Gender Prediction Using Hanja-based Korean Names

KDI SCHOOL
KDI School of Public Policy and Management

David Woo, MPP • woo@kdis.ac.kr • Machine Learning for Social Scientists, Fall 2022

The Introduction + Goal

Korean names are often **derived from Hanja**, a **writing system** that is **based on Chinese characters**. Naturally, the name itself also has specific **sounds** when spoken.

Q: Can this pattern be used to predict gender?

Surnames are **not** expected to have an **effect** on **gender** and thus are disregarded.

Law (법도 도)	度 도	熙 희	Shining (빛날 희)

Dohui

Data Collection and Processing

Scraping (대법원: 'LOL no') → **Cleaning** (Removing Long Names) → **Processing** (Names → Syllables) → **Hanja Check** (Syllable vs. Hanja Dict.) → **Cleaning** (Removing Non-Hanja) → **Feature Engineering + One-hot Encoding** → **Machine Learning Applications**

건우, 미켈, 필립세라핀드 → 건우, 미켈, 필립세라핀드 → 건 우 / 미 켈 → 건 우 / 미 켈 → 건 우 → ㄱㅓㄴㅇㅜ → ㄱ = x0, ㅓ = x1

Machine Learning Models: The Decision Trees (because **Classification**)

Decision Tree
(18 hours, GridSearchCV)

- **Basic Tree**
- Easy to interpret and understand
- Overfitted data, hard to generalize
- Does not handle 3 character names well

Random Forest
(16 hours, GridSearchCV)

- **Bagging Ensemble**
- Better results, but still has a heavy bias on more represented data
- Does not handle 3 character names well

XGBoost
(21 hours, Bayesian Optimization via Hyperopt)

- **('eXtreme') Boosting Ensemble**
- Appropriately handles missing values and uses regularization to avoid overfitting/bias
- **Best Results!**

The Details

Prepared Dataset

- 5,398,828 Observations
- 155 Features (One-hot Encoding)
- Position-important
- 6.2 GB Mem. Usage
- 260 MB CSV File

Train/Test Split

- **Train:** 0.50
  - **Validation**
  - Test: 0.50 (0.25)
  - Train: 0.50 (0.25)
- **Test:** 0.50

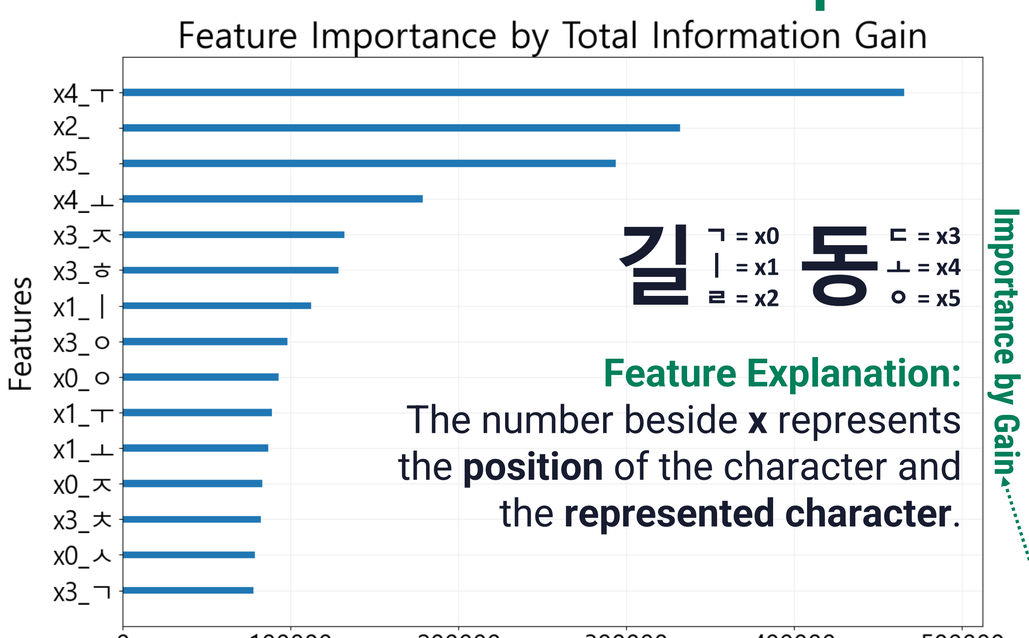
Lots of data. Tested with 0.75/0.25. The results are extremely similar.

