Pattern Recognition and Machine Learning Bonus Project Report

Personality Detection

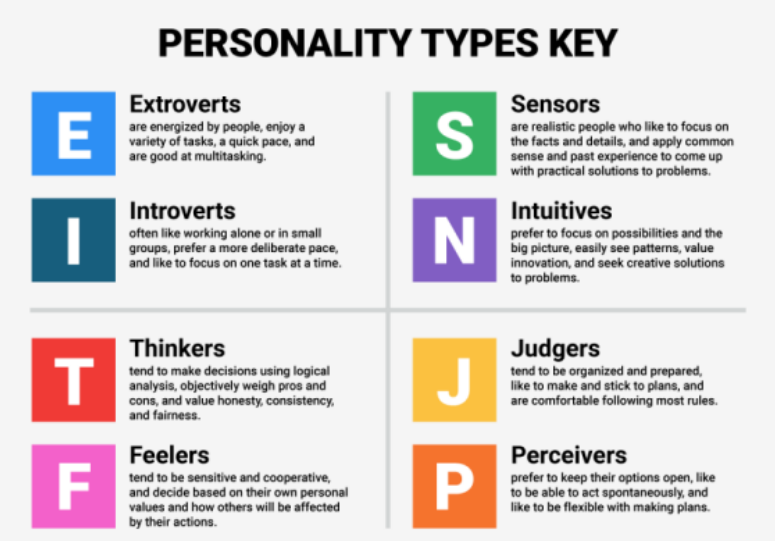
Siddharth Singh (B20EE067)

Abstract

The Myers-Briggs Type Indicator (MBTI)—also referred to as the "Myers-Briggs personality test" or simply the "Myers-Briggs test"—is a self-reported questionnaire. The test helps people assess their personality using four specific dichotomies, or scales: introversion-extraversion, sensing-intuition, thinking-feeling and judging-perceiving. The MBTI was first developed in the 1940s by Isabel Briggs Myers and her mother Katharine Cook Briggs, and it's based on psychologist Carl Jung's theory of psychological types. The purpose of the test is to "make the theory of psychological types described by C.G. Jung understandable and useful in people's lives," according to the Myers Briggs Foundation.

1 Introduction

In this project, I have developed an MBTI personality classifier that uses machine learning models to predict a person's personality based on the 50 recent social media posts per user as input. I find correlations between a person's MBTI personality type and writing style. The classifier also demonstrates the validity of the MBTI test.For the project we had used dataset provided in Link for dataset.



2 Data Description, Visualization ,Preprocessing and Feature Extraction

2.1 Data Description

The dataset has 8675 rows and 2 columns, namely- type and posts. The data in column 'post' contains 50 recent social media posts for each user. There are 16 unique labels in column 'type' with no NULL values, each representing 16 MBTI type indicators.

2.2 Data Preprocessing

For better feature extraction, some preprocessing is performed on the textual data in column 'posts':

- To lower case
- Removal of URL/links
- Removal of special characters and numbers
- Removal of extra space
- Removal of stopwords
- Removal of MBTI personality names
- Lemmatization

2.3 Data Visualization

Our dataset is unbalanced for a few personality types, implying some words appear more often and carry little meaning and information about the data. The MBTI classifier has four main dimensions, namely 'Introversion-Extraversion' (IE), 'Intuition-Sensing' (NS), 'Thinking-Feeling' (TF), 'Judging-Perceiving' (JP). Four more columns are added to the dataset. In each column, '1' represents the first part of each dimension (I, N, T, J), whereas '0' represents the second part of each dimension (E, S, F, P), respectively.

