Analysis of Ireland Property Prices

Ginu Varghese   
Department of Computing Science and Mathematics *Dundalk Institute of Technology*Dundalk, Co.Louth, Ireland  
d00251842@student.dkit.ie

# Litrature Review

Several studies have been performed in the past to analyze and predict the real estate and residential property prices to aid the customers as well as the developers. Aoife K. Hurley & James Sweeney studied the impact of post codes and address in Irish property prices based on the data from January 2018 to November 2018(Hurley et al. 2022). They focused on analyzing the property prices in Dublin with the dataset of 5028 properties for the development of geospatial statistical models where the post codes and addresses are being mislabeled. They also performed text mining to create additional variables that describes the features of the properties and its surroundings. A spatial hedonic regression model was used to separate the spatial and non-spatial contributions of property features to the resale value. Generalized additive models (GAM), regression kriging and geographically weighted regression (GWR) have also been used because these methods provide greater interpretability with smaller data requirements. Three different Machine learning (ML) methods like Decision Tree, Random Forest and K-nearest neighbor algorithms were also applied on the data to evaluate how these ML models perform with smaller datasets.

Their models shown a reduction in median absolute percentage error with an increasing model complexity i.e., 12% for the hedonic model and 9.6% for linear model with spatial surface. And the authors also stated that although ML models are widely used for property prices prediction, they did not have the probability-based uncertainty intervals and the interpretability of the statistical spatial models. They also added that the random forest models may not fit well with the areas outside Dublin since the housing turnover is low outside Dublin and here statistical spatial modelling will work more efficiently. In Ireland, where property valuations are currently based on comparison to recently sold neighboring properties, the authors claim their model has higher applicability and will help to improve property tax computations and site value estimates because the model is not only based on the property value but uses the spatial location scaling.

Machine learning models were used to predict house prices in Godavari district of Andhra Pradesh, India by M. Thamarai and Malarvizhi(2020). The models were built to help the people buy suitable houses for their needs. Decision tree regression, decision tree classification and linear regression models were performed on the data based on the attributes of the property like number of bedrooms, age of the house, availability of school near the house and shopping malls available nearby the house location. Attribute selection algorithms were used to remove the redundant features to reduce the impurity in the process before splitting the data for modelling.

To predict the availability of houses according to the requirement of the user, they used the decision tree classifier which gives responses like yes or no to show whether a house is available or not. Along with this, regression methods like decision tree regression and multiple linear regression were used to predict the prices of the houses. The dataset used for the modelling was a real-time data acquired from the Godavari district with all the attributes of the house and modelled using the scikit learn, a machine learning tool.

The main dataset was divided into train and test data and the decision tree classifier is performed using the training dataset. The accuracy of the model is then tested using the test data. Mean Squared Error (MSE), Mean Absolute Error (MAE) and root mean squared error (RMSE) were used to evaluate the performance of the classification model. The house price prediction with decision tree algorithm produced an output with some data record prices predicted with lesser deviations. From the multiple linear regression, it is found that the number of bedrooms is the feature having high influence on house price and age of the house is the feature having less influence on price of house. Like the classification model, Mean Squared Error (MSE), Mean Absolute Error (MAE) and root mean squared error (RMSE) were used to evaluate the performance of both the regression models and it is found that the prediction of house price using multiple linear regression has higher performance than the prediction using the decision tree regression. The authors also stated that the developed model can be used to predict the availability and prices of houses for any new attributes according to the users and the overall accuracy can be increased in the future with a large dataset and by identifying the best features.

In 2020, Quang and his fellow authors conducted a study to compare housing prices prediction using traditional and advance machine learning methods. They used three different machine learning models: Random Forest, XGBoost and LightGBM. Further, two machine learning techniques are used, Hybrid Regression and Stacked Generalization Regression for prediction. For the analysis, the Beijing house price data from 2009 and 2018 was used and feature engineering was performed to select the appropriate features. In addition to this, exploratory data analysis was done to discover the patterns in the data. The machine learning models were applied on the training data which is split from the actual data and Root Mean Square Logarithmic Error (RMSLE) was used for evaluating the performance of the models. Results from the study shows that the Random Forest method is prone to overfitting even though it has lowest errors. The Hybrid regression method is better performing among all three methods. They also stated that although the Stacked regression method has the worst time complexity, it is the best choice when accuracy is considered(Quang et al. 2020).

To understand the factors influencing the property prices, Decision Tree (DT), Random Forest (RF) and linear regression (LR) methods were used by (Yee et al. 2021) using the Malaysian dataset from April 2017 to December 2019, to help the buyers and sellers who wish to spend in the property market. The accuracy of the models is evaluated based on the R squared, RMSE and MAE values. Natural Language Processing was used to transform the data so that it is readable by the machine. The outcome of the study shown that the RF is the better model with higher accuracy compared to the DT and LR.

Alen Ihre (2019) discussed the machine learning algorithms K-Nearest Neighbour (K-NN) and Random Forest (RF) regression to predict the house prices. Five-fold cross validation was performed on the dataset to minimize the bias and the RF was found to be the best performing model based on the MAE values (Ihre 2019).

