DATS 6450: Time Series Analysis & Modeling

Instructor: Dr. Reza Jafari

Term Project Report

Arathi Nair AR – December 15, 2021

Table of Contents

ABSTRACT

This is a Time Series analysis for forecasting temperature based on Air Quality Data. The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city. Data were recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer.
The dataset is evaluated and discovered to be non-stationary and highly seasonal. Seasonal differencing procedures are performed to improve model fit. Base models show that the data set is a poor fit with RMSE scores. Holt Winters predictions do not yield significantly better results either. Multiple Linear Regression models show that Temperature is significant related to the Humidity, Benzene and Nitrogen Dioxide levels. But the model did not yield good predictions. ARMA and SARIMA models were found to be far better predictors of the data. An SARIMA model is finally used to predict values and make observations with confidence intervals provided.

INTRODUCTION

Time series analysis objective is to provide an insight for an ordered series of data by making predictions using different modelling methods. Data usually involves different patterns, cyclic and seasonal effects that need to be adjusted before conducting any time series analysis. Forecasting requires an in depth understanding of the data to deliver accurate predictions and drawing conclusions from past behavior. This report attempts to capture an overview of the processes involving in analyzing effect of air quality data on temperature over time and utilizes a variety of modeling options to achieve this objective.

Preprocessing
Prior to making predictions, the dataset is cleaned and processed to achieve the optimal result by eliminating any issues that may arises in the data. This done by conducting exploratory data analysis to handle missing values, resampling, and understanding how each feature correlates to each other. Finally, the dataset is split into training and testing sets for further analysis.

Stationarity
We will evaluate stationarity of the dataset with the use of rolling mean and rolling variance statistics to check the variability of the dependent variable through time. Differencing will be explored as an option if the data set is found to be non-stationary. An Augmented Dickey–Fuller test (ADF) test and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test are performed on the data to confirmed if the data does not have a unit root and is stationary.

Time Series Decomposition
To quantify the effect of trend and seasonality of the data, Seasonal and Trend decomposition using Loess (STL decomposition) is used to separate trend and seasonality from original data. The

strength of trend and seasonality is then calculated between the seasonally adjust and detrended data compared to the original data.

Feature Selection
Selecting important features that will improve model's performance is a critical procedure in time series analysis. Features that do not aid in providing enough information and that have high collinearity are removed from the dataset to avoid reduce performance of the model. Feature selection is performed with backward stepwise regressions until the best fit is found with the use of Adjusted R2, AIC and BIC. Multi-collinearity is detected by including SVD analysis and Condition Number.

Basic Models & Holt-Winters Model
Different methods like the Average method, the Naïve method, the Drift method, the SES method, and the Holt-Winters method are explored for forecasting. Base models are computationally inexpensive models that provide a baseline for comparing with more complex models. These models fail to capture the seasonality in the dataset, unlike the Holt-Winters method that incorporates trend and seasonality into the forecast and generally perform better than the base models.

Multiple Linear Regression
After performing feature selection, the model is checked for its fit and accuracy. We use an ordinary least square (OLS) multiple linear regressor to check the accuracy of the model. Hypothesis test using F-test and T-test is done to analyze the goodness of fit for the model.

ARMA, ARIMA, SARIMA Models
Unlike a multiple linear regression, ARMA models only require past values of the dependent variable to produce forecasts, allowing forecasts generated with a minimal number of parameters. ARIMA models are variants of the ARMA model that have been extended to accommodate trend in the data, and similarly SARIMA models accommodate both trend and seasonality. The best order for model will be determined using Generalized Partial Autocorrelation (GPAC) table and the ACF/PACF plots.

Levenberg Marquardt Algorithm
After determining the model order, significance of the model order will be checked by estimating the coefficients of the ARMA/ARIMA/SARIMA models using Levenberg Marquardt Algorithm (LM). The coefficients falling to fit between the confidence interval will be determined as insignificant.

Diagnostic Analysis
To determine the quality of the prediction and the goodness of fit of the model, diagnostic analyses will be reported for all utilized models. This includes determining if the roots of the coefficients are as simplified as they can be, using a zero-pole cancellation, performing a chi-squared test on the residuals to assess if they are capturing all information (white), and checking for bias in the prediction using the mean value of the residuals. All models will report RMSE values to evaluate the best model.

Final Model selection & Forecasting
Using the quantitative measurements taken in previous steps, the model that fits the data the best is determined and selected for further processing. The forecasting function is created using the coefficients from the final model. This forecasting function is then used to make a forecast about the future values of the testing set.

DATASET DESCRIPTION

The Air Quality dataset contains the responses of a gas multisensory device deployed on the field in an Italian city. Hourly responses averages are recorded along with gas concentrations references from a certified analyzer deployed on the polluted areas in an Italian city. The dataset includes hourly responses of the pollutant gas and additional features as listed below.

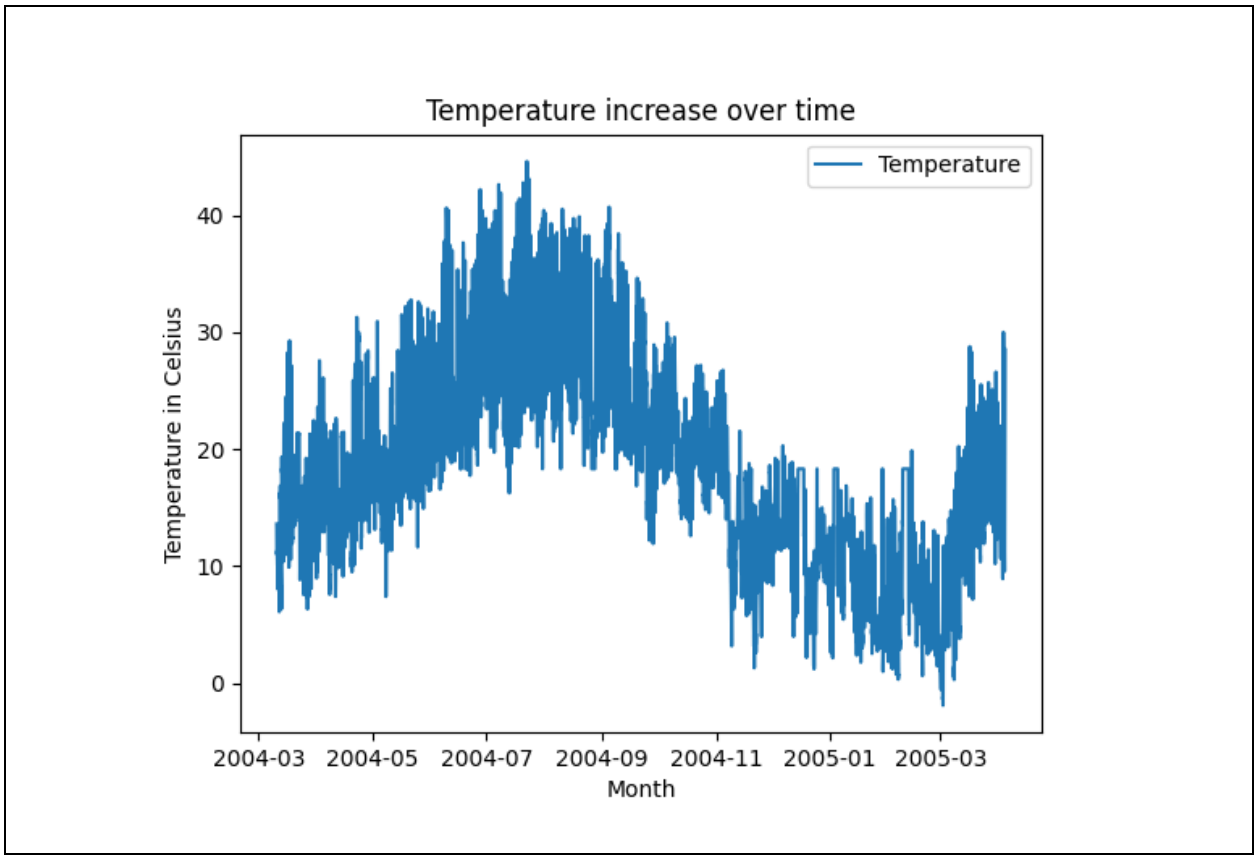
ATTRIBUTE	DESCRIPTION
DATE	Date (DD/MM/YYYY)
TIME	Time (HH.MM.SS)
CO(GT)	True hourly averaged concentration CO in mg/m^3
PT08.S1(CO)	PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
NMHC(GT)	True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3
C6H6(GT)	True hourly averaged Benzene concentration in microg/m^3
PT08.S2(NMHC)	PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
NOX(GT)	True hourly averaged NOx concentration in ppb (reference analyzer)
PT08.S3(NOx)	PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
NO2(GT)	True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
PT08.S4(NO2)	PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
PT08.S5(O3)	PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
T	Temperature in °C

