FAME OR FLOP:

PREDICTING THE SUCCESS OF A FILM USING NATURAL LANGUAGE PROCESSING & MACHINE LEARNING

AUG 14 2019

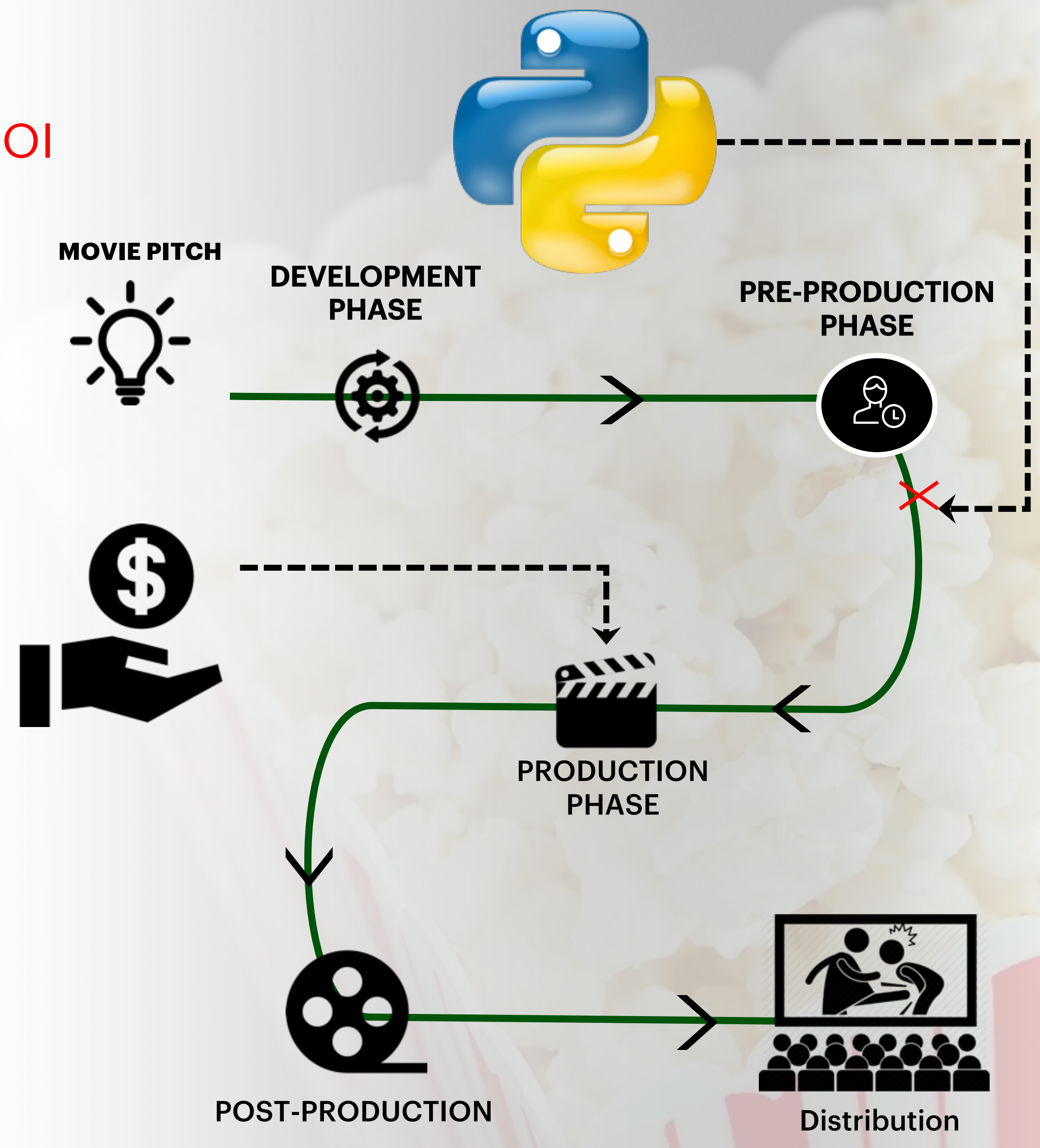
CONTENTS

PROJECT MOTIVATION

EARLY PREDICTION OF MOVIE SUCCESS AND ROI

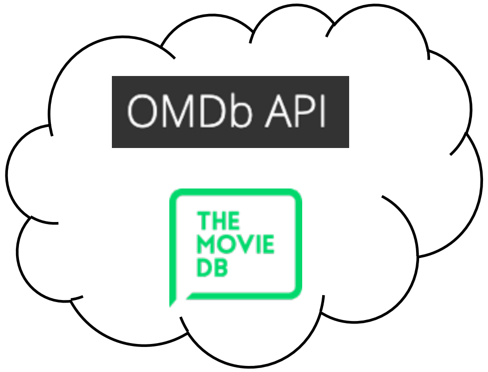
Current Challenge:

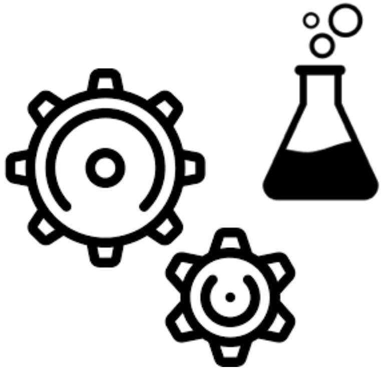
- **In Hollywood, the stakes are high:** Sustained success can mean fame and fortune, but 70% of movies flop!
- **What if you could know with more certainty if a movie has a real shot at success before your money is sunk?**
- **Hollywood already does this.** *Movies tend to follow a formula*. What is that formula? That's what I set out to find!



