CSE 6242 A
Team 20
Aditya Vikram, Kshitij Pathania, Shrestha Mishra, Sultan Syed

CineSage : A Predictive and Visualization Tool for Filmmakers

Problem

Production houses allocate significant time, energy and financial resources to a movie's preproduction phase like:

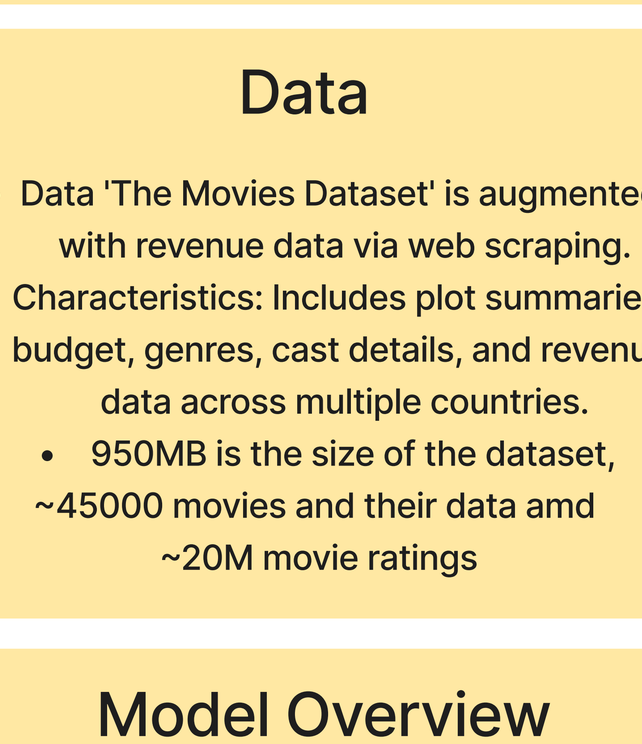
- Budget Planning
- Set Design
- Casting
- Script Analysis

Why Care?
Production houses make decisions based on intuition, past experiences and subjectivity, thus overshadowing data driven decision making. Our approach leverages ML accurate data driven revenue prediction model.

Data

- Data 'The Movies Dataset' is augmented with revenue data via web scraping.
- Characteristics: Includes plot summaries, budget, genres, cast details, and revenue data across multiple countries.
- 950MB is the size of the dataset, ~45000 movies and their data amd ~20M movie ratings

Model Overview



Approach

What are they?
Advanced data-driven visualization suite using BERT for text analysis and neural networks for revenue prediction leveraged by clustering techniques to categorize movies reviewers.

How do they work?
BERT model processes plot summaries; neural networks analyze budget, cast, and genre data for revenue prediction. Clustering techniques uses movie critics ratings for various movies genre and use them to categorize based on their rating trends.

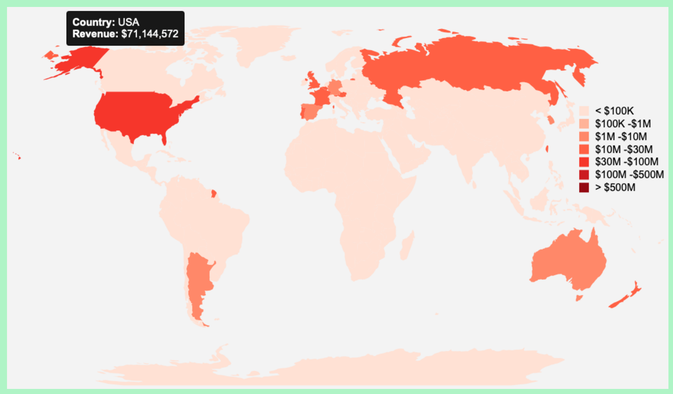
Why effective?
Combines linguistic analysis with statistical modeling for accurate, comprehensive insights.

What is new?
Real-time predictive choropleth map and audience preference clustering using t-SNE visualization

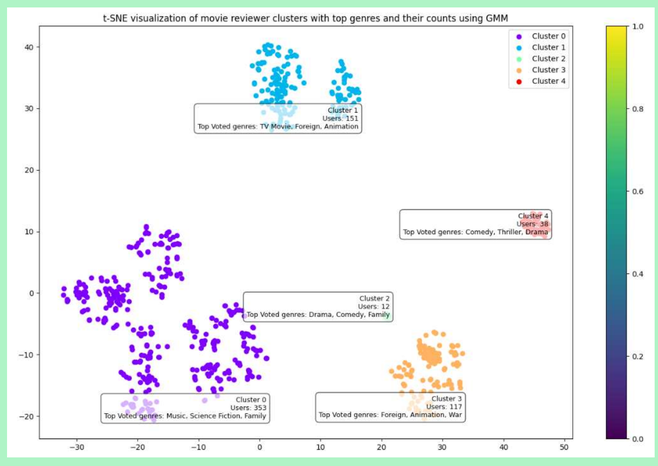
Experiments, Results and Comparison

Predictive Revenue Visualization
Our model projects revenue distribution across countries using BERT-generated plot embeddings, genre vectors, and language data. The architecture includes Linear layers and ReLU activation, optimized for each country's data

Choropleth map with Revenue Predictions



t-SNE plot showing clusters



Visualizing Clusters Using t-SNE
t-SNE embedding visualizes movie critics clusters in two dimensions. Node size indicates cluster size, and distance represents similarity in preferences. Interacting with the nodes reveals detailed statistics about preferred genres among reviewers and target section for filmmakers.

Experimental Evaluation/ Comparison
We evaluated our tools for effectiveness, accuracy, and usability in the filmmaking process. This involved testing the revenue prediction model and audience clustering visualization.

Loss curves for different hyperparameter combinations