Deep Learning & AutoML in h2o

Set up

```
# Load libraries
library(h2o)
library(tidyverse)
library(wesanderson)
library(knitr)

# Disable progress bar in document
h2o.no_progress()

# Start h2o cluster
h2o.init(nthreads = -1,
         max_mem_size = '4G')
```

```
##
## H2O is not running yet, starting it now...
##
## Note:  In case of errors look at the following log files:
##     C:\Users\User\AppData\Local\Temp\RtmpaSjdSg\file23f42ada253a/h2o_User_started_from_r.out
##     C:\Users\User\AppData\Local\Temp\RtmpaSjdSg\file23f4343678e4/h2o_User_started_from_r.err
##
##
## Starting H2O JVM and connecting:  Connection successful!
##
## R is connected to the H2O cluster:
##     H2O cluster uptime:         2 seconds 939 milliseconds
##     H2O cluster timezone:       Europe/Warsaw
##     H2O data parsing timezone:  UTC
##     H2O cluster version:        3.40.0.1
##     H2O cluster version age:    1 month and 6 days
##     H2O cluster name:           H2O_started_from_R_User_ltv033
##     H2O cluster total nodes:    1
##     H2O cluster total memory:   3.98 GB
```

```
##     H2O cluster total cores:    16
##     H2O cluster allowed cores:  16
##     H2O cluster healthy:        TRUE
##     H2O Connection ip:          localhost
##     H2O Connection port:        54321
##     H2O Connection proxy:       NA
##     H2O Internal Security:      FALSE
##     R Version:                  R version 4.2.2 (2022-10-31 ucrt)
```

```r
# Load file
mushrooms <- read.csv('https://tinyurl.com/hmkhs9au')

# Transform data set for further analysis
mushrooms <- mushrooms %>%
  # Remove not needed characters
  mutate(across(1:23, ~ substr(.x, 3,3))) %>%
  # Change columns from strings to factors
  mutate(across(everything(), as.factor))

# Load data to h2o cluster
mushrooms_hex <- as.h2o(mushrooms, destination_frame = 'mushrooms_hex')

# Split data set to train (75%) & test (25%)
mushrooms_split <- h2o.splitFrame(data = mushrooms_hex, ratios = 0.75)

mushrooms_train <- mushrooms_split[[1]]
mushrooms_test <- mushrooms_split[[2]]
```

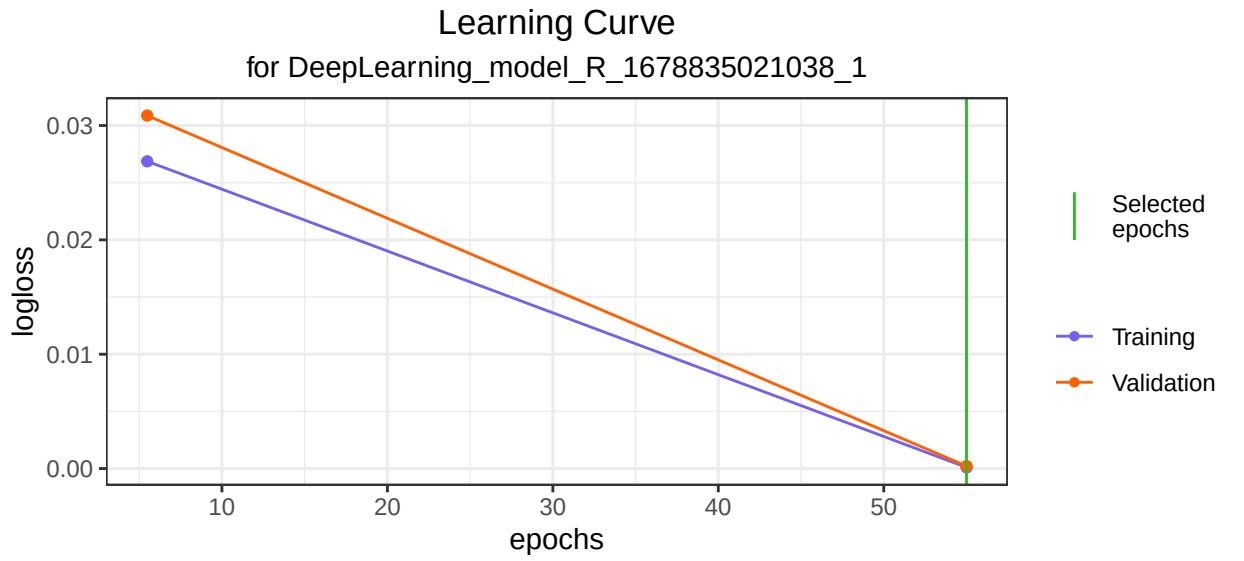
Data

Data set contains 8124 observations of 23 species from Agaricus and Lepiota families. Beside class (e - edible/p - poisonous or uknown edibility) there are 22 physical attributes.

