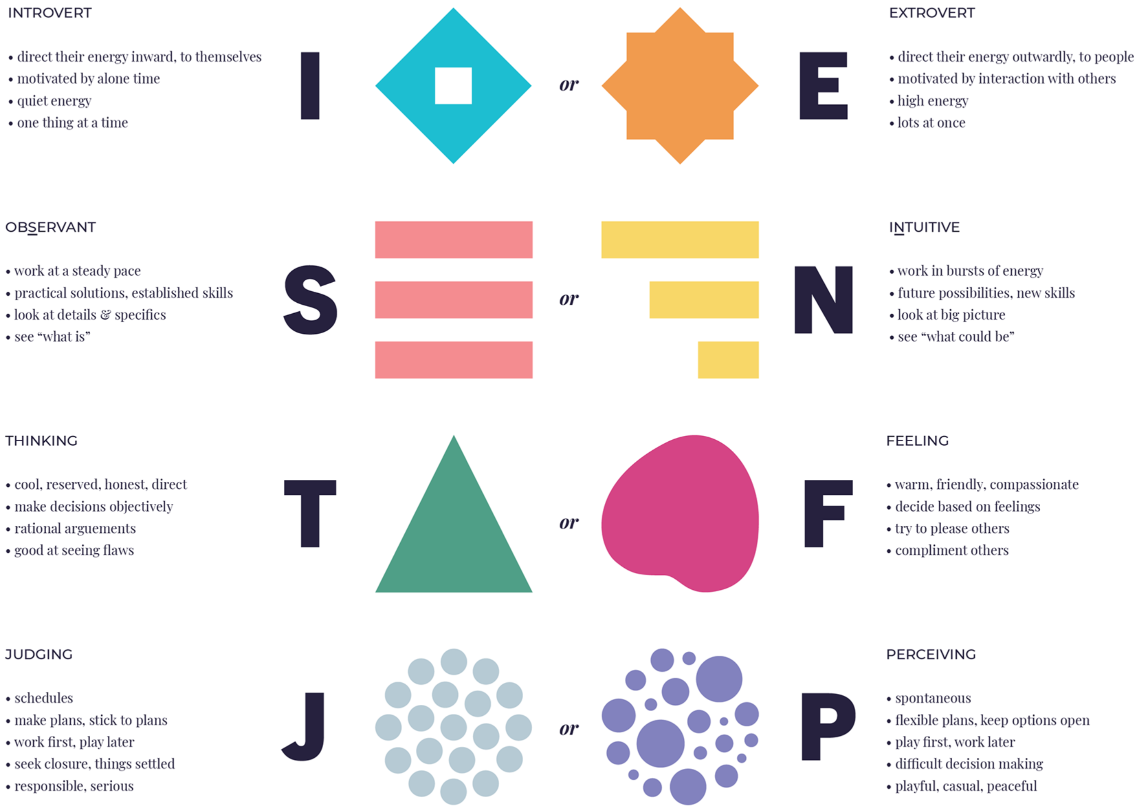
Predicting Myers-Briggs Personality Type

Ankita Budhraja

Introduction

The Myers-Briggs Type Indicator (MBTI) is a personality assessment tool used to determine personality types. The test categorizes people into 16 personality types based on four dichotomies: introversion/extroversion, intuition/sensing, thinking/feeling, and judging/perceiving.


INTROVERT
• direct their energy inward, to themselves
• motivated by alone time
• quiet energy
• one thing at a time
I
or
E
EXTROVERT
• direct their energy outwardly, to people
• motivated by interaction with others
• high energy
• lots at once
OBSERVANT
• work at a steady pace
• practical solutions, established skills
• look at details & specifics
• see "what is"
S
or
N
INTUITIVE
• work in bursts of energy
• future possibilities, new skills
• look at big picture
• see "what could be"
THINKING
• cool, reserved, honest, direct
• make decisions objectively
• rational arguements
• good at seeing flaws
T
or
F
FEELING
• warm, friendly, compassionate
• decide based on feelings
• try to please others
• compliment others
JUDGING
• schedules
• make plans, stick to plans
• work first, play later
• seek closure, things settled
• responsible, serious
J
or
P
PERCEIVING
• spontaneous
• flexible plans, keep options open
• play first, work later
• difficult decision making
• playful, casual, peaceful