THE DATA PIPELINE & MACHINE LEARNING

EARLY PREDICTION OF MOVIE SUCCESS AND ROI





Gather and clean data

Data is ingested through two APIs:

- Open Movie Database (OMDB) API directors, actors, writers and awards
- The Movie Database (TMBD) API budget and revenue information

- Define Success:

  *Profit = Revenue – (3 x Budget)**

- *Other ways to define?*

Feature Engineer

Features are derived and added to the data. Some examples:

- Director/Writer/Actor popularity
- Last movie' Oscar Performance
- Studio Performance
- Historical Chemistry
- Release proximity to other films

Train The Topic Model

Plot Synopsis text: fed to a Latent Drichlet Allocation Algorithm. Producing:

- 20 general topics
- Percentage contribution scores to each topic

Train The Classifier Model

Supervised learning model is trained using ~250 features:

- GradientBoosting Classifier
- XGBoost & Random Forest yield similar results

FEATURE ENGINEERING

A DRILL DOWN OF EACH FEATURE

1. **add_star_power()** #=> Actor popularity

2. **add_writer_power()** #=> Writer popularity

3. **add_director_power()** #=> Director popularity

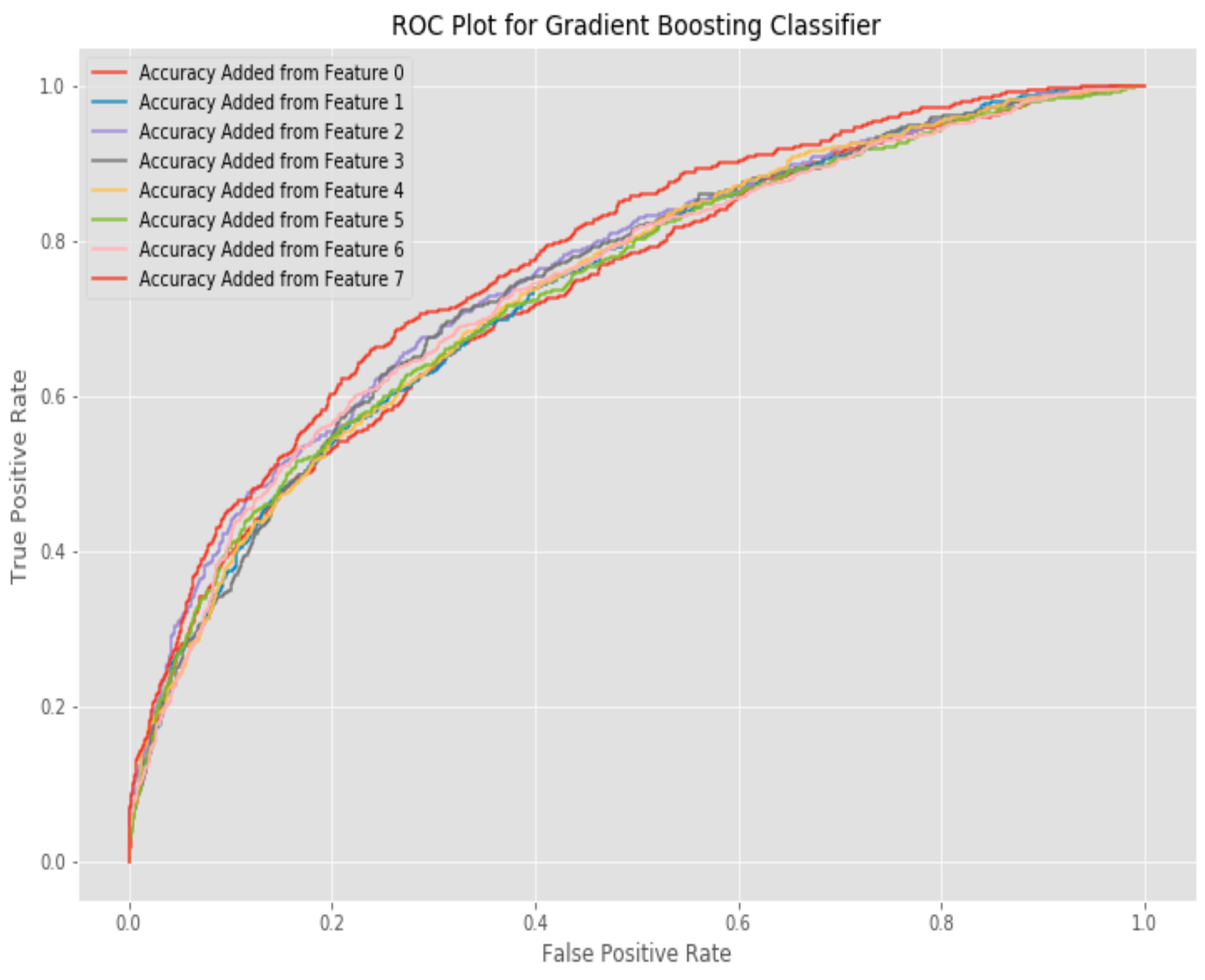
4. **last_movie_award()** #=> Oscars from Actors/Directors/Writers' last movie

