Classification of Stars, Galaxies and Quasars

Enio Linhares Junior

5/19/2019

Contents

1 Introduction

1.1 Content

The Sloan Digital Sky Survey (SDSS) offers public data of space observations and the task here is to build a model that is able to predict the different classes of objects (Stars, Galaxies and Quasars) based on the data acquired through the scientific equipment. The data consists of 10,000 observations of space taken by the SDSS. Every observation is described by 17 feature columns and 1 class column which identifies it to be either a star, galaxy or quasar.
Our model achieved an **accuray** of 0.9915019

1.2 Feature Description

The table results from a query which joins two tables (actuaclly views): "PhotoObj" which contains photometric data and "SpecObj" which contains spectral data.

During our data exploratory analysis we will be explaining the features as they appear.

The data released by the SDSS is under public domain. Its taken from the current data release RD14.

2 Downloading Installing and Starting R

2.1 Installing the DataSet

The url can be found here: https://www.kaggle.com/lucidlenn/sloan-digital-sky-survey/version/1

2.2 Libraries used

```
library(tidyverse)
library(data.table)
library(caret)
library(ggplot2)
```

```
library(ggcorrplot)
library(RSNNS)
library(randomForest)
library(ggcorrplot)
library(kernlab)
library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 3.5.2
```

3 Load The Data

```
sky.df <- fread(file = "Skyserver_SQL2_27_2018 6_51_39 PM.csv", sep=",") # load the data and save for u
```

3.1 Create a Validation Dataset

```
# create a list of 80% of the rows in the original dataset we can use for training
index <- createDataPartition(sky.df$class, p=0.8, list=FALSE)
# select 20% of the data for validation
validation <- sky.df[-index,]
# use the remaining 80% of data to training and testing the models
sky.train <- sky.df[index,]
```

4 Summarize Dataset

4.1 Structure, summary, NA's check, and dimensions

```
str(sky.train)
```

```
## Classes 'data.table' and 'data.frame':   8001 obs. of  18 variables:
##  $ objid    :integer64 1237648704577142822 1237648704577208477 1237648704577273907 1237648704577273907
##  $ ra       : num  184 184 184 184 184 ...
##  $ dec      : num  0.0897 0.1262 0.0499 0.1026 0.1737 ...
##  $ u        : num  19.5 19.4 17.8 17.6 19.4 ...
##  $ g        : num  17 18.2 16.6 16.3 18.5 ...
##  $ r        : num  15.9 17.5 16.2 16.4 18.2 ...
##  $ i        : num  15.5 17.1 16 16.6 18 ...
##  $ z        : num  15.2 16.8 15.9 16.6 18 ...
##  $ run      : int  752 752 752 752 752 752 752 752 752 752 ...
##  $ rerun    : int  301 301 301 301 301 301 301 301 301 301 ...
##  $ camcol   : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ field    : int  267 268 269 269 269 269 270 270 270 271 ...
##  $ specobjid: chr  "3722360139651588096" "323274319570429952" "3722365362331820032" "3722365912087633
##  $ class    : chr  "STAR" "GALAXY" "STAR" "STAR" ...
##  $ redshift : num  -8.96e-06 1.23e-01 -1.11e-04 5.90e-04 3.15e-04 ...
##  $ plate    : int  3306 287 3306 3306 324 3306 323 288 3306 3306 ...
##  $ mjd      : int  54922 52023 54922 54922 51666 54922 51615 52000 54922 54922 ...
##  $ fiberid  : int  491 513 510 512 594 515 595 400 506 544 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
summary(sky.train)
```

```
##      objid                             ra               dec
##  Min.   :1237646798138245746   Min.   :  8.235   Min.   :-5.3788
##  1st Qu.:1237648705652326512   1st Qu.:157.057   1st Qu.:-0.5261
##  Median :1237648722296111157   Median :180.312   Median : 0.4150
##  Mean   :1237649701292874031   Mean   :175.444   Mean   :15.0368
##  3rd Qu.:1237651191890509846   3rd Qu.:201.704   3rd Qu.:37.2457
##  Max.   :1237651540334280888   Max.   :260.851   Max.   :68.5406
##        u                g                r                i
##  Min.   :12.99   Min.   :12.80   Min.   :12.43   Min.   :11.95
##  1st Qu.:18.17   1st Qu.:16.80   1st Qu.:16.17   1st Qu.:15.85
##  Median :18.85   Median :17.49   Median :16.86   Median :16.56
##  Mean   :18.61   Mean   :17.37   Mean   :16.84   Mean   :16.58
##  3rd Qu.:19.26   3rd Qu.:18.01   3rd Qu.:17.52   3rd Qu.:17.26
##  Max.   :19.60   Max.   :19.74   Max.   :24.80   Max.   :28.18
##        z              run             rerun         camcol
##  Min.   :11.61   Min.   : 308.0   Min.   :301   Min.   :1.000
##  1st Qu.:15.62   1st Qu.: 752.0   1st Qu.:301   1st Qu.:2.000
##  Median :16.39   Median : 756.0   Median :301   Median :4.000
##  Mean   :16.42   Mean   : 984.1   Mean   :301   Mean   :3.667
##  3rd Qu.:17.14   3rd Qu.:1331.0   3rd Qu.:301   3rd Qu.:5.000
##  Max.   :22.83   Max.   :1412.0   Max.   :301   Max.   :6.000
##      field        specobjid            class              redshift
##  Min.   : 11.0   Length:8001        Length:8001        Min.   :-0.004136
##  1st Qu.:183.0   Class :character   Class :character   1st Qu.: 0.000079
##  Median :298.0   Mode  :character   Mode  :character   Median : 0.042671
##  Mean   :301.1                                         Mean   : 0.143781
##  3rd Qu.:411.0                                         3rd Qu.: 0.092037
##  Max.   :768.0                                         Max.   : 5.353854
##      plate           mjd           fiberid
##  Min.   : 266   Min.   :51578   Min.   :   1.0
##  1st Qu.: 301   1st Qu.:51900   1st Qu.: 191.0
##  Median : 441   Median :51997   Median : 354.0
##  Mean   :1453   Mean   :52936   Mean   : 355.4
##  3rd Qu.:2559   3rd Qu.:54468   3rd Qu.: 512.0
##  Max.   :8410   Max.   :57481   Max.   :1000.0
```

```r
colSums(is.na(sky.train)) # any NA's?
```

```
##     objid        ra       dec         u         g         r         i
##         0         0         0         0         0         0         0
##         z       run     rerun    camcol     field specobjid     class
##         0         0         0         0         0         0         0
##  redshift     plate       mjd   fiberid
##         0         0         0         0
```

```r
dim(sky.train)
```

```
## [1] 8001   18
```

4.2 Types of attributes

```
sapply(sky.train, class) # checking the class of every feature
```

```
##       objid          ra         dec           u           g           r
## "integer64"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric"
##           i           z         run       rerun      camcol       field
##   "numeric"   "numeric"   "integer"   "integer"   "integer"   "integer"
##   specobjid       class    redshift       plate         mjd     fiberid
## "character" "character"   "numeric"   "integer"   "integer"   "integer"
```

The "class" column is our response variable. Since this is a classification problem, we will transform it in a factor with three levels:

```
sky.train$class <- as.factor(sky.train$class)
levels(sky.train$class)
```

```
## [1] "GALAXY" "QSO"    "STAR"
```

```
validation$class <- as.factor(validation$class)
levels(validation$class)
```

```
## [1] "GALAXY" "QSO"    "STAR"
```

4.3 Class distribution

Summarize the class distribution:

```
percentage <- prop.table(table(sky.train$class)) * 100
cbind(freq=table(sky.train$class), percentage=percentage)
```

```
##        freq percentage
## GALAXY 3999  49.981252
## QSO     680   8.498938
## STAR   3322  41.519810
```

4.4 Statistical Summary

```
summary(sky.train)
```

```
##      objid                           ra              dec
##  Min.   :1237646798138245746   Min.   :  8.235   Min.   :-5.3788
##  1st Qu.:1237648705652326512   1st Qu.:157.057   1st Qu.:-0.5261
##  Median :1237648722296111157   Median :180.312   Median : 0.4150
##  Mean   :1237649701292874031   Mean   :175.444   Mean   :15.0368
##  3rd Qu.:1237651191890509846   3rd Qu.:201.704   3rd Qu.:37.2457
##  Max.   :1237651540334280888   Max.   :260.851   Max.   :68.5406
##        u               g               r               i
##  Min.   :12.99   Min.   :12.80   Min.   :12.43   Min.   :11.95
##  1st Qu.:18.17   1st Qu.:16.80   1st Qu.:16.17   1st Qu.:15.85
##  Median :18.85   Median :17.49   Median :16.86   Median :16.56
##  Mean   :18.61   Mean   :17.37   Mean   :16.84   Mean   :16.58
##  3rd Qu.:19.26   3rd Qu.:18.01   3rd Qu.:17.52   3rd Qu.:17.26
##  Max.   :19.60   Max.   :19.74   Max.   :24.80   Max.   :28.18
##        z              run             rerun         camcol
```

```
##  Min.   :11.61   Min.   : 308.0   Min.   :301   Min.   :1.000
##  1st Qu.:15.62   1st Qu.: 752.0   1st Qu.:301   1st Qu.:2.000
##  Median :16.39   Median : 756.0   Median :301   Median :4.000
##  Mean   :16.42   Mean   : 984.1   Mean   :301   Mean   :3.667
##  3rd Qu.:17.14   3rd Qu.:1331.0   3rd Qu.:301   3rd Qu.:5.000
##  Max.   :22.83   Max.   :1412.0   Max.   :301   Max.   :6.000
##      field        specobjid            class         redshift
##  Min.   : 11.0   Length:8001        GALAXY:3999   Min.   :-0.004136
##  1st Qu.:183.0   Class :character   QSO   : 680   1st Qu.: 0.000079
##  Median :298.0   Mode  :character   STAR  :3322   Median : 0.042671
##  Mean   :301.1                                    Mean   : 0.143781
##  3rd Qu.:411.0                                    3rd Qu.: 0.092037
##  Max.   :768.0                                    Max.   : 5.353854
##      plate           mjd           fiberid
##  Min.   : 266   Min.   :51578   Min.   :   1.0
##  1st Qu.: 301   1st Qu.:51900   1st Qu.: 191.0
##  Median : 441   Median :51997   Median : 354.0
##  Mean   :1453   Mean   :52936   Mean   : 355.4
##  3rd Qu.:2559   3rd Qu.:54468   3rd Qu.: 512.0
##  Max.   :8410   Max.   :57481   Max.   :1000.0
```

We observe here two different things that the data preparation and exploratory analysis tells us:
- The class distribution is not even and this could be solved later using the SMOTE function which equalizes the classes proportion.
- The numeric columns are not on the same scale.
As a first step we decided to explore the data "as it is" and later, when we build the model, we evaluate the use of the SMOTE function and equalize the classes.

