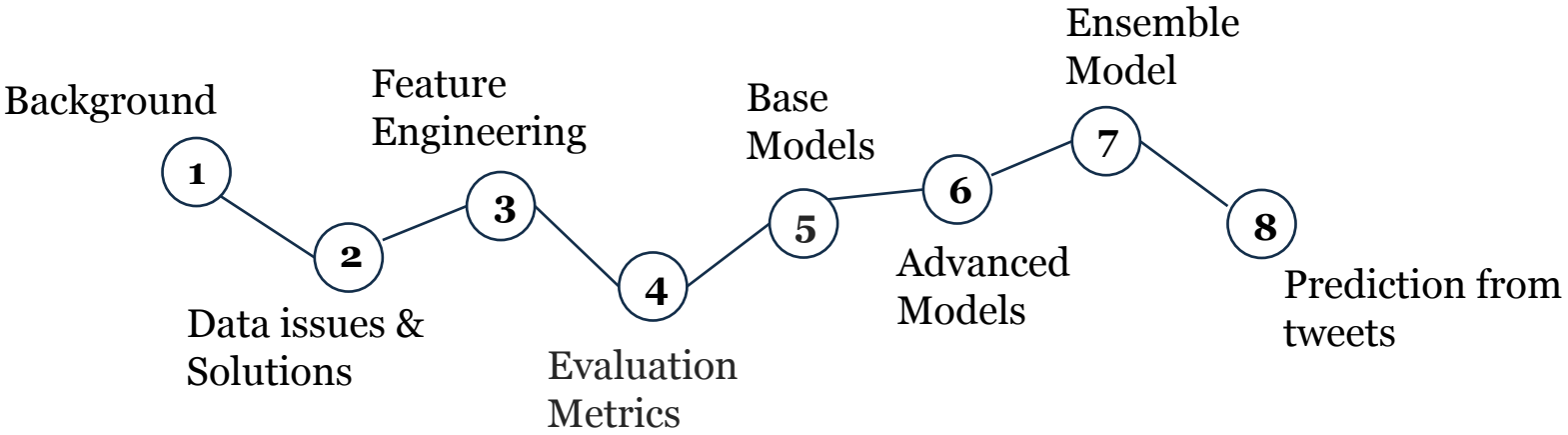
Predicting MBTI Personality from Forum Posts

