

Time Series Analysis on Daily Website Visit

Time Series Forecasting Final Project

Seungheon Han, Alexis Yang, Qunzhe Ding, Chengshu Yang

The George Washington University
Master of Science in Business Analytics

< 1. Introduction and Overview >

The dataset we chose to use is “website_visit.csv”. This dataset provides daily observations from September 14, 2014 to August 19 2020 regarding the daily visit of the website: “<http://people.duke.edu/~rnau/411home.htm>”. The variables being involved in this project are shown below:

- 1) **First.Time.Visits:** Number of unique visitors who do not have a cookie identifying them as a previous customer.
- 2) **Returning.Visits:** Daily number of visitors from whose IP addresses there haven't been hits on any page in over 6 hours minus first time visitors.
- 3) **Page.Loads:** Daily number of pages loaded.

Our interest is to select the best forecasting model among the below time-series models. The link of the dataset is (<https://www.kaggle.com/bobnau/daily-website-visitors>). We are going to use forecast horizon: 12, Hold-out Sample: 300

< 2. Univariate Time-series models >

Fitting the univariate model with the variable FIRST.TIME.VISIT

series:

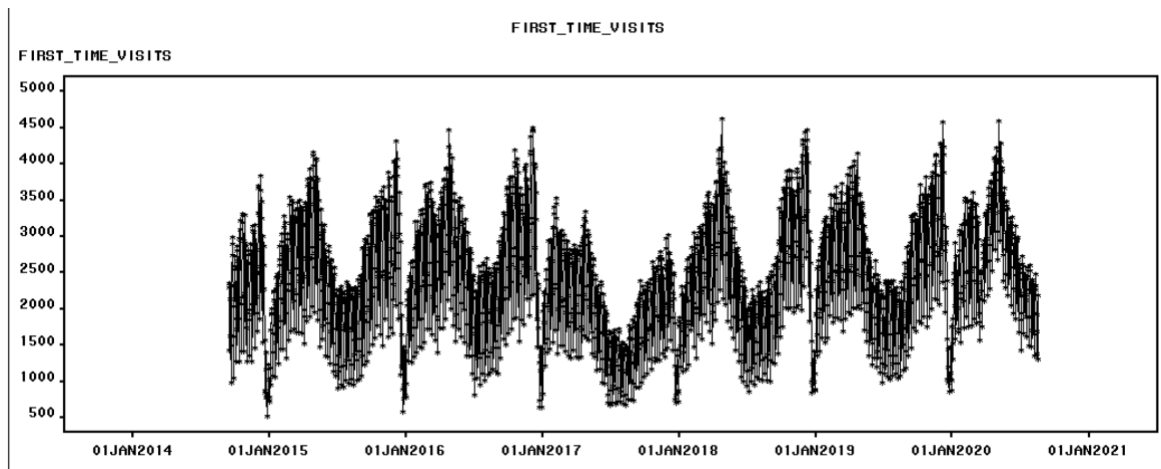


Figure1-1

As shown in the figure1-1, the overall series is characterized by a yearly-based seasonal pattern created by the daily data. The variability of the seasonal patterns look similar and no significant trend is observed.

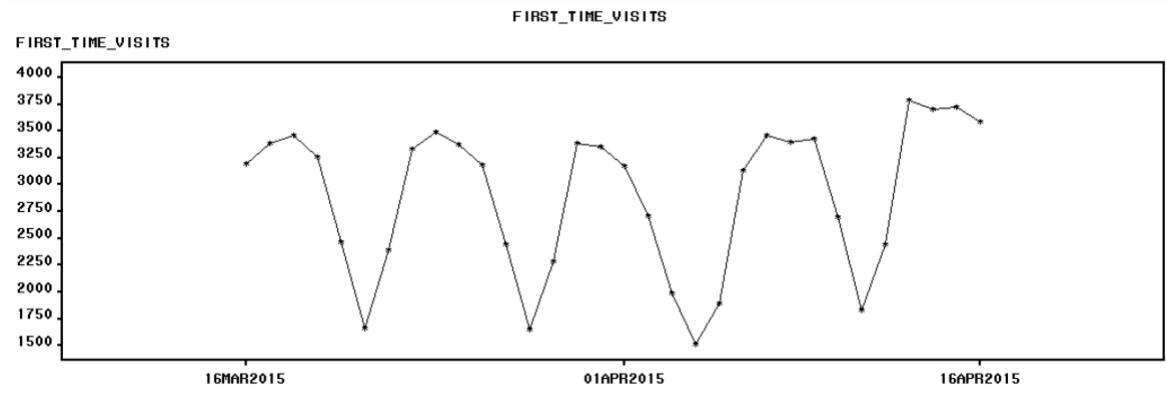


Figure1-2

The close-up series in Figure1-2 demonstrates the data also has the weekly pattern without trend.

2.1 Deterministic Time Series Models and Error model.

Forecast horizon: 1, Hold-out Sample: 300

- **Seasonal Dummies and Trend**

Seasonal dummies and linear trend

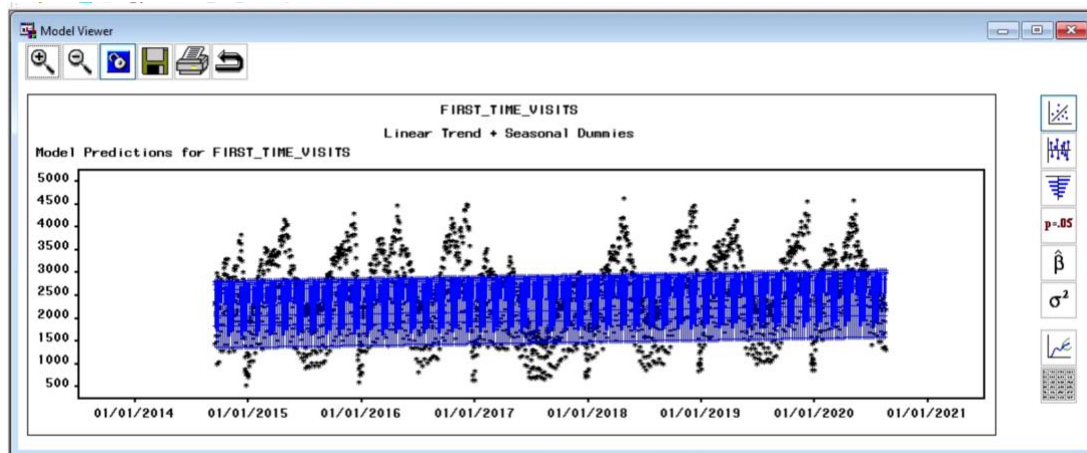


Figure 2-1

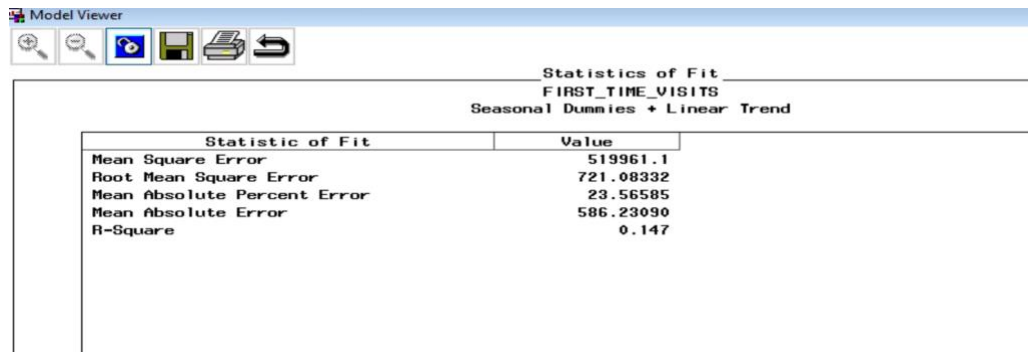


Figure 2-2

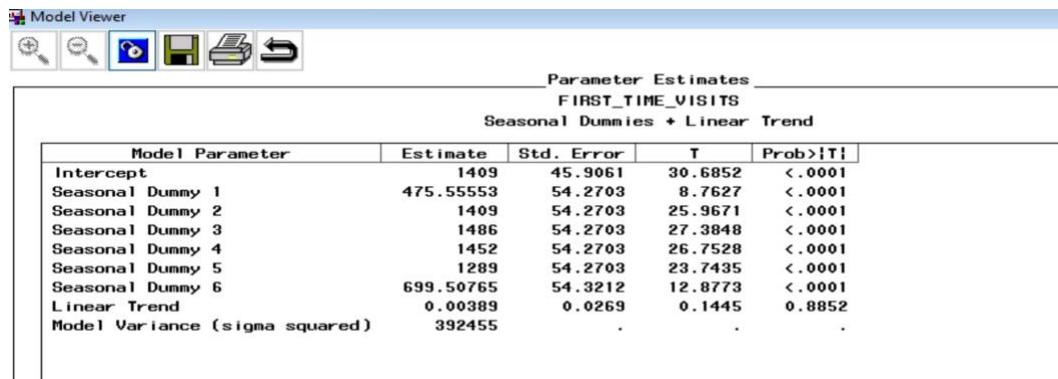


Figure 2-3

The seasonal dummies and linear trend model does not fit well to the series, mean absolute percent error is 23.56% and R-square is as low as 0.147. However, when we check the coefficient of estimate, we see the linear trend p value is 0.8852, which is greater than 0.05. We conclude that the linear trend is not statistically significant.

