submission 8

Shannon Rush

July 17, 2014

Submission 8 - Bin on jets, random forest on bins

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")
```

Clean and Transform Data

```
    set.seed(123)
    training.indices <- createDataPartition(train$Label, p=0.6, list=F)
    training <- train[training.indices,]
    validation <- train[-training.indices,]
```

```
    train.jets <- split(train, train$PRI_jet_num)
    for (i in 0:3) { assign(paste0("train.jets",i),train.jets[[i+1]]) }
    val.jets <- split(validation, validation$PRI_jet_num)
    for (i in 0:3) { assign(paste0("val.jets",i),val.jets[[i+1]]) }
    test.jets <- split(test, test$PRI_jet_num)
    for (i in 0:3) { assign(paste0("test.jets",i),test.jets[[i+1]]) }
```

0 Jets

```
    summary(train.jets0)
```

```
##     EventId        DER_mass_MMC    DER_mass_transverse_met_lep
##  Min.   :100003   Min.   :-999.0   Min.   :  0.0
##  1st Qu.:162434   1st Qu.:-999.0   1st Qu.: 35.1
##  Median :224898   Median :  96.5   Median : 62.1
##  Mean   :224893   Mean   :-172.1   Mean   : 58.8
##  3rd Qu.:287392   3rd Qu.: 126.8   3rd Qu.: 79.7
##  Max.   :349999   Max.   : 863.6   Max.   :570.1
##   DER_mass_vis       DER_pt_h      DER_deltaeta_jet_jet DER_mass_jet_jet
##  Min.   :   7.1   Min.   :   0.0   Min.   :-999         Min.   :-999
##  1st Qu.:  60.9   1st Qu.:   2.3   1st Qu.:-999         1st Qu.:-999
##  Median :  75.5   Median :   6.7   Median :-999         Median :-999
##  Mean   :  81.9   Mean   :  13.8   Mean   :-999         Mean   :-999
##  3rd Qu.:  94.1   3rd Qu.:  24.2   3rd Qu.:-999         3rd Qu.:-999
##  Max.   :1349.4   Max.   :2835.0   Max.   :-999         Max.   :-999
##  DER_prodeta_jet_jet DER_deltar_tau_lep   DER_pt_tot       DER_sum_pt
##  Min.   :-999        Min.   :0.277      Min.   :   0.0   Min.   :  46.1
##  1st Qu.:-999        1st Qu.:2.338      1st Qu.:   2.3   1st Qu.:  61.3
##  Median :-999        Median :2.822      Median :   6.7   Median :  71.6
##  Mean   :-999        Mean   :2.665      Mean   :  13.8   Mean   :  76.4
##  3rd Qu.:-999        3rd Qu.:3.095      3rd Qu.:  24.2   3rd Qu.:  85.2
##  Max.   :-999        Max.   :5.684      Max.   :2835.0   Max.   :1324.7
##  DER_pt_ratio_lep_tau DER_met_phi_centrality DER_lep_eta_centrality
##  Min.   : 0.127       Min.   :-1.41          Min.   :-999
##  1st Qu.: 0.968       1st Qu.:-1.40          1st Qu.:-999
##  Median : 1.318       Median :-1.36          Median :-999
##  Mean   : 1.393       Mean   :-0.91          Mean   :-999
##  3rd Qu.: 1.728       3rd Qu.:-1.12          3rd Qu.:-999
##  Max.   :10.571       Max.   : 1.41          Max.   :-999
##    PRI_tau_pt     PRI_tau_eta       PRI_tau_phi        PRI_lep_pt
##  Min.   : 20.0   Min.   :-2.4990   Min.   :-3.1420   Min.   : 26.0
##  1st Qu.: 23.7   1st Qu.:-0.9600   1st Qu.:-1.5860   1st Qu.: 31.9
##  Median : 29.3   Median :-0.0430   Median :-0.0490   Median : 39.0
##  Mean   : 34.0   Mean   :-0.0249   Mean   :-0.0157   Mean   : 42.4
##  3rd Qu.: 39.8   3rd Qu.: 0.9000   3rd Qu.: 1.5560   3rd Qu.: 49.0
##  Max.   :764.4   Max.   : 2.4940   Max.   : 3.1420   Max.   :560.3
##   PRI_lep_eta       PRI_lep_phi         PRI_met        PRI_met_phi
##  Min.   :-2.5050   Min.   :-3.1420   Min.   :   0.1   Min.   :-3.1420
##  1st Qu.:-1.1210   1st Qu.:-1.5260   1st Qu.:  18.0   1st Qu.:-1.5830
##  Median :-0.1040   Median : 0.0850   Median :  29.5   Median :-0.0510
##  Mean   :-0.0523   Mean   : 0.0424   Mean   :  31.5   Mean   :-0.0244
##  3rd Qu.: 0.9890   3rd Qu.: 1.6210   3rd Qu.:  42.3   3rd Qu.: 1.5460
##  Max.   : 2.5030   Max.   : 3.1420   Max.   :2842.6   Max.   : 3.1420
##  PRI_met_sumet     PRI_jet_num PRI_jet_leading_pt PRI_jet_leading_eta
##  Min.   :  13.7   Min.   :0    Min.   :-999       Min.   :-999
##  1st Qu.:  88.0   1st Qu.:0    1st Qu.:-999       1st Qu.:-999
##  Median : 119.5   Median :0    Median :-999       Median :-999
##  Mean   : 125.9   Mean   :0    Mean   :-999       Mean   :-999
##  3rd Qu.: 156.4   3rd Qu.:0    3rd Qu.:-999       3rd Qu.:-999
##  Max.   :1391.5   Max.   :0    Max.   :-999       Max.   :-999
##  PRI_jet_leading_phi PRI_jet_subleading_pt PRI_jet_subleading_eta
##  Min.   :-999        Min.   :-999          Min.   :-999
##  1st Qu.:-999        1st Qu.:-999          1st Qu.:-999
##  Median :-999        Median :-999          Median :-999
##  Mean   :-999        Mean   :-999          Mean   :-999
```

