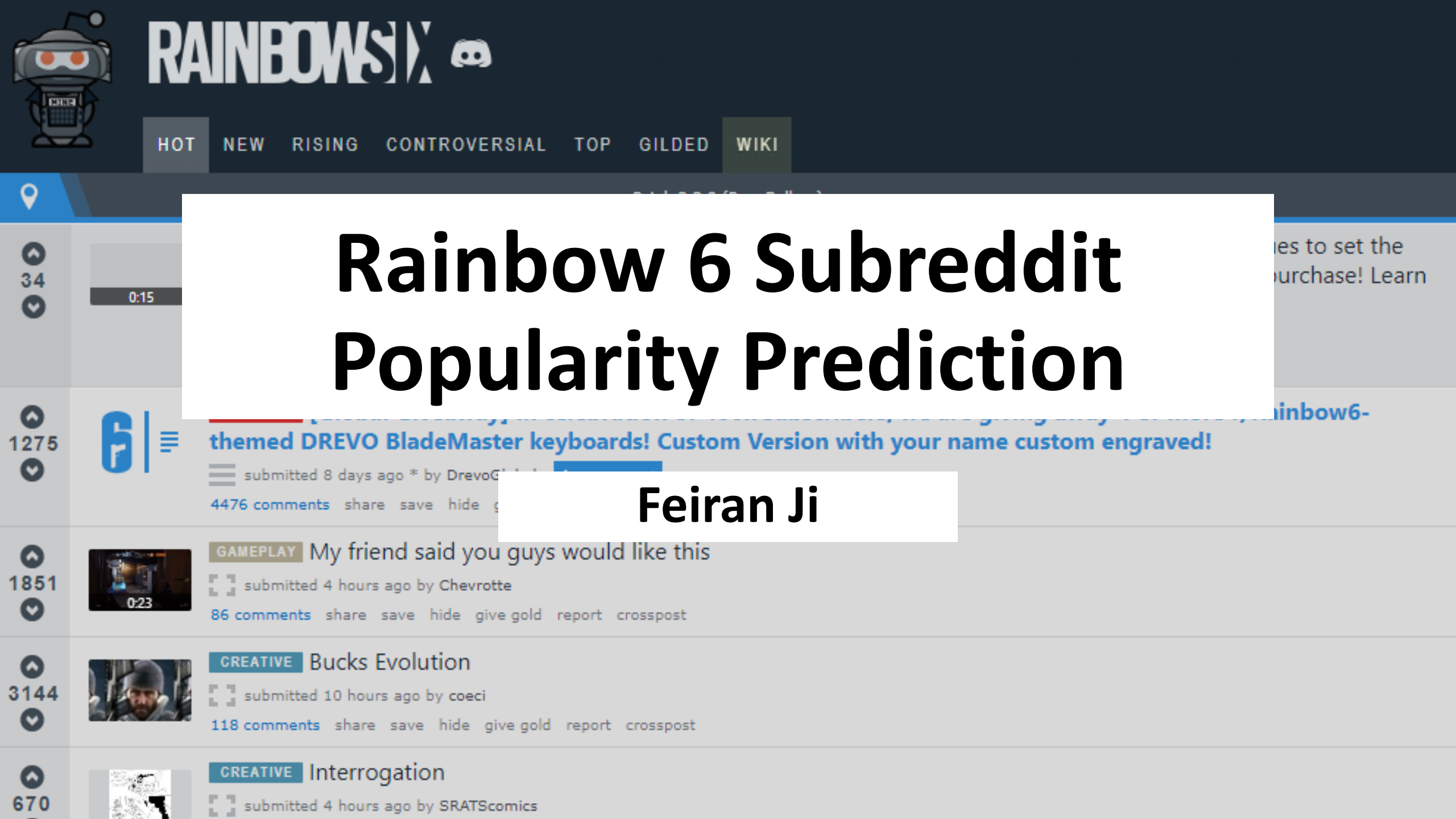
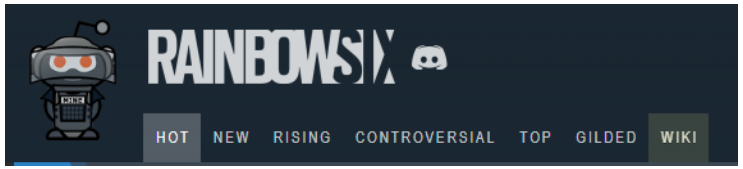
Rainbow 6 Subreddit Popularity Prediction

Feiran Ji

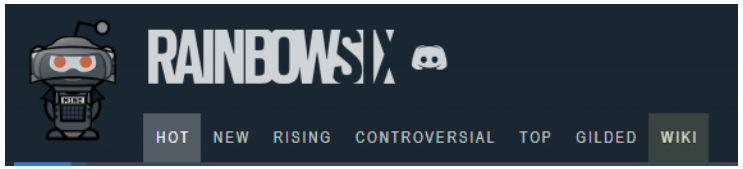


Background & Motivation



- **Ubisoft**, a French video game publisher, is my practicum company and I have been working here for 10 months.
- **Rainbow 6 Siege** is an online tactical shooter video game developed and published by Ubisoft. It has the biggest community (> 35 million) and generates most revenue among all the games in Ubisoft.
- **Rainbow6 subreddit** (https://www.reddit.com/r/Rainbow6/) is a community for R6 fans to discuss Rainbow 6 Siege.
- Understanding R6 Subreddit will be helpful for the Ubisoft R6 brand team, marketing team and product team to develop marketing strategy, improve the game experience and create a better relationship with players.
- Here, I want to investigate the popular posts (number of comments > 10) and try to gain some insights about what make some posts popular.