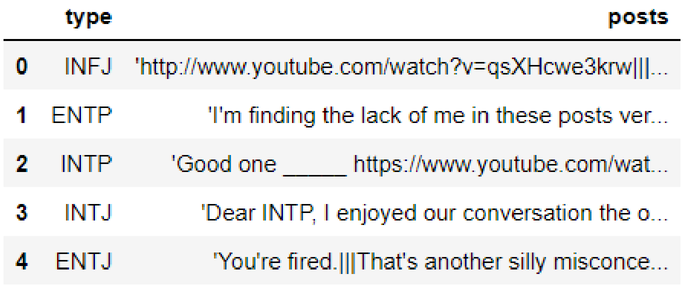
Goal

The overall goal for the outcome is to predict a person's personality type based on their social media posts and trying to address the following questions:
- Can Myers-Briggs personality type be predicted based on social media activity?

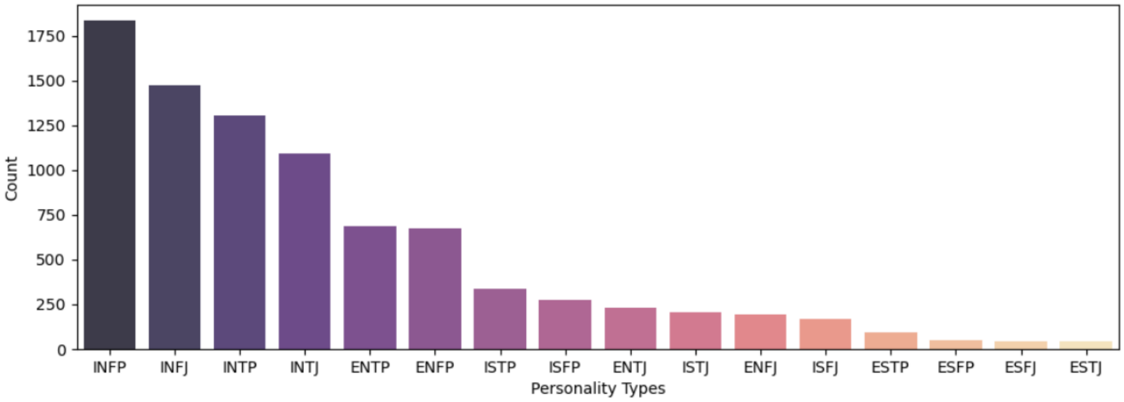
- Which Model performs the best and gives the most accurate results?
- What are the essential features of this data?
- Where can this model be used?

About the Dataset

Kaggle Dataset: This dataset contains over 8600 rows. Each row has a person's personality Type (MBTI code) and a list of the last 50 things they've posted.



	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	ENTP	'I'm finding the lack of me in these posts ver...
2	INTP	'Good one _____ https://www.youtube.com/wat...
3	INTJ	'Dear INTP, I enjoyed our conversation the o...
4	ENTJ	'You're fired.|||That's another silly misconce...

