Hi everyone, I am Liang. Today I am going to present time series analysis for temperature prediction.

Here is the list of my content today.

The first part is the overview.

The raw dataset, the Jena Weather dataset was recorded every 10 minutes and covers data from January 1st, 2004, to December 31st, 2020. I chose the data from January 1st, 2018, to December 31st, 2020 (3-Year), and averaged the data in each hour. The dataset has 22 columns including” date” and 21 numerical variables. The dependent variable is the temperature in Celsius.

In term of preprocessing the data, I used the drift method to fill in the missing values. I checked the histogram plots and statistical information for all independent variables. Only three variables, wind speed, max PAR, and CO2, have negative outliers. The average method was used to fix outliers in wind speed and CO2 and the naive method was used for max PAR.

This is the plot of the temperature in Celsius versus time, which shows obvious trends and seasonality.

The ACF shows seasonality order at 24 from which is consistent with the characteristics of hourly temperature data.

After splitting train and test. 14,036 in the train set (80%) and 3,509 in the test set (20%).

The target variable passes the ADF test with a p-value of 0.00 but fails to pass the KPSS test with a p-value of 0.02(the p-value threshold is 0.05 for both tests). The rolling mean and variance stabilize once all samples are included. Thus, the dataset is weak-stationary.

I used STL method to decompose the dataset. The strength of the trend is 94.37%, and the strength of the seasonality is 74.79%, so highly trended and seasonal.

The Holt-Winters method considers the level, trend, and seasonality of time series data. I used 744 (monthly seasonality) as the seasonal period. This method captures most seasonality but not the trend.

I normalized the data before running the regression model. All the singular values are greater than 0, but the last few singular values are relatively small compared to the first largest one. The condition number is highly greater than 1,000. Both results from singular value and condition number indicate severe co-linearity among some independent variables.

The threshold for the PCA feature selection is a variance ratio of less than 0.95. 7 features are fitted into the OLS model. The adjusted …….

For the Backwards stepwise regression, I started with the model containing all independent variables, removed one predictor with the highest p-value at a time. 3 features were deleted. Then I found the confidence interval for some coefficients is very small, and I decided to remove those variables. 8 features are left. The model result ……

The threshold for the VIF value is 10. Again, I started with the model containing all independent variables, removed one predictor with the highest VIF value at a time (deleted 9 features). Then I deleted one feature with a p-value greater than 0.05, and 6 features with small coeﬀicients. 3 features left. The mode result……

The model derived from VIF has fewer features and no multi-collinearity problem, so this model would be the final Multiple Linear Regression. The forecast shows most trend and seasonality were captured by this model.

In this case, the p-values from T-test are less than 0.05, so we can reject the null hypothesis and conclude the coefficients are not 0. The p-value for the F-test here is significantly less than 0.05, so we can reject the null hypothesis and conclude my model provides a better fit than the intercept-only model.

The consistency of the metrics across different folds of cross-validation suggests that the model is stable and generalizes well to different subsets of the data.

This plot shows the h-step prediction results based on based models (Average, Naive, Drift, SES). All four models have low mean of error. The average method shows the best performance here compared to other three.

Next core part is the SARIMA model. I fitted the raw data into GPAC and plotted the ACF/PACF. The ACF tails off and PACF cuts off, which indicates the AR process, and with a seasonal pattern at order 24. The GPAC shows a pattern at na equals 1 and nb equals 0 or 1. I tried both nb and then I found nb equals 1 has better performance. This is the result of 1-step prediction. The ACF or residual still shows seasonality. The ACF/PACF of residual shows a large correlation at lag equals 24, so I chose 1 as the seasonal order of AR and MA (1). The GPAC indicates na equals 1 or 2.

I fitted non-seasonal order (1,0,3) and seasonal order (1,0,1,24) to the model this time. The ACF/PACF shows faint significance in the second seasonal order (48). Then I fitted (1,0,3) and seasonal order (2,0,2,24) into the model. The ACF for residual shows most autocorrelation between residuals is removed. The performance of 1-step prediction is good. I finally chose this model as my final SARIMA model.

The model captures some seasonality but still unable to capture the trend.

I performed two tests to check if the residual error was white noise. Unfortunately, both tests failed. The residual errors are not white noise. The estimated mean of the forecast error is -1.64, so the derived mode is biased. The variance of the residual errors is 1.35, and the variance of forecast errors is 47.65. I performed a zero-pole cancellation operation and there is no zero cancellation.

The average improvement of variance of forecast errors by the SARIMA model is 16.06%. The average improvement of MSE is 30.77%. The table shows detailed comparison of MSE and variance of forecast error between base models and SARIMA. In conclusion, SARIMA model has much better performance than the base models.