TESTING THE GENERALIZATION OF AUTOMATED REAL ESTATE PROPERTY EVALUATION MODELS

Eric TZIMAS[1]

Manolis KRITIKOS[2]

Abstract

The goal of this paper is to analyze the implementation of an automation valuation model in real estate and provide insight regarding its behavior when faced with real world data. An automated valuation model was implemented using two different datasets from Ames Iowa and Athens Greece. The models implemented were a KNeighborsRegressor, a GradientBoostingRegressor, a DecisionTreeRegressor, a Random Forest Regressor, a Stacked Regressor, and a Neural Network. The best scoring model for both datasets was the Random Forest Regressor. Two different methods were used for the evaluation of the above models. These methods include testing using twenty percent of the starting dataset and testing using a custom dataset created by authorized property appraisers. In both techniques, the models scored similarly, with only a three percent difference in accuracy, showcasing the rigidity and robustness of the valuation model when faced with external and quality assured data.

Keywords: Machine learning, real estate, property evaluation, price prediction

JEL Classification: R320 Other Spatial Production and Pricing Analysis

1. Introduction

The real estate sector which is divided into residential real estate, commercial real estate, and industrial real estate, stands as one of the biggest markets in the world and according to recent studies, the residential market was approximated at more than $33,8 trillion [1]. This specific market has been widely affected by the uprising of big data in the last years. Businesses use large volumes of records to assess the state of the mortgage industry and to assess insurance risk. Property evaluations have also played a major role in assessing price trends and geographical fluctuations. Businesses can assess financial risks and pinpoint high yielding investments [2]. Residential property prediction has lately been a topic of interest for scientists since it is a component of critical financial decisions for potential buyers as well as real estate brokers. Individuals and investors want to invest in the housing sector and observe the fluctuations in the real estate market, while these trends also tend to mirror the economic state of any developing country [3]. Due to the latest advances in Big Data technologies as well as Statistics and Machine Learning, real estate property predictions has become a valid field of study where those issues can be resolved by applying deep learning or machine learning algorithms on real transaction data. Application

[1] CTO-Co-Founder Hobsido, Athens, email: errikos.tzimas@gmail.com

[2] Management Science Laboratory, Athens University of Economics and Business, kmn@aueb.gr

of big data techniques in the real estate field can be divided into two categories, forecasting the house price index and real estate price prediction [4]. Mass appraisal is now widely used by real estate companies all around the world for business decisions and the most important asset of this system is the automated valuation model [5].

The first step for creating a machine learning estimator is the collection of data. The real transactional data is hard to find and collect. Scraping techniques can be used, but then arises the problem of duplicate entries, when data is integrated from multiple sources [5]. These duplications can disrupt the training of the algorithm by indicating that these duplicate values have more importance. If duplicates exist in training and testing datasets it can bias the prediction towards these false, duplicate entries [6].

Additionally, it is difficult to collect sufficient data regarding the features of the house, as well as location data and macro variables like mean neighborhood income rate [5]. Data Selection plays a big role in the final accuracy of residential real estate price prediction models as well as the various ETL techniques that should be used to face the problem of automated estimation. The three strongest indicators of a house's price come from durability, heterogeneity, and spatial fixity. Durability of a property indicates the duration of which a house can keep its price due to good construction and a good aging process in the market. Heterogeneity stems from all the factors that differentiate one property from another such as numbers of bathrooms or bedrooms. Spatial fixity refers to the location of the property. The five aspects of locations are the distance from the analyzed property, the socioeconomic character of the location, the natural traits of a location and the local government that prevails in the area. Some basic macroeconomic factors that should be taken into consideration in conjunction with real transaction data is the GDP of a country, the GNP, and the consumer price index [7].

Analyzing this problem at scale, it should be noted that integration of heterogeneous data from different sources should be considered as the source of information for a scalable system. Building a robust model requires data in high velocity and volume to keep up with fluctuation in market prices and updates in the real estate sector as well as evaluating the constantly changing macroeconomic features of a country. A rise in the real estate market can be attributed to the increasing income of the inhabitants of a certain area, but this can change over time. Careful analysis has led us to the conclusion that the factors that indicate a real estate property's price can change over time [3].

