Supply Chain HW I3

Brett Scroggins

Due: 11/29/2017

Question 1

```
x_train <- xm[1:94,]
y_train <- y[1:94]
x_test <- xm[95:104,]
y_test <- y[95:104]


set.seed(1)
grid = 10^seq(10,-2,length = 100)
cv_lasso = cv.glmnet(x_train, y_train, alpha=1, lambda=grid,
                     thresh = 1e-12,nfolds=10, type.measure = "mse")

min_lambda = cv_lasso$lambda.min
lasso_model = glmnet(x_train, y_train, alpha=1,
                     lambda = min_lambda, thresh=1e-12)

predict(lasso_model, s=min_lambda, type="coefficients")
```

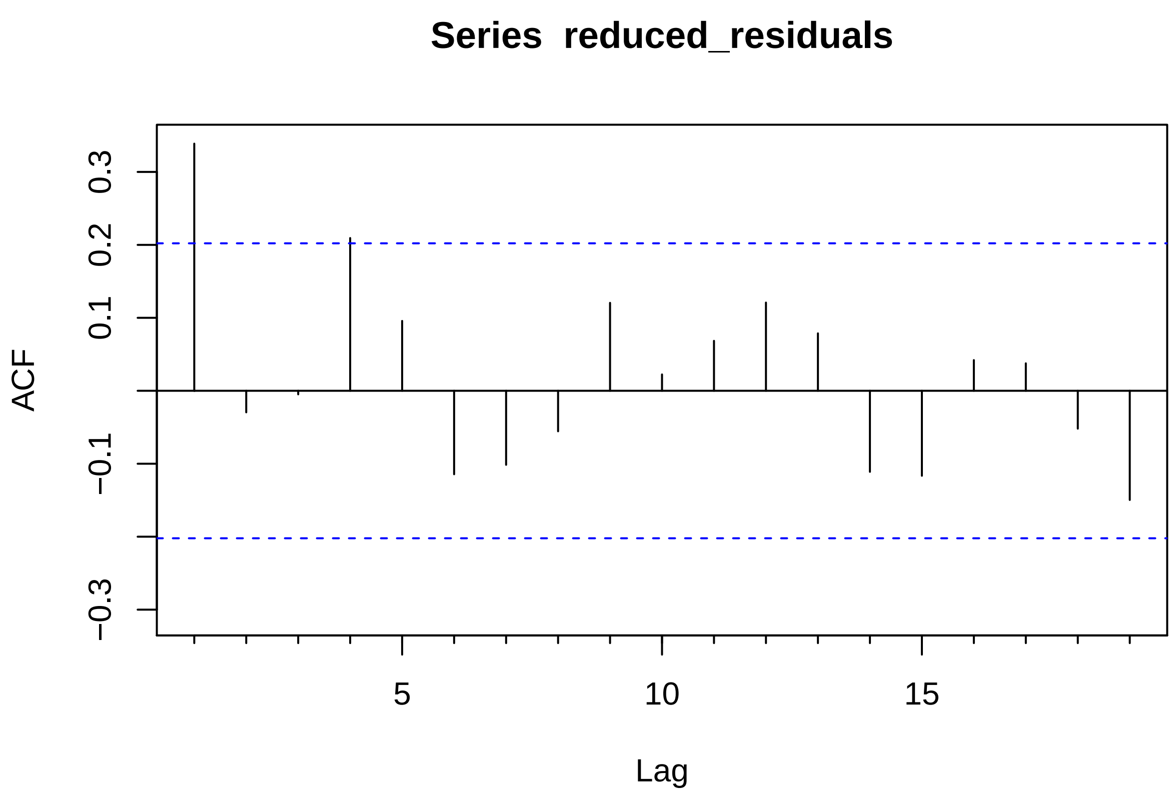
```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                      1
## (Intercept)  6.7524532
## S.WF_LSA     .
## S.WF_MSA     .
## S.WF_SSA     .
## S.WD_MIN     0.4758104
## S.WD_MAJ     .
## A.PPU       -2.4165558
## S.WF_LSA.1   .
## S.WF_MSA.1   .
## S.WF_SSA.1   .
## S.WD_MIN.1   .
## S.WD_MAJ.1   .
## A.PPU.1      .
## S.WF_LSA.2   .
## S.WF_MSA.2   .
## S.WF_SSA.2   .
## S.WD_MIN.2   .
## S.WD_MAJ.2   .
## A.PPU.2      .
```

