Prediction of Housing Prices Using Machine Learning, Time Series ARIMA Model and Artificial Neural Network

Shivani Mangaleswaran[1], Vigneshwari.S.[1]

[1]Sathyabama Institute of Science and Technology, School of Computing, Department of Computer Science and Engineering, Chennai, India
shivanimangaleswaran@gmail.com,
vigneshwari@sathyabama.ac.in

Abstract. Volatility of housing prices is difficult to predict and varies based on several factors like area and year of purchase. This paper summarizes the results of the prediction of housing prices through Linear Regression, K-Means Clustering, Logistic Regression, Auto Regressive Integrated Moving Average and an Artificial Neural Network. The neural network can classify whether the price of a particular house is above or below the average market value. The performance of machine learning algorithms, is measured using metrics like Precision, Recall, F1-Score, while the ANN described predicts the price range of a newly entered data point, hence paving the way for potential real-time prediction.

Keywords: Prediction, Classification, Regression, Machine Learning, Artificial Neural Networks

1 Introduction

The prices of houses sold between May 2014 and May 2015 in King County, USA have been analyzed using exploratory data analysis techniques. The data obtained from Kaggle contains 21,613 observations [1]. Data analysis has been extensively used in several other potential prediction fields like weather prediction [2]. Supervised and unsupervised learning algorithms, ARIMA modelling and ANN were used to make predictions. In determining the prices of homes, the developer must calculate carefully and determine the appropriate method because property prices always increase continuously and almost never fall in the long term or short [3]. Hence, different models have been executed to determine which performs the best, with least error.

2 Overview of Prediction Models

Performance of several supervised and unsupervised machine learning models in the field of health care, to classify genes and predict infected cells, threw light on robustness of different machine learning algorithms, as presented by E. Moses et al. and S. Vigneshwari et al. [4-5]. Unsupervised Learning algorithms are used when no classifier label is present in the dataset and deriver some structure from the input by manipulating the similarities between different features and try to develop classes. K-Means clustering is one such unsupervised learning algorithm that forms clusters, each with its own centroid, based on feature similarity. Supervised Learning is prediction based on a given set of input values and corresponding output labels, and is broadly divided into classification and regression. Linear Regression is the prediction of real valued continuous outputs, while classification, also known as Logistic Regression is the used to predict which discrete class data belongs to based on class labels. David B. Alencar et al. [6] present a hybrid forecasting approach based on Seasonal Auto Regression Integrated Moving Average (SARIMA) and Neural Network for accurate wind speed forecasting using explanatory variables. Similarly, ARIMA was used for time series prediction of housing prices, with three parameters p, d, q. Parameter p corresponds to auto regression, which is computed by performing regression on the values as a factor of time lag, parameter q corresponds to moving average which uses the error generated in Auto Regressive Model to form a regression curve. The parameter ‘d’ stands for the non-seasonal differences required to achieve stationarity, where stationarity is achieved if the mean, variance and covariance of a given data is constant over a period of time. The success of neural networks has become evident over the years. Convolutional Neural Networks have proved to be robust for image classification, as presented by Sankari.A and Vigneshwari.S [7], However, neural networks require higher computational resources. Vijai Chandra Prasad.R et al. [8] presented facial recognition using Principle Component Analysis (PCA), on low-memory Raspberry Pi single board computer. An Artificial Neural Network has been designed and evaluated to predict housing price ranges, as described in this paper.

3 Results and Comparison of Model Performance

K-Means clustering was first implemented to predict housing prices. To improve the accuracy of predictions, the data was feature-engineered to include a classifier label. Time series prediction using tradition ARIMA modelling was implemented by adding datetime indexing to the data. Finally, to avoid overfitting and make new predictions with high accuracy, an artificial neural network was designed.

3.1 Linear Regression

Linear Regression model resulted in extremely high error rate. A graphical representation of test prices vs predicted prices illustrates the differences in predicted prices from the actual expected test prices, presented in Fig. 1. Hence this model failed to make useful predictions.