The information needed to provide accurate real estate pricing predictions is mostly derived from quantitative and qualitative data. Thus, it is important to include enough of both feature classes in the analyzed dataset. Quantitative data consists of macroeconomic factors like GDP per country, Business Cycles, and unemployment metrics. Qualitative metrics are composed from building styles and living environment, but it has been proven to be scarce in availability and various problems seem to arise in the collection of these attributes [8]. In his analysis of random forest models for the mass appraisal of residential property in South Korea, Jengei Hong used the following variable groups to train the model, Structural attributes, Neighborhood attributes, Locational attributes, and Macro variables. These variables comprise a succinct template for a basic model as they sufficiently fulfill the prerequisites for macroeconomic features as well as estate features. These variables can also address the durability of the property, in regard to the construction year, the macro

variables (transaction period, land price fluctuation etc.). Regarding heterogeneity the chosen features might not be sufficient [9].

Identifying the correct dataset or after extracting and loading from various data sources the data should be explored and visualized to study its characteristics, something that can critically affect the model's success. All different parameters and relations should be investigated and evaluated, and tests should be run to identify possible outliers or invalid values [10]. After visualization, data pre-processing must take place. Data should be cleaned, and the outliers should be removed. The outliers could be removed using the interquartile range technique [8].

Regarding missing values, we use the values of the records with the most closely related features or, if dealing with categorical features with uneven distribution, replace null values with the most highly frequent category [11]. Additionally, numerical values can be replaced by the mean of the feature [12]. For models that do not directly work with categorical variables, categorical variables should be transformed using one-hot encoding. This method transforms n categories into n new binary features [12]. The variables could also be scaled on a scale of 1-5 based on a table of assumption connected with the significance of each variable to cover the set of features that is the most valuable [7]. Finally, data can be normalized by subtracting the mean value of each feature and dividing it with its standard deviation. This technique can speed up learning and lead to more accurate predictions [12].

2. Proposed methodology

The aim of this prototype is to assess the efficacy of a real estate prediction model and its robustness when faced with high variance data. The tools used for this prototype are Python and specifically the Pandas library, for data exploration, data cleaning, model creation and model performance assessment. Specifically for the creation of machine learning models the Sci-kit python library was used. For visualization matplotlib and seaborn libraries were used. The models are going to be tested using two different datasets from different countries and cities, containing different amounts of records. These two datasets are going to be opposite in terms of shape, meaning that the first one is going to contain a good amount of features while being small in record size and the second one is going to contain a large number of records with a limited feature size. Various techniques are going to be tested in terms of data pre-processing and model tuning but only the most successful techniques are going to be presented. Finally, the prototype is going to be tested, using 20% of the starting dataset, and using real world estimations done by human appraisers to test its generalization and rigidity.

The datasets used are two vastly different sets of real transactional data from different countries and cities. The first dataset comes from the website kaggle.com and it contains a wide range of characteristics for each property in Ames, Iowa for houses sold between 2006 and 2010. This dataset contains a small number of records but with many features. Namely, 1460 house transactions with 78 features per record.

The second dataset comes from multiple website listings in Athens Greece. It contains 69.823 records of website listings with 27 features for each record. The dataset is composed only of apartment records and the features mostly describe the aspects of the property.

Beginning with data exploration on the Ames housing dataset, each continuous feature of its records was analyzed regarding the count of occurrences in the dataset, the mean, the standard deviation, the range, the maximum and minimum values and the 25th, 50th and 75th percentiles. The mean sale price for the dataset is 180.921 dollars. The frequency of the houses compared to a continuous variable was visualized to comprehend the various distributions of our records. In Figure 1, the frequency of the records fitted into different bins based on their sale price is presented, most properties are priced between 100.000 and 150.000 Euros.