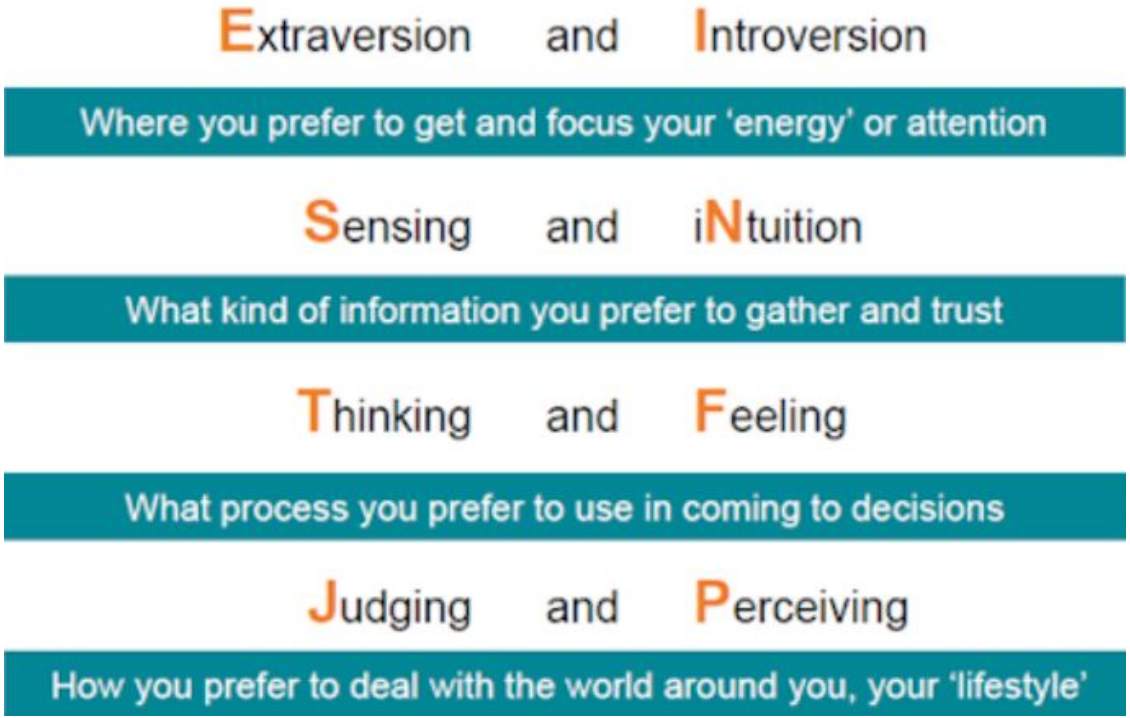
The Introverts

Tomo, Zack, Ben, Nan, & Donya

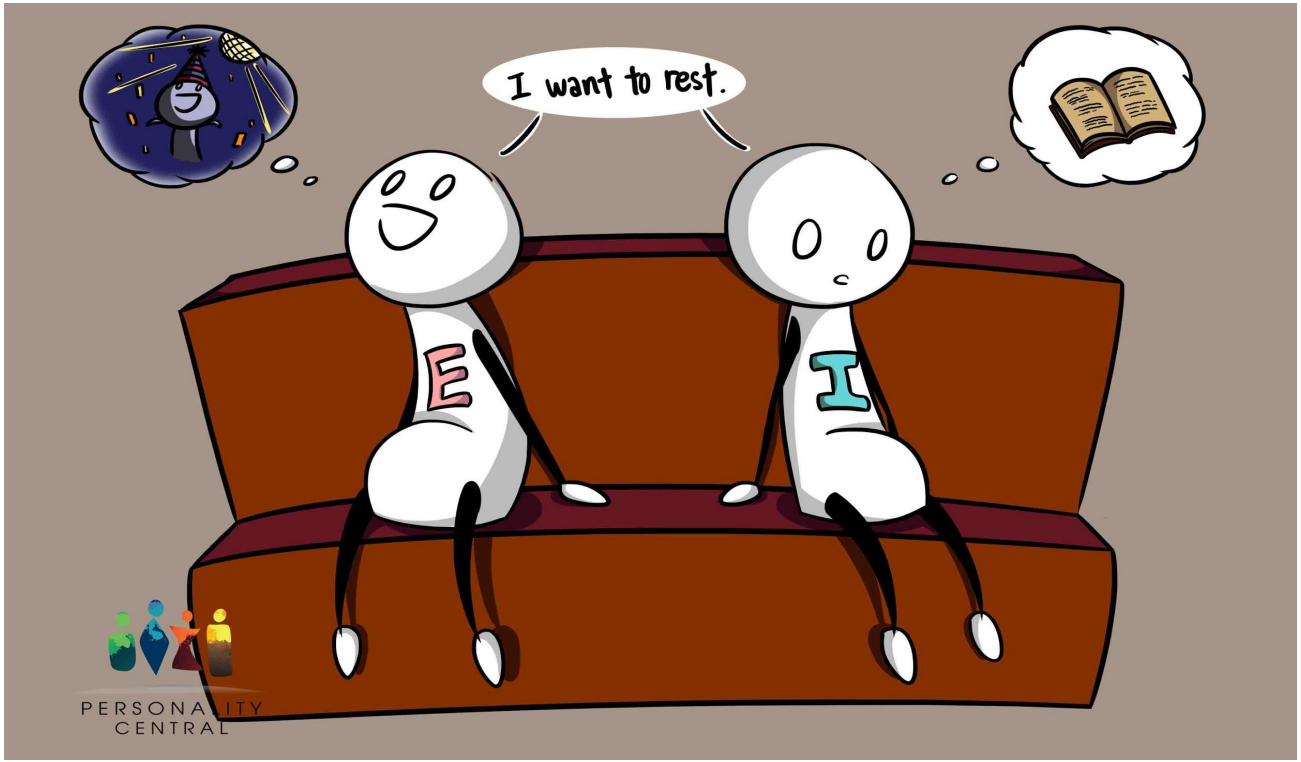
Agenda


Background
1
Data issues & Solutions
2
Feature Engineering
3
Evaluation Metrics
4
Base Models
5
Advanced Models
6
Ensemble Model
7
Prediction from tweets
8

What are Myers-Briggs Type Indicators (MBTI)?



