Analyzing Movies to Predict their Commercial Viability for Producers

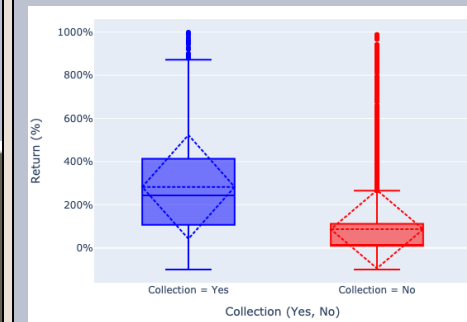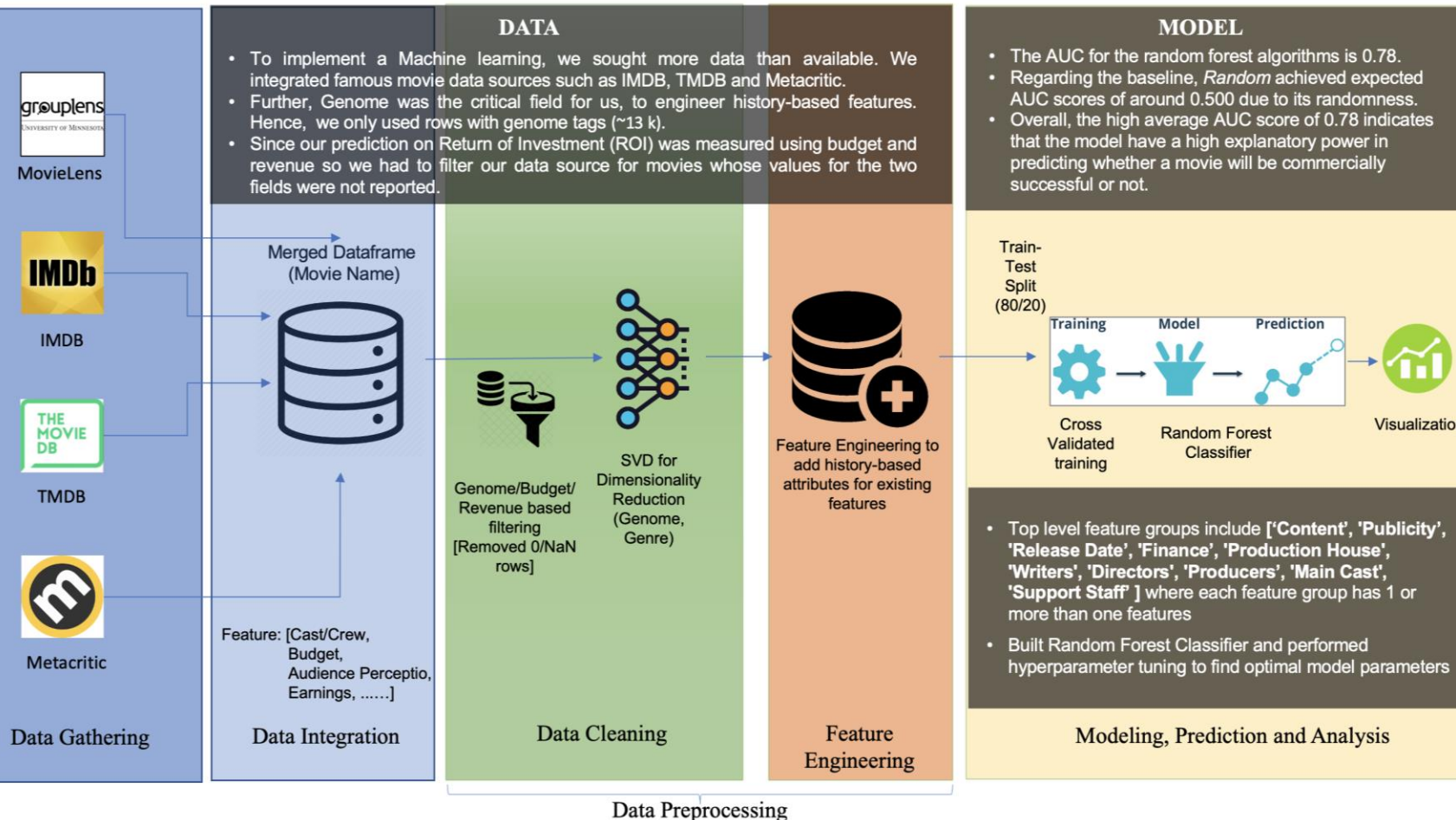Devendra Swami | Yash Phogat | Aadiraj Batlaw | Ashwin Goyal

