Predictive Analytics For Business with H2O in R

Mahin Anwar

12/1/2020

```r
#Import Libraries
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.2     v dplyr   1.0.2
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

```
## -- Conflicts ------------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(readxl)
library(h2o)
```

```
## Warning: package 'h2o' was built under R version 4.0.3
```

```
##
## ----------------------------------------------------------------------
##
## Your next step is to start H2O:
##     > h2o.init()
##
## For H2O package documentation, ask for help:
##     > ??h2o
##
## After starting H2O, you can use the Web UI at http://localhost:54321
## For more information visit https://docs.h2o.ai
##
## ----------------------------------------------------------------------
```

```
##
## Attaching package: 'h2o'
```

```
## The following objects are masked from 'package:stats':
##
##     cor, sd, var
```

```
## The following objects are masked from 'package:base':
##
##     %*%, %in%, &&, ||, apply, as.factor, as.numeric, colnames,
##     colnames<-, ifelse, is.character, is.factor, is.numeric, log,
##     log10, log1p, log2, round, signif, trunc
```

```r
#Read Excel Sheets
path <- 'UCI_bank_marketing.xlsx'
sheets <- excel_sheets(path)
```

```r
#Explore Data In Each Sheet
sheets %>%
  map(~ read_excel(path = path, sheet = .)) %>%
  set_names(sheets)
```

```
## New names:
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * `` -> ...6
## * ...
```

```
## New names:
## * `` -> ...2
## * `` -> ...4
```

```
## $PROCEDURE
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 14 x 1
##    `BANK MARKETING ANALYSIS PROCEDURE`
##    <chr>
##  1 <NA>
##  2 STEP 1: COLLECT INFORMATION
##  3 1) CLIENT INFORMATION: AGE, JOB, MARITAL STATUS, EDUCATION LEVEL
##  4 2) CLIENT LOAN HISTORY: DEFAULT HISTORY, HOME LOAN, PERSONAL LOAN, CURRENT B~
##  5 3) MARKETING HISTORY: CONTACT TYPE, DAY LAST CONTACT, MONTH LAST CONTACT, LA~
##  6 4) SUBSCRIPTION HISTORY: ENROLLED IN TERM LOAN? (Y/N)
##  7 <NA>
##  8 STEP 2: MERGE INFORMATION
##  9 1) PERFORM VLOOKUP
## 10 <NA>
## 11 STEP 3: MARKETING ANALYSIS
## 12 1) DAILY RANGE: WHAT IS NORMAL HIT RATE?
## 13 2) WHAT FEATURES CONTRIBUTE TO TERM LOAN ENROLLMENT?
## 14 - Job Analysis
##
## $`DATA DESCRIPTION`
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 70 x 1
##    bank_info
##    <chr>
##  1 Citation Request:
##  2 This dataset is public available for research. The details are described in ~
##  3 Please include this citation if you plan to use this database:
##  4 <NA>
##  5 [Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining fo~
##  6 In P. Novais et al. (Eds.), Proceedings of the European Simulation and Model~
##  7 <NA>
##  8 Available at: [pdf] http://hdl.handle.net/1822/14838
##  9 [bib] http://www3.dsi.uminho.pt/pcortez/bib/2011-esm-1.txt
## 10 <NA>
## # ... with 60 more rows
##
## $`Step 1 - Collect Information`
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 1 x 2
##    Step Description
##   <dbl> <chr>
## 1     1 Collect Client Information
##
## $CLIENT_INFO
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 45,211 x 5
##    ID      AGE JOB          MARITAL  EDUCATION
##    <chr> <dbl> <chr>        <chr>    <chr>
##  1 2836     58 management   married  tertiary
##  2 2837     44 technician   single   secondary
##  3 2838     33 entrepreneur married  secondary
##  4 2839     47 blue-collar  married  unknown
##  5 2840     33 unknown      single   unknown
##  6 2841     35 management   married  tertiary
##  7 2842     28 management   single   tertiary
##  8 2843     42 entrepreneur divorced tertiary
##  9 2844     58 retired      married  primary
## 10 2845     43 technician   single   secondary
## # ... with 45,201 more rows
```

