**IE5202 Project 2**

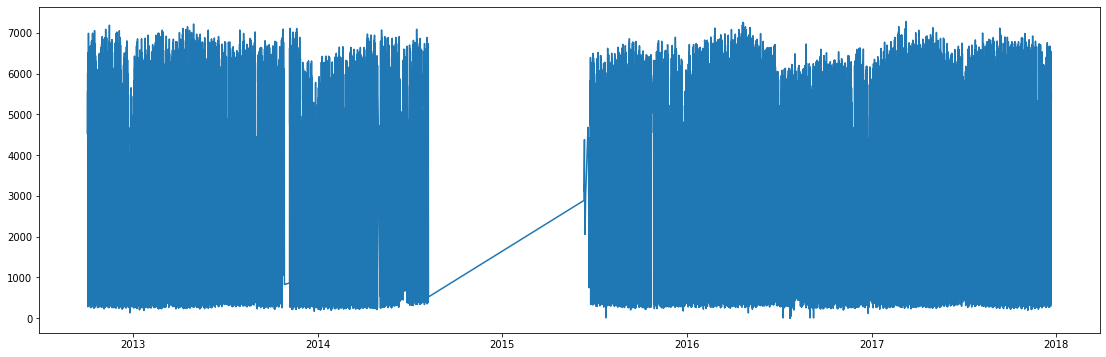
Dong Zheng

A0119545B

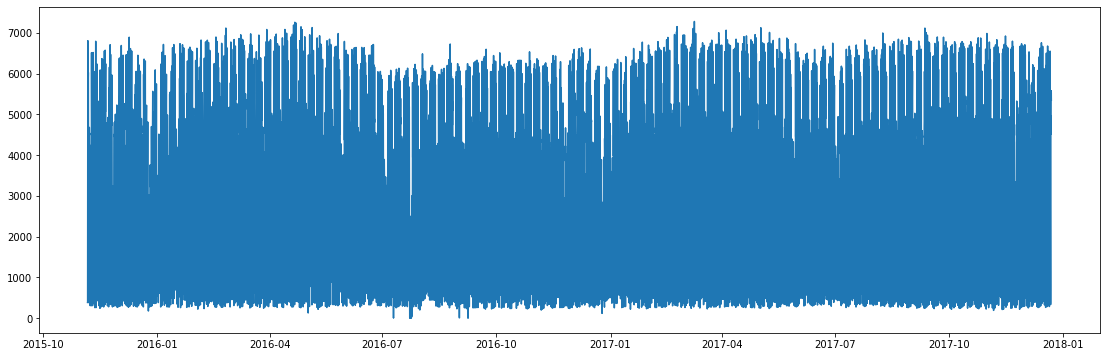
In this project, we would like to predict traffic volume in 2018, given 5 years historical data of hourly traffic volume and weather information.

**1: Data Exploration**

First of all we plot the given training data set from 2012 to 2017 as Figure below.

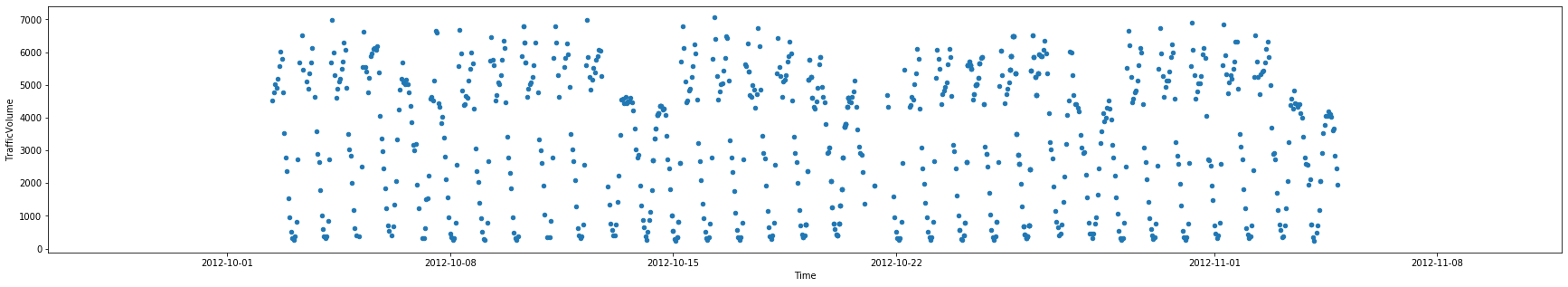


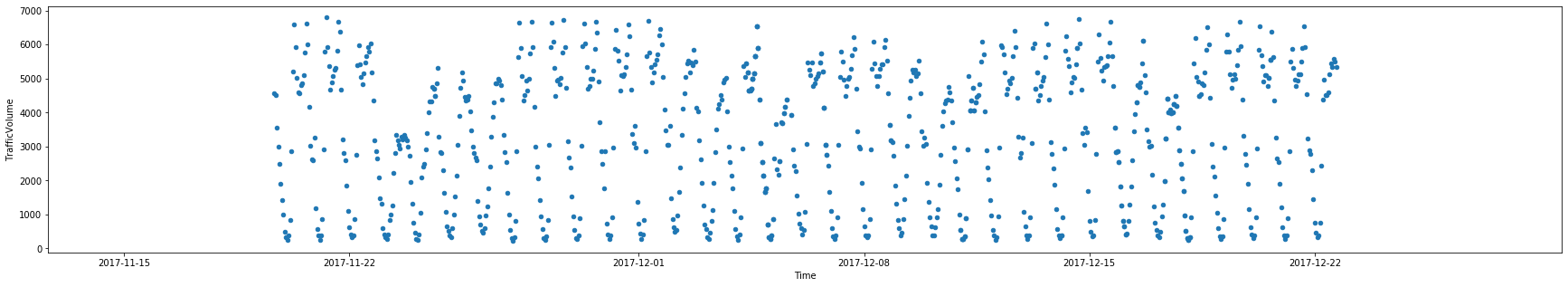
*Figure 1.1 Training data from 2012 to 2017*



*Figure 1.2 Training data from 2015 to 2017*

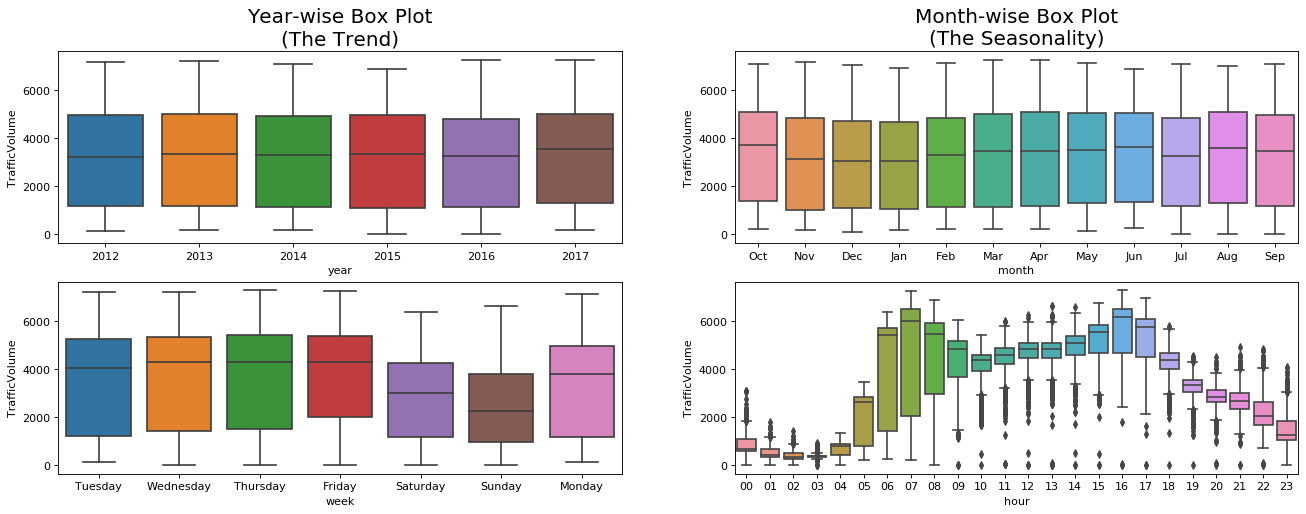
Notice that there is a big gap of missing data from 2014-08-01 to 2015-08-01. We will first focus on right part of the training data set because it is nearer to the test data set. Zoom into a smaller time interval, you can observe that there are obvious daily and weekly seasonal components, and trend movement is trivial.





