Time Series Final Project

Efrem Ghebreab, Okeefe Niemann, Patrick Poon

March 2021

For this project, we are tasked with making a forecast of monthly median sold prices for housing in California during 2016. We are given given three features between the years 2008-2015: Unemployment Rate, Median Mortgage Rate, and Median Sold Price. In total, six model types were prepared and tested against a validation set created with the values between 12-1-2014 to 12-30-2015. The TES Univariate Exponential Smoothing model was the most accurate using the metric root mean square error, and was chosen to forecast the median price houses were sold between 1-31-2016 and 12-31-2016.

1 EDA of Zillow Dataframe

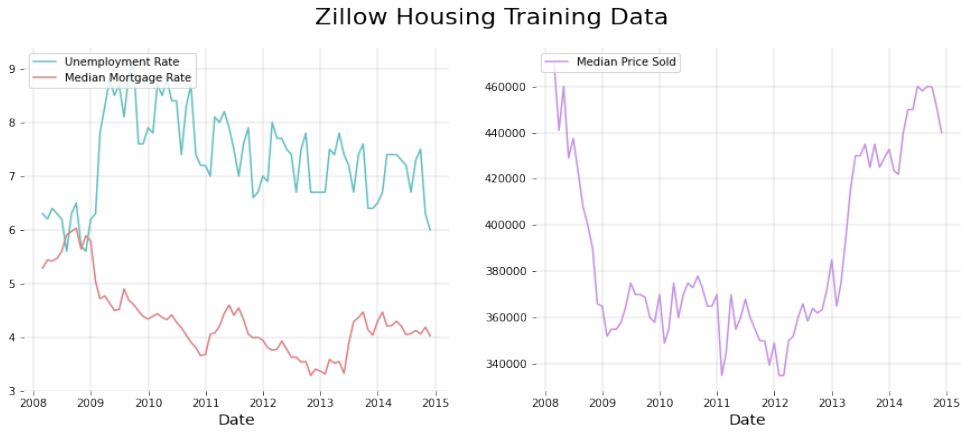

Figure 1: Decomposition for the training portion of the Zillow dataframe $2/2008 - 11/2014$. The domain for features Unemployment Rate and Median Mortgage Rate are between $3 - 10$ while Median Price Sold is of order hundreds of thousands.

An initial glance at the given features show a linear trend for Unemployment Rate and a quadratic trend for the other two. This informs us that we need to difference the data at least once to turn them stationary. In terms of seasonality, Median Mortgage Rate shows seasonality of 2 years ($m = 24$) where the others exhibit 1 year ($m = 12$).

2 Candidate - SARIMA

A classical uni-variate model was fitted to provide a baseline of model performance (root mean squared error). From inspecting the median sold price (Figure 1), it was found that there was no consistent seasonal pattern at first glance.

First, auto-correlation function (acf) and augmented Dickey-Fuller (adf) tests were performed; acf test provides us potential orders for the SARIMA model and adf test can check stationary condition. The adf test of the original data results in a p-value of 0.95, which explains that the series is not stationary. Thus, data differencing was needed. After differencing the time series by one time interval (one month), the p-value from the adf test is 0.027, which is less than 0.05. This satisfies the stationary condition for performing SARIMA models.

From the inspection of the acf graph, there is a slight seasonal pattern; therefore a seasonal component (m = 12) was included. To select the orders for the best SARIMA model, step-wise search using Akaike information criterion (AIC) as a comparison metric was conducted. The best model (lowest AIC) of SARIMA was (3,1,4) (0,1,0,12).

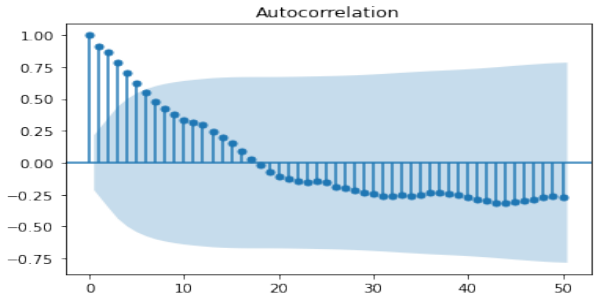

Figure 2: ACF plot of time series showing seasonal pattern of m = 12

SARIMA (3,1,4) (0,1,0,12) was used on the validation set of the median housing prices to understand the model performance. The root mean square error (RMSE) was 9991. The validation prediction contains similar up and down trends as the actual validation set, but do not match quite much with the time.

