

Personality Prediction with ML: MBTI Classification

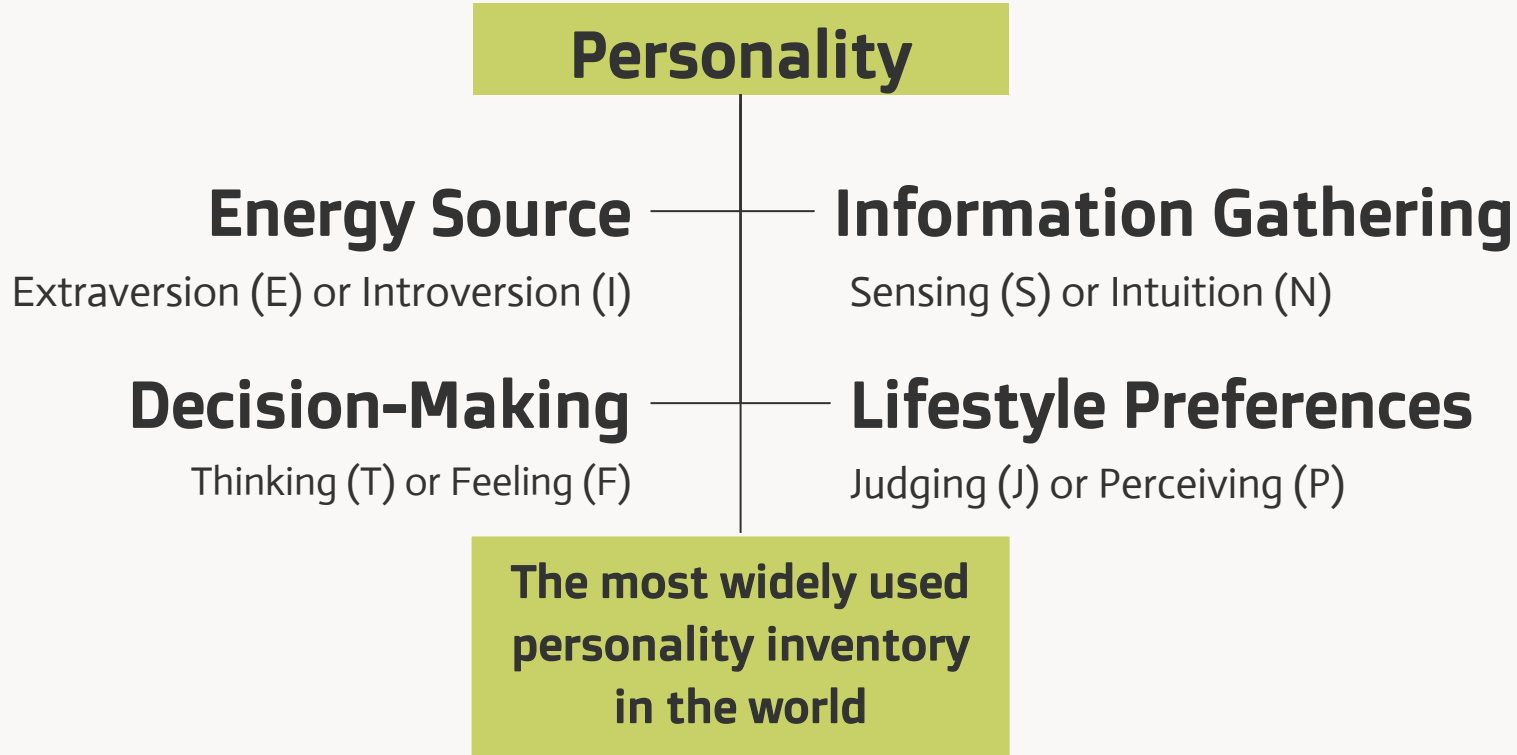
Vicki YE

CIS 413-PRACTICAL DATA SCIENCE USING MACHINE LEARNING





What is **MBTI**?





What is **MBTI**?

E

Extroverts

are energized by people, enjoy a variety of tasks, a quick pace, and are good at multitasking.

S

Sensors

are realistic people who like to focus on the facts and details, and apply common sense and past experience to come up with practical solutions to problems.

I

Introverts

often like working alone or in small groups, prefer a more deliberate pace, and like to focus on one task at a time.

