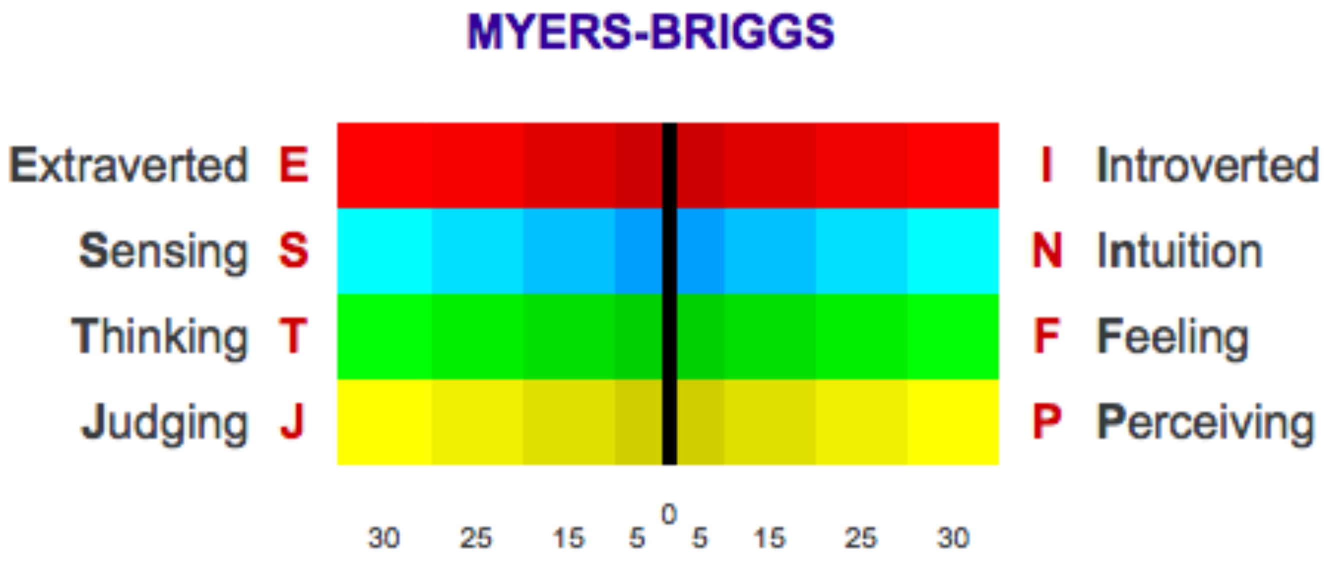
Adventures in P-Typing

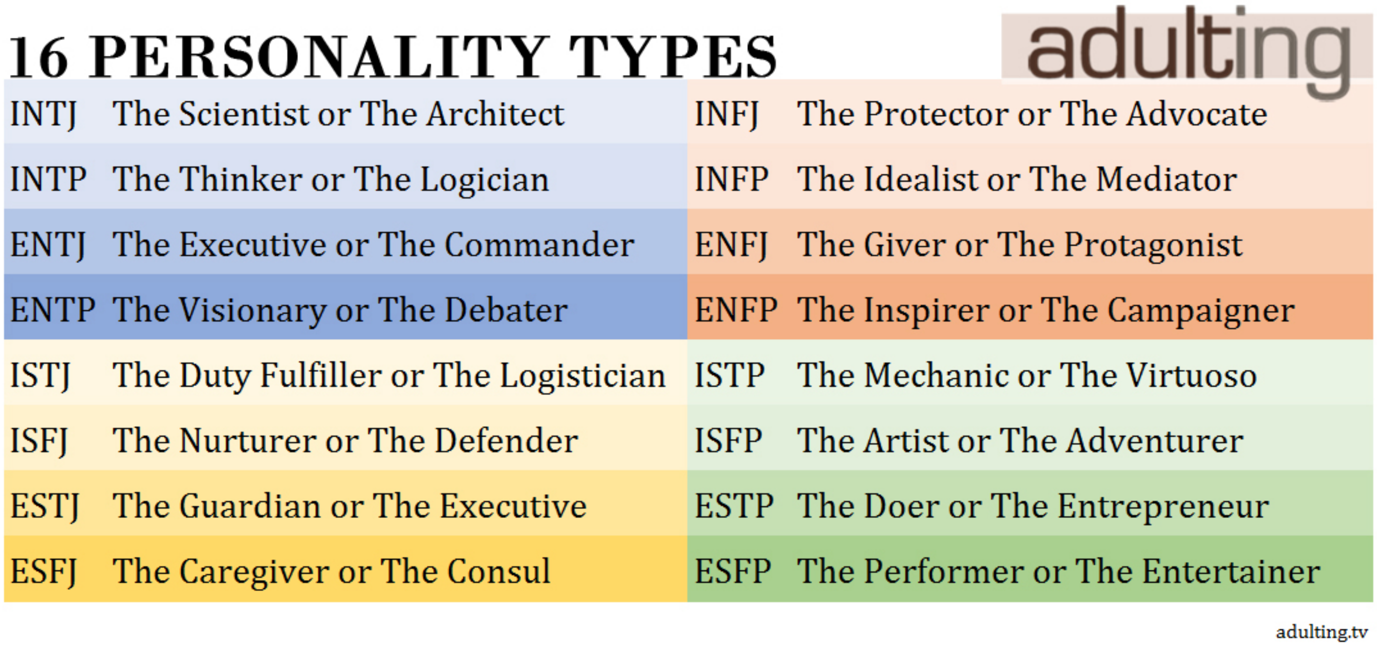
Using Natural Language Processing and Data Science to Explore the Myers-Briggs Personality Test

The Myers-Briggs

The Creators

- Mother/daughter team of Katharine Briggs and Isabel Briggs Myers
- Based off of the work of Karl Jung
- Help people figure out the right careers for themselves



