26 SEASONS OF SIMPSONS:

PREDICTING LOW, MEDIUM, OR HIGH EPISODE IMBD RATING

- LAURA HOWARD

564 FULL EPISODE SCRIPTS

Including IMBD rating per episode

1,306,503 WORDS

Plus additional scene direction notes

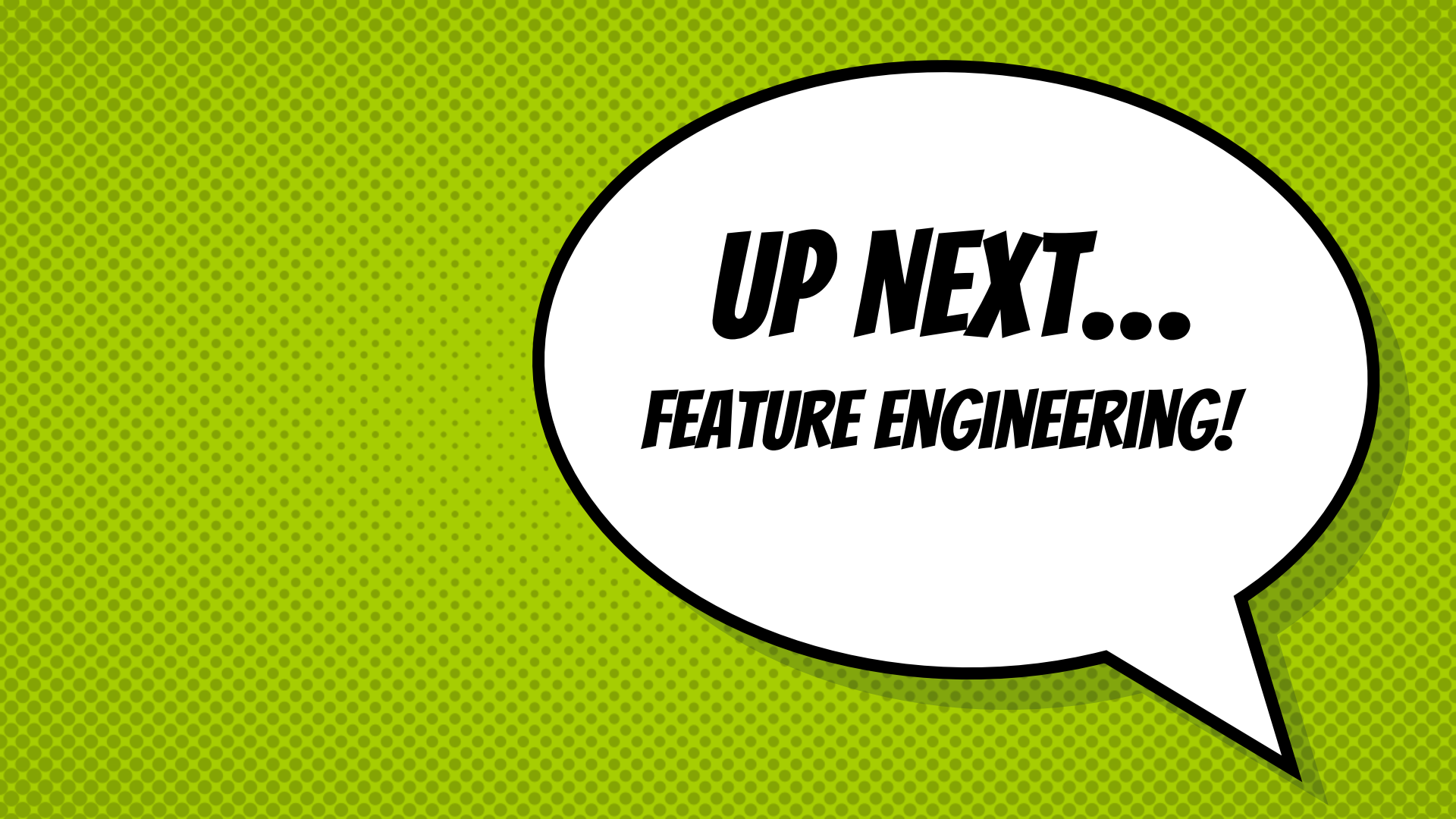
6,722 CHARACTERS

That's a lot of guest stars!

