NLP Machine Learning Model

- Subreddit Classifier Prototype

Young Park
Data Scientist

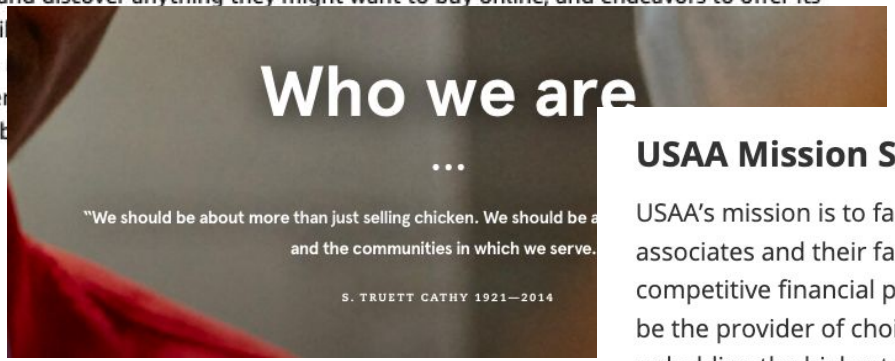
Agenda

- Inspiration
- Process
- Performance Summary
- Recommendations/Applications
- Q/A

Inspiration

Earth's most customer-centric company

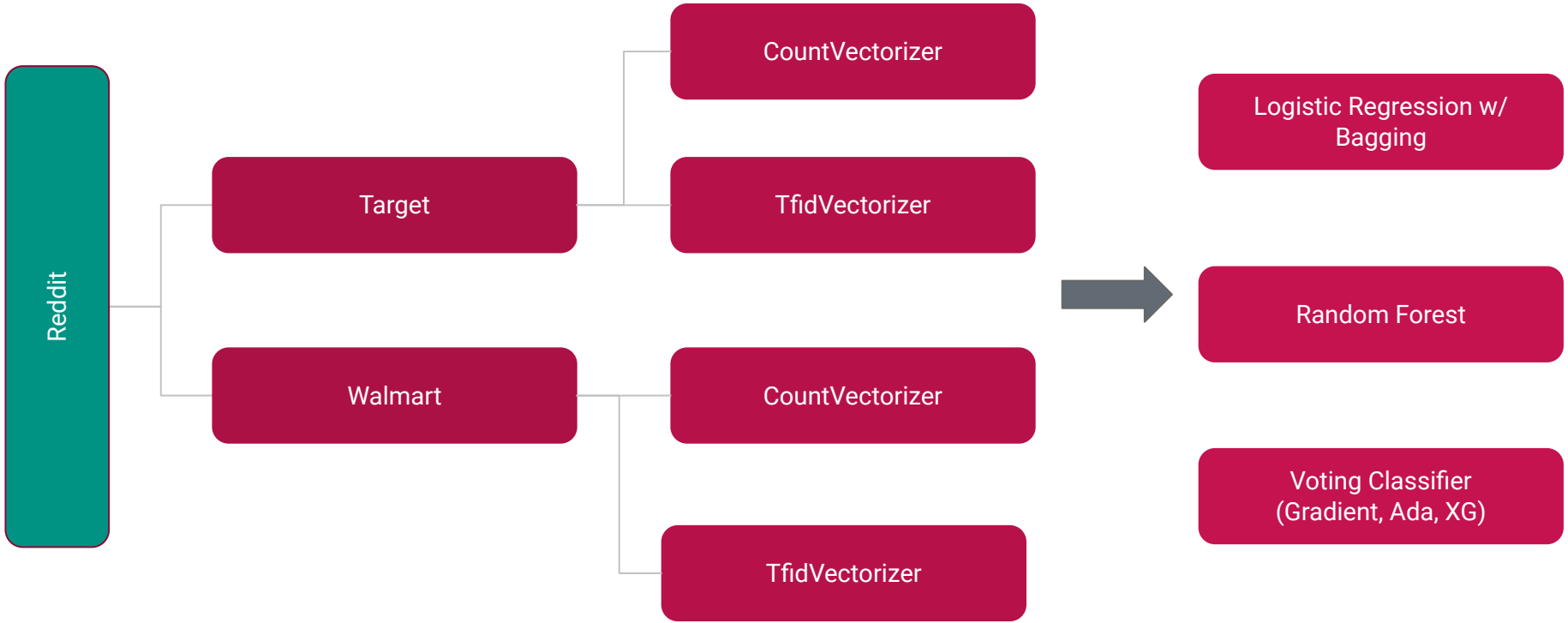
When Amazon.com launched in 1995, it was with the mission "to be Earth's most customer-centric company, where customers can find and discover anything they might want to buy online, and endeavors to offer its customers the lowest possi
and have grown to include
of these groups has differe
make things easier, faster, b



USAA Mission Statement

USAA's mission is to facilitate the financial security of its members, associates and their families by providing a full range of highly competitive financial products and services. In so doing, we seek to be the provider of choice for the military community. We do this by upholding the highest standards and ensuring that our corporate business activities and individual employee conduct reflect good judgement and common sense, and are consistent with our core values of: Service, Loyalty, Honesty and Integrity.

