Results: Compare and Contrast

3/3/2020

Results: Compare and Contrast

This next section is my attempt to compare and contrast the ML Algorithms prepared in this report.

As you will remember, I have studied 6 different M.L. algorithms using protein amino acid percent composition data from two classes. Class number 1 is my positive control which is a set of Myoglobin proteins, while the second class is a control group of human proteins that do not have Fe binding centers.

Group	Class	N of Class	Range of Groups
Controls	0 or (-)	1217	1, ..., 1217
Myoglobin	1 or (+)	1124	1218, ..., 2341

The Six M.L Algorithms consist of:

Name	Type	Output Used For Graphing
Principal Component Analysis	Unsupervised	Anomalies > Abs(3σ)
Logistic Regression	Supervised	FP & FN
SVM-linear	Supervised	FP & FN
SVM-polynomial kernel	Supervised	FP & FN
SVM-radial basis function kernel	Supervised	FP & FN
Neural Network	Supervised	FP & FN

==========

Scatter Plots of Anomalies Vs. FP & FN Outputs

To obtain False-Positive (fp) from sets: {-}

1. False-Postives $\stackrel{\text{def}}{=}$ $\{\text{obs} = 0 \quad \wedge \quad \text{pred} = 1\}$
2. False-Negatives $\stackrel{\text{def}}{=}$ $\{\text{obs} = 1 \quad \wedge \quad \text{pred} = 0\}$

Anomalies Inner Joined with PC

```
## Load Libraries
Libraries = c("knitr", "readr")

for(p in Libraries){
    library(p, character.only = TRUE)
```

```
}
opts_chunk$set(cache = TRUE, fig.align = "center")
```

Prepare PCA: PC1 and PC2 for all 2340 proteins

```
norm_c_m_20aa <- read_csv("./00-data/03-ml_results/norm_c_m_20aa.csv")

pca_values <- prcomp(norm_c_m_20aa)

row_pc12 <- cbind(rowNum = 1:2340, PC1 = pca_values$x[,1], PC2 = pca_values$x[,2])
# dim(row_pc12)
```

