

Submission 11

Shannon Rush

July 21, 2014

Fitting a generalized linear model with 60% of training data

Load and Source

```
source("../helpers/predictions.R")
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(doMC)
```

```
## Loading required package: foreach
```

```
## Loading required package: iterators
```

```
## Loading required package: parallel
```

```
train <- read.csv("../data/processed/processed_train.csv")
```

```
test <- read.csv("../data/original/test.csv")
```

Create Validation

```
set.seed(1111)
```

```
training.indices <- createDataPartition(train$Label, p=0.6, list=F)
```

```
training <- train[training.indices,]
```

```
validation <- train[-training.indices,]
```

Fit GLM

```
predictors <- training[,setdiff(names(training),c("EventId","Label"))]
```

```
registerDoMC(cores=4)
```

```
glm.fit <- train(x=predictors, y=training$Label, method="glm")
```

```
glm.fit
```

```
## Generalized Linear Model
```

```
##
```

```
## 150001 samples
```

```
##      30 predictors
```

```
##      2 classes: 'b', 's'
```

```
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
##
## Summary of sample sizes: 150001, 150001, 150001, 150001, 150001, 150001, ...
##
## Resampling results
##
##   Accuracy  Kappa  Accuracy SD  Kappa SD
##   0.8       0.4    0.002         0.004
##
##
```

```
str(glm.fit)
```

```
## List of 19
## $ method      : chr "glm"
## $ modelInfo    :List of 14
## ..$ label     : chr "Generalized Linear Model"
## ..$ library    : NULL
## ..$ loop       : NULL
## ..$ type       : chr [1:2] "Regression" "Classification"
## ..$ parameters:'data.frame': 1 obs. of 3 variables:
## .. ..$ parameter: Factor w/ 1 level "parameter": 1
## .. ..$ class    : Factor w/ 1 level "character": 1
## .. ..$ label    : Factor w/ 1 level "parameter": 1
## ..$ grid       :function (x, y, len = NULL)
## .. ..- attr(*, "srcref")=Class 'srcref' atomic [1:8] 8 26 8 82 26 82 8 8
## .. ..- attr(*, "srcfile")=Classes 'srcfilecopy', 'srcfile' <environment: 0x7fc0e5fb8608>
## ..$ fit        :function (x, y, wts, param, lev, last, classProbs, ...)
## .. ..- attr(*, "srcref")=Class 'srcref' atomic [1:8] 9 25 30 19 25 19 9 30
## .. ..- attr(*, "srcfile")=Classes 'srcfilecopy', 'srcfile' <environment: 0x7fc0e5fb8608>
## ..$ predict    :function (modelFit, newdata, submodels = NULL)
## .. ..- attr(*, "srcref")=Class 'srcref' atomic [1:8] 31 29 42 19 29 19 31 42
## .. ..- attr(*, "srcfile")=Classes 'srcfilecopy', 'srcfile' <environment: 0x7fc0e5fb8608>
## ..$ prob       :function (modelFit, newdata, submodels = NULL)
## .. ..- attr(*, "srcref")=Class 'srcref' atomic [1:8] 43 26 50 19 26 19 43 50
## .. ..- attr(*, "srcfile")=Classes 'srcfilecopy', 'srcfile' <environment: 0x7fc0e5fb8608>
## ..$ varImp     :function (object, ...)
## .. ..- attr(*, "srcref")=Class 'srcref' atomic [1:8] 51 28 58 19 28 19 51 58
## .. ..- attr(*, "srcfile")=Classes 'srcfilecopy', 'srcfile' <environment: 0x7fc0e5fb8608>
## ..$ predictors:function (x, ...)
## .. ..- attr(*, "srcref")=Class 'srcref' atomic [1:8] 59 32 59 67 32 67 59 59
## .. ..- attr(*, "srcfile")=Classes 'srcfilecopy', 'srcfile' <environment: 0x7fc0e5fb8608>
## ..