IMDb

Predicting US Movies Gross

USING WEB SCRAPING AND REGRESSION

-HIMANI KAUSHIK

Introduction

- Find the best ML Model to predict US movie gross earnings to decrease financial risk

Methodology

01 Data Scraping

- Scraping IMDb Website
- 3000 records and 9 features

02 Data Cleansing

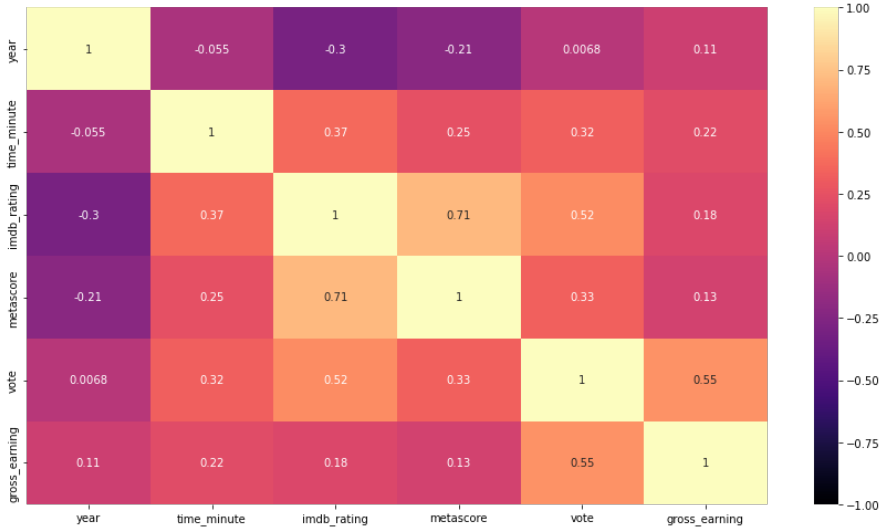
- Disintegrating some columns
- Detecting and handling missing data and outliers

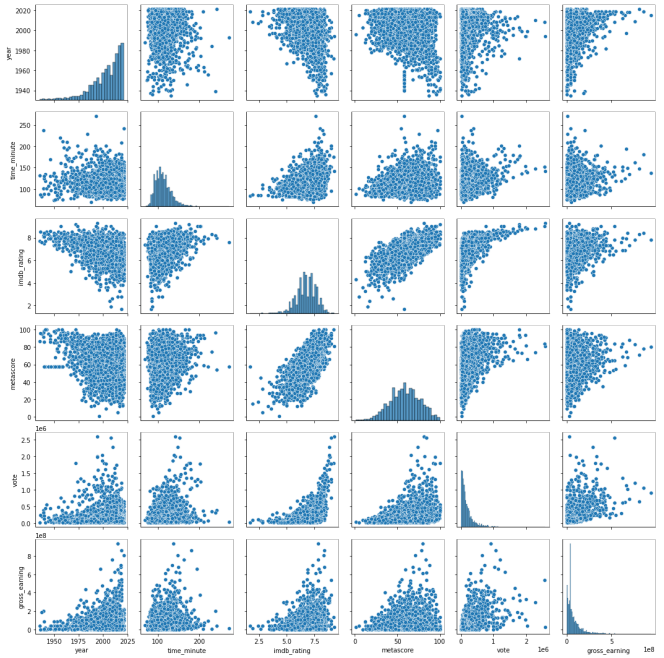
03 Exploratory Data Analysis

- Looking into feature correlations, pairplots, etc.
- Understanding the transformations needed to make data suitable for regression model

04 Building Models

- Creating different regression models.
- Using cross-validations, and using evaluation metrics to select the final model