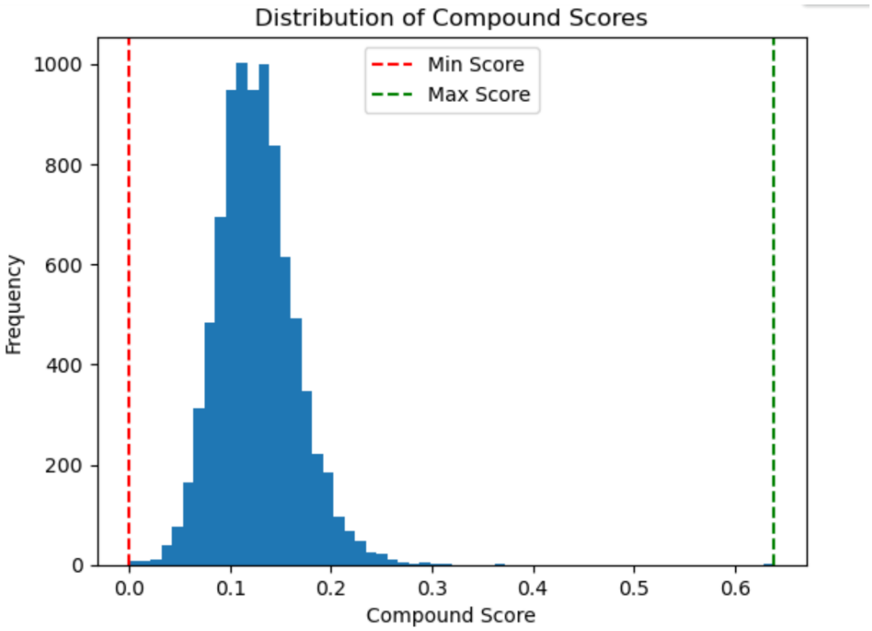


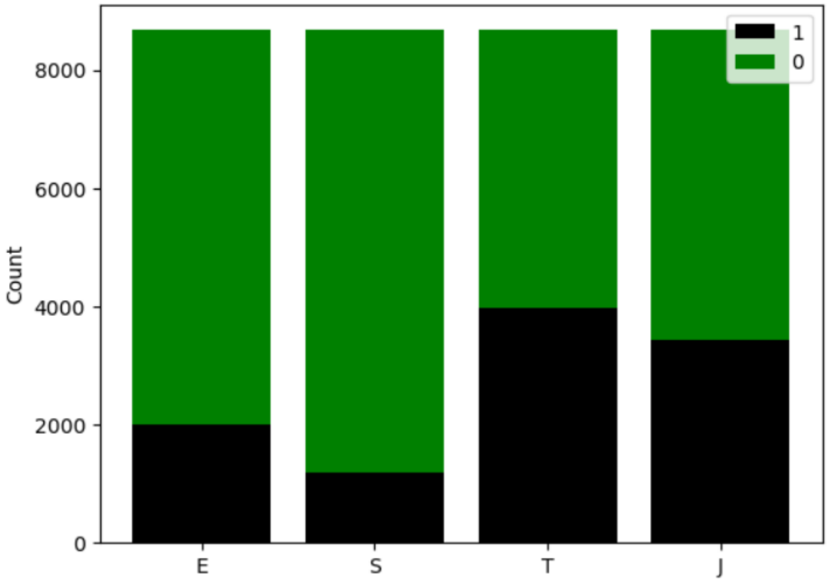
Data Preprocessing

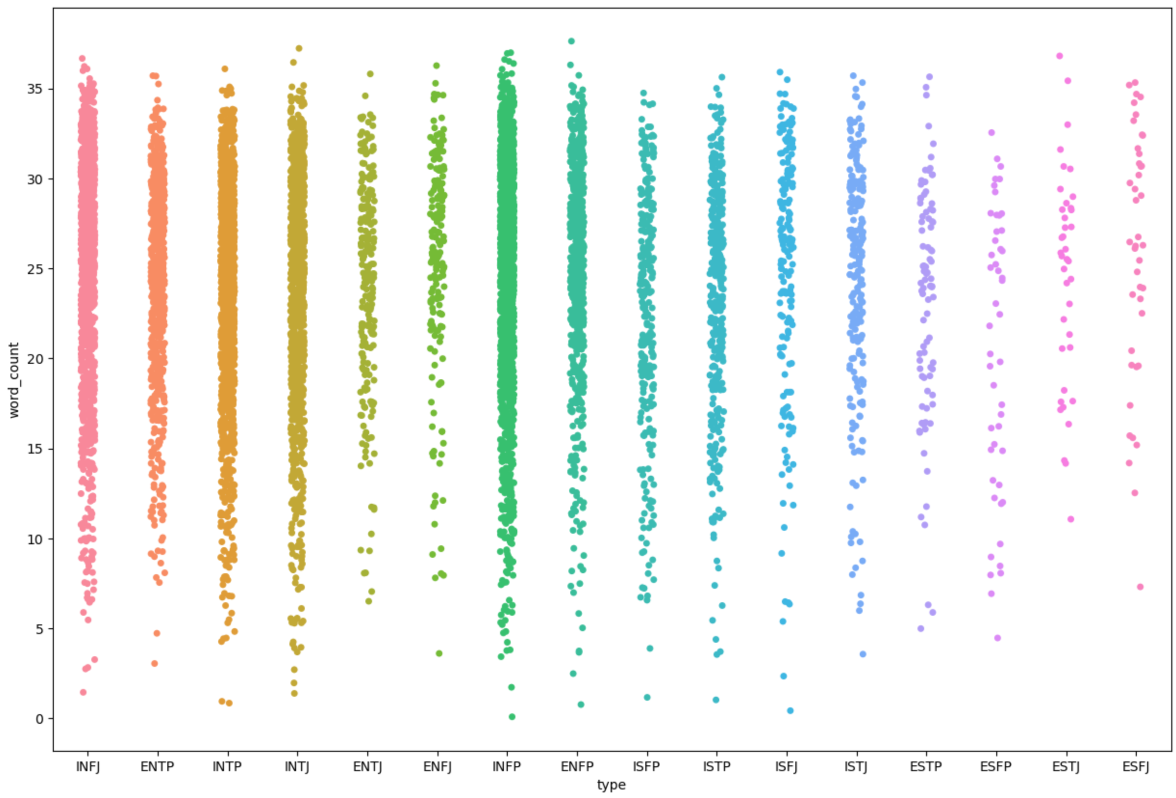
1. Remove URLs, numbers, extra white spaces, special characters, etc
2. Apply lemmatizer.lemmatize() to return the lemma form of each word to get the tokenized text