Extroverted vs Introverted

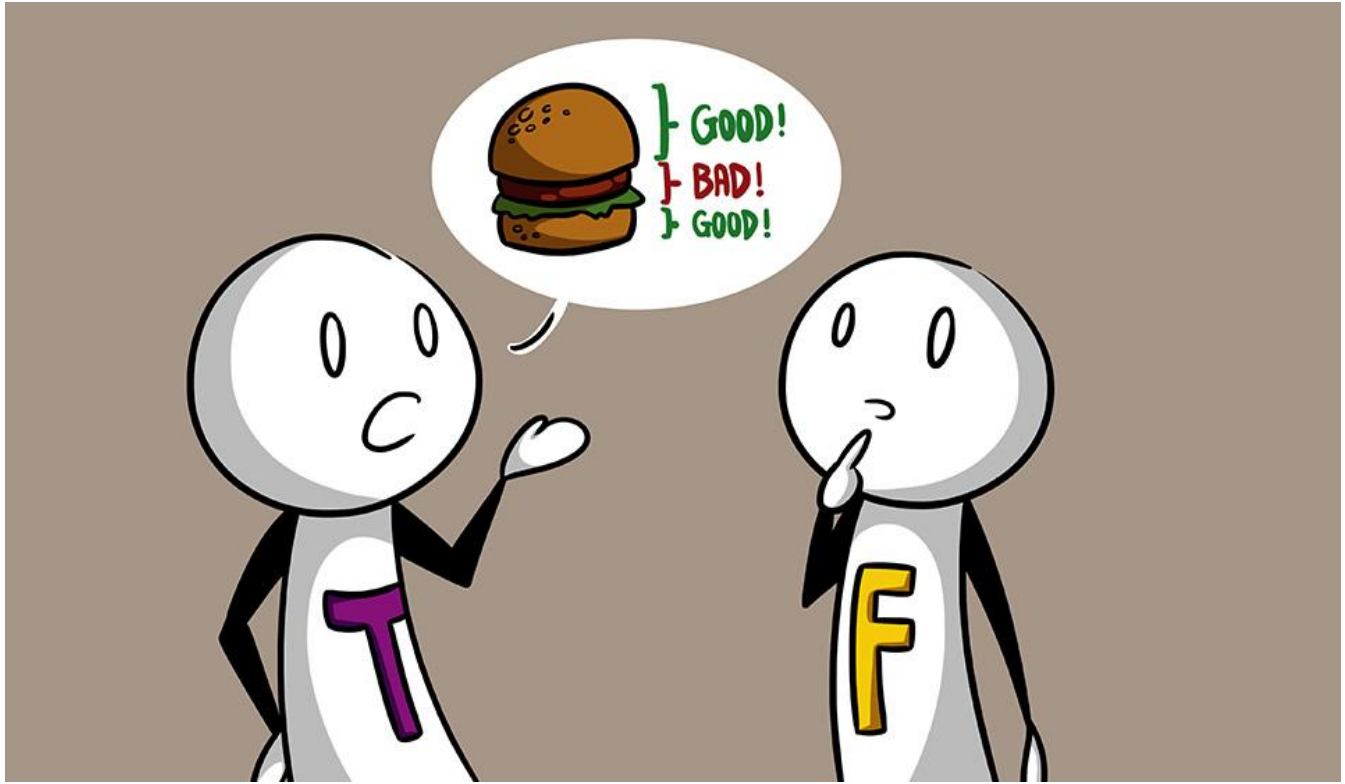


How a person gets energy

iNtuition vs Sensing

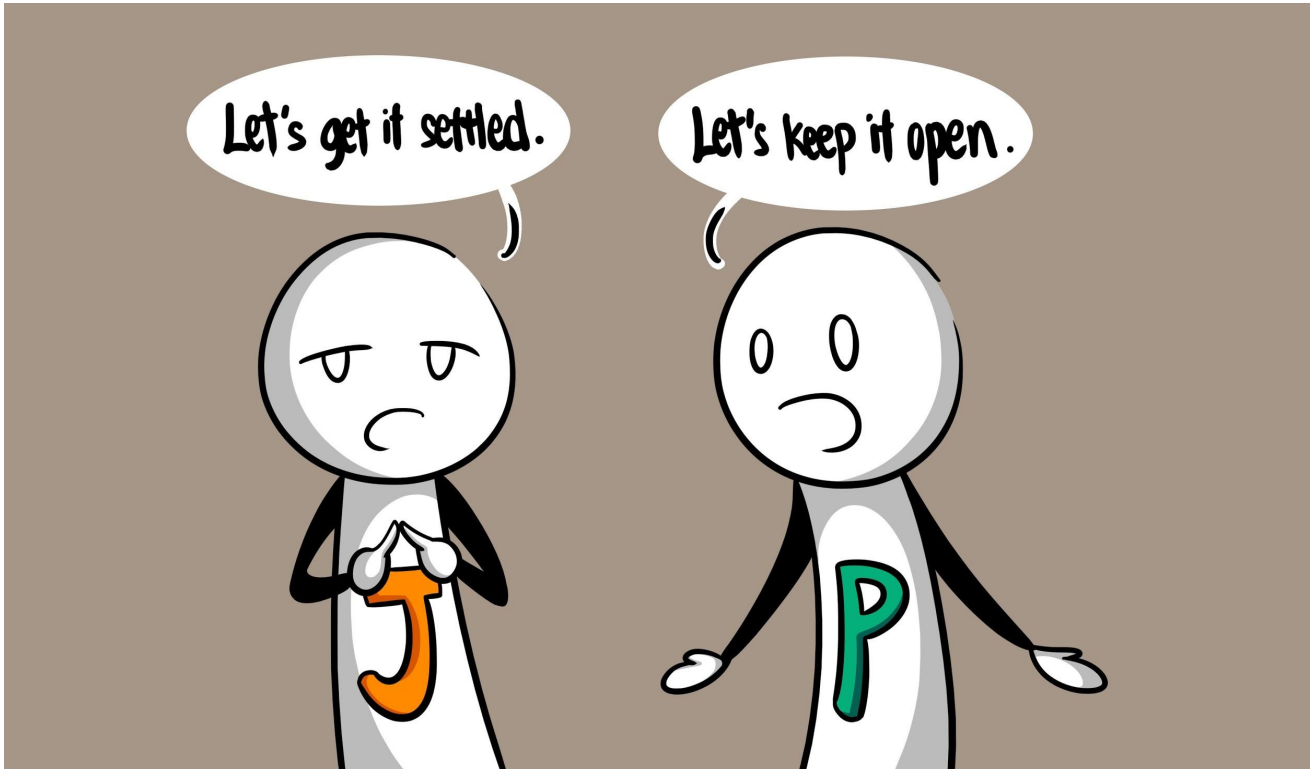
How a person takes in information

Thinking vs Feeling



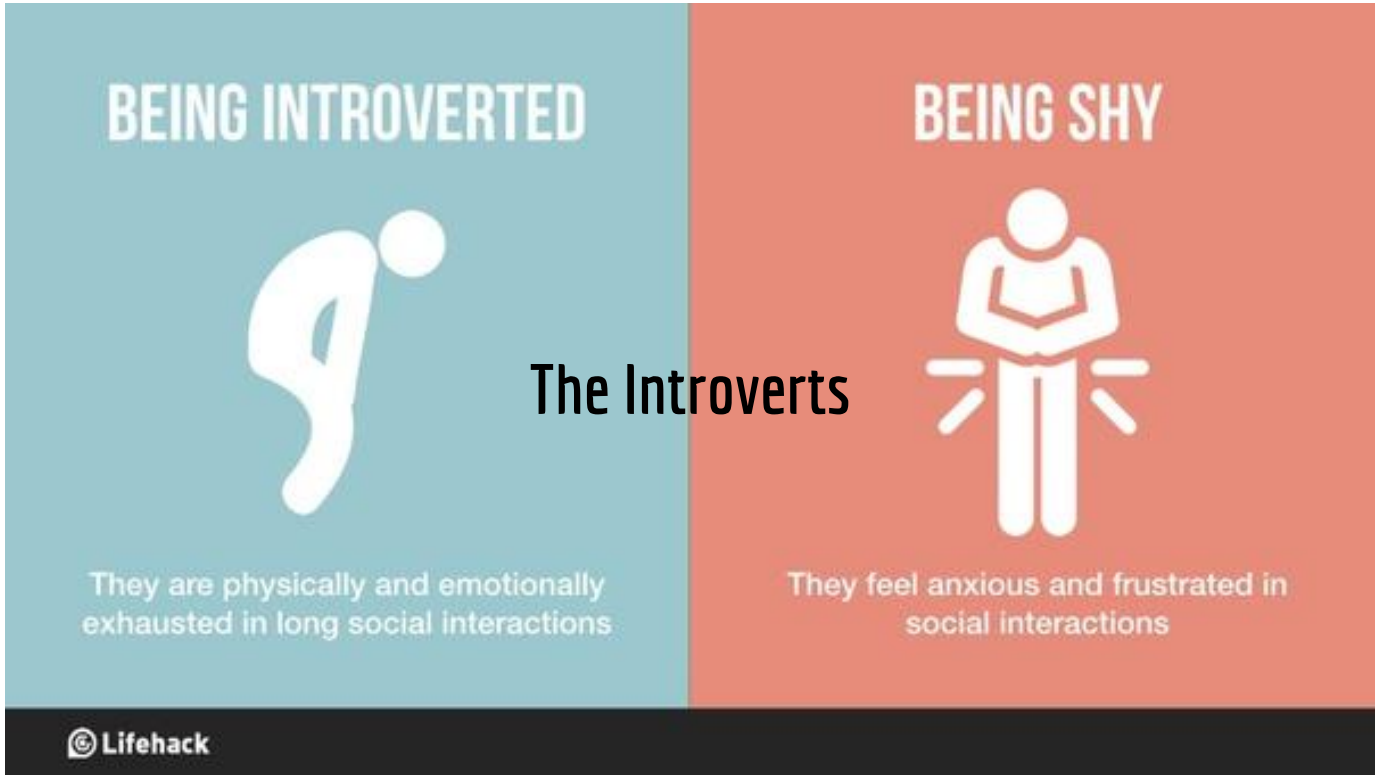
How a person makes decisions

Judging vs Perceiving



The speed with which a person makes decisions

We are all Introverted…



The Introverts

Objective

