ABC Toy Co.

Where the Babies Get Their Stuff

Alexander Gluck

Data Scientist
General Assembly

Problem Statement

Marketers at ABC Toy Co have discovered that their target demo: parents of small children, and dog owners talk about their... dependents in similar ways. I have been commissioned to build a model that can distinguish those with actual human children, and those whose babies are of the fur variety

Data

Reddit Posts using Pushshift/api

r/Dogs

r/Parenting

- Chosen because tone is similar. Posts largely concern practical issues
  - Assembled a sample of 4,000 total posts, 2,000 from each

Model Types

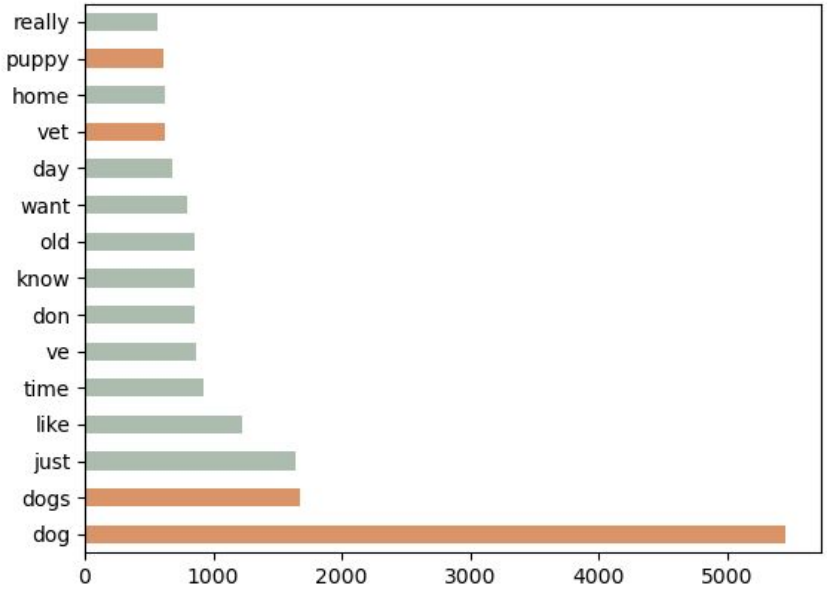
01 Logistic Regression

02 Random Forest

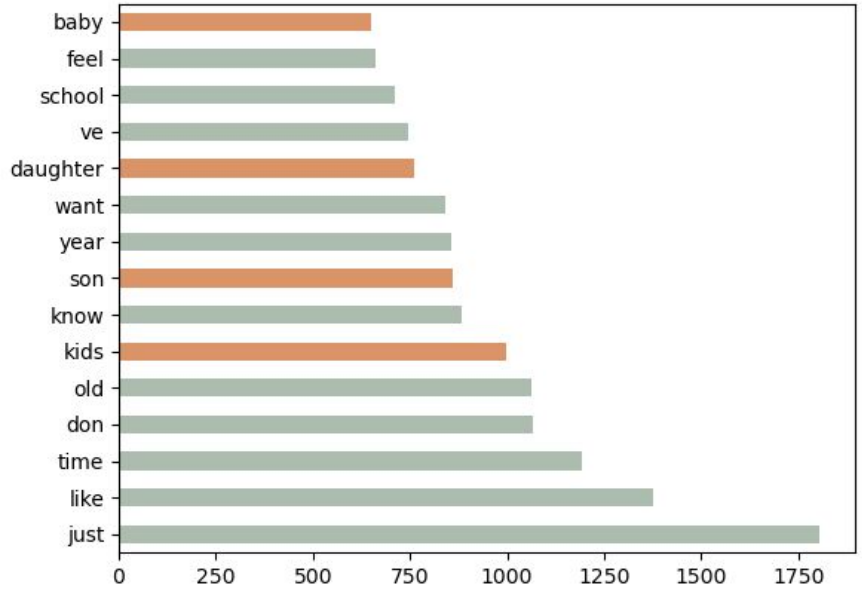
03 Gradient Boost

04 Support Vector Machine

r/Dogs Most Common Words



r/Parenting Most Common Words