$ levels     :function (x)
## .. ..- attr(*, "srcref")=Class 'srcref' atomic [1:8] 60 28 60 93 28 93 60 60
## .. ..- attr(*, "srcfile")=Classes 'srcfilecopy', 'srcfile' <environment: 0x7fc0e5fb8608>
## ..$ tags       : chr [1:2] "Generalized Linear Model" "Linear Classifier"
## ..$ sort       :function (x)
## .. ..- attr(*, "srcref")=Class 'srcref' atomic [1:8] 62 26 62 38 26 38 62 62
## .. ..- attr(*, "srcfile")=Classes 'srcfilecopy', 'srcfile' <environment: 0x7fc0e5fb8608>
## $ modelType    : chr "Classification"
## $ results      :'data.frame': 1 obs. of 5 variables:
## ..$ parameter : Factor w/ 1 level "none": 1
```

```

## ..$ Accuracy : num 0.75
## ..$ Kappa : num 0.417
## ..$ AccuracySD: num 0.00164
## ..$ KappaSD : num 0.00357
## $ pred : NULL
## $ bestTune : 'data.frame': 1 obs. of 1 variable:
## ..$ parameter: Factor w/ 1 level "none": 1
## $ call : language train.default(x = predictors, y = training$Label, method = "glm")
## $ dots : list()
## $ metric : chr "Accuracy"
## $ control :List of 21
## ..$ method : chr "boot"
## ..$ number : num 25
## ..$ repeats : num 25
## ..$ p : num 0.75
## ..$ initialWindow : NULL
## ..$ horizon : num 1
## ..$ fixedWindow : logi TRUE
## ..$ verboseIter : logi FALSE
## ..$ returnData : logi TRUE
## ..$ returnResamp : chr "final"
## ..$ savePredictions : logi FALSE
## ..$ classProbs : logi FALSE
## ..$ summaryFunction :function (data, lev = NULL, model = NULL)
## ..$ selectionFunction: chr "best"
## ..$ preProcOptions :List of 3
## .. ..$ thresh : num 0.95
## .. ..$ ICAcomp: num 3
## .. ..$ k : num 5
## ..$ index :List of 25
## .. ..$ Resample01: int [1:150001] 1 3 4 5 5 6 6 6 8 8 ...
## .. ..$ Resample02: int [1:150001] 1 1 2 2 3 3 4 4 4 5 ...
## .. ..$ Resample03: int [1:150001] 1 3 4 4 6 6 7 7 9 10 ...
## .. ..$ Resample04: int [1:150001] 5 6 6 7 8 8 9 10 11 13 ...
## .. ..$ Resample05: int [1:150001] 1 2 4 5 6 7 7 7 10 12 ...
## .. ..$ Resample06: int [1:150001] 2 4 4 7 7 8 8 9 10 10 ...
## .. ..$ Resample07: int [1:150001] 1 1 3 5 7 8 9 10 12 12 ...
## .. ..$ Resample08: int [1:150001] 3 3 5 5 6 7 8 10 11 13 ...
## .. ..$ Resample09: int [1:150001] 1 4 4 6 7 8 10 10 11 12 ...
## .. ..$ Resample10: int [1:150001] 1 1 2 2 4 7 8 8 9 9 ...
## .. ..$ Resample11: int [1:150001] 3 6 7 8 8 9 9 11 13 13 ...
## .. ..$ Resample12: int [1:150001] 1 2 3 4 5 5 7 8 8 9 ...
## .. ..$ Resample13: int [1:150001] 1 1 3 4 4 4 7 8 9 13 ...
## .. ..$ Resample14: int [1:150001] 1 3 6 7 8 9 10 11 11 11 ...
## .. ..$ Resample15: int [1:150001] 1 2 5 5 6 7 7 8 9 9 ...
## .. ..$ Resample16: int [1:150001] 1 1 2 5 6 7 8 8 10 10 ...
## .. ..$ Resample17: int [1:150001] 2 2 2 7 7 7 8 8 8 10 ...
## .. ..$ Resample18: int [1:150001] 1 1 1 3 3 5 6 7 8 10 ...
## .. ..$ Resample19: int [1:150001] 1 3 3 5 5 6 6 8 10 11 ...
## .. ..$ Resample20: int [1:150001] 1 1 2 3 4 5 5 7 8 8 ...
## .. ..$ Resample21: int [1:150001] 4 4 5 5 5 6 7 9 13 14 ...
## .. ..$ Resample22: int [1:150001] 2 2 2 2 3 3 4 5 6 6 ...
## .. ..$ Resample23: int [1:150001] 3 3 3 4 5 6 6 8 9 10 ...
## .. ..$ Resample24: int [1:150001] 3 4 5 8 8 8 9 9 10 10 ...

