Pattern Recognition

**Milestone (1) report**

Movie Revenue Prediction

horizontal line

**Team Members**

# Abdelrahman Hassan Soliman

# 20191700336

# Department: CS

# Youssef Sherif Ahab

# 20191700773

# Department: CS

# Badr Mohamed Ashraf

# 20191700179

# Department: CS

# Seif Mokhtar Ibrahim

# 20191700310

# Department: CS

# Abdelrahman Omar Ibrahim

# 20191700343

# Department: CS

**Preprocessing**

1. Imputation:
   1. Missing data were filled in using TMDB API. These data are revenue, genres, directors, release dates and MPAA ratings.
2. Encoding:
   1. Rating: Three encoding types were used; Label encoding, One-hot encoding and Ordinal encoding. The encoder with the best results was the Ordinal encoder.
   2. Genre: Three encoding types were used; Label encoding, One-hot encoding and a modified Ordinal encoding, which was discarded due to the genres not having a logical order. The weight of each genre was calculated by finding the mean revenue of each genre.
   3. Directors: Two encoding types were used; Target encoding and a modified One-hot encoding. The best result was achieved by applying Target encoding.
3. Feature Selection:
   1. Dropped characters and voice actors columns due to insufficient data.
   2. Dropped movie title column due to not having a relation with revenue.
4. Data Cleaning:
   1. Dropped duplicate data.
   2. Dropped unfillable data; some movie titles were incomplete and there were a couple of tv series and tv episodes, therefore not all revenues and/or directors of some data could be found.
5. Feature Scaling:
   1. Applied min-max scaler.
6. Feature Extraction:
   1. Date column was expanded into a day, month and weekday columns.
   2. Created a “new movie” column based on the year of the release date; to indicate whether the movie was released later than 2005.

Regression Techniques:

1. Polynomial Regression: We tried degrees in the range of (2, 5). The best results yielded an MSE of ????
2. Ridge Regression: the model yielded an MSE of ????

Training time of each model was nearly instantaneous.

Used/Discarded Features

In the final model, the following features were handled accordingly:

| Used | Discarded |
| --- | --- |
| Directors (Target encoded) | Characters |
| Release date (Created five columns: day, month, year, weekday and new\_movie) | Voice actors |
| Rating (Ordinal encoded) | Movie title |
| Genre (Mean-weighted) |  |
| Genre (Ordinal encoded) |  |

Train-Test split: A distribution of 80%-20% was applied.