Pattern Recognition & Machine Learning BONUS PROJECT

Instructor: Dr. Richa Singh

Name : Ajay Nirmal Roll no. : B20AI002

1 Introduction

The aim of the project is to predict the personality of a specific person. The project incudes building and training of a model capable of predicting a person's MBTI label using only what people post in online forums.

This project uses Natural Language processing (NLP) as the processed data is used to classify and assign MBTI labels to a person's online forum posts.

2 Data Analysis

2.1 Dataset Used

The dataset contains 8675 rows of data. On each row, there is:

- person's type in the form of 4 letter code
- The posts made by the person

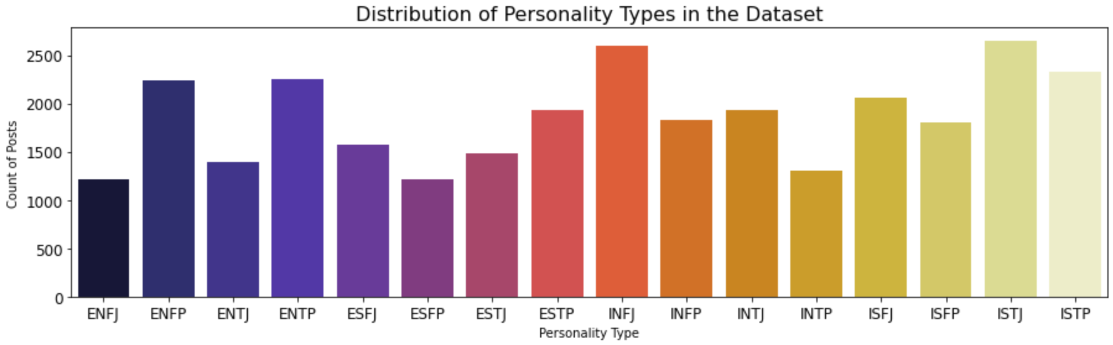
The four-letter code is the combination of 8 variables (four binary variables).

2.2 Exploring personality types

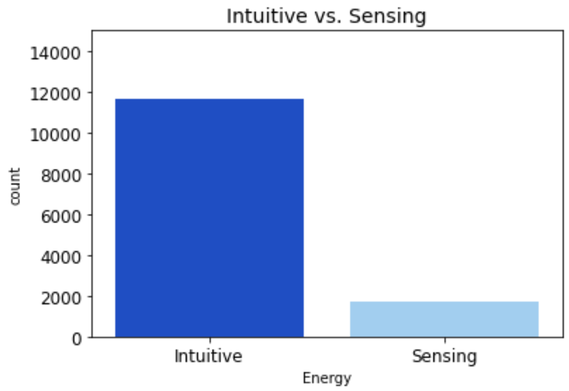
Each personality type consists of four binary variables i.e the four labels, they are:

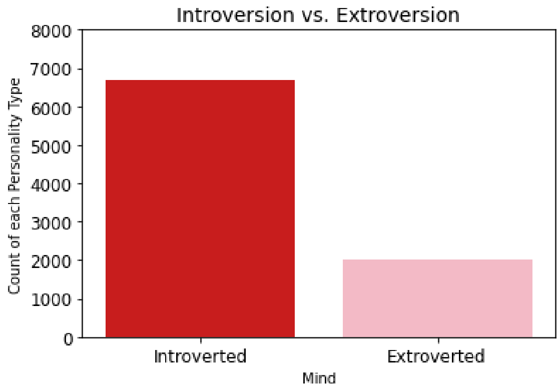
- Mind: Introverted (I) or Extraverted (E)
- Energy: Sensing (S) or Intuitive (N)
- Nature: Feeling (F) or Thinking (T)
- Tactics: Perceiving (P) or Judging (J)

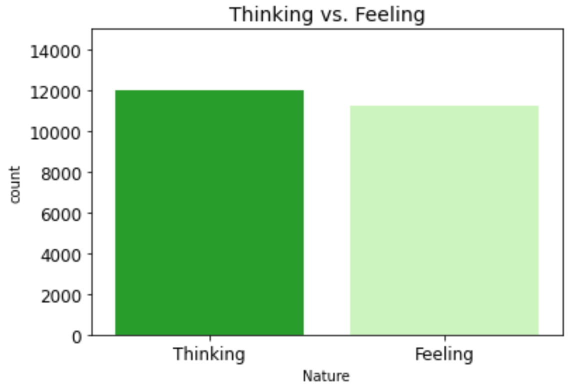
The combination of these eight classes gives the final personality type. The distribution of the personality types is shown in the figure below:

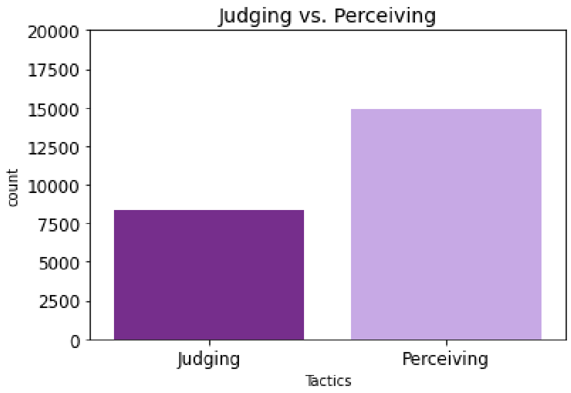


2.3 Plots for class-wise distribution of personality features



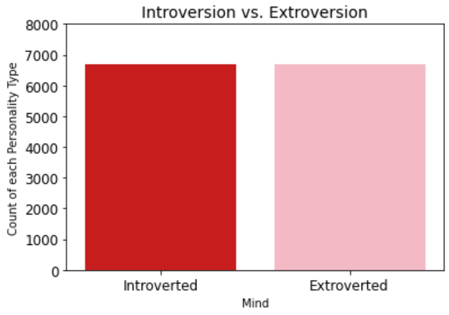


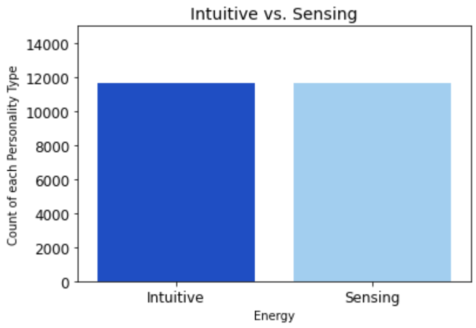


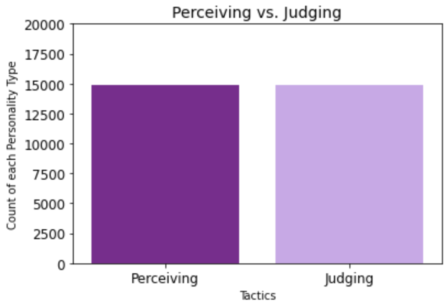


2.4 Handling data imbalance

The data in classes like Mind and Energy is highly imbalanced. That will affect the accuracy and evaluation metrics of the model. Thus, that data was re-sampled to remove the data imbalance. The graphs for the same can be seen below:







3 Approach

3.1 Goals

The aim is to classify the personality type of a person on the MBTI labels based upon the posts available to us.

3.2 Tasks

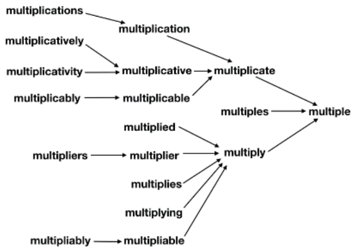
- Removing data imbalance
- Splitting the dataset into test and train set
- Pre-processing the dataset for further analysis
- Trying various models and find the best one for personality prediction
- Training the model class wise
- Prediction made on the test data for each of the four characteristics (classes)

4 Experiments

4.1 Data pre-processing

Pre-processing or cleaning of data leads to a more efficient model and higher accuracy. We have created a pipeline for the same including various functions which are listed below:

- Removal of html tags and URLs : www.harvard.edu:80
- Removal of delimiters
- Lowercase all sentences
- Removal of punctuations and numbers: ‘!()*+,-./:;?@[]‘— ’
- Lemmatization:



- Remove Stop-Words: “a”, “the”, “is”, “are”

4.2 Finding the best model

I tried to build four model on this dataset. The comparison of the same on all four labels have been done. Results of one of them i.e. Mind model can be seen in the table below:

Models	Accuracy score	Precision	Recall	F1-score
Logistic Regression	0.84	0.87	0.84	0.86
MultinomialNB	0.76	0.91	0.73	0.81
AdaBoost Classifier	0.82	0.76	0.83	0.85
LGBM Classifier	0.98	0.98	0.99	0.99

As observed from the table above, Logistic Regression model performs the best. The scores of the LGBM classifier look fairly good. But on testing it on the test data, it was the case of overfitting.