USE CASE

Can producers determine prior to full production if viewers will like an episode based on the script?

- Cut episode scripts predicted to be low
- Budget episode spend based on higher expected ratings

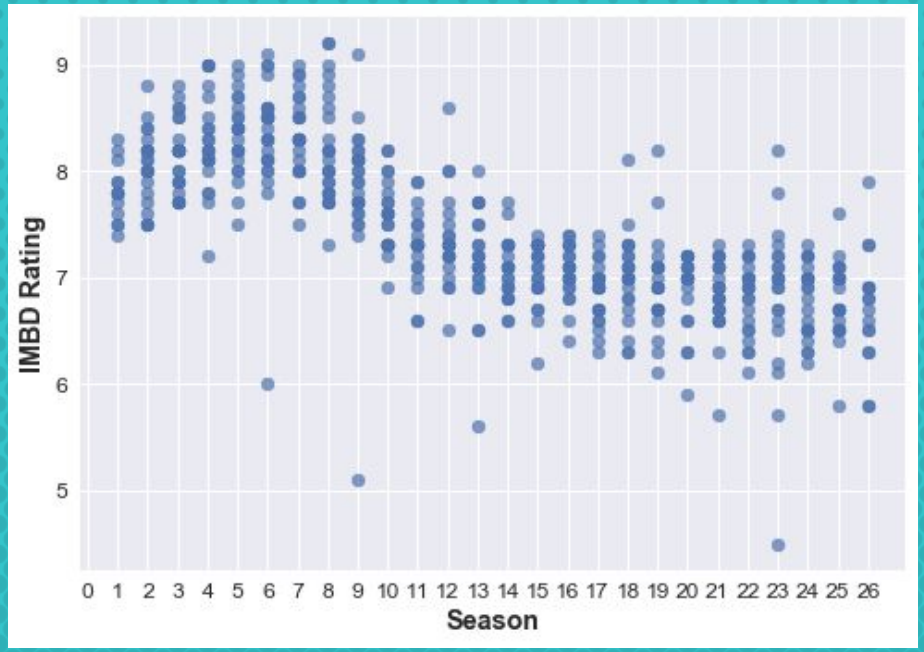

UP NEXT...
FEATURE ENGINEERING!

