Data, Algorithms and Meaning – Autumn 2020

Assignment 2 – Building a Recommendation Engine

There are two key deliverables for this assignment, Part A and Part B. You can work in teams of your choice for part A of this assignment. Part B is an individual assignment. Please refer to the ***assessment guide*** for further details on marking criteria.

PART A – Recommendation Engine

The Business Brief

You are working as a Data Scientist for an online streaming entertainment company. The content management executive team has hired you to look at their data and build a recommendation engine that is able to recommend movies to their users that they will enjoy. The board is very interested in getting the right content to their users as well as any insights you have about their consumer's current watching habits. There has been very little deep analytical work done on this data so far.

The Dataset

The folder you have been given has been prepared by their previous analyst and contains the following data sources:

Dataset	Description
train_raw.rds	Raw training data containing information about the user, the item (movie) they rated and their rating.
test_raw.rds	The same as train_raw.rds but this is the testing data. Hence it does not contain the rating.
scrape.rds	Extra data on the movies. This has been scraped from IMDB and detailed further below.
train_students.rds	The training and test sets for this project. After applying the AT2_prep_students.R code to the raw training and testing sets, these are written out. Below outlines further the treatment applied to the data sources to arrive at this data.
test_students.rds	

The original dataset is taken from the famous MovieLens 100k dataset.

https://grouplens.org/datasets/movielens/100k/

A more detailed data dictionary on each of these items is given in the appendix. Users are identified by a user_id, movies by a movie_id (or item_id, these are the same) and the target column is 'ratings' as these are the ratings (out of 5) that a user has given to a movie.

A number of variables have already been engineered for you, however there is a lot of scope to engineer more variables for this task. Some specific questions that will guide discussion on feature engineering are detailed below.

- How do users generally rate movies or specific genres of movies?
- How does a user differ from the population in the above? Are they generally harsher or easier on movies or specific types of movies?
    - How do you define population?
- Do users of specific demographic groupings tend to rate movies (or types of movies) a certain way?
- Are there attributes about movies that make them more likely to be rated higher or lower in general or by certain people?

You have been provided with the script **AT2_prep_students.R** which outlines how some first features were made and joined together to give you a head start.

Hint 1: You must be careful with how you generate features to ensure there is no data leakage. Close attention will be paid to this.

Hint 2: There are also features in this dataset you should not use in your modelling.

Some background reading

Some papers and notes are given below to assist you to understand traditional approaches. This is not needed to undertake your machine learning task, however it assists to set the scene of what the standard approaches have traditionally been.

Some key topics you may want to learn about:

1. User-user and item-item collaborative filtering
2. Content-based collaborative filtering.

Paper Reference	PDF Link	Notes
Linden, G., Smith, B., York, J., 2003. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing 7, 76–80. https://doi.org/10.1109/MIC.2003.1167344	Link	A very well-cited paper by Amazon that discusses collaborative filtering
Castelluccio, M., 2006. The music genome project. Strategic Finance 57–59.	Link	A paper on Pandora's content based approach, one of the early examples of using item features to assist predictions.
Gunawardana, A., Shani, G., 2009. A survey of accuracy evaluation metrics of recommendation tasks. Journal of Machine Learning Research 10, 2935–2962.	Link	A paper that discusses various metrics, their trade-offs and importances for recommendation tasks.

* Note some of the links will require a UTS sign on.

You are free to explore previous work that has been completed on this data, from either the original data set or kaggle competitions. However, caution is advised in using other people's solutions since this new data has proprietary variables, a proprietary data scrape from IMDB and hence requires some new interpretation and a newly trained model.

The Competition

The data and submission process are managed via a kaggle competition. There will also be a live leaderboard. The link to the competition is here:

https://www.kaggle.com/t/105a4ec334a1481d95f9fc92fa2ba3f8

https://www.kaggle.com/c/dam-aut-20-at2a

The first link is a privacy link required to ensure the competition is not open to the public. **You must go through the first link** to register for the competition. You may use the second link when you come back other times to the competition after having registered. The competition will close at midnight the night before the assessment is due.

You will need to each join the competition and then merge a team. This link should assist.

The performance of your model will be evaluated using RMSE on the Kaggle competition as this is a regression task. You do not have access to the rating column for the test set as Kaggle will evaluate the results you upload and provide your score. Therefore, you must do your work on train_students.rds and then submit predictions for the ratings that users gave to items in test_students.rds. You must therefore be careful in engineering variables to ensure your model will have access to these when undertaking the prediction for Kaggle.

Part of the validation data is used to create your public score but a second part is private and will be used to create a private leaderboard that is only seen by the administrators and withheld until the assignment is finished. This is to assist with 'gaming' the public leaderboard with submissions. You are also limited to 4 submissions per day so make sure you put your best feet (and by that we mean models) forward!

You can see the AT2_ sample_UPLOAD.csv file for how to format your submissions. The user_id and item_id do not matter for Kaggle so should not be included. You are advised to concatenate user_id and item_id then drop them prior to writing out for submission. Kaggle will take the user_item column and the rating column to do its calculation.