```
##
## $LOAN_HISTORY
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 45,211 x 5
##    ID    DEFAULT BALANCE HOUSING LOAN
##    <chr> <chr>     <dbl> <chr>   <chr>
##  1 2836  no         2143 yes     no
##  2 2837  no           29 yes     no
##  3 2838  no            2 yes     yes
##  4 2839  no         1506 yes     no
##  5 2840  no            1 no      no
##  6 2841  no          231 yes     no
##  7 2842  no          447 yes     yes
##  8 2843  yes           2 yes     no
##  9 2844  no          121 yes     no
## 10 2845  no          593 yes     no
## # ... with 45,201 more rows
##
## $`MARKETING HISTORY`
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 45,211 x 9
##    ID    CONTACT   DAY MONTH DURATION CAMPAIGN PDAYS PREVIOUS POUTCOME
##    <chr> <chr>   <dbl> <chr>    <dbl>    <dbl> <dbl>    <dbl> <chr>
##  1 2836  unknown     5 may        261        1    -1        0 unknown
##  2 2837  unknown     5 may        151        1    -1        0 unknown
##  3 2838  unknown     5 may         76        1    -1        0 unknown
##  4 2839  unknown     5 may         92        1    -1        0 unknown
##  5 2840  unknown     5 may        198        1    -1        0 unknown
##  6 2841  unknown     5 may        139        1    -1        0 unknown
##  7 2842  unknown     5 may        217        1    -1        0 unknown
##  8 2843  unknown     5 may        380        1    -1        0 unknown
##  9 2844  unknown     5 may         50        1    -1        0 unknown
## 10 2845  unknown     5 may         55        1    -1        0 unknown
## # ... with 45,201 more rows
##
## $`SUBSCRIPTION HISTORY`
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
```

```
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 45,211 x 2
##    ID    TERM_DEPOSIT
##    <chr> <chr>
##  1 2836  no
##  2 2837  no
##  3 2838  no
##  4 2839  no
##  5 2840  no
##  6 2841  no
##  7 2842  no
##  8 2843  no
##  9 2844  no
## 10 2845  no
## # ... with 45,201 more rows
##
## $`Step 2 - Merge Information`
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 1 x 2
##    Step Description
##   <dbl> <chr>
## 1     2 Perform Data Merge
##
## $CLIENT_MERGE
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 10,006 x 20
##    `VLOOKUP MERGE ~ ...2  ...3  ...4  ...5  ...6  ...7  ...8  ...9  ...10 ...11
##    <chr>            <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
##  1 1. DIFFICULT TO~ <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  2 2. COMPUTATIONA~ <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  3 3. EVERY CELL C~ <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  4 <NA>             <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  5 <NA>             CLIE~ <NA>  <NA>  <NA>  LOAN~ <NA>  <NA>  <NA>  MARK~ <NA>
##  6 <NA>             2.0   3.0   4.0   5.0   2.0   3.0   4.0   5.0   2.0   3.0
##  7 ID               AGE   JOB   MARI~ EDUC~ DEFA~ BALA~ HOUS~ LOAN  CONT~ DAY
##  8 2836             58    mana~ marr~ tert~ no    2143  yes   no    unkn~ 5
##  9 2837             44    tech~ sing~ seco~ no    29    yes   no    unkn~ 5
## 10 2838             33    entr~ marr~ seco~ no    2     yes   yes   unkn~ 5
## # ... with 9,996 more rows, and 9 more variables: ...12 <chr>, ...13 <chr>,
## #   ...14 <chr>, ...15 <chr>, ...16 <chr>, ...17 <chr>, ...18 <chr>,
## #   ...19 <chr>, ...20 <chr>
##
## $`Step 3 - Marketing Analysis`
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 1 x 2
##    Step Description
##   <dbl> <chr>
## 1     3 Perform Marketing Analysis
##
## $`DAILY RANGE`
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 28 x 4
##    `HIT RATE` ...2  `DAILY SUMMARY`     ...4
##         <dbl> <lgl> <chr>              <dbl>
##  1     0.0386 NA    MEAN             0.0351
##  2     0.0360 NA    MEDIAN           0.0362
##  3     0.0551 NA    SD               0.0138
##  4     0.0613 NA    LOWER CONF       0.00755
##  5     0.0427 NA    UPPER CONF       0.0627
##  6     0.0391 NA    <NA>            NA
##  7     0.0451 NA    <NA>            NA
##  8     0.0166 NA    <NA>            NA
##  9     0.0222 NA    <NA>            NA
## 10     0.0179 NA    <NA>            NA
## # ... with 18 more rows
##
## $`JOB ANALYSIS`
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 0 x 0
```

