DataFest 2019 Modeling Exercises

3/25/2019

Intro and Data Processing

Some text and introduction to the dataset.

```
library(tidyverse)
library(lubridate)
cities <- read.csv("/Users/mm06832/Documents/Projects/datafest/2019/data/city_attributes.csv")
humidity <- read.csv("/Users/mm06832/Documents/Projects/datafest/2019/data/humidity.csv") %>%
  gather(2:37, key = "city", value = "humidity")
pressure <- read.csv("/Users/mm06832/Documents/Projects/datafest/2019/data/pressure.csv") %>%
  gather(2:37, key = "city", value = "pressure")
temperature <- read.csv("/Users/mm06832/Documents/Projects/datafest/2019/data/temperature.csv") %>%
  gather(2:37, key = "city", value = "temperature")
weather <- read.csv("/Users/mm06832/Documents/Projects/datafest/2019/data/weather_description.csv") %>%
  gather(2:37, key = "city", value = "weather")
wind_direction <- read.csv("/Users/mm06832/Documents/Projects/datafest/2019/data/wind_direction.csv") %
  gather(2:37, key = "city", value = "wind_direction")
wind_speed <- read.csv("/Users/mm06832/Documents/Projects/datafest/2019/data/wind_speed.csv") %>%
  gather(2:37, key = "city", value = "wind_speed")

#join tables into single dataset
dat <- humidity %>%
  inner_join(pressure, by = c("datetime", "city")) %>%
  inner_join(temperature,  by = c("datetime", "city")) %>%
  inner_join(weather, by = c("datetime", "city")) %>%
  inner_join(wind_direction, by = c("datetime", "city")) %>%
  inner_join(wind_speed, by = c("datetime", "city")) %>%
  left_join(cities, by = c("city" = "City"))
```

Let's take a look at the final dataset - what do you observe that could impact a model's performance? Feel free to perform additioal EDA as required before beginning the modeling stage.

```
glimpse(dat)
```

```
## Observations: 1,629,108
## Variables: 11
## $ datetime       <fct> 2012-10-01 12:00:00, 2012-10-01 13:00:00, 2012-...
## $ city           <chr> "Vancouver", "Vancouver", "Vancouver", "Vancouv...
## $ humidity       <dbl> NA, 76, 76, 76, 77, 78, 78, 79, 79, 80, 81, 81,...
## $ pressure       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ temperature    <dbl> NA, 284.6300, 284.6290, 284.6270, 284.6250, 284...
## $ weather        <chr> "", "mist", "broken clouds", "broken clouds", "...
## $ wind_direction <dbl> NA, 0, 6, 20, 34, 47, 61, 75, 89, 102, 116, 130...
## $ wind_speed     <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Country        <fct> Canada, Canada, Canada, Canada, Canada, Canada,...
## $ Latitude       <dbl> 49.24966, 49.24966, 49.24966, 49.24966, 49.2496...
## $ Longitude      <dbl> -123.1193, -123.1193, -123.1193, -123.1193, -12...
```

Classification

Since the `weather` variables has over 50 different values (run `unique(dat$weather)` to see them all for yourself!) we need to do some preprocessing in order to make this a more suitable target for modeling efforts.

```
#binary classification
dat <- dat %>%
  mutate("weather_binary" = ifelse(weather %in% c("sky is clear", "broken clouds", "few clouds",
                                                  "scattered clouds", "overcast clouds"),
                                   "good",
                                   "bad"))

#multiclass classification
dat <- dat %>%
  mutate("weather_broad" = case_when(
                                   weather %in% c("drizzle", "freezing_rain", "heavy intensity drizzle",
                                                  "heavy intensity rain", "heavy intensity shower rain",
                                                  "light intensity drizzle", "light intensity drizzle ra
                                                  "light intensity shower rain", "light rain",
                                                  "light shower rain", "moderate rain", "proximity modera
                                                  "ragged shower rain", "shower drizzle",
                                                  "very heavy rain", "proximity shower rain") ~ "rain",
                                   weather %in% c("broken clouds", "overcast clouds",
                                                  "scattered clouds", "few clouds") ~ "cloudy",
                                   weather %in% c("heavy snow", "light rain and snow", "light shower sle
                                                  "light snow", "rain and snow", "shower snow",
                                                  "sleet", "snow", "heavy shower snow") ~ "snow",
                                   weather %in% c("thunderstorm with drizzle", "thunderstorm with heavy
                                                  "thunderstorm with light drizzle",
                                                  "thunderstorm with rain", "thunderstorm with light rai
                                                  "heavy thunderstorm", "proximity thunderstorm",
                                                  "proximity thunderstorm with drizzle", "proximity thun
                                                  "proximity thunderstorm",
                                                  "thunderstorm", "ragged thunderstorm") ~ "thunder",
                                   weather %in% c("sky is clear") ~ "clear",
                                   TRUE ~ "Other"
                               )
  )
```

Exercise 1 Fit a classification model of your choice to predict `weather_binary`, and interpret its confusion matrix when applied to held-out data. What is your interpretation of the model's performance?

Exercise 2 Fit a classification model of your choice to predict `weather_broad`. Explain the model's performance, and how your interpretation of its confusion matrix changes in a multi-class scenario.

Regression

For this portion of the example, choose a continuous variable of interest (for example, `temperature`) then predict it using a regression technique.

Exercise 3* For your variable of choice, fit a linear regression model using only the single variable that best explains it (how could you figure this out from the data at hand?). Perform any necessary diagnostics or analysis required to make a conclusion about its interpretation and performance. How does this change if more variables are added?

Exercise 4 Try picking a different variable than the one selected before, and use a regression method of your choice to predict it. What is the MSE on test data, and how do you interpret it?

Conclusion

Exercise 5 In the