Classification of Jobs in High Performance Computing Environment

OR 568-Applied Predictive Analytics

Final Project Report

Team Members

1) **Venkata Jayandra Kumar Lade** **G01046700**
2) **Arun Reddy Bollam** **G01040932**
3) **Vamsi Krishna Reddicherla** **G01039349**
4) **Vaibhav Mukesh Trivedi** **G01009749**

Problem Description:

High Performance computing is used to facilitate large scale computations in highly reliable and sophisticated environment. It can handle many programs simultaneously. Scheduling of the jobs is extremely crucial in such an environment. The objective of our model is to predict the class of the job.

Dataset:

Data was obtained directly from a package called Applied Predictive Modelling in R. Data consists of 4331 records with 8 attributes (Protocol, Compounds, Input Fields, Iterations, Numpending, Hour, Day, and Class). The Class attribute is our response variable.

Technical Approach:

Data is divided into training and testing in the ratio 60:40. Classification models built are

1) Multinomial Logistic Regression
2) Random Forest
3) CART
4) Bagging
5) Boosting
6) Support vector machine
7) Naïve Bayes

Multinomial Logistic Regression:

We implemented Multinomial Logistic Regression with all the seven attributes as predictors and Class as response variable.

The picture below shows the formula and coefficients for each variable.

```
Call:
multinom(formula = Class ~ Protocol + Compounds + InputFields +
    Iterations + NumPending + Hour + Day, data = training)

Coefficients:
  (Intercept)  ProtocolC  ProtocolD  ProtocolE ProtocolF ProtocolG ProtocolH ProtocolI ProtocolJ  ProtocolK  ProtocolL
F  -1.908818  -8.096433  -6.526574  -1.422737 -0.840063 -1.763532  1.445252  -8.45086 -1.430965  -2.042535  -5.552802
M  -4.538881 -10.630176 -30.659195  -3.174965 -1.656474 -2.330201  1.643222 -64.69962 -2.898075 -18.016706 -11.459503
L  -6.947444 -14.202986 -25.653805 -18.511353 -1.620110 -4.070341  2.009509 -80.11145 -3.616129 -19.085198 -76.320904
  ProtocolM ProtocolN ProtocolO   Compounds  InputFields Iterations   NumPending        Hour     DayTue     Daywed     DayThu
F  1.677575 -1.261116  2.441019 0.002585436 0.0003530193 0.04578449 0.0007131671 0.02146812 0.07816788 -0.7043564 -0.2278998
M  1.764395 -2.195979  2.167061 0.003871288 0.0006267509 0.06183542 0.0013996060 0.05296045 0.60601272 -0.9065842  0.2567241
L  1.644456 -4.054406  3.284073 0.004565396 0.0011768361 0.07721219 0.0009888205 0.05376981 0.43012311 -1.4505903 -0.6867003
       DayFri      DaySat      DaySun
F -0.008185119  0.21220863   0.6216711
M  0.086950188 -7.89107749 -13.6812926
L -0.341349392  0.04986013 -10.5900781
```

Confusion Matrix:

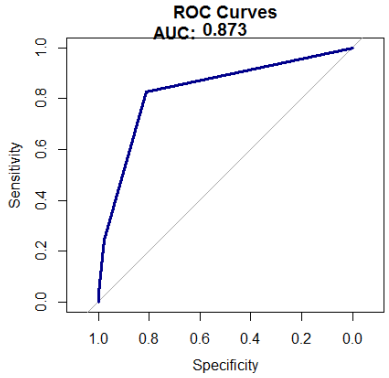
```
Confusion Matrix and Statistics

          Reference
Prediction  VF   F   M   L
        VF 755 154  19   0
        F  110 373 132  23
        M    3  18  52  14
        L    2   3  17  58

Overall Statistics

               Accuracy : 0.7143681
                 95% CI : (0.6924617, 0.7355464)
    No Information Rate : 0.5020196
    P-Value [Acc > NIR] : < 0.00000000000000022204
```

Roc Curve:



Area under curve area is 0.873

Multinomial Logistic Regression with backward Selection:

Backward selection is done the model obtained above.