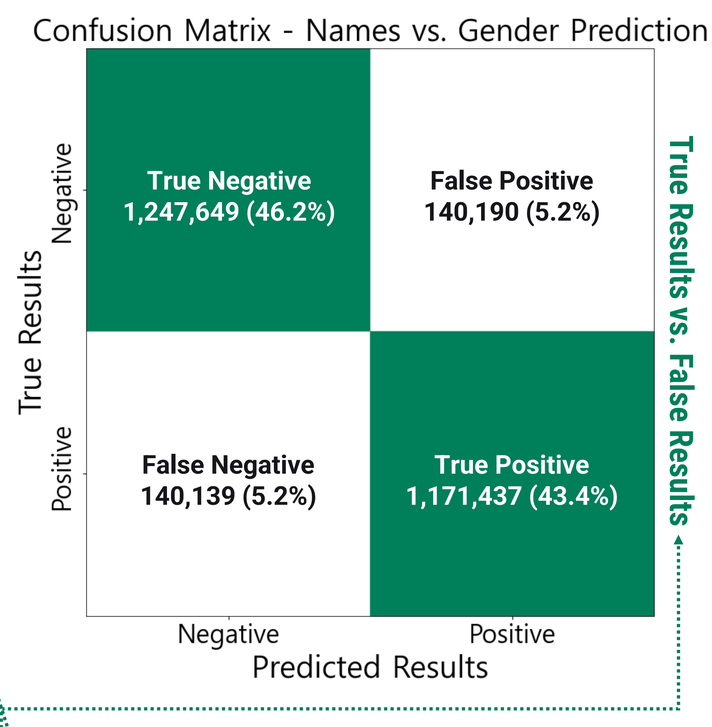
Prediction Conditions

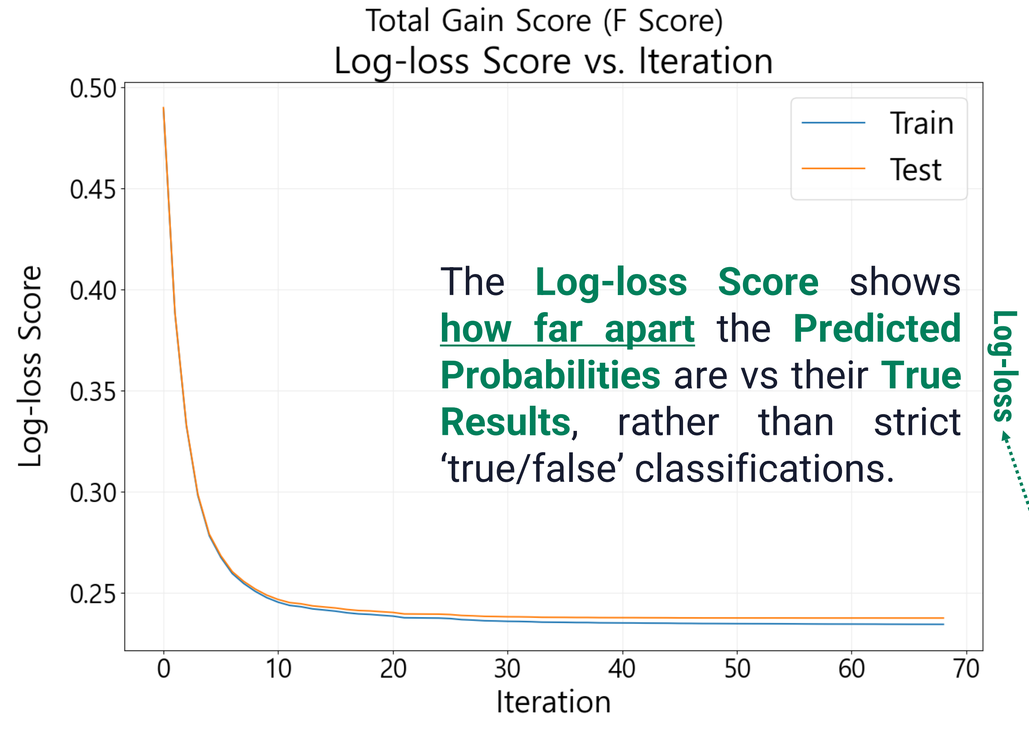
The Name must…

- Be in Korean
- Be Hanja-based
- Be 3 or fewer syllables
- Avoid unusual characters (ex. ㅃ, ㄹㅁ, ㄸ, etc.)

The Technical Interpretations







These 15 features have offered the **most discriminating information (gain)** in gender classification.

This is **different** from the **number of occurrences**, which may not tell us much.

Key Note about Names vs. Genders:
Perfect classification is **impossible.**

ex. 시현 = 9,462 Males (51.9%) and 8,746 Females (48.1%).
There are 672 'permutations' of that name: 시: 28, 현: 24

Good enough is the desired goal.

The References

• Park, Jaehyuk. "Machine Learning for Social Scientists." 2022, Sejong, KDI School of Public Policy and Management. • Hu, Yifan, et al. "What's in a Name? -- Gender Classification of Names with Character Based Machine Learning Models." arXiv, 2021, https://doi.org/10.48550/arXiv.2102.03692. Accessed 2 Nov. 2022. • "Hanja." Wikipedia, Wikimedia Foundation, 23 Nov. 2022, en.wikipedia.org/wiki/Hanja. Accessed 7 Nov. 2022. • "다음 한자사전 (Daum 漢韓辭典)." Daum 사전, https://dic.daum.net/index.do?dic=hanja. • "대한민국 이름 통계 서비스.", https://www.name-ranking.com/. • Google + Stack Overflow et. al – For any technical information regarding machine learning, decision trees, XGBoost, parameter optimization, etc. • And more. So, so much more. – ☺