RH	Relative Humidity (%)
AH	AH Absolute Humidity

a. Pre-processing Dataset

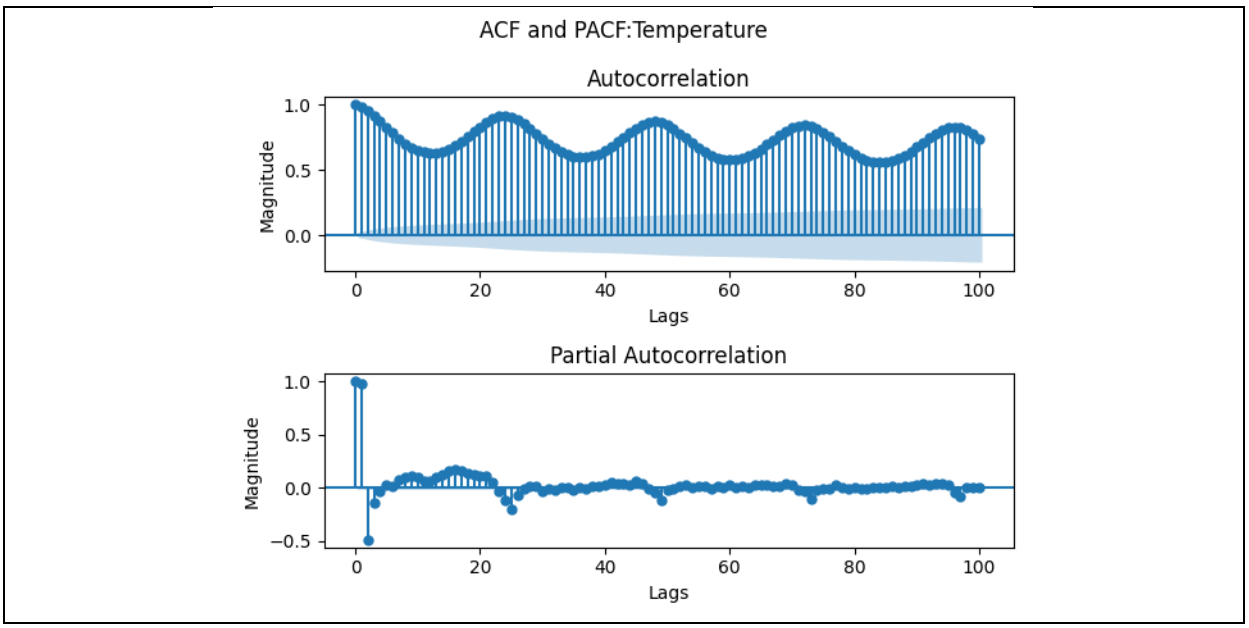
As the first step the dataset was cleaned i.e., removing unwanted data, or identifying null values. The missing values in the original dataset were tagged with -200 value and were replaced with mean values of the feature and the duplicates were removed.

b. Plot Dependent Variable Versus Time



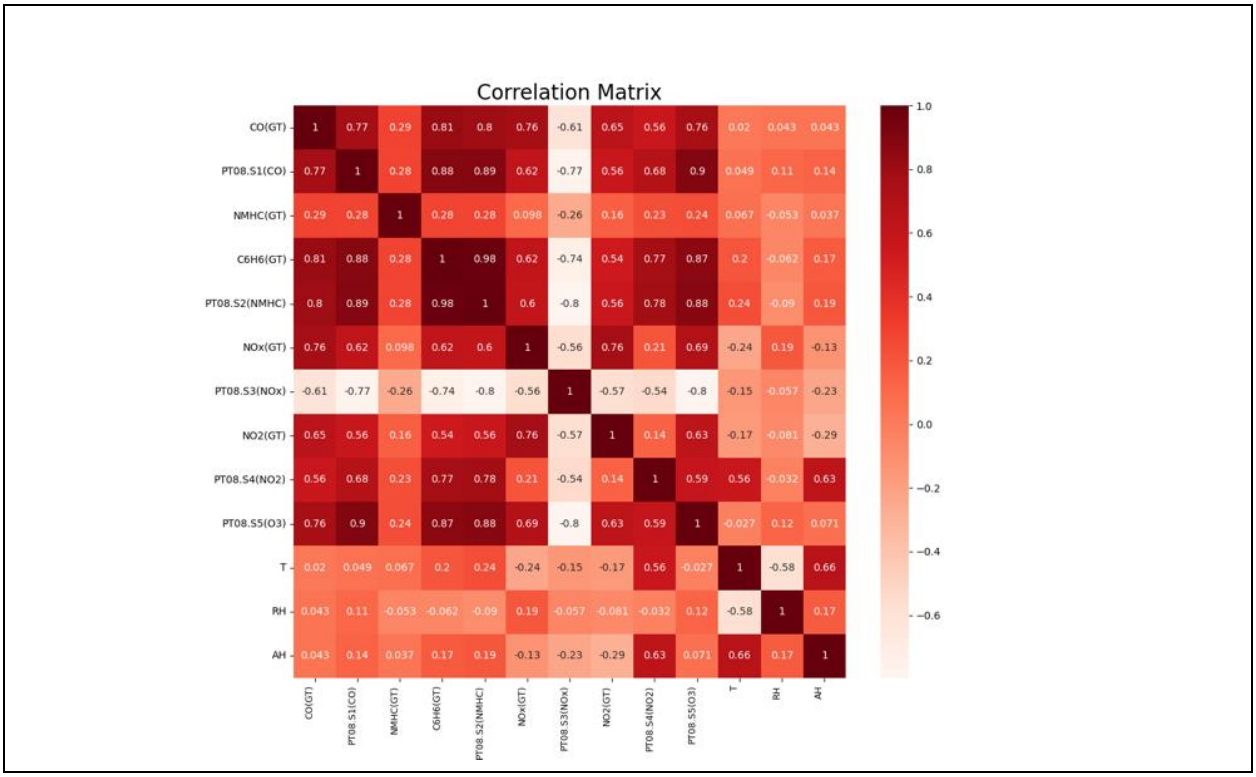
Observations: Above plot shows the temperature increase over time recorded from March 2004 to February 2005 (one year). The data appears to be highly seasonal, and we can notice a cyclic pattern. We also notice an increasing or decreasing trend overtime.

c. Plot of Autocorrelation Function and Partial Autocorrelation Function



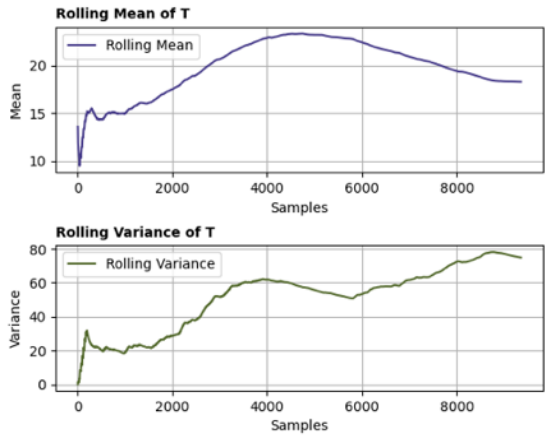
Observations: The Autocorrelation plot shows significant lags. The lags stay significant throughout the plot and do not trail off. The Partial autocorrelation plot shows three major lags after the initial spike and then mildly significant lags at period of 24.

d. Correlation Matrix

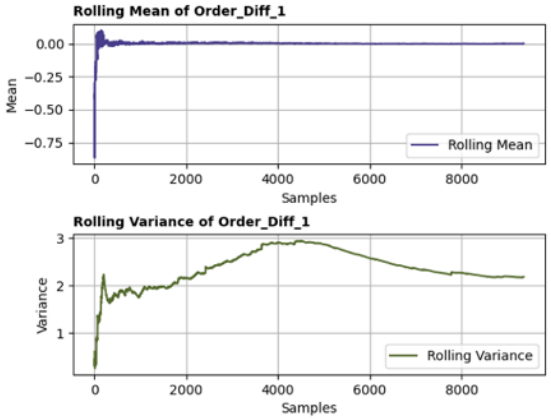


Observations: The Correlation matrix shows that the pollutant gas variables are strongly correlated with each other except with NMHC(GT). However, there are not strongly correlated with the dependent variable Temperature. Absolute Humidity (Pearson c = 0.66), Relative humidity (Pearson c = -0.58) and pollutant gas PT08.S2(NO2) (Pearson c = 0.56) show a moderate correlation with Temperature.

STATIONARITY




```
ADF Statistic for T: -3.621766
p-value: 0.005358
Critical Values:
    1%: -3.431
    5%: -2.862
    10%: -2.567
Results of KPSS Test for T:
Test Statistic            7.205413
p-value                   0.010000
Lags Used                55.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
```

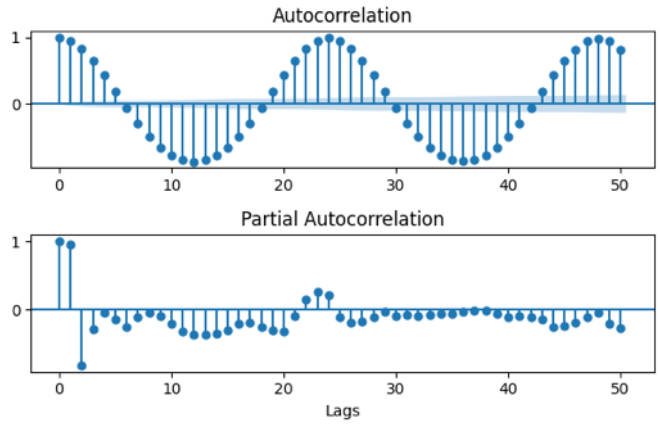



```
ADF Statistic for Order_Diff_1: -18.451598
p-value: 0.000000
Critical Values:
    1%: -3.431
    5%: -2.862
    10%: -2.567
Results of KPSS Test for Order_Diff_1:
Test Statistic           0.002747
p-value                  0.100000
Lags Used                9.000000
Critical Value (10%)     0.347000
Critical Value (5%)      0.463000
Critical Value (2.5%)    0.574000
Critical Value (1%)      0.739000
```