5. **top_production_score()** #=> Best prod studios

6. **add_chem_factor(verbose)** #=> count of past collaborations between crew-member pairs

7. **add_macro_trend()** #=> avg revenue in last year

8. **add_release_proximity()** #=> how close to o

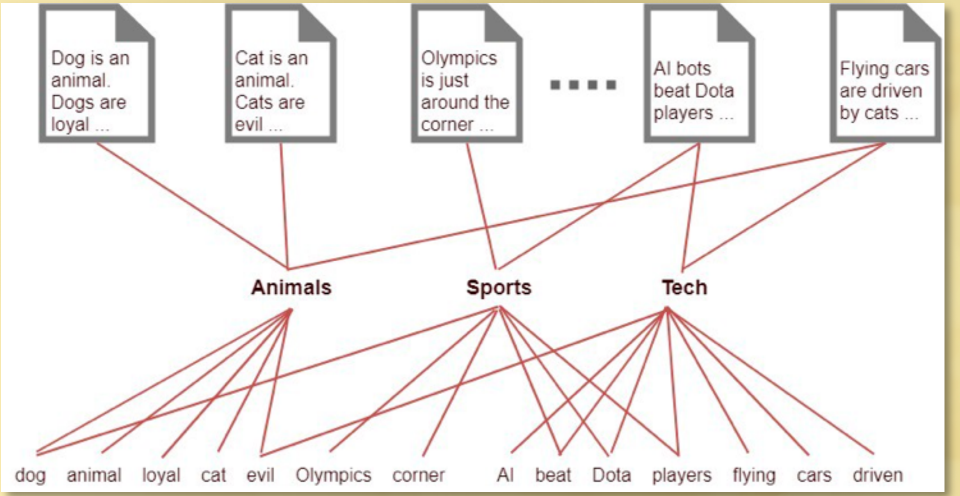


TOPIC MODELING

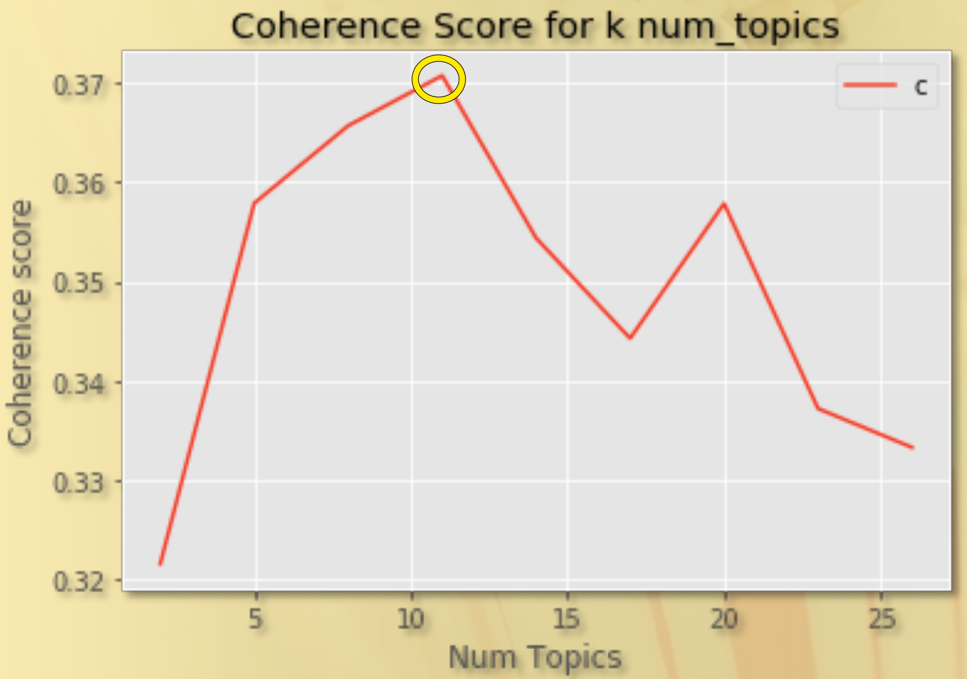
LATENT DRICHLET ALLOCATION (LDA)

<u>What is LDA / Topic Modelling?</u>

- Topic modelling refers to the task of identifying topics that best describes a set of documents.
- LDA represents documents as **mixtures of topics** represented as words with certain probabilities
- Assumes documents (plots) are generated probabilistically from these topics
- Algorithm requires a parameter k for # of latent topics I had to set
  - Coherence score
  - Perplexity score
  - Optimal=11



$$\gamma^\star, \phi^\star, \lambda^\star = \text{argmin}_{(\gamma,\phi,\lambda)} D(q(\theta, \mathbf{z}, \beta|\gamma, \phi, \lambda)||p(\theta, \mathbf{z}, \beta|\mathcal{D}; \alpha, \eta))$$