The following code will assist:

df$user_item <- paste(df$user_id, df$item_id, sep="_")

An important note on metrics:

This assignment is set up as a regression task for the purpose of the Kaggle competition and this should be your primary aim and you must submit a regression-based model and RMSE score. However, you are free to reinterpret the brief in the spirit of the business brief (Classification? Different regression metric or scoring function?). Wider reading may find a better approach to deliver on the business brief. However, be careful of losing focus and spreading yourself too thin!

Deliverables:

The assessment brief for DAM outlines that a report is also due for this component as well as a statement of contributions. Therefore, there are three deliverables for this task (these only need to be submitted once per team).

1. Complete, commented R-code
2. Report following the CRISP-DM framework including:
    a. The business problem
    b. The available data
    c. Your data preparation process
    d. Any particular insights you discovered about the data
    e. Details of your model training, including the assumptions you made with a rationale for why you adopted this process
    f. Your evaluation methodology
    g. Preliminary results (kaggle public evaluation measures)
    h. Consideration of ethical issues
3. A statement outlining contributions of each team member for this assignment

The board are reasonable tech savvy and have listened to analyst presentations before. However, they have not undertaken any work on recommendation engines or machine learning.

Part B – Management Presentation (INDIVIDUAL ASSIGNMENT)

Each member of the team is required to submit a management presentation on your approach. You should reduce and *alter* your report to align to this different audience. It is **crucial** to ensure that your presentation is appropriate for senior management that are largely non-technical in background.

Your presentation should be short and concise, no more than 10 slides.

Appendix 1 – IMDB Web Scrape Data:

Webscraped IMDB feature	Description
Rating of ten	The global rating out of 10 displayed on IMDB
Count Ratings	Global number of ratings of the movie displayed on IMDB
Mature rating	Rating of a film's suitability for certain audiences based on its content - 13 levels but some of them are duplicates according to different standards
Length	Length of movie in minutes
Males votes	Number of votes by males
Males average vote	Average rating by males
Females votes	Number of votes by females
Females average votes	Average rating by females
Aged under 18 votes	Number of votes by those under 18 years old
Aged under 18 average	Average rating by those under 18 years old
Males under 18 votes	Number of votes by males under 18
Males under 18 average	Average rating by males under 18
Females under 18 votes	Number of votes by females under 18
Females under 18 average	Average rating by females under 18
Aged 18-29 votes	Number of votes by those under between 18-29 years old
Aged 18-29 average	Average rating by those between 18-29 years old
Males 18-29 votes	Number of votes by males between 18-29
Males 18-29 average	Average rating by males between 18-29
Females 18-29 votes	Number of votes by females between 18-29
Females 18-29 average	Average rating by females under between 18-29
Aged 30-44 votes	Number of votes by those under between 30-44 years old
Aged 30-44 average	Average rating by those between 30-44 years old
Males 30-44 votes	Number of votes by males between 30-44
Males 30-44 average	Average rating by males between 30-44
Females 30-44 votes	Number of votes by females between 30-44
Females 30-44 average	Average rating by females under between 30-44
Aged 45+ votes	Number of votes by those over 45 years old
Aged 45+ average	Average rating by those over 45 years old
Males 45+ votes	Number of votes by males over 45
Males 45+ average	Average rating by males over 45
Females 45+ votes	Number of votes by females over 45
Females 45+ average	Average rating by females over 45
Imdb staff votes	Number of votes by IMDB staff
IMDB staff average	Average rating by IMDB staff
Top 1000 voters votes	Number of votes by top 1K voters

Top 1000 voters average	Average rating by top 1K voters
US users votes	Number of votes by US users
US users average	Average rating by US users
Non US users votes	Number of votes by non-US users
Non US users average	Average rating by non-US users

* Some of these will be in the training dataset prepended by 'item_imdb_'

Appendix 2 – Extra Engineered Variables:

Derived feature	Description
item_mean_rating	The mean rating for the item across all its recorded ratings
age_band	From a user's age, classified them into either: under_18, 18_to_29, 30_to_44, or 45_and_over
user_gender_item_mean_rating	The mean rating for the item across the given user's gender cohort
user_age_band_item_mean_rating	The mean rating for the item across the given user's age band cohort
user_gender_item_imdb_mean_rating	The mean IMDB rating (from webscraping) for the item across the given user's gender cohort
user_gender_item_imdb_votes	The number of IMDB votes (from webscraping) for the item across the given user's gender cohort
user_age_band_item_imdb_mean_rating	The mean IMDB rating (from webscraping) for the item across the given user's age band cohort
user_age_band_item_imdb_votes	The number of IMDB votes (from webscraping) for the item across the given user's age band cohort
user_gender_age_band_item_imdb_mean_ratin g	The mean IMDB rating (from webscraping) for the item across the given user's gender and age band cohort
user_gender_age_band_item_imdb_votes	The number of IMDB votes (from webscraping) for the item across the given user's gender and age band cohort

Appendix 3 – User, Item, Ratings Variables:

Feature	Description
user_id	Unique ID of a user
item_id	Unique ID of an item (movie). This is the same as movie_id
rating	Rating of 1-5 that a user gave a movie
timestamp	Timestamp when the user submitted the rating for the movie.
age	Age of the user in years
gender	Gender of a user (M or F)
occupation	Users occupation. A variety of options.

movie_title	Title of the movie
release_date	Release date of the movie
video_release_date	Release date of the video of the movie
unkown….western	Boolean true/false if the movie was this category. A number of columns.