h2o's Deep Learning

```r
# Build and train Deep Learning model
mushrooms_dl <- h2o.deeplearning(
                        y = 23,
                        x = 1:22,
                        training_frame = mushrooms_train,
                        validation_frame = mushrooms_test,
                        distribution = 'multinomial',
                        activation = 'RectifierWithDropout',
                        hidden = c(100, 200, 100),
                        input_dropout_ratio = 0.2,
                        l1 = 1e-5,
                        epochs = 55,
                        variable_importances = TRUE,
                        # Set seed for reproducible results:
                        seed = 123)
```

```
# Learning curve plot
h2o.learning_curve_plot(mushrooms_dl)
```



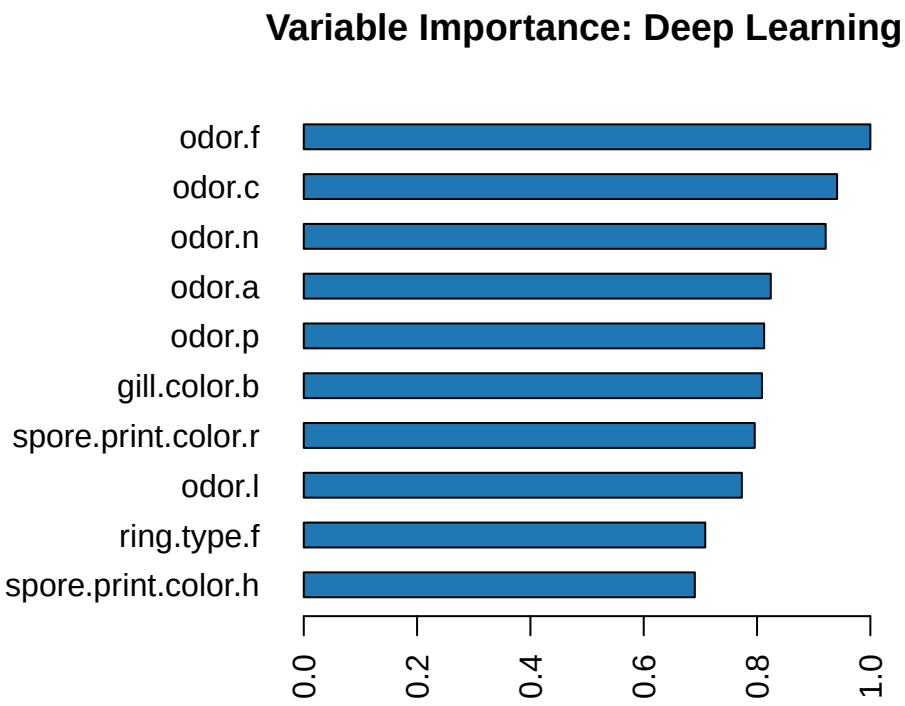

```
# Confusion matrix
h2o.confusionMatrix(mushrooms_dl, mushrooms_test, valid = FALSE, xval = FALSE)
```

```
## Confusion Matrix (vertical: actual; across: predicted)  for max f1 @ threshold = 0.995779564041497:
##            e    p    Error     Rate
## e       1050    0 0.000000  =0/1050
## p          0 1005 0.000000  =0/1005
## Totals  1050 1005 0.000000  =0/2055
```