Question 2

```r
reduced_x_train = data.frame(x_train[,c('S.WD_MIN', 'A.PPU')])
reduced_x_test = data.frame(x_test[,c('S.WD_MIN', 'A.PPU')])
reduced_model = lm(y_train~., data = reduced_x_train)
summary(reduced_model)
```

```
##
## Call:
## lm(formula = y_train ~ ., data = reduced_x_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87268 -0.19213  0.00467  0.19773  0.89398
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.8563     0.1248  54.950  < 2e-16 ***
## S.WD_MIN      0.5741     0.0947   6.063 2.99e-08 ***
## A.PPU        -2.5493     0.1279 -19.937  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3202 on 91 degrees of freedom
## Multiple R-squared:  0.8423, Adjusted R-squared:  0.8388
## F-statistic:   243 on 2 and 91 DF,  p-value: < 2.2e-16
```

```r
reduced_residuals = residuals(reduced_model)
Acf(reduced_residuals)
```



Comments on the fit of the model and examine the auto-correlations of the residuals of this model.

Solution

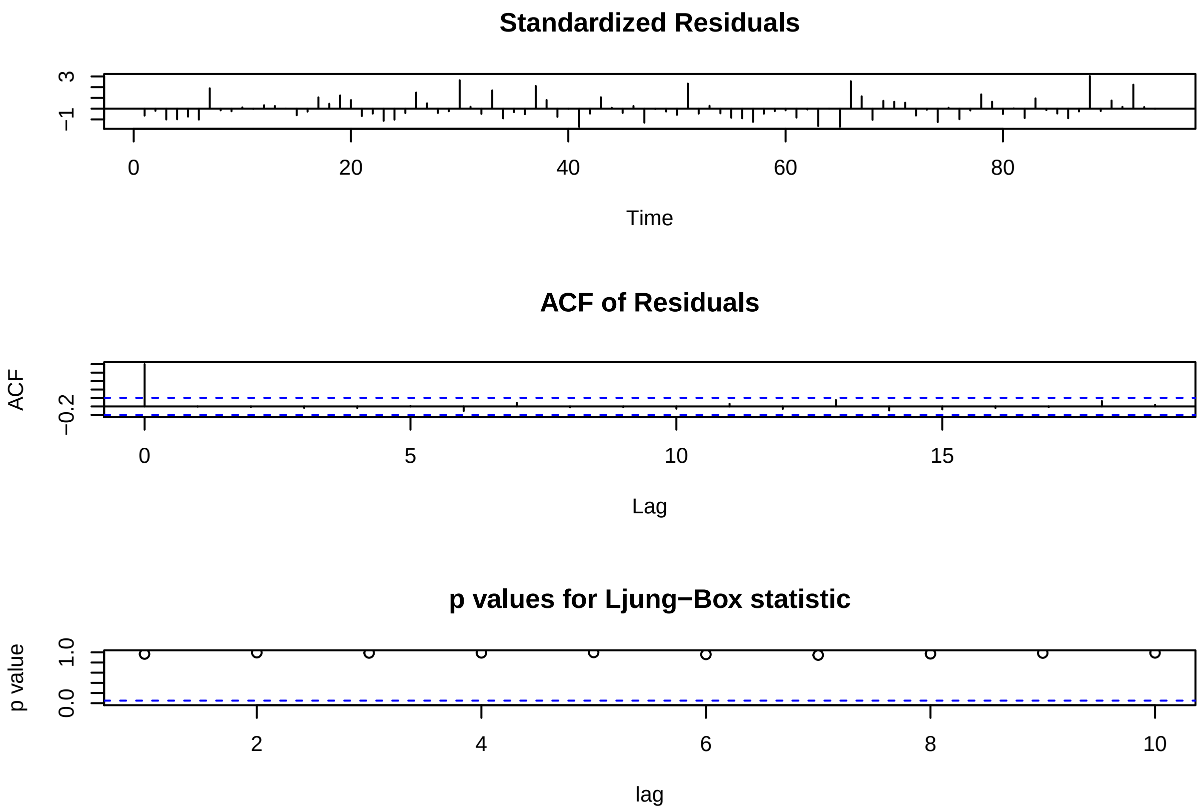
After utilizing the LASSO to reduce variables, the model fit was done and fits the data well. There are no significant auto-correlations past the expected zero time residual, and therefore shows that it is an effective model. This model output an 84% fit and had significant variables.