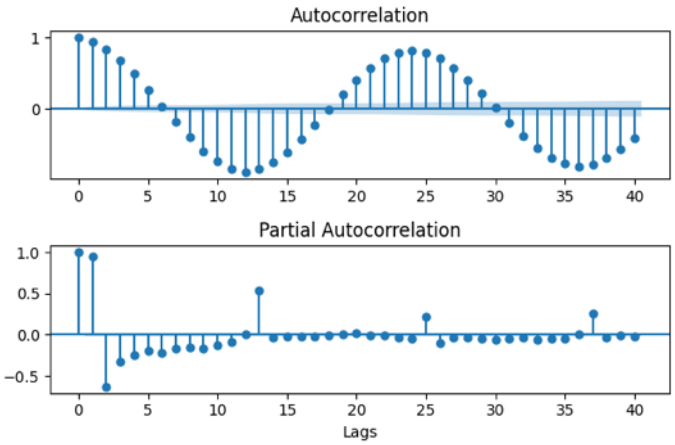
Observations: The original dataset is nonstationary. The calculation of the rolling mean and variance show that there is a steady increase in trend. An Augmented Dickey Fuller Test (ADF) also shows that the null hypothesis can be rejected with a significant p-value of $p<0.01$. However, the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test shows that the p-value is significant with $p<0.01$, rejecting the null hypothesis and suggests that the dataset is nonstationary. To make the data stationary, first order differencing was performed on the dataset and the ADF test shows the p-value is significant while the KPSS test also shows the p-values is not significant suggesting that the dataset has been stationarized. However, an ACF and PCF analysis suggests that the data still

might not be stationary. The ACF and PCF plots below shows that the sample has strong autocorrelations.

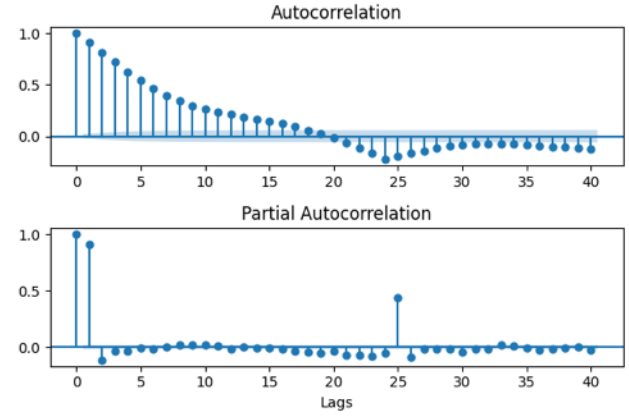
First Order Differencing

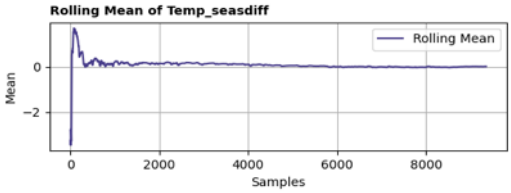


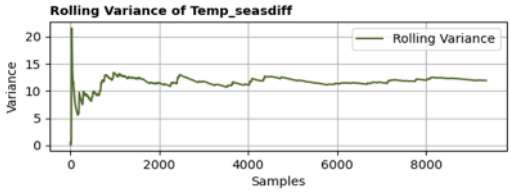
Seasonal Differencing 12 hours



Seasonal Differencing 24 hours



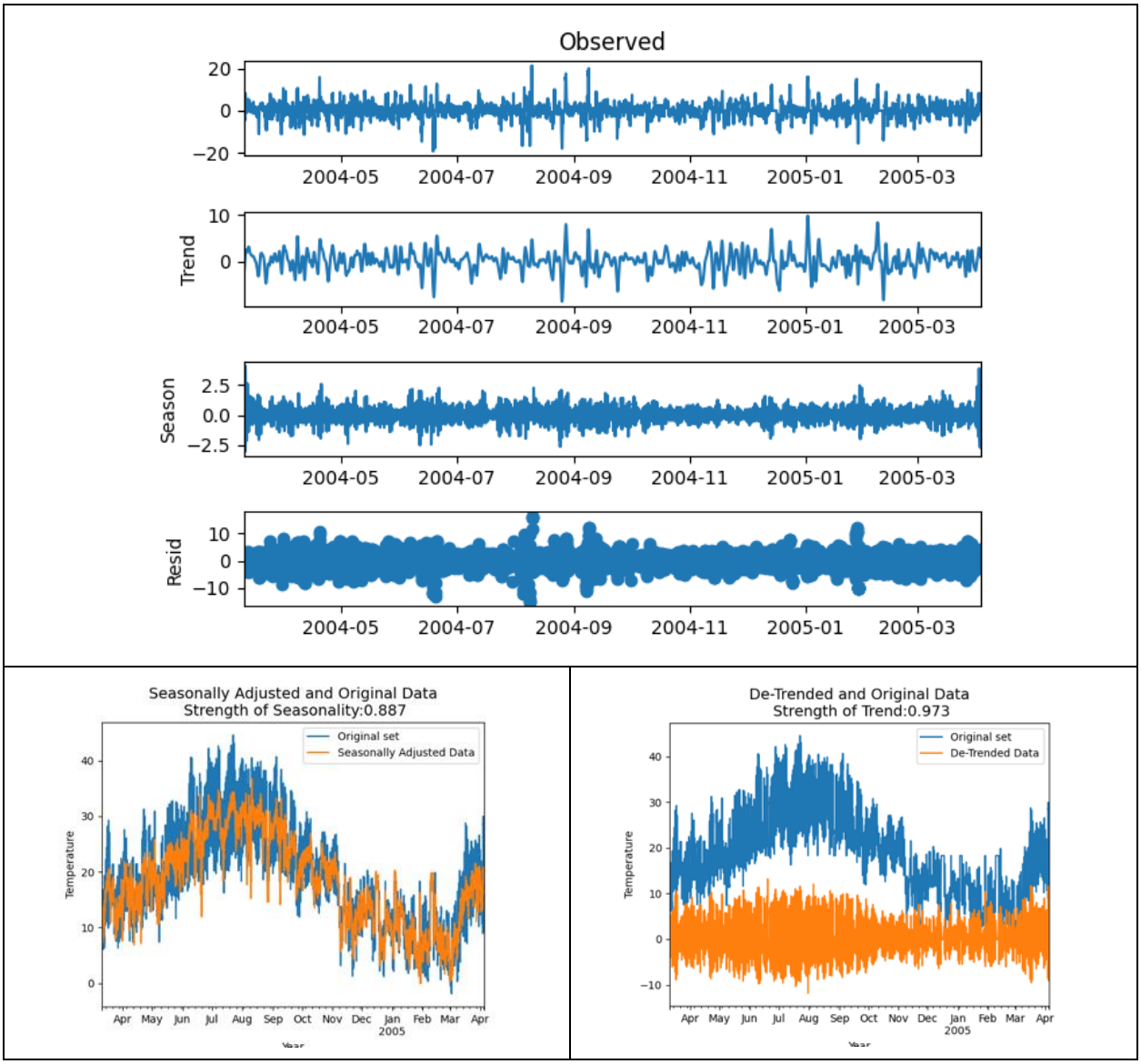





```
ADF Statistic for Temp_seasdiff: -15.993908
p-value: 0.000000
Critical Values:
    1%: -3.431
    5%: -2.862
    10%: -2.567
Results of KPSS Test for Temp_seasdiff:
Test Statistic            0.048617
p-value                   0.100000
Lags Used                51.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
dtype: float64
```

Observations: The ACF & PACF plots for the first-order differencing, shows a sinusoidal ACF plot indicative of a large degree of seasonality in the data. To make the data stationary, a seasonal differencing was performed. To select the interval to difference, several seasons were tested and visualized with an ACF/PACF plot. The 12-hour differencing did not reduce the oscillatory seasonality in the autocorrelation and failed to make the data stationary. A 24-hour seasonal differencing was then tested considering Temperature reading for each day. This seasonal differencing led to an ACF that trails off and remains non-significant without oscillations, indicating a removal of seasonality from the data. The PACF cuts off after 2 significant lags, with another significant at lag=24. The rolling mean and rolling variance show the trend and seasonality is consistent throughout and the ADF & KPSS tests also suggests that the data has been completely stationarized.

TIME SERIES DECOMPOSITION

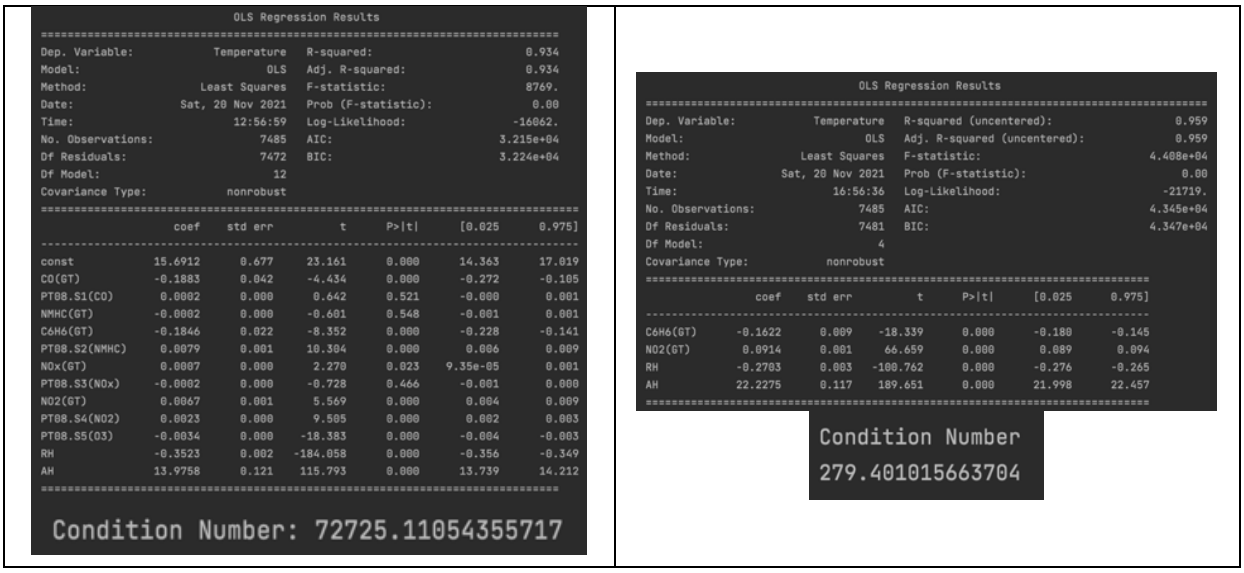


Observations: The STL decomposition results show that the data is stationary. Over the course of the data, the trend line remains at a consistent level, with fluctuations that always remain in the same span. The seasonality is likewise consistent, with no shifting trend or variability. The comparison between the original data and seasonally adjusted data shows the strength of seasonality is extremely high with 0.8 on scale. The detrend and original data also shows the strength of trend is also high with 0.9 on scale.

