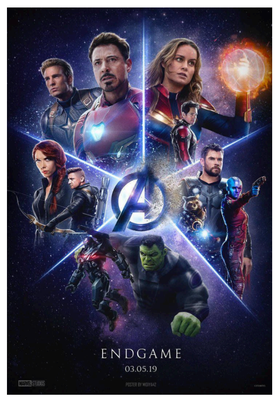
How about Avengers: Endgame?
Movie revenue prediction with hierarchical model

Jiazhao Li[1], Yuan Gao[1]
Department of EECS, University of Michigan

Problem and Motivations

1. Will this movie be good ?
2. Will this cast and crew succeed when they cooperated?
3. How should I manage my movie schedule to balance interests of customers and increase occupancy rate.

INPUT Given movie information.
OUTPUT Prediction revenue of movies



Dataset

Movie dataset crawled from IMDb.com from 2008 to 2018.

1. **3258** movies (average 300 per year) released in the United States from the whole world.
2. Actors/Actress: **5,147**
3. Directors: **2,073**
4. Writers : **2,184**

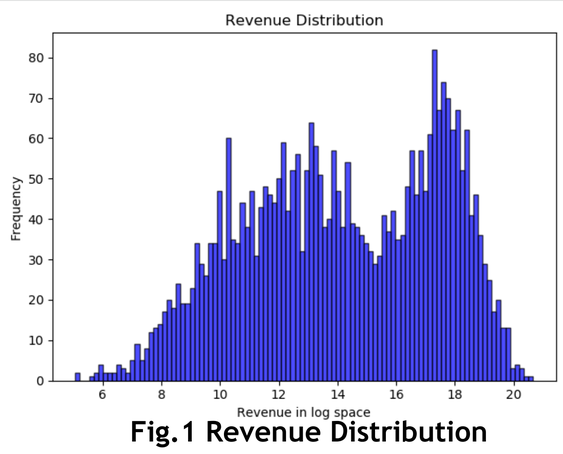
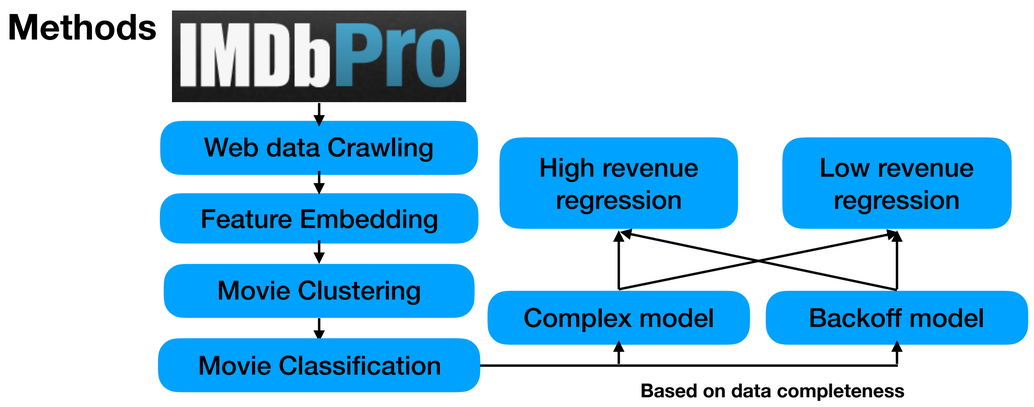


Fig.1 Revenue Distribution

Methods



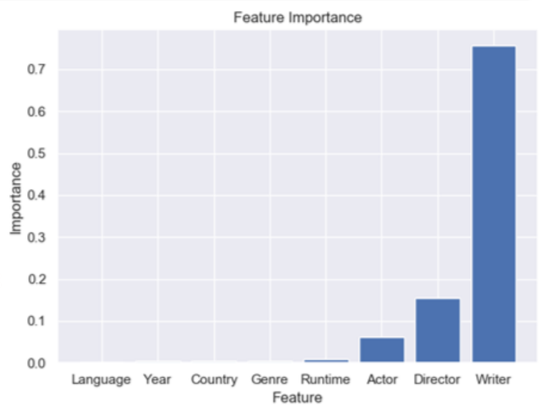
STEP 1: Feature Engineering

Numerical

Categorical

runtime, year,
genres, language, country
actors/actresses, directors, writers

Sparse

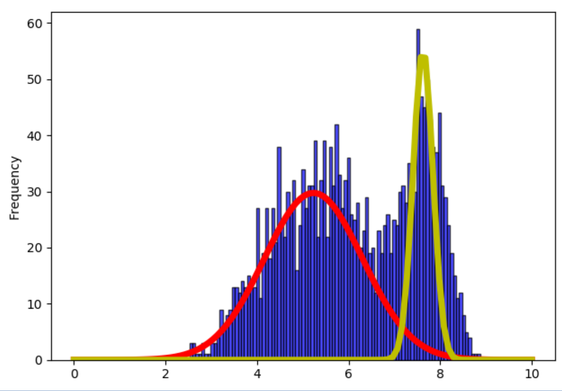


Encoding:
Categorical Features : **One-hot** encoding
Sparse Feature: Using **historical revenue**

Fig.2 Feature Importance using Gradient Boosting Regression

STEP 2: GMM Clustering and RFC Classification

- Two Gaussian distribution clusters
- **Gaussian Mixture Model** (GMM) clustering into two classes: **High revenue** and **Low revenue**
- **Random Forest Classification** based on clustering result.



