NCU IM5033 Data Science and Machine Learning

Gender Classification Using Tree-Based Models and Text Features

KAI-XIANG, CHANG | ZEI-WEI, XIE | KAI-SONG, KUO

*contact information: Dept. of Information Management | Email: youto201266@gmail.com

Introduction

We present a three-stage workflow for gender classification, starting with data cleaning, followed by feature engineering, and concluding with model selection. By refining data quality and comparing tree-based models, we significantly improve classification accuracy and robustness.

Data Preparation & Feature Engineering

Remove Outlier

Outliers in the training dataset were identified using the interquartile range (IQR) method and set to NaN to reduce distortion in modeling.

Missing-Value Analysis

As shown in figure (**Fig. 1**) , all features had a similar missing rate. Figure (**Fig. 2**) confirms the absence of significant correlations among missing values. Features with low data completeness were excluded from further modeling.

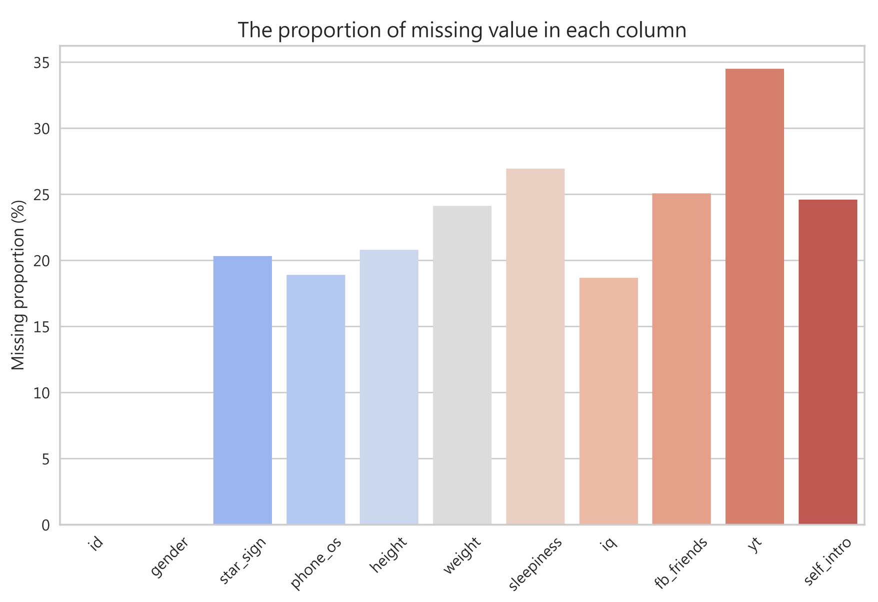

Figure 1. *Missing Proportion of Each Feature*

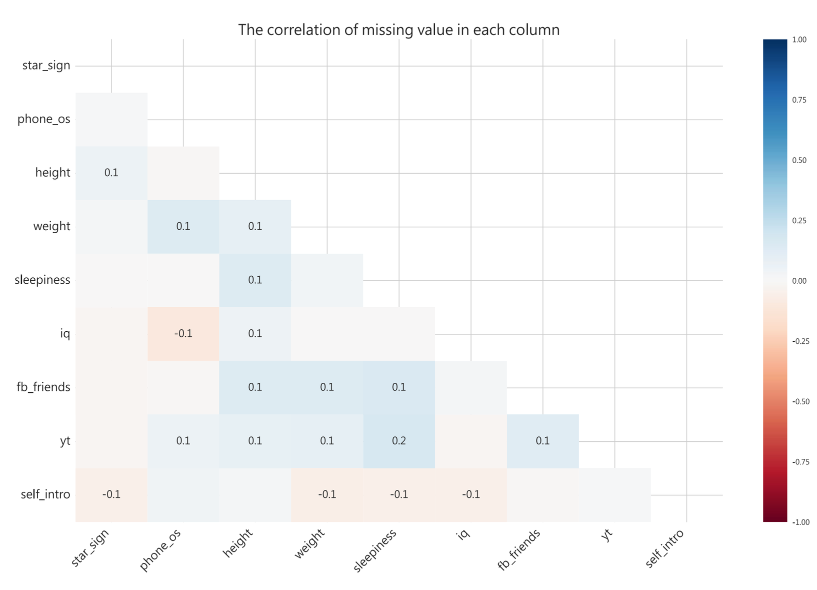

Figure 2. *Missing Correlation*

Numerical Features

Figure (**Fig. 3**) demonstrates a clear distinction between male and female groups—males had higher values on average, confirming these features' high discriminative power.