5 Visualization

5.1 Grouped Features

The **Thuan-Gunn** astronomic magnitude system. u, g, r, i, z represent the response of the 5 bands of the telescope:

```
thuan_gunn <- c("u", "g", "r", "i", "z")
```

- u = better of DeV/Exp magnitude fit
- g = better of DeV/Exp magnitude fit
- r = better of DeV/Exp magnitude fit
- i = better of DeV/Exp magnitude fit
- z = better of DeV/Exp magnitude fit

Field:
Run, rerun, camcol and field are features which describe a field within an image taken by the SDSS. A field is basically a part of the entire image corresponding to 2048 by 1489 pixels. A field can be identified by: - run number, which identifies the specific scan, - the camera column, or "camcol," a number from 1 to 6, identifying the scanline within the run, and - the field number. The field number typically starts at 11 (after an initial rampup time), and can be as large as 800 for particularly long runs. - An additional number, rerun, specifies how the image was processed.

```
field_feat <- c("run", "rerun", "camcol", "field")
```

- run = Run Number
- rereun = Rerun Number
- camcol = Camera column

- field = Field number

Skies or Sky:
Right ascension (abbreviated RA) is the angular distance measured eastward along the celestial equator from the Sun at the March equinox to the hour circle of the point above the earth in question. When paired with declination (abbreviated dec), these astronomical coordinates specify the direction of a point on the celestial sphere (traditionally called in English the skies or the sky) in the equatorial coordinate system. Source.

```
skies <- c("ra", "dec")
```

Remaining features:

- redshift = Final Redshift
- plate = plate number
- mjd = MJD of observation
- fiberid = fiber ID
  In physics, redshift happens when light or other electromagnetic radiation from an object is increased in wavelength, or shifted to the red end of the spectrum.

Each spectroscopic exposure employs a large, thin, circular metal plate that positions optical fibers via holes drilled at the locations of the images in the telescope focal plane. These fibers then feed into the spectrographs. Each plate has a unique serial number, which is called plate in views such as SpecObj in the CAS.

Modified Julian Date, used to indicate the date that a given piece of SDSS data (image or spectrum) was taken.

The SDSS spectrograph uses optical fibers to direct the light at the focal plane from individual objects to the slithead. Each object is assigned a corresponding fiberID.

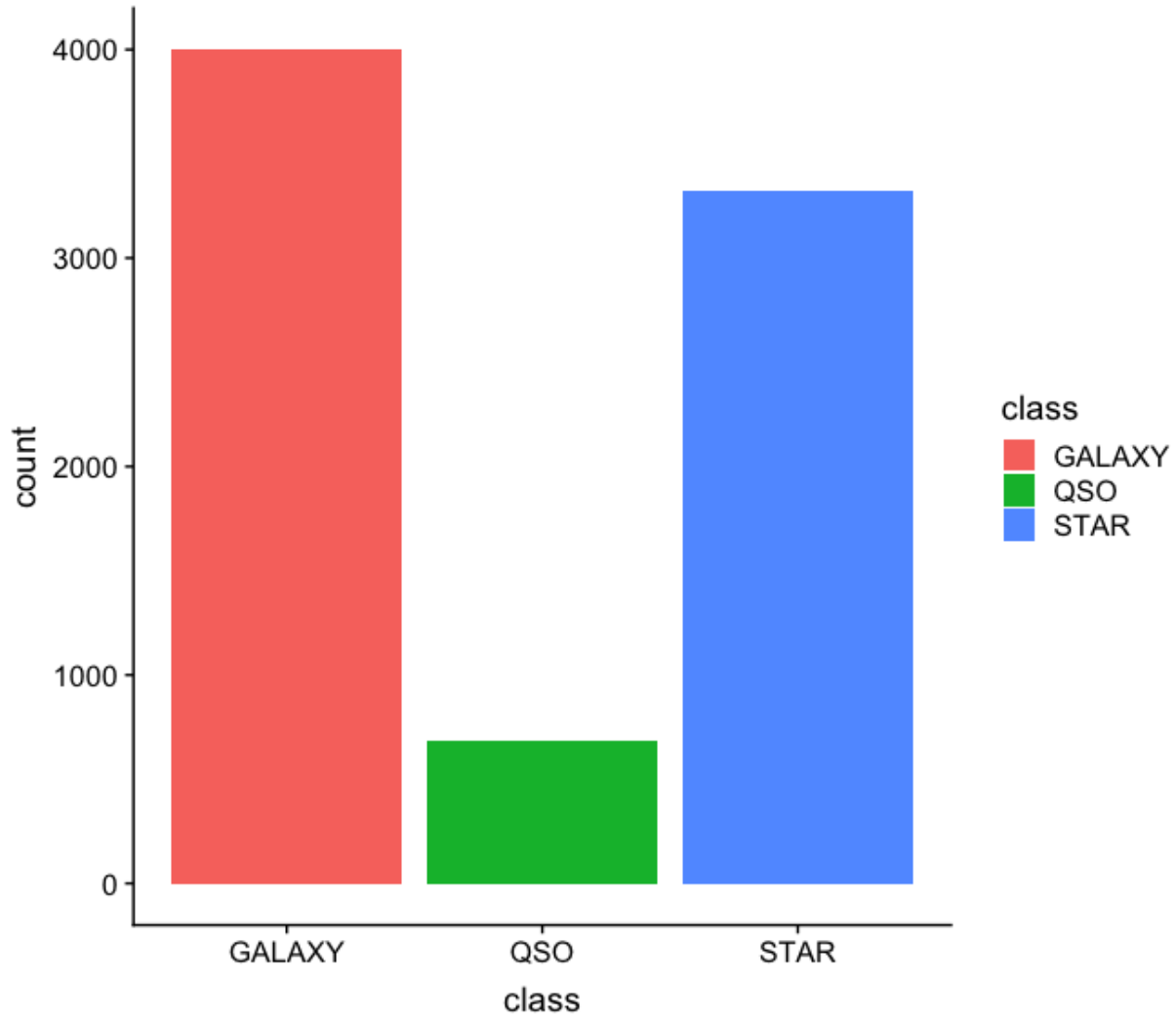
```
equipment_feat <- c("redshift", "plate", "mjd",
                    "fiberid") # These features are related to the measurement equipment
```

Class and Identification:
The class identifies an object to be either a galaxy, star or quasar. This will be the response variable which we will be trying to predict. - View "SpecObj" - specobjid = Object Identifier - class = object class (galaxy, star or quasar object)

5.2 Univariate Plots

The *class* distribution can now be visualized on the training set:

```
x <- sky.train[,-c(1, 13, 14, 16, 17)]
y <- sky.train$class # split input and output
ggplot(sky.train, aes(class, fill = class)) +
  geom_bar() # plot the classes
```



We need a different approach now: normalize the data to have all the numeric features between 0 and 1 to be able to compare them and understand better the data.

```
# normalize using the package 'RSNNS'
set.seed(2205)
sky.train.norm <- normalizeData(sky.train[,-c(1, 13, 14)], type = "0_1")
sky.train.norm <- as.data.frame(sky.train.norm)
summary(sky.train.norm) # check the normalization
```

```
##        V1                V2                V3                V4
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.5891   1st Qu.:0.06565   1st Qu.:0.7834   1st Qu.:0.5769
##  Median :0.6812   Median :0.07838   Median :0.8865   Median :0.6762
##  Mean   :0.6619   Mean   :0.27619   Mean   :0.8508   Mean   :0.6585
##  3rd Qu.:0.7659   3rd Qu.:0.57664   3rd Qu.:0.9484   3rd Qu.:0.7513
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##        V5                V6                V7                V8
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3023   1st Qu.:0.2406   1st Qu.:0.3576   1st Qu.:0.4022
##  Median :0.3582   Median :0.2841   Median :0.4265   Median :0.4058
##  Mean   :0.3564   Mean   :0.2857   Mean   :0.4291   Mean   :0.6124
```

```
##  3rd Qu.:0.4111   3rd Qu.:0.3275   3rd Qu.:0.4931   3rd Qu.:0.9266
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##        V9            V10              V11              V12
##  Min.   :301   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000000
##  1st Qu.:301   1st Qu.:0.2000   1st Qu.:0.2272   1st Qu.:0.0007868
##  Median :301   Median :0.6000   Median :0.3791   Median :0.0087359
##  Mean   :301   Mean   :0.5333   Mean   :0.3832   Mean   :0.0276068
##  3rd Qu.:301   3rd Qu.:0.8000   3rd Qu.:0.5284   3rd Qu.:0.0179494
##  Max.   :301   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000000
##       V13                V14               V15
##  Min.   :0.000000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.004298   1st Qu.:0.05455   1st Qu.:0.1902
##  Median :0.021488   Median :0.07098   Median :0.3534
##  Mean   :0.145786   Mean   :0.22998   Mean   :0.3548
##  3rd Qu.:0.281557   3rd Qu.:0.48958   3rd Qu.:0.5115
##  Max.   :1.000000   Max.   :1.00000   Max.   :1.0000
```