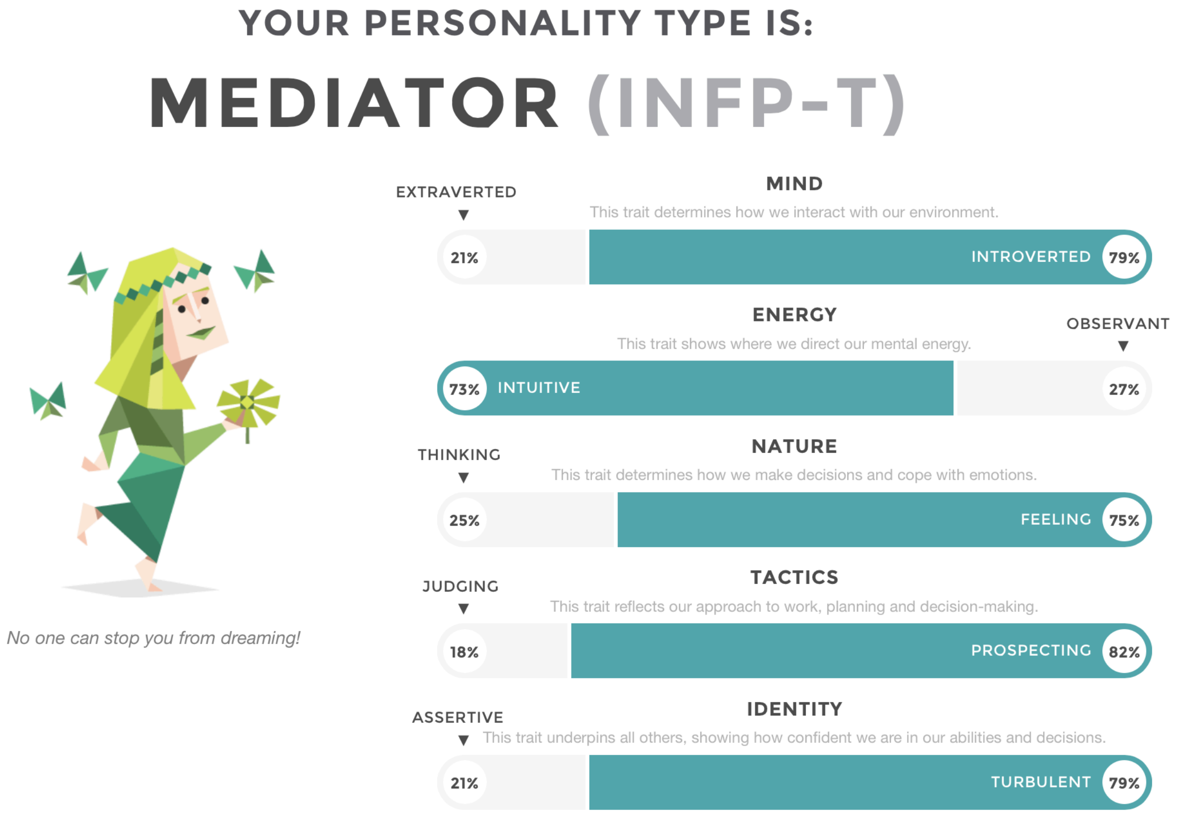


16 PERSONALITY TYPES

adulting

INTJ	The Scientist or The Architect	INFJ	The Protector or The Advocate
INTP	The Thinker or The Logician	INFP	The Idealist or The Mediator
ENTJ	The Executive or The Commander	ENFJ	The Giver or The Protagonist
ENTP	The Visionary or The Debater	ENFP	The Inspirer or The Campaigner
ISTJ	The Duty Fulfiller or The Logistician	ISTP	The Mechanic or The Virtuoso
ISFJ	The Nurturer or The Defender	ISFP	The Artist or The Adventurer
ESTJ	The Guardian or The Executive	ESTP	The Doer or The Entrepreneur
ESFJ	The Caregiver or The Consul	ESFP	The Performer or The Entertainer

adulting.tv



YOUR PERSONALITY TYPE IS:

MEDIATOR (INFP-T)

No one can stop you from dreaming!

MIND

This trait determines how we interact with our environment.

EXTRAVERTED 21% — INTROVERTED 79%

ENERGY

This trait shows where we direct our mental energy.

INTUITIVE 73% — OBSERVANT 27%

NATURE

This trait determines how we make decisions and cope with emotions.

THINKING 25% — FEELING 75%

TACTICS

This trait reflects our approach to work, planning and decision-making.

JUDGING 18% — PROSPECTING 82%

IDENTITY

This trait underpins all others, showing how confident we are in our abilities and decisions.

ASSERTIVE 21% — TURBULENT 79%

Eerily Acurate:

- Idealistic
- Career must serve the greater good
- Creative
- Great writer/communicator
- Bad with data!?!?!?

Dataset

I found this dataset (https://www.kaggle.com/datasnaek/mbti-type) on the Kaggle website.

It is a collection of the last 50 forum posts from 8675 members of the Personality Cafe (http://personalitycafe.com) community.

The Personality Cafe:

'A community dedicated to helping you develop your personality through interactions with people who have the same personality as you.'

```
(8675, 2)
```

	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||http://41.media.tumblr.com/tumblr_lfouy03PMA1qa1ro...
1	ENTP	'I'm finding the lack of me in these posts very alarming.|||Sex can be boring if it's in the sam...
2	INTP	'Good one _____ https://www.youtube.com/watch?v=fHiGbolFFGw|||Of course, to which I say I kno...
3	INTJ	'Dear INTP, I enjoyed our conversation the other day. Esoteric gabbing about the nature of th...
4	ENTJ	'You're fired.|||That's another silly misconception. That approaching is logically is going to b...

Project

- Create a model that can predict a person's personal type based off of their writing.
- Find an easy way for teachers to use this model.

Concerns going into it:

- Is there enough data?
- Will the model be able to predict 16 different personality types accurately?
- Will it work?
- Will the results even make sense?

```
posts    'I'm finding the lack of me in these posts very alarming.|||S
ex can be boring if it's in the same position often. For example me an
d my girlfriend are currently in an environment where we have to creat
ively use cowgirl and missionary. There isn't enough...|||Giving new m
eaning to 'Game' theory.|||Hello *ENTP Grin*  That's all it takes. Tha
n we converse and they do most of the flirting while I acknowledge the
ir presence and return their words with smooth wordplay and more cheek
y grins.|||This + Lack of Balance and Hand Eye Coordination.|||Real IQ
test I score 127. Internet IQ tests are funny. I score 140s or higher.
Now, like the former responses of this thread I will mention that I do
n't believe in the IQ test. Before you banish...|||You know you're an
ENTP when you vanish from a site for a year and a half, return, and fi
nd people are still commenting on your posts and liking your ideas/tho
ughts. You know you're an ENTP when you...|||http://img188.imageshack.
us/img188/6422/6020d1f...
Name: 1, dtype: object
```

Processing the Data

As with all text samples, there is some data processing to be done.

