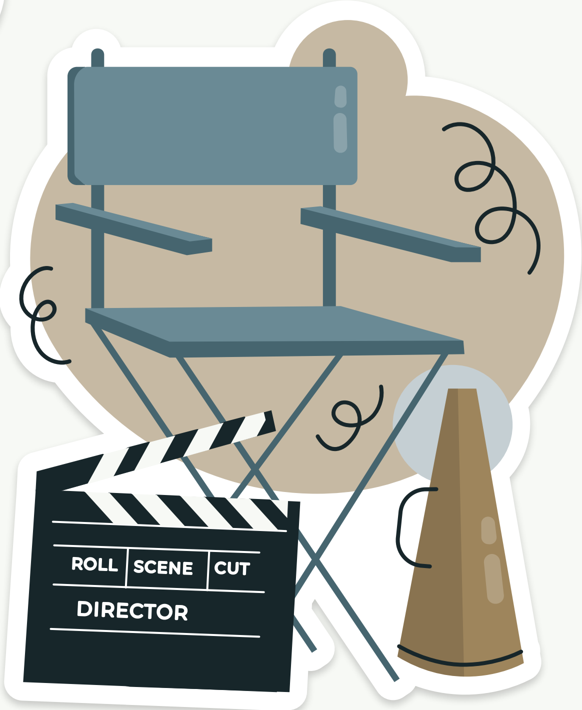
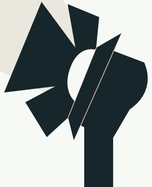
Movie Revenue Prediction

Ashley Lopez and Alice Liu



01

Introduction



Our questions are...

How can we predict movie revenue?
What factors impact a movie's revenue?

Data

Size

[Kaggle: Getting Started with a Movie Recommendation System]
~5800 rows
24 columns

Columns

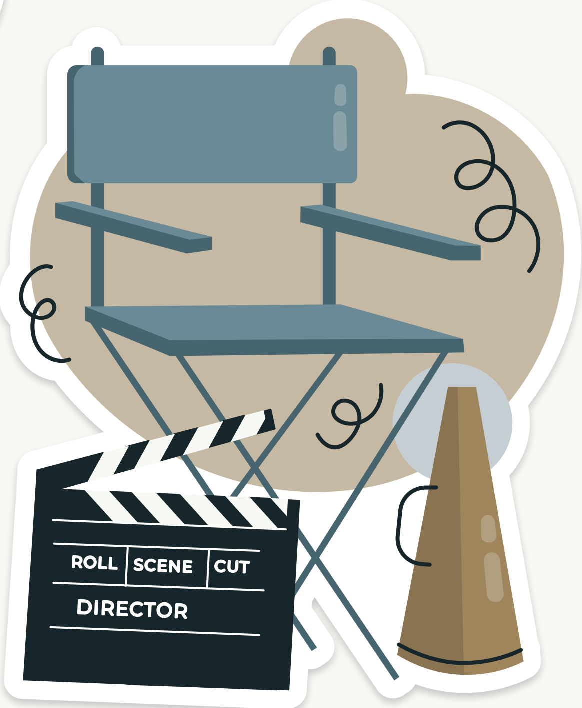
- budget
- genre
- original_language
- popularity
- runtime
- ...

Features

- adult (categorical)
- original_language (categorical)
- genres (categorical)
- budget (numerical)
- popularity (numerical)
- runtime (numerical)
- vote_average (numerical)
- revenue (target)

02

Methods



Data Preprocessing

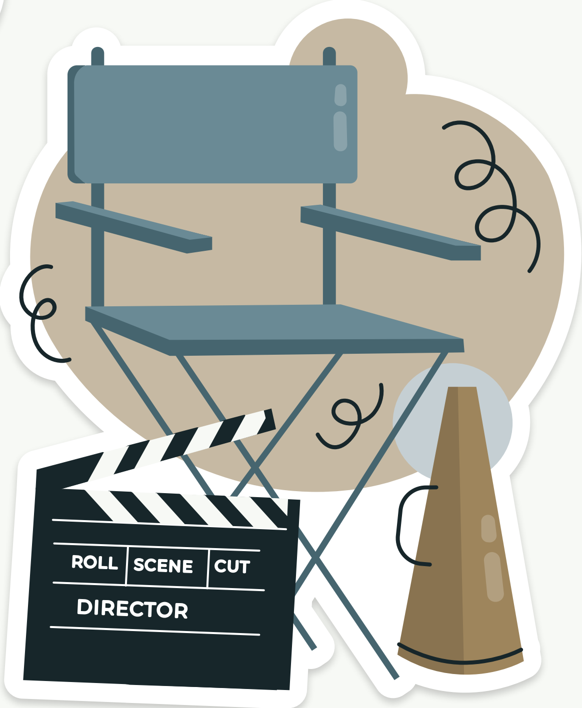
One-hot encoding categorical values	Retained the top five languages (English, Hindi, French, Russian, and Japanese) and categorized all others under “other”, due to high cardinality of the *original_language* feature (40 unique values) Utilized MultiLabelBinarizer to encode the *genres* feature which transformed the genre lists into a binary matrix 31 columns after encoding
Scaling data	StandardScaler
Removing the adult feature	Removed *adult* feature due to insufficient variations in the values
Removing invalid data	Removed NaN and outlier data

Different Models

- Linear Regression
- Regularized Linear Regression (Lasso & Ridge)
- Polynomial Regression (degree 2, 3, 4)

Model Evaluation

- Mean Squared Error, R^2
- Cross validation (RidgeCV, LassoCV, etc.) to select the best hyperparameters
- Visualization: Scatter plots and convergence plots



03

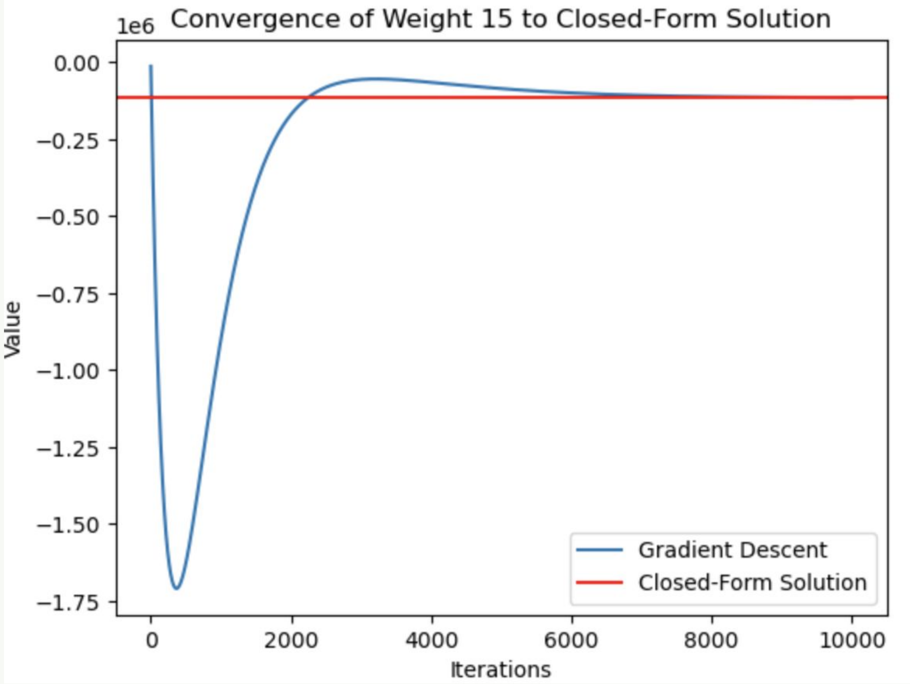
Results

From Linear Regression

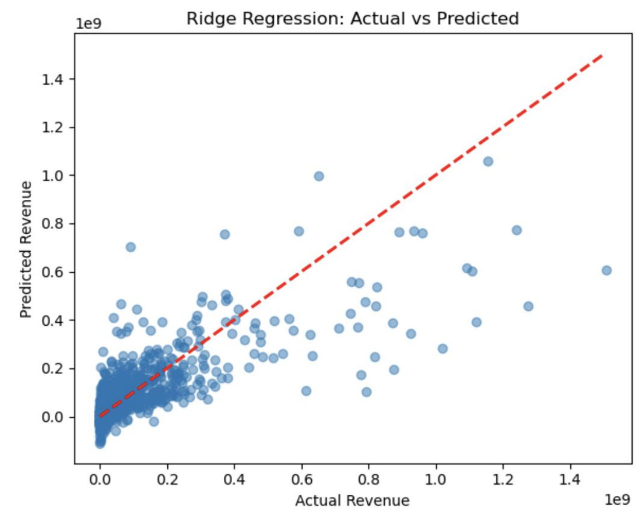
The gradient descent method and closed-form solutions show similar results.

Linear regression effectively captures a significant portion of the variance in the revenue data.

- **MSE:** 1.073×10^{16}
- **R^2 Score:** 0.6126



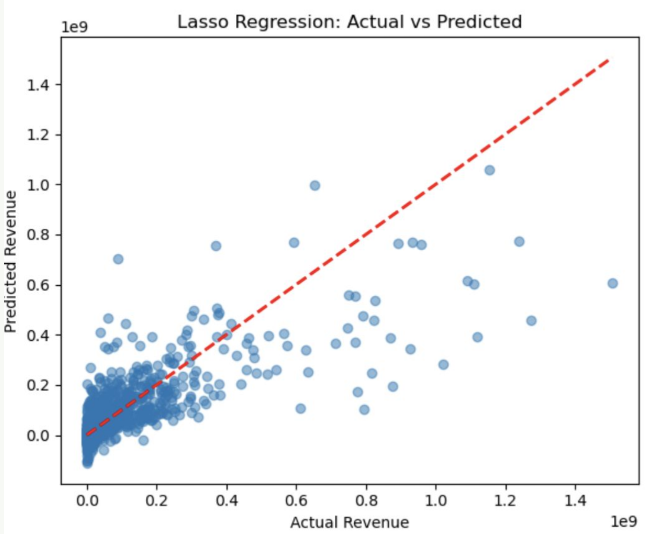
From Lasso & Ridge Models



- **MSE:** 1.165×10^{16}
- **R^2 Score:** 0.5848

Ridge and Lasso slightly underperform compared to linear regression. These models help mitigate overfitting by penalizing large coefficients.

The importance of features like **budget**, **popularity**, and **vote average** is consistent across both models.



From Polynomial Regression

Best Degree (2):

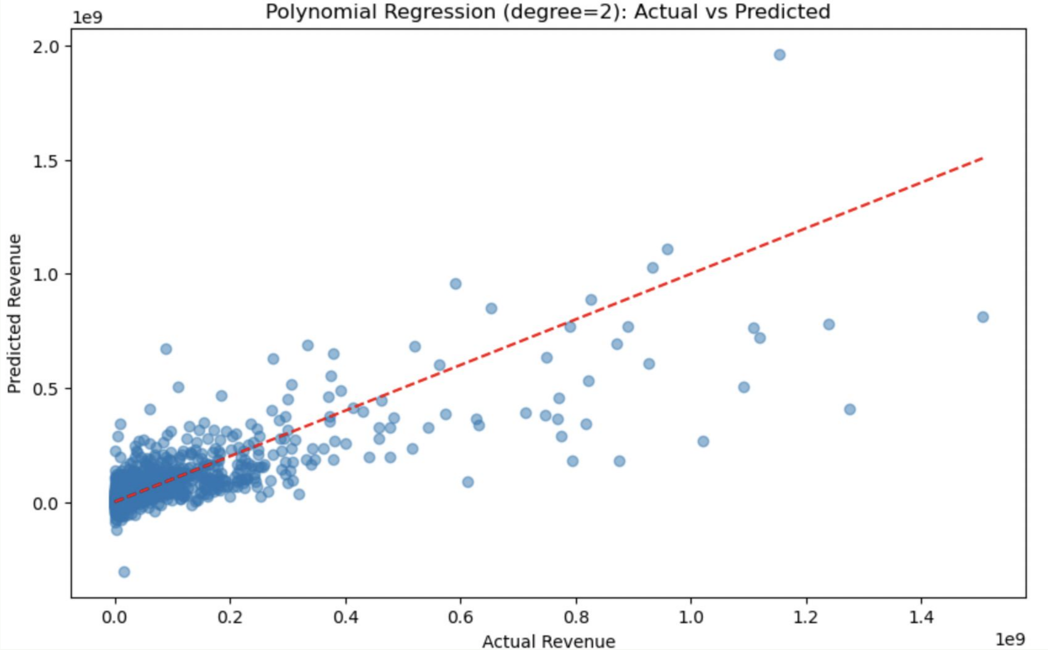
- **R^2 Score:** 0.6121 (highest performing model in polynomial regression models)

Higher Degrees (3 & 4):

- **R^2 Scores:** Negative (indicating overfitting)

