

Make your Reddit post hot! - The next big hit at FiveThirtyEight!

Youn Hee Pernling Frödin



Reddit bridges communities and individuals with ideas, the latest digital trends, and breaking news (...okay, and maybe cats). Our mission is to help people discover places where they can be their true selves, and empower our community to flourish.



Share - Vote - Discuss



900,000,000+
comments in 2017

Source: https://www.reddit.com/r/help/comments/7zvcs1/how_many_comments_are_made_on_reddit_each_day/

Scraping the data

- Look at Reddit and see that the page is ok to scrape
- Scrape the data
 - 22 pages
 - Reddit API also said to put a user name in
 - Print text to see the progress
 - Make sure you get all pages. Try for each page until you do not get “Too Many Request”.
 - Scrape with BeautifulSoup.
 - Make sure to find next page to scrape through the page that you are scraping
- Make the time for sleep between each loop
- Save data to Pandas data frame
- Save the data frame into a .csv file





Titles

Time up

Subreddits

Comments

Data munging, EDA and NLP

- Create dummies for subreddits
- CountVectorizer on the titles.
- Create binary variable:
 - Class 1 > median (47 %)
 - Class 0 <= median (53 %)

We do not have to balance classes
- Independent variables not highly correlated (good)
- Dependent and independent variables not either highly correlated (less good)



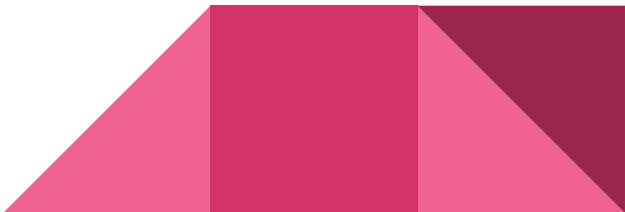
- Time_up
- Time_up + All words
- Time_up + All subreddit
- Time_up + All words + All subreddit
- Time_up + 10 most common words
- Time_up + 10 most common subreddits
- Time_up + 10 most common words + 10 most common subreddits
- Time_up + 10 most important words
- Time_up + 10 most important subreddits
- Time_up + 10 most important words + 10 most important subreddits
- Time_up + 5 most important words + 5 most important subreddits
- Time_up + 4 most important words + 4 most important subreddits
- Time_up + Different combinations with words and/or subreddits

Other tools

Logistic Regression

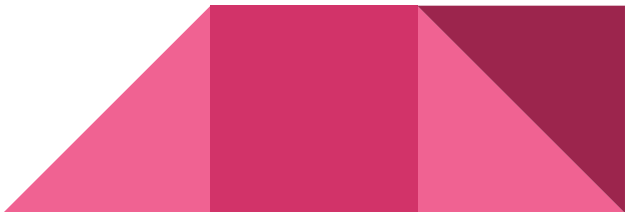
- Feature importance (words and subreddits)
- Tune hyperparameter (C)
- Scaling
- Max_iter, Solver, Penalty

Random Forest Classifier

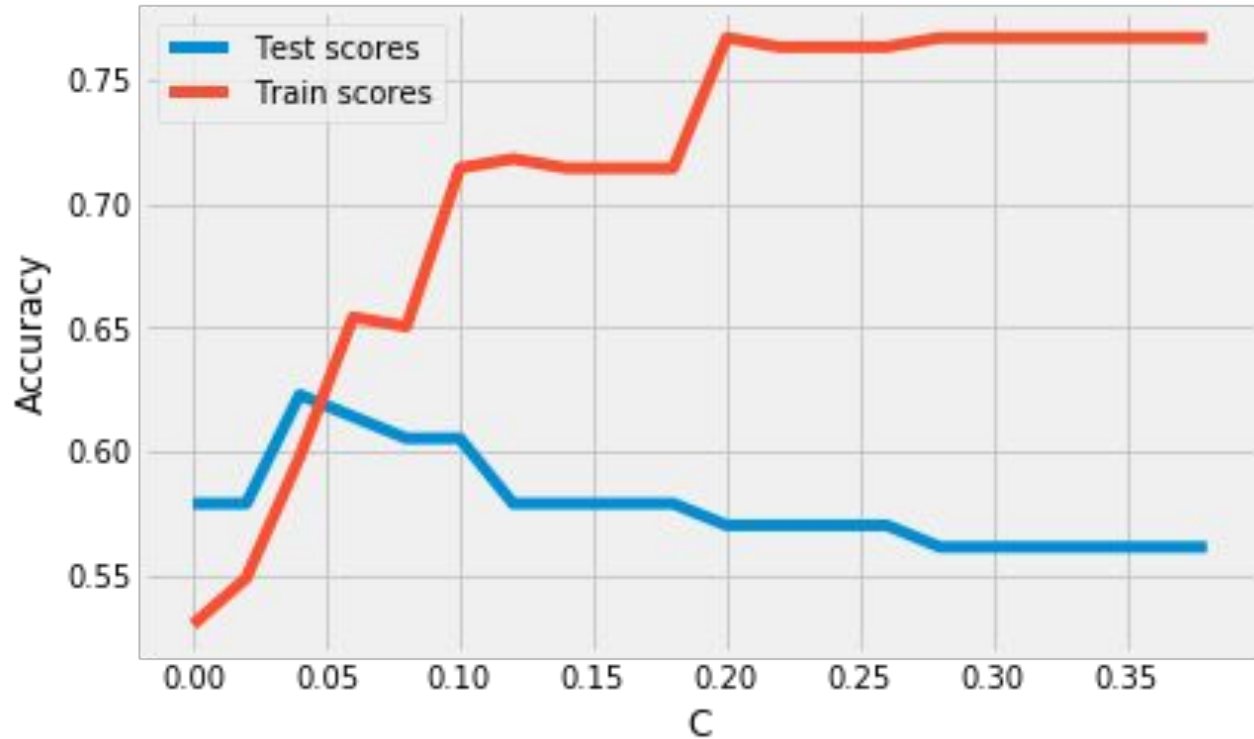
- Feature importance (words and subreddits)
 - Tune hyperparameters (max_depth)
 - N-estimators
- 