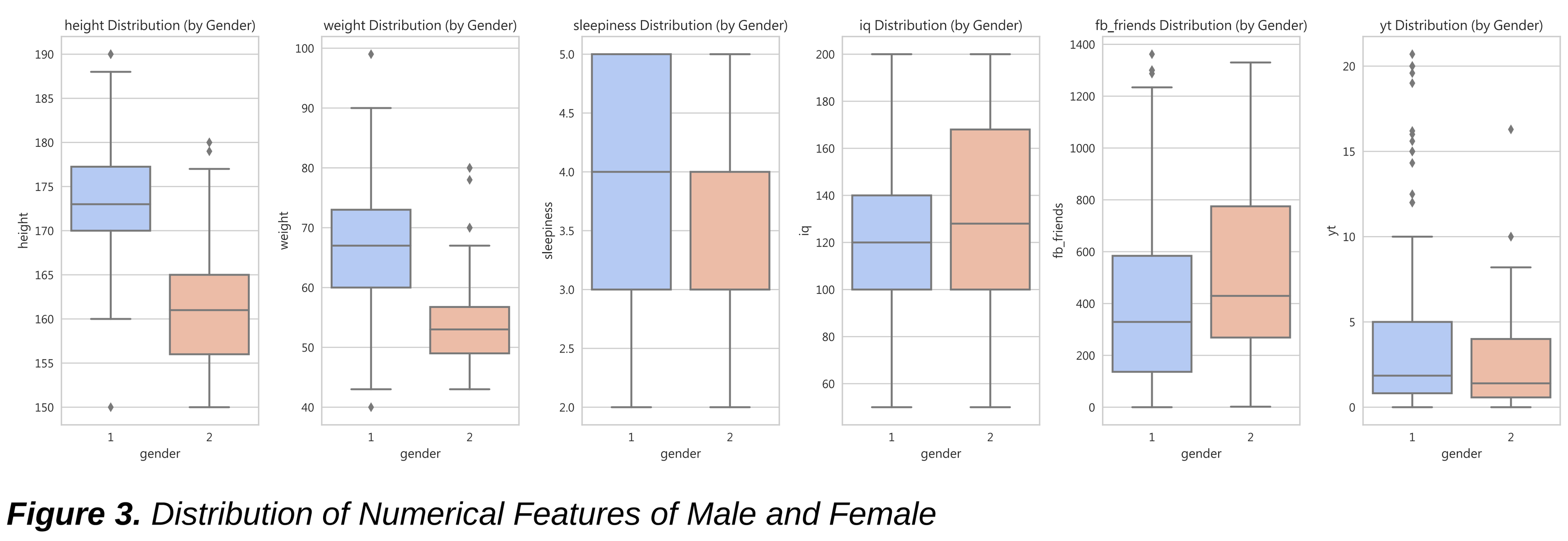

Figure 3. *Distribution of Numerical Features of Male and Female*

Categorical Features

Figure (**Fig. 4**) illustrates that "phone_os" is biased similarly to the gender ratio (~3:1) and lacks discriminative capability. Similarly, the "star_sign" feature appeared random and uninformative.

