INDR 372 Homework 3 Part 2 Report

Umur Berkay Karakaş
69075

a) I have reduced the model in Part 1 to 4 predictors using Lasso regression, the implementation can be seen in the code.

b) Four predictors that I obtained using Lasso are Monday dummy, Saturday dummy, Sunday dummy and seven-day lag. Initially, I put all predictors to the Lasso regression model. I put not 6 but 7 dummies for the days of the week because even though only six of them are independent, some days could have been more significant than the others, therefore I put a dummy for all of them and the model would keep the significant ones in the model. To find a value for alpha that keeps only 4 predictors in the model, I tried different alphas by hand, and I came up with a value of 100 for alpha. Also, the model wasn't accepting NaN values in the training and until 7th data point, I had NaN values because of seven-day lag predictor, therefore I started my training set from 7th index.

c) As it can be observed from Figure 1, the regression is significant as F-test has a P-value of 2.43e-117 which is less than 1% and 5%. R-squared is 0.891 and adjusted R-squared is 0.889. Both values are high, and it appears to be that my predictors are good at predicting the data.

d)

```
RMSE of reduced model on training set: 6520.0963
MAPE of reduced model on training set: 0.0922
MAE of reduced model on training set: 4488.0725
```

e) From Figure 1, it can be observed that each predictor is significant since their P-values are less than 5%.

f)

```
RMSE of reduced model on test set: 16702.3141
MAPE of reduced model on test set: 0.2244
MAE of reduced model on test set: 10471.7548
```

RMSE is almost 2.5 times of the RMSE for training set. It is not desirable, but it is acceptable since MAPE is 0.22.

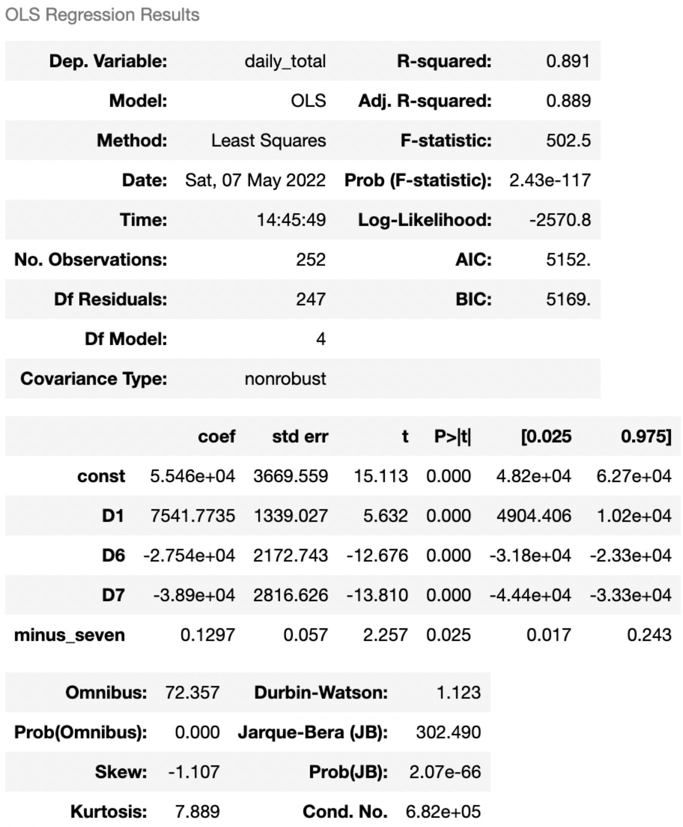
g) In part 1, the RMSE of full model on training set and test set were 5352.27704 and 16966.24709, respectively. In part 2, they are 6520.0963 and 16702.3141. Full model has less RMSE on training set as expected but reduced model has less RMSE on test set, which clearly shows that there is an overfitting in full model.

Also in part 1, R-squared and adjusted R-squared were 0.926 and 0.924 in full model. In part 2, they are 0.891 and 0.889, which are still high and very close to those for full model. We can infer that 4 predictors explain the variability of the model as almost good as full model.

Considering my discussion above, my suggestion to call center manager would be to use the reduced model on their predictions. In full model, there is clearly an overfitting issue and reduced model is better at predicting the test set than the full model. It would be inefficient to use the full model which has 6 more predictors than the reduced model to get worse results than a model only with 4 predictors.

Appendix

Figure 1: Summary of the reduced regression model on training set



OLS Regression Results

Dep. Variable:	daily_total	**R-squared:**	0.891
Model:	OLS	**Adj. R-squared:**	0.889
Method:	Least Squares	**F-statistic:**	502.5
Date:	Sat, 07 May 2022	**Prob (F-statistic):**	2.43e-117
Time:	14:45:49	**Log-Likelihood:**	-2570.8
No. Observations:	252	**AIC:**	5152.
Df Residuals:	247	**BIC:**	5169.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	5.546e+04	3669.559	15.113	0.000	4.82e+04	6.27e+04
D1	7541.7735	1339.027	5.632	0.000	4904.406	1.02e+04
D6	-2.754e+04	2172.743	-12.676	0.000	-3.18e+04	-2.33e+04
D7	-3.89e+04	2816.626	-13.810	0.000	-4.44e+04	-3.33e+04
minus_seven	0.1297	0.057	2.257	0.025	0.017	0.243

Omnibus:	72.357	**Durbin-Watson:**	1.123
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	302.490
Skew:	-1.107	**Prob(JB):**	2.07e-66
Kurtosis:	7.889	**Cond. No.**	6.82e+05