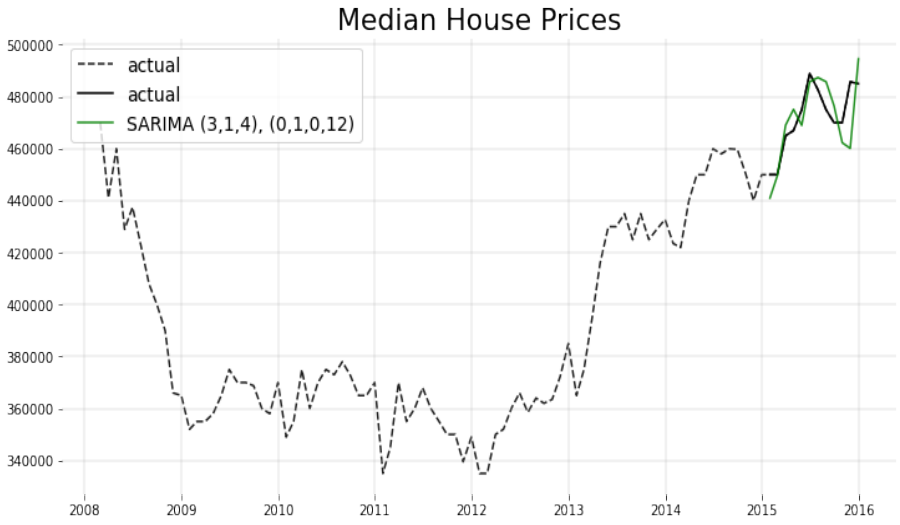

Figure 3: SARIMA validation for Median Housing Prices in the year 2015

3 Candidate - Prophet

Prophet was developed by Facebook for analyzing and forecasting time series data based on mostly additive models. For this task, two models were developed using Prophet: 1) Prophet Model with Median Mortgage Rate, and 2) Prophet Model with Median Mortgage Rate and Unemployment Rate. Essentially, we are comparing the Prophet Model with one and two regressors.

The training data was further divided up into train data and validation data. The data from 2008-2015 was considered as training data, and data from 2015-2016 was for validation. Similar to the previous comparison metric, RMSE was used.

	Model 1	Model 2
Regressors	Median Mortgage Rate	Median Mortgage Rate and Unemployment Rate
Validation RMSE (in $)	7257	8804

From the table above, it is shown that Prophet model with both unemployment rate and median mortgage rate included as regressors perform better than the baseline ARIMA model. Both graphs were plotted (fig:prophet) below:

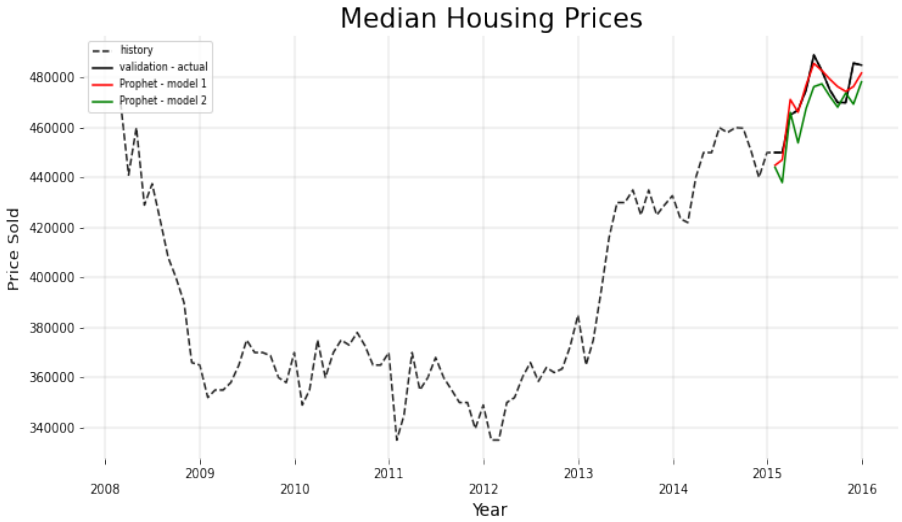

Figure 4: Prophet models (Prophet - model 1 represents Prophet with Median Mortgage Rate Regressor; Prophet - model 2 represents Prophet with Median Mortgage Rate and Unemployment Rate Regressors) Validation for Median Housing Prices in the year 2015.

4 Candidate: Vector Auto Regression (VAR)

Looking at the features given in the dataset, intuition told me that they all could possibly influence each other. With the traits of a potentially endogenous relationship, VAR seemed to be the best first attempt of the more primitive models.

To use the VAR model, we first had to verify that all features were stationary. If they weren't, the feature would be differenced with one earlier time-step of itself. The feature "Unemployment Rate" had to be differenced twice to pass the Augmented Dickey-Fuller test, with the p-value on the first difference being well above a 5% threshold ($p = 0.39$). All three features had to be twice-differenced as a result to stay in the same differenced domain. The results are shown below.

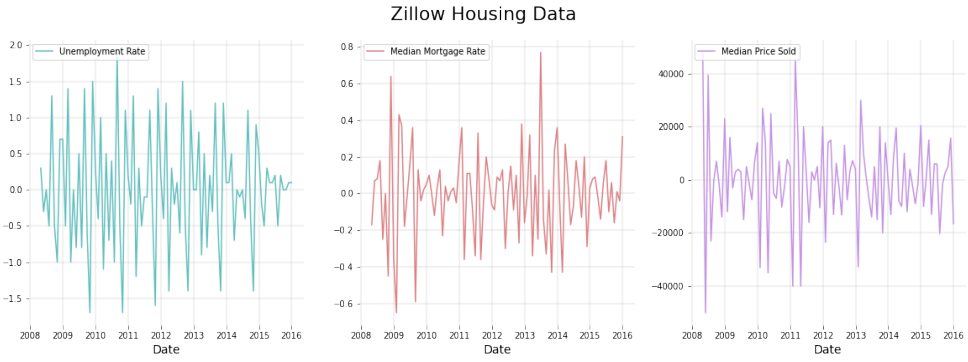

Figure 5: Twice-Differenced Data. ADF Fuller confirms intuition that all components are now stationary.

In addition, the values were all scaled between the values 0 and 1. The distributions were left as they were un-scaled. This is to eliminate bias in the feature weights. When running this model, we obtain the following root mean squared error. Refer to the figure below for graphical representation.

$$RMSE = 22956.7$$

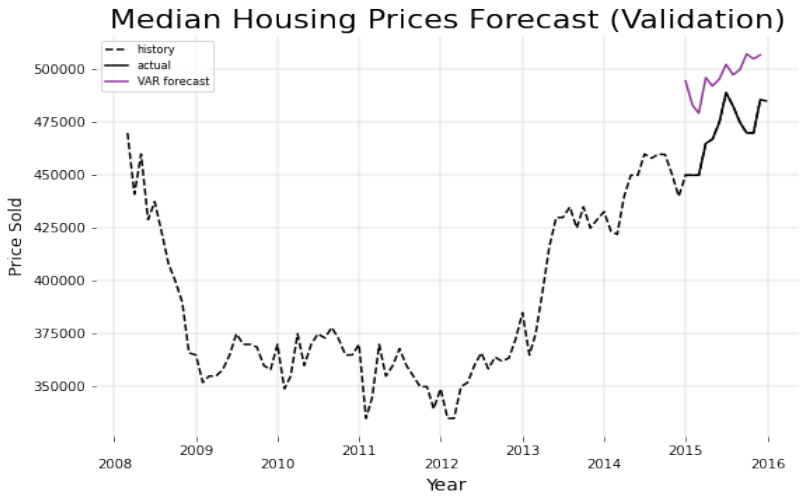

Figure 6: VAR Forecast for Median Housing Prices in the year 2016. Forecast slightly overestimates, comes close to accurate by the end of Quarter 1, before ending on a deviated trajectory.

