# Python projects course

.

## Home Work

| DEADLINE | TOPICS |
|---|---|
| 1402/12/27 | پروژه دوم دیتا |

لینک گیت هاب پروژتون در شیت لینک گیت هاب بارگذاری کنید.

# TMDB Movie Dataset EDA, Modelling, and Recommender System

## About Dataset

What can we say about the success of a movie before it is released? Are there certain companies (Pixar?) that have found a consistent formula? Given that major films costing over $100 million to produce can still flop, this question is more important than ever to the industry. Film aficionados might have different interests. Can we predict which films will be highly rated, whether or not they are a commercial success?

This is a great place to start digging in to those questions, with data on the plot, cast, crew, budget, and revenues of several thousand films. For further information about dataset you can refer here.

## Project Description

The aim of this project is to gain some insight about the data for further modeling it to for investigating the profitability of the movies. That is, your job is to predict whether based on some features a movie is profitable or not. Obviously, you need to gain inight about your data, so before modeling it is necessary to do a comprehensive Exploratory Data Analysis, and then based on your knowledge dive into the modeling phase.
After the modeling you will be asked to design a simple recommendation system. Do a little research about different recommendation system schemes (e.g. Collaborative filtering, content-based filtering) and design a simple recommender system.

The suggest path toward this project is described as follow:

1. Data loading and gain info
2. Data cleaning and Preprocessing
3. Exploratory Data Analysis (EDA)
4. Data Modeling (Using different ML models)
5. Evaluation
6. Recommender System

As we are all familiar with the first two parts, we won't discuss it here.

**Exploratory Data Analysis:**

Based on the features and different questions that may arise in the dataset perform different analysis and gain more exploratory insight about the data. For each analysis write your query, output, and your conclusion. Your creativity and the questions you design is important.

At the final stage, summarise your conclusions.

**Data Modeling:**

Based on the profit column define a target indicating whether a movie has been profitable or not.

For the next stage, perform feature engineerig and select a subset of useful features for training a model.

Perform prediction based on the following models and evaluate your model:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. XGBoost

For each model report 4 metrics (Accuracy, Precision, Recall, F1-Score) and further plot ROC-AUC plot for each model.

Compare differernt models and their evaluations and analyse the results.

**Recommender System:**

Generally, a recommender system performs under certain similarity measures between the prefered items by different users. They are used to predict the rating or preference that a user would give to an item. Almost every major tech company has applied them in some form or the other: Amazon uses it to suggest products to customers, YouTube uses it to decide which video to play next on autoplay, and Facebook uses it to recommend pages to like and people to follow. Moreover, companies like Netflix and Spotify depend highly on the effectiveness of their recommendation engines for their business and sucees.

The similarity mesuare is generally, cosine similarity

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

For this project, develop a recommender system based on different relevant tags exist in your dataset. Different tags should mapped to a vectorized space and for this manner you can use TF-IDF vectorizer, Count vectorizer or any other relevant model (even a trainable model).
Finally, by calculating the similarity measure of the vectors, write a function that gives 5 movie suggestions based on the input movie.

**Notes:**

1. Your creativity in developing the above model is the most important aspect of this project.
2. Your project should have at least one .ipynp file (the main executable file)
3. Write proper **markdowns** in your notebook and try to be as comprehensive as possible
4. Write a report on the project and explain different phases of the project and further their outcomes and conclusions.
5. Any other additional works will have bonus point.