THE EFFECT OF GOOGLE DRIVE DISTANCE AND DURATION IN RESIDENTIAL PROPERTY IN SYDNEY, AUSTRALIA

MEHRDAD ZIAEE NEJAD

Faculty of Engineering and Information Technology, University Technology of Sydney, Decision Systems and e-Service Intelligence Lab, Centre for Quantum Computation and Intelligent Systems (QCIS), Ultimo NSW 2007, Australia

JIE LU

Faculty of Engineering and Information Technology, University Technology of Sydney, Decision Systems and e-Service Intelligence Lab, Centre for Quantum Computation and Intelligent Systems (QCIS), Ultimo NSW 2007, Australia

POOYAN ASGARI

Data Products and Insights, Domain Group, FairFax Media, 100 Harris St., Pyrmont NSW 2009, Australia

VAHID BEHBOOD

Faculty of Engineering and Information Technology, University Technology of Sydney, Decision Systems and e-Service Intelligence Lab, Centre for Quantum Computation and Intelligent Systems (QCIS), Ultimo NSW 2007, Australia

Predicting the market value of a residential property accurately without inspection by professional valuer could be beneficial for vary of organization and people. Building an Automated Valuation Model could be beneficial if it will be accurate adequately. This paper examined 47 machine learning models (linear and non-linear). These models are fitted on 1967 records of units from 19 suburbs of Sydney, Australia. The main aim of this paper is to compare the performance of these techniques using this data set and investigate the effect of spatial information on valuation accuracy. The results demonstrated that tree models named eXtreme Gradient Boosting Linear, eXtreme Gradient Boosting Tree and Random Forest respectively have best performance among other techniques and spatial information such drive distance and duration to CBD increase the predictive model performance significantly.

Keywords: Property Valuation, Machine Learning, Regression, Pre-processing

1. Introduction

The market value of a residential property depends on its structural and locational characteristics. To value a residential property accurately, these characteristics should be involved in valuation process. In traditional way, an inspector or valuer visits the property to take into account these characteristics visually and using the "*compare*" approach to value the property. These approach is time consuming and expensive while novel approaches which use machine learning (ML) techniques in the core of automated valuation models (AVMs), work with recorded transactions (real estate databases) to predict the value of property faster and more cost effective than traditional approach. It should be noted that the accuracy of AVMs might be less than inspection with professional valuer because of lacking structural and spatial information in databases such as type of used material in the building and distance of property to amenities like central business district (CBD). Also, the relationship between property characteristics and its value is not crystal clear and it is difficult to measure. Forming the appropriate predictive model to value a property has been challenging area for academic and industry. To investigate the mentioned relationship and estimate the market price of residential property, researchers have applied various statistical and machine learning techniques. The statistical techniques include linear regression and hedonic regression. For instance, Smith [1] uses the multiple regression analysis (MRA) and Henry [2] applies hedonic regression to value private properties. The machine learning techniques such as neural networks ([3], [4], [5], [6], [7] and [8]), Support Vector Machine ([9]) and tree based models ([10] and [11]) have also been studied to value a residential property. In recent years, McCluskey [10] investigates the performance of boosted regression trees (BRT) to estimate the price of residential properties in Malaysia. To evaluate the performance of the BRT model, they compare it with two specific multiple regression analysis

(MRA) models (linear and non-linear). Based on the presented results, the BRT model has lower error rate to predict the price of residential effectively compared to the two other models but it is not as interpretable as the other two because of its complexity. Antipov and Pokryshevskaya [11] test the Random Forest method to value apartments in Saint Petersburg. The Random Forest method is compared with CHAID, CART, KNN, MRA, NN and BRT. Based on the results of the comparative study, the random forest method in most criteria has the best performance compared to other methods. In addition, they show that all the algorithms perform better if the price per meter is predicted at first and then the price of property is calculated based on it. McCluskey et al. [13] compare NN method to the several MRA techniques to estimate the price of residential properties accurately in mass appraisal. The applied data sample consists of 2694 sales registered within the Lisburn District Council area (Northern Ireland) during the period 2002-2004. The ordinary least square (OLS) regression is used as a linear model and NN and two non-linear regression models are used as non-linear models. This research shows that the accuracy of a non-linear regression model is higher than the NN.

