Predicting Motorcycle Accident Severity with GBM in H2O

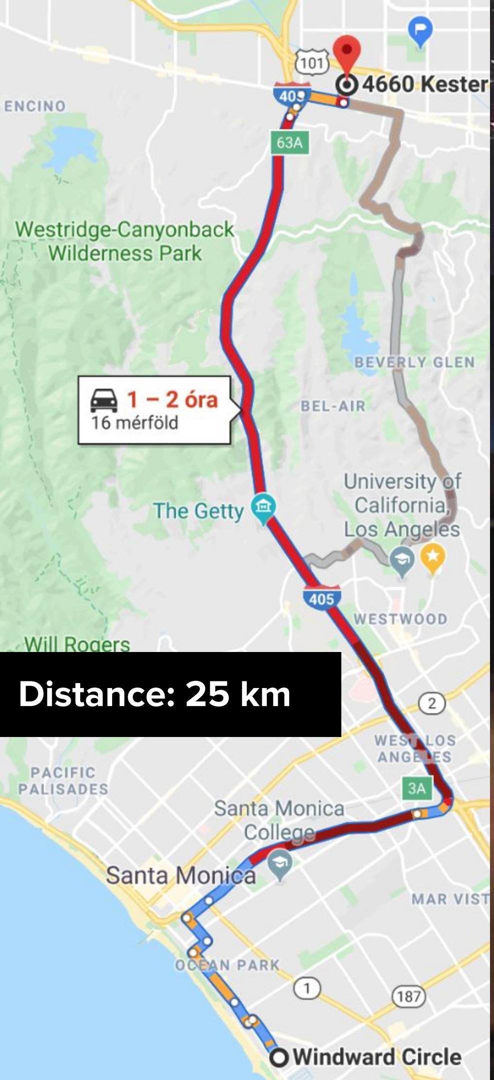
Alex Trickey -- Budapest BI Fórum -- 2019

Motivations

Explore H2O features on an Open Dataset

Understand to what extent dangerous circumstances can be forecasted (and therefore avoided).

Make my commute a little bit safer.



Distance: 25 km

390.000 vehicles/day

Commute Time:

Car: 45 mins - 2.5 hours

Motorcycle: 30 - 45 mins

Savings: ~2 hours per day

Times Hit by Car: 2

The Data

Open Data Source:

- Statewide Integrated Traffic Records System (SWITRS)
- 10.533 collisions involving motorcycles (2012-2017)

Outcome:

- Accident Severity (0: minor injuries only 1: If hospitalization or fatality)
- Unbalanced: Only 1.533 rated severe

Features: Date/time, other vehicles, weather, traffic violations, etc.