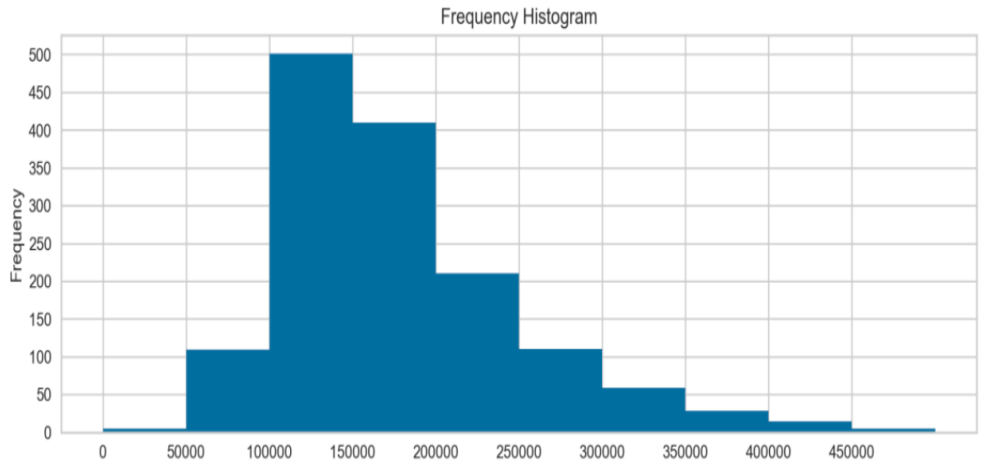

Figure 1. Frequency of properties based on price range for Ames dataset

The features of the dataset were segmented based on their type. The columns containing categorical features were factorized to deal only with numerical values from now on. To deal with null values, all empty values were replaced with the mode of the feature, if the feature was categorical, or the mean of the feature if it was continuous. Furthermore, as proposed in international literature, the categorical features were transformed using one-hot encoding. This means that a feature like MSZoning that identifies the general zoning classification of the sale and can take values: A, C, FV, I, RH, RL, RP and RM for Agriculture, Commercial, Floating Village Residential, Industrial, Residential High Density, Residential Low Density, Residential Low-Density Park and Residential Medium Density respectively, we transpose one column with n different values to n different columns, one for each value.

3. Computational Results

In this section, the models that were implemented are going to be presented and assessed, based on Mean Absolute Error, Mean Square Error and Accuracy. Accuracy is considered as 1 - MAPE, with MAPE the Mean Absolute Percentage Error. The algorithms used were KNeighborsRegressor, GradientBoostingRegressor, DecisionTreeRegressor, Random

Forest Regressor, Stacked Regressor, Deep Learning technique, and an Artificial Neural Network.

It should be noted that the data split for the models was 80% training data and 20% data used for testing. The first algorithm implemented was KNeighborsRegressor. The number of neighbors for each iteration was set to 20 while the metric that was used was the Euclidean. The model achieved mediocre results with an explained variance for the testing set of 0.6, a maximum error for the testing set of 501.857 dollars, a Mean Absolute Error of 32.060 dollars, r-squared of 59% and accuracy of 80.72%. The second algorithm implemented was a Decision Tree Regressor. The algorithm was tuned using the cross-validation Grid Search to increase its performance, since it was one of the highest-ranking algorithms. The criterion used was the Mean Square Error, while the trees were limited to a maximum depth of 9. Additionally, every leaf of the trees was limited to containing a minimum of 20 samples. As mentioned before, this algorithm performed sufficiently with an explained variance for testing of 0.8, a mean absolute error of 24.723 dollars, r-squared for testing 80% and an accuracy of 85.18%. In the Figure 2, the distribution of records with their residuals is depicted.

