MATH 11205: Machine Learning in Python 2023-2024
Project 1 Description

We will be using data collected by the SHARE project, a survey on health, ageing, and retirement collected across multiple European countries. A simplified version of the data, easySHARE, will be used for this project, and the data have been provided in the file `easyshare.csv`, after some initial cleaning steps (described below). We have also included the file `easyshare_all.csv` for those who wish to explore additional variables.

Assignment Goal

For the purpose of the project, consider yourself a **Data Scientist Consultant** who has been hired by European Union to improve understanding of factors associated with higher dementia risk and potential strategies to reduce risk. Dementia is a major international public health concern, and the cost of dementia to governments, social services and individuals has reached staggering figures. In the absence of a cure, effective prevention strategies are critical to reduce dementia risk and lessen the burden of dementia at all levels of society.

Towards this aim, you have been asked to use these data to build a predictive model of the cognitive score of individuals that captures its underlying relationship with the various factors that have been collected. In this setting, the cognitive score is used a proxy for dementia severity and is a clinical measure that combines results from tests designed to assess cognitive function (namely, two numeracy tests, two word recall tests, and an orientation test). Your model should then be used to gain insights to advise on potential risk factors for dementia. In particular, the government officials and carer providers are interested in identifying *modifiable* risk factors that are amenable to lifestyle interventions. As such, you should use your model to suggest potential lifestyle and/or government/societal interventions to reduce dementia risk. In summary, you need to develop an **explainable, validated** model for cognitive score as the outcome of interest using features derived from the data provided and any additional sources you would like to use.

It is important that any conclusions you draw from your model are well supported and sound and that you understand limitations of the model and the data. We explicitly **do not want a blackbox model** - you should be able to explain and justify your modeling choices and your model's predictions.

Your model may use as few or as many of the provided factors, and you may transform and manipulate these factors in any way that you want to generate additional features.

We have covered a number of models and modeling approaches in the lectures and workshops, and you should explore a variety of different approaches for this particular task. However, your ultimate goal is to deliver a **single** model. These are competing interests, and it is up to you to find a reasonable balance between exploring different models and selecting your proposed model; some of your marks will be based on how well you accomplish this. In addition, you should compare the performance of your model against a baseline model(s); although the main focus should be the description of your model (not the baseline).

Working as a team

This project may be completed by a team of up to 4 students (minimum of 1 student). Feel free to create your own team during workshop hours, building on the pairs for the workshop assignments. Since we are not assigning teams, if you are a team that is looking for more members or someone looking for a team please use the pinned post on Piazza to find each other.

After the assignment is completed we will distribute a brief peer evaluation survey - members who contributed significantly less than their peers will potentially have their overall mark penalized.

Dataset Details

These are the available variables given in the dataset `easyshare.csv`:

- `mergeid` - person identifier
- `wave` - wave identifier
- `country` - country identifier
- `country_mod` - modified country identifier
- `female` - dummy encoded gender with 0 for male and 1 for female
- `age` - age at interview
- `birth_country` - country of birth
- `citizenship` - citizenship of respondent
- `isced1997_r` - ISCED-97 encoding of education (6 levels - see pg. 11 of data guide)
- `eduyears_mod` - years of education
- `eurod` - depression scale ranging from 0 "not depressed" to 12 "very depressed"
- `bmi` - body mass index
- `bmi2` - categorized body mass index
- `smoking` - smoke at present time
- `ever_smoked` - ever smoked daily
- `br010_mod` - drinking behavior
- `br015_` - vigorous activities
- `casp` - CASP-12 score measures quality of life and is based on four subscales on control, autonomy, pleasure and self-realization, ranges from 12 to 48
- `chronic_mod` - number of chronic diseases
- `sp008_` - gives help to others outside the household
- `ch001_` - number of children
- `cogscore` - measure of cognitive function combining results from two numeracy tests, two word recall tests, and an orientation test.

For additional information on each variable, please refer the the easySHARE data guide included in the project materials.

