Zadanie domowe 1

Wojciech Kretowicz

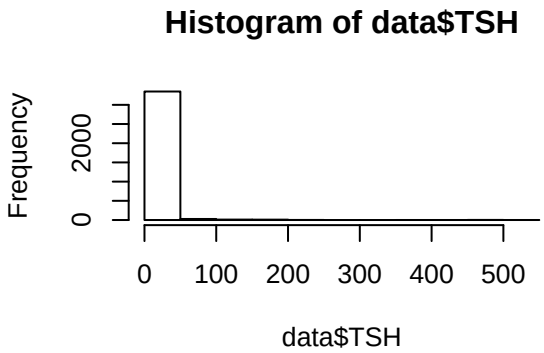
17 kwietnia 2020

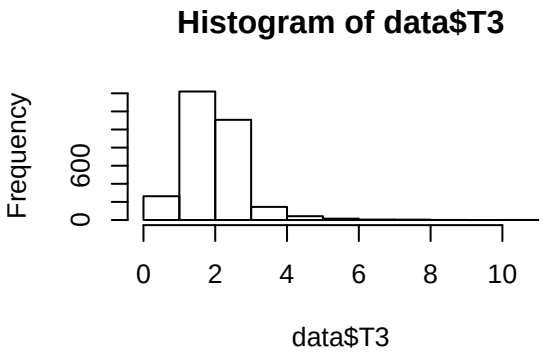
Glance at the data

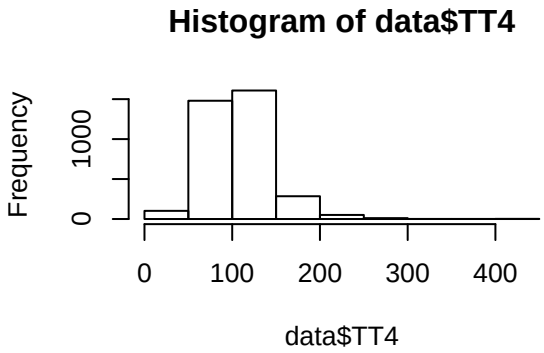
```
## Data '38' file 'description.xml' found in cache.
```

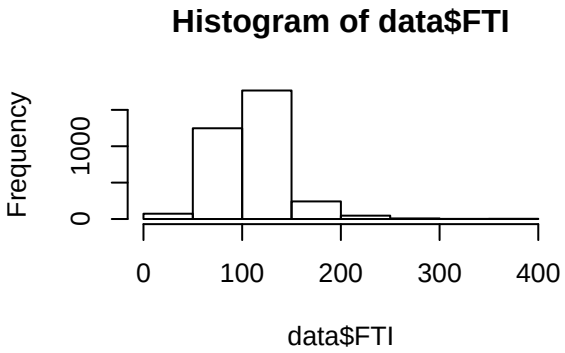
```
## Data '38' file 'dataset.arff' found in cache.
```

```
par(mfrow=c(2,2))
hist(data$TSH)
hist(data$T3)
hist(data$TT4)
hist(data$FTI)
```









