r/todayilearned & r/askreddit
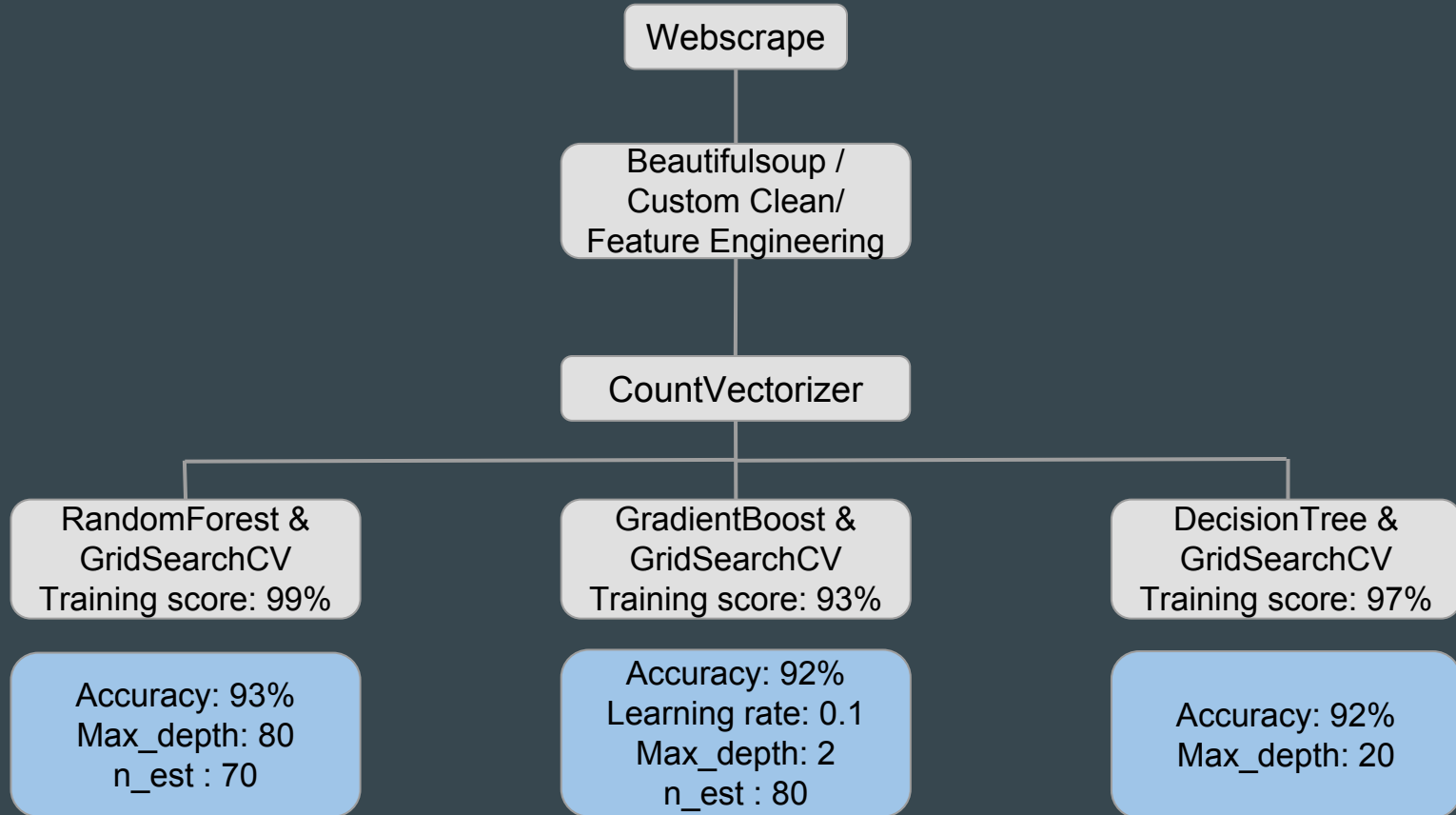
• • •

Author: Will J. Suh

Background

- r/todayilearned is a subreddit about people sharing cool things they learned today.
  - All TIL posts start with the tag 'TIL'

- r/askreddit is a subreddit dedicated to Q & A.
  - All AskReddit posts ended with '?'

Process

```
                    ┌─────────────┐
                    │  Webscrape  │
                    └─────────────┘
                           │
               ┌───────────────────────┐
               │   Beautifulsoup /     │
               │   Custom Clean/       │
               │  Feature Engineering  │
               └───────────────────────┘
                           │
                  ┌──────────────────┐
                  │  CountVectorizer │
                  └──────────────────┘
```

RandomForest & GridSearchCV Training score: 99%	GradientBoost & GridSearchCV Training score: 93%	DecisionTree & GridSearchCV Training score: 97%
Accuracy: 93% Max_depth: 80 n_est : 70	Accuracy: 92% Learning rate: 0.1 Max_depth: 2 n_est : 80	Accuracy: 92% Max_depth: 20

GA

Custom cleaning and Feature Engineering

```
In [81]:  # Remove "TIL" from concat_df['interact'] because it would be too easy.
          # Remove "?" from concat_df['interact'] because almost all of AskReddit ends in a question mark.
          # Remove other punctuation marks and lower case everything.

          artifact = ['\n\n', '\n', '\t', '#', r'r/', 'https://', 'www.', 'http://', 'reddit.com', 'wiki',
                      '.', ',', '-',';',''"', ':', "[", "[", "(", ")", '[/', '/', '?', "'", '*', '$', '&',
                      'TIL', 'the', 'reddit', 'it', 'be']
          for i in artifact:
              concat_df['interact'] = concat_df['interact'].map(lambda x: x.replace(i, ''))

          concat_df['interact'] = concat_df['interact'].str.lower()
```

```
In [73]:  til.isnull().sum()

Out[73]:  post     611
          title      0
          true_y     0
          dtype: int64
```

```
In [74]:  ask.isnull().sum()

Out[74]:  post     954
          title      0
          true_y     0
          dtype: int64
```

Feature Importance

- Top three important words: **What**, **What's**, **How**
  - A lot of question words were used to determine whether a post belonged to r/TIL or r/AskReddit

- Other interesting words that were in the top 100:
  - Serious
  - Mexico
  - Life
  - Earth
  - WW2

- Judging from keywords, there were a lot of questions relating to scientific, historical, and societal questions.

Posted by u/Critical-Lion 1 day ago

33.4k **What's something that looked fun, but you realized it was horrible once you tried it?**

18.2k Comments   Share   Save   ...
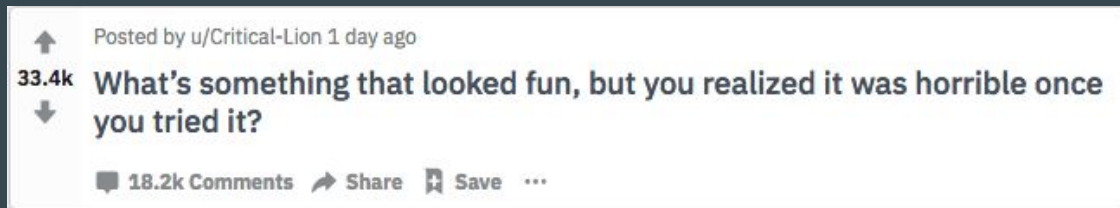
Next Steps

1. **Reddit Gold**
   a. See if there is a correlation with feature important words and gilded posts.
   b. Experiment posting on r/todayilearned on topics related to feature important words.