WEB API & CLASSIFICATION MODEL

BY: EVELYN LI

AGENDA

- SUBREDDIT SELECTION
- MODELS
- KEY FEATURES
- KEY TAKEAWAYS

SUBREDDIT SELECTION

ASK MEN

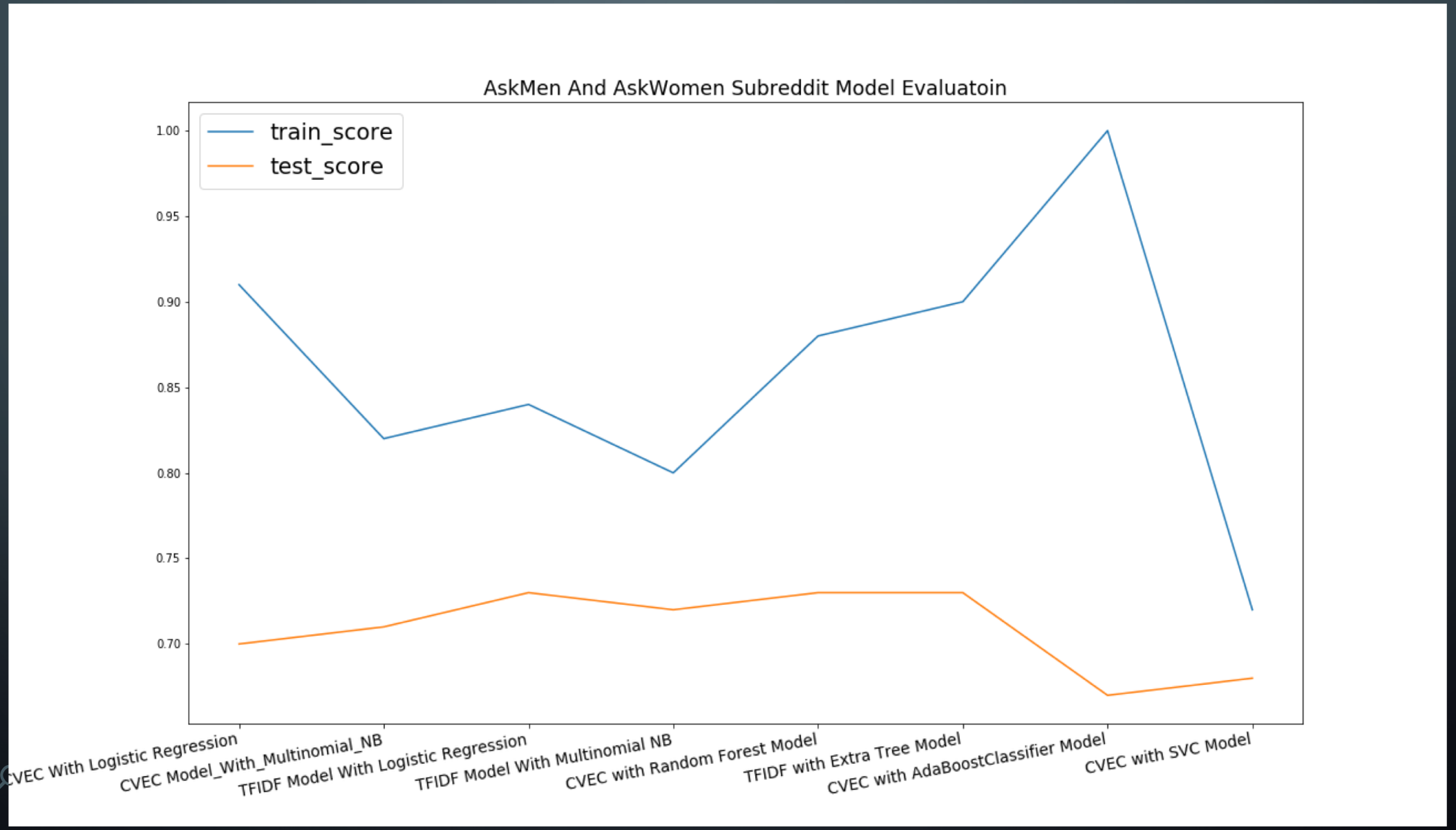
Title + Self Text

ASK WOMEN

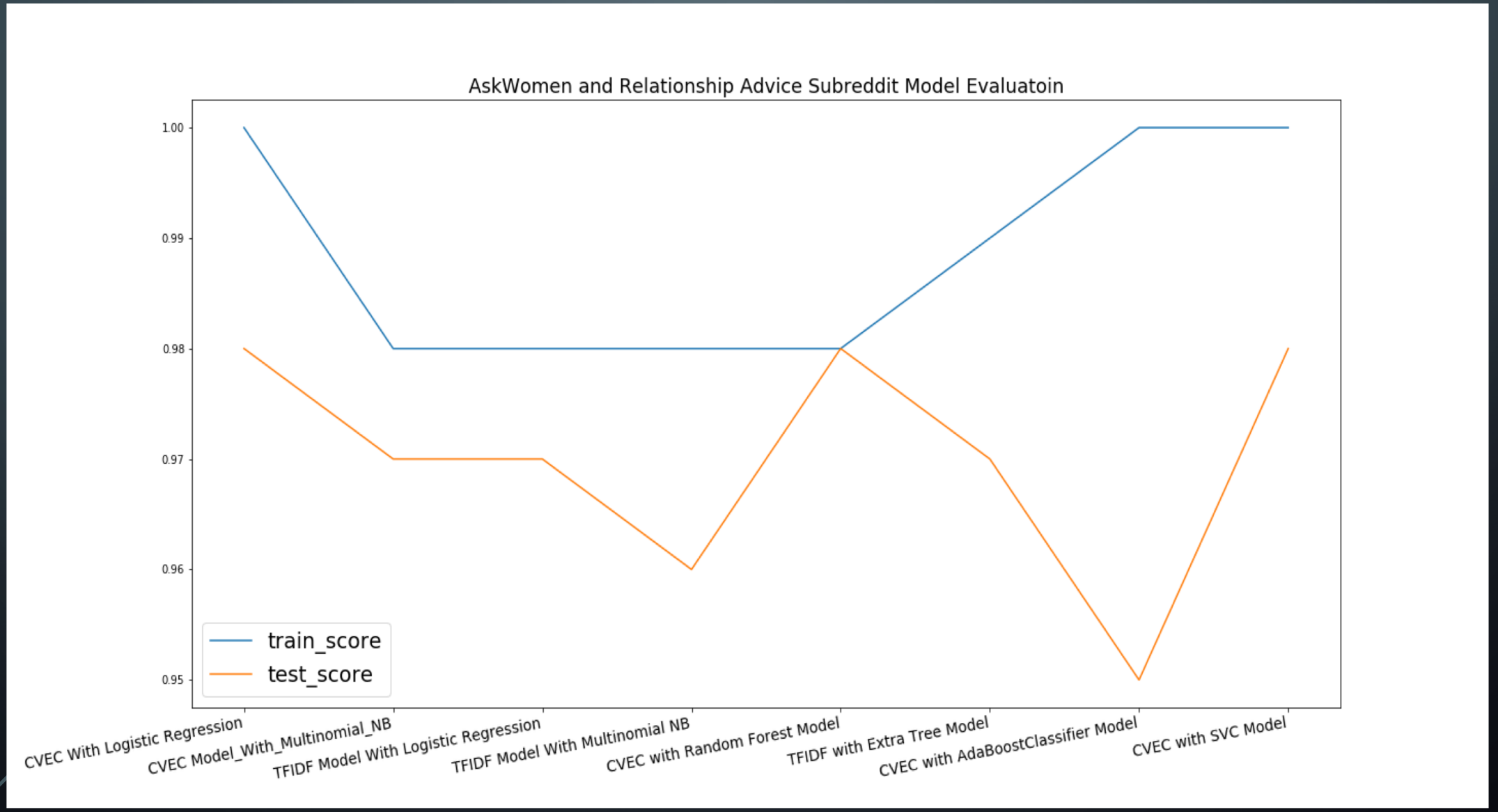
RELATIONSHIP ADVICE

Title + Self Text

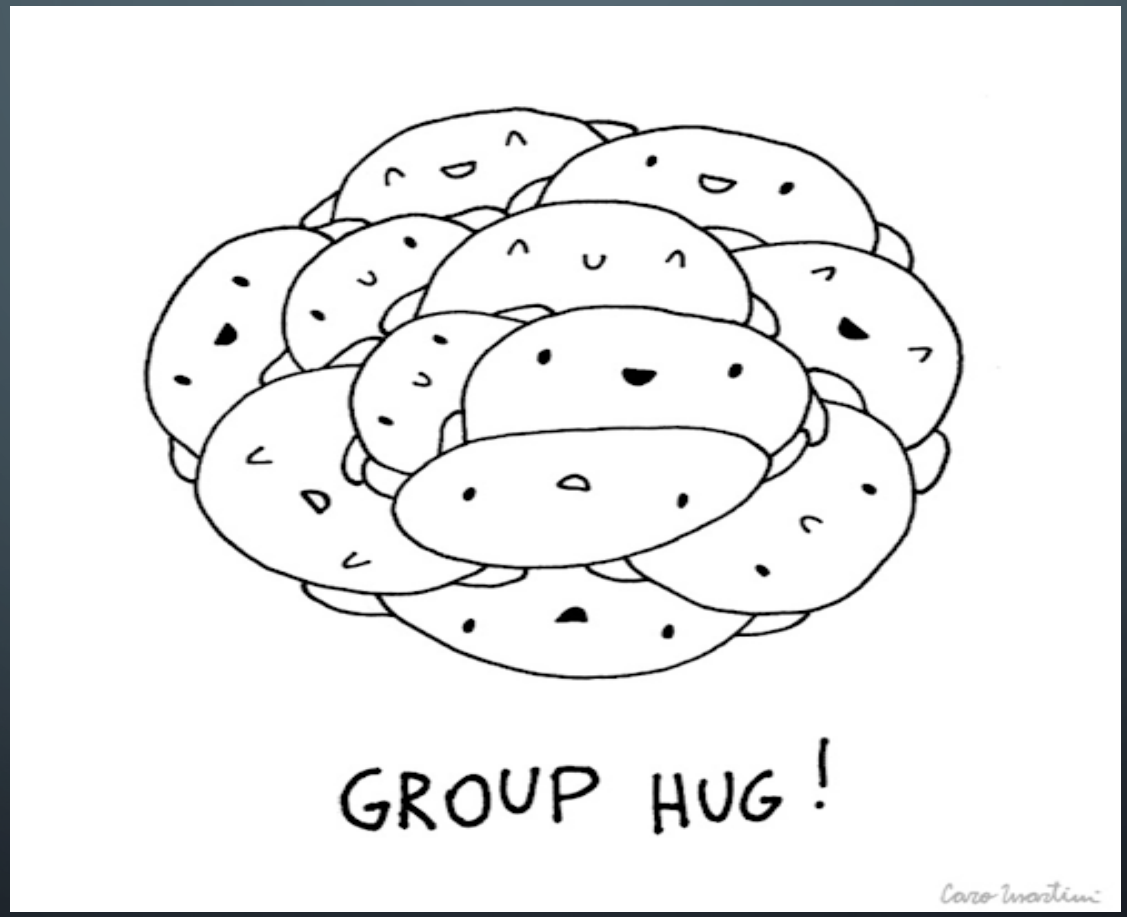
MODELS FOR ASK MEN ASK WOMEN



MODELS FOR ASK WOMEN RELATIONSHIP ADVICE



WHY NOT COMBINE ALL OF THEM?



AND THIS IS THE MODEL:

About 400 Models

```
# models with Best Parameters from previous notebook
mnb = MultinomialNB()
logit = LogisticRegression()
rf = RandomForestClassifier(max_depth = 100,
                            min_samples_leaf= 2,
                            n_estimators= 100)