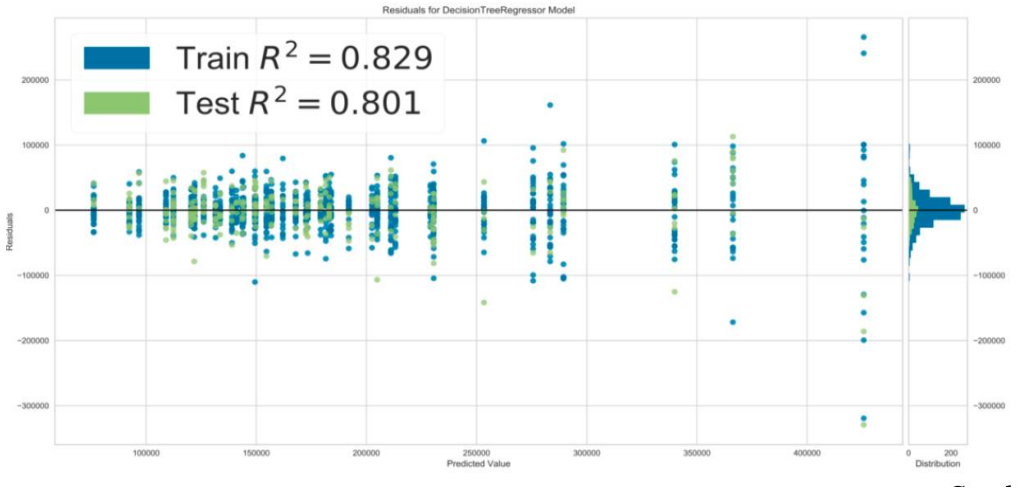

Figure 2. Price distribution and residuals for Decision Tree Regressor

The next algorithm implemented was a Gradient Boosting Regressor, improving the performance of a Decision Tree Regressor. The algorithm displayed an explained variance for testing of 0.88, a maximum error for testing 262.635 dollars, a mean absolute error for testing of 17.785 dollars, mean square error 908.281.353 dollars, r-squared of 88% and accuracy of 89.74%. These are surprisingly good results, without adding any macro - economic factors into our dataset. The predictions on the testing dataset were close to the original values, except for some properties that may be outliers. In the graph below the original and predicted price of each property is presented in blue and green respectively. Additionally, the distribution of the percentage error seems to be gravitating towards zero which indicates a good fit of the model. The graph in Figure 3 shows the frequency for the percentage error of our records, the distribution seems to gravitate towards zero, indicating a good model fit.

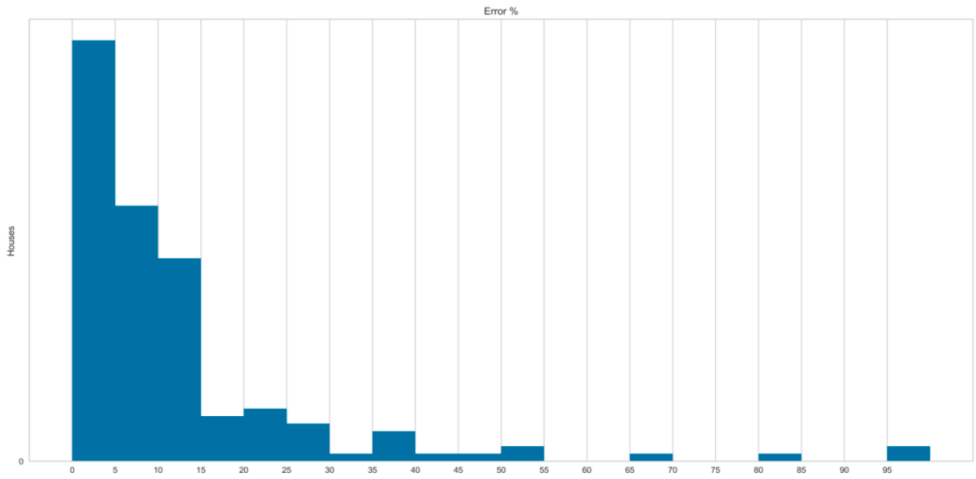

Figure 3. Frequency of properties based on Error Percentage for Gradient Boosting Regressor

The next algorithm implemented was Random Forest. The criterion used for this algorithm was mean absolute error and the number of estimators was set to 150. This algorithm achieved high performance, as expected, since this was the algorithm that the literature indicated as the most suitable for a real estate price prediction model. The algorithm achieved an explained variance for testing of 0.89, a maximum error of 221.737 dollars, a mean absolute error of 17.431 dollars, an r-squared of 0.89 and an accuracy of 89.42%.