```

```

## ..$ Resample25: int [1:150001] 1 2 3 3 4 4 6 6 6 7 ...
## ..$ indexOut      :List of 25
## ..$ Resample01: int [1:55164] 2 7 13 16 17 18 22 23 25 27 ...
## ..$ Resample02: int [1:55251] 10 12 18 21 22 24 26 29 34 38 ...
## ..$ Resample03: int [1:55114] 2 5 8 11 12 17 19 20 24 27 ...
## ..$ Resample04: int [1:55199] 1 2 3 4 12 14 17 19 20 21 ...
## ..$ Resample05: int [1:55255] 3 8 9 11 13 14 17 18 27 37 ...
## ..$ Resample06: int [1:55282] 1 3 5 6 11 15 16 17 21 22 ...
## ..$ Resample07: int [1:55218] 2 4 6 11 13 14 17 18 19 21 ...
## ..$ Resample08: int [1:55106] 1 2 4 9 12 16 17 22 26 28 ...
## ..$ Resample09: int [1:55177] 2 3 5 9 13 15 17 21 22 23 ...
## ..$ Resample10: int [1:55208] 3 5 6 10 14 19 21 25 30 32 ...
## ..$ Resample11: int [1:55266] 1 2 4 5 10 12 14 15 19 21 ...
## ..$ Resample12: int [1:55264] 6 12 13 20 21 22 28 29 31 38 ...
## ..$ Resample13: int [1:55040] 2 5 6 10 11 12 19 23 25 26 ...
## ..$ Resample14: int [1:55346] 2 4 5 16 17 18 25 27 28 29 ...
## ..$ Resample15: int [1:55360] 3 4 13 14 16 24 30 31 33 34 ...
## ..$ Resample16: int [1:55250] 3 4 9 11 16 19 28 31 32 33 ...
## ..$ Resample17: int [1:55208] 1 3 4 5 6 9 11 20 28 31 ...
## ..$ Resample18: int [1:55349] 2 4 9 13 20 21 24 25 26 32 ...
## ..$ Resample19: int [1:55278] 2 4 7 9 15 18 19 28 35 38 ...
## ..$ Resample20: int [1:54992] 6 9 10 13 15 16 18 20 24 25 ...
## ..$ Resample21: int [1:55100] 1 2 3 8 10 11 12 18 20 22 ...
## ..$ Resample22: int [1:55310] 1 7 8 9 10 11 24 25 26 27 ...
## ..$ Resample23: int [1:55188] 1 2 7 11 18 19 25 30 32 33 ...
## ..$ Resample24: int [1:55035] 1 2 6 7 12 13 20 21 23 26 ...
## ..$ Resample25: int [1:55105] 5 8 10 12 14 17 18 20 22 25 ...
## ..$ timingSamps      : num 0
## ..$ predictionBounds : logi [1:2] FALSE FALSE
## ..$ seeds             : logi NA
## ..$ allowParallel     : logi TRUE
## $ finalModel :List of 33
## ..$ coefficients      : Named num [1:31] -1.329674 0.000554 -0.018491 -0.026153 0.003207 ...
## ..$ - attr(*, "names")= chr [1:31] "(Intercept)" "DER_mass_MMC" "DER_mass_transverse_met_lep" "DER_
## ..$ residuals         : Named num [1:150001] -2.3 -1.11 -1.48 -2.36 -1.37 ...
## ..$ - attr(*, "names")= chr [1:150001] "2" "3" "6" "12" ...
## ..$ fitted.values     : Named num [1:150001] 0.566 0.101 0.323 0.577 0.27 ...
## ..$ - attr(*, "names")= chr [1:150001] "2" "3" "6" "12" ...
## ..$ effects           : Named num [1:150001] 75.3 69.2 -87.9 -17.4 22.9 ...
## ..$ - attr(*, "names")= chr [1:150001] "(Intercept)" "DER_mass_MMC" "DER_mass_transverse_met_lep"
## ..$ R                  : num [1:31, 1:31] -158 0 0 0 0 ...
## ..$ - attr(*, "dimnames")=List of 2
## ..$ : chr [1:31] "(Intercept)" "DER_mass_MMC" "DER_mass_transverse_met_lep" "DER_mass_vis" .
## ..$ : chr [1:31] "(Intercept)" "DER_mass_MMC" "DER_mass_transverse_met_lep" "DER_mass_vis" .
## ..$ rank               : int 31
## ..$ qr                 :List of 5
## ..$ qr : num [1:150001, 1:31] -1.58e+02 1.91e-03 2.96e-03 3.13e-03 2.81e-03 ...
## ..$ - attr(*, "dimnames")=List of 2
## ..$ : chr [1:150001] "2" "3" "6" "12" ...
## ..$ : chr [1:31] "(Intercept)" "DER_mass_MMC" "DER_mass_transverse_met_lep" "DER_mass_vis"
## ..$ rank : int 31
## ..$ qraux: num [1:31] 1 1.01 1 1 1 ...
## ..$ pivot: int [1:31] 1 2 3 4 5 6 7 8 9 10 ...
## ..$ tol : num 1e-11

