Gender Classification

Experimental process overview

We propose a concise three-stage workflow to optimize predictive performance, beginning with data preprocessing to detect and correct inconsistencies, missing values, and outliers; followed by feature engineering to transform domain-specific variables for modeling compatibility; and concluding with model selection by comparing multiple tree-based algorithms to identify the best performer. Our experiments demonstrate that this rigorous approach to data preparation and targeted feature engineering substantially improves model accuracy and robustness.

Data observation and pre-processing

1. Remove Outlier

In the training set, outliers were identified using the interquartile range (IQR) method and subsequently marked as NaN to minimize their influence on downstream analyses.

2. Missing-Value Analysis

Missing values were then examined to determine their overall proportion and inter-feature correlations. The results indicated that each feature exhibited a similar missing-rate (Figure 1) and that there were no significant correlations among features with missing values (Figure 2). Accordingly, features with low data availability were earmarked for removal in order to preserve a larger pool of non-null samples, pending a more detailed evaluation of feature utility.

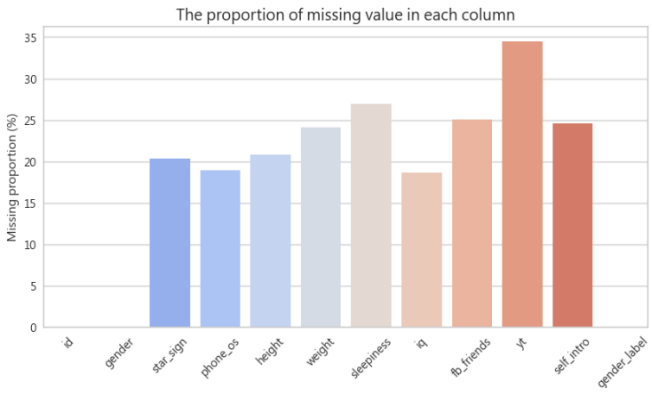

Figure 1. Missing ratio of each feature

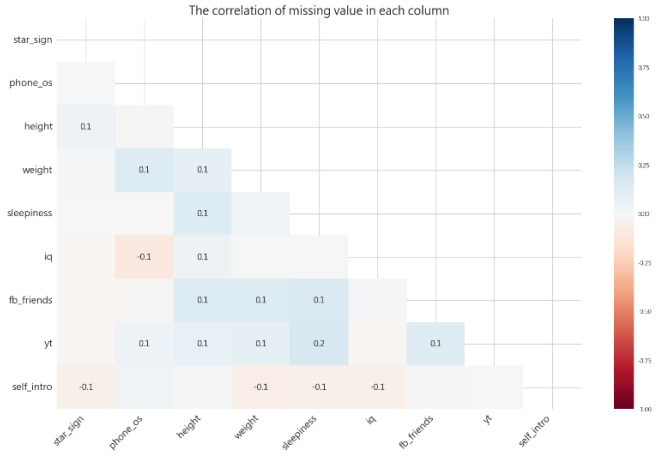

Figure 2. Missing Correlation

3. Feature Selection

Feature-selection analysis (Figure 3) revealed significant differences in height and weight between male and female subgroups. Specifically, both the height and weight distributions for males were shifted higher than those for females, demonstrating that these two features possess strong discriminative power.

