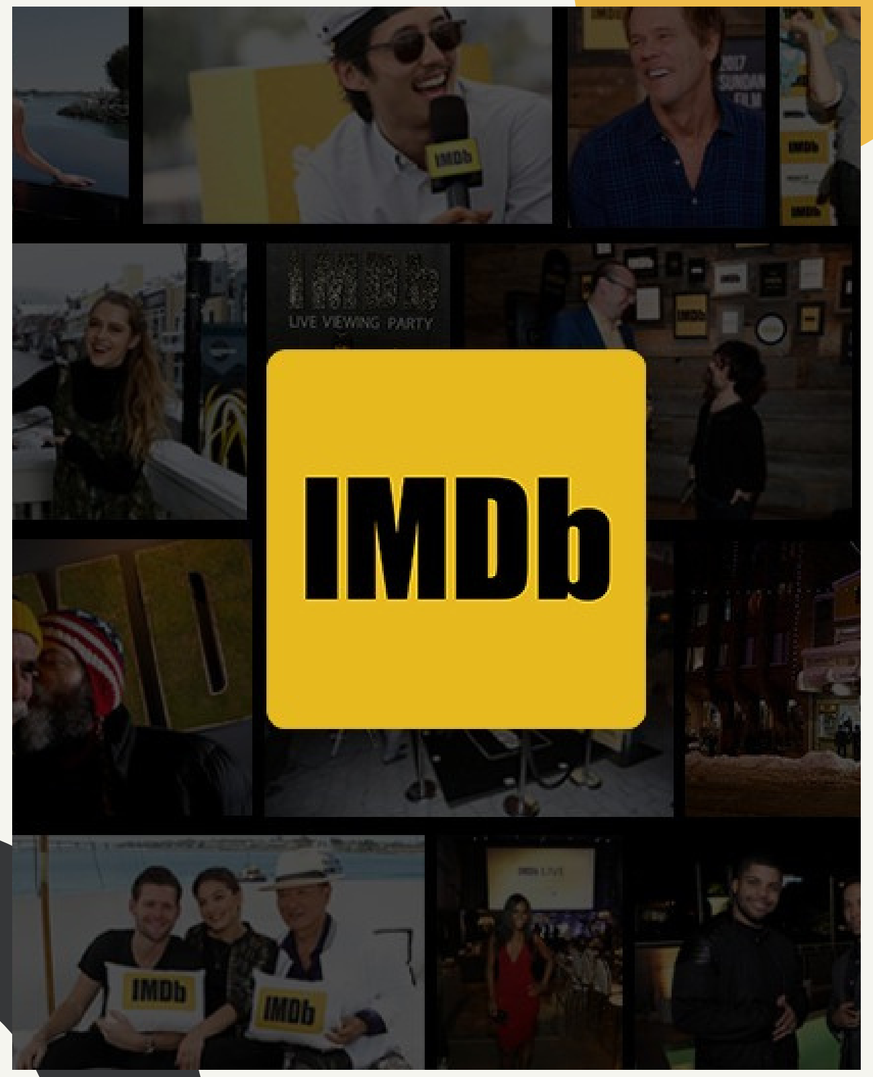
Predicting US Movies Gross

Using Web Scraping and Regression

Leena AlQasem Randa Almohammadi



INTRODUCTION

Business objective:

- Regression Approach
- Find the best ML model to predict US movie gross

Dataset:

Scarping IMDb website

3000 records and 9 features

METHODOLOGY

Gathering Data

Using web scraping

Exploratory Data Analysis

- Cleaning
- Analysis and visulasing

Data preparation and Regression

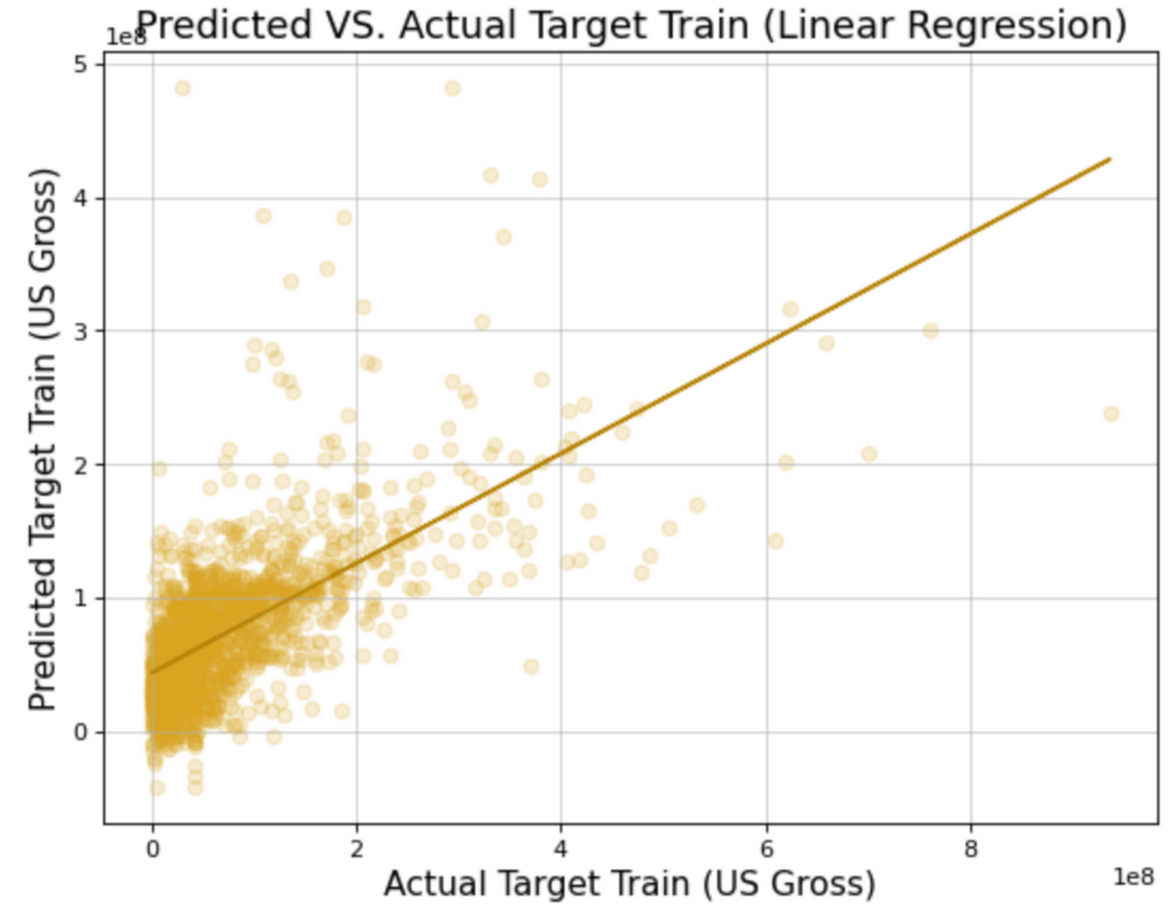
DATA PREPARATION

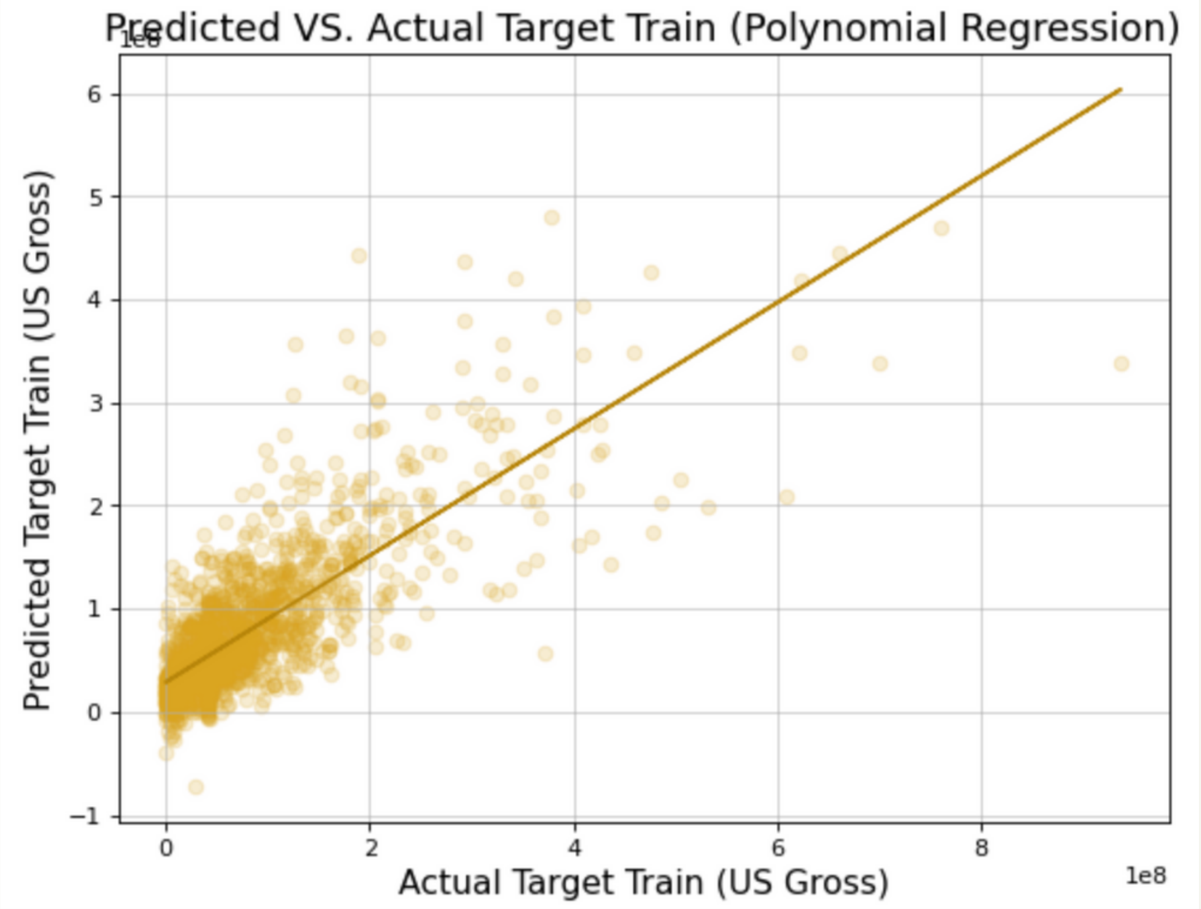
- **Feature Selection**
- **Feature Engineering**
    - Encoding (Dummy variables)
    - Substracting interaction terms
    - 2902 records × 20 features
- **Splitting Data**
    - 60% train, 20% validation, and 20% test
- **Regression models**

ANALYSIS AND RESULTS

Regression Models	Training Score	Validation Score
Normal Linear Regression	0.410840302	0.404234396
K-fold Linear Regression	0.410840303	0.37299705
Polynomial Regression Degree 2	**0.61368493**	**0.59439209**
Ridge Regression	0.41084005	0.40425610
Ridge Regression Cross-Validation	0.41066877	0.3736137
Lasso Regression Cross-Validation	0.410840303	0.372997054

ANALYSIS AND RESULTS





Testing Score: 0.6724790

CONCLUSION

Polynomial Regression shows the best prediction of "Gross."

Future Work:

- Scaling: Normalization and Standardization
- Other algorithms: Ex. Decision tree, Random forest

THANK YOU!

We hope you enjoy it!