Next, we are interested in switching to the first different trend to see how the model performs as the original series has an obvious seasonal trend.

- Seasonal dummies and first difference trend

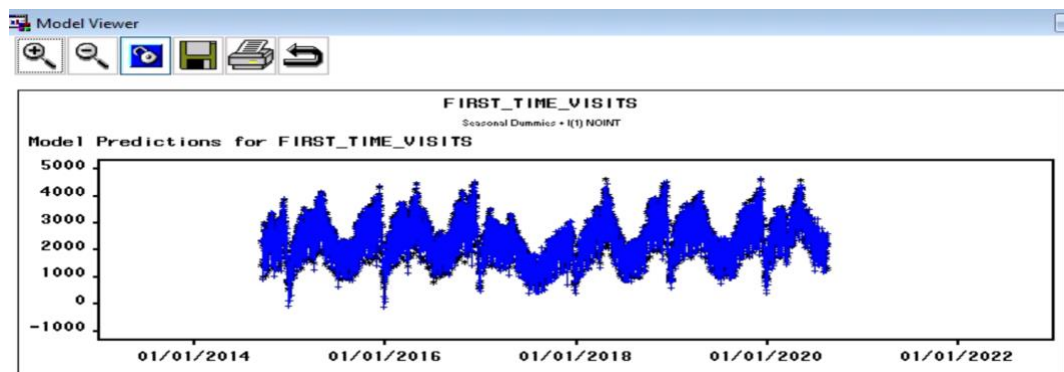
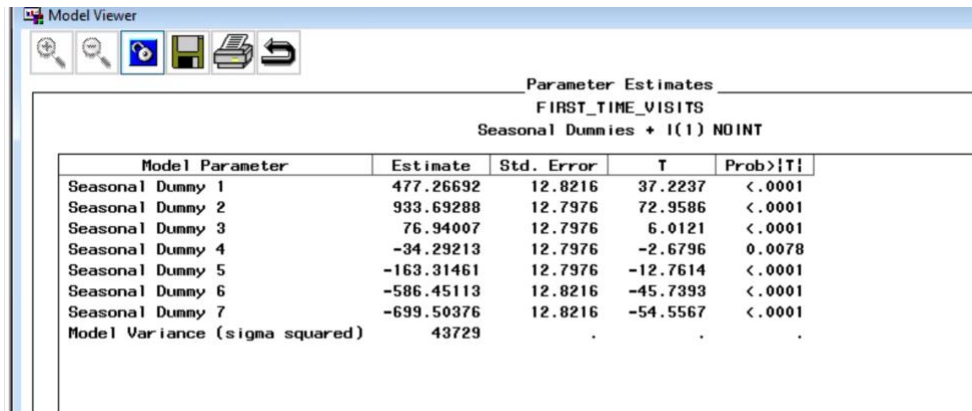


Figure 2-4

According to the picture above, by switching from linear trend to first difference the model fits much better to the series.

Parameter Estimate:

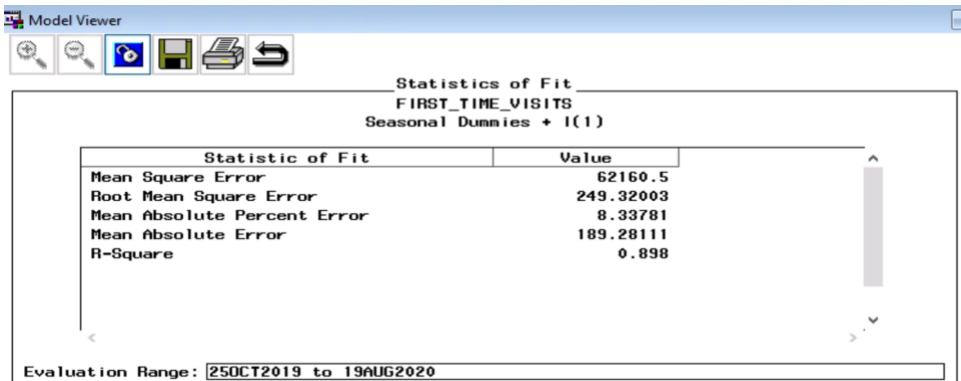


Model Parameter	Estimate	Std. Error	T	Prob> T
Seasonal Dummy 1	477.26692	12.8216	37.2237	<.0001
Seasonal Dummy 2	933.69288	12.7976	72.9586	<.0001
Seasonal Dummy 3	76.94007	12.7976	6.0121	<.0001
Seasonal Dummy 4	-34.29213	12.7976	-2.6796	0.0078
Seasonal Dummy 5	-163.31461	12.7976	-12.7614	<.0001
Seasonal Dummy 6	-586.45113	12.8216	-45.7393	<.0001
Seasonal Dummy 7	-699.50376	12.8216	-54.5567	<.0001
Model Variance (sigma squared)	43729	.	.	.

Figure 2-5

Seasonal dummies represent Monday through Sunday, p value of each seasonal dummy is smaller than 0.05, meaning they are statistically significant. We can see the difference of first-time-visits between Monday and Tuesday, the difference between Tuesday and Wednesday, and the difference between Wednesday and Thursday have accelerated positive relations, the difference for the rest of days have accelerated negative influence.

Statistics of Fit:



Statistic of Fit	Value
Mean Square Error	62160.5
Root Mean Square Error	249.32003
Mean Absolute Percent Error	8.33781
Mean Absolute Error	189.28111
R-Square	0.898

Evaluation Range: 25OCT2019 to 19AUG2020

Figure 2-6

The mean absolute percent error is 8.48% and R-square is 89.9%. The seasonal dummies and first difference does a good job on predicting.

Stationary and White Noise of residual:

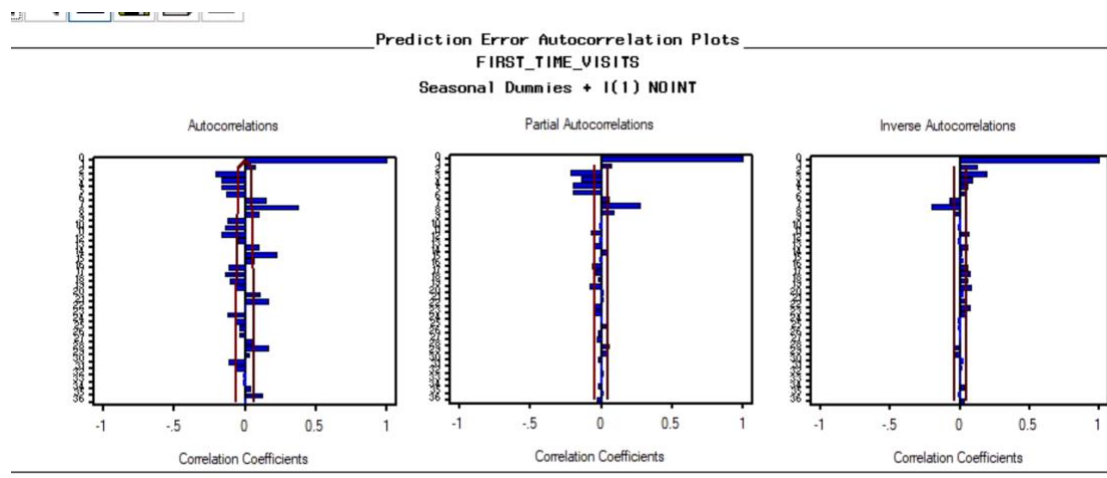


Figure 2-7

Regarding the above graph, we can see lag 7, 14, 21, 28, 36, ect are decaying slowly, and the non-seasonal lags decay exponentially which finally is within the 2.s.e bound. There is a seasonal AC behavior, we are going to dig into more on the seasonal lags in the following models.

- **Cyclical model**

Identify the 10 harmonics (for purposes of parsimony) with the highest amplitudes to include in a cyclical trend model.

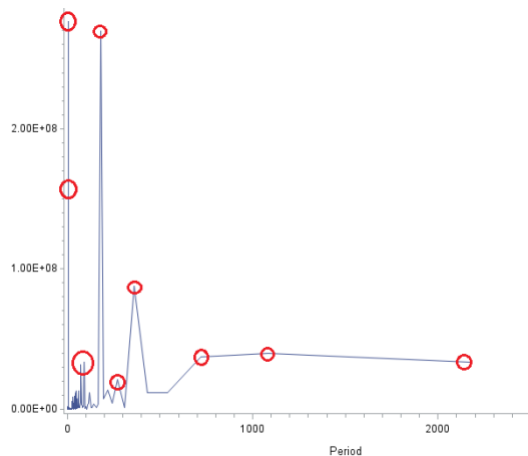


Figure 3-1

Harmonics	Period	P_01
310	6.99	276216135.08
12	180.58	269348629.99
309	7.01	166216950.42
6	361.17	87627465.39
2	1083.50	39702304.86
3	722.33	37356809.81
24	90.29	33607895.52
1	2167.00	33392841.83
30	72.23	31777528.60
8	270.88	20810670.32

Table 1

Create the necessary sine and cosine pairs we have identified from the periodogram. Plot of actual versus predicted values.

