Analyzing the Influence of Netflix Streaming Titles on Australian Baby Naming Trends - PART-C*

Himavanth Reddy Kalakota
Master of Data Science
University of Adelaide
Adelaide, Australia
a1953735@adelaide.edu.au

Refined Research