Developing classifiers that are able to identify **personality** traits from text

Expectation

The predictions of the developed classifiers will correlate with true labels, but **can not be nearly perfect**

Assumption

Personality types in each of the four dimensions can be derived **independently.**

About the dataset

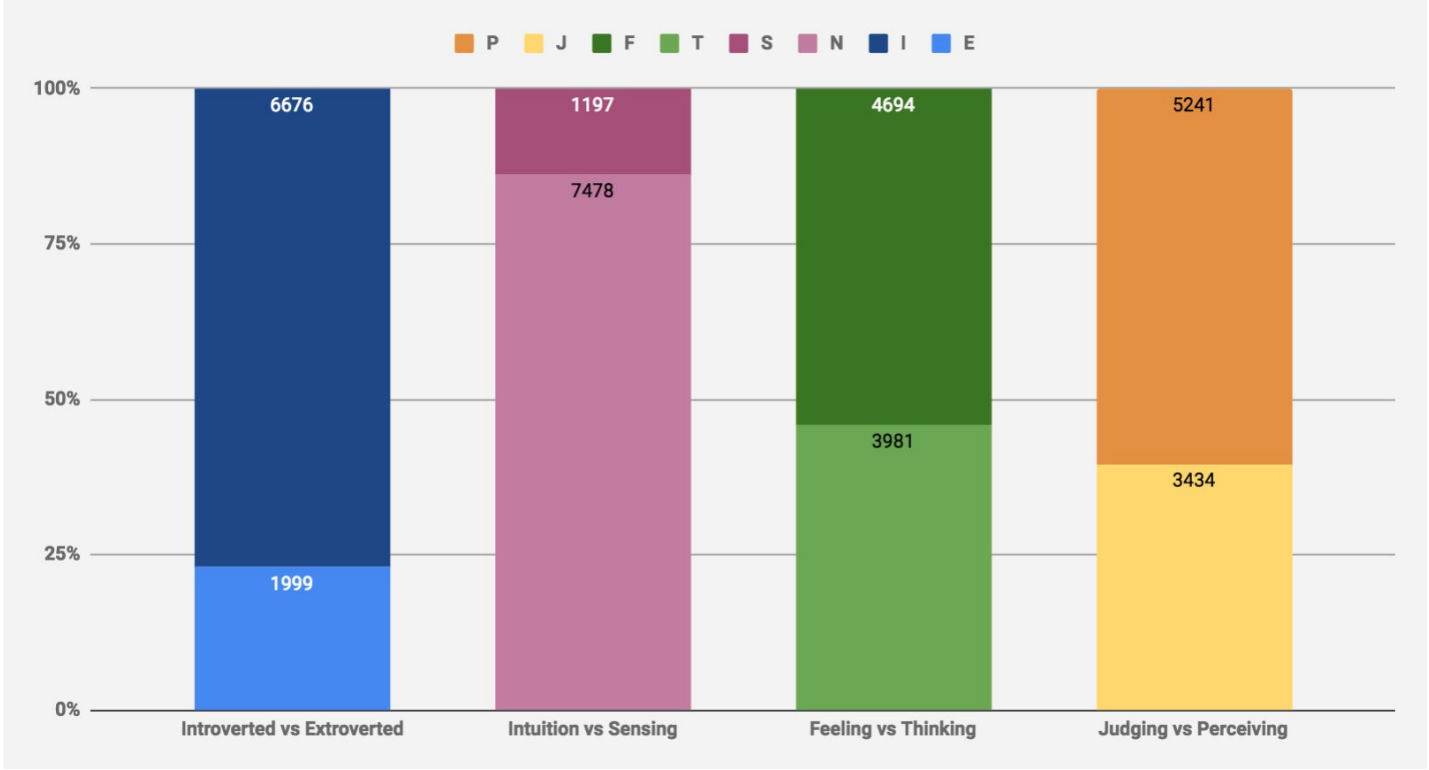
- 50 posts per user from 8600 users on a forum
- 200 character limit for each post
- Labeled with MBTI personality types

