# Comprehensive Evaluation of Predictive Models and Feature Engineering in Financial Forecasting.

Yamuna Dhungana

I created a model using the "Weekly" dataset and fitted it with the MclustDA function from the "mclust" library. My objective was to select the most appropriate model based on the Bayesian Information Criterion (BIC).

Subsequently, I calculated key metrics including the true positive rate, true negative rate, training error, and test error. These measurements are crucial for evaluating the model's performance and its ability to correctly classify data points as positive or negative.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df       BIC
##       -2129.439 985 10 -4327.804
##
## Classes   n     % Model G
##    Down 441 44.77     V 2
##    Up   544 55.23     V 2
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down   76 365
##   Up     70 474
## Classification error = 0.4416
## Brier score          = 0.2452
```

```
##          train_error
## accuracy       55.84
## TPR            87.13
## TNR            17.23
```

```
##          test_error
## accuracy      54.81
## TPR           85.25
## TNR           11.63
```

```
## ## With all the variables
```

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
```

```
## MclustDA model summary:
##
##  log-likelihood   n  df       BIC
##       -12477.54 985 286 -26926.37
##
## Classes   n     % Model G
##    Down 441 44.77   VVV 4
##    Up   544 55.23   VVV 4
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down  251 190
##   Up    160 384
## Classification error = 0.3553
## Brier score          = 0.2177
```

```
##          train_error
## accuracy       64.47
## TPR            70.59
## TNR            56.92
```

```
##          test_error
## accuracy      53.85
## TPR           85.25
## TNR            9.30
```

In the preceding analysis, I identified Lag2 as the most significant variable. Nevertheless, I conducted two separate model runs: one exclusively with Lag2 and the other encompassing all variables. My primary aim was to assess the performance of the model when considering all variables.

In the single-variable model, the model is characterized by variable variance, rendering it one-dimensional and applicable to two distinct groups. Conversely, in the model with all variables, the structure is described as ellipsoidal, demonstrating varying volume, shape, and orientation. Notably, the Bayesian Information Criterion (BIC) for the single-variable model is higher than that for the model employing all variables.

Furthermore, I generated tables for both the test and train datasets, encompassing these two model scenarios, to facilitate a comprehensive comparison.

Now, I'm re-running the MclustDA analysis, but this time I'm specifying modelType = "EDDA". I'm going through the same process of selecting the best model based on the Bayesian Information Criterion (BIC). Additionally, I'll calculate the true positive rate, true negative rate, training error, and test error.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -15029.33 985 49 -30396.39
##
## Classes   n     % Model G
##    Down 441 44.77   VVE 1
##    Up   544 55.23   VVE 1
##
## Training confusion matrix:
```

```
##       Predicted
## Class  Down  Up
##   Down   98 343
##   Up     90 454
## Classification error = 0.4396
## Brier score          = 0.2497
```

```
##          train_error
## accuracy       56.04
## TPR            83.46
## TNR            22.22
```

```
##          test_error
## accuracy      46.15
## TPR           14.75
## TNR           90.70
```

I attempted to fit the model using both all variables and a single variable. However, it's important to note that the single-variable model failed to converge. Subsequently, I created tables to summarize the train and test errors for these models.

Upon examining the Mclust documentation, I discovered that specifying "EDDA" as the model type enforces a single component in each class with the same covariance structure. This single component exhibited an ellipsoidal structure with equal orientation, denoted as VVE.

Now Comparing the results,

Table 1: MsclustDA Test Accuracy with single variable

| | train_error | test_error |
|---|---|---|
| accuracy | 55.84 | 54.81 |
| TPR | 87.13 | 85.25 |
| TNR | 17.23 | 11.63 |

Table 2: MsclustDA Test Accuracy with all variables

| | train_error | test_error | train_error | test_error |
|---|---|---|---|---|
| accuracy | 64.47 | 53.85 | 56.04 | 46.15 |
| TPR | 70.59 | 85.25 | 83.46 | 14.75 |
| TNR | 56.92 | 9.30 | 22.22 | 90.70 |

Table 3: MsclustDA with EDDA Test Accuracy with all variables

| | train_error | test_error |
|---|---|---|
| accuracy | 56.04 | 46.15 |
| TPR | 83.46 | 14.75 |
| TNR | 22.22 | 90.70 |

Table 4: Logreg Accuracy measures with single variable

| | Train_error | Test_error |
|---|---|---|
| accuracy | 55.53 | 62.50 |
| TPR | 96.32 | 91.80 |
| TNR | 5.22 | 20.93 |

Table 5: LDA Accuracy measures with single variable

| | Train_error | Test_error |
|---|---|---|
| accuracy | 55.43 | 62.50 |
| TPR | 96.32 | 91.80 |
| TNR | 4.99 | 20.93 |

Table 6: QDA Accuracy measures with single variable

| | Train_error | Test_error |
|---|---|---|
| accuracy | 55.23 | 58.65 |
| TPR | 100.00 | 100.00 |
| TNR | 0.00 | 0.00 |

Table 7: KNN Accuracy measures with single variable

| | Test_error |
|---|---|
| accuracy | 50.96 |
| TPR | 52.46 |
| TNR | 49.00 |

In this context, I've compiled tables summarizing the results from all the methods we applied, both in the current analysis and previous ones. A quick glance at these tables reveals a range of test data accuracy, which spans from 62.50% to 46.15%. Similarly, training accuracy varies between 64.47% and 50.0%.

It's worth noting that the logistic regression model, in particular, stands out as highly accurate. Furthermore, the Linear Discriminant Analysis (LDA) model also demonstrates a commendable accuracy rate.

In this stage of the analysis, I took the original model variables and created a new set of variables. I then fitted a model using `MclustDA` and replicated the previous steps. The objective was to assess whether these new variables led to an improvement in error rates when compared to the previous models.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df       BIC
##       -4220.728 985 18 -8565.523
##
## Classes   n     % Model G
```

```
##    Down 441 44.77   VII 3
##    Up   544 55.23   VII 2
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down  156 285
##   Up    137 407
## Classification error = 0.4284
## Brier score          = 0.243
```