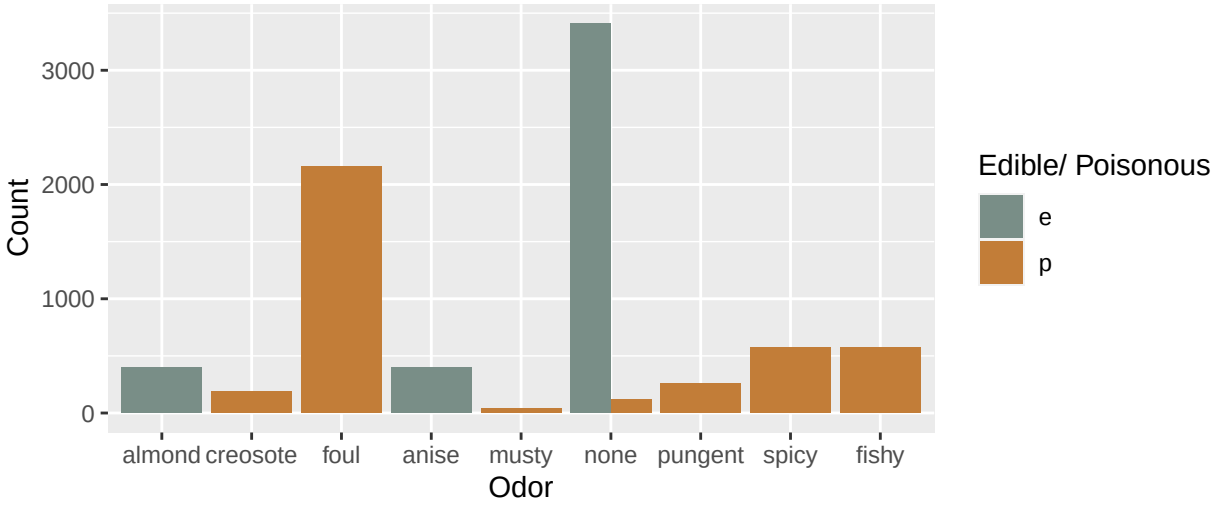
```
# If needed, predictions can be done on test data set
# (which was used for validation in model training)
mushroom_pred <- h2o.predict(object = mushrooms_dl,
                             newdata = mushrooms_test)

# Confusion matrix as simple table: test vs predictions
table(as.data.frame(mushrooms_test[,23])[,1],
      as.data.frame(mushroom_pred[,1])[,1])
```

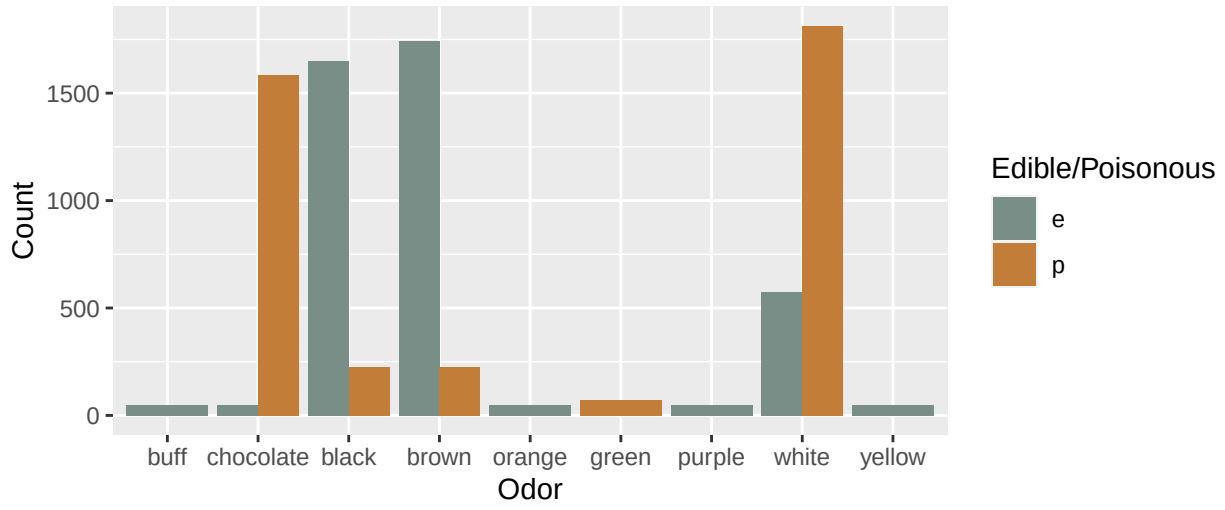
```
##
##        e    p
##   e 1050    0
##   p    0 1005
```

```
# Importance of parameters
h2o.varimp_plot(mushrooms_dl)
```