Procedure

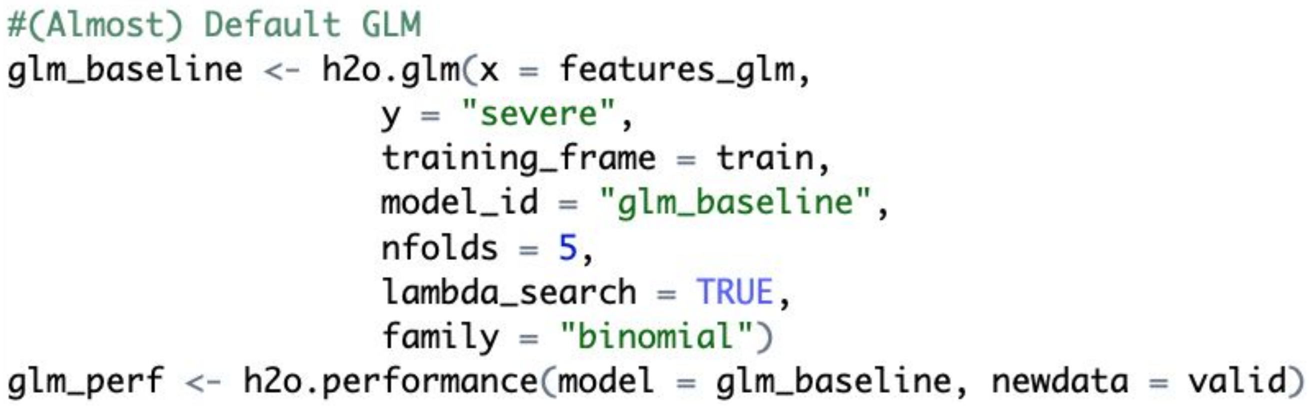
Clean / Transform Data

Split into training, testing, and validation sets

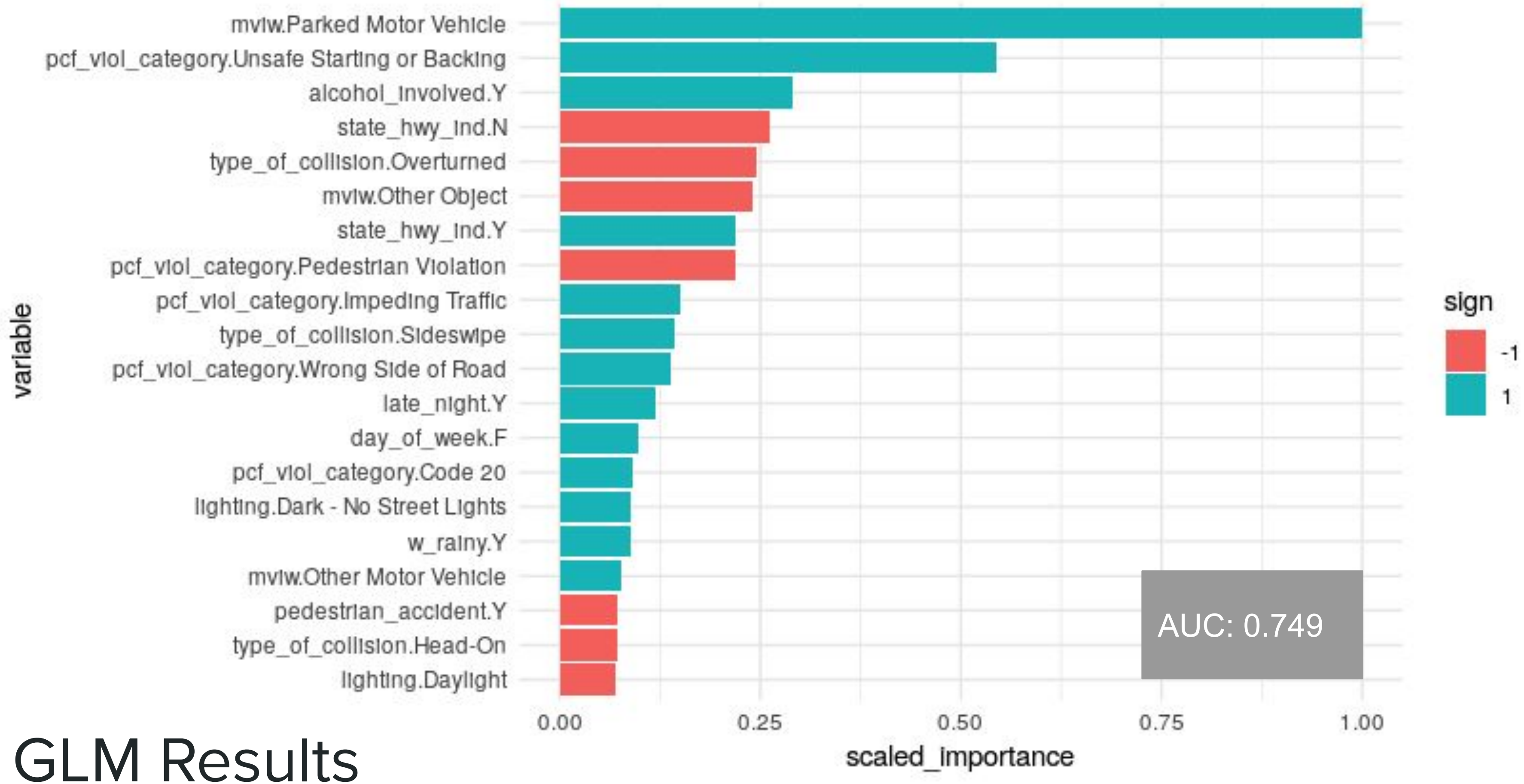
Explore data and set a baseline (GLM, visualizations)