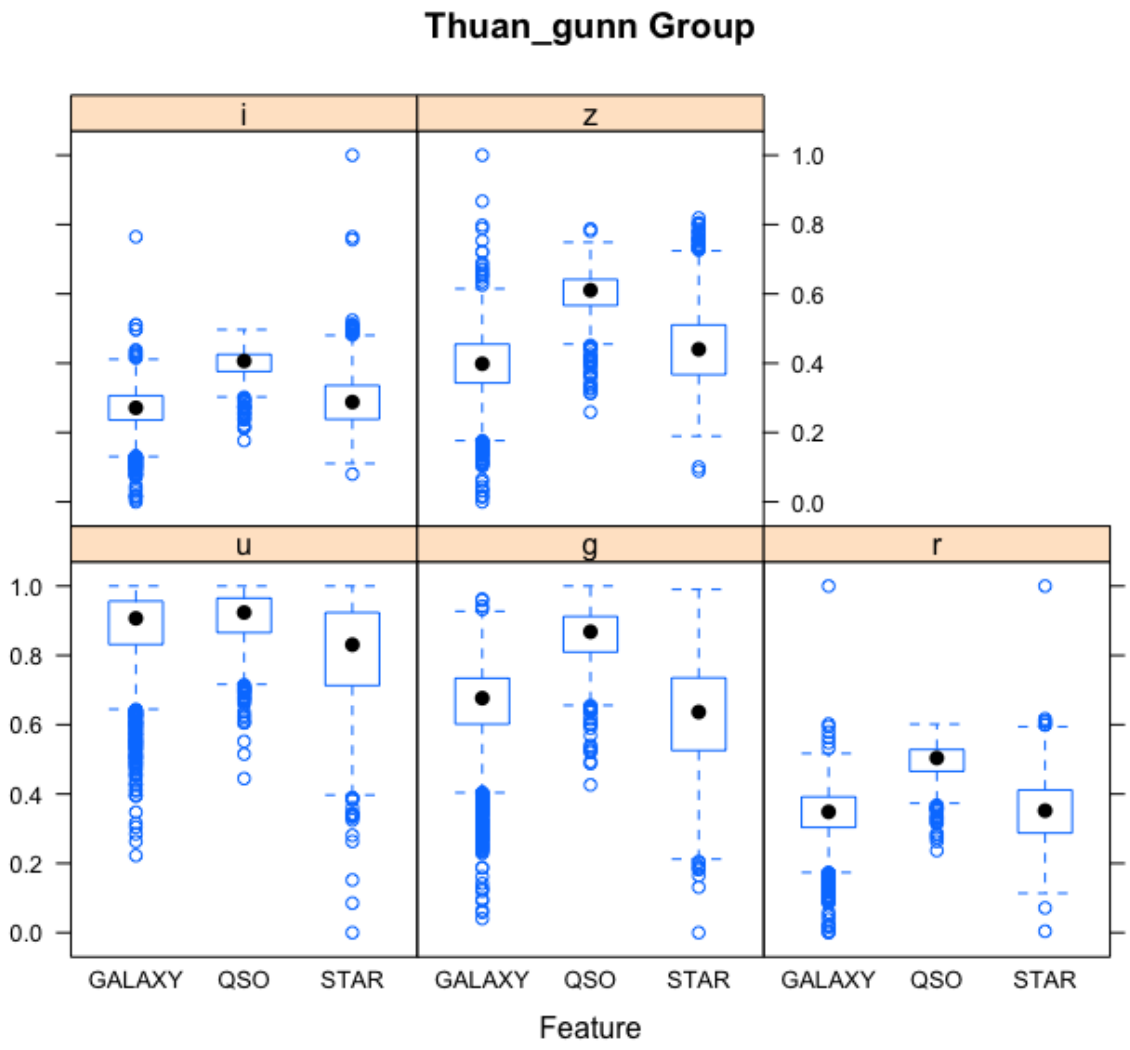
Tidying the normalized data:

```
names_sky.train <- names(sky.train[,-c(1, 13, 14)]) # add the names non-numeric columns back to the df.
names(sky.train.norm) <- names_sky.train
# now we add back the columns that were not included in the normalization.
sky.train.norm <- add_column(sky.train.norm, objid = sky.train$objid)
sky.train.norm <- add_column(sky.train.norm, specobjid = sky.train$specobjid)
sky.train.norm <- add_column(sky.train.norm, class = sky.train$class)
head(sky.train.norm, 2)
```

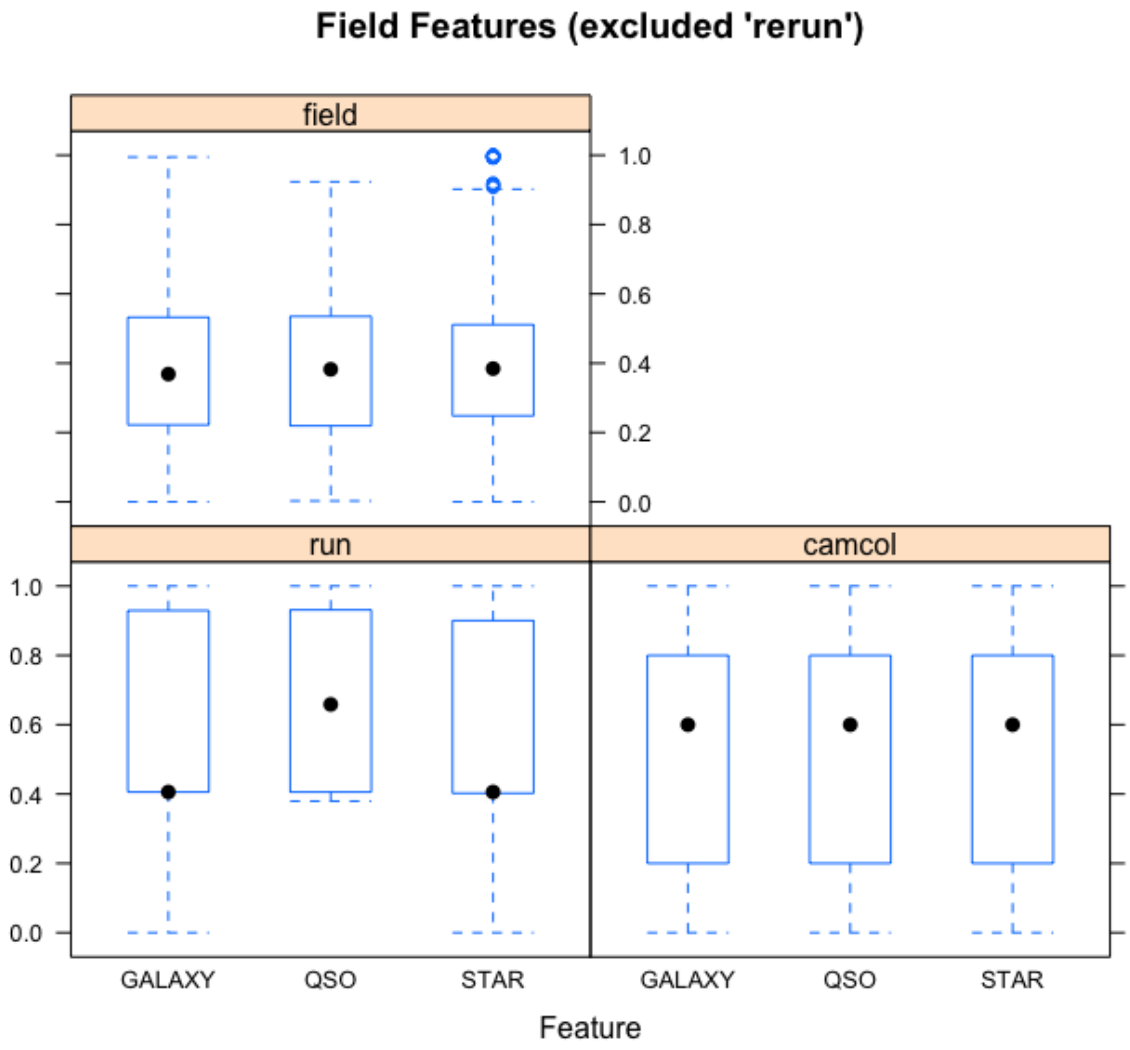
```
##          ra        dec         u         g         r         i         z
## 1 0.6939242 0.07397909 0.9809931 0.6114374 0.2841766 0.2190807 0.3222841
## 2 0.6945136 0.07447276 0.9672155 0.7770617 0.4076395 0.3166570 0.4627861
##         run rerun camcol    field     redshift       plate       mjd
## 1 0.4021739   301    0.6 0.338177 0.0007702743 0.373280943 0.5664916
## 2 0.4021739   301    0.6 0.339498 0.0237490694 0.002578585 0.0753854
##     fiberid               objid           specobjid  class
## 1 0.4904905 1237648704577142822 3722360139651588096   STAR
## 2 0.5125125 1237648704577208477  323274319570429952 GALAXY
```

plotting with the normalized data:

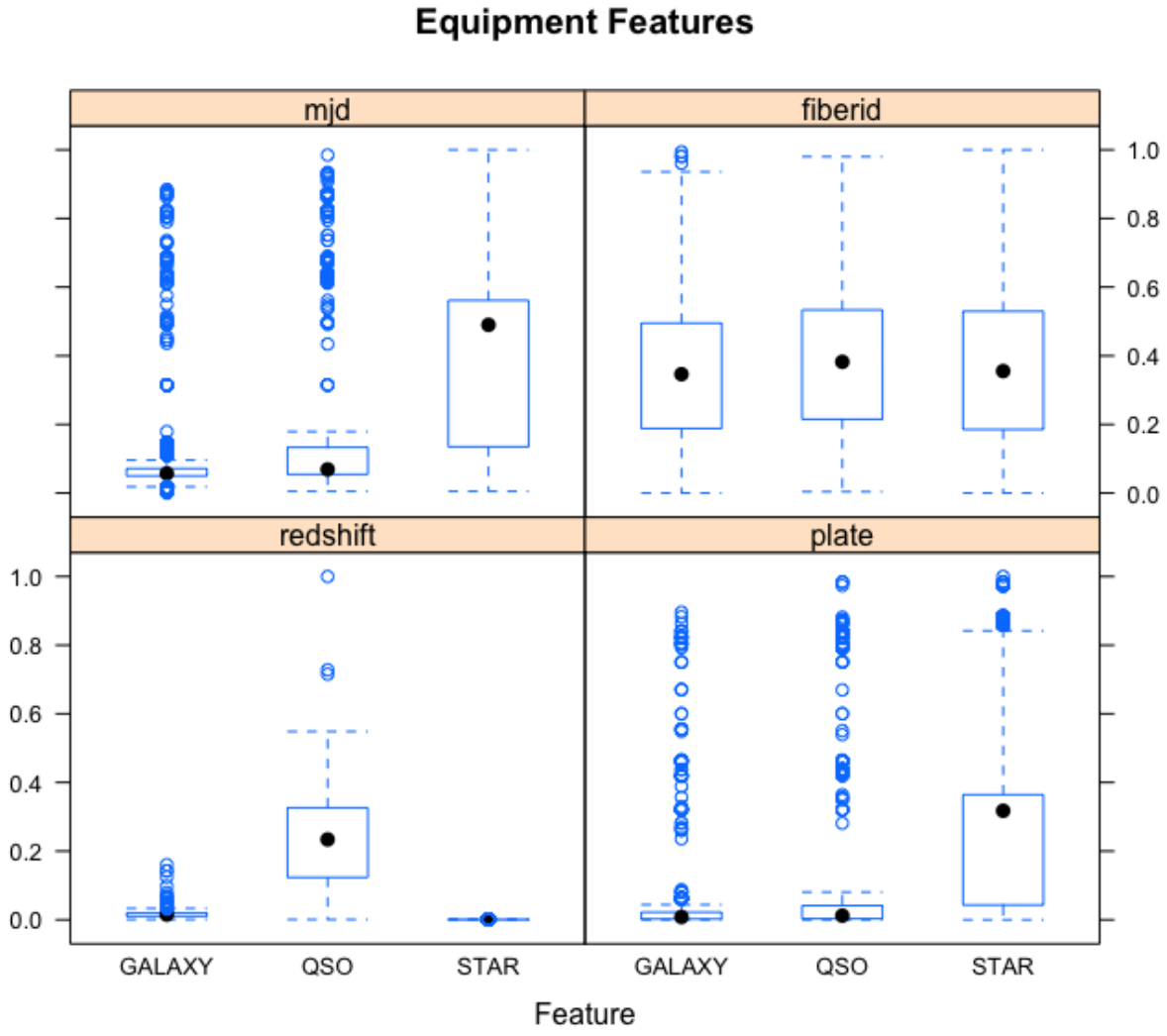
```
x <- sky.train.norm[,-c(9, 16:18)]
y <- sky.train.norm$class
featurePlot(x = sky.train.norm[, c("u", "g", "r", "i", "z")], y = y, plot="box",
            main = "Thuan_gunn Group")
```



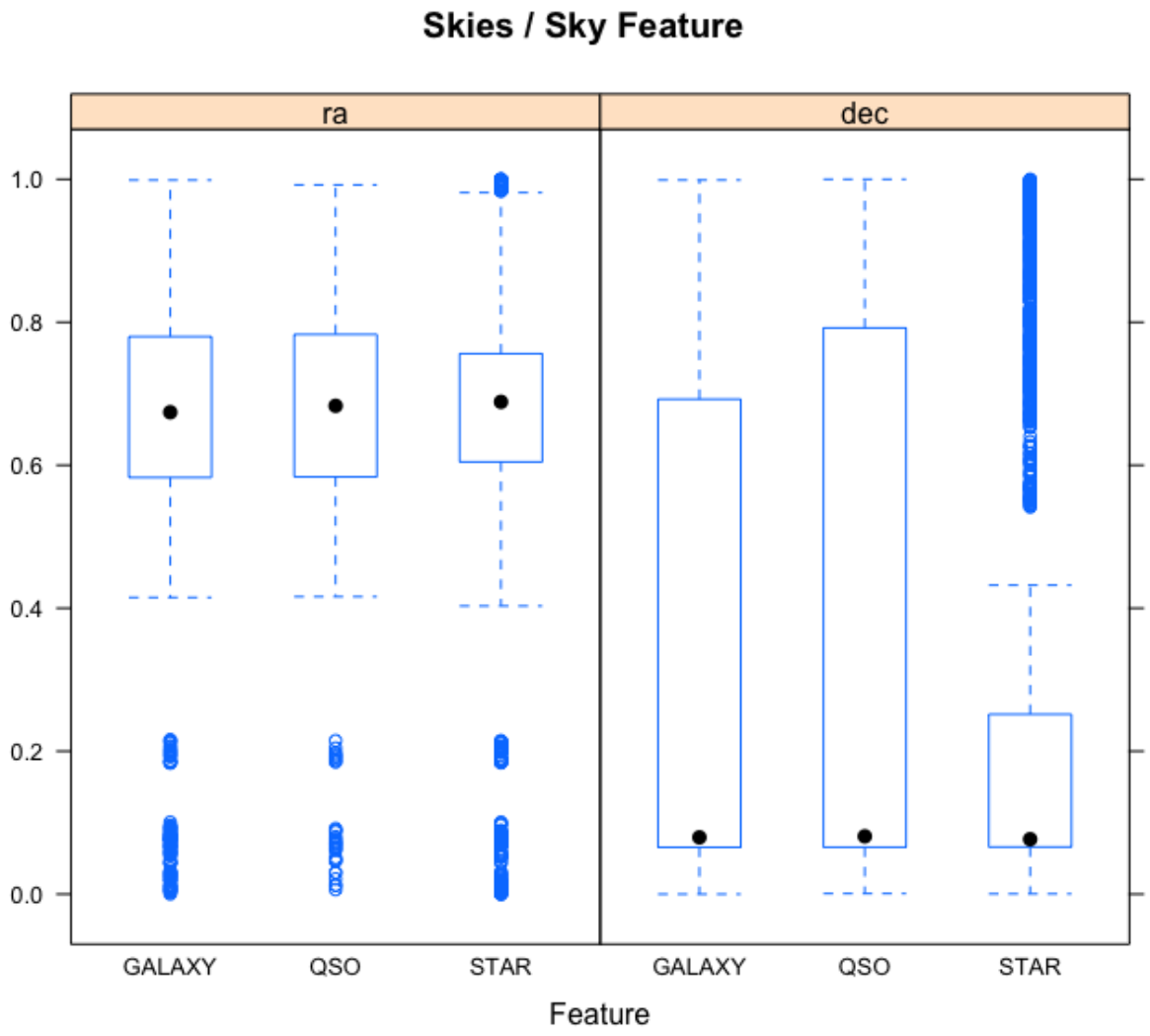

```
featurePlot(x = sky.train.norm[, c("run", "camcol", "field")], y = y, plot="box",
            main = "Field Features (excluded 'rerun')")
```




```
featurePlot(x = sky.train.norm[, c("redshift", "plate", "mjd", "fiberid")], y = y, plot="box",
            main = "Equipment Features")
```




```
featurePlot(x = sky.train.norm[, c("ra", "dec")], y = y, plot="box",
            main = "Skies / Sky Feature")
```

