HOUSE PRICE PREDICTION

Arjun Avadhani, Disha Singh , Vaishnavi R Bhat

Abstract–We study the problem of efficient prediction of house prices based on various physical factors, locations and years of establishment. By relying on expensive sampling techniques or computationally heavy pre/post-processing steps, most existing approaches are only able to be trained and operate over small-scale house datasets. In this paper, we explore various research papers and the different methods that they have used in prediction. From simple regression techniques to complex deep neural networks, any method can be used in house price prediction.

Index Terms– Linear Regression, Machine learning, Support Vector Machines, random forest, Boosting, LSTM

1 INTRODUCTION

EFFICIENT prediction of house price is a fundamental and essential necessity for affordable and reasonable housing across the world. In the past years, Machine learning has proven to be able to solve real world problems using various algorithms. It plays a major role in advances of medical imaging, spam and fraud detection, enhancements in automobile industry, security alerts and Business Analysis. In this paper, we have used machine learning algorithms to perform predictive analysis of house prices to provide an overview of real estate businesses and property demand. Data is the most important part for analysis of any problem. It provides the information in a detailed format which is able to be understood by machines.

In the contemporary digital era, useful information of the society can be retrieved from a wide variety of sources and stored in the form of structured, unstructured and semi-structured formats. In the analysis of economic phenomena or social observations, advancement of innovative technology makes it possible to systematically extract the relevant information, transform them into complex data formats and structures, and then perform suitable analyses. Because of these new circumstances, traditional data processing and analytical tools may not be able to capture, process and analyse highly complex information in the social and economic worlds. New techniques have been developed in response to the treatment of the colossal amount of available data.

In this paper, we choose three algorithms as examples to illustrate how machine learning algorithms can be utilised to predict housing prices with the input of conventional housing attributes. We use these techniques for a number of reasons.

- *https://www.tandfonline.com/doi/full/10.1080/09599916.2020.1832558#:~:text=Based%20on%20conventional%20evaluation%20criteria,technique%20to%20predict%20property%20prices..*
- *Corresponding author: Bo Sing Tang , Siu Wai Wong*

First, SVM is used because it works well with high dimensional data, semi-structured and unstructured data, and is seen as a more powerful tool to make accurate predictions (Hassanien et al., 2018). In the commercial world, the use of SVM is very popular in predicting a company's sale volume and revenue. Second, due to many decision trees participating in the process, random forest (RF) is used because it can reduce over-fitting of data. Previous property research has also demonstrated that random forest is a robust algorithm which provides accurate predictions (Mohd et al., 2019; Mullainathan & Spiess, 2017; Pérez-Rave et al., 2019). Third, gradient boosting machine (GBM) is also used because it is considered as an emerging machine learning algorithm with highly flexibility. It also provides better accuracy than many other machine learning algorithms, as suggested by Kaggle Data Science Competition (Kaggle, 2019).

2 CRITIQUE

The papers which we have reviewed use advanced regression methods and neural networks to solve the house rent prediction problem. Some examples of that include Lasso Regression, Gradient Boosting Method, Tree Regression, Random Forest, Stacked LSTM, Stacked generalization, etc.

Assumptions: No assumptions made.

3 SUMMARY

Standard regression techniques like Simple and Multiple Linear regressions have a lot of limitations. Therefore, Machine Intelligence algorithms are used which helps in improving the efficiency of the model and gives us more efficient and better results. It also helps in reduction of the error factor. It provides more accuracy.

We can use textual information for our house rent prediction analysis as well.