CHARACTERS - TALKING TIME & APPEARANCES

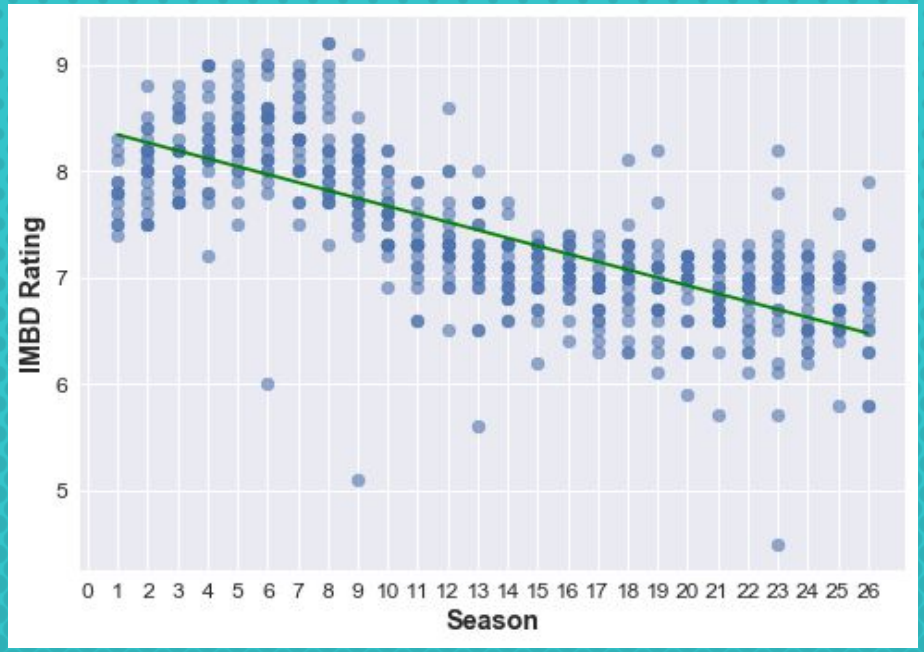
LOCATIONS - TIME SPENT

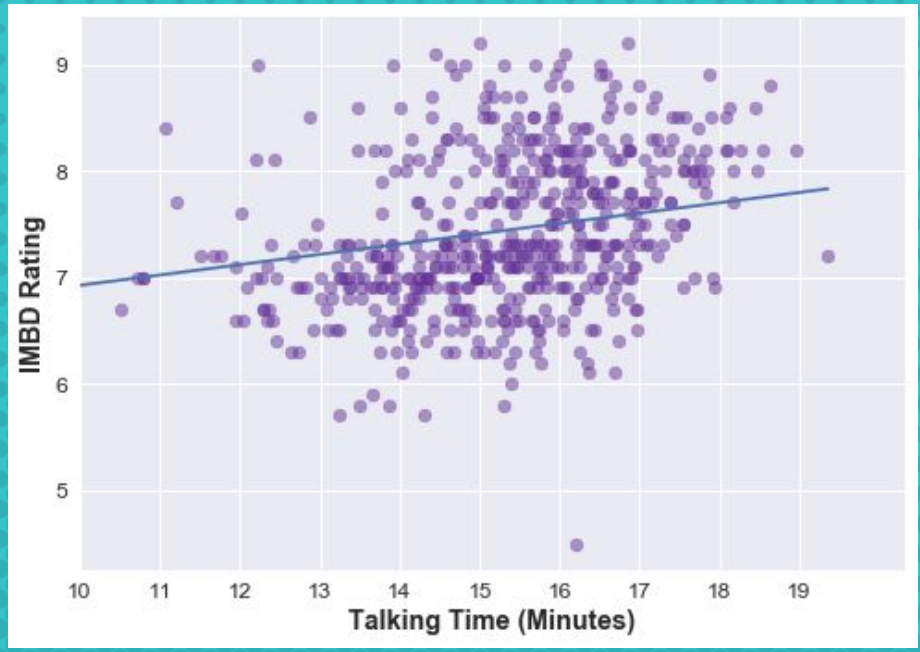
1000 +

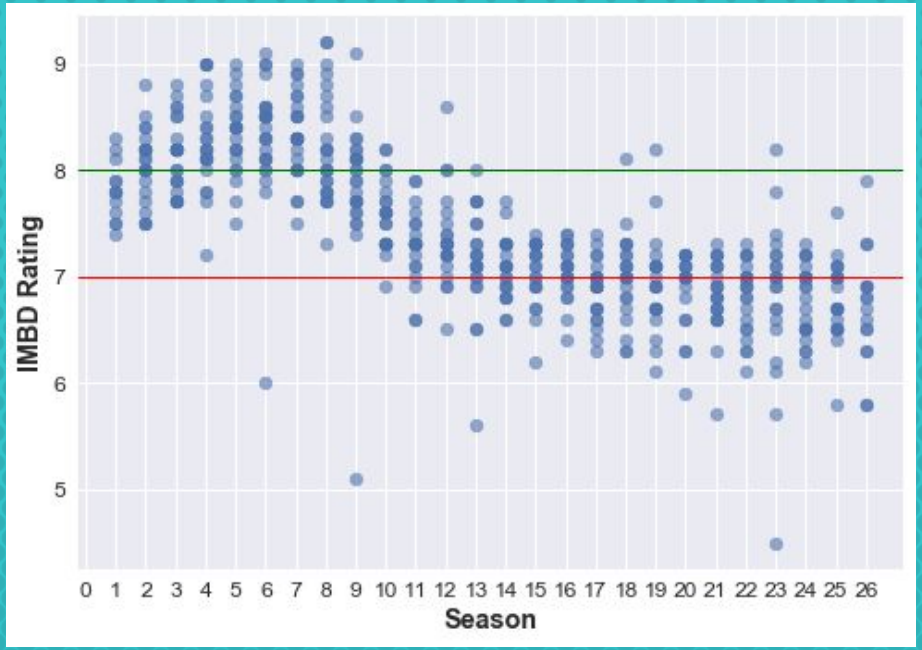
FAMOUS GUEST STARS


IMBD Rating
9
8
7
6
5
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Season

IMBD Rating
9
8
7
6
5
10 11 12 13 14 15 16 17 18 19
Talking Time (Minutes)


IMBD Rating
9
8
7
6
5
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Season


IMBD Rating
9
8
7
6
5
10 11 12 13 14 15 16 17 18 19
Talking Time (Minutes)


IMBD Rating
9
8
7
6
5
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Season

IMBD Rating
9
8
7
6
5
10 11 12 13 14 15 16 17 18 19
Talking Time (Minutes)