For clean up, I will do the following:

1. replace urls with 'https'
2. punctuation removal (and other random symbols)
3. remove digits
4. lowercase and stop word removal
5. remove excess white space
6. remove the types

```
posts    finding lack posts alarming sex boring position often example
girlfriend currently environment creatively use cowgirl missionary eno
ugh giving new meaning game theory hello  grin takes converse flirting
acknowledge presence return words smooth wordplay cheeky grins lack ba
lance hand eye coordination real iq test score internet iq tests funny
score higher like former responses thread mention believe iq test bani
sh know  vanish site year half return find people still commenting pos
ts liking ideas thoughts know  https think things sometimes go old she
rlock holmes quote perhaps man special knowledge special powers like r
ather encourages seek complex cheshirewolf tumblr com post really neve
r thought e j p real functions judge use use   dominates  emotions rar
ely  also use  due strength know though ingenious saying really want t
ry see happens playing first person shooter back drive around want see
look rock paper one best makes lol guys lucky really high tumblr syste
m hear new first pers...
Name: 1, dtype: object
```

Creating TF-IDF Word Matrix

- TF-IDF (Term Frequency-Inverse Document Frequency) Vectorizer.
- Similiar to a 'bag of words' matrix.
- Each word is weighted depending on how often it shows up in the ENTIRE sample (corpse).
- The more often a word appears, the less distinct it is and so the less weight it gets.
- "Bi-grams"

```
from sklearn.feature_extraction.text import TfidfVectorizer

tf = TfidfVectorizer(analyzer='word', ngram_range=(1,2), min_df = 0.02,
                     stop_words = 'english', norm='l2')

tfidf_matrix = tf.fit_transform(data.posts)

print('Number of documents:', tfidf_matrix.shape[0])
print('Number of features:', tfidf_matrix.shape[1])

print('\nNote: According to my mentor, this is still not a lot of words/features')
```

```
Number of documents: 8675
Number of features: 3459

Note: According to my mentor, this is still not a lot of words/feature
s
```

	abilities	ability	able	absolute	absolutely	absolutely love	abstract	absurd	abuse	abusiv
0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0
1	0.0	0.000000	0.031013	0.000000	0.000000	0.0	0.0	0.0	0.0	0
2	0.0	0.108889	0.037769	0.000000	0.091012	0.0	0.0	0.0	0.0	0
3	0.0	0.000000	0.064068	0.060195	0.000000	0.0	0.0	0.0	0.0	0
4	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0

5 rows × 3459 columns

Predictive Models

Chosen:

- Random Forests: Feature Importance method
- Logistic Regression: Gave me the best accuracy scores

Tried but passed on:

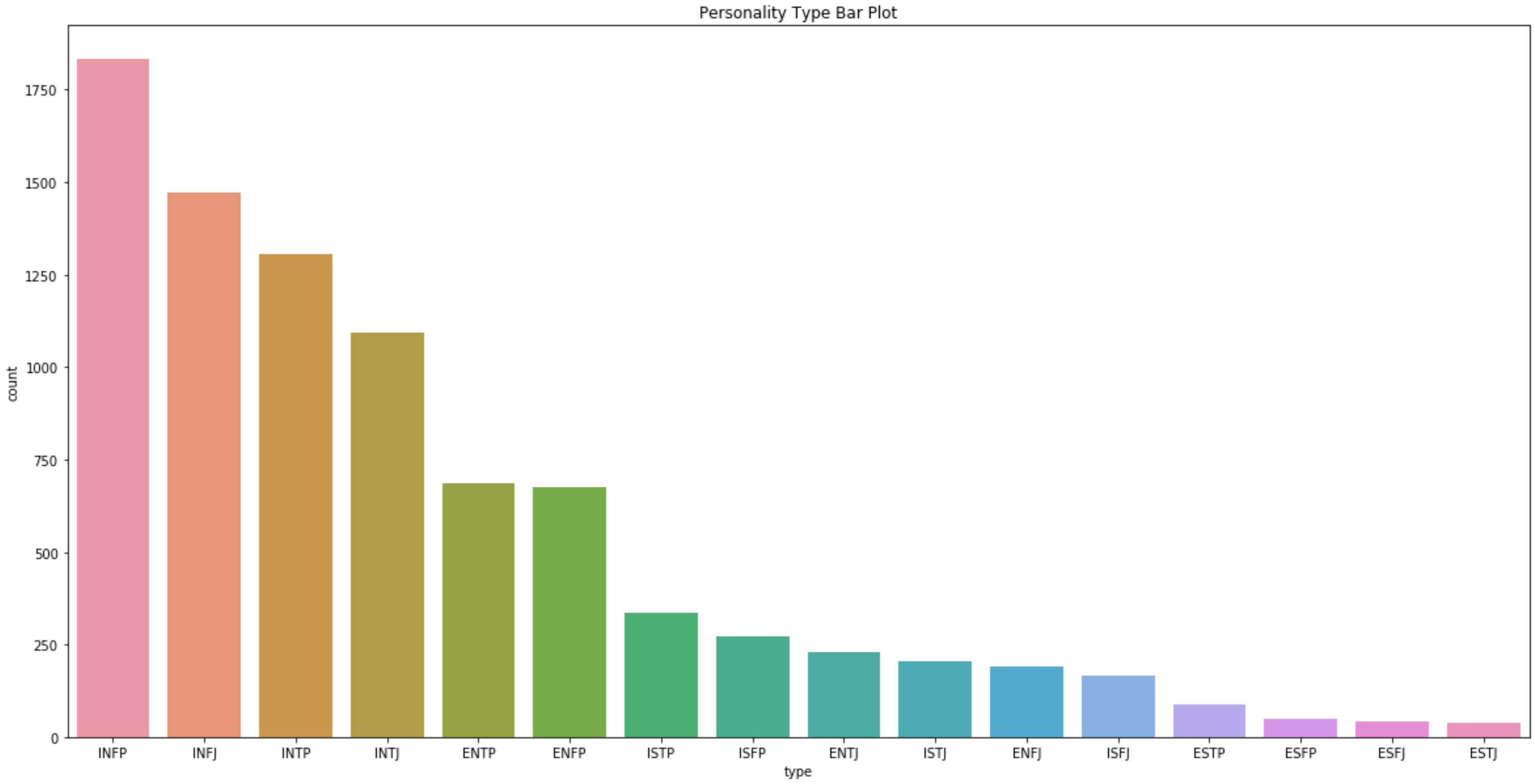
- Support Vector Machine: Took 15 minutes and same accuracy scores as Logistic Regression
- Gradient Boosting: Timed out

Features:

- TF-IDF Matrix

Target:

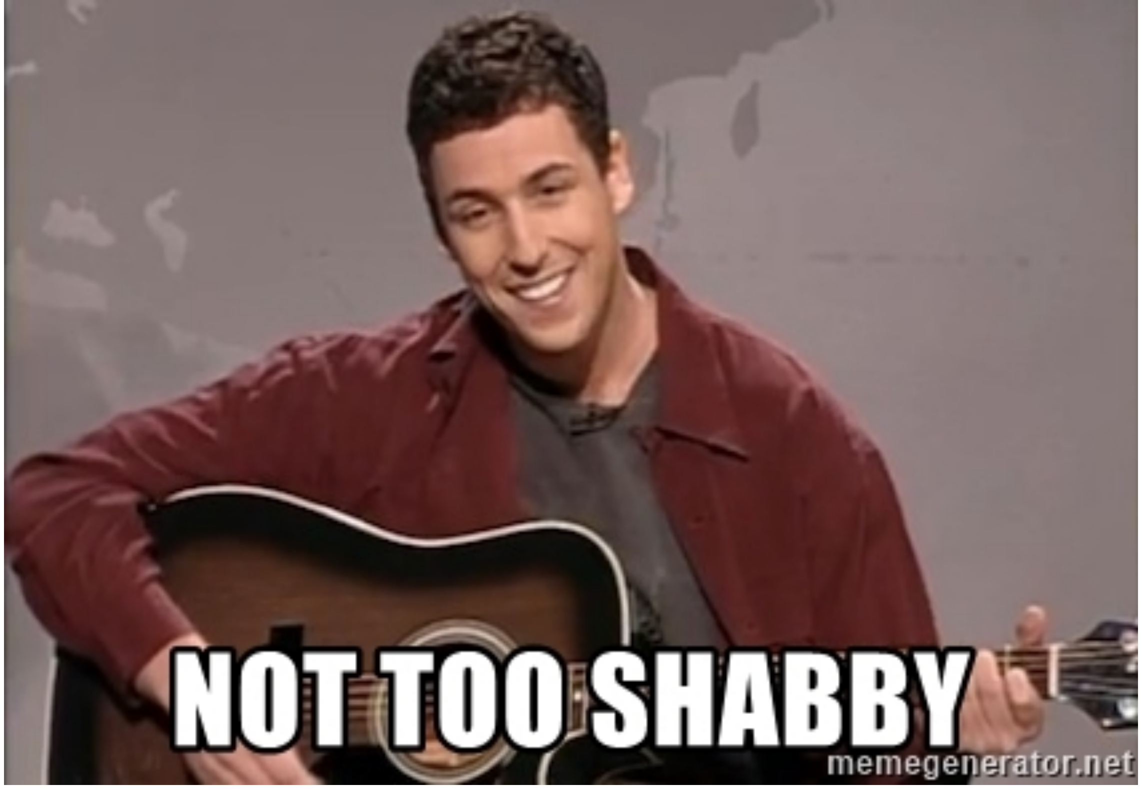
- Personality Types

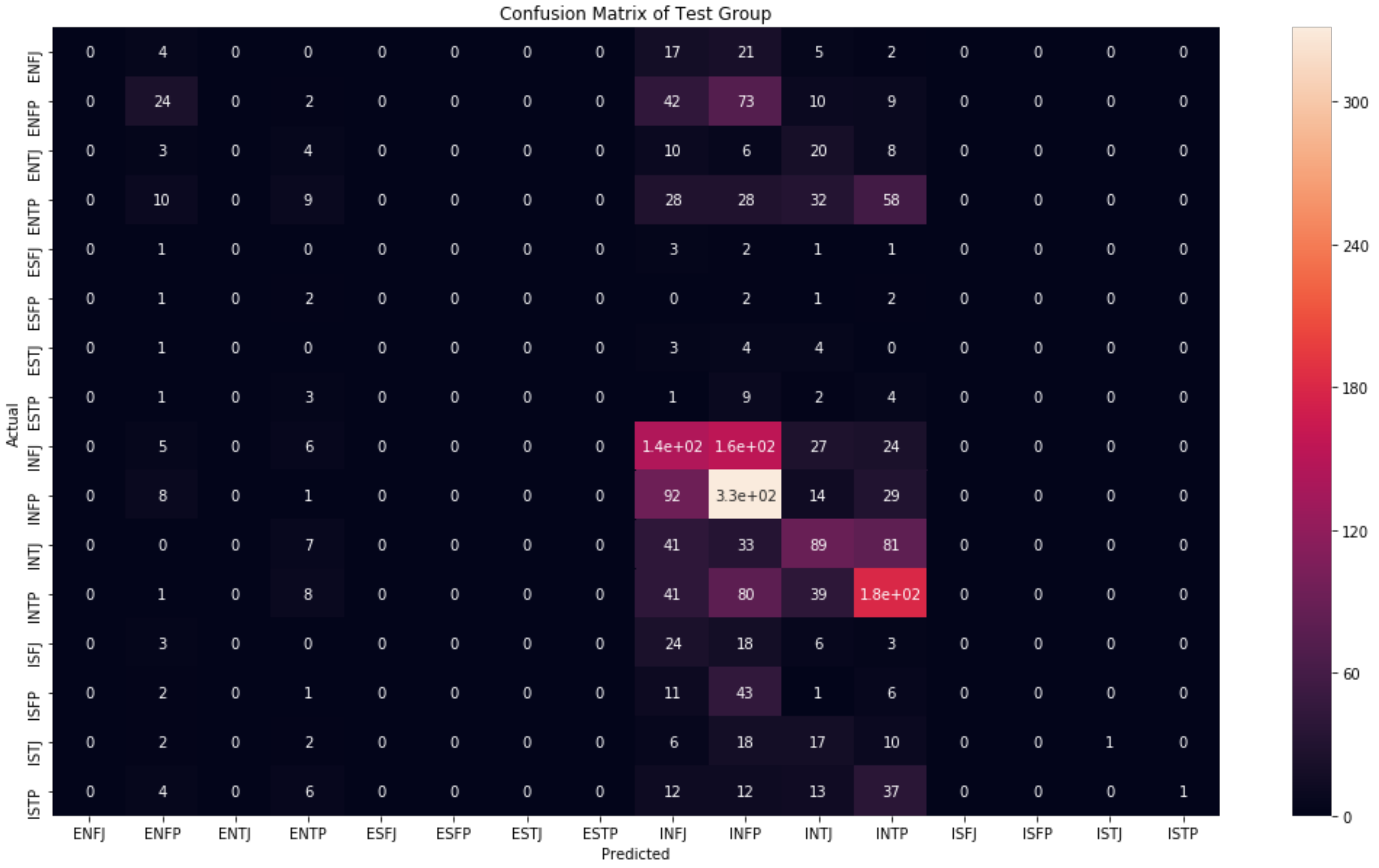


```
Random Forest Classifier:
Training set score: 0.993083307716
Test set score: 0.20378054403

Logistic Regression:
Training set score: 0.583922533046
Test set score: 0.359612724758
```

Still better than completely random:

- 1/16 = 0.0625




Confusion Matrix of Test Group

Actual \ Predicted	ENFJ	ENFP	ENTJ	ENTP	ESFJ	ESFP	ESTJ	ESTP	INFJ	INFP	INTJ	INTP	ISFJ	ISFP	ISTJ	ISTP
ENFJ	0	4	0	0	0	0	0	0	17	21	5	2	0	0	0	0
ENFP	0	24	0	2	0	0	0	0	42	73	10	9	0	0	0	0
ENTJ	0	3	0	4	0	0	0	0	10	6	20	8	0	0	0	0
ENTP	0	10	0	9	0	0	0	0	28	28	32	58	0	0	0	0
ESFJ	0	1	0	0	0	0	0	0	3	2	1	1	0	0	0	0
ESFP	0	1	0	2	0	0	0	0	0	2	1	2	0	0	0	0
ESTJ	0	1	0	0	0	0	0	0	3	4	4	0	0	0	0	0
ESTP	0	1	0	3	0	0	0	0	1	9	2	4	0	0	0	0
INFJ	0	5	0	6	0	0	0	0	1.4e+02	1.6e+02	27	24	0	0	0	0
INFP	0	8	0	1	0	0	0	0	92	3.3e+02	14	29	0	0	0	0
INTJ	0	0	0	7	0	0	0	0	41	33	89	81	0	0	0	0
INTP	0	1	0	8	0	0	0	0	41	80	39	1.8e+02	0	0	0	0
ISFJ	0	3	0	0	0	0	0	0	24	18	6	3	0	0	0	0
ISFP	0	2	0	1	0	0	0	0	11	43	1	6	0	0	0	0
ISTJ	0	2	0	2	0	0	0	0	6	18	17	10	0	0	1	0
ISTP	0	4	0	6	0	0	0	0	12	12	13	37	0	0	0	1

