Introduction

House price has a great impact on everyone's life. The goal of this report is to research how we can predict the median future sold price for housing in California based on time series models. Here, we are using the Zillow dataset recorded Feb 2008- Dec 2015 monthly as our training set, including median sold price for housing in California, median mortgage rate, and unemployment rate, and trying to predict the Jan- Dec 2016 median sold price for housing in California.

1.1 Data Exploration

	Date	MedianSoldPrice	MedianMortageRate	UnemploymentRate
0	2008-02-29	470000.0	5.29	6.3
1	2008-03-31	441000.0	5.44	6.2
2	2008-04-30	460000.0	5.42	6.4
3	2008-05-31	429000.0	5.47	6.3
4	2008-06-30	437500.0	5.60	6.2

(fig. 1.1 The first five rows of the training data)

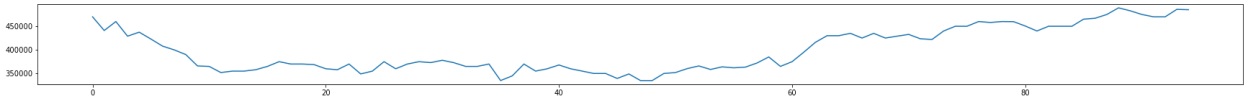
Variable	Number of records	Data type
Date	107	datetime
MedianSoldPrice	95	float64
MedianMortageRate	107	float64
UnemploymentRate	107	float64

The dataset consists of 95 observations with 4 columns:

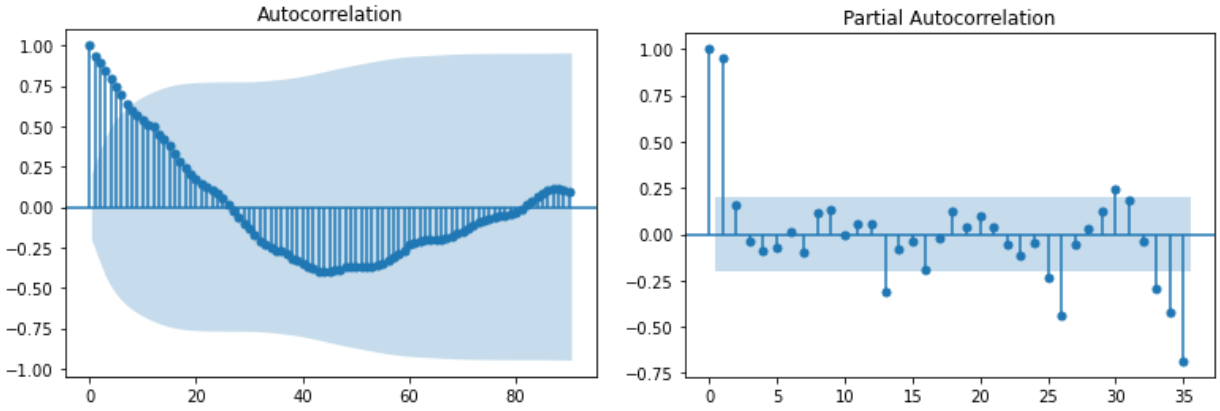
- **Date** gives information about the date of each record.
- **MedianSoldPrice** gives information on the median sold price of houses in California.
- **MedianMortageRate** is the median mortgage rate of the houses.
- **UnemploymentRate** is the unemployment rate for that month.

Firstly, we would be performing exploratory data analysis to get some insights from the data to understand it better, and find some guidelines for choosing models.