Accurate prediction of housing sales price is most important in the operation of the housing market. Having an accurate estimation of market price of residential property is important for different users such as investors who face choices among housing securities and other investment opportunities aim to have a precise price estimate. Some researches investigate the risk of investment in real estate industry that they mention the importance of accurate estimation of real estate value ([14], [15] and [16]).

In this study, we conduct a comprehensive experiment to investigate and analysis existing popular machine learning models for property valuation on real world data samples. We examine 47 ML techniques to predict sale price of about 2000 apartments located in 19 suburbs in Sydney, Australia. These ML techniques consist of six main groups (Linear Regression, Tree Based Regression, Neural Network, K-Nearest Neighbors, Support Vector Machines, and Multivariate Adaptive Regression Spline). This study aims to 1) investigate the performance and reliability of these ML techniques; and 2) measure the influence of spatial data such as drive distance and duration to CBD on the accuracy of predictive models. The results prove strong positive impact of spatial data on underlying models particularly non-linear ones.

This paper is organized as follows. Section 2 describes data set, data preparation process and predictive models which have been applied to data. The results and analysis are presented in section 3 and Section 4 covers the conclusion and future works.

2. Data and Models

2.1. *Data*

The sample data include 3775 records of unit sales between 01/05/2015 until 30/11/2015 (7 months) from PriceFinder dataset that PriceFinder is an online property search application offering a range of tools to assist in generating timely information for all residential, commercial and rural properties. These observations are from 18 suburbs around the central business district (CBD) of Sydney Australia. As shown in Table 1, the data set includes 7 variables.

Table 1. Definition of variables in data set.

Variable	Definition
Address	The full address of apartment
Post code	The inclusive number for each suburb
Sale price	The sale price transaction (AUD)
Date	The date of transaction
#Bedrooms	The number of bedrooms
#Bathrooms	The number of bathrooms
#Car parks	The number of car parks

2.2. *Spatial data*

Since residential properties are a heterogeneous product, these predictors are very poor to build an accurate model to estimate unit's price in Sydney. To build an accurate model we need more information about the property such as date of construction, quality of applied material, air-conditions, lift and spatial information. The lack of vital information directs us to extract spatial information to enhance the accuracy of automated valuation model. The drive distance and duration to CBD is one of spatial information. Calculating the direct line for distance between an observation's address and CBD is not accurate and practical, because in real world we do not have such road. Therefore, we calculate the drive distance and duration using Google map API and add these two variables into the data set using “ggmap” package in R [17]. The R is a free software environment for statistical computing and graphics. As a result, two datasets are created, first dataset without spatial data with 22 predictors (# Bedrooms, #Bathrooms, # Car parks plus 19 dummy variables regarded to suburb’s post codes) and another dataset including spatial data with 24 predictors.

2.3. *Data Preparation*

Dealing with missing values: the observations that have missing values in 'Bedrooms', 'Bathrooms' and 'Car Parks' are removed from data frame, because these variables are very important, so imputation may increase the error of prediction, the minimum meaningful price in Sydney is 200000 AUD and every unit has at least one bathroom. Dealing with outliers: the auto detect outliers based on Grubbs test [18] is used for sale price, number of bedrooms, number of bathrooms and number of car parks respectively. After removing outliers and observation with missing value, the number of observation drops to 1967 records. Table 2 shows the summary of variables after data cleansing.

Table 2. Summary of price, #bedrooms, #bathrooms, and #car parks.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Box-Cox Lambda
Sales price	220K	735K	914K	1053K	1250K	3010K	0
# Bedrooms	0	1	2	1.88	2	4	NA
# Bathrooms	1	1	1	1.46	2	3	-1.1
# Car parks	0	1	1	1.10	1	3	NA

To reduce the skewness and preparing data for models that work with variables between 0 and 1, scaling, centering and the box-cox [19] transformation are done for variables. Dummy variables are used for categorical variables like post codes that they regarded to suburbs.

2.4. *Models*

In this study, we examine six main groups of machine learning models as: 1) Linear model; 2) Tree models; 3) Neural networks; 4) Support Vector Machines; 5) K-Nearest Neighbors and 6) Multivariate adaptive regression splines with 10 fold cross validation. Each group consist several methods that are showed in figures1-11

3. Results

Models are examined with the two datasets by using “caret” package in R [20], first 24 predictors (including spatial information) and another with 22 predictors (without spatial information). In each group, the predictive models are compared based on RMSE and R^2 criteria [21] and shown in figurers [22].