type,posts
INFJ,"'http://www.youtube.com/watch?v=qsXHcwe3krw|||http://41.media.tumblr.com/tumblr_lfouy03PMA1qa1rooo1_500.jpg|||enfp and intj moments https://www.youtube.com/watch?v=iz7lE1g4XM4 sportscenter not top ten plays https://www.youtube.com/watch?v=uCdfze1etec pranks|||What has been the most life-changing experience in your life?|||http://www.youtube.com/watch?v=vXZeYwwRDw8 http://www.youtube.com/watch?v=u8ejam5DP3E On repeat for most of today.|||May the PerC Experience immerse you.|||The last thing my INFJ friend posted on his facebook before committing suicide the next day. Rest in peace~ http://vimeo.com/22842206|||Hello ENFJ7. Sorry to hear of your distress. It's only natural for a relationship to not be perfection all the time in every moment of existence. Try to figure the hard times as times of growth, as...|||84389 84390 http://wallpaperpassion.com/upload/23700/friendship-boy-and-girl-wallpaper.jpg http://assets.dornob.com/wp-content/uploads/2010/04/round-home-design.jpg ...|||Welcome and stuff.|||http://playeressence.com/wp-content/uploads/2013/08/RED-red-the-pokemon-master-32560474-450-338.jpg Game. Set. Match.|||Prozac, wellbrutin, at least thirty minutes of moving your legs (and I don't mean moving them while sitting in your same desk chair), weed in moderation (maybe try edibles as a healthier alternative...|||Basically come up with three items you've determined that each type (or whichever types you want to do) would more than likely use, given each types' cognitive functions and whatnot, when left by...|||All things in moderation. Sims is indeed a video game, and a good one at that. Note: a good one at that is somewhat subjective in that I am not completely promoting the death of any given Sim...|||Dear ENFP: What were your favorite video games growing up and what are your now, current favorite video games? :cool:|||https://www.youtube.com/watch?v=QyPqT8umzmY|||It appears to be too late. :sad:|||There's someone out there for everyone.|||Wait... I thought confidence was a good thing.|||I just cherish the time of solitude b/c i revel within my inner world more whereas most other time i'd be workin... just enjoy the me time while you can. Don't worry, people will always be around to...|||Yo entp ladies... if you're into a complimentary personality,well, hey.|||... when your main social outlet is xbox live conversations and even then you verbally fatigue quickly.|||http://www.youtube.com/watch?v=gDhy7rdfm14 I really dig the part from 1:46 to 2:50|||http://www.youtube.com/watch?v=msqXffgh7b8|||Banned because this thread requires it of me.|||Get high in backyard, roast and eat marshmellows in backyard while conversing over something intellectual, followed by massages and kisses.|||http://www.youtube.com/watch?v=Mw7eoU3BMbE|||http://www.youtube.com/watch?v=4V2uYORhQOk|||http://www.youtube.com/watch?v=SlVmgFQQ0TI|||Banned for too many b's in that sentence. How could you! Think of the B!|||Banned for watching movies in the corner with the dunces.|||Banned because Health class clearly taught you nothing about peer pressure.|||Banned for a whole host of reasons!|||http://www.youtube.com/watch?v=IRcrv41hgz4|||1) Two baby deer on left and right munching on a beetle in the middle. 2) Using their own blood, two cavemen diary today's latest happenings on their designated cave diary wall. 3) I see it as...|||a pokemon world an infj society everyone becomes an optimist|||49142|||http://www.youtube.com/watch?v=ZRCEq_JFeFM|||http://discovermagazine.com/2012/jul-aug/20-things-you-didnt-know-about-deserts/desert.jpg|||http://oyster.ignimgs.com/mediawiki/apis.ign.com/pokemon-silver-version/d/dd/Ditto.gif|||http://www.serebii.net/potw-dp/Scizor.jpg|||Not all artists are artists because they draw. It's the idea that counts in forming something of your own... like a signature.|||Welcome to the robot ranks, person who downed my self-esteem cuz I'm not an avid signature artist like herself. :proud:|||Banned for taking all the room under my bed. Ya gotta learn to share with the roaches.|||http://www.youtube.com/watch?v=w8IgImn57aQ|||Banned for being too much of a thundering, grumbling kind of storm... yep.|||Ahh... old high school music I haven't heard in ages. http://www.youtube.com/watch?v=dcCRUPCdB1w|||I failed a public speaking class a few years ago and I've sort of learned what I could do better were I to be in that position again. A big part of my failure was just overloading myself with too...|||I like this person's mentality. He's a confirmed INTJ by the way. http://www.youtube.com/watch?v=hGKLI-GEc6M|||Move to the Denver area and start a new life for myself.'"
ENTP,"'I'm finding the lack of me in these posts very alarming.|||Sex can be boring if it's in the same position often. For example me and my girlfriend are currently in an environment where we have to creatively use cowgirl and missionary. There isn't enough...|||Giving new meaning to 'Game' theory.|||Hello *ENTP Grin* That's all it takes. Than we converse and they do most of the flirting while I acknowledge their presence and return their words with smooth wordplay and more cheeky 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 don'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 find people are still commenting on your posts and liking your ideas/thoughts. You know you're an ENTP when you...|||http://img188.imageshack.us/img188/6422/6020d1f9da6944a6b71bbe6.jpg|||http://img.adultdvdtalk.com/813a0c6243814cab84c51|||I over think things sometimes. I go by the old Sherlock Holmes quote. Perhaps, when a man has special knowledge and special powers like my own, it rather encourages him to seek a complex...|||cheshirewolf.tumblr.com So is I :D|||400,000+ post|||Not really; I've never thought of E/I or J/P as real functions. I judge myself on what I use. I use Ne and Ti as my dominates. Fe for emotions and rarely Si. I also use Ni due to me strength...|||You know though. That was ingenious. After saying it I really want to try it and see what happens with me playing a first person shooter in the back while we drive around. I want to see the look on...|||out of all of them the rock paper one is the best. It makes me lol. You guys are lucky :D I'm really high up on the tumblr system.|||So