```
##  3rd Qu.:-999        3rd Qu.:-999          3rd Qu.:-999
##  Max.   :-999        Max.   :-999          Max.   :-999
##  PRI_jet_subleading_phi PRI_jet_all_pt Label
##  Min.   :-999           Min.   :0      b:74421
##  1st Qu.:-999           1st Qu.:0      s:25492
##  Median :-999           Median :0
##  Mean   :-999           Mean   :0
##  3rd Qu.:-999           3rd Qu.:0
##  Max.   :-999           Max.   :0
```

Since we're going to train 4 separate random forests based on jet bin we can eliminate or transform variables in each group separately.

A number of variables are set to -999 when there are no jets present. These should be eliminated in all the jets0

```
    remove<-apply(train.jets0[1:nrow(train.jets0),]==-999, 2, all)
    train.j0 <- train.jets0[,!remove]
    val.j0 <- val.jets0[,!remove]
    test.j0 <- test.jets0[,c("EventId",setdiff(names(!remove),"Label"))]
```

Now let's see if there's any additional variables to eliminate

```
    summary(train.j0)
```

```
##     EventId        DER_mass_MMC    DER_mass_transverse_met_lep
##  Min.   :100003   Min.   :-999.0   Min.   :  0.0
##  1st Qu.:162434   1st Qu.:-999.0   1st Qu.: 35.1
##  Median :224898   Median :  96.5   Median : 62.1
##  Mean   :224893   Mean   :-172.1   Mean   : 58.8
##  3rd Qu.:287392   3rd Qu.: 126.8   3rd Qu.: 79.7
##  Max.   :349999   Max.   : 863.6   Max.   :570.1
##   DER_mass_vis       DER_pt_h      DER_deltar_tau_lep   DER_pt_tot
##  Min.   :   7.1   Min.   :   0.0   Min.   :0.277      Min.   :   0.0
##  1st Qu.:  60.9   1st Qu.:   2.3   1st Qu.:2.338      1st Qu.:   2.3
##  Median :  75.5   Median :   6.7   Median :2.822      Median :   6.7
##  Mean   :  81.9   Mean   :  13.8   Mean   :2.665      Mean   :  13.8
##  3rd Qu.:  94.1   3rd Qu.:  24.2   3rd Qu.:3.095      3rd Qu.:  24.2
##  Max.   :1349.4   Max.   :2835.0   Max.   :5.684      Max.   :2835.0
##    DER_sum_pt     DER_pt_ratio_lep_tau DER_met_phi_centrality
##  Min.   :  46.1   Min.   : 0.127       Min.   :-1.41
##  1st Qu.:  61.3   1st Qu.: 0.968       1st Qu.:-1.40
##  Median :  71.6   Median : 1.318       Median :-1.36
##  Mean   :  76.4   Mean   : 1.393       Mean   :-0.91
##  3rd Qu.:  85.2   3rd Qu.: 1.728       3rd Qu.:-1.12
##  Max.   :1324.7   Max.   :10.571       Max.   : 1.41
##    PRI_tau_pt     PRI_tau_eta       PRI_tau_phi        PRI_lep_pt
##  Min.   : 20.0   Min.   :-2.4990   Min.   :-3.1420   Min.   : 26.0
##  1st Qu.: 23.7   1st Qu.:-0.9600   1st Qu.:-1.5860   1st Qu.: 31.9
##  Median : 29.3   Median :-0.0430   Median :-0.0490   Median : 39.0
##  Mean   : 34.0   Mean   :-0.0249   Mean   :-0.0157   Mean   : 42.4
##  3rd Qu.: 39.8   3rd Qu.: 0.9000   3rd Qu.: 1.5560   3rd Qu.: 49.0
##  Max.   :764.4   Max.   : 2.4940   Max.   : 3.1420   Max.   :560.3
```

```
##   PRI_lep_eta       PRI_lep_phi         PRI_met        PRI_met_phi
##  Min.   :-2.5050   Min.   :-3.1420   Min.   :   0.1   Min.   :-3.1420
##  1st Qu.:-1.1210   1st Qu.:-1.5260   1st Qu.:  18.0   1st Qu.:-1.5830
##  Median :-0.1040   Median : 0.0850   Median :  29.5   Median :-0.0510
##  Mean   :-0.0523   Mean   : 0.0424   Mean   :  31.5   Mean   :-0.0244
##  3rd Qu.: 0.9890   3rd Qu.: 1.6210   3rd Qu.:  42.3   3rd Qu.: 1.5460
##  Max.   : 2.5030   Max.   : 3.1420   Max.   :2842.6   Max.   : 3.1420
##  PRI_met_sumet     PRI_jet_num PRI_jet_all_pt Label
##  Min.   :  13.7   Min.   :0    Min.   :0      b:74421
##  1st Qu.:  88.0   1st Qu.:0    1st Qu.:0      s:25492
##  Median : 119.5   Median :0    Median :0
##  Mean   : 125.9   Mean   :0    Mean   :0
##  3rd Qu.: 156.4   3rd Qu.:0    3rd Qu.:0
##  Max.   :1391.5   Max.   :0    Max.   :0
```

PRI_jet_all_pt is defined as all zeros, and of course so is PRI_jet_num so we'll eliminate those

```
keep <- setdiff(names(train.j0),c("PRI_jet_num","PRI_jet_all_pt"))
train.j0 <- train.j0[,keep]
val.j0 <- val.j0[,keep]
test.j0 <- test.j0[,c("EventId",setdiff(keep,"Label"))]
```

1 Jet

