**1. PROBLEM DEFINITION**

Multi-class classification on Predicting MBTI Personality Types

1. **Client:**
2. **Relationship:**

Type differences in relationships can be a source of growth and/or conflict. However, there are no best or more successful combinations of types in relationships. Two persons who share all four preferences are as likely to get along easily as are a couple who share only one or two preferences. Understanding and applying type theory to relationships can enhance communication, provide people with a better understanding of how they deal with conflict, and provide tools for a variety of situations including successfully making decisions and engaging in activities together.

1. **Careers:**

Research has clearly shown that people are attracted to careers that allow them to make use of their natural type preferences. Though all four letters of your type can affect the kind of career that interests you, the two middle letters (ST, SF, NF, or NT) of your type have a particular importance for your career choice.

1. **Education:**

Type can tell us many things about the way people prefer to learn. For example, people with a preference for Extraversion often prefer learning situations that allow them to talk with others and to become physically engaged with the environment. Those with a preference for Introversion often prefer learning environments that allow them to engage in quiet reflection, where they can process their thoughts internally until they are more developed. An understanding of type leads to the appreciation that there are many different and equally valuable ways to learn. Type can also help you identify some of your strengths and challenges as you approach studying and learning.

1. **Workplace and Organizations:**

Type is widely used in the workplace and in organizational settings. As we have already noted, different types are clearly drawn to different careers. However, within any specific job or work setting many different types are represented. This diversity in types is healthy and stimulating, but it can also lead to misunderstandings and frictions in the workplace. Understanding and applying type to the workplace can result in increased communication, more effective teams, and more satisfied employees and customers.

1. **Data Set:**

The data has been obtained from the "Kaggle" (<https://www.kaggle.com/datasnaek/mbti-type>)

Data includes 8675 different users’ posts.

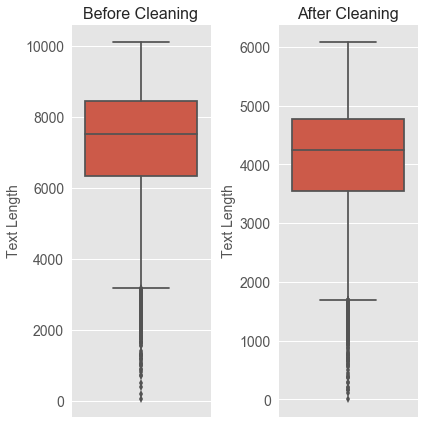
Our target column is type column in the dataset.

**2. DATA WRANGLING**

In the context of corpus normalization:

* Lowercase
* Removing links
* Removing punctuations
* Removing stop words
* Using Stemmer
* Expanding contractions
* Removing numbers

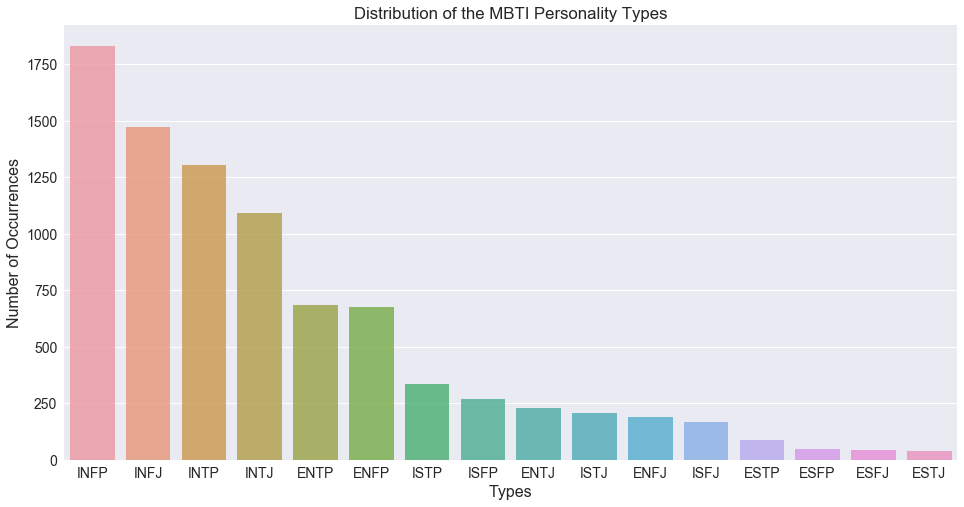
As a result, the post lengths were reduced around 40%.