```

```

## ..- attr(*, "class")= chr "qr"
## ..$ family :List of 12
## ..$ family : chr "binomial"
## ..$ link : chr "logit"
## ..$ linkfun :function (mu)
## ..$ linkinv :function (eta)
## ..$ variance :function (mu)
## ..$ dev.resids:function (y, mu, wt)
## ..$ aic :function (y, n, mu, wt, dev)
## ..$ mu.eta :function (eta)
## ..$ initialize: expression({ if (NCOL(y) == 1) { if (is.factor(y)) y
## ..$ validmu :function (mu)
## ..$ valideta :function (eta)
## ..$ simulate :function (object, nsim)
## ..- attr(*, "class")= chr "family"
## ..$ linear.predictors: Named num [1:150001] 0.264 -2.183 -0.741 0.309 -0.996 ...
## ..- attr(*, "names")= chr [1:150001] "2" "3" "6" "12" ...
## ..$ deviance : num 149244
## ..$ aic : num 149306
## ..$ null.deviance : num 192839
## ..$ iter : int 5
## ..$ weights : Named num [1:150001] 0.2457 0.0911 0.2186 0.2441 0.197 ...
## ..- attr(*, "names")= chr [1:150001] "2" "3" "6" "12" ...
## ..$ prior.weights : Named num [1:150001] 1 1 1 1 1 1 1 1 1 1 ...
## ..- attr(*, "names")= chr [1:150001] "2" "3" "6" "12" ...
## ..$ df.residual : int 149970
## ..$ df.null : int 150000
## ..$ y : Named num [1:150001] 0 0 0 0 0 0 0 0 0 0 ...
## ..- attr(*, "names")= chr [1:150001] "2" "3" "6" "12" ...
## ..$ converged : logi TRUE
## ..$ boundary : logi FALSE
## ..$ model :'data.frame': 150001 obs. of 31 variables:
## ..$ .outcome : Factor w/ 2 levels "b","s": 1 1 1 1 1 1 1 1 1 1 ...
## ..$ DER_mass_MMC : num [1:150001] 160.9 -999 89.7 114.7 82.5 ...
## ..$ DER_mass_transverse_met_lep: num [1:150001] 68.8 162.2 13.6 10.3 31.7 ...
## ..$ DER_mass_vis : num [1:150001] 103.2 126 59.1 75.7 64.1 ...
## ..$ DER_pt_h : num [1:150001] 48.15 35.63 116.34 30.82 8.23 ...
## ..$ DER_deltaeta_jet_jet : num [1:150001] -999 -999 2.64 2.56 -999 ...
## ..$ DER_mass_jet_jet : num [1:150001] -999 -999 285 253 -999 ...
## ..$ DER_prodetta_jet_jet : num [1:150001] -999 -999 -0.54 -1.4 -999 ...
## ..$ DER_deltar_tau_lep : num [1:150001] 3.47 3.15 1.36 2.89 2.82 ...
## ..$ DER_pt_tot : num [1:150001] 2.08 9.34 61.62 36.74 8.23 ...
## ..$ DER_sum_pt : num [1:150001] 125.2 197.8 278.9 239.8 58.6 ...
## ..$ DER_pt_ratio_lep_tau : num [1:150001] 0.879 3.776 0.588 1.061 1.303 ...
## ..$ DER_met_phi_centrality : num [1:150001] 1.414 1.414 0.479 1.364 -1.414 ...
## ..$ DER_lep_eta_centrality : num [1:150001] -999 -999 0.975 0.769 -999 -999 -999 -999 -999 ...
## ..$ PRI_tau_pt : num [1:150001] 42 32.2 53.7 36 25.5 ...
## ..$ PRI_tau_eta : num [1:150001] 2.039 -0.705 0.371 -0.669 -0.654 ...
## ..$ PRI_tau_phi : num [1:150001] -3.011 -2.093 1.329 -0.342 -2.99 ...
## ..$ PRI_lep_pt : num [1:150001] 36.9 121.4 31.6 38.2 33.2 ...
## ..$ PRI_lep_eta : num [1:150001] 0.501 -0.953 -0.884 -0.165 -1.665 ...
## ..$ PRI_lep_phi : num [1:150001] 0.103 1.052 1.857 2.502 -0.354 ...
## ..$ PRI_met : num [1:150001] 44.7 54.3 40.7 22.4 12.4 ...
## ..$ PRI_met_phi : num [1:150001] -1.92 -2.19 2.24 2.15 1.43 ...

