Analysing Real Estate Ads with Data Science and Inferring Patterns

2021-01-11

Abstract

In the 21st century, with the ever-increasing demand for properties in Singapore, it is becoming increasingly difficult to accurately determine the value of properties based on the variables of a property, such as location and price per square feet (psf). Besides these variables, the physical appearance of a property, often represented by images, often plays a crucial role in determining the value of a property. In our project, we create a predictive regression model using machine learning to predict the value of a property based on its features, location, amenities and images. We also create a similar model without the image data to determine the effectiveness of using images in increasing the accuracy of our predictions. (results). (further discussion)

Introduction

The aim of our project is to use machine learning (ML) to analyze real estate advertisements and determine property prices.

While the value of properties can be determined manually by a professional, this process is highly tedious and costly. Therefore, we seek to create a machine learning algorithm which would allow people to determine the value of a property in a reasonably accurate, cost- and time-effective way.

For our project, we limit our scope to properties within Singapore only. We acquire a database of property listings, and their respective images, scraped from PropertyGuru. The data is then used to train and evaluate two machine-learning algorithms which will be designed to predict the value of a property.

Methodology

We acquired a dataset of property listings with their images from PropertyGuru, which is then split into two subsets. The first subset, containing 16,000 listings, is used to train an algorithm to make inferences about the value of a property based on its images, and the second subset, containing 4,000 listings, is used to train a regression model to predict property prices.

Our image regression model first extracts image-property value pairs from the first dataset. The images in each pair are transformed into feature vectors using a multichannel Histogram of Oriented Gradients (HOG) transformer from the scikit-image library. Missing property prices are imputed with the mean value, and outlier prices are detected using IsolationForest from the scikit-learn library and removed from the dataset. The resulting dataset is then fitted into a KerasRegressor, which is a customisable perceptron model from the TensorFlow library. From the second dataset, basic features (see above) are extracted, along with the respective images for each listing. The

trained KerasRegressor model is used to assign a value to each image. This value represents the predicted value of the property from the image. The mean value of all images of a property is then calculated.

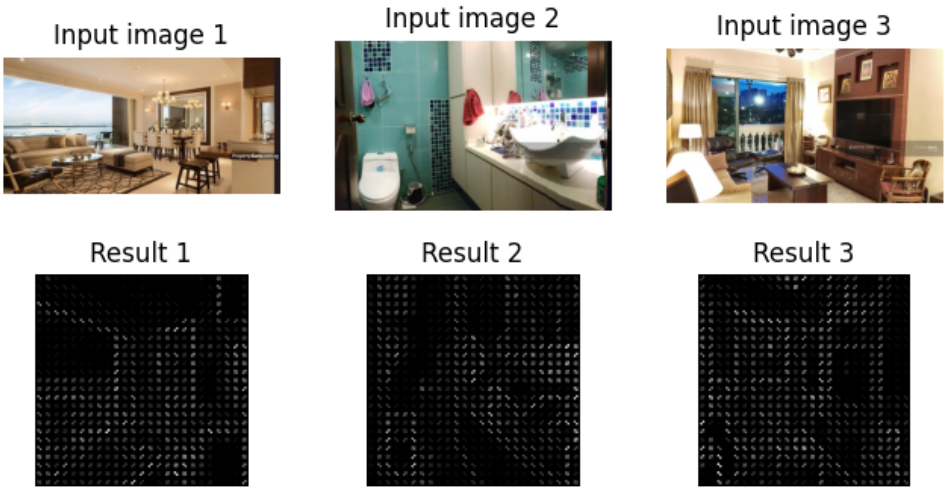

Figure 1: Transformation of JPG images to HOG feature vectors

The mean values, along with the basic features, form the feature vectors. Missing data is imputed, either with the mean value, in the case of continuous variables, or the most frequent value, in the case of categorical variables. Categorical variables are encoded using a one-hot encoder, in the case of nominal variables, or an ordinal encoder, in the case of ordinal variables, in order to form a feature vector for each listing. Each feature in the feature vector is then scaled to zero mean and unit variance. A DecisionTreeRegressor from scikit-learn is used to train the model to predict the price of the property from its features as well as the image value.