****

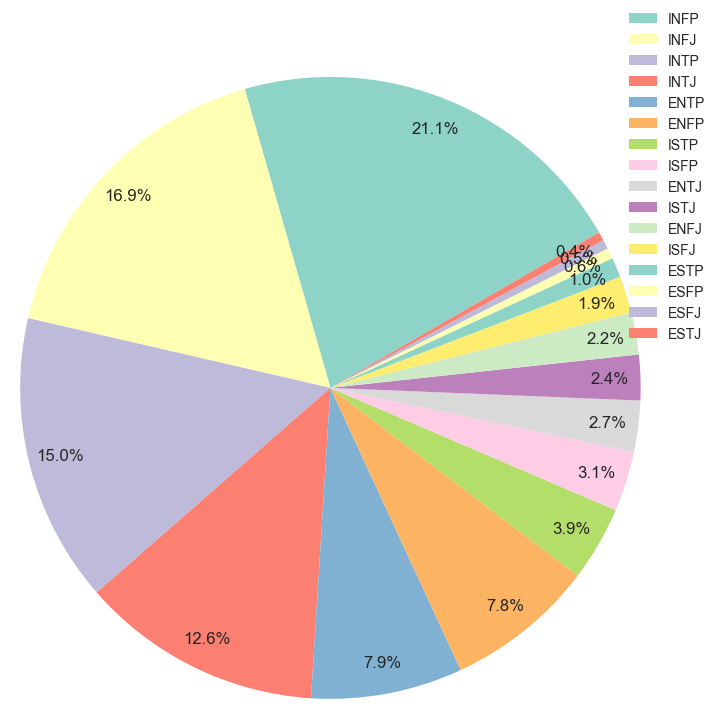
### 

**3. EXPLORATORY DATA ANALYSIS (EDA)-DATA VISUALIZATION**

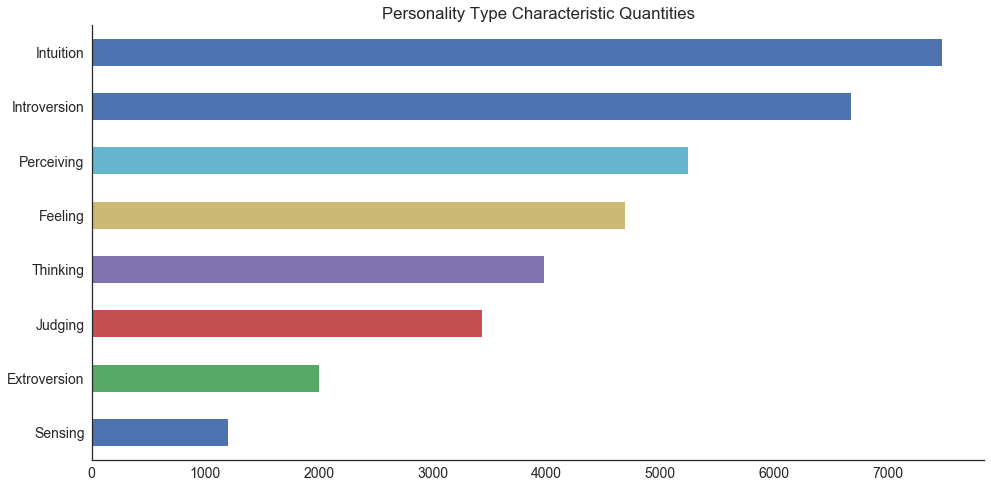
**a. MBTI Personality Type Distribution:**

****

Most common one is INTP,   
while least common one is ESTJ

****

**b. Personality Type Quantities:**

****

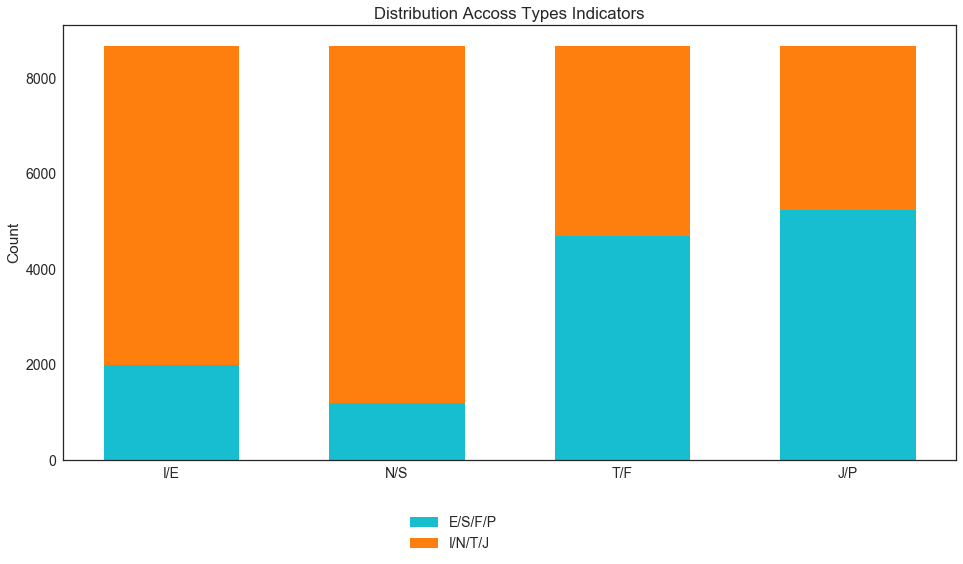
The graphic shows personality type quantities for each characteristic.

The most common is Intuition and the least common one is Sensing in our data set.

The graphic shows flight numbers for each month.

