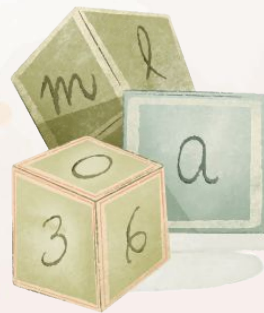




BABIES

Fur and Otherwise





ALEXANDER GLUCK

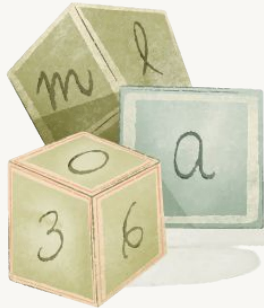
Data Scientist
General Assembly





PROBLEM STATEMENT

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



DATA

REDDIT POSTS USING PUSHSHIFT/API



R/DOGS



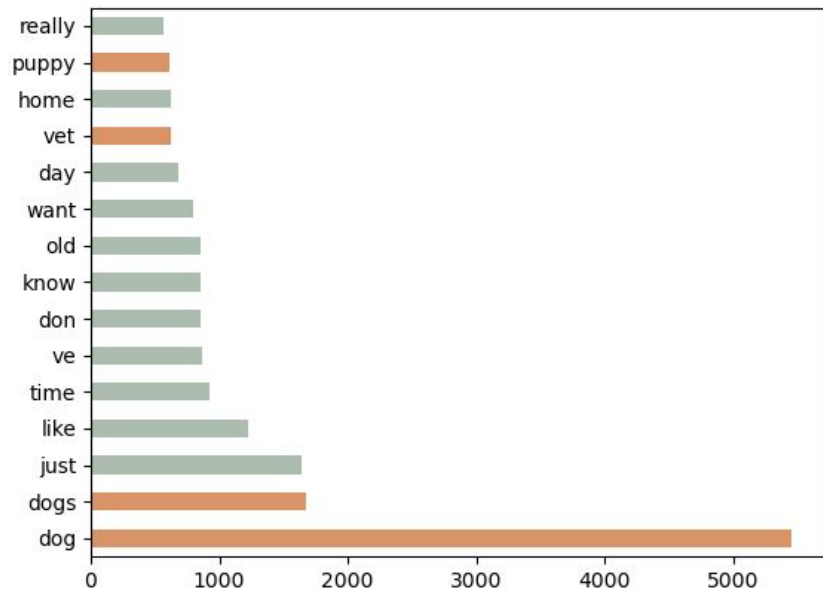
R/PARENTING



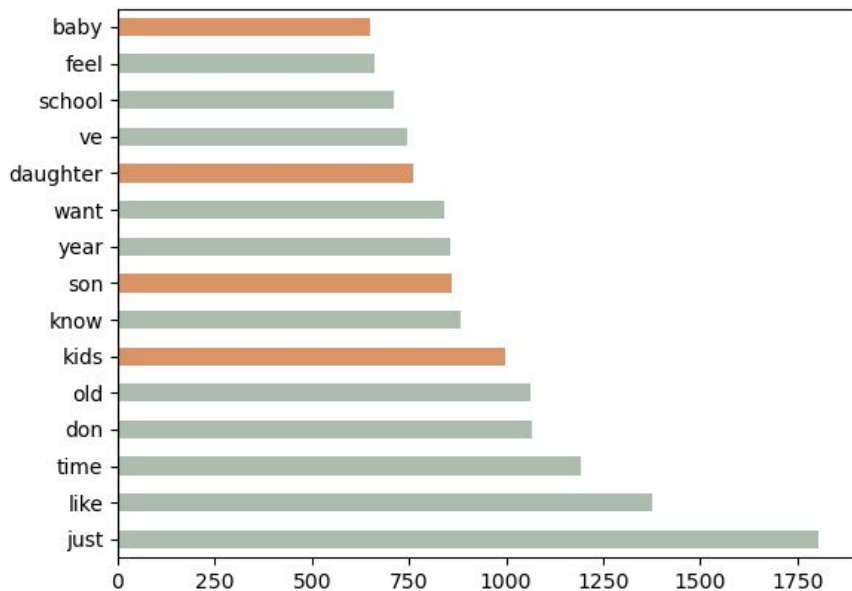
- Chosen because tone is similar. Posts largely concern practical issues
 - Assembled a sample of 4,000 total posts, 2,000 from each