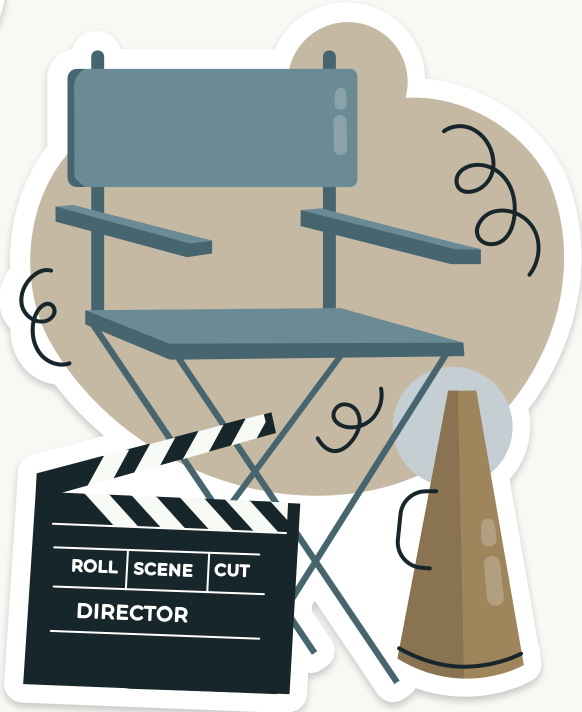
Polynomial regression with degree 2 captures the non-linear relationships between features and revenue effectively.

Higher-degree polynomials (3 & 4) suffer from overfitting, as seen in the negative R^2 scores.



Comparison of Models

Model	MSE	R^2 Score
Linear Regression	1.073×10^{16}	0.6126
Ridge Regression	1.165×10^{16}	0.5848
Lasso Regression	1.165×10^{16}	0.5848
Polynomial (Degree 2)	1.088×10^{16}	0.6121
Polynomial (Degree 3)	-	-35.6270
Polynomial (Degree 4)	-	-147066.7



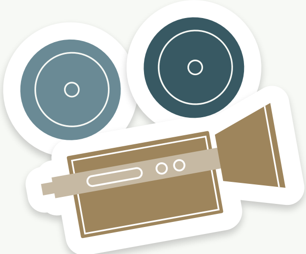
04

Conclusions

Key Findings

- Linear Regression provided the most accurate predictions, outperforming polynomial and regularized regression models by bit.
- Ridge and Lasso regression helped reduce overfitting but were slightly less predictive than polynomial regression.
- Higher-degree polynomials (3, 4) resulted in overfitting, causing performance degradation.

Insights and Implications



Feature Importance: Budget, popularity, and vote average are significant predictors of revenue, as seen in all models.

Top 10 Features based on Absolute Ridge and Lasso Coefficients:

	Feature	Ridge_Abs	Lasso_Abs
0	budget	1.064755e+08	1.090277e+08
1	popularity	3.451692e+07	3.453333e+07
3	vote_average	2.469727e+07	2.538111e+07
16	Drama	7.097812e+06	7.096639e+06
21	Horror	4.726447e+06	5.078946e+06
24	Romance	4.617406e+06	4.881683e+06
11	Adventure	4.324433e+06	4.161654e+06
20	History	4.003916e+06	4.036695e+06
29	Western	3.939193e+06	3.940520e+06
27	Thriller	3.772481e+06	3.984897e+06

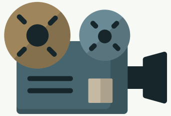
Top 10 Features based on Absolute Polynomial Coefficients:

	Feature	Poly_Coef	Poly_Abs
1	budget	8.693650e+07	8.693650e+07
2	popularity	5.011849e+07	5.011849e+07
34	budget vote_average	3.805600e+07	3.805600e+07
4	vote_average	3.099525e+07	3.099525e+07
327	Animation Crime	-2.082980e+07	2.082980e+07
77	popularity Family	1.935112e+07	1.935112e+07
78	popularity Fantasy	1.868019e+07	1.868019e+07
363	Crime Family	1.579712e+07	1.579712e+07
37	budget original_language_hi	1.329187e+07	1.329187e+07
364	Crime Fantasy	1.270900e+07	1.270900e+07

Linearity: The relationship between features and revenue is linear, which is why linear regression performed better.

Overfitting Issues: While the polynomial models explain certain patterns, the right model complexity is crucial for the best results and preventing overfitting.

Ideas for Future Work



Feature Engineering

Further exploration of feature selection and engineering could improve predictive accuracy.

Model Complexity

Looking into alternative models, such as ensemble methods, may help balance model complexity and accuracy.

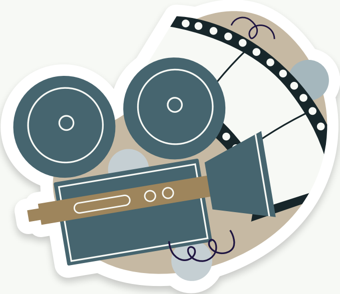
Final Thoughts

Impact

Accurate movie revenue predictions can significantly help in decision-making in the film industry, from production budgeting to marketing strategies

Takeaway

Linear regression appears to be the most effective approach based on our analysis, though further refinements and model testing could help improve our findings.



Thank you!