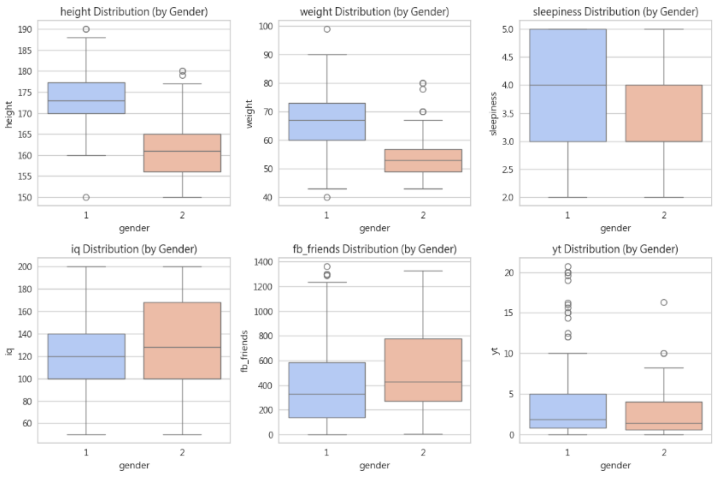

Figure 3. Distribution of numerical features of male and female

For the categorical feature “phone_os” (Figure 4), its distribution was observed to be similar to the overall gender ratio in the sample (approximately 3:1)(Figure 6). Among Android users, the male-to-female ratio is 4:1, and among Apple (iOS) users it is also 4:1. Regardless of the operating system, the proportion of males is consistently higher, so this feature lacks discriminative power. In addition, the “star_sign” feature exhibits high randomness with no significant patterns. In summary, height and weight are determined to be the most discriminative features.

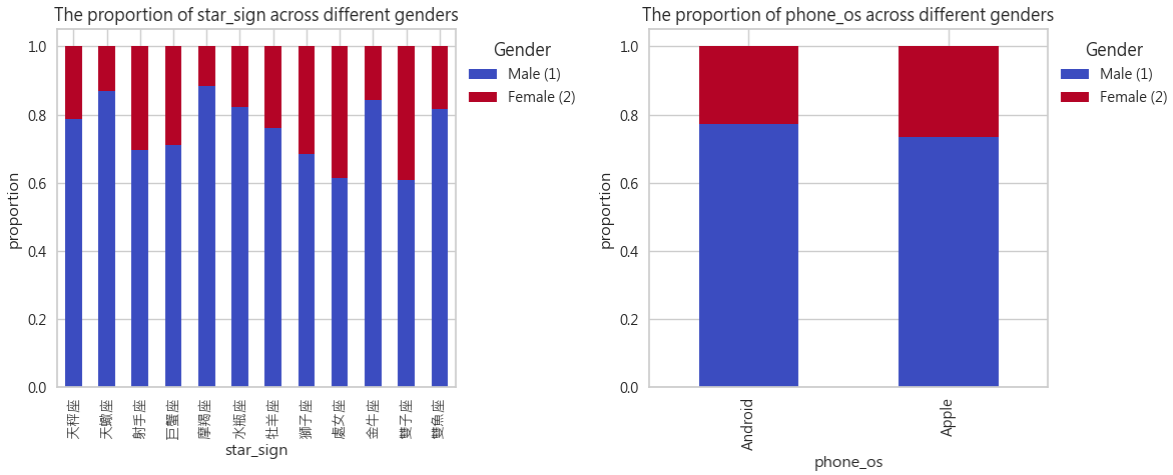

Figure 4. Distribution of numerical features of male and female

4. Data Imputation

Given the high correlation between height and weight (Figure 5), a mutual-reference imputation approach was adopted. Missing values in each feature were imputed using a random forest regression model trained on the other feature, thereby improving imputation accuracy.

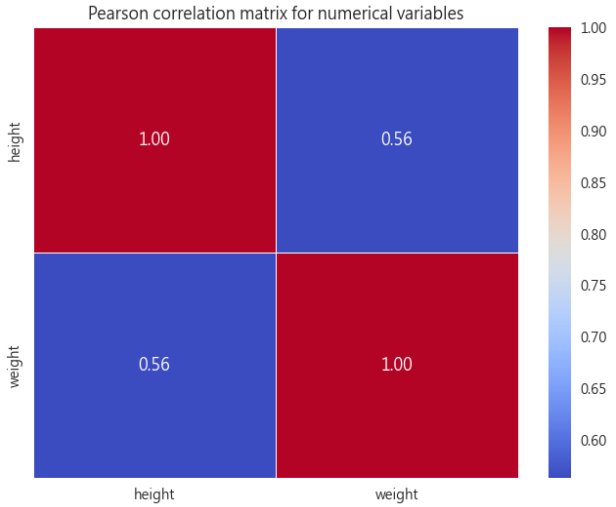

Figure 5. Correlation of numerical features in the training set

5. Imbalanced Sampling

To address class imbalance, we referred to our Kaggle competition’s best submission, which maintained a male-to-female ratio similar to our training set (Figure 6 and Figure 7). Consequently, no sampling adjustments were made in this experiment, preserving the original imbalance so that model predictions reflect real-world distributions.