*Figure 1.3 Weekly training data in 2012 and 2017*

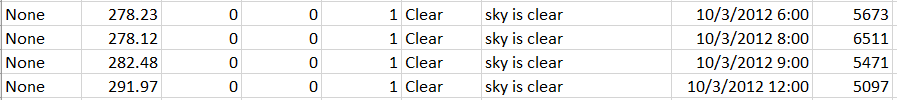
In order to further investigate seasonality, new columns of [year, month, week, hour] information are generated for each observation in the data. From the following box plots of traffic volumes within each year, month, week and hour. We can see that hourly and weekly seasonal patterns are significant, while year-wise trend and monthly seasonal effects are not significant.

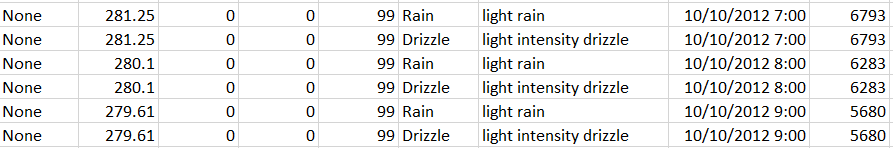


*Figure 1.4 box plots of traffic volumes by year, month, week and hour*

In the training dataset, some hourly data are missing and some other hourly data are duplicated if weather changes. These ununiform time indexes will cause problem in some time-series model, such as exponential smoothing model. Therefore, we have done some data pre-processing and data cleaning. Index duplicated data are aggregated by their mean, and missing hourly data are interpolated by their preceding and succeeding observations. Moreover, Holiday information is also converted to Boolean data type for later use.

*Table 1.5 Missing and duplicated data within an hour*





**2: Regression on Time**

After data exploration, in this section we need to build a regression model by only using 'Time' and the response value 'TrafficVolume'. Two different methods used for model building are seasonal factor approach and trigonometric functions approach.

* 1. **Seasonal Factor Model**

Recall from previous section that the data exhibits seasonal variation on daily and weekly basis. Therefore, a regression model of the following form can be used:

yt = TRt + SNt + εt;

The linear trend part (TRt) can be expressed as , where t is the time delta to the time zero, corresponding to ‘T\_difference’ column of the training dataframe. The seasonality part (SNt) has three components: hour, week and month. Then, we can plug in the following formula to train the OLS regression model:

TrafficVolume ~T\_difference+C(hour)+C(week)+C(month)

The regression result is in Appendix 1 and shows that model has a R-squared of 0.838, which indicates that 83.8% of the variability of the response variable can be explained by the seasonal factor model.

* 1. **Trigonometric Function Model**

Another common way to model seasonal time series data is to use the trigonometric functions. Here we use collections of periodic functions with 4 different types of frequencies. Hence Sin1, Cos1, Sin2, Cos2, Sin3, Cos3, Sin4, Cos4 values are calculated and stored in the dataframe. We use period (L) of 24 hours and fit above trigonometric values into the model and estimated coefficients are as below:

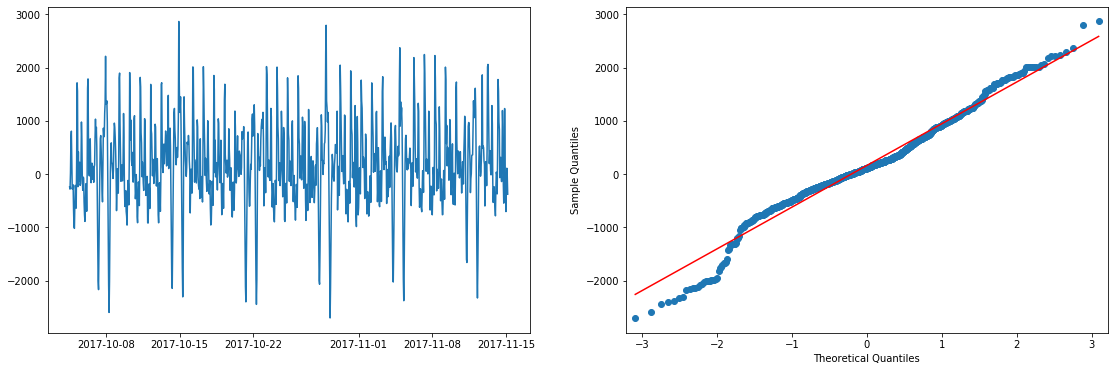
*Table 2.2.1 Coefficients for trigonometric function model*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Sin1** | -699.6669 | 8.522 | -82.105 | 0.000 | -716.370 | -682.964 |
| **Cos1** | -2133.0414 | 8.521 | -250.323 | 0.000 | -2149.744 | -2116.339 |
| **Sin2** | -328.8191 | 8.521 | -38.589 | 0.000 | -345.521 | -312.117 |
| **Cos2** | -638.6561 | 8.522 | -74.945 | 0.000 | -655.359 | -621.953 |
| **Sin3** | -347.2996 | 8.521 | -40.756 | 0.000 | -364.002 | -330.597 |
| **Cos3** | 468.4080 | 8.521 | 54.969 | 0.000 | 451.705 | 485.111 |
| **Sin4** | -11.7784 | 8.521 | -1.382 | 0.167 | -28.481 | 4.924 |
| **Cos4** | 153.8245 | 8.521 | 18.051 | 0.000 | 137.122 | 170.527 |

