Movies Lifetime Gross Prediction Report

Prepared by :

Lubna Almuammar

and Alanoud Alhwairany

Abstract

The goal is to Create linear regression model that can predict the lifetime gross of a movie Using information we scrape from the web, and the dataset we have from IMBD we can we build linear regression models using the features we have to predict the target which is the lifetime gross where the Gross refers to all earnings of a film from all revenue sources.

Design

- web scraping & data gathering
- eda & feature engineering
- model building & training

Data

we used BeautifulSoup to scrape movie data from box mojo and obtained 5 features for 1617 movies. The features include the Title, Rank, Lifetime Gross, Overall Rank and Year. And joined the scraped data with a dataset we took from Kaggle with the features Title, original_title, year, date_published, genre, duration, country, language, director, writer ,production_company, actors, description, avg_vote, votes ,budget.

Dataset source: https://www.kaggle.com/rounakbanik/the-movies-dataset

Website link: https://www.boxofficemojo.com/

Algorithm

- **Data Collection:**

  Collect data using web scraping and datasets.

- **Data Understanding:**

  Understand the columns and what they represent.

- **Data Cleaning:**

  Deleted the duplicated values.
  Delete null values.

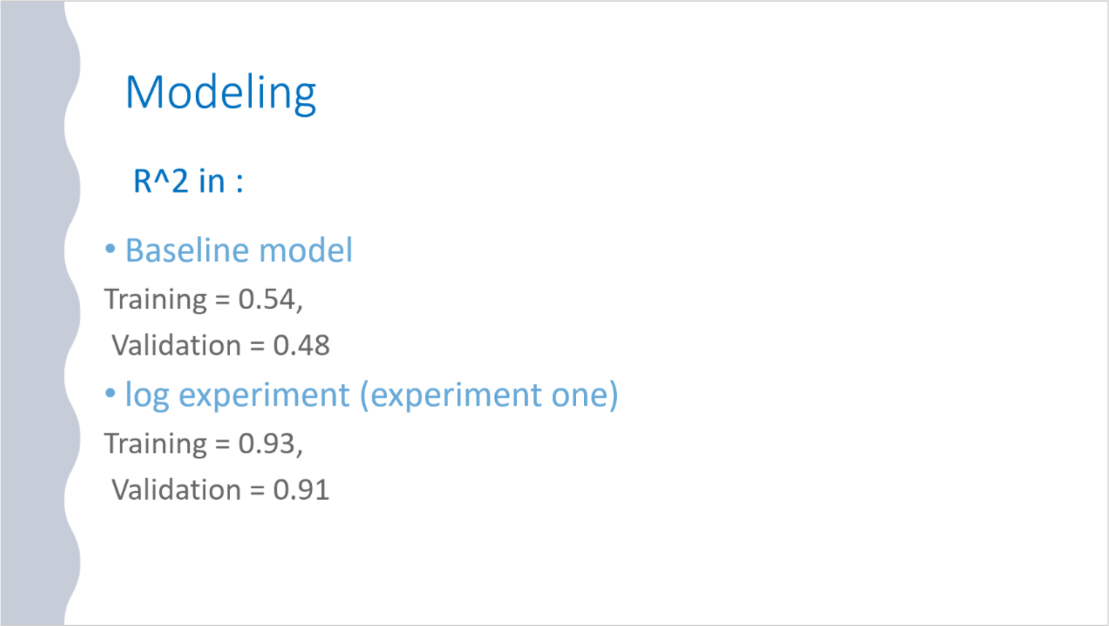
- **Feature Engineering**:
  turn the categorical values to numeric values

- **modeling and training:**
  create a linear regression model to predict the lifetime gross of a movie.

Tools

- technologies: Jupiter notebook.
- libraries: scikit-learn,Numpy, Pandas, matplotlib, BeautifulSoup

Communication



Modeling

R^2 in :