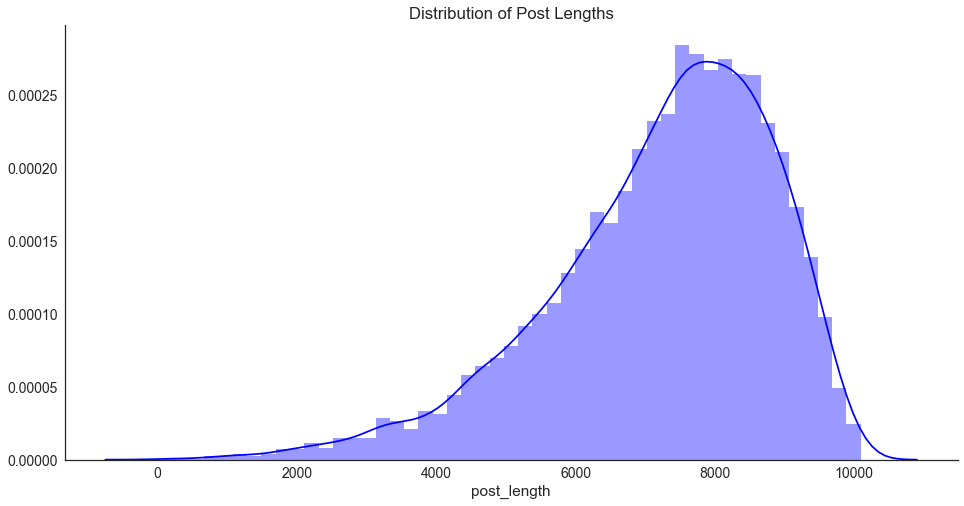
The flights are almost fairly distributed among the mounths. As we would expect, due to being the shortest month February has the least flight number.

**c. Distribution Across Personality Types:**

****

The graphic shows the number of Personality types. Intuition vs Sensing and Introversion vs Extroversion ones are imbalanced.

**d. Distribution of Post Lengths:**

****

The post lengths seem slightly skewed Normal Distribution

**4. MACHINE LEARNING MODELS**

This is a supervised multi-class classification problem. We are trying to predict the personality types among 16 classes from social media posts. We used Python’s scikit learn libraries to solve our problem. In this context, we implemented Logistic Regression, Linear SVM and XGBOOST algorithms.

In this study, in Part-1 we predicted personality types across 16 classes. The original data set had MBTI labels in post to representing some persons. We did not remove those labels in this part.

In the Part-2, we removed those labels and predicted personality types across 16 classes.

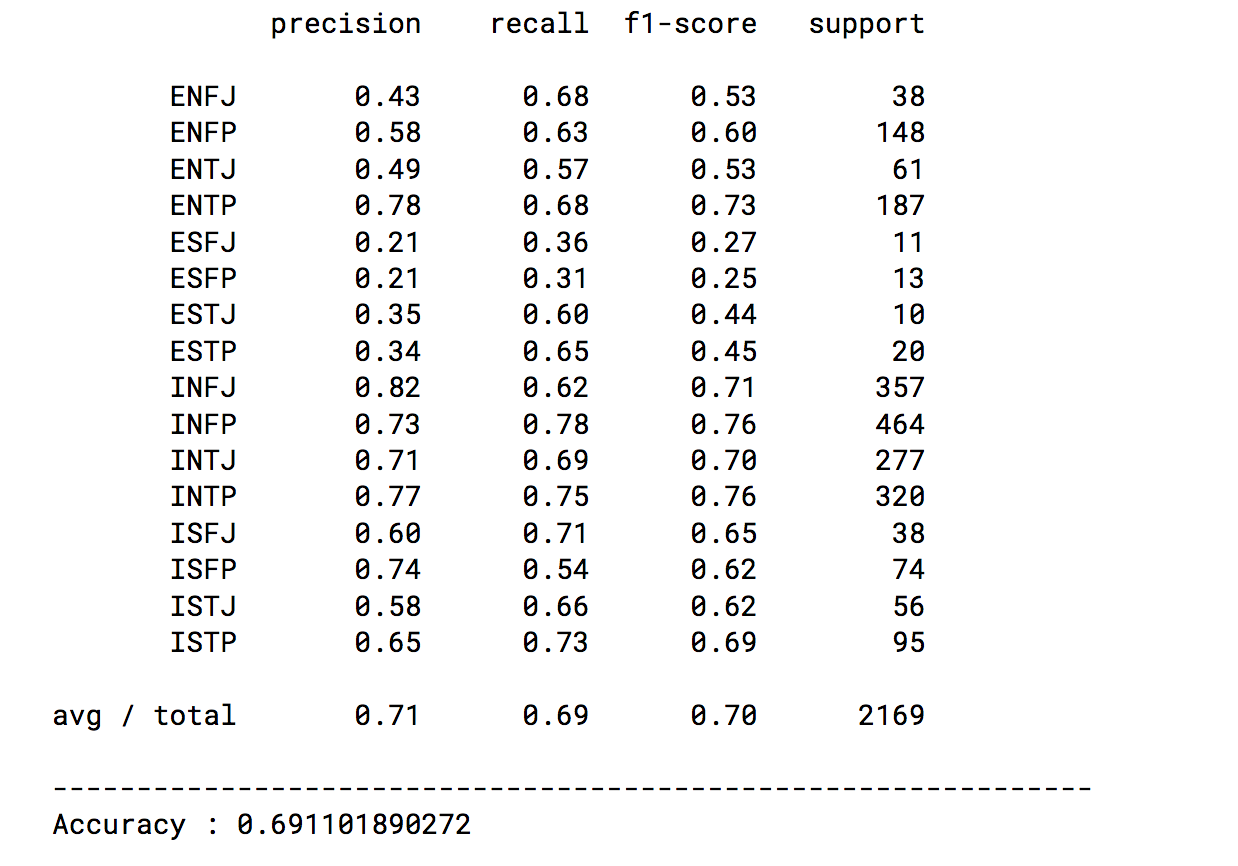
In Part-3, we predict only opposite classes such as Introversion vs Extroversion. We did not remove the MBTI labels from the posts.

