Final Assignment: AI methods for Business 25/26

Dr. Debarati Bhaumik and Dr. Riccardo Pinosio

2.1 General guidelines

In the final assignment you are required to implement several machine learning models to answer a business question for the provided dataset (the *Metacritic dataset* you already are familiar with from the database course). The goal is for you to consolidate the knowledge you have acquired in the lectures and tutorial sessions by building a variety of machine learning models independently.

2.2 The problems

Your group is required to formulate one business question and related sub-questions for the *Metacritic dataset* which is available on Brightspace.

The business question must be of relevance, and it must have a suitable number (three to five) research sub-questions that need to be answered. The business questions and sub-questions must be clear and answerable based on the data and must be of sufficient complexity to warrant the development of machine learning models to answer them.

Each model you build needs to satisfy the quality criteria discussed in section 2.4 below.

2.3 Deliverables

The deliverables on which you will be graded on as a group effort are:

1. A zipfile containing the python code (jupyter notebooks) with the implementation of the cleaning, feature extraction, modelling, evaluation and explainability code.
2. A group presentation video of a **maximum of 10 minutes** of the product developed including the answers to the questions in the assignment, description of the methodology deployed, results along with model assessment and an indication of each individual participation in the assignment. *If the presentation is above 10 minutes it is an automatic fail for the assignment.*
3. The powerpoint for the video presentation, submitted as a separate ppt file.

All deliverables must be submitted on Brightspace.

Although in principle grading is applied to a group as opposed to an individual student, the lecturers reserve the right to individually assess a student based on their contributions and understanding and adjust the grade for individuals within a group.

2.4 Machine learning analysis of the Metacritic dataset

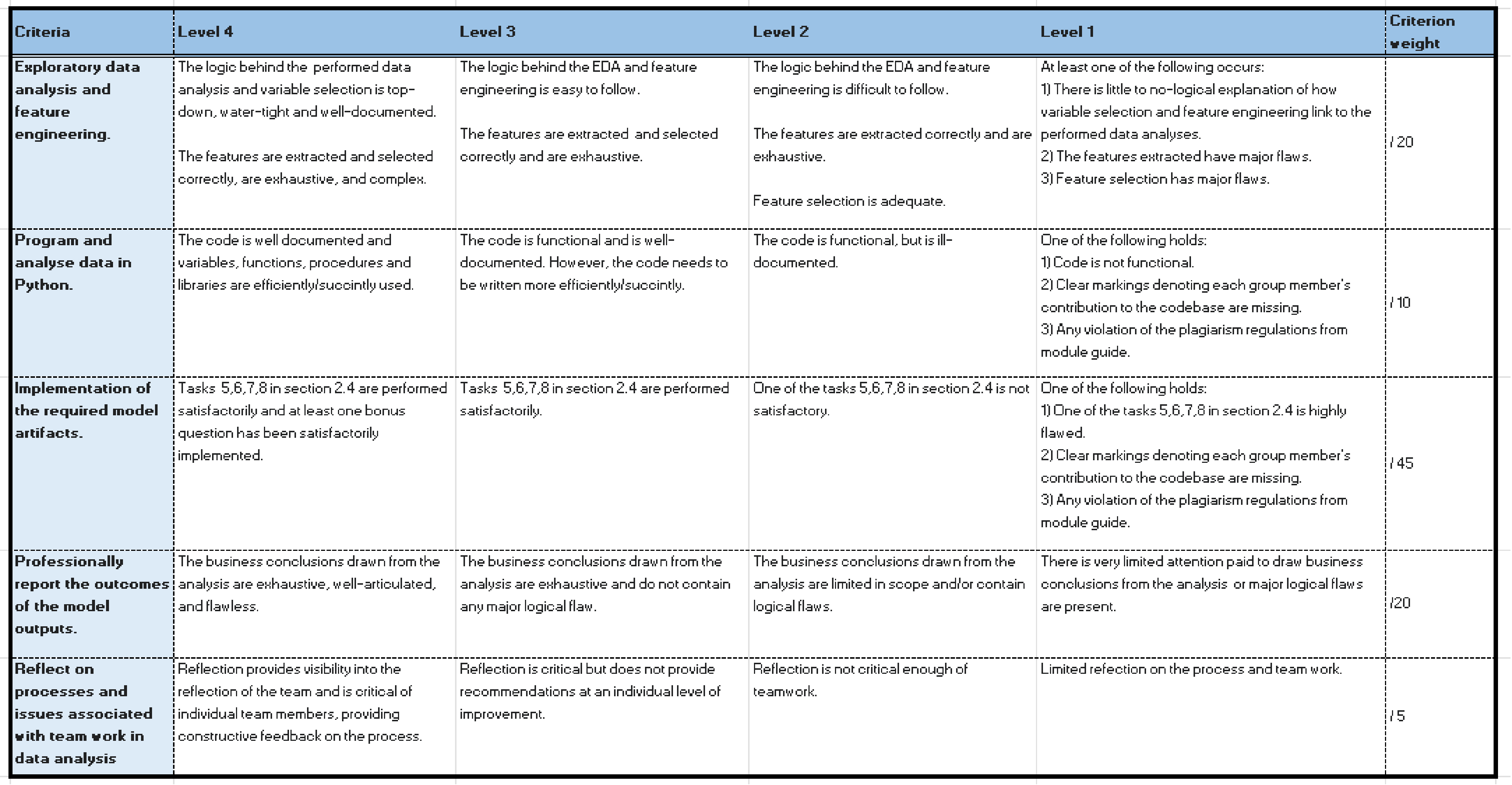
To answer the research question and sub questions on the Metacritic dataset you are required to implement the following:

1. **Exploratory data analysis and data cleaning of the dataset:** Typically you will use here the techniques you have already learned in statistics and database management, such as correlation analysis and descriptive statistics, distribution visualizations, outlier removal, missing values imputation, …

2. **Feature engineering:** To train your models, you will need to create a feature dataset from the clean data. You might have to perform some or all of the following: feature scaling/normalization, feature selection based on the EDA results, categorical variable encoding, feature transformation and new feature creation.
    a. You are required to use one or more transformer models from the Hugging Face repository to extract features from the Metacritic dataset. For instance, you can calculate the vector embeddings of the review titles and use the embeddings as features for the supervised models.
3. **Feature selection**: You are required to use a matrix decomposition method to reduce the dimensionality of the dataset and perform feature selection.
4. **Datasets generation:** Generate train, validation, and test datasets as explained during the lectures and labs.
5. **Model fitting:** Perform fitting and grid-search hyperparameter tuning on the train and validation datasets to find the best parameters for the following models:
    a. A baseline model of your choice. You must justify why you choose that specific model as a baseline,
    b. A random forest model, and
    c. A neural network model

    You should also produce plots to investigate how the loss function of the neural network evolves over the epochs on the train and validation datasets. You should do this to identify the best number of training epochs for the set of hyperparameters you identified.
6. **Model evaluation:** Evaluate the fit of each model on the test set using an appropriate performance metric that is suitable for the specific problem. Here you should also consider whether the chosen metric is easy to communicate to non-technical stakeholders. You might want to consider multiple metrics for this purpose. You should also produce some plots investigating in more detail the models' accuracy on the test set: for instance, actual vs. predicted plot, confusion matrix.
7. **Model selection:** Compare the performance of the models you have trained: which one is better?
8. **Explainability:** you must explain the predictions of **the best performing model** based on both SHAP and counterfactuals. A clear interpretation of the explainability results must be provided.
9. **Bonus**: can you identify the main topics in the movie reviews using a transformer-based clustering model like BERTopic? Are there topics that are good predictors for your target?
10. **Bonus**: can you perform a cluster analysis of the dataset and identify meaningful clusters?

2.5 Grading rubric



Criteria	Level 4	Level 3	Level 2	Level 1	Criterion weight
Exploratory data analysis and feature engineering.	The logic behind the performed data analysis and variable selection is top-down, water-tight and well-documented. The features are extracted and selected correctly, are exhaustive, and complex.	The logic behind the EDA and feature engineering is easy to follow. The features are extracted and selected correctly and are exhaustive.	The logic behind the EDA and feature engineering is difficult to follow. The features are extracted correctly and are exhaustive. Feature selection is adequate.	At least one of the following occurs: 1) There is little to no-logical explanation of how variable selection and feature engineering link to the performed data analyses. 2) The features extracted have major flaws. 3) Feature selection has major flaws.	/ 20
Program and analyse data in Python.	The code is well documented and variables, functions, procedures and libraries are efficiently/succintly used.	The code is functional and is well-documented. However, the code needs to be written more efficiently/succintly.	The code is functional, but is ill-documented.	One of the following holds: 1) Code is not functional. 2) Clear markings denoting each group member's contribution to the codebase are missing. 3) Any violation of the plagiarism regulations from module guide.	/ 10
Implementation of the required model artifacts.	Tasks 5,6,7,8 in section 2.4 are performed satisfactorily and at least one bonus question has been satisfactorily implemented.	Tasks 5,6,7,8 in section 2.4 are performed satisfactorily.	One of the tasks 5,6,7,8 in section 2.4 is not satisfactory.	One of the following holds: 1) One of the tasks 5,6,7,8 in section 2.4 is highly flawed. 2) Clear markings denoting each group member's contribution to the codebase are missing. 3) Any violation of the plagiarism regulations from module guide.	/ 45
Professionally report the outcomes of the model outputs.	The business conclusions drawn from the analysis are exhaustive, well-articulated, and flawless.	The business conclusions drawn from the analysis are exhaustive and do not contain any major logical flaw.	The business conclusions drawn from the analysis are limited in scope and/or contain logical flaws.	There is very limited attention paid to draw business conclusions from the analysis or major logical flaws are present.	/20
Reflect on processes and issues associated with team work in data analysis	Reflection provides visibility into the reflection of the team and is critical of individual team members, providing constructive feedback on the process.	Reflection is critical but does not provide recommendations at an individual level of improvement.	Reflection is not critical enough of teamwork.	Limited refection on the process and team work.	/ 5