In order to measure the effectiveness of including image data in our machine learning algorithm, we also include a control group by repeating the same procedure, except without the image values. The overall effectiveness of the full machine learning algorithm, as well as the control group, are compared using cross-validation. KFold splitting is used to create 5 train-test sets for each method, which are evaluated iteratively. Thereafter, the R^2 value for each train-test set is determined. The mean R^2 value for each method is then calculated and compared.

Results

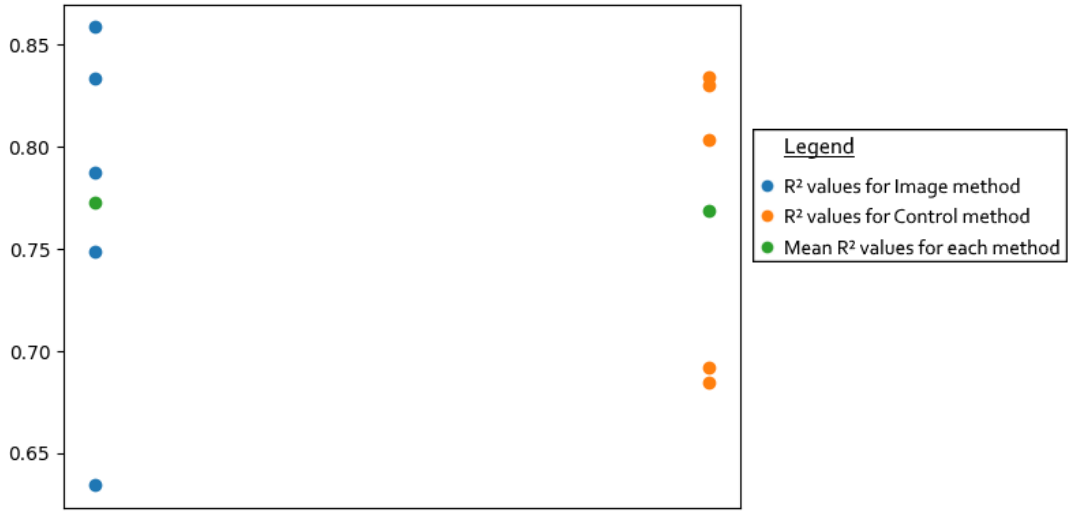

Figure 2: The final R^2 values for each method.

Figure 2 shows R^2 values for each of the 5 folds for both prediction methods, along with the mean R^2 value. The mean R^2 value for the Image method is 0.773, and that for the Control method is 0.769; there is a difference of 0.369% between the 2 means.

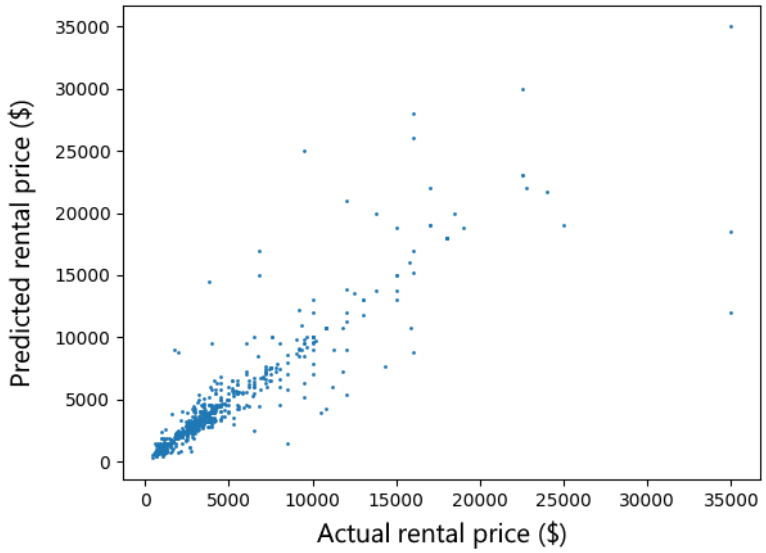

Figure 3: Graph of predicted rental price against actual rental price

Figure 3 demonstrates the ability of the predictive model to accurately predict property prices from its features and images. Though there exists several outliers as can be seen from the above figure, the predicted rental price is mostly consistent with the actual rental price with high levels of accuracy.