In the last part, we removed those labels again and predicted the opposite personality types like in Part-3.

We split our data set into training set (75%) and test set (25%).

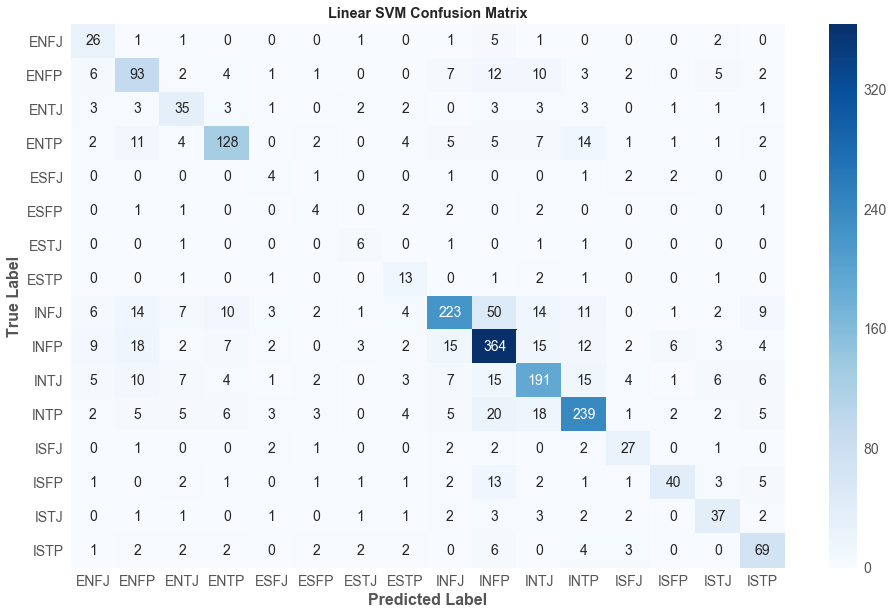
Additionally, we used Grid Search method with 5-fold cross validation technique to get rid of overfitting problem. As an evaluation metric we used accuracy.

1. **Part-1: Prediction Across 16 Classes (w MBTI labels)**

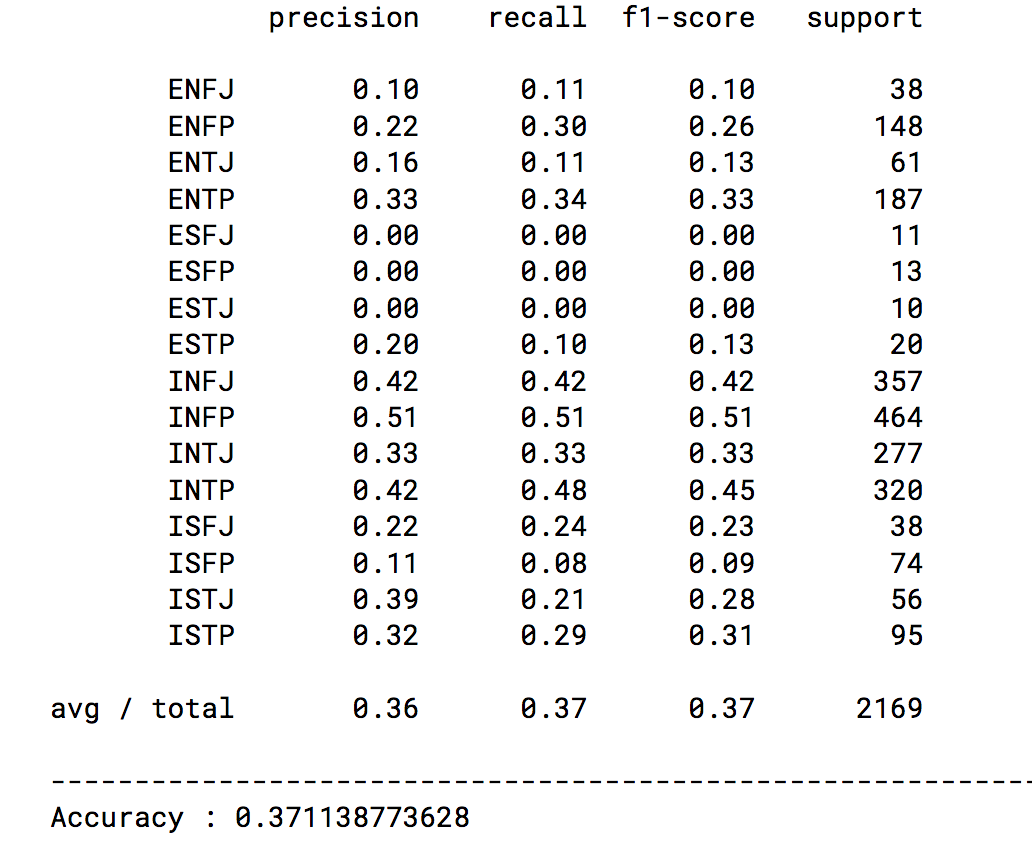
****

The best score is given by **Linear SVM with Count Vectorizer.**

The accuracy is **69%** while, baseline accuracy is 21%.

