Small data regression

Load some packages

```
library(data.table)
library(caret)
library(caTools)
```

Problem statement

- 2,000 rows
- 500 columns
- 200 of the columns are actually noise
- 50% of the data is the positive class

```
nrows <- 2000
ncols <- 500
noise_vars <- 200
pct <- .50
```

Build the dataset

Simulate the data:

```
set.seed(42)
X <- matrix(rnorm(nrows*ncols*3), ncol=ncols)
CF <- runif(ncols, min=-1, max=1)
CF[sample(1:ncols, noise_vars)] <- 0
Y <- X %*% CF + rnorm(nrows, sd=10)
```

Split the classes:

```
cutoff <- quantile(Y, 1-pct)
Y <- as.integer(Y > cutoff)
```

Create 2 training sets and ONE test set:

```
dat <- data.table(Y=Y, X)
train_a <- dat[1:nrows,]
train_b <- dat[(nrows+1):(nrows*2),]
test <- dat[(nrows*2+1):(nrows*3),]
```

Fit models and predict on the test set

```
model_a <- glm(Y ~ ., train_a, family='binomial')
model_b <- glm(Y ~ ., train_b, family='binomial')

pred_a <- predict(model_a, test, type = 'response')
pred_b <- predict(model_b, test, type = 'response')
```

Compare results

Both models are good models, with a high AUC:

```
preds <- cbind(pred_a, pred_b)
colAUC(preds, test$Y)
```

```
##            pred_a    pred_b
## 0 vs. 1 0.7461173 0.7337159
```

However, the 2 models do not agree with each other. They have a 100% mis-match between their predictions:

```
sum(round(pred_a, 6) == round(pred_b, 6))
```

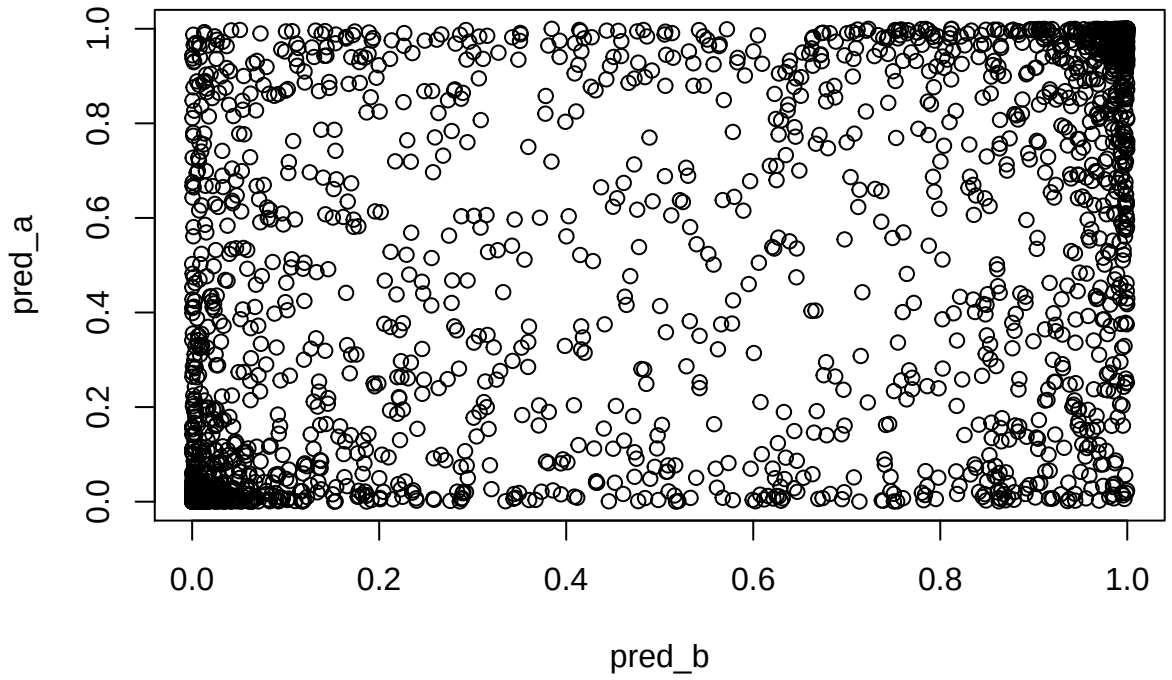
```
## [1] 0
```

The correlation between their predicitons is also low:

```
cor(pred_a, pred_b)
```

```
## [1] 0.5024911
```

```
plot(pred_a ~ pred_b)
```



In *theory*, these models should be identical. In *practice*, the predicted probabilities differ by up to 99.1%.