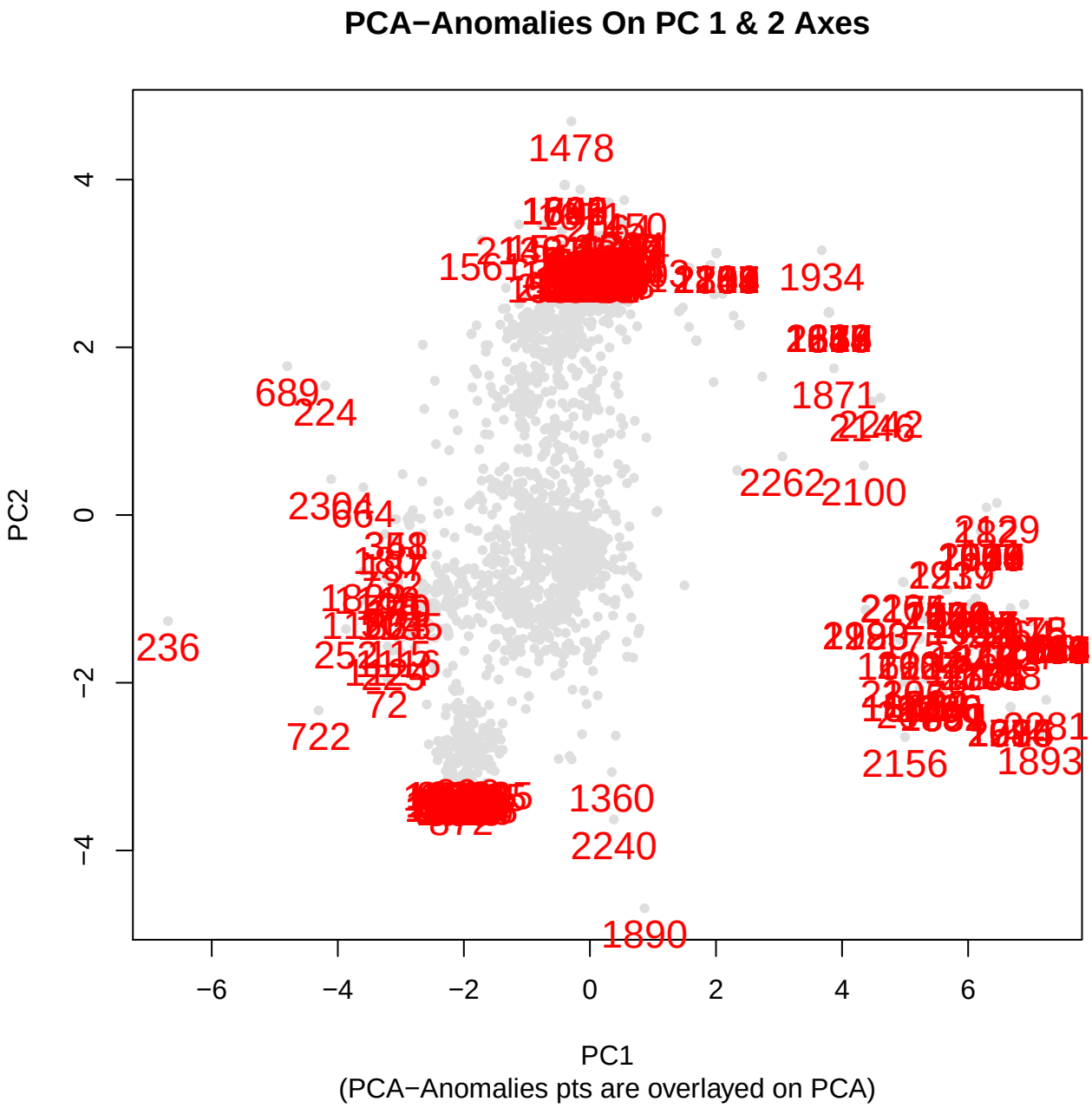
Load Files & change column names

Inner-join (merge) with PC 1&2 (row_pc12)