Best models part 1

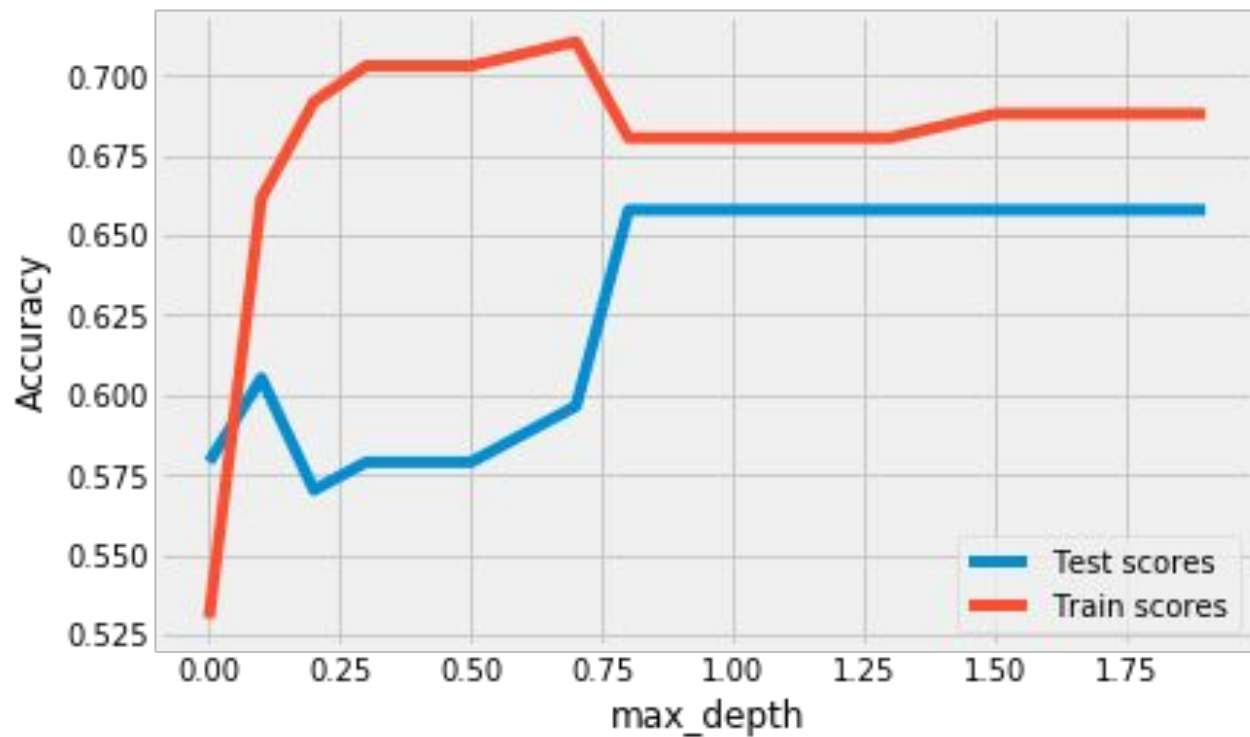
Model	Independent variables	Best hyperparameter	Score	Prediction
Logistic Regression	Time_up + all subreddit dummies	$C = 1e-06$	65.79 %	0.8158
Random Forest C	Time_up + 10 most important words and 10 most important subreddits	Max_depth = 7	59.65%	0.6842