```
summary(train.jets1)
```

```
##     EventId        DER_mass_MMC    DER_mass_transverse_met_lep
##  Min.   :100001   Min.   :-999.0   Min.   :  0.0
##  1st Qu.:162818   1st Qu.:  84.5   1st Qu.: 16.3
##  Median :225050   Median : 107.9   Median : 40.5
##  Mean   :225144   Mean   :  12.8   Mean   : 46.1
##  3rd Qu.:287516   3rd Qu.: 132.9   3rd Qu.: 70.0
##  Max.   :349997   Max.   :1192.0   Max.   :571.9
##   DER_mass_vis      DER_pt_h     DER_deltaeta_jet_jet DER_mass_jet_jet
##  Min.   :  6.3   Min.   :  0.0   Min.   :-999         Min.   :-999
##  1st Qu.: 59.8   1st Qu.: 37.1   1st Qu.:-999         1st Qu.:-999
##  Median : 73.9   Median : 53.1   Median :-999         Median :-999
##  Mean   : 82.2   Mean   : 65.9   Mean   :-999         Mean   :-999
##  3rd Qu.: 92.7   3rd Qu.: 79.0   3rd Qu.:-999         3rd Qu.:-999
##  Max.   :959.6   Max.   :753.7   Max.   :-999         Max.   :-999
##  DER_prodeta_jet_jet DER_deltar_tau_lep   DER_pt_tot      DER_sum_pt
##  Min.   :-999        Min.   :0.208      Min.   :  0.0   Min.   :  77
##  1st Qu.:-999        1st Qu.:1.855      1st Qu.:  2.9   1st Qu.: 111
##  Median :-999        Median :2.404      Median : 10.7   Median : 132
##  Mean   :-999        Mean   :2.340      Mean   : 16.6   Mean   : 150
##  3rd Qu.:-999        3rd Qu.:2.855      3rd Qu.: 26.2   3rd Qu.: 168
##  Max.   :-999        Max.   :5.655      Max.   :330.5   Max.   :1215
##  DER_pt_ratio_lep_tau DER_met_phi_centrality DER_lep_eta_centrality
##  Min.   : 0.083       Min.   :-1.414         Min.   :-999
##  1st Qu.: 0.868       1st Qu.:-0.975         1st Qu.:-999
##  Median : 1.268       Median : 0.626         Median :-999
```

```
##  Mean   : 1.444       Mean   : 0.236         Mean   :-999
##  3rd Qu.: 1.797       3rd Qu.: 1.324         3rd Qu.:-999
##  Max.   :16.776       Max.   : 1.414         Max.   :-999
##    PRI_tau_pt      PRI_tau_eta       PRI_tau_phi        PRI_lep_pt
##  Min.   : 20.0   Min.   :-2.4980   Min.   :-3.1410   Min.   : 26.0
##  1st Qu.: 24.8   1st Qu.:-0.9220   1st Qu.:-1.5650   1st Qu.: 32.4
##  Median : 32.2   Median :-0.0110   Median :-0.0330   Median : 40.6
##  Mean   : 38.6   Mean   :-0.0017   Mean   :-0.0073   Mean   : 46.8
##  3rd Qu.: 45.3   3rd Qu.: 0.9110   3rd Qu.: 1.5660   3rd Qu.: 54.1
##  Max.   :505.1   Max.   : 2.4970   Max.   : 3.1410   Max.   :426.4
##   PRI_lep_eta       PRI_lep_phi         PRI_met       PRI_met_phi
##  Min.   :-2.4940   Min.   :-3.1420   Min.   :  0.3   Min.   :-3.1410
##  1st Qu.:-0.9750   1st Qu.:-1.5220   1st Qu.: 21.8   1st Qu.:-1.5790
##  Median : 0.0160   Median : 0.0940   Median : 34.6   Median :-0.0110
##  Mean   : 0.0067   Mean   : 0.0449   Mean   : 40.1   Mean   :-0.0056
##  3rd Qu.: 0.9842   3rd Qu.: 1.6182   3rd Qu.: 51.1   3rd Qu.: 1.5680
##  Max.   : 2.5020   Max.   : 3.1420   Max.   :536.5   Max.   : 3.1420
##  PRI_met_sumet     PRI_jet_num PRI_jet_leading_pt PRI_jet_leading_eta
##  Min.   :  21.1   Min.   :1    Min.   : 30.0      Min.   :-4.499
##  1st Qu.: 148.4   1st Qu.:1    1st Qu.: 37.5      1st Qu.:-1.346
##  Median : 189.5   Median :1    Median : 50.0      Median :-0.002
##  Mean   : 203.3   Mean   :1    Mean   : 65.0      Mean   :-0.001
##  3rd Qu.: 240.7   3rd Qu.:1    3rd Qu.: 74.6      3rd Qu.: 1.347
##  Max.   :1383.6   Max.   :1    Max.   :743.2      Max.   : 4.492
##  PRI_jet_leading_phi PRI_jet_subleading_pt PRI_jet_subleading_eta
##  Min.   :-3.1420     Min.   :-999          Min.   :-999
##  1st Qu.:-1.5940     1st Qu.:-999          1st Qu.:-999
##  Median :-0.0390     Median :-999          Median :-999
##  Mean   :-0.0151     Mean   :-999          Mean   :-999
##  3rd Qu.: 1.5640     3rd Qu.:-999          3rd Qu.:-999
##  Max.   : 3.1410     Max.   :-999          Max.   :-999
##  PRI_jet_subleading_phi PRI_jet_all_pt  Label
##  Min.   :-999           Min.   : 30.0   b:49834
##  1st Qu.:-999           1st Qu.: 37.5   s:27710
##  Median :-999           Median : 50.0
##  Mean   :-999           Mean   : 65.0
##  3rd Qu.:-999           3rd Qu.: 74.6
##  Max.   :-999           Max.   :743.2
```

There are a number of variables that are undefined (all -999s) when there is 1 jet observed.