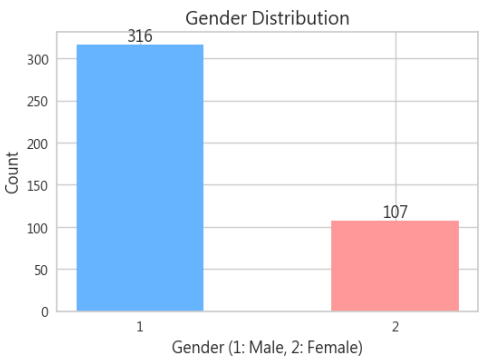

Figure 6. Distribution of male and female samples(train)

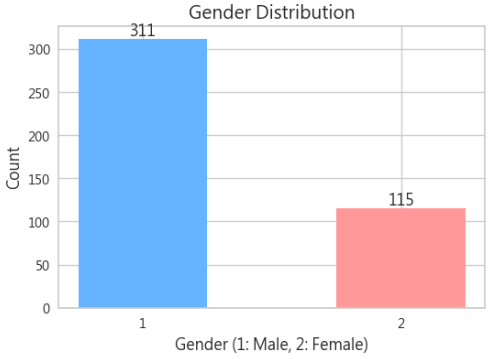

Figure 7. Distribution of male and female samples(best submission)

Test set is processed in the same way as the training set, including outliers (using the criteria of the training set), feature selection and filling methods, based on the similar feature correlation between the two to ensure consistency between the training and testing phases.

Text feature processing

1. Text Cleaning

First, inputs were checked to ensure they were strings; any non-string inputs were converted to strings. The text was then converted to lowercase to standardize the format. Using regular expressions, all punctuation was removed—retaining only letters, numbers, and spaces—and the cleaned text was split into tokens based on whitespace. Stopwords were then applied to filter out common words lacking substantive meaning. After text cleaning, we calculated word frequencies and use word cloud for male and female to visually present differences in word usage patterns between the sexes (Figure 8).

Figure 8. Male and Female Word Cloud

2. Weight Assignment

We conducted a detailed analysis of the most frequently used words by male and female users in the training set (Figure 9). A noticeable drop in word frequency was observed beyond the 28th-ranked term, which included both gender-specific words and those shared by both genders. To highlight gender-specific vocabulary, we further extracted the top thirty most frequent words for each gender and removed those common to both groups. Based on these observations, we assigned weights to the remaining vocabulary in proportion to their frequencies, and calculated a total weighted score for each instance. Instances with missing self-introduction entries (i.e., NaN) were assigned a score of zero.