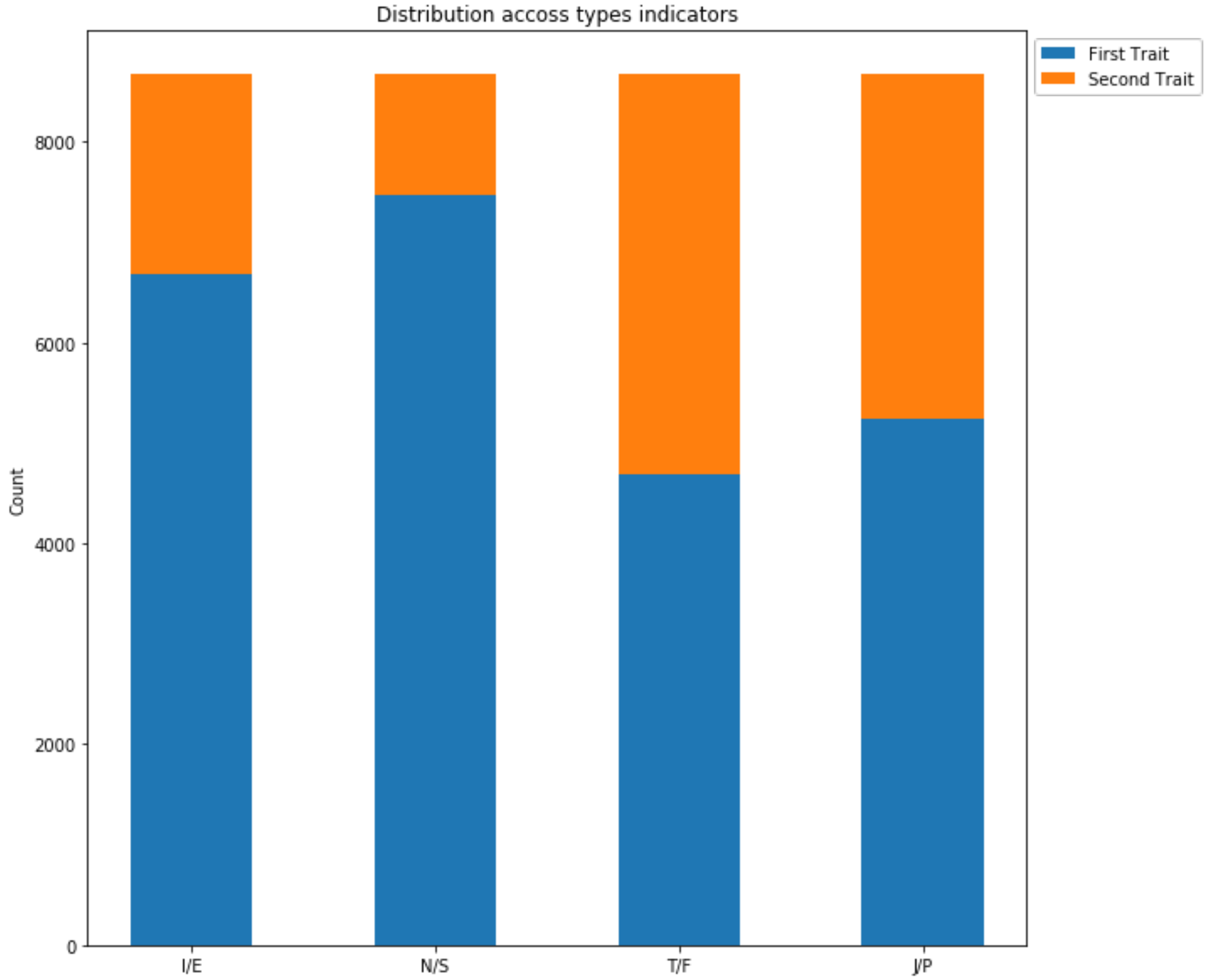
```
             precision    recall  f1-score   support

       ENFJ       0.00      0.00      0.00        49
       ENFP       0.34      0.15      0.21       160
       ENTJ       0.00      0.00      0.00        51
       ENTP       0.18      0.05      0.08       165
       ESFJ       0.00      0.00      0.00         8
       ESFP       0.00      0.00      0.00         8
       ESTJ       0.00      0.00      0.00        12
       ESTP       0.00      0.00      0.00        20
       INFJ       0.30      0.40      0.34       361
       INFP       0.40      0.70      0.51       475
       INTJ       0.32      0.35      0.33       251
       INTP       0.40      0.52      0.45       350
       ISFJ       0.00      0.00      0.00        54
       ISFP       0.00      0.00      0.00        64
       ISTJ       1.00      0.02      0.04        56
       ISTP       1.00      0.01      0.02        85

avg / total       0.34      0.36      0.30      2169
```

Predictive Models Based on Individual Traits

- New columns that would identify each sample by their individual traits: Introvert/Extrovert, Intuitive/Sensing, Thinking/Feeling, and Judging/Prospecting.
- Each model will now be binary instead of having to choose between 16 different categories.
- Take the predictions and put them back together to get the personality type.

	abilities	ability	able	absolute	absolutely	absolutely love	abstract	absurd	abuse	abusiv
0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0
1	0.0	0.000000	0.031013	0.000000	0.000000	0.0	0.0	0.0	0.0	0
2	0.0	0.108889	0.037769	0.000000	0.091012	0.0	0.0	0.0	0.0	0
3	0.0	0.000000	0.064068	0.060195	0.000000	0.0	0.0	0.0	0.0	0
4	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0

5 rows × 3463 columns



- I will take sample sizes to balance the groups.
- In order from smallest to biggest sample sizes...