TOPIC MODELING (CON'T)

LATENT DRICHLET ALLOCATION (LDA)

Topic Model Output:

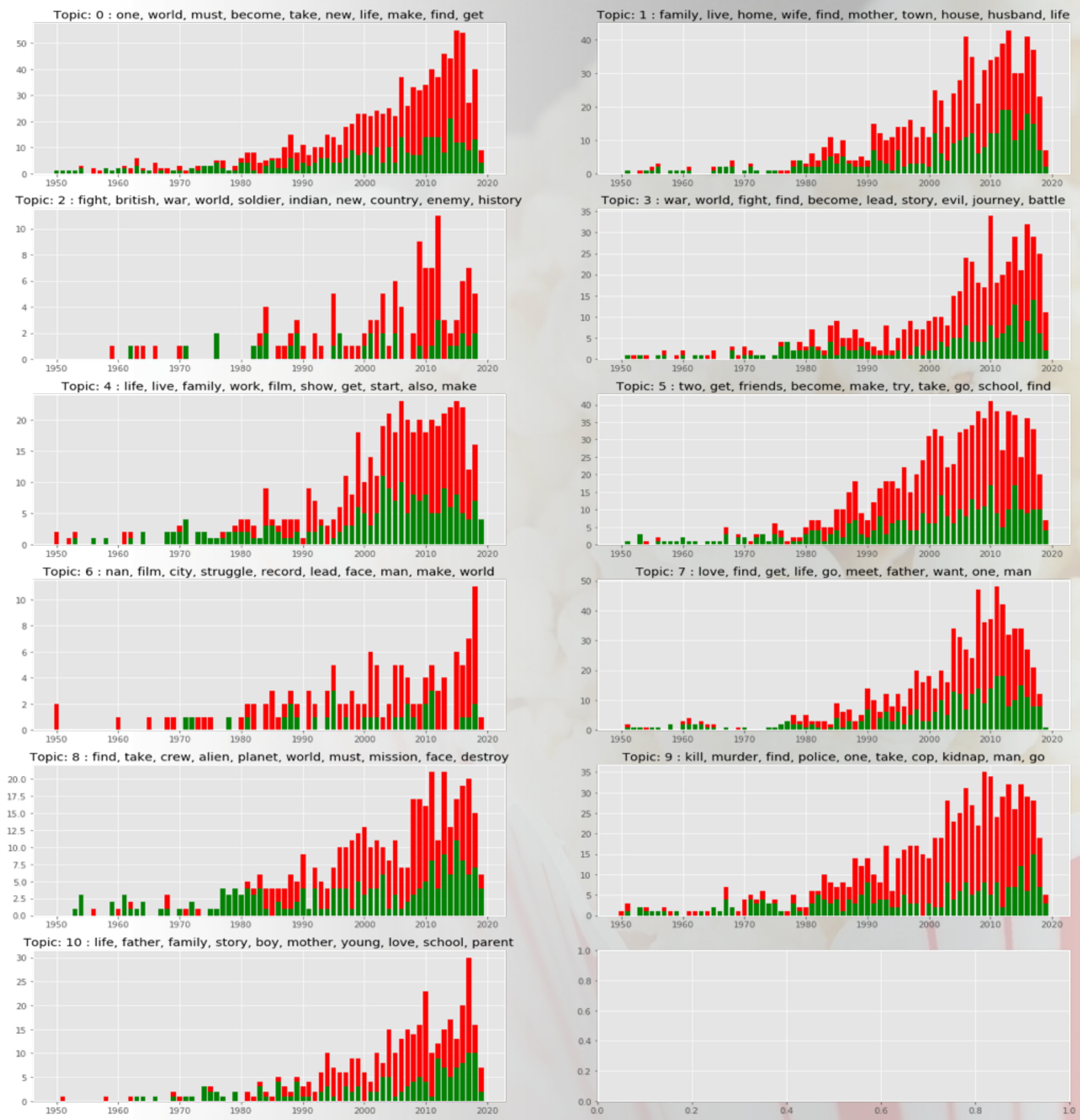
```
[(0, '0.008*"one" + 0.008*"world" + 0.007*"must" + 0.006*"become" + 0.006*"take" ' '+ 0.006*"new”'),

(1, '0.023*"family" + 0.012*"live" + 0.010*"home" + 0.010*"wife" + 0.009*"find" ' '+ 0.009*"mother”'),

(2, '0.008*"fight" + 0.008*"british" + 0.008*"war" + 0.007*"world" + ' '0.006*"soldier" + 0.006*"indian”'),

(3, '0.009*"war" + 0.008*"world" + 0.008*"fight" + 0.007*"find" + 0.007*"become" ' '+ 0.007*"lead”'),

(4, '0.014*"life" + 0.012*"live" + 0.009*"family" + 0.006*"work" + 0.006*"film" ' '+ 0.006*"show”'),

(5, '0.011*"two" + 0.010*"get" + 0.007*"friends" + 0.006*"become" + 0.006*"make" ' '+ 0.006*"try”'),

(6, '0.018*"nan" + 0.016*"film" + 0.006*"city" + 0.006*"struggle" + ' '0.005*"record" + 0.005*"lead”'),

(7, '0.015*"love" + 0.013*"find" + 0.012*"get" + 0.012*"life" + 0.011*"go" + ' '0.009*"meet”'),

(8, '0.011*"find" + 0.008*"take" + 0.008*"crew" + 0.007*"alien" + 0.007*"planet" ' '+ 0.007*"world"'),

(9, '0.018*"kill" + 0.012*"murder" + 0.012*"find" + 0.008*"police" + 0.007*"one" ' '+ 0.007*"take”'),

(10, '0.024*"life" + 0.010*"father" + 0.010*"family" + 0.010*"story" + ' '0.009*"boy" + 0.007*"mother"')]
```