N

Intuitives

prefer to focus on possibilities and the big picture, easily see patterns, value innovation, and seek creative solutions to problems.

T

Thinkers

tend to make decisions using logical analysis, objectively weigh pros and cons, and value honesty, consistency, and fairness.

J

Judgers

tend to be organized and prepared, like to make and stick to plans, and are comfortable following most rules.

F

Feelers

tend to be sensitive and cooperative, and decide based on their own personal values and how others will be affected by their actions.

P

Perceivers

prefer to keep their options open, like to be able to act spontaneously, and like to be flexible with making plans.

Business Problem & Workflows



Problem

Personality measure is widely used for recruitment and recommendation system.

BUT traditional assessments rely on subjective self-reports, prone to bias



Goal

Automate and objectively **predict personality types** from user-generated text



Process

Data Cleaning – Tokenization, removing noise

Feature Engineering – TF-IDF & Word Embeddings

Model Training – Comparing traditional & deep learning approaches

Evaluation & Validation – Accuracy, Precision, F1-score.

Deployment Considerations – API-based predictions for real-time use





App. in Behavioral Analysis

Predicts personality traits for candidate-
job fit

AI-driven hiring platforms (HireVue)



Analyzes text sentiment for early
intervention

Mental health applications (Woebot)



Suggests content or ADs based on
personality-driven user behavior.

Personalized recommendations (Ins, Twitter)



Models Used & Their Trade-offs

Logistic Regression

A statistical model based on linear relationships in text features

Random Forest

An ensemble learning algorithm that makes predictions by aggregating multiple decision trees

RNNs

Neural networks designed to process sequential text data and capture long-term dependencies

Strength

Interpretable, fast computation.

Handles non-linearity, robust to overfitting

Efficient for sequential text, ideal for long text

Weakness

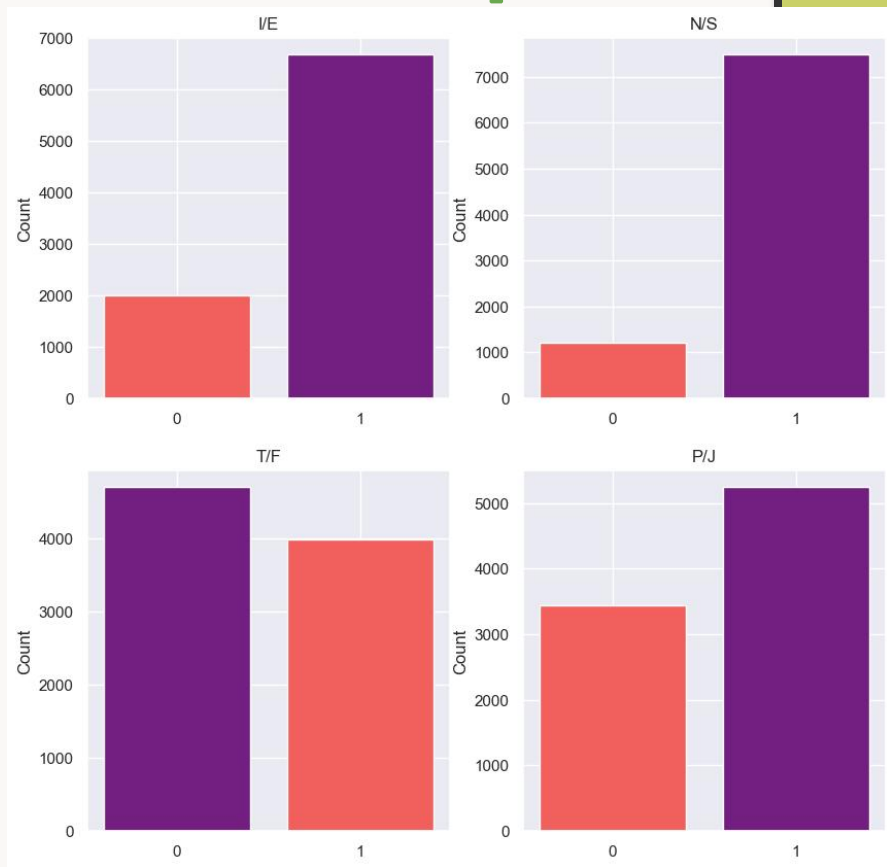
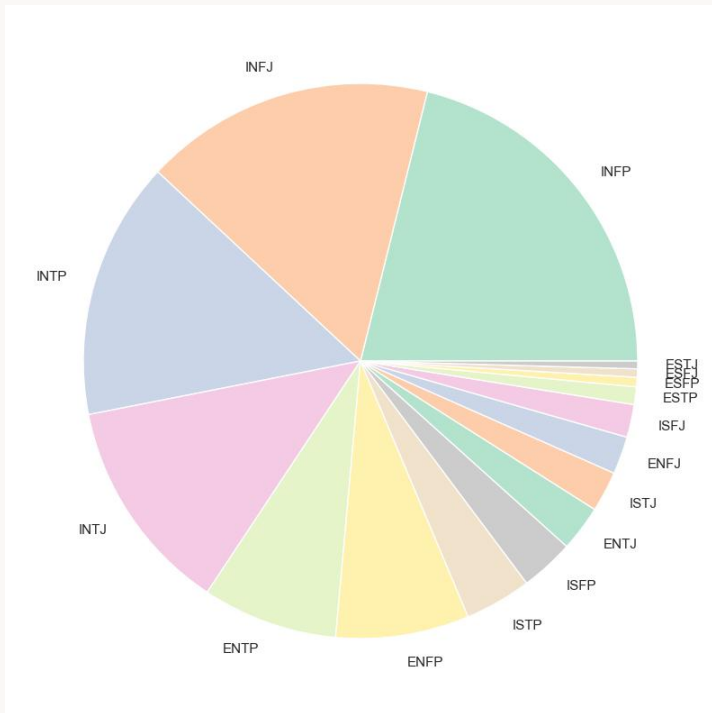
Struggles with non-linear text relationships.