Data Preparation

Feature Selection

Feature Enginering

- Encoding (Dummy variables)
- Subtracting interaction terms

Splitting Data

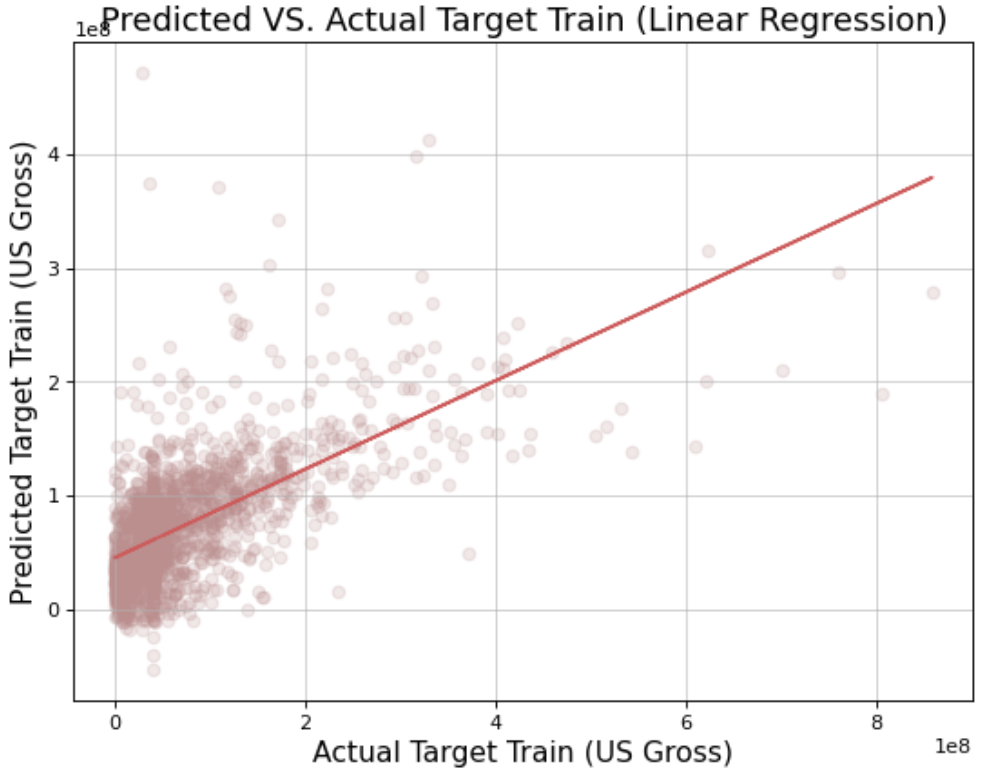
- 60% Train
- 20% Validation
- 20% Test

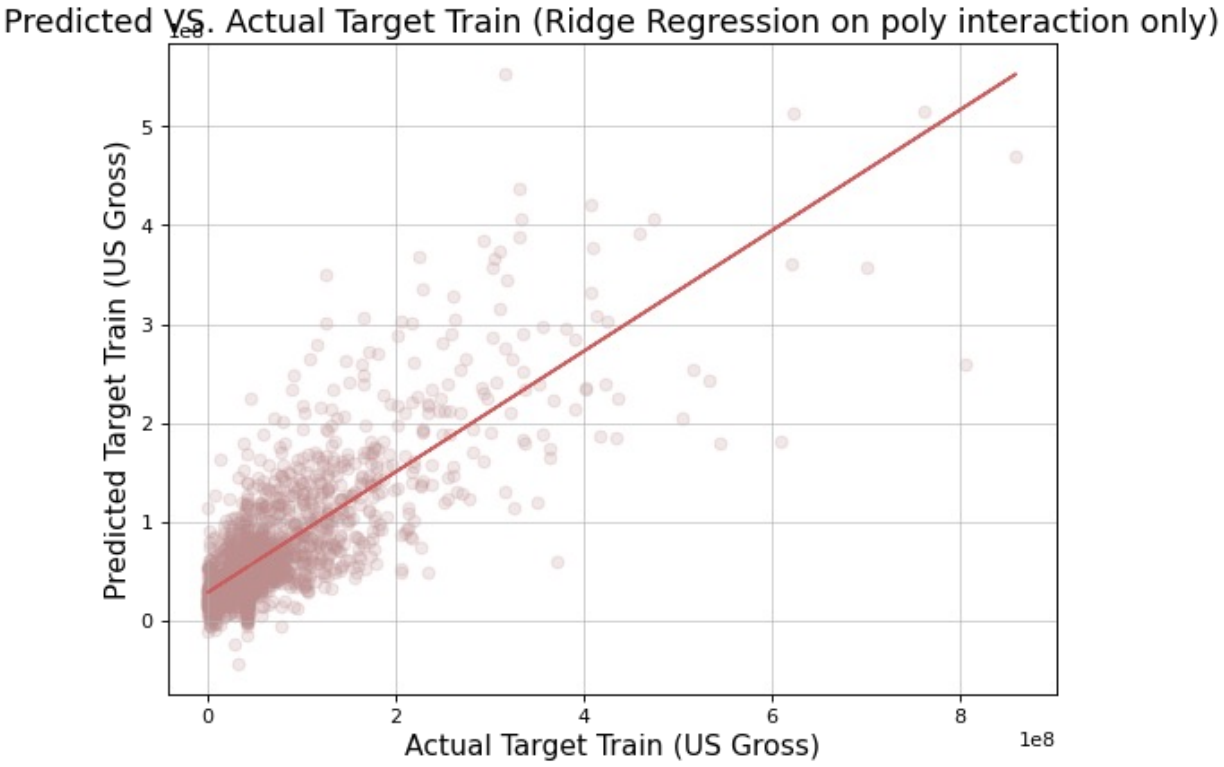
Regression Models

Analysis And Results

Regression Models	Training Score	Validation Score
Normal Linear Regression	0.389285939	0.461151975
K-fold Linear Regression	0.389285939	0.379048417
Polynomial Regression Degree 2	0.00471088	-0.07468319
Polynomial Regression Interaction	0.50838276	0.3800856
Ridge Regression	0.46113795	0.38928325
Ridge Regression Cross-Validation	0.389283247	0.379102316
Lasso Regression Cross-Validation	0.389285939	0.379048764
