TMDB Movie Revenue Prediction

Hyoin An, Jonghoo Lee, Hyeran Cho

2023 Erdosbootcamp

December 1, 2023

Data Gathering

- Dataset from TMDB API
- About 6500 movies available in the database with valid revenue data
- Features include genres, budget, production countries/companies, release date, cast, popularity, etc.

Data Cleaning

- Genres:
  The raw data contains a column of lists of genres.

['horrors', 'comedy'] →

genre_horrors	genre_comedy	genre_drama
True	True	False

- Cast:
  There are total of 130,005 actors/actresses.

0 ['Robert Duvall', 'Al Pacino', ...]

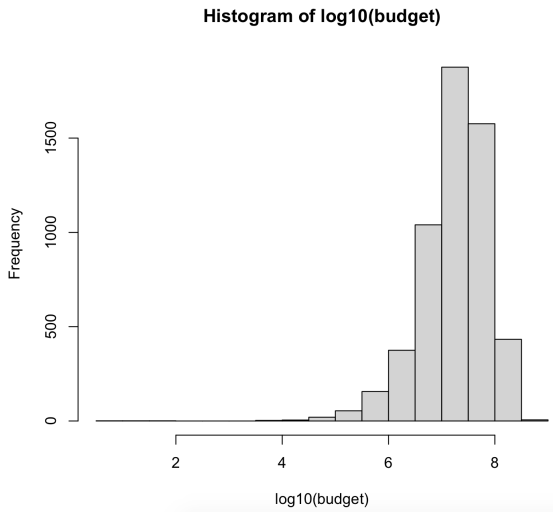
1 ['Leonardo DiCaprio', 'Kate Winslet', ...]

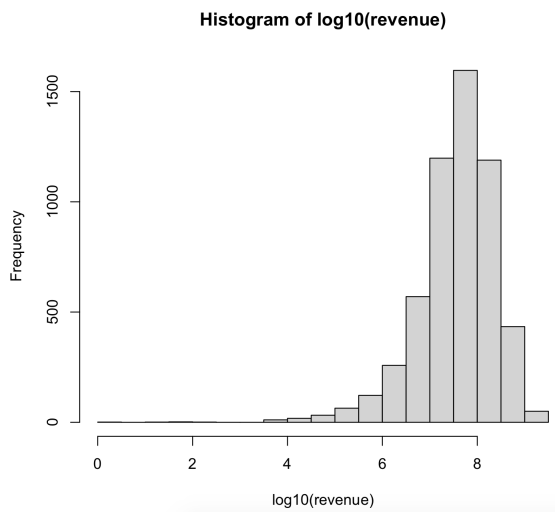
Index	...	Robert Duvall	Al Pacino	Leonardo DiCaprio	Kate Winslet
0	...	True	True	False	False
1	...	False	False	True	True