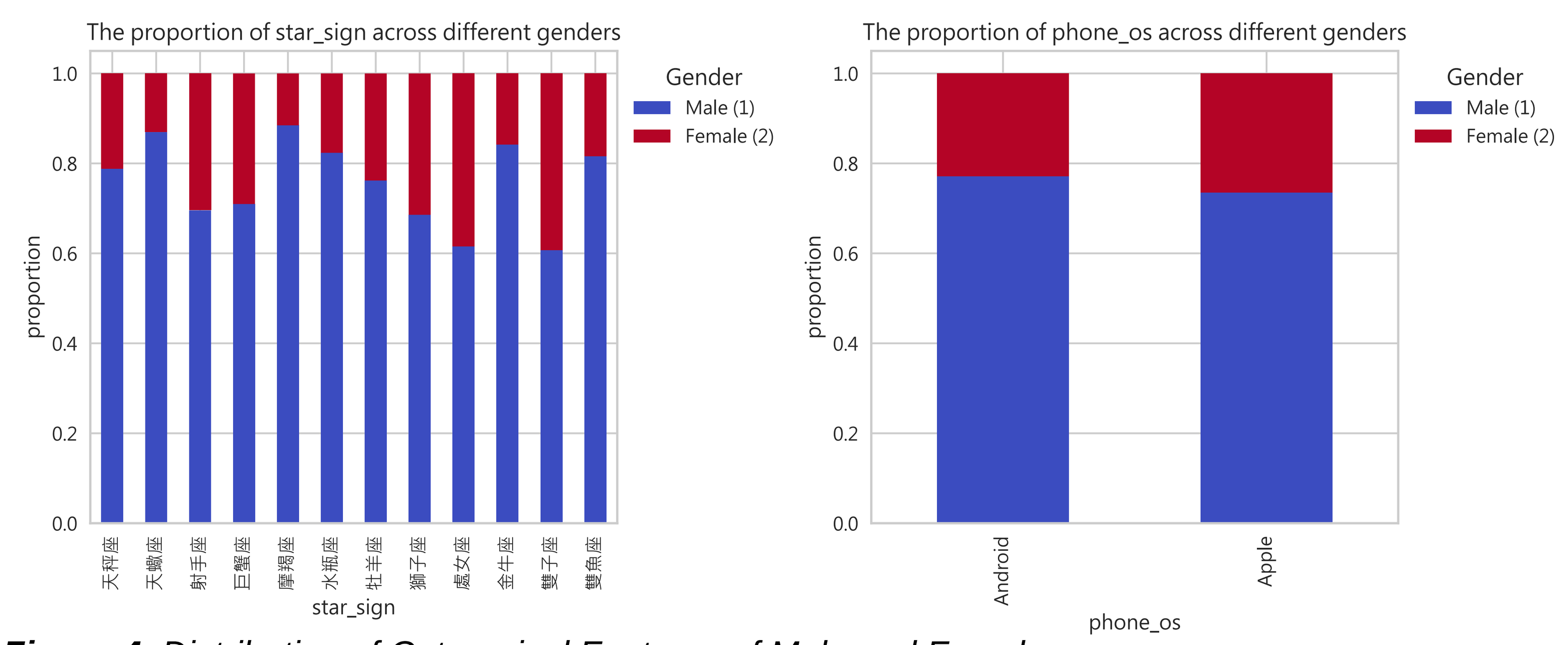

Figure 4. *Distribution of Categorical Features of Male and Female*

Data Imputation

Based on a strong correlation between height and weight, missing values in one feature were imputed using a random forest model trained on the other.

Class Imbalance

The class distribution was left unchanged based on the gender ratio observed in both training and best-submission sets, shown in figures (**Fig. 6, Fig. 7**) .