1.1.1 Initial plot of median sold price for housing in California



(fig. 1.1.1-1 initial line plot of median sold price for housing in California)



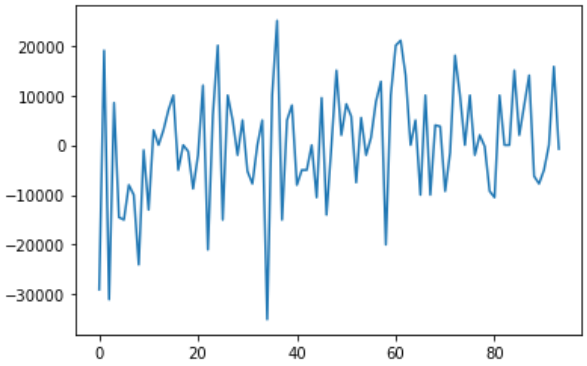

```
Test Statistic   -0.058792
p-value           0.953391
dtype: float64
```

(fig. 1.1.1-2 ACF and PACF plot of median sold price for housing in California)

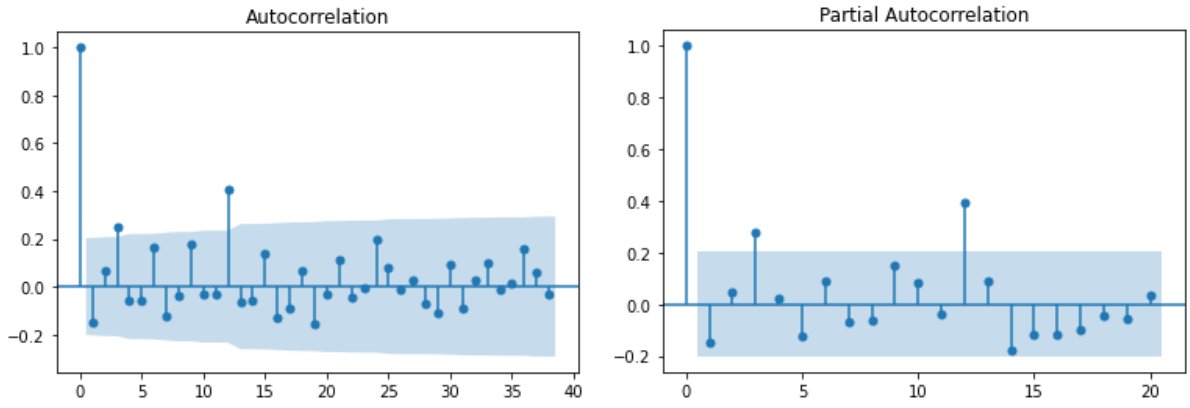
From the initial plot, we can see that the housing price data showed a trend of first decline and then rise and it seems that the data is not stationary. From the ACF and PACF plot we don't see any obvious seasonality pattern in the plot, but we will explore it further. We proceeded the ADF test and got the p-value with 0.95339, which was bigger than 0.05 and confirmed that the initial data is not statistically significant stationary, we need to do differencing.

1.1.2 Differencing the initial data

We chose the differencing order to be 1, and drew the plots again.



(fig. 1.1.2-1 line plot of differenced median sold price for housing in California with order 1)




```
Test Statistic   -3.088139
p-value           0.027443
dtype: float64
```

(fig. 1.1.2-2 ACF and PACF plots of differenced median sold price for housing in California with order 1)

From the plots we see the time series plot gets a bit flatter, we also do the ADF test and get the p-value with 0.027443. We see from the ADF test that p value=0.027443 < 0.05. So we can reject

the null hypothesis and accept that the given process is stationary. Therefore, we considered the differenced median sold price for housing in California with order 1 as stationary. From the line plot, ACF plot, and PACF plot, we cannot find obvious seasonality.

1.2 Model Selection

Because we did not find seasonality in the above plots, firstly we decided to choose the ARIMA model instead of SARIMA. And we know that while linear exponential smoothing models are special cases of ARIMA models, the non-linear exponential smoothing models have no equivalent ARIMA counterparts. Therefore, the ETS model is also worth fitting. In addition, we also choose the Prophet model. Although the Prophet model is most appropriate for daily or sub daily time-series dataset, we still want to take more possible models into consideration to see if we can get smaller RMSE. At last, apart from the target time series variable MedianSoldPrice, we also have MedianMortageRate and UnemploymentRate, which may have stable impacts on housing price. Hence, we also chose SARIMAX to fit multivariate time series models to see whether we can get better predictions. As a result, ARIMA, ETS, Prophet, SARIMAX were chosen to fit models and do cross validation.

About Candidate Models:

For our Analysis, we are broadly using three different types of model families.