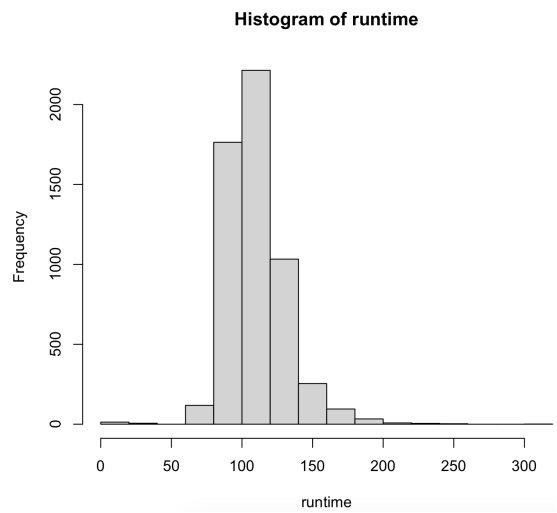
Data Cleaning

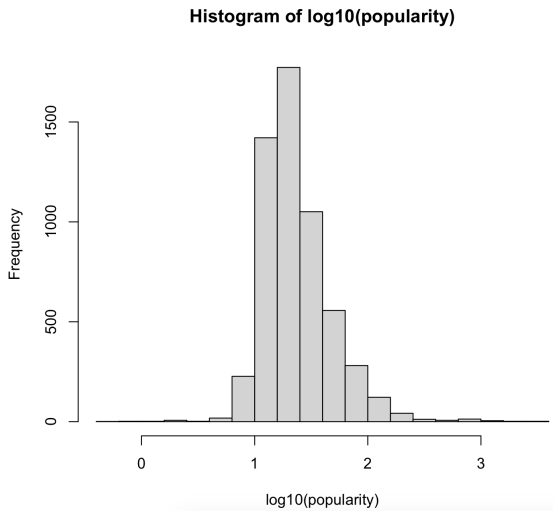
- Release Data
  The raw data is in Datetime format. We normalized them by linearly mapping onto $[-1, 1]$. We also added columns (dummy variables) storing the month of the movie's release date.
- Production Countries
  We added the column indicating whether the movie is produced in US or not.

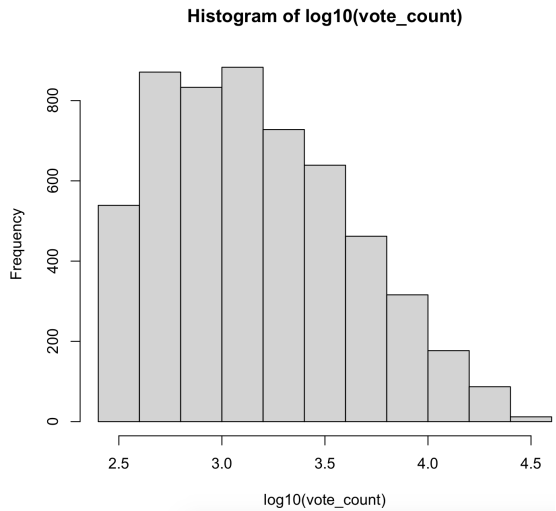
Exploratory Data Analysis (EDA)

