Final Project - GDA Track 2

Going to the movies

Building a full Machine Learning Classification Project

The goal of this challenge is to do a complete machine learning project for the prediction of a movie genre. We are using a real-world dataset of the IMDB website. Using the description, the main actors, the rating and other information we predict the movie genres.

Main Question

How precisely can you predict the genre of a movie provided the other information in the IMDB dataset?

- One challenge is the fact that each movie can be classified into up to three genres. There are several options on how to deal with this. As data scientists, it's your task to formulate some strategy!
- This is a real-world dataset, success is no guarantee. If no good model can be found, describe your attempts and explain why they bore no fruit.
- Also it might be a good idea to start with a reduced form (only a couple of genres, only a fraction of the rows etc…)

TASK 1 - Data Gathering

- Download the data from this Kaggle site
- Import it into pandas dataframes
- Check for missing values

TASK 2 - Basic Data Exploration & Train-Validation-Split

- Explore the data and try to understand the features
- Merge the tables as you deem it necessary
- Engineer the target value - depending on your strategy
- By using your prior knowledge, formulate a hypothesis on the most important features
- Perform a data exploration and show your findings in different plots
- Split the dataset into training, validation and test set

TASK 3 - Model Training and Tuning

- Train a classification model of your choice
- Pick a performance metric and print the results for the training dataset as well as the validation dataset
- Repeat this process for at least two other algorithms of your choice
- Tune different hyperparameters for each of the algorithms
- Add a k-fold Cross-Validation to each of the algorithms
- Explore different effects of feature engineering
- Integrate a Hyperparameter-GridSearch
- Perform an analysis of feature importance
- In the end, run your best algorithm/hyperparameters on the test set

Possible Add-Ons

- Find other interesting patterns and facts in the dataset
- Formulate ideas for possible follow-up projects

TASK 4 - Present your findings

- Present your findings in the form of slides as well as your jupyter notebook (with structure and explanations)
- Focus on the workflow of your project. In this case, the main point is to showcase how a classification machine learning model is done, what the main steps are and how training and tuning work.

Good Luck!