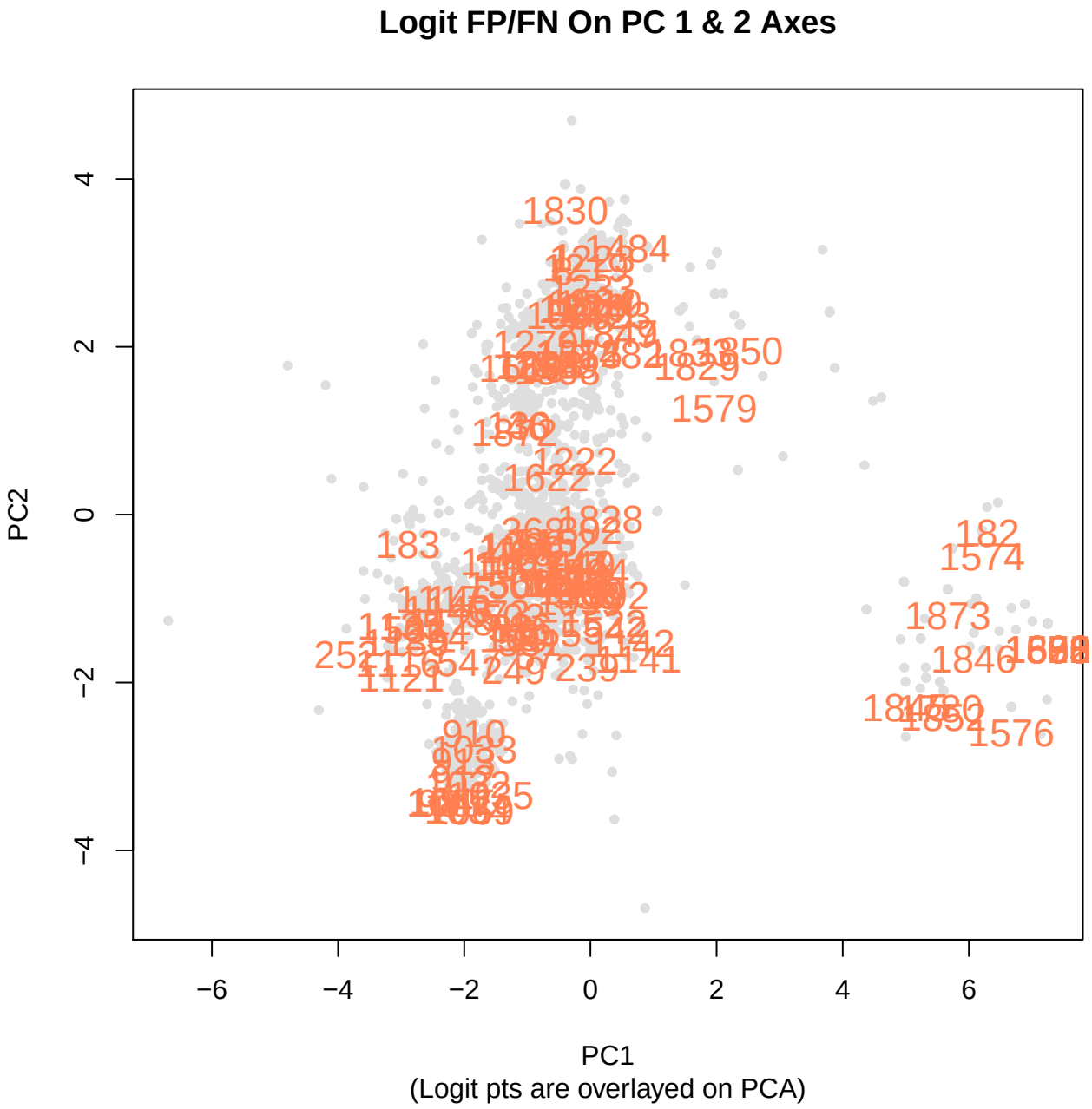
```
logit_pc    <- merge(x = logit_nums,    y = row_pc12, by = "rowNum")
pca_pc      <- merge(x = pca_outliers,  y = row_pc12, by = "rowNum")
# rf_pc       <- merge(x = rf_nums,       y = row_pc12, by = "rowNum")
svm_lin_pc  <- merge(x = svm_lin_nums,  y = row_pc12, by = "rowNum")
svm_poly_pc <- merge(x = svm_poly_nums, y = row_pc12, by = "rowNum")
svm_rbf_pc  <- merge(x = svm_rbf_nums,  y = row_pc12, by = "rowNum")
```

