NEXfy

Subreddit Problem

SAIFUL HASAN

NEXfy

Problem Statement

I am a cofounder of a small tech company which wants to develop an online streaming platform naming 'NeXfy' with a desire to compete with Netflix and Spotify. We want to serve the consumer with both movies and music on a same online platform. We also want to place a blog section where user can leave their reviews, stories and start any discussion.

As a part of it we want to establish a model which can segregate the posts based on their title so when someone searches for ‘movies’ or ‘music’, the relevant posts show up for them. Also, the model will help us to understand the users’ usage pattern, demand in future fore further business expansion. As a part of this project we want to try different classification models and select the best model that satisfy our target. We have chosen to use classification metric, accuracy score to select the best performing model.

Work Flow

- Acquire Data
- Clean Data
- Model Preparation
- Modelling
- Model Selection
- Model evaluation
- Conclusion
- Recommendation

NEXfy

Data

Submission from 2 subreddits

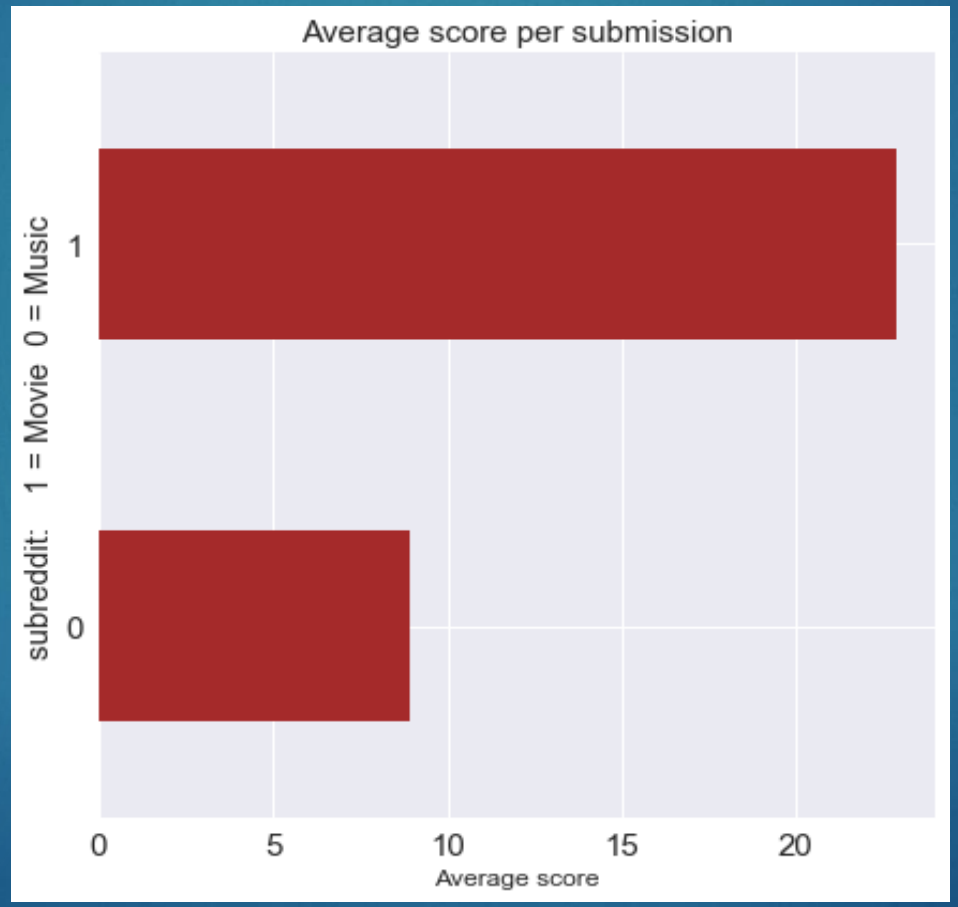
1. Movies
2. Music
3. Total submissions (around 4800)
4. Target variable (Title of the submissions)

Cleaning

- Remove non-letters
- Convert to lower case
- Remove hashtags
- Remove HTML special entities (e.g. &)
- punctuation
- Remove hyperlinks
- Remove whitespace (including new line characters)
- lemmatizer