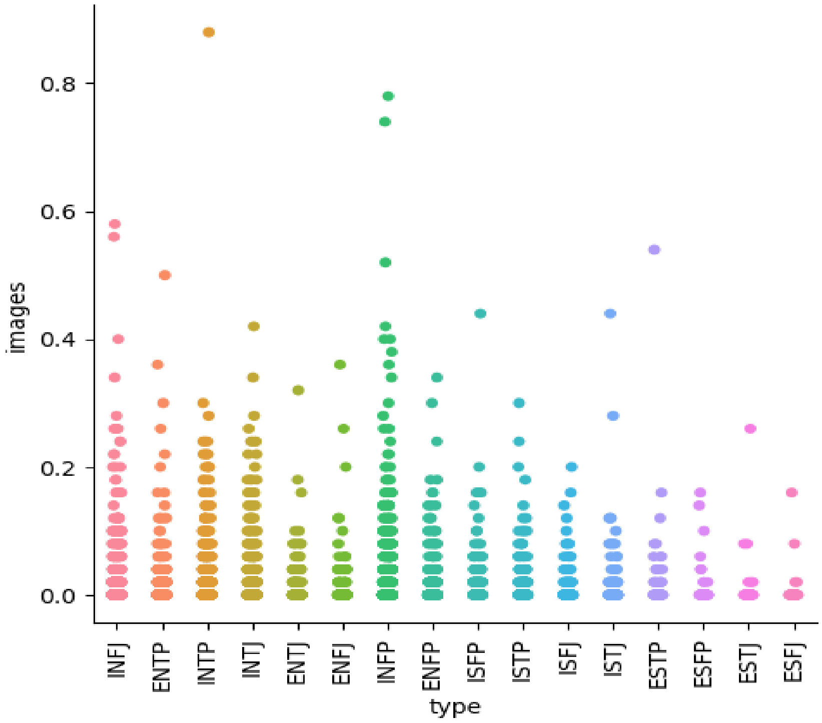
3. Remove default English stopwords
4. Performed sentiment analysis (using TextBlob) of the clean tokenized text (giving each a compound score for their 50 posts together)