Loses sequential text information.

Longer training time, requires large datasets.



From Raw Text to Model Input



From Raw Text to **Model Input**

Removed URLs, special characters, stopwords & Split user posts separated

Data Cleaning

Stratified train/test split to address class imbalance

Data Splitting

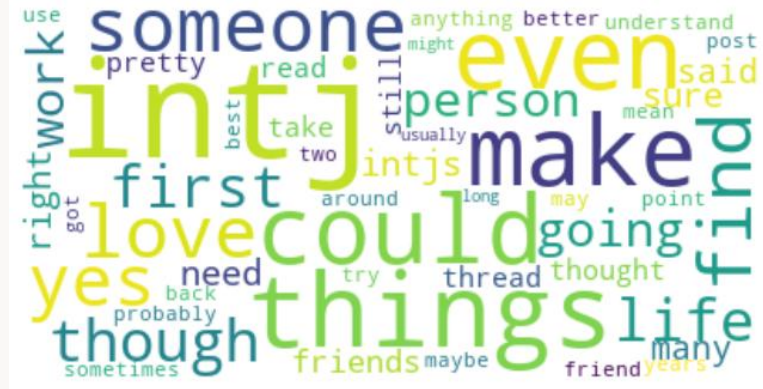
Feature Engineering

TF-IDF for traditional models & Tokenization

Wordcloud

Remove redundant words & show the most frequent words per MBTI type



A stylized illustration of a man's head and shoulders. He has brown hair, a full brown beard, and a mustache. He is wearing a light blue shirt. The background is a solid light blue color.

Why Accuracy, Precision, and F1?

Accuracy

- The proportion of correctly classified personality types out of all predictions
- Measures overall correctness but **doesn't account for class imbalance.**

Precision

- The proportion of correctly predicted personality types out of all instances classified as that type.
- **Important for avoiding misclassification** (like wrong personality type in hiring).

F1-score

The harmonic mean of precision and recall, balancing both for imbalanced datasets





