Kimberly Hall-Martin
Week 6- Project 2 White Paper
7/7/2025

Business Problem

The business problem at stake here is that ABC corporation is trying to improve team building and efficiency amongst employees working on projects together. They believe that if employees with like personalities are paired together, they will have better computability and outcomes on the projects they are working on. They are challenging the data science team to create a model that can use employee personality results to predict what type of personality category (introvert, extrovert, ambivert) the employee will fall into, so they can create teams accordingly.

Background/History

ABC corporation values their employees and collaborative group projects. They also place a big emphasis on efficiency and good communication. With these in mind they decided to give all 20,000 of their employees a personality test where they rate themselves on a scale from 1 (low- this does not apply to them) to 10 (high-this does apply to them) of how they perceive themselves in different categories. There were over 30 categories to rate themselves in some of which include creativity, organization, and leadership.

Data Explanation (Data Prep/Data Dictionary/etc)

This data was fairly clean and required limited data preparation. Personality_type was the only column that was not numeric, so a label encoder was used to change that. The data was split into training and tests sets, and a standard scaler was applied so the logistic, k nearest neighbors and random forest models could be fitted.

Kimberly Hall-Martin
Week 6- Project 2 White Paper
7/7/2025

Methods

The methods being used to predict personality types on this data are logistic regression, k nearest neighbors (KNN) and random forest regression. These are classification models that will predict personality type. Logistic regression is its own category of classifier while KNN and random forest are nonlinear classifiers (GeeksforGeeks). Exploratory data analysis (EDA) was also used to create visuals that show the relationships between traits and personality types. The last method used was features importance to see what features impact personality type the most.

Analysis

The analysis of this data set showed many things. To measure how well each model fit accuracy and F1 scores were used to evaluate performance. Each model performed very well. All models produced accuracy and F1 scores close to one, which is what is expected of a well fit model. The correlation heat map and boxplot visuals (Image 1 and 2) highlighted which traits were most often associated with each personality type. These images bring to attention that creativity, stress handling and emotional stability are not strongly correlated with one particular personality type. The top features in predicting personality type were party liking, public speaking comfort and excitement seeking (Image 3).

Conclusion

After analyzing this data, it was discovered that social interests and factors are what most affect a person's personality type. Factors regarding a person's character (empathy, stress handling, emotional stability, and creativity) had little impact on their personality type. All the characteristics that would theoretically make a good employee, do not impact personality type. Social traits, which assumedly have less impact on work ethic, are what decide personality type the most.

Assumptions

To be totally transparent with the audience, some assumption that were made in this project were that the data used is accurate and complete. It is also assumed that the data follows a normal distribution.

Challenges and Limitations

Some challenges faced were the validity of the data. It is important to remember that this data is coming from people rating themselves and how they feel about certain situations. There could be a lot of human bias, therefore potentially not giving totally accurate and reliable results. There was not much to be done about this challenge. It is a part of working with human opinion provided data.

Future Uses/Additional Applications

Some future uses of this application could be to curate specific groups based on the features. There may be specific projects that the company believes need individuals with specific personality traits, not specifically introvert, extrovert or ambivert. They may need to find people who specifically have high deep reflection, creativity, and listening skills, but also have low party liking and are not excitement seeking.

Kimberly Hall-Martin
Week 6- Project 2 White Paper
7/7/2025

Recommendations/ Implementation Plan

Recommendations would be to take this information and create your teams based on the personality result types. It is recommended that this data not be shared with employees, simply label groups as something inconspicuous such as group A, group B and group C.

To implement this plan, it is recommended to start with the employees you already have and categorize them into new teams. As you hire new employees, during the on boarding process have them take the personality test so they can be sorted into their teams at the very beginning.

It is recommended to analyze performance after one year to see if sorting teams by personality type is successful in increasing communication and teamwork. It is also suggested to have employees retake their personality test every year. Many factors can affect human mood and behavior and retesting yearly can keep up with those changes.

Ethical Assessment

The biggest ethical consideration for this project was the psychological impact it can have. It was feared that some employees could receive their scores and form a negative opinion of themselves. For example, some people could believe introverts to be boring and anti-social. When an employee has it revealed to them that they are an introvert, they could also believe those biases and begin to think poorly of themselves. With this in mind, the labeling of results was a major consideration. To maintain transparency when presenting to leaders, labels of introvert, extrovert and ambivert were not changed. When the employees were presented the groups, they were assigned, these labels were changed to A for introvert, B for extrovert and C for ambivert.