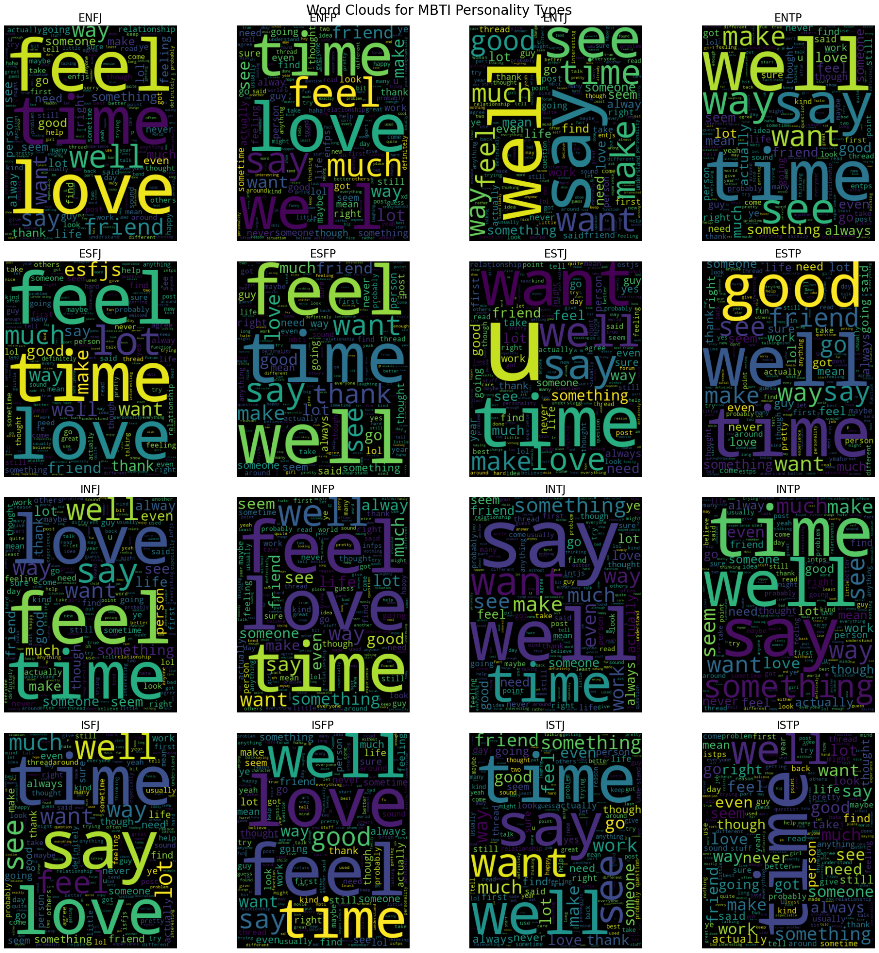
EDA






word_count
35
30
25
20
15
10
5
0
INFJ
ENTP
INTP
INTJ
ENTJ
ENFJ
INFP
ENFP
ISFP
ISTP
ISFJ
ISTJ
ESTP
ESFP
ESTJ
ESFJ
type

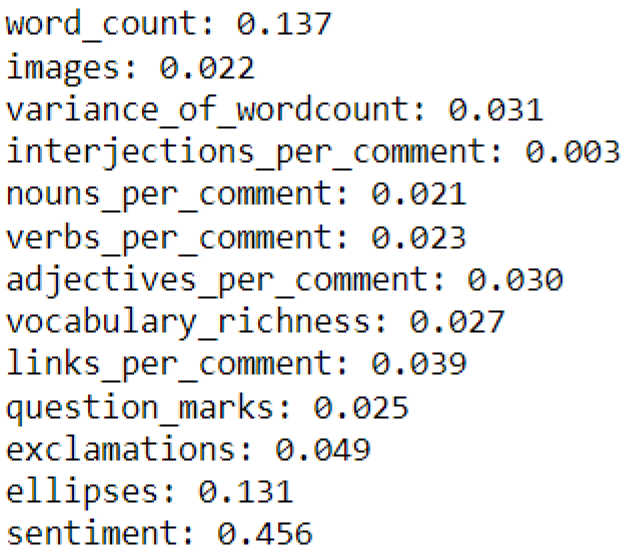

images
0.8
0.6
0.4
0.2
0.0
INFJ
ENTP
INTP
INTJ
ENTJ
ENFJ
INFP
ENFP
ISFP
ISTP
ISFJ
ISTJ
ESTP
ESFP
ESTJ
ESFJ
type



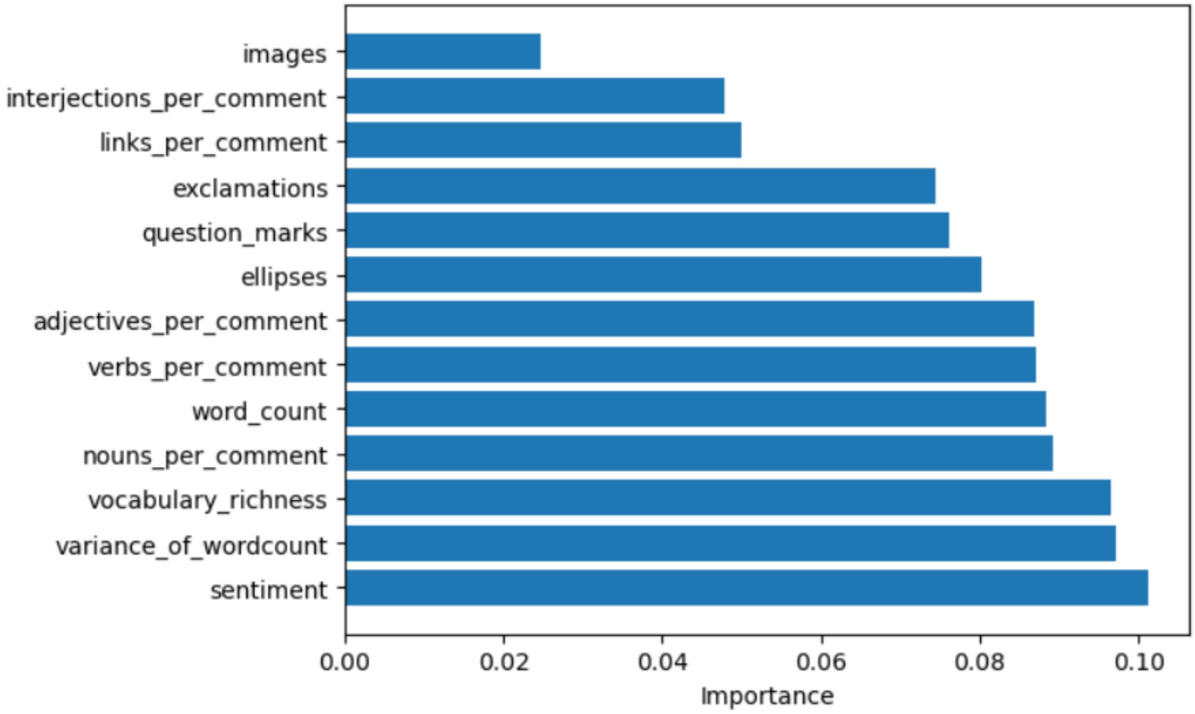
Feature Engineering

1. Word count
2. Variance of word count
3. Vocabulary richness (no. of unique words / total number of words)
4. Nouns, verbs, adjectives, interjections (using pos tag)
5. Average links and images per post
6. Average Question_marks, exclamations, ellipses per post
7. 4 new categories for each pair - E/I, S/N, T/F, J/P
8. Sentiment polarity score

Feature Importance using PCA and random forest