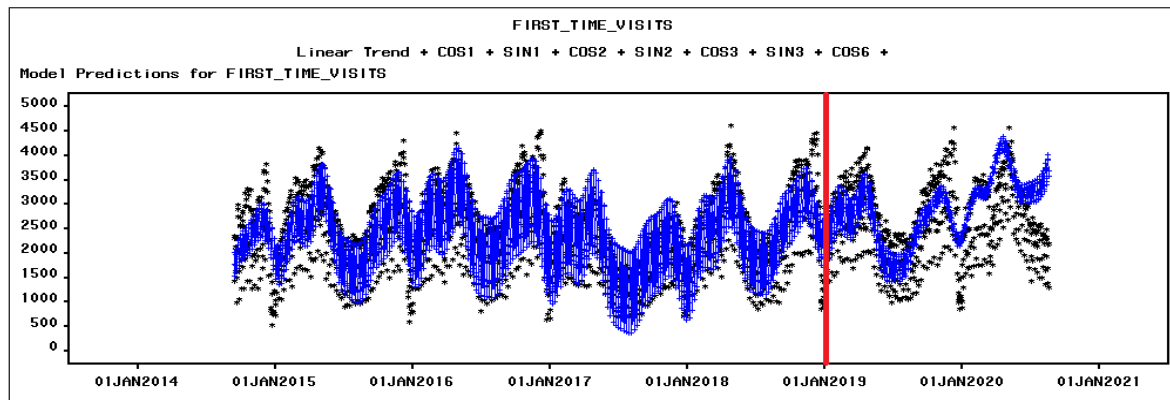


Figure 3-2

Taking the red line in the figure 3-2 as the dividing line, the data before 2019 shows that the cyclical trend model fits the data very well, but after that, the predicted value will have a higher probability of error and will be different than the actual ones.

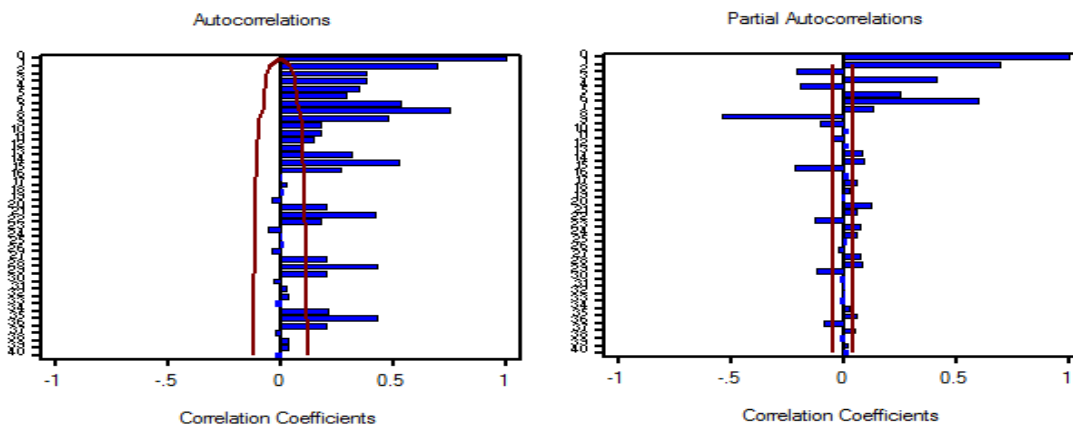


Figure 3-3

From figure 3-3, we can see that most lags are out of 2.s.e error bounds. There are seasonal and non-seasonal parts in both ACF and PACF, and seasonal lags are at 7, 14, 21, and 28. In ACF, series dropped exponentially in both seasonal and non-seasonal lags even though most lags still out of 2.s.e error bounds. In PACF, We will explore more in the cyclical model with ARMA and seasonal difference to find out ways to make the plots look better.

Mean Square Error	930515.4
Root Mean Square Error	964.63229
Mean Absolute Percent Error	36.68834
Mean Absolute Error	777.38584

Figure 3-4

Model Parameter	Estimate	Std. Error	T	Prob> T
Intercept	1458	144.0527	10.1184	<.0001
Linear Trend	0.96997	0.1537	6.3111	<.0001
COS1	316.13249	46.3355	6.8227	<.0001
SIN1	593.28059	92.6195	6.4056	<.0001
COS2	-65.92117	36.4842	-1.8068	0.0719
SIN2	186.12661	32.5491	5.7183	<.0001
COS3	269.42337	25.0717	10.7461	<.0001
SIN3	62.76514	18.3508	3.4203	0.0007
COS6	-277.42246	14.7547	-18.8023	<.0001
SIN6	97.90791	14.9448	6.5513	<.0001
SIN8	-76.62258	15.1262	-5.0655	<.0001
COS8	32.44749	14.5862	2.2245	0.0269
COS12	-39.16259	14.3291	-2.7331	0.0067
SIN12	518.67546	14.3270	36.2026	<.0001
COS24	10.09704	14.2015	0.7110	0.4777
SIN24	-172.54576	14.2325	-12.1234	<.0001
COS30	25.22275	14.2112	1.7749	0.0770
SIN30	162.98162	14.1998	11.4777	<.0001
COS309	-49.09848	14.3475	-3.4221	0.0007
SIN309	388.74725	14.3390	27.1112	<.0001
COS310	-212.89964	14.3427	-14.8438	<.0001
SIN310	-451.58049	14.3439	-31.4823	<.0001
Model Variance (sigma squared)	187385	.	.	.

Figure 3-5

From figure 3-4, MAPE is much higher than the seasonal model we introduced before. The coefficients for cos2, cos24, and cos30 are not statistically significant. For the following part, we combine the cyclical model and ARIMA to get better prediction.

- **Cyclical model + ARIMA(1,0,0) (1,1,0)**

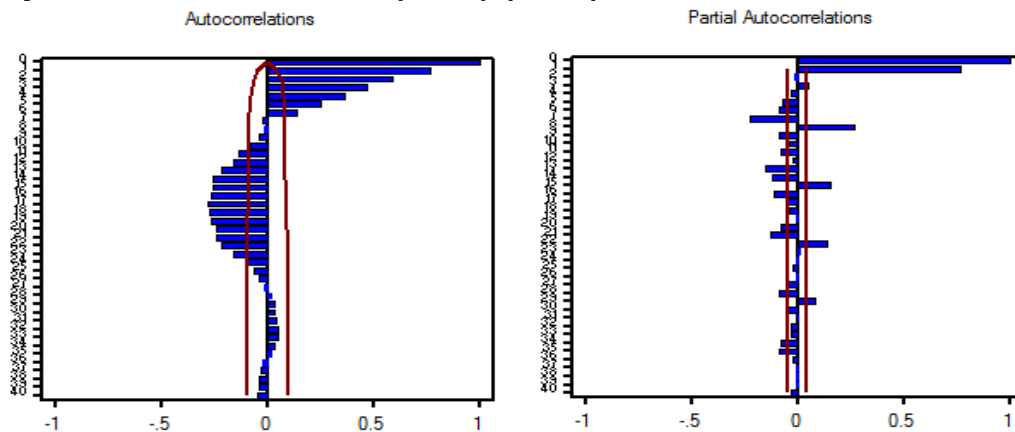


Figure 3-6 Seasonal difference

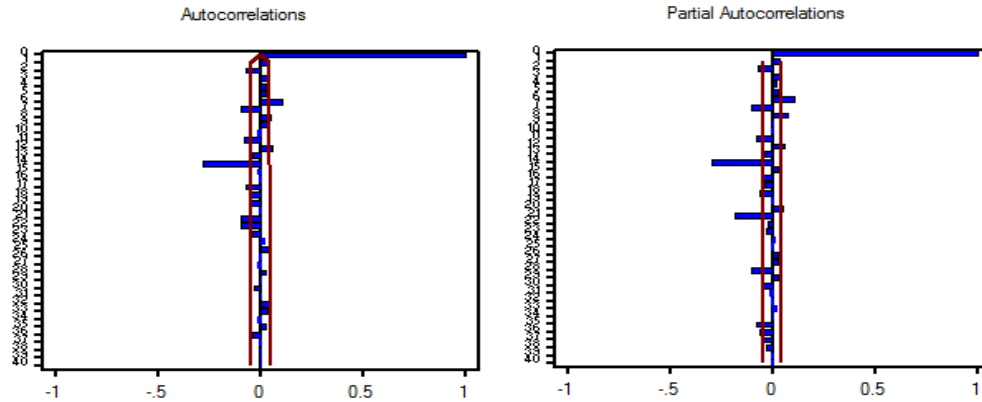


Figure 3-7 ARIMA(1,0,0)(1,1,0)

From figure 3-6, we first added the seasonal difference, and ACF seems to decay exponentially and PACF dropped to 0 after lag 1 and seasonal lags. So we add the first order of the AR process for both parts. From figure 3-7, even though all lags do not perfectly stay within 2.s.e error bounds, they are really close to the bounds except for some seasonal lags. The plot seems better compared with the only cyclical model.