FEATURE SELECTION

To build a model that can predict accurately, we require to select important features from the dataset. We first performed an SVD analysis and examined the condition number of the full dataset. We then perform backward stepwise regression to eliminate features with those with the largest p-values and that increased the adjusted R squared value. A final SVD and Condition Number analysis was performed for this reduced dataset.

CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOX(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	RH	AH	AIC	BIC	Adj R^2
1	1	1	1	1	1	1	1	1	1	1	1	3.215e+04	3.224e+04	0.934
1	1	0	1	1	1	1	1	1	1	1	1	3.215e+04	3.223e+04	0.934
1	0	0	1	1	1	1	1	1	1	1	1	3.215e+04	3.222e+04	0.934
1	0	0	1	1	1	0	1	1	1	1	1	3.215e+04	3.221e+04	0.934
1	0	0	1	1	0	0	1	1	1	1	1	3.215e+04	3.221e+04	0.934
0	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**4.345e+04**	**4.347e+04**	**0.959**



OLS Regression Results

Dep. Variable:	Temperature	R-squared:	0.934
Model:	OLS	Adj. R-squared:	0.934
Method:	Least Squares	F-statistic:	8769.
Date:	Sat, 20 Nov 2021	Prob (F-statistic):	0.00
Time:	12:56:59	Log-Likelihood:	-16062.
No. Observations:	7485	AIC:	3.215e+04
Df Residuals:	7472	BIC:	3.224e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	15.6912	0.677	23.161	0.000	14.363	17.019
CO(GT)	-0.1883	0.042	-4.434	0.000	-0.272	-0.105
PT08.S1(CO)	0.0002	0.000	0.642	0.521	-0.000	0.001
NMHC(GT)	-0.0002	0.000	-0.601	0.548	-0.001	0.001
C6H6(GT)	-0.1846	0.022	-8.352	0.000	-0.228	-0.141
PT08.S2(NMHC)	0.0079	0.001	10.304	0.000	0.006	0.009
NOx(GT)	0.0007	0.000	2.270	0.023	9.35e-05	0.001
PT08.S3(NOx)	-0.0002	0.000	-0.728	0.466	-0.001	0.000
NO2(GT)	0.0067	0.001	5.569	0.000	0.004	0.009
PT08.S4(NO2)	0.0023	0.000	9.505	0.000	0.002	0.003
PT08.S5(O3)	-0.0034	0.000	-18.383	0.000	-0.004	-0.003
RH	-0.3523	0.002	-184.058	0.000	-0.356	-0.349
AH	13.9758	0.121	115.793	0.000	13.739	14.212

Condition Number: 72725.11054355717

OLS Regression Results

Dep. Variable:	Temperature	R-squared (uncentered):	0.959
Model:	OLS	Adj. R-squared (uncentered):	0.959
Method:	Least Squares	F-statistic:	4.408e+04
Date:	Sat, 20 Nov 2021	Prob (F-statistic):	0.00
Time:	16:56:36	Log-Likelihood:	-21719.
No. Observations:	7485	AIC:	4.345e+04
Df Residuals:	7481	BIC:	4.347e+04
Df Model:	4		
Covariance Type:	nonrobust		

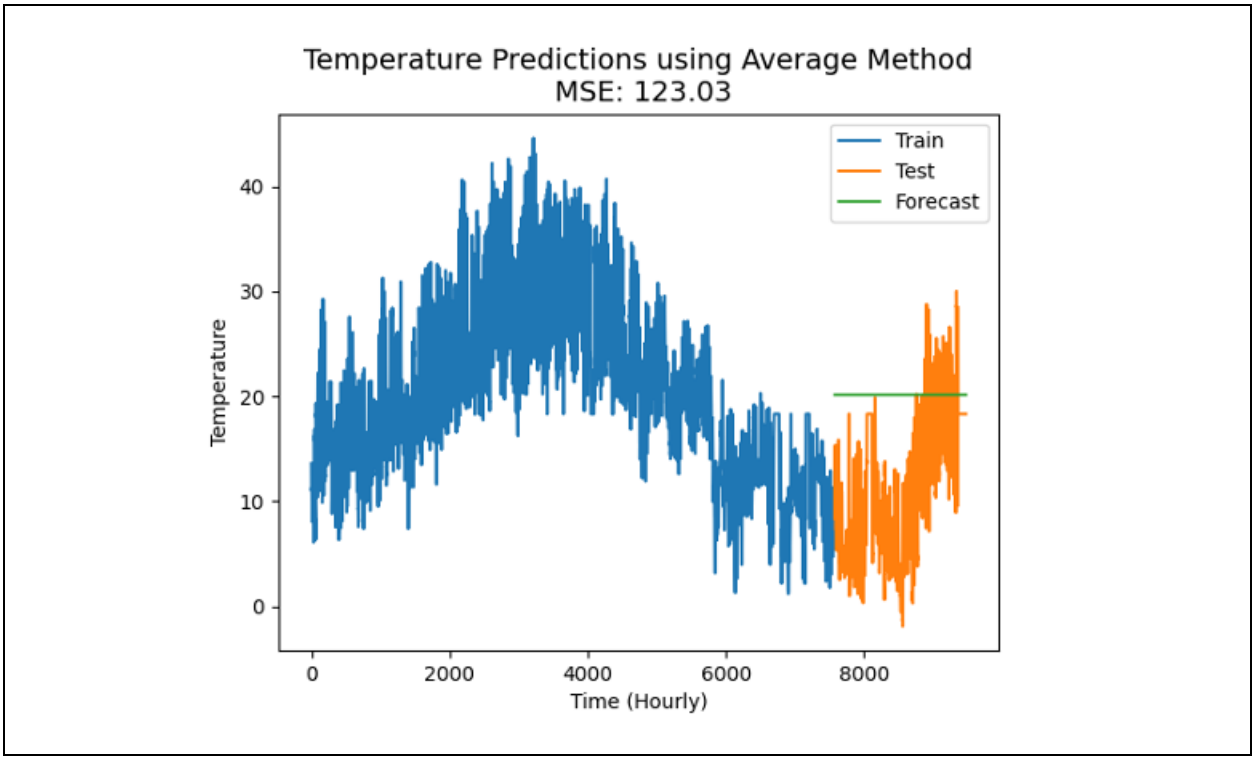
	coef	std err	t	P>|t|	[0.025	0.975]
C6H6(GT)	-0.1622	0.009	-18.339	0.000	-0.180	-0.145
NO2(GT)	0.0914	0.001	66.659	0.000	0.089	0.094
RH	-0.2703	0.003	-100.762	0.000	-0.276	-0.265
AH	22.2275	0.117	189.651	0.000	21.998	22.457