```
##
## $Sheet3
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 0 x 0
```

```r
#Join Data by ID Column (VLOOKUP Equivalent)
data_joined <- sheets[4:7] %>%
  map(~ read_excel(path = path, sheet = .)) %>%
  reduce(left_join)
```

```
## Joining, by = "ID"
## Joining, by = "ID"
## Joining, by = "ID"
```

```r
#Start H2O Cluster
h2o.init()
```

```
##  Connection successful!
##
## R is connected to the H2O cluster:
##     H2O cluster uptime:         1 hours 2 minutes
##     H2O cluster timezone:       Asia/Karachi
##     H2O data parsing timezone:  UTC
##     H2O cluster version:        3.32.0.1
##     H2O cluster version age:    1 month and 22 days
##     H2O cluster name:           H2O_started_from_R_Mahin_bgk343
##     H2O cluster total nodes:    1
##     H2O cluster total memory:   0.79 GB
##     H2O cluster total cores:    4
##     H2O cluster allowed cores:  4
##     H2O cluster healthy:        TRUE
##     H2O Connection ip:          localhost
##     H2O Connection port:        54321
##     H2O Connection proxy:       NA
##     H2O Internal Security:      FALSE
##     H2O API Extensions:         Amazon S3, Algos, AutoML, Core V3, TargetEncoder, Core V4
##     R Version:                  R version 4.0.2 (2020-06-22)
```

```r
#Data Preperation
data_joined <- data_joined %>%
  mutate_if(is.character, as.factor)

train <- as.h2o(data_joined)
```

```
## Warning in use.package("data.table"): data.table cannot be used without R
## package bit64 version 0.9.7 or higher. Please upgrade to take advangage of
## data.table speedups.
```

```
##   |                                                                      |
y <- 'TERM_DEPOSIT'
x <- setdiff(names(train), c(y, 'ID'))
```

```
#H2O AutoML Training
aml <- h2o.automl(
  x = x,
  y = y,
  training_frame = train,
  max_runtime_secs = 600,
  balance_classes = TRUE
)
```

