Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia

Süleyman Ali Burak Çınar, Atakan Tekoğlu
Department of Computer Engineering
Yıldız Technical University, 34220 Istanbul, Turkey

***Abstract*—House price forecasting is an important topic of real estate. The literature attempts to derive useful knowledge from historical data of property markets. Machine learning techniques are applied to analyze historical property transactions in Australia to discover useful models for house buyers and sellers. Revealed is the high discrepancy between house prices in the most expensive and most affordable suburbs in the city of Melbourne. Moreover, experiments demonstrate that the combination of machine learning models that is based on mean squared error measurement is a competitive approach.**

***Keywords*—House Price Prediction, Regression, Linear Regression, Machine Learning**

I. Introduction

Buying a house is undoubtedly one of the most important decisions one makes in his life. The price of a house may depend on a wide variety of factors ranging from the house's location, its features, as well as the property demand and supply in the real estate market. The housing market is also one crucial element of the national economy. Therefore, forecasting housing values is not only beneficial for buyers, but also for real estate agents and economic professionals.

Studies on housing market forecasting investigate the house values, growth trend, and its relationships with various factors. The improvement of machine learning techniques and the proliferation of data or big data available have paved the way for real estate studies in recent years. There is a variety of research leveraging statistical learning methods to investigate the housing market. In these studies, the most popular investigated locations are the United States [1], Europe [2]; as well as China [3]; and Taiwan [4]. However, research on the housing market by applying data analytics with machine learning algorithms in Australia is rare, or elusive to find.

The goal of this study is through analyzing a real historical transactional dataset to derive valuable insight into the housing market in Melbourne city. It seeks useful models to predict the value of a house given a set of its characteristics. Effective models could allow home buyers, or real estate agents to make better decisions. Moreover, it could benefit the projection of future house prices and the policy making process for the real estate market.

II. RELATED WORK

Previous studies on the real estate market using machine learning approaches can be categorized into two groups: the trend forecasting of house price index, and house price valuation. Literature review indicates that studies in the former category deem predominant.

In the house growth forecasting, researchers try to find optimal solutions to predict the movement of housing market using historical growth rates or house price indices, which are often calculated from a national average house price [5], or the median house price [6]. house growth forecasting could act as a leading indicator for policymakers to assess the overall economy. Factors that affect house price growth tend to be macroeconomic features such as income, credit, interest rates, and construction costs. In these papers, Vector Auto regression (VAR) model was commonly applied in earlier periods [7] while Dynamic Model Averaging (DMA) has become more popular in recent years [4].

On the other hand, house price valuation studies focus on the estimation of house values [8]. These studies seek useful models to predict the house price given its characteristics like location, land size, and the number of rooms. Support Vector Machine (SVM) and its combination with other techniques have been commonly adopted for house value prediction.

It is underscored that Neural Network (NN) and SVM has recently been applied in a wide variety of applications across numerous industries. Neural Networks has been further developed to become deep networks or Deep learning method. Besides, the advance of SVM deems to achieve by integrating it with other algorithms.

III. DATA PREPARATION AND EXPLORATION

A. Original Data

The data implemented in this study is the Melbourne Housing Market dataset downloaded from Kaggle website [9]. The original dataset has 34,857 observations and 21 variables. Each observation presents a real sold house transaction in the city of Melbourne from 2016 to 2018. These variables can be categorized into 3 groups:

- Transactional variables include Price, Date, Method, Seller, Property count.
- Related location predictors which contain Address, Suburbs, Distance to CBD, Postcode, Building Area, Council Area, Region name, Longitude, Latitude
- Other house features such as House Type, Number of Bedrooms, Number of Bathrooms, Number of Car slots, and Land size

The outcome of house value prediction is the price which is a continuous value, and predictors consist of other features with both numeric and categorical types.

B. Data Preparation

Before applying models for house price prediction, the dataset needs to be pre-processed. The investigation of missing data is at first performed. Several missing patterns are assessed rigorously since they play an important role in deciding suitable methods for handling missing data [10], [11], [12].

Columns with more than 55% values missing are removed from the original dataset since it is difficult to impute these missing values with an acceptable level of accuracy. In addition, there are many rows with missing values of the outcome variable (Price). Since the imputation of these values could increase bias in input data, observations with missing values of the Price column are deleted.

Furthermore, outliers are also discovered and addressed. Outliers are defined as an observation which seems to be inconsistent with the remainder of the dataset [13]. Outliers may stem from factors such as human errors, relationship with probability models, and even structured situations [14]. For instance, total area of less than 10 square meters are removed.

