TomaszBiegusCoffeeR

July 8, 2018

1 Tomasz Biegus Coffee data analysis

In analysis I decided to use great library DALEX which is designed exactly for tasks like this one. Here's the site of project: https://github.com/pbiecek/DALEX DALEX calculate how much we loose in accuracy if certain variable is permuted, which means this variable havent hold any valuable information anymore.

```
In [16]: options(warn=-1)
```

```
In [17]: library(tidyverse)
         library(randomForest)
         library(DALEX)
```

Read the data and deal with na's.

```
In [18]: coffee_data = read.csv("coffee_data.csv")
         set.seed(222)
         coffee_imputed  <- rfImpute(mark ~ ., coffee_data);
         set.seed(333)
```

```
     |      Out-of-bag   |
Tree |      MSE  %Var(y) |
 300 |    1.076    68.26 |
     |      Out-of-bag   |
Tree |      MSE  %Var(y) |
 300 |    1.062    67.38 |
     |      Out-of-bag   |
Tree |      MSE  %Var(y) |
 300 |     1.08    68.49 |
     |      Out-of-bag   |
Tree |      MSE  %Var(y) |
 300 |    1.079    68.42 |
     |      Out-of-bag   |
Tree |      MSE  %Var(y) |
 300 |    1.066    67.64 |
```

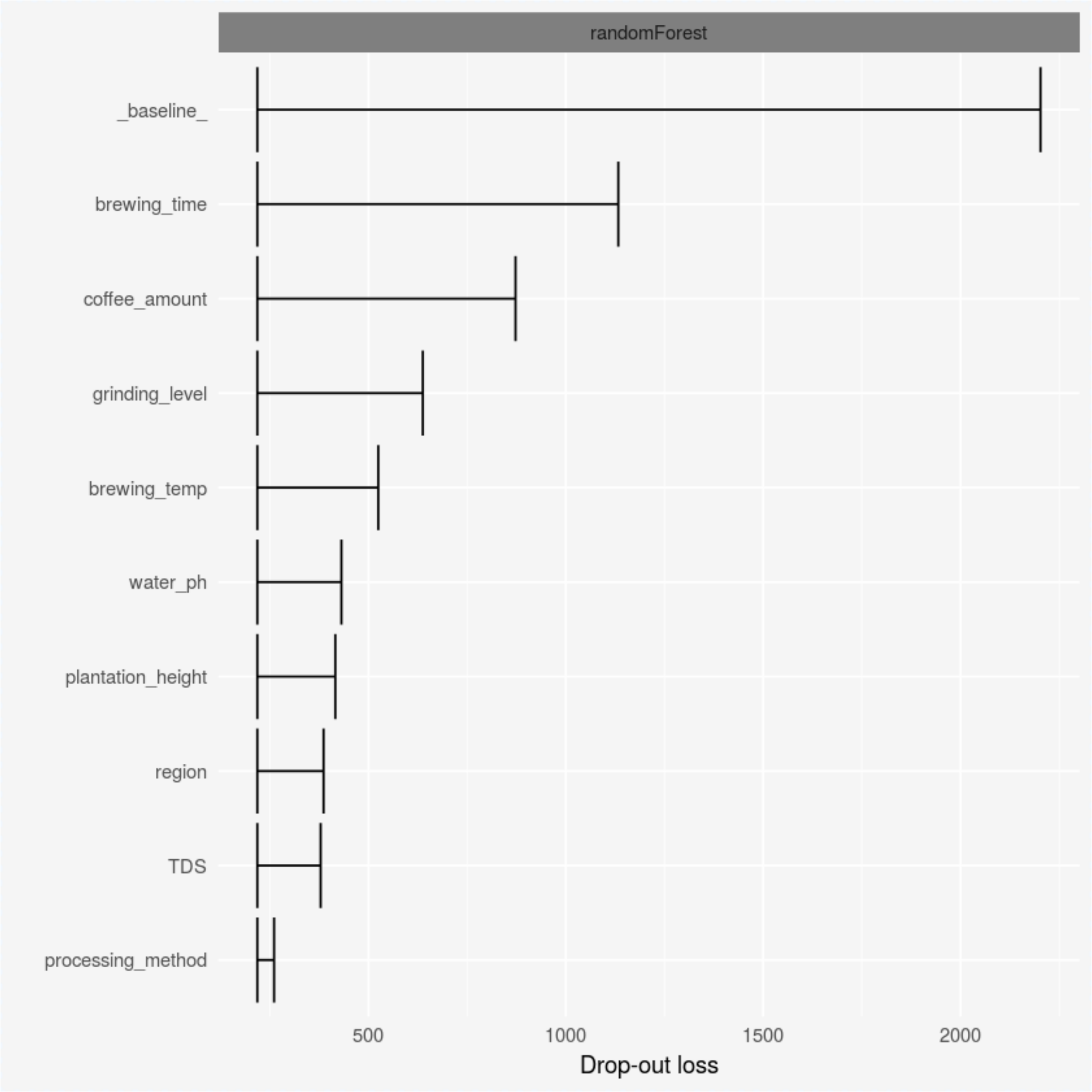
Build random forest model.

```
In [22]: coffee_rf <- randomForest(mark~., data=coffee_imputed, ntree=100)
```

Explain importance of variables using DALEX.

```
In [23]: explainter_rf <- explain(coffee_rf, data = coffee_imputed, y = coffee_impu
         variable_importance_rf <- variable_importance(explainter_rf, type = "raw")
         variable_importance_rf
         plot(variable_importance_rf)
```

variable	dropout_loss	label
_full_model_	218.2816	randomForest
mark	218.2816	randomForest
preinfusion	255.4084	randomForest
processing_method	260.7525	randomForest
TDS	378.7894	randomForest
region	386.1740	randomForest
plantation_height	415.9573	randomForest
water_ph	431.4057	randomForest
brewing_temp	524.7559	randomForest
grinding_level	637.6321	randomForest
coffee_amount	872.5146	randomForest
brewing_time	1132.9263	randomForest
baseline	2202.6463	randomForest



As we can see, the most important variable is brewing_time followed by coffee_amount and grinding_level.