TOPIC MODELING (CON'T)

LATENT DRICHLET ALLOCATION (LDA)

How does this help the classification?

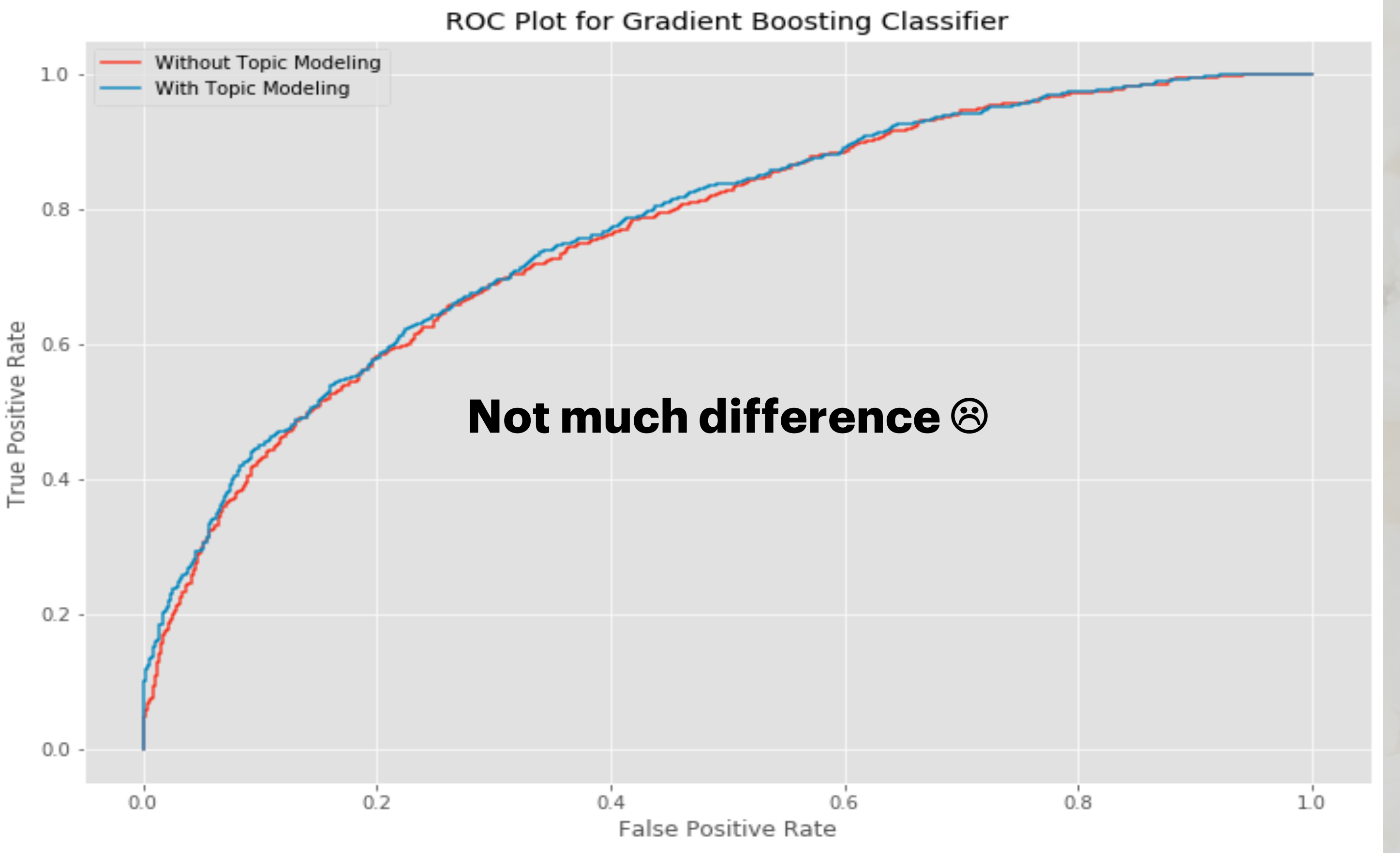
- *Each movie contributes to a topic by some percentage*
- ***In theory*** some topics are more profitable than others
- ***Conclusion*:**
  - Its interesting but doesn't help much
  - Need to segment the data better (i.e. last 5 years)
- ***Challenges:*** environmental issues using LDA Mallet