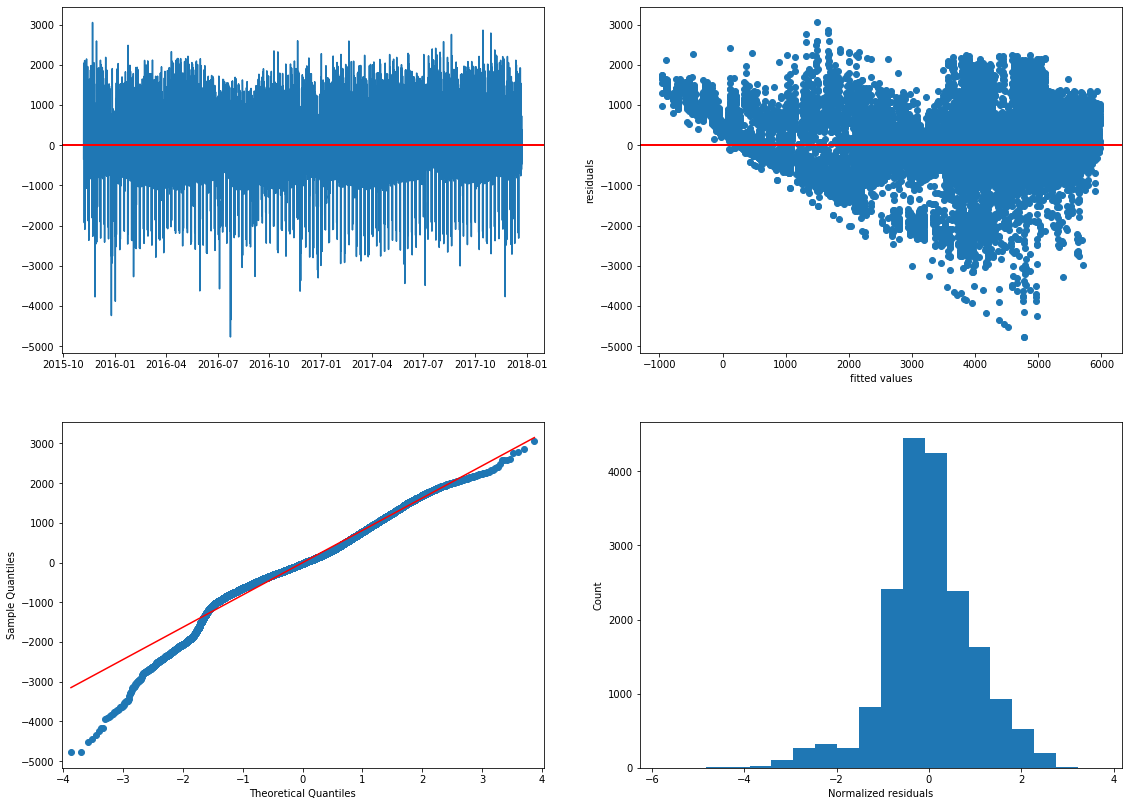
We can see that most of the trigonometric terms are significant as their p-values are small. The R-squared and Adj. R-squared of the trigonometric model are 0.823 (Appendix 2), which is similar to 0.838 of the seasonal factor model. We can expect trigonometric model’s R-square is less than that of seasonal factor model, because trigonometric model has less parameters in the model. Here we achieved similar R-square value with less by using Trigonometric Function Model here.

* 1. **Model Diagnostic**

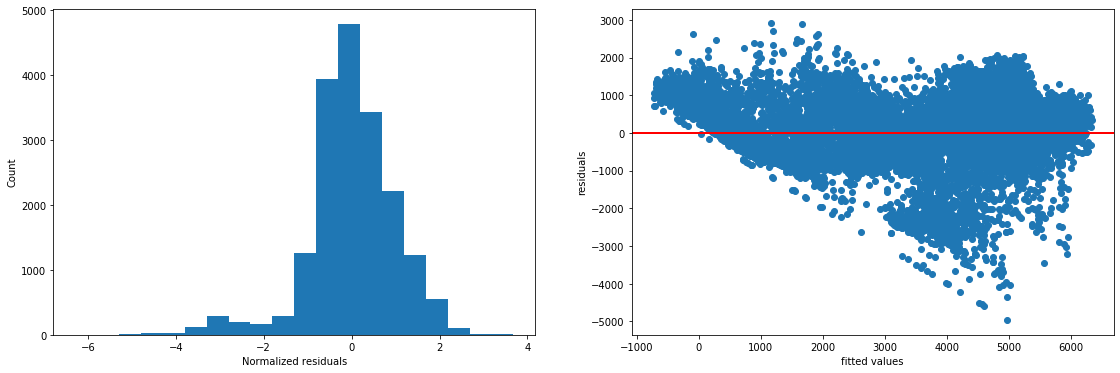
After we trained our Regression on Time model, we need to diagnose the models and check if the model assumptions are satisfied. Regression assumptions made for error terms are 1. Zero mean 2. Constant variance 3. No correlation with X 4. No autocorrelation 5. Normally distributed. Here we will use residuals from seasonal factor model to do the diagnostic.

*Figure 2.3.1 Residual plot for seasonal factor model*

First, we plot the residuals of the training data set as below to check assumption 1, 2 & 4. From the residual plot, we can see that the mean of the error term is around zero and variance are in general constant. However, we can observe that there is an obvious seasonal pattern in the residual plot. Thus, assumption 4 is violated. We need to adopt other method in the later section to tackle autocorrelation property of the residuals.

Secondly, to check the normality of the residuals, quantile-quantile (QQ) plot and histogram of normalized residuals are displayed as below:

*Figure 2.3.2 Histogram and QQ plot for normalized residuals*

From the right side of the histogram and QQ plot, we can see that residuals are quite normally distributed. However, on the left extreme, the residuals deviates far from normal distribution. It may be due to the non-negative constrain of traffic volume, while our regression model may predict negative result for extreme low traffic volume.

*Figure 2.3.3 Residuals versus fitted value*

The residuals against fitted value plot valid our previous guess: when the fitted values are less than zero, residuals are always positive. There is some correlation between fitted values and residuals. From the bottom slash line we can infer that, as the fitted value goes large, our model is more likely to over-estimate the traffic volume. Therefore, it give us motivation to find a better approach to model residuals in the later part.

**3: Exponential smoothing**

In this section, we will explore different kinds of exponential smoothing models, which includes simple exponential smoothing, double exponential smoothing and Holt-Winters method. We will use training dataset to select the best model and then use the select the model to make predictions.

Firstly we use simple exponential smoothing which follows a constant trend. Yt = β0 + ϵt. The result obtained from the package is shown as below:

*Table 3.1 Results for simple exponential smoothing*

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | TrafficVolume | **No. Observations:** | 1000 |
| **Model:** | SimpleExpSmoothing | **SSE** | 760388237.529 |
| **Optimized:** | True | **AIC** | 13545.584 |
| **Trend:** | None | **BIC** | 13555.400 |
| **Seasonal:** | None | **AICC** | 13545.625 |

The optimal alpha obtained is 0.995, means that the new observation is highly relevant to the latest observation. It makes the model hard to predict precisely in the long term. Moreover, since the model only assume a constant trend, it is not able to catch the trend and seasonal information as shown in the results from package.

Secondly, we fit our data in to double exponential smoothing smoothing, the Holt’s Trend Model Results is as below:

*Table 3.2 Results for double exponential smoothing*

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | TrafficVolume | **No. Observations:** | 1000 |
| **Model:** | Holt | **SSE** | 637642390.091 |
| **Optimized:** | True | **AIC** | 13373.533 |
| **Trend:** | Additive | **BIC** | 13393.164 |
| **Seasonal:** | None | **AICC** | 13373.617 |

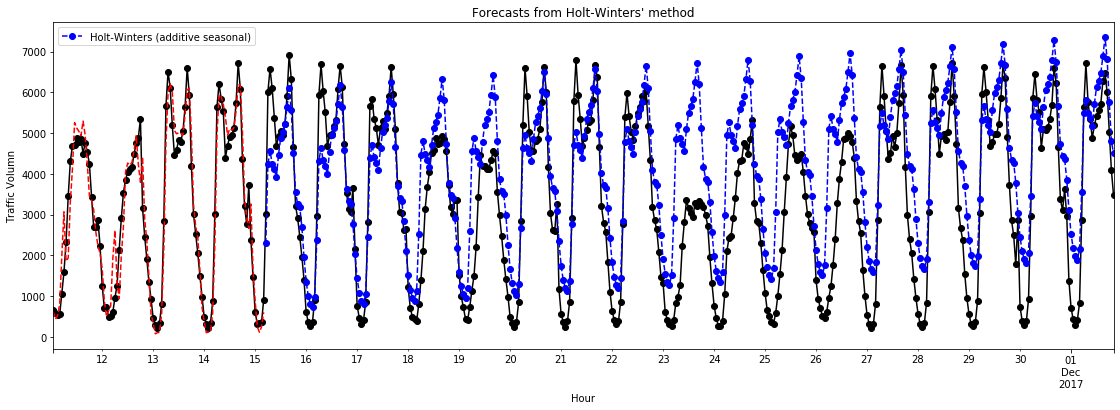
Again, the level smoothing parameter (β0) and growth rate smoothing parameter (β1) obtained are both 0.995. It means that both level and trend of training data are changing drastically. The Holt’s Trend model can capture the trend information but is not able to capture the seasonality in the data set, because it has a form of Yt = β0 + β1t + ϵt. We will compare its results with Holt-Winters Method later on to see which is the better model.