The below results show the formula and Coefficients for each variable. We can observe that only attribute that doesn't have significant contribution to previous model is the HOUR attribute.

```
Call:
multinom(formula = Class ~ Protocol + Compounds + InputFields +
    Iterations + NumPending + Day, data = training)

Coefficients:
  (Intercept)  ProtocolC  ProtocolD  ProtocolE ProtocolF ProtocolG ProtocolH ProtocolI ProtocolJ  ProtocolK  ProtocolL
F  -1.616086  -8.031563  -6.455007  -1.383672 -0.7987949 -1.747757  1.468121  -8.343253 -1.388544  -1.932513  -5.497481
M  -3.832256 -10.612589 -38.730215  -3.056274 -1.5809276 -2.296695  1.695312 -58.663258 -2.809238 -17.337866 -11.373510
L  -6.234285 -14.177589 -22.345558 -12.191181 -1.5316073 -4.032320  2.065051 -71.885622 -3.522967 -17.398119 -69.144229
  ProtocolM ProtocolN ProtocolO   Compounds  InputFields Iterations   NumPending     DayTue     Daywed     DayThu     DayFri
F  1.684171 -1.245568  2.480515 0.002535040 0.0003501228 0.04568964 0.0007339888 0.08344137 -0.7146019 -0.2389874 -0.03939609
M  1.811293 -2.181279  2.257296 0.003801674 0.0006258956 0.06201522 0.0014350104 0.63124809 -0.9385292  0.2287605  0.01345774
L  1.698764 -4.037049  3.372541 0.004496736 0.0011739306 0.07742126 0.0010239172 0.45225376 -1.4791075 -0.7176339 -0.40593986
       DaySat      DaySun
F  0.1258508   0.6537942
M -13.6294779 -11.4102348
L  -0.4624328  -8.6396046
```

Confusion Matrix:

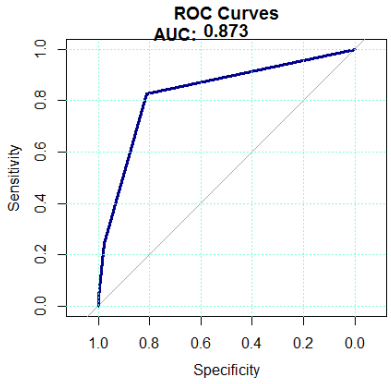
```
Confusion Matrix and Statistics

          Reference
Prediction  VF   F   M   L
        VF 755 148  19   0
        F  110 381 131  23
        M    3  16  51  14
        L    2   3  19  58

Overall Statistics

               Accuracy : 0.7184
                 95% CI : (0.6966, 0.7395)
    No Information Rate : 0.502
    P-Value [Acc > NIR] : < 2.2e-16
```

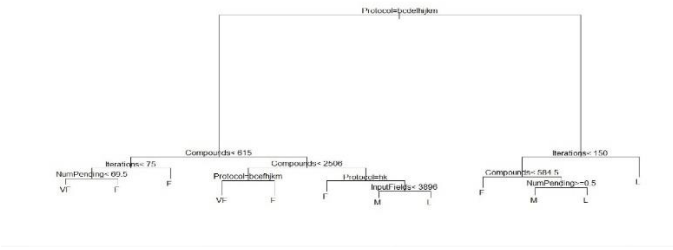
Roc Curve:



Area under curve area is 0.873

CART:

Classification Tree is generated for the data.



Confusion Matrix:

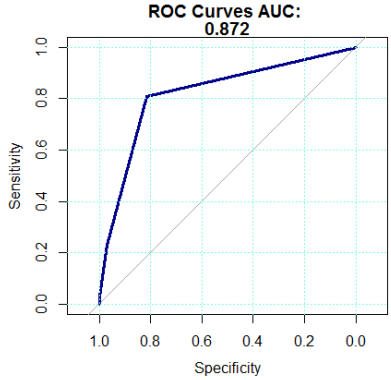
```
Confusion Matrix and Statistics

          Reference
Prediction  VF   F   M   L
        VF 737 139  28   0
        F  129 390 135  19
        M    3  13  44  16
        L    1   6  13  60

Overall Statistics

               Accuracy : 0.7103289
                 95% CI : (0.6883403, 0.7316031)
    No Information Rate : 0.5020196
    P-Value [Acc > NIR] : < 0.00000000000000022204
```

Roc Curve:



Area under curve area is 0.872

Random Forest:

```
Call:
 randomForest(formula = Class ~ ., data = training, mtry = round(sqrt(7)))
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3
```

Confusion Matrix:

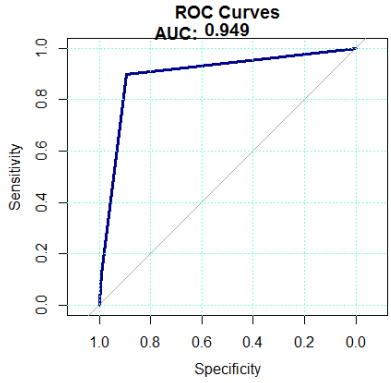
```
Confusion Matrix and Statistics

          Reference
Prediction  VF   F   M   L
        VF 809  88   6   0
        F   58 438  60   8
        M    3  22 150   8
        L    0   0   4  79

Overall Statistics

               Accuracy : 0.8517023
                 95% CI : (0.8340871, 0.868117)
    No Information Rate : 0.5020196
    P-Value [Acc > NIR] : < 0.00000000000000022204
```

Roc Curve:



Area under curve area is 0.949

Random Forest with 5-folds cross validations:

```
$n.var
[1] 7 4 1

$error.cv
          7           4           1
0.1526206419 0.1770953590 0.3588085892
```

Error rates obtained on the cross validation for 4 variables split is 17.7%

which is very close to obtained above error rate from the model.

Bagging:

Bagging done using adabag package in R.

Confusion Matrix:

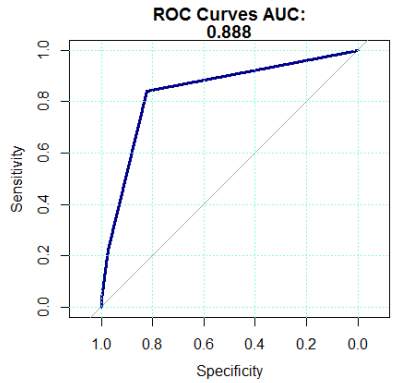
```
Confusion Matrix and Statistics

          Reference
Prediction  VF   F   M   L
        VF 772 143  23   0
        F   96 377 118  18
        M    2  21  63   6
        L    0   7  16  71

Overall Statistics

               Accuracy : 0.7403347
                 95% CI : (0.7190074, 0.760845)
    No Information Rate : 0.5020196
    P-Value [Acc > NIR] : < 0.00000000000000022204
```

Roc Curve:



Area under curve area is 0.888

Bagging with 5-folds cross-validation:

```
$error
[1] 0.2620641884
```

Error rates obtained on cross validation is 26.2% is very close to the error rate obtained above.

Boosting:

➔ Boosting done using adabag package in R.

Confusion Matrix:

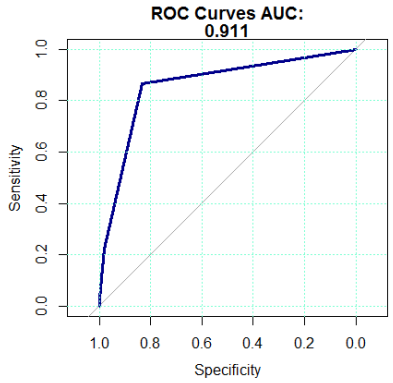
```
Confusion Matrix and Statistics

          Reference
Prediction  VF   F   M   L
        VF 789 139  19   0
        F   79 386 115  19
        M    2  19  83   9
        L    0   4   3  67

Overall Statistics

               Accuracy : 0.7645701
                 95% CI : (0.7438682, 0.7843723)
    No Information Rate : 0.5020196
    P-Value [Acc > NIR] : < 0.00000000000000022204
```

Roc Curve:



Area under curve area is 0.911

Boosting with 5-folds cross-validation:

```
$error
[1] 0.2318171323
```

Error rates obtained on cross validation is 23.1% is very close to the error rate obtained above.

Support Vector Machine:

SVM performed on all 7 predictors and 1 Response variable with default Parameters.

