Predictions of Film Success Based on Script Content

Flatiron School Capstone Project

By: William Newton

Business Problem

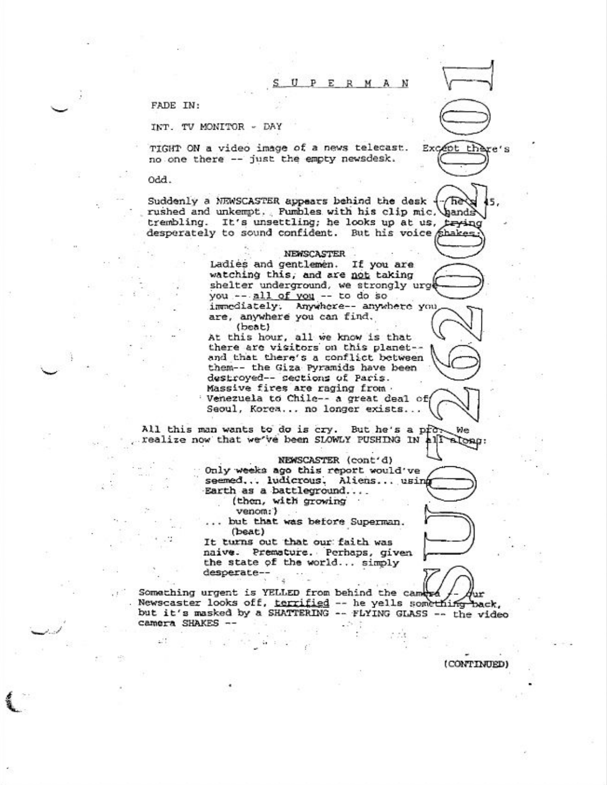
- 50,000 scripts are registered with WGA per year [1]
  - 150 movies are released per year
  - 0.3% of screenplays are made into films which leaves 49,850 unproduced screenplays
- Scripts for a 2-hour plus movie are between 7,500 and 20,000 words long
- Average person reads at 300 words per minute
  - ~45 minutes to read screenplay, and up to 3 hours to complete report on screenplay [2]
- Almost 200,000 man hours per year wasted on reading unproduced screenplays
  - Freelance script readers make ~ $50/hour [2]
- That is $10 million in wasted expense every year

1. How Hollywood Chooses Scripts: The Insider List That Led to 'Abduction', *The Atlantic*, https://www.theatlantic.com/entertainment/archive/2011/09/how-hollywood-chooses-scripts-the-insider-list-that-led-to-abduction/245541/
2. How to Become a Script Reader, *StudioBinder*, https://www.studiobinder.com/blog/how-to-become-a-script-reader/

Solution:

Build Model To Read Scripts and Determine if They Are a Good Investment

Data and Methodology



SUPERMAN

FADE IN:

INT. TV MONITOR - DAY

TIGHT ON a video image of a news telecast. Except there's no one there -- just the empty newsdesk.

Odd.

Suddenly a NEWSCASTER appears behind the desk -- he's 45, rushed and unkempt. Fumbles with his clip mic, hands trembling. It's unsettling; he looks up at us, trying desperately to sound confident. But his voice shakes:

NEWSCASTER

Ladies and gentlemen. If you are watching this, and are not taking shelter underground, we strongly urge you -- all of you -- to do so immediately. Anywhere-- anywhere you are, anywhere you can find.

(beat)

At this hour, all we know is that there are visitors on this planet-- and that there's a conflict between them-- the Giza Pyramids have been destroyed-- sections of Paris. Massive fires are raging from Venezuela to Chile-- a great deal of Seoul, Korea... no longer exists...

All this man wants to do is cry. But he's a pro. We realize now that we've been SLOWLY PUSHING IN all along:

NEWSCASTER (cont'd)

Only weeks ago this report would've seemed... ludicrous. Aliens... using Earth as a battleground....

