Business Analytics Lab (Data 1): Predictive Modeling using Forest Fire Data

Professor Jeffrey Yau

10/21/2017

Background

In some areas, forest fires are a major environmental concern, endangering human lives and causing substantial economic damage. The recent wild fire in Northern California is one such example.

Imagine that your team has been hired by the a government agency that wants to develop an early warning system to identify particularly damaging forest fires.

Data

You are provided with the file **forestfires.csv**, containing measurements taken of recent fires in a Portuguese park.

As a proxy for how damaging a fire is, you should use the **area** variable, representing the region burned in hectares. That is, the dependent (or target) variable in this data is named **"area"**.

The dataset includes a number of meteorological variables. Some of these come from the forest Fire Weather Index (FWI), a Canadian system for rating fire danger. These include Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), and Initial Spread Index (ISI).

The following codebook summarizes each variable:

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: "jan" to "dec"
4. day - day of the week: "mon" to "sun"
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in hectares): 0.00 to 1090.84

Objectives

1. Descriptive Analysis: Conduct an exploratory analysis with the aim of understanding what factors lead to particularly damaging forest fires.
2. Predictive Analysis: Build a machine learning model to predict forest fires

** NOTE: More instructions will be given in the following weeks. Before the next class (on Novemberr 4), try to start exploring the dataset, understand each of the variables, clean the data if necessary, identify any anomalous values in each of the variables, and conduct a thorough EDA. **

Other Instructions:

- **Report is due on December 13, 2017 Presentation is due one day before the last class: December 15, 2017**
- Submission:
  - Each group only needs to make 1 submission via the course page
  - Submit 3 files:
    1. A pdf file including the summary, the details of your analysis, and all the R codes used to produce the analysis. **Please do not suppress the codes in your pdf file.**
    2. R markdown file used to produce the pdf file
    3. A final presentation in powerpoint format
  - Each group only needs to submit one set of files via the course website
  - Use the following file naming convensation; fail to do so will receive 10% reduction in the grade:
    - Student1LastNameFirstName_Student2LastNameFirstName_Student3LastNameFirstName.fileExtension
    - For example, you have a group of three students named John Smith, Jane Doe, and Scott Jones, you should name your file the following
      - SmithJohn_DoeJane_JonesScott.Rmd
      - SmithJohn_DoeJane_JonesScott.pdf
      - SmithJohn_DoeJane_JonesScott.pptx
  - Although it sounds obvious, please write the name of each members of your group on page 1 of your report.
  - This is a group lab, and it can be completed in a group of 3 or 4 people. Each group only needs to make one submission. them take away your own opportunity to learn.
  - In your report, make sure your answer to each of the tasks given below can be easily identified. Highlight or circle your answer would be helpful. I will not spend time searching for your answers.