Conclusion

We find that machine learning is a viable method for predicting property prices based on its features. While our regression model was capable of making inferences about the value of a property from its images, this only accounted for a small increase in the accuracy of our model.

Discussion

Manual labelling of images was attempted, however this method was not effective. The process is tedious and time consuming, and with the time and manpower available to us, would not produce sufficient data to produce a model that could accurately label images. Furthermore, even if image labels were accurate, they would not be good predictors of property value, since they do not take into account the quality, design or style of each item in the image, which is more important in determining the value of a property than the label itself.

Due to the nature of Singapore's real estate market and that of PropertyGuru listings, many outliers were present in our dataset, which significantly reduced the measured accuracy of our model. While it would be possible to remove outliers from our dataset through an algorithm such as Isolation Forest, we did not attempt this for our final predictive model as this could have resulted in significant overfitting and inflated accuracy.

Future work

Singapore's real estate market is relatively small as compared to other big cities and countries, and is different as well. Trends, demand and supply and certain other characteristics of the Singapore market may not apply to other markets. In general, Singapore properties are also significantly more expensive than those in other places. By expanding our scope to other markets, analyzing their trends and characteristics, and understanding how their markets are different from Singapore's, we will be able to apply our machine learning algorithm to properties in those places. This way, fewer people will need to rely on professionals for property valuation, thus buying and selling of properties would be made easier. At the same time, comparing other markets with Singapore's will also give us a more diverse and open understanding of the real estate market in general, and can be useful in improving our machine learning algorithm in future updates.

References

[1] Law, S., Paige, B., & Russell, C. (2019, October 21). Take a Look Around: Using Street View and Satellite Images to Estimate House Prices. Retrieved October 31, 2020, from `https://arxiv.org/abs/1807.07155`

[2] Naik, N., Raskar, R., & Hidalgo, C. (2016, May). Cities Are Physical Too: Using Computer Vision to Measure the Quality and Impact of Urban Appearance. Retrieved October 31, 2020, from `https://www.aeaweb.org/articles?id=10.1257%2Faer.p20161030`

[3] Glaeser, E., Kincaid, M., & Naik, N. (2018, October 22). Computer Vision and Real Estate: Do Looks Matter and Do Incentives Determine Looks. Retrieved October 31, 2020, from `https://www.nber.org/papers/w25174`

[4] Poursaeed, O., Matera, T., & Belongie, S. (2018, October 03). Vision-based Real Estate Price Estimation. Retrieved October 31, 2020, from `https://arxiv.org/abs/1707.05489`

[5] S. Bell, K., Pagourtzi, E., J. Benjamin, G., S. McGreal, A., C. Bagnoli, H., Byrne, P., . . . LC. Chen, G. (2018, April 03). Vision-based real estate price estimation. Retrieved October

31, 2020, from
https://link.springer.com/article/10.1007/s00138-018-0922-2

[6] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[7] Van der Walt, S., Schönberger, Johannes L, Nunez-Iglesias, J., Boulogne, Franccois, Warner, J. D., Yager, N., ... Yu, T. (2014). scikit-image: image processing in Python. PeerJ, 2, e453.

[8] Jason Brownlee: Your First Machine Learning Project in Python Step-By-Step
https://machinelearningmastery.com/machine-learning-in-python-step-by-step/

[9] Gilbert Tanner: Introduction to data visualization in Python
https://link.medium.com/U8I5oL2OL8