(then, with growing venom:)

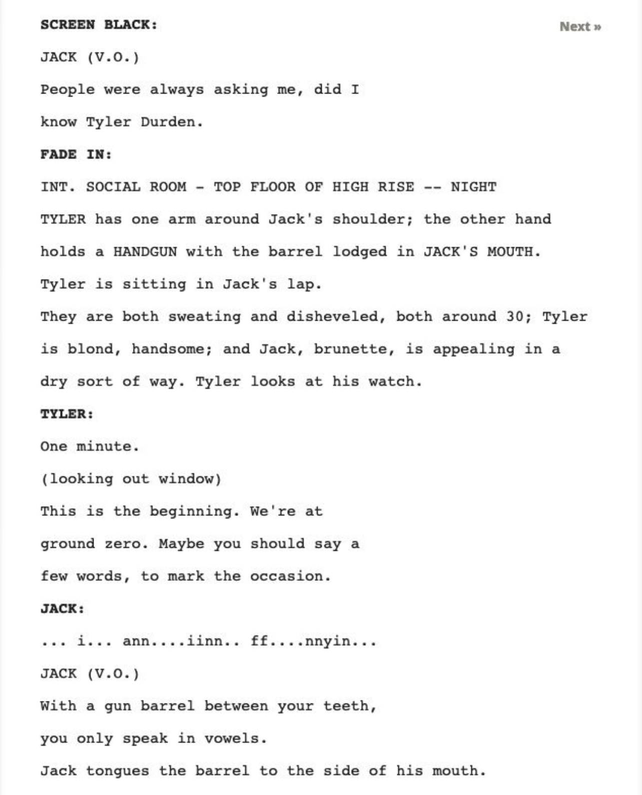
... but that was before Superman.

(beat)

It turns out that our faith was naive. Premature. Perhaps, given the state of the world... simply desperate--

Something urgent is YELLED from behind the camera -- our Newscaster looks off, terrified -- he yells something back, but it's masked by a SHATTERING -- FLYING GLASS -- the video camera SHAKES --

(CONTINUED)



Next »

SCREEN BLACK:

JACK (V.O.)

People were always asking me, did I

know Tyler Durden.

FADE IN:

INT. SOCIAL ROOM - TOP FLOOR OF HIGH RISE -- NIGHT

TYLER has one arm around Jack's shoulder; the other hand

holds a HANDGUN with the barrel lodged in JACK'S MOUTH.

Tyler is sitting in Jack's lap.

They are both sweating and disheveled, both around 30; Tyler

is blond, handsome; and Jack, brunette, is appealing in a

dry sort of way. Tyler looks at his watch.

TYLER:

One minute.

(looking out window)

This is the beginning. We're at

ground zero. Maybe you should say a

few words, to mark the occasion.

JACK:

... i... ann....iinn.. ff....nnyin...

JACK (V.O.)

With a gun barrel between your teeth,

you only speak in vowels.

Jack tongues the barrel to the side of his mouth.

