











Classify Dating groups based on Reddit Text

Krishna Devabhaktuni



I am a freelance Data Scientist, one of the entrepreneurs approached me and they want to start a Dating website targeting a particular age group, they gave me text from Reddit and wanted me to classify the text as reddit only gave them the text. They gave me the text and wanted me to classify if the *text* belongs to over 30 age group or over 40.



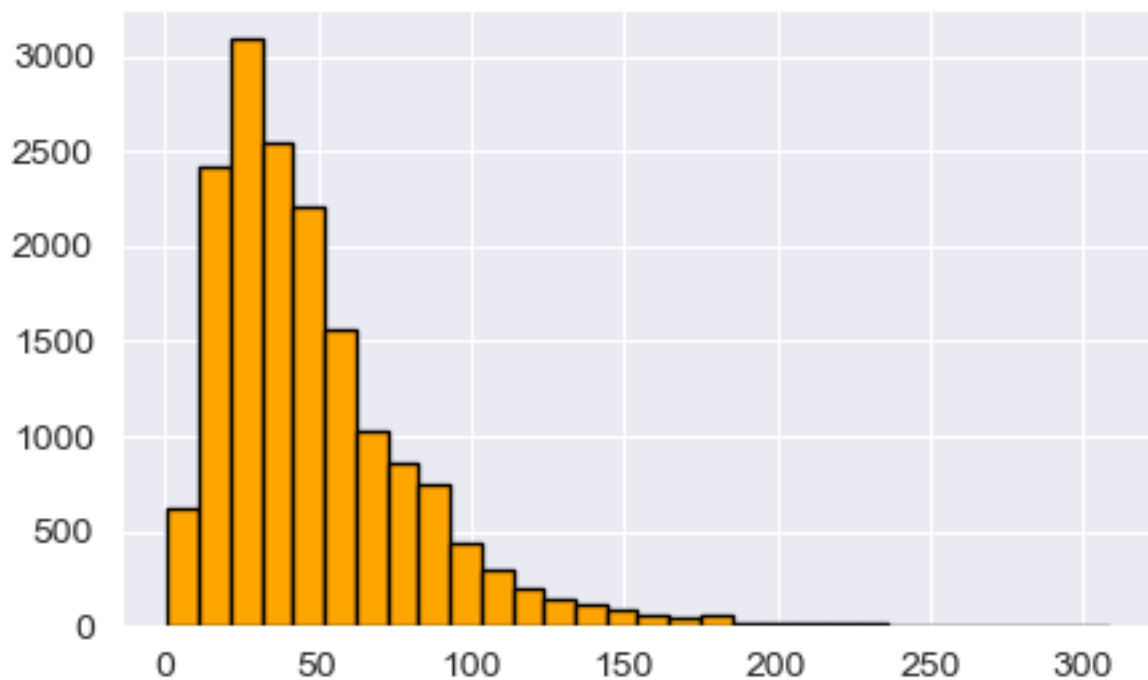


Finding a partner helps many peoples' lives feel complete. Sometimes finding the perfect date means seeking assistance from an experienced matchmaker or relationship expert