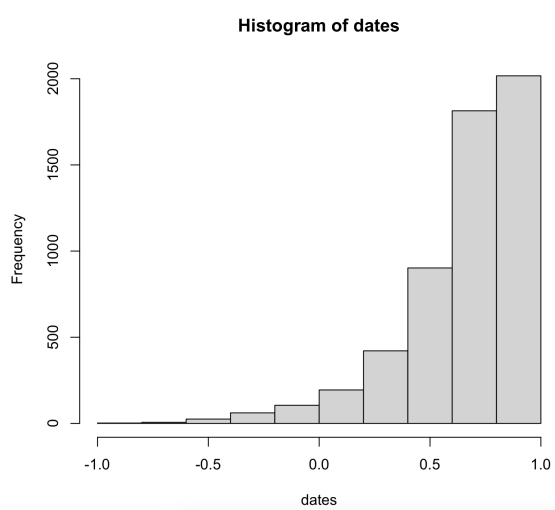




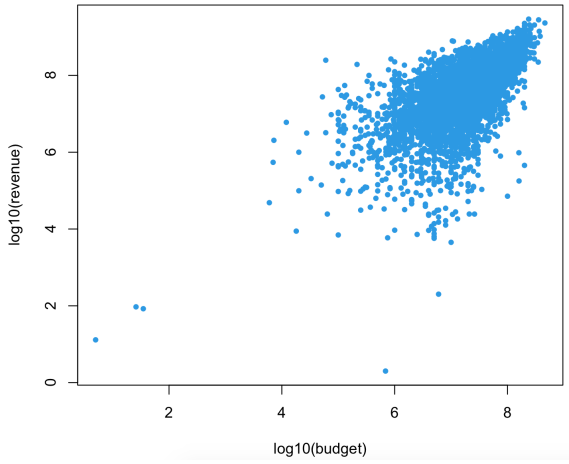




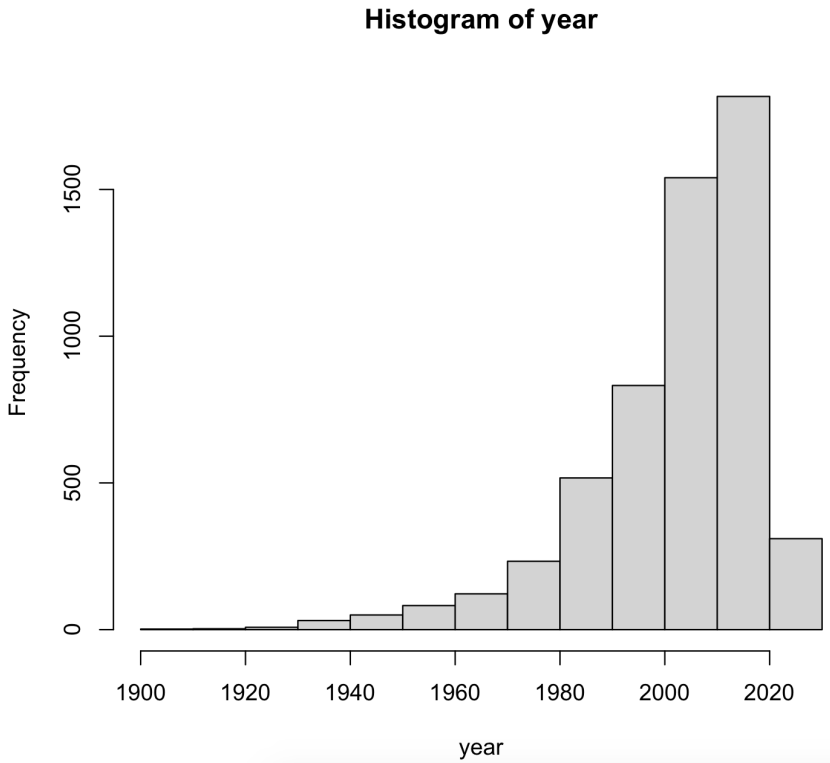


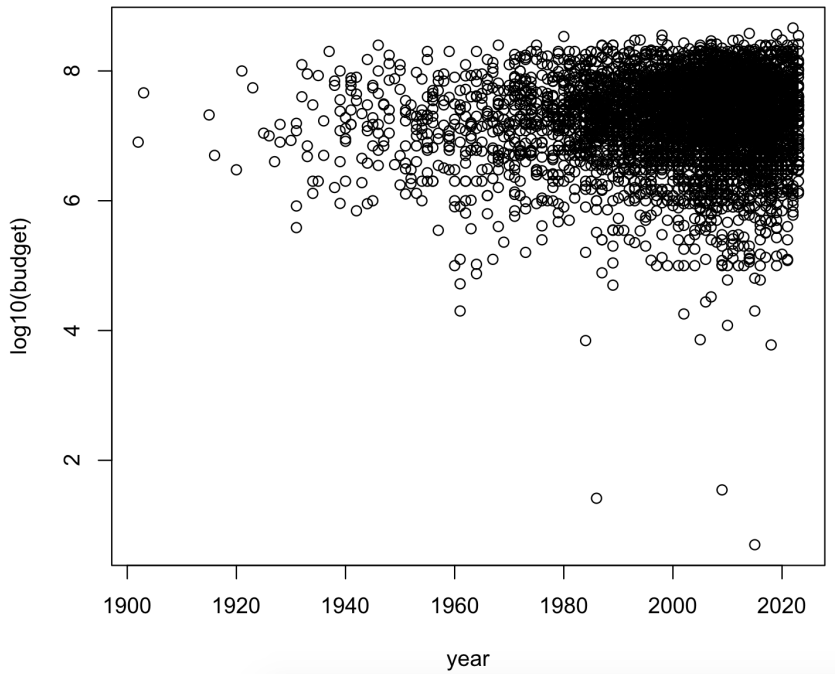


Exploratory Data Analysis (EDA)