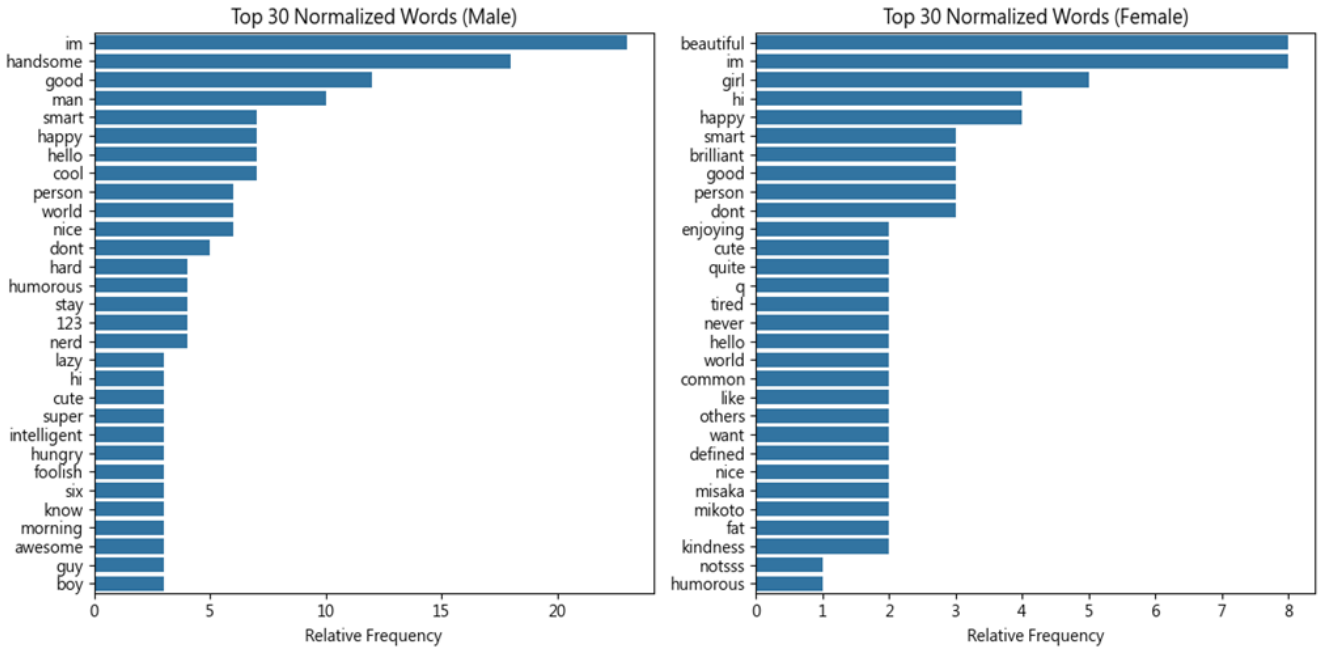

Figure 9. Words Used by Males and Females

Model Selection

In our study, we incorporated professor's recommendations from class and employed tree-based decision models. Given the diversity of tree models available in the literature, we compared the performance of various tree-based models for this task and selected the best-performing model as the standard for prediction. The models compared included the Decision Tree Classifier, Random Forest Classifier, Extra Trees Classifier, Gradient Boosting Classifier, Light Gradient Boosting Machine, and AdaBoost Classifier. Ultimately, we chose the Light Gradient Boosting Machine as our final model because it exhibited the best performance following parameter tuning.

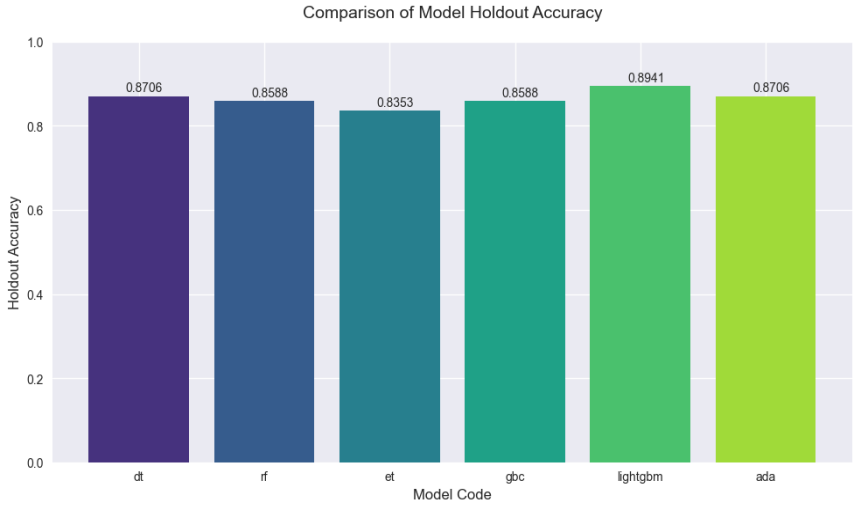

Figure 10. Comparison of the accuracy of various tree models

Teamwork Division

Name	Work description	Contribution
KAI-XIANG, CHANG 113423045	Develop model plan	33%
ZEI-WEI, XIE 113423036	Data observation and pre-processing	33%
KAI-SONG, KUO 113423027	Text feature processing	33%