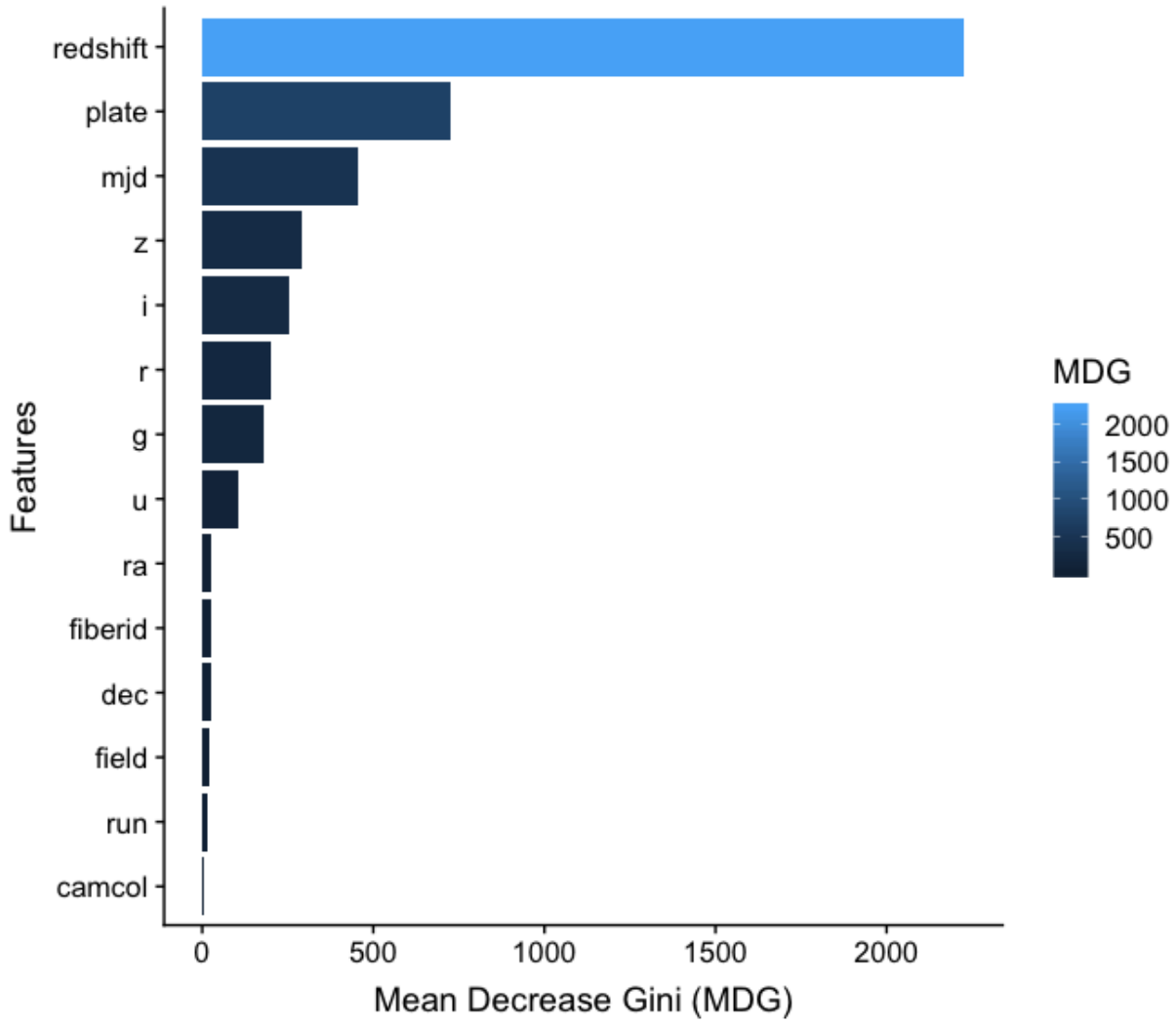


This visualization is useful for us to notice that there are clearly different distributions of the attributes for each class value and to identify the outliers (noise).
There seems to be great variability in the 'mjd' and 'plate' parameters.
The variability found in the 'Thuann_gunn' group will be kept 'as it is' and we will deal with it only in case we need to improve our model.

5.3 Variables importance

Running a model with *randomForest* to check the variables importance:

```
set.seed(2205)
rf.sky.train <- randomForest(class ~ ., data = sky.train[, -c(1, 10, 13)])
imp.df <- importance(rf.sky.train) # importance of the features
imp.df <- data.frame(features = row.names(imp.df), MDG = imp.df[,1])
imp.df <- imp.df[order(imp.df$MDG, decreasing = TRUE),]
ggplot(imp.df, aes(x = reorder(features, MDG), y = MDG, fill = MDG)) +
  geom_bar(stat = "identity") + labs (x = "Features", y = "Mean Decrease Gini (MDG)") +
  coord_flip()
```



The *plate* seems to have importance when predicting. This could be a noisy parameter since we have different plates measuring the waves.
The *MJD* seems to be important as well. Could those two features make the difference in the model?

Running a PCA analysis to check the variables importance (we are excluding 2 constant features and the factors)

```
PCA.sky.train <- prcomp(sky.train[, -c(1, 10, 13, 14)])
summary(PCA.sky.train)
```

```
## Importance of components:
##                             PC1       PC2       PC3      PC4      PC5
## Standard deviation     2324.3892 308.30502 277.45749 195.6791 140.3412
## Proportion of Variance    0.9589   0.01687   0.01366   0.0068   0.0035
## Cumulative Proportion     0.9589   0.97577   0.98944   0.9962   0.9997
##                             PC6      PC7   PC8   PC9   PC10   PC11   PC12
## Standard deviation     36.34377 14.05613 2.217 1.391 0.6682 0.3439 0.1649
## Proportion of Variance  0.00023  0.00004 0.000 0.000 0.0000 0.0000 0.0000
## Cumulative Proportion   0.99996  1.00000 1.000 1.000 1.0000 1.0000 1.0000
##                           PC13    PC14
## Standard deviation     0.1343 0.08122
```

```
## Proportion of Variance 0.0000 0.00000
## Cumulative Proportion  1.0000 1.00000
```

We notice that the first 6 components respond by 99.99% of the data.
Considering that the number of features is not so big, we could use all the features already considered by the PCA analysis or use only the first 6 features.

The MDG definition: "Because Random Forests are an ensemble of individual Decision Trees, Gini Importance can be leveraged to calculate Mean Decrease in Gini, which is a measure of variable importance for estimating a target variable.
Mean Decrease in Gini is the average (mean) of a variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each individual decision tree in the random forest. This is effectively a measure of how important a variable is for estimating the value of the target variable across all of the trees that make up the forest. A higher Mean Decrease in Gini indicates higher variable importance. Variables are sorted and displayed in the Variable Importance Plot created for the Random Forest by this measure.
The most important variables to the model will be highest in the plot / list and have the largest Mean Decrease in Gini Values, conversely, the least important variable will be lowest in the plot, and have the smallest Mean Decrease in Gini values.

```
rownames(imp.df) <- NULL
imp.df %>% knitr::kable(caption = "Importance")
```

Table 1: Importance

features	MDG
redshift	2227.91988
plate	726.91154
mjd	457.11171
z	290.15545
i	254.03472
r	203.28377
g	179.68961
u	108.07015
ra	25.94476
fiberid	25.06785
dec	24.61466
field	21.55145
run	13.97996
camcol	6.76516

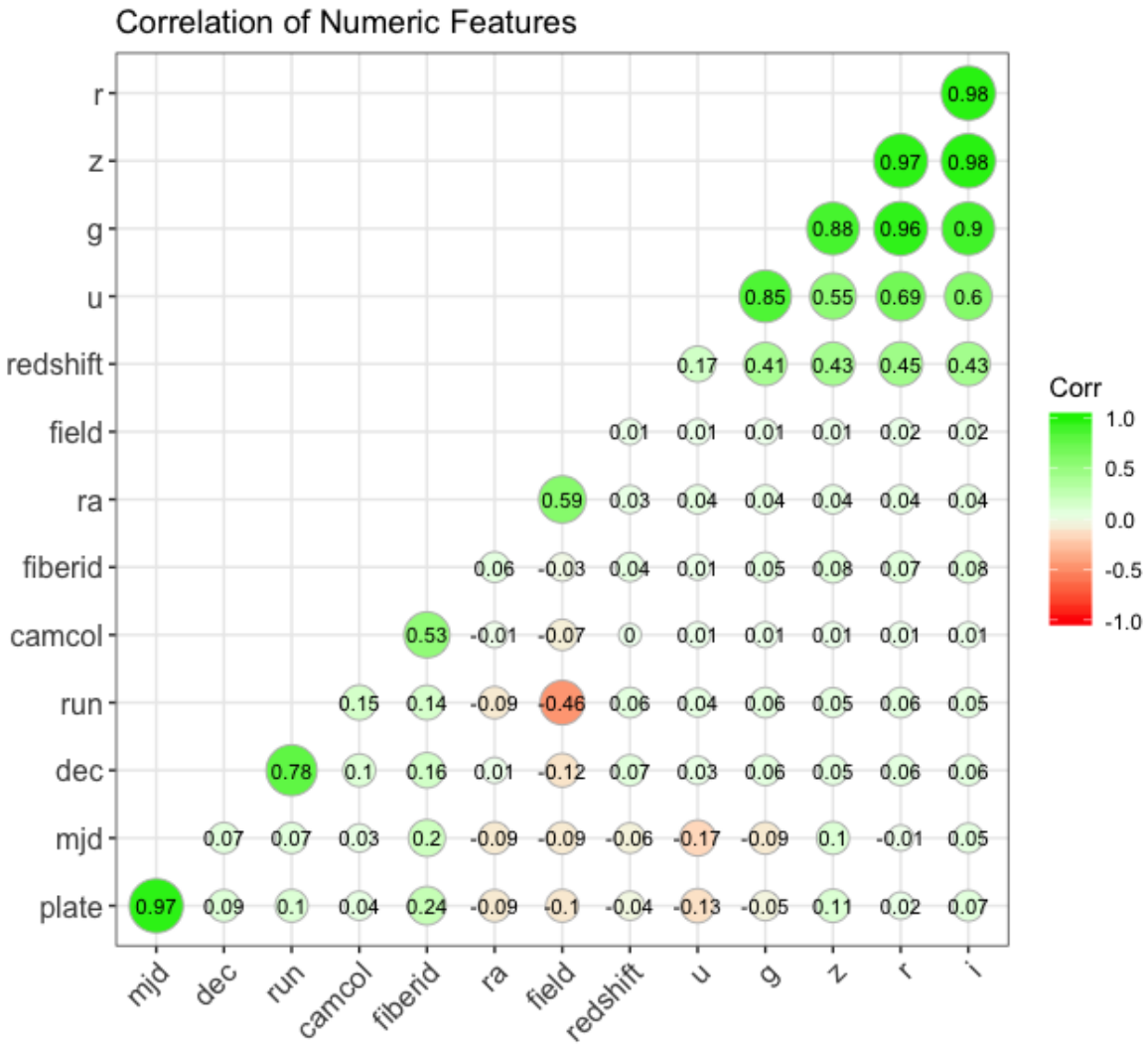
6 Defining the dataset for the model:

6.1 Features Selection

We will select the features to be part of the dataset that will be evaluated.