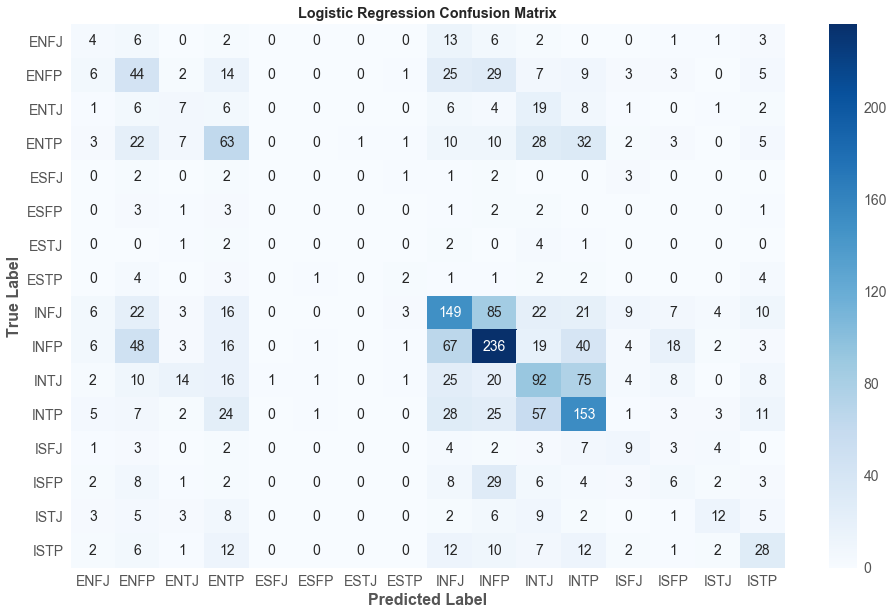
****

1. **Prediction Across 16 Classes (w/o MBTI labels)**

****

The best score is given by **Logistic Regression with Tf-Idf Vectorizer.**

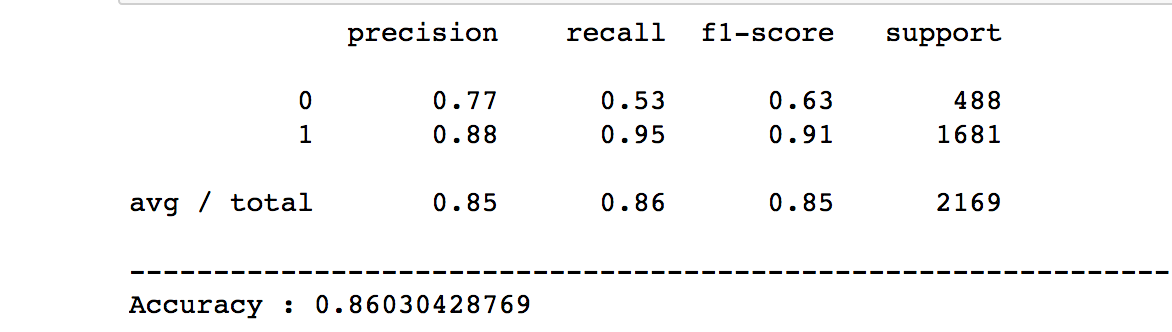
The accuracy is **37%** while, baseline accuracy is 21%.

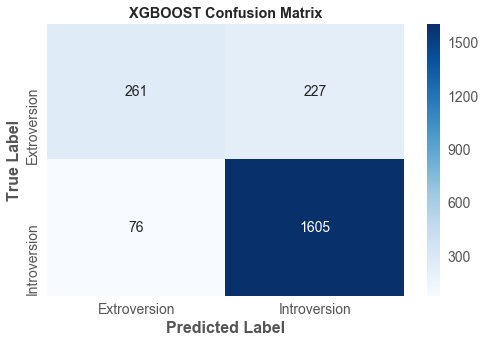
****

1. **Part 3: Prediction Opposite Personality Types (w MBTI Labels)**
2. **Introversion vs Extroversion:**

The best score is given by **XGBOOST  
 with Tf-Idf Vectorizer.**

The accuracy is **86%** while, baseline accuracy is 77.5%.