```

```

## ..$ PRI_met_sumet : num [1:150001] 165 260 283 291 163 ...
## ..$ PRI_jet_num : int [1:150001] 1 1 3 3 0 0 1 0 0 0 ...
## ..$ PRI_jet_leading_pt : num [1:150001] 46.2 44.3 90.5 76.8 -999 ...
## ..$ PRI_jet_leading_eta : num [1:150001] 0.725 2.053 -2.412 -0.79 -999 ...
## ..$ PRI_jet_leading_phi : num [1:150001] 1.158 -2.028 -0.653 0.303 -999 ...
## ..$ PRI_jet_subleading_pt : num [1:150001] -999 -999 56.2 56.9 -999 ...
## ..$ PRI_jet_subleading_eta : num [1:150001] -999 -999 0.224 1.773 -999 ...
## ..$ PRI_jet_subleading_phi : num [1:150001] -999 -999 3.11 -2.08 -999 ...
## ..$ PRI_jet_all_pt : num [1:150001] 46.2 44.3 193.7 165.6 0 ...
## ..- attr(*, "terms")=Classes 'terms', 'formula' length 3 .outcome ~ DER_mass_MMC + DER_mass_transverse_met_lep
## ..- attr(*, "variables")= language list(.outcome, DER_mass_MMC, DER_mass_transverse_met_lep, DER_pt_h)
## ..- attr(*, "factors")= int [1:31, 1:30] 0 1 0 0 0 0 0 0 0 0 ...
## ..- attr(*, "dimnames")=List of 2
## ..$ : chr [1:31] ".outcome" "DER_mass_MMC" "DER_mass_transverse_met_lep" "DER_pt_h"
## ..$ : chr [1:30] "DER_mass_MMC" "DER_mass_transverse_met_lep" "DER_pt_h" "DER_deltaeta_jet_jet"
## ..- attr(*, "term.labels")= chr [1:30] "DER_mass_MMC" "DER_mass_transverse_met_lep" "DER_pt_h" "DER_deltaeta_jet_jet"
## ..- attr(*, "order")= int [1:30] 1 1 1 1 1 1 1 1 1 1 ...
## ..- attr(*, "intercept")= int 1
## ..- attr(*, "response")= int 1
## ..- attr(*, ".Environment")=<environment: 0x7fc0e5f91e78>
## ..- attr(*, "predvars")= language list(.outcome, DER_mass_MMC, DER_mass_transverse_met_lep, DER_pt_h)
## ..- attr(*, "dataClasses")= Named chr [1:31] "factor" "numeric" "numeric" "numeric" ...
## ..- attr(*, "names")= chr [1:31] ".outcome" "DER_mass_MMC" "DER_mass_transverse_met_lep" "DER_pt_h"
## ..$ formula :Class 'formula' length 3 .outcome ~ .
## ..- attr(*, ".Environment")=<environment: 0x7fc0e5f91e78>
## ..$ terms :Classes 'terms', 'formula' length 3 .outcome ~ DER_mass_MMC + DER_mass_transverse_met_lep
## ..- attr(*, "variables")= language list(.outcome, DER_mass_MMC, DER_mass_transverse_met_lep, DER_pt_h)
## ..- attr(*, "factors")= int [1:31, 1:30] 0 1 0 0 0 0 0 0 0 0 ...
## ..- attr(*, "dimnames")=List of 2
## ..$ : chr [1:31] ".outcome" "DER_mass_MMC" "DER_mass_transverse_met_lep" "DER_pt_h"
## ..$ : chr [1:30] "DER_mass_MMC" "DER_mass_transverse_met_lep" "DER_pt_h" "DER_deltaeta_jet_jet"
## ..- attr(*, "term.labels")= chr [1:30] "DER_mass_MMC" "DER_mass_transverse_met_lep" "DER_pt_h" "DER_deltaeta_jet_jet"
## ..- attr(*, "order")= int [1:30] 1 1 1 1 1 1 1 1 1 1 ...
## ..- attr(*, "intercept")= int 1
## ..- attr(*, "response")= int 1
## ..- attr(*, ".Environment")=<environment: 0x7fc0e5f91e78>
## ..- attr(*, "predvars")= language list(.outcome, DER_mass_MMC, DER_mass_transverse_met_lep, DER_pt_h)
## ..- attr(*, "dataClasses")= Named chr [1:31] "factor" "numeric" "numeric" "numeric" ...
## ..- attr(*, "names")= chr [1:31] ".outcome" "DER_mass_MMC" "DER_mass_transverse_met_lep" "DER_pt_h"
## ..$ data :'data.frame': 150001 obs. of 31 variables:
## ..$ DER_mass_MMC : num [1:150001] 160.9 -999 89.7 114.7 82.5 ...
## ..$ DER_mass_transverse_met_lep : num [1:150001] 68.8 162.2 13.6 10.3 31.7 ...
## ..$ DER_mass_vis : num [1:150001] 103.2 126 59.1 75.7 64.1 ...
## ..$ DER_pt_h : num [1:150001] 48.15 35.63 116.34 30.82 8.23 ...
## ..$ DER_deltaeta_jet_jet : num [1:150001] -999 -999 2.64 2.56 -999 ...
## ..$ DER_mass_jet_jet : num [1:150001] -999 -999 285 253 -999 ...
## ..$ DER_prodetajet : num [1:150001] -999 -999 -0.54 -1.4 -999 ...
## ..$ DER_deltatarau : num [1:150001] 3.47 3.15 1.36 2.89 2.82 ...
## ..$ DER_pt_tot : num [1:150001] 2.08 9.34 61.62 36.74 8.23 ...
## ..$ DER_sum_pt : num [1:150001] 125.2 197.8 278.9 239.8 58.6 ...
## ..$ DER_pt_ratio_lep_tau : num [1:150001] 0.879 3.776 0.588 1.061 1.303 ...
## ..$ DER_met_phi centrality : num [1:150001] 1.414 1.414 0.479 1.364 -1.414 ...
## ..$ DER_lep_eta centrality : num [1:150001] -999 -999 0.975 0.769 -999 -999 -999 -999 -999
## ..$ PRI_tau_pt : num [1:150001] 42 32.2 53.7 36 25.5 ...

