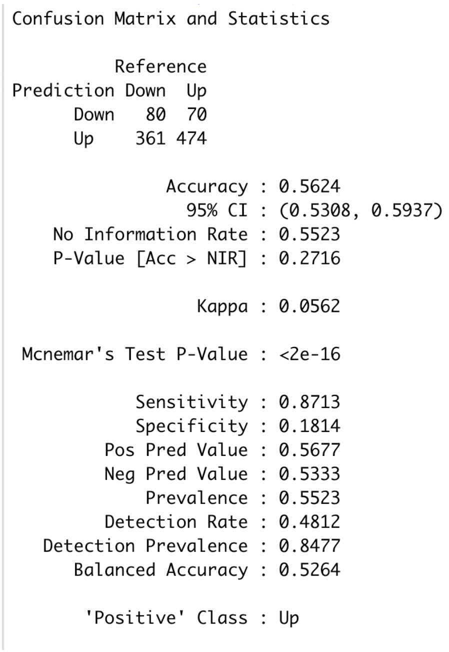
(c) Summary of the logistic regression model built with 5 Lag and Volume input variables.

```
Call:
glm(formula = Direction ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7186  -1.2498   0.9823   1.0841   1.4911

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.33258    0.09421   3.530 0.000415 ***
Lag1        -0.06231    0.02935  -2.123 0.033762 *
Lag2         0.04468    0.02982   1.499 0.134002
Lag3        -0.01546    0.02948  -0.524 0.599933
Lag4        -0.03111    0.02924  -1.064 0.287241
Lag5        -0.03775    0.02924  -1.291 0.196774
Volume      -0.08972    0.05410  -1.658 0.097240 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1354.7  on 984  degrees of freedom
Residual deviance: 1342.3  on 978  degrees of freedom
AIC: 1356.3

Number of Fisher Scoring iterations: 4
```

Training data:

- Confusion matrix:

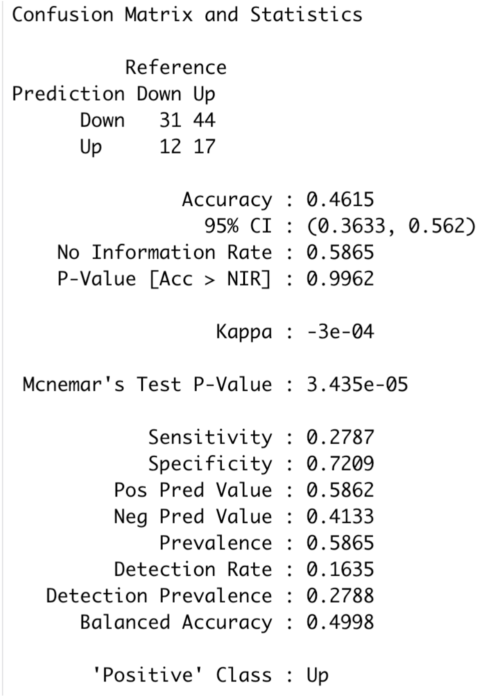


```
Confusion Matrix and Statistics

          Reference
Prediction Down  Up
      Down   80  70
      Up    361 474

               Accuracy : 0.5624
                 95% CI : (0.5308, 0.5937)
    No Information Rate : 0.5523
    P-Value [Acc > NIR] : 0.2716

                  Kappa : 0.0562

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8713
            Specificity : 0.1814
         Pos Pred Value : 0.5677
         Neg Pred Value : 0.5333
             Prevalence : 0.5523
         Detection Rate : 0.4812
   Detection Prevalence : 0.8477
      Balanced Accuracy : 0.5264

       'Positive' Class : Up
```

- Accuracy: 0.5624365

Test data:

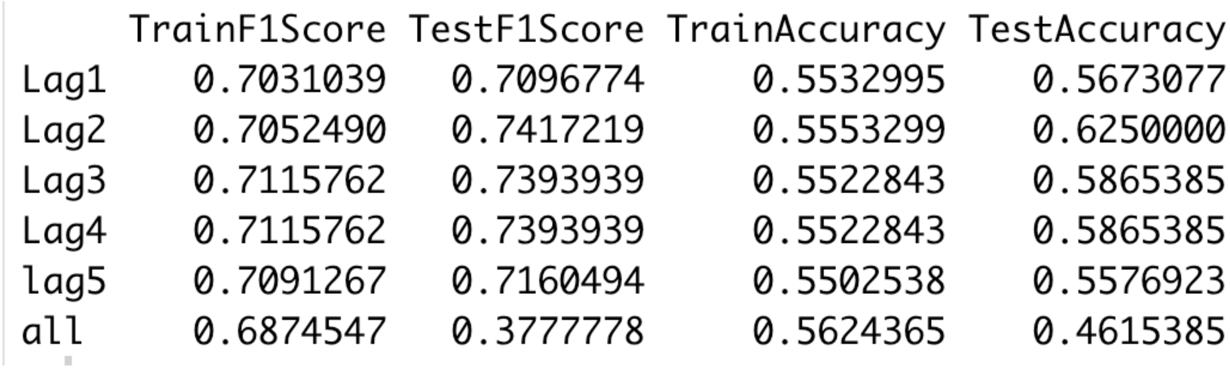
- Confusion Matrix:



```
Confusion Matrix and Statistics

          Reference
Prediction Down Up
      Down   31 44
      Up     12 17

               Accuracy : 0.4615
                 95% CI : (0.3633, 0.562)
    No Information Rate : 0.5865
    P-Value [Acc > NIR] : 0.9962

                  Kappa : -3e-04

 Mcnemar's Test P-Value : 3.435e-05

            Sensitivity : 0.2787
            Specificity : 0.7209
         Pos Pred Value : 0.5862
         Neg Pred Value : 0.4133
             Prevalence : 0.5865
         Detection Rate : 0.1635
   Detection Prevalence : 0.2788
      Balanced Accuracy : 0.4998

       'Positive' Class : Up
```

- Accuracy: 0.4615385

(d) Comparison table for all the evaluation metrics across the 6 models:

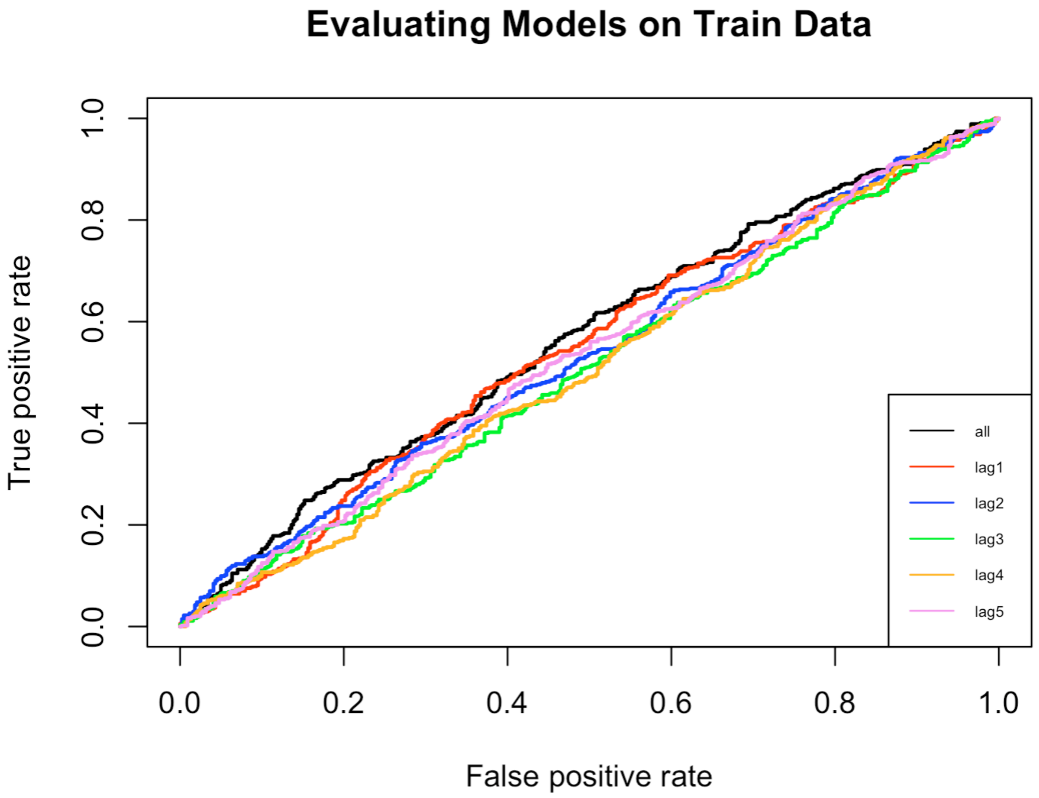


```
     TrainF1Score TestF1Score TrainAccuracy TestAccuracy
Lag1    0.7031039   0.7096774     0.5532995    0.5673077
Lag2    0.7052490   0.7417219     0.5553299    0.6250000
Lag3    0.7115762   0.7393939     0.5522843    0.5865385
Lag4    0.7115762   0.7393939     0.5522843    0.5865385
lag5    0.7091267   0.7160494     0.5502538    0.5576923
all     0.6874547   0.3777778     0.5624365    0.4615385
```