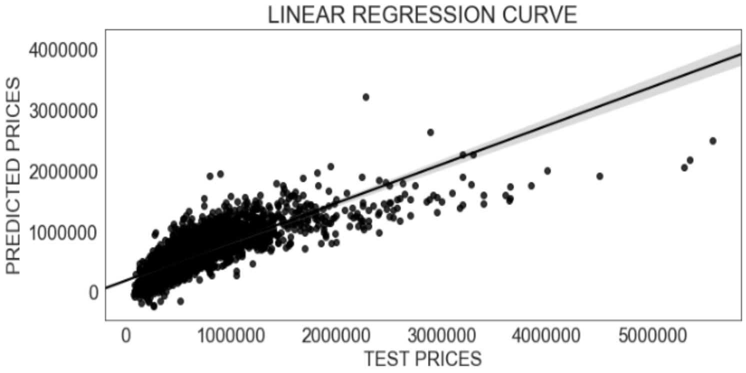

Fig. 1. Linear Relationship between Test Prices and Predicted Prices

Regression is commonly used model to predict prices, as it returns real-valued predictions. However, when used directly, it does not always yield good results for real-world data. Modifications have been made to the traditional linear regression model to improve its efficiency. One such method is Particle Swarm Optimization combined with Hedonistic Pricing, as proposed by Adyan Nur Alfiyatin[9].

3.2 K-Means Clustering and Logistic Regression

Next, K-means clustering was programmed to return two clusters. A visual representation of the results is depicted in Fig. 2. The white circles represent the cluster centers. There is no clear boundary that separates the two clusters, as indicated by overlapping cluster centers. Hence, this model did not perform very well. However, to further interpet the results through a numeric approach, basic feature engineering, a process of using domain knowledge of data to create features that make machine learning algorithms work better, was implemented to create a new classfier label.

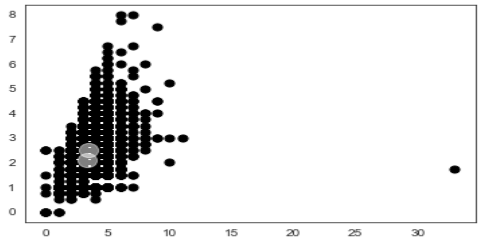

Fig. 2. Overlapping Clusters from K-Means Model indicates unsatisfactory performance

The mean price of the houses was calculated to split the data into two classes. Class labels were assigned iteratively and this new feature was added to the data under a column called Classifier. ‘1’ denoted an above average price, while ‘0’ denoted below average price. Fig. 3 illustrates the proportion of houses above and below mean price.

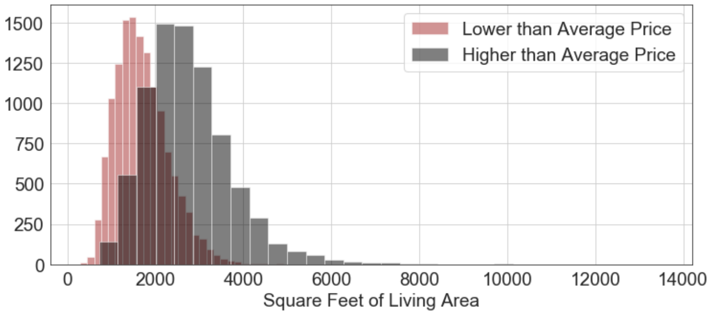

Fig. 3. Histogram of Living Area, based on Mean Price Classifier, showing houses with average, below and above average selling prices.

Using this classifier column, Logistic Regression, otherwise known as classification model, was executed. The clusters formed in K-means clustering algorithm were compared with the new classifier label's values. The results of K-Means Clustering and Logistic Regression are compared in Table 1.

Table 1. Comparison of Performance after Adding a Classifier Label

Algorithm	Precision	Recall	F1-Score
1. K-Means Clustering	0.61	0.64	0.51
2. Logistic Regression	0.80	0.81	0.80

3.3 Time Series Modelling Using ARIMA

The year in which the house was sold had to be dropped as date could not be converted to an integer or floating-point form, as required by algorithms like clustering and regression. To overcome this, the data was modified to include a Date-Time index. Date-Time indexing is the first step to perform any time series prediction. A steep increase in prices was evident from 2014 to 2015 as depicted in Fig. 4. The prices of houses sold during May 2014 – May 2015 is visualized as a time series plot in Fig. 5.