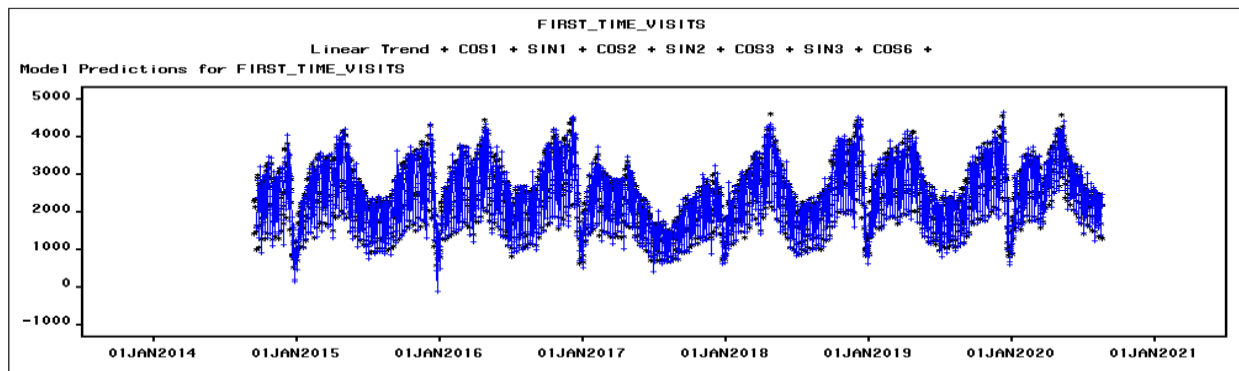


Figure 3-8

From the picture above, data fits the model much better when comparing with the only cyclical model.

Statistic of Fit	Value
Mean Square Error	65811.4
Root Mean Square Error	256.53737
Mean Absolute Percent Error	8.01014
Mean Absolute Error	186.96030
R-Square	0.892

Figure 3-9

MAPE for cyclical + ARIMA(1,0,0)(0,1,0) is 8.01014

MAPE for cyclical model is 36.68834

Model Parameter	Estimate	Std. Error	T	Prob> T
Autoregressive, Lag 1	0.81944	0.0136	60.1158	<.0001
Seasonal Autoregressive, Lag 7	-0.34451	0.0224	-15.3984	<.0001
Linear Trend	0.00561	0.0251	0.2231	0.8236
COS1	111.24727	1411	0.0788	0.9372
SIN1	514.26714	2004	0.2566	0.7977
COS2	-124.60030	784.1196	-0.1589	0.8739
SIN2	152.45492	835.7253	0.1824	0.8554
COS3	241.23967	556.2726	0.4337	0.6649
SIN3	44.37432	480.7733	0.0923	0.9265
COS6	-283.27286	238.0644	-1.1899	0.2351
SIN6	89.22412	236.5683	0.3772	0.7063
SIN8	-83.56441	174.7175	-0.4783	0.6328
COS8	28.57285	173.8566	0.1643	0.8696
COS12	-40.98122	114.4872	-0.3580	0.7206
SIN12	513.92385	114.5569	4.4862	<.0001
COS24	9.71359	55.8318	0.1740	0.8620
SIN24	-174.28067	55.6737	-3.1304	0.0019
COS30	24.21570	43.9941	0.5504	0.5825
SIN30	161.63401	44.1373	3.6621	0.0003
COS309	-34.50999	545.6469	-0.0632	0.9496
SIN309	-33.57953	545.6387	-0.0615	0.9510
COS310	23.02642	726.2441	0.0317	0.9747
SIN310	37.59823	726.1861	0.0518	0.9587
Model Variance (sigma squared)	42437	.	.	.

Figure 3-10

From figure 3-10, we can see a huge problem that except for the coefficient of autoregressive at lag 1, all other coefficients of sin and cos pairs are not statistically significant. Even though we got lower MAPE and model variance, the higher p-value indicates that their coefficients are equal to 0 and only the ARIMA matters. According to this, no matter which model we look at, the only cyclical model or the cyclical model with ARIMA(1,0,0)(1,1,0), the cyclical model doesn't fit the data very well if we want to make more accurate predictions.

- Error model

Forecast horizon: 1, Hold-out Sample: 300

Based on the attribute of the model, we thought it will be a good way to discover the dataset by firstly model it with seasonal dummies, and then see how its autocorrelation distributes.

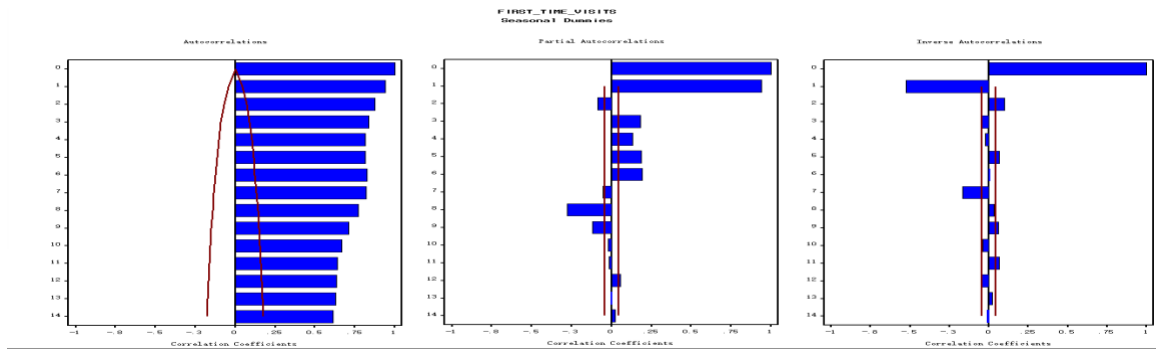


Figure 4-1

The Figure 4-1 is the autocorrelation plot of the seasonal dummies model, for the first ACF graph, we can see it didn't decay exponentially. However, when moving to PACF and IACF, we can see the performance of them are becoming better, especially for PACF, the seasonal lags are chopped off after 2 and even they are perfectly staying inside the bounds but are very close to it; since the complexity and huge amount of data, we think it would be acceptable. Based on their overall performance, we decided to use seasonal dummies combined with AR(1) for our next model.

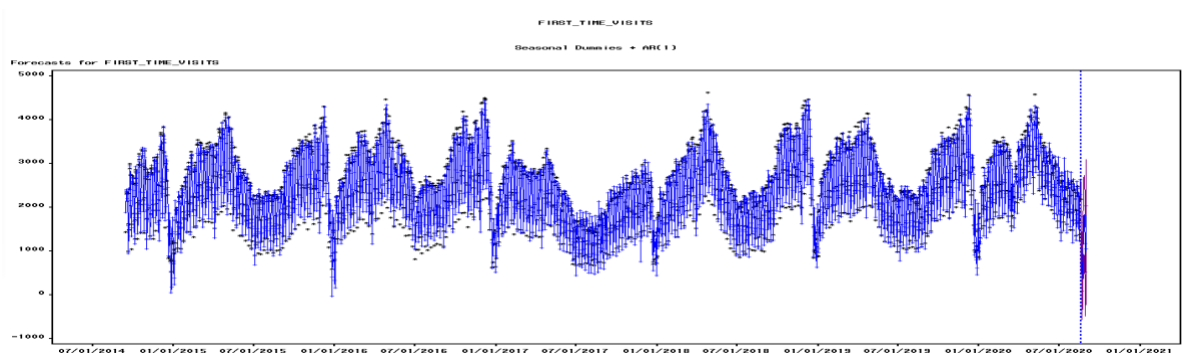


Figure 4-2

The estimated Error model of first time looks fit well to the data

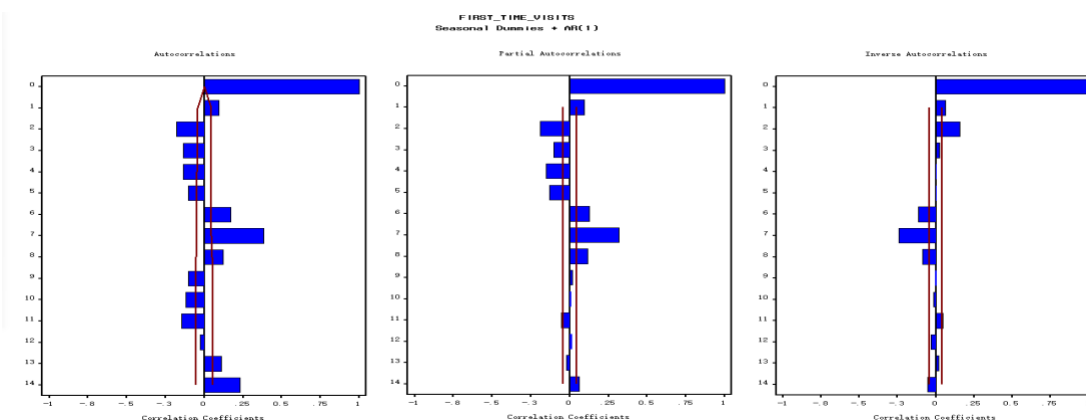


Figure 4-3

The Figure 4-3 demonstrates that the distribution of ACF, PACF, and IACF, we can see the performance of them are becoming better than the seasonal dummies model, even some of lags do not perfectly stay inside the bonds, but most of them are very close to.