Condition Number
279.401015663704

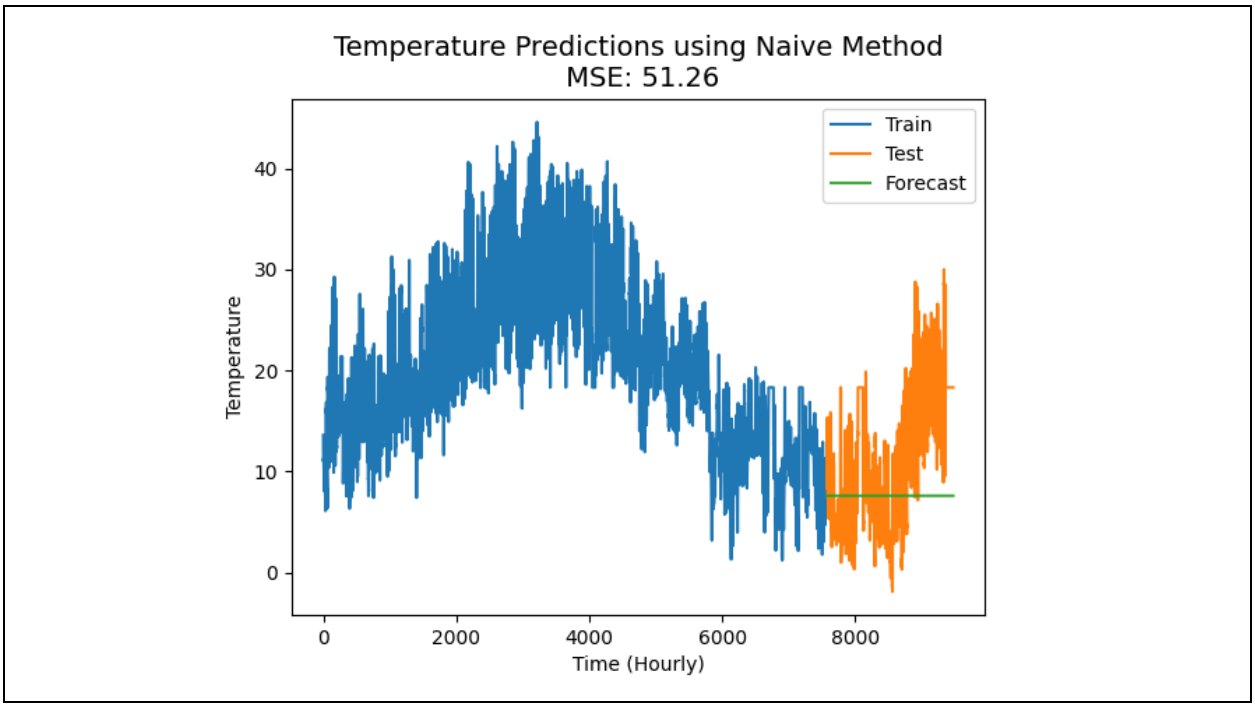
Observation: In comparison to the condition number (72725) of the original model including all the features, we see a significant drop in the condition number (279) after performing feature selection of the model. Also, the selected features containing C6H6(GT), NO2(GT), RH and AH yielded an improved adjusted R^2 at 0.95. This suggests that only 5% of the variance in the model remains unexplained.

BASIC MODELS

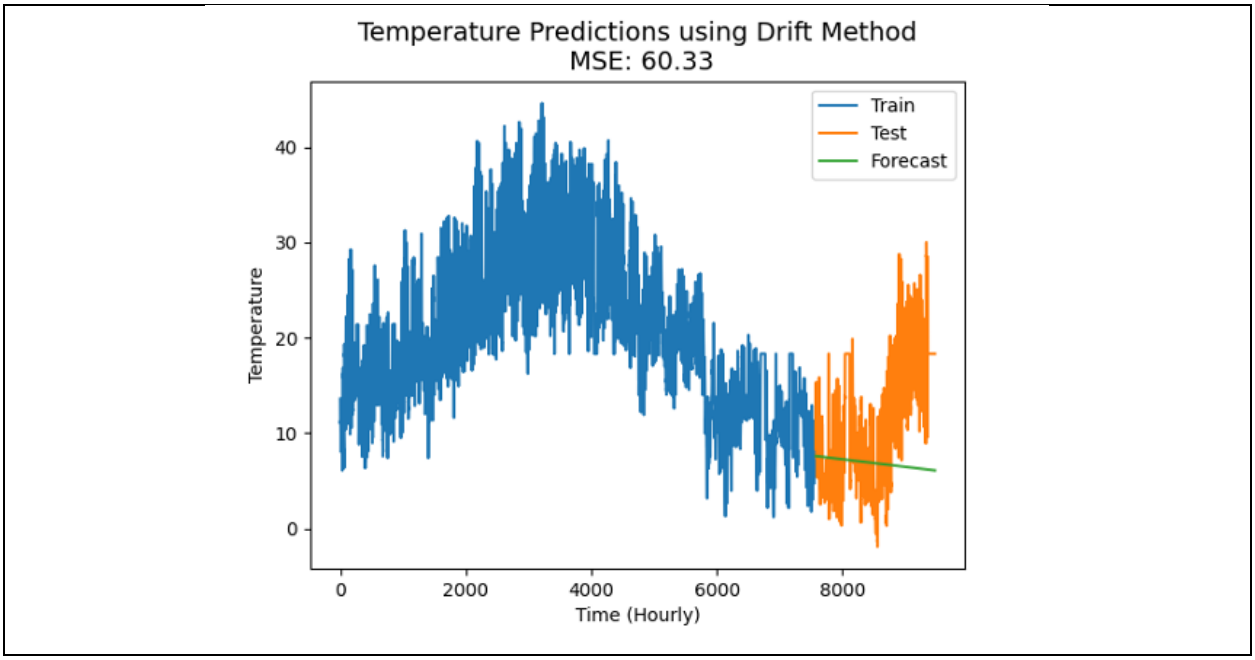
For each of the basic models we have done one-step ahead forecasting of the testing data, which was plotted for comparison.



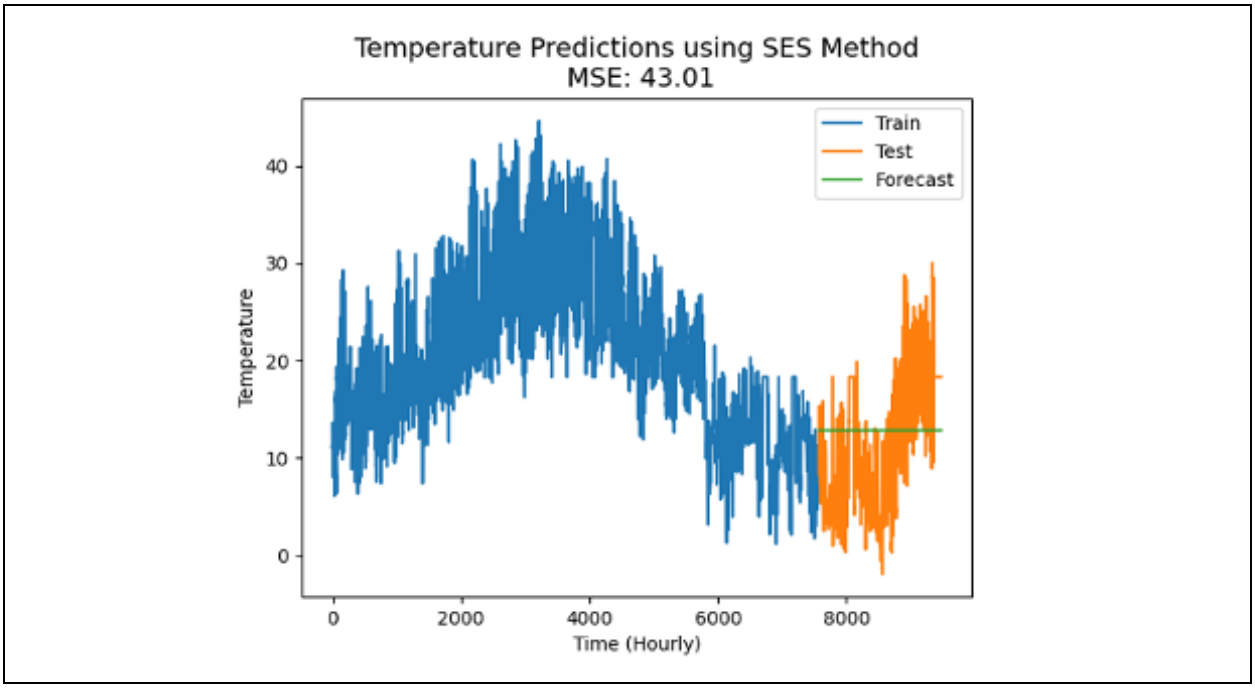
Observation: The Average basic model gives all datapoints equal weight, and in essence takes an average of all the datapoints prior to the one you are predicting. As seen above, all forecasted values are the same, and are equal to the mean value of the training set. The MSE of the forecast errors is 123 and the Q-score calculated on the residuals is 96722.34, indicating that there is a large amount of information not captured by the model.



Observation: The Naïve basic model applies zero weight to any datapoint beyond the most recent datapoint, which in the case of the testing set is the final datapoint of the training set. As with the average model, this single value is applied to all forecasted values. The MSE of the forecast errors is 51.26. The predictive power of this model is slightly better than the previous model. The Q-score calculated on the residuals is 10942.05, indicating that there is a large amount of information not captured by the model.



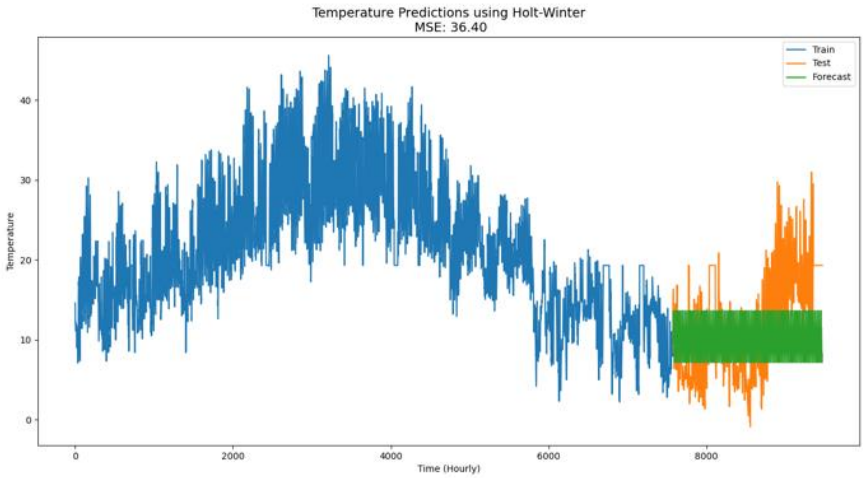
Observation: The Drift method is slightly more sophisticated than the average of the naïve method. It operates by applying weight to the first and last points, and extrapolating a slope from them, then extrapolating all forecasted values onto that slope. While the fit looks slightly better, in this case the MSE of the forecast errors is 60.33 which is slightly worse than the Naïve method. The Q-score calculated on the residuals is 10939.88, indicating that there is a large amount of information not captured by the model.



