MATH 11205: Machine Learning in Python 2022-2023
Project 1 Description

We will be using data about the American TV show The Office. These data are provided as 'the_office.csv' in this repository and are meant to give you a reasonable starting point for this assignment. These data were derived from the data available in the https://pypi.org/project/schrutepy/ package. The package provides a data frame containing the entire text transcripts from all episodes of the show.

Assignment Goal

For the purpose of the project, consider yourself a **Data Scientist contractor** who has been hired by NBC Universal to advise on the creation of a special reunion episode of The Office. Your employers are particularly interested in understanding what made some episodes more popular than others. As such, your task is to use these data (or any other) to build a predictive model that captures the underlying relationships between these features and the audience ratings, and then use the insights you gain from this model to advise what NBC Universal should do to produce the highest rated reunion episode possible. In other words, you need to develop an **understandable, validated** model for The Office's **'imdb_rating'** 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 this model be **accurate and reliable** and any conclusions you draw well supported and sound. 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 features, and you may transform and manipulate these features 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 the two models; some of your marks will be based on how well you accomplish this.

Working as a team

This project may be completed by a team of up to 3-4 people (at most 4). Feel free to create your own team during workshop hours and do self-enrolment in one of the previously created Assignment Groups. Since we are not assigning or forming 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.

Data Set Details

These are the available variables given in the data set. You can benefit other related data sets if you feel it is a good plan.

- ‘season‘ - Season number of the episode
- ‘episode‘ - Episode number within a season
- ‘episode_name‘ - Episode name
- ‘director‘ - Episode director(s), names are separated by ‘;‘
- ‘writer‘ - Episode writer(s), names are separated by ‘;‘
- ‘imdb_rating‘ - Episode rating on IMDB
- ‘total_votes‘ - Number of ratings for episode on IMDB
- ‘air_date‘ - Original air date of episode
- ‘n_lines‘ - Number of spoken lines in episode
- ‘n_directions‘ - Number of lines containing a stage direction
- ‘n_words‘ - Number of dialog words in episide
- ‘n_speak_char‘ - Number of different characters with spoken lines in episode
- ‘main_chars‘ - Main characters appearing in episode (main characters were determined to be characters appearing in more than 1/2 of the episodes)

Required Structure

A Jupyter notebook template called ‘project.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 ‘project.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. For the name of your .ipynb file, use the convention of ‘name_surname.ipynb (ie. ozan_evkaya.ipynb).

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. For a group submission - all contributors should be added to the assignment on Gradescope. At the beginning of project '.ipynb' notebook, the contributor name places are available to fill out !

Getting Help

- Project Q&A Online Meetings: See details on course information tab on the Learn page.
- 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