Data and Methodology

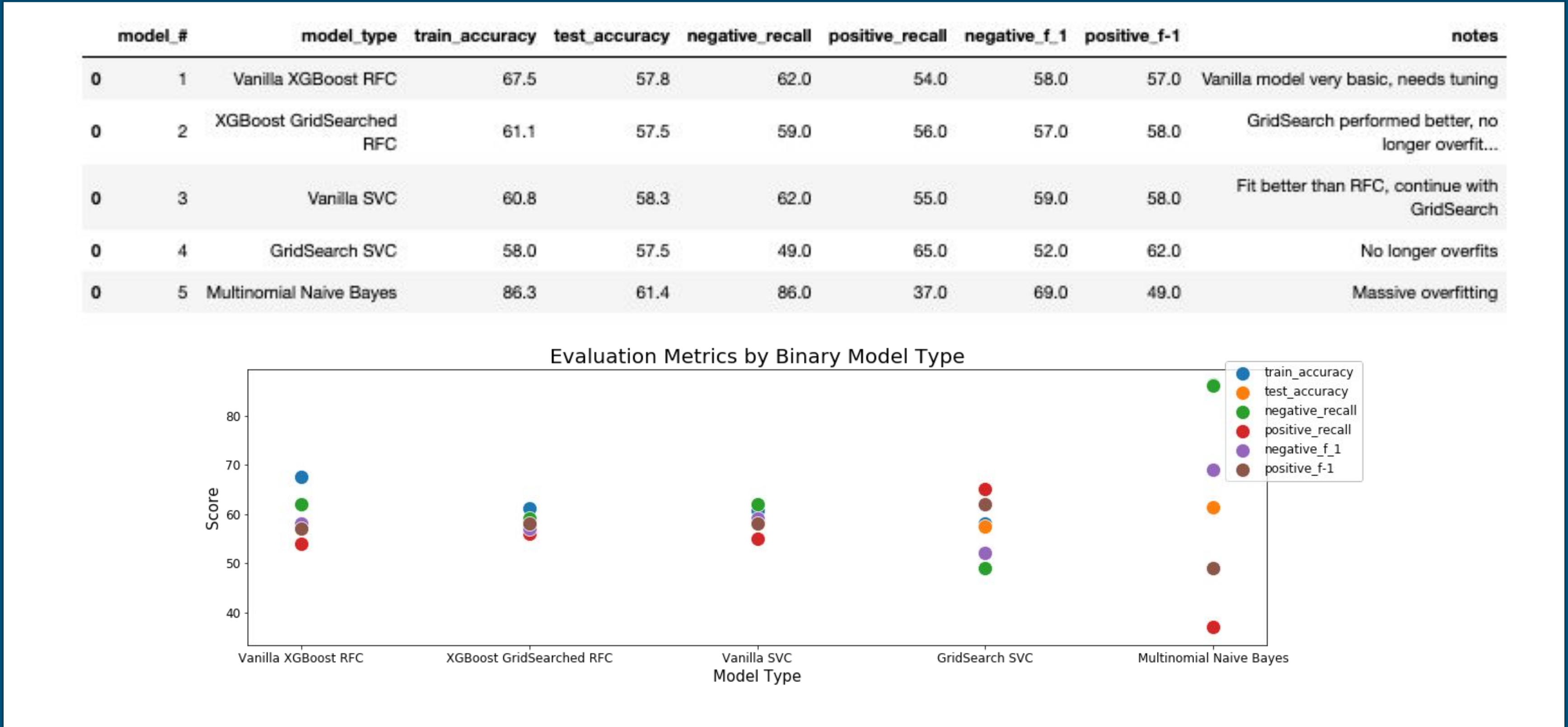
- Data containing film production budget and box office gross was obtained from TheNumbers.com and OMDBapi.com
- Script text data was web scraped from Internet Movie Scripts Database, Scripts.com, and SubsLikeScript.com
    - Final features that made it into the model were...
        - Raw text data converted to numerical values
        - Total word count
        - Words per minute
        - Unique word count
        - Vocabulary Diversity
- Film was considered success or failure based on profit, critic ratings, and audience ratings
- Methodology for the project outline was ROSE-MED

What Constitutes a Film’s Success?

- 65% ROI Metric
  - Profit = Worldwide Box Office Gross - Production Budget
  - ROI = Profit / Production Budget
- 20% Audience Score
  - User score from IMDB.com
- 15% Critic Score
  - Aggregate score from Metacritic.com
    - Combines review scores from dozens of established critics

