Machine Learning for Home Valuation in New Taipei City

Andrew Bourgeois
Mathematical and Computational Sciences
University of Prince Edward Island
Charlottetown, Canada
abourgeois93@upei.ca

1 Abstract

Home pricing is varied and complex, but it is important both to buyers and sellers to do it fairly. Models for predictions of home prices and for analyzing feature importance have been created with datasets in other parts of the world. One study [3] created a model that could perform a prediction of home values in London. Another study [4] created a neural network to predict home prices in Merced County, California, United States, and analyzed the statistical significance of features using a significance test. To expand on this research, this paper uses data from homes in Sindian Dist., New Taipei City, Taiwan to produce various machine learning models that can predict the cost of a home based on prices of surrounding homes and information about those homes. The models are analyzed to understand the relationship between different aspects of a home, such as location, age, etc. in relation to the price of the home, and the importance of each variable. The Jupyter Notebook created and the data file are accessible through the author's GitHub repository [5].

2 Introduction

The prices of houses vary widely, and it can be difficult to tell if a given home is overpriced or underpriced. The goal of this research project is to produce a model that can accurately predict the price of a house given a set of datapoints, using data from Sindian Dist., New Taipei City, Taiwan to expand on existing research in other parts of the world. The machine learning model will also allow us to find out what variables determine the prices of houses, and to understand how they impact pricing.

3 Background

3.1 Dataset Description

A dataset from the UCI repository was used. It is a real estate valuation dataset from Sindian Dist., New Taipei City, Taiwan [1]. It has six variables:

X1 = the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)

X2 = the house age (unit: year)

X3 = the distance to the nearest MRT station (unit: meter)

X4 = the number of convenience stores in the living circle on foot (integer)

X5 = the geographic coordinate, latitude. (unit: degree)

X6 = the geographic coordinate, longitude. (unit: degree)

The output variable Y is the price in 10000's of New Taiwan Dollars, of one Ping (3.3 metres squared).

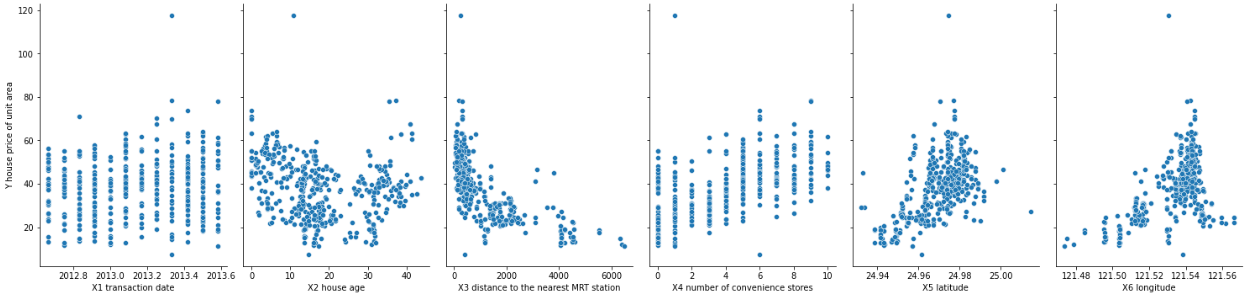

Figure 1: Each variable plotted linearly against house price.

The relationships are diverse and complex. X2, X3, X5 and X6 appear to have curved relationships, whereas X1 and X4 are linear.

This dataset was chosen because it is clean, and it is from a reputable source, so it is expected to be of high quality.

3.2 Dataset Optimization

The Python library scikit-learn (abbreviated sklearn) version 1.0.2 was used to manage the data, and to create and test different models [2]. Dimensionality reduction using Principal Component Analysis (PCA) and feature selection were applied to find out if some of the data could be omitted for performance improvements and reduction in overfitting.

