Appendix Part II: Data Processing

Introduction

In this section, we will try to process the data for further work in predictive modeling.

Data processing

First we load the downsampled data as is mentioned in Appendix I.

```
# drop rows with price being NA
rideshare = read.csv("data/sampled_rideshare.csv")
rideshare = rideshare[is.na(rideshare$price) == F & is.na(rideshare$cab_type) == F, ]

# is there any NA values after we dropped the NAs in price and cab_type?
sum(is.na(rideshare))
```

```
## [1] 0
```

```
# sample the data to take a peek at the predictors
head(rideshare)
```

```
##        X                                    id  timestamp hour day month
## 1  94613 abb1cae0-8ce3-41ed-8ee3-d83fa972eccb 1543647484    6   1    12
## 2 665025 49ddacaf-a32d-44c2-9376-1bbfa5b6427c 1543444688   22  28    11
## 3  99257 6e2b6b90-bfff-42a3-9ec6-51aaebe4adc7 1543486083   10  29    11
## 4 340794 c50c80d8-9e61-4278-9fe5-9db9b125dce8 1544989213   19  16    12
## 5 552051 8997949b-4983-470f-8937-de58c07131ba 1543684681   17   1    12
## 6 490295 05cc2fab-3b8f-4acc-9311-60e9a307d075 1543425068   17  28    11
##              datetime         timezone             source        destination
## 1 2018-12-01 06:58:03 America/New_York  Boston University Financial District
## 2 2018-11-28 22:38:07 America/New_York      South Station   Theatre District
## 3 2018-11-29 10:08:03 America/New_York   Theatre District   Haymarket Square
## 4 2018-12-16 19:40:13 America/New_York      North Station      South Station
## 5 2018-12-01 17:18:00 America/New_York      North Station   Haymarket Square
## 6 2018-11-28 17:11:07 America/New_York Financial District             Fenway
##   cab_type                            product_id         name price distance
## 1     Lyft                         lyft_luxsuv Lux Black XL  45.5     4.36
## 2     Uber 997acbb5-e102-41e1-b155-9df7de0a73f2     UberPool   8.0     1.30
## 3     Lyft                           lyft_line       Shared   9.0     1.72
## 4     Lyft                            lyft_lux    Lux Black  19.5     1.77
## 5     Lyft                        lyft_premier          Lux  10.5     0.77
## 6     Lyft                         lyft_luxsuv Lux Black XL  38.0     3.68
##   surge_multiplier latitude longitude temperature apparentTemperature
```

```
## 1               1  42.3647  -71.0542       34.59              32.16
## 2               1  42.3647  -71.0542       40.43              34.59
## 3               1  42.3519  -71.0643       37.92              32.00
## 4               1  42.3644  -71.0661       43.06              38.26
## 5               1  42.3505  -71.1054       41.89              41.89
## 6               1  42.3644  -71.0661       40.77              35.14
##     short_summary                              long_summary precipIntensity
## 1       Overcast   Light rain in the morning and overnight.               0
## 2       Overcast          Mostly cloudy throughout the day.               0
## 3  Partly Cloudy          Partly cloudy throughout the day.               0
## 4       Overcast                   Rain throughout the day.               0
## 5  Partly Cloudy   Light rain in the morning and overnight.               0
## 6       Overcast          Mostly cloudy throughout the day.               0
##   precipProbability humidity windSpeed windGust windGustTime visibility
## 1                 0     0.78      3.07     3.07   1543672800      9.945
## 2                 0     0.64      9.08    12.72   1543431600     10.000
## 3                 0     0.67      8.11    12.38   1543514400      9.995
## 4                 0     0.71      8.12    13.44   1545015600     10.000
## 5                 0     0.57      2.51     4.03   1543672800      9.953
## 6                 0     0.63      8.76    14.90   1543431600     10.000
##   temperatureHigh temperatureHighTime temperatureLow temperatureLowTime
## 1           44.66          1543690800          35.04         1543712400
## 2           42.61          1543438800          37.60         1543489200
## 3           44.80          1543510800          28.70         1543579200
## 4           43.74          1544990400          34.07         1545044400
## 5           44.54          1543690800          34.74         1543712400
## 6           42.57          1543438800          37.37         1543489200
##   apparentTemperatureHigh apparentTemperatureHighTime apparentTemperatureLow
## 1                   43.99                  1543690800                  35.69
## 2                   36.57                  1543438800                  32.12
## 3                   38.51                  1543510800                  26.30
## 4                   38.36                  1544986800                  28.17
## 5                   43.87                  1543690800                  35.39
## 6                   36.55                  1543438800                  31.91
##   apparentTemperatureLowTime                 icon dewPoint pressure
## 1                 1543712400               cloudy     28.47  1019.21
## 2                 1543478400               cloudy     29.27   994.99
## 3                 1543575600  partly-cloudy-night     27.86  1003.62
## 4                 1545044400               cloudy     34.25  1015.00
## 5                 1543712400    partly-cloudy-day     27.66  1022.54
## 6                 1543478400               cloudy     29.17   991.33
##   windBearing cloudCover uvIndex visibility.1 ozone sunriseTime sunsetTime
## 1         296       0.94       0        9.945 287.3  1543665331 1543698851
## 2         295       1.00       0       10.000 354.8  1543405936 1543439716
## 3         306       0.22       0        9.995 341.1  1543492402 1543526097
## 4          71       1.00       0       10.000 322.7  1544962122 1544994842
## 5         325       0.34       2        9.953 275.8  1543665341 1543698866
## 6         303       1.00       1       10.000 352.4  1543405938 1543439719
##   moonPhase precipIntensityMax uvIndexTime temperatureMin temperatureMinTime
## 1      0.82             0.0000  1543683600          31.71         1543658400
## 2      0.72             0.0000  1543420800          33.85         1543399200
## 3      0.75             0.0000  1543510800          35.02         1543550400
## 4      0.30             0.1246  1544979600          38.88         1544954400
## 5      0.82             0.0000  1543683600          31.31         1543662000
```

```
## 6      0.72             0.0000  1543420800          33.70         1543399200
##   temperatureMax temperatureMaxTime apparentTemperatureMin
## 1          44.66         1543690800                  28.06
## 2          42.61         1543438800                  30.03
## 3          44.80         1543510800                  30.81
## 4          43.74         1544990400                  33.68
## 5          44.54         1543690800                  28.10
## 6          42.57         1543438800                  29.94
##   apparentTemperatureMinTime apparentTemperatureMax apparentTemperatureMaxTime
## 1                 1543658400                  43.99                 1543690800
## 2                 1543399200                  36.57                 1543438800
## 3                 1543550400                  38.51                 1543510800
## 4                 1545019200                  38.36                 1544986800
## 5                 1543662000                  43.87                 1543690800
## 6                 1543399200                  36.55                 1543438800
```

```
# number of records
nrow(rideshare)
```

```
## [1] 45984
```

