Improving Property Price Predictions Through Features Extracted From Satelite Images

Jakub Miksa [13372734]
University of Amsterdam

Abstract

Real estate property values may depend on many different assets. Often similar properties may differ based on the visual aspects. In this study, features extracted from satellite data using Convolutional Neural Network were used to verify whether they can increase the robustness of the house price prediction model.

It was found that those features reduced the prediction error by 10% on New York City properties data.

Introduction

Estimating real estate prices are one of the most popular problems in statistics. However, traditional approaches do have some limitations. Properties that have similar tangible assets may differ in price due to some context information. Housing prices datasets often do not contain information about the visual aspects of the properties that could affect the prices. Another factor may be the area surrounding the property that is not taken under consideration in traditional models.

The focus of this paper is building a robust model predicting property prices and then illustrating that using features taken from satellite images has a positive effect on model predictions.

Data taken from New Your City properties are first sampled into the train, test and validations set. Several models are then used to predict prices from which two best are selected. Finally, image features learned from Convolutional Neural Networks are added to show their positive impact.

Data

Real estate sales data is taken from the public NYC Geodatabase with geocoded data of 84,768 real estate sales in New York City in 2015[1]. The original data set contains 30 attributes describing the property and its sale price. By comparing the correlation of the attributes with the price as well as deleting those with too many missing values, 9 categories were selected: zip, residential units, commercial units, total units, land square feet, total square feet, tax class and building class at the moment of sale. Dummy variable were created for categorical variables before putting them into the model.

56,959 properties were found to have 0 total square feet, total units or price. Those values were deleted from the datasets. Additionally, 596 price outliers were deleted by excluding rows deviating from mean by more than one standard deviation. Lastly, 655 outliers were removed by visualizing scatterplot of price against other categories. After deleting the outliers correlation between categories and price increased.

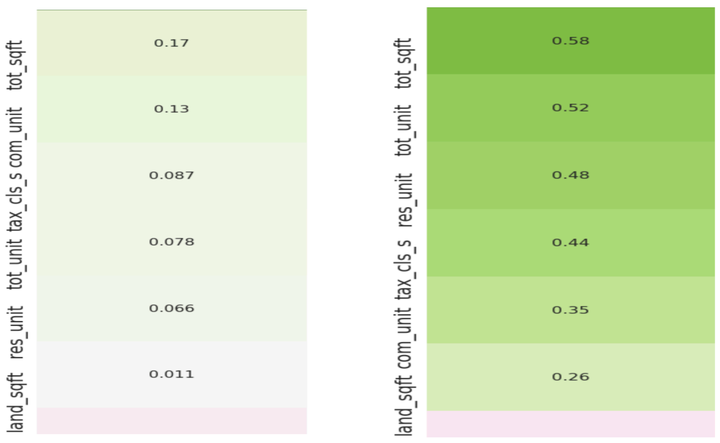

Figure 1: Correlation for 6 highest correlated categories before and after cleaning the data

Satellite image data is provided by the New York City Department of Information Technology & Telecommunications. The images were taken in 2018[2]. The images show each property in the centre and its surrounding withing 500ft radius.

Figure 2: Satelite RGB Image

Methodology

Several regression models were tested from which LGBMR Regressor[3][4] and Ridge Regression with cross-validation[5] were chosen.

First technique use Random Forest, a machine learning technique that builds multiple decision trees and combines them to create more accurate predictions.

Second model uses Ridge regularization on regression model to avoid overfitting. 10-fold cross validation was used to choose the λ parameter.

State-of-the-art Convolutional Neural Network model Inception-V3[6] was used To extract image feature. The model was pretrained on ImageNet[7], a dataset containing more than 14 million images. The input was the satellite image of the property and the output array of 32 features that could be stored in the tabular format for the regression models.

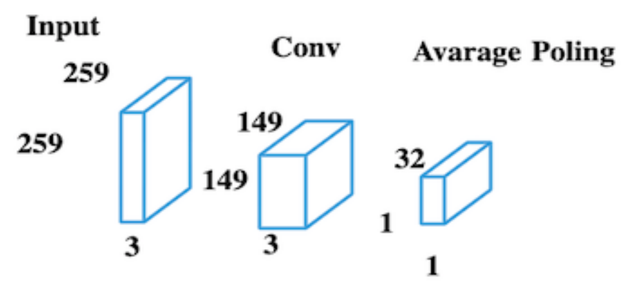

Figure 3: Feature extractor

Results

Model performance was measured using Mean Absolute Error that represents the sum of absolute values of the residuals divided by the number of datapoints and R^2 – the proportion of the variance of the dependent variable that is explained by the independent variables.

For the tabular data without satellite images, LGBMR Regressor got MAE of $322k and R^2 of 0.84. Ridge CV Regressor for the same data got MAE of $350k and 0.73 R^2.

When features from satellite images were added to the dataset, both models performed better with MAE $284k, 0.89 R^2 and $340, R^2 0.75 respectively.

	LGBMR	Ridge CV	LGBMR Satelite	Ridge CV Satelite
MAE	$322k	$350k	$294k	$340
R^2	0.84	0.73	0.88	0.75

Table 1: Results before and after aplying satelite images

Conclusions

This research had shown that using satellite images may be beneficial for machine learning models used to predict property prices. Inception-V3 model pre-trained on ImageNet has been used to extract features from images that later been used in the gradient boosting models. While this approach was successful in this scenario, further research is needed to test whether it can be applied to other datasets.

Contact

Jakub Miksa
University of Amsterdam
Email: miksa.kuba@gmail.com
Website: github.com/kmiksa

References:

1. GIS Lab, Newman Library, Baruch CUNY. 2015 New York City Real Estate Sales. https://geo.nyu.edu/catalog/nyu-2451-34678.
2. E. Kamptner. NYC Orthoimagery. GitHub repository. https://github.com/CityOfNewYork/nyc-geo-metadata/blob/master/ Metadata/Metadata_AerialImagery.md
3. Gulion Ke et al.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Microsoft Research, NIPS 2017
4. Bruno Klaus De Aquino Afonso et al. Housing Prices Prediction with a Deep Learning and Random Forest Ensamble, September 2019
5. Kamil Khalid: Modelling House Price Using Ridge Regression and Lasso Regression, International Journal of Engineering And Technology, November 2018
6. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. IEEE, 2016
7. Deng, J el al.. ImageNet: A Large-Scale Hierarchical Image Database, CVPR09, 2009

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com