```
remove <- apply(train.jets1[1:nrow(train.jets1),]==-999, 2, all)
train.j1 <- train.jets1[,!remove]
val.j1 <- val.jets1[,!remove]
test.j1 <- test.jets1[,c("EventId",setdiff(names(!remove),"Label"))]
```

```
summary(train.j1)
```

```
##     EventId        DER_mass_MMC    DER_mass_transverse_met_lep
##  Min.   :100001   Min.   :-999.0   Min.   :  0.0
##  1st Qu.:162818   1st Qu.:  84.5   1st Qu.: 16.3
##  Median :225050   Median : 107.9   Median : 40.5
##  Mean   :225144   Mean   :  12.8   Mean   : 46.1
```

```
##  3rd Qu.:287516   3rd Qu.: 132.9   3rd Qu.: 70.0
##  Max.   :349997   Max.   :1192.0   Max.   :571.9
##   DER_mass_vis      DER_pt_h     DER_deltar_tau_lep   DER_pt_tot
##  Min.   :  6.3   Min.   :  0.0   Min.   :0.208      Min.   :  0.0
##  1st Qu.: 59.8   1st Qu.: 37.1   1st Qu.:1.855      1st Qu.:  2.9
##  Median : 73.9   Median : 53.1   Median :2.404      Median : 10.7
##  Mean   : 82.2   Mean   : 65.9   Mean   :2.340      Mean   : 16.6
##  3rd Qu.: 92.7   3rd Qu.: 79.0   3rd Qu.:2.855      3rd Qu.: 26.2
##  Max.   :959.6   Max.   :753.7   Max.   :5.655      Max.   :330.5
##    DER_sum_pt   DER_pt_ratio_lep_tau DER_met_phi_centrality
##  Min.   :  77   Min.   : 0.083       Min.   :-1.414
##  1st Qu.: 111   1st Qu.: 0.868       1st Qu.:-0.975
##  Median : 132   Median : 1.268       Median : 0.626
##  Mean   : 150   Mean   : 1.444       Mean   : 0.236
##  3rd Qu.: 168   3rd Qu.: 1.797       3rd Qu.: 1.324
##  Max.   :1215   Max.   :16.776       Max.   : 1.414
##    PRI_tau_pt     PRI_tau_eta       PRI_tau_phi        PRI_lep_pt
##  Min.   : 20.0   Min.   :-2.4980   Min.   :-3.1410   Min.   : 26.0
##  1st Qu.: 24.8   1st Qu.:-0.9220   1st Qu.:-1.5650   1st Qu.: 32.4
##  Median : 32.2   Median :-0.0110   Median :-0.0330   Median : 40.6
##  Mean   : 38.6   Mean   :-0.0017   Mean   :-0.0073   Mean   : 46.8
##  3rd Qu.: 45.3   3rd Qu.: 0.9110   3rd Qu.: 1.5660   3rd Qu.: 54.1
##  Max.   :505.1   Max.   : 2.4970   Max.   : 3.1410   Max.   :426.4
##   PRI_lep_eta       PRI_lep_phi         PRI_met       PRI_met_phi
##  Min.   :-2.4940   Min.   :-3.1420   Min.   :  0.3   Min.   :-3.1410
##  1st Qu.:-0.9750   1st Qu.:-1.5220   1st Qu.: 21.8   1st Qu.:-1.5790
##  Median : 0.0160   Median : 0.0940   Median : 34.6   Median :-0.0110
##  Mean   : 0.0067   Mean   : 0.0449   Mean   : 40.1   Mean   :-0.0056
##  3rd Qu.: 0.9842   3rd Qu.: 1.6182   3rd Qu.: 51.1   3rd Qu.: 1.5680
##  Max.   : 2.5020   Max.   : 3.1420   Max.   :536.5   Max.   : 3.1420
##  PRI_met_sumet     PRI_jet_num PRI_jet_leading_pt PRI_jet_leading_eta
##  Min.   :  21.1   Min.   :1    Min.   : 30.0      Min.   :-4.499
##  1st Qu.: 148.4   1st Qu.:1    1st Qu.: 37.5      1st Qu.:-1.346
##  Median : 189.5   Median :1    Median : 50.0      Median :-0.002
##  Mean   : 203.3   Mean   :1    Mean   : 65.0      Mean   :-0.001
##  3rd Qu.: 240.7   3rd Qu.:1    3rd Qu.: 74.6      3rd Qu.: 1.347
##  Max.   :1383.6   Max.   :1    Max.   :743.2      Max.   : 4.492
##  PRI_jet_leading_phi PRI_jet_all_pt  Label
##  Min.   :-3.1420     Min.   : 30.0   b:49834
##  1st Qu.:-1.5940     1st Qu.: 37.5   s:27710
##  Median :-0.0390     Median : 50.0
##  Mean   :-0.0151     Mean   : 65.0
##  3rd Qu.: 1.5640     3rd Qu.: 74.6
##  Max.   : 3.1410     Max.   :743.2
```

Just one more to eliminate: PRI_jet_num

```
    keep <- setdiff(names(train.j1),c("PRI_jet_num"))
    train.j1 <- train.j1[,keep]
    val.j1 <- val.j1[,keep]
    test.j1 <- test.j1[,c("EventId",setdiff(keep,"Label"))]
```

2 Jets