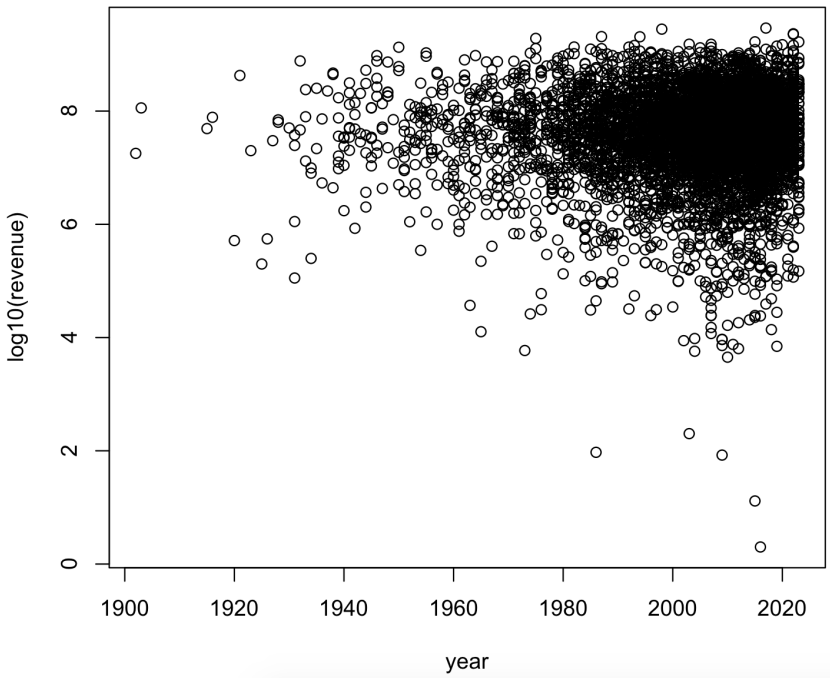


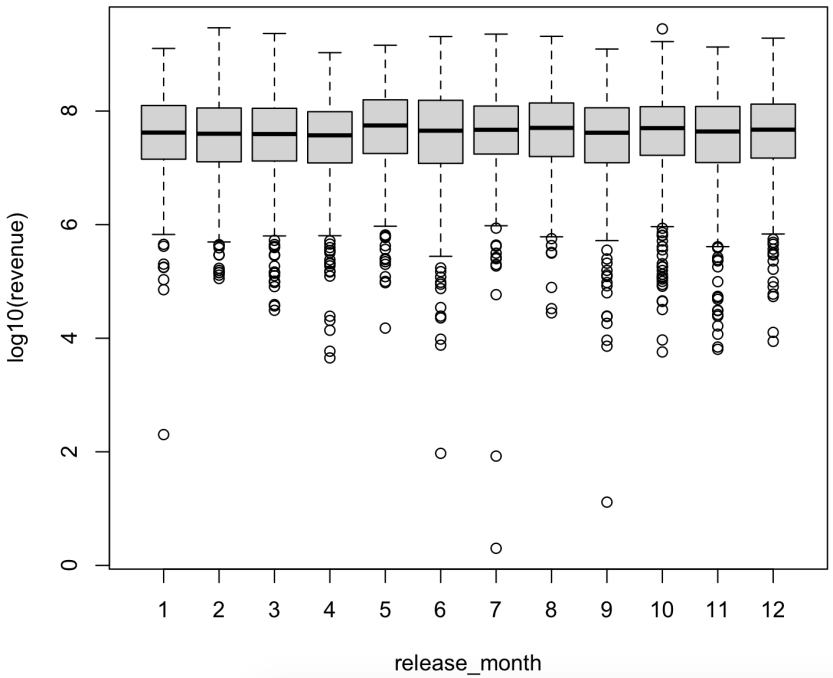
Exploratory Data Analysis (EDA)