Since the main strength of ML approaches is their predictive capability, it outperforms OLS regression especially when number of determinants is large in the case of in-sample prediction. The Boosting , Forest and bagging models have an excellent overall performance score of 100% while the tree model is at 33%. The ridge and LASSO models fail miserably at 0%. In the case of out-of-sample performances (accounting for spatial lag autocorrelation) the performances are : Ridge : 83% LASSO : 83% Tree : 33% Bagging : 83% Forest : 83% Boosting : 100% For out-of-sample predictions (accounting for spatial error autocorrelation , Ridge - 66% LASSO-66% Tree-33% Bagging-100% Forest-83% Boosting-100%. For out-of-sample predictions(accounting for spatial autocorrelation) the performances are Ridge - 83% LASSO-83% , Tree - 33%, Bagging -83%, Forest - 83%, Boosting-100%

The prediction in baseline model is much better than baseline ARIMA, about 90 percent reduce on MSE. The stateful LSTM model has bad prediction on validation data set. Which indicate that the essence of the problem may not suitable for the application of stateful LSTM. The stacked LSTM has similar accuracy with Basic LSTM. In fact, deeper neural network is expected to give a better result than the simple one, which was not seen in evaluation. Which inspire us to do further exploration on finding better structure and parameter for stacked LSTM and improve the accuracy. As we found several problems. In future, we will do more research on stacked LSTM, which is just simply implemented using fixed structure and parameters. Given more time, the even comprehensive study on stacked LSTM may be done. At the same time. Poor data quality and amount limited the accuracy of the model. By introducing bigger dataset and handle the data smarter, the prediction may be improved dramatically.

In Random Forest Modeling, a perfect prediction model must accommodate the specific variables that influences the price of a house in the region being considered. This study has further affirmed the prowess of random forest machine learning technique in predicting the prices of a house based on variables made available in Boston housing dataset. A comparison of the predicted and actual prices sown in Table 1 revealed that the model achieved a prediction difference of ±5. This showed that the model can be used to predict house prices. Several other machine learning models especially deep learning models can also be explored for house price prediction.

The best results belong to Random Forest for the training set and Stacked GeneralizationRegression for the test set. Since 49 of 58 features of the one-hot-encoded dataset were boolean values, it is reasonablethat the Random Forest worked well on this dataset. Unlike the three traditional machine learning methods, Hybrid Regression and Stacked Generalization Regressionwere neither tuned nor implemented sophisticatedly but delivered promising results on the training set and test set. Since Random Forest was proven to be overfitting, Hybrid Regression could be considered as the best model ontraining set where the RSMLE is 0.14969. Surprisingly, Stacked Generalization Regression did not work well on thetraining set as Hybrid Regression, but this model did exceptionally on the test set.

All algorithms except SVR show excellent performance . Among tree based algorithms RFR and ETR perform better than the others all the rmse %rmse mae %mae show an increase whic says the predictions made are valid and credible we also find out the relative order of importance of all the determinants. Machine learning techiques can achieve high level of accuracy . textual information leads to lower overall errors performance of models with textual information was weaker overall fit was better.

4 LIMITATIONS

The journal where we predict house rents does not calculate error, accuracy, precision etc and hence does not include it in its findings. Many iterations of performance tuning were done to find the optimal solution of each model. Random Forest Regression, XGBoost, and LightGBM were intensively tuned by function GridSearchCV provided by scikit-learn which increases run time. Random Forest was prone to overfitting.

In the modelling of house rentals in Atlanta, there is no seperate error / accuracy checking mentioned . it is mostly embedded in the various MI algorithms that have been used for the analysis process.

A comparison of the predicted and actual prices reveal that the model achieves a prediction difference of ±5. There are other models that give accurate prediction. Stateful LSTM model has bad prediction on validation data set which indicate that the essence of the problem may not suitable for the application of stateful LSTM. The result in some districts are not ideal. The possible reasons are low frequency of data and the loss of data in several months.

LACUNA— the journal does not any form of missing data and our data set has no missing values.

5 PROBLEM STATEMENT WITH SPECIFIC ISSUE WE INTEND TO ADDRESS

We are taking into consideration all the various parameters given in our kaggle dataset and building a model whch will help us to predict the price of any given house .

6 HOW IS OUR APPROACH DIFFERENT FROM OTHERS?

we are using a combination of different methods that we have learnt through our paper reviews . The various Machine learning algorithms will help us t achieve more efficient results .