Most important parameters are odor and spore print color:





AutoML

```
# Run AutoML: max 100 models, max 60 seconds
mushrooms_auoml <- h2o.automl(
                        y = 23,
                        x = 1:22,
                        training_frame = mushrooms_train,
                        max_runtime_secs = 60,
                        max_models = 100)
```

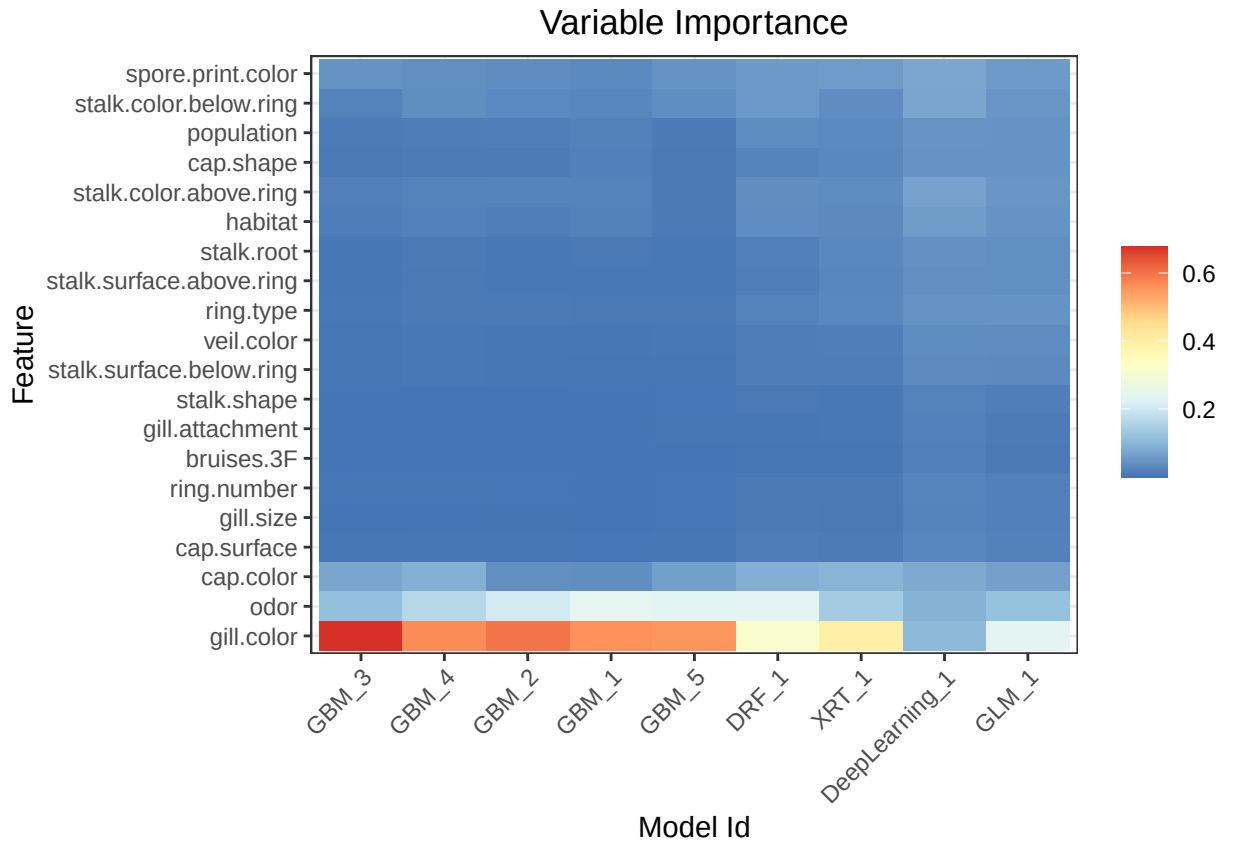
```
##
## 00:04:11.44: AutoML: XGBoost is not available; skipping it.
## 00:04:11.72: _train param, Dropping bad and constant columns: [veil.type]
## 00:04:14.402: _train param, Dropping bad and constant columns: [veil.type]
## 00:04:31.965: _train param, Dropping bad and constant columns: [veil.type]
## 00:04:32.977: _train param, Dropping bad and constant columns: [veil.type]
## 00:04:41.334: _train param, Dropping bad and constant columns: [veil.type]
## 00:04:50.140: _train param, Dropping bad and constant columns: [veil.type]
## 00:05:02.23: _train param, Dropping bad and constant columns: [veil.type]
## 00:05:02.932: _train param, Dropping bad and constant columns: [veil.type]
## 00:05:09.915: _train param, Dropping bad and constant columns: [veil.type]
```

```
# Leader board
df <- h2o.get_leaderboard(object = mushrooms_auoml)
df <- as.data.frame(df)
df$model_id <- substr(df$model_id,1,5)
kable(df)
```

