Machine Learning Report

The objective of this project is to forecast the future price of houses in the San Francisco Bay Area by using time series data of housing prices by zip code in the Bay Area. As discussed in earlier reports, the housing market in the Bay Area is more aggressive when compared to other parts of the United States meaning the prices are higher and seem to still be growing.

Since the data is time series, methods learned up to the point in the course, such as linear regression, weren't appropriate for this project. Instead, an ARIMA model was used to forecast the data. ARIMA stand for autoregressive integrated moving average, which means it uses a number of different parameters to model data to control a certain aspect of the ARIMA model. One parameter represents an autoregressive lag (*p*), another the differencing order (*d*), and lastly the moving average order (*q*). The ARIMA model is most effective if the data being modeled is considered stationary, that means that over time, the mean, variance, and autocorrelation are constant over time. If your data isn't stationary, there are a number of methods to induce to become stationary.

There are two methods to test whether your data is stationary. The first is to visually plot your data and visually inspect if there's any pattern to the mean and variance over time. If the plots look relatively flat and without form, then you could say your data is stationary. For a numerical/statistical approach, the Dickey-Fuller test could be used. This test returns a statistic and p-value which could be used with a predetermined confidence level to conclude is your data is stationary.

The most effective way to induce a stationary dataset is to difference out trends such as the moving average, weighted moving average, shifted data, and seasonality. The the number of times the data is differenced will determine you parameter *d*. Parameters *p* and *q* are determined by plotting the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the stationary data. Where these plots reach a predetermined significance level will determine your parameters *p* and *q*.

After a model is fitted to the data and given a RSS/RSME score, it is used to forecast the housing price for the next period in time with a 95% confidence interval. Functions were created in Python to help automate the process.

For purposes of this project, the data was aggregated by county and the average of the prices were taken. That left a group of the nine Bay Area counties with housing price data dating back to January 2012. January of 2012 was taken instead of 2010 as in the previous reports because it was proving difficult to induce the data to become stationary. The data was then split 70:30 into training and testing sets. The first 70% of the data was taken chronologically to be the training set and the last 30% to be the testing set.

The logarithm was taken of each of the 9 data sets to reduce the scale and then each was passed to the four differencing techniques mentioned above to determine which produced the most stationary dataset.

For the Alameda training set, differencing provided a significant enough p-value (0.023) to consider it stationary. Fitting the ARIMA model to the Alameda training set with parameters according to the first order differenced stationary data, it produced a reasonable model with

roughly normalized residuals and a RMSE of $17.44. When forecasting the next 24 months and comparing it to the test data, the RMSE inflates to $32,834, however the overall trend of the data seems to be captured as the actual data lies within the 95% confidence interval predicted by the model.

The data for Contra Costa county proved to be difficult to get to become stationary with the weighted moving average technique providing the smallest p-value (0.064) and most stationary result. The fitted ARIMA model produced a roughly normalized residuals and a RMSE of $165.71 on the actual data. When forecasting data and comparing it to the test set, the values tend to diverge from each other with the actual values outside the 95% predicted confidence interval after about 12 months. More fine tuning of the model will need to be considered to capture the trend better or perhaps more differencing to produce a more stationary training set.

Unlike Contra Costa county, Marin county didn't pose as many problems to induce it into becoming stationary. The decomposition technique produced a residuals that gave a p-value much smaller than 0.01. Since the decomposition technique removes the overall trend and seasonality trend, the differencing order becomes 2. However, the resulting model didn't produce normalized residuals of the fitted values and stationary values; the plot seems to be right skewed. The model still performs well producing a RMSE of $59.55 on the training set and $15,022 on the test set capturing the trend of the data. One thing to note though is the exponentially increasing confidence interval of the model which signifies the models short term value to forecast data reliably.

Napa county training data produced the most stationary data set when differencing the exponentially weighted moving average (p-value $\ll$ 0.01). The Napa county model predicted the training set with a RMSE of $219.45 and test set with $59,141. The model was also able to predict the test set within the 95% confidence interval.

The only method that produced stationary values of the San Francisco training set was the decomposition technique but even then, the residual plot doesn't appear to be normal and skewed to the right. The following ARIMA model predicted the training set with a RMSE of $30.21 but miscalculated the test set forecast by giving a $275,505 RMSE. By looking at the forecast plot, even though the test data was captured by the models 95% confidence interval, the model predicted a negative trend in the forecast and exponentially increasing confidence intervals signifying either poor fit or unpredictability in the housing price in San Francisco county.

The model for San Mateo county was used to fit the residuals of the decomposed San Mateo data. It gave an output that predicted the training set with a RMSE of $507.42 and while it scored a RMSE of $63,676 on the test set, it correctly predicted the data set with 95% confidence given the volatility of the data in recent months.

Double differencing the Santa Clara county training set produced the most stationary data I was able to work with for this instance. The model performed well on the training set scoring a RMSE of $56.32 but when it came to the test set, like San Francisco county, it scored a high RMSE of $186,182 while correctly predicting the test data within the 95% confidence interval of the model. That's partly because the model has an exponentially increasing confidence interval which alludes to the short term value of this model for this data set.

For Solano county, double differencing also produced the most stationary data set resulting in an ARIMA model that scored a RMSE of $33.63 on the training set. For Solano county's test set, the model was able to capture the data within it's 95% confidence interval and score a RMSE of $12,061.

Lastly, for Sonoma county, decomposing it produced stationary residuals that the model was able to use to predict the training set with a RMSE of $113.03. Forecasting it and comparing it to the test set scored a RMSE of $5,106 which is exceptionally well when compared to the other Bay Area counties.