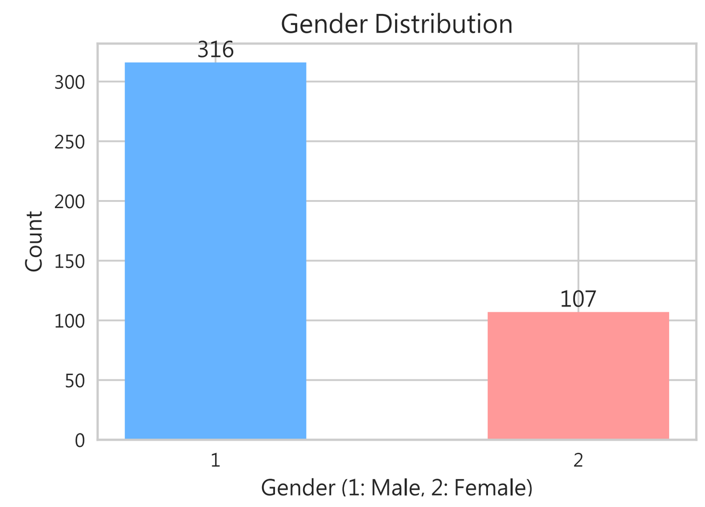

Figure 6. *Distribution of Male and Female Samples (train)*

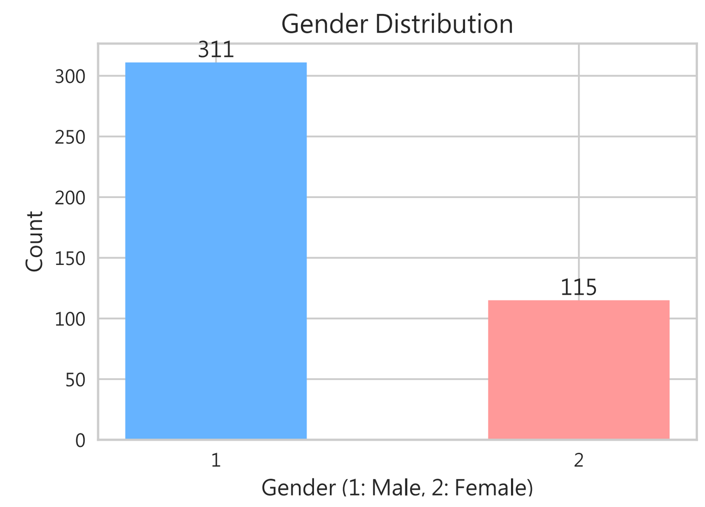

Figure 7. *Distribution of Male and Female Samples (best-submission)*

Text Cleaning & Weight Assignment

Included lowercasing, removal of punctuation via regex, and stopword filtering. Top 30 frequent words for each gender were analyzed (**Fig. 8**). Words common to both genders were excluded, and the remaining words were weighted based on frequency to compute a text score per instance.

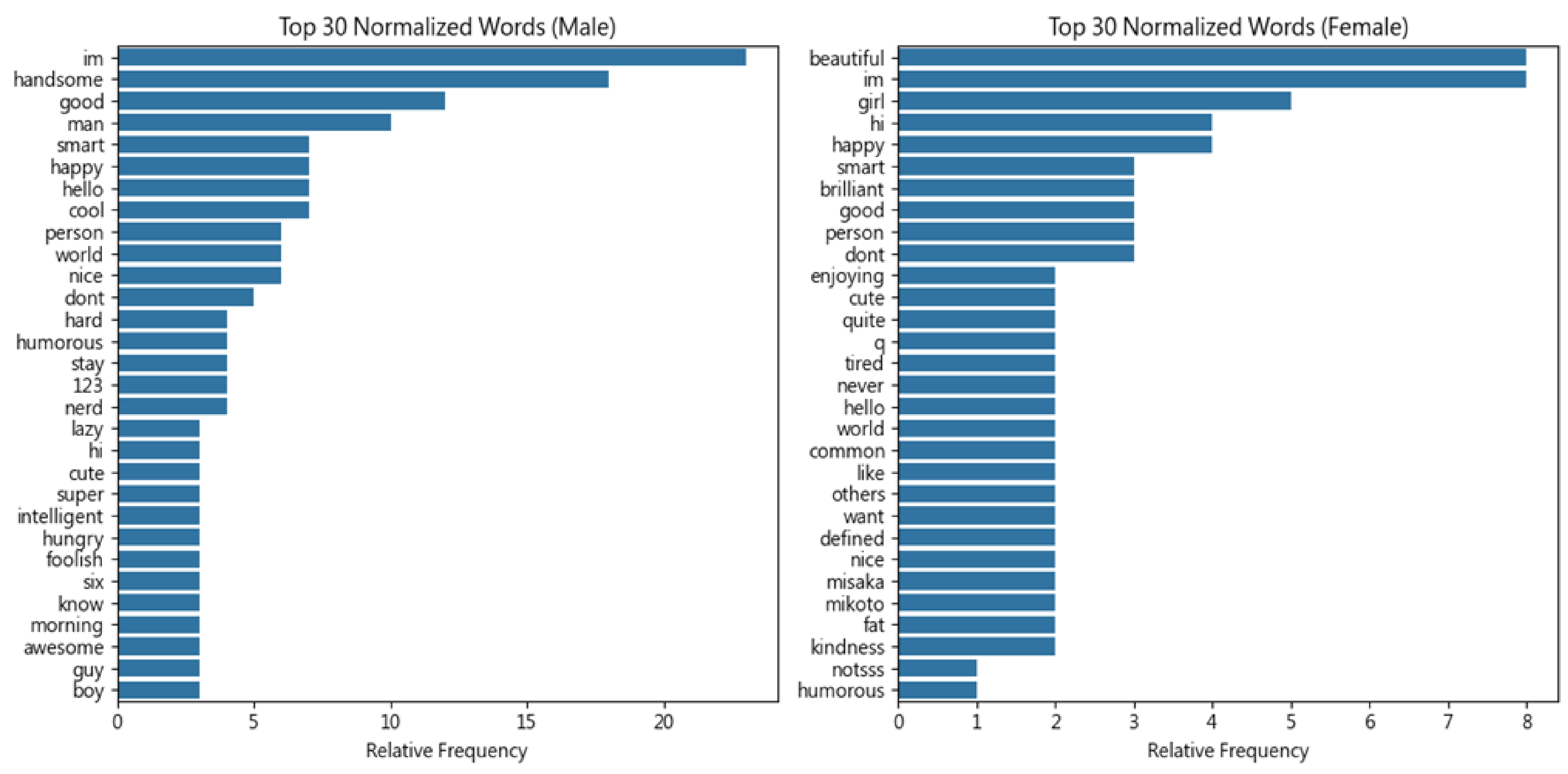

Figure 8. *Words Used by Male and Female*

Model Selection & Result

We evaluated six tree-based classifiers (**Fig. 9**)—Decision Tree, Random Forest, Extra Trees, Gradient Boosting, LightGBM, and AdaBoost—with LightGBM achieving the highest accuracy and selected as the final model. It achieved 88.73% accuracy on the public dataset and 87.79% on the private dataset.