Preprocessing

Laboratory tests

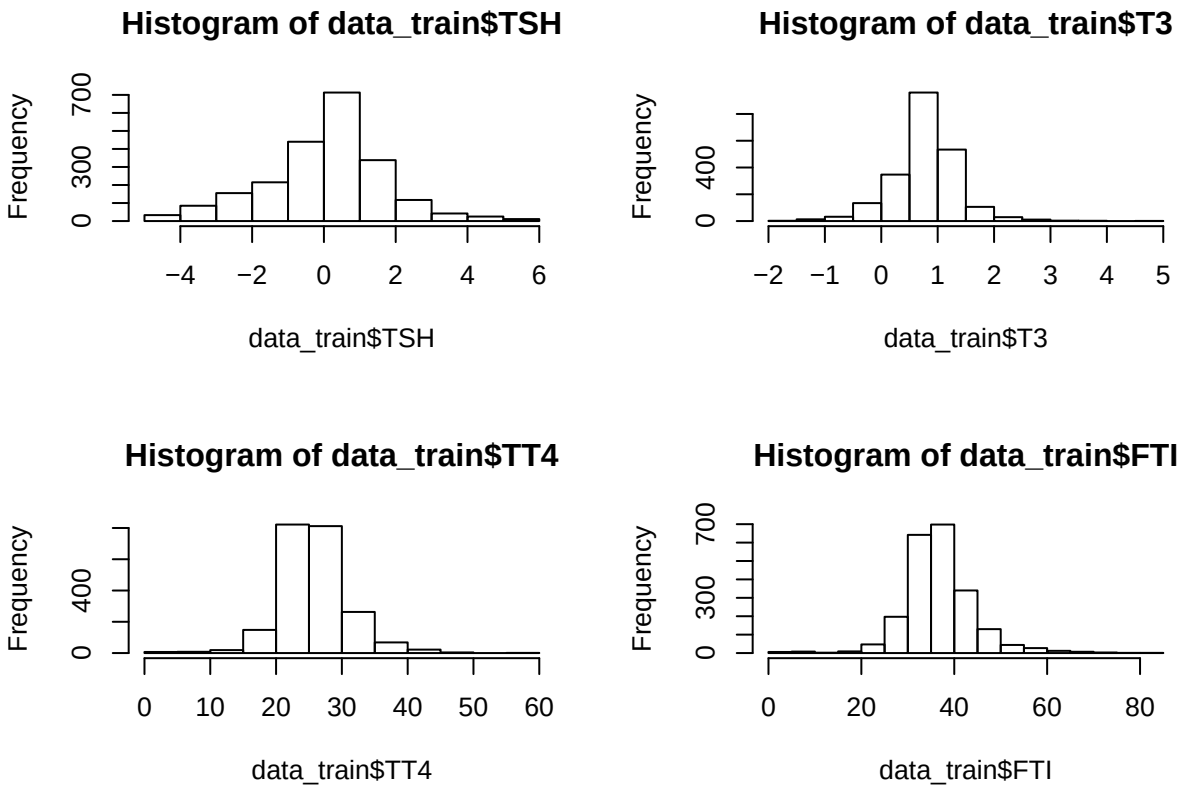
First, I checked laboratory tests to find out, what are possible values of hormones in a human body. Taking very large margin, I chose following values:

- $TSH < 100$
- $T3 < 30$
- $TT4 < 400$
- $FTI < 400$

Distributions

Furthermore, I transformed all numerical values with boxcox trasformation, resulting in these distributions:

```
par(mfrow=c(2,2))
hist(data_train$TSH)
hist(data_train$T3)
hist(data_train$TT4)
hist(data_train$FTI)
```



Missing values

At this moment missing values and features with suffix “measured” are negligible - there are only 2 features with missing values with aroung 3% of missing values. I used mice package with ‘pmm’ method to fill these. Them I dropped all features with suffix ‘measure’.

Black box

I have constructed black box with "ranger" to have a comparison and explained it to make accumulated dependency plot and feature importance. It showed me the most important variables in this task.

```
library(DALEX)
```

```
## Welcome to DALEX (version: 0.4.9).
## Find examples and detailed introduction at: https://pbiecek.github.io/PM_VEE/
## Additional features will be available after installation of: ALEPlot, breakDown, pdp, factorMerger, g
## Use 'install_dependencies()' to get all suggested dependencies
```

```
library(mlr)
black_box = mlr::makeLearner('classif.ranger', predict.type = 'prob')
```