Intuitive/Sensing Model:

Intutive:

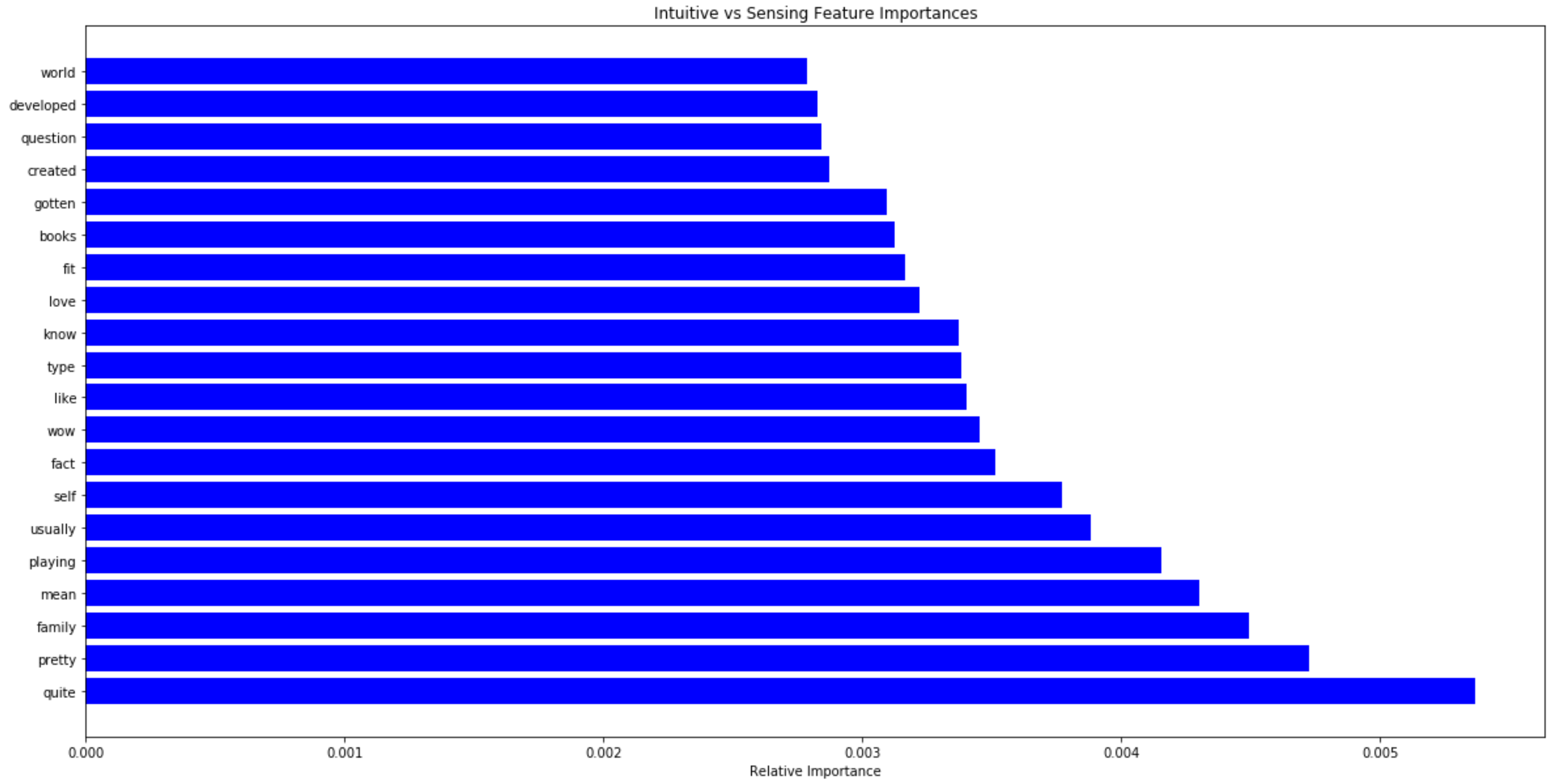
- Likes abstract thinking
- Looks beyond the physical

Sensing:

- Likes concrete information
- Relies mostly on their senses

```
Random Forest Classifier:
Training set score: 0.989415041783
Cross Validation Scores:
 [ 0.53333333  0.53333333  0.55020921  0.53138075  0.51046025]
Average score:  0.531743375174

Logistic Regression:
Training set score: 0.899164345404
Cross Validation Scores:
 [ 0.67708333  0.68125     0.67782427  0.71966527  0.66317992]
Average score:  0.68380055788
```

- Relatively unstable Random Forest Feature Importance



- Reoccuring words include: believe, fact, idea, world, think.

Introverted/Extroverted Model:

Introverted:

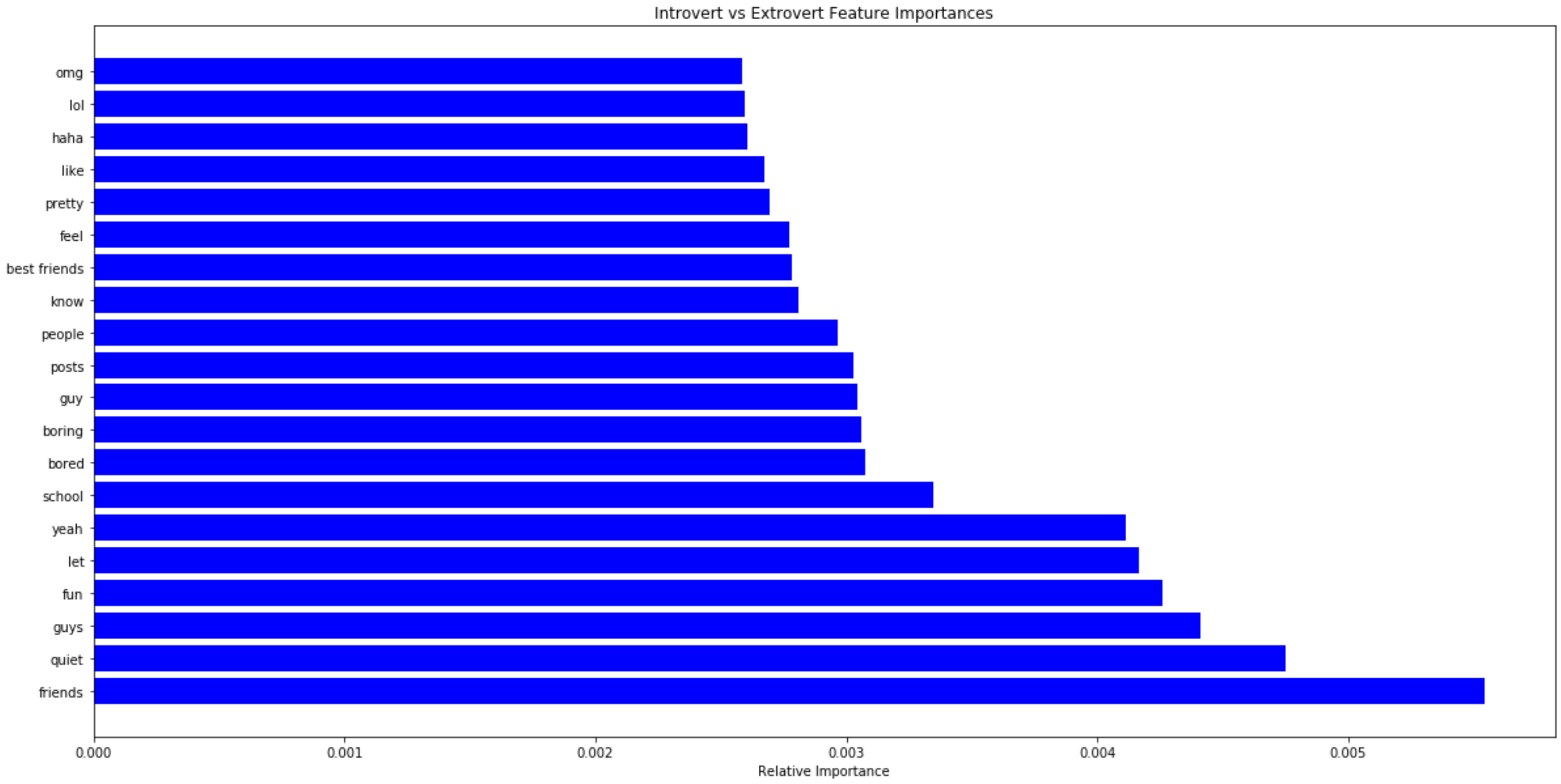
- Low threshold for stimulation
- Gets tired easily being around other people

