House Prices Forecasting Using Time Series Modeling

Ina Tayal
001494253
INFO 6105, Spring 2019
College of Engineering,
Northeastern University

Prima Aranha
001425555
INFO 6105, Spring 2019
College of Engineering,
Northeastern University

Sindhu Raghavendra
001490547
INFO 6105, Spring 2019
College of Engineering,
Northeastern University

***Abstract* - In this paper, historical data for the past eight years from Zillow has been used to forecast the median listing price per SqFt of houses for the next few years. As part of this research idea, a hybrid method that combines the time series model ARIMA (Auto Regressive Integrated Moving Average) and ANN (Artificial Neural Network) has been proposed to make better predictions as compared to the predictions from individual ARIMA and ANN models. Experiments on some real datasets has proven that this hybrid model takes in the strength of both ARIMA and ANN models in terms of dealing with linearity and non-linearity in data respectively and gives a combined model that is effective in improving the forecasting accuracy.**

I. Introduction

The real estate market is a very integral contributor to the US economy. Affordable real estate and the changes in house prices have a direct impact on wealth and consumer spending. The real estate market has always been through highs and lows in terms of house prices.

Real estate analysis and house price forecasting is important as it enables better decision-making related to investments by consumers. Forecasting is the process of utilizing historical data to make informed predictions for the future.

In this paper, the dataset that has been used is from the website named Zillow. Zillow is an online real estate marketplace that contains over 110 million listings of houses across North America which are available for sale and rent. In this analysis, the listing price per sqft of houses will be forecasted for next two years based on eight years of historical data. This forecasting will be done through time series analysis and will enable consumers to make informed decisions and best choices related to their real estate investments.

II. Dataset

The Zillow Economics data consists of 86 variables which contain values pertaining to the historical and current listing prices of houses in the United States across all fifty states. The variables comprise of median listing prices per sqft and various types of houses available that is one bedroom, two bedroom, three bedroom, four bedroom, five bedroom or more, condo, duplex and single-family residence. Historical data starting from the year 2010 was selected for this analysis.

III. Models and Methods

Exploratory data analysis was the first step in the approach. The target variable is the median listing price per sqft. It was observed that there were ups and downs trends in the median listing per sqft of the houses that are available for sale. There is also a seasonality in these trends. It was observed that house prices decreased from year 2010 and further decreased in year 2011. House prices increased in the year 2012 past June and were at a peak during June through October (inclusive) across all years. Based on these trends and seasonality, time series modelling was used to forecast the prices of houses.

After the exploratory data analysis was performed, the time series models were used to

train the data and make predictions. The four models used were ARIMA, SARIMA, Prophet and an ARIMA – ANN Hybrid Model.
ARIMA stands for Auto Regressive Integrated Moving Average. It consists of three parameters – p, d, q where p is the number of lag observations included in the model, d is the difference of past values, and q is the linear combination of error terms.

The p value of 15 was selected by plotting the autocorrelation graph. The value of d was chosen as 2 and the q value was chosen as 0. The model was fitted using the selected values for the p, d, q parameters. The residuals were plotted. The plots showed that there were residual errors in the forecast.
The data was split into training and test with 60 percent training data and 40 percent test data and then validated using the forecast() method. After forecasting the median listing price per sqft for the next 12 months, it was observed that there was a seasonality present. Hence, the next model that was fitted was SARIMA.

SARIMA is an extension of ARIMA. In addition to the parameters used by ARIMA, it also has four hyperparameters P, D, Q and m.

The hyperparameters was chosen through grid search method. The model was fitted using SARIMAX function and the hyperparameters. The model was then validated and split into training and test. The actual vs forecasted values graph was plotted which showed little variance between the two.

Prophet is a time series forecasting model that works well with data that has many seasons of historical data as well as strong seasonal effects. This model is very robust and handles outliers and noisy data very well.

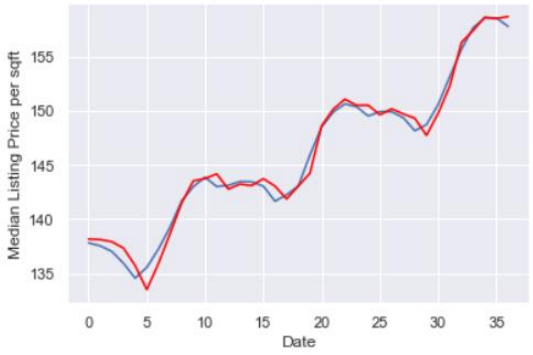
The Prophet model was fitted on the data frame that contained the date ranges for which forecasting was to be made.