Results

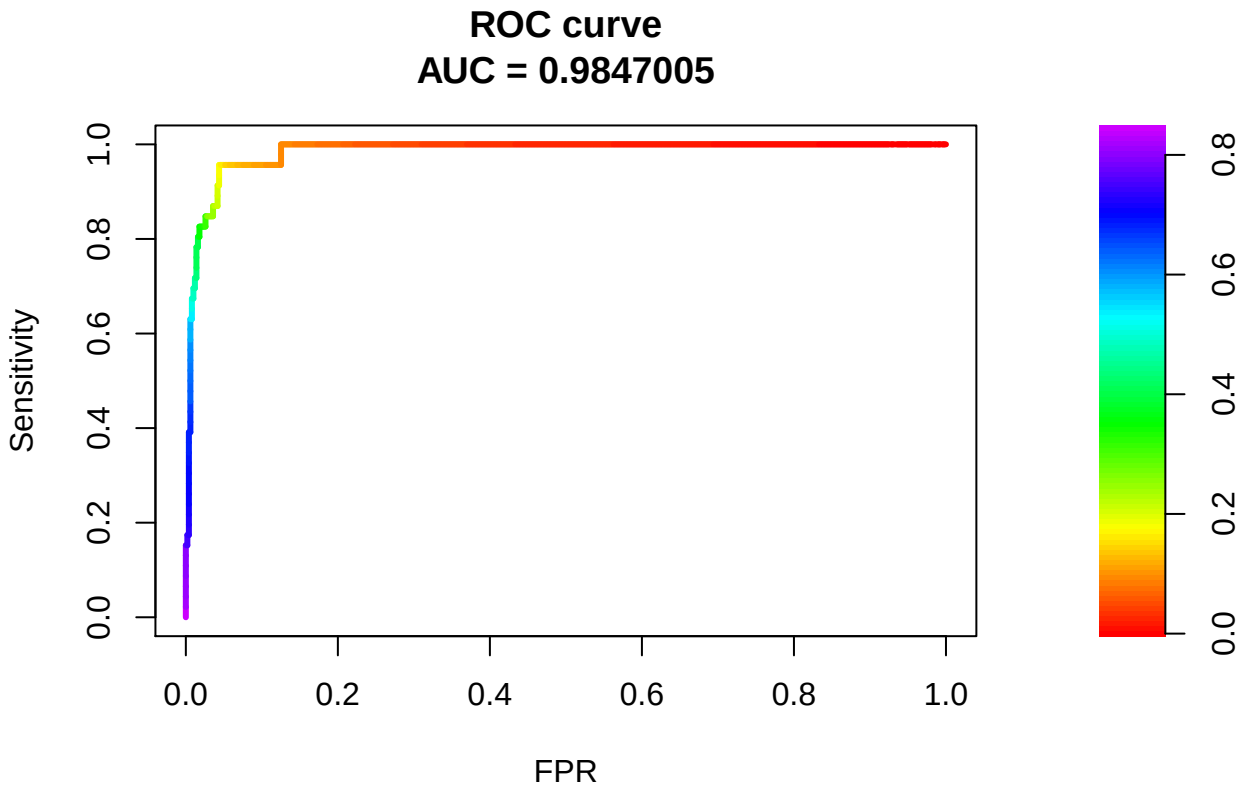
On 5-cv:

- AUPRC: $95.9 + -1.4$
- AUROC: $99.5 + -2.5$

On test data:

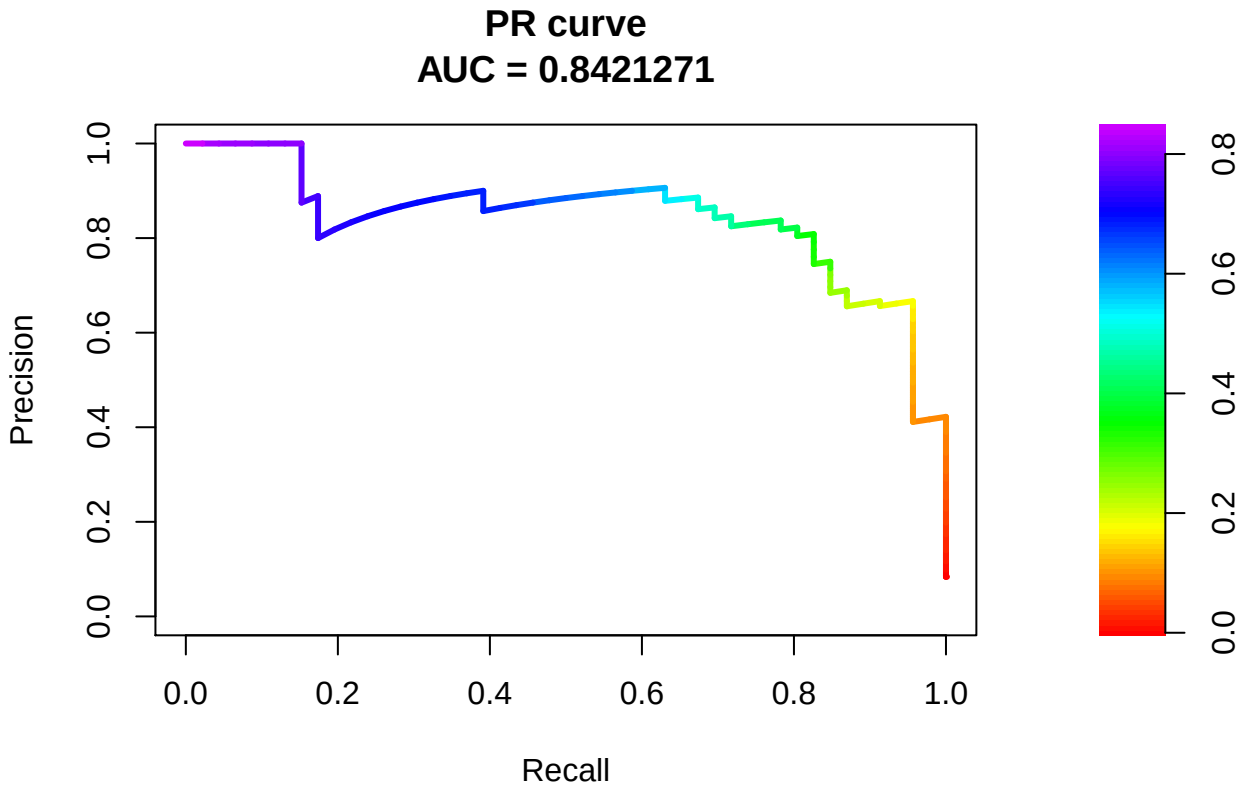
- AUPRC: 98.5
- AUROC: 84.0

```
roc <- roc.curve(scores.class0 = fg, scores.class1 = bg, curve = T)
plot(roc)
```



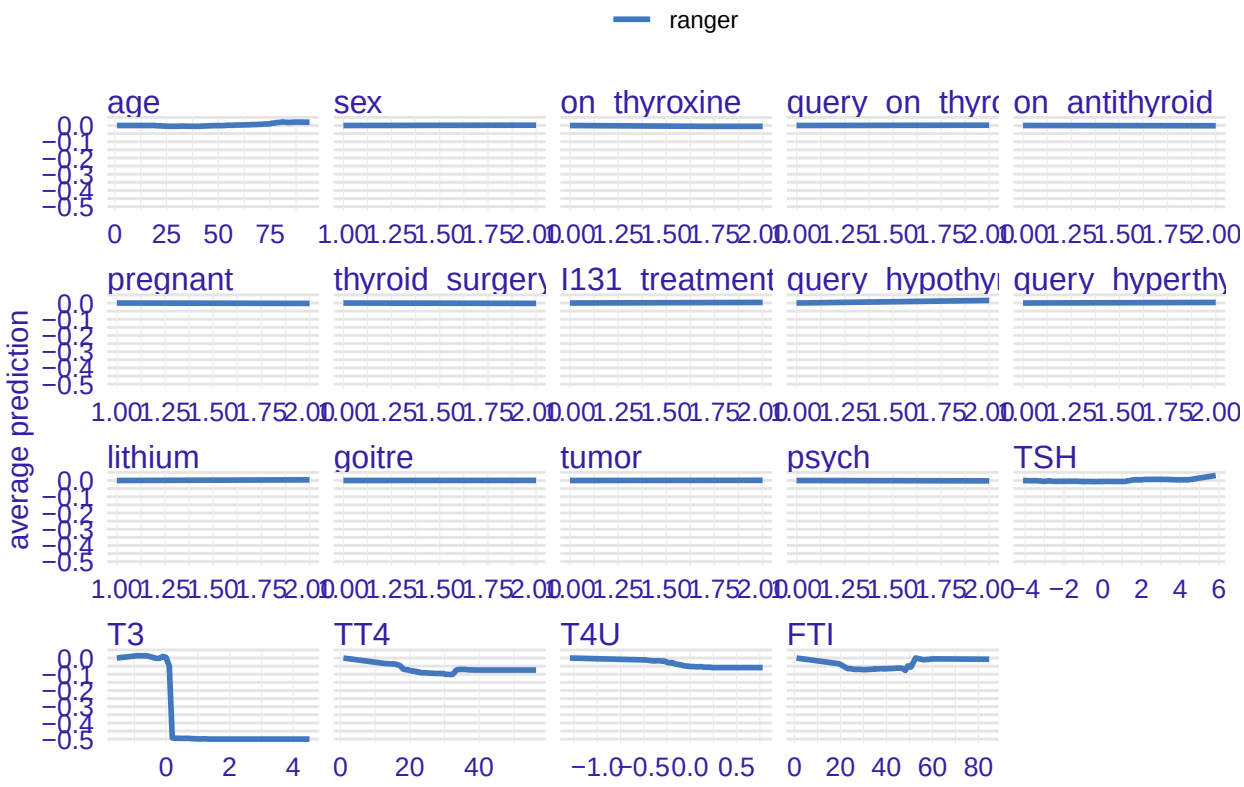
```r
set.seed(77)

pr <- pr.curve(scores.class0 = fg, scores.class1 = bg, curve = T)
plot(pr)
```

