Uber Hackaton

Nick McCubbin

2023-07-27

```
library(caret)
library(lubridate)
library(gbm)
```

Data dictionary

- `pickup_dt`: Time period of the observations.
- `borough`: NYC's borough.
- `pickups`: Number of pickups for the period.
- `spd`: Wind speed in miles/hour.
- `vsb`: Visibility in Miles to nearest tenth.
- `temp`: temperature in Fahrenheit.
- `dewp`: Dew point in Fahrenheit.
- `slp`: Sea level pressure.
- `pcp01`: 1-hour liquid precipitation.
- `pcp06`: 6-hour liquid precipitation.
- `pcp24`: 24-hour liquid precipitation.
- `sd`: Snow depth in inches.
- `hday`: Being a holiday (Y) or not (N).

```
# Load in data
load("UBERHACKATHON.RData")

# Convert pickup time to date object
TRAIN$pickup_dt <- ymd_hms(TRAIN$pickup_dt)
KAGGLE$pickup_dt <- ymd_hms(KAGGLE$pickup_dt)

# Extract out potentially useful date information
# hour of day
TRAIN$hour <- factor(hour(TRAIN$pickup_dt))
KAGGLE$hour <- factor(hour(KAGGLE$pickup_dt))

# month
TRAIN$month <- month(TRAIN$pickup_dt, label = TRUE)
KAGGLE$month <- month(KAGGLE$pickup_dt, label = TRUE)

# day of week
TRAIN$wday <- wday(TRAIN$pickup_dt, label = TRUE)
KAGGLE$wday <- wday(KAGGLE$pickup_dt, label = TRUE)
```

```
# Drop date column, no longer useful
TRAIN$pickup_dt <- NULL
KAGGLE$pickup_dt <- NULL

# Remove ID column from training data
TRAIN$IDno <- NULL
```

- Build predictive models to predict `logpickups`
- Create predictions

```
dim(TRAIN)
```

```
## [1] 10000    15
```

```
head(TRAIN, 3)
```

```
##   logpickups       borough spd vsb temp dewp    slp pcp01 pcp06 pcp24 sd
hday
## 1   2.437751        Queens   3  10 51.0   31 1005.0     0 0.000 0.905  0
N
## 2   3.049218     Manhattan   5  10 41.0   25 1008.8     0 0.000 0.000  0
N
## 3   0.698970 Staten Island   0  10 64.5   55 1014.0     0 0.055 0.080  0
N
##   hour month wday
## 1   13   Apr  Mon
## 2    4   Mar  Sat
## 3    0   Jun  Sat
```

```
fit <- gbm(
  logpickups ~ .,
  data = TRAIN,
  shrinkage = 0.01,
  interaction.depth = 20,
  n.minobsinnode = 3,
  n.trees = 3750)
```

```
## Distribution not specified, assuming gaussian ...
```

```
y_pred <- predict(fit, newdata = KAGGLE)
```

```
## Using 3750 trees...
```

```
sd(y_pred)
```

```
## [1] 1.216969
```

```
predictions <- data.frame(
  IDno = KAGGLE$IDno,
  logpickups = y_pred
)
```

```r
write.csv(
  predictions,
  file = "GBM40.csv",
  row.names = FALSE
)
```