1. ARIMA Family

The ARIMA family is the largest with possibly infinite candidate models. Arima models assume that the future values of our time series data depend on the past data points, the periodic changes and the shocks from previous measurements/values. Hence, for any Time Series forecasting, we divide the problem into three sub-problems:

1) estimating the general trend of our data
2) estimating the seasonality and
3) adding the remaining noise part to the model.

ARIMA(p, d, q) models are composed of the Auto-Regressive component (AR(p)) and the Moving Average component MA(q). The order of the AR component (p) is responsible for determining the extent of dependency on previous data points and the order of the MA component (q) determines the

dependency on the shocks from the previous data points. In other words, p and q determine how far back in history data the model utilizes for forecasting. 'd' determines the order of differencing needed to make the time series stationary. For our initial analysis, we ignored the seasonality component in the median housing prices time series and looked at the trend and the remaining noise part for model simplicity.

SARIMA: In order to add the effect of Seasonality, we use the SARIMA model family. The SARIMA model has an additional set of parameters (P, D, Q) to incorporate the seasonality of the time series. Here, P and Q signify the order of Seasonal AR and MA components and D signifies the number of times seasonal differencing needs to be done to make the time series stationary. We also define a seasonality period (m) to specify the number of intervals after which we see a repeat in pattern. According to some studies conducted by the National Association of Realtors, Median housing prices do show some seasonality with prices increasing in Summers and dropping in the months of December and January. Though we can't observe a strong seasonality pattern in our data, slight seasonal patterns can be seen towards the later half of our dataset. Hence, we can include the SARIMA models as well in our analysis.

Adding a further layer of complexity, we use SARIMA models with Exogenous variables. Exogenous variables are other time series variables which may or may not directly impact the behaviour of our target time series variable. In our case, we use Unemployment rate and Mortgage rate as our exogenous variables to predict median housing prices. It is intuitive that the change in these variables may affect the future housing prices. Hence, we include these two variables as our exogenous variables.

2. Exponential Smoothing Models

Most Exponential Smoothing models are a subset of the ARIMA family. Exponential smoothing methods are similar to ARIMA models since prediction is again a weighted sum of past observations, However the model explicitly uses exponentially decaying weights as we include further historical observations to our model. The past observations are given weights in a geometrically decreasing ratio. For our Analysis, we consider all possible combinations of trend and seasonality types in the Holt-Winter Exponential Smoothing function. Different

permutations of no trend, additive and multiplicative trend along with similar seasonality types cover all possibilities of the single, double and triple exponential models.

3. Prophet Model

The prophet Model is ideal for daily or sub daily time-series dataset. It performs extremely well on daily/sub-daily business time-series data. Hence, it is not a suitable model for our case. However, as we are considering different possible candidate models to select our final model, we decided to also include the prophet model in our analysis. Prophet generally uses additive models and fits non-linear trends with yearly, weekly, and daily seasonality, plus holiday effects.

Model Selection Process:

The Algorithm that we use to select the best candidate model can be summarized in these three steps:

Step 1: In order to find a list of candidate models, we perform grid-search on a set of candidate model parameters.

For this purpose, we use Akaike information criterion (AIC), Bayesian information criterion (BIC) and OOB (Out-of-bag) as our information criteria. The model parameters which result in the minimum information criteria are taken as our candidate models. The model parameters signify the number of additional independent variables (terms) that we are adding to our model for better prediction. The best-fit model according to these information criteria is the one that explains the greatest amount of variation using the fewest possible independent variables.

Step 2: We perform one-step cross validation to find the prediction performance of our candidate models. We use an initial train-test split of 67% - 33%.

We use RMSE (Root Mean Squared Error) and MAPE (Mean Absolute Percentage Error) as our performance criteria to compare the performance of candidate models. Both these performance criteria have individual advantages and limitations. RMSE gives an overall better estimate for errors in model predictions, since it penalizes large deviations more and diminishes smaller deviations. However, RMSE is sensitive in case of outliers. MAPE is considered a better

candidate if we have outliers or a large range of values in our dataset. However, MAPE has a significant disadvantage that it gives undefined values when we have data close to zero, which is rarely the case in housing prices data.