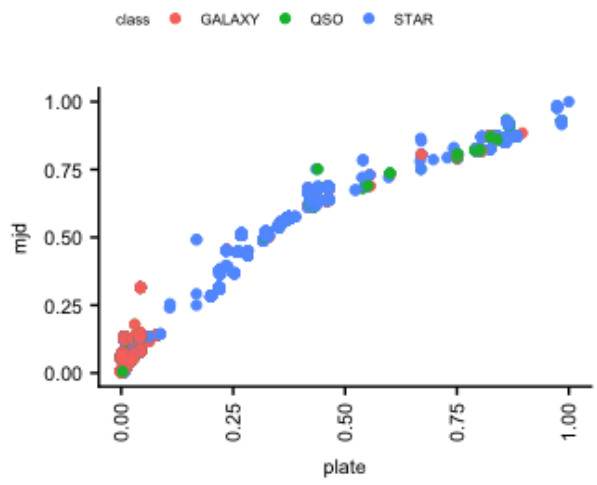
```
sky.model <- sky.train[, -c(1, 10, 13)]
corr <- round(cor(sky.model[, -11], use = "complete.obs"), 2)
ggcorrplot(corr, hc.order = TRUE,
           type = "lower",
           lab = TRUE,
           lab_size = 3,
           method="circle",
```

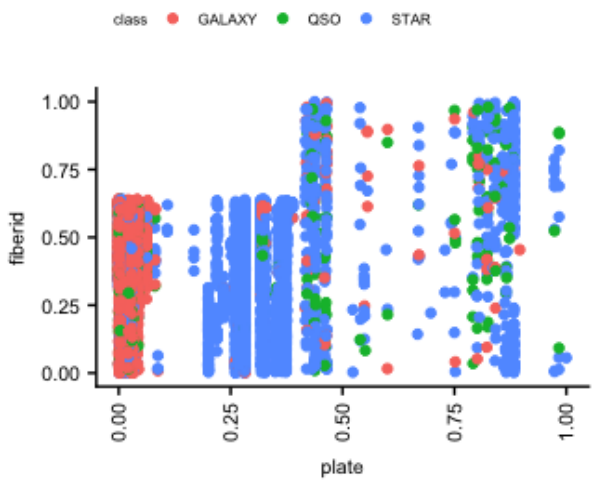
```
          colors = c("red1", "honeydew", "green2"),
          title="Correlation of Numeric Features",
          ggtheme=theme_bw)
```

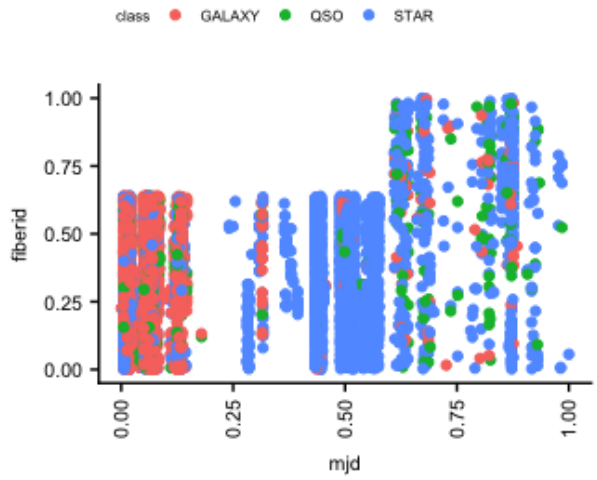


Although the *plate* and *mjd* feature show with a high importance in the MDG evaluation, their correlation is based on the multicollinearity adn they are not independent because every *plate* is linked to a *fiberid* and a *mjd*.
For this reason we decided to **exclude** these features from the dataset that will be used on the model.

```
theme1 <- theme(axis.text.x = element_text(size = 8, angle = 90, hjust = 0.5, vjust = 0.5),
                axis.text.y = element_text(size = 8, angle = 0, hjust = 0.5, vjust = 0.5),
                axis.title.x = element_text(size = 8, angle = 0, hjust = 0.5, vjust = 0.5),
                axis.title.y = element_text(size = 8, angle = 90, hjust = 0.5, vjust = 0.5),
               legend.position="top", legend.text = element_text(size = 6),
               legend.title = element_text(size = 6))
plot_grid(ggplot(sky.train.norm, aes(plate, mjd, col = class)) + geom_point() + theme1,
          ggplot(sky.train.norm, aes(plate, fiberid, col = class)) + geom_point() + theme1,
          ggplot(sky.train.norm, aes(mjd, fiberid, col = class)) + geom_point() + theme1,
          align = "h")
```





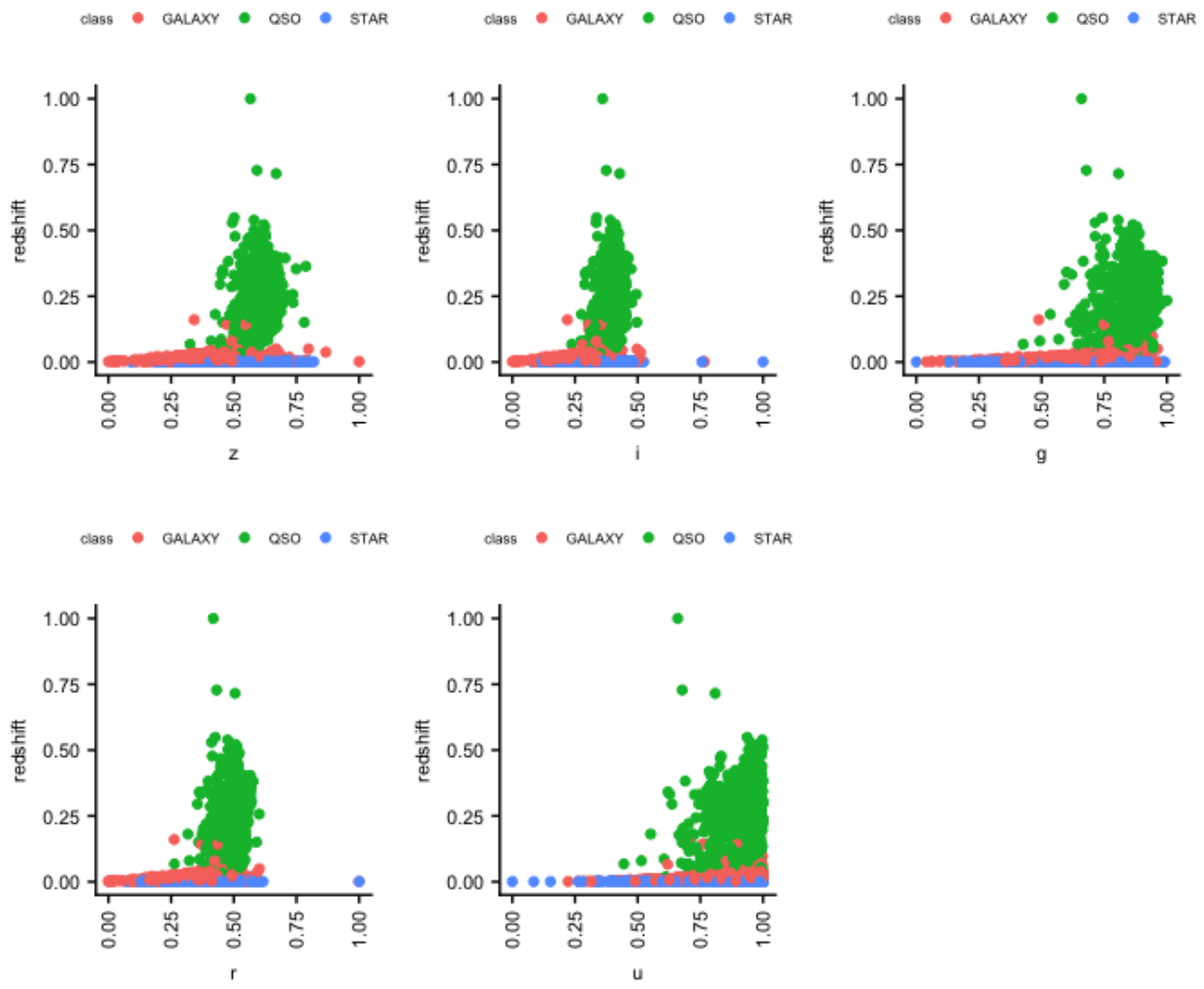


Selecting the columns:

```
sky.model <- sky.train[,c("redshift", "z", "i", "g", "r", "u", "class")] # selecting the columns we nee
```

As a last check we will visualize the **redshift** variable because it showed a greater importance compared to the other variables.
This is the interaction between "redshift" and the other selected features (note: for better visualization we used the normalized dataset to make it easier the comparison).

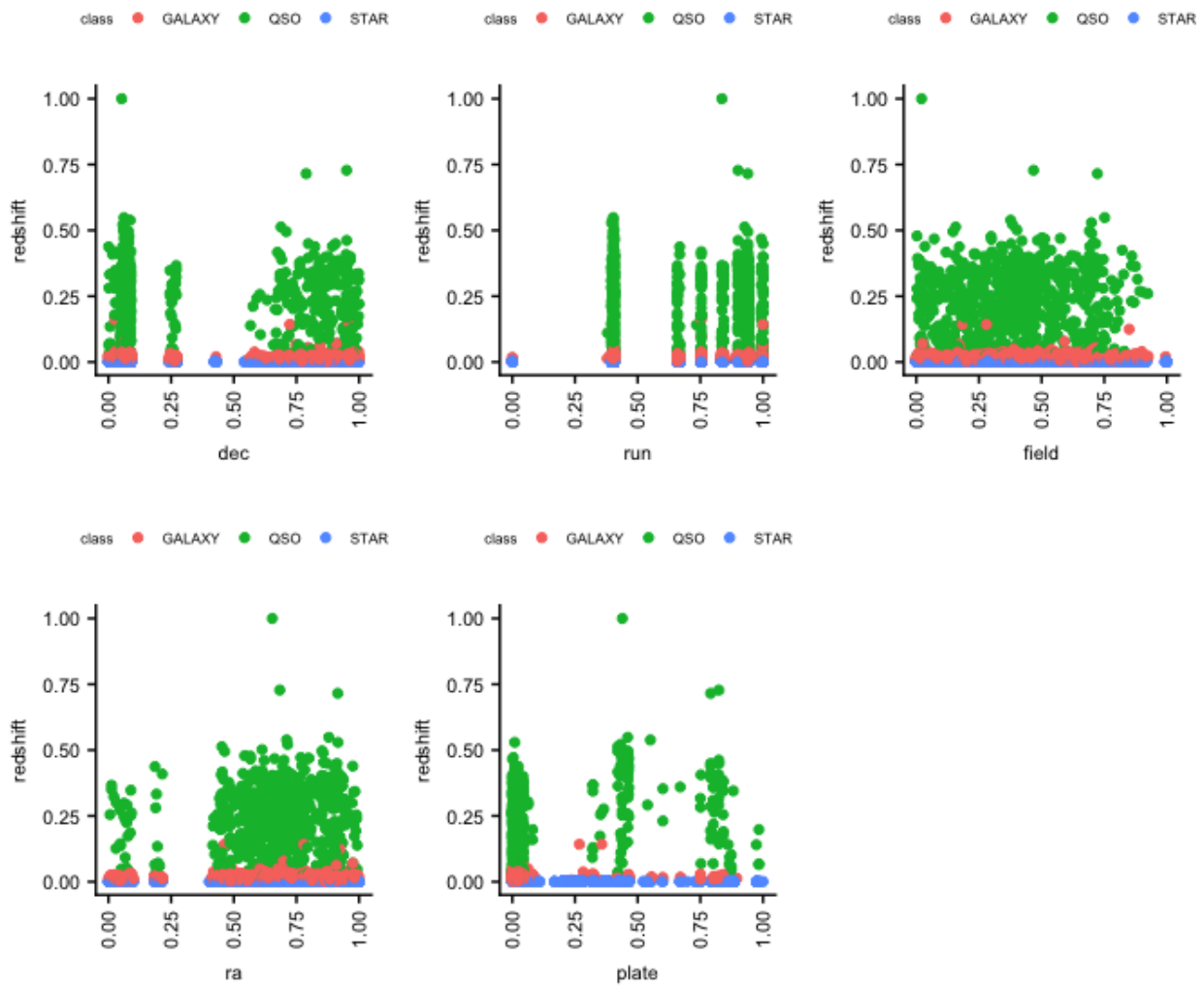
```
plot_grid(ggplot(sky.train.norm, aes(z, redshift, col = class)) + geom_point() + theme1,
          ggplot(sky.train.norm, aes(i, redshift, col = class)) + geom_point() + theme1,
          ggplot(sky.train.norm, aes(g, redshift, col = class)) + geom_point() + theme1,
          ggplot(sky.train.norm, aes(r, redshift, col = class)) + geom_point() + theme1,
          ggplot(sky.train.norm, aes(u, redshift, col = class)) + geom_point() + theme1,
          align = "h") # Thuann_Gunn group
```