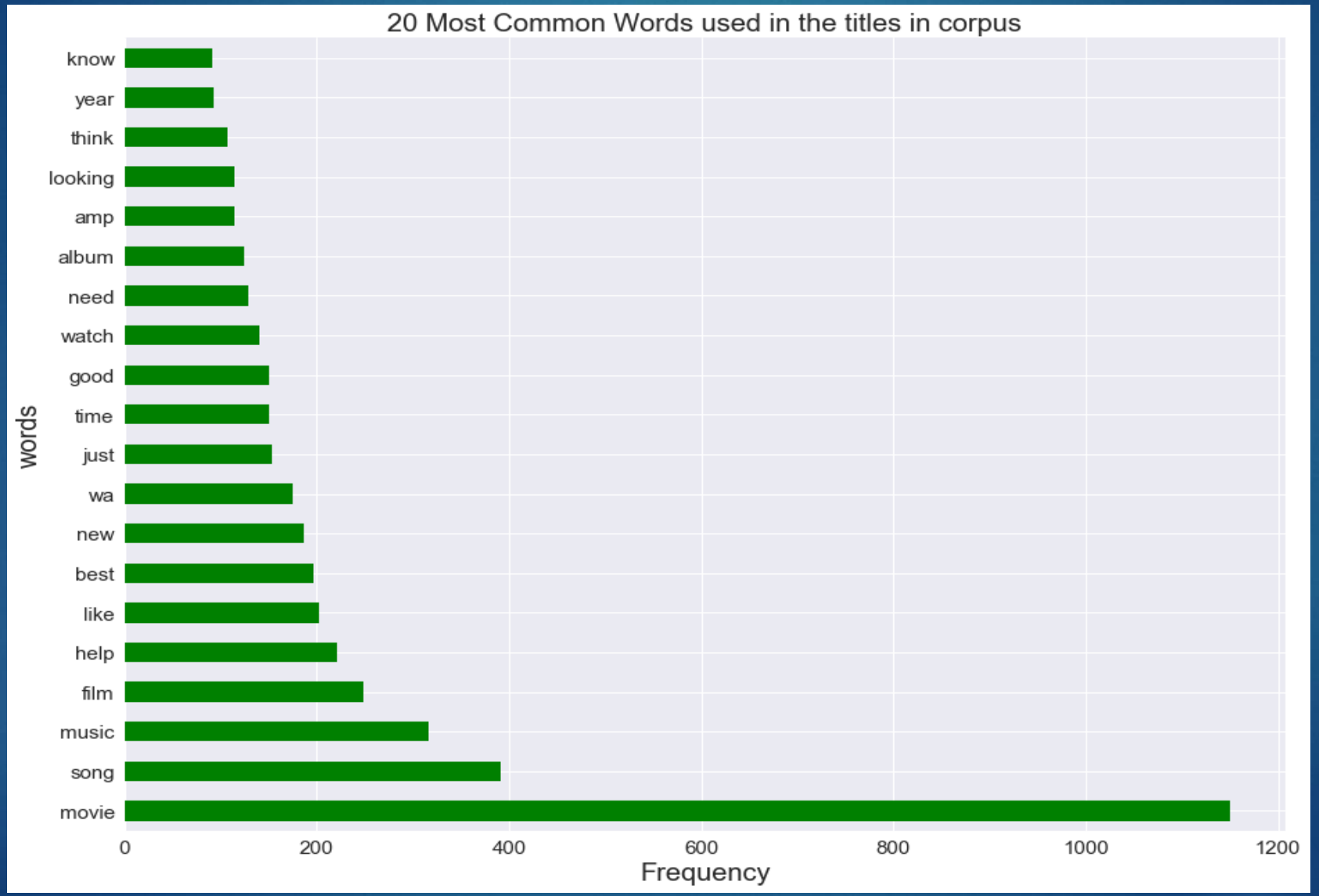
NEXfy

Average Scores



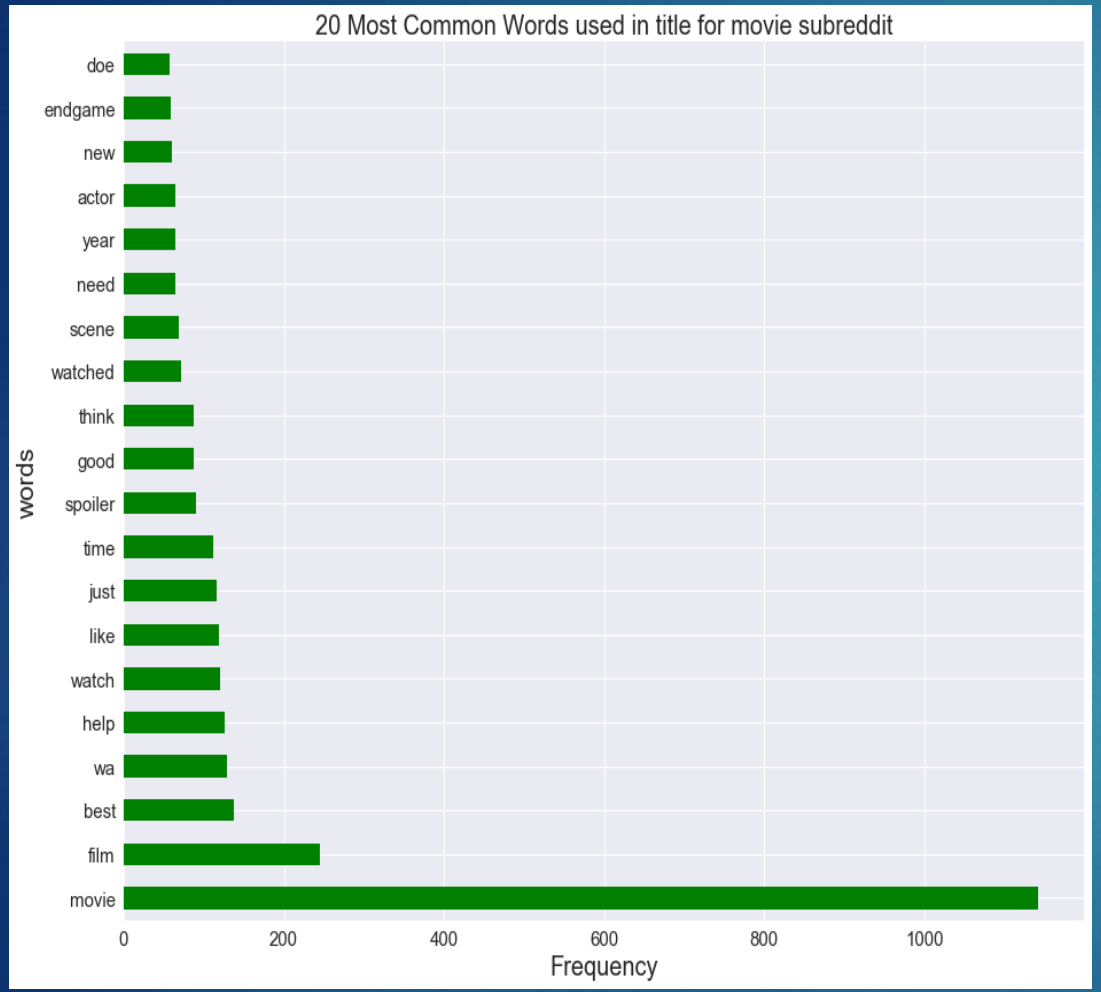
NEXfy

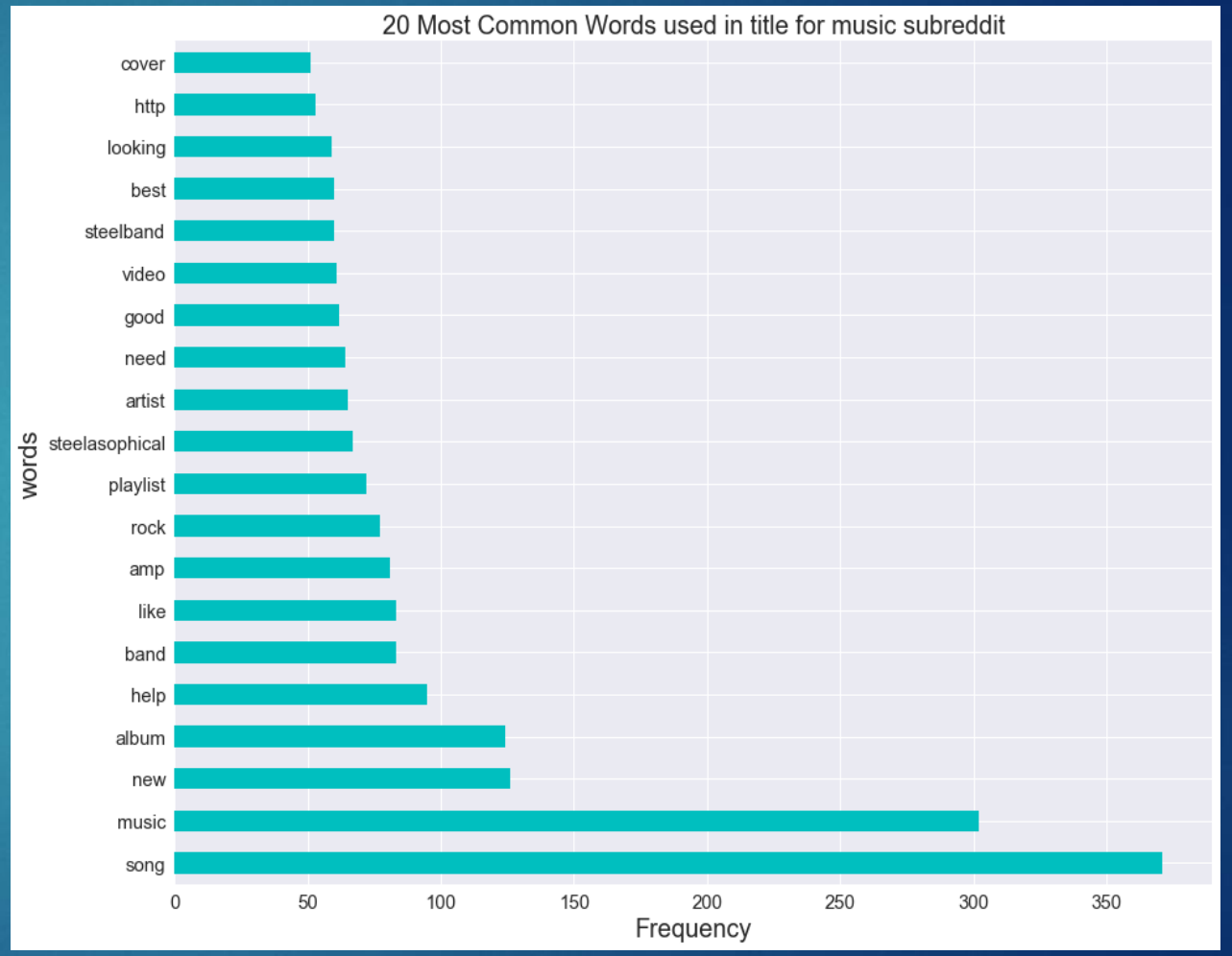
Frequency of words



NEXfy

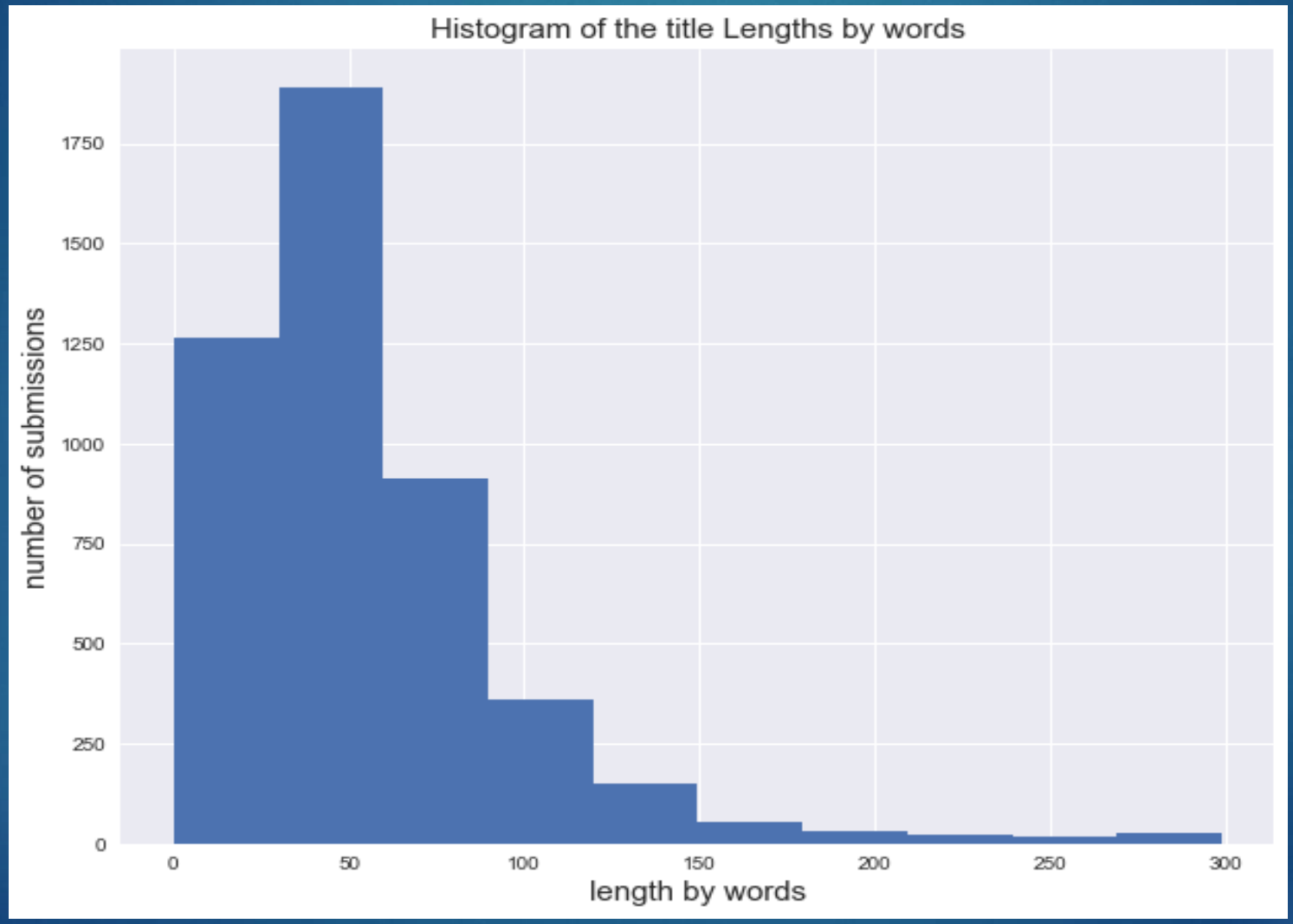
Frequency of words





NEXfy

Histogram of the title length



NEXfy

Models

1. Baseline Model
2. CVEC + Logistic Regression
3. TF-IDF + Logistic Regression
4. CVEC + KNN
5. TF-IDF + KNN
6. CVEC + Naive bayes (Multinomial)
7. TD-IDF + Naive bayes (Gaussian)
8. CVEC + Decision Tree
9. CVCE + Bagging Classifier
10. CVCE + Randomforest

NEXfy

Selection

Model	Accuracy Score
1. Baseline Model	57%
2.CVEC + Logistic Regression	87%
3. TF-IDF + Logistic Regression	88.39%
4. CVEC + KNN	81%
