Reddit classifier: NLP and ML models

Udayshankar Menon

Objective

- Compare different classification models to classify tea and coffee subreddit posts (169172, 285090)
- EDA: To gain some perspective on the information at hand
- Fit and score various models
- Can we predict the number of upvotes based on the information at hand?
- Can we define a basis vector for each subreddit?

Methodology

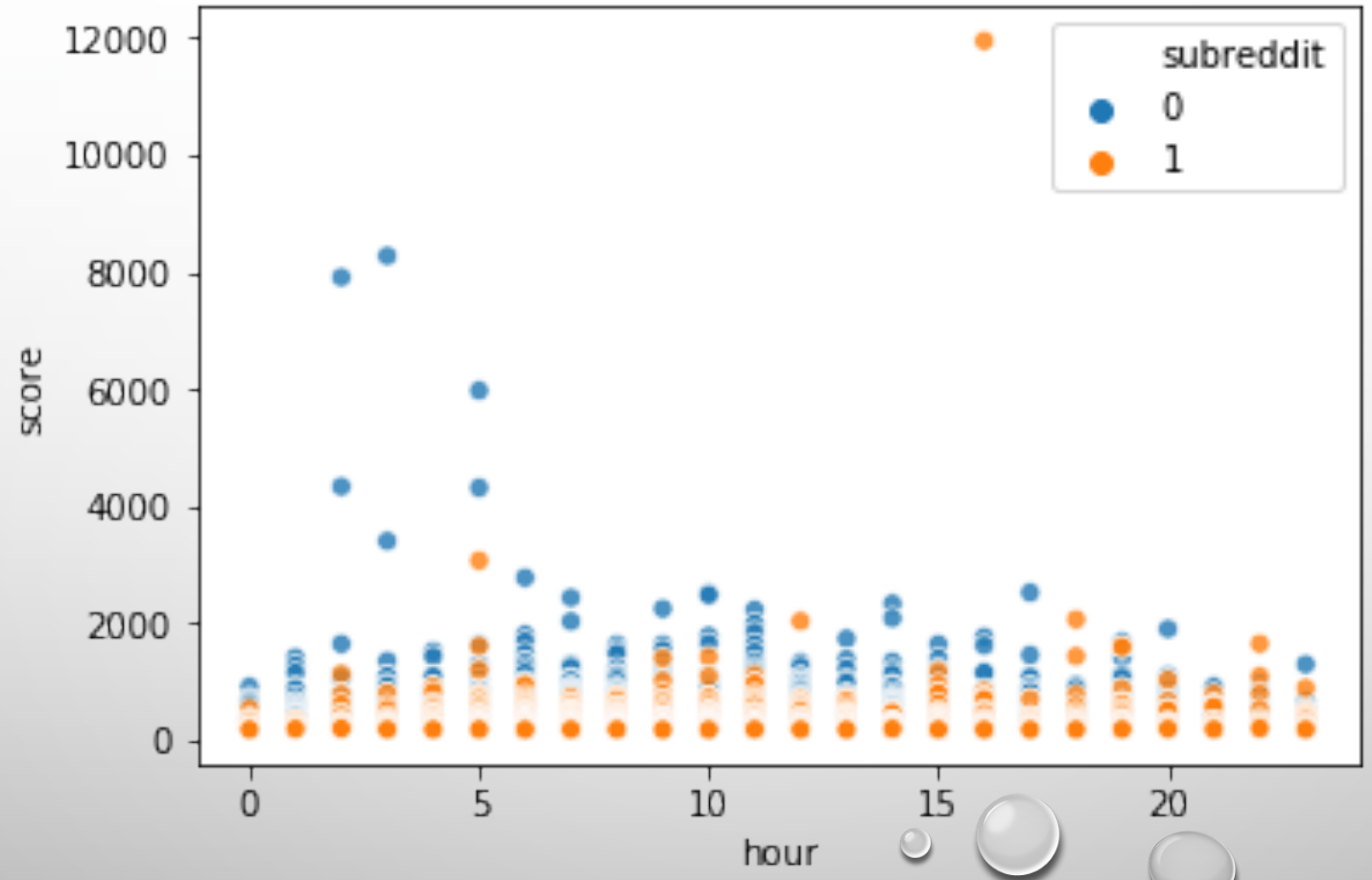
- Train and test split the combined data set with id for each home subreddit
- Fit various classifiers and text vectorizer models
- Features considered in this study:
  - Post titles
  - Month of posting
  - Time of posting
  - Upvotes on title