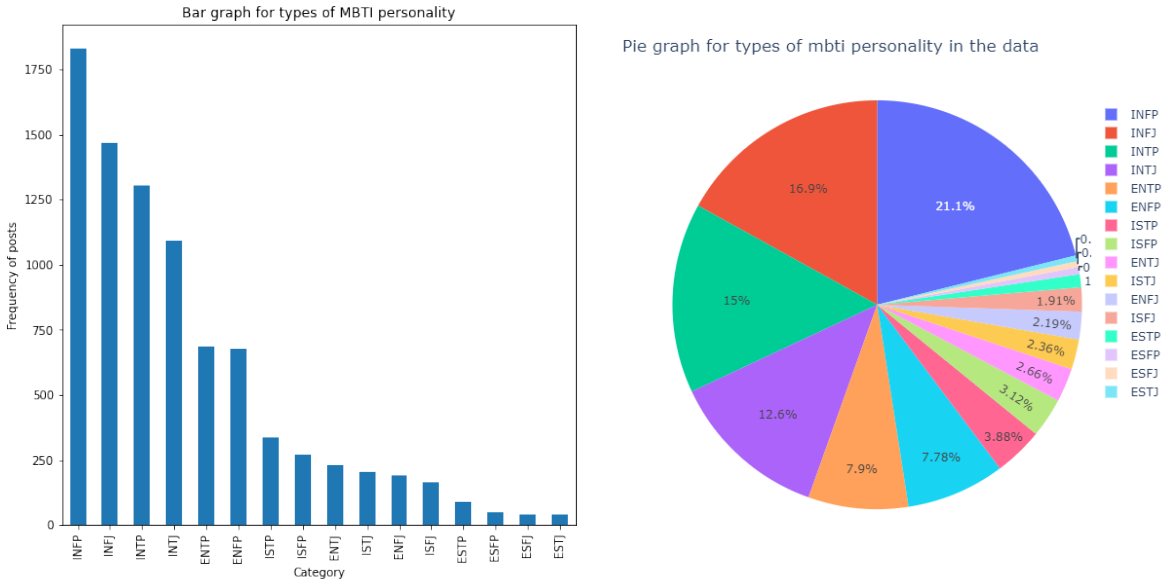


Figure 1: *Dataset distribution for different classes available*

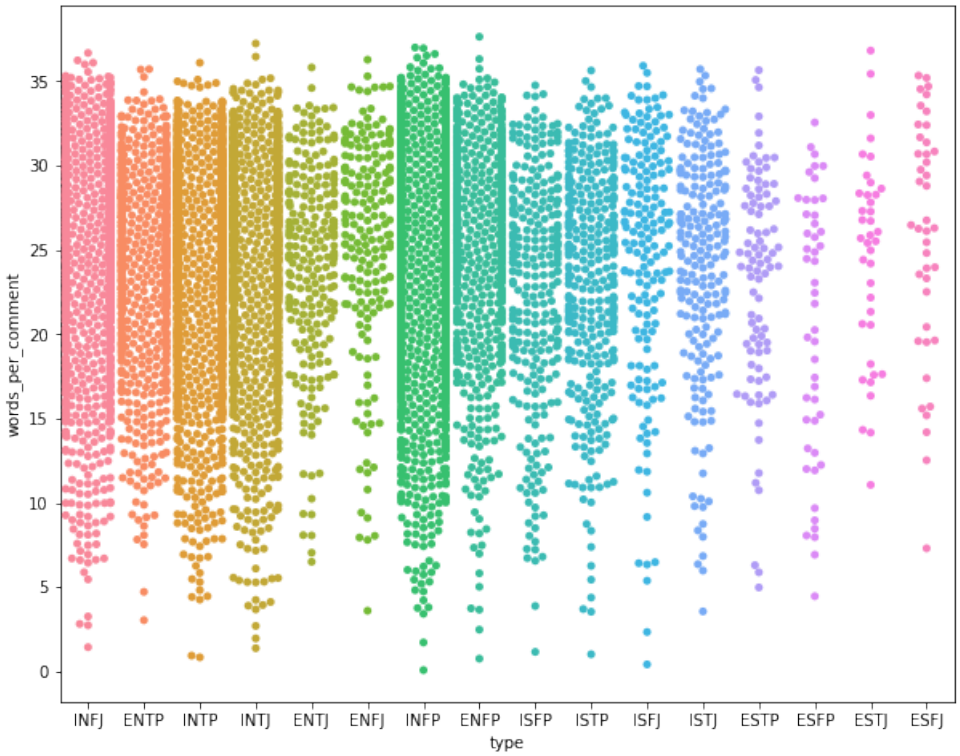


Figure 2: *Swarm plot of words per comment vs type*

2.4 Feature Extraction

TF-IDF and count vectorizer is used to convert text into features, providing a more focused text view. First, vectorize the data using countVectorizer and convert the post into the matrix of token counts for the model. Then TF-IDF normalization is used to scale the feature from the count vectorizer into floating-point values. TF-IDF analyzes how much a word is relevant to a corpus in a corpus collection and provides the importance of words in data. After vectorizing, the dataset had 1500 features for each user post.