3.1. *Analysis of Spatial data impact*

As we can see in figure 1, in *Linear* group, the Lasso fitted with 22 predictors, Generalized Linear Model fitted with 22 predictors and Ridge Regression (figure 2) with Variable Selection with 24 predictors (including spatial information) have the best three performance respectively. It seems the spatial information do not have positive effect on linear models.

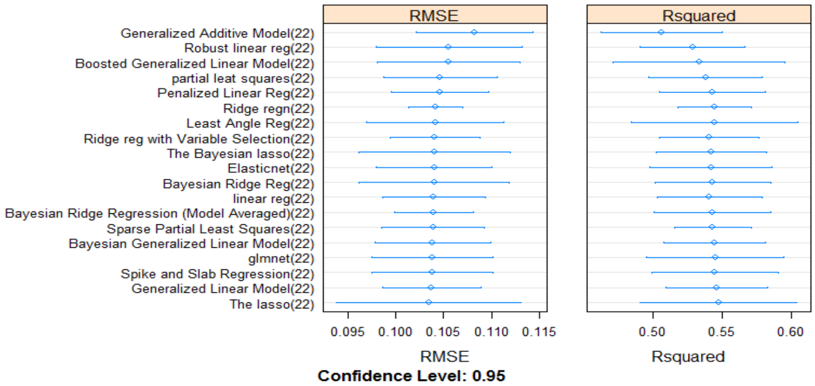

Figure 1. Performance of linear models based on datasets without spatial information.

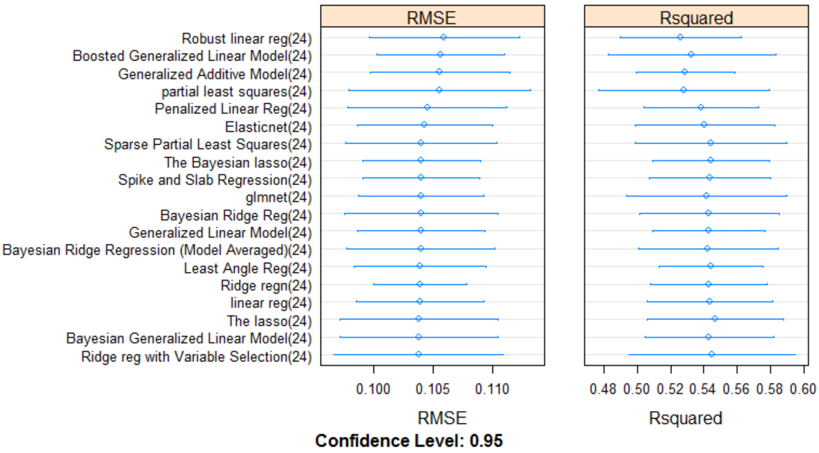

Figure 2. Performance of linear models based on datasets with spatial information.

For the *Tree* group, results are different from the linear group (figure 3). Almost all tree models that fit with dataset including spatial information have less RMSE and higher R^2 compare to models that fit with dataset without spatial data. Boosting models and Random Forest have best performance among other models in this group.

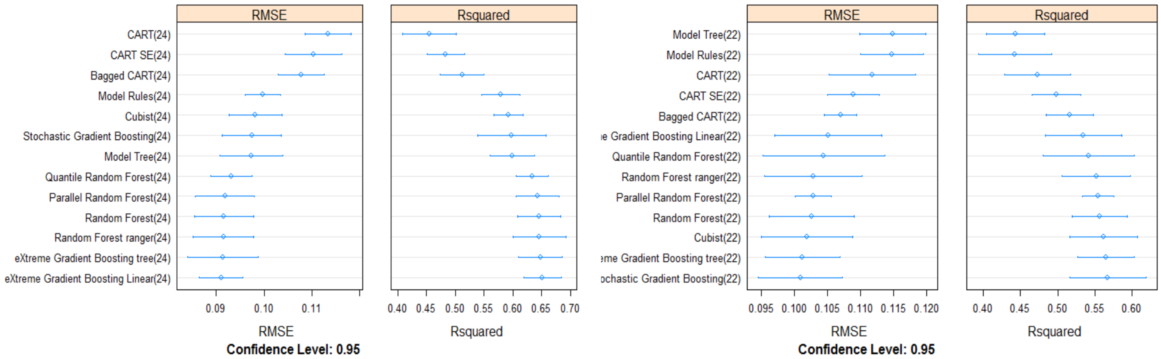

Figure 3. Performance of tree models based on datasets with and without spatial information.