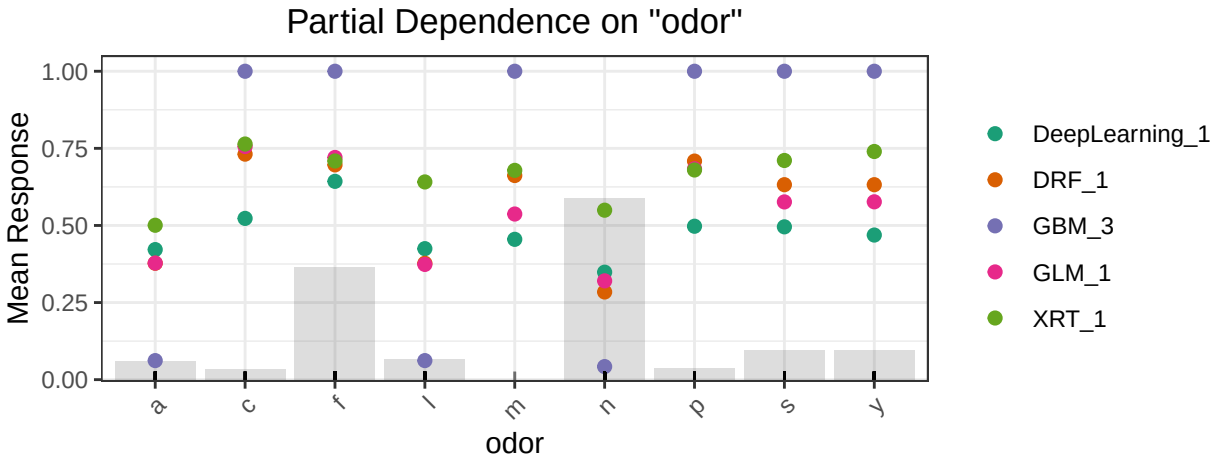
model_id	auc	logloss	aucpr	mean_per_class_error	rmse	mse
DeepL	1.0000000	0.0002825	1.0000000	0.0000000	0.0074360	0.0000553
GBM_3	1.0000000	0.0000000	1.0000000	0.0000000	0.0000000	0.0000000
GBM_4	1.0000000	0.0000000	1.0000000	0.0000000	0.0000000	0.0000000

model_id	auc	logloss	aucpr	mean_per_class_error	rmse	mse
GBM_1	1.0000000	0.0000000	1.0000000	0.0000000	0.0000000	0.0000000
DRF_1	1.0000000	0.0020105	1.0000000	0.0000000	0.0120680	0.0001456
GBM_5	1.0000000	0.0014990	1.0000000	0.0000000	0.0207927	0.0004323
GBM_2	1.0000000	0.0000000	1.0000000	0.0000000	0.0000000	0.0000000
GLM_1	1.0000000	0.0016588	1.0000000	0.0000000	0.0123610	0.0001528
XRT_1	0.9999735	0.2961661	0.9999712	0.0022703	0.2830245	0.0801029

```
# Variables importance heatmap for different AutoML models
h2o.varimp_heatmap(mushrooms_auoml)
```




```
# Effect of odor variable for each model
h2o.pd_multi_plot(mushrooms_auoml, mushrooms_test, "odor")
```



References

1. **Data set**: https://www.kaggle.com/datasets/ulrikthygepedersen/mushroom-attributes
2. **Agaricus family graphic**: https://en.wikipedia.org/wiki/Agaricus
3. **Mushrooms graphics**: <https://biolwww.usask.ca/fungi/glossary.html>
4. **h2o Deep Learning**: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/deep-learning.html
5. **h2o AutoML**: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html
6. **h2o models explainability**: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/explain.html#explanation-plotting-functions