Model Results

Model Results



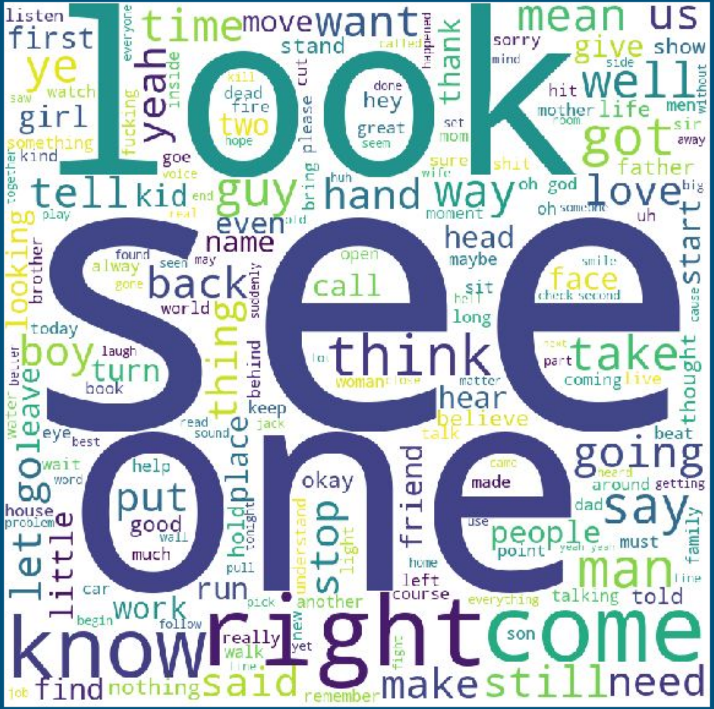
	model_#	model_type	train_accuracy	test_accuracy	negative_recall	positive_recall	negative_f_1	positive_f-1	notes
0	1	Vanilla XGBoost RFC	67.5	57.8	62.0	54.0	58.0	57.0	Vanilla model very basic, needs tuning
0	2	XGBoost GridSearched RFC	61.1	57.5	59.0	56.0	57.0	58.0	GridSearch performed better, no longer overfit...
0	3	Vanilla SVC	60.8	58.3	62.0	55.0	59.0	58.0	Fit better than RFC, continue with GridSearch
0	4	GridSearch SVC	58.0	57.5	49.0	65.0	52.0	62.0	No longer overfits
0	5	Multinomial Naive Bayes	86.3	61.4	86.0	37.0	69.0	49.0	Massive overfitting