Lastly, we have Holt-Winters Method of the form: Yt = β0 + β1t + SNt + ϵt. Three smoothing parameters β0, β1, SNt need to be updated, to return series of smoothed points. We also need to decide the seasonal periods manually. Here we have two choice, 24 or 24\*7 corresponding to hours in a day and week respectively.

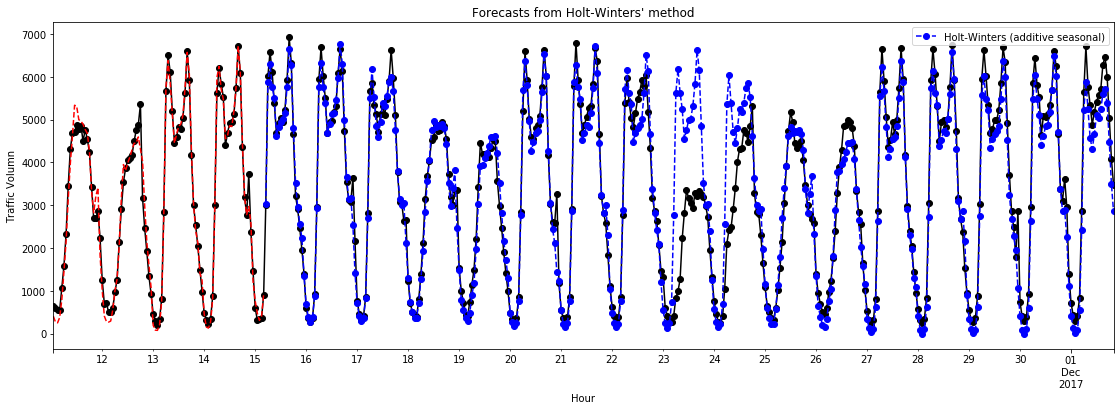
*Table 3.3 Results for Holt-Winters method (L=24\*7)*

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | TrafficVolume | **No. Observations:** | 1000 |
| **Model:** | ExponentialSmoothing | **SSE** | 46020454.315 |
| **Optimized:** | True | **AIC** | 11080.841 |
| **Trend:** | Additive | **BIC** | 11924.975 |
| **Seasonal:** | Additive | **AICC** | 11154.659 |

When seasonal periods is 24, The level smoothing, trend smoothing and seasonal smoothing coefficient obtained are 0.95, 0.0001, and 0.04 respectively. This implies the tanning data has some seasonal component while trend component is not obvious. While seasonal periods is 24\*7, alpha, beta and gamma are 240.3975957, 0.0025796 0.0039728.



*Figure 3.1 Predicted vs. True value for Holt-Winters method (L=24)*



*Figure 3.2 Predicted vs. True value for Holt-Winters method (L=24\*7)*

The blue lines display the forecasting results of Holt-Winters. When seasonal\_periods is 24, we can see that the even though at the beginning, the forecasted values are close to blue true value, as time span gets longer, the forecasted value start to deviate from the true value. Especially after one weekend, the model tends to overestimate the traffic volume. Therefore, we change seasonal\_periods to 24\*7, and the forecasted result are the chart below. Most of time the forecasted results look okay, but there is a large discrepancy on 25th Dec, which further causes the future values to deviate.

Comparing SSE and AIC of four exponential smoothing models below, we can see that Holt-Winters Model (24\*7) has the lowest SSE and AIC. Therefore we choose Holt-Winters (24\*7) Model as the best exponential smoothing models for forecasting.

*Table 3.2 Comparing model results for all exponential smoothing models*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Simple Exponential smoothing | Double Exponential Smoothing | Holt-Winters 24 | Holt-Winters  24\*7 |
| SSE | 7.60388237e8 | 6.37642390e8 | 2.760015e8 | 4.6020454e7 |
| AIC | 13545 | 13373 | 12584 | 11080 |

In summary, exponential smoothing models are good at explaining historical data set, but its prediction power for unknown future are quite limited. It is also unable to consider special case such as holidays. Therefore, we will explore more advance models in the next section.

**4: Free form forecasting**

In this section, we will use multivariable regression model to make primary prediction of the missing values in test data, and then use SARIMA model to refine our predictions.

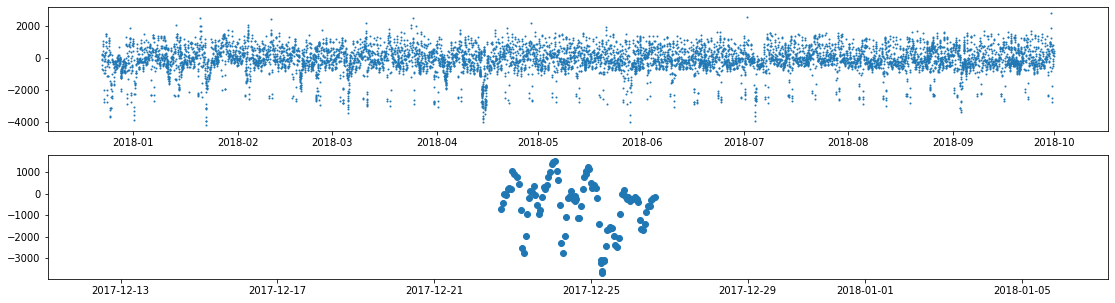
**4.1 Methodology**

From the model diagnostic part in step 1, residuals display some seasonal patterns. Therefore, ARIMA could be a good choice to model the residuals. Moreover, since other information such as weather and holidays in training data could possibly influence traffic volume, we would like to incorporate them into the preliminary multivariable regression model. Best subset selection method picks the final multivariable regression model:

TrafficVolume~C(hour)+C(week)+C(month)+WeatherMain+Temp+CloudsAll+year+IsHoliday

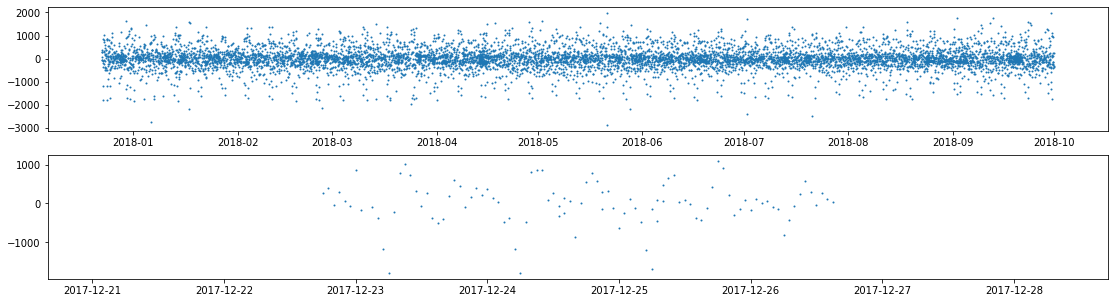
**4.2 Time Series Stationarity**

The regression result in Appendix 4 shows that the model’s adjusted-R square value is 0.841, which is slightly higher than regression on time model in the section 1. The trained model is then applied to the testing data set to get the initial predicted value. Meanwhile, the training residuals from this multivariable regression model are obtained and plotted as below:



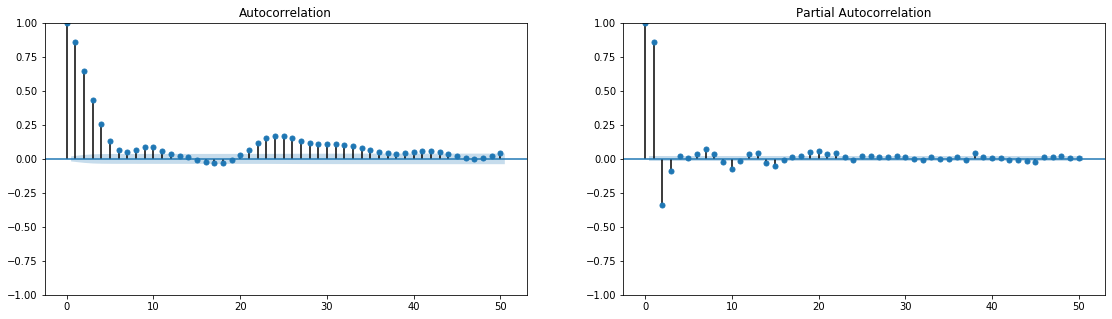
*Figure 4.1 Original residuals for stationarity check*

The residual data is non-stationary, because its variance changes over time. A first order non-seasonal differencing is applied to the residual data.

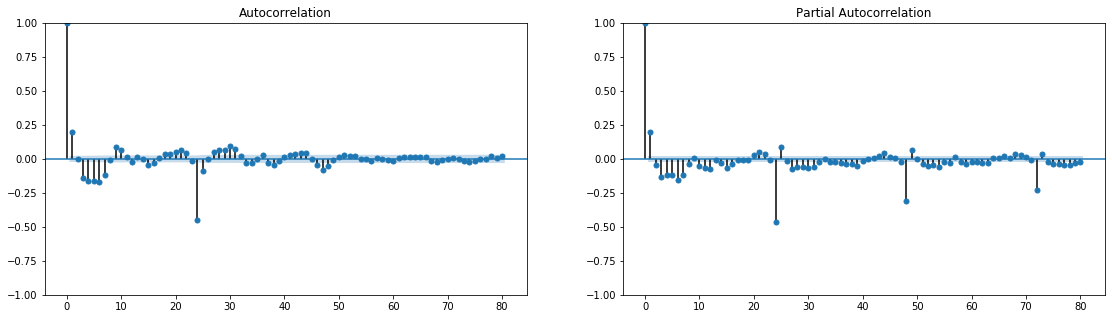


*Figure 4.2 Residuals after first order non-seasonal differencing*

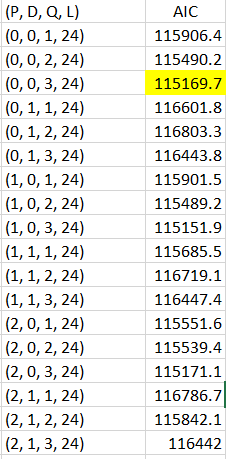
**4.3 ACF, PACF and Parameter Tuning**

The data now appears more stationary after one non-seasonal differencing. Next, ACF and PACF of the differentiated residual are plotted to identify parameter p and q in ARIMA model.

*Figure 4.3: ACF and PACF for residual data*



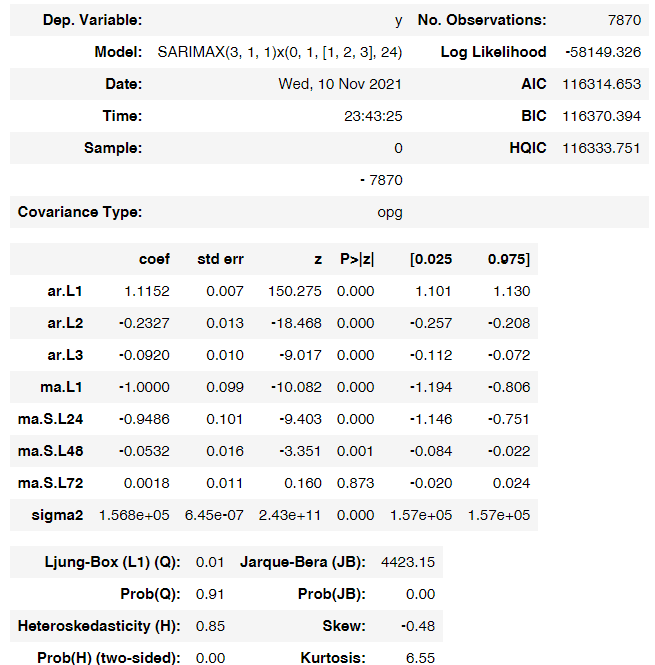
*Figure 4.4: ACF and PACF after first-order differencing*

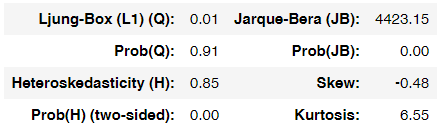
Observed from the non-seasonal autocorrelation plot (ACF) and Partial Auto-Correlation Function (PACF), ACP dies down and PACF cut off at 3, suggesting a MA(3) model. For seasonal peaks, PACF dies down and ACF cuts off after one peaks at 24, suggesting an seasonal AR(2) model. In order to validate our observation and find best ARIMA Model hyperparameters. We grid search the optimal p, q combination as below:

*Table 4.1 Grid Search Results for p,q*

|  |  |  |
| --- | --- | --- |
| AIC | p | q |
| 117479.33 | 1 | 0 |
| 117479.449 | 1 | 1 |
| 116695.07 | 1 | 2 |
| 116442.888 | 1 | 3 |
| 117469.663 | 2 | 0 |
| 116427.549 | 2 | 1 |
| 116368.428 | 2 | 2 |
| 116379.013 | 2 | 3 |
| 117305.486 | 3 | 0 |
| 116360.709 | 3 | 1 |

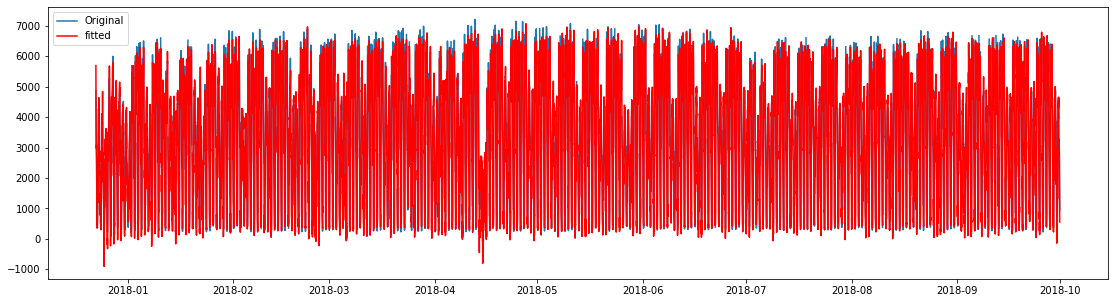
The grid search result indicate the optimal (p,d,q) to be (3,1,1). This combination has the lowest AIC value of 116360. It is also consistent with our previous ACF and PACF plots. For the seasonal hyper-parameters, the gird search shows a combination of (0,0,3,24) has the lowest AIC value. Therefore, a SARIMA model is built with SARIMA(3,1,1)(0,0,3)24 and then applied on test dataset.