Question 3

```r
auto_arima = auto.arima(y_train)
summary(auto_arima)
```

```
## Series: y_train
## ARIMA(2,0,1) with non-zero mean
##
## Coefficients:
##          ar1      ar2     ma1    mean
##       0.0363  -0.3600  0.5797  4.6322
## s.e.  0.1711   0.1271  0.1818  0.0778
##
## sigma^2 estimated as 0.4153:  log likelihood=-90.51
## AIC=191.02   AICc=191.7   BIC=203.73
##
## Training set error measures:
##                         ME      RMSE      MAE       MPE     MAPE     MASE
## Training set -0.0001286775 0.6305537 0.474108 -1.696712 10.05719 0.726594
##                     ACF1
## Training set -0.004242972
```

```r
tsdiag(auto_arima)
```



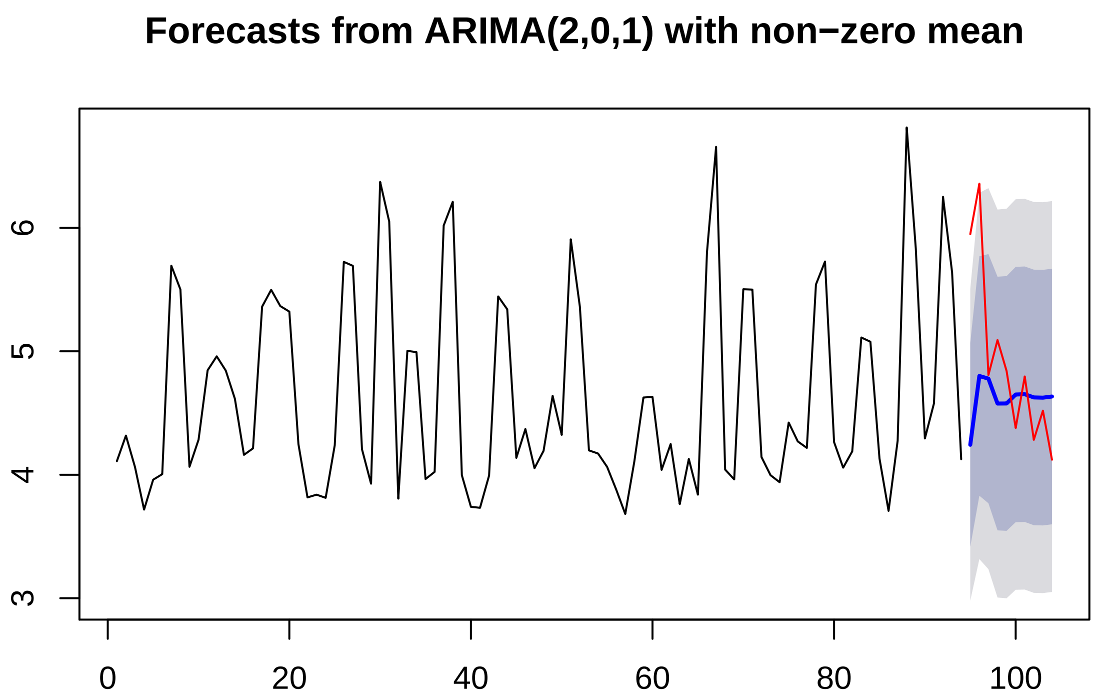
Comments on the model's validity.

Solution

In this model, the data is fit well and outputs an MASE under one (0.726) and therefore outperforms a naive forecast. This yields that this is a good predictor to use as a forecast going foreward.

Question 4

```
aa_forecast = forecast(auto_arima, h=10)
plot(aa_forecast)
lines(x = seq(95,104), y = y_test, col='red')
```



Comments on the usefulness of this model in terms of precision and confidence interval.

Solution

What can be seen by this diagram is that this is not a great predictor of our testing set. This could be from the fact that the testing set is a small sample, but going foreward this model would need to be adjusted prior to putting it into use.

