Creating a Good Post for r/datascience

Collecting Data

- Data was collected using Python Reddit API Wrapper.
- Collected almost 1000 posts sorted by top and another 1000 sorted by new

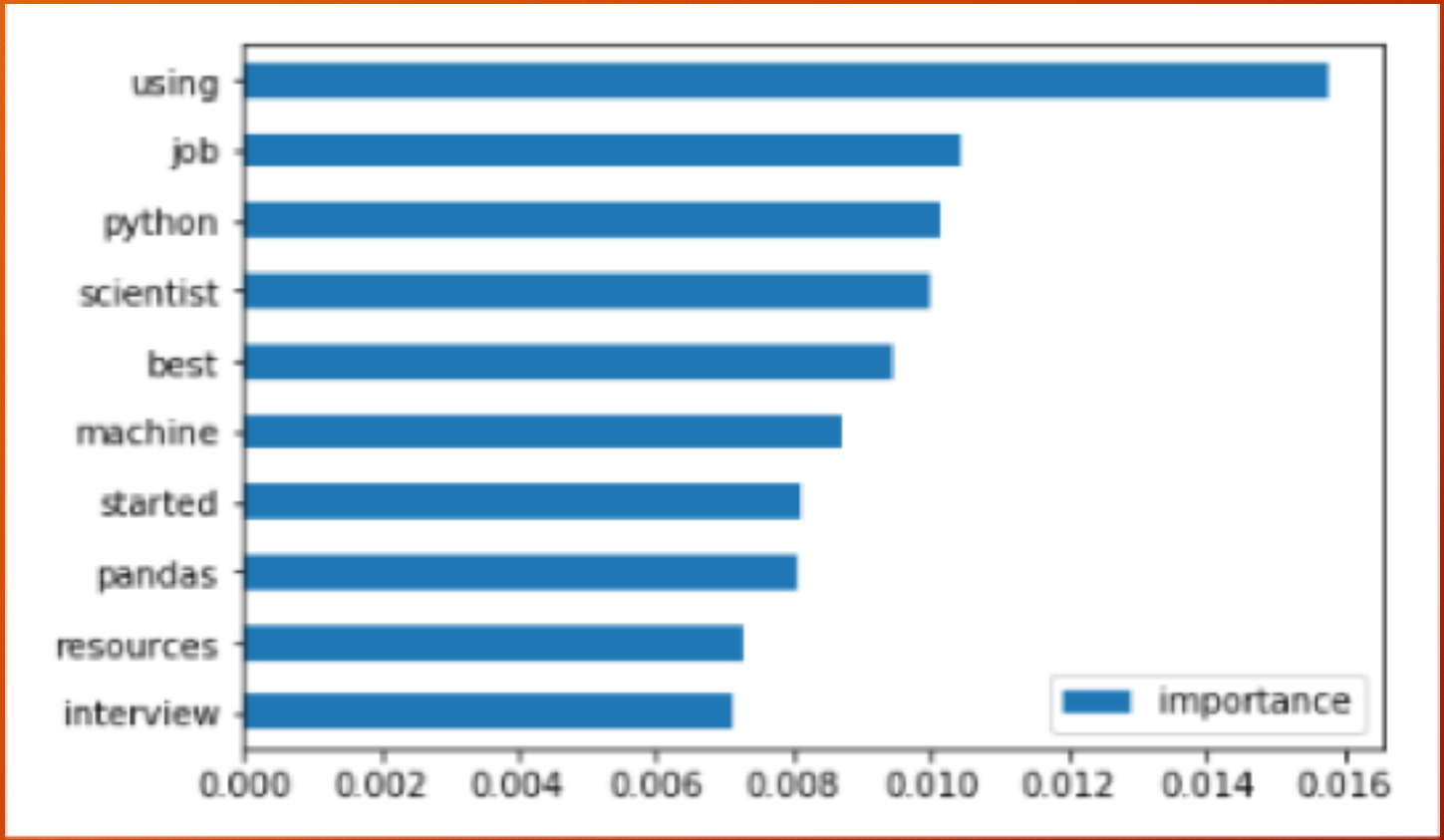
The Problem

- What words have the greatest impact when creating a successful post on the data science subreddit?
- Success is defined here as having comments and score above the 50th percentile. This is equivalent to receiving 15 or more comments and having a score of 55 or higher.