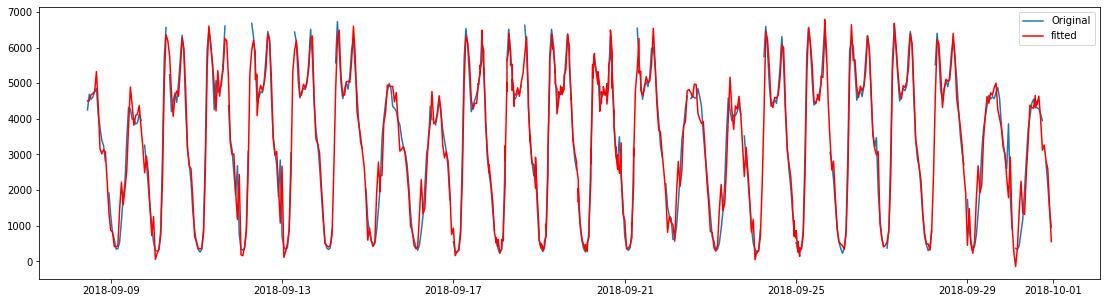
*Table 4.2 Results for SARIMA(3,1,1)x(0,0,3)24 Model*

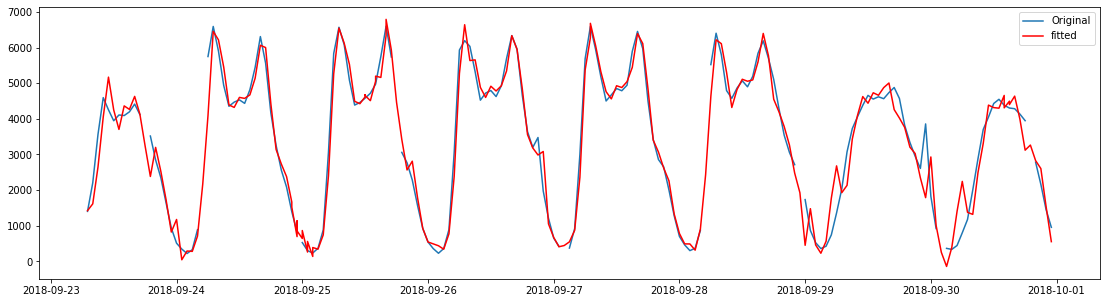


**4.4 Model Forecasting**

One-Step Out-of-Sample forecast is use to forecast the residuals of missing value in the test data set, which means that only one prediction is made at one time, and only values before the prediction time are used. After each prediction, the forecasted results are recorded in the dataframe and SARIMA model is also updated for next prediction. One-step forecasting usually performs better than multi-step in out-of-sample forecasting. It will also make full use of test data before the prediction time point. The final predicted traffic volume is a sum of predicted values from multivariable regression model and predicted residuals from SARIMA model. Results as below:

*Figure 4.5: Fitted Value vs. True Values in Test Data*

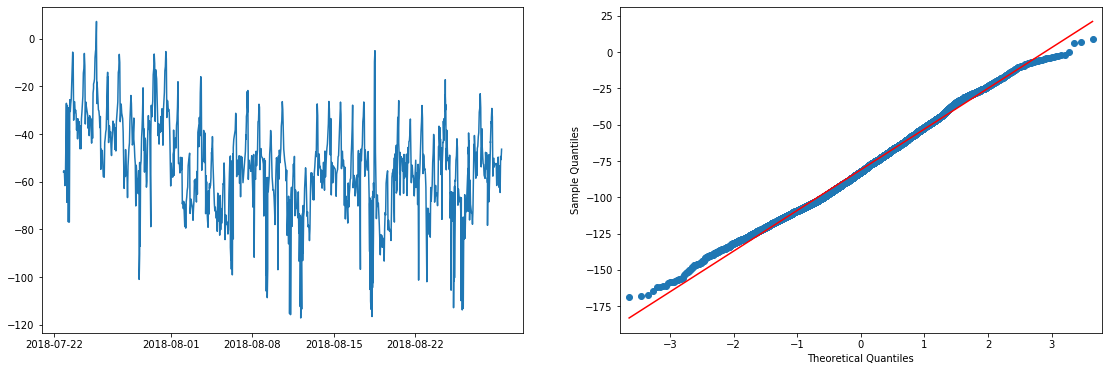


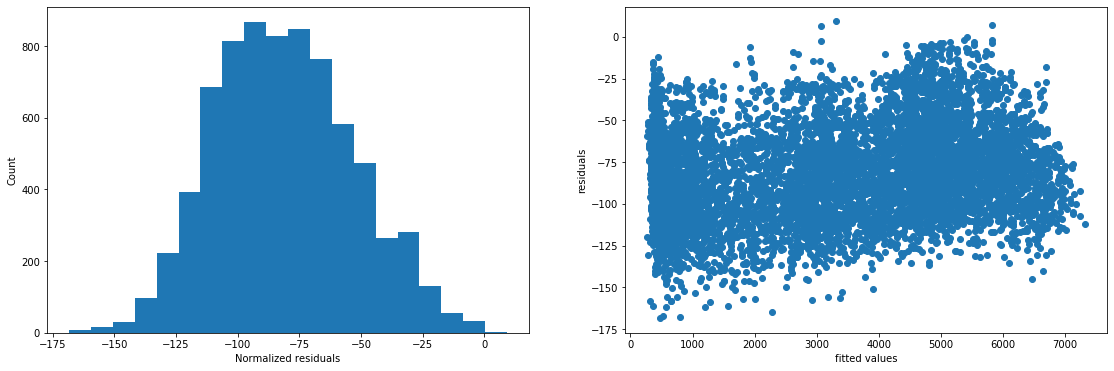


**4.5 Model Diagnostics**

We have made the model diagnostic plot as in the graph below:

*Figure 4.5: Diagnostic plots for final result’s residuals*





From QQ plot and histogram of the residuals plots, we can see that residuals are quite normally distributed. However, it can be observed from 1st graphs that the average value of the residuals is -75 (less than zero), which indicates that our model is very likely to overestimate the traffic volumes. Also from the residuals vs fitted values plot, we can observe that when the fitted traffic volumes are big, our model predicts more accurately. One possible explanation of this situation is there are some unknown factors that limit traffic volume in some days, such as road breakdown, which make prediction for low traffic volume harder.

**4.6 Future Improvement**

There are also some ways that could further optimize our multivariable regression model and SARIMA model. First we can decomposite time series into different components (level, trend, seasonal ) and model them separately. Secondly, employing Augmented Dickey-Fuller (ADF) test for stationarity check after visual check would help to identify better ordered of differencing. Lastly, tuning seasonal hyperparameters P,D,Q and even L values could further optimizes model performance in test dataset.