```
summary(train.jets2)
```

```
##     EventId        DER_mass_MMC    DER_mass_transverse_met_lep
##  Min.   :100000   Min.   :-999.0   Min.   :  0.0
##  1st Qu.:162596   1st Qu.:  91.6   1st Qu.: 11.6
##  Median :225414   Median : 113.0   Median : 28.5
##  Mean   :225130   Mean   :  56.9   Mean   : 38.3
##  3rd Qu.:287656   3rd Qu.: 132.4   3rd Qu.: 56.5
##  Max.   :349994   Max.   : 967.0   Max.   :595.8
##   DER_mass_vis       DER_pt_h      DER_deltaeta_jet_jet DER_mass_jet_jet
##  Min.   :   7.3   Min.   :   0.1   Min.   :0.000        Min.   :  14
##  1st Qu.:  58.5   1st Qu.:  54.1   1st Qu.:0.984        1st Qu.: 107
##  Median :  72.3   Median :  87.7   Median :2.383        Median : 229
##  Mean   :  79.2   Mean   : 103.0   Mean   :2.607        Mean   : 391
##  3rd Qu.:  89.5   3rd Qu.: 134.1   3rd Qu.:4.004        3rd Qu.: 514
##  Max.   :1051.4   Max.   :1053.8   Max.   :8.503        Max.   :4975
##  DER_prodeta_jet_jet DER_deltar_tau_lep   DER_pt_tot      DER_sum_pt
##  Min.   :-18.066     Min.   :0.228      Min.   :  0.0   Min.   : 111
##  1st Qu.: -3.251     1st Qu.:1.485      1st Qu.:  2.8   1st Qu.: 178
##  Median : -0.465     Median :2.020      Median :  9.6   Median : 219
##  Mean   : -1.115     Mean   :2.061      Mean   : 17.3   Mean   : 246
##  3rd Qu.:  0.869     3rd Qu.:2.632      3rd Qu.: 26.0   3rd Qu.: 283
##  Max.   : 16.690     Max.   :5.579      Max.   :513.7   Max.   :1282
##  DER_pt_ratio_lep_tau DER_met_phi_centrality DER_lep_eta_centrality
##  Min.   : 0.047       Min.   :-1.414         Min.   :0.000
##  1st Qu.: 0.767       1st Qu.:-0.017         1st Qu.:0.014
##  Median : 1.202       Median : 1.051         Median :0.551
##  Mean   : 1.453       Mean   : 0.566         Mean   :0.493
##  3rd Qu.: 1.820       3rd Qu.: 1.340         3rd Qu.:0.904
##  Max.   :19.773       Max.   : 1.414         Max.   :1.000
##    PRI_tau_pt     PRI_tau_eta       PRI_tau_phi        PRI_lep_pt
##  Min.   : 20.0   Min.   :-2.4950   Min.   :-3.1410   Min.   : 26.0
##  1st Qu.: 26.4   1st Qu.:-0.8750   1st Qu.:-1.5570   1st Qu.: 32.9
##  Median : 36.2   Median :-0.0080   Median : 0.0010   Median : 42.6
##  Mean   : 44.7   Mean   :-0.0008   Mean   : 0.0069   Mean   : 50.9
##  3rd Qu.: 53.1   3rd Qu.: 0.8680   3rd Qu.: 1.5780   3rd Qu.: 59.7
##  Max.   :622.9   Max.   : 2.4910   Max.   : 3.1420   Max.   :447.9
##   PRI_lep_eta       PRI_lep_phi         PRI_met       PRI_met_phi
##  Min.   :-2.4870   Min.   :-3.1420   Min.   :  0.2   Min.   :-3.1420
##  1st Qu.:-0.9020   1st Qu.:-1.5160   1st Qu.: 26.9   1st Qu.:-1.5605
##  Median :-0.0090   Median : 0.0850   Median : 44.2   Median :-0.0030
##  Mean   :-0.0035   Mean   : 0.0468   Mean   : 53.8   Mean   : 0.0033
##  3rd Qu.: 0.8980   3rd Qu.: 1.6170   3rd Qu.: 69.0   3rd Qu.: 1.5760
##  Max.   : 2.4990   Max.   : 3.1410   Max.   :951.4   Max.   : 3.1420
##  PRI_met_sumet     PRI_jet_num PRI_jet_leading_pt PRI_jet_leading_eta
##  Min.   :  34.3   Min.   :2    Min.   :  30.2     Min.   :-4.497
##  1st Qu.: 220.2   1st Qu.:2    1st Qu.:  56.6     1st Qu.:-1.416
##  Median : 274.7   Median :2    Median :  80.5     Median : 0.015
##  Mean   : 296.0   Mean   :2    Mean   :  98.5     Mean   :-0.002
##  3rd Qu.: 346.4   3rd Qu.:2    3rd Qu.: 120.4     3rd Qu.: 1.403
##  Max.   :1364.6   Max.   :2    Max.   :1120.6     Max.   : 4.499
```

```
##  PRI_jet_leading_phi PRI_jet_subleading_pt PRI_jet_subleading_eta
##  Min.   :-3.1420     Min.   : 30.0         Min.   :-4.500
##  1st Qu.:-1.5780     1st Qu.: 35.2         1st Qu.:-1.757
##  Median :-0.0320     Median : 43.6         Median :-0.013
##  Mean   :-0.0133     Mean   : 51.7         Mean   :-0.014
##  3rd Qu.: 1.5530     3rd Qu.: 58.7         3rd Qu.: 1.725
##  Max.   : 3.1410     Max.   :464.3         Max.   : 4.500
##  PRI_jet_subleading_phi PRI_jet_all_pt   Label
##  Min.   :-3.1420        Min.   :  60.2   b:24645
##  1st Qu.:-1.5685        1st Qu.:  96.9   s:25734
##  Median : 0.0120        Median : 128.5
##  Mean   : 0.0061        Mean   : 150.2
##  3rd Qu.: 1.5890        3rd Qu.: 180.0
##  Max.   : 3.1420        Max.   :1173.6
```

For now only PRI_jet_num can be eliminated when num jets is 2

```
    keep <- setdiff(names(train.jets2),c("PRI_jet_num"))
    train.j2 <- train.jets2[,keep]
    val.j2 <- val.jets2[,keep]
    test.j2 <- test.jets2[,c("EventId",setdiff(keep,"Label"))]
```

3 Jets