```

```

## ..$ PRI_tau_eta : num [1:150001] 2.039 -0.705 0.371 -0.669 -0.654 ...
## ..$ PRI_tau_phi : num [1:150001] -3.011 -2.093 1.329 -0.342 -2.99 ...
## ..$ PRI_lep_pt : num [1:150001] 36.9 121.4 31.6 38.2 33.2 ...
## ..$ PRI_lep_eta : num [1:150001] 0.501 -0.953 -0.884 -0.165 -1.665 ...
## ..$ PRI_lep_phi : num [1:150001] 0.103 1.052 1.857 2.502 -0.354 ...
## ..$ PRI_met : num [1:150001] 44.7 54.3 40.7 22.4 12.4 ...
## ..$ PRI_met_phi : num [1:150001] -1.92 -2.19 2.24 2.15 1.43 ...
## ..$ PRI_met_sumet : num [1:150001] 165 260 283 291 163 ...
## ..$ PRI_jet_num : int [1:150001] 1 1 3 3 0 0 1 0 0 0 ...
## ..$ PRI_jet_leading_pt : num [1:150001] 46.2 44.3 90.5 76.8 -999 ...
## ..$ PRI_jet_leading_eta : num [1:150001] 0.725 2.053 -2.412 -0.79 -999 ...
## ..$ PRI_jet_leading_phi : num [1:150001] 1.158 -2.028 -0.653 0.303 -999 ...
## ..$ PRI_jet_subleading_pt : num [1:150001] -999 -999 56.2 56.9 -999 ...
## ..$ PRI_jet_subleading_eta : num [1:150001] -999 -999 0.224 1.773 -999 ...
## ..$ PRI_jet_subleading_phi : num [1:150001] -999 -999 3.11 -2.08 -999 ...
## ..$ PRI_jet_all_pt : num [1:150001] 46.2 44.3 193.7 165.6 0 ...
## ..$ .outcome : Factor w/ 2 levels "b","s": 1 1 1 1 1 1 1 1 1 1 ...
## ..$ offset : NULL
## ..$ control :List of 3
## ..$ epsilon: num 1e-08
## ..$ maxit : num 25
## ..$ trace : logi FALSE
## ..$ method : chr "glm.fit"
## ..$ contrasts : NULL
## ..$ xlevels : Named list()
## ..$ xNames : chr [1:30] "DER_mass_MMC" "DER_mass_transverse_met_lep" "DER_mass_vis" "DER_
## ..$ problemType : chr "Classification"
## ..$ tuneValue : 'data.frame': 1 obs. of 1 variable:
## ..$ parameter: Factor w/ 1 level "none": 1
## ..$ obsLevels : chr [1:2] "b" "s"
## ..$ attr(*, "class")= chr [1:2] "glm" "lm"
## $ preprocess : NULL
## $ trainingData: 'data.frame': 150001 obs. of 31 variables:
## ..$ DER_mass_MMC : num [1:150001] 160.9 -999 89.7 114.7 82.5 ...
## ..$ DER_mass_transverse_met_lep: num [1:150001] 68.8 162.2 13.6 10.3 31.7 ...
## ..$ DER_mass_vis : num [1:150001] 103.2 126 59.1 75.7 64.1 ...
## ..$ DER_pt_h : num [1:150001] 48.15 35.63 116.34 30.82 8.23 ...
## ..$ DER_deltaeta_jet_jet : num [1:150001] -999 -999 2.64 2.56 -999 ...
## ..$ DER_mass_jet_jet : num [1:150001] -999 -999 285 253 -999 ...
## ..$ DER_prodelta_jet_jet : num [1:150001] -999 -999 -0.54 -1.4 -999 ...
## ..$ DER_deltar_tau_lep : num [1:150001] 3.47 3.15 1.36 2.89 2.82 ...
## ..$ DER_pt_tot : num [1:150001] 2.08 9.34 61.62 36.74 8.23 ...
## ..$ DER_sum_pt : num [1:150001] 125.2 197.8 278.9 239.8 58.6 ...
## ..$ DER_pt_ratio_lep_tau : num [1:150001] 0.879 3.776 0.588 1.061 1.303 ...
## ..$ DER_met_phi_central : num [1:150001] 1.414 1.414 0.479 1.364 -1.414 ...
## ..$ DER_lep_eta_central : num [1:150001] -999 -999 0.975 0.769 -999 -999 -999 -999 -999 -999
## ..$ PRI_tau_pt : num [1:150001] 42 32.2 53.7 36 25.5 ...
## ..$ PRI_tau_eta : num [1:150001] 2.039 -0.705 0.371 -0.669 -0.654 ...
## ..$ PRI_tau_phi : num [1:150001] -3.011 -2.093 1.329 -0.342 -2.99 ...
## ..$ PRI_lep_pt : num [1:150001] 36.9 121.4 31.6 38.2 33.2 ...
## ..$ PRI_lep_eta : num [1:150001] 0.501 -0.953 -0.884 -0.165 -1.665 ...
## ..$ PRI_lep_phi : num [1:150001] 0.103 1.052 1.857 2.502 -0.354 ...
## ..$ PRI_met : num [1:150001] 44.7 54.3 40.7 22.4 12.4 ...
## ..$ PRI_met_phi : num [1:150001] -1.92 -2.19 2.24 2.15 1.43 ...