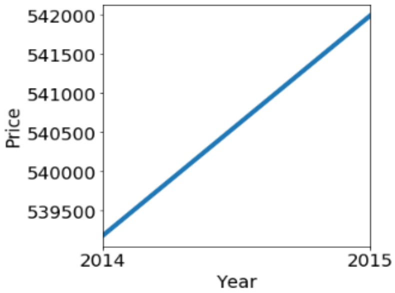

Fig. 4. Increase in housing prices from 2014-2015.

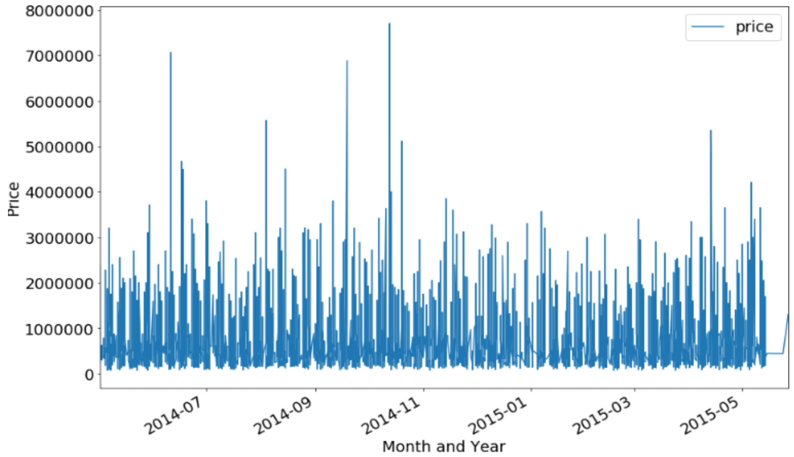

Fig. 5. Time Series Visualization of House Prices from May 2014 to May 2015

After time series indexing, the traditional ARIMA model was developed to predict housing prices based on the date sold. Four possible combinations of p,q,r parameter values were tested, and the least AIC obtained was 93.9211 for the order p=0, q=1, r=0.

3.4 Artificial Neural Network

Finally, an Artificial Neural Network was designed. The input test data was scaled after splitting the data into train and test portions. A visual depiction of the neural network designed is presented in Fig. 6.

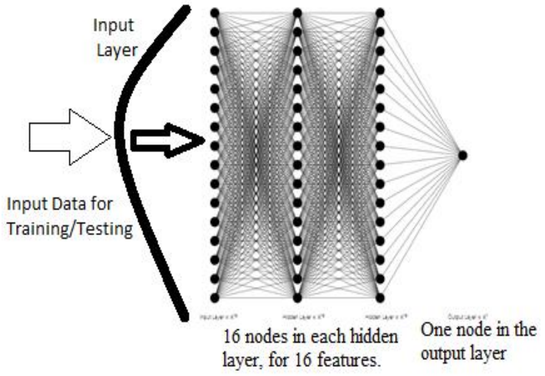

Fig. 6. Architecture of the Designed Artificial Neural Network

An artificial neural network was designed with 16 features in each hidden layer, and a total of 3 hidden layers. Training data was split into a batch size of 10, which corresponds to the input layer with 10 nodes per epoch. The output layer has one node, which predicts the probability that a particular input data has a price range either above or below the mean price. These probabilities are transformed into binary values to correspond with the original data format. The activation function used in the first and second hidden layers was Rectified Linear Unit (reLu), while the third hidden layer used the Sigmoid activation function. The ANN was trained by splitting the data into 0.25% for testing and remaining 0.75% for training. The training data was split into 10 epochs with two batch sizes of 5 and 50 to compare performance. One epoch is when the entire dataset is passed through a neural network, with back-propagations included. However, in cases of big data, it is nearly impossible to pass the entire dataset through an ANN in one epoch. Lixian He et al. [10] proposed a Python based batch processing method for to deal with large volumes of urban spatial data. Batch-wise training was hence used to divide the data into several batches, each with a uniform batch size, to reduce the computational power. A comparison of large and small batch sizes, for a constant number of 10 epochs, is made in Table 2.