Parameter Estimate:

FIRST_TIME_VISITS Seasonal Dummies + AR(1)				
Model Parameter	Estimate	Std. Error	T	Prob> T
Intercept	1414	85.3015	16.5808	<.0001
Autoregressive, Lag 1	0.94421	0.0076	123.8413	<.0001
Seasonal Dummy 1	476.58189	12.0249	39.6331	<.0001
Seasonal Dummy 2	1410	15.5033	90.9289	<.0001
Seasonal Dummy 3	1486	16.9762	87.5386	<.0001
Seasonal Dummy 4	1451	16.9786	85.4732	<.0001
Seasonal Dummy 5	1287	15.5112	82.9946	<.0001
Seasonal Dummy 6	700.19571	12.0249	58.2287	<.0001
Model Variance (sigma squared)	42531	.	.	.

Figure 4-4

All of the Seasonal Dummy are statistically significant since their p-values are smaller than 0.05. The model variance is 42531, it shows that our model has a very good estimate of parameter (Figure 4-4)

Statistics of Fit:

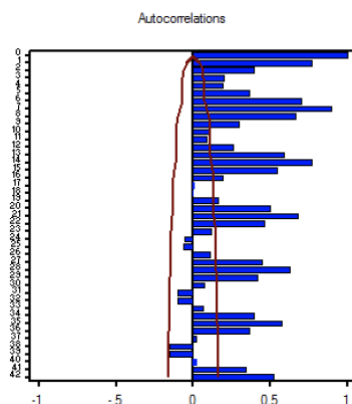
FIRST_TIME_VISITS Seasonal Dummies + AR(1)	
Statistic of Fit	Value
Mean Square Error	60656.0
Root Mean Square Error	246.28434
Mean Absolute Percent Error	8.27386
Mean Absolute Error	187.15137

Figure 4-5

According to the statistic of fit in the figure 4-5, we can see the mean absolute error is equal to 8.27386 which is relatively small, which also provides the reliability of our error model.

2.2 ARIMA models (with seasonal ARIMA components if relevant)

Forecast horizon: 1, Hold-out Sample: 300



Regarding the stationarity of the original series (Figure5-1), The ACF of the series presents that the autocorrelations of the seasonal lags (weekly) are decaying slowly and non-seasonal lags can be interpreted as being decaying exponentially but the seasonal and non-seasonal lags are not explicitly distinguishable.

So, we can first take a look at the seasonal difference model:

Figure 5-1

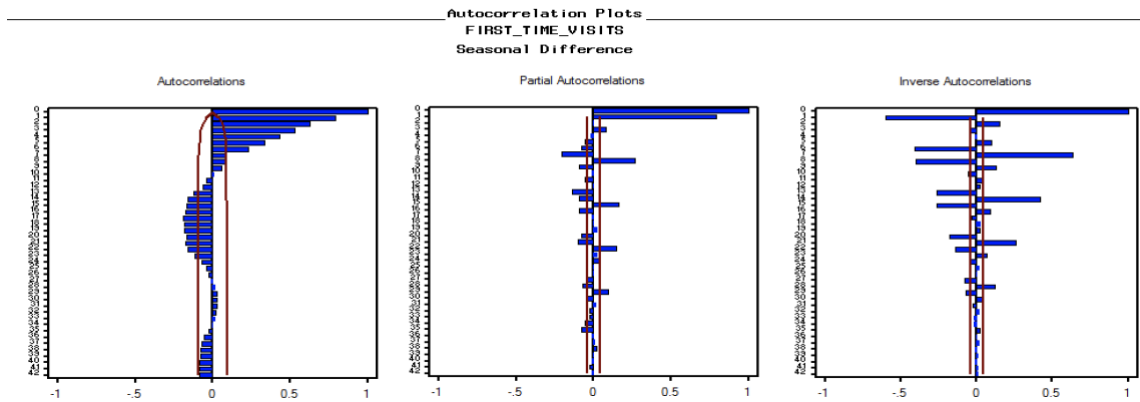


Figure 5-2

The first picture in the figure 5-2 is the ACF of the seasonal difference model. It seems to be decaying exponentially but the sinusoidal ACF contains a number of autocorrelations that are not significantly equal to 0. In addition, the PACF (2nd picture) consists of the seasonal lags decaying slowly and lots of non-seasonal lags whose PACs are not 0. Based on these features, we can decide to take a non-seasonal first-differencing process with the seasonal difference.

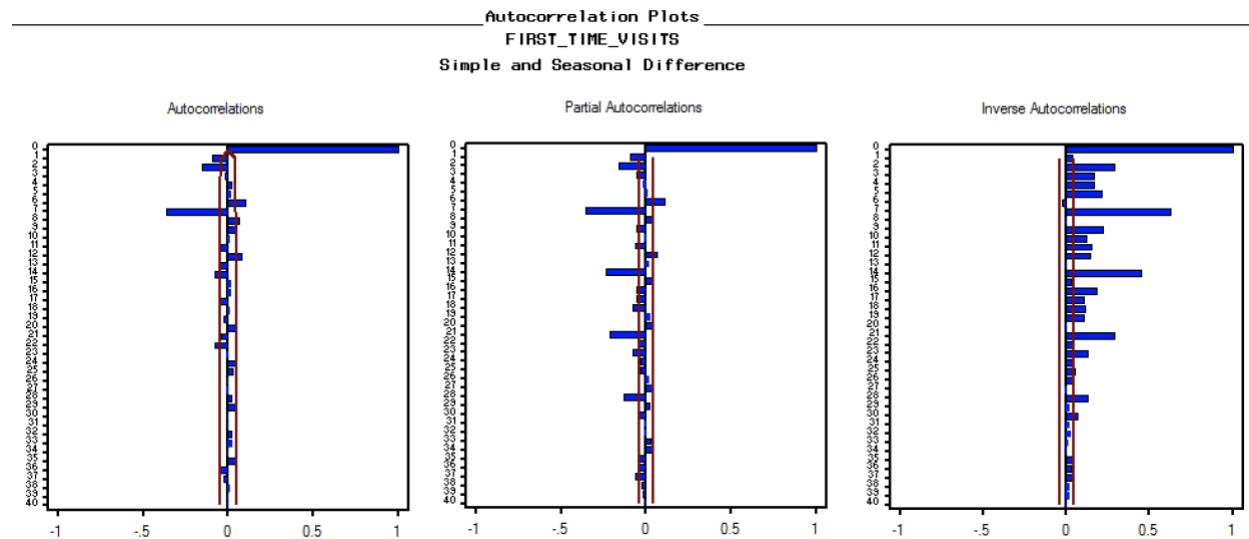


Figure5-3

Now, the ACF, PACF, and IACF in the Figure5-3 look more interpretable.

Non-seasonal lags: the ACF and PACF (the 1st and 2nd pictures in Figure5-3) are chopped off to 0 after lag=2. Even though the autocorrelations and partial autocorrelations of several non-seasonal lags are different than 0, those are not considerably far off from bounds. In the non-seasonal part, the IACF (the 3rd picture in Figure5-3) slowly decays which implies that the non-seasonal part is more likely to be an MA process.

Seasonal lags: after the first seasonal lag (lag = 7), the autocorrelations are chopped off to 0. In the PACF, we can see that the PACs of the seasonal lags are slowly decaying with 7-days terms. The IACF at the seasonal lags are also decaying slowly. The overall shapes of the ACF,

PACF, and IACF are closer to the MA model in spite of the fact that the speed of the PACF decaying towards 0 does not perfectly satisfy the general characteristics for the MA process.

- **Expected Model: ARIMA(0,1,2)(0,1,1)_s**

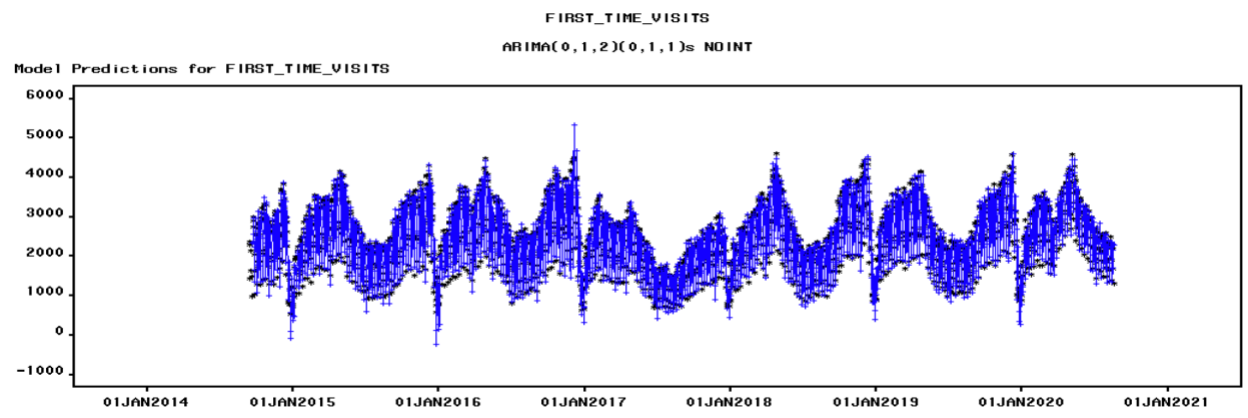


Figure5-4

The estimated ARIMA model seems to fit well the data (Figure5-4).