```
word_count: 0.137
images: 0.022
variance_of_wordcount: 0.031
interjections_per_comment: 0.003
nouns_per_comment: 0.021
verbs_per_comment: 0.023
adjectives_per_comment: 0.030
vocabulary_richness: 0.027
links_per_comment: 0.039
question_marks: 0.025
exclamations: 0.049
ellipses: 0.131
sentiment: 0.456
```



Models

Created a pipeline using imblearn package:

- Created a TfidfVectorizer for the tokenized text column.
- Used SelectKBest using the f_classif scoring function and the MinMaxScaler
- Used Under sampling because of class imbalance

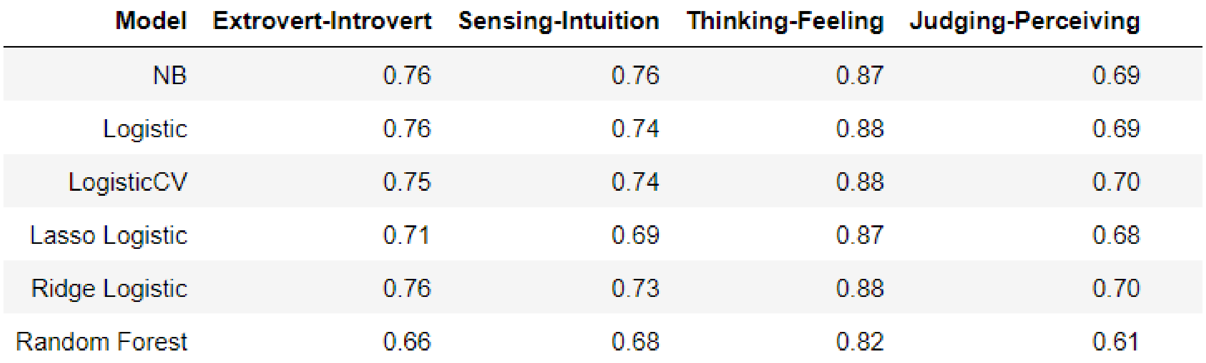
- Added more stop words as part of preprocessing (as seen in WordCloud)

Types of Models

- Naive Bayes
- Logistic Regression
    - Lasso
    - Ridge
    - LogisticCV
- Random Forest