Confusion Matrix:

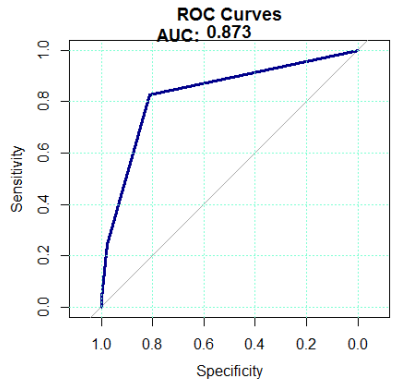
```
Confusion Matrix and Statistics

          Reference
Prediction  VF   F   M   L
        VF 846 177  29   1
        F   61 329 109  35
        M    3  13  50  12
        L    0   9   6  53

Overall Statistics

               Accuracy : 0.7374
                 95% CI : (0.7161, 0.758)
    No Information Rate : 0.5251
    P-Value [Acc > NIR] : < 2.2e-16
```

Roc Curve:



Area under curve area is 0.873

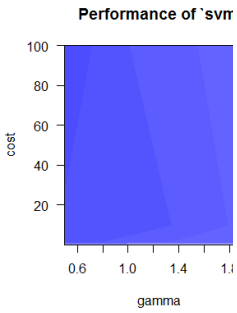
Support Vector Machine – Optimized:

SVM is performed using best parameters, best parameters are obtained by tuning the model using GRID search.

```
Parameter tuning of 'svm':

- sampling method: 10-fold

- best parameters:
 gamma cost
   0.5  100
```



Confusion Matrix:

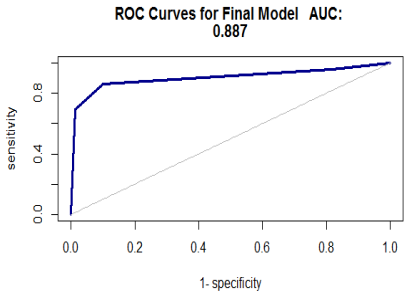
```
Confusion Matrix and Statistics

          Reference
Prediction  VF   F   M   L
        VF 804 104  15   5
        F   99 375  67  12
        M    6  42 106  14
        L    1   7   6  70

Overall Statistics

               Accuracy : 0.7819
                 95% CI : (0.7617, 0.8011)
    No Information Rate : 0.5251
    P-Value [Acc > NIR] : < 2e-16
```

Roc Curve:



Area under curve area is 0.887

Naïve Bayes:

Confusion Matrix:

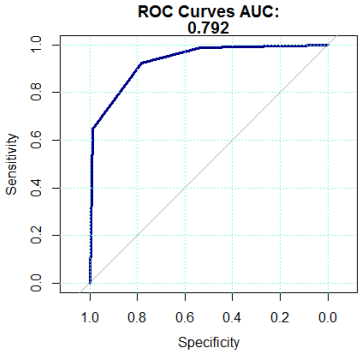
```
Confusion Matrix and Statistics

          Reference
Prediction  VF   F   M   L
        VF  1  18  22  47
        F  824 352  84  14
        M    5   9  33  11
        L   40 169  81  23

Overall Statistics

               Accuracy : 0.2360069
                 95% CI : (0.2161869, 0.2567248)
    No Information Rate : 0.5020196
    P-Value [Acc > NIR] : 1
```

ROC Curve:



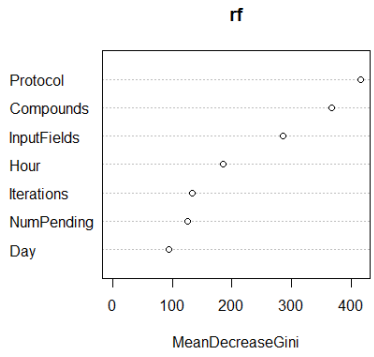
Area under curve area is 0.792

Model Selection:

Random forest model is clearly the best model among all of models based on accuracy.

Conclusion:

Based on the variable importance plot of the random forest, we can safely conclude that Protocol, Compounds, InputFields have the highest influence on the class of the job.



The hardware that is being used in such an environment is being updated every year. The results will change when the jobs are done on the updated hardware. Therefore, whenever there is a change in the hardware, new models must be introduced so that get to know the factors affecting the Class [Speed of computation].

References:

[1] http://appliedpredictivemodeling.com/data/

[2] http://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/

[3] https://cran.r-project.org/web/packages/adabag/adabag.pdf

[4] http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf

Contribution:

Arun Reddy Bollam	Project selection, Naïve bayes, Boosting, Report
Vamsi Krishna Reddicherla	Project selection, Support Vector Machine, CART, Report
Vaibhav Trivedi	Project selection, Bagging, Power Point Presentation, Report
Venkata Jayandra Kumar Lade	Project selection, Multinomial Logistic Regression, Random Forest, Report