The objective of this challenge is to assess your ability to:

- perform basic data manipulation and data pre-processing
- demonstrate awareness of the computations involved
- perform feature engineering
- train and tune ML models
- asses performance of the ML models
- obtaining clear, useful, and business driven insights from data and models

Below you will find the instructions for this challenge. Good Luck!

Data
The data employed was obtained from a public Kaggle dataset related to movie ratings. Feel free to have a look at the data description on Kaggle, and please, download from there the following files:

- genome_scores.csv
- genome_tags.csv
- link.csv
- movie.csv
- rating.csv
- tag.csv

Instructions

0.1 Technology used

We suggest you work on this challenge using Python, although other languages are not excluded if you think there is a justification for using them.

0.2 Presentation format
You can present your challenge in a Jupyter Notebook. If this is not the best option for you, you can use a different tool, but be sure we have access to your code.

1. Exploratory analysis
You might perform exploratory analysis on this data, but you are not required to present it to us, we will focus mainly on the feature engineering section of this challenge.

2. Binary classification
The objective of the model you create will be to predict whether a client will rate as “high” a movie or not.

3. Modeling structure
Create a dataframe where each instance (row) corresponds to a rating of some movie made

by some user at a given point in time. Note in particular that if a user has several ratings, then each of her ratings must appear on a different row. Each column will correspond to a predictive variable (below we give instructions on the predictive variables). Then, create a column with the response variable for your model. This response variable is defined as:

- 1 in case the rating is >= 4 (flag for "high" rating)
- 0 in case the rating is < 4

Feature engineering show feature importance

The main part of this challenge will consist in performing feature engineering. Implement a series of features that you think will have a high predictive power. Be creative, and explore all the ideas you might have on what information could be useful to predict the rating of a client.

Important Note: When creating the features that you propose, that predict the rating that a user will give to some movie:
Assume that this model will be used to generate online predictions on a production setting, and be aware of the implications of that, and put special attention for data leakage.

Model implementation

Implement a ML model which predicts your response variable using the predictive features you created. Explain the process you followed to generate/choose the model. Do not invest too much time training/tuning your model. It will be enough for us if you choose an algorithm and a configuration of hyperparameters you have seen in the past to work well for this type of problems. Please, explain and justify your selection of the algorithm and hyperparameters.

Feature importance

Give an explanation of the importance of each feature, and show us which of the features you created had the highest impact on your model. Explain and justify your choice of the importance metric.

Important note

Even though your model predicts whether a client will rate as “high” a movie or not, we will not look into your skills building recommendation systems (like collaborative filtering). As we mentioned, we are interested in your feature engineering and modeling skills, using the modeling structure defined above.

4. Conclusions

Add some comments summarizing your work. Also, add comments on how you would improve it if further time was given to you.

We wish you success with this challenge!