Observation: The Simple Exponential Smoothing (SES) method works as a compromise between Average and Naïve by putting a large amount of weight on the most recent point, but still applying a steadily decreasing weight to historical data. Despite the increase in complexity of the model, the SES method did not predict the data well, possibly due to a fluctuation in the data near the end of the training set. The MSE of the forecast errors is 43.01 and the calculated Q-score of 29685.38 indicates that there is a large amount of information not captured by the model.

HOLT WINTER

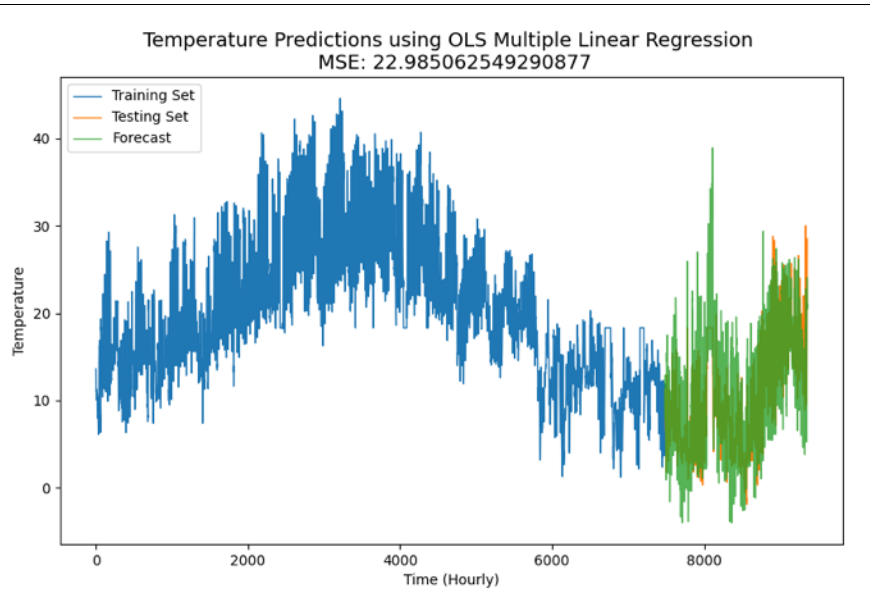
The Holt-Winters' seasonal method is an extension of the original Holt method which aimed to capture seasonality. It comprises of a forecast equation and three smoothing equations. An additive variation of this method is utilized and suggested when constant variations are present. In this dataset, since seasonal variations appear to be high, we utilize the multiplicative method instead of an additive method.



Observations: The Holt Winter provides a base for trying to find the best fit with the training set while accounting for seasonality. When the Holt-Winter method is plotted against the test set for a prediction, we notice that the seasonal components are not fully accounted for by the model. The Mean Square Error (MSE) is 36.40 and the Q-score is 548.22. Compared to the base models, Holt-Winter has the better predicting power for the dataset

MULTIPLE LINEAR REGRESSION

The OLS multiple linear regression is a linear regressor that takes multiple features as input, in addition to the dependent predictor variable. For this analysis, we use not only the temperature to make our predictions, as with the previous models, but also the features we selected during the feature selection step.

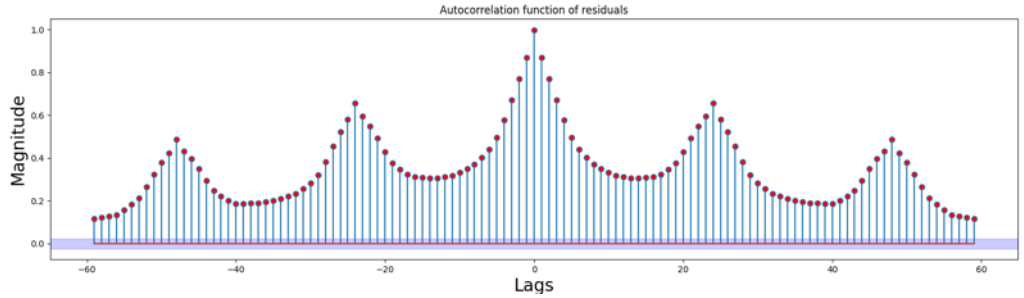


OLS Regression Results

Dep. Variable:	Temperature	R-squared (uncentered):	0.959
Model:	OLS	Adj. R-squared (uncentered):	0.959
Method:	Least Squares	F-statistic:	4.431e+04
Date:	Sun, 05 Dec 2021	Prob (F-statistic):	0.00
Time:	09:27:26	Log-Likelihood:	-21968.
No. Observations:	7576	AIC:	4.394e+04
Df Residuals:	7572	BIC:	4.397e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
C6H6(GT)	-0.1628	0.009	-18.563	0.000	-0.180	-0.146
NO2(GT)	0.0903	0.001	66.767	0.000	0.088	0.093
RH	-0.2704	0.003	-101.784	0.000	-0.276	-0.265
AH	22.3269	0.115	193.797	0.000	22.101	22.553

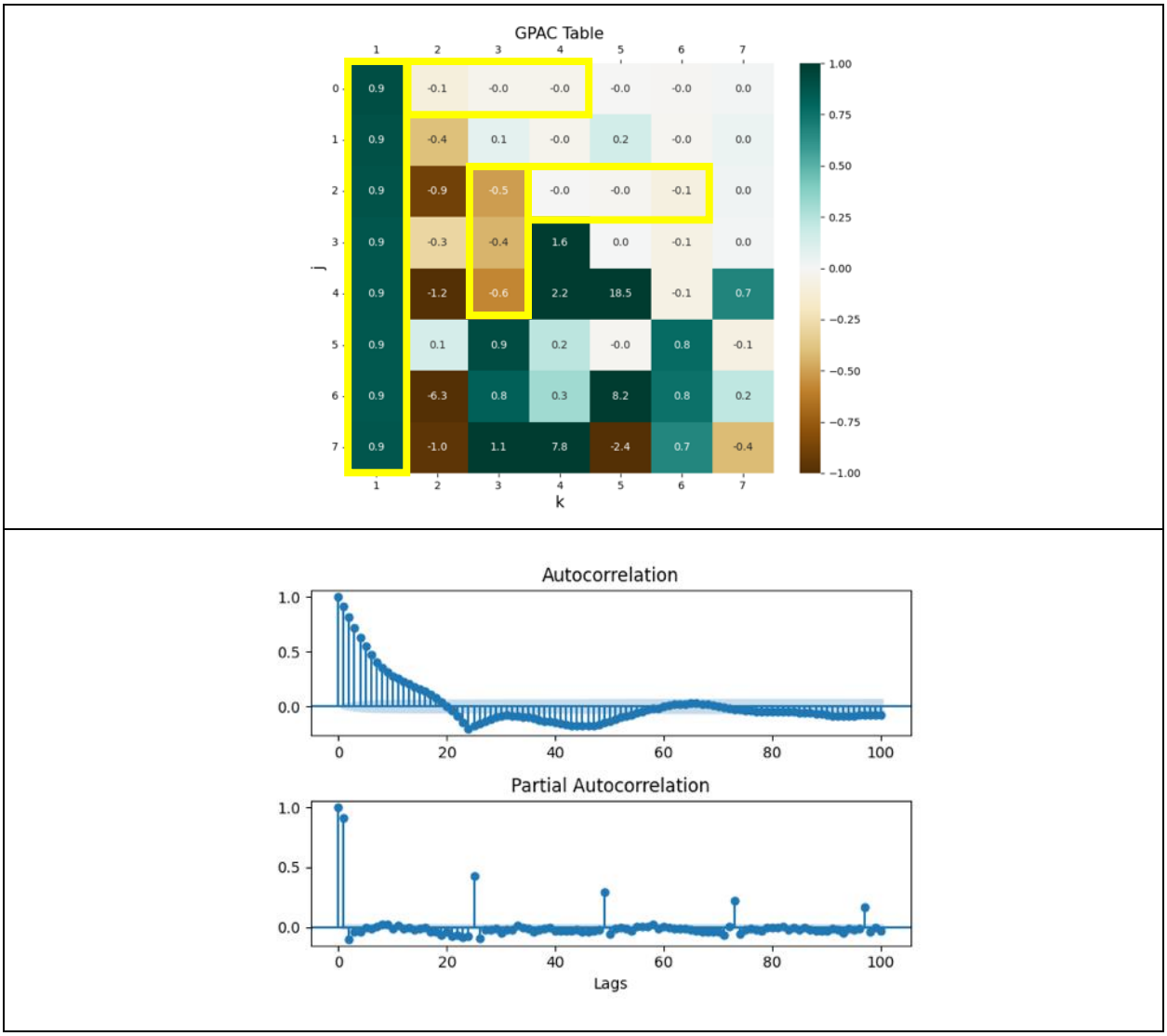
Omnibus:	248.301	Durbin-Watson:	0.254
Prob(Omnibus):	0.000	Jarque-Bera (JB):	395.865
Skew:	0.305	Prob(JB):	1.09e-86
Kurtosis:	3.939	Cond. No.	279.




```
The Residual errors are NOT RANDOM for this model

The Critical value:  7855.107012083004
The Q value:  [16169.37832241]
The P value:  [0.]
The variance of the residuals are:  17
The mean of the residuals are:  -1
```

Observations: The r-squared and adjusted r-squared are both high at 95%, which indicates a good overall fit. As the r-squared and adjusted r-squared are the same value, we can conclude that the number of features we retained is appropriate. All t-test values are significant and satisfy the criteria of $p < 0.05$. F-test value is 4.431 e+04 which is also significant with a $p < 0.01$. The MSE of 22 is higher than the models tested so far, and we can conclude stepwise regression model is a better fit. Analyzing the ACF plot shows that the model is not well explained. The residuals are still correlated suggesting that other factors are at play beyond the features that predict this model and cannot be sufficiently explained by a regression model. The Q value is sufficient to state that the residual errors are not random white noise.