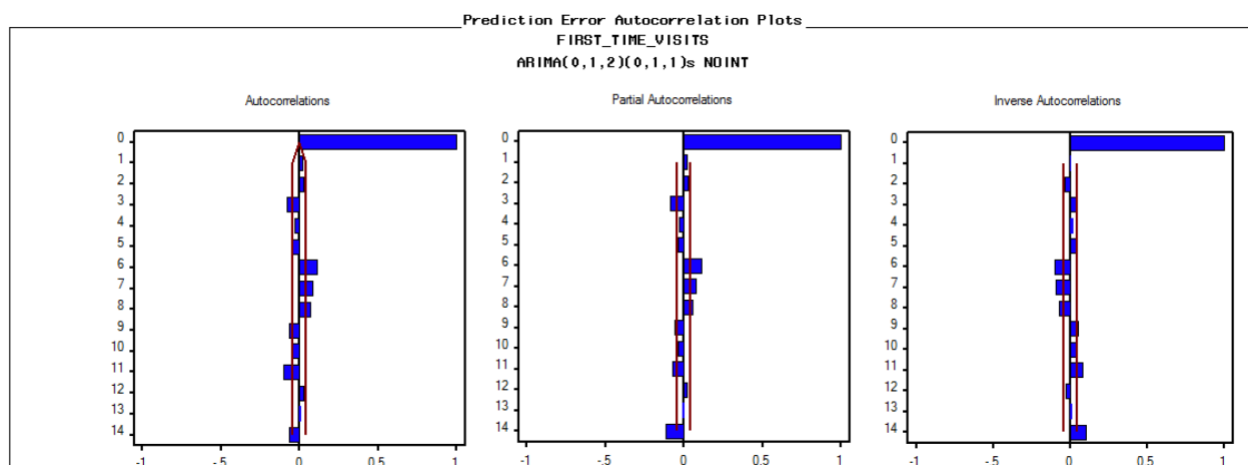


Figure5-5

The Figure5-5 demonstrates that the ACF, PACF, and IACF are not perfectly staying inside the bounds but these are closely wandering around the bounds.

FIRST_TIME_VISITS Seasonal Dummies + AR(1)	
Statistic of Fit	Value
Mean Square Error	60656.0
Root Mean Square Error	246.28434
Mean Absolute Percent Error	8.27386
Mean Absolute Error	187.15137
R-Square	0.900

Figure5-6

The two non-seasonal MA coefficients and the seasonal MA coefficient are all statistically significant with the p-values smaller than 0.05.

Statistic of Fit	Value
Mean Square Error	59655.6
Root Mean Square Error	244.24496
Mean Absolute Percent Error	7.79166
Mean Absolute Error	178.12891

Figure5-7

Lastly, the values in the Figure5-7 provide several measures for the goodness-of-fit of the ARIMA(0,1,2)(0,1,1)s model.

2.3 Comparison of models (in terms of fit and validation)

MODEL	Model Variance	MAPE	MAE	P-values
Seasonal Dummies	43729	8.33781	189.28111	Every coefficient is statistically significant
Cyclical Model	468228	36.688	777.386	Most coefficient is statistically significant except cos2, cos24, cos30
Cyclical + ARIMA (1,0,0)(0,1,0)	42437	8.01014	201.8778	Only autoregressive at lag 1 is statistically significant.
Error	42531	8.27386	187.15137	Every dummy is significant
ARIMA	40615	7.79166	178.12891	Every coefficient is statistically significant

< 3. Multivariate Time Series Models >

3.1 Regression model and analysis of regression residuals

Forecast horizon: 1, Hold-out Sample: 300

Dependent variable: First.Time.Visit

Predictors: Returning.Visits + Page.Loads

Parameter Estimates				
FIRST_TIME_VISITS				
Page_Loads + Returning_Visits				
Model Parameter	Estimate	Std. Error	T	Prob> T
Intercept	-25.37859	8.8811	-2.8576	0.0046
Page_Loads	0.67582	0.0050	134.9625	<.0001
Returning_Visits	-0.70254	0.0397	-17.7082	<.0001
Model Variance (sigma squared)	14239	.	.	.

Figure 6-1

- The multivariate regression model with First.Time.Visits as the dependent variable and the two independent variables produces the statistically significant coefficients across all the variables.
- The model variance is 14239 which should be compared based on relative criteria for the model's goodness-of-fit. In this regard, the residuals can provide the clue for building an improved model comparable to this model in terms of the model fit performance.

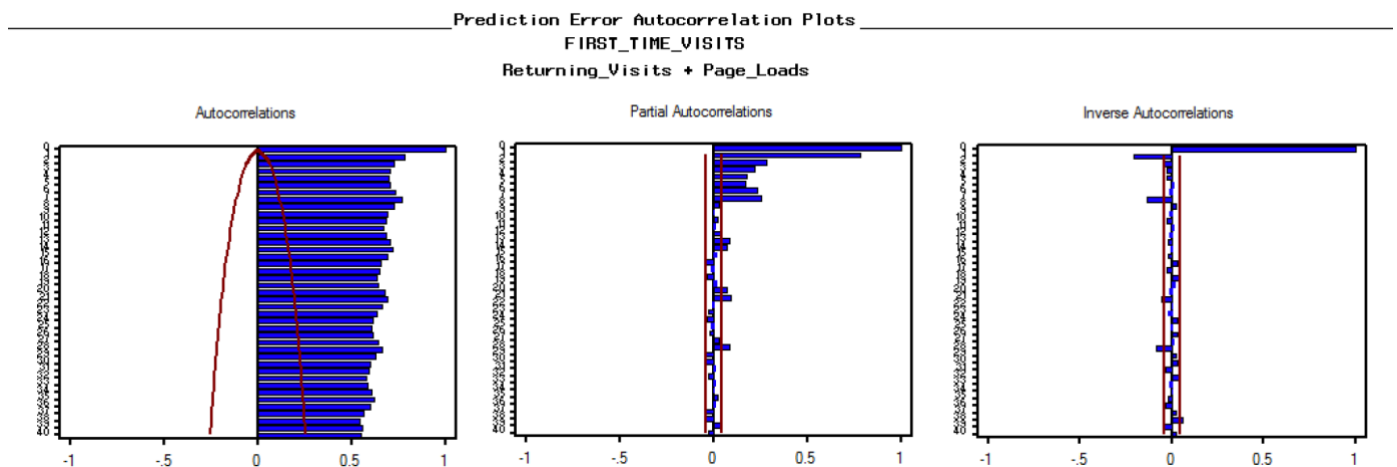


Figure 6-2

- The residual ACF (1st picture in the Figure 6-2) of the regression model can be separated into two parts one of which is a non-seasonal part and the other of which is a seasonal part with seven days terms. Both parts are decaying very slowly, which implies that the seasonal and non-seasonal parts are non-stationary. In order for the series to become stationary, differencing is requisite for both seasonal and non-seasonal lags.