Expectations

- Correlation of post timing, subreddit, and score?
- Correlation of title length and upvotes?
- Predicting upvotes based on length of posts and other features?

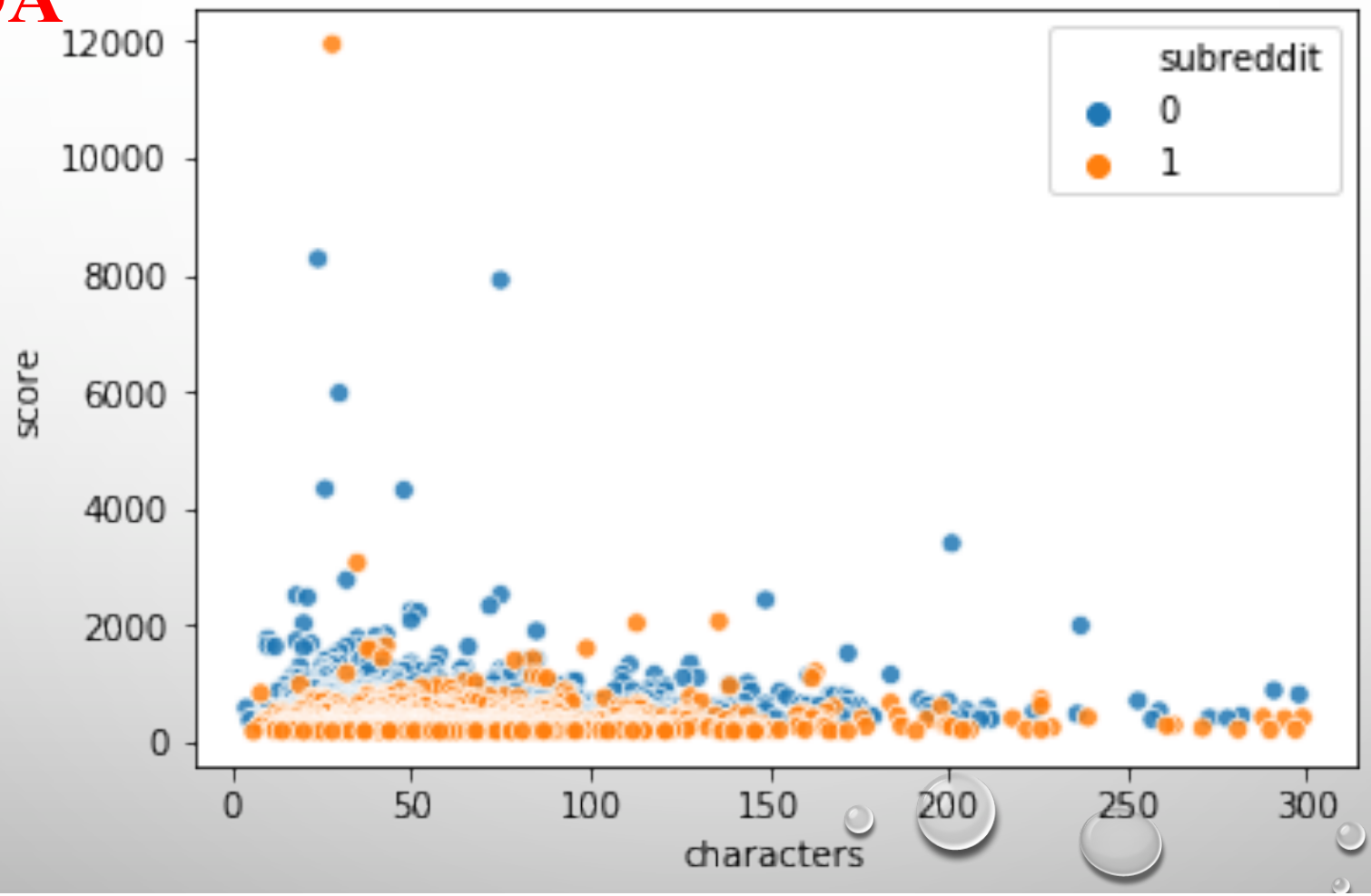
EDA

No significant or observable correlation of post timing, subreddit, and score



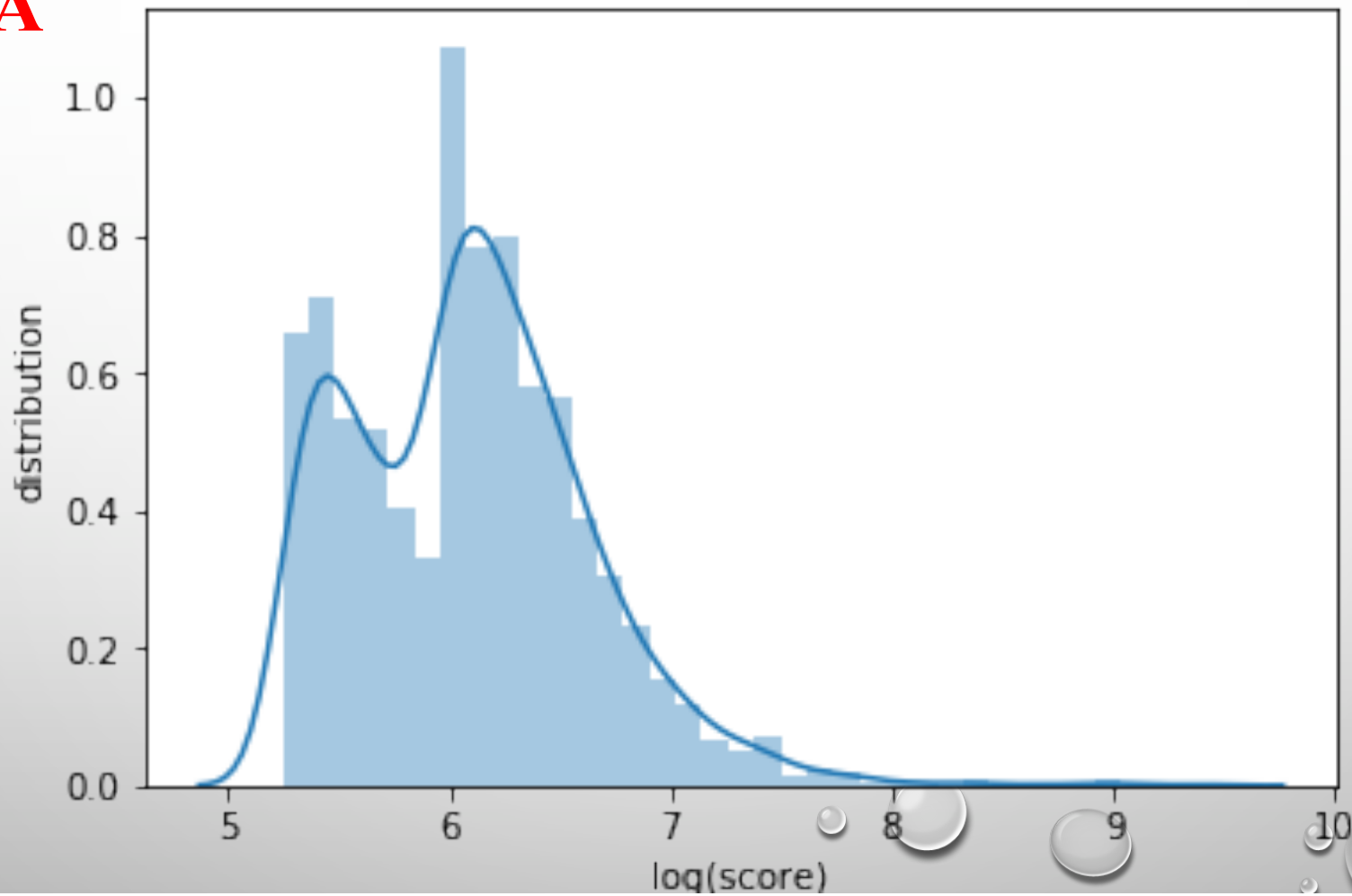
EDA

"Maybe" some correlation of score and title length



EDA

Score distribution has some pattern to it



Baseline accuracy

- Baseline accuracy: 50.40 %