Observations: From the above GPAC table we can identify the likely ARMA model orders from the patterns present. As shown with the yellow outlines, there are two likely ARMA models representing our data. The first model of order appears at j=0, k=1 representing an ARMA (1,0) model. The next model order appears at j=3, k=2, representing an ARMA (3,2) model.
The ACF and PACF plot, we can see that the PACF plot cuts off at lag=2 and there is a small, significant lag at lag = 24, which may be an indicator of some remaining seasonality in the data. To address the seasonal component of the data, a SARIMA (2,0,0) (2,0,0)24 model was also fit.

LEVENBERG MARQUARDT ALGORITHM

ARMA(1,0)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0041	0.184	-0.022	0.982	-0.364	0.356
ar.L1	0.9112	0.003	313.486	0.000	0.905	0.917
sigma2	2.0144	0.010	194.609	0.000	1.994	2.035

ARMA(3,2)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0082	0.153	-0.053	0.958	-0.309	0.292
ar.L1	0.5443	0.071	7.634	0.000	0.405	0.684
ar.L2	0.8801	0.005	175.749	0.000	0.870	0.890
ar.L3	-0.5253	0.061	-8.583	0.000	-0.645	-0.405
ma.L1	0.4811	0.074	6.501	0.000	0.336	0.626
ma.L2	-0.5188	0.074	-7.024	0.000	-0.664	-0.374
sigma2	1.9446	0.012	159.399	0.000	1.921	1.969

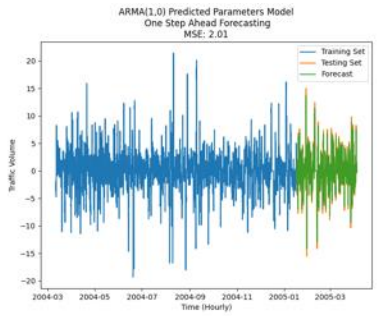
SARIMA(2,0,0)(2,0,0)24

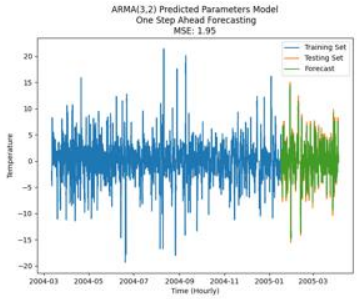
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0047	0.107	-0.044	0.965	-0.215	0.206
ar.L1	1.0609	0.008	135.298	0.000	1.046	1.076
ar.L2	-0.1301	0.008	-16.451	0.000	-0.146	-0.115
ar.S.L24	-0.6019	0.005	-111.320	0.000	-0.612	-0.591
ar.S.L48	-0.3002	0.006	-53.102	0.000	-0.311	-0.289

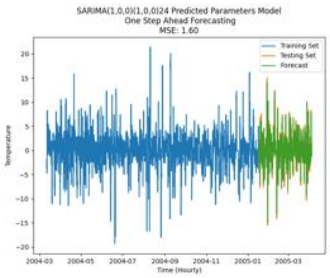
Observations: All three of our models were fit using a LM algorithm, and the coefficients of the parameters were obtained. The confidence interval for each coefficient was analyzed and checked for significance by determining if the interval spanned across 0. For the ARMA (1,0), ARMA (3,2) and SARIMA (1,0,0) (1,0,0)24, all coefficients were significant.

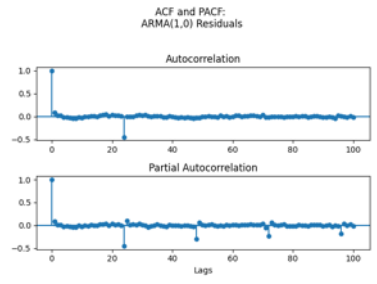
DIAGNOSTIC ANALYSIS

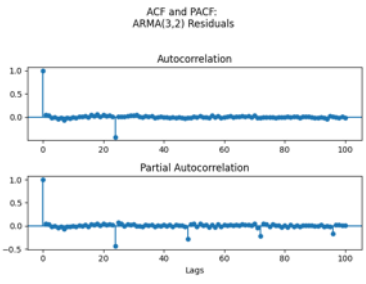
Model	Q-Score	Q-Critical	Result	Mean of the Residuals	Biased/ Unbiased	Variance of error	Variance of residual errors	Variance of forecast error
ARMA(1,0)	1693.91	41.63	Not Random	0.00	No	1.817	2.016	22.157
ARMA(3,2)	1542.27	36.19	Not Random	0.00	No	1262.178	1.948	22.228
SARIMA(2,0,0)(2,0,0)24	143.94	40.28	Not Random	0.00	No	2.475	1.438	22.762