Step 3: We compare the cross-validation performance criteria and select the best performing model among our different candidate models for final prediction.

2. Model estimation and validation

UNIVARIATE MODELS

From the observation of time series plots, ACF, and PACF plots, no seasonality can be seen. However it is also reasonable to assume a seasonality of 12 for housing prices. Hence, our general idea of establishing univariate models is to consider the situation with or without seasonal factors respectively.

When modeling only considering trends, we select candidate ARIMA models based on our intuition (observation), AIC, and BIC separately. Then use RMSE and MAPE as our criteria to choose the best model for this case.

When modeling with both trends and seasonality, we apply ARIMA family, ETS, and Prophet to fit models separately and then compare the candidate models using RMSE and MAPE.

2.1 Case 1: Only trend no seasonality

Candidate 1 - ARIMA(0, 1, 0)(0, 0, 0)[0]

According to the time series plot and ACF , PACF plots after differencing(fig. 1.1.2-2 ACF and PACF plot), we can see that there is a clear shutdown in the ACF plot and PACF plot at order 0. Hence, our first candidate model is ARIMA(0,1,0)(0,0,0)[0].

Candidate 2 - ARIMA(3, 1, 0)(0, 0, 0)[0]

Instead of selecting the order of the ARIMA model based on observation, we use AIC to choose our candidate 2. Here we set max_d equals 2 to let the auto_arima method determine ‘d’ and set seasonal to false. Auto_arima function gives us the second candidate, ARIMA (3, 1, 0)(0, 0, 0) [0].

Candidate 3 - ARIMA(0, 2, 4), (0, 0, 0) [0]

Because AIC and BIC could give different selections for models, here we use bic_sarima to do model selection and select our third candidate. Here we set d values with range [0,2] to let the bic_sarima method to select the differencing order.

Then we perform one step cross validation to calculate the RMSE value and MAPE value to choose the best model from these three candidates, with 33% of the whole data to be our validation data set. Below are the results.

Candidates	Method Used	Parameters Used	Candidate Model	RMSE	MAPE
Candidate 1	observation	From plots	(0,1,0)(0,0,0)[0]	8530.276959	0.014816
Candidate 2	auto_arima	Max_p=7 Max_q=7 m=1 d=None Seasonal=False	(3,1,0)(0,0,0)[0]	8658.186728	0.015044
Candidate 3	bic_sarima	p=range(0,5) d=range(0,3) q=range(0,5) Seasonal components=None	(0,2,4)(0,0,0)[0]	11221.627693	0.020938

(fig. 2.1 CV results of case 1- only trend no seasonality)

From the table above, we can see The RMSE of candidate 1 is 8530.27 and the MAPE of candidate 1 is 0.014816, which are both smaller than the criterias value of candidate 2 and 3. Therefore, our choice of case 1 is ARIMA (0, 1, 0)(0, 0, 0)[0].

2.2 Case 2: With trend and seasonality

Based on our daily life experience and the law of housing price changes, we assume that the period of seasonal changes in housing prices is one year. Hence, we let m equals 12 to all the models below.

Candidate 1 & 2 - ARIMA(3,1,0)(1,0,1)[12] & ARIMA(1,2,2)(0,1,2)[12]

Same as case 1, here we use auto_arima and bic_sarima methods to select the trend order and seasonality order of our candidate models. The differences are we set m equals 12 and Seasonality to be true. AIC gives the model ARIMA(3,1,0)(1,0,1)[12], while BIC recommends ARIMA(1,2,2)(0,1,2)[12]. Below are the results.

Candidates	Method Used	Parameters Used	Candidate Model	RMSE	MAPE
Candidate 1	auto_arima	Max_p=7 Max_q=7 m=12 d=None Seasonal=True, max_P=4, max_Q=4	(3,1,0)(1,0,1)[12]	8403.862462	0.015245
Candidate 2	bic_sarima	p=range(0,4) d=range(0,3) q=range(0,4) P=range(3) Q=range(3) m=12 D=1	(1, 2, 2), (0, 1, 2) [12]	14014.179070	0.025861

(fig. 2.2-1 CV results of case 2- ARIMA model with both trend and seasonality)

We can see the RMSE of candidate 1 is 8403.86 and the MAPE of candidate 1 is 0.015, which are all smaller than candidate 2. Therefore, we choose **ARIMA(3,1,0)(1,0,1)[12]** to be our choice for the ARIMA family in case 2.

