CAS MA 575 / SPH BS 755 – Linear Models

Boston University, Spring 2024

Data Descriptions for Lab Project

Contents

1 Data Sets

For this project, you should work on **one** data set from the following three choices:

1. **Bike Sharing.** Data on the usage of bike sharing resources (e.g., Bluebikes system in Boston).[1]

2. **BMW Pricing Challenge.** Data on the prices of used BMW cars. This dataset contains various features of more than 3000 cars, the times they were sold in an auction, and the highest bid. You will explore pricing of used BMW cars.

3. **Facebook Social Media Metrics** Data related to posts published during 2014 on the Facebook page of a renowned cosmetics brand. This data set contains 500 of the 790 rows and part of the features analyzed by Moro et al. (2016). [2]

All 3 datasets can be downloaded from the Blackboard site at
`Content` → `Project and Lab Materials` → `Project Data`.
Following are brief descriptions of each of the three data sets. **More details can be found in the `DataSetDescription.txt` file in the corresponding data folders.**

[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] 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 worldwide. 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.

You can download the data from the course website on Blackboard. The data are also available at the UCI Machine Learning Repository at

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

This website has additional information on the study and 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 BMW Pricing Challenge

Estimating the price at which an item will be sold in an auction is one of the main everyday challenges in business. In this project you are asked to analyze market prices for used cars and to understand how prices change depending on car characteristics, market trends, etcetera.

The data were already somewhat cleaned-up, with severely damaged cars excluded. There is no information on minor damages. The price shown in the data is the highest bid that was reached during the auction. The dataset has several car features: type of car (such as convertible or SUV), engine power, model, mileage (they are used cars), registration date, and color. In addition, there are 8 features (feature_1, feature_2, . . . , feature_8) that are each either present in a car or not, but the dataset does not reveal what those features are (they could be leather seats, air conditioning, fancy wheels etc.).

From a business point of view, it is of interest to be able to accurately predict the value of a used car. At a minumum we would like you to find the main determinants of the price of a used car depending on the basic description and the 8 provided features.

You can download the data from the course website on Blackboard. The BMW data set can also be downloaded from Kaggle at

`https://www.kaggle.com/datasets/danielkyrka/bmw-pricing-challenge`

There you can find additional information on the study and the data, as well as the data file itself. The data can be loaded into R with the following command:

```
BMWdata <- read.csv("BMWpricing.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 was 0.97 billion and was projected to increase to 2.44 billion users by 2018, i.e. an increase of around 300% in 8 years. Social media may become the most important media channel for brands to reach their clients. Moro et al. study the effect on the *Lifetime.Post.Consumers* with respect to the 7 input features: *category*, *page total likes*, *type*, *month*, *hour*, *weekday*, *paid.*

You can download the data from the course website on Blackboard. The data is available at

`https://archive.ics.uci.edu/ml/datasets/Facebook+metrics`

The data can be loaded into R with the following command:

```
FBdata <- read.csv("dataset_Facebook.csv",header=T,sep=";")
```

2 Team structure dynamics

Team interaction and coordination is going to be crucial as you work on the project. Sometimes it can be difficult to coordinate the work among the team members. There are many ways to run an effective team and even more for an unsuccessful one.[3]

The following are some suggested roles for the team members. However, feel free to use the team structure that you think will is best suited for the group.

- **Coordinator:** Team leader who will be assigning roles to the different team members. The team leader will coordinate between all the members to produce a well written lab report.

- **Recorder/Writer:** This person will be effectively write the lab report in coordination with the leader and the rest of team.

- **Modeler:** The modeler will be in charge of formulating/producing the regression models and interpreting the results.

- **Coder:** This person will write the `R` code and produce the results.

[3] Please watch the presentations on effective teamwork available at

`http://groupsurvivalguide.babson.edu/`

and read the paper *Successful teamwork: A case study.* It is available in the Lab #1 folder on Blackboard course page or you can download it from

`https://ro.ecu.edu.au/cgi/viewcontent.cgi?article=5007&context=ecuworks`. In particular, look at page 643 to 645 for a description of a successful and unsuccessful team.

- **Monitor/Checker:** This is a very important role. This person will check on the work done from the coder and modeler. Feedback should be given back to the modeler and coder to help them produce a better model.

It is suggested that these roles be rotated between lab reports and for the final lab the best person for the role is assigned. Of course everyone should think about how best to solve the problem and share ideas with the group. Note that for the first report does not have any `R` coding.

2.1 Author contribution statement

Each lab report must include an author contribution statement. Some potential roles to include are

- data acquisition, data processing, conception and design, methodology (modeling), formal analysis (coding), interpreting results, writing (original draft), writing (review and editing), visualization, project coordination

For example, if there are 4 authors Author A, Author B, Author C and Author D, the author contribution statement should look as follows:

A.A.: data processing, formal analysis (coding), writing (original draft).
A.B.: conception and design, interpreting results, writing (original draft), visualization.
A.C.: methodology (modeling), formal analysis (coding), writing (review and editing), project coordination.
A.D.: data acquisition, methodology (modeling), interpreting results, writing (review and editing).