Generate FP/FN & PCA-Anomalies on PCA Plots

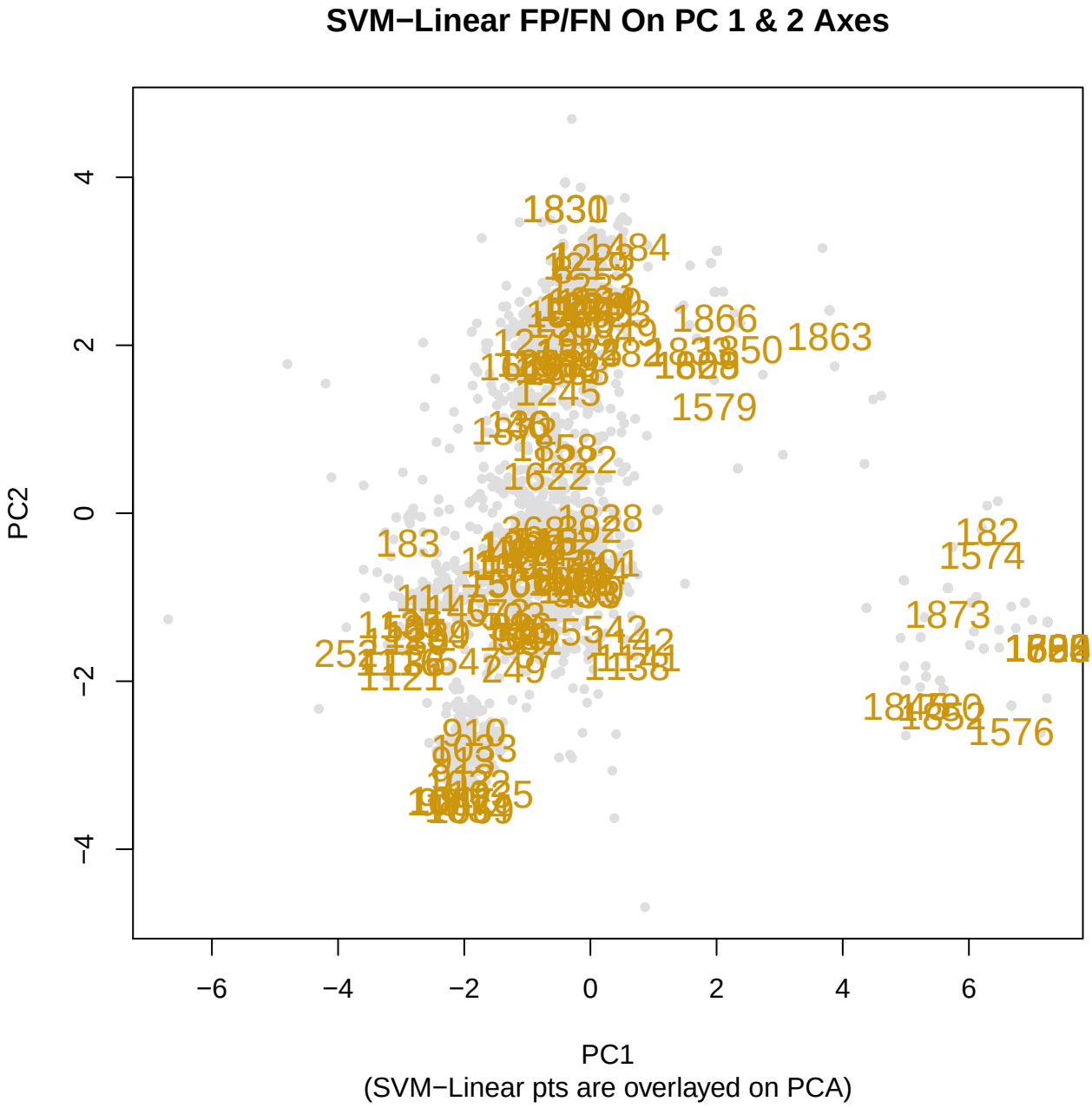
PCA-Anomalies Plot



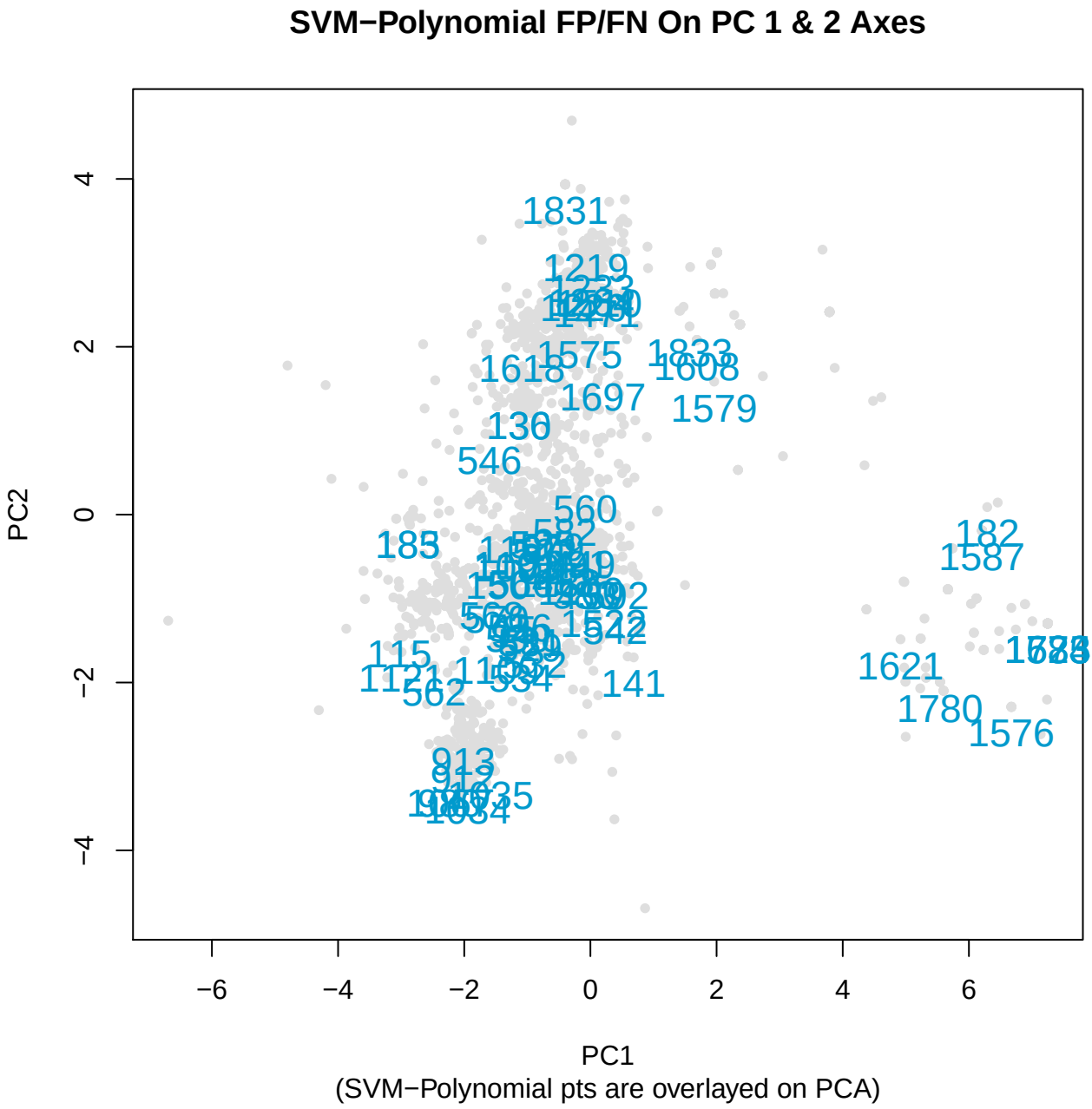
Logit Plot


Logit FP/FN On PC 1 & 2 Axes
PC2
PC1
(Logit pts are overlayed on PCA)
−6
−4
−2
0
2
4
6
4
2
0
−2
−4