**Appendix**

Appendix 1. Regression results for Seasonal Factor Model

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | TrafficVolume | **R-squared:** | 0.838 |
| **Model:** | OLS | **Adj. R-squared:** | 0.838 |
| **Method:** | Least Squares | **F-statistic:** | 2240. |
| **Date:** | Sun, 07 Nov 2021 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 12:18:58 | **Log-Likelihood:** | -1.5097e+05 |
| **No. Observations:** | 18664 | **AIC:** | 3.020e+05 |
| **Df Residuals:** | 18620 | **BIC:** | 3.024e+05 |

Appendix 2. Regression results for Trigonometric Factor Model

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | TrafficVolume | **R-squared:** | 0.823 |
| **Model:** | OLS | **Adj. R-squared:** | 0.823 |
| **Method:** | Least Squares | **F-statistic:** | 3464. |
| **Date:** | Sat, 13 Nov 2021 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 12:56:04 | **Log-Likelihood:** | -1.5181e+05 |
| **No. Observations:** | 18664 | **AIC:** | 3.037e+05 |
| **Df Residuals:** | 18638 | **BIC:** | 3.039e+05 |
| **Df Model:** | 25 |  |  |

Appendix 3. Results for triple exponential smoothing (L=24hrs)

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | TrafficVolume | **No. Observations:** | 1000 |
| **Model:** | ExponentialSmoothing | **SSE** | 276001574.456 |
| **Optimized:** | True | **AIC** | 12584.162 |
| **Trend:** | Additive | **BIC** | 12721.579 |
| **Seasonal:** | Additive | **AICC** | 12586.081 |

Appendix 4. Regression results for Multivariable Regression Model

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | TrafficVolume | **R-squared:** | 0.841 |
| **Model:** | OLS | **Adj. R-squared:** | 0.840 |
| **Method:** | Least Squares | **F-statistic:** | 1636. |
| **Date:** | Wed, 10 Nov 2021 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 18:40:46 | **Log-Likelihood:** | -1.3318e+05 |
| **No. Observations:** | 16471 | **AIC:** | 2.665e+05 |
| **Df Residuals:** | 16417 | **BIC:** | 2.669e+05 |

1. Use automatic way to verify p,d,q values, as well as
2. Able to catch the trend level more efficiently

that traffic tends to be over or under predicted at extreme edge.

This suggests that the

and find out that

some non-linearly in the data. Therefore, including interaction terms into our advanced model would be a good improvement.

In order to efficiently find the best interaction for the model, we need to reduce the number of predictors considered for interaction effects.

The correlation matrix of 52 variables exhibits there is some collinearity in our data set. As you can observe from the correlation heatmap (Appendix 2), white color girds indicate highly correlated covariate pairs. To reduce highly correlated predictors in the model, we use SelectNonCollinear function in ‘Collinearity’ package to remove them.

The resulted correlation heat map displays that there are no more highly correlated covariates in the data set.

**Model Building**

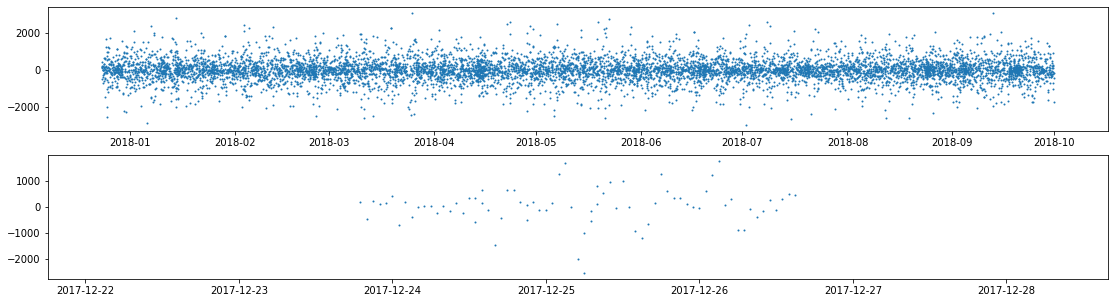
|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | y | **No. Observations:** | 8204 |
| **Model:** | SARIMAX(3, 1, 1)x(0, 1, [1, 2, 3], 24) | **Log Likelihood** | -58018.015 |
| **Date:** | Thu, 11 Nov 2021 | **AIC** | 116052.030 |
| **Time:** | 00:56:22 | **BIC** | 116108.104 |
| **Sample:** | 0 | **HQIC** | 116071.202 |
|  | - 8204 |  |  |
| **Covariance Type:** | opg |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **ar.L1** | 1.1337 | 0.007 | 155.734 | 0.000 | 1.119 | 1.148 |
| **ar.L2** | -0.2398 | 0.013 | -18.879 | 0.000 | -0.265 | -0.215 |
| **ar.L3** | -0.0899 | 0.010 | -8.864 | 0.000 | -0.110 | -0.070 |
| **ma.L1** | -1.0000 | 0.212 | -4.707 | 0.000 | -1.416 | -0.584 |
| **ma.S.L24** | -0.8935 | 0.214 | -4.177 | 0.000 | -1.313 | -0.474 |
| **ma.S.L48** | -0.1184 | 0.027 | -4.451 | 0.000 | -0.171 | -0.066 |
| **ma.S.L72** | 0.0120 | 0.011 | 1.056 | 0.291 | -0.010 | 0.034 |
| **sigma2** | 1.475e+05 | 1.47e-06 | 1.01e+11 | 0.000 | 1.47e+05 | 1.47e+05 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Ljung-Box (L1) (Q):** | 0.00 | **Jarque-Bera (JB):** | 4983.84 |
| **Prob(Q):** | 0.95 | **Prob(JB):** | 0.00 |
| **Heteroskedasticity (H):** | 0.85 | **Skew:** | -0.50 |
| **Prob(H) (two-sided):** | 0.00 | **Kurtosis:** | 6.69 |

for future forecasting.

P,D,Q

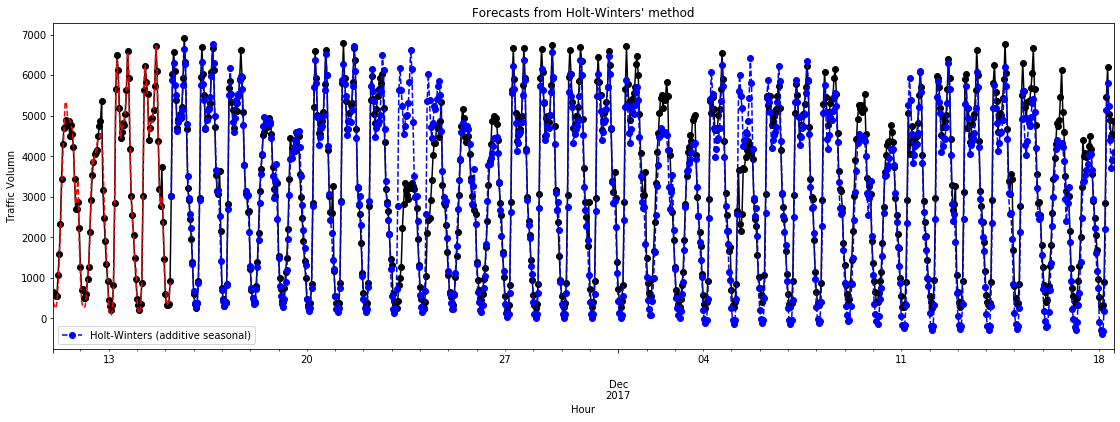


The residual are plotted as below. First we need to check

For seasonal peaks, PACF dies down and ACF cuts off after two peaks at 24 and 48, suggesting parameters to be.

**2.3.2 Result Interpretations and Main Findings**

**2.3.4 Result Interpretations and Main Findings**



|  |  |  |
| --- | --- | --- |
| **coeff** | **code** | **optimized** |
| **smoothing\_level** | 0.9596429 | alpha | True |
| **smoothing\_trend** | 0.0001 | beta | True |
| **smoothing\_seasonal** | 0.0403571 | gamma | True |

Th

Moreover,

Even though,

Trigonometric functions: A collection of periodic functions with a variety of frequencies:TrafficVolume~T\_difference+Sin1+Cos1+Sin2+Cos2+Sin3+Cos3+Sin4+Cos4

**Yt = β0 + β1t + SNt + ϵt**

The method for building Primitive Regression Model can be broken down into three sub-steps: The first is to identify which 5 predictors can explain most of the response value ‘total votes change percentage’. Then we need to analyze each predictor and its residuals to identify the area of improvement. Lastly, we try different power transformations on original variables, to reduce the root mean residual sum of squares and improve the model prediction power.