TOPIC MODELING (CON'T)

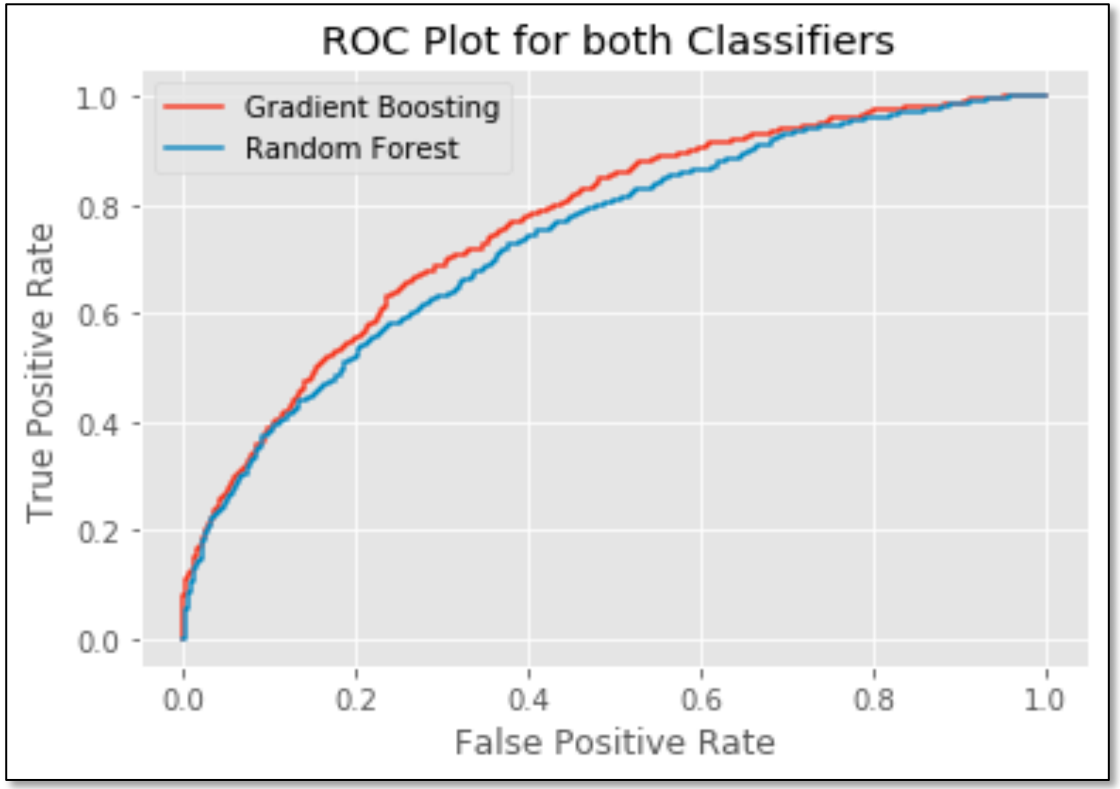
LATENT DRICHLET ALLOCATION (LDA)

How does this help the classification?