The ARIMA – ANN hybrid has been constructed using the ARIMA and Artificial Neural Networks (ANN). ANN is a framework which has many machine learning algorithms which work together to process complex data. It works well with non – linear data. This hybrid model is constructed by feeding the residuals obtained from fitting the ARIMA model as an input to the ANN. These residuals are non-linear in nature and have not been captured by the ARIMA model. The ANN then uses this input to make predictions about the ARIMA error terms.

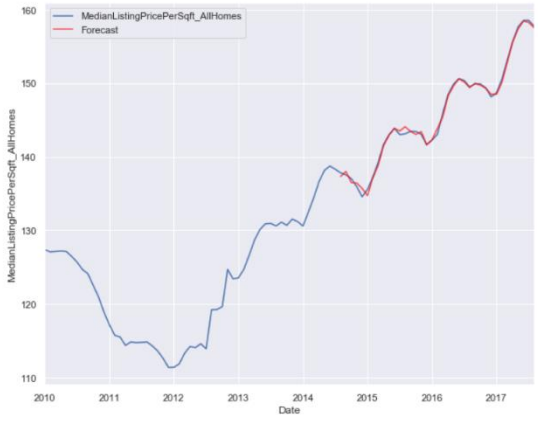
The model made accurate forecasting with the actual vs predicted values having very little variance.

IV. Results

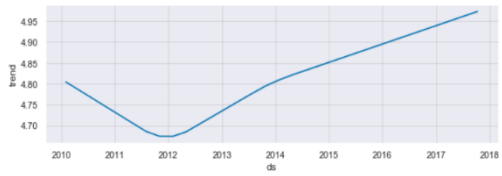
Forecasting results from ARIMA model is given below:



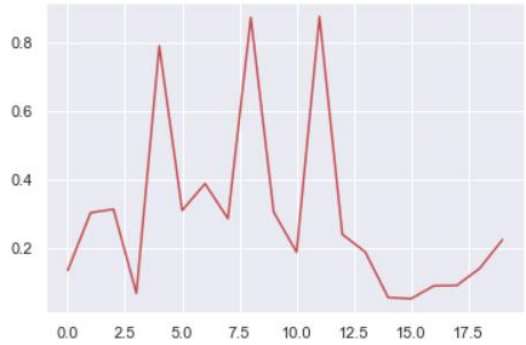
Forecasting results from the best SARIMA model:



Forecasting results from Prophet model:



Forecasting results from hybrid ARIMA – ANN:



V. Discussion

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) have been used as evaluation metrics to estimate the accuracy of the forecasted results for ARIMA and SARIMA. It has also been used to compare the predicted values of the fitted models to determine which model performed the best in terms of forecasting.

ARIMA model with p, d, q values of (15, 2, 0) gave an RSME value of 0.81. Among the three SARIMA models, the model with order = (2, 0, 4) and seasonal_order = (3, 1, 2,12)12 gave the lowest MSE value of 0.174 and was the best model.

The forecasted values in these models were very close to the actual values and did not have a lot of variance. The ARIMA – ANN hybrid model's output graphs showed an even closer match between the actual and predicted values.

VI. Scope

The domain of the research project was to analyze the effectiveness of a hybrid model of ARIMA and ANN (Artificial Neural Networks) over individual models. We have analyzed the results and accuracy of each individual model as well as the hybrid model and concluded that the forecasting accuracy of the hybrid model was better than that of the individual models.

VII. Context

For this research project, we have referred the following notebook (https://www.kaggle.com/rgrajan/time-series-exploratory-data-analysis-forecast) where time series models ARIMA and SARIMA were implementing for forecasting. We have generated new models for SARIMA by hyperparameter tuning and implemented other time series models like Prophet and the hybrid ARIMA – ANN model.

VIII. Conclusion

The findings demonstrate that the ARIMA – ANN Hybrid model fits the data well and makes more accurate predictions than the individual models. It handles both linearity and non – linearity in the data and for complex problems which have them both, the hybrid model will be effective for forecasting.

IX. References

[1] Data source: https://www.kaggle.com/zillow/zecon
[2] G.Peter Zhang, Time series forecasting using a hybrid ARIMA and neural network model, https://www.sciencedirect.com/science/article/pii/S0925231201007020
[3] https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-arima-in-python-3
[4] https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/
[5] https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
[6] https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-prophet-in-python-3
[7] https://www.kaggle.com/rgrajan/time-series-exploratory-data-analysis-forecast