To find the top 5 predictors, we standardize all 52 quantitative predictors and rank them by their feature importance. The result (in Appendix 1) shows that the top 5 important predictors are 'RHI225214', 'RHI125214', 'RHI325214', 'PST045214' and 'POP010210'. We should not use all of them in our primitive model, because they could be highly correlated. Here we only pick the most important feature 'RHI225214'. The categorical covariate ‘state\_abbr’ in the first column should also be included, then we have three more predictors to choose. The best way is to enumerate all possible 3 predictors combinations and compare their AIC. Thus, we enumerated and compared 20825 models of 5 predictors. The ‘getAll’ function result exhibits that (RHI225214 , state\_abbr, RHI325214, LFE305213, HSG495213) are the best 5 predictors in terms of AIC value, and (RHI225214 , state\_abbr, POP645213, LFE305213, HSG495213) are the second best model predictors.

**2.2 Justifications and Findings**

**1.2 Result Interpretations and Main Findings**

Now let us build up our primitive regression model using formula: **Total\_votes\_change\_percentage ~ RHI225214 + C(State\_abbr) + RHI325214 + LFE305213 + HSG495213 + 1**

The Adj. R-squared value (0.372) is not very high, suggesting that there is room for improvement. F-statistic (29.43) determines that at least one of the covariate’s coefficient is not zero. Look into P-value list, RHI225214, RHI325214, LFE305213, HSG495213’s P-values are all very small, indicates that they are all statistically significant variables. Categorical variable ‘State\_abbr’has different P-values for different states. States with small P-values indicates that these states have significant impact on ‘total votes change percentage’. While in other states with large P-values, they are not significantly different from the average response value.

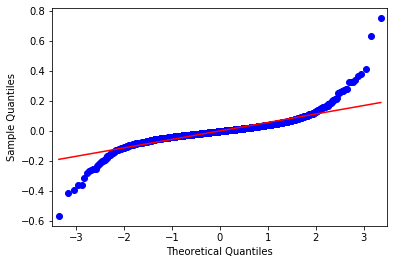
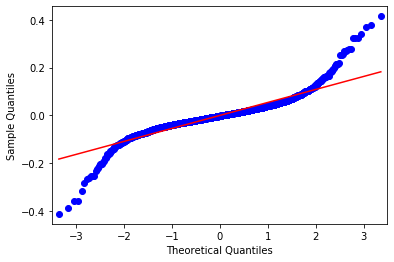
Let us then look at what are these 5 variables mean and what insight we can get from them.

*Table 1.1: Top 5 Important Predictors*

|  |  |
| --- | --- |
| Predictors | Description |
| RHI225214 | Black or African American alone, percent, 2014 |
| RHI325214 | American Indian and Alaska Native alone, percent, 2014 |
| HSG495213 | Median value of owner-occupied housing units, 2009-2013 |
| LFE305213 | Mean travel time to work (minutes), workers age 16+, 2009-2013 |

RHI225214 and RHI325214 are two variables that related to ethnic composition of a state

**4. Model Diagnostics**

In this section we diagnose our primitive regression model and find ways to improve it. The three assumptions made for error terms in linear regression are constant variance, independence of variables and normality of the distribution. These assumptions need to be checked and see if our model violates any of the them. First, we plot the normal qq-plot to check the normality of the data.

*Figure 1.1 Normal QQ-plot before and after removing 3 Outliers*

From the plot we can see that within the middle part, our data is quite normally distributed. However, at two extremes our model data deviate and do not follow normal distribution. we can easily identify two outliers at top right corner and one at lower left corner. After checking original data set, we found that these outliers are all from Texas. In the future model, we can create a new dummy variable called ‘isTexasOutlier’ to consider these special data, or probably remove them from our training set because there could be leverage and influence points.

*Table 4.1: Extreme Outlier ID*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Row ID | State\_abbr | |  | | --- | | Total\_votes\_change\_percentage | | Dem\_change\_percentage |
| 578 | TX | 0.78785468 | -0.045181407 |
| 2274 | TX | 0.65608987 | 0.005040385 |
| 785 | TX | -0.53045616 | -0.042610196 |

Secondly, the residuals against fitted values plot is displayed as below to check the constant variance assumption. We can see that despite there are a few outliers in the middle part, the overall residuals variance are quite constant along the x-axis.

*Figure 1.2 residuals against fitted values*

Lastly we also need to check model residuals against each predictor to see if they are independent of each variable. Regression plots for RHI225214, RHI325214, LFE305213, HSG495213 are listed below in sequence:

From RHI225214 and RHI325214 Regression plots, we can see that there is a decreasing trend of variance along x-axis, and there is a little non-linearity in the RHI225214 scatter plot. Thus, we may consider applying a log transformation for RHI225214 and RHI225214 data to see if that can make residual terms more constant.

In terms of the LFE305213, the residual versus LFE305213 plots shows that variance are generally constant. Therefore, we do not need to apply any transformation to LFE305213 data. Lastly, residuals versus HSG495213 plot shows that variances are not so normally distributed and we may consider applying Box-Cox transformation to the data.

In conclusion, assumptions made for linear regression model are generally satisfied, except that some extreme data points may not follow normal distribution. The partial regression plots also imply that some systematic variance in our preliminary regression model could be reduced by including higher-order terms, such as interactions and transformations. We will explore that in our Step 2 Advanced Regression Model.

**Step 2 Advanced Regression Model**

In step two, we can include interaction and any new variables in order to build the best regression model for predicting the responses.

**2.1 Methodology**

From the model diagnostic part in step 1, we observed that there are some non-linearly in the data. Therefore, including interaction terms into our advanced model would be a good improvement. Since we have 53 predictors, which means that there will be more than thousands of two-factor interactions for us to consider. Thousands of interactions are a little bit redundant for the forward selection algorithm and it will take very long computation time.

In order to efficiently find the best interaction for the model, we need to reduce the number of predictors considered for interaction effects. The correlation matrix of 52 variables exhibits there is some collinearity in our data set. As you can observe from the correlation heatmap (Appendix 2), white color girds indicate highly correlated covariate pairs. To reduce highly correlated predictors in the model, we use SelectNonCollinear function in ‘Collinearity’ package to remove them. The resulted correlation heat map displays that there are no more highly correlated covariates in the data set.

There are around 30 predictors remained which is still too much for interaction terms. Then sorting the remaining predictors by their VIF (Variance Inflation Factor) score and filtering out predictors, that have a VIF Score higher than 10, will give us final 20 predictors list as below:

*Table 2.1 List of Predictors to be Considered for Interaction Effects*

Then

. **2.1 Methodology**

**2.2 Justifications and Findings**

By including second-order interaction terms and their main effect, our advance-regression model for predicting ‘total\_votes\_change\_percentage’ results are shown in Table 2.2 below. The Adj. R-squared increase from 0.372 to 0.536 and the AIC value decrease from -6670 to -7400. If it is also true in the test data set, our advanced model’s prediction power has increased around 50% when comparing with our primitive regression model.

*Table 2.2 OLS Regression Results for Advanced Regression Model – R1*

Our advanced regression model for predicting ‘dem\_change\_percentage ‘ results are shown in Table 2.3 below. The Adj. R-squared value is 0.849 indicates that the statistical patterns in ‘dem\_change\_percentage’ are more obvious and it could be easier for us to predict ‘dem\_change\_percentage’ than ‘total\_votes\_change\_percentage’.