```
    summary(train.jets3)
```

```
##     EventId        DER_mass_MMC    DER_mass_transverse_met_lep
##  Min.   :100005   Min.   :-999.0   Min.   :  0.0
##  1st Qu.:161858   1st Qu.:  85.0   1st Qu.: 13.4
##  Median :224412   Median : 104.6   Median : 32.2
##  Mean   :224680   Mean   :  48.4   Mean   : 42.1
##  3rd Qu.:287700   3rd Qu.: 130.2   3rd Qu.: 62.2
##  Max.   :349993   Max.   : 988.2   Max.   :690.1
##   DER_mass_vis       DER_pt_h     DER_deltaeta_jet_jet DER_mass_jet_jet
##  Min.   :  10.3   Min.   :  0.8   Min.   :0.000        Min.   :  17
##  1st Qu.:  54.0   1st Qu.: 65.5   1st Qu.:0.714        1st Qu.: 124
##  Median :  68.0   Median :108.9   Median :1.635        Median : 221
##  Mean   :  78.9   Mean   :126.1   Mean   :1.943        Mean   : 327
##  3rd Qu.:  87.8   3rd Qu.:166.4   3rd Qu.:2.877        3rd Qu.: 412
##  Max.   :1329.9   Max.   :762.8   Max.   :7.877        Max.   :4062
##  DER_prodeta_jet_jet DER_deltar_tau_lep   DER_pt_tot      DER_sum_pt
##  Min.   :-15.347     Min.   :0.379      Min.   :  0.0   Min.   : 146
##  1st Qu.: -1.345     1st Qu.:1.230      1st Qu.: 33.2   1st Qu.: 255
##  Median : -0.015     Median :1.767      Median : 45.7   Median : 321
##  Mean   : -0.154     Mean   :1.885      Mean   : 53.5   Mean   : 358
##  3rd Qu.:  1.119     3rd Qu.:2.491      3rd Qu.: 66.1   3rd Qu.: 417
##  Max.   : 14.772     Max.   :5.505      Max.   :466.5   Max.   :1852
##  DER_pt_ratio_lep_tau DER_met_phi_centrality DER_lep_eta_centrality
##  Min.   : 0.081       Min.   :-1.414         Min.   :0.000
##  1st Qu.: 0.782       1st Qu.: 0.015         1st Qu.:0.000
##  Median : 1.261       Median : 0.966         Median :0.235
```

```
##  Mean   : 1.581       Mean   : 0.545         Mean   :0.380
##  3rd Qu.: 1.984       3rd Qu.: 1.320         3rd Qu.:0.785
##  Max.   :19.672       Max.   : 1.414         Max.   :1.000
##    PRI_tau_pt      PRI_tau_eta       PRI_tau_phi        PRI_lep_pt
##  Min.   : 20.0   Min.   :-2.4960   Min.   :-3.1410   Min.   : 26.0
##  1st Qu.: 26.3   1st Qu.:-0.9190   1st Qu.:-1.5920   1st Qu.: 34.2
##  Median : 36.1   Median :-0.0050   Median :-0.0285   Median : 45.7
##  Mean   : 46.8   Mean   :-0.0039   Mean   :-0.0118   Mean   : 55.9
##  3rd Qu.: 55.3   3rd Qu.: 0.9110   3rd Qu.: 1.5750   3rd Qu.: 65.9
##  Max.   :449.6   Max.   : 2.4970   Max.   : 3.1410   Max.   :461.9
##   PRI_lep_eta       PRI_lep_phi         PRI_met       PRI_met_phi
##  Min.   :-2.4850   Min.   :-3.1420   Min.   :  0.2   Min.   :-3.141
##  1st Qu.:-0.9292   1st Qu.:-1.5232   1st Qu.: 32.1   1st Qu.:-1.548
##  Median : 0.0110   Median : 0.0650   Median : 53.2   Median : 0.009
##  Mean   : 0.0003   Mean   : 0.0367   Mean   : 65.8   Mean   : 0.008
##  3rd Qu.: 0.9330   3rd Qu.: 1.6050   3rd Qu.: 85.1   3rd Qu.: 1.562
##  Max.   : 2.4970   Max.   : 3.1400   Max.   :695.5   Max.   : 3.141
##  PRI_met_sumet     PRI_jet_num PRI_jet_leading_pt PRI_jet_leading_eta
##  Min.   :  22.7   Min.   :3    Min.   : 31.3      Min.   :-4.439
##  1st Qu.: 306.5   1st Qu.:3    1st Qu.: 71.5      1st Qu.:-1.179
##  Median : 382.9   Median :3    Median :101.7      Median :-0.022
##  Mean   : 414.9   Mean   :3    Mean   :123.1      Mean   :-0.014
##  3rd Qu.: 486.1   3rd Qu.:3    3rd Qu.:150.2      3rd Qu.: 1.163
##  Max.   :2004.0   Max.   :3    Max.   :760.8      Max.   : 4.476
##  PRI_jet_leading_phi PRI_jet_subleading_pt PRI_jet_subleading_eta
##  Min.   :-3.141      Min.   : 30.2         Min.   :-4.491
##  1st Qu.:-1.559      1st Qu.: 46.0         1st Qu.:-1.326
##  Median :-0.018      Median : 59.8         Median : 0.000
##  Mean   :-0.001      Mean   : 71.2         Mean   :-0.008
##  3rd Qu.: 1.578      3rd Qu.: 82.9         3rd Qu.: 1.326
##  Max.   : 3.141      Max.   :721.5         Max.   : 4.500
##  PRI_jet_subleading_phi PRI_jet_all_pt   Label
##  Min.   :-3.1420        Min.   :  92.8   b:15433
##  1st Qu.:-1.5933        1st Qu.: 167.9   s: 6731
##  Median :-0.0410        Median : 222.2
##  Mean   :-0.0191        Mean   : 255.3
##  3rd Qu.: 1.5480        3rd Qu.: 304.9
##  Max.   : 3.1410        Max.   :1633.4
```

As when jets equals 2, the only feature to eliminate at this stage is PRI_jet_num

```
    keep <- setdiff(names(train.jets3),c("PRI_jet_num"))
    train.j3 <- train.jets3[,keep]
    val.j3 <- val.jets3[,keep]
    test.j3 <- test.jets3[,c("EventId",setdiff(keep,"Label"))]
```

Models

0 Jets

Fit a random forest with train.j0

```
    x <- setdiff(names(train.j0),c("EventId","Label"))
    set.seed(4646)
    trainj0.fit <- train(x=train.j0[,x], y=train.j0$Label, method="rf", proxy=T)
```

```
## Loading required package: randomForest
## randomForest 4.6-7
## Type rfNews() to see new features/changes/bug fixes.
```

```
    trainj0.fit
```

```
## Random Forest
##
## 99913 samples
##    18 predictors
##     2 classes: 'b', 's'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
##
## Summary of sample sizes: 99913, 99913, 99913, 99913, 99913, 99913, ...
##
## Resampling results across tuning parameters:
##
##   mtry  Accuracy  Kappa  Accuracy SD  Kappa SD
##   2     0.8       0.6    0.002        0.005
##   10    0.8       0.6    0.002        0.004
##   20    0.8       0.6    0.002        0.004
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was mtry = 2.
```

