Reddit Posts Prediction Model

ANKITA PATIL

Problem Statement

Can we identify different used by novice vs experts in the Engineering field and use them to predict subject matter difficulty level of a post?

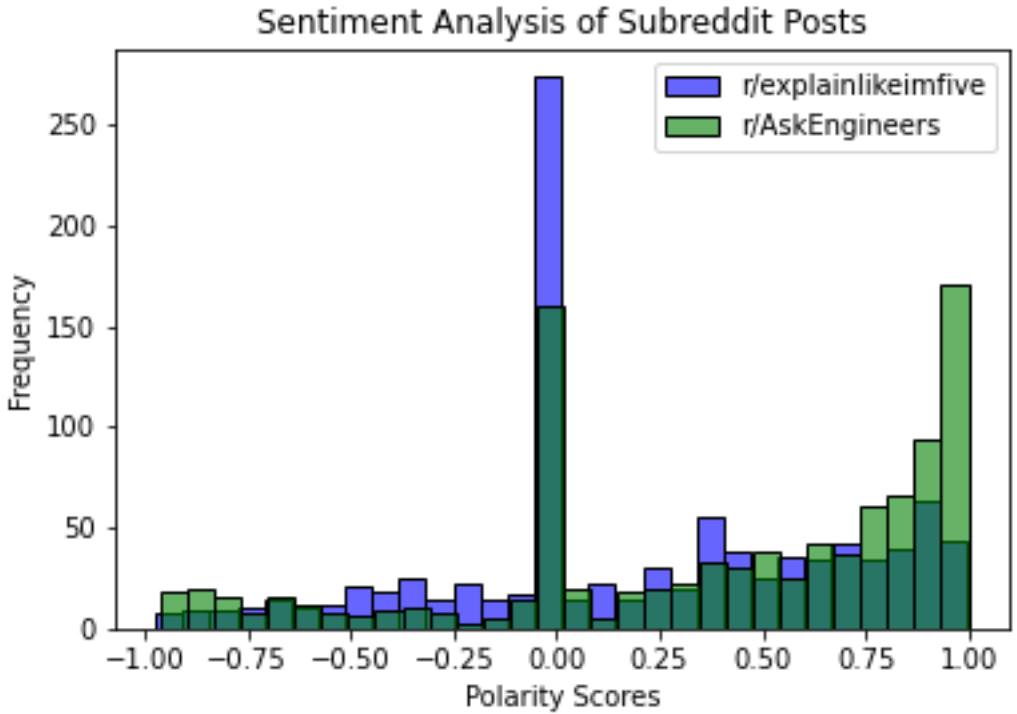
Data Gathering

- Scraped 2000 Reddit Posts - Pushshift API - r/explainlikeimfive and r/AskEngineers
- Language - layperson-friendly vs technical
- Data – title, selftext, subreddit, created_utc