The next model implemented was a stacked regressor, using multiple regression algorithms stacked 'on top' of each other. The first regressor used was a Decision Tree Regressor with the same tuning of the hyperparameters as before. Mainly, the criterion used was the Mean Square Error, while the trees were limited to a maximum depth of 9. Additionally, every leaf of the trees was limited to a minimum of 20 samples. The model produced an explained variance of 0.77, a maximum error of 396525, a mean absolute error of 26.521 dollars, r-squared of 77% and accuracy of 84.66%.

Finally, the last model implemented was an Artificial Neural Network, with three hidden layers. The input dimension was set to 96, for the 96 different features which were augmented due to the one-hot encoding. The activation function of the first and second layer was reLU while for the last layer was linear. The loss metric for each epoch was set to mean squared error and the optimizer was adam. The model was trained for 100 epochs because of the size of the dataset as well as to avoid overfitting.

Regarding the models trained with 69.823 apartment website listings from Athens, Greece, the same process as for the Iowa dataset was followed for the most part. Specifically, the data preparation was significantly more tedious due to the format of the data and the fact that, unlike the Iowa dataset, this dataset required extensive cleaning to be useful. It should also be noted that the addresses of the properties were transformed to coordinates. The two datasets, apart from containing a symmetrically opposite shape, were also trained to predict

different values. In the Iowa case, the property's price was the dependent variable. In this case the price per square meter is going to be the dependent variable.

The best performing algorithm for both Ames and Athens datasets using the 80/20 split as testing was Random Forest. In addition to the testing that was the 20% of our starting dataset, we outsourced 30 apartments to real estate brokers, to create a small custom testing set that will not be subjective to the seller's opinion, as it is, when sellers upload their properties on the web.

In Table 1, the results from the custom testing as well as the testing using the training data for the Random Forest Regressor for Athens are presented:

Testing	**MAE**	**MSE**	**R-SQUARED**	**ACCURACY**
RF model	225.2	94046.49	71%	87.02%
Custom Testing (External brokers)	**MAE**	**MSE**	**R-SQUARED**	**ACCURACY**
RF model	257.88	122932.36	65%	84.74%

Table 1. Random Forest Accuracy using two testing techniques

Both testing techniques proved that the Random Forest algorithm can achieve almost the same accuracy with significantly less features when faced with a sizable dataset. Although the data was collected from website listings, which reduced the quality compared to the Iowa dataset, the accuracy of the model in Athens was only 3% less than the one from Iowa. Additionally, the accuracy from the testing set produced by the real estate brokers, performed right in par with the 20% testing of our data, proving that the model generalizes its predictions, while it can be improved by tuning the algorithm using Grid Search or other tuning - boosting techniques and cleaning the data furthermore.

4. Conclusions

Data collection for transactional data can prove to be very tedious, because of this, most companies tend to collect data from websites containing property listings. The problems that arise is that data needs extensive cleaning, and the price can be subjective or plainly inaccurate. Collecting data is crucial for the training of the models as well as testing the performance of them. Thus, it is important to deal with these issues optimally. After thoroughly cleaning the data, various thresholds can be set in order to identify records that do not pertain to valid properties. The thresholds can be implemented by using limitations in different combinations of features, such as size and price. Additionally, in order to accurately assess the generalizability of the models, custom datasets should be set, by

creating pipelines that connect to various other sources of data. These sources can come from transactional data, or new appraisals implemented by brokers and not systems. This study indicated that the performance of the models increases vastly when the dataset is limited to locations that are beyond a certain density threshold. The density of a location indicates the number of properties that reside on it. Data Preprocessing is also a factor that will critically affect the performance of the automated valuation models. The random forest model seems to improve when we transform any categorical value that indicates a spatial location to a number that correlates with all other classes. The algorithm also improves when applying one-hot encoding to all categorical features. Models like the Random Forest, Gradient Boost and Decision Tree, seem to achieve good results when faced with an automated valuation problem, the Neural Network, although it performed poorly, it achieved much better results when trained with the more sizable dataset from Athens. These models can greatly improve if more data can be collected to fulfill the density of the locations, as well as the variety of the features, or by applying boosting techniques and feature selection algorithms. In very large datasets, distributed ways of handling data can be introduced, to distribute computational load across different processing nodes and increase the speed of ETL processes of a systems pipeline as well as the training of the models.