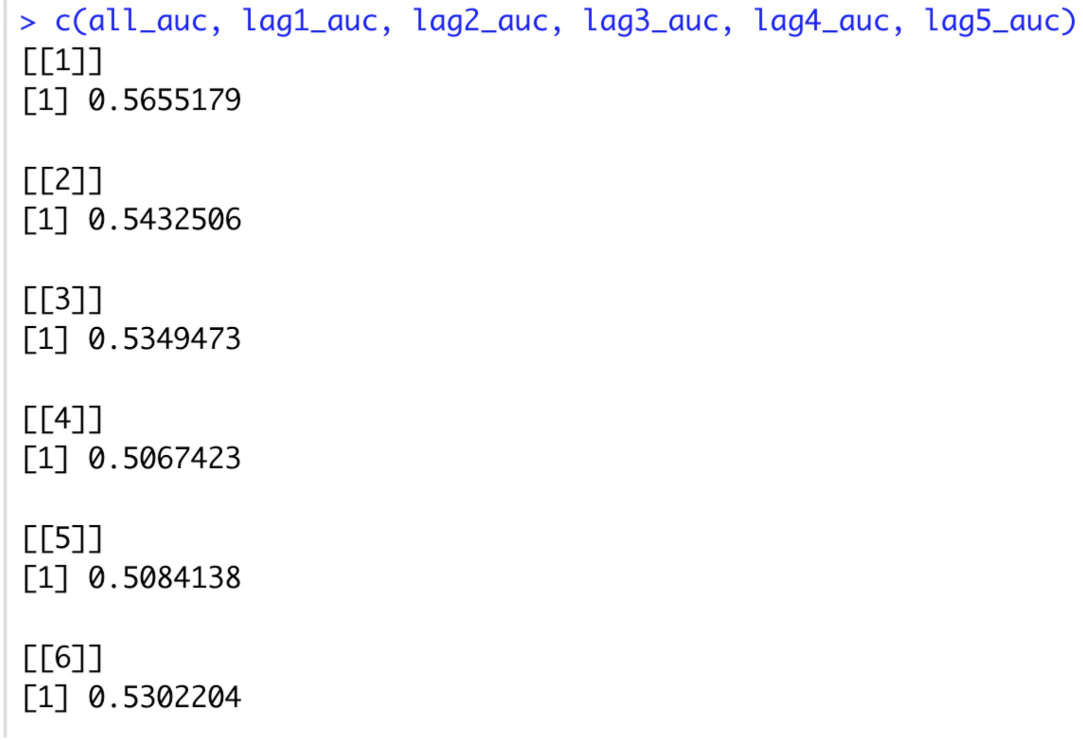
The best model in terms of accuracy and F1 score is Lag2 (the model with only Lag2 as independent variable), based on the test data.

No, Lag2 model is not the best model on training data as it does not achieve the highest accuracy and F1 Score. On training data, the model with all variables achieves the highest accuracy and the model with Lag 3/ Lag 4 have the highest F1-Score on train data.

(e) Evaluating models on Train data:

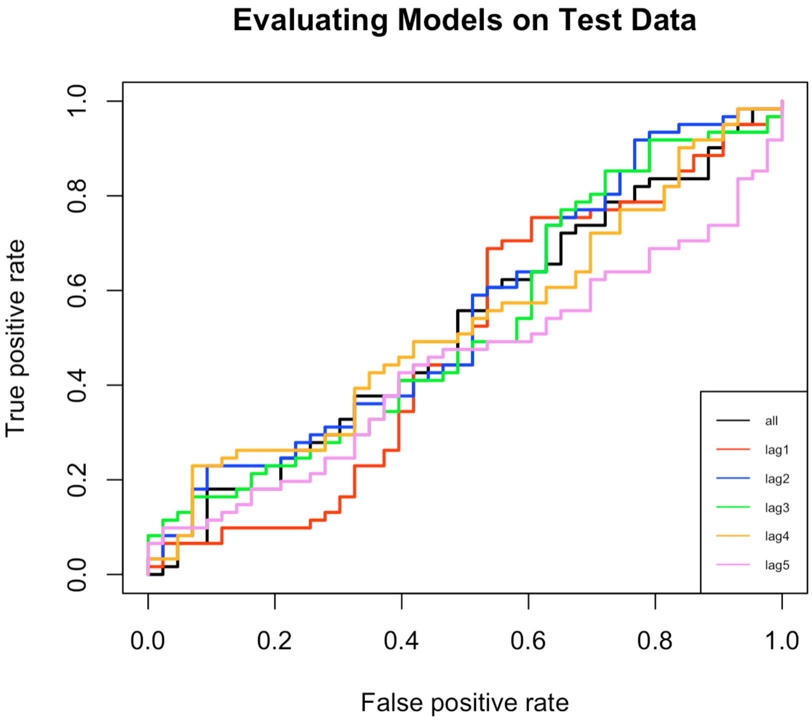


Area under the curve for training data for each model:



```
> c(all_auc, lag1_auc, lag2_auc, lag3_auc, lag4_auc, lag5_auc)
[[1]]
[1] 0.5655179

[[2]]
[1] 0.5432506

[[3]]
[1] 0.5349473

[[4]]
[1] 0.5067423

[[5]]
[1] 0.5084138

[[6]]
[1] 0.5302204
```