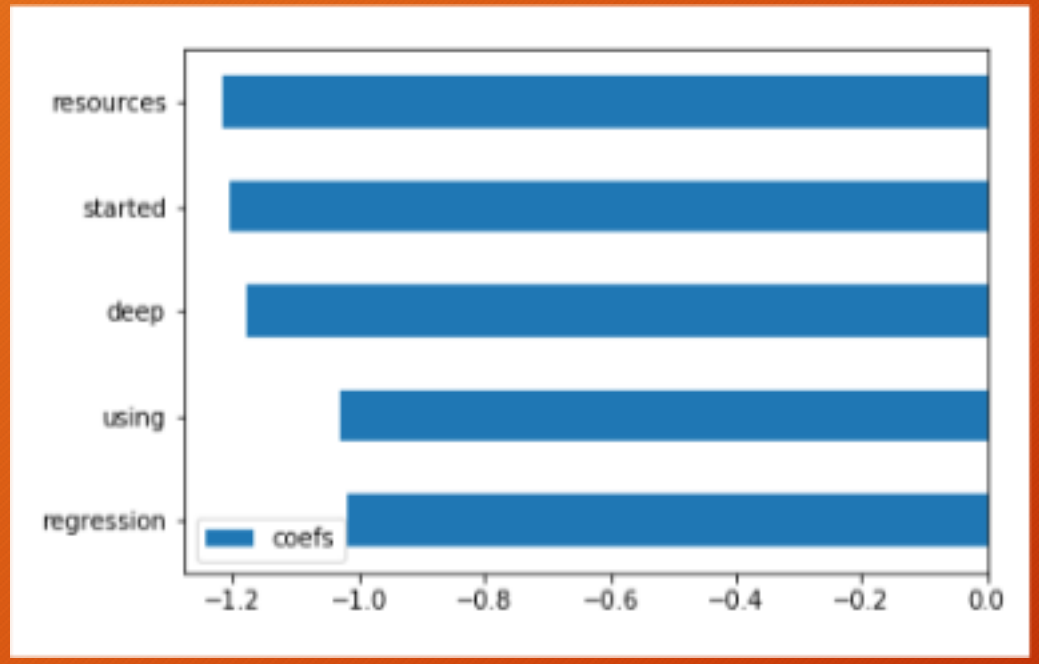
Modelling

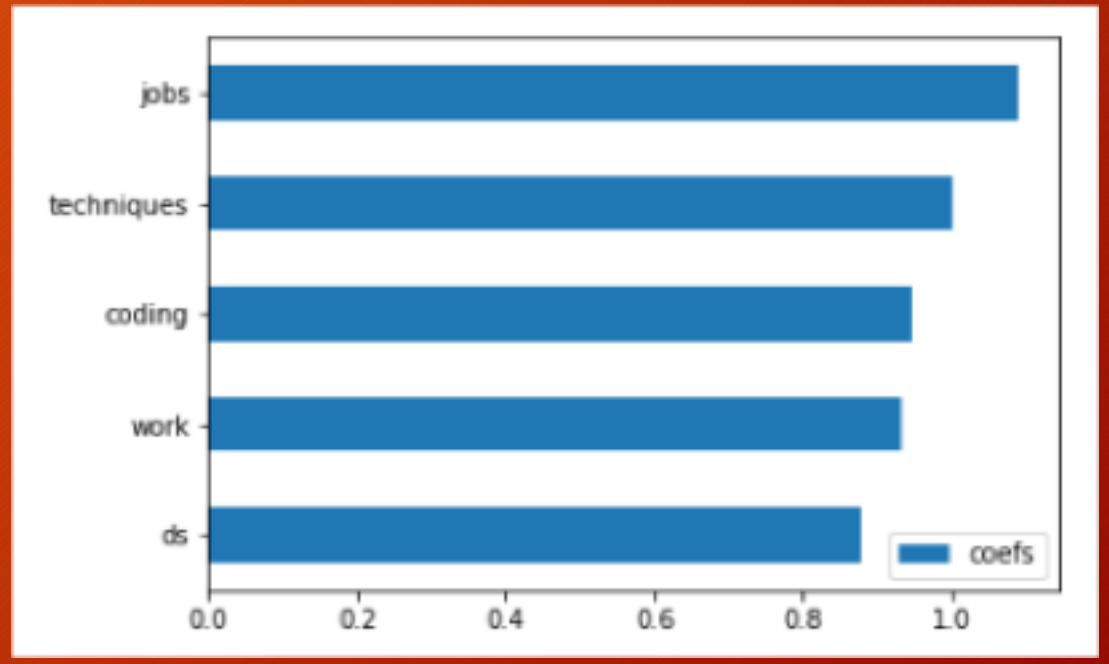
- Count vectorization with random forest and logistic regression
- Term frequency – inverse document frequency with random forest and logistic regression
- Ultimately decided on count vectorization and logistic regression
  - Received an accuracy score of 0.58 which is 8% above the baseline.

Random Forest Features



Logistic Regression Features





Conclusion

- This subreddit appears to be not as friendly to beginners
  - The words started and resources appear to negatively impact a posts' chances of success.
- This subreddit prefers work practical examples or career oriented questions.
  - This is evidenced by the words jobs and work having a high impact on success.

Further Research

- Collect a greater assortment of posts not sorted by top.
  - The data was heavily skewed as a result of 50 posts that had extremely high scores and comments.
  - These posts may have reached r/all and have uncharacteristically high scores and comments
- Try Naïve Bayes as opposed to Logistic Regression
- GridSearch