**Project Speech: "A Personality Trait Prediction Model Based on Handwritten Features"**

Dear teachers and classmates, hello everyone:

My name is Zhao Pengyu, and today I will show you my final project in CS422 Data Mining course - "Personality Trait Prediction Model Based on Handwritten Features".

**Part One: Research Background and Problem Statement**

Personality assessment has a wide range of applications in fields such as psychology, education, and human resources. The current mainstream evaluation methods often rely on questionnaires and subjective self-assessment, which are easily influenced by factors such as emotions and environment.

So I posed a new question: Can a person's score in the Big Five personality model be predicted through more objective handwriting behavior characteristics? That is to say, can we reconstruct a person's personality traits solely through a handwritten sample?

This is an innovative attempt, but also faces challenges such as:

* Multi dimensional objective regression is difficult;
* Features may be highly correlated with each other;
* Significant individual differences and large prediction bias.

**Part 2: Methodology and Technical Path**

In terms of methodology, I used a typical machine learning process. The project dataset is from Kaggle and contains a total of 4182 handwritten sample records, with each sample corresponding to 5 target variables, namely Openness, Conscientiousness, Extraversion, Agroableness, and Neurology.

Each piece of data contains multiple numerical features, such as letter spacing, stroke inclination, letter height, pen pressure, and so on. In terms of category characteristics, the main focus is on gender.

In terms of data processing:

* I used StandardScaler to standardize numerical features;
* Use OneHotEncoder to encode gender;
* Integrating these preprocessing steps together using ColumnTransformer improves the modularity and reproducibility of the process.

In terms of models, I have tried three typical regression models:

* Random Forest Regression
* XGBoost Regression (Gradient Boosting)
* Linear Regression

Calculate the R ² and MSE values of each model through cross validation to evaluate their performance in five dimensions.

**Part Three: Results and Visualization Display**

From the results, the XGBoost model performs better than other models overall, especially in the Openness and Conscientiousness dimensions, with ideal prediction performance.

However, it was also found that some personality dimensions had poor predictive performance, such as Extraversion, whose R ² value was even negative, indicating that the model's predictive accuracy in this area was not as good as the simple average.

To further improve the performance of the model, I conducted grid search parameter tuning on XGBoost. The adjusted parameters include the number of trees, maximum depth, learning rate, etc. The optimal parameter combination was ultimately determined and more stable performance was achieved on the test set.

In addition, I used SHAP tool for interpretability analysis, which not only ranked the impact of each feature on various dimensions, but also demonstrated how the model made judgments through visual charts. Among them, it was found that variables such as "letter height", "spacing consistency", and "tilt angle" have a significant impact on personality dimensions.

**Part Four: Conclusion and Prospect**

Overall, the optimized XGBoost model has achieved a good balance between performance and interpretability, making it the best performing model in this project.

The main contribution of this project is:

* Exploratory modeling of handwriting features for personality prediction has been achieved;
* Six new features were proposed, including the first systematic quantification of "letter spacing consistency";
* The single sample prediction speed only takes 0.8 seconds, far faster than the traditional questionnaire's 5-10 minutes.

Of course, there are still many areas worth improving in this project:

* In the future, each dimension can be independently modeled to improve personalized prediction accuracy;
* Introducing image recognition and deep learning models to directly extract features from handwritten images;
* Conduct generalization validation and fairness analysis across different populations, age groups, and cultural backgrounds.

The above is my presentation content. Thank you for listening. If you are interested in this project, please feel free to communicate and discuss. You can also contact me through my email z1603799337@hotmail.com Contact me. thank you!