2.3 Candidate 3 - Holt Winters Exponential Smoothing

Moving on to the ETS model, because we can not observe obviously additive semble or multiplicative semble in the original time series plot, we set additive, multiplicative, and None to the trend order range and seasonality order range respectively. Still let 33% of the train data be our initial validation set and perform one step cross validation. We get the following results.

Trend	Seasonality	RMSE	MAPE
Additive	Additive	12473.662181	0.022000

Additive	Multiplicative	32843.755670	0.037153
Additive	None	9949.351595	0.018259
Multiplicative	Additive	3.852408e+61	1.602391e+55
Multiplicative	Multiplicative	6.181354e+46	3.506324e+40
Multiplicative	None	9846.882342	0.018000

(fig. 2.2-2 CV results of case 2- ETS model with both trend and seasonality)

Here, we can see the model with multiplicative trend order and additive seasonality order failed to converge. The best model for this scenario based on lowest rmse score is **Trend: Additive, Seasonality: None**. But the RMSE is 9949, which is bigger than the RMSE of candidate 1 with the ARIMA model.

2.4 PROPHET

Our last candidate for the univariate model is fitted by the Prophet. As we talked above, the Prophet model performs best for daily data. But in order to take more possible candidates into consideration, we still fit Prophet models to see if we can get a smaller RMSE.

Here we fit two candidates with the Prophet model. Candidate 4 is fitted with the default Prophet. We fit this candidate based on 67% of the training data, and do one time validation on the left 33% data set. For candidate 5, cross validation is utilized. We set 'freq' equals 'M' which means the model is fitting on monthly data, and 'period' equals 1 to let the model predict 1 step forward each time.

Below are the results.

Model	Model Parameters	RMSE	MAPE
Candidate 4	Default Prophet	22482.005808	0.039515
Candidate 5	Prophet Cross Validation	11513.81296	0.02156

(fig. 2.4 Results of case 2- Prophet model)

We can see that both RMSE and MAPE of candidate 5 are smaller than candidate 4. Therefore, the model fitted based on cross validation is better, but the RMSE is still bigger than candidate 1 and 3, we may not select it as our final choice

2.5 Multivariate Time Series Models

SARIMAX

For the next part of our analysis, we also include the exogenous variables that we have for the given time period. Unemployment rate and Mortgage rate are the exogenous variables provided for the analysis. It is intuitive that both of these can affect the future housing prices. Hence, we can say that there is a one-sided causal relationship between Unemployment rate and Mortgage rate to the Median Housing Prices. Therefore, SARIMAX models are ideal for such scenarios. We did not consider the VAR (Vector Autoregression) model in this case as we only observe one-sided causal relationships between the exogenous variables and target time series.

The basic algorithm to find the best fit SARIMAX model is similar to the univariate case. Inorder to find a list of candidate models, we perform grid-search on a set of candidate model parameters. For this purpose, we use AIC, BIC and OOB as information criteria.

We individually perform grid-search for three different combinations of exogenous variables. First we consider the Unemployment Rate and Mortgage rate as individual exogenous variables. We find the best models for both cases using appropriate Information Criteria. Then we find a candidate model by considering both Unemployment Rate and Mortgage rate together as two exogenous variables. We then perform one-step cross validation for each of the three candidate models obtained. The results are summarized in the table below.

Method Used	Model Parameters	Endogenous Variable	Exogenous Variable	RMSE	MAPE
auto_arima	(2,0,0), (0,1,1,12)	MedianSoldPrice	UnemploymentRate	8927.396107	0.016699
auto_arima	(1,1,2)(1,0,0)[12]	MedianSoldPrice	MedianMortageRate	9274.288249	0.016436
auto_arima	(1,0,0)(0,1,2)[12]	MedianSoldPrice	UnemploymentRate, MedianMortageRate	9806.958925	0.017882
Same model for bic sarima					

We observe that the best cross-validation **RMSE (~8927.39)** is obtained when we only take Unemployment rate as the exogenous variable. The model which gives us the best cross-validation RMSE is **SARIMA(2,0,0)(0,1,1,12)**.