```
##   |                                                                               |
## 18:12:52.845: AutoML: XGBoost is not available; skipping it.  |
```

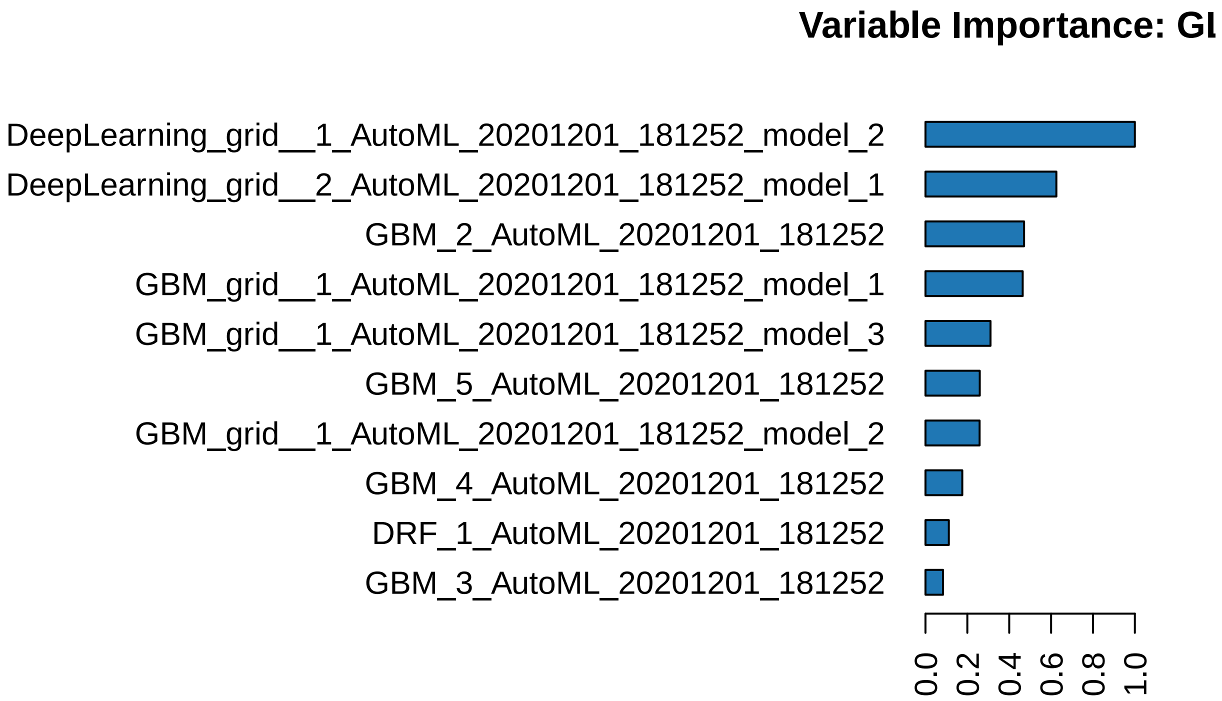
```
#View AutoML Leaderboard
lb <- aml@leaderboard
print(lb, n = nrow(lb))
```

```
##                                                model_id       auc   logloss
## 1  StackedEnsemble_BestOfFamily_AutoML_20201201_181252 0.9303744 0.2242502
## 2           GBM_grid__1_AutoML_20201201_181252_model_2 0.9285158 0.2055691
## 3           GBM_grid__1_AutoML_20201201_181252_model_1 0.9280281 0.2261159
## 4                         GBM_2_AutoML_20201201_181252 0.9279082 0.2283889
## 5     StackedEnsemble_AllModels_AutoML_20201201_181252 0.9259490 0.2110710
## 6                         GBM_3_AutoML_20201201_181252 0.9252512 0.2420418
## 7                         GBM_1_AutoML_20201201_181252 0.9248350 0.2343442
## 8                         GBM_4_AutoML_20201201_181252 0.9234814 0.2485037
## 9                         GBM_5_AutoML_20201201_181252 0.9231539 0.2635884
## 10          GBM_grid__1_AutoML_20201201_181252_model_3 0.9216991 0.2594821
## 11                        GLM_1_AutoML_20201201_181252 0.9066907 0.2397973
## 12                        DRF_1_AutoML_20201201_181252 0.9021722 0.4807522
## 13 DeepLearning_grid__1_AutoML_20201201_181252_model_1 0.8924632 0.2822809
## 14                        XRT_1_AutoML_20201201_181252 0.8923253 0.3645295
## 15 DeepLearning_grid__1_AutoML_20201201_181252_model_2 0.8882014 0.3864282
## 16              DeepLearning_1_AutoML_20201201_181252 0.8613606 0.2978449
## 17 DeepLearning_grid__2_AutoML_20201201_181252_model_1 0.8602353 1.1630130
## 18          GBM_grid__1_AutoML_20201201_181252_model_4 0.7867482 0.3489532
##        aucpr mean_per_class_error      rmse        mse
## 1  0.6174270            0.1711389 0.2581269 0.06662950
## 2  0.6044228            0.1848626 0.2538347 0.06443204
## 3  0.6039276            0.1841835 0.2648567 0.07014906
## 4  0.5969950            0.1746795 0.2679831 0.07181495
## 5  0.6199293            0.1860954 0.2531437 0.06408175
## 6  0.5892246            0.1685691 0.2751572 0.07571148
## 7  0.5854732            0.1808183 0.2708603 0.07336528
## 8  0.5894715            0.1750221 0.2793287 0.07802454
## 9  0.5721913            0.1584124 0.2858204 0.08169331
## 10 0.5844173            0.1850892 0.2845901 0.08099152
## 11 0.5507272            0.2075865 0.2667996 0.07118201
## 12 0.5496305            0.1972503 0.2845060 0.08094365
## 13 0.5010746            0.2083141 0.2785832 0.07760861
## 14 0.5298567            0.1937692 0.2853172 0.08140591
## 15 0.4928066            0.2376992 0.3432755 0.11783805
## 16 0.4641415            0.2607457 0.2838183 0.08055282
## 17 0.4661204            0.2489415 0.5639029 0.31798651
## 18 0.4205574            0.2949039 0.3174832 0.10079556
##
## [18 rows x 7 columns]
```

