Group Project Assignment

Data Analysis for Business (DAB25)

Nina Deliu

Due Sunday, April 6th 2025 23:59 on LUISS Learn

General Instructions

I expect you to upload your solutions on LUISS Learn as a **single running** `R Markdown` file (`.rmd`) + its `html` output, both named `ProjectWork_xx`, with `xx` the group name of your choice (communicated on LUISS Learn). Assignments are to be performed in groups of maximum 4 students, and individual projects are also welcomed. The report must be submitted by the group leader using the dedicated form on the LUISS learn course page.

You will give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Your responses must be supported by both *textual explanations and the code* you generate to produce your results. *Just examining your various objects in the "Environment" section of RStudio is insufficient – you must use scripted commands and functions.*

The final grade will be based on the methodological correctness and the arguments you give in the answers. Higher grades will be granted for insightful comments, smart coding and clever data visualization.

R Markdown Test

To be sure that everything is working fine, start `RStudio` and create an empty project called `ProjectWork_DAB25`. Now open a new `R Markdown` file (`File > New File > R Markdown...`); set the output to `HTML mode`, press `OK` and then click on `Knit HTML`. This should produce a web page with the knitting procedure executing the default code blocks. You can now start editing this file to produce your project submission.

Important Info

- For more info on `R Markdown`, check the support webpage that explains the main steps and ingredients: R Markdown from RStudio.
- For more info on how to write math formulas in LaTex: Wikibooks.
- **Policy on collaboration**: *Collaboration on homework assignments with fellow students is* ***accepted****. However, such collaboration should be clearly acknowledged, by listing the group of the students with whom you have had discussions concerning your solution. You may* ***not****, however, share written work or code after discussing a problem with others. The solutions are specific to* ***your group****, and any violation will include a grade penalty.*
- **Policy on AI**: *Any support from AI systems, such as* ***CheatyGPT****, should be acknowledged and justified. Omission of acknowledgement will result in a penalty proportional to the extent of their use.*

Problem: Cross-validate this multinomial model

Hitters Data This exercise is based on Hitters dataset, containing information of a major league baseball players from the 1986 and 1987 seasons. Data are provided to you in the ‘Hitters.csv‘ file. The following variables on individual performances and salaries are recorded:

AtBat Number of times at bat in 1986

Hits Number of hits in 1986

HmRun Number of home runs in 1986

Runs Number of runs in 1986

RBI Number of runs batted in in 1986

Walks Number of walks in 1986

Years Number of years in the major leagues

CAtBat Number of times at bat during his career

CHits Number of hits during his career

CHmRun Number of home runs during his career

CRuns Number of runs during his career

CRBI Number of runs batted in during his career

CWalks Number of walks during his career

League A factor with levels A and N indicating player’s league at the end of 1986

Division A factor with levels E and W indicating player’s division at the end of 1986

PutOuts Number of put outs in 1986

Assists Number of assists in 1986

Errors Number of errors in 1986

Salary 1987 annual salary on opening day in thousands of dollars

NewLeague A factor with levels A and N indicating player’s league at the beginning of 1987

Your task

In this exercise, you will apply a multinomial logistic regression model to predict the salary level of baseball players, given their full set of performance statistics (full model). You will then quantify the prediction error of your procedure using cross-validation approaches, and will try to identify a better model.

a. As a first step, formalize this model in a generic format (i.e., with no link to the above data). How does it differ from a simple logistic regression?

b. Now, import the `Hitters` dataset in `R` using the `Hitters.csv` file attached to your assignment. Inspect the data with the appropriate commands, and report the total number of units and measurements. Perform any data pre-processing you deem important (e.g., handle missing values, normalize/standardize predictors if necessary) and justify your actions.

c. To apply this model in R, we first need a categorical response variable (ideally with more than two levels). However, in the original dataset the response Salary is quantitative; therefore, not providing a categorical variable directly. We will derive our Salary levels (or categories) by defining appropriate salary tertiles, e.g., LowSalary, MediumSalary, and HighSalary. This can be done in R by using the command provided below.

```
Hitters$SalaryLevel = cut(Hitters$Salary, breaks = quantile(Hitters$Salary, probs = c(0,
    1/3, 2/3, 1), na.rm = TRUE), labels = c("LowSalary", "MediumSalary", "HighSalary"),
    include.lowest = TRUE)
```

Bonus. You can also define more than three categories and/or more appropriate cutoffs, based on the data distribution. Feel free to choose your option or stick to the proposed one. Any alternative option will be rewarded with a bonus.

d. Take an appropriate visual inspection at your data, focusing on the newly created categorical variable and other variables that may have a predictive power. What type of plot would you use? (*You can get some insights from the Project Guidelines file*)

e. Time to run your model. More than one function is available in R to implement a multinomial regression. Based on your Lab classes, report the function and associated library you will use for this task. Now you can perform a multinomial regression using as response variable the one defined in point c. above and the full set of predictors (full model). Once you get your result, provide a detailed description of the output that you get and *interpret* your results.

f. Time to evaluate the *accuracy* or equivalently the *overall error* of your prediction. Bear in mind that our interest goes beyond the training error.

- First, define the error measure that you will use in this case. *Hint*: We are not in a regression setting, but rather a classification one (you can also refer to Chapter 5 of your book).
- Second, evaluate the full model by using one of the following cross-validation (CV) methods covered in class: 1. Vanilla validation set, 2. LOO-CV, 3. K-fold CV (with K = 5), 4. K-fold CV (with K = 10). The choice of the specific method will be based on a random sample of the values 1:4 according to a seed specified by the date of birth (format DDMMYYYY) of the group member, with the closest date of birthday to August 21. For example, if the group has 3 members with dates of birth 23-07-2000, 14-12-2005, and 09-04-1999, then the one with birthday closest to 21-08 is the first one, and the seed is set to 23072000. You shall also report the name of the student with associated date of birth. The following code will produce the CV method:

```
set.seed(23072000)  # report the name of the student with this date of birth
cv_methods = c("1. Vanilla validation set", "2. LOO-CV", "3. K-fold CV (with K = 5)",
    "4. K-fold CV (with K = 10)")
sample(cv_methods, 1)
```

```
## [1] "3. K-fold CV (with K = 5)"
```

For groups of 3-4 students. Implement a second cross-validation method of your choice.

g. The full model may not be the ideal one especially if it based on a large number of (irrelevant) predictors. Your goal now is to identify a better model (with lowest *test* error) for predicting the individual salary level. You may use any strategy that you think is useful, including but not limited to stepwise procedures for Multiple Regression you have seen during the Lab sessions.