# Group Projects

## Introduction to the continuous assessment project

The main aims of this assessment are to enable you to gain some experience in working as part of a small team ($\sim$4 students) on a research project. Together with your team you will be applying and comparing a suite of classification methods to answer a question of interest about a specific dataset.

A group report containing all findings and analysis, as well as the associated `R` code, should be submitted by **1pm on Friday 24th March**. The report and `R` code should be submitted by **only one** person from each group.

A **Contribution Form** and a **Declaration of Originality Form** should be submitted by **all** members of a group, also by **1pm on Friday 24th March**. In the Contribution Form, you will provide information on how much you feel you personally contributed to the project, as well as the contributions of the other group members. Your final grade for the group project will take your contribution into account.

Details on submitting the report, Contribution Form and Declaration of Originality Form are given at the end of this handout.

## Projects

This section provides information on the different group projects. Your group should **only** analyse the dataset **allocated** to you; no mark will be given for submitting a report on another dataset. Information on project allocation is available from Moodle ("Group project allocation.pdf").

### Project 1: IMDB movie dataset (for groups 1-10)

What can we say about the success of a movie before it is released? Are there certain companies (Pixar?) that have found a consistent formula? Given that major films costing over $100 million to produce can still flop, this question is more important than ever to the industry. Can we predict which films will be highly rated, whether or not they are a commercial success?

To dig into these questions, the IMDB movie dataset is collected. It originally contains metadata information on 5043 movies, such as genre, director name and actor name; a full list of variable description is given in Table 1.

For the project, 2000 movies are randomly allocated to each group; you can find the allocated dataset on Moodle.

**Question of interest**

Based on the metadata, it would be interesting to understand what are the important factors that make a movie more successful than others. You will need to define your own notion of success, e.g. films with high gross earnings are successful and those with low gross earnings are unsuccessful, films with high IMDB scores are successful and those with low scores are unsuccessful. Regardless of which variable you use to define success, this is a binary classification task, i.e. successful *versus* unsuccessful.

It is not necessary to use all of the variables, but you may if you think appropriate. You may also need to do some data manipulation/transformation to get it into a format appropriate for the question you wish to answer.

| Variable Name | Description |
|---|---|
| movie_title | Title of the Movie |
| duration | Duration in minutes |
| director_name | Name of the Director of the Movie |
| director_facebook_likes | Number of likes of the Director on his Facebook Page |
| actor_1_name | Primary actor starring in the movie |
| actor_1_facebook_likes | Number of likes of the Actor_1 on his/her Facebook Page |
| actor_2_name | Other actor starring in the movie |
| actor_2_facebook_likes | Number of likes of the Actor_2 on his/her Facebook Page |
| actor_3_name | Other actor starring in the movie |
| actor_3_facebook_likes | Number of likes of the Actor_3 on his/her Facebook Page |
| num_user_for_reviews | Number of users who gave a review |
| num_critic_for_reviews | Number of critical reviews on imdb |
| num_voted_users | Number of people who voted for the movie |
| cast_total_facebook_likes | Total number of facebook likes of the entire cast of the movie |
| movie_facebook_likes | Number of Facebook likes in the movie page |
| plot_keywords | Keywords describing the movie plot |
| facenumber_in_poster | Number of the actor who featured in the movie poster |
| color | Film colorization. 'Black and White' or 'Color' |
| genres | Film categorization like 'Animation', 'Comedy', 'Romance', 'Horror', 'Sci-Fi', 'Action', 'Family' |
| title_year | The year in which the movie is released |
| language | English, Arabic, Chinese, French, German, Danish, Italian etc |
| country | Country where the movie is produced |
| content_rating | Content rating of the movie |
| aspect_ratio | Aspect ratio the movie was made in |
| movie_imdb_link | IMDB link of the movie |
| gross | Gross earnings of the movie in Dollars |
| budget | Budget of the movie in Dollars |
| imdb_score | IMDB score of the movie on IMDB |

Table 1: Variable description of IMDB dataset

## Project 2: Bank marketing dataset (for groups 11-20)

A Portuguese banking institution launched directed marketing campaigns to promote their products. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be subscribed or not.

The dataset collected is related to 17 campaigns that occurred between May 2008 and November 2010, corresponding to a total of 41,118 clients. For each client, a number of attributed was stored (see Table 2) and if there was a success (the target variable).