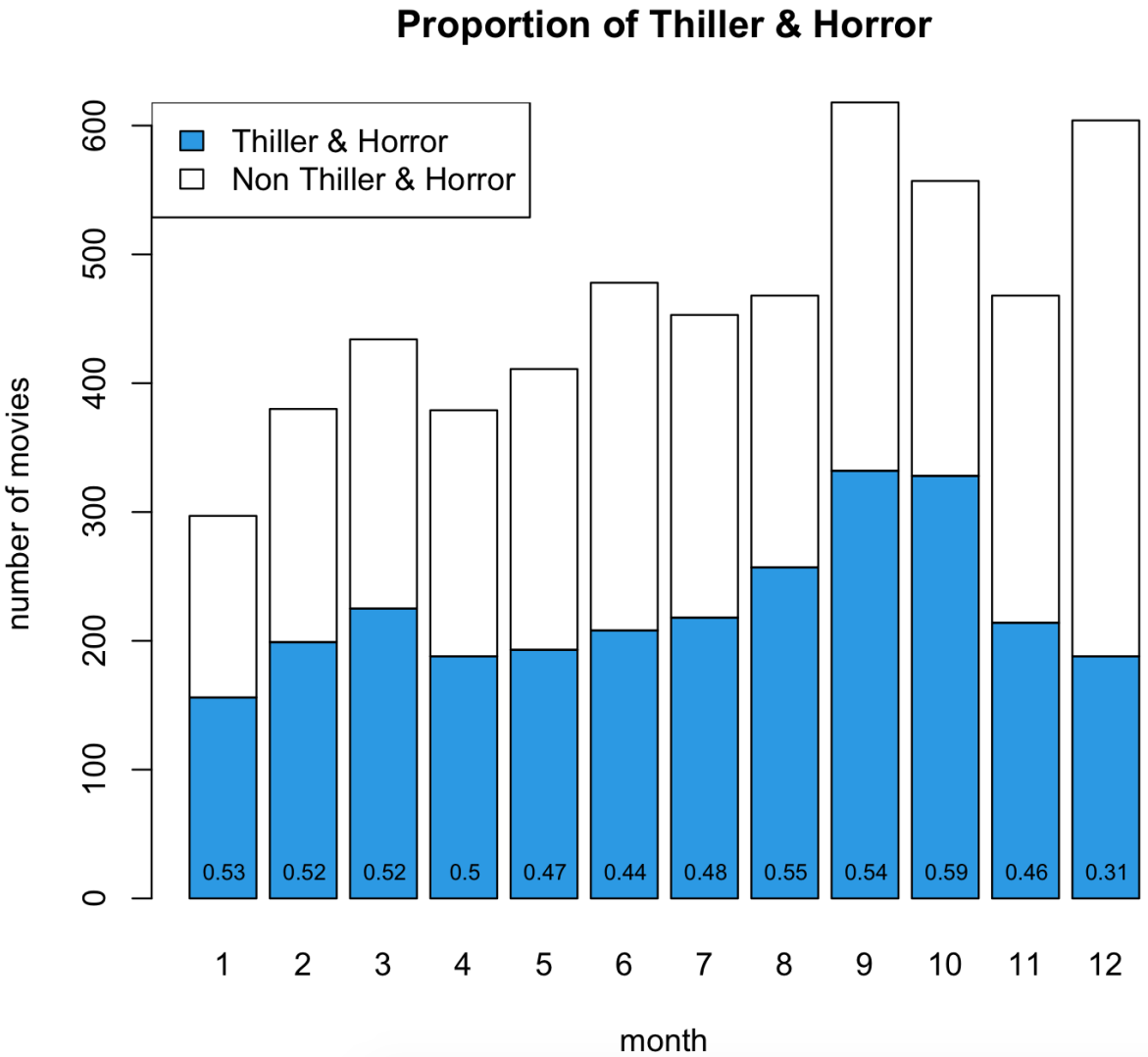


Exploratory Data Analysis (EDA)