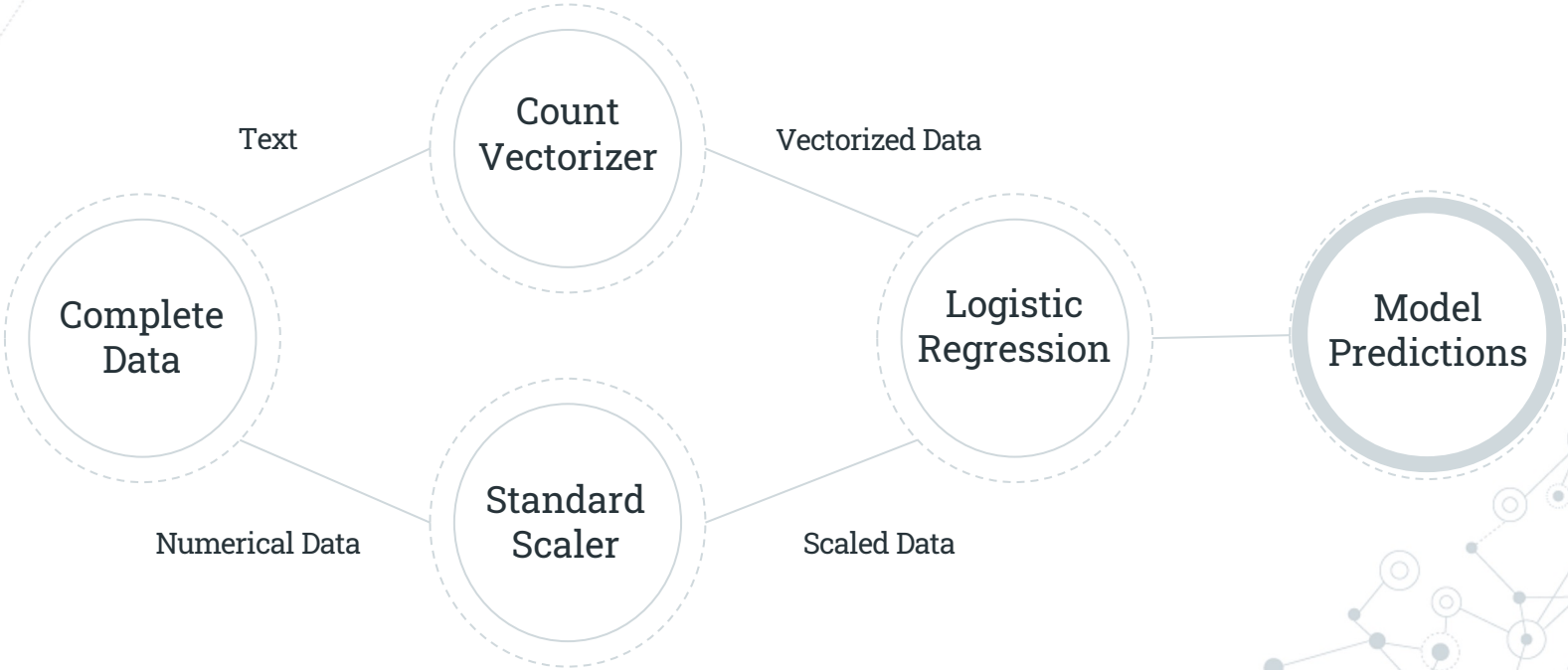
Data Processing

- Data Cleaning
- Feature Engineering
- Natural Language Pre-Processing

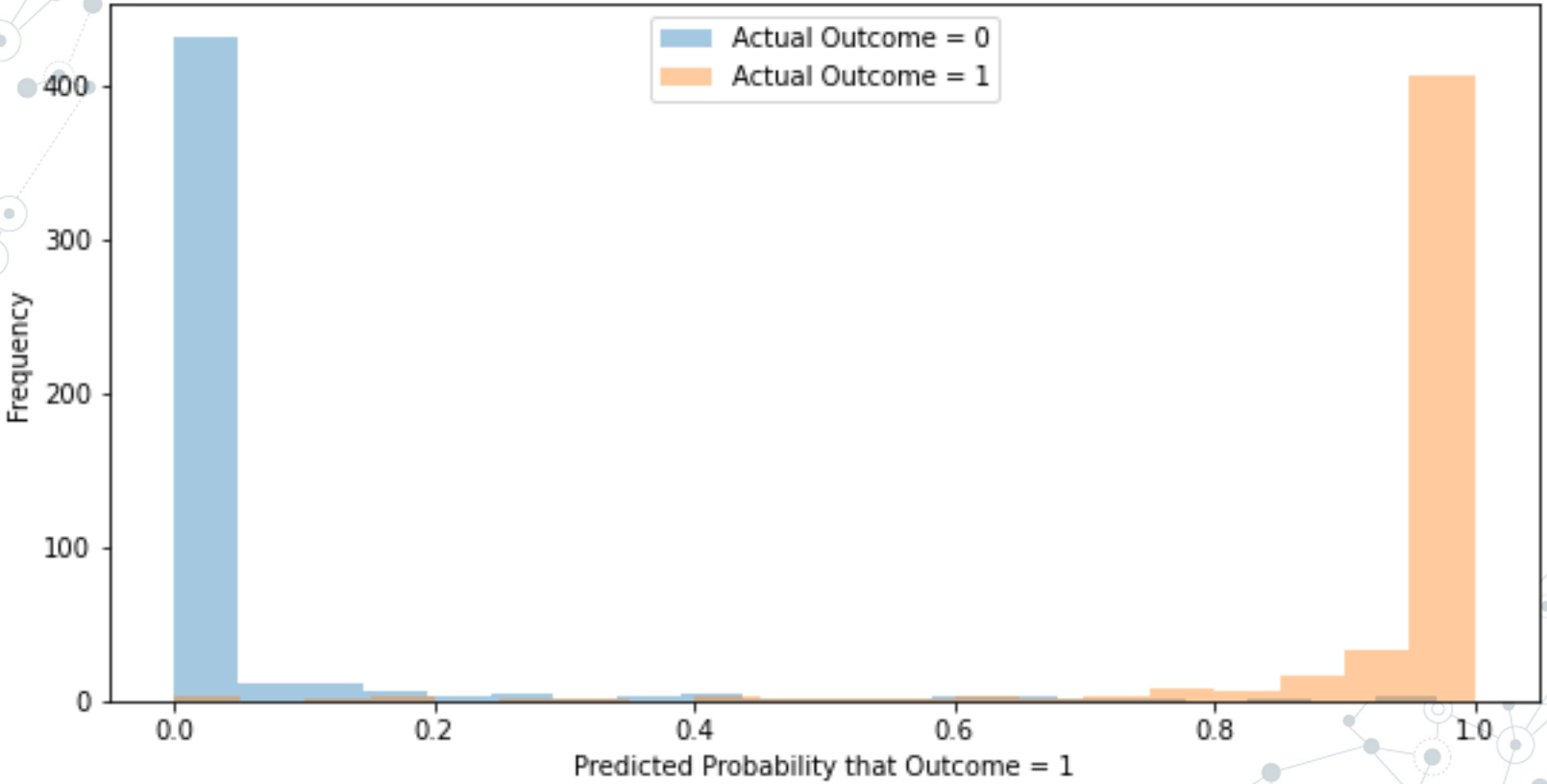
Sentiment Analysis



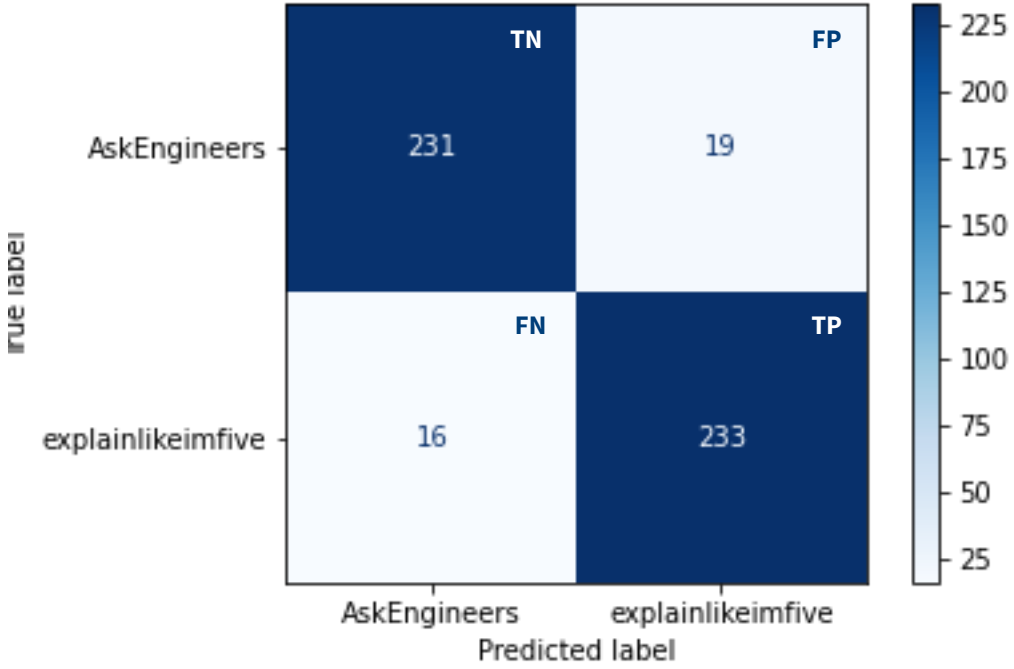