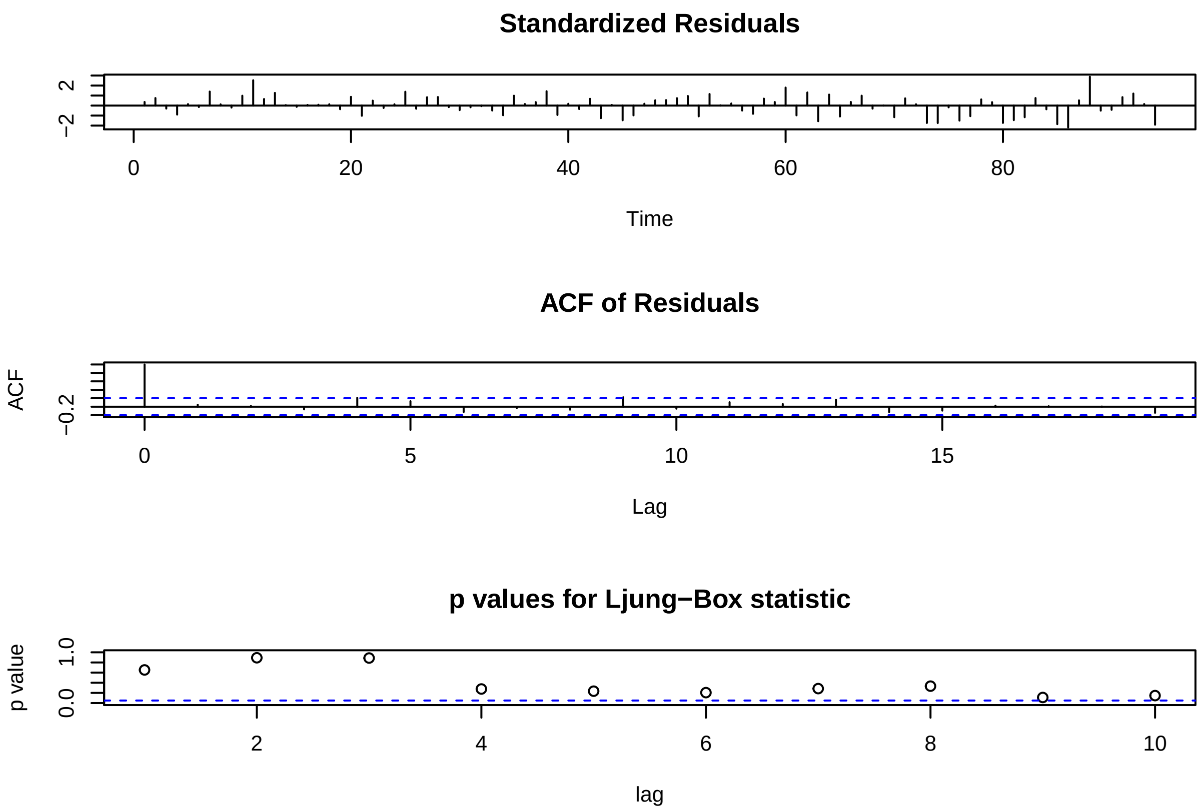
Question 5

```
reduced_arima = auto.arima(y_train, xreg = as.matrix(reduced_x_train))
summary(reduced_arima)
```

```
## Series: y_train
## Regression with ARIMA(0,0,1) errors
##
## Coefficients:
##          ma1  intercept  S.WD_MIN    A.PPU
##       0.5109     6.9871    0.3591  -2.6548
## s.e.  0.0942     0.1340    0.0928   0.1341
##
## sigma^2 estimated as 0.08464:  log likelihood=-15.43
## AIC=40.86   AICc=41.54   BIC=53.58
##
## Training set error measures:
##                        ME      RMSE       MAE        MPE     MAPE
## Training set -0.000157446 0.2846773 0.2242023 -0.3514288 4.941856
##                   MASE       ACF1
```

```
## Training set 0.3436012 0.04571852
```

```r
tsdiag(reduced_arima)
```



Comments on its validity.

