**CAS MA 575 – Linear Models.**

Boston University, Fall 2018

Course Project Data Description

# 1 Background

As part of your work in this course, you are asked to complete a course project. For this project, you should work on *one* data set from the following 3 choices:

- **Bike Sharing.** For this project, you will use data on the usage of bike sharing resources (e.g., like the Hubway system in Boston).[1]
- **Forest Fires.** In this project you will use data from Portugal to predict how susceptible an area is to forest fires. [2]
- **Facebook Social Media Metrics** The data is related to posts published during the year of 2014 on the Facebook's page of a renowned cosmetics brand. This dataset contains 500 of the 790 rows and part of the features analyzed by Moro et al. (2016). [3]

All 3 datasets can be downloaded from the UCI Machine Learning Repository or the Bulearn site at *Content* → *Project and Lab Materials* → *Project Data*.

[1] Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', *Progress in Artificial Intelligence* (2013): pp. 1-15, Springer Berlin Heidelberg.

[2] If this is your project choice, please cite: P. Cortez and A. Morais. *A Data Mining Approach to Predict Forest Fires using Meteorological Data.* In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimares, Portugal, pp. 512-523, 2007.

[3] Moro, S., Rita, P., & Vala, B. (2016). *Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach.* Journal of Business Research, 69(9), 3341-3351.

## 1.1 Bike Sharing

Bike sharing has become a phenomenon seen world-wide. Large numbers of bikes are made available for rental throughout major metropolitan areas. In addition, and importantly, the system by which these bikes may be rented is automated and, moreover, the data gathered from these automated systems can easily be coupled with other sensor data, measuring things like temperature and other weather characteristics. From a business point of view, it is of interest to be able to accurately predict the level at which bike resources are likely to be used on any given day. That is, at a minimum it is of interest to predict the numbers of bike rentals in an area on a daily (or even hourly) basis.

The data are available at the UCI Machine Learning Repository. See

`http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#`

There you can find additional information on the study and the data, as well as the data files themselves. The data come in the form of two CSV files, one for hourly counts of bike rentals and the other for daily counts. You may load it into R using the command

```
bikedata <- read.csv("day.csv",header=T)
```

## 1.2 Forest Fires

Forest fires are a major environmental issue, creating economical and ecological damage while endangering human lives. Fast detection is a key element for controlling such phenomenon. To achieve this, one alternative is to use automatic tools based on local sensors, such as provided by meteorological stations. In effect, meteorological conditions (e.g. temperature, wind) are known to influence forest fires and several fire indexes, such as the forest Fire Weather Index (FWI), use such data.

The forest Fire Weather Index (FWI) is the Canadian system for rating fire danger and it is formed from five components: Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), and Buildup Index (BUI). The first three are related to fuel codes: the FFMC denotes the moisture content surface litter and influences ignition and fire spread, while the DMC and DC represent the moisture content of shallow and deep organic layers, which affect fire intensity. The ISI is a score that correlates with fire velocity spread, while BUI represents the amount of available fuel. The FWI index is an indicator of fire intensity and it combines the two previous components.

The data is available at

`https://archive.ics.uci.edu/ml/datasets/forest+fires`

and can be loaded into R with the following command

```
Data <- read.csv("forestfires.csv",header=T)
```

## 1.3 Facebook Social Media Metrics

The worldwide dissemination of social media was triggered by the exponential growth of Internet users, leading to a completely new environment for customers to exchange ideas and feedback about products and services. The number of social network users in 2010 is 0.97 billion. This is projected to increase to 2.44 billion users by 2018, i.e. an increase of around 300% in 8 years. Considering its rapid development, social media may become the most important media channel for brands to reach their clients in the near future.

In this study the authors are interested in studying the effect on the *Total interactions* (The sum of *likes*, *comments*, and *shares* of the post) with respect to the 7 input features: *category*, *page total likes* , *type*, *month*, *hour*, *weekday*, *paid*.

The data is available at `https://archive.ics.uci.edu/ml/datasets/Facebook+metrics` or at the Bulearn site.