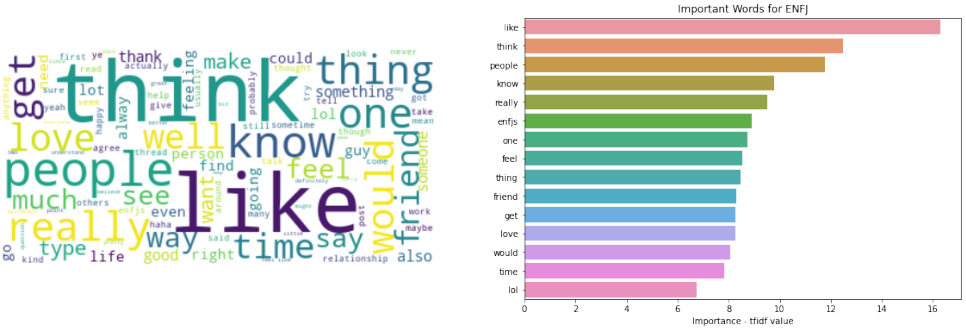


Figure 3: *Words cloud for ENFJ*

3 Models

3.1 Logistic Regression

It is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.

Logistic Regression without LDA

Class	Accuracy	Precision	Recall	F1 Score
Extroversion(E)/ Introversion(I)	0.81	0.815	0.975	0.887
Sensing(S)/ Intuition(I)/	0.867	0.86	0.997	0.928
Feeling(F)/ Thinking(T)	0.805	0.8	0.768	0.783
Perceiving(P)/ Judging(J)	0.722	0.710	0.5036	0.589

Logistic Regression with LDA

Class	Accuracy	Precision	Recall	F1 Score
Extroversion(E)/ Introversion(I)	0.86	0.889	0.947	0.917
Sensing(S)/ Intuition(I)/	0.915	0.9276	0.977	0.9514
Feeling(F)/ Thinking(T)	0.862	0.856	0.450	0.589
Perceiving(P)/ Judging(J)	0.795	0.763	0.6972	0.7290

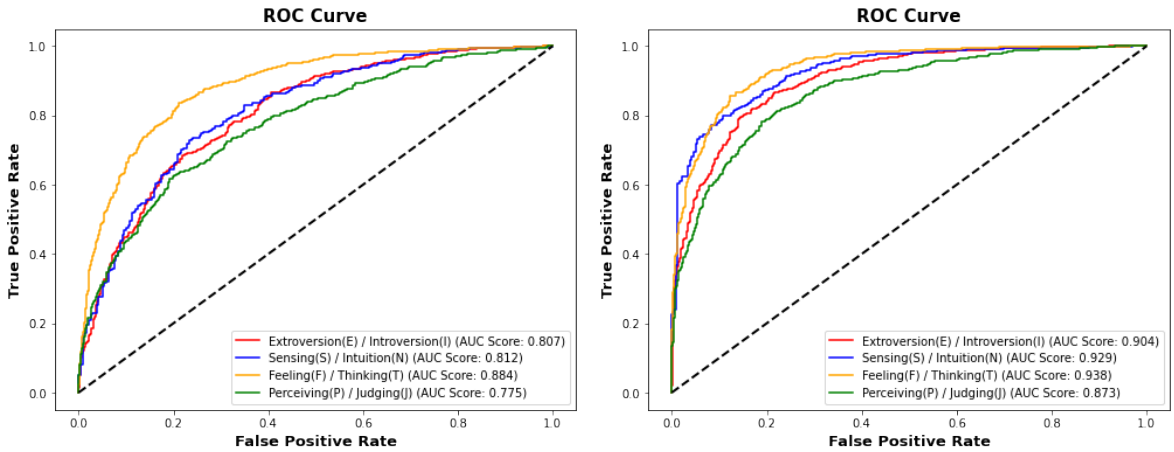


Figure 4: *ROC AUC curve for Logistics and Logistics + LDA*

3.2 Support Vector Machine

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression.The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

Class	Accuracy	Precision	Recall	F1 Score
Extroversion(E)/ Introversion(I)	0.802	0.804	0.981	0.88
Sensing(S)/ Intuition(I)/	0.866	0.866	0.997	0.928
Feeling(F)/ Thinking(T)	0.801	0.790	0.770	0.779
Perceiving(P)/ Judging(J)	0.721	0.720	0.483	0.579