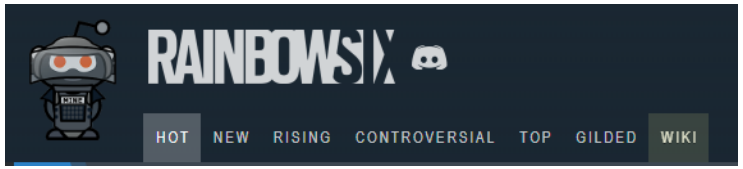
Data



- R6 Subreddit has more than 404,000 subscribers. Although it's a small proportion of the 35 million players, they should be some active and representative players.
- From 2015 to 2018.2, R6 Subreddit has 478,460 posts. It's too much for this project, so we subset the posts from 2017.12 – 2018.2
- A lot of posts are too short or non-text (e.g. picture, video, etc.). We decided to filter out them and only keep posts longer than 50 words.

num_comments	score	title	selftext	created
12	0	The problem that is ruining ranked multiplayer...	So let's allow people to make as many fucking ...	2018-02-11 00:10:10
9	2	Jackal,Buck,Blackbeard or Capitão	Rank the following attackers from best to wors...	2018-02-11 00:22:00
2	0	Vigil's elite animation	He should disappear and then make his team dis...	2018-02-11 00:25:58
4	2	What Channel do we need to watch on Twitch for...	I linked my Twitch and Ubisoft account and I w...	2018-02-11 00:30:49

Class Imbalance

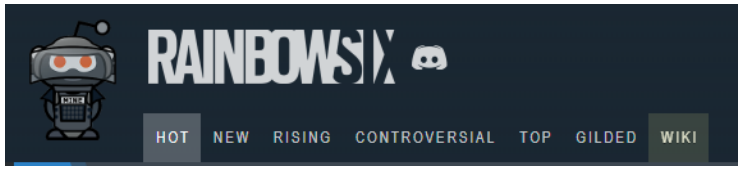


- Use F1 score to evaluate the classification model, also see precision and recall for insights.
- Try different threshold in predicted probability for classification.

```
r6s['popular'].value_counts()
```

```
False    23602
True      5468
Name: popular, dtype: int64
```

Models



- Here, we chose CountVectorizer as our vectorizing technique because we want to have more insights about which words are important in classifying popular and non-popular posts.
- The modeling techniques we have tried include:
  - Logistic Regression
  - Random Forest
  - Gradient Boosting
  - Naïve Bayes
- The best model: Random Forest (n_estimators=50, min_samples_leaf=3)

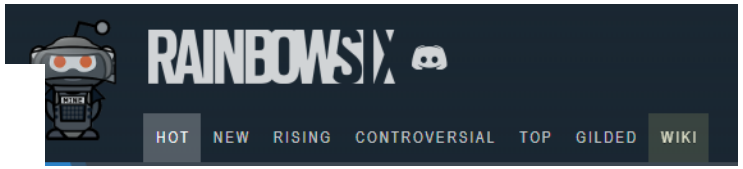
Models

Model	F1 score	Precision	Recall
Logistic Regression	0.241	0.182	0.353
Logistic Regression (After tuning)	0.242	0.183	0.357
Random Forest	0.295	0.315	0.277
Random Forest (After tuning)	0.377	0.275	0.598
Gradient Boosting	0.304	0.292	0.316
Gradient Boosting (After tuning)	0.356	0.301	0.436
Naïve Bayes	0.204	0.184	0.229

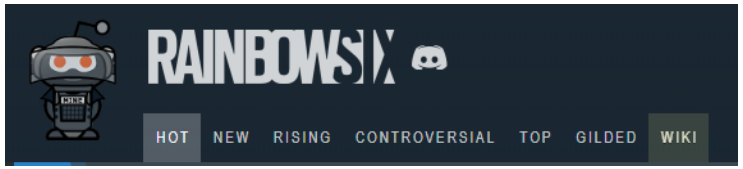
Insights

	cols	imp
1380	edit	0.009468
4449	think	0.008569
2998	operator	0.006630
2511	like	0.005861
554	buff	0.005852
3135	people	0.005850
1816	game	0.005417
320	bad	0.005176
183	anyone	0.005130
2634	make	0.004992
1399	ela	0.004564
4826	want	0.004406
1790	fucking	0.004240
2881	nerf	0.004096
3520	recoil	0.004079

	cols	imp
1380	edit	0.043679
1790	fucking	0.018078
1681	fix	0.017476
4449	think	0.015828
1816	game	0.015599
4108	spawn	0.014629
554	buff	0.013266
1610	favela	0.012882
183	anyone	0.012785
4566	tried	0.012357
320	bad	0.012278
3227	please	0.011839
2877	needs	0.011309
220	arguably	0.010695
2511	like	0.010158



Future Work



- 1. Use Word2vec as vectorizer. We chose countvectorizer as the vectorizing technique before because we want some insights, we can also try word2vec and see if it generates a higher accuracy.
- 2. Create more features, for example, if the post contains a picture or video. ('is_self' column in the original dataset indicates if the post contains non-text component.)
- 3. Here, we use a threshold of probability to solve the class imbalance. We can also try upsampling or downsampling.
- 4. We can try more modeling techniques like SGD, SVM, XGBoost, etc.