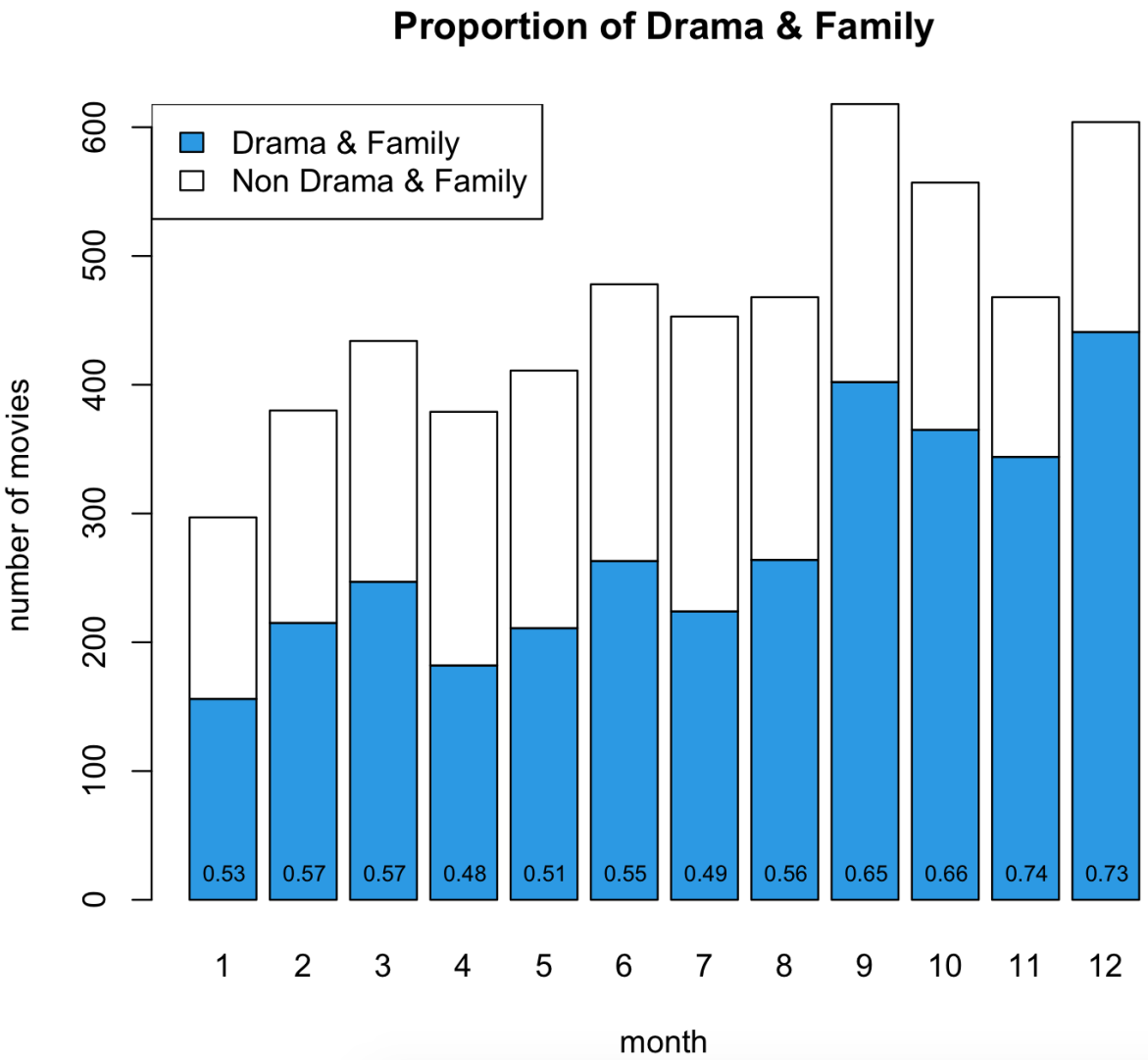




Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)