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df       BIC
##       -5545.222 985 90 -11710.78
##
## Classes   n     % Model G
##    Down 441 44.77   VEV 5
##    Up   544 55.23   VVV 5
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down  204 237
##   Up    200 344
## Classification error = 0.4437
## Brier score          = 0.2945
```

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df       BIC
##       -2314.625 985 58 -5029.024
##
## Classes   n     % Model G
##    Down 441 44.77   VVV 5
##    Up   544 55.23   VVV 5
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down  125 316
##   Up    128 416
## Classification error = 0.4508
## Brier score          = 0.3041
```

Table 8: Accuracy measures using MclustDA

| | tr.error modd1 | tt.error modd1 | tr.error modd2 | tt.error modd2 | tr.error modd3 | tt.error modd3 |
|---|---|---|---|---|---|---|
| accuracy | 57.16 | 50.96 | 55.63 | 55.77 | 54.92 | 57.69 |
| TPR | 74.82 | 70.49 | 63.24 | 60.66 | 76.47 | 77.05 |
| TNR | 35.37 | 23.26 | 46.26 | 48.84 | 28.34 | 30.23 |

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -4408.247 985  5 -8850.957
##
## Classes   n     % Model G
##    Down 441 44.77   EII 1
##    Up   544 55.23   EII 1
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down   50 391
##   Up     52 492
## Classification error = 0.4497
## Brier score          = 0.2454
```

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -7896.142 985 18 -15916.35
##
## Classes   n     % Model G
##    Down 441 44.77   VVV 1
##    Up   544 55.23   VVV 1
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down  311 130
##   Up    355 189
## Classification error = 0.4924
## Brier score          = 0.256
```

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
```

```
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -6144.173 985 10 -12357.27
##
## Classes   n     % Model G
##    Down 441 44.77   VVV 1
##    Up   544 55.23   VVV 1
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down   14 427
##   Up     19 525
## Classification error = 0.4528
## Brier score          = 0.2558
```

Table 9: Accuracy measures using MclustDA EDDA

| | tr.error.1ed | tt.error.1ed | tr.error.2ed | tt.error.2ed | tr.error.3ed | tt.error.3ed |
|---|---|---|---|---|---|---|
| accuracy | 55.03 | 56.73 | 50.76 | 46.15 | 54.72 | 62.50 |
| TPR | 90.44 | 83.61 | 34.74 | 40.98 | 96.51 | 95.08 |
| TNR | 11.34 | 18.60 | 70.52 | 53.49 | 3.17 | 16.28 |

Table 10: Accuracy measures using LDA

| | Trn_err.lda1 | Tt_err.lda1 | Trn_err.lda2 | Tt_err.lda2 | Trn_err.lda3 | Tt_err.lda3 |
|---|---|---|---|---|---|---|
| accuracy | 54.82 | 57.69 | 54.82 | 57.69 | 54.82 | 61.54 |
| TPR | 91.54 | 86.89 | 91.54 | 86.89 | 96.69 | 93.44 |
| TNR | 9.52 | 16.28 | 9.52 | 16.28 | 3.17 | 16.28 |

Table 11: Accuracy measures using QDA

| | Trn_err.qda1 | Tt_err.qda1 | Trn_err.qda2 | Tt_err.qda2 | Trn_err.qda3 | Tt_err.qda3 |
|---|---|---|---|---|---|---|
| accuracy | 55.33 | 55.77 | 50.76 | 46.15 | 54.72 | 62.50 |
| TPR | 84.93 | 83.61 | 34.74 | 40.98 | 96.51 | 95.08 |
| TNR | 18.82 | 16.28 | 70.52 | 53.49 | 3.17 | 16.28 |

I created three models with the following specifications: 1. `Direction~Lag1+Lag2` 2. `Direction~Lag1+Lag2+Lag1*Lag2` 3. `Direction~Lag2+I(Lag2^2)`

For these models, I conducted fitting alongside all the previous models we've explored. In the case of the models with `MclustDA`, both exhibited a spherical structure with unequal volume, featuring three groups for "down" and two for "up" in the first model. The second model displayed an ellipsoidal shape with equal orientation for "down" and "up," including ellipsoidal structures with varying volume, shape, and orientation across five groups. The third model was also ellipsoidal with varying volume, shape, and orientation across five groups.

Regarding the first model with "EDDA," it featured one group with a spherical structure and equal volume. The second and third "EDDA" models adopted ellipsoidal models with varying volume, shape, and orientation, each with one group.

To summarize the performance of all these models, I created a combined table that presents the accuracy for both the test and training errors.