3.2.1 Principal Component Analysis

Dimensionality reduction was applied using PCA from the machine learning package sklearn. Keeping 95% of the variance, the number of features was reduced from 6 to 5. However, the performance benefit of dropping one dimension in this case would not be significant: it would save a few seconds when training one of the models below, and it would only save about 4Kb of training data (one column = 414 values of size 64 bits). Since a PCA reduction would complicate the interpretation of the models and their results, all 6 dimensions were kept.

3.2.2 Feature Selection

Feature selection was also tested. As seen in figure 1, transaction date does not seem to have a linear correlation with house price. This feature was removed from the dataset temporarily and a linear regression model was trained and tested using a train/test split of 75%/25%. The MAE (mean absolute error) of the model’s predictions was slightly higher than the MAE of same model trained with all 6 dimensions of the data, so the transaction date dimension was kept.

3.3 Models Used

Four different models were trained, tested, and compared: Multiple Linear Regression (MLR), Polynomial Fitting with many different degrees tested, Decision Tree Regressor, and Random Forest Regressor.

The MLR model was trained using 75% of the data from the dataset, and was tested on the remaining 25%.

The polynomial fitting model was trained, and the best degree was selected through k-fold cross-validation with 4 folds (sklearn.model_selection.cross_val_score) using 80% of the data. The model was tested with the final 20% of the data. Polynomial degrees 1 to 10 were compared.

The Decision Tree Regressor and Random Forest Regressor were created using the default parameters of sklearn. They were then trained and tested through k-fold cross-validation using 4 folds.

4 Results and Analysis

4.1 Model Results

4.1.1 Multiple Linear Regression

After being trained on 75% of the data, MLR had the coefficients shown in figure 2.

X1 transaction date	5.1310563902535336
X2 house age	-0.23891787209106435
X3 distance to the nearest MRT station	-0.004891043430633901
X4 number of convenience stores	1.070950942996594
X5 latitude	216.89220859325224
X6 longitude	-39.17243057683667

Figure 2: Linear coefficients of the MLR model.

The model produced a Mean Absolute Error (MAE) of 5.57 when tested against the final 25% of the data.

4.1.2 Polynomial Fitting

Figure 3 shows the MAE values of the various degrees selected (degrees 1 to 10). Degrees 8 and higher were heavily overfitting as seen by the high MAE revealed through k-fold cross validation with 4 folds. Degree 2 had the lowest MAE at 5.24.

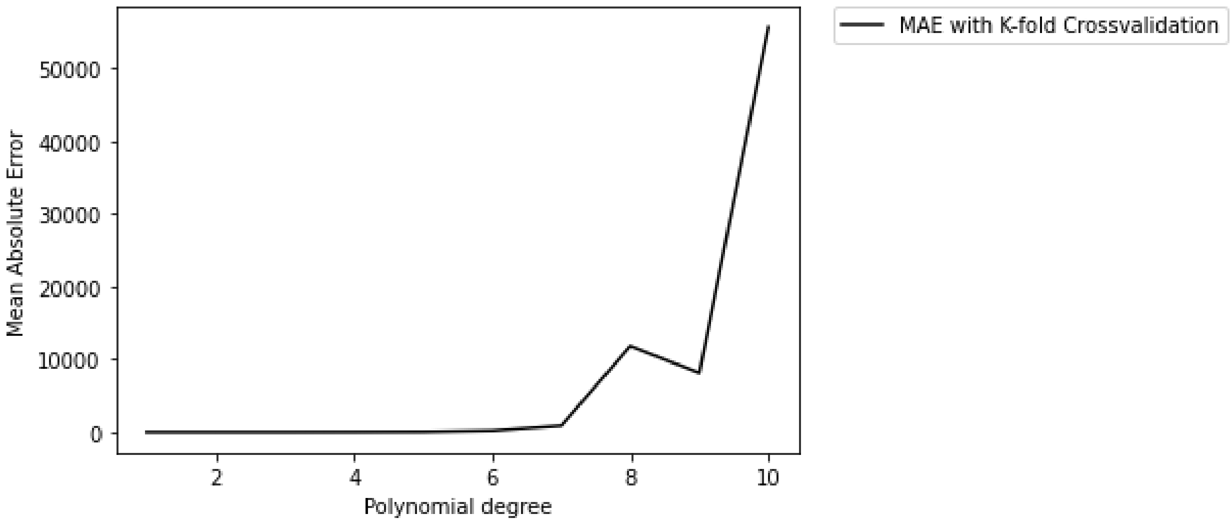

Figure 3: Mean Absolute Error of polynomial fitting with various degrees.