Model and Predictors

Models

Our model: XGBoost

Reference model: $\hat{y} = \log(\text{budget}+1)$

Response: log revenue

Predictors: Total 79 variables

- 10 key variables: budget, popularity, runtime, video, vote average, vote count, number of genres, collection, homepage, release date
- 59 dummy variables: genre (15), language (32), month (12)
- 10 principal components: cast, director

Model Training

Train/Test set split

- Training: $n = 4800$ (movies from 1902-04-17 to 2017-09-22)
- Test: $n = 747$ (movies 2017-09-22 to 2023-10-25)

Parameter Tuning using 5-fold Cross Validation

- Training/Validation set size: 3840/960
- Objective: Minimize average RMSE $\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n}$
- Hyperparameters for Tuning
  - max depth: [1, 3, 5, 10, 30, 100],
  - learning rate: [0.001, 0.005, 0.01, 0.05, 0.1, 0.2],
  - n estimators: [500, 1000, 2000]

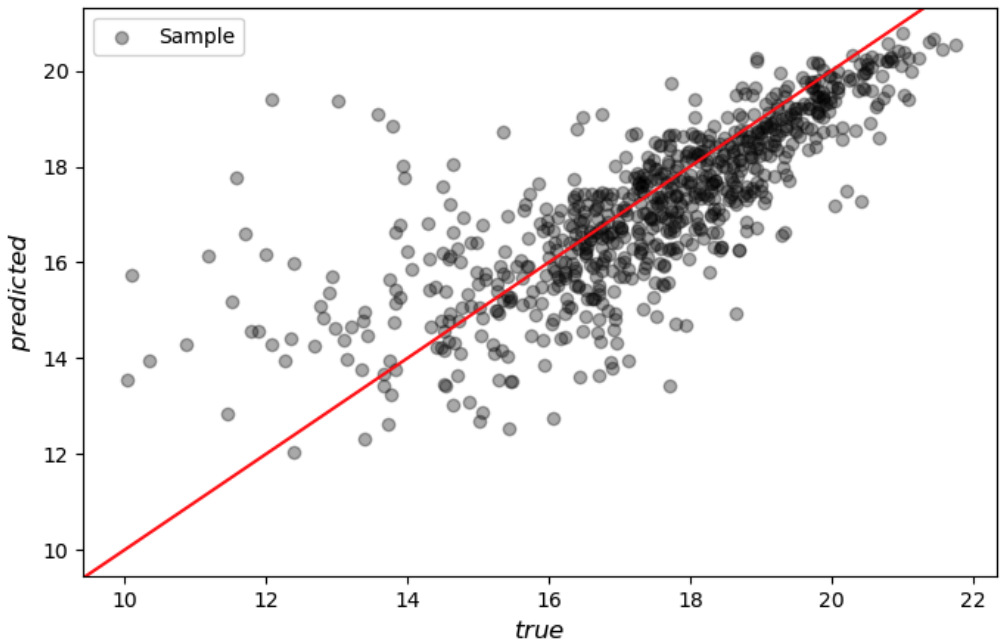
  [Selected] n estimators: 2000, max depth: 10, learning rate: 0.005
- Fixed Hyperparameters
  - colsample bytree: 0.7, min child weight: 1.5, reg alpha: 0.75, reg lambda: 0.45, subsample: 0.8, early stopping rounds: 100

Model Result

RMSE for Test data (n=747)