Process


Reddit
Target
Walmart
CountVectorizer
TfidVectorizer
CountVectorizer
TfidVectorizer
Logistic Regression w/ Bagging
Random Forest
Voting Classifier (Gradient, Ada, XG)

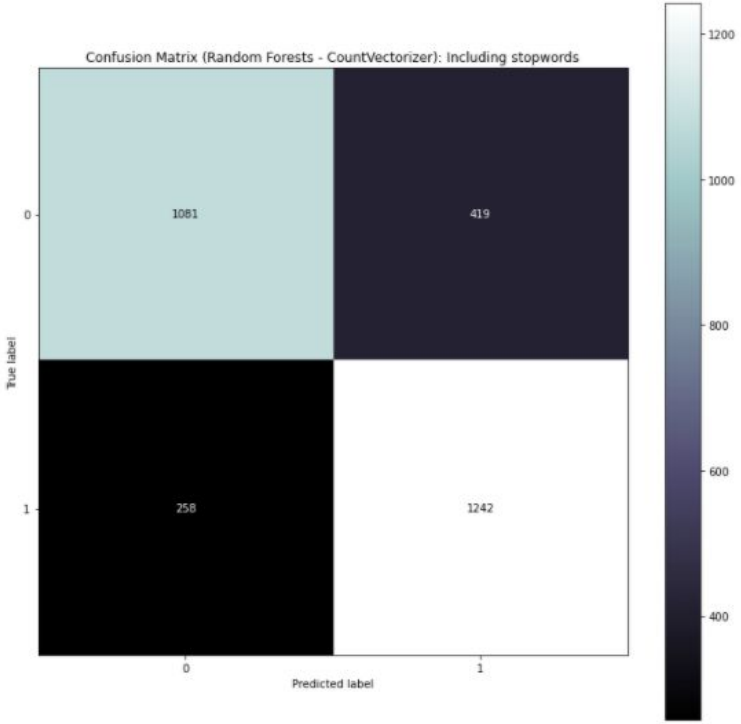
Hyperparameters

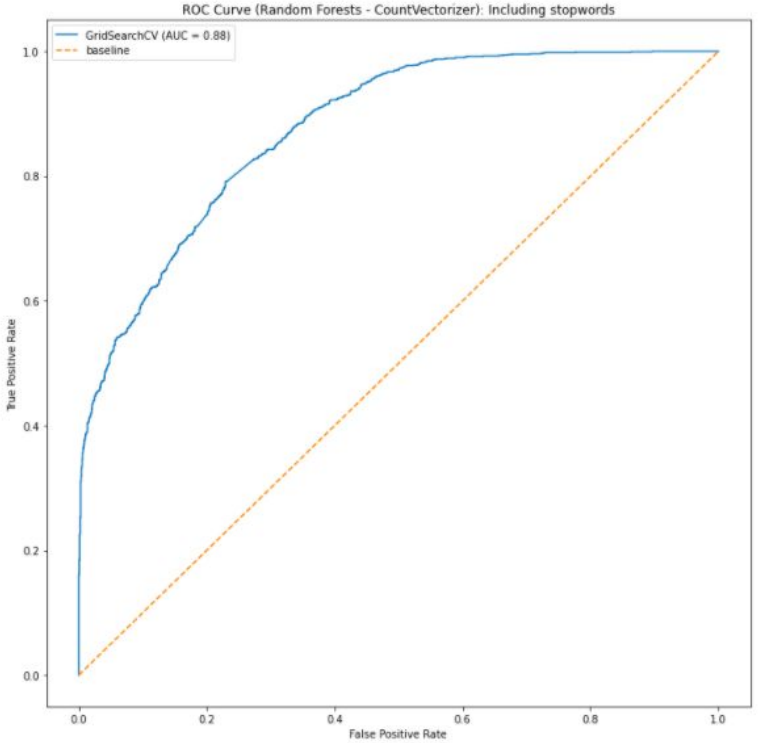
Logistic Regression	• BaggingClassifier • n_estimators = 100
Random Forest	• n_estimators = 250, 350, 450 • max_depth = None, 2, 3, 4 • max_features = auto, log2 • min_samples_split = 4, 5, 6 • min_samples_leaf = 2, 3
Voting Classifier	• gb/ada/xg__n_estimators = 500 • gb/ada/xg__learning_rate = 0.01, 0.05 • gb__max_features = log2 • xg__max_depth = 2, 3

Performance Summary

Model	Transformer	Stopwords	Accuracy	Recall	Precision	F1	ROC
Log	Cvec	Y	0.761667	0.762000	0.761492	0.761746	0.865672
Log	Tfid	Y	0.772333	0.762667	0.777702	0.770111	0.867852
Log	Cvec	N	0.731000	0.742000	0.726027	0.733927	0.817595
Log	Tfid	N	0.734000	0.732667	0.734626	0.733645	0.822621
Rf	Cvec	Y	0.774333	0.828000	0.747742	0.785827	0.877035
Rf	Tfid	Y	0.770333	0.809333	0.750773	0.778954	0.870309
Rf	Cvec	N	0.730000	0.806000	0.699653	0.749071	0.827448
Rf	Tfid	N	0.716000	0.801333	0.684510	0.738329	0.813973
Vote	Cvec	Y	0.765667	0.874000	0.718356	0.788571	0.875278
Vote	Tfid	Y	0.762000	0.840667	0.726382	0.762000	0.868837
Vote	Cvec	N	0.713667	0.833333	0.672405	0.744269	0.815984
Vote	Tfid	N	0.703667	0.841333	0.659697	0.739525	0.812748

Classification Metrics





Recommendations

- While the best performing model achieved accuracy score of 77%, which is 27 pts higher than the baseline of 50%, there is room for improvement by better fine-tuning hyperparameters using powerful computers that are designed to handle big data.

- For demonstration and learning purposes, I selected two subreddits that I thought were similar in nature. While I was limited to 10,000 rows of text data in total, with access to much more powerful computers, more data can certainly help to build better model.

- Potential application for this type of machine learning:
    - Pinpoint topic of interests. What are you customers interested in?
    - Deeper sentiment analysis. How are your customers feeling?
    - Real-time response algorithm. Respond to your customers right away.

Q&A