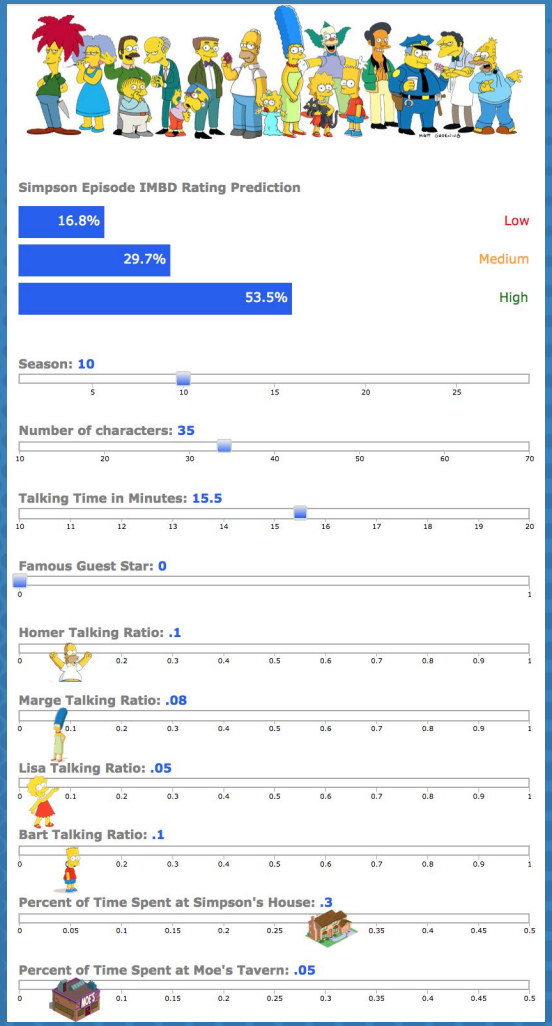
"Well, it doesn't matter how you live or what you did wrong.

As long as you're on TV, people will respect you."

- Bart Simpson

COMPARING CROSS VALIDATED MODELS

Accuracy Scores	tf-idf Features	Non-Text Features	Both Feature Sets
Logistic Regression	48%	66%	66%
Random Forest	43%	67%	65%
Gradient Boosted	NA	66%	66%



RANDOM FOREST CV MODEL

How well can we predict?

- Accuracy ~ 67%
- F1 Score ~ 70%
- About twice as good as chance

CONCLUSIONS

Most Important Features:

- Season
- Total Talking Time
- Marge
- Homer
- Lisa
- Bart

What Next?

- Better guest star feature
- Acquire scriptwriter data
- Add additional text analysis features

Any questions?