Extroverted:

- High threshold for stimulation
- Gets energized around others

```
Random Forest Classifier:
Training set score: 0.984656437625
Cross Validation Scores:
 [ 0.52625     0.55        0.56625     0.5375      0.53258145]
Average Score:  0.542516290727

Logistic Regression:
Training set score: 0.847898599066
Cross Validation Scores:
 [ 0.6575      0.6625      0.70125     0.6825      0.67919799]
Average Score:  0.676589598997
```

- Slightly better Random Forest accuracy
- Slightly worse Logistic Regression accuracy
- Slightly more stable Feature Importance



- Reoccuring words: fun, friends, bored, lol, haha, think, and quiet.
- Might be easier to distinguish Introvert vs Extrovert words.

Judging/Prospecting Model:

Judging:

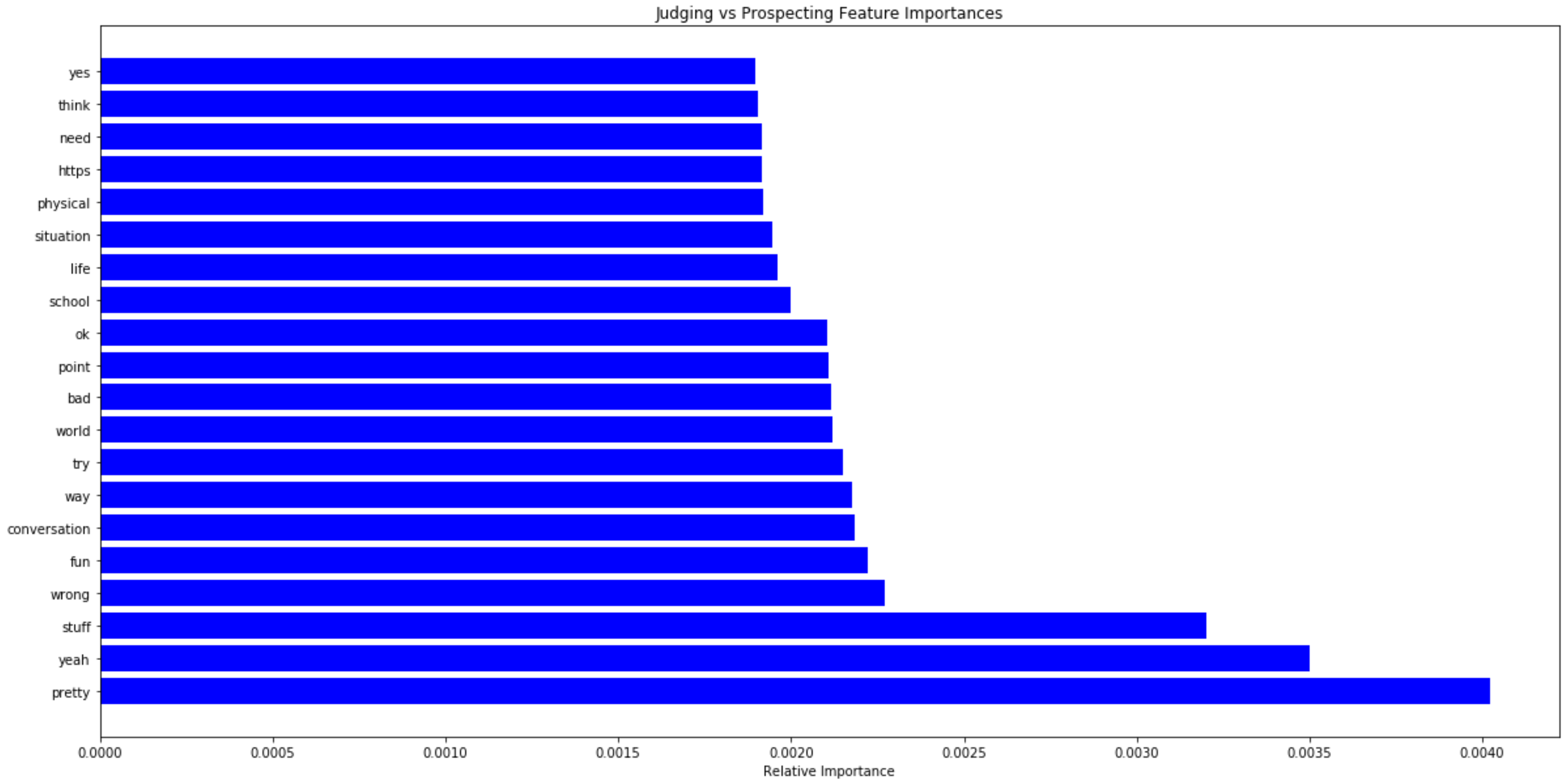
- Likes structure

Prospecting:

- Likes spontaneity

```
Random Forest Classifier:
Training set score: 0.9846631722
Cross Validation Scores:
 [ 0.52256186  0.50727802  0.53275109  0.51237263  0.53717201]
Average Scores:  0.522427124312

Logistic Regression:
Training set score: 0.80198019802

Cross Validation Scores:
 [ 0.63027656  0.61135371  0.62299854  0.62736536  0.62026239]
Average Scores:  0.622451313651
```

- Overall worst accuracy scores
- Random Forest: about 1% worse
- Logistic Regression: about 5% worse
- Very unstable Feature Importance



- Reoccuring words: plan, simply, time, situation, try
- Maybe the structure of the sentences would be more telling than word choice.