Evaluation Metrics by Binary Model Type

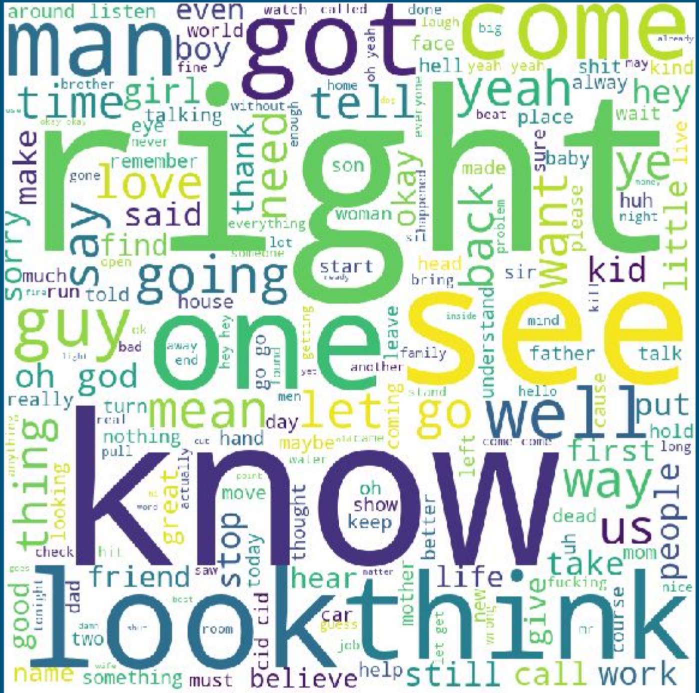
Exploring the Data

Word Clouds

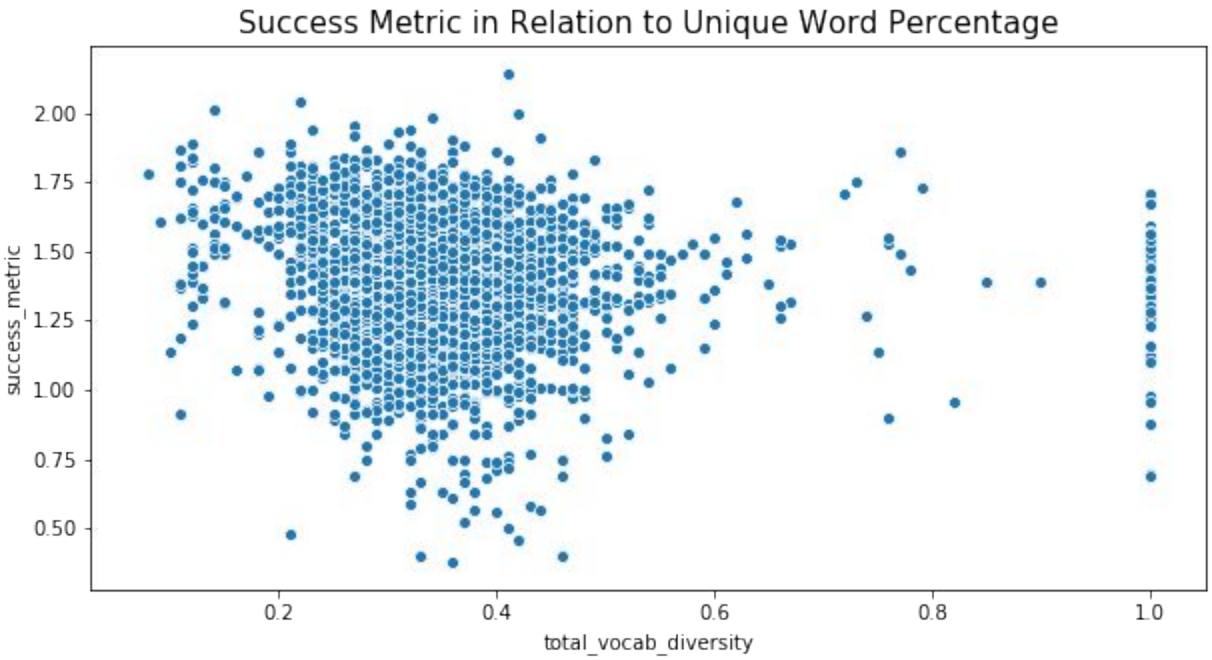
Success



Failure

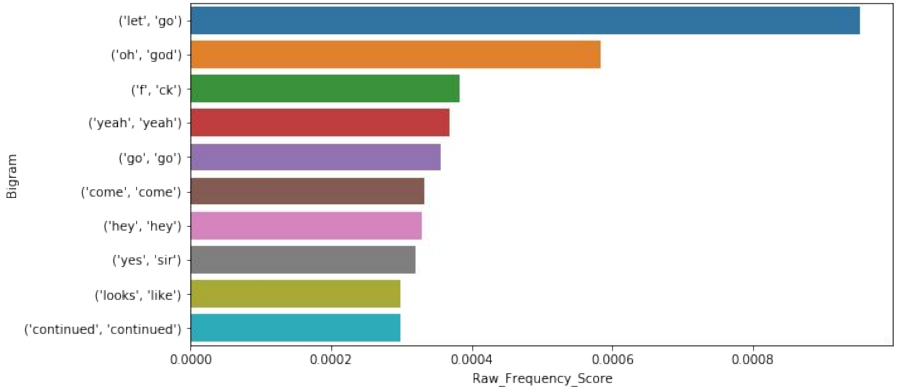


Success Metric / Vocabulary Diversity

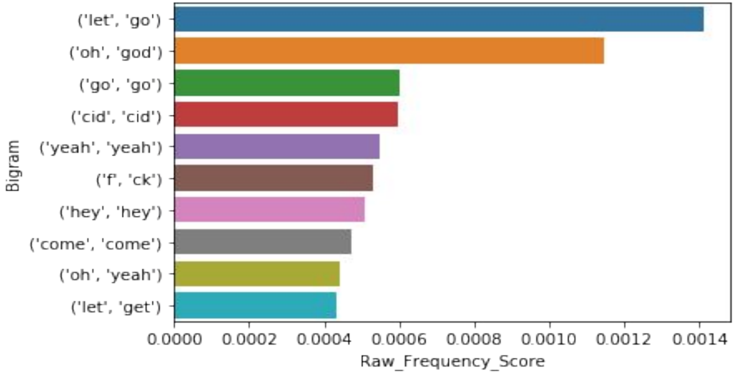


Bigrams

Success



Failure



Conclusion

Conclusion & Future Work

- Conclusion
    - The dataset that I gathered for this project was not large enough to reliably predict whether a film would be a future success
    - The model's output however leads me to believe that with additional data, a successful model could be built
- Future Work
    - Gather more data and re-train model
    - Break down scripts on a scene-by-scene level to model accurate future production budget prediction
    - Explore additional modeling options with neural networks to see if that boosts accuracy

Thank you and I look forward to working together!