STEP 3: Regression with original/ back-off models

Cold start problem: No reference revenue data for first appearing actors. Back-off model

Data with missing features: → Back-off regression without 'Sparse feature'.
Data with all features: → Original regression with all features.

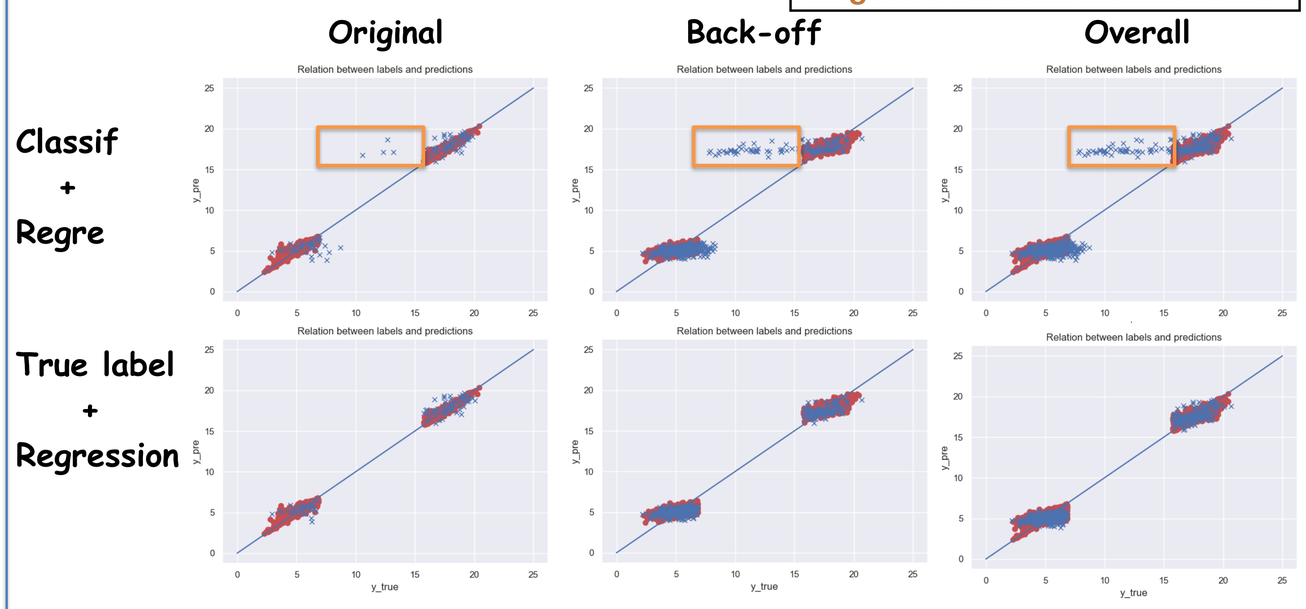
Gradient Boosting Regression (GBR) performed best.

Experimental Results

Ablation experiment:

How about performance of each component?

Y axis: prediction revenue
x axis: ground true revenue
Red points: train samples
Blue crosses: test samples
Original box: miss classification



Model	Test ROC score
Original model	0.88
Back-off model	0.73

Observation:
1. Most of instances have missing features
2. Under true classification, regression model works perfectly.
3. More features → the better of classification

Result

Train set	Test set	MAE	SMAPE
2008-14	2015-18	$23.6M	0.905

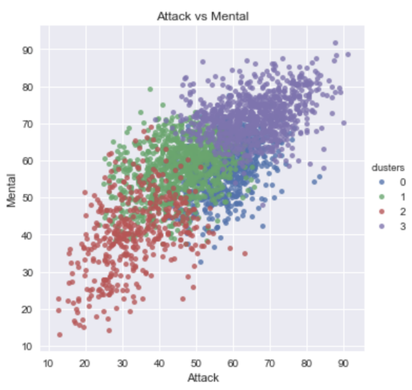
Observations:
More information contributes to model.

More information help?

Train set	Test Set	MAE	SMAPE
2008-14	2017-18	$15.2M	1.47
2008-15	2017-18	$14.9M	1.41
2008-16	2017-18	$13.4M	1.38

Model Generalization

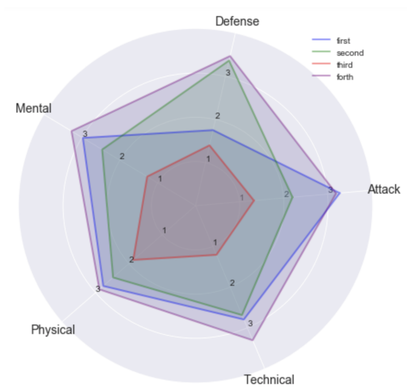
Can our model works on other dataset?



New dataset: **Europe Soccer**

Features: Crossing, Short-passing, Dribbling, Shot power, Penalties……
5 categories : Technical, Attack, Physical, Defense, Mental.

0: Attack players, 1. Defense players, 3. All-star players, 2. Goalkeeper



This is a final project presentation of SI 699: Big Data Analytics. We thank Prof. Qiaozhu for advise and supervision.

Contact:jiazhaol@umich.edu