. **Table 2.** Comparison of ANN performances for relatively low and high batch sizes

Batch Size	Average Time per Epoch (s)	Accuracy
5	5.5	84.94
50	0.7	84.84

4 Conclusions

To conclude, the ANN made the most accurate classification predictions with 85% accuracy, closely followed by Logistic Regression with 80% precision. Comparison of the algorithms tested is depicted in Fig. 7.

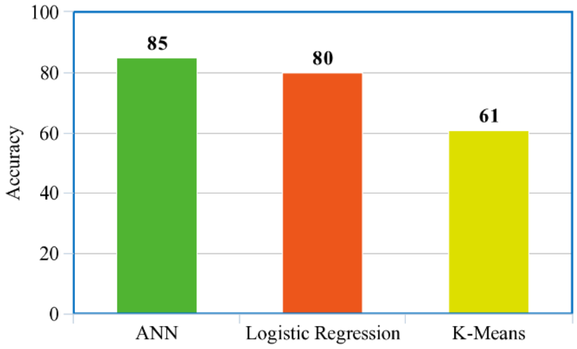

Fig. 7. Comparison of Different Algorithm’s Performances

ANN performed better than traditional machine learning algorithms like Linear Regression and K-Means clustering, while ARIMA model returned significant error. When ANN’s are applied to big data sets, the execution time may be high, but the accuracy of the predictions yielded are also likely to be high. A slight increase in accuracy is observed when the batch sizes increase, for the same ANN. This can be attributed to the small size of the dataset tested. A larger difference in performances may be evident for big data.

References

1. Kaggle, “House Sales in King County, USA”, https://www.kaggle.com/harlfoxem/housesalesprediction

2. Saranya, M. Sri, and S. Vigneshwari. "Analysis of Weather Datasets Using Data Mining Techniques."International Journal of Control Theory and Applications, Vol. 10, No. 14, 97-106, Google Scholar (2017)
3. Y. Feng and K. Jones,Comparing multilevel modelling and artificial neural networks in house price prediction, 2015 2nd IEEE Int. Conf. Spat. Data Min. Geogr. Knowl. Serv., 108–114, (2015).
4. E. Moses, P. Melwin, S.Vigneshwari, “Instant answering for health care system by machine learning approach”, International Conference on Circuit, Power and Computing Technologies (ICCPCT),1-7, Scopus and Web of Science (2016).
5. S. Vigneshwari, B. Bharathi, A. Sivasangari, study on the application of Machine Learning algorithms using R programming in gene classification and disease identification process in complex PubChem datasets, 2nd Global Conference on Computing and Media Technology, Asia Pacific University, Malaysia on Nov 14& 15, Scopus (2018).
6. D.B.Alencar, C.M Affonso, R.C.L. Oliveira, J.C.R Filho, “Hybrid Approach Combining SARIMA and Neural Networks for Multi-Step Ahead Wind Speed Forecasting in Brazil” IEEE Access, vol.6, (2018).
7. Sankari. A, Vigneshwari. S, “Automatic tumour segmentation using CNN”,Third IEEE international conference on Science Technology, Engineering Management- ICONSTEM, Chennai, Scopus (2017).
8. Vijai Chandra Prasad.R, S.Yashwanth M,Niveditha P.R, Dr.Sasipraba .T, Dr.Vigneswari S and S.Gowri ,”Low Cost Automated Facial Recognition System”,2017 Second IEEE International Conference on Electrical, Computer and Communication Technologies, SVS Engineering College, Coimbatore, Scopus (2017).
9. Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Adyan Nur Alfiyatin Faculty of Computer Science Brawijaya University, Malang, Indonesia et al. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, (2017).
10. Lixian He, Shu Gan, Yingyue Chen, “Spatial data processing based on Python language,” Value engineering, vol.36, 207-209, (2014).