et = ExtraTreesClassifier( max_depth = 50,
                           min_samples_leaf = 2,
                           n_estimators = 100)
ada = AdaBoostClassifier(base_estimator=DecisionTreeClassifier(max_depth = 50),
                         learning_rate = 1.0,
                         n_estimators = 200)
svc = svm.SVC(C= 1.5, kernel= 'rbf')
```

```
('tfidf2', TfidfVectorizer(max_df= 0.95,
                           max_features= 1500,
                           min_df= 2,
                           ngram_range= (1,2))),
```

MODEL EVALUATION

Ask Men Ask Women Dataset

- Train Score 0.92
- Test Score 0.733
- Train Precision = 0.88
- Test Precision = 0.709
- Train ROC AUC score = 0.905
- Test ROC AUC score = 0.699

Relationship Advice Ask Women Dataset

- Train Score 0.99
- Test Score 0.98
- Train Precision = 1.0
- Test Precision = 0.99
- Train ROC AUC score = 0.98
- Test ROC AUC score = 0.98

WHY IS THIS HAPPENING?

HYPOTHESIS:

LARGE TRAIN TEST SCORE GAP
=> MORE OVERLAP IN TOP FEATURES

SMALL TRAIN TEST SCORE GAP
=> LESS OVERLAP IN TOP FEATURES

A QUICK WALKTHROUGH OF THE STEPS

- FLIP THE TARGETS
  - {Ask Men : 0, Ask Women :1} → {Ask Men : 1, Ask Women: 0 }
  - {Relationship : 0, Ask Women :1} → {Relationship: 1, Ask Women : 0 }
- COLLECT THE COUNTS FROM CVEC/TFIDF
  - Top 100 to 500 most common words and find the % of overlap
- EXTRACT COEFFICENTS OF THE MODELS
  - Find top100 to 500 features and find the % of overlap