For the project, 10,000 records are randomly allocated to each group; you can find the allocated dataset on Moodle.

**Question of interest**

Based on the dataset, predict if the client will subscribe (yes/no) a term deposit (variable y).

It is not necessary to use all of the variables, but you may if you think appropriate.

## Project 3: Drug consumption dataset (for groups 21 and beyond)

The drug consumption dataset contains records for 1885 respondents. For each respondent 12 attributes are known: Personality measurements which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), level of education, age, gender, country of residence and ethnicity. All input attributes are originally categorical and are quantified. After quantification values of all input features can be considered as real-valued. In addition, participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse) and one fictitious drug (Semeron) which was introduced to identify over-claimers. For each drug they have to select one of the answers: never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day. Detailed description of the dataset is given in Table 3.

For the project, each group is given a dataset with 1885 rows and 14 columns, where the first 12 columns are the 12 attributes, 13th column is the fictitious drug, and the last column is one of the legal or illegal drugs.

**Question of interest**

Based on the dataset, discriminate participants into different categories of drug use. The categories are: (1) 'Never Used', (2) 'Used over a Decade Ago', (3) 'Used in Last Decade', (4) 'Used in Last Year', (5) 'Used in Last Month', (6) 'Used in Last Week', and (7) 'Used in Last Day'. Each participant belongs to only one class, that is, if the participant used the drug in the last day, they will be categorised into 'Used in Last Day', rather than 'Used in Last Week', 'Used in Last Month', etc.

You may perform a seven-class classification using the categories above. Alternatively, you may merge the seven categories into three and then perform three-class classification, e.g. classifying between *never used* (i.e. 1), *used over a decade over* (i.e. 2) and *used in last decade* (i.e. merging 3–7), or between *never used* (i.e. 1), *used over a year ago* (i.e. merging 2 and 3) and *used in last year* (i.e. merging 4–7). It is best to look at the dataset allocated to you and then make the decision about the number of classes (between 3 and 7) and the class name. Note that you are **NOT** allowed to merge into two classes and perform binary classification.

It is not necessary to use all of the variables, but you may if you think appropriate. You may also need to do some data manipulation/transformation to get it into a format appropriate for the question you wish to answer.

| Variable Name | Description |
|---|---|
| *target variable:* | |
| y | has the client subscribed a term deposit? (binary: 'yes', 'no') |
| *bank client data:* | |
| age | age at the contact date |
| job | type of job, e.g. 'admin.', 'blue-collar', 'entrepreneur', 'self-employed', 'services', 'student', 'technician', 'unemployed' |
| marital | marital status, incl. 'divorced', 'married', 'single' (note: 'divorced' means divorced or widowed) |
| education | education level, e.g. 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree' |
| default | has credit in default? |
| housing | has housing loan? |
| loan | has personal loan? |
| *related with the last contact of the current campaign:* | |
| contact | contact communication type ('cellular', 'telephone') |
| month | last contact month of year, e.g. 'jan', 'feb' |
| day_of_week | last contact day of the week, e.g. 'mon', 'tue' |
| duration | last contact duration, in seconds. **Important note**: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model. |
| *other attributes:* | |
| campaign | number of contacts performed during this campaign and for this client |
| pdays | number of days that passed by after the client was last contacted from a previous campaign |
| previous | number of contacts performed before this campaign and for this client |
| poutcome | outcome of the previous marketing campaign, incl. 'failure', 'nonexistent', 'success' |
| *social and economic context attributes:* | |
| emp.var.rate | employment variation rate – quarterly indicator |
| cons.price.idx | consumer price index – monthly indicator |
| cons.conf.idx | consumer confidence index – monthly indicator |
| euribor3m | euribor (Euro Interbank Offered Rate) 3 month rate – daily indicator |
| nr.employed | number of employees – quarterly indicator |