- Baseline model

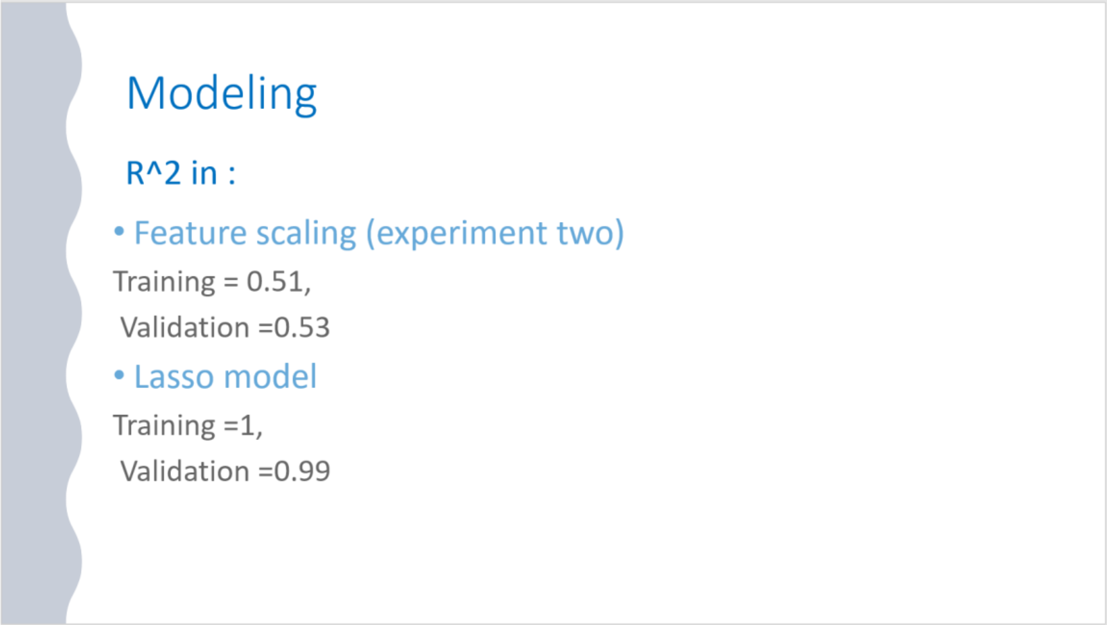
Training = 0.54,

Validation = 0.48

- log experiment (experiment one)

Training = 0.93,

Validation = 0.91



Modeling

R^2 in :

- Feature scaling (experiment two)

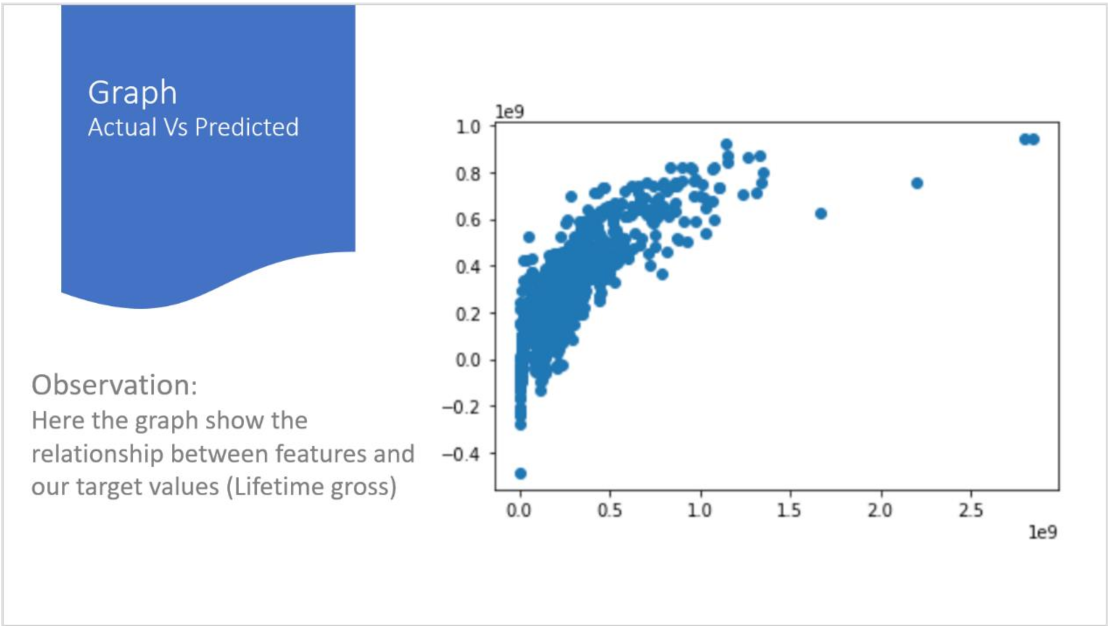
Training = 0.51,

Validation =0.53

- Lasso model

Training =1,

Validation =0.99



Graph

Actual Vs Predicted

Observation:

Here the graph show the relationship between features and our target values (Lifetime gross)