References

[1] Depersio, G. (2021). What Are the Main Segments of the Real Estate Sector? Available at https://www.investopedia.com/ask/answers/052715/what-are-main-segments-real-estate-sector.asp

[2] Treistman, H. (2020). How Big Data Is Transforming Real Estate, 1.

Available at https://brightdata.com/blog/leadership/how-big-data-is-transforming-real-estate

[3] Imran. (2021). Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data. *Soft Computing and Machine Intelligence Journal*, *1*(1), 11-23.

[4] Mohamad, J., (2020). Heritage property valuation using machine learning algorithms Available at http://www.prres.net/papers/Mohamad_Heritage_Property_Valuation_Using_Machine_Learning_Algorithms.pdf

[5] Niu, J., & Niu, P. (2019). An Intelligent Automatic Valuation System for Real Estate Based on Machine Learning. *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, (12), 1-6.

[6] Bhushan Jha, S., (2020) Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study Available at https://www.researchgate.net/publication/342302491_Housing_Market_Prediction_Problem_using_Different_Machine_Learning_Algorithms_A_Case_Study

[7] K.C. Lam, C.Y. Yu & C.K. Lam (2009) Support vector machine and entropy based decision support system for property valuation, *Journal of Property Research*, 26(3), 213-233.

[8] Pai, P., & Wang, W. (2020). Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices, 1-11. Available at https://www.mdpi.com/2076-3417/10/17/5832/htm

[9] Hong, J., Choi, H., & Kim, W. (2019). A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management*, *24*(3), 140-152.

[10] Louati, A. (2021). Price forecasting for real estate using machine learning: A case study on Riyadh city, 1-16. Available at https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.6748

[11] Fan, C., Cui, Z., & Zhong, X. (2018). House Prices Prediction with Machine Learning Algorithms, *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, 6-10.

[12] Yazdani, M. (2021). Machine Learning, Deep Learning, and Hedonic Methods for Real Estate Price Prediction, 1-33. Available at https://arxiv.org/abs/2110.07151

Bibliography

Bhushan Jha, S., (2020) Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study Available at https://www.researchgate.net/publication/342302491_Housing_Market_Prediction_Problem_using_Different_Machine_Learning_Algorithms_A_Case_Study

Depersio, G. (2021). What Are the Main Segments of the Real Estate Sector? Available at https://www.investopedia.com/ask/answers/052715/what-are-main-segments-real-estate-sector.asp

Fan, C., Cui, Z., & Zhong, X. (2018). House Prices Prediction with Machine Learning Algorithms, *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, 6-10.

Treistman, H. (2020). How Big Data Is Transforming Real Estate, 1.

Available at https://brightdata.com/blog/leadership/how-big-data-is-transforming-real-estate

Imran. (2021). Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data. *Soft Computing and Machine Intelligence Journal*, *1*(1), 11-23.

K.C. Lam, C.Y. Yu & C.K. Lam (2009) Support vector machine and entropy based decision support system for property valuation, *Journal of Property Research*, 26(3), 213-233.

Louati, A. (2021). Price forecasting for real estate using machine learning: A case study on Riyadh city, 1-16. Available at https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.6748

Mohamad, J., (2020). Heritage property valuation using machine learning algorithms. Available at http://www.prres.net/papers/Mohamad_Heritage_Property_Valuation_Using_Machine_Learning_Algorithms.pdf

Niu, J., & Niu, P. (2019). An Intelligent Automatic Valuation System for Real Estate Based on Machine Learning. *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, (12), 1-6.

Pai, P., & Wang, W. (2020). Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices, 1-11. Available at https://www.mdpi.com/2076-3417/10/17/5832/htm

Hong, J., Choi, H., & Kim, W. (2019). A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management*, *24*(3), 140-152.

Yazdani, M. (2021). Machine Learning, Deep Learning, and Hedonic Methods for Real Estate Price Prediction, 1-33. Available at https://arxiv.org/abs/2110.07151