```

```

## ..$ PRI_met_sumet : num [1:150001] 165 260 283 291 163 ...
## ..$ PRI_jet_num : int [1:150001] 1 1 3 3 0 0 1 0 0 0 ...
## ..$ PRI_jet_leading_pt : num [1:150001] 46.2 44.3 90.5 76.8 -999 ...
## ..$ PRI_jet_leading_eta : num [1:150001] 0.725 2.053 -2.412 -0.79 -999 ...
## ..$ PRI_jet_leading_phi : num [1:150001] 1.158 -2.028 -0.653 0.303 -999 ...
## ..$ PRI_jet_subleading_pt : num [1:150001] -999 -999 56.2 56.9 -999 ...
## ..$ PRI_jet_subleading_eta : num [1:150001] -999 -999 0.224 1.773 -999 ...
## ..$ PRI_jet_subleading_phi : num [1:150001] -999 -999 3.11 -2.08 -999 ...
## ..$ PRI_jet_all_pt : num [1:150001] 46.2 44.3 193.7 165.6 0 ...
## ..$ .outcome : Factor w/ 2 levels "b","s": 1 1 1 1 1 1 1 1 1 1 ...
## $ resample : 'data.frame': 25 obs. of 3 variables:
## ..$ Accuracy: num [1:25] 0.75 0.751 0.75 0.75 0.747 ...
## ..$ Kappa : num [1:25] 0.419 0.418 0.418 0.417 0.412 ...
## ..$ Resample: chr [1:25] "Resample01" "Resample02" "Resample03" "Resample04" ...
## $ resampledCM : 'data.frame': 25 obs. of 6 variables:
## ..$ cell1 : num [1:25] 31185 31369 31149 31343 31202 ...
## ..$ cell2 : num [1:25] 5093 4936 5047 4913 5071 ...
## ..$ cell3 : num [1:25] 8677 8839 8747 8876 8890 ...
## ..$ cell4 : num [1:25] 10209 10107 10171 10067 10092 ...
## ..$ parameter: Factor w/ 1 level "none": 1 1 1 1 1 1 1 1 1 1 ...
## ..$ Resample : chr [1:25] "Resample01" "Resample02" "Resample03" "Resample04" ...
## $ perfNames : chr [1:2] "Accuracy" "Kappa"
## $ maximize : logi TRUE
## $ yLimits : NULL
## $ times :List of 3
## ..$ everything:Class 'proc_time' Named num [1:5] 14.301 0.843 160.291 471.802 33.81
## .. ..- attr(*, "names")= chr [1:5] "user.self" "sys.self" "elapsed" "user.child" ...
## ..$ final :Class 'proc_time' Named num [1:5] 12.668 0.491 13.242 0 0
## .. ..- attr(*, "names")= chr [1:5] "user.self" "sys.self" "elapsed" "user.child" ...
## ..$ prediction: logi [1:3] NA NA NA
## - attr(*, "class")= chr "train"