Next we process some of the special predictors

```
nsamples = nrow(rideshare)

# day
rideshare$daycount = rideshare$day
for(i in 1 : nsamples) {
  if(rideshare[i, "month"] == "12"){
    rideshare[i, "daycount"] = rideshare[i, "daycount"] + 30
  } else {
    rideshare[i, "daycount"] = rideshare[i, "day"]
  }
}

# hour -> time periods
rideshare$period = rep(NA, nsamples)
for (i in 1 : nsamples){
  if ((rideshare$hour[i] > 6) && (rideshare$hour[i] <= 12)) {
    rideshare$period[i] = "morning"
  }
  if ((rideshare$hour[i] > 12) && (rideshare$hour[i] <= 18)) {
    rideshare$period[i] = "afternoon"
  }
  if ((rideshare$hour[i] > 18) && (rideshare$hour[i] <= 23)) {
    rideshare$period[i] = "evening"
  }
  if ((rideshare$hour[i] > 23) || (rideshare$hour[i] <= 6)) {
    rideshare$period[i] = "midnight"
  }
}

# check again for NA values
sum(is.na(rideshare))
```

```
## [1] 0
```

```
# combine some of the descriptions
rideshare$weather = rideshare$short_summary
for(i in 1 : nsamples) {
  if(rideshare$short_summary[i] %in% c(" Overcast ", " Partly Cloudy ", " Mostly Cloudy ")){
    rideshare$weather[i] = " Cloudy "
  } else if (rideshare$short_summary[i] %in% c(" Rain ", " Light Rain ", " Possible Drizzle ", " Drizzl
    rideshare$weather[i] = " Rain "
  } else {
    rideshare$weather[i] = rideshare$short_summary[i]
  }
}
```

Then select needed predictors, perform train / test split, and save the processed data to result

```
# select columns
rideshare = rideshare[c(
  "distance", "cab_type",
  "daycount", "weather",
  "source", "destination", "name",
  "temperature", "humidity", "windSpeed", "visibility", "pressure",
  "period",
  "price"
)]

# take a peek
head(rideshare, 5)
```

```
##   distance cab_type daycount  weather            source        destination
## 1     4.36     Lyft       31  Cloudy  Boston University Financial District
## 2     1.30     Uber       28  Cloudy       South Station   Theatre District
## 3     1.72     Lyft       29  Cloudy   Theatre District   Haymarket Square
## 4     1.77     Lyft       46  Cloudy      North Station      South Station
## 5     0.77     Lyft       31  Cloudy      North Station   Haymarket Square
##           name temperature humidity windSpeed visibility pressure    period
## 1 Lux Black XL       34.59     0.78      3.07      9.945  1019.21  midnight
## 2     UberPool       40.43     0.64      9.08     10.000   994.99   evening
## 3       Shared       37.92     0.67      8.11      9.995  1003.62   morning
## 4    Lux Black       43.06     0.71      8.12     10.000  1015.00   evening
## 5          Lux       41.89     0.57      2.51      9.953  1022.54 afternoon
##   price
## 1  45.5
## 2   8.0
## 3   9.0
## 4  19.5
## 5  10.5
```

```
# train / test split
shuffled.rideshare = rideshare[sample(1:nrow(rideshare)), ]
rideshare.train = shuffled.rideshare[1:40000, ]
rideshare.test = shuffled.rideshare[40001:45984, ]
```

```r
# save to csv
write.csv(rideshare.train, "data/rideshare_train.csv", row.names=F)
write.csv(rideshare.test, "data/rideshare_test.csv", row.names=F)
```