Artificial Neural Network (ANN) is used for analysing the real estate price by (Shi and Li 2009) to analyse the house price determinants. An improved genetic algorithm was used to optimize the weights of the neural network. The results of the study shown that the GA-ANN is more capable of determining the house price determinants more time efficiently and the errors of the HGA-ANN model was found to be lower than the back propagation (BP) and the genetic algorithm (GA) models.

Support Vector Machine (SVM) has been employed for predicting and forecasting house prices by Jingyi Mu et al. (2014), Phan (2018), Winky K.O. Ho et al. (2020), Jieh-Haur Chen et al. (2017), Gu Jirong et al. (2011). and Xibin Wang et al. (2014).

House value forecasting based machine learning methods by Jingyi Mu, Fang Wu and Aihua Zhang (2014) aimed at helping the developers and government to take decisions regarding developing real estate in the Boston area. The authors collected the data from the UCI data sets and support vector machine (SVM), least squares support vector machine (LSSVM), and partial least squares (PLS) algorithms were applied on the training data to predict the housing value. The Mean Square Error (MSE) was used for evaluating the performance of the models. From the prediction results it is found that the SVM and LSSVM has better efficiency with the nonlinear data. PLS algorithm is better for linear data due to the simplicity of the algorithm. They also added that to achieve best forecasting effect and an optimal solution, SVM can be used (Mu et al. 2014).

Phan (2018) utilized the SVM technique to predict the house prices in the Melbourne city of Australia to help the house buyers and sellers. Neural network (NN), Polynomial regression, Linear regression and Regression Tree models were also developed along with SVM to identify the best fit. The data used in the study was downloaded from the Kaggle website and it has the house sold houses transaction from 2016 to 2018. Data imputation and descriptive analysis techniques were performed on the data prior to modelling. Along with this, principal component analysis (PCA) was also performed to select the desired features and stepwise method was utilized for subset selection. MSE was used to evaluate the performance of the models as lower error shows the higher accuracy. The results shown that the SVM with the subset selection method gives the best efficiency. When regression tree and linear regression delivered almost equal prediction result, the polynomial regression gave better accuracy with lower errors. The neural network seemed to be not working well with the available dataset. The authors also stated that regression tree and neural network worked more faster than the SVM, where PCA with SVM took more time than SVM with stepwise (Phan 2018).

SVM methods were employed by Winky K.O. Ho, Bo-Sin Tang and Siu Wai Wong (2020) for predicting the property prices in Hong Kong. Random forest (RF) and Gradient Boosting Machine (GBM) were also utilized along with SVM to compare the performance of algorithms. The dataset used was a sample data with over 40000 housing transactions in a time of 18 years. Correlation matrix was used to determine the features which should be included in the models. MSE, RMSE and Mean Absolute Percentage Error (MAPE) were the performances matrices used for evaluating the models. The results from the performance metrices revealed that the RF and GBM were able to estimate the house prices better than the SVM with smaller errors. They also found that the ML algorithms need more computation time than the traditional Hedonic pricing model and among the three models used, SVM is the better choice for forecasting when speed is the priority and RF and GBM should be used if the accuracy is considered (Ho et al. 2020).

(Chen et al.) implemented SVM models to forecast the residential housing prices in the Taipei city of Hong Kong from 2007 to 2010. By using stepwise multi regression, the support vectors were found, and a SVM hedonic price model was built using the support vectors, the structural and the spatial variables to predict the house prices in the Taipei city. The SVM model is then developed based on the identified support vectors to forecast the future housing prices. One of the advantages in using SVM is that it is not depending on the probability distribution assumption and hence it could plot the input variables into a high dimensional feature space. To compensate the bias variance trade-off, five-fold cross validation has been used for testing and training in the analysis. The outcome from the study points out that the SVM can be considered as a superior approach which legitimise the issues in the multiple regression analysis and combining the hedonic approach with SVM is feasible for non-linear modelling (Chen et al. 2017).

(Gu et al. 2011) used the SVM methods along with hybrid genetic algorithm methods to forecast the house prices in China. SVM has proven to be one of the best algorithms in both classification and regression in lots of applications. In the study, genetic algorithm (GA) has been used instead of grid algorithm to optimize the parameters of the SVM since, GA is more time efficient and the G-SVM is developed. The results of the study revealed that the G-SVM method giving more accuracy than the Grey Algorithm (GM) which is used in the past to predict the house prices (Gu et al. 2011).

Machine learning algorithms has used to forecast the real estate prices by many researchers and (Wang et al. 2014) used SVM to forecast the real estate price with particle swarm optimization (PSO) in the Chongqing city in China. One of the reasons to choose SVM is its ability to conquer the ‘Curse of dimensionality’. To identify the parameters of SVM, PSO method is used instead of GA and grid algorithm since it is easy to enforce. The actual data was divided into train and test data for the modelling. The study shown that the PSO-SVM model is performing better than the BP neural network methods used by other researchers.