Another more common ethical consideration was that this was personal employee data. It is assumed that the company obtained this information in good faith and with the right permissions.

Image 1: Correlation Heat Map

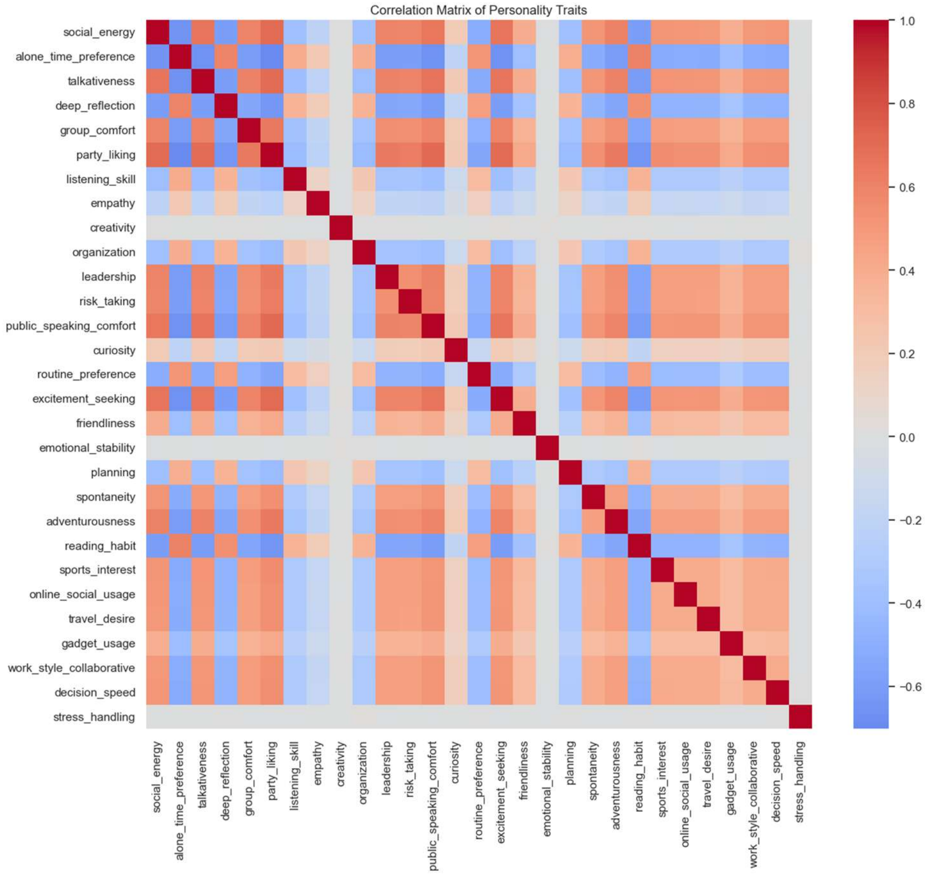


Image 2: Box Plot

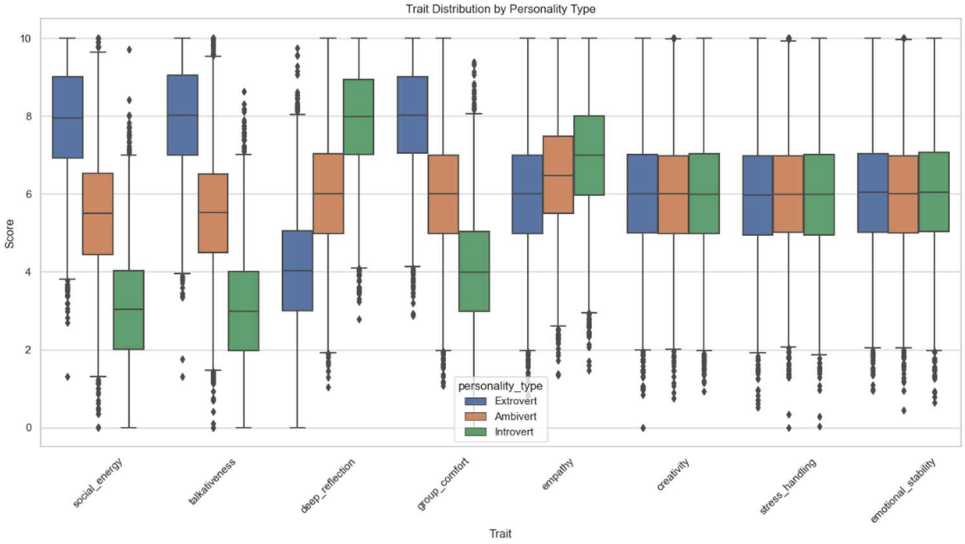
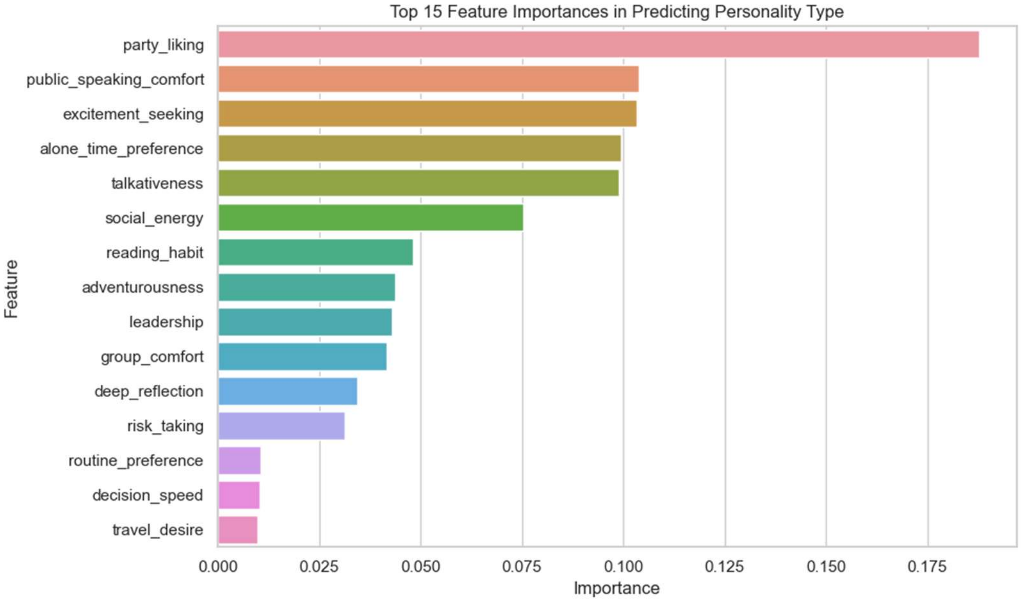


Image 3: Features Importance



Kimberly Hall-Martin
Week 6- Project 2 White Paper
7/7/2025

Question and Answer

1. Where did the data come from? Is it real or synthetic?
    a. This was synthetic data obtained from Kaggle.
2. Are these standardized psychological traits?
    a. These are not standardized psychological traits, as far as we are aware.
3. Why did you choose the models that you did?
    a. All models were chosen because of how well they fit the data. Random forest regression was used in particular to obtain features importance.
4. How well did each model perform on each personality type?
    a. No model performed exceptionally better on a certain personality type, they were all very similar.
5. Could this model be misused?
    a. Yes, there is possibility for this model to be misused, especially in the hiring process. This model should not be used when hiring individuals. People should be fairly and appropriately assessed based on their skills and experience, not their personality.
6. Would people agree with how the model classifies them?
    a. Based on average knowledge of these 3 personality types, I think people would agree with how the model classified them, but there are always going to be participants who might think differently
7. Could this model adapt to different cultural or age groups?
    a. This information was not provided to us in the original dataset, but it could be applied if given. This is not recommended in a professional environment.
8. What would happen if you added noise or real-world imperfections?
    a. Noise was added and the models were retested. They still performed very well, the accuracy and F1 scores only decreased slightly.
9. Why did you not use just accuracy to test model fit?
    a. The more metrics you use to evaluate model performance, the better chance you have of uncovering and unfit model.
10. Did you use cross-validation or just a single train-test split?
    a. Just a single train- test split was used.

Kimberly Hall-Martin
Week 6- Project 2 White Paper
7/7/2025

Appendix:

- The personality data set was obtained from Kaggle
    - https://www.kaggle.com/datasets/miadul/introvert-extrovert-and-ambivert-classification/data
- GitHub Link for Code
    - https://github.com/kimhall1/Personality-Type-Prediction-Project-2.git

Kimberly Hall-Martin
Week 6- Project 2 White Paper
7/7/2025

References:

GeeksforGeeks. (2025, January 20). *Getting started with Classification*. https://www.geeksforgeeks.org/machine-learning/getting-started-with-classification/