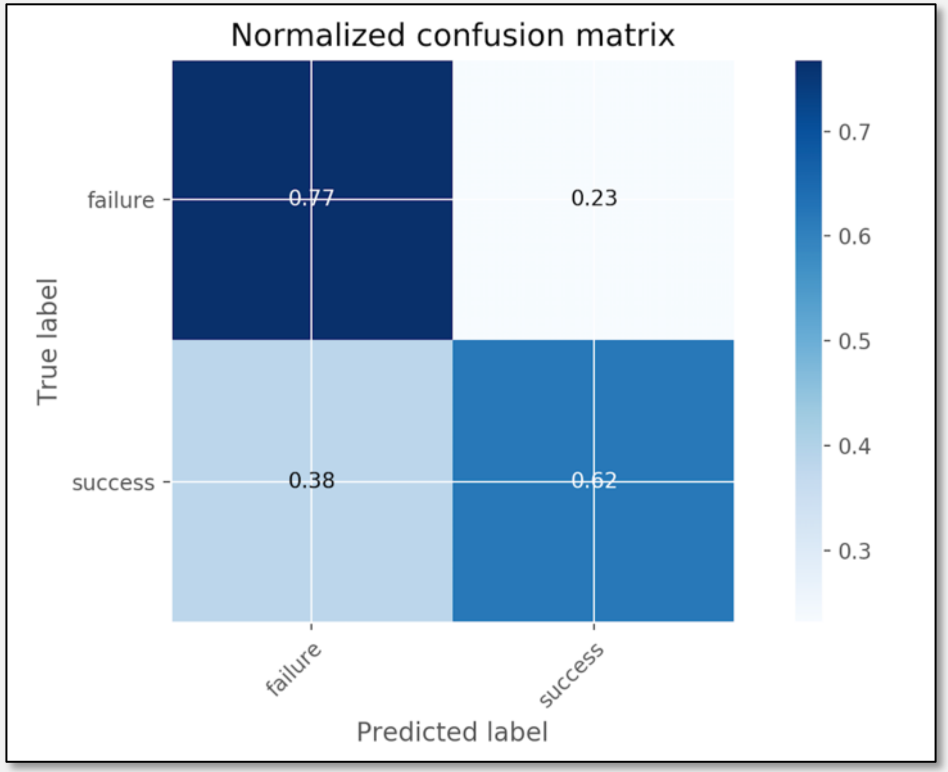


EVALUATION

EARLY PREDICTION OF MOVIE SUCCESS AND ROI




```
Accuracy : 0.7231
ROC AUC  : 0.7761
Recall   : 0.6675
Precision: 0.6132
```



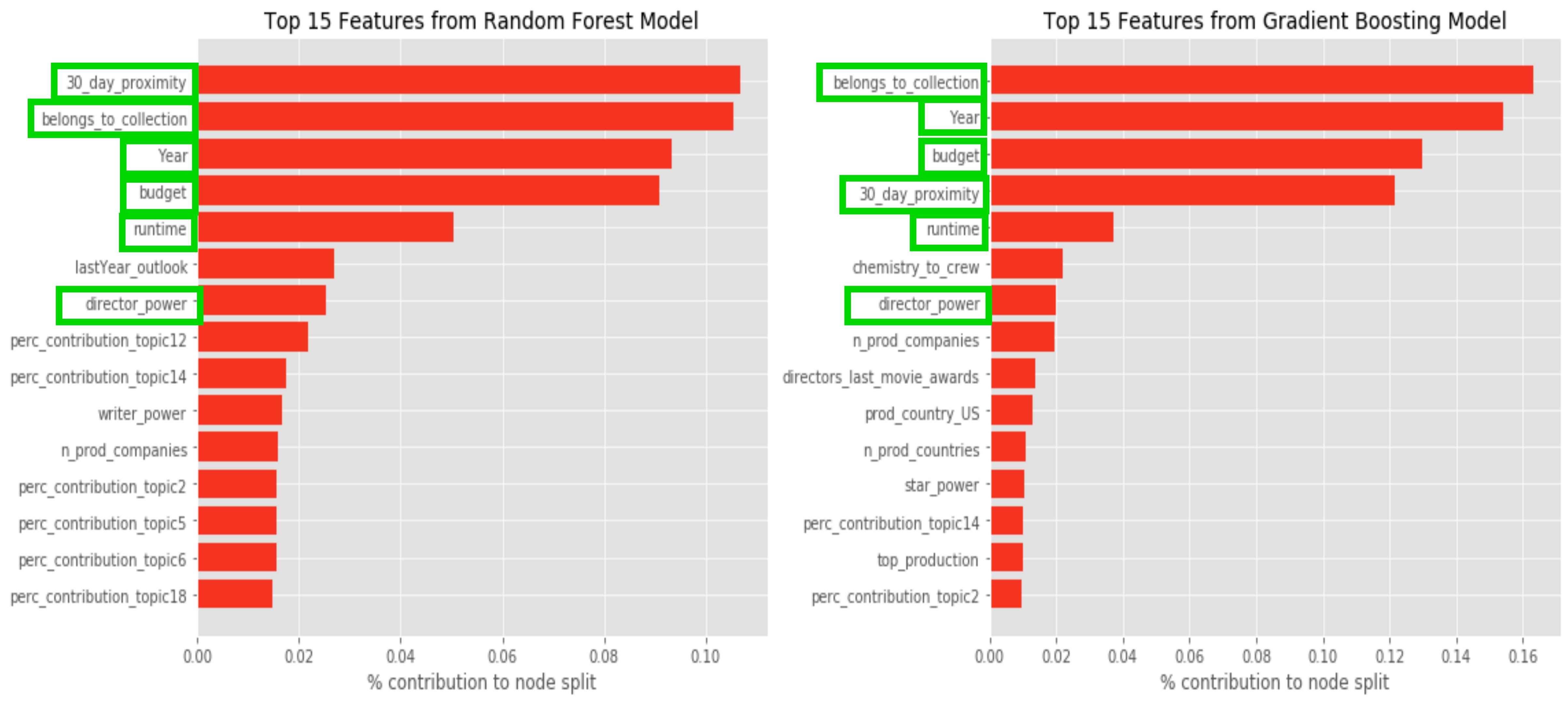
- ***Use Gradient Boosting!***:
  - *Exhaustive GridSearchCV*
  - *Sample_weights / Class_weights*
- ***Challenges:*** Time series data leakage
- ***Conclusion***: Need a high threshold – model is better at finding the successes than avoiding false negatives