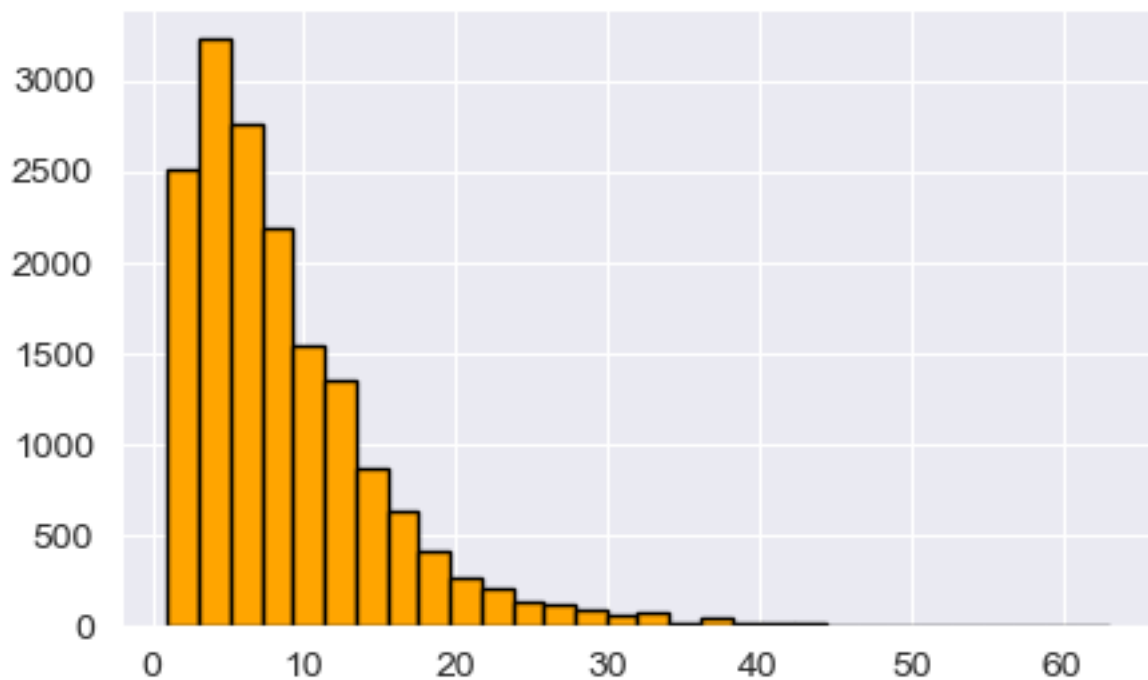
Data

	title	title_length	title_word_count
10959	What dating advice would you give to your youn...	55	10
10960	Dating fulltime after 9 years, but still alone	46	8
10961	Does the desire to date or have a relationship...	84	16
10962	What to do, what to do...	25	6
10963	How hard it is to find a partner in 40's	40	10

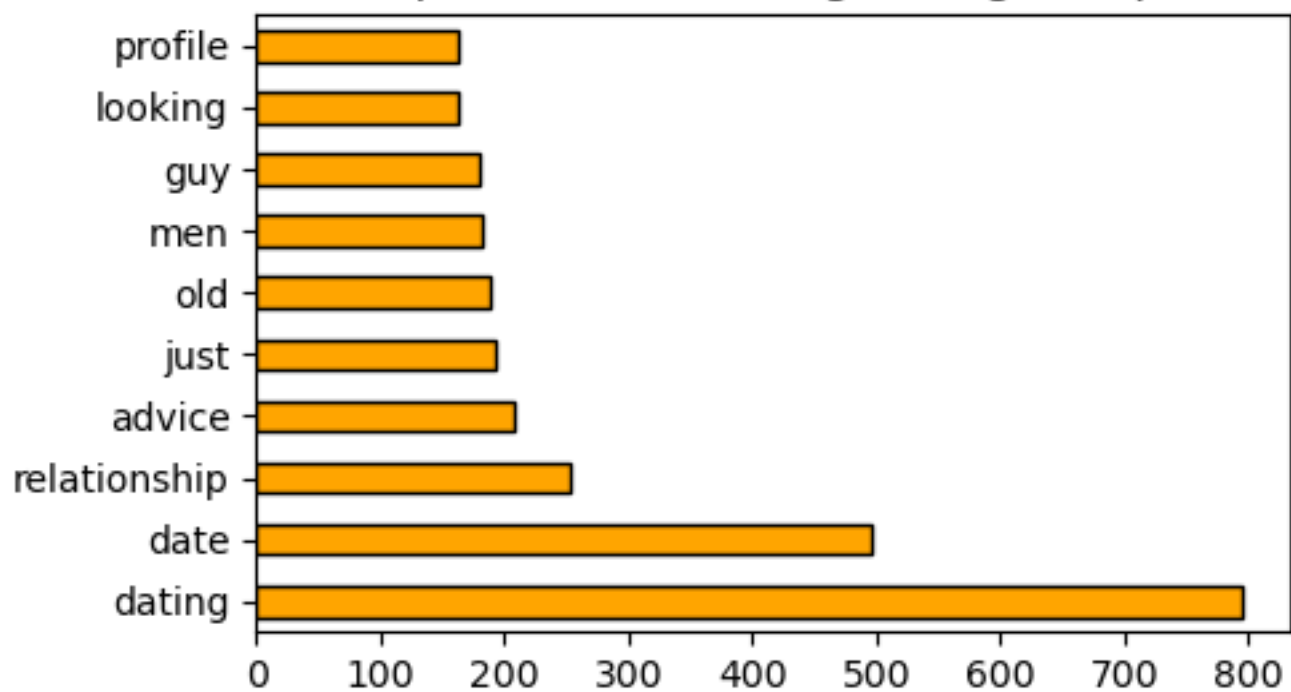
Distribution of Title Length





Distribution of Word Counts




Popular Words Among Dating Groups





I've used four models to classify, firstly I used Countervectorizer with Naive Bayes Classifier, the model returned a training score of 0.77 and the testing score of 0.66. I then used Naive Bayes with TFIDF vecorizer, then I used two more models using support vector machine and xgboost classifier which returned the results below:





Training Score Counter Vectorizer Naive Bayes : 0.77

Testing Score Counter Vectorizer Naive Bayes : 0.66

Training Score TFIDF Naive Bayes : 0.74

Testing Score TFIDF Naive Bayes : 0.67

Training Score TFIDF SVM : 0.91





Testing Score TFIDF SVM : 0.67



Training Score CounterVectorizer, XGBoost Classifier : 0.75

Training Score CounterVectorizer, XGBoost Classifier : 0.66



- 
- 
- 
- 
- All the Models were overfitting
 - A model with Countervectorizer and Naïve Bayes Classifier predicted less false positive but predicted more false negatives
 - Data is imbalanced
 - Surprisingly both groups have similar vocabulary



Next Steps

- **Collect More Data**
 - **Balance the data**
 - **Extract emoji's and hashtags to find better signals to differentiate**
- 