As we can see in the *Neural Network* group at figure 4, again models that fitted by dataset including spatial data have better performances. Lowest RMSE measures are gained by Averaged Neural Network, Bayesian Regularized Neural Network and basic Neural Network respectively.

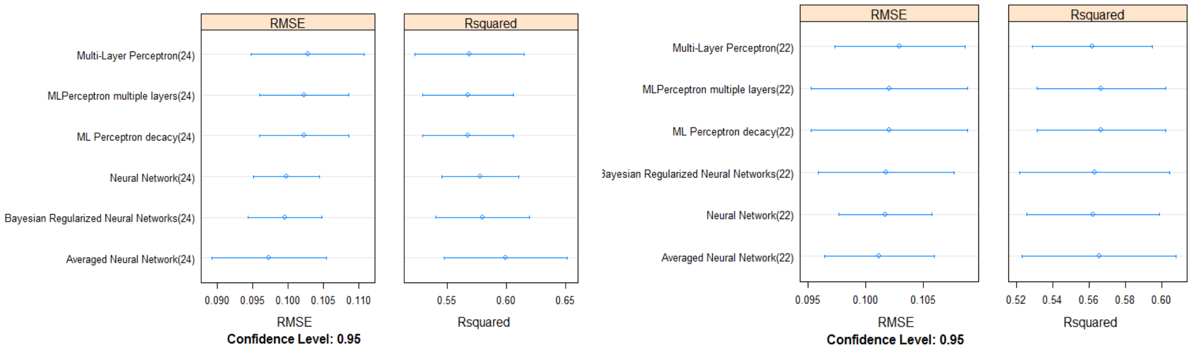

Figure 4. Performance of Neural Networks based on datasets with and without spatial information.

In the Support Vector Machine group (figure 5), non-linear SVM models with spatial information have better performance. The SVM with Radial Basis Function Kernel and Polynomial Kernel are the best models in this group.

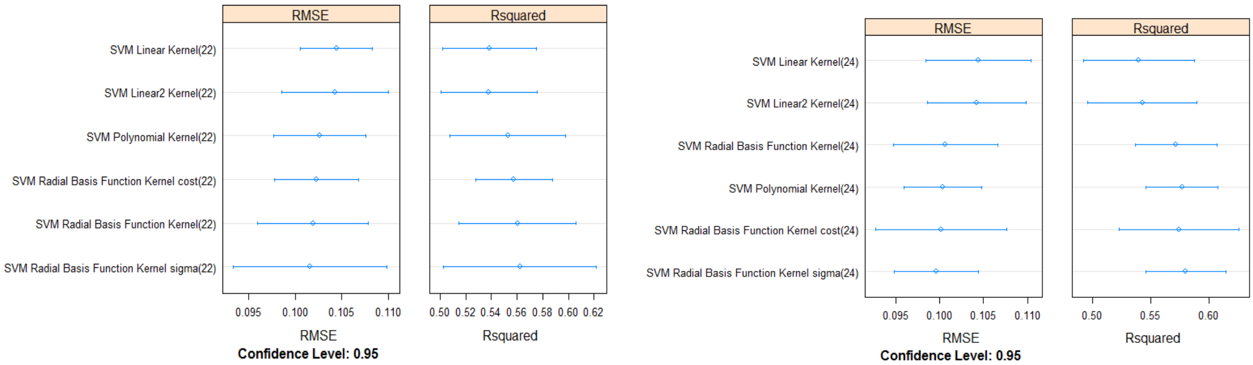

Figure 5. Performance of SVM based on datasets with and without spatial information.

Based on figure 6, related to MARS models, these models are not as good as the SVM, Neural Networks or Tree models. Here, again we can see the positive effect of drive distance and duration on model performance.

