School of Computing and Information Systems
The University of Melbourne
COMP30027, Machine Learning, 2024

Project 2: IMDB Movie Rating Prediction

Task:	Build a classifier to predict the rating of movie based on information features on movie.
Due:	Group Registration: Friday 03 May, 5pm
	Project: Friday 17 May, 5pm
Submission:	Report (PDF) and code to Canvas; test outputs to Kaggle in-class competition
Marks:	The Project will be marked out of 20, and will contribute 20% of your total mark.
Groups:	Groups of 1 or 2, with commensurate expectations for each (see Sections 2 and 5).

1 Overview

The goal of this project is to build and critically analyze supervised machine learning methods for predicting IMDB movie ratings based on various predictor variables that include movie-title, duration, director and actor(s) names and facebook likes, keywords, genre, country, budget, and others. There are five possible outcomes 0 being the lowest and 4 being the highest.

This assignment aims to reinforce the largely theoretical lecture concepts surrounding data representation, classifier construction, evaluation and error analysis, by applying them to an open-ended problem. You will also have an opportunity to practice your general problem-solving skills, written communication skills, and critical thinking skills.

2 Deliverables

The deliverables of the project are listed as follows. More details about deliverables are given in the Submission: 5

1. **Report**: a written report, of 1,300-1,800 words (for a group of one person) or 2,000-2,500 words (for a group of two people).
2. **Output**: the output of your classifiers, comprising the label predictions for test instances, submitted to the Kaggle[1] in-class competition described below.
3. **Code**: one or more programs, written in Python, which implement feature selection and machine learning models to make predictions and evaluate the results.

3 Data

The movie dataset is derived from the IMDB 5000 Movie Dataset[2], focusing on movie rating prediction based on various predictors associated with movies.

The aim of gathering this data is to design a classifier (or predictor) that will use the predictor variables (features) and predict the rating for each movie as 4 being the highest.

In our dataset, each entry represents infromation of the movie and contains:

[1] `https://www.kaggle.com/`

[2] `https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset`

- Predictors: director_name, num_critic_for_reviews, duration, director_facebook_likes, actor_3_facebook_likes, actor_2_name, actor_1_facebook_likes, gross, genres, actor_1_name, movie_title, num_voted_users, cast_total_facebook_likes, actor_3_name, facenumber_in_poster, plot_keywords, num_user_for_reviews, language, country, content_rating, title_year, actor_2_facebook_likes, movie_facebook_likes, title_embedding, average_degree_centrality, andimdb_score_binned
- Movie Rating: Indicates binned rating of the movies imdb_rating_binned (5 possible values, 0, 1, 2, 3, and 4).

You will be provided with a training set and a test set. The training set (consisting of 2997 rows) contains features and the imdb_rating_binned, which is the "class label" of our task. The test set (consists of 759 rows) only contains the features without labels.

Students are provided with a training set and a test set:

- *movies_train.csv:* the features and class label for training instances.
- *movie_test.csv:* the features without class labels for test instances.
- *movie_text_features_*.zip*: the preprocessed text features for training and test sets, 1 zipped file for each text encoding method. Details about these text features are provided in README.

4 Task

You are expected to develop machine learning models to predict the rating of movie based on predictor variables (e.g., movie title, duration, director name, genre, etc.). You will **explore effective features, employ feature selection/ feature engineering, implement and compare different machine learning models and conduct error analysis** for this task.

Various machine learning techniques have been (or will be) discussed in this subject (0R, Naive Bayes, Decision Trees, kNN, SVM, neural network, etc.); many more exist. You may use any machine learning method you consider suitable for this problem. *You are strongly encouraged to make use of machine learning software and/or existing libraries (such as sklearn) in your attempts at this project.*

In addition to different learning algorithms, consider different ways to preprocess and utilize the predictor variables. For instance, feature engineering techniques like normalization, standardization, and handling missing values can significantly impact model performance. Feel free to apply any preprocessing steps or feature engineering techniques that you think could improve your model's ability to generalize from the training set to the test set.

You are expected to complete the following two phases for this task:

- **Training-evaluation phase**: the holdout or cross-validation approaches can be applied on the training data provided.
- **Test phase**: the trained classifiers will be evaluated on the unlabeled test data. The predicted labels of test cases should be submitted as part of deliverable.

5 Submission

The report, code, peer reviews and reflections should be submitted via Canvas; the predictions on test data should be submitted to Kaggle.

5.1 Individual vs. Team Participation

You have the option of participating individually, or in a group of two. In the case that you opt to participate individually, you will be required to implement **at least 2 and up to 4** distinct Machine Learning models. Groups of two will be required to implement **at least 4 and up to 5** distinct Machine Learning models, of which *one is to be an ensemble model – stacking based on the other models*. The report length requirement also differs, as detailed below:

Group size	Distinct models required	Report length
1	2–4	1,300–1,800 words
2	4–5	2,000–2,500 words

Group Registration

If you wish to form a group of 2, **only one** of the members needs to register by **Friday 03 May 5:00pm**, via the form "**Project 2 Group Registration**" on Canvas. For a group of 2, **only one** of the members needs to submit deliverables.

Note that after the registration deadline, you will not be allowed to change groups. If you do not register before the deadline above, we will assume that you will be completing the assignment as an individual, even if you were in a two-person group for Assignment 1.