Fit a GBM and dissect results

Fitting a GLM in H2O

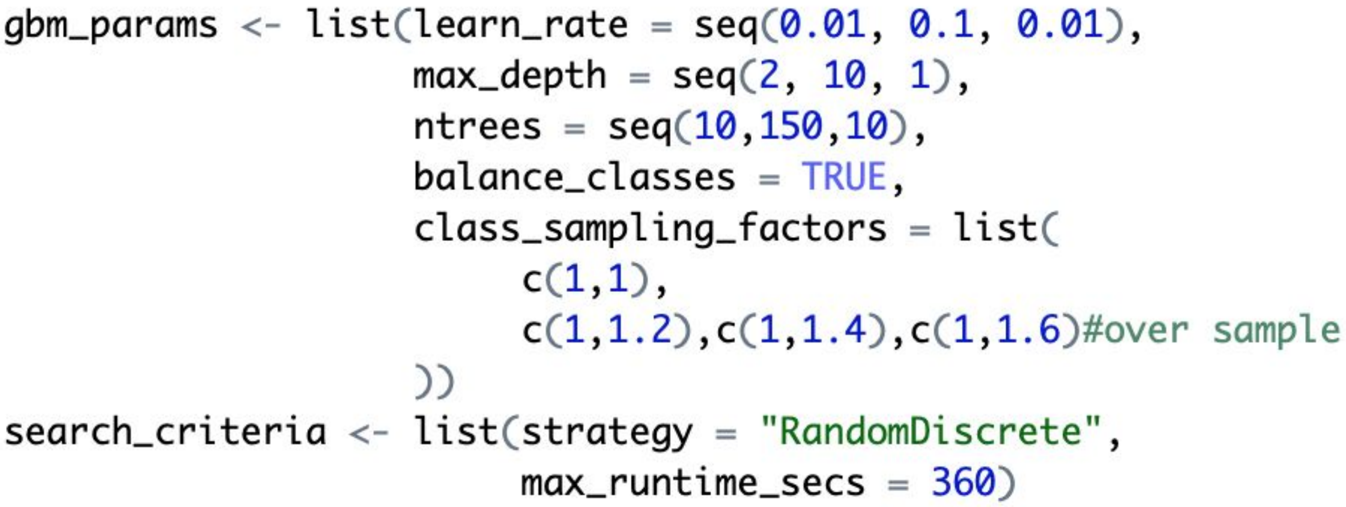


```
#(Almost) Default GLM
glm_baseline <- h2o.glm(x = features_glm,
                        y = "severe",
                        training_frame = train,
                        model_id = "glm_baseline",
                        nfolds = 5,
                        lambda_search = TRUE,
                        family = "binomial")
glm_perf <- h2o.performance(model = glm_baseline, newdata = valid)
```


mviw.Parked Motor Vehicle
pcf_viol_category.Unsafe Starting or Backing
alcohol_involved.Y
state_hwy_ind.N
type_of_collision.Overturned
mviw.Other Object
state_hwy_ind.Y
pcf_viol_category.Pedestrian Violation
pcf_viol_category.Impeding Traffic
type_of_collision.Sideswipe
pcf_viol_category.Wrong Side of Road
late_night.Y
day_of_week.F
pcf_viol_category.Code 20
lighting.Dark - No Street Lights
w_rainy.Y
mviw.Other Motor Vehicle
pedestrian_accident.Y
type_of_collision.Head-On
lighting.Daylight
variable
0.00
0.25
0.50
0.75
1.00
scaled_importance
sign
-1
1
AUC: 0.749