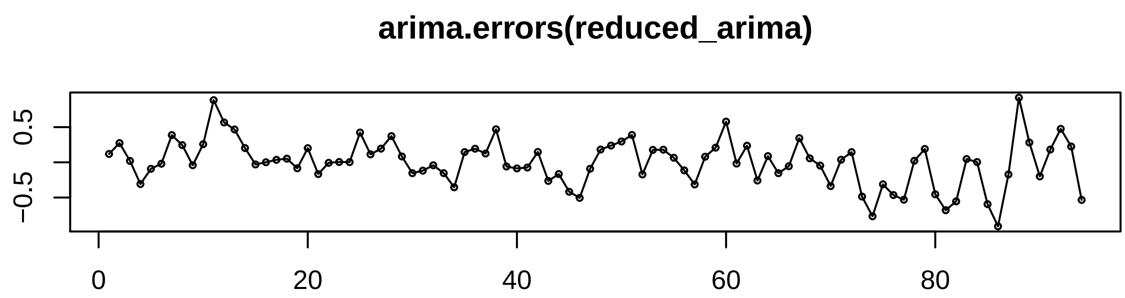
Solution

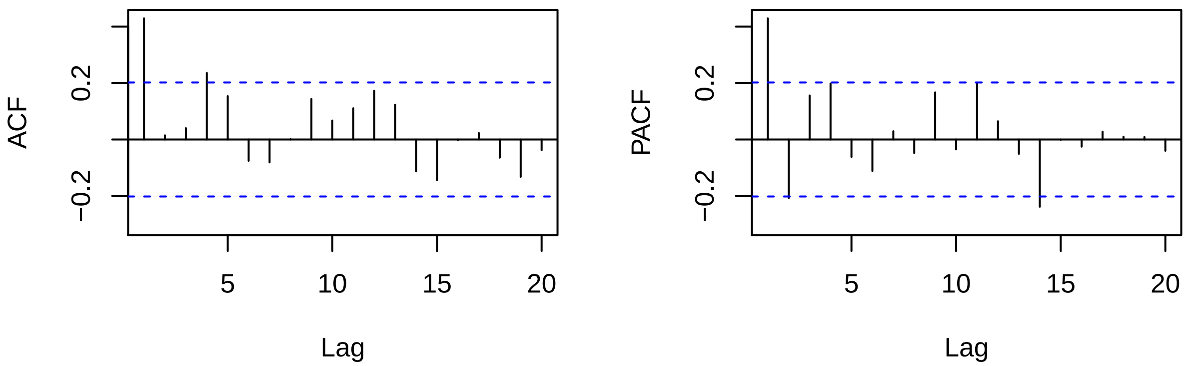
Using our reduced x training set, the auto arima was fitted and produced a model that fits the data well. The MASE tells that this is a much better fit than a naive forecast and therefore should be tried as a model to predict the training set data to assess validity.

Question 6

```r
tsdisplay(arima.errors(reduced_arima))
```

```
## Deprecated, use residuals.Arima(object, type='regression') instead
```






```r
# p = 2, q = 2

dynamic_reg = Arima(y_train, xreg = as.matrix(reduced_x_train),
                    order = c(2,0,2))
summary(dynamic_reg)
```

```
## Series: y_train
## Regression with ARIMA(2,0,2) errors
##
## Coefficients:
##           ar1      ar2     ma1     ma2  intercept  S.WD_MIN    A.PPU
##       -0.0169  -0.3824  0.5764  0.4203     6.9081    0.3521  -2.5671
## s.e.   0.4878   0.2870  0.5129  0.1722     0.1527    0.0934   0.1584
##
## sigma^2 estimated as 0.08369:  log likelihood=-13.4
## AIC=42.79   AICc=44.49   BIC=63.14
##
## Training set error measures:
##                        ME      RMSE       MAE        MPE     MAPE     MASE
## Training set 0.0004954658 0.2783076 0.2180301 -0.3524652 4.816171 0.334142
##                    ACF1
## Training set 0.02196659
```

Compare the coefficients of the explanatory variables in the Lasso model, unrestricted model of Question 2, and this model.

Solution

When comparing all three models built so far, it seems that the dynamic regression built in this question fits the data better than the unrestricted model fit in Question 2 and the Lasso reduced model. The AICc and BIC are both lower when looking at the Lasso reduced model in compared to the dynamic regression, but the RMSE (which is applicable because both datasets are similar) and MASE are both lower, and these more descriptive statistics yield that this is a better model than the one with the reduced x training set.