Logistic Regression



Random Forest Classifier



Best models part 2

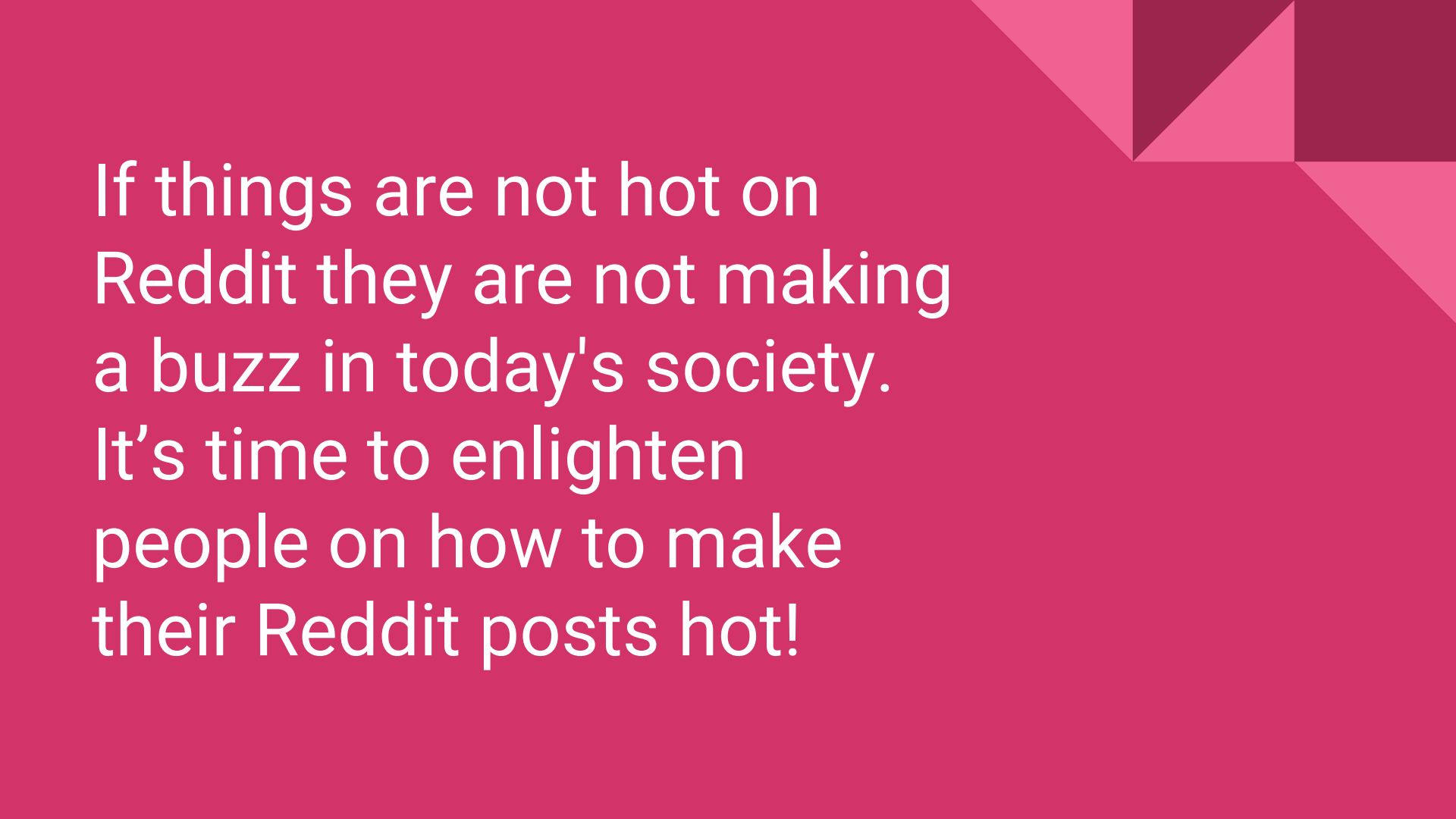
Model	Independent variables	Best hyperparameter	Score	Prediction
Logistic Regression	Time_up + all subreddit dummies	C = 0.4	65.79 %	0.6842
Random Forest C	Time_up + 10 most important words and 10 most important subreddits	Max_depth = 1	59.65%	0.6842



Longer time up

All subreddit dummies

You might want to choose subreddits that's not so dependent on
what's happening in the world



If things are not hot on
Reddit they are not making
a buzz in today's society.
It's time to enlighten
people on how to make
their Reddit posts hot!

Further studies

- This was just a prediction made from one scraping (hot posts at one time). To get an even better picture you need to do more scraping over a longer period of time.
- Look at 2-grams instead of 1-grams
- Is number of comments the best way to measure popularity? Likes better?





Thank you!