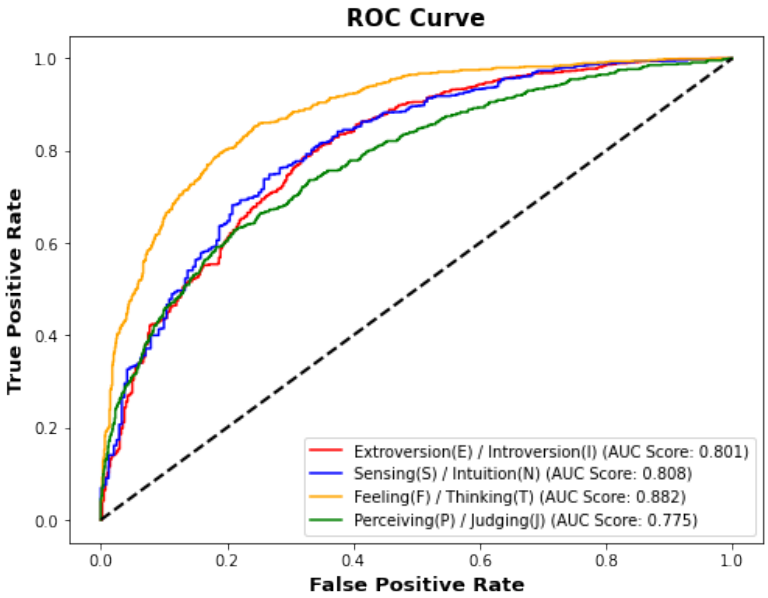


Figure 5: *ROC AUC curve for SVM*

3.3 XGBoost

XGboost stands for eXtreme Gradient Boosting.It is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. I have taken max depth = 5.

Class	Accuracy	Precision	Recall	F1 Score
Extroversion(E)/ Introversion(I)	0.814	0.819	0.972	0.889
Sensing(S)/ Intuition(I)/	0.877	0.879	0.993	0.933
Feeling(F)/ Thinking(T)	0.783	0.768	0.755	0.761
Perceiving(P)/ Judging(J)	0.725	0.731	0.483	0.581

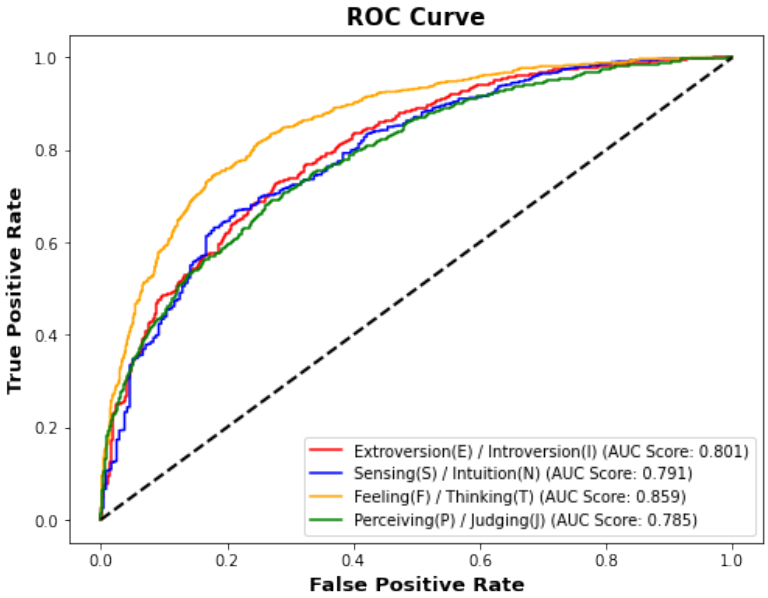


Figure 6: *ROC AUC curve for XGBoost*

3.4 Random Forest Classifier

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.I have found the best max depth = 100 and n estimators = 100 using GridSearchCv.

Random Forest Classifier without LDA

Class	Accuracy	Precision	Recall	F1 Score
Extroversion(E)/ Introversion(I)	0.772	0.771	1.00	0.870
Sensing(S)/ Intuition(I)/	0.862	0.862	1.00	0.9259
Feeling(F)/ Thinking(T)	0.745	0.75	0.667	0.706
Perceiving(P)/ Judging(J)	0.725	0.731	0.483	0.581

Random Forest Classifier with LDA

Class	Accuracy	Precision	Recall	F1 Score
Extroversion(E)/ Introversion(I)	0.817	0.875	0.889	0.881
Sensing(S)/ Intuition(I)/	0.870	0.922	0.927	0.925
Feeling(F)/ Thinking(T)	0.802	0.852	0.446	0.586
Perceiving(P)/ Judging(J)	0.736	0.667	0.665	0.666