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



MODEL TYPES

LOGISTIC REGRESSION

01

03

GRADIENT BOOST

RANDOM FOREST

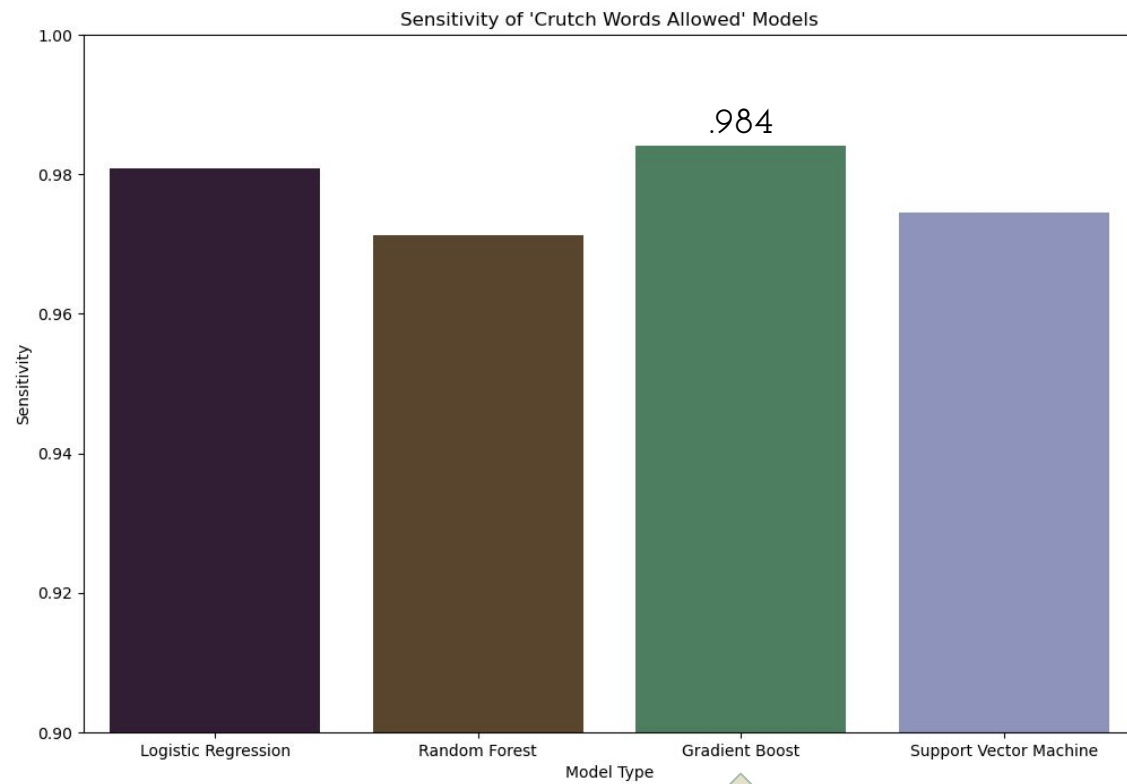
02

04

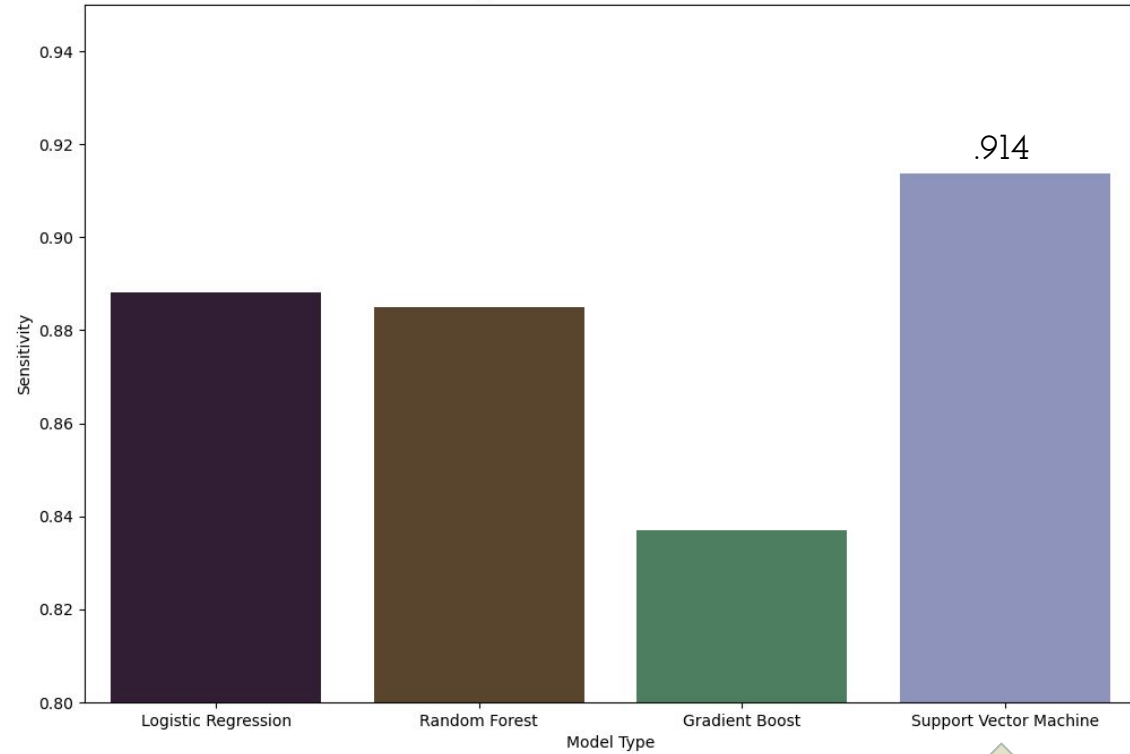