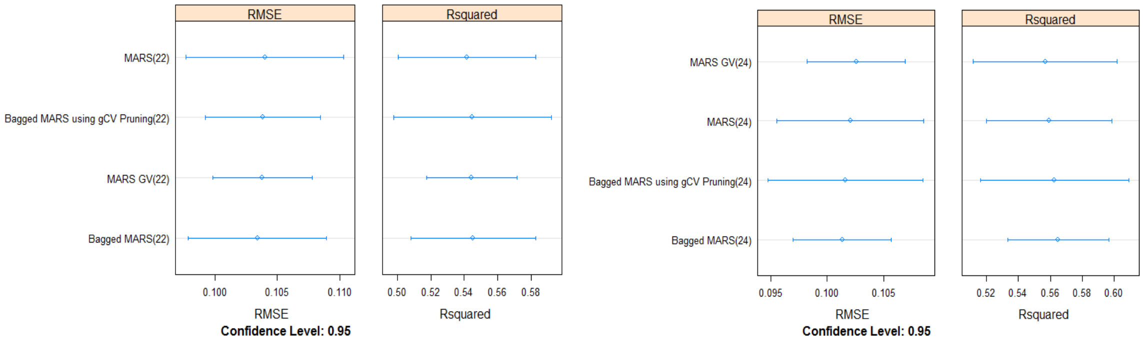

Figure 6. Performance of MARS models based on datasets without spatial information.

As we can see in figure 7, the basic KNN model fitted with dataset including 24 predictors has also less error.

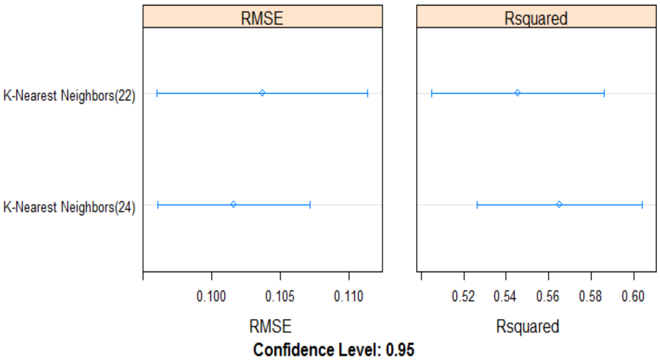

Figure 7. Performance of KNN on datasets with and without spatial information

3.2. *Comparative Analysis*

We select 10 models with highest performance to compare. As we expect there is no linear regression model because of the non-linear relationship between predictors and sale price (figure 8). The nine out of ten best models fitted with data set with spatial information. The four best model are *eXtreme Gradient Boosting Linear*, *eXtreme Gradient Boosting Tree*, *Random Forest* and *Parallel Random Forest* respectively.

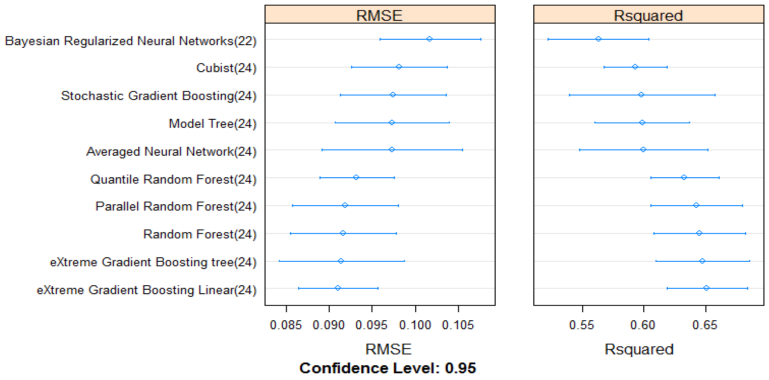

Figure 8. The 10 best models

4. Conclusion and future works

In this paper, vary machine-learning regression techniques were tested on real estate data regarded to 1967 apartments in 19 suburbs in Sydney. Based on each unit's address, the drive distance and duration to CBD was calculated and added to data to enhance the accuracy of sale price prediction. The experiments show the positive effect of drive distance and duration on models accuracy. Based on results, tree models have best performance among other techniques regarded to Sydney real estate data. For future work, distances to amenities can be added to model to reduce the error of prediction.