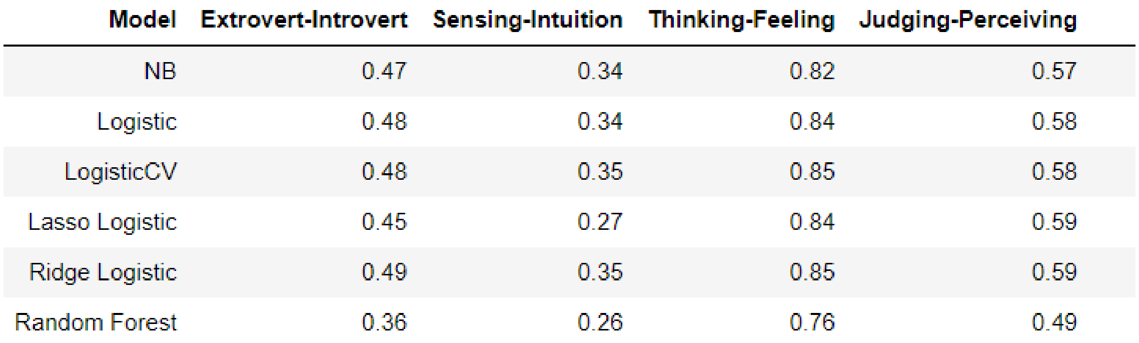
Evaluation Metrics used to compare models

1. ROC-AUC Score



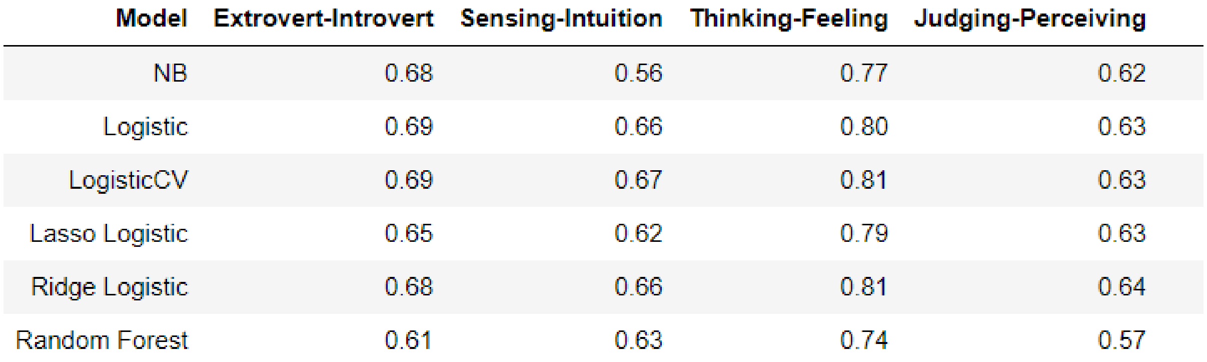
Model	Extrovert-Introvert	Sensing-Intuition	Thinking-Feeling	Judging-Perceiving
NB	0.76	0.76	0.87	0.69
Logistic	0.76	0.74	0.88	0.69
LogisticCV	0.75	0.74	0.88	0.70
Lasso Logistic	0.71	0.69	0.87	0.68
Ridge Logistic	0.76	0.73	0.88	0.70
Random Forest	0.66	0.68	0.82	0.61

2. Precision-Recall Score



Model	Extrovert-Introvert	Sensing-Intuition	Thinking-Feeling	Judging-Perceiving
NB	0.47	0.34	0.82	0.57
Logistic	0.48	0.34	0.84	0.58
LogisticCV	0.48	0.35	0.85	0.58
Lasso Logistic	0.45	0.27	0.84	0.59
Ridge Logistic	0.49	0.35	0.85	0.59
Random Forest	0.36	0.26	0.76	0.49

3. Geometric Mean Score



Model	Extrovert-Introvert	Sensing-Intuition	Thinking-Feeling	Judging-Perceiving
NB	0.68	0.56	0.77	0.62
Logistic	0.69	0.66	0.80	0.63
LogisticCV	0.69	0.67	0.81	0.63
Lasso Logistic	0.65	0.62	0.79	0.63
Ridge Logistic	0.68	0.66	0.81	0.64
Random Forest	0.61	0.63	0.74	0.57

Conclusion

- Heavily Imbalanced Data
- Didn’t work well for Extrovert-Introvert and Sensing-Intuition as most of them identified as Introvert and Intuitive.
- Regularization didn’t improve scores by a lot, but LogisticCV Regression worked best for this dataset.
- In the future:
    - also try Neural Networks which could improve the scores and can skip feature engineering by a lot.
    - Could be used in improving marketing campaigns, understanding social media behavior, and styles of each type