We can see on the plot above that the "redshift" how important it is, especially in the low wave length values showing a higher concentration forthe QSO making this feature an important observation for the "class" of the object to be predicted.

Another comprison between redshift and features that have shown to be correlated in the previous analysis that have not been considered for the final model but those features have shown some correlation with the redshift feature:

```
plot_grid(ggplot(sky.train.norm, aes(dec, redshift, col = class)) + geom_point() + theme1,
          ggplot(sky.train.norm, aes(run, redshift, col = class)) + geom_point() + theme1,
          ggplot(sky.train.norm, aes(field, redshift, col = class)) + geom_point() + theme1,
          ggplot(sky.train.norm, aes(ra, redshift, col = class)) + geom_point() + theme1,
          ggplot(sky.train.norm, aes(plate, redshift, col = class)) + geom_point() + theme1,
          align = "h")
```



In all these visualizations we have noted the data "redshift" is concentrated for the GALAXY and STAR classes and that for the QSO class it covers a broader range showing that the data has a greater variability when we consider it as a QSO class feature. Despite the small class imbalance we considered this not to be enough difference to make use of the *SMOTE* library from the **DMwR** package to equalize the classes proportion.

7 Evaluate some algorithms

7.1 Testing Harness

Run algorithms using 20-fold cross validation

```
control <- caret::trainControl(method="cv", number=20)
metric <- "Accuracy"
```

7.2 Build Models

The model to be evaluated will be:
- Linear - nonLinear - Advanced

```
# a) linear algorithms
set.seed(2205)
fit.lda <- caret::train(class ~ . , data=sky.model, method="lda", metric=metric, trControl=control)
# b) nonlinear algorithms
# CART
set.seed(2205)
fit.cart <- caret::train(class ~ . , data=sky.model, method="rpart", metric=metric, trControl=control)
# kNN
set.seed(2205)
fit.knn <- caret::train(class ~ . , data=sky.model, method="knn", metric=metric, trControl=control)
# c) advanced algorithms
# SVM
set.seed(2205)
fit.svm <- caret::train(class ~ . , data=sky.model, method="svmRadial", metric=metric, trControl=contro
# Random Forest
set.seed(2205)
fit.rf <- caret::train(class ~ . , data=sky.model, method="rf", metric=metric, trControl=control)
```

7.3 Select best model

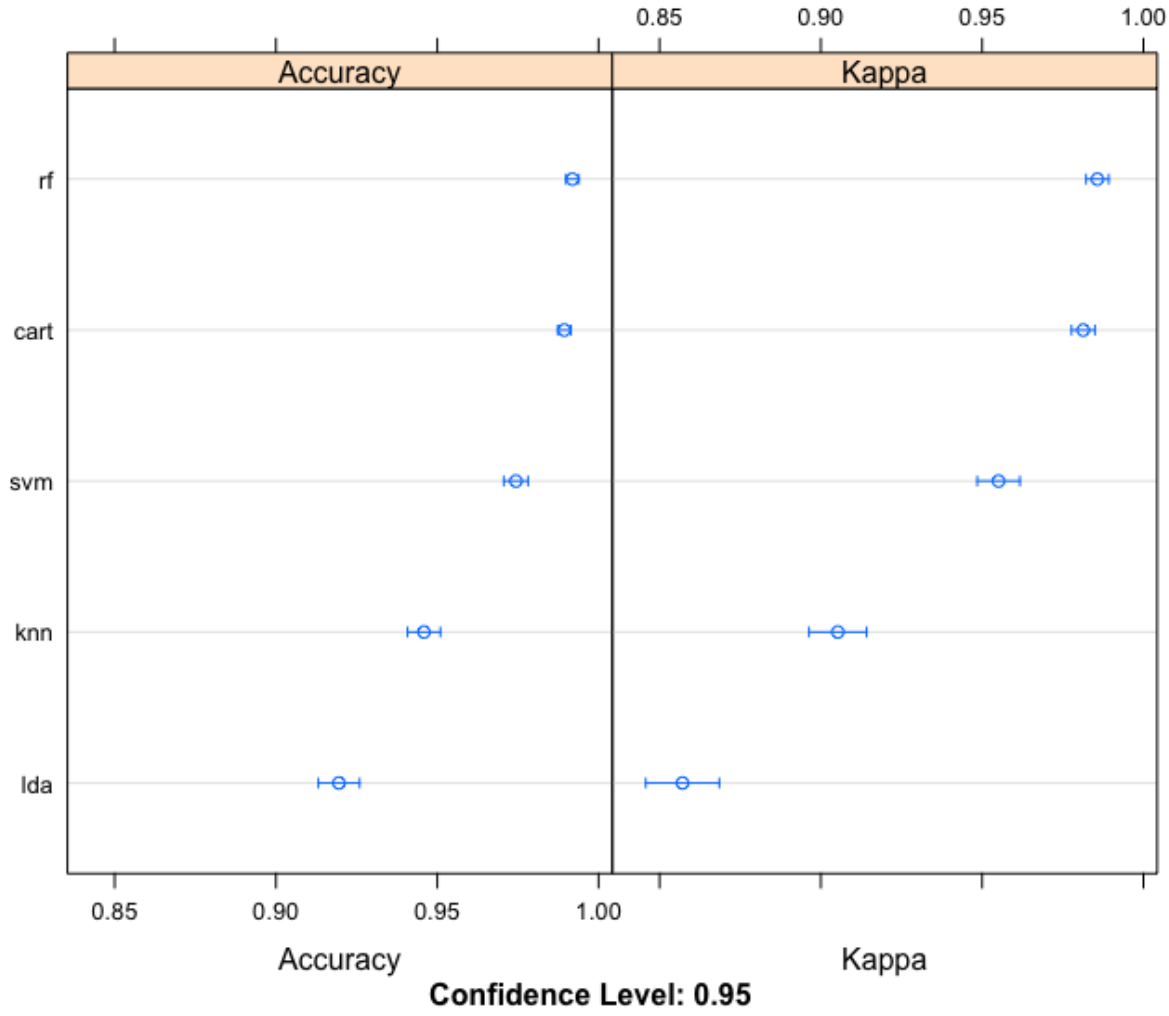
Summarizing the accuracy of the models:

```
results <- resamples(list(lda=fit.lda, cart=fit.cart, knn=fit.knn, svm=fit.svm,
                          rf=fit.rf))
summary(results)
```

```
##
## Call:
## summary.resamples(object = results)
##
## Models: lda, cart, knn, svm, rf
## Number of resamples: 20
##
## Accuracy
##        Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lda  0.8950 0.9100000 0.9162500 0.9195171 0.9262500 0.9548872    0
## cart 0.9825 0.9868750 0.9887375 0.9893762 0.9931250 0.9975062    0
## knn  0.9175 0.9400000 0.9475000 0.9458812 0.9532138 0.9625000    0
## svm  0.9575 0.9718750 0.9725000 0.9743794 0.9800000 0.9899749    0
## rf   0.9825 0.9899937 0.9925000 0.9918756 0.9950000 0.9975062    0
##
## Kappa
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lda  0.8143810 0.8403909 0.8510749 0.8571316 0.8684659 0.9203434    0
## cart 0.9692633 0.9769201 0.9802139 0.9813341 0.9879674 0.9956218    0
## knn  0.8560052 0.8951720 0.9081886 0.9052629 0.9180869 0.9345635    0
## svm  0.9254893 0.9505588 0.9519440 0.9550926 0.9649885 0.9824272    0
## rf   0.9692720 0.9823488 0.9868163 0.9857321 0.9912365 0.9956313    0
```