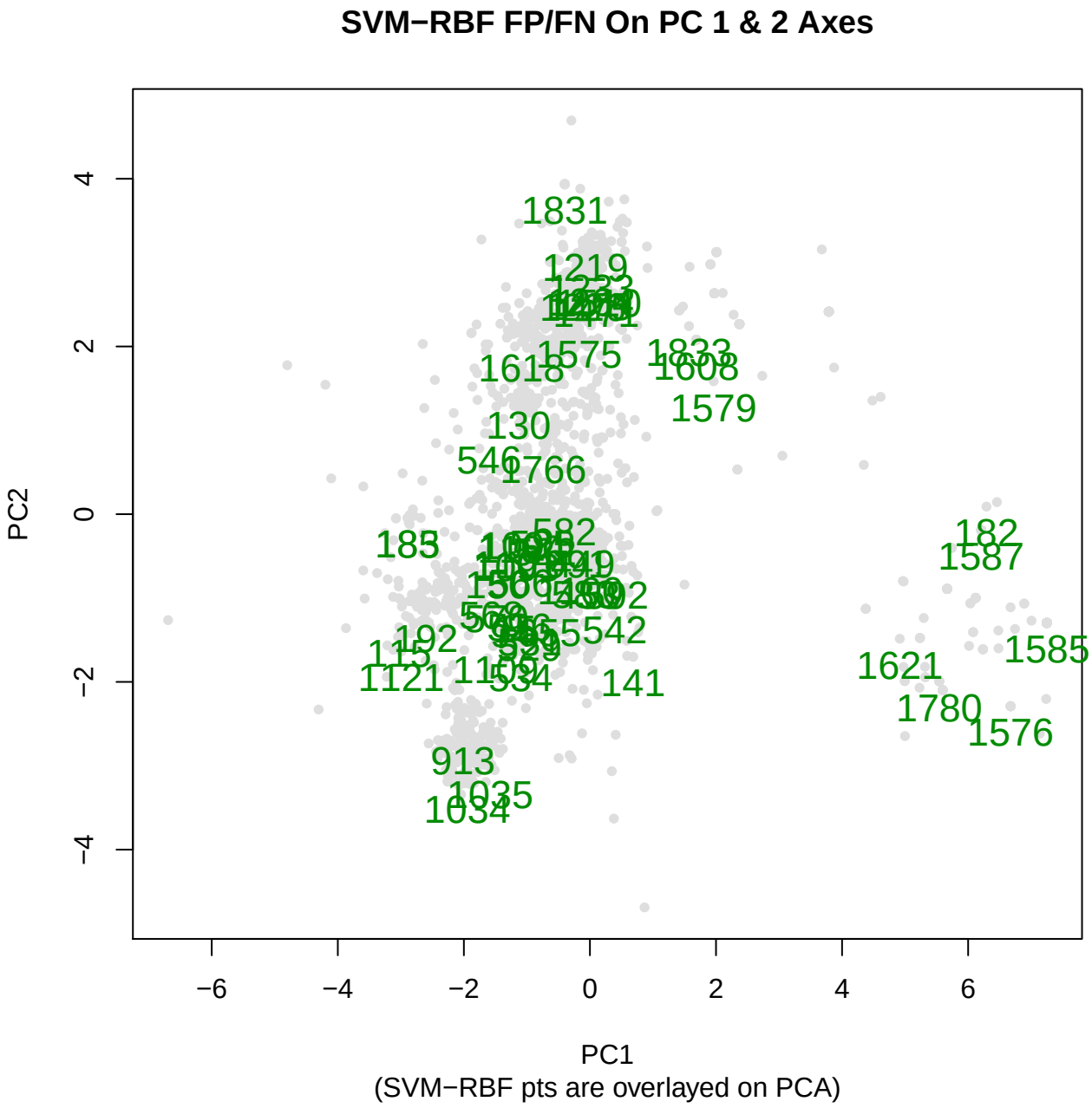
SVM-Linear Plot



SVM-Polynomial Plot



SVM-Radial Basis Function Plot



Neural Network Function Plot

Statistical Learning Method Vs Total Number of FP/FN

Statistical Method	Total Number Produced	Unique	Total/Unique
Principal Componnent Analysis	461	460	1.002
Logit	537	119	4.51
SVM Linear	496	125	3.97

Statistical Method	Total Number Produced	Unique	Total/Unique
SVM Polynomial	278	70	3.97
SVM Radial Basis Function	244	58	4.21
Random Forest	190	46	4.13
Deep Learning	347	133	2.61

==========

```
# Load Libraries
Libraries <- c("doMC", "knitr", "readr", "caret", "nnet", "caretEnsemble", "e1071", "kernlab")
for (p in Libraries) {
    library(p, character.only = TRUE)
}
```