5 Overall Results

5.1 Multi label classification

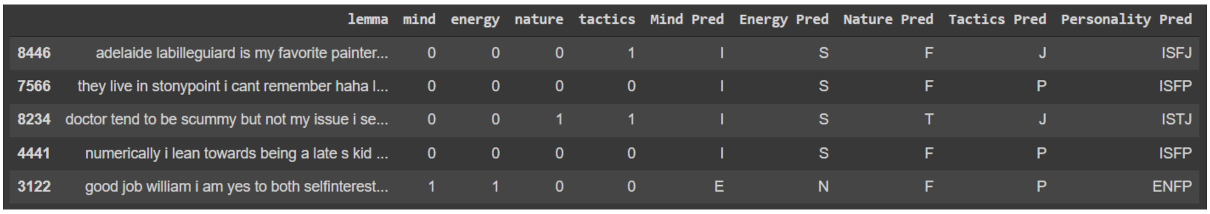
In our binary classification job, Logistic Regression outperformed all other models, attaining the greatest accuracy, precision, and F1 scores. The post was classified majorly in four labels: Mind, Energy, Nature and Tactics.

5.2 Multi-label binary class classification

The posts are finally classified into 8 classes in groups of 2 each. Using the Logistic Regression model, prediction results on the test data have been shown ahead label wise:

Labels	Accuracy score	Precision	Recall	F1-score	Log loss
Mind					
• Introvert(I)	0.83	0.84	0.87	0.85	5.72
• Extrovert(E)	0.83	0.83	0.79	0.81	5.72
Energy					
• Sensing(S)	0.84	0.83	0.84	0.84	5.65
• Intuitive(N)	0.84	0.84	0.83	0.84	5.65
Nature					
• Feeling(F)	0.95	0.95	0.93	0.92	1.77
• Thinking(T)	0.95	0.95	0.94	0.91	1.77
Tactics					
• Perceiving(P)	0.81	0.82	0.81	0.82	6.42
• Judging(J)	0.81	0.81	0.81	0.81	6.42

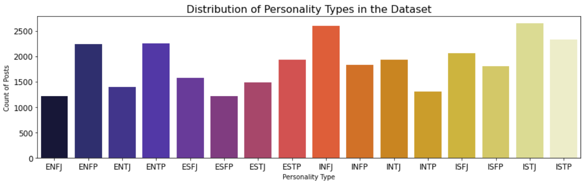
Further the final predicted data has been converted to a dataframe. The predicted personalities of the posts made by the 5 people can be seen. Screenshot of which is attached below:

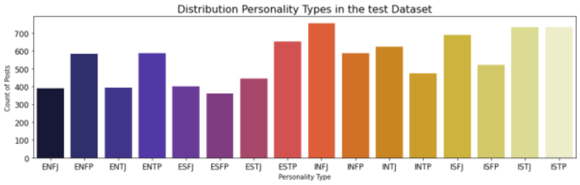


	lemma	mind	energy	nature	tactics	Mind Pred	Energy Pred	Nature Pred	Tactics Pred	Personality Pred
8446	adelaide labilleguiard is my favorite painter...	0	0	0	1	I	S	F	J	ISFJ
7566	they live in stonypoint i cant remember haha l...	0	0	0	0	I	S	F	P	ISFP
8234	doctor tend to be scummy but not my issue i se...	0	0	1	1	I	S	T	J	ISTJ
4441	numerically i lean towards being a late s kid ...	0	0	0	0	I	S	F	P	ISFP
3122	good job william i am yes to both selfinterest...	1	1	0	0	E	N	F	P	ENFP

5 Conclusion

The Logistic Regression worked the best and gave appropriate results during Multi label binary class classification. It outperformed all the model giving good accuracy. At the end I have plotted the distribution of final predicted data (on test data) and the initial dataset. It can be seen from the plot that predictions done are right up to a significant level.


Distribution of Personality Types in the Dataset
Count of Posts
ENFJ
ENFP
ENTJ
ENTP
ESFJ
ESFP
ESTJ
ESTP
INFJ
INFP
INTJ
INTP
ISFJ
ISFP
ISTJ
ISTP
Personality Type


Distribution Personality Types in the test Dataset
Count of Posts
ENFJ
ENFP
ENTJ
ENTP
ESFJ
ESFP
ESTJ
ESTP
INFJ
INFP
INTJ
INTP
ISFJ
ISFP
ISTJ
ISTP
Personality Type