SUPPORT VECTOR
MACHINE

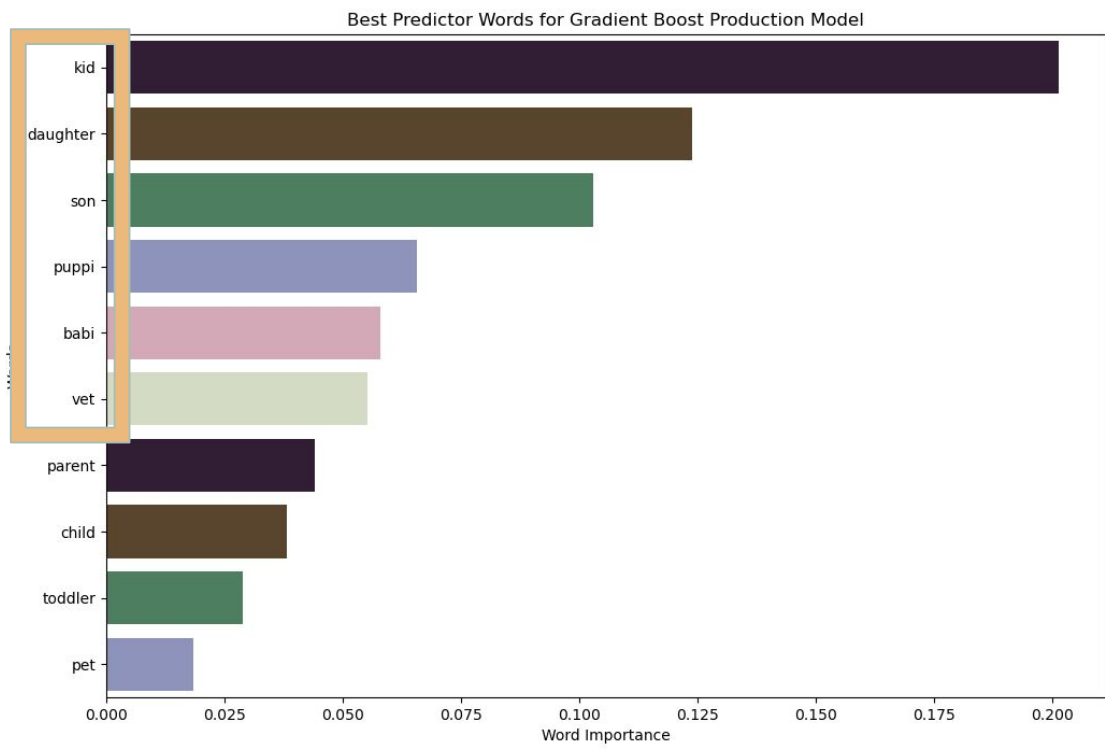
OPTIMIZE FOR SENSITIVITY





Sensitivity of 'Crutch Words Removed' Models







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?