Data Issues

1. *Dataset is not large*
2. Data labels are imbalanced
3. Text Data is messy
4. No descriptive features

Imbalanced Classes



Solutions

1. Parameters in models
2. Performance metrics: AUC-ROC, PR-ROC, F1-Score
3. Resampling

Data Issues

1. *Dataset is not large*
2. Data labels are imbalanced
3. Text Data is messy → Iteratively data processing
4. No descriptive features

Data Preprocessing

1. Removed:
    a. links
    b. punctuation
    c. MBTI-types words
    d. stop words
2. Lowercase
3. Lemmatized words
4. Words like “LOL, fun, guy”

Iterate… Iterate… Iterate……

Data Issues

1. *Dataset is not large*
2. Data labels are imbalanced
3. Text Data is messy
4. No descriptive features → Feature engineering

Two Sets of Features

01 Hand Crafted Features

- Sentiment Score
- Ellipses Count
- Exclamation Count
- Question Mark Count
- Upper Case Count
- Link Count
- Picture Count
- Emojis Count :)

02 Vectorizers

- TFIDF
- Counter

Modeling

Base Models:

- Naive Bayes
- Logistic Regression
- Random Forest

Advanced Models:

- Xgboost
- LightGBM

Ensemble:

- VotingClassifier

Multinomial Naive Bayes

- Discrete features
- Effective and Mathematically Sensible

Model	E or I		N or S		T or F		J or P		Execution Time(s)
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	
MNB	0.60	0.40	0.56	0.26	0.69	0.67	0.57	0.75	3.91

Logistic Regression

- Also do well in text
- Simple and interpretable

Green means the best in that column

Model	E or I		N or S		T or F		J or P		Execution Time(s)
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	
MNB	0.60	0.40	0.56	0.26	0.69	0.67	0.57	0.75	**3.91**
LR	**0.74**	**0.50**	**0.74**	**0.42**	**0.86**	**0.78**	**0.71**	**0.77**	40.34

Random Forest

- Bagging makes it stable
- Always good to try

Model	E or I		N or S		T or F		J or P		Execution Time(s)
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	
MNB	0.60	0.40	0.56	0.26	0.69	0.67	0.57	0.75	3.91
LR	0.74	0.50	0.74	0.42	0.86	0.78	0.71	0.77	40.34
RF	0.71	0.47	0.69	0.33	0.82	0.75	0.67	0.76	162.46

Xgboost (eXtreme Gradient Boosting)

- Boosting trees
- Famous in Kaggle

Model	E or I		N or S		T or F		J or P		Execution Time(s)
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	
MNB	0.60	0.40	0.56	0.26	0.69	0.67	0.57	0.75	3.91
LR	0.74	0.50	0.74	0.42	0.86	0.78	0.71	0.77	40.34
RF	0.71	0.47	0.69	0.33	0.82	0.75	0.67	0.76	162.46
XGB	0.72	0.48	0.68	0.34	0.84	0.76	0.70	0.77	1339.4

LightGBM (Light Gradient Boosting Machine)

- **Much** faster than XGB
- Low memory usage

Model	E or I		N or S		T or F		J or P		Execution Time(s)
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	
MNB	0.60	0.40	0.56	0.26	0.69	0.67	0.57	0.75	3.91
LR	0.74	0.50	0.74	0.42	0.86	0.78	0.71	0.77	40.34
RF	0.71	0.47	0.69	0.33	0.82	0.75	0.67	0.76	162.46
XGB	0.72	0.48	0.68	0.34	0.84	0.76	0.70	0.77	1339.4
LGBM	0.72	0.49	0.69	0.35	0.85	0.77	0.69	0.77	221.14

Hyperparameter Tuning for LightGBM

- Random Search (n_interations=60)

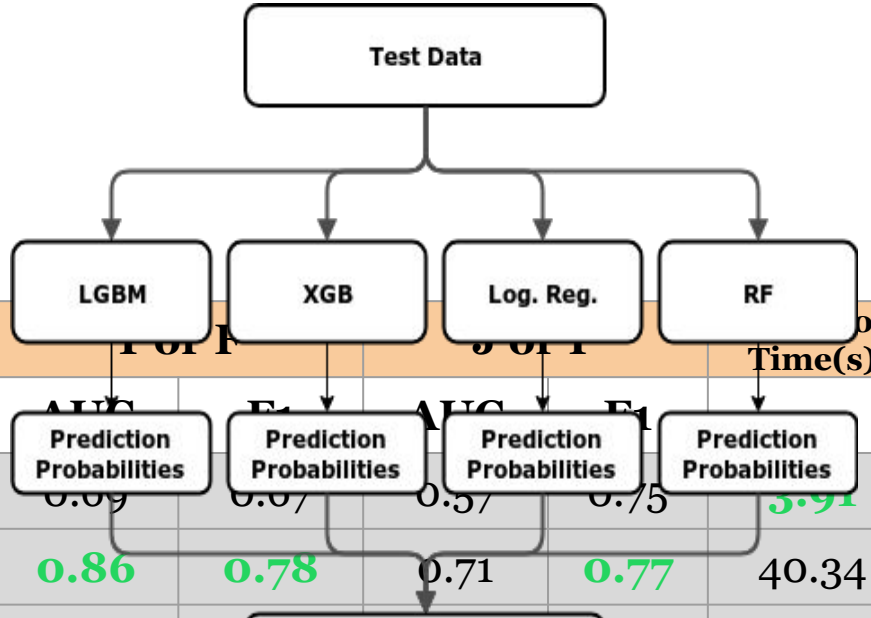
```
for i in range(0, n_iterations):

    param_dist = {'n_es                    50,300,350,400,450]),
                  'bagging_                5, 0.7, 0.8, 0.9]),
                  'learning                 0.1, 0.3, 0.5]),
                  'is_unbal
                  'max_bin'                5, 18, 20, 25]),
                  'boosting                ', 'dart']),
                  'max_dept                ,
                  'feature_                7, 0.8, 0.9]),
                  'lambda_l                , 30, 40]),
                  'objectiv
                  'metric':
```