Thinking/Feeling Model:

Thinking:

- Base decisions on facts
- Very logical

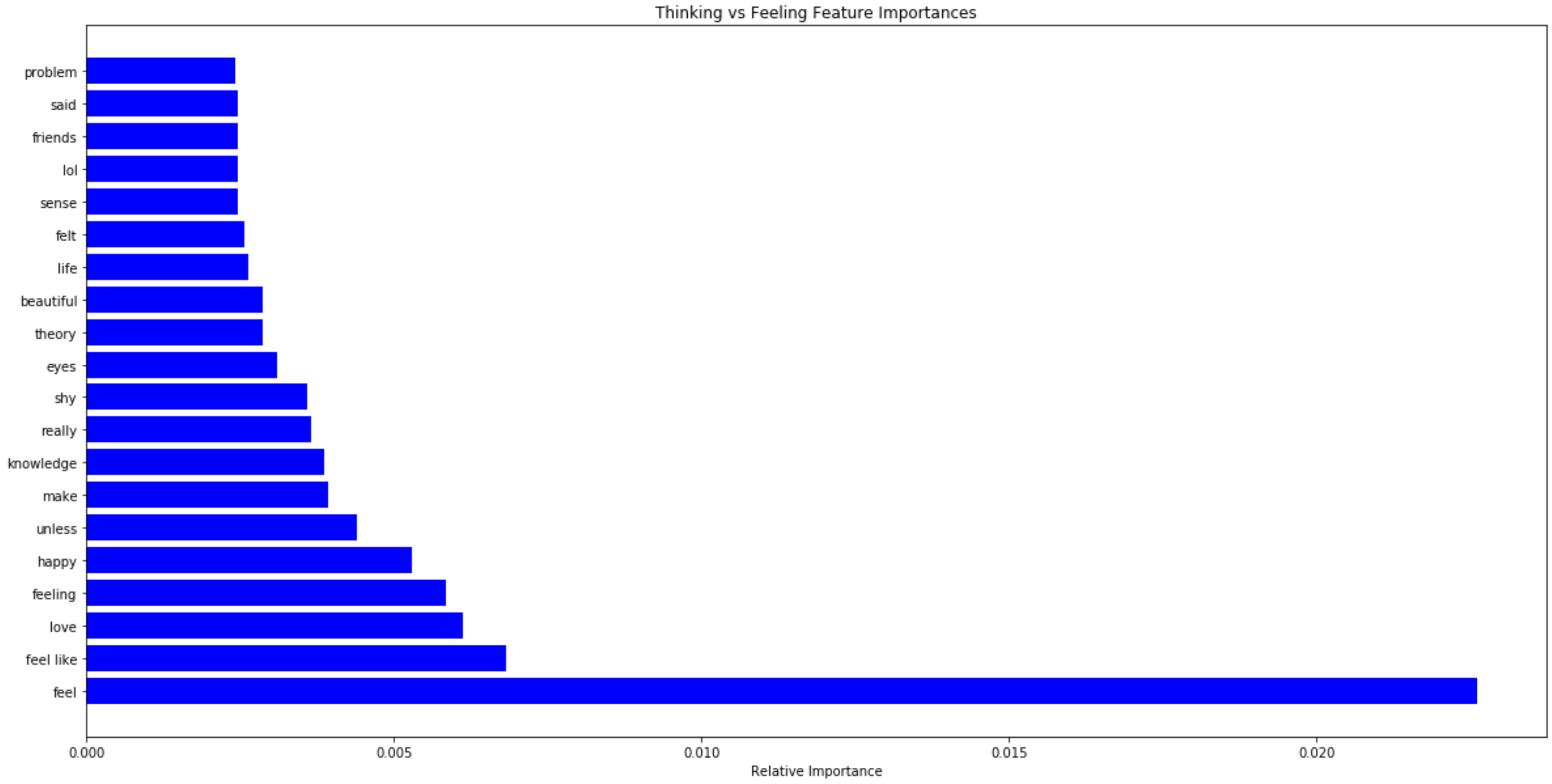
Feeling:

- Base decisions of emotions

- Very empathetic

```
Random Forest Classifier:
Training set score: 0.9886258838
Cross Validation Scores:
 [ 0.63421659  0.62708934  0.61325648  0.63631124  0.63956171]
Average Score:  0.630087071483

Logistic Regression:
Training set score: 0.861512450046
Cross Validation Scores:
 [ 0.81278802  0.78847262  0.78270893  0.78789625  0.78200692]
Average Score:  0.790774549729
```

- Best Random Forest accuracy: 10% jump
- Best Logistic Regression accuracy: 10% jump
- Most stable Feature Importance



- Reoccuring words: feel, love, feel like, beautiful, thank, life, happy, hope, felt, heart

Twitter Application

```
Using Tweepy I searched Twitter posts with #INFP
and scraped the following Twitter handles:

['Buffy A Summers #SlipKid', 'Moonlight Night', 'ج .', 'JustPlainJane'
, 'Obaasant @144p', 'Joyce', '🌿 MICHI 🌿', 'VirtualOfficeSales', '🌈'
, 'ExerciseinFrugality', '¹⁶', '#INFP .♡', '´', '- لـوسا\u200fن', 'Ely
Bakouche', 'Gary Smith', 'Rebecca']
```

```python
def get_twitter_type(twitter_handle):
    auth.set_access_token(atoken, asecret)

    api = tweepy.API(auth)

    stuff = api.user_timeline(screen_name = twitter_handle, count = 300, include_rts

    twitter_text = ''
    for status in stuff:
        twitter_text += status.text
        twitter_text += ' '

    twitter_text = remove_url(twitter_text)
    twitter_text = remove_puncuation(twitter_text)
    twitter_text = remove_digits(twitter_text)
    twitter_text = remove_stop_words(twitter_text)
    twitter_text = remove_extra_white_space(twitter_text)
    twitter_text = remove_types(twitter_text)

    my_tfidf_matrix2 = tf.transform([twitter_text])

    my_tfidf_feature_matrix2 = pd.DataFrame(my_tfidf_matrix2.toarray(), columns=tf.
    my_tfidf_feature_matrix2.head()

    print(lr_IE.predict(my_tfidf_feature_matrix2)[0]
          + lr_NS.predict(my_tfidf_feature_matrix2)[0]
          + lr_TF.predict(my_tfidf_feature_matrix2)[0]
          + lr_JP.predict(my_tfidf_feature_matrix2)[0])
```

get_twitter_type(YourTwitterHandleHere)

```
get_twitter_type('VincentCleopeGo')
get_twitter_type('Vivianne Ouya')
get_twitter_type('Moonlight Night')
get_twitter_type('Eliza Kinde')
get_twitter_type('inkandstars')
```

```
ESFP
ESFP
ESFP
ESFP
INFP
```

- The first 2 letters were the models with the LEAST amount of samples.
- The last letter had the worse accuracy scores but still got all of them correct.
- Third letter had the highest accuracy score and were all correct.
- 'inkandstars' was the most self aware and had the most Tweets that talked about their personality type and what it meant, similiar to the posts found on Personality Cafe.

Summary

- It is possible to make it work.
- Need more samples.
- Sample text from people who are not self-aware of their type.
- Add other features.