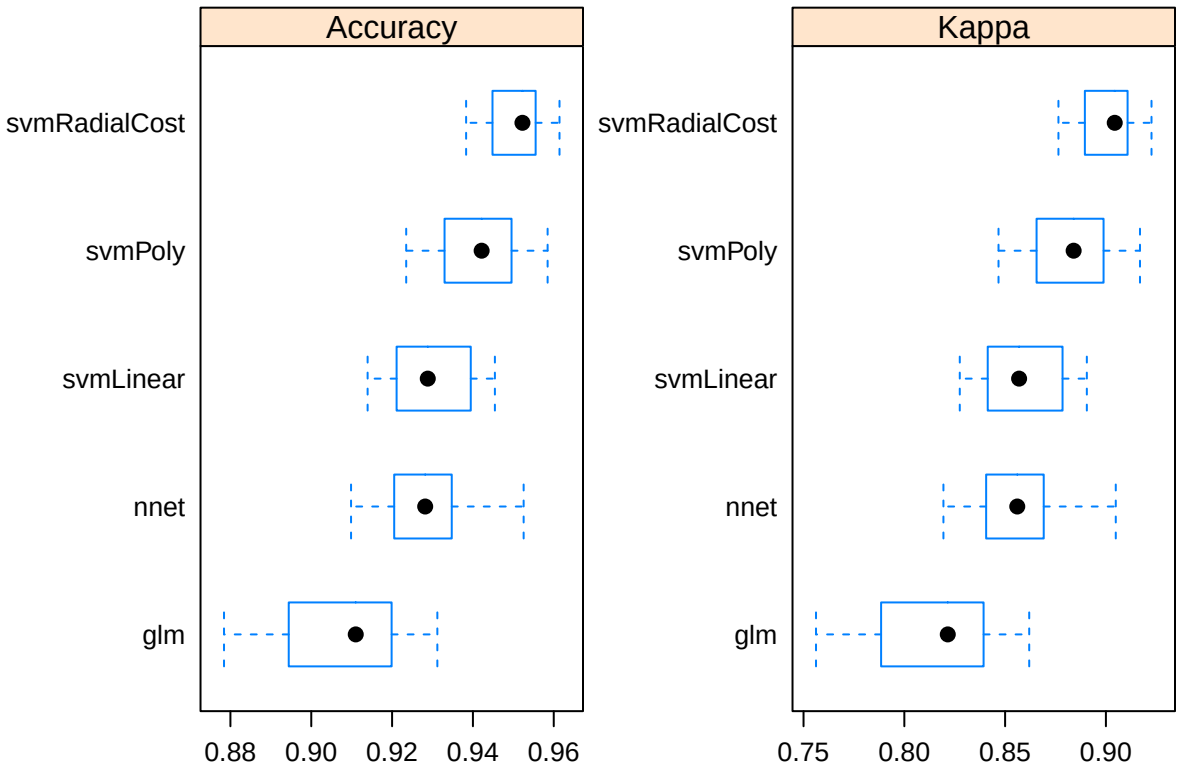
```
# Import data & data handling
c_m_TRANSFORMED <- read_csv("./00-data/02-aac_dpc_values/c_m_TRANSFORMED.csv",
                            col_types = cols(Class = col_factor(levels = c("0", "1")),
                                             PID = col_skip(),
                                             TotalAA = col_skip()))
```

```
# Partition data into training and testing sets
set.seed(1000)
index <- createDataPartition(c_m_TRANSFORMED$Class, p = 0.8, list = FALSE)

training_set <- c_m_TRANSFORMED[ index,]
test_set     <- c_m_TRANSFORMED[-index,]

Class_test <- as.factor(test_set$Class)
```

Stacking Algorithms - Run multiple algorithms in one call.

Plot the resamples output to compare the models.

```
# Box plots to compare models
scales <- list(x = list(relation = "free"),
               y = list(relation = "free"))
bwplot(results, scales = scales)
```


Accuracy
svmRadialCost
svmPoly
svmLinear
nnet
glm
0.88
0.90
0.92
0.94
0.96
Kappa
svmRadialCost
svmPoly
svmLinear
nnet
glm
0.75
0.80
0.85
0.90