Model Performance & Business Insights

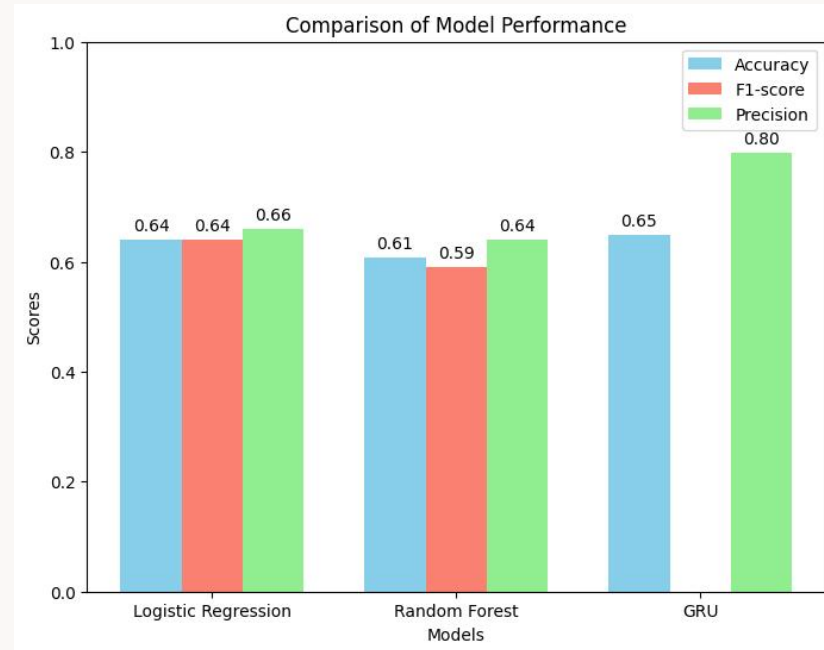
Logistic Regression	Accuracy: 0.6406 Weighted Average F1-score: 0.64 (Average of all types) Weighted Average Precision: 0.66
Random Forest	Accuracy: 0.6083 Weighted Average F1-score: 0.59 (Average of all types) Weighted Average Precision: 0.64
GRU	Categorical accuracy: 0.6486 Precision: 0.7974

- GRU achieves the highest accuracy
- Logistic regression is the best non-deep-learning model



Model Performance & Business Insights

- GRU achieves the highest accuracy
- Logistic regression is the best non-deep-learning model





Model Performance & Business Insights



Logistic Regression	Accuracy: 0.6406 F1-score: 0.64 (Weighted Average of all types) Precision: 0.66
Random Forest	Accuracy: 0.6083 F1-score: 0.59 (Weighted Average of all types) Precision: 0.64
GRU	Categorical accuracy: 0.6486 Precision: 0.7974



- Logistic Regression: Fast but struggles with nuanced text (e.g., sarcasm). Best for **quick screening**.
- Random Forest: Robust for short-text analysis; ideal for **social media snippets**.
- GRU: Captures **context in long posts (e.g., Reddit threads)**; optimal for high-stakes roles.

Next Steps to Improve Performance



Data-Level

- Apply bigger dataset (balance minority MBTI types like INFJ)
- Address **class imbalance** with synthetic data generation (SMOTE)



Model-Level

- Experiment with **pre-trained models** (BERT, RoBERTa)
- Hyperparameter tuning (GRU layers, dropout rates)



Business-Level

- Incorporate **user behavior data** (likes/shares) to enrich features.
- Integrate models into social platforms to enhance user recommendations.





References

- [1] Mushtaq, Z., Ashraf, S., & Sabahat, N. (2020, November). Predicting MBTI personality type with K-means clustering and gradient boosting. In 2020 IEEE 23rd International Multitopic Conference (INMIC) (pp. 1-5). IEEE.
- [2] Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., ... & Bühner, M. (2020). Personality research and assessment in the era of machine learning. *European Journal of Personality*, 34(5), 613-631.
- [3] Fernau, D., Hillmann, S., Feldhus, N., & Polzehl, T. (2022, September). Towards Automated Dialog Personalization using MBTI Personality Indicators. In INTERSPEECH (pp. 1968-1972).
- [4] Ryan, G., Katarina, P., & Suhartono, D. (2023). MbtI personality prediction using machine learning and smote for balancing data based on statement sentences. *Information*, 14(4), 217.
- [5] Nisha, K. A., Kulsum, U., Rahman, S., Hossain, M. F., Chakraborty, P., & Choudhury, T. (2022). A comparative analysis of machine learning approaches in personality prediction using MBTI. In *Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2021* (pp. 13-23). Springer Singapore.
- [6] Amirhosseini, M. H., & Kazemian, H. (2020). Machine learning approach to personality type prediction based on the myers–briggs type indicator®. *Multimodal Technologies and Interaction*, 4(1), 9.



Thanks!

**Have any questions?
You can write it under my post**

Feel free to check out my code—I've put a lot of thought into it :)