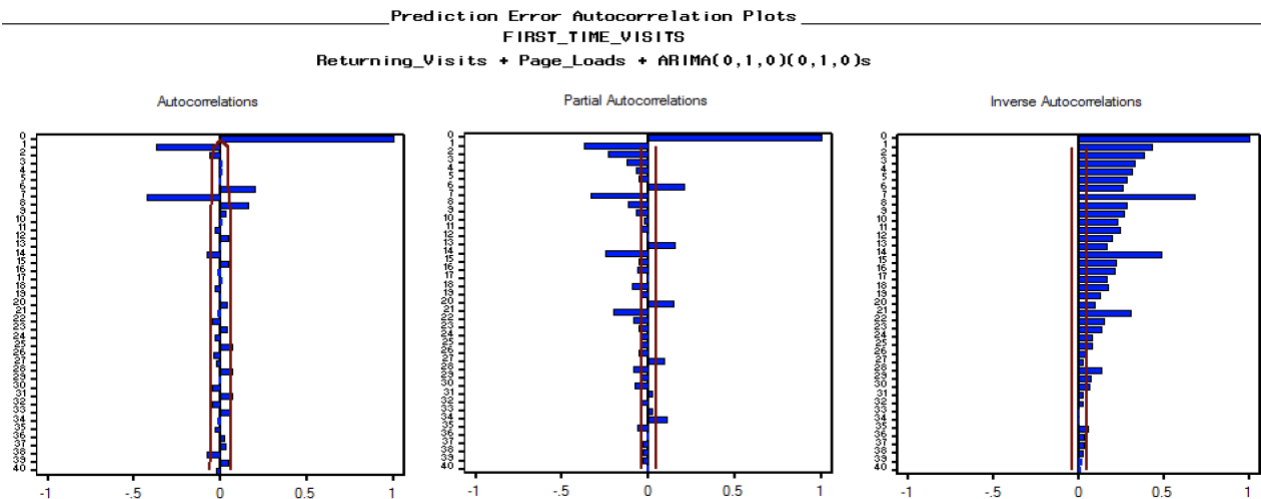


Figure 6-3

- The residual ACF of the edited series (Figure 6-3) is now stationary at the seasonal and non-seasonal lags.
- The residual PACF (2nd picture of Figure 6-3) explicitly shows that the seasonal and non-seasonal parts are decaying exponentially to zero as lags increase.
- The residual IACF (3rd picture of Figure 6-3) is decaying slowly at seasonal and non-seasonal lags. However, the IACF in this case does not necessarily have to be considered because the shapes of the ACF and PACF are interpretable enough to identify the appropriate forecasting model.

3.2 Error model using regression residuals

According to the residual ACF in the Figure 6-3, the series adjusted to the seasonal and non-seasonal differences generates the seasonal residuals whose autocorrelations are chopped off to zero after the first seasonal lag (lag=7) and the non-seasonal residuals whose autocorrelations also drop to zero immediately after the first lag (lag=1). In light of the exponentially decaying PACF, the possible best way of improving the regression model should be to build **ARIMA(0,1,1)(0,1,1)s without intercept**.

Parameter Estimates				
FIRST_TIME_VISITS				
Returning_Visits + Page_Loads + ARIMA(0,1,1)(0,1,1)s NOINT				
Model Parameter	Estimate	Std. Error	T	Prob> T
Moving Average, Lag 1	0.78248	0.0150	52.2656	<.0001
Seasonal Moving Average, Lag 7	0.92930	0.0091	102.2179	<.0001
Returning_Visits	-0.34615	0.0589	-5.8728	<.0001
Page_Loads	0.62389	0.0075	83.3729	<.0001
Model Variance (sigma squared)	6062	.	.	.

Figure 7-1

- According to the p-values appearing in Figure7-1, the MA coefficients of both seasonal and non-seasonal parts are statistically significant. The coefficients of the two predictors are also statistically significant.
- The difference of Returning_Visits has the negative relationship with the First_Time_Visits and the difference of Page_Loads is positively correlated with the First_Time_Visits. In other words, as the difference of the returning visitors increases, we can find the number of the first time visitors decreases. In addition, the greater difference of pages loaded on the daily basis tends to attract more visitors who have not yet explored the website.
- The error model presents the even better performance of the model fit. The model variance has been improved from 14239 to 6062 which is more than half.

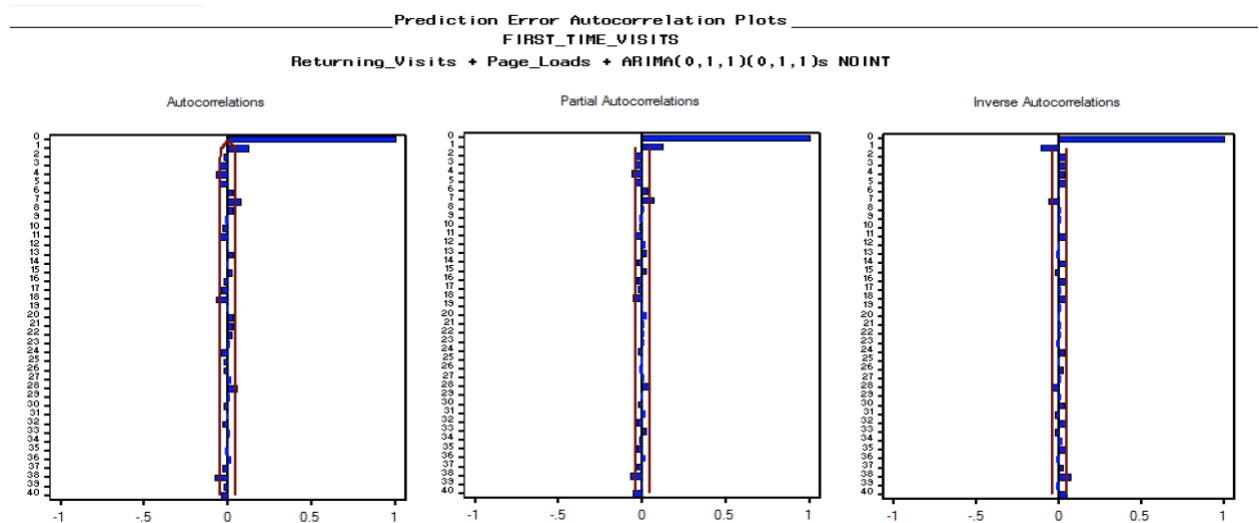


Figure 7-2

- The residual ACF of the **ARIMA(0,1,1)(0,1,1)s** model (Figure 7-2) presents that it barely has significant autocorrelations at both seasonal and non-seasonal lags. In particular, the autocorrelations at lag1 and lag 7 which were large have become significantly small now.
- Several lags at which the autocorrelations are observed to be located outside the bounds result from the large number of observations included in the model.

Statistics of Fit	
FIRST_TIME_VISITS	
Returning_Visits + Page_Loads + ARIMA(0,1,1)(0,1,1)s NOINT	
Statistic of Fit	Value
Mean Square Error	6071.0
Root Mean Square Error	77.91632
Mean Absolute Percent Error	2.46197
Mean Absolute Error	60.37272

Figure 7-3

- Figure 7-3 shows several values to measure the model's performance when it comes to its estimation. The ARIMA(0,1,1)(0,1,1)s model can be considered to be performing well on the basis of its small MAPE.

3.3 Cross correlation analysis to identify lagged values of predictors and use them as predictors in the model.

In order to identify the appropriate multivariate regression model, we need to find the relationships between a dependent variable and independent variables by different lags. We can find the relationships by looking at the Cross-Correlation Function, which will help us determine the predictors' lags correlated to the dependent variable that should be incorporated into the model.

Prior to creating the CCF, we need to make sure that the dependent and the two independent variables are stationary:

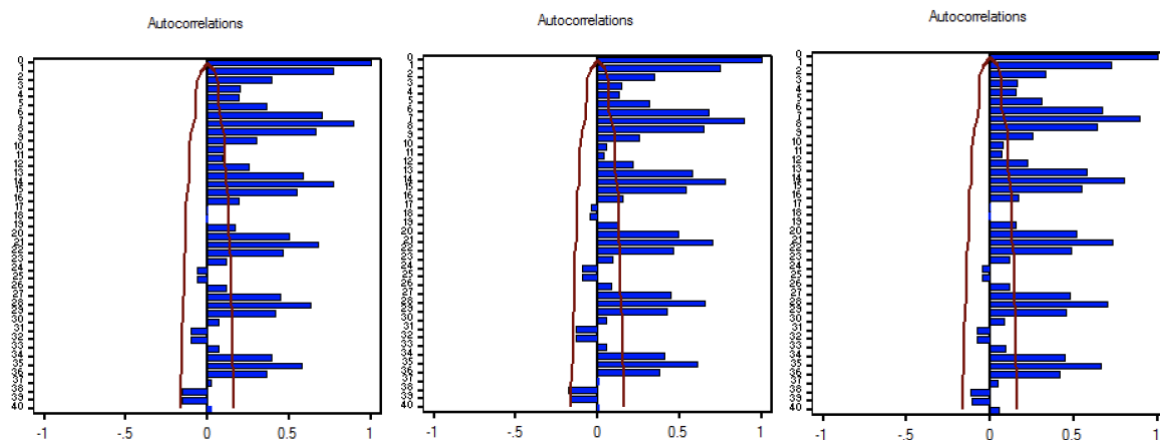


Figure 8-1

The ACFs in the figure 8-1 correspond to the First.Time.Visits, Page.Loads, and Returning.Visits in order. Since all the three ACFs have slowly decaying autocorrelations at both seasonal and non-seasonal lags (7-days term), we need to take seasonal and non-seasonal differencing to have them stationary.