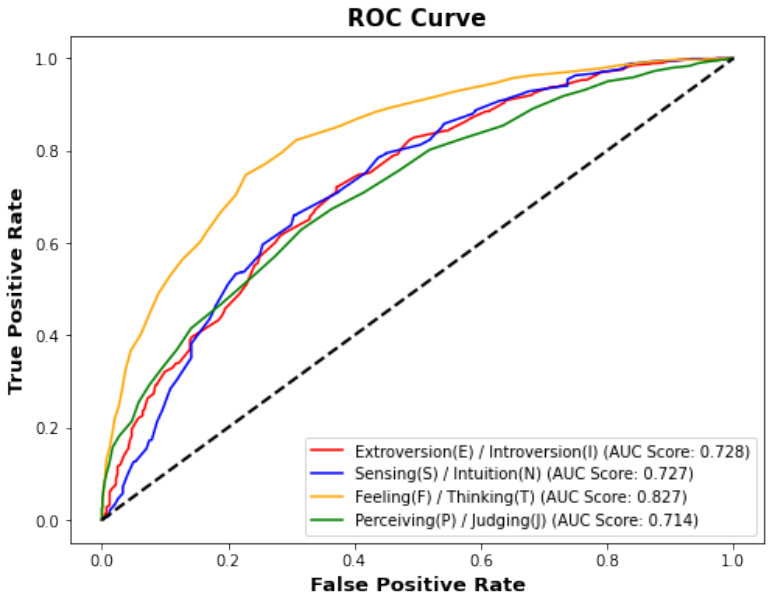


Figure 7: *ROC AUC curve for Random Forest*

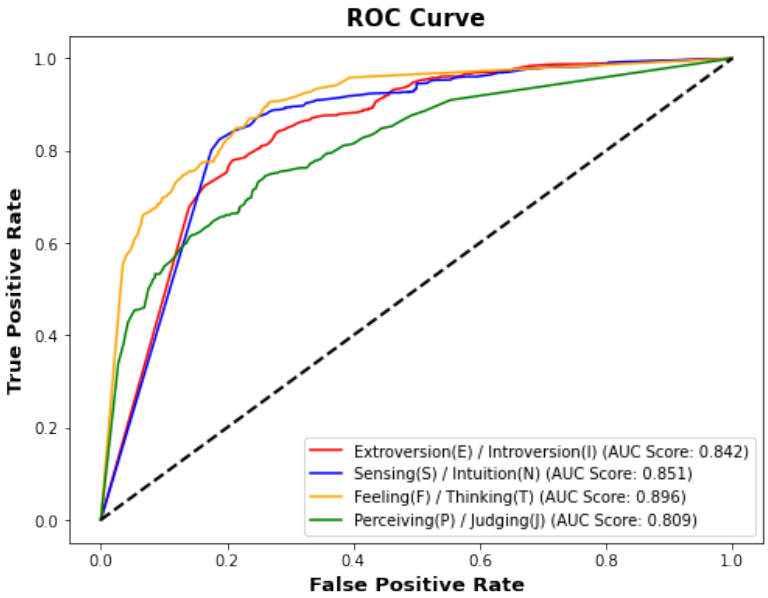


Figure 8: *ROC AUC curve for Random Forest + LDA*

Accuracy Table

Models	I/E	N/S	F/T	J/P
Logistics Regression	0.817	0.867	0.805	0.722
Logistics Regression + LDA	0.86	0.915	0.862	0.795
Support Vector Machine	0.802	0.866	0.801	0.721
XGBoost	0.814	0.877	0.783	0.725
Random Forest	0.772	0.862	0.745	0.725
Random Forest + LDA	0.817	0.870	0.802	0.736

4 Conclusion

The above table shows that all the classifiers performed nearly equal.Our model accurately predicted MBTI personality based on social media posts using all six supervised machine learning algorithms.Dimensionality reduction techniques like LDA helped in speeding up random forest classification and Logistics Regression classification. It also improved accuracy.We conclude that the Logistic Regression + LDA model performs the best for personality classification based on The Myers Briggs Personality Model. The ROC curve for logistic regression + LDA also supports the fact that it results in better performance.We can get more precise results by training models on a larger and more accurate dataset.This system can assist in the development of better recommendation systems.

5 References

- Understanding Support Vector Machine(SVM) algorithm | by Sunil | Analytics Vidya
- NLP Text Preprocessing: A Practical Guide and Template | by Jiahao Weng | Towards Data Science
- Boosting in Machine Learning and the Implementation of XGBoost in Python | Evan Lutins | Towards Data Science
- Hyperparameter Tuning the Random Forest in Python| by Will Koehrsen
- Pattern Classification -Book by David G. Stork, Peter E. Hart, and Richard O. Duda