Table 2: Variable description of bank marketing dataset

| Variable Name | Description | Value | Meaning |
|---|---|---|---|
| ID | the index of record in original database | | |
| Age | age of participant | -0.95197 | 18-24 |
| | | -0.07854 | 25-34 |
| | | 0.49788 | 35-44 |
| | | 1.09449 | 45-54 |
| | | 1.82213 | 55-64 |
| | | 2.59171 | 65+ |
| Gender | gender of participant | 0.48246 | Female |
| | | -0.48246 | Male |
| Education | level of education of participant | -2.43591 | Left school before 16 years |
| | | -1.73790 | Left school at 16 years |
| | | -1.43719 | Left school at 17 years |
| | | -1.22751 | Left school at 18 years |
| | | -0.61113 | Some college or university, no certificate or degree |
| | | -0.05921 | Professional certificate/ diploma |
| | | 0.45468 | University degree |
| | | 1.16365 | Masters degree |
| | | 1.98437 | Doctorate degree |
| Country | country of current residence of participant | -0.09765 | Australia |
| | | 0.24923 | Canada |
| | | -0.46841 | New Zealand |
| | | -0.28519 | Other |
| | | 0.21128 | Republic of Ireland |
| | | 0.96082 | UK |
| | | -0.57009 | USA |
| Ethnicity | ethnicity of participant | -0.50212 | Asian |
| | | -1.10702 | Black |
| | | 1.90725 | Mixed-Black/Asian |
| | | 0.12600 | Mixed-White/Asian |
| | | -0.22166 | Mixed-White/Black |
| | | 0.11440 | Other |
| | | -0.31685 | White |
| *personality measurements:* | | | |
| Nscore | NEO-FFI-R Neuroticism | | |
| Escore | NEO-FFI-R Extraversion | | |
| Oscore | NEO-FFI-R Openness to experience | | |
| Ascore | NEO-FFI-R Agreeableness | | |
| Cscore | NEO-FFI-R Conscientiousness | | |
| Impulsive | impulsiveness measured by BIS-11 | | |
| SS | sensation seeing measured by ImpSS | | |
| *drug_name* (Target) | Legal or illegal drug, e.g. alcohol, cocaine | CL0 | Never Used |
| | | CL1 | Used over a Decade Ago |
| | | CL2 | Used in Last Decade |
| | | CL3 | Used in Last Year |
| | | CL4 | Used in Last Month |
| | | CL5 | Used in Last Week |
| | | CL6 | Used in Last Day |

Table 3: Variable description of drug consumption dataset

## Tasks

After creating the training, validation (if applicable) and test data sets, you should:

- Perform exploratory analysis on the training data set.
- Apply five to seven classification techniques that you have learned during weeks 3 to 6 (e.g. k-nearest neighbours, linear/quadratic discriminant analysis, tree-based methods, support vector machines, neural networks). Include a brief[1] description of the method and justification for selecting the method.
- Study one of the following methods and implement the method on your dataset. Include a description of the method in your own words.
  - t-distributed stochastic neighbor embedding (a dimensionality reduction method which may be used for visualisation)
  - Adaptive boosting (a classification method)
  - Gradient boosting (a classification method)
  - Some regularisation techniques, optimisation techniques, or hyperparameter selection approaches for deep neural networks[2]
- Create appropriate graphs or summaries that communicate the results from these methods.
- Comment on the graphs/summaries and interpret the results.
- Compare the results to choose a final, "best" classification model, with justification given for the choice and comment on this model's overall performance using appropriate evaluation measures.

## Details and submission instructions of report, R code, contribution form and declaration of originality forms

### Report

The **maximum number of pages** that the report can be is **12** (excluding title page and reference if any), with minimum size 12 font, margin size 1 inch/2.54cm, 1.5 line spacing. Every part or full page over that limit will incur a 10% penalty taken off your final project mark (i.e. 2 pages over will result in 20% being taken off your mark).

The report should have a clear structure, e.g. dividing into different sections, such as Introduction, Exploratory Analysis, Methods, Results and Conclusions, labelling graphs and tables (if any).

Include reference where appropriate to avoid plagiarism. This applies to any direct quotation or paraphrasing, from any material, including but not limited to textbooks, online articles and videos.

Do not simply copy and paste `R` output into your report; instead, create proper tables. You can copy plots from `R` into your report.

The language used in your report should be formal. Don't write the report like a story (e.g. "I did this and then I did this and then I did that"). Use the third person where possible.

The report should be written in language interpretable by a layperson as much as possible.

Remember to include the project number in the report.

[1] Do not go into too much detail describing the methods you used (i.e. don't quote all the notes back to me).

[2] See Chapters 7, 8 and 12 in the textbook 'Deep Learning'. https://www.deeplearningbook.org/

Please save your report as a **PDF** document in the form

`projectnumber.pdf`

One student from each group should be allocated the job of submitting the report to Moodle. Submit by clicking on the link "Submit report and R code".