EVALUATION

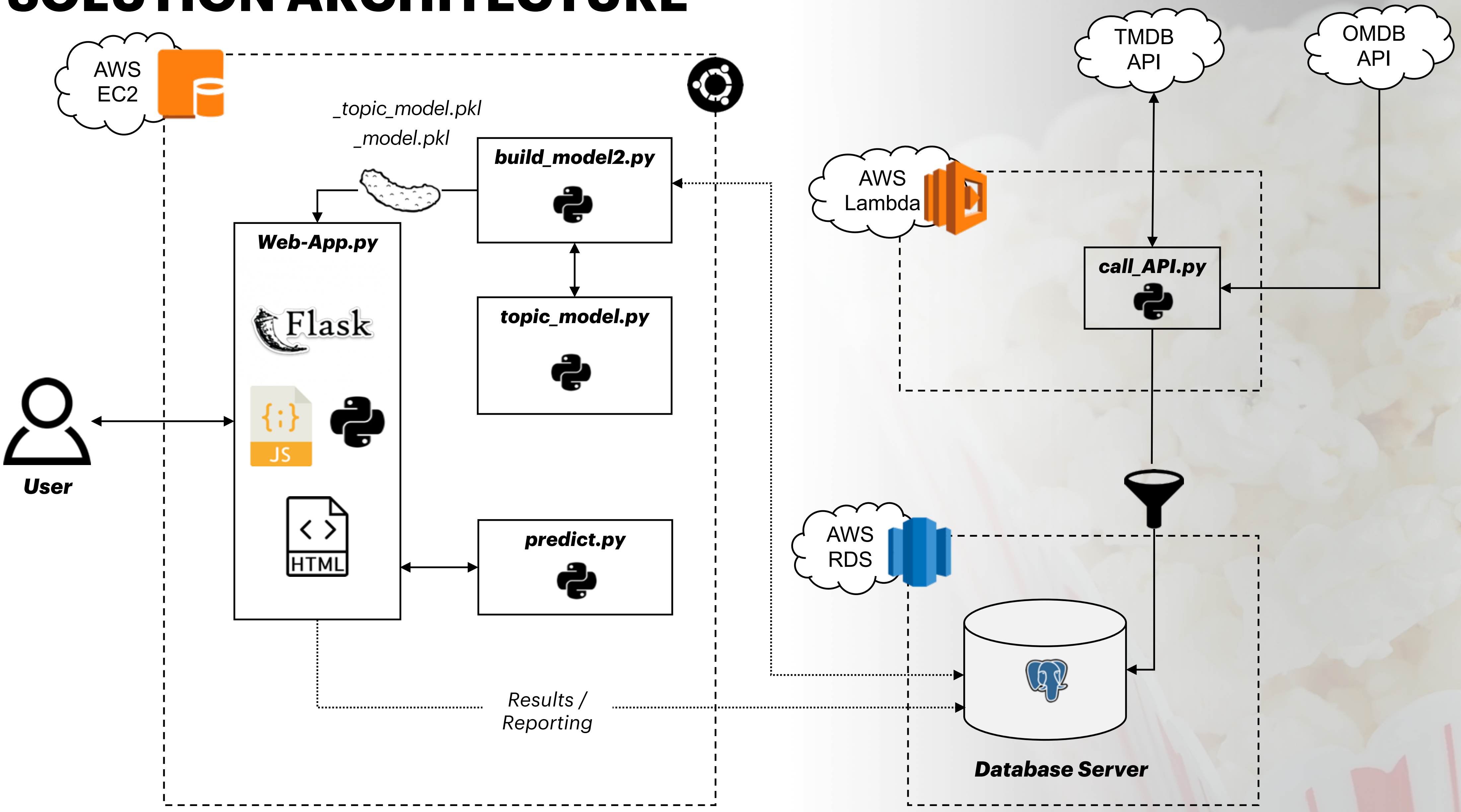
EARLY PREDICTION OF MOVIE SUCCESS AND ROI

Demo: Lets make a [successful] movie!!

Feature Importance – 'The Formula' to a successful movie!



SOLUTION ARCHITECTURE



NEXT STEPS

PHASE 2.0:

1. Clean the data better (e.g. remove wrestling matches)
2. Optimize model with a more exhaustive grid search (XGBoost)
3. Add more features:
   - Measuring success within genre
   - NLP / Topic model entire movie scripts
4. Improve the Topic modeling with more meticulous NLP
5. Fork the data pipeline into two sets of data and separate classifiers: one topic models just the movies in the last 5 years
6. Build the rest of the architecture and database servers to continuously ingest new movie data and update model monthly
7. More analysis on the textual data I'm feeding to the LDA algorithm
8. Improve the Web App:
   - Make case-insensitive
   - Set range parameters (e.g. on Budget, Avg Revenue etc.)
   - Make prettier results page
   - Make form more intuitive (i.e. Genre checklist, specify which fields are optional)

ABOUT ME

MAX BAMBERGER

Summary

I'm a Management Consultant with with 7+ years of experience as a professional consultant in the Cable, Telecom and Media/Advertising industries.

github.com/MaxBamberger

max.bamberger@gmail.com

linkedin.com/in/max-bamberger

Work cited:

Entertainment Industry Economics (Vogel).

https://www.quora.com/What-is-the-average-return-on-investment-for-a-Hollywood-movie

https://www.quora.com/What-is-a-good-explanation-of-Latent-Dirichlet-Allocation

https://www.quora.com/What-makes-a-film-a-box-office-success

https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158