When building your model, it may be helpful to consider possible risk factors that have been identified in literature, specifically, low education, midlife hearing loss, obesity and hypertension, late-life depression, smoking, physical inactivity, diabetes, and social isolation (Livingstone et al. (2020)).

It is also possible to extract further variables from the full easySHARE dataset contained in the file `easyshare_all.csv`; if you choose to do so, you will need to merge based on `mergeid` and `wave`. Note that the SHARE project collects longitudinal data, that is multiple observations (waves) for

each individual. We are focusing on a simplified version of the data, with only one observation per individual included in `easyshare.csv` (i.e. the data has been cleaned so that `mergeid` is unique across rows). Note that each person may have been interviewed in different years (waves).

The target variable `cogscore` combines the results from five cognitive function tests available in the full easySHARE dataset (two numeracy tests, two word recall tests, and an orientation test).

If you wish, you may also access additional data collected in the full SHARE database (from the website). However, you would need to sign a waiver to access the data. In addition, under data agreements and waivers that I have signed on your behalf, the data provided for this project must NOT be shared publicly (e.g. if your team is using GitHub, keep the repo private).

You may even choose to utilise additional data sources, if you feel it is a good plan.

Required Structure

A Jupyter notebook template called 'project1.ipynb' has been provided. It includes the required sections along with brief instructions on what should be included in each section. Your completed assignment must follow this structure - **you should not add or remove any of these sections, if you feel it is necessary you may add extra subsections within each**. Please remove the instructions for each section in the final document.

All of your work must be contained in the 'project1.ipynb' notebook, we will only mark what is included in this file (both the write-up and relevant coding). You may work on the notebook in whichever environment you prefer, but please ensure that the final pdf file includes all necessary parts of your writing.

Our expectation is that most projects will be roughly 20-25 pages in length at most including text & figures, but excluding the related code. Overall, there is an **upper limit of 30 pages** including the coding part. Your notebook must include all of your work, but make sure that you are only retaining required components, e.g. remove unused code and figures (if a figure is not explicitly discussed in the text it should not be in the final document). **So, there is a trade-off between the length of your text and coding snippets while constructing your report.** Overall, your project will be partially assessed on your organization / presentation of the document - it should be as polished and streamlined as possible. **Try to be as concise as possible while creating your write-up. We highly recommend that you check the appearance of your rendered PDF before submitting, as its appearance can differ significantly from the notebook**.

You are expected to submit your completed work. For this, please submit your final PDF of project report (generated from a Jupyter notebook) to the Project assignment on Gradescope. Please ensure that you **tag all groups members** on Gradescope, and also add all group member names either in the notebook metadata or in additional markdown cell block at the beginning of the file.

Getting Help

- **Project Q&A Online Meeting:** We will hold an online Q&A meeting to answer any questions at the end of Week 6. Date and time to be determined, please keep an eye on Piazza where we will post a poll to find the best date that works for most students.

- **Piazza:** This forum will be used as the central location for all course related discussions and questions, and should be used over emailing course staff directly. The course lecturers will monitor and respond to questions, but feel free to provide some constructive responses to peer's questions. You can access Piazza from the course LEARN page or sign-up at:

  https://piazza.com/ed.ac.uk/spring2023/math11205

  Also, see the good practice guide for how to use piazza most effectively:

  https://teaching.maths.ed.ac.uk/main/undergraduate/studies/learning-advice/piazza

- You can also ask questions at the end of lectures during any Q&A time or during workshops.

Further References

We have provided additional resources in the project materials. For further information on the dataset and variables, see:

- *Guide to easySHARE* (2022)
- *SHARE Release Guide* (2022)

For further information, on dementia risk factors:

- Livingston et al (2020). *Dementia prevention, intervention, and care: 2020 report of the Lancet Commission.* Lancet. 396(10248):413-446.
- Livingston (2017). *Dementia prevention, intervention, and care.* Lancet. 390(10113):2673-2734.