Highlights

- ❖ A Machine Learning model predicting Return on Investment (ROI) slab for an upcoming movie based on its pre-release attributes.
- ❖ The study provides a set of individual and groups of features, that contribute majorly to a production's success.
- ❖ The features considered in this study may have potential application in solving other movie industry problems

Background

Producing major films costing $100M+, needs an estimate of expected positive cashflow in returns, for investors to make the decisions. Understanding features contributing towards commercial success of a production can help make data driven decisions.

Architecture

DATA

- To implement a Machine learning, we sought more data than available. We integrated famous movie data sources such as IMDB, TMDB and Metacritic.
- Further, Genome was the critical field for us, to engineer history-based features. Hence, we only used rows with genome tags (~13 k).
- Since our prediction on Return of Investment (ROI) was measured using budget and revenue so we had to filter our data source for movies whose values for the two fields were not reported.

MODEL

- The AUC for the random forest algorithms is 0.78.
- Regarding the baseline, *Random* achieved expected AUC scores of around 0.500 due to its randomness.
- Overall, the high average AUC score of 0.78 indicates that the model have a high explanatory power in predicting whether a movie will be commercially successful or not.

Data Gathering — MovieLens (grouplens, University of Minnesota), IMDB, TMDB (THE MOVIE DB), Metacritic

Data Integration — Merged Dataframe (Movie Name). Feature: [Cast/Crew, Budget, Audience Perceptio, Earnings, ...]

Data Cleaning — Genome/Budget/Revenue based filtering [Removed 0/NaN rows]. SVD for Dimensionality Reduction (Genome, Genre)

Feature Engineering — Feature Engineering to add history-based attributes for existing features

Modeling, Prediction and Analysis — Train-Test Split (80/20). Training, Model, Prediction. Cross Validated training. Random Forest Classifier. Visualization

Data Preprocessing

- Top level feature groups include ['Content', 'Publicity', 'Release Date', 'Finance', 'Production House', 'Writers', 'Directors', 'Producers', 'Main Cast', 'Support Staff'] where each feature group has 1 or more than one features
- Built Random Forest Classifier and performed hyperparameter tuning to find optimal model parameters

Collection (Is the movie part of a series) is the most important feature identified for prediction. This curve demonstrates that productions that are part of a series have seen higher ROI on average.

Movie per month, demonstrates the competition for a movie. As expected, when there is less competition the movies get a good ROI.
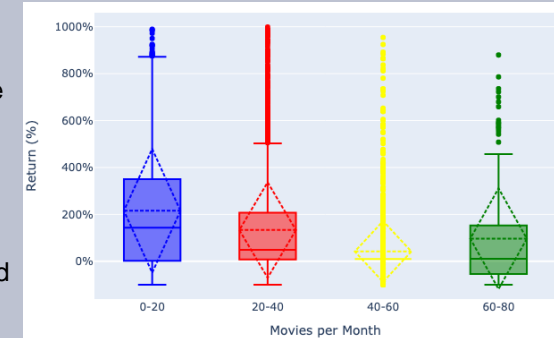
Distribution of Expected return vs Time Discounted Budget
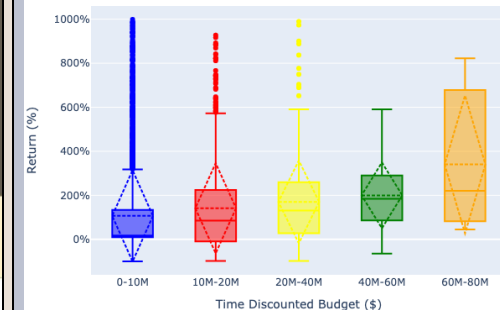
Again, as expected, Average ROI over given Time Discounted Budget (handling inflation over time) of a movie is directly proportional. Higher Budget, higher ROI.