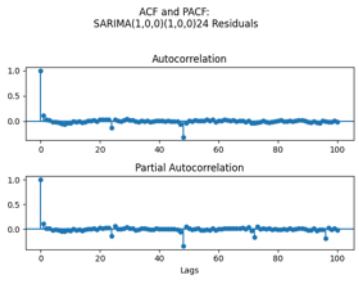


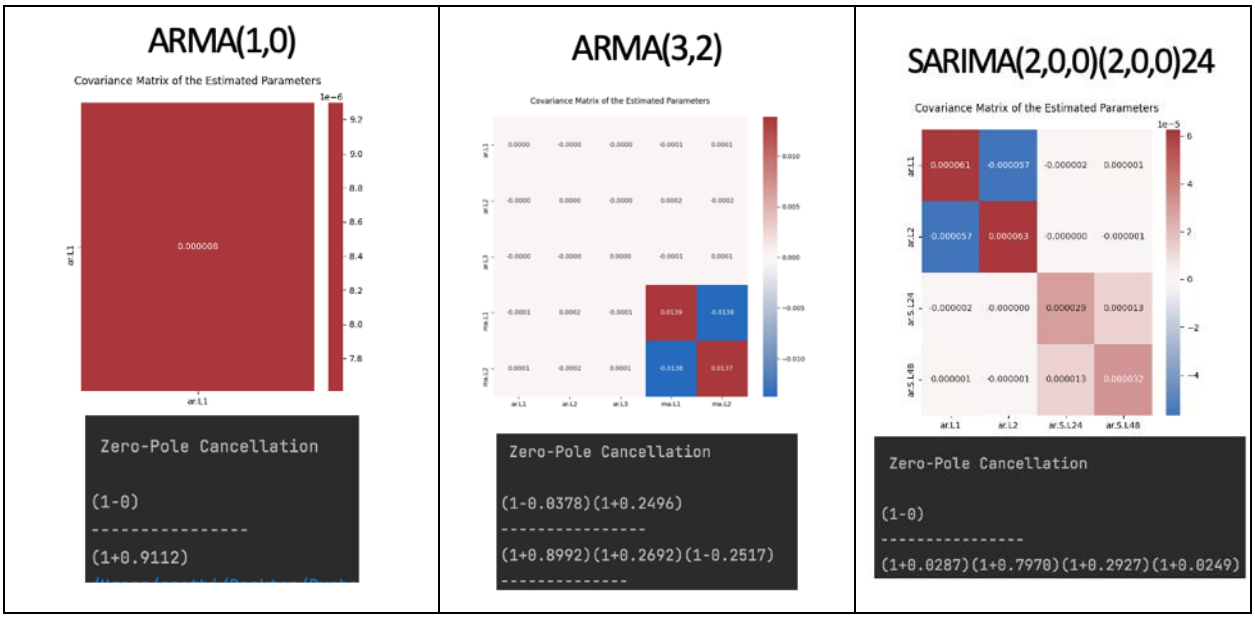










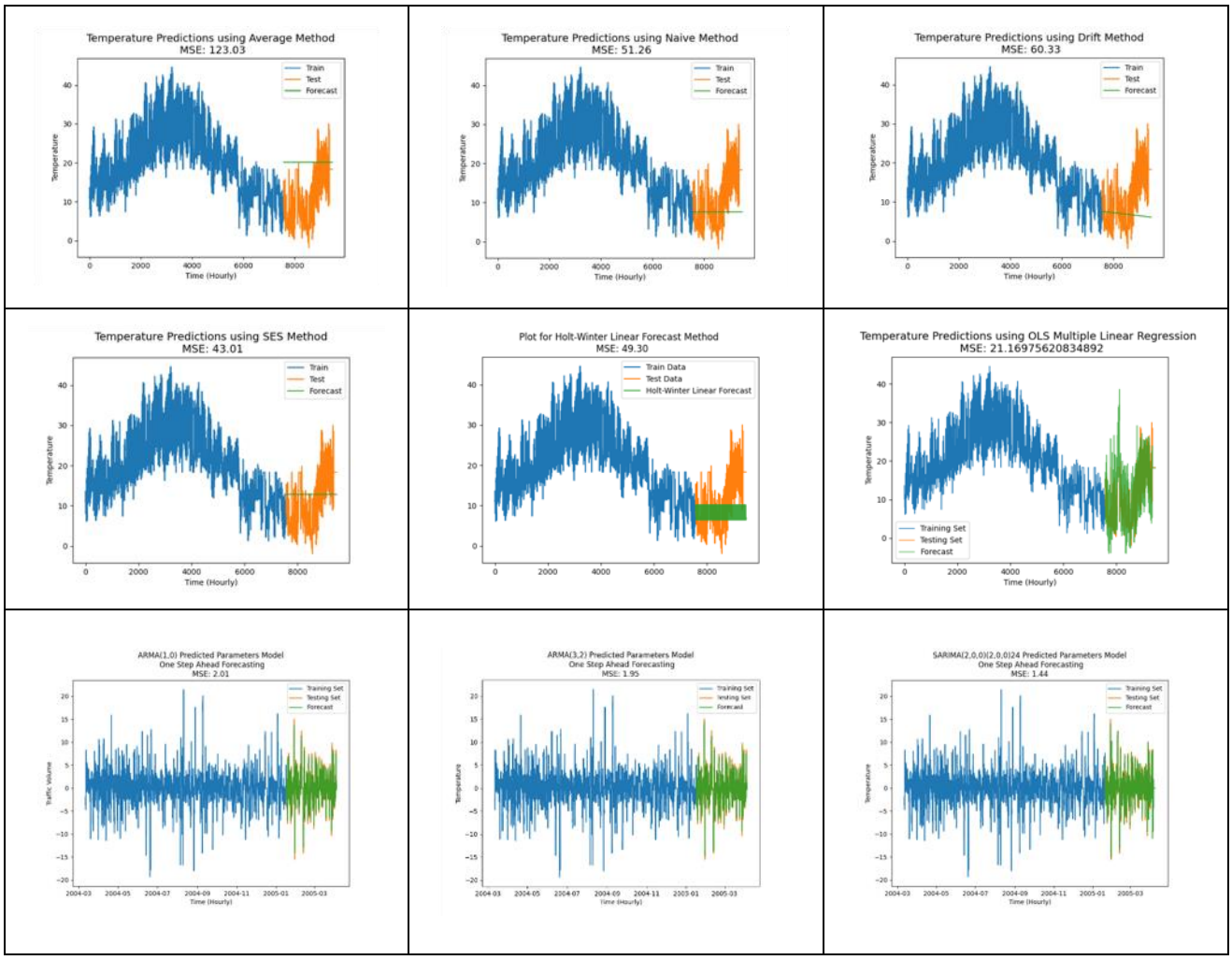


Observation: As shown in the above table, all three models were checked for best fit to proceed as the final model by performing diagnostic analysis. We calculate the Q score of the residuals to determine if they were white noise, estimate variance of the error and covariance of the parameters, variance of residual errors and variance of forecasted errors and determine if the selected model is unbiased.

From the above analysis, SARAMIA has the lowest Q score and variance of residual errors while ARMA (1,0) has the lowest variance of the error and variance of the forecasted errors. All three models have zero mean of residuals indicating the models are unbiased.

All three models are not completely white indicating some of the information is still not captured by the models. However, when we compare the ACF and PACF plots the magnitude of the data uncovered is very small. The one step ahead prediction plots show SARIMA model has the lowest MSE score of 1.60 suggesting the model could be a better fit to make future predictions.

FINAL MODEL SELECTION



Model	Average	Naive	Drift	SES	Holt-Winter	OLS	ARMA (1,0)	ARMA (3,2)	SARIMA (2,0,0) (2,0,0)24
RSME	11.09	7.16	7.77	6.56	7.02	4.6	1.41	1.39	1.19

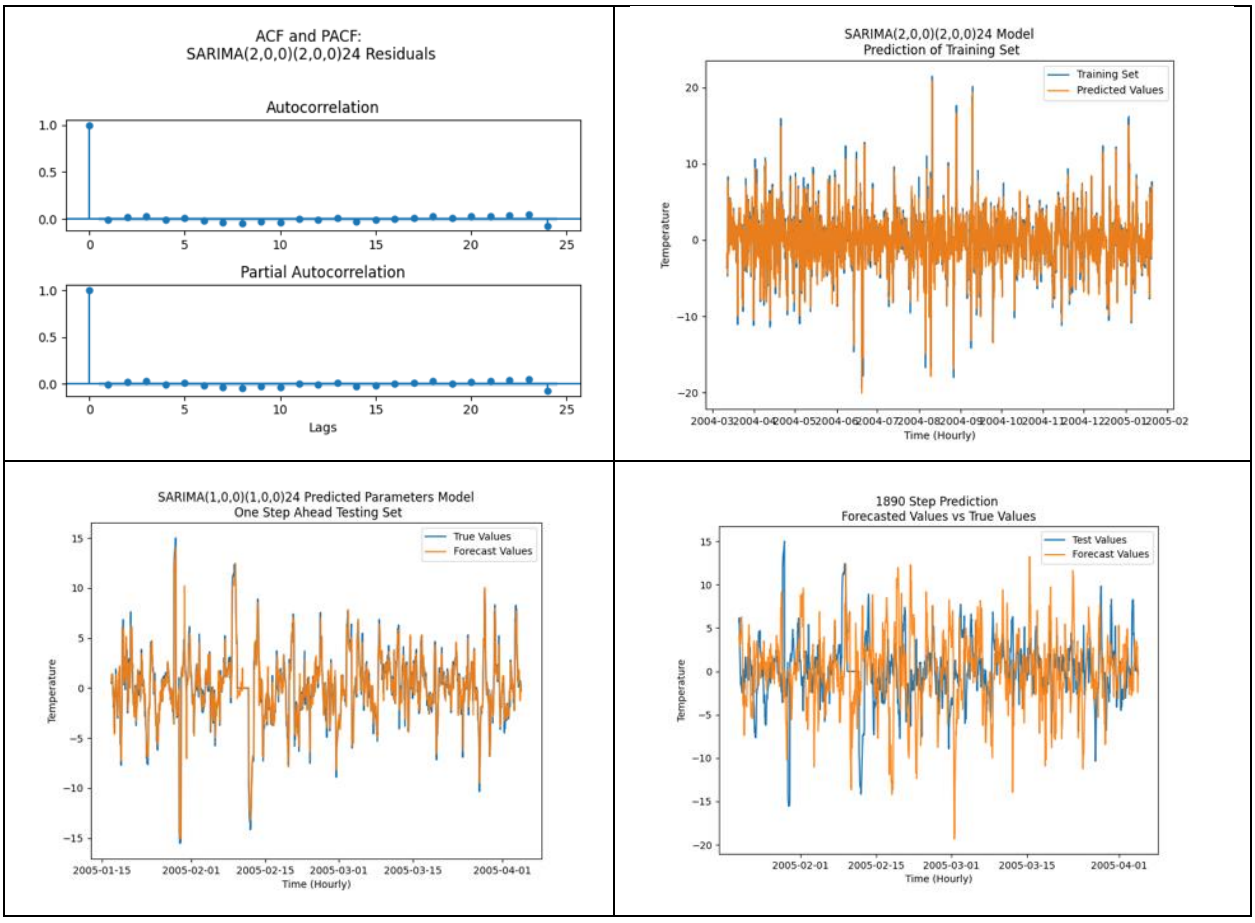
Observation: Out of the basic models, SES had the lowest MSE & RSME scores of the forecasting, indicating a good fit. Out of the moderately complex model, OLS model performed the best with MSE (21.16) and RSME (4.6), followed by the Holt-Winter mode with MSE of 49.5 and RSME (7.02). Out of the ARMA models, SARIMA (2,0,0) (2,0,0)24 had the lowest MSE (1.44) and RSME (1.19) and was selected as the final model to perform forecasting on the test set.

FORECAST FUNCTION

The forecast function for the SARIMA (2,0,0) (2,0,0)24 model can be written as:

$$\mathbf{y(t+h) = -1.0609 * y(t+h-1) + 0.1301 * y(t+h-2) + 0.6019 * y(t+h-24) + 0.3002 * y(t+h-48) + e(t)}$$

H-STEP AHEAD PREDICTIONS



Observation: After performing forecast function on the test set, we observe the SARIMA (2,0,0) (2,0,0)24 model produces more accurate results for one-step prediction. However, the model does not yield a good prediction for h-step prediction. Further exploration is required to remove the residual seasonality and eliminate nonrandom noise.

SUMMARY AND CONCLUSION

Time Series analysis for this dataset was very complex due to high seasonality noticed in the temperature variable over time. In this analysis, the temperature data was seasonally adjusted and fit to a variety of models, including basic models (Average, Naïve, Drift, SES), moderately complex holt-winter and OLS models, as well as ARMA models. After conducting diagnostic analysis and comparing the RSME values from one-step forecasting, SARIMA (2,0,0) (2,0,0)24 model was selected as the most well-suited model at capturing the data within the dataset.
Base models produced poor predictions while holt winter model provided insight into seasonality but did not yield accurate results. Linear methods like multiple regression were also not equipped with the necessary features to be able to accurately predict the time series data of temperature. ARMA and SARIMA models showed significant improvements compared to the prior models and provide necessary input and insight into time series data. The Seasonal Additive ARIMA model provide greater flexibility for seasonality adjustments. However, it did not provide appropriate results for long term forecasting.

REFERENCES

Jafari, Reza (2021, Fall Session). Lecture 1-12 PowerPoint Slides. Department of Data Science, The George Washington University.

Hyndman, R. J., & Athanasopoulos, G. (2021, May 5). *Forecasting: Principles and Practice (2nd ed)*. Forecasting: Principles and Practice: https://otexts.com/fpp2/regression.html

UCI Machine Learning Repository: Air Quality Dataset: https://archive.ics.uci.edu/ml/datasets/air+quality