5 Candidate: SARIMAX

For this next model, I decided to be a little more close minded and create a SARIMAX model. This model is best suited for exogenous relationships, where past rates will be the best way to inform the median price houses were sold at throughout 2016.

For preparation, the data only needed to be scaled down to the values between zero and one before being fed into the SARIMAX model. Though it was easier to execute, it had a slightly higher mean squared error than Vector Auto Regression:

$$MSE = 11280$$

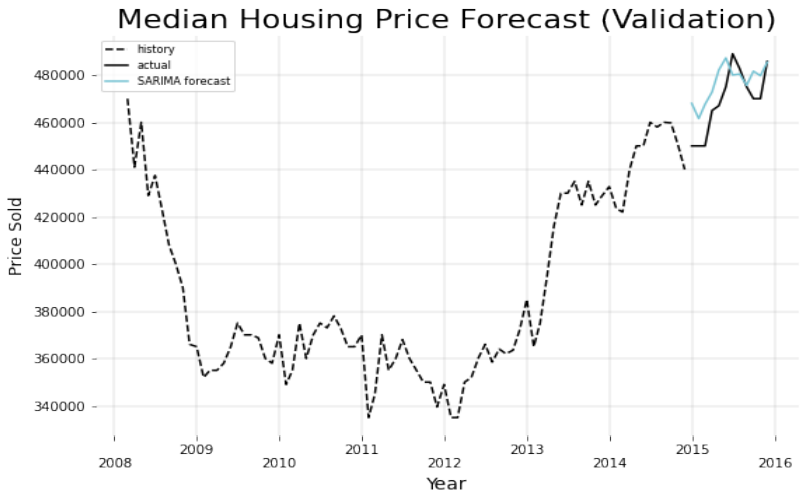

Figure 7: SARIMA Forecast for Median Housing Prices in the year 2016. Forecast overestimates more often than not, but does a decent job at identifying the general trend of the housing market between $2015 - 2016$.

6 Candidate: TES Univariate Exponential Smoothing

TES Univariate Exponential smoothing was performed on the data. The Data was split initially in validation and training sets for initial evaluation of the models. Five different ETS models were tried with different trends and seasonality. Model-1 is an additive trend and seasonality, Model-2 with additive trend and multiplicative seasonality. Model-3 is multiplicative in both trend and seasonality, Model-4 multiplicative in trend with additive seasonality and Model-5 has no trend and additive seasonality. Forecasting was tried in different ways, one taking 90 percent data for training and 10 percent for validation from the original zillow training set. The following results of RMSE were retrieved.

Model	Validation Score
rmse-1	9411
rmse-2	12763.9
rmse-3	14429.094
rmse-4	7164.610
rmse-5	16131.057

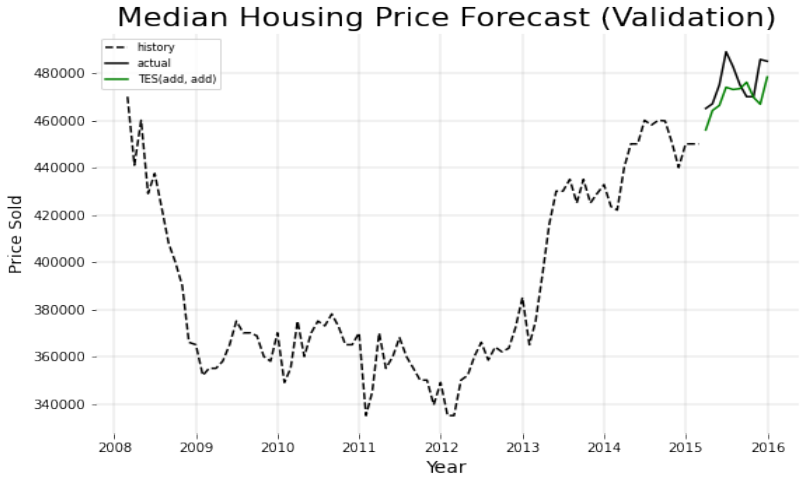

Figure 8: TES Forecast for Median Housing Prices depending on the validation set in 2015 compared to the actual results. Forecast represented in green from the best selected model seems to have a close estimate to the actual price.