Best Model

Our best model is a multivariate model with one variable (UnemploymentRate) as exogenous variables. We got this model using auto arima on the following modelling parameters:

We choose this from all the above mentioned candidate models on the basis of lowest RMSE score. This model seems to have a comparatively lower MAPE score as well. From the results on the cross validation data it seems that both the variables are a significant predictor of the target median sold price of the houses.

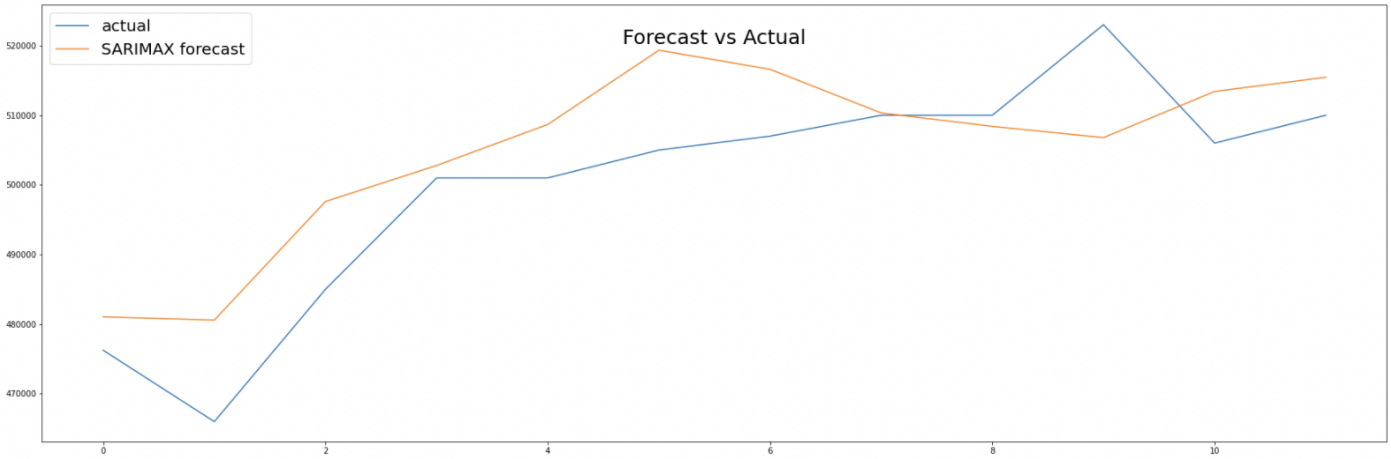
Method Used	Model Parameters	Endogenous Variable	Exogenous Variable	RMSE	MAPE
auto_arima	(2,0,0), (0,1,2,12)	MedianSold Price	UnemploymentRate	8927.39	0.016699

We will now test out our best model on the test data to measure our model's performance.