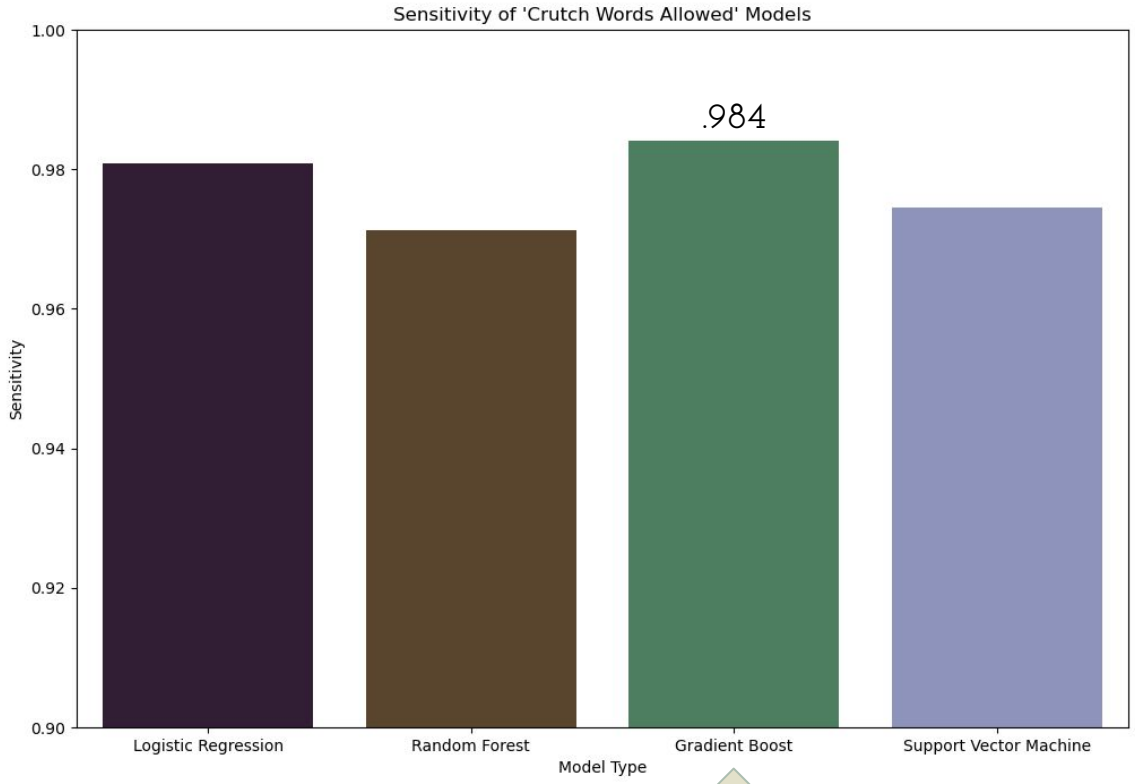
Custom Stop Words - Crutch Words

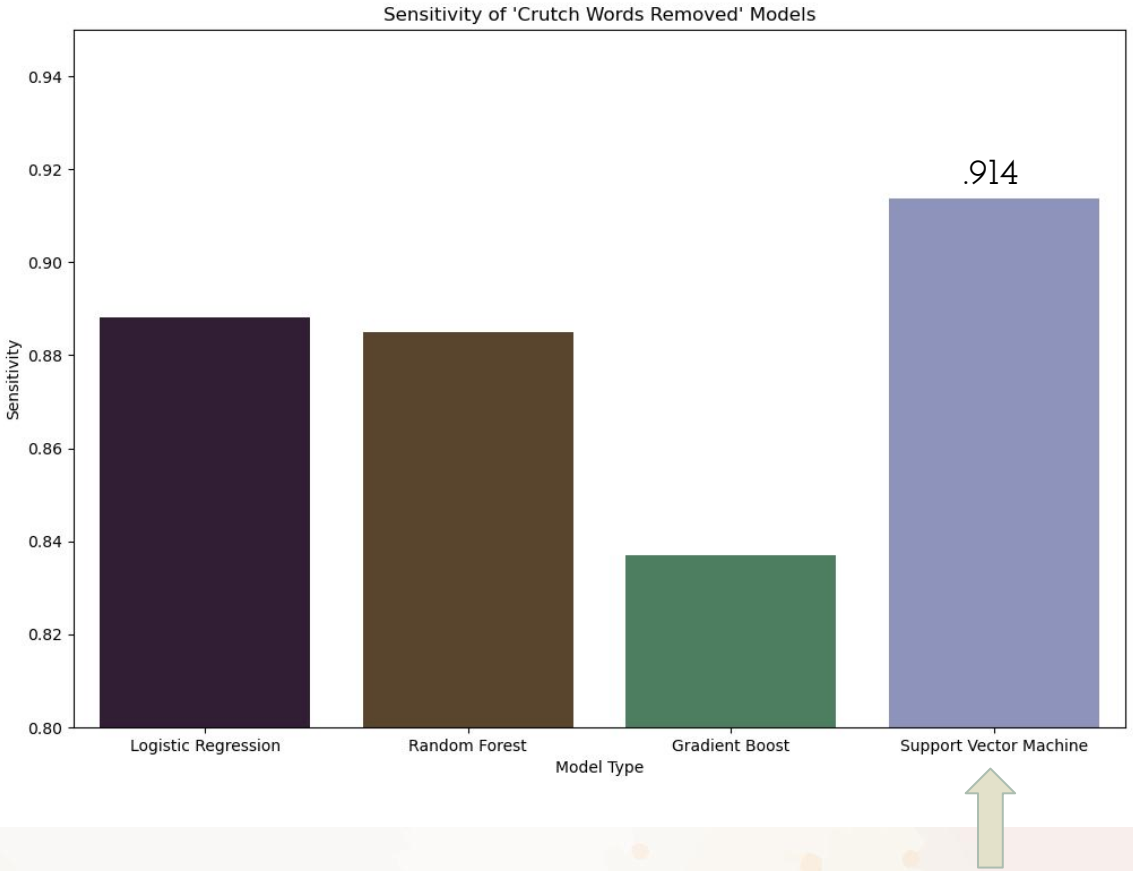
Crutch Words Allowed

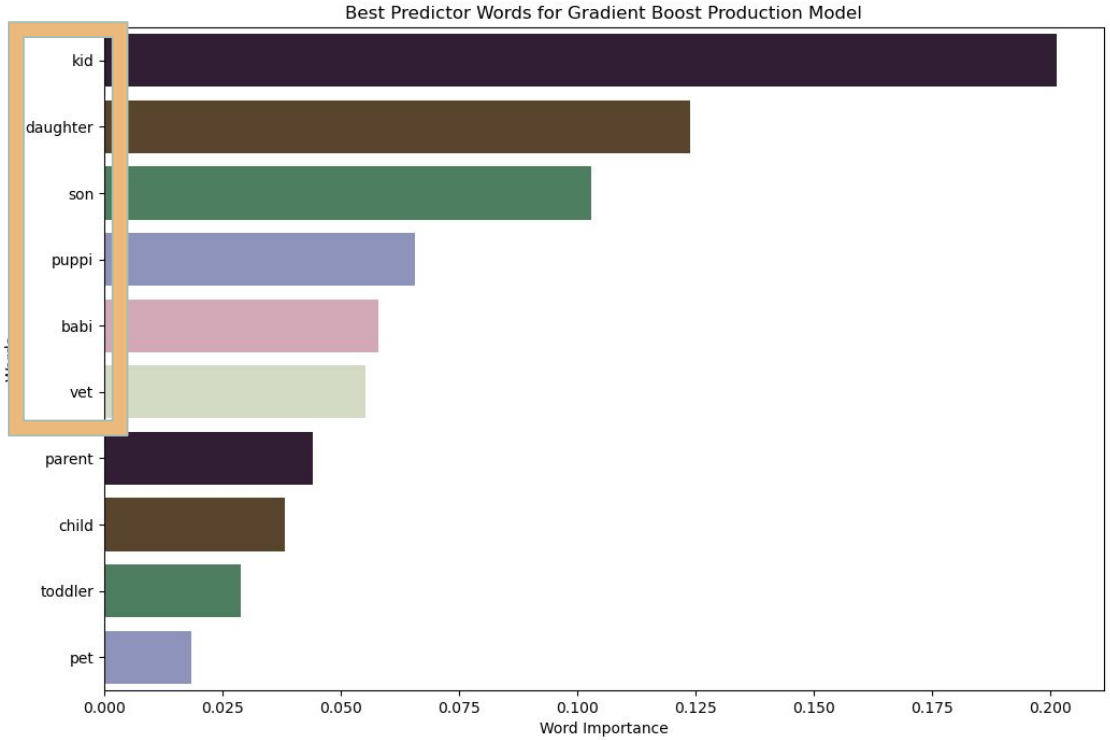
One set of models included the 'Crutch Words' in the Corpus

Crutch Words Removed

One set of models had 'Crutch Words' removed by adding them to stop words


Sensitivity of 'Crutch Words Allowed' Models
1.00
0.98
0.96
0.94
0.92
0.90
Sensitivity
.984
Logistic Regression
Random Forest
Gradient Boost
Support Vector Machine
Model Type


Sensitivity of 'Crutch Words Removed' Models
Sensitivity
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
.914
Logistic Regression
Random Forest
Gradient Boost
Support Vector Machine
Model Type


Best Predictor Words for Gradient Boost Production Model
kid
daughter
son
puppi
babi
vet
parent
child
toddler
pet
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Word Importance

Conclusions and Recommendations

Conclusions:

- Several of the models built are capable of helping ABC to better identify and reach their target demographic
- **Gradient Boost** was the best candidate for production based on performance on **sensitivity**
- While it does not make business sense to productionize any of the **Crutch Words** models, the **Support Vector Machine** model's performance was notable

Recommendations:

- The performance of the **Gradient Boost** model could be further improved in a couple of ways:
  - **More Data**: Models are more effective the more data they have to train on, and Reddit is a treasure trove
  - **Ensembling:** Though **Gradient Boost** was the best performer, many of the other models also performed well. These models could be combined to achieve greater performance

Questions?


slidesgo