B Notation of model obtained

β_0 + β_1*$\mathbf{S.WD\_MIN}$ + β_2*$\mathbf{A.PPU}$

β_0 = **6.9081**

β_1 = **0.3521**

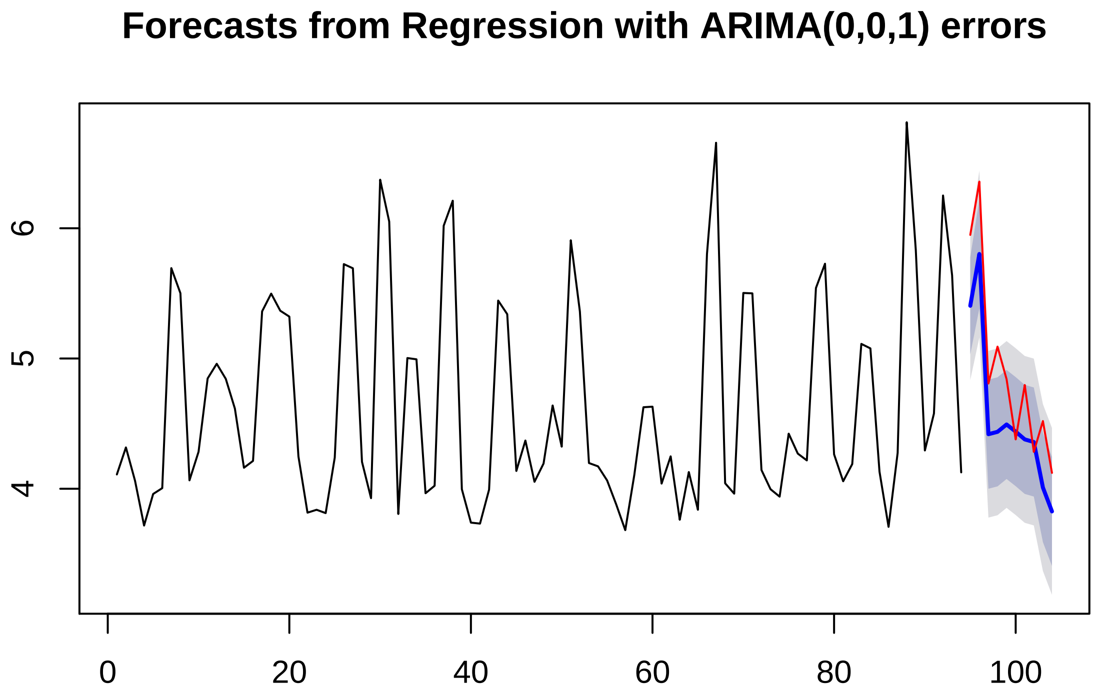
β_2 = **-2.5671**

Which equates to:

6.9081 + 0.3521 * S.WD_MIN - 2.5671 * A.PPU

Question 7

```
aa_red_forecast = forecast(reduced_arima, xreg = as.matrix(reduced_x_test),h = 10)
plot(aa_red_forecast)
lines(x=seq(95,104), y=y_test, col = 'red')
```



Comments on the usefulness of this model in terms of precision and confidence interval relative to the model without explanatory variables in Question 3.

Solution

The model seems to be a much better fit that the previous forecast utilizing the Lasso reduced training set. The testing set falls within the confidence interval, although it is still not a perfect fit. Again, with a more robust training set, this could have been remedied, but as this data seems very sparatic, this may be the best fit that we have in this scenario.

Question 8

Comment on the training and testing fit statistics and discuss how do you think you could improve on the performance of the model in terms of (a) additional data, (b) different pre-processing of the existing data, and (c) different modeling choices. Discuss your assessment of the potential for improvement (ex-ante priorities) for the different improvement options you suggest.

Solution

The fit statistics for the best two models here seemed to perform very well on the training set, but did struggle when used to forecast the testing set. To better improve this fit, the first priority would be to collect more data. This could be beneficial to help train our model as long as the information in relevent and contributes to the model fitting. Next, as this data is quite volitile, additional computing time would be beneficial to hopefully fitting this better. Lastly, during pre-processing, it would potentially be beneficial to view sales differently instead of simply aggregating them, and that would be an area of exploration that could help our fit.