- "*If the posts were randomly assumed to be from subreddit tea, there is a 50% possibility of it being a correct assumption.*"
- Our classes seem very well balanced.

You don't need a $100, 000 data scientist, a dollar and ignorance is enough!

Classification models

Id	Model	Vectorizer
1	Logistic Regression	Tfidf Vectorizer
2	Bernoulli Naïve Bayes	Count Vectorizer
3	Logistic Regression	Count Vectorizer
4	K-Nearest Neighbors	Tfidf Vectorizer
5	Decision Tree Classifier	Tfidf Vectorizer
6	Extra Tree Classifier	Tfidf Vectorizer
7	Bagged Decision Tree	Tfidf Vectorizer
8	Random Forest	Tfidf Vectorizer

Best parameters across models

- Vectorizers:

  Maximum features: 75
  N word grams : (1,1)
  Stop words : English

- K-Nearest neighbors: 5

- Tree classifiers:

  Max depth, min leaves, min split

Classification models

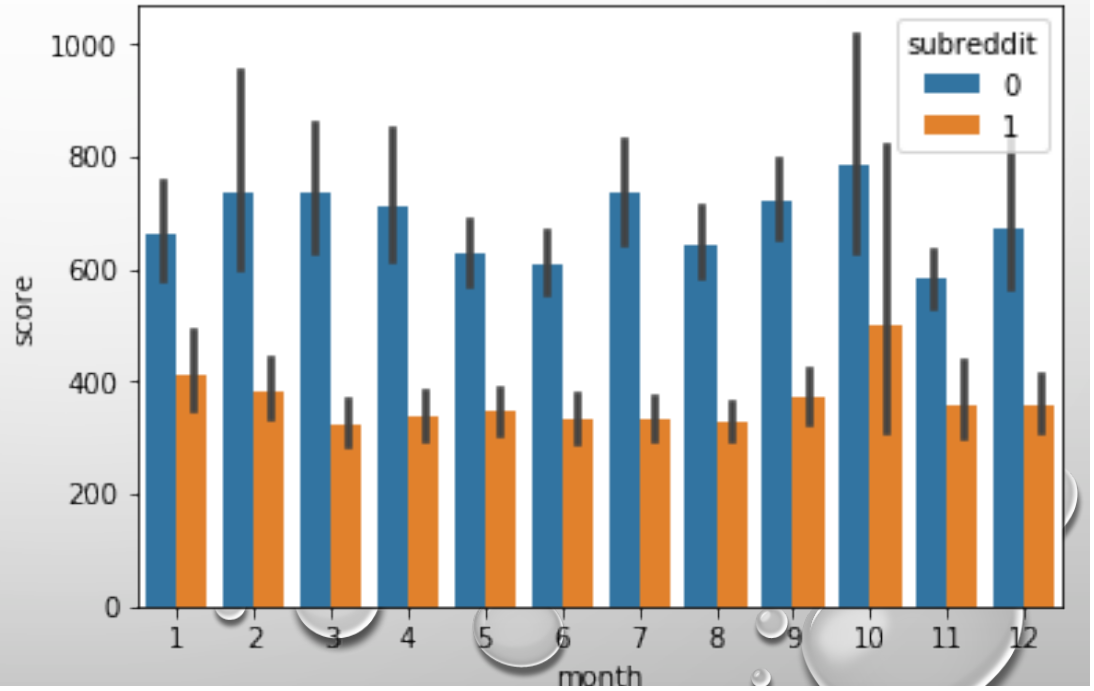
Model	Training score	Testing score
Logistic Regression	87.64	84.88
Bernoulli Naïve Bayes	86.59	85.04
Logistic Regression	87.87	85.19
K-Nearest Neighbors	85.15	79.39
Decision Tree Classifier	84.10	85.34
Extra Tree Classifier	82.14	82.75
Bagged Decision Tree	85.00	84.88
Random Forest	87.42	83.97