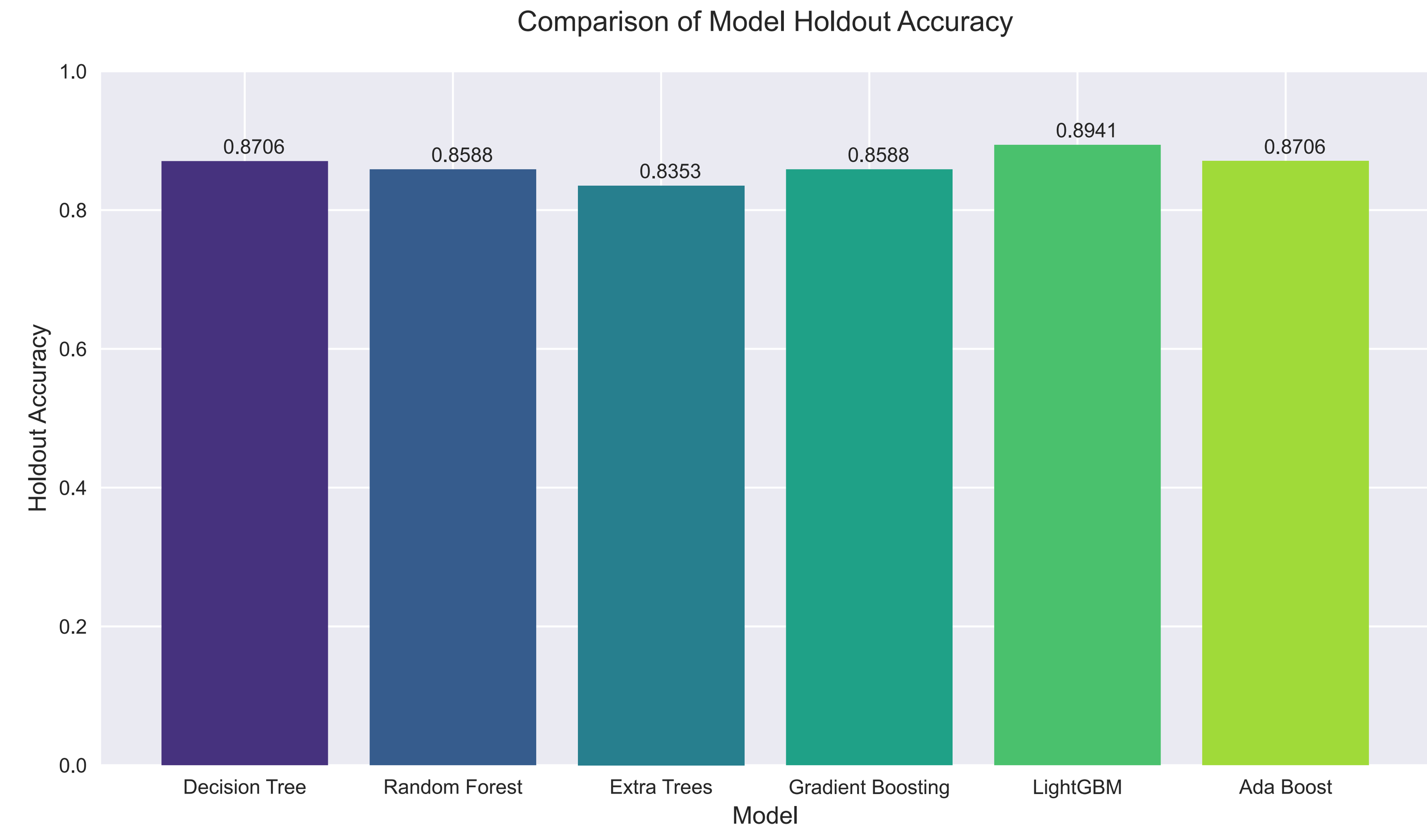

Figure 9. *Comparison of The Accuracy of Various Tree Models*