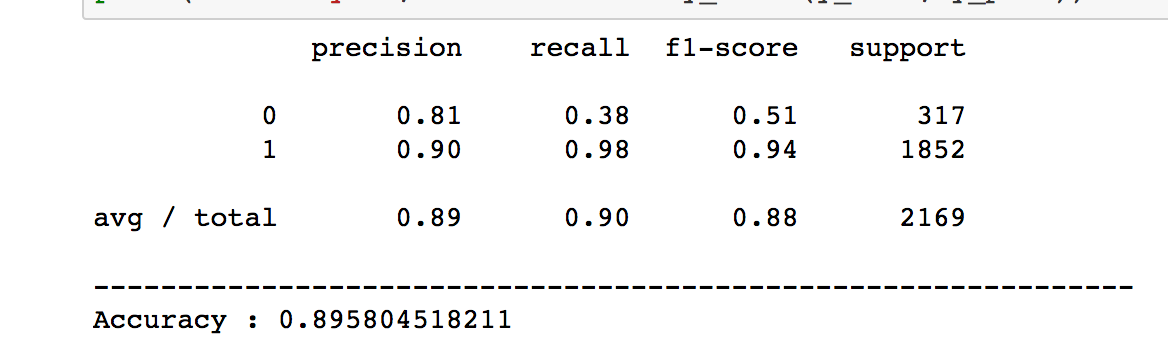


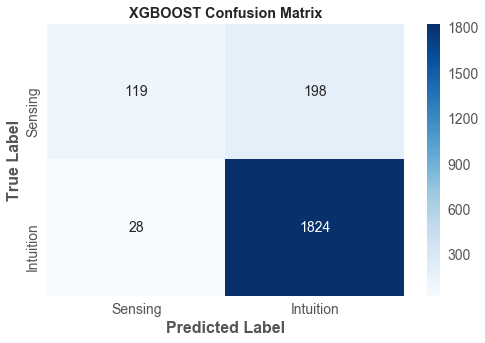


1. **Intuition vs Sensing:**

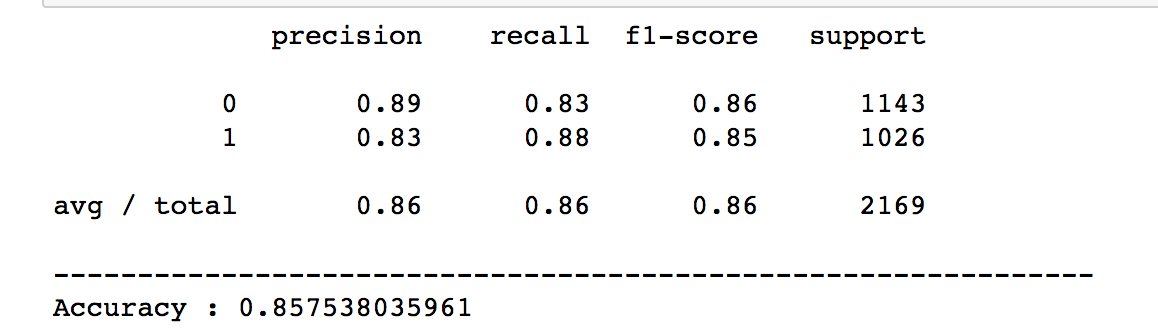
The best score is given by **XGBOOST  
 with Tf-Idf Vectorizer.**

The accuracy is **89.5%** while, baseline accuracy is 85.3%.

****

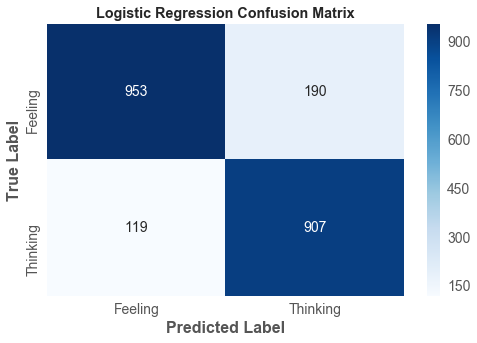
****

1. **Thinking vs Feeling:**



The best score is given by **Logistic Regression with Tf-Idf Vectorizer.**

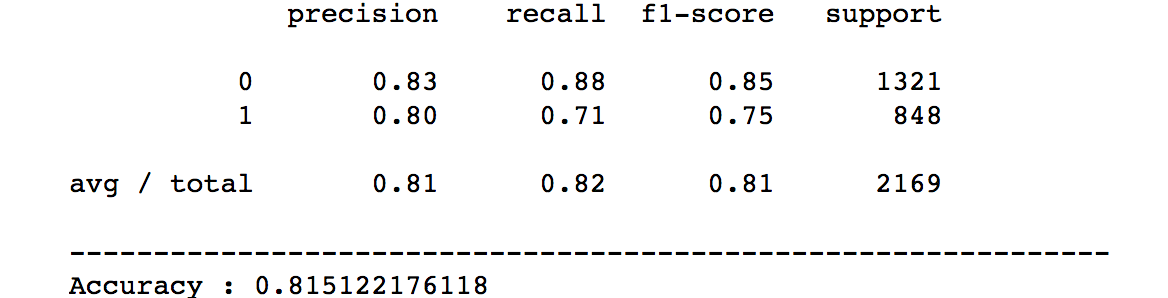
The accuracy is **85.7%** while, baseline accuracy is 52.7%.

