11. Nghiep, Nguyen, and Cripps Al. "Predicting housing value: A comparison of multiple regression analysis and artificial neural networks." Journal of real estate research 22.3 (2001): 313-336.
12. Pace, R. Kelley. "Parametric, semiparametric, and nonparametric estimation of characteristic values within mass assessment and hedonic pricing models." The Journal of Real Estate Finance and Economics 11.3 (1995): 195-217.
13. Worzala, Elaine, Margarita Lenk, and Ana Silva. "An exploration of neural networks and its application to real estate valuation." Journal of Real Estate Research 10.2 (1995): 185-201.
14. Zurada, J.M., Levitan, A.S. &G uan, J., 2011. Non-conventional approaches to property value assessment. Journal of Applied Business Research (JABR), 22(3).
15. Biau, Gasrard. "Analysis of a random forests model." Journal of Machine Learning Research 13.Apr (2012): 1063-1095.