The report must be uploaded to the Moodle site by **1pm on Friday 24th March**. The standard penalties for late hand-in apply.

**This report contributes 20% to the total module mark.**

### R code

The R code should be self-contained, i.e. that is allow the examiner to:

- install and load any packages you might have used;
- reproduce any models you have worked on; and
- recreate any of the graphs and summaries you have on your report.

Make sure `set.seed` is included, e.g. before data splitting, before classification methods that are sensitive to initialisation, in order to ensure that all reproduced results are same as the ones on your report.

The R code should be well enough commented that the code is understandable to one who hasn't written it.

Please save the R code by either using the R script, i.e. `.r` file, or using the Rmarkdown, i.e. `.rmd` file. Again, name the file by the project number, i.e.

`projectnumber.r` or `projectnumber.rmd`

One student from each group should be allocated the job of submitting R code to Moodle. Submit by clicking on the link "Submit report and R code".

The R code must be uploaded to the Moodle site by **1pm on Friday 24th March**. The standard penalties for late hand-in apply.

### Contribution form and declaration of originality forms

All students from the group should submit a Contribution Form and Declaration of Originality Form.

Please submit the completed forms as PDF documents via the link "Submit contribution form and declaration form".

Both files must be uploaded to the Moodle site by **1pm on Friday 24th March**.

## Marking rubrics

As shown in the table overleaf, the report will be marked according to four sections ("style/presentation", "approach", "quality of explanation", "statistical programming"), and an individual mark between 0 and 22 will be given for each section based on the headings (Poor, Weak, etc.) and the heading descriptors. The final report mark is the sum of the individual section marks multiplied by their associated weights.

| | *Criterion* | *Poor*<br>*0-5* | *Weak*<br>*6-8* | *Acceptable*<br>*9-11* | *Moderate*<br>*12-14* | *Good*<br>*15-17* | *Very Good*<br>*18-22* |
|---|---|---|---|---|---|---|---|
| **CORE** | **STYLE AND PRESENTATION (weight = 10%)**<br>e.g.<br>• clarity of style<br>• structure and balance between sections<br>• use of notation<br>• quality of diagrams and tables<br>• correct referencing style | **Unclear with errors.** | **Unclear.** | **Variable clarity.** | **Generally clear and sound.** | **High quality.** | **Generally excellent.** |
| | **APPROACH (weight = 30%)**<br>e.g.<br>• appropriateness of choice of statistical methods, including visualizations, data pre-processing, model fitting, model selection, etc.<br>• correctness of application of statistical methods<br>• scale of ambition | **No understanding apparent. Incorrect and inaccurate use of statistical techniques.** | **Little understanding apparent. Inappropriate and/or inaccurate use of statistical techniques.** | **Sensible but inadequate understanding. Not always appropriate choice of statistical techniques.** | **Understanding of issues and theory. Generally appropriate and sound use of statistical techniques.** | **Appreciable depth of understanding of issues and theory. Appropriate, accurate and sound use of statistical techniques.** | **Exceptionally assiduous, precise and concise; very high quality throughout.** |
| | **QUALITY OF EXPLANATION (weight = 45%)**<br>e.g.<br>• description of project background, data, statistical methods<br>• justification for choice of statistical methods<br>• interpretation of results reported<br>• understanding of implications and limitations<br>• appropriateness of conclusions drawn | **Incoherent or incorrect descriptions, interpretations and/or conclusions.** | **Poor exposition; substantial defects in descriptions and/or interpretations.** | **Coherent but sketchy exposition; some defects in descriptions and/or interpretations.** | **Fairly clear and coherent exposition. Generally sound descriptions and appropriate interpretations.** | **Mostly clear exposition, with clear indications of thought. Appreciable depth of understanding of theory.** | **Very clear, concise exposition, with clear indications of outstandingly good thought. A very sound understanding of theory.** |
| | **STATISTICAL PROGRAMMING (weight=15%)**<br>e.g.<br>• correctness of R code<br>• quality of reporting R code | **No code or missing code for a substantial amount of methods.** | **Substantial errors in code for a few methods; missing code for some methods.** | **Major errors in code for a small amount of methods; very little comments included.** | **Minor errors in code for a small amount of methods; some comments included.** | **Correct code for almost all methods with mostly clear comments.** | **Correct, concise code for all methods with very clear comments.** |