```
    varImp(trainj0.fit)
```

```
## rf variable importance
##
##                              Overall
## DER_mass_transverse_met_lep 100.000
## DER_pt_ratio_lep_tau         57.229
## DER_mass_MMC                 56.786
## PRI_met                      52.063
## PRI_tau_pt                   51.441
## DER_mass_vis                 45.764
## DER_deltar_tau_lep           43.398
## DER_sum_pt                   19.912
## PRI_lep_pt                   18.480
## DER_met_phi_centrality        6.300
## DER_pt_tot                    5.582
## DER_pt_h                      5.491
## PRI_lep_eta                   4.777
## PRI_met_sumet                 4.763
## PRI_tau_eta                   2.369
## PRI_met_phi                   0.636
```

```
## PRI_lep_phi                   0.459
## PRI_tau_phi                   0.000
```

```
    pred.val0 <- predict(trainj0.fit, val.j0)
    confusionMatrix(pred.val0, val.j0$Label)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     b     s
##          b 29629     0
##          s     0 10173
##
##                Accuracy : 1
##                  95% CI : (1, 1)
##     No Information Rate : 0.744
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 1
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.000
##             Specificity : 1.000
##          Pos Pred Value : 1.000
##          Neg Pred Value : 1.000
##              Prevalence : 0.744
##          Detection Rate : 0.744
##    Detection Prevalence : 0.744
##       Balanced Accuracy : 1.000
##
##        'Positive' Class : b
##
```

```
    pred.test0 <- predict(trainj0.fit, test.j0, type="prob")
    pred.0 <- PrepPrediction(pred.test0, test.j0)
```

```
    pred.80 <- pred.0
    save(pred.80, file="RData/pred80.RData")
```

1 Jet

Fit a random forest with train.j1. Using 8 cores to decrease fit time.

```
    predictors <- train.j1[,setdiff(names(train.j1),c("EventId","Label"))]
    set.seed(4646)
    registerDoMC(cores = 8)
    trainj1.fit <- train(x=predictors, y=train.j1$Label, method="rf", proxy=T,
                         trControl=trainControl(allowParallel=T))
```

```
    trainj1.fit
```

```
## Random Forest
##
## 77544 samples
##    22 predictors
##     2 classes: 'b', 's'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
##
## Summary of sample sizes: 77544, 77544, 77544, 77544, 77544, 77544, ...
##
## Resampling results across tuning parameters:
##
##   mtry  Accuracy  Kappa  Accuracy SD  Kappa SD
##   2     0.8       0.6    0.002        0.004
##   10    0.8       0.6    0.002        0.004
##   20    0.8       0.6    0.002        0.004
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was mtry = 12.
```

```
    varImp(trainj1.fit)
```

```
## rf variable importance
##
##   only 20 most important variables shown (out of 22)
##
##                              Overall
## DER_mass_MMC                  100.00
## DER_mass_transverse_met_lep    43.63
## DER_mass_vis                   39.00
## PRI_jet_leading_eta            17.52
## DER_met_phi_centrality         17.06
## PRI_tau_pt                     16.78
## DER_deltar_tau_lep              8.02
## DER_pt_tot                      6.89
## PRI_lep_eta                     6.58
## PRI_tau_eta                     5.43
## DER_sum_pt                      5.27
## PRI_met                         5.08
## PRI_jet_leading_phi             4.58
## DER_pt_ratio_lep_tau            4.46
## PRI_met_sumet                   4.00
## PRI_tau_phi                     3.86
## PRI_lep_phi                     3.65
## PRI_met_phi                     3.38
## DER_pt_h                        2.88
## PRI_lep_pt                      1.73
```

```
    pred.val1 <- predict(trainj1.fit, val.j1)
    confusionMatrix(pred.val1, val.j1$Label)
```

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction     b     s
##          b 19880     0
##          s     0 11079
##
##                Accuracy : 1
##                  95% CI : (1, 1)
##     No Information Rate : 0.642
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 1
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.000
##             Specificity : 1.000
##          Pos Pred Value : 1.000
##          Neg Pred Value : 1.000
##              Prevalence : 0.642
##          Detection Rate : 0.642
##    Detection Prevalence : 0.642
##       Balanced Accuracy : 1.000
##
##        'Positive' Class : b
##
```

```
    pred.test1 <- predict(trainj1.fit, test.j1, type="prob")
    pred.1 <- PrepPrediction(pred.test1, test.j1)
```

2 Jets

Fit a random forest with train.j2, register 8 cores

```
    getDoParWorkers()
```

```
## [1] 1
```

```
    registerDoMC(cores=8)
    getDoParWorkers()
```

```
## [1] 8
```

```
    predictors <- train.j2[,setdiff(names(train.j2),c("EventId","Label"))]
    set.seed(4646)
    getDoParWorkers()
```

```
## [1] 8
```

```
    trainj2.fit <- train(x=predictors, y=train.j2$Label, method="rf", proxy=T)
```

```
    trainj2.fit
```

```
## Random Forest
##
## 50379 samples
##    29 predictors
##     2 classes: 'b', 's'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
##
## Summary of sample sizes: 50379, 50379, 50379, 50379, 50379, 50379, ...
##
## Resampling results across tuning parameters:
##
##   mtry  Accuracy  Kappa  Accuracy SD  Kappa SD
##   2     0.8       0.7    0.003        0.005
##   20    0.8       0.7    0.002        0.005
##   30    0.8       0.7    0.003        0.005
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was mtry = 15.
```

```
    varImp(trainj2.fit)
```

```
## rf variable importance
##
##   only 20 most important variables shown (out of 29)
##
##                              Overall
## DER_mass_MMC                  100.00
## DER_mass_jet_jet               25.14
## DER_mass_vis                   22.67
## DER_mass_transverse_met_lep    12.88
## DER_deltaeta_jet_jet           12.74
## DER_lep_eta_centrality         10.15
## DER_met_phi_centrality          8.46
## PRI_tau_pt                      6.87
## DER_pt_tot                      6.45
## DER_deltar_tau_lep              5.30
## DER_prodeta_jet_jet             4.82
## PRI_met                         3.03
## DER_pt_h                        2.49
## PRI_tau_eta                     1.97
## PRI_jet_leading_phi             1.90
## DER_pt_ratio_lep_tau            1.85
## PRI_jet_subleading_phi          1.82
## PRI_jet_subleading_pt           1.76
## PRI_met_sumet                   1.68
## PRI_tau_phi                     1.60
```

```
    pred.val2 <- predict(trainj2.fit, val.j2)
    confusionMatrix(pred.val2, val.j2$Label)
```