PSO has used by (Alfiyatin et al. 2017) for predicting the house prices to help the builders to decide the selling price of the house and to help the buyers to set the right time to buy the house. PSO and regression analysis was implemented on the houses data in the Malang city of Indonesia within 2014-2017. Hedonic regression was chosen as the regression prediction model and PSO to select the appropriate features. Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) were used to test the model. The error prediction values are found to be higher for the regression model compared to the PSO-regression model. Hence the study proved that the combination of PSO with regression can give the minimum prediction error.

Support Vector Machine Regression (SVR) has used by several researchers to predict the real estate and housing prices due to its efficiency and application. A SVR model was used by (Li et al. 2009) to understand the possibility of predicting real estate prices in China from 1998-2008. The results of the model are compared with the Back Propagation Neural Network (BPNN) model to analyze the performance and MAE, MAPE and RMSE were used to evaluate the performance of the model. Several indicators like CPI, loan interest rate, real estate price, real estate investment, income etc. were used to forecast the real estate prices. The analysis demonstrated that the SVR model works better than the BPNN model for real estate forecasting based on the MAE, MAPE and RMSE. The study also proved that the SVR method is an efficient approach to forecast the real estate price (Li et al. 2009).

SVM regression with Gaussian kernel was utilized by (Miao et al. 2021) to predict the property prices in Boston area. The aim of the study was to help the people to estimate the price of the house based on the properties of the house. They selected the most important features using the decision tree with ID3 algorithm, divided the data to train and test data and SVR is employed on the data with the gaussian kernel to predict the prices and compared the model with different regression methods to analyze the performance. The study proved that the SVR with gaussian kernel has more efficiency than the KNN, decision tree and SVR with linear kernel. But the model still has some drawbacks which are mentioned by the authors and the important one is that some of the factors that impact the house prices cannot be measured such as the cultural tolerance of the neighbors and so on.

Jiao Yang Wu also conducted a study (2017) to analyse the house price prediction using SVR based on the housing sales data of Kings County, USA with an aim to help the buyers and sellers. Feature selection methods like Random Forest selector, Lasso ridge and Recursive Feature Extraction and the feature extraction method PCA is done prior to building the model. Further to this, parameter tuning, and transformation techniques have been used to improve the accuracy of the model. MSE was employed as a performance metrics for evaluating the model and the lowest MSE was found to be 0.04. But the results from the experiments shown that there is not much difference in the performance of the model with feature extraction and feature selection. The Radial Basis Function (RBF) kernel with SVR was found to be the best one among the performed combinations. The author also pointed out that in future other machine learning models like XGBoost and other feature engineering methods can be applied on the data (Wu 2017).

Hedonic Regression is one of the estimation and prediction method preferred by the researchers when it comes to prices. It is used in most of the scenarios when a price variable is to be considered. Property prices in Croatia is studied by (Kunovac and Zagreb 2019) using Hedonic regression based on the data collected from different sources. One of the goals of the research was to propose how the hedonic models can be used for the evaluation of residential property. The hedonic model built allows the evaluation of some attributes of the property such as age, location and so on. The results of the analysis shown that the micro location of the property should also be considered to the hedonic models and to commonly used other models to improve the prediction of residential house prices.

(Abdulai and Owusu-Ansah 2011) used hedonic regression to determine the house price determinants in Liverpool, The United Kingdom over a period of 18 years from 1990-2008. They also analysed how the past and present buyers valued the property features. The regression model was able to explain almost 75% of variation in the housing prices and all the variables in the dataset was found to be statistically significant. Another thing they found was that the price of the new properties is almost double than the old properties and the detached houses are more expensive than the flats. When the past buyers before 1999 focused more on the number of bedrooms, bathrooms and detached houses, the buyers after 2000 more value the number of floors, gardens, and showers.

(Selim 2008) identified the house price determinants in Turkey using hedonic regression using the 2004 household survey data. Environmental factors were not included in the data for analysis and natural logarithm of the house price is treated as the dependent variable. To estimate the hedonic model, ordinary least square method is utilized. The heteroscedasticity present in the model was eliminated using the White’s heteroscedasticity consistent coefficient covariance matrix. And the results of the analysis shown most of the variables to be significant and the house prices seems to be higher for houses with more rooms. The variables such as water system, pool, type of the house, number of rooms, size of house, locational attributes and building structure are the found to be the most significant variables that influence the house prices.

compared the hedonic regression with Artificial Neural Network (ANN) model on house price prediction based on a sample dataset of 200 New Zealand houses. The age of the house, size of house, bedrooms, bathrooms, and other features were considered for the experiment. The heteroscedasticity consistent coefficient covariance matrix and weighted least squares method were used instead of the ordinary least square method to eliminate the heteroscedasticity issues. However, the heteroscedasticity problem was not completely removed. The results from the study revealed that even though the hedonic model was able to explain almost 70% of variance in the model, it did not outperform the neural network model. The authors mentioned that the small dataset and the lack of environmental features may be some reasons for the poor performance of the hedonic model.