```

```
glm.fit$finalModel
```

```

##
## Call: NULL
##
## Coefficients:
## (Intercept) DER_mass_MMC
## -1.329674 0.000554
## DER_mass_transverse_met_lep DER_mass_vis
## -0.018491 -0.026153
## DER_pt_h DER_deltaeta_jet_jet
## 0.003207 -0.172408
## DER_mass_jet_jet DER_prodeteta_jet_jet
## 0.002505 -0.042881
## DER_deltar_tau_lep DER_pt_tot
## 1.304524 -0.000245
## DER_sum_pt DER_pt_ratio_lep_tau
## 0.584055 -0.902684
## DER_met_phi_centralty DER_lep_eta_centralty
## 0.190496 0.264355
## PRI_tau_pt PRI_tau_eta
## -0.560305 -0.004847

```



```
##          PRI_tau_phi          PRI_lep_pt
##          -0.001217          -0.537755
##          PRI_lep_eta          PRI_lep_phi
##          0.006568          0.001984
##          PRI_met          PRI_met_phi
##          0.007278          -0.003898
##          PRI_met_sumet          PRI_jet_num
##          -0.001644          -0.465372
##          PRI_jet_leading_pt          PRI_jet_leading_eta
##          0.000530          -0.000999
##          PRI_jet_leading_phi          PRI_jet_subleading_pt
##          0.001325          -0.002614
##          PRI_jet_subleading_eta          PRI_jet_subleading_phi
##          -0.012912          -0.035235
##          PRI_jet_all_pt
##          -0.588419
##
## Degrees of Freedom: 150000 Total (i.e. Null); 149970 Residual
## Null Deviance: 193000
## Residual Deviance: 149000 AIC: 149000
```

```
varImp(glm.fit)
```

```
## glm variable importance
##
## only 20 most important variables shown (out of 30)
##
## Overall
## DER_deltar_tau_lep 100.00
## DER_mass_transverse_met_lep 97.45
## DER_mass_vis 90.41
## DER_pt_ratio_lep_tau 62.94
## DER_mass_jet_jet 40.82
## DER_met_phi_centrality 40.64
## DER_mass MMC 28.54
## PRI_met 26.55
## PRI_met_sumet 18.88
## DER_lep_eta_centrality 18.12
## PRI_jet_num 15.81
## DER_deltaeta_jet_jet 14.42
## DER_pt_h 12.53
## DER_prodeteta_jet_jet 11.22
## PRI_jet_subleading_phi 8.15
## PRI_jet_subleading_pt 3.97
## PRI_jet_subleading_eta 3.22
## PRI_met_phi 1.55
## PRI_lep_eta 1.40
## PRI_jet_leading_pt 1.02
```

```
pred.val <- predict(glm.fit, validation)
confusionMatrix(pred.val, validation$Label)
```

```
## Warning: NAs produced by integer overflow
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    b    s
##           b 56884 15957
##           s  8849 18309
##
##           Accuracy : 0.752
##           95% CI : (0.749, 0.755)
##       No Information Rate : 0.657
##       P-Value [Acc > NIR] : <2e-16
##
##           Kappa : NA
##  McNemar's Test P-Value : <2e-16
##
##       Sensitivity : 0.865
##       Specificity : 0.534
##       Pos Pred Value : 0.781
##       Neg Pred Value : 0.674
##       Prevalence : 0.657
##       Detection Rate : 0.569
##       Detection Prevalence : 0.728
##       Balanced Accuracy : 0.700
##
##       'Positive' Class : b
##
```

```
pred.test <- predict(glm.fit, test, type="prob")
pred.final <- PrepPrediction(pred.test, test)
WriteSubmission(pred.final, 11)
```

Results

2.00083

Experimentation

```
predictors <- training[,setdiff(names(training),c("EventId","Label"))]
registerDoMC(cores=4)
glm.pca <- train(x=predictors, y=training$Label, method="glm", preProcess="pca")
```

```
glm.pca
```

```
## Generalized Linear Model
##
## 150001 samples
##    30 predictors
##    2 classes: 'b', 's'
##
## Pre-processing: principal component signal extraction, scaled, centered
```

```
## Resampling: Bootstrapped (25 reps)
##
## Summary of sample sizes: 150001, 150001, 150001, 150001, 150001, 150001, ...
##
## Resampling results
##
##   Accuracy   Kappa   Accuracy SD   Kappa SD
##   0.7        0.3     0.002         0.005
##
##

pred.pca.val <- predict(glm.pca, validation[,setdiff(names(training),c("EventId","Label"))])
confusionMatrix(pred.pca.val, validation$Label)

## Warning: NAs produced by integer overflow

## Confusion Matrix and Statistics
##
##           Reference
## Prediction      b      s
##           b 55997 18445
##           s  9736 15821
##
##           Accuracy : 0.718
##           95% CI : (0.715, 0.721)
##           No Information Rate : 0.657
##           P-Value [Acc > NIR] : <2e-16
##
##           Kappa : NA
##           McNemar's Test P-Value : <2e-16
##
##           Sensitivity : 0.852
##           Specificity : 0.462
##           Pos Pred Value : 0.752
##           Neg Pred Value : 0.619
##           Prevalence : 0.657
##           Detection Rate : 0.560
##           Detection Prevalence : 0.744
##           Balanced Accuracy : 0.657
##
##           'Positive' Class : b
##
```