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction     b     s
##          b  9916     0
##          s     0 10276
##
##                Accuracy : 1
##                  95% CI : (1, 1)
##     No Information Rate : 0.509
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 1
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.000
##             Specificity : 1.000
##          Pos Pred Value : 1.000
##          Neg Pred Value : 1.000
##              Prevalence : 0.491
##          Detection Rate : 0.491
##    Detection Prevalence : 0.491
##       Balanced Accuracy : 1.000
##
##        'Positive' Class : b
##
```

```
    pred.test2 <- predict(trainj2.fit, test.j2, type="prob")
    pred.2 <- PrepPrediction(pred.test2, test.j2)
```

3 Jets

```
    predictors <- train.j3[,setdiff(names(train.j3),c("EventId","Label"))]
    set.seed(4646)
    registerDoMC(cores=8)
    getDoParWorkers()
```

```
## [1] 8
```

```
    trainj3.fit <- train(x=predictors, y=train.j3$Label, method="rf", proxy=T)
```

```
    trainj3.fit
```

```
## Random Forest
##
## 22164 samples
```

```
##    29 predictors
##     2 classes: 'b', 's'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
##
## Summary of sample sizes: 22164, 22164, 22164, 22164, 22164, 22164, ...
##
## Resampling results across tuning parameters:
##
##   mtry  Accuracy  Kappa  Accuracy SD  Kappa SD
##   2     0.8       0.6    0.003        0.008
##   20    0.8       0.6    0.003        0.008
##   30    0.8       0.6    0.003        0.007
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was mtry = 15.
```

```
    varImp(trainj3.fit)
```

```
## rf variable importance
##
##   only 20 most important variables shown (out of 29)
##
##                              Overall
## DER_mass_MMC                  100.00
## DER_mass_vis                   23.93
## DER_mass_transverse_met_lep    14.18
## DER_deltar_tau_lep             13.25
## DER_met_phi_centrality         10.03
## PRI_tau_pt                      8.38
## DER_pt_h                        6.23
## DER_deltaeta_jet_jet            5.96
## DER_prodeta_jet_jet             5.47
## DER_lep_eta_centrality          4.85
## DER_mass_jet_jet                4.31
## PRI_met                         4.24
## DER_pt_tot                      3.77
## DER_pt_ratio_lep_tau            2.65
## PRI_tau_eta                     2.58
## PRI_jet_subleading_phi          2.38
## PRI_tau_phi                     2.14
## PRI_jet_leading_phi             2.12
## PRI_jet_leading_eta             2.05
## PRI_lep_eta                     2.04
```

```
    pred.val3 <- predict(trainj3.fit, val.j3)
    confusionMatrix(pred.val3, val.j3$Label)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    b    s
```

```
##          b 6308    0
##          s    0 2738
##
##                Accuracy : 1
##                  95% CI : (1, 1)
##     No Information Rate : 0.697
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 1
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.000
##             Specificity : 1.000
##          Pos Pred Value : 1.000
##          Neg Pred Value : 1.000
##              Prevalence : 0.697
##          Detection Rate : 0.697
##    Detection Prevalence : 0.697
##       Balanced Accuracy : 1.000
##
##        'Positive' Class : b
##
```

```
    pred.test3 <- predict(trainj3.fit, test.j3, type="prob")
    pred.3 <- PrepPrediction(pred.test3, test.j3)
```

Submission

```
    pred <- rbind(pred.0,pred.1,pred.2,pred.3)
    WriteSubmission(pred, 8)
```

Mini Models

In order to speed up random forest processing time I'm going to take just 10% of the jet0 training set and try to optimize for speed

```
    mini.indices <- createDataPartition(train.j0$Label, p=0.1, list=F)
    train.minij0 <- train.j0[mini.indices,]
    dim(train.minij0)
```

```
## [1] 9993   20
```

```
    getDoParWorkers()
```

```
## [1] 1
```

```
    predictors <- train.minij0[,setdiff(names(train.j0),c("EventId","Label"))]
    set.seed(4646)
    Sys.time()
```

```
## [1] "2014-07-18 14:38:11 MDT"
```

```
    w1.fit <- train(x=predictors, y=train.minij0$Label, method="rf", proxy=T)
    Sys.time()
```

```
## [1] "2014-07-18 15:01:58 MDT"
```

```
    registerDoMC(cores=8)
    getDoParWorkers()
```

```
## [1] 8
```

```
    set.seed(4646)
    Sys.time()
```

```
## [1] "2014-07-18 15:04:55 MDT"
```

```
    w8.fit <- train(x=predictors, y=train.minij0$Label, method="rf", proxy=T)
    Sys.time()
```

```
## [1] "2014-07-18 15:12:06 MDT"
```

```
    registerDoMC(cores=8)
    getDoParWorkers()
```

```
## [1] 8
```

```
    set.seed(4646)
    Sys.time()
```

```
## [1] "2014-07-18 15:28:26 MDT"
```

```
    w8.allow.fit <- train(x=predictors, y=train.minij0$Label, method="rf", proxy=T,
                          trControl=trainControl(allowParallel=T))
    Sys.time()
```

```
## [1] "2014-07-18 15:35:38 MDT"
```

RData

```
    dir.create("RData")
```

```
## Warning: 'RData' already exists
```

```r
save(trainj0.fit, file="RData/trainj0.RData")
save(trainj1.fit, file="RData/trainj1.RData")
save(trainj2.fit, file="RData/trainj2.RData")
save(trainj3.fit, file="RData/trainj3.RData")
```