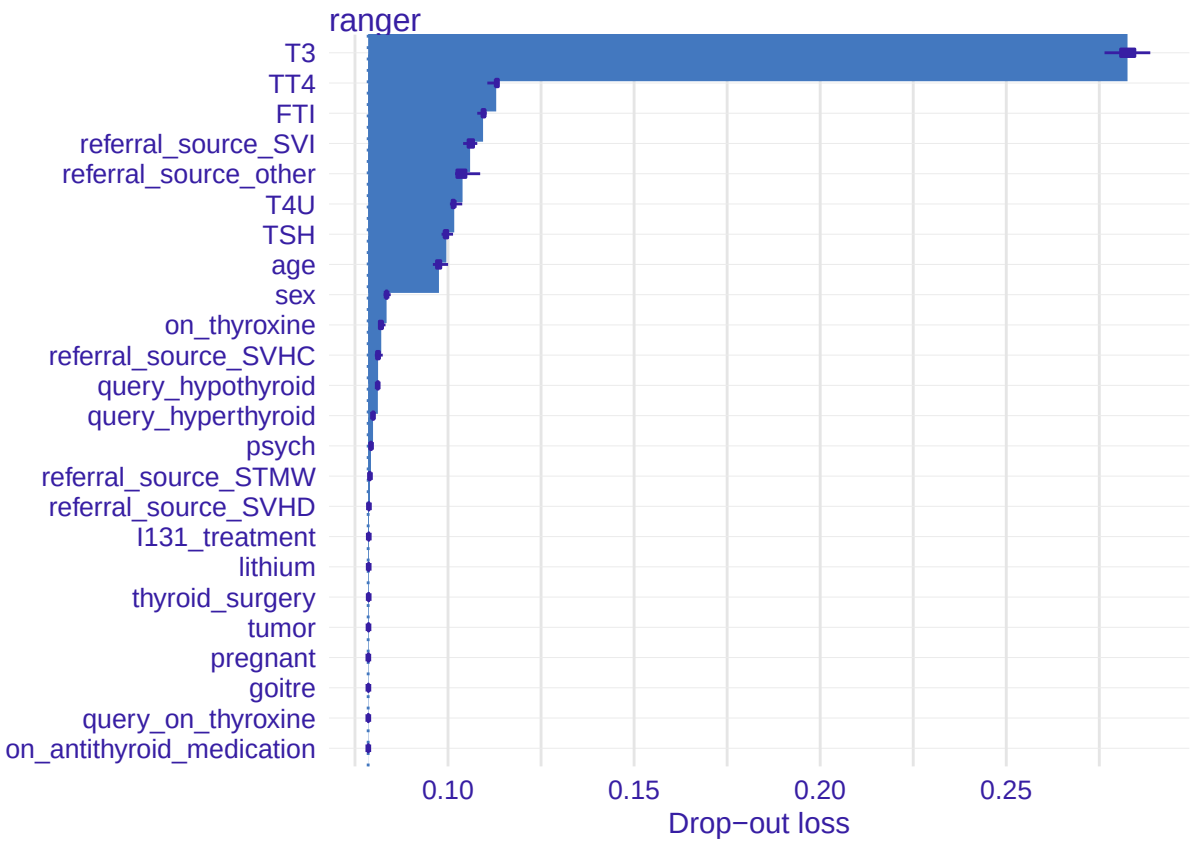


Explaining black box

```r
exp = explain(black_box, data_train[,-25],data_train$Class, verbose=FALSE)
library(ingredients)
acc = ingredients::accumulated_dependency(exp, variables = c("age", "sex", "on_thyroxine", "query_on_th
plot(acc)
```




```
fi = ingredients::feature_importance(exp)
plot(fi)
```



Model

I have chosen logistic regression and rpart for modelling, because both are easily interpretable. However, rpart had much better results.