Compare the accuracy of the models:

```
dotplot(results)
```



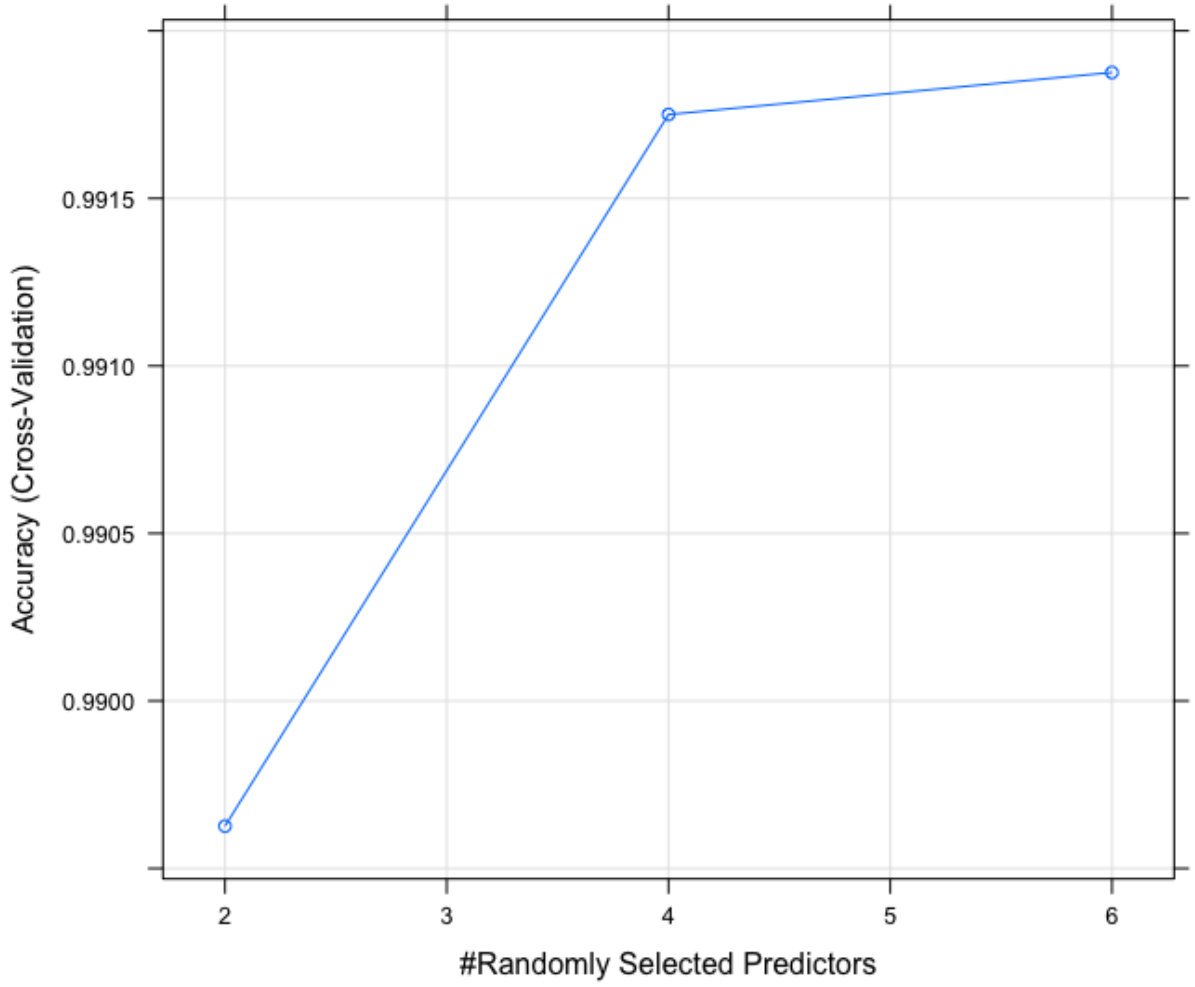

```
print(fit.rf)
```

```
## Random Forest
##
## 8001 samples
##    6 predictor
##    3 classes: 'GALAXY', 'QSO', 'STAR'
##
## No pre-processing
## Resampling: Cross-Validated (20 fold)
## Summary of sample sizes: 7601, 7601, 7601, 7601, 7601, 7601, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##   2     0.9896262  0.9817642
##   4     0.9917503  0.9855026
##   6     0.9918756  0.9857321
##
```

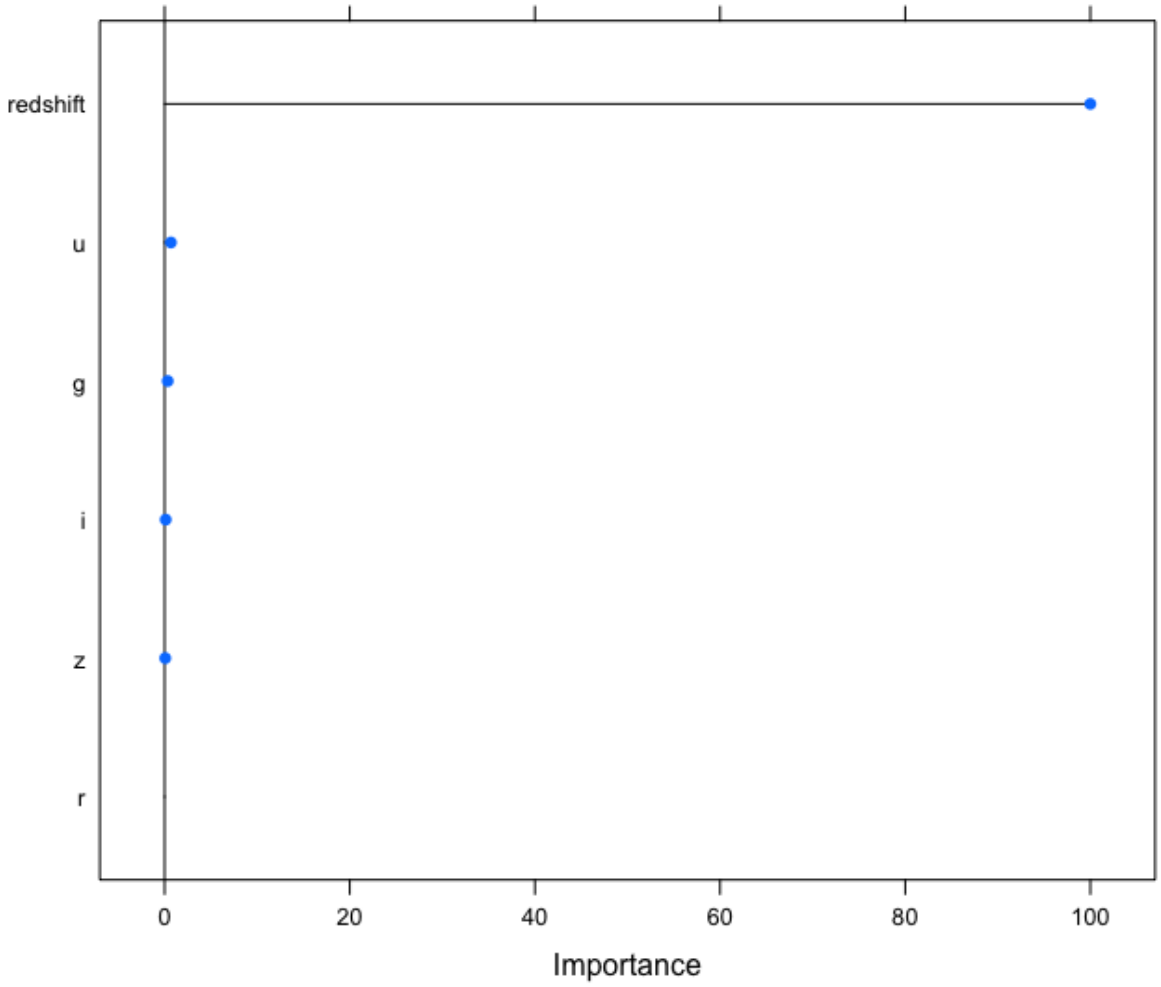
```
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 6.
```

The Accuracy vs. predictors and variable importance:

```
plot(fit.rf)
```




```
plot(varImp(fit.rf))
```



8 Make Predictions

8.1 Estimating the skill of RF (randomForest) on the validation dataset.

columns to be discarded:

```
validation.model <- validation[,c("redshift", "z", "i", "g", "r", "u", "class")]
```

Predict:

```
set.seed(2205)
predictions <- predict(fit.rf, validation.model)
caret::confusionMatrix(predictions, validation.model$class)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction GALAXY QSO STAR
##     GALAXY    991   4    2
##     QSO         7 166    0
```

```
##     STAR        1   0  828
##
## Overall Statistics
##
##                Accuracy : 0.993
##                  95% CI : (0.9883, 0.9962)
##     No Information Rate : 0.4997
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9877
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: GALAXY Class: QSO Class: STAR
## Sensitivity                 0.9920    0.97647      0.9976
## Specificity                 0.9940    0.99617      0.9991
## Pos Pred Value              0.9940    0.95954      0.9988
## Neg Pred Value              0.9920    0.99781      0.9983
## Prevalence                  0.4997    0.08504      0.4152
## Detection Rate              0.4957    0.08304      0.4142
## Detection Prevalence        0.4987    0.08654      0.4147
## Balanced Accuracy           0.9930    0.98632      0.9984
```

9 Results and Conclusion

The conclusion we came to is that the validation set cases are not so hard to predict and the data is well clustered around the classes.
We tried to discard the "redshift" feature in order to maximize the weight of the other features in previuos models but we only achieved an Accuracy = 0.9305 in the train set and Accuracy = 0.9235 in the validation (test) set having many cases of False Positives around 10% FP's when predicting STAR (predicted GALAXY instead) and 9% FP's when predicting QSO in relation to the other two classes.
The decision is to keep the "redshift" feature along with the Thuann_gunn group.
The final Accuracy is 0.9915019

Notes:
- Time to run the whole code 6.31 mins
- Computer used:
- MacBook Pro
- Processor: 2.4 GHz Intel Core i5
- Memory: 8 GB 1600 MHz DDR3
- Graphics: Intel Iris 1536 MB