Logistic Smile

Mod									Execution Time(s)
MN									3.91
LR									40.34
RF									162.46
XG									1339.4
LGBM	0.72	0.49	0.			0.77	0.69	0.77	221.14
LGBM-T	0.73	0.50	0.72	0.38	0.85	0.77	0.70	0.77	90.56

Ensemble Model

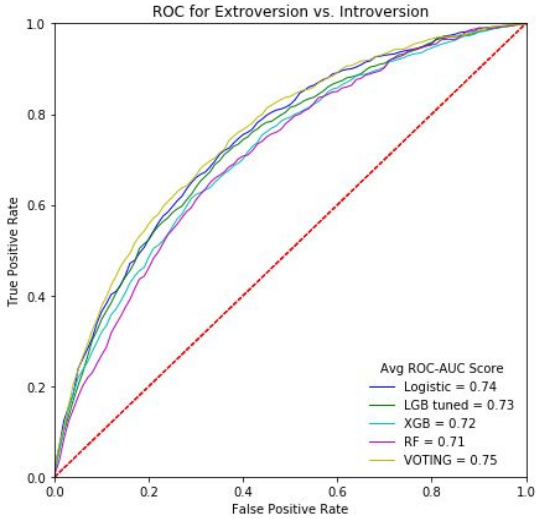
- Combining advantages
- Stable

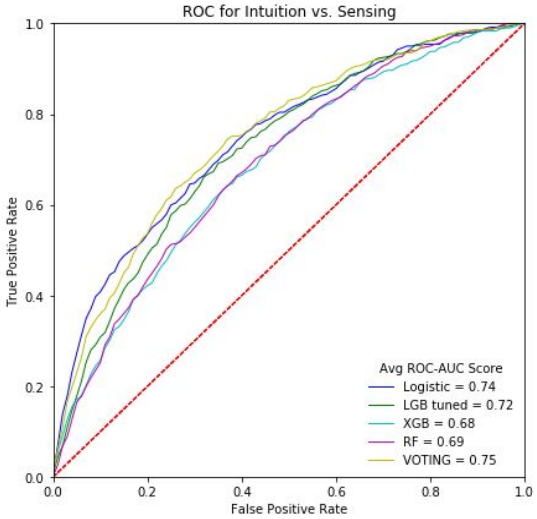


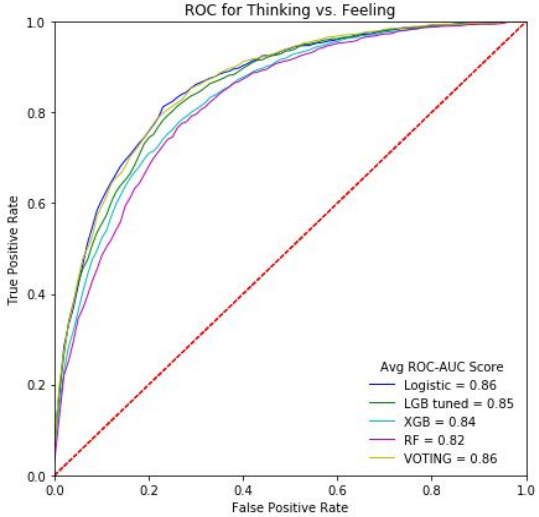
Model	E or I		N or S					Time(s)
	AUC	F1	AUC	F1				
MNB	0.60	0.40	0.56	0.26				
LR	0.74	0.50	0.74	0.42	0.86	0.78	0.71	0.77	40.34
RF	0.71	0.47	0.69	0.33	0.82			0.76	162.46
XGB	0.72	0.48	0.68	0.34	0.84			0.77	1339.4
LGBM	0.72	0.49	0.69	0.35	0.85			0.77	221.14
LGBM-T	0.73	0.50	0.72	0.38	0.85			0.77	90.56
Ensemble	0.75	0.51	0.75	0.41	0.86	0.78	0.72	0.77	1784.9

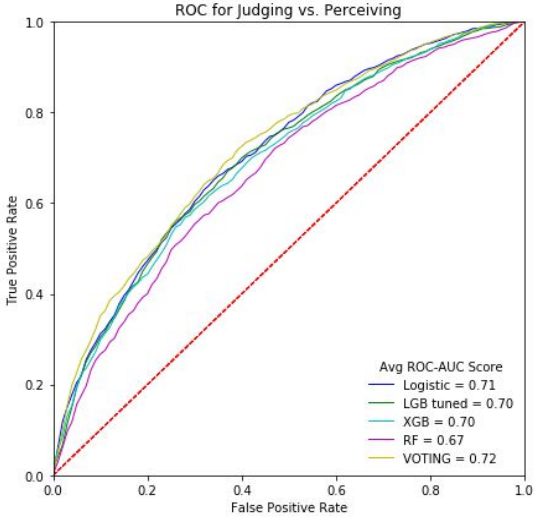
ROC-AUC

- x-axis False Positive Rate
- y-axis True Positive Rate (Recall)