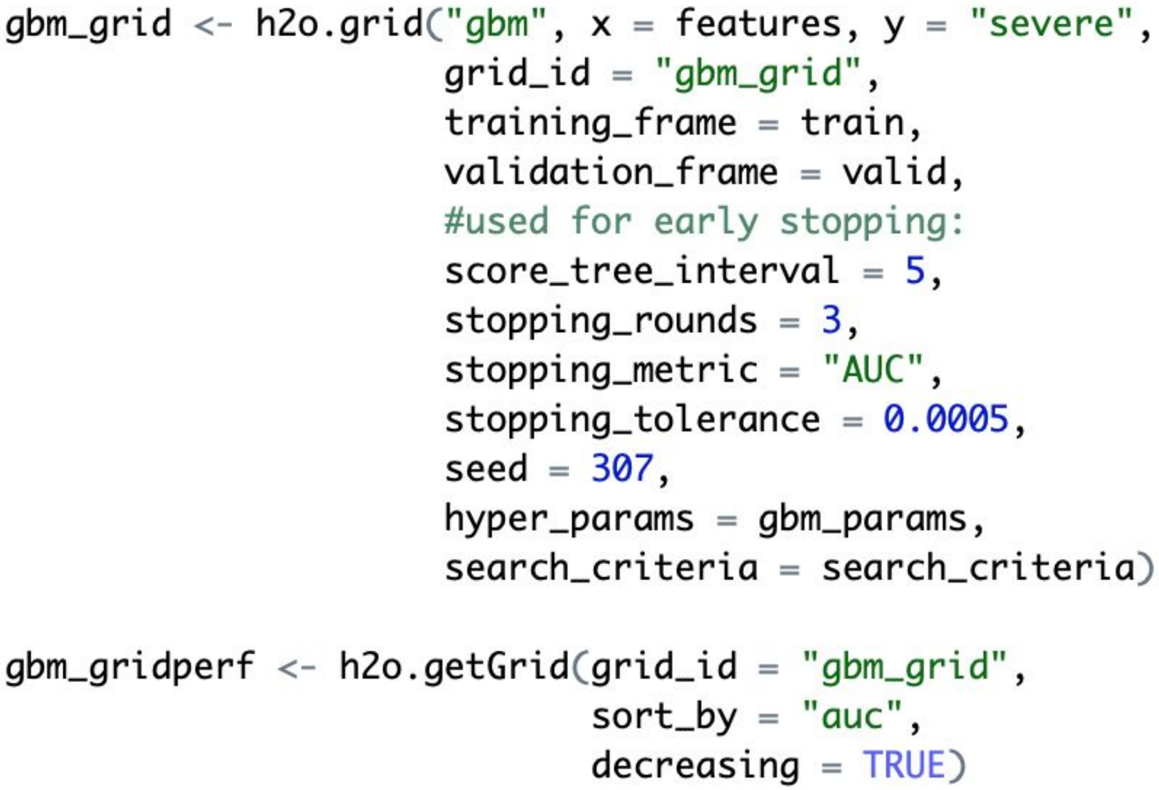
GLM Results

GBM Grid Search in H2O



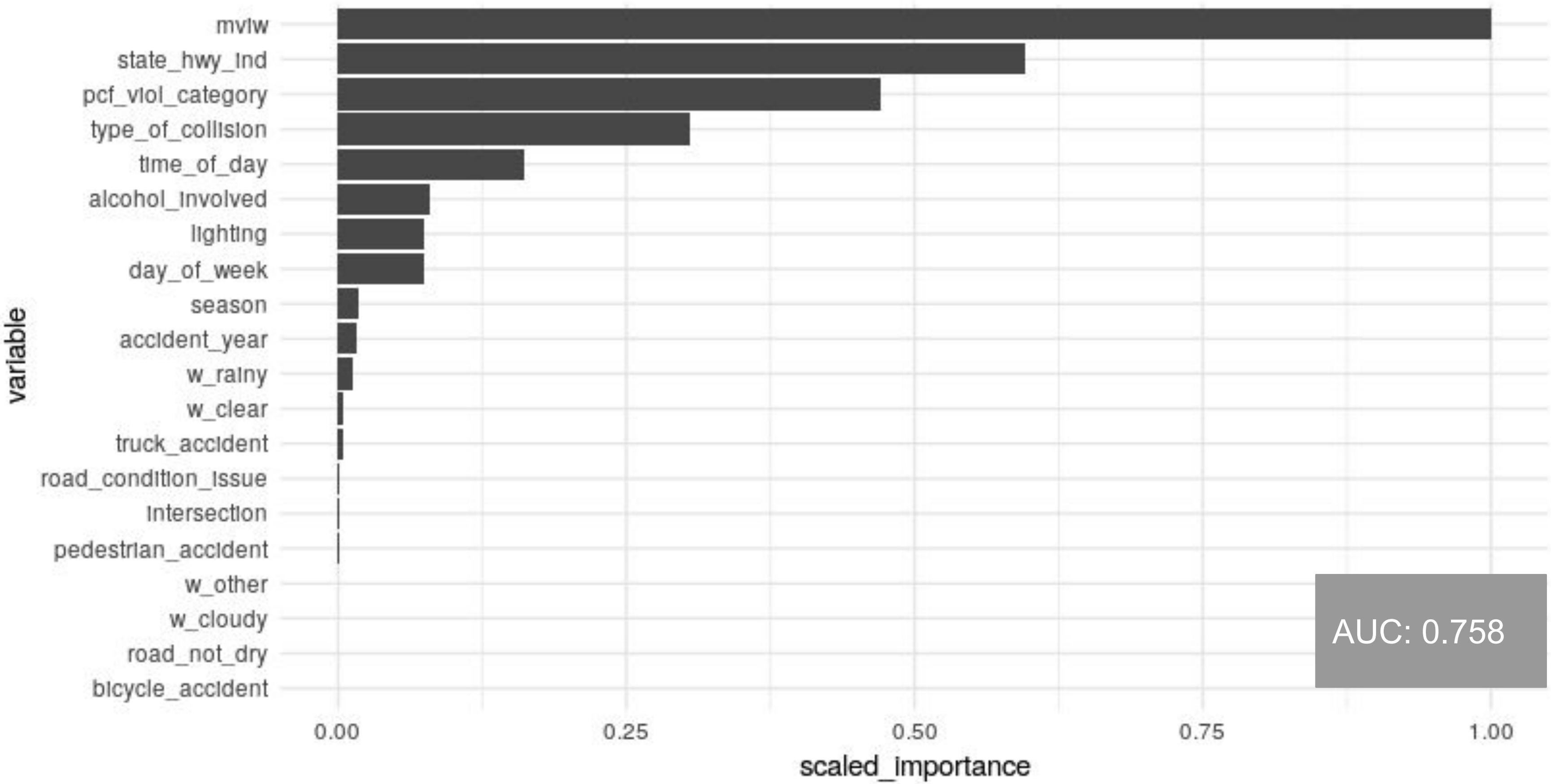
```
gbm_params <- list(learn_rate = seq(0.01, 0.1, 0.01),
                   max_depth = seq(2, 10, 1),
                   ntrees = seq(10,150,10),
                   balance_classes = TRUE,
                   class_sampling_factors = list(
                        c(1,1),
                        c(1,1.2),c(1,1.4),c(1,1.6)#over sample
                   ))
search_criteria <- list(strategy = "RandomDiscrete",
                        max_runtime_secs = 360)
```