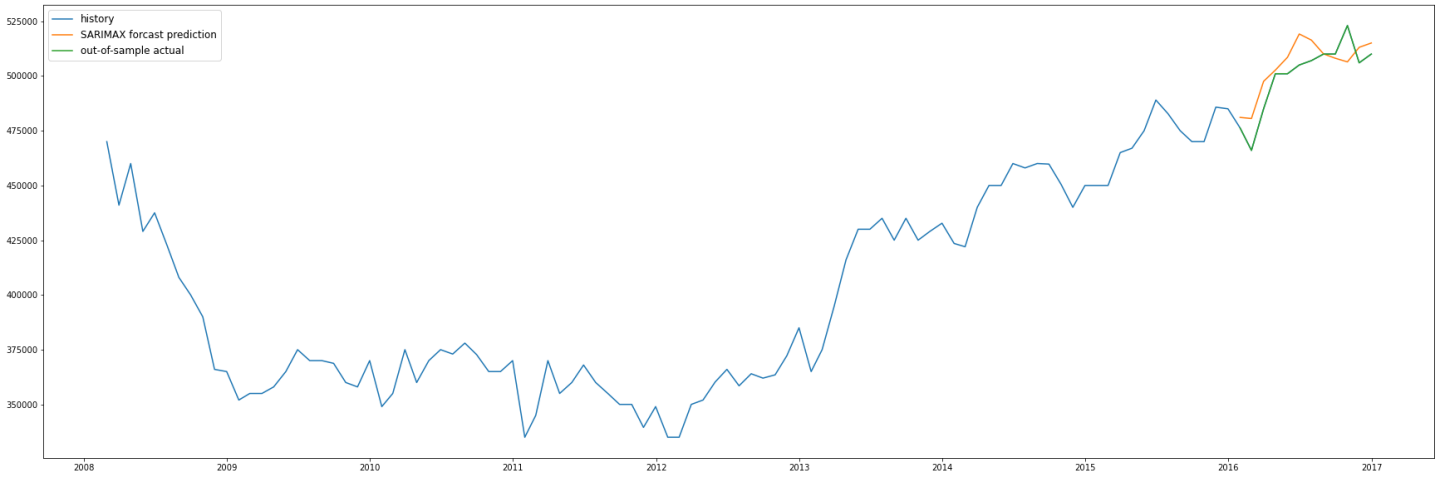
Prediction

The selected best model's performance on test data is mentioned below. It seems our model gives a fairly decent prediction of the test data as well with an RMSE score of 9596 and a low MAPE score of 0.016.

Method Used	Model Parameters	Endogenous Variable	Exogenous Variable	RMSE	MAPE
auto_arima	(2,0,0), (0,1,2,12)	MedianSold Price	UnemploymentRate	9596.3179	0.016



From the forecast plot we see that our predicted values are closer to the actual values. There seems to be some deviation at the end of the plot but for most of the plot, it seems to have forecasted a very close approximation of the actual value.



Conclusion

We attempt to predict the Median Housing Prices for the near future using the history data and exogenous variables. From our initial exploratory analysis, we find that the time series data shows an obvious trend. Some seasonality can be seen towards the later half of the data.

For our analysis, we used a variety of candidate time series forecasting models ranging from ARIMA family to prophet. For initial univariate analysis, we start with low complexity models by just considering simple ARIMA models with only trend components and no seasonality. We selected the

best candidate models with and without seasonality using different information criteria. Comparing the one-step cross validation RMSE of each of these candidate models, we found that ARIMA(3,1,0)(1,0,1)[12] is our best model for the univariate ARIMA family. Next, we performed a similar procedure for Exponential Smoothing models. We considered all the possible combinations for trend and seasonality and we found that the Exponential Smoothing model with Multiplicative trend and no seasonality gives the best cross validation RMSE. We did not get good cross-validation results for the Prophet model. This is expected since the Prophet model is not ideal for time series data with monthly frequency.

Next, we analysed SARIMA models with different combinations of exogenous variables. On comparing the cross-validation RMSE, we found that SARIMAX(2,0,0),(0,1,2)[12] with Unemployment rate as the exogenous variable gives the best results. Comparing the cross-validation RMSE values for our best models from different cases, we find that SARIMAX(2,0,0),(0,1,2)[12] has the lowest RMSE and hence we chose it as our best model. On the test data, we get a prediction RMSE value of 9596.31for our best model.

Further Improvements

For SARIMA models with exogenous variables, the exogenous variables for the test set are generally unavailable in a real world setting. Hence, ideally we need to consider the prediction of the exogenous variables for the test set as a separate univariate time series analysis problem. Once we have the test predictions for exogenous variables, we can use those predictions in the SARIMAX model to predict the target time series variable.

Also, the auto-arima function uses an information criterion to find the best candidate model parameters. However, auto-arima doesn't cover the entire parameter space and hence we may miss out on the best model. So, we can use a different model selection function like "bic_sarima" which covers the entire parameter space. Using the current version of bic_sarima is not feasible because it runs very slow.

For better forecasting predictions with exogenous variables, we can also explore deep learning models like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU).