ML Final Project

Group SZA
Zachary Zambrana, Savannah Day, Amer Amer, Dhivya Venkatachalam

Problem Statement & Business Context

The Challenge

Movie studios invest millions in production before knowing if a film will succeed financially.

Our Solution

Predict box office revenue using machine learning to inform:

- **Budget allocation decisions**
- **Marketing strategy investments**
- **Project greenlight choices**
- **International distribution planning**

Industry Impact

- Average film budget: $65M (production) + $35M (marketing)
- Box office failure rate: ~80% of films don't break even
- **Our models help reduce financial risk**

Methodology

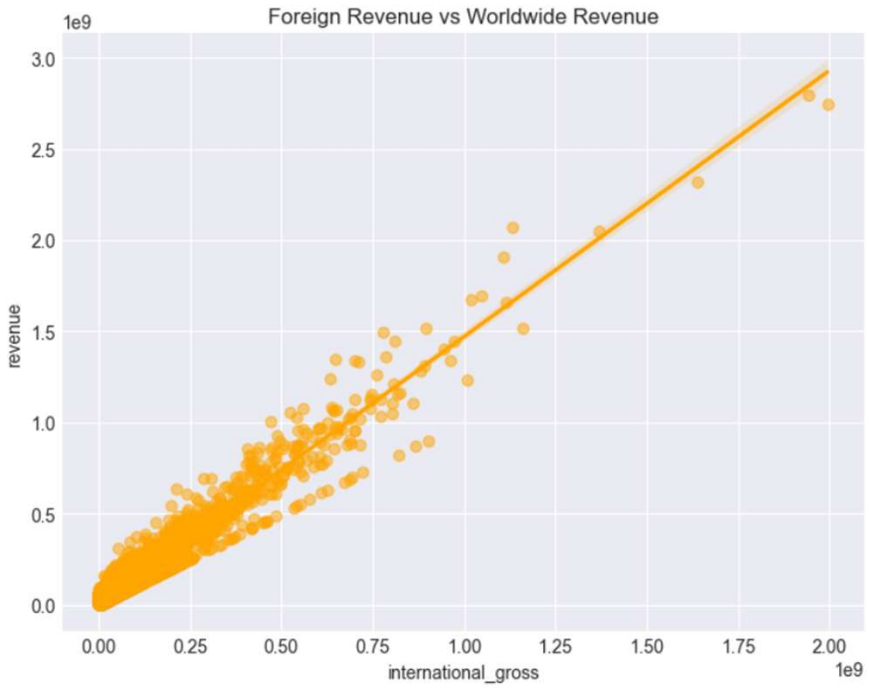
- Retrieved and cleaned Kaggle movie data to ensure consistent revenue, genre, and rating values.
- Analyzed patterns that influence box office success.
- Used feature engineering to encode categorical data and scale numerical features.
- Trained and compared Linear Regression, Decision Trees, and KNN models for profitability and rating predictions.
- Evaluated models using metrics like accuracy, recall, and F1-score.
- Used results to guide data-driven decisions for film studios’ next steps.

Key Findings

- Most movies earn very little; revenue is heavily skewed.
- Domestic and international revenue both strongly correlate with worldwide revenue.
- Ratings correlate with revenue but are weaker predictors.
- KNN model performed well ($R^2 \approx 0.91$), though outliers reduced accuracy.
- Using too many neighbors in KNN decreases performance.
- Classification model had ~64% accuracy and struggled to predict “Liked” movies.
- vote_count was the most important feature in predicting movie success.

Model Comparisons

Linear Regression

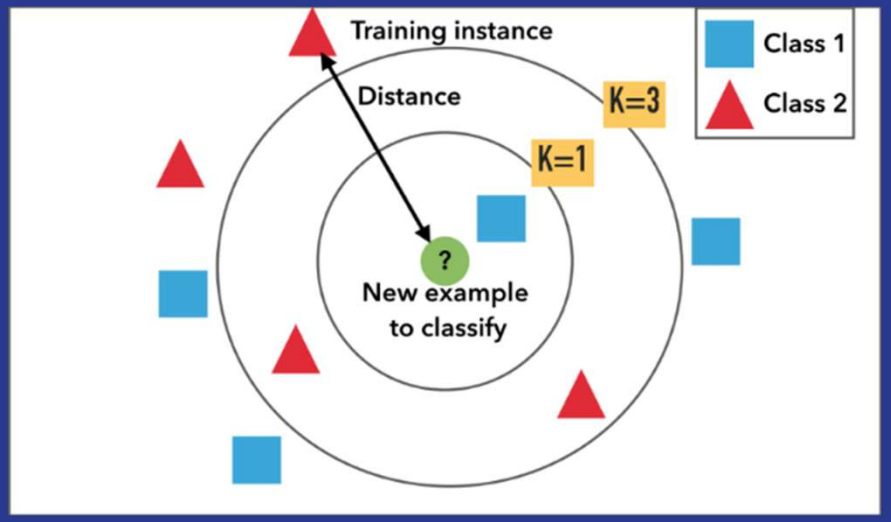


Summary of Findings

- Domestic and foreign revenue strongly predict worldwide revenue.
- Ratings have a smaller impact, helpful but not strong predictors.
- Revenue varies by year, showing the importance of release timing.
- High-earning genres: Adventure, Action, Animation, Sci-Fi.
- Lower-earning genres: Documentary, Western, Romance.
- Ratings matter more for some genres (Music, History) than others.
- Simple regressions are limited, advanced models can capture more factors and improve predictions.

KNN

Regression Based



Why use KNN?

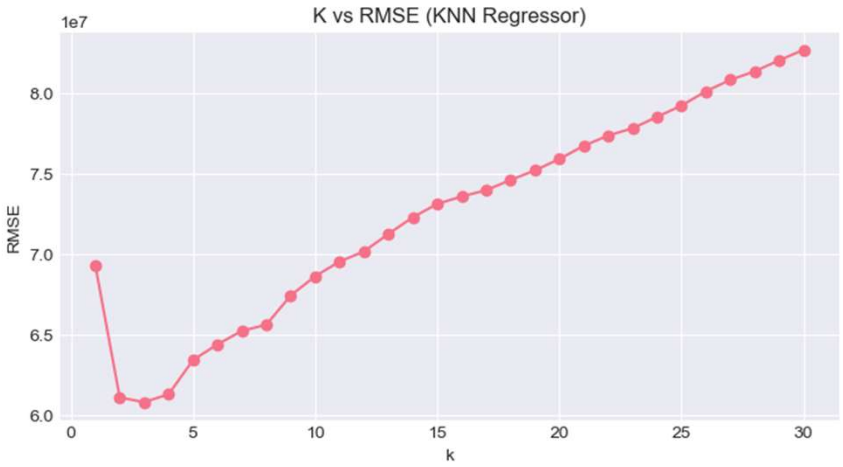
Captures complex patterns: It can model nonlinear relationships between movie traits (genre, rating, release year) and revenue.

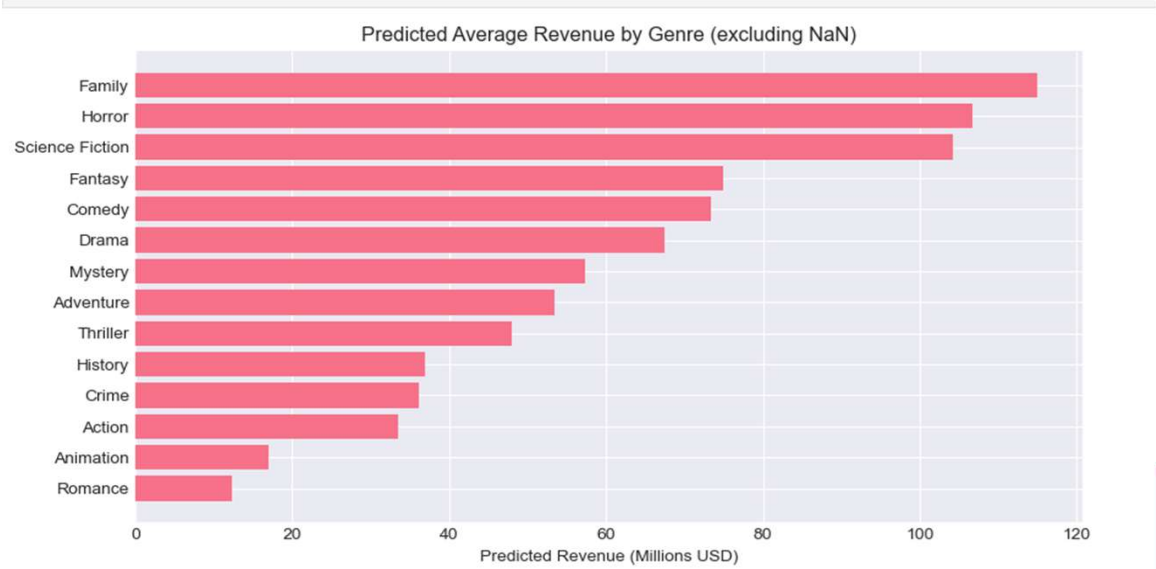
Easy to interpret: predictions are based on the average revenue of the closest “similar” movies.

Non-parametric model: makes no assumptions about linearity between genre features and revenue.

Works well with messy or noisy data: smooths predictions by averaging over nearest neighbors.

Here's The Results



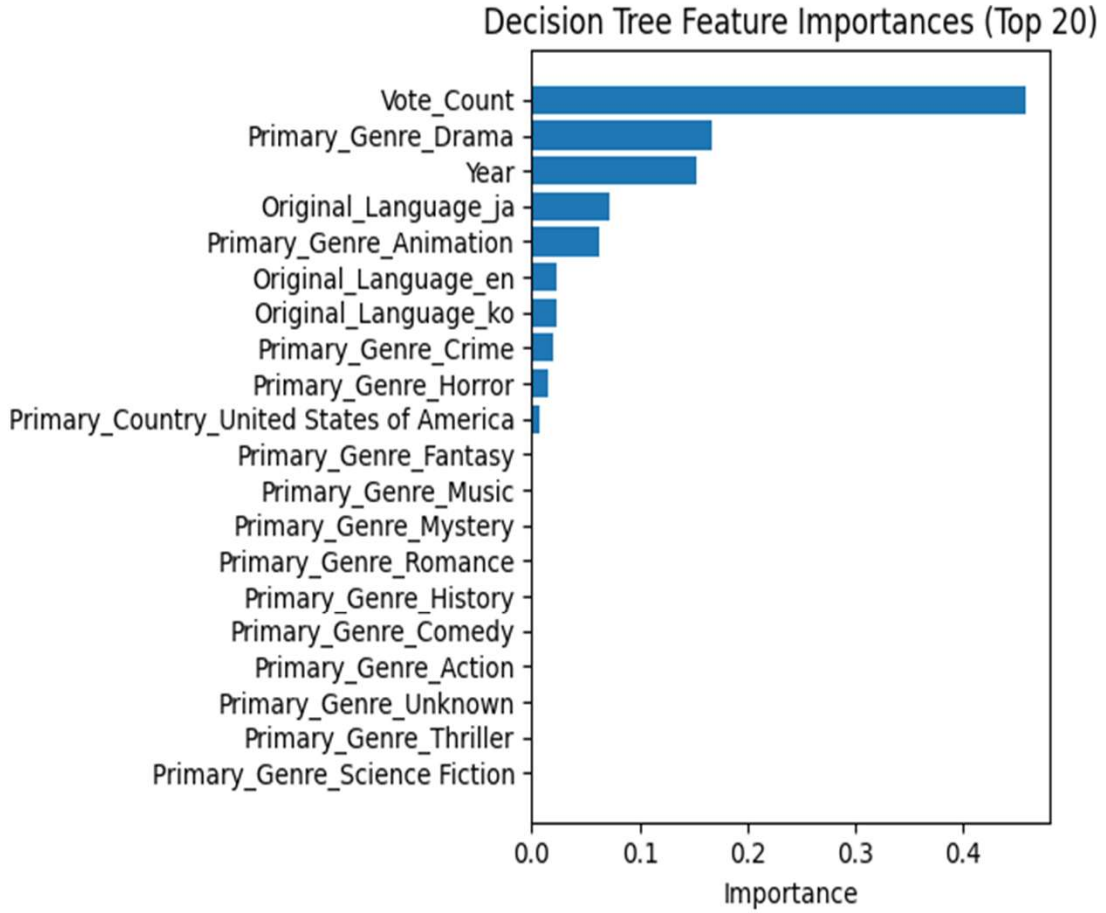


Choose Decision Tree Because…

- Genre, budget, rating, production, and country predict movie success, allowing budgeting.
- Patterns of high-budget action/adventure films and english-language productions performed best, while low-budget dramas underperformed.
- Revealed interactions in high budget + action genre, mid-budget + PG/PG-13 rating, non-US films needing higher budgets.
- Modeled complex interactions like budget × genre and budget × rating
- Less preprocessing and immediate business rules like budget > $70M AND Action => high success likelihood.
- Recall and F1, generalization, and actionable feature importance rankings like budget, genre, and rating
    - Studio makes better decisions

Results

```
Sample of Tree Rules For The First 5 Levels:
|--- Vote_Count <= 7097.00
    |-=-- Primary_Genre_Drama < 0.50
    |   |-=- Year < 2016.50
    |   |   |-=- Original_Language_ja <= 0.50
    |   |   |   |-=- class: 0
    |   |   |-=- Primary_Genre_Crime > 0 < 0.50
    |   |       |-=- class: 0
    |   |-=- Primary_Genre Anmation <= 0.50
    |-=-- Vote_Count > 305.00
  |-=- Vote_Count > Drama > 0.50
  |   |-=- Vote_Count |< 2467.00
  |   |   |-=- Primary_Gontte gn 2021.01 !.050
  |   |   |   |-=- Year < 2021.50 | 0.50
  |   |   |   |   |Original_Languageen en <= 1 050
  |   |   |   |-=- class: 0
  |   |   |-=- Vote_Count > 2021.50
  |   |-=- Vote_Count > 2467.00
  |-=- Vote_Count > Vote Count= 0.50
  |   |-=- Primary_Genre_Drama > 0.50
  |   |   |-=- Vote_Count | < 2467.00 | 167.00
  |   |   |   |-=- Original_Contrry United Staesica ≤ 50
  |   |   |   |   |-=- class: 1
  |   |   |   |-=- class: 1
  |   |-=- Vote_Count > 0 2467.00
  |-=- Vote_Count > 7097.00
  |   |-=- Primary Country United States of America ≤ 50
  |   |   |-=- Vote_Count > 9057.50
  |   |   |   |-=- class: 0
  |   |-=- Vote_Count > 10184.50
```



Performance Evaluation and Metrics

WHY THESE METRICS?

R^2 (R-Squared) - Coefficient of Determination

What it measures: % of revenue variance explained by the model (0-1 scale)
Business Insights: "How much of box office variability can we actually predict?" Higher R^2 = more reliable forecasts for strategic planning

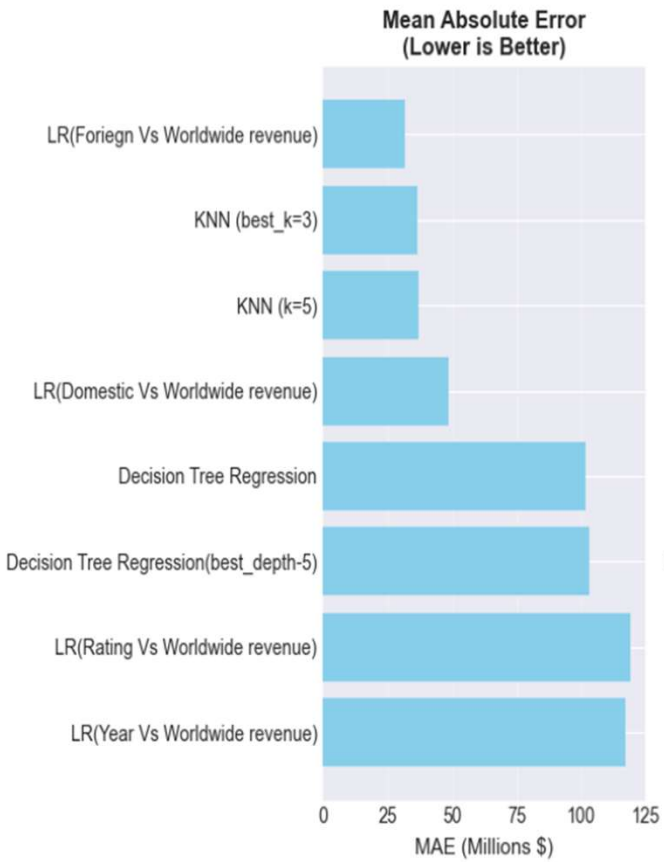
MAE (Mean Absolute Error)

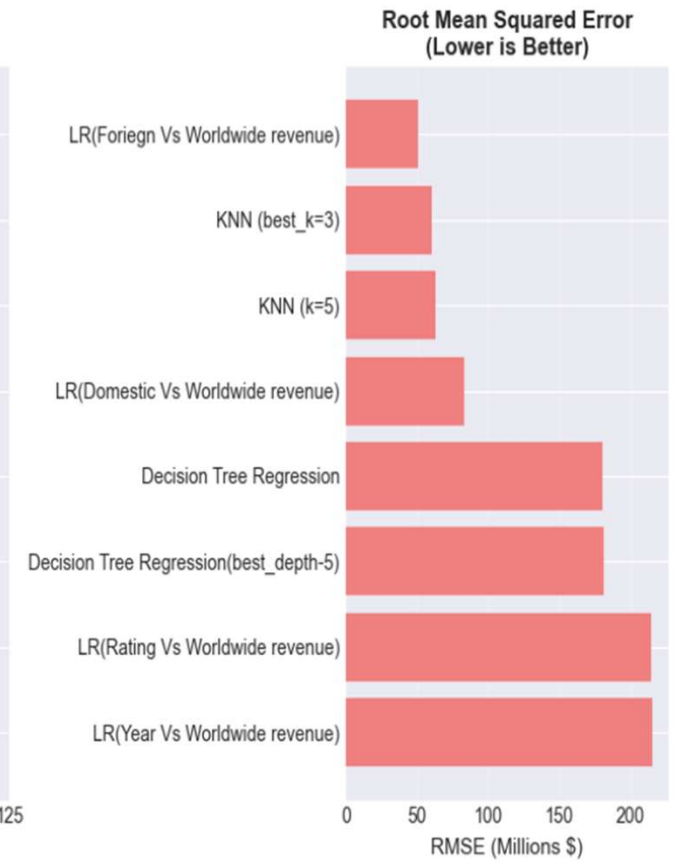
What it measures: Average prediction error in dollars
Business Insights: "On average, how far off are our predictions?" Easy to understand: "We're typically off by $X million"