GBM Grid Search in H2O



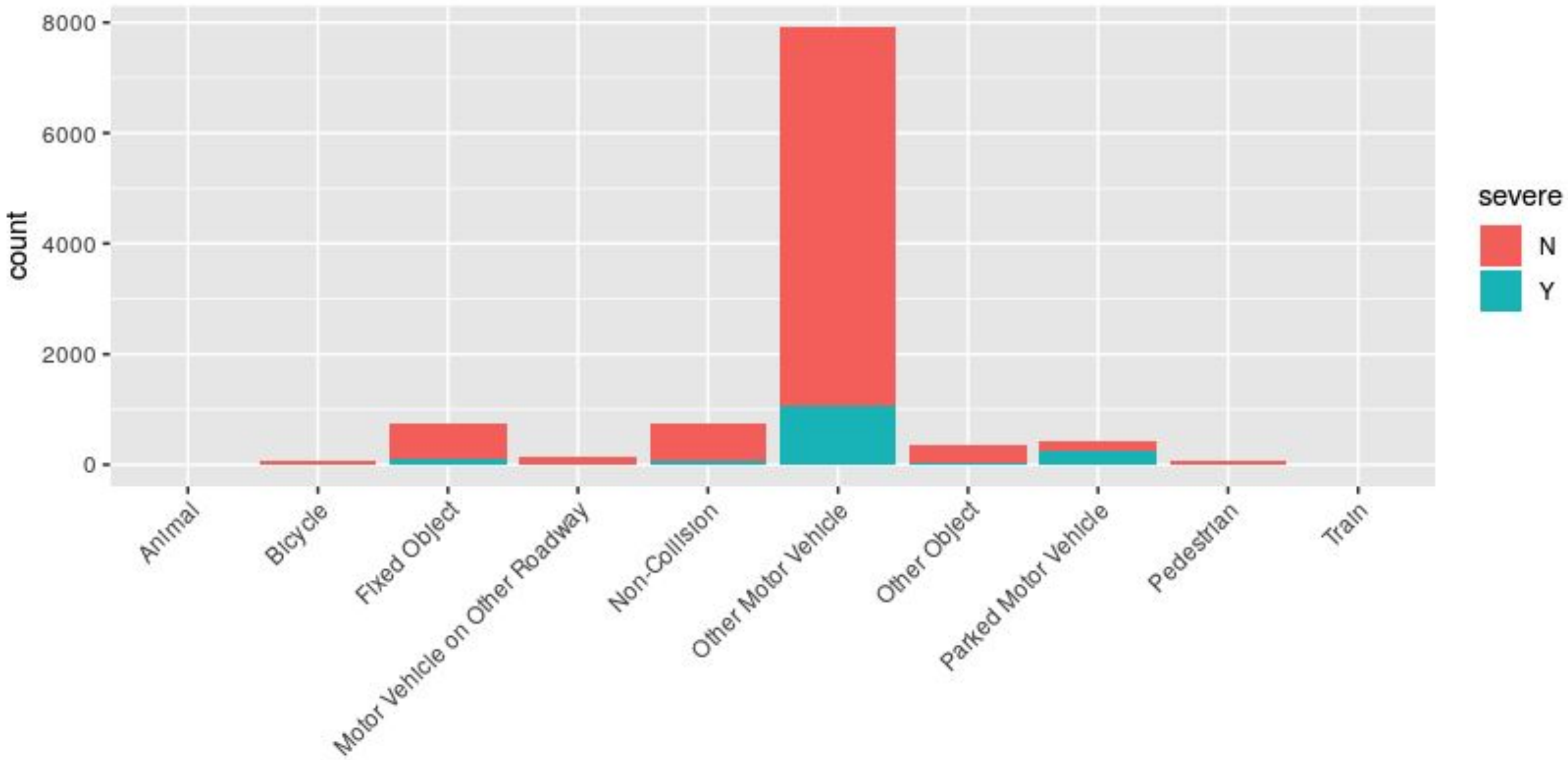
```r
gbm_grid <- h2o.grid("gbm", x = features, y = "severe",
                     grid_id = "gbm_grid",
                     training_frame = train,
                     validation_frame = valid,
                     #used for early stopping:
                     score_tree_interval = 5,
                     stopping_rounds = 3,
                     stopping_metric = "AUC",
                     stopping_tolerance = 0.0005,
                     seed = 307,
                     hyper_params = gbm_params,
                     search_criteria = search_criteria)

gbm_gridperf <- h2o.getGrid(grid_id = "gbm_grid",
                            sort_by = "auc",
                            decreasing = TRUE)
```