Model Walkthrough



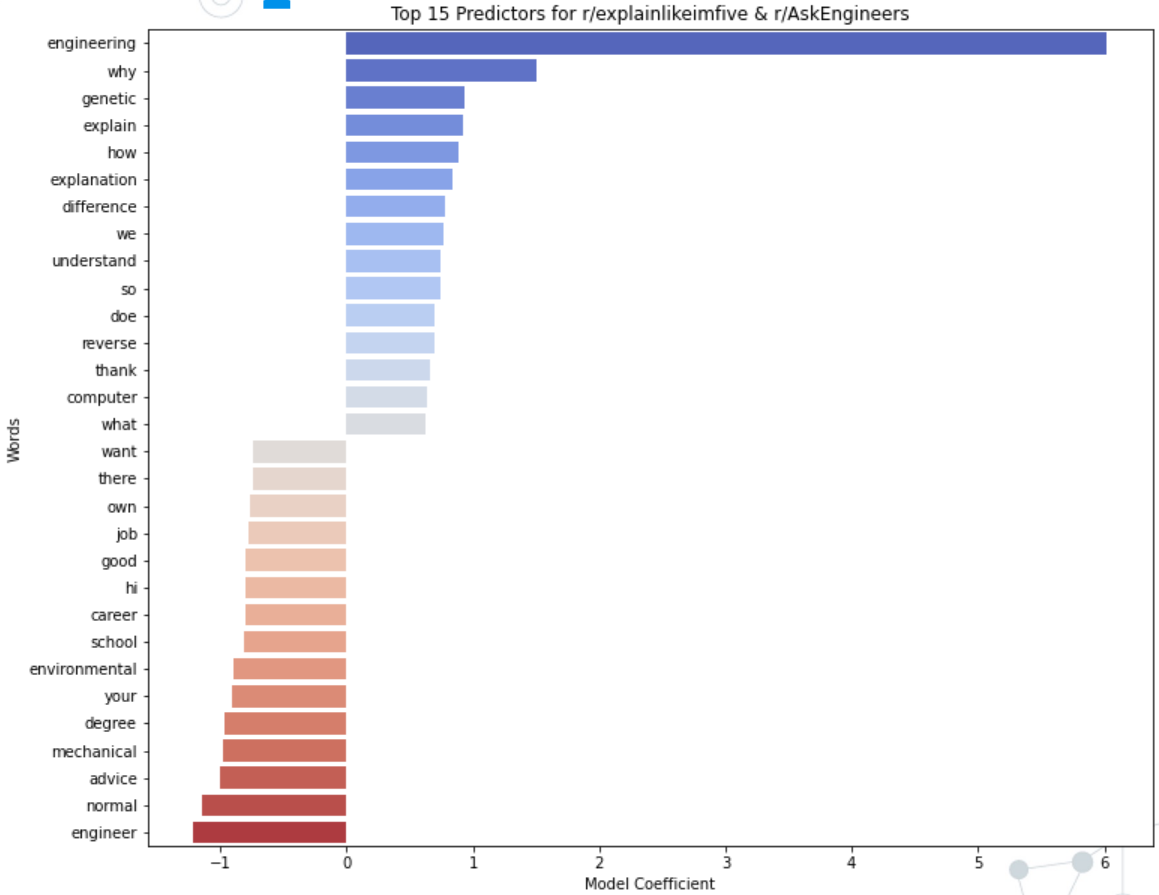
Predictions of the Model



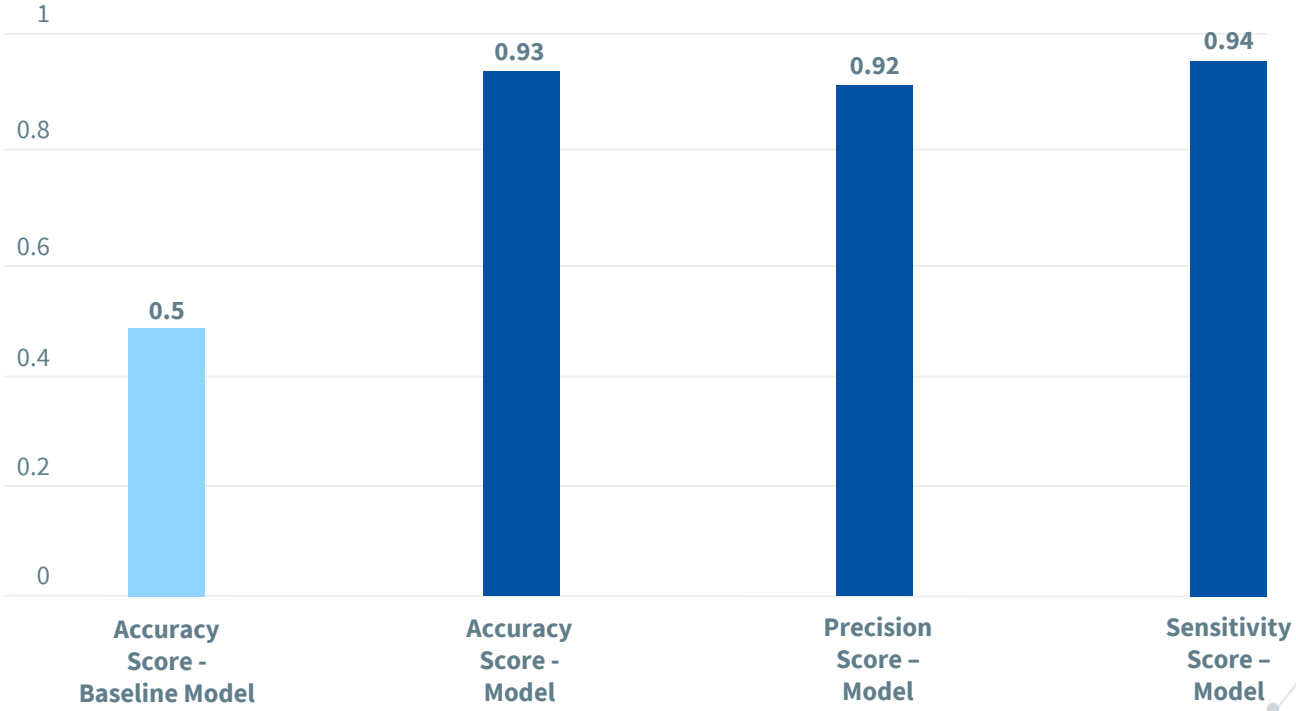
Confusion Matrix



Top Model Predictors



Evaluation Metrics



Conclusion

- The model accuracy - 93%
- r/explainlikeimfive - trying to understand things with why, how and what
- r/AskScience - specific to the field
- Based on the words, a post can be categorized into novice or expert level

Recommendation

- Model can be improved by incorporating multiple algorithms
- Model can be scaled and used to categorize articles on the level of subject matter difficulty, could be used alongside *reading time* feature.

Thanks!

Any questions?