Ridge Regression on polynomial interaction only	**0.61451453**	**0.54744346**

Analysis And Results



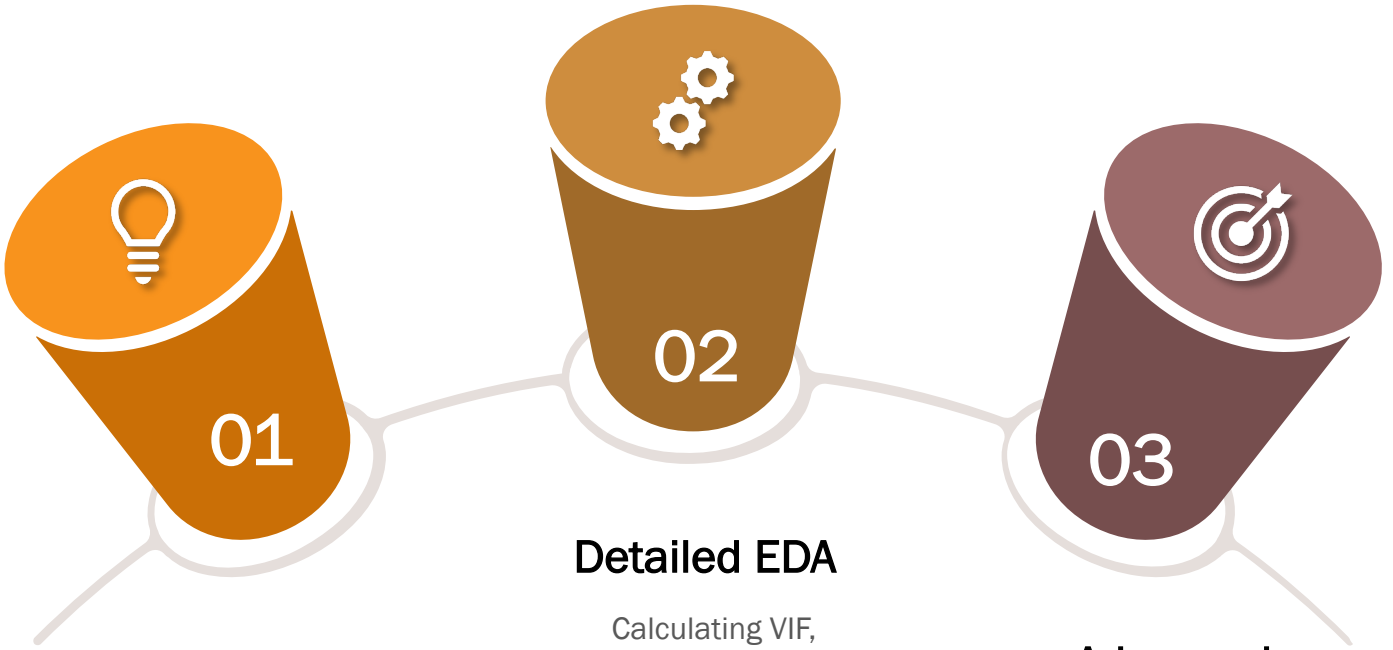


Testing Score: 0.59082610

Conclusion

- The objective of this project was to predict US movie gross earnings to decrease financial risk
- Thus, based on the evaluation metrics, Ridge Regression on polynomial interaction only features is the best model for prediction of gross earnings.

Future Work



More Data

Data from multiple sources

Detailed EDA

Calculating VIF, muticollinearity, etc

Advanced Models

Other alogorithms like gradient boost, XGBoost regressors, etc


THANK YOU

Appendix

Analysis And Results