```
#Ensemble Exploration
model_ids <- as.data.frame(aml@leaderboard$model_id)[,1]
se <- h2o.getModel(grep('StackedEnsemble_AllModels', model_ids, value = TRUE)[1])
metalearner <- h2o.getModel(se@model$metalearner$name)
h2o.varimp(metalearner)
```

```
##                                                variable relative_importance
## 1  DeepLearning_grid__1_AutoML_20201201_181252_model_2         0.502362923
## 2  DeepLearning_grid__2_AutoML_20201201_181252_model_1         0.314089413
## 3                         GBM_2_AutoML_20201201_181252         0.236596984
## 4           GBM_grid__1_AutoML_20201201_181252_model_1         0.233236557
## 5           GBM_grid__1_AutoML_20201201_181252_model_3         0.156026361
## 6                         GBM_5_AutoML_20201201_181252         0.130276423
## 7           GBM_grid__1_AutoML_20201201_181252_model_2         0.129843594
## 8                         GBM_4_AutoML_20201201_181252         0.088326214
## 9                         DRF_1_AutoML_20201201_181252         0.056021856
## 10                        GBM_3_AutoML_20201201_181252         0.042177510
## 11                        GBM_1_AutoML_20201201_181252         0.041992347
## 12                        XRT_1_AutoML_20201201_181252         0.009623663
## 13                        GLM_1_AutoML_20201201_181252         0.000000000
## 14 DeepLearning_grid__1_AutoML_20201201_181252_model_1         0.000000000
## 15               DeepLearning_1_AutoML_20201201_181252         0.000000000
## 16          GBM_grid__1_AutoML_20201201_181252_model_4         0.000000000
##    scaled_importance  percentage
## 1         1.00000000 0.258873386
## 2         0.62522411 0.161853884
## 3         0.47096825 0.121921144
## 4         0.46427900 0.120189478
## 5         0.31058495 0.080402177
## 6         0.25932730 0.067132938
## 7         0.25846572 0.066909896
## 8         0.17582152 0.045515513
## 9         0.11151670 0.028868706
## 10        0.08395825 0.021734556
## 11        0.08358966 0.021639139
## 12        0.01915679 0.004959184
## 13        0.00000000 0.000000000
## 14        0.00000000 0.000000000
## 15        0.00000000 0.000000000
## 16        0.00000000 0.000000000
```