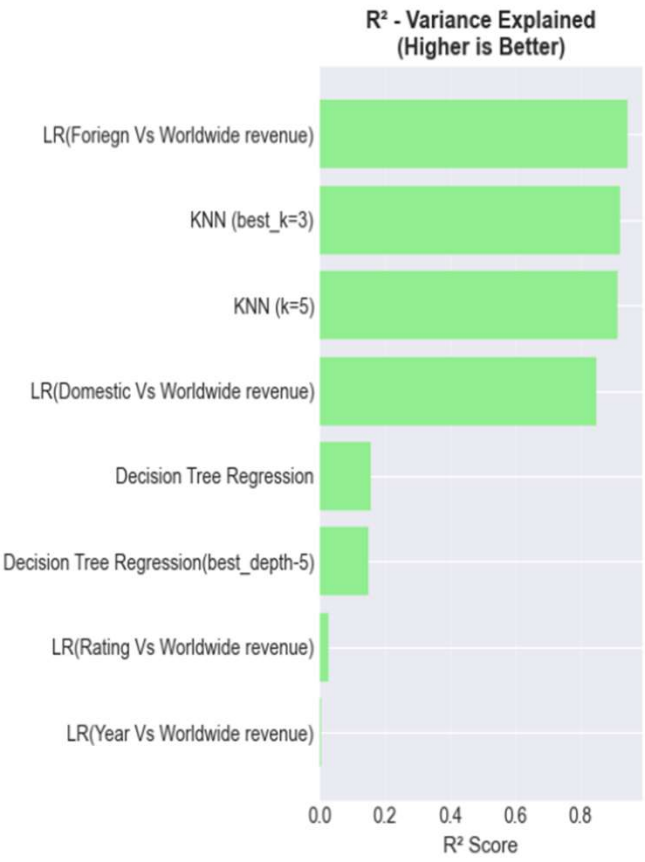
RMSE (Root Mean Squared Error)

What it measures: Prediction error that heavily penalizes large mistakes
Business Insights: "How much financial risk from outliers?" Critical for identifying potentially disastrous predictions

Performance Evaluation and Metrics - Continued…







Business Insights and Implications

Strategic Recommendations for Studios

1. Production Strategy

- **Prioritize high-ROI genres:** Family ($115M), Horror ($107M), Sci-Fi ($104M)
- **Use KNN comps** before greenlighting (find 3-5 similar films)
- **Invest in pre-release buzz:** Vote count drives 77% of predictions

2. Marketing Focus

- **International markets critical:** Foreign revenue = 94.4% R^2 predictor
- **Allocate 60%+ budget** to international distribution
- **Don't rely on ratings:** Only 2.5% predictive power

3. Financial Planning

- **Budget Formula:** Max Budget = (Predicted Revenue - RMSE) × 0.4
- **Risk Tiers:**
  - Low: Revenue > 3× budget
  - Medium: Revenue = 2-3× budget
  - High: Revenue < 2× budget

4. Model Usage by Stage

- **Pre-production:** KNN for comp analysis (92% R^2)
- **Post-opening:** Linear Regression for tracking (94% R^2)
- **Board meetings:** Decision Tree for interpretability

🎯 Key Executive Takeaways

1. International markets drive success → Prioritize global distribution
2. Pre-release buzz predicts revenue → Invest in social media
3. Genre matters more than quality → Choose commercial genres
4. Use right model at right time → Stage-specific tools
5. Manage expectations → Works best for typical films, not mega-hits

Limitations and Next Steps

⚠️ Current Limitations

1. Data Constraints

- Selection bias: Only top 200 films/year (missing failures and indie films)
- Revenue estimates, not exact figures
- No production budget or marketing spend data

2. Model Weaknesses

- Underestimates mega-hits >$1B (insufficient training examples)
- Best model requires opening weekend data (limited pre-production utility)
- Missing features: star power, director reputation, competition

3. External Factors Not Modeled

- Competitive releases same weekend
- Economic conditions (recession, pandemic)
- Streaming platform impact

🚀 Recommended Next Steps

Phase 1: Immediate (0-3 months)

- Add production budget and marketing spend features
- Include social media metrics (Twitter sentiment, trailer views)
- Implement ensemble modeling (KNN + LR combined)

Phase 2: Advanced (3-6 months)

- Deep learning for complex interactions
- External data integration (Google Trends, IMDb Pro)
- Regional models (domestic, China, Europe)

Phase 3: Deployment (6-12 months)

- Real-time prediction dashboard
- "What-if" scenario planning tools
- Automated retraining pipeline

Thank You!