Based up on the predictions from the validations in the uni-variate Exponential smoothing model, the lowest RMSE score was achieved from Model-4 with multiplicative in trend and additive in seasonality.The RMSE score was 7164.6.

7 Choosing a Model, Final Forecast (Jan - Dec 2016)

After training six separate models and testing them on the validation set consisting of median housing price in California in the year 2015, the following Root-Mean-Square-Error values were found:

Model	Validation Score
SARIMA (3,1,4)(0,1,0,12)	9991
TES Univariate Exponential Smoothing	7164
Vector Auto Regression (VAR)	28932
SARIMAX(3, 1, 0)(0, 1, 1, 12)	11280
Prophet (Median Mortgage Rate)	7257
Prophet (Median Mortgage Rate, Unemployment Rate)	8804

These validation scores show that Uni-variate Exponential Smoothing is the most accurate when using to forecast median housing price in California. As a result, we will use it to forecast against the test set. Univariate Exponential

Smoothing multiplicative in trend and additive in seasonality was tested on unseen test set to forecast median housing price and RMSE score of 13913.4 was reported. As shown below in the graph it was able to predict very similar to the actual prices in the most part of 2016 with minor difference on the last quarter of 2016.

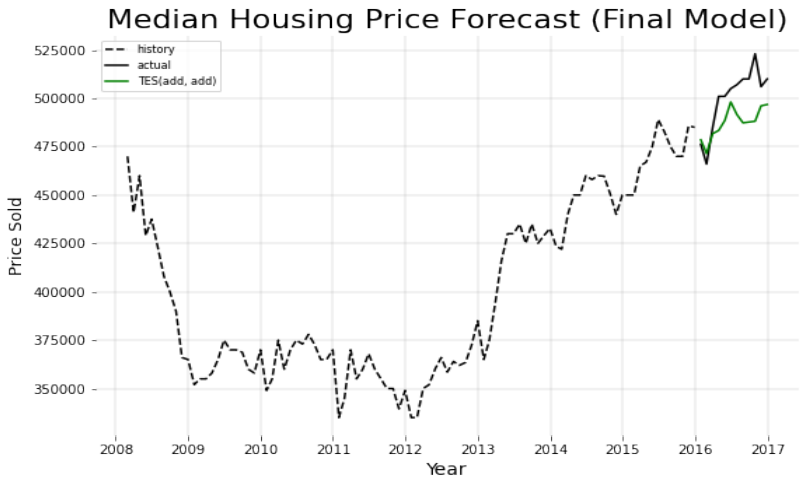

Figure 9: Final TES Forecast for Median Housing Prices depending on the test set in 2016 compared to the actual results. Forecast represented in green from the best selected model seems to have a close estimate to the actual price.

$$\boxed{\text{RMSE} = 13913.4}$$

8 Summary

The goal of this project is to use time series forecasting methods to predict the monthly median sold price for housing in California from January 2016 to December 2016. In addition to median sold price, unemployment rate and median mortgage rate were provided as well. The total time duration of data spans from 2008 to 2016. For this analysis, training data was from 2008 to end of 2014. Validation period was from January 2015 - December 2015.

Multiple time series models were utilized in this analysis, such as SARIMA(X), TES, Prophet, and VAR. To choose the final candidate model, RMSE was chosen to be the metric as the validation score. It was shown that TES with Univariate Exponential Smoothing performs best on the validation set.

Lastly, this model was used on the test set to observe how well it performs. Similar to the validation score, RMSE was used. An RMSE of 13913.4 was calculated from the test set. The TES forecast matches well in the first two-third of 2016, but there was slight change in the later third.