THE TRUTH IS … HYPOTHESIS IS CORRECT

Ask Men Ask Women Dataset

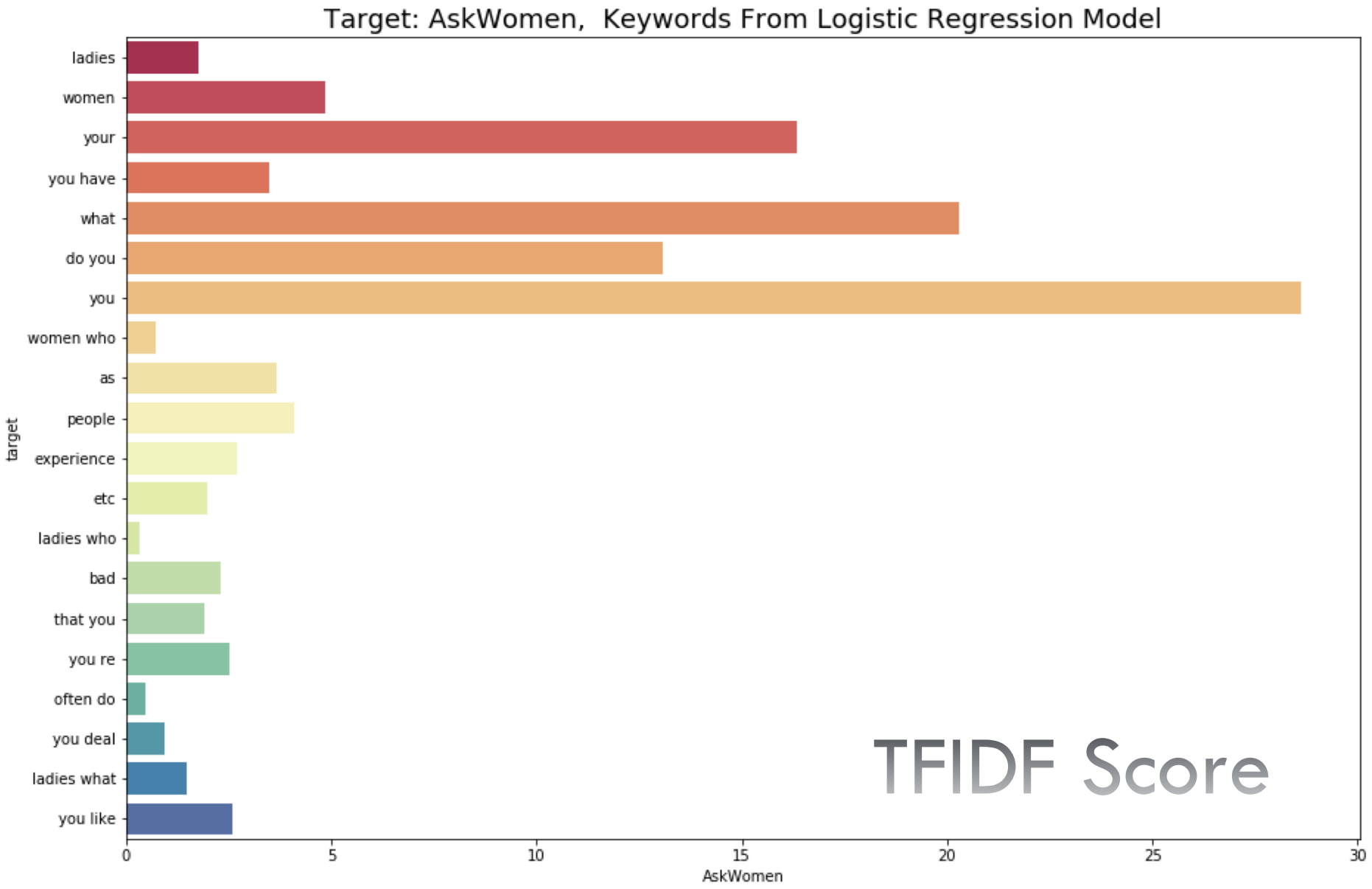
Top Common Words	Common Words Overlap
100	0.81
300	0.727
500	0.676

# Of Top Features	Log Reg Overlap	NB Overlap
100	0.0	0.71
200	0.0	0.685
300	0.02	0.646
400	0.0375	0.64
500	0.062	0.624

Relationship Advice Ask Women Dataset

Top Common Words	Common Words Overlap
100	0.92
300	0.905
1000	0.878

# Of Top Features	Log Reg Overlap	NB Overlap
100	0.00	0.43
200	0.00	0.5
300	0.013	0.503
400	0.035	0.53
500	0.046	0.566


Target: AskWomen, Keywords From Logistic Regression Model
ladies
women
your
you have
what
do you
you
women who
as
people
experience
etc
ladies who
bad
that you
you re
often do
you deal
ladies what
you like
target
0
5
10
15
20
25
30
AskWomen
TFIDF Score

KEY TAKEAWAYS

LARGE TRAIN TEST SCORE GAP
=> MORE OVERLAP IN TOP FEATURES

SMALL TRAIN TEST SCORE GAP
=> LESS OVERLAP IN TOP FEATURES

<u>SOLUTION:</u>
INCREASE SIZE OF THE DATA

THANK YOU