Course Project

Introduction

The course project is designed to provide students with machine learning modeling experience on real-world data. You will be given a prediction task where the training data set is provided. You will use the data science concepts learned during the lectures and labs to train machine learning models to deliver predictions on a given test set, whose actual labels are, however, not seen.

In the data preparation step, you will practice handling missing data and encoding non-numerical data for model input. In the model training step, you will practice choosing an appropriate machine learning model, training, and fine-tuning the model to achieve the best prediction performance. Note that software development and implementation code should only be based on Jupyter notebooks.

Although machine learning modeling can be done easily with the use of publicly available programming libraries like `scikit-learn`, project implementation which usually involves handling unseen input data and organizing various experiments can be time-consuming and error-prone if not well planned. Hence, apart from the data science principles, you should also practice analyzing and designing software systems and software architecture on this coursework-size project as part of the software development process before you are exposed to a larger-scale machine learning project at work. It is important to note that your project structure might differ from those learned during the labs, however, your software development and implementation should strictly follow the data science pipeline.

Data

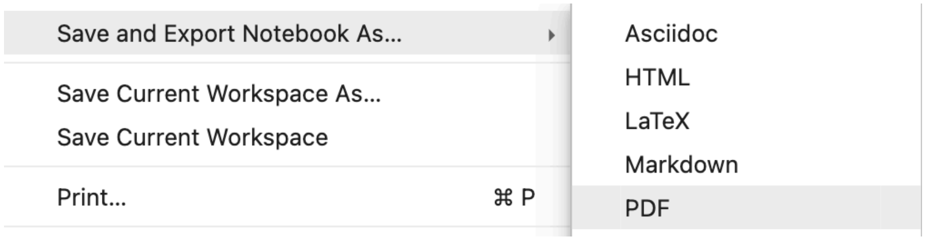
1. The file `high_salary.csv` is the training data set. Each row represents an individual person and the information about the person. The column `label` has values 0 or 1, representing whether the person has a low salary or a high salary, respectively. Your task is to train a model to predict the column `label`.

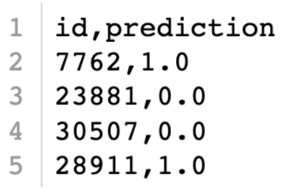
2. The file `high_salary.live.csv` is the live data set. Likewise, each row represents an individual person and the information about the person. Note that the column `label` in the live data set is not given. You will use the trained model to predict whether the person in the live data set has a low or high salary. This simulates using the trained model in a real-world scenario where actual labels are not known at the prediction time.

The predictions on the live set, however, are to be submitted as part of the project and graded toward the final project score. Hence, it is highly recommended that you split the training data to create your own test data. This will allow you to evaluate your prediction results on unseen data.

Submission Requirements

File	Requirements	Mark scheme
1. Presentation and slides	A PDF export of the presentation slides This file should be named: **`{group_name}_presentation.pdf`** - Do not submit `.ppt` or `.pttx` files as they will not be graded The presentation must cover all of the following topics: - Describe the specification of the prediction task	10 marks The project follows the data science pipeline The presentation covers all the topics

	- Describe the problems found in the input data - Describe the steps used to process the data - Explain why the chosen model is appropriate - Explain why the chosen model configuration is appropriate - If you do experiments before choosing, clearly describe the methodology - If you use any specific techniques, clearly describe the techniques - Provide the performance metrics on the training data set - Provide supporting evidence/metrics that the model should perform well on the live data	The presentation clearly addresses all the topics
2. Code	PDF exports of all Python notebook files used in the project  Save and Export Notebook As... Save Current Workspace As... Save Current Workspace Print... ⌘ P Asciidoc HTML LaTeX Markdown PDF These files should be named: `step***.pdf` Compress all the files into a `.zip` file named: **`{group_name}_code.zip`** You may organize the files in subdirectories as you need - Do not compress `.ipynb` files as they will not be graded - Do not submit `.rar` or other extensions as they will not be graded - Do not submit malicious files or any other files than specified as the entire project will not be graded and you are responsible for the violation of the computer crime act	5 marks The code follows the data science pipeline The code follows the presentation The students can explain how the code correctly

3. Predictions	A CSV file, containing the predictions on the live data
	This file should be named: **`{group_name}_predictions.live.csv`**
	The file content should look like this:
	
	`id,prediction`
	`7762,1.0`
	`23881,0.0`
	`30507,0.0`
	`28911,1.0`
	The predictions should always preserve the original sample IDs. There shall be no excuse for mistaken IDs as there never are in real-world situations.
		5 marks
		F1 scores are sorted in descending order (no decimal rounding).
		5.0 marks to {1st, 2nd, 3rd}
		4.5 marks to {4th, 5th,}
		4.0 marks to {6th, 7th, 8th}
		3.5 marks to {9th,10th}
		…
		1.5 marks to {19th,20th}
		1 mark to the rest
		0 mark if no submission, submissions with an invalid format, submissions that cannot be graded

Submission Deadlines

Deadline	Session	Details
15 November 2024 12:00 (before noon) via LMS	Submission of code and predictions	Following the submission requirements, you are required to submit all the code and the predictions. Late submissions will not be graded.
16 November 2024 09:00 (before class) via LMS	Submission of presentation slides	Following the submission requirements, you are required to submit the presentation slides before class. You should submit the presentation slides at the same time as the code and the predictions. This extra time is kindly given to those in need.
16 November 2024 09:00 - 12:00, 13:00 - 16:00 via WebEx	Final presentation	Each group is given a 10-minute presentation session to present the course project to the class. The presentation and the presentation slides will be graded with respect to the specifications, together with the code and predictions.