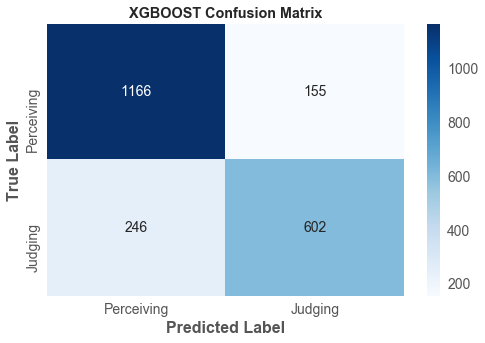


1. **Judging vs Perceiving:**

The best score is given by **XGBOOST with Count Vectorizer.**

The accuracy is **81.5%** while, baseline accuracy is 60.9%.

****

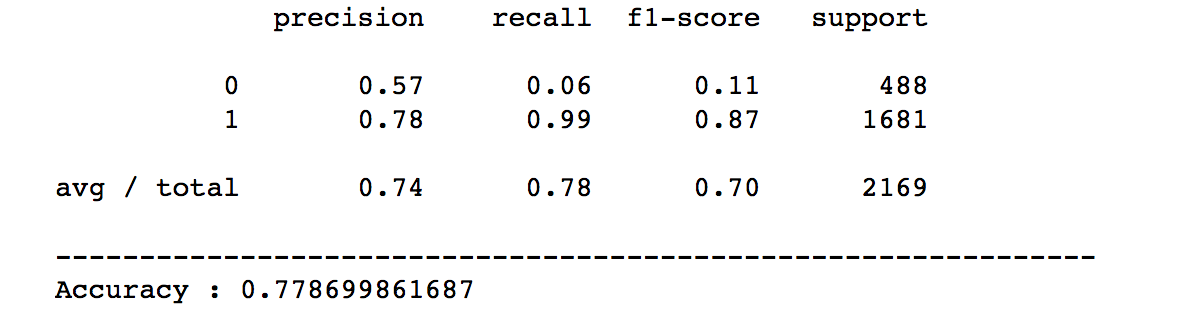
****

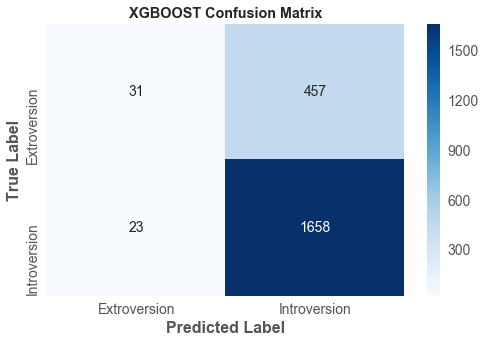
**d. Part 4: Prediction Opposite Personality Types (w/o MBTI Labels)**

**(1) Introversion vs Extroversion:**

The best score is given by **XGBOOST with Count Vectorizer.**

The accuracy is **77.8%** while, baseline accuracy is 77.5%.

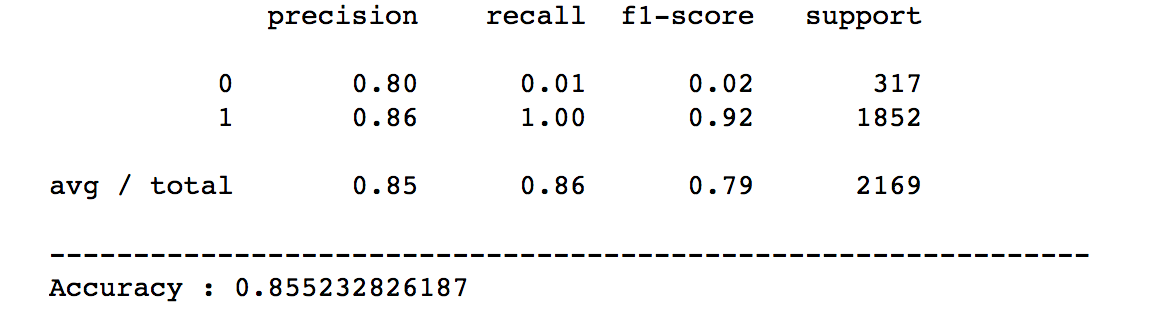
****

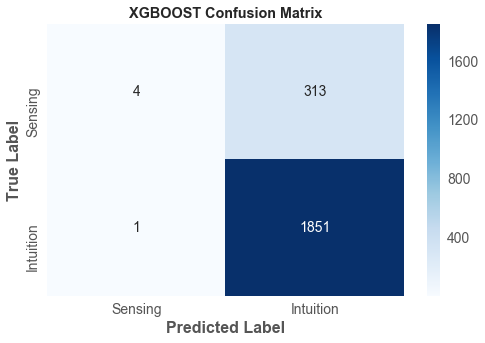
****

**(2) Intuition vs Sensing:**

The best score is given by **XGBOOST  
 with Tf-Idf Vectorizer.**

The accuracy is **85.5%** while, baseline accuracy is 85.3%.