As a result, the cleaned data, which is used to build and evaluate models, has 8 variables (in Table 1) and more than 20 thousand observations.

Table 1 FEATURES DESCRIPTION

Name	Type
Price	Numerical
Location	Categorical
Status	Categorical
Distance	Numerical
Rooms	Numerical
Bathroom	Numerical
Garage	Numerical
Total Area	Numerical

C. Descriptive Exploration

This section only presents the most important findings.The data summary information and other informative figures are allocated in the project report.

Median price is roughly 900 thousand dollars. Fig. 1 and 2 indicate the histograms of Price and log(Price). While the range of Price values varies widely with a long tail, log(Price) deems to have a normal distribution. Thus, log(Price) will be used as the output in model building and evaluation phases.

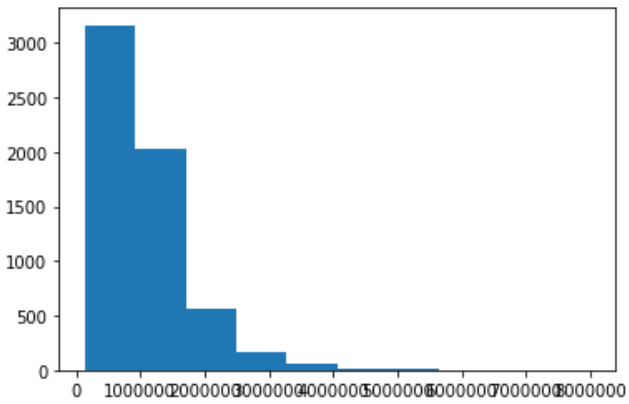

Figure 1 Histogram of Price

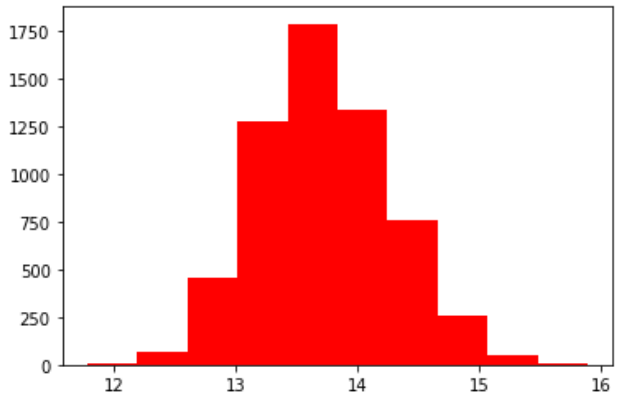

Figure 2 Histogram of Log Price

IV. METHODOLOGIES

A. Data Reduction and Transformation

In order to improve the interpretability and enhance the performance of prediction models, data reduction techniques applied.

One hot encoding is a powerful technique to transform categorical data into a numerical representation that machine learning algorithms can utilize to perform optimally without falling into the misrepresentation issue previously mentioned.

One hot encoding is the technique to convert categorical values into a 1-dimensional numerical vector. The resulting vector will have only one element equal to 1 and the rest will be 0. The 1 is called Hot and the 0's are Cold.

In order to apply one hot encoding we can now use the pandas method to convert categorical variables into dummy/indicator variables with the **get_dummies** function.

B. Model Selection and Evaluation

The paper implements different regression models to find the useful ones.

An attribute subset from Stepwise will be inputted in Linear Regression, Polynomial Regression, Regression Tree, as well as Neural Network. Let's briefly consider these models.

1) Linear regression: It will be used as a baseline for model evaluation, which based on Mean Squared Error (MSE) measured on an evaluation dataset. MSE is the most popular tool to measure the quality of fit [15]. It is calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)^2\right)$$

2) Lasso regression: Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of muticollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.[16] It is calculated as:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|$$

3) Regression Trees: Decision Trees is a widely known methodology for classification; and Regression Trees, which use for continuous outcome prediction, is a special case of Decision Trees. Each leaf contains the prediction value which is the mean of prices of all observations in that leaf. The feature selection as a node in a Regression Tree will be based on the goal of minimizing the Residual Sum of Squares (RSS)[15].

$$RSS = \sum_{i=1}^{j}\sum i \epsilon R_j (y_i - y(R_j))^2$$

Before fitting data into models, the cleaned dataset is divided into train and evaluation data. The evaluation set will be kept isolated from model building, and only used for model evaluation. The model fitting process utilizes train data using ten folds cross-validation. It is noted that cross-validation is applied in both data reduction and model construction stage.