5.2 Report

You report is expected to demonstrate the knowledge that you have gained and the critical analysis you have conducted in a manner that is accessible to a reasonably informed reader.

The report should be 1,300-1,800 words (individual) or 2,000-2,500 words (groups of two people) in length excluding reference list, figure captions and tables. The report should include the following sections:

1. Introduction: a basic description of the task and a short summary of your report.
2. Methodology: what you have done, including any learners that you have used, and features that you have engineered. *This should be at a conceptual level; a detailed description of the code is not appropriate for the report. The description should be similar to what you would see in a machine learning conference paper.*
3. Results: performance of your classifiers, in terms of evaluation metric(s) and, ideally include figures and tables.
4. Discussion and Critical Analysis: this section should include a more *detailed discussion* which contextualises the behaviour of the method(s), in terms of the theoretical properties we have identified in the lectures and error analysis of the method(s). This is the most important section of the report.
5. Conclusion: demonstrates your identified knowledge about the problem.
6. Reference: reference of related work.

Note that we are more interested in seeing evidence that you have thought about the task and investigated the reasons for the relative performance of different methods, rather than in the raw accuracy/scores of different methods. This is not to say that you should ignore the relative performance of different runs over the data, but rather that you should think beyond simple numbers to the reasons that underlie them, and connect these to the theory that we have discussed in this subject.

We provide LaTeXand Word style files that we would prefer that you use in writing the report. Reports must be submitted in the form of **a single PDF file**. If a report is submitted in any format other than PDF, we reserve the right to return the report with a mark of 0.

5.3 Predictions of test data

To give you the possibility of evaluating your models on the test set, we will be setting up a Kaggle in-class competition for this project. You can submit results on the test set there, and get immediate feedback on your model's performance. There is a Leaderboard, that will allow you to see how well you are doing as compared to other classmates participating online. The Kaggle in-class competition URL and instructions are provided on the LMS.

You will receive marks for submitting at least one set of predictions for the unlabelled test set into the competition; and get good accuracy (higher than 50%). The focus of this assignment is on the quality of your critical analysis and your report, rather than the performance of your Machine Learning models.

6 Assessment Criteria

The Project will be marked out of 20, and is worth 20% of your overall mark for the subject. The mark breakdown will be:

Report	18 marks
Performance of classifier / Kaggle	2 marks
TOTAL	20 marks

The report will be marked according to the rubric, which is published on the Canvas. You have to submit your code that supports the results presented in your report. If you do not submit an executable code that supports your findings, you will receive a mark of 0 for the "Report".

The performance of classifier is for submitting at least one set of model predictions to the Kaggle competition (1 mark); and getting better accuracy (higher than 50%) (1 mark).

Since all the data exist on the World Wide Web, it is possible to "cheat" and identify some of the class labels from the test data using non-Machine Learning methods. If there is any evidence of this, the performance of classifier will be ignored, and you will instead receive a mark of 0 for the "Performance of classifier".

7 Using Kaggle

The Kaggle in-class competition URL will be announced on Canvas shortly. To participate the competition:

- Each student should create a Kaggle account (unless they have one already) using your Student-ID
- You may make up to 8 submissions per day. An example submission file can be found on the Kaggle site.
- Submissions will be evaluated by Kaggle for accuracy, against just 50% of the test data, forming the public leaderboard.
- Prior to competition close, you may select a final submission out of the ones submitted previously – by default the submission with highest public leaderboard score is selected by Kaggle.
- After competition close, public test scores will be replaced with the private leaderboard test scores (100% test data).

8 Assignment Policies

8.1 Terms of Use

Please note that the dataset is a sample of actual data posted to the World Wide Web. As such, it may contain information that is in poor taste, or that could be considered offensive. We would ask you, as much as possible, to look beyond this to the task at hand. For example, it is generally not necessary to read individual records.

The opinions expressed within the data are those of the anonymised authors, and in no way express the official views of the University of Melbourne or any of its employees; using the data in an educative capacity does not constitute endorsement of the content contained therein.

If you object to these terms, please contact Basim Azam (basim.azam@unimelb.edu.au) as soon as possible.

8.2 Changes/Updates to the Assignment Specifications

We will use Canvas to advertise any (hopefully small-scale) changes or clarifications in the assignment specifications. Any addenda made to the assignment specifications via Canvas will supersede information contained in this version of the specifications.

8.3 Late Submissions

You are strongly encouraged to submit by the date and time specified above. If circumstances do not permit this, then the marks will be adjusted as follows:

- Each day (or part of the day) that the report is submitted after the specified due date and time for Stage I, 10% will be deducted from the marks available, up until 7 days (1 week) has passed, after which regular submissions will no longer be accepted.

8.4 Academic Honesty

While it is acceptable to discuss the assignment with others in general terms, excessive collaboration with students outside of your group is considered collusion. Your submissions will be examined for originality and will invoke the University's Academic Misconduct policy (`https://academicintegrity.unimelb.edu.au/`) where either inappropriate levels of collaboration or plagiarism are deemed to have taken place.

Since Kaggle competition performance contributes to the final mark on this assignment, your submissions to Kaggle must be your own original work and must abide by the rules of the Kaggle competition. Submitting predictions to Kaggle which are not the results of your own models (or allowing another student to submit your model's predictions under their Kaggle account) will be considered a breach of the university's Academic Misconduct policy, as will any attempts to circumvent the rules of the Kaggle competition (for example, exceeding the competition's daily submission limit).