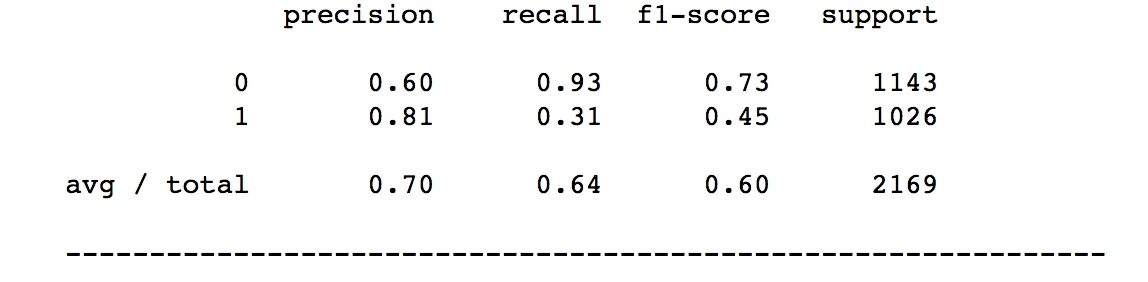
****

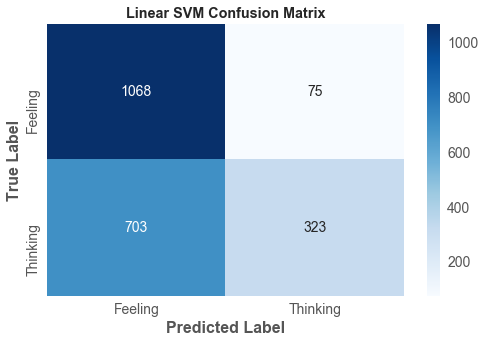


**(3) Thinking vs Feeling:**

The best score is given by **Linear SVM  
 with Tf-Idf Vectorizer.**

The accuracy is **78.2%** while, baseline accuracy is 52.1%.

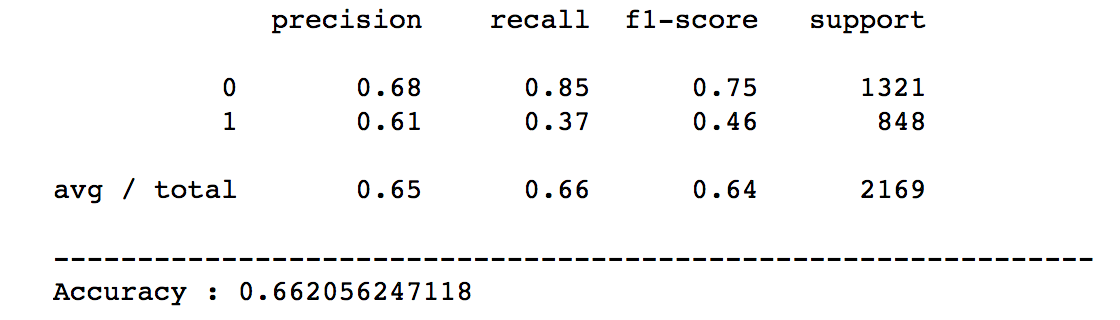




**(4) Judging vs Perceiving:**

The best score is given by **XGBOOST  
 with Count Vectorizer.**

The accuracy is **66.2%** while, baseline accuracy is 60.1%.



**5. CONCLUSION:**

In this study, we tried to predict the personality type of persons from their social media posts. In Part-1 our goal was to classify a person as one of the 16 classes of type for a given his/her post. In this context, in the Part-1, we predicted the personality type considering 16 classes while MBTI labels are in the posts. We obtained approximately 69% accuracy with MBTI labels through our best model. It seems a quite good result for 16 classes classification problem.

After that in the Part-2, we removed the MBTI labels from the posts and did the same exact thing in Part 1. Namely, the only difference between Part-1 and Part-2 is whether having MBTI labels in the posts or not. As a result, we concluded that MBTI labels are very important features to get higher accuracy for personality type prediction. We can roughly say that those labels affects almost 30% accuracy increment. But at the same time we know that those labels are artificial and in the real world we don't expect them in our data set.

In the Part-3, we predicted only opposite personality types such as Introversion vs Extroversion or Thinking vs Feeling etc. Like in Part-1 we kept MBTI labels in the posts data set. In this part of the study , we improved the baseline accuracy 9% for Introversion-Extroversion, 4.5% for Intuition-Sensing, 33% for Thinking-Feeling and 21% for Judging-Perceiving.

In the last part of the study, we removed the MBTI labels and predicted opposite personality types. As we figured out in the Part-2, MBTI labels are having an important role for prediction. As a result, we improved the baseline accuracy 8.5% for Introversion-Extroversion, 25% for Thinking-Feeling and 6% for Judging-Perceiving. Unfortunately, our models for Intuition-Sensing could perform as good as baseline accuracy. The reason for that is the data is fairly imbalanced. This issue is a future study subject.

During this study one of the problems we suffered was run time. For example, the mean run time for Part-1 and Part-2 is around 18 hours. If we slightly change something in the code we had to wait at least 18 hours to see the results for only that part or if we changed a single parameter affecting our model workflow then we had to wait at least 3 days to obtain results.

**6. FUTURE STUDY:**

As a future study, we will concentrate on the topics mentioned below:

* Some personality types are imbalanced. We used class\_weight to overcome this issue but we will use Upsampling techniques such as SMOTE to overcome the issue and get better accuracy results.
* We will implement Deep Learning models to get better results.
* We will use word2vec technique in NLP part.
* We will implement Dask library for parallel processing to decrease run time.
* After decreasing run time, we will focus on hyperparameter tuning more.