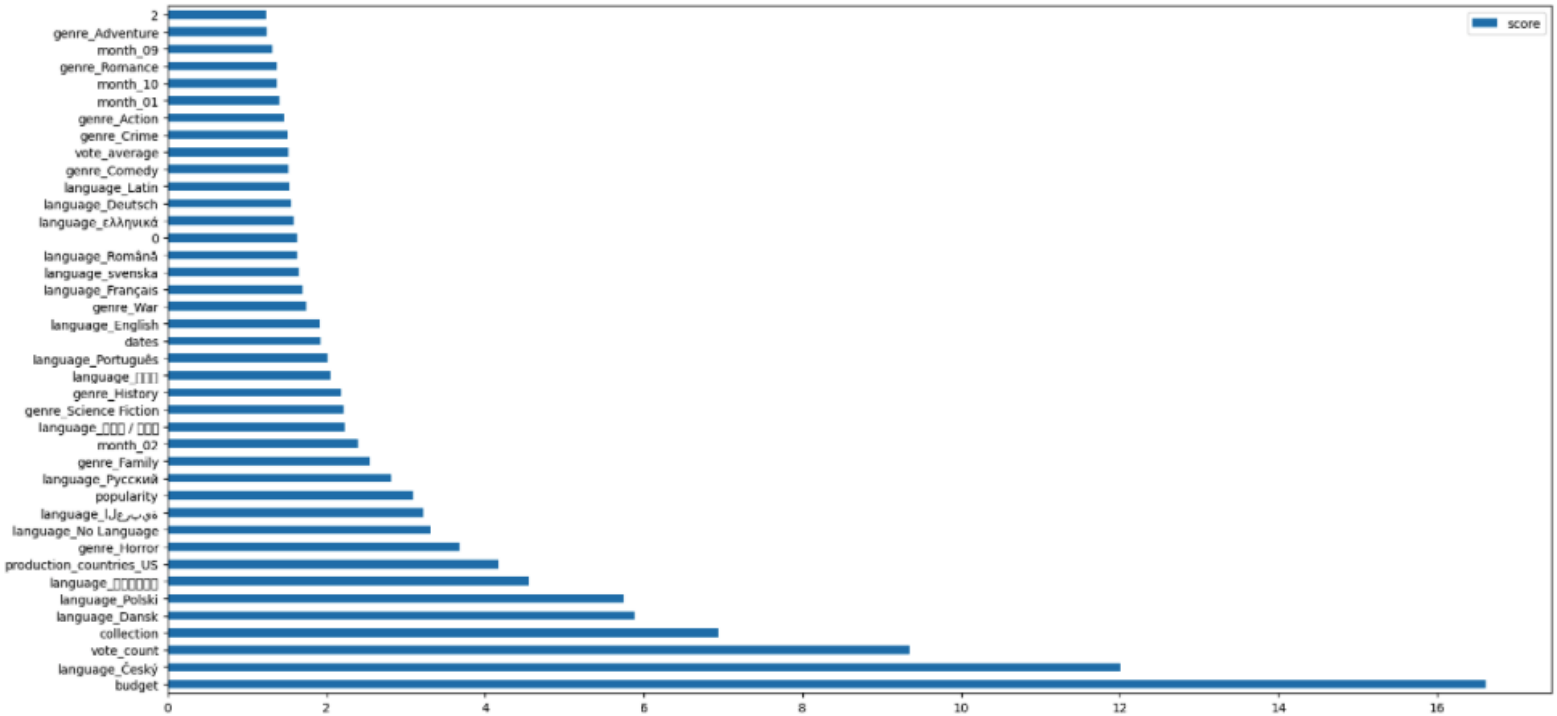
	Our Model	Reference
RMSE	1.3118	1.6552

Predicted vs. True plot



Model Result

Variable Importance plot



Discussion

- Utilizing movie title and overview information would be helpful for improving prediction
- Other ways of feature engineering for cast and director information could be considered