```
h2o.varimp_plot(metalearner)
```

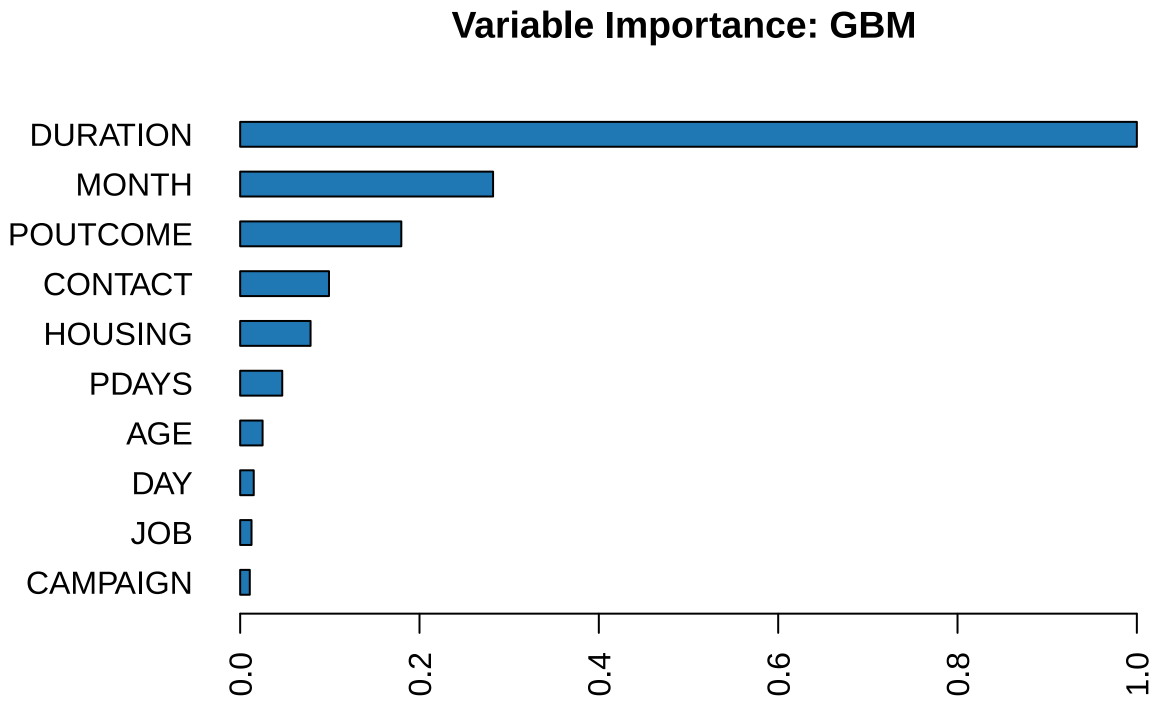


```r
#Baselearner Variable Importance

gb <- h2o.getModel(grep('GBM', model_ids, value = TRUE)[1])
h2o.varimp(gb)
```

```
## Variable Importances:
##     variable relative_importance scaled_importance percentage
## 1   DURATION        28162.781250          1.000000   0.561947
## 2      MONTH         7942.854492          0.282034   0.158488
## 3   POUTCOME         5060.835938          0.179699   0.100981
## 4    CONTACT         2789.882324          0.099063   0.055668
## 5    HOUSING         2210.745850          0.078499   0.044112
## 6      PDAYS         1321.715698          0.046931   0.026373
## 7        AGE          704.127747          0.025002   0.014050
## 8        DAY          425.929138          0.015124   0.008499
## 9        JOB          356.550690          0.012660   0.007114
## 10  CAMPAIGN          302.324585          0.010735   0.006032
## 11   BALANCE          283.221436          0.010057   0.005651
## 12      LOAN          189.409912          0.006726   0.003779
## 13  PREVIOUS          135.869446          0.004824   0.002711
## 14   MARITAL          112.696335          0.004002   0.002249
## 15 EDUCATION          108.803421          0.003863   0.002171
## 16   DEFAULT            8.721868          0.000310   0.000174
```

```r
h2o.varimp_plot(gb)
```


Variable Importance: GBM
DURATION
MONTH
POUTCOME
CONTACT
HOUSING
PDAYS
AGE
DAY
JOB
CAMPAIGN
0.0
0.2
0.4
0.6
0.8
1.0