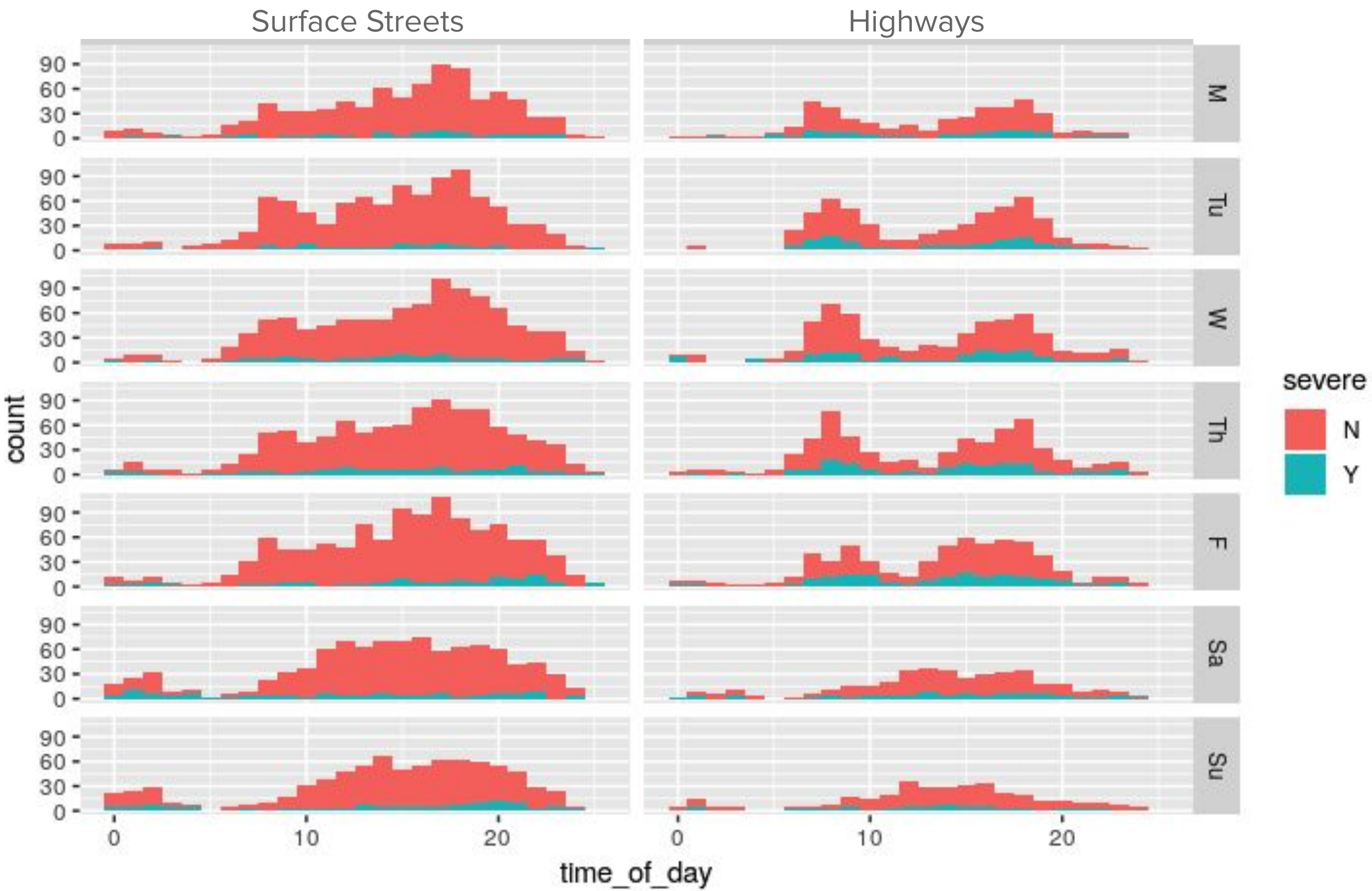

mviw
state_hwy_ind
pcf_viol_category
type_of_collision
time_of_day
alcohol_involved
lighting
day_of_week
season
accident_year
w_rainy
w_clear
truck_accident
road_condition_issue
intersection
pedestrian_accident
w_other
w_cloudy
road_not_dry
bicycle_accident
variable
0.00
0.25
0.50
0.75
1.00
scaled_importance
AUC: 0.758

GBM Results

Following Up - Objects we shouldn't drive into...



Following Up - Date / Time Patterns



Thank You!