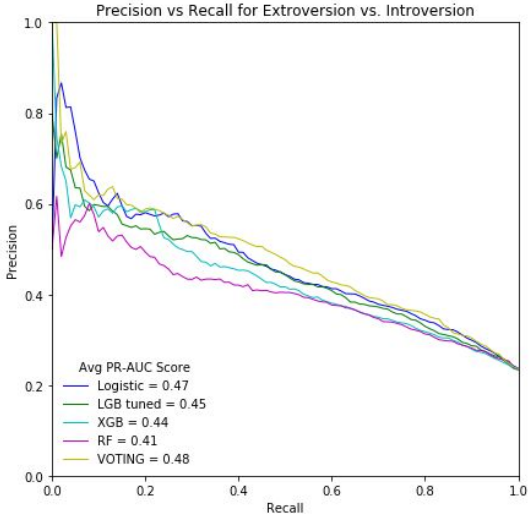


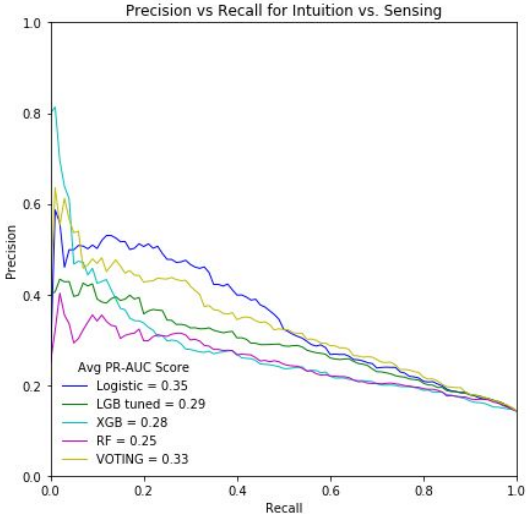


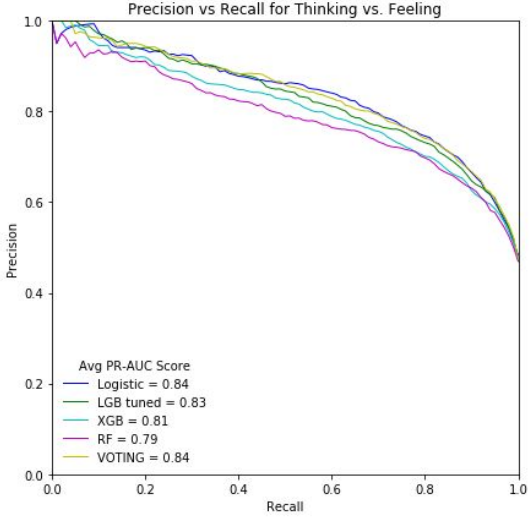


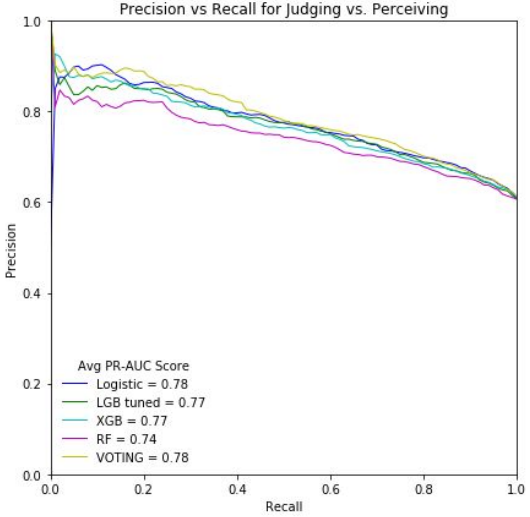
PR-AUC

- x-axis Recall
- y-axis Precision









Feature Importance for Extroversion or Introversion

Feature Importance for Extroversion or Introversion

Feature Importance for Extroversion or Introversion

Feature Importance for Extroversion or Introversion

Posts with exclamation marks:

> *“Did not realize all INFJs were so adorable and attractive!!!! :)”*

> *“Cause it's dope lol. Not so much a good lyrical song but it bumps like a motherfuggah!”*

Feature Importance for Extroversion or Introversion

Posts with exclamation marks:

Introvert

> *“Did not realize all INFJs were so adorable and attractive!!!! :)”*

Extrovert

> *“Cause it's dope lol. Not so much a good lyrical song but it bumps like a motherfuggah!”*

Predicting Celebrities

Arnold
@Schwarzenegger

Barack Obama
@BarackObama

kina grannis
@kinagrannis

Lady Gaga
@ladygaga

Donald J. Trump
@realDonaldTrump

Predicting Celebrities

Arnold
@Schwarzenegger

Barack Obama
@BarackObama

kina grannis
@kinagrannis

Lady Gaga
@ladygaga

Donald J. Trump
@realDonaldTrump

MBTI Type:	**INTJ**	**ENFJ**	**INFP**	**INFJ**	**ESTP**
Predicted:	**ISFJ**	**INFJ**	**INFP**	**INFP**	**INFJ**

Predicting Celebrities

Arnold
@Schwarzenegger

Barack Obama
@BarackObama

kina grannis
@kinagrannis

Lady Gaga
@ladygaga

Donald J. Trump
@realDonaldTrump

MBTI Type:	**INTJ**	**ENFJ**	**INFP**	**INFJ**	**ESTP**
Predicted:	**ISFJ**	**INFJ**	**INFP**	**INFP**	**INFJ**

And someone you know...

And someone you know...

MBTI Type: **INTJ**

Predicted: **ISTJ**

Takeaways

1. Resampling can be ineffective because it does not add any additional variability to the training set.

Takeaways

1. Resampling can be ineffective because it does not add any additional variability to the training set.
2. Evaluation metrics can be misleading.

Improvements

1. Gather more data for Extroverts and Sensing types to introduce more variability to the training set.

Improvements

1. Gather more data for Extroverts and Sensing types to introduce more variability to the training set.
2. Try more advanced algorithms like deep learning which are better known for NLP problems.

Thanks!

-The Introverts