Feature engineering

I added few more features based on cross val scores of auprc and black box explanation:

- $T3^2$
- $T4U^2$
- $TT4^2$
- $T3/FTI$

Tuning

I used random search for tuning rpart.

Results

After all I have got on 5-cv:

- AUPRC: $92.9 + -4.1$
- AUROC: $96.2 + -2.9$

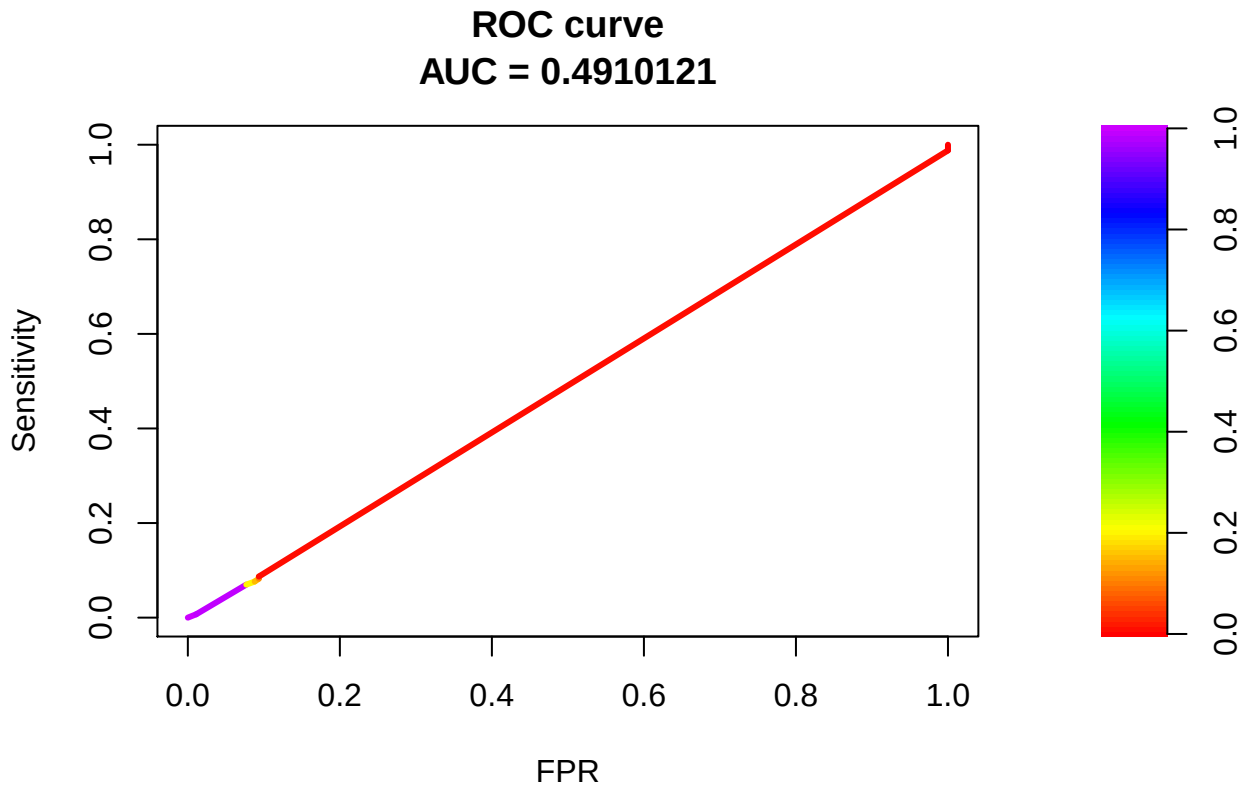
And on the test set:

- AUPRC: 91.3

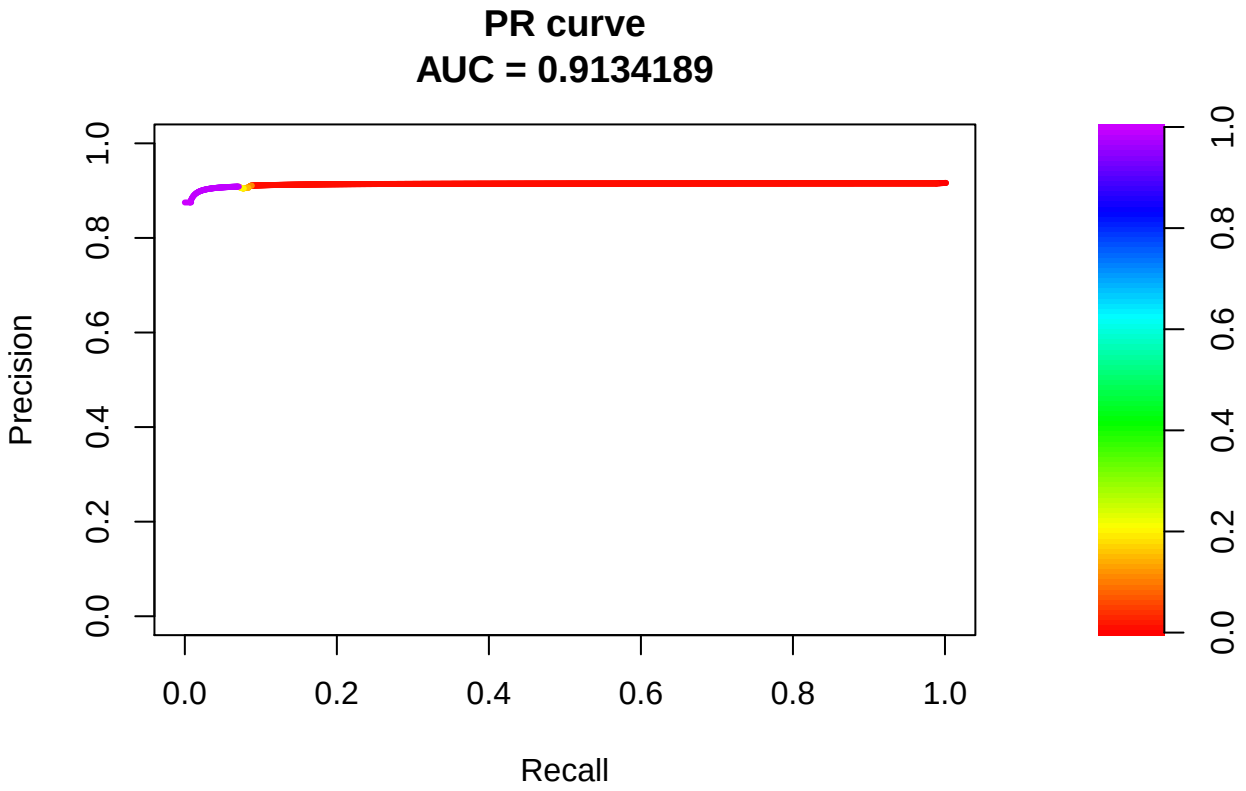
- AUROC: 49.1?

```
## $minsplit
## [1] 9
##
## $minbucket
## [1] 3
##
## $cp
## [1] 0.01444444
##
## $maxcompete
## [1] 6
##
## $usesurrogate
## [1] 0
##
## $maxdepth
## [1] 10
```

```
# ROC Curve
roc <- roc.curve(scores.class0 = fg, scores.class1 = bg, curve = T)
plot(roc)
```




```
# PR Curve
pr <- pr.curve(scores.class0 = fg, scores.class1 = bg, curve = T)
plot(pr)
```



Conclusions

Final model, `rpart` is worse than `ranger` but also achieves very good results and not much different. `rpart` is on average 3 points worse on `AUPRC` and 3 points worse on `AUROC`. However, `rpart` is easily understandable by humans. This difference in this case in negligible.