Classification models

Model	Training score	Testing score
Logistic Regression	87.64	84.88
Bernoulli Naïve Bayes	86.59	85.04
Logistic Regression	87.87	85.19
K-Nearest Neighbors	85.15	79.39
Decision Tree Classifier	84.10	85.34
Extra Tree Classifier	82.14	82.75
Bagged Decision Tree	85.00	84.88
Random Forest	87.42	83.97

Conclusions and Recommendations

- There is no significant correlation of scores and timing of post.
- Tea subreddit has consistent higher scores

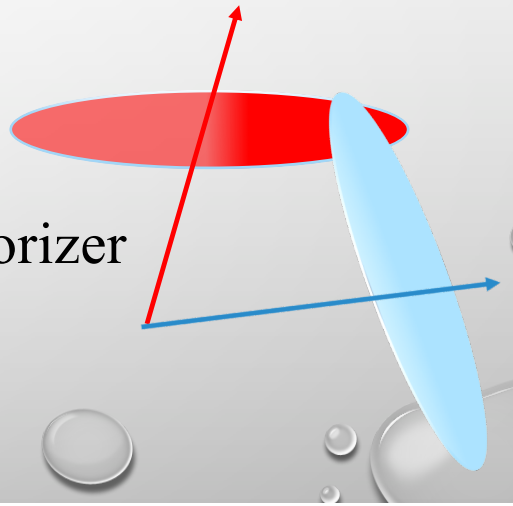


Conclusions and Recommendations

- There is no significant correlation of scores and timing of post.
- Tea subreddit has consistent higher scores
- Developed models that fit title text to subreddit class
- Developed and compared various classification models
- Improved prediction accuracy to 85% from a baseline accuracy of 50%
- Attempted at developing a Poisson regression model to predict the scores, though it provided poor results

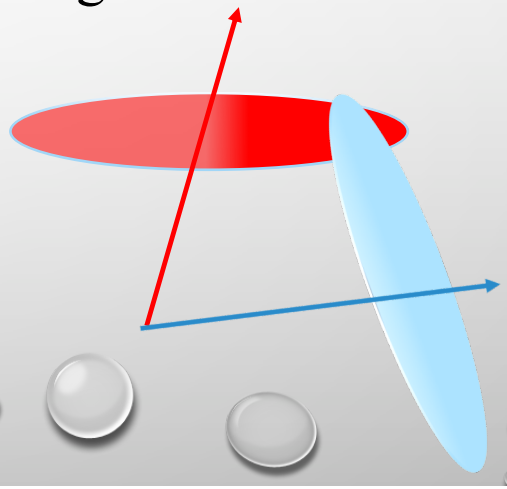
Alignment of vectors

- Each subreddit must have a unit vector
- Posts from that subreddit must be closely aligned to this unit vector
- How separated are the two subreddit post unit vectors?
- Fit a vectorizer on one kind of subreddit
- Transform other subreddit based on previous vectorizer



Alignment of vectors

- Unit vector representative of subreddit: Adding all the vectors and normalizing it
- Features are chosen based on one kind of subreddit from vectorization
- Tea and coffee unit vectors were almost orthogonal
- $\theta \approx 85°$
- Can we use this to model a classifier?
- Does it already exist?



Thank you