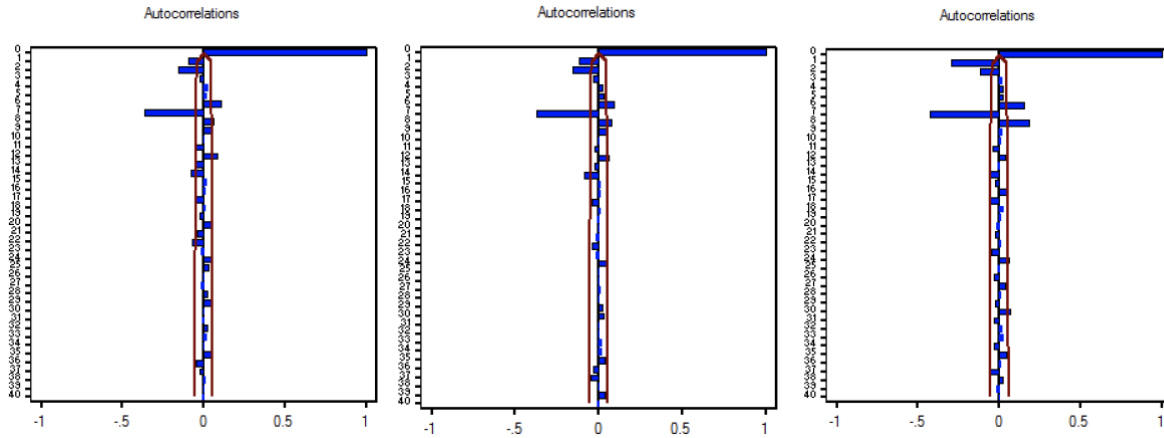


Figure 8-2

After taking seasonal and non-seasonal differencing, we obtained the stationary series (Figure 8-2). Now, we are ready to fit the cross correlation functions with Page.Loads and Returning.Visits respectively.

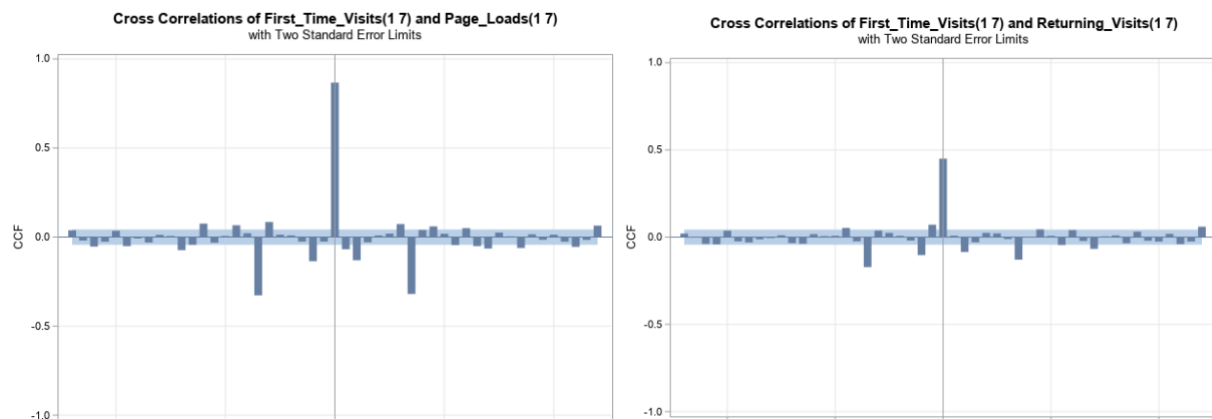


Figure 8-3

- We're only focusing on the positive lags to figure out the influences that the independent variables at the present or previous lags have on the dependent variable.
- PAGE_LOADS at lag 0 is positively correlated with the FIRST_TIME_VISITS, which means if PAGE_LOADS increases at period t , FIRST_TIME_VISITS would be positively affected. In addition, PAGE_LOADS at lag 2 and 7 are negatively correlated with the dependent variable, which means if PAGE_LOADS increases at period $t-2$ and $t-7$, FIRST_TIME_VISITS would be negatively affected.
- Likewise, RETURNING_VISITS at lag 0 is positively correlated with the FIRST_TIME_VISITS, which means if RETURNING_VISITS increases at period t , FIRST_TIME_VISITS would be positively affected; and RETURNING_VISITS at lag 2 and 7 are negatively correlated, which means if RETURNING_VISITS

increases at period t-2 and t-7, FIRST_TIME_VISITS would be negatively affected.

- As a result, the predictors that should be included in the regression model are “PAGE_LOADS(t), PAGE_LOADS(t-2), PAGE_LOADS(t-7), RETURNING_VISITS(t), RETURNING_VISITS(t-2), RETURNING_VISITS(t-7)”.

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Variable	Shift
MA1,1	0.24511	0.01368	17.91	<.0001	1	First_Time_Visits	0
MA1,2	0.71426	0.01349	52.96	<.0001	7	First_Time_Visits	0
NUM1	0.57779	0.0076259	75.77	<.0001	0	Page_Loads	0
NUM2	0.02289	0.0075377	3.04	0.0024	0	LAG1PAGE	0
NUM3	0.03104	0.0073359	4.23	<.0001	0	LAG7PAGE	0
NUM4	-0.42619	0.05633	-7.57	<.0001	0	Returning_Visits	0
NUM5	0.04449	0.05596	0.79	0.4266	0	LAG1RETURN	0
NUM6	-0.0086612	0.05374	-0.16	0.8720	0	LAG7RETURN	0
				Variance Estimate	7975.918		

Figure 8-4

- When building the MA(0,1,1)(0,1,1) error regression model with the six predictors found above, the p-values of the variables “RETURNING_VISITS(1)” and “RETURNING_VISITS(7)” are greater than 0.05 according to the figure 8-4. So, these insignificant variables should be deleted.
- The variance of the model is 7975.918 which is greater than the one of the regression model without lagged variables.

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Variable	Shift
MA1,1	0.23947	0.01361	17.59	<.0001	1	First_Time_Visits	0
MA1,2	0.71870	0.01344	53.48	<.0001	7	First_Time_Visits	0
NUM1	0.57896	0.0073715	78.54	<.0001	0	Page_Loads	0
NUM2	0.02731	0.0052255	5.23	<.0001	0	LAG1PAGE	0
NUM3	0.03017	0.0052854	5.71	<.0001	0	LAG7PAGE	0
NUM4	-0.43920	0.05325	-8.25	<.0001	0	Returning_Visits	0
				Variance Estimate	7970.567		

Figure 8-5

- After removing the insignificant lagged variables, we obtained the predictors all of which p-values are smaller than 0.05. In addition, the variance of the model has become slightly smaller than 7975.918 but it is still bigger than the one of the regression model without the lagged values.

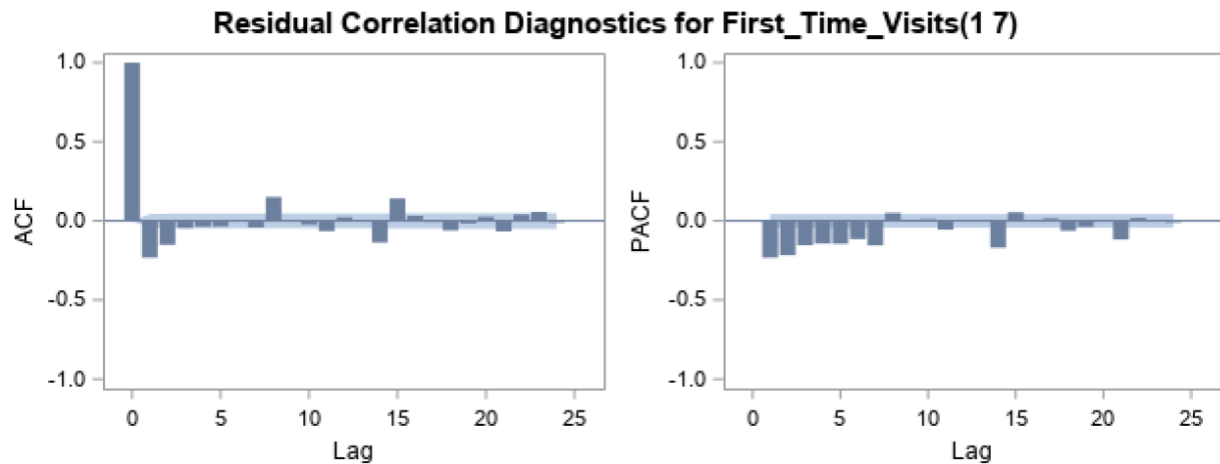


Figure 8-6

- The figure 8-6 is the final model's residual ACF. This shows pretty small autocorrelations and partial autocorrelations at the seasonal and non-seasonal lags. Accordingly, the error term of the model can be regarded to be close to white noise but not perfect due to several lags slightly exceeding the bounds.

< 4. Conclusions >

Based on the above models we can conclude that:

- Univariate Models

ARIMA(0,1,2)(0,1,1) is the best model, which results in the lowest model variance, MAE and MAPE, and all the coefficients are statistically significant.

- Multivariate Models

Page-loads and Returning-visit are two good predictors, and have the positive and negative relationship with First-time-visit (our objective) respectively; and ARIMA(0,1,2)(0,1,1) model is the best model among all models we tried cause it has the best parameter performance. In Cross Correlation Analysis, we determine the predictors' lags correlated to the dependent variable that should be incorporated into the model, which are "PAGE_LOADS(t), PAGE_LOADS(t-2), PAGE_LOADS(t-7), RETURNING_VISITS(t), RETURNING_VISITS(t-2), RETURNING_VISITS(t-7)".