Evaluating models on Test data:

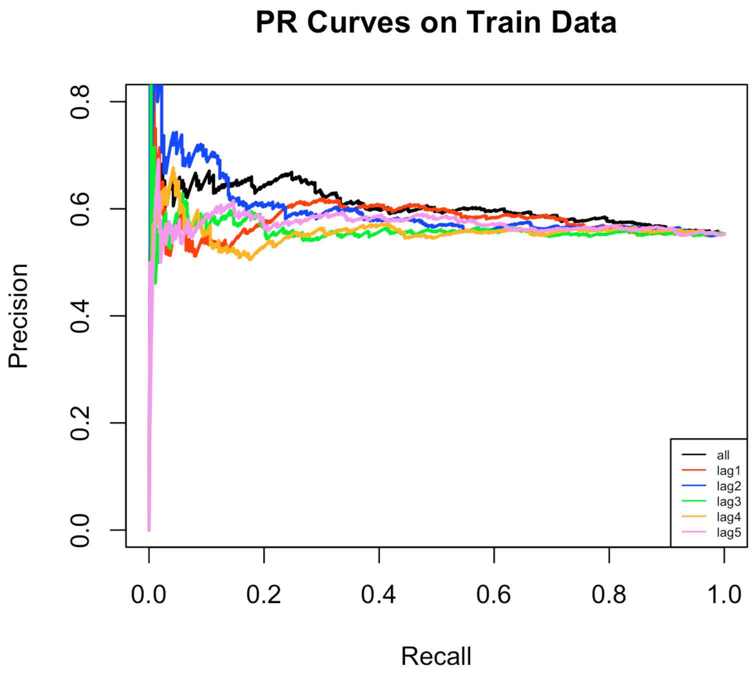


Area under the curve for test data for each model:

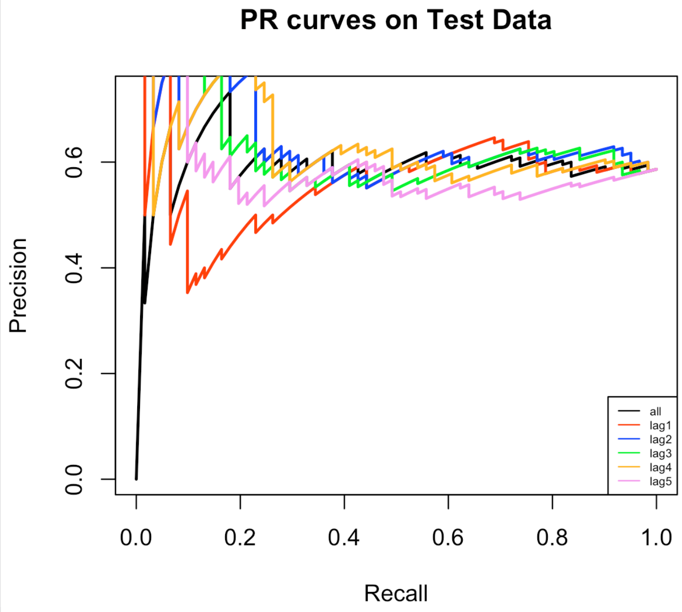
```
> c(all_auc, lag1_auc, lag2_auc, lag3_auc, lag4_auc, lag5_auc)
[[1]]
[1] 0.5177278

[[2]]
[1] 0.4864659

[[3]]
[1] 0.546321

[[4]]
[1] 0.5242089

[[5]]
[1] 0.5257339

[[6]]
[1] 0.4422417
```

The best model on the basis of AUC is Lag2 on test data.
(Whereas on train data, best AUC is of the model with all the variables)

(f) Precision Recall curves on Train data:



Precision Recall curves on Test data:



The best model according to the PR curve is Lag2.

(g) Accuracy on the train data is maximum for the model with all variables due to more data capturing. But on the other hand, this model’s accuracy is the worst on the test data. Overall, Lag2 model is the best based on a higher F1 Score on test data. Even though, it also has the highest accuracy, it does not matter for this case because we have imbalanced classes. On the same line, PR Curve’s area under the curve (AUC) makes more sense than ROC’s AUC in this case.