V. RESULTS

The experiments have been deployed in R on a Window system. The Mean Squared Error (MSE) of both train and evaluation datasets are presented in TABLE 2. As in the previous discussion, linear regression will act as the baseline for model comparison. The evaluation ratio of each model is equal to its evaluation MSE divides to the evaluation MSE of Linear regression. The smaller evaluation ratio, the higher accuracy of the model's prediction.

Table 2 MODEL AND SCORES

Name	Type
Linear Regression	0.600619
Lasso Regression	0.600605
Decision Tree	0.457164

It can be seen from TABLE 2 that Lasso Regression delivers a prediction result as good as Linear Regression, while Decision Tree results is higher errors.

VI. MOBILE IMPLEMENTATION

The purpose of this section is to demonstrate how a model can be consumed and used by an Android app. The following steps are the required items:

- Downloading and installing Android Studio
- Creating a new Android project with a single screen
- Designing the layour of the screen
- Adding a functionality to accept input
- Adding a functionality to consume the RESTful API that serves the model

After creating an Android project, it has three main folders:

- **manifests:**This folder contains the manifests file used for permissions and application versioning.
- **java:** This folder has all the JAVA code files.
- **res:**This folder has all the layout files and media files used in the application.

After creating the model, we need to implement this model through the RESTful API. For doing this, we use Flask Server,it is a framework of the Python language. Then we get an output our model created in Jupyter environment in JSON format so that building an API.

The content of the JSON file can be passed as the data in the POST API request. After that, to hit the API with the estimate button, the following helper functions are triggered.

- **ByPostMethod:**Accepts the URL as a String and returns the response as an InputStream. This function takes the server URL string that we created using the FLASK framework and returns the response from the server as an input stream.
- **ConvertStream:**This function accepts InputStream and returns a String of the response. The input stream returned from the previous function is processed as a string object.
- **makeJSON:** This function accepts the values from the edit boxes and returs the JSON object so that pass as data in API call.

REFERENCES

[1] S. M. Rangan Gupta, Alain Kabundi, "Forecasting the us real house price index: Structural and non-structural models with and without fundamentals," *Economic Modelling*, vol. 28, no. 4, pp. 2013–2021, 2011.

[2] M. K. Marian Risse, "Forecasting house-price growth in the euro area with dynamic model averaging," *The North American Journal of Economics and Finance*, vol. 38, pp. 70–85, 2016.

[3] Y. C. Yu Wei, "Forecasting house prices using dynamic model averaging approach: Evidence from china," *Economic Modelling*, vol. 61, pp. 147–155, 2017.

[4] C.-C. L. Pei-Fen Chen, Mei-Se Chien, "Dynamic modeling of regional house price diffusion in taiwan," *Journal of Housing Economics*, vol. 20, no. 4, pp. 315–332, 2011.

[5] P. G. T. Vasilios Plakandaras, Rangan Gupta, "Forecasting the u.s. real house price index," *Economic Modelling*, vol. 45, pp. 259–267, 2015.

[6] S. M. M. Mehmet Balcilar, Rangan Gupta, "The out-of-sample forecasting performance of nonlinear models of regional housing prices in the us," *Applied Economics*, vol. 47, pp. 2259–2277, 2015.

[7] Y. H. C. Jing Li, "What pushes up china's real estate price?" vol. 5, no. 2, pp. 161–176, 2012.

[8] L. Z. Chuan-Fang Ong, "Forecasting spatial dynamics of the housing market using support vector machine," *International Journal of Strategic Property Management*, vol. 21, no. 3, pp. 273–283, 2017.

[9] T. Pino. (2016) Melbourne housing market. [Online]. Available: https://www.kaggle.com/anthonypino/melbourne-housing-market

[10] codeBasics. (2017) Handling missing values. [Online]. Available: https://www.youtube.com/watch?v=EaGbS7eWSs0

[11] ——. (2017) Handling missing values. [Online]. Available: https://www.youtube.com/watch?v=XOxABiMhG2U

[12] S. Sarkar. (2017) Handling missing values. [Online]. Available: https://github.com/SubhamIO/House-Price-Prediction

[13] J. . J-Secur1ty. (2018) Data cleaning in python. [Online]. Available: https://www.youtube.com/watch?v=rzR$_c$$KnkD$18

[14] D. B. R. Roderick J. A. Little, "Outliers in statistical data," in *Statistical Analysis with Missing Data, 3rd Edition*. Wiley, 2019.

[15] W. D. H. T. T.-R. James, G., "Statistical learning," in *An Introduction to Statistical Learning*, 2013.

[16] S. Glen. (2015) Lasso regression: Simple definition. [Online]. Available: https://www.statisticshowto.com/lasso-regression/