References

1. Smith, T.R., *Multiple Regression and the Appraisal of Single Family Residential Properties.* The Appraisal Journal, 1971. **39**(2, April): p. 277-84.
2. Mok, H.M., P.P. Chan, and Y.-S. Cho, *A hedonic price model for private properties in Hong Kong.* The Journal of Real Estate Finance and Economics, 1995. **10**(1): p. 37-48.
3. Lewis, O.M., J.A. Ware, and D. Jenkins, *A novel neural network technique for the valuation of residential property.* Neural Computing & Applications, 1997. **5**(4): p. 224-229.
4. McGreal, S., et al., *Neural networks: the prediction of residential values.* Journal of Property Valuation and Investment, 1998. **16**(1): p. 57-70.
5. García, N., M. Gámez, and E. Alfaro, *ANN+ GIS: An automated system for property valuation.* Neurocomputing, 2008. **71**(4): p. 733-742.
6. Hamzaoui, Y.E. and J.A.H. Perez. *Application of Artificial Neural Networks to Predict the Selling Price in the Real Estate Valuation Process*. in *Artificial Intelligence (MICAI), 2011 10th Mexican International Conference on*. 2011. IEEE.
7. Kontrimas, V. and A. Verikas, *The mass appraisal of the real estate by computational intelligence.* Applied Soft Computing, 2011. **11**(1): p. 443-448.
8. Schulz, R., M. Wersing, and A. Werwatz, *Automated valuation modelling: a specification exercise.* Journal of Property Research, 2014. **31**(2): p. 131-153.
9. Zurada, J., A. Levitan, and J. Guan, *A comparison of regression and artificial intelligence methods in a mass appraisal context.* Journal of Real Estate Research, 2011. **33**(3): p. 349-387.
10. *Boosted regression trees: An application for the mass appraisal of residential property in Malaysia.* Journal of Financial Management of Property and Construction, 2014. **19**(2): p. 152-167.
11. Antipov, E.A. and E.B. Pokryshevskaya, *Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics.* Expert Systems with Applications, 2012. **39**(2): p. 1772-1778.
12. Kilpatrick, J., *Expert systems and mass appraisal.* Journal of Property Investment & Finance, 2011. **29**(4/5): p. 529-550.
13. McCluskey, W., et al., *The potential of artificial neural networks in mass appraisal: the case revisited.* Journal of Financial Management of Property and Construction, 2012. **17**(3): p. 274-292.
14. Demong, N.A.R., Lu, J. & Hussain, F.K. 2014, 'Personalised Property Investment Risk Analysis Model in the Real Estate Industry', Human-Centric Decision-Making Models for Social Sciences, Springer, pp. 369-90.
15. Lu, J., Shambour, Q., Xu, Y., Lin, Q. & Zhang, G. 2013, 'a web-based personalized business partner recommendation system using fuzzy semantic techniques', Computational Intelligence, vol. 29, no. 1, pp. 37-69.
16. Demong, N.A.R., Lu, J. & Hussain, F.K. 2012, 'Multidimensional and Data Mining Analysis for Property Investment Risk Analysis', Proceedings of World Academy of Science, Engineering and Technology, World Academy of Science, Engineering and Technology (WASET), p. 608.
17. Kahle, D. and H. Wickham, *ggmap: Spatial Visualization with ggplot2.* The R Journal, 2013. **5**(1): p. 144-161.
18. Grubbs, F.E., *Sample criteria for testing outlying observations.* The Annals of Mathematical Statistics, 1950: p. 27-58.
19. Box, G.E. and D.R. Cox, *An analysis of transformations.* Journal of the Royal Statistical Society. Series B (Methodological), 1964: p. 211-252.
20. Engelhardt, M.K.C.f.J.W.a.S.W.a.A.W.a.C.K.a.A., *caret: Classification and Regression Training*. 2012.
21. Eugster, M.J., T. Hothorn, and F. Leisch, *Exploratory and inferential analysis of benchmark experiments.* Ludwigs-Maximilians-Universität München, Department of Statistics, Tech. Rep, 2008. **30**.
22. Hothorn, T., et al., *The design and analysis of benchmark experiments.* Journal of Computational and Graphical Statistics, 2005. **14**(3): p. 675-699.