The degree 2 polynomial model produced a MAE of 5.99 when tested against the final 20% of the data.

4.1.3 Decision Tree Regressor

After 4-fold cross-validation, the model reached an MAE of 6.17.

4.1.4 Random Forest Regressor

After 4-fold cross-validation, the model reached an MAE of 4.87. The model was then fit directly to all of the training examples + labels, and feature importances (Gini importance) were retrieved. They are shown in figure 4.

X1 transaction date	0.04107669989724715
X2 house age	0.18633711327988953
X3 distance to the nearest MRT station	0.5748198788212079
X4 number of convenience stores	0.0224021949401898
X5 latitude	0.09739407904615952
X6 longitude	0.07797003401530611

Figure 4: Feature importances in the Random Forest Regressor model.

4.2 Model Comparison

The results were close between MLR and the polynomial model, with MAEs of 5.57 and 5.99 respectively. The Decision Tree Regressor was also close with an MAE of 6.17. The Random Forest Regressor was significantly ahead of the others, making it the most performant model with an MAE of 4.87.

5 Conclusions and Discussion

The dataset relates six variables to the price per unit area of homes in a part of Taiwan. It was analyzed to find their effect on price.

5.1.1 The Best Model

Multiple linear regression (MLR), polynomial fitting, decision tree, and random forest models were trained and tested. The Random Forest Regressor had a much better performance than the others, making it the strongest model found. It had a Mean Absolute Error (MAE) of 4.87, which can be interpreted to mean that the average error in predicting a home's price is approximately 4.87 * 10000 = 48700 New Taiwan Dollars. It was expected that the Random Forest model would perform better than the Decision Tree model given that it had the compounded benefit of using multiple decision trees.

5.1.2 What the Models Reveal About the Data

As seen in Figure 2, transaction date, number of convenience stores, and latitude have positive coefficients so higher values of these variables are correlated with higher home valuation. House age, distance to nearest MRT station, and longitude have negative coefficients so homes with lower values of these variables were observed to be a lesser value. These correlations apply in Sindian Dist., New Taipei City, Taiwan.

Figure 4 reveals that the most important features by far in determining home price in Sindian Dist. are "distance to nearest MRT station" and then "house age".

5.1.3 Future work

Future expansion of this research could be done by trying more sophisticated machine learning models such as deep learning to learn the complex and varied relationships between each variable and the home price. Machine learning models could also be applied to datasets in different parts of the world to understand and predict house pricing in those areas.

6 REFERENCES

[1] *UCI machine learning repository: Real estate valuation data set data set.* Retrieved March 22, 2022, from https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set

[2] *Scikit-learn: Machine learning in python.* Retrieved March 22, 2022, from https://scikit-learn.org/stable/index.html

[3] Carpentier, A., & Schluter, T. (2016). *Learning relationships between data obtained independently* Journal of Machine Learning Research. Retrieved from http://proceedings.mlr.press/v51/carpentier16b.pdf

[4] Horel, E., & Giesecke, K. (2020). *Significance tests for neural networks* Journal of Machine Learning Research. Retrieved from https://www.jmlr.org/papers/volume21/19-264/19-264.pdf

[5] Bourgeois, A. (2022). *Machine-Learning-for-Home-Valuation-in-New-Taipei-City*. https://github.com/apbourgeois/Machine-Learning-for-Home-Valuation-in-New-Taipei-City