5. TF-IDF + KNN	73%
6. CVEC + Naive bayes (Multinomial)	88.02%
7. TD-IDF + Naive bayes (Gaussian)	77%
8. CVEC + Decision Tree	78%
9. CVCE + Bagging Classifier	84%
10. CVCE + Randomforest	79%

We have chosen accuracy score as a matric to select our model. Accuracy score refers to the percentage of observations the model predicts correctly.

NEXfy

Confusion Matrix

- True Negative: 718
- True Positive: 957
- False Negative: 132
- False Positive: 88

132 False negative score suggests that our model predicted 132 observations as music subreddits while they were actually movie subreddits. on the other hand, False Positive scores suggests that our model predicted 88 models as movie subreddits while they were actually music subreddits

NEXfy

Coefficients

Movie	9.545260
film	4.404036
spoiler	2.122880
watched	1.935990
actor	1.921349
trailer	1.555715
Watch	1.529863
scene	1.511531
Endgame	1.441977

genre	-1.706972
pop	-1.717974
artist	-2.090984
playlist	-2.434930
rock	-2.714338
band	-2.765547
album	-3.785590
music	-5.602808
song	-5.769362

The coefficient for 'movie' word is 4.064 which refers that for a unit increase of the presence of word 'movie' in title, an observation is $e\beta 1 = e9.55 = 14045 e\beta 1 = e9.55 = 14045$ TIMES AS LIKELY to be a movie subreddit. So basically as the word 'movie' occurrence increases by one unit in the title, an observation is 14045 TIMES AS LIKELY to be a movie subreddit.

On the other hand, The coefficient for 'song' word is -5.769 which refers that for a unit increase of the presence of word 'song' in title, an observation is $e\beta 1 = e - 5.769 = 99.7 e\beta 1 = e - 5.769 = 99.7$ percent LESS LIKELY to be a movie subreddit.

NEXfy

Conclusion & Recommendation

From the findings of the TF-IDF with a logistic Regression model I have come to these recommendations:

1. We can start our blog posting project on the basis of our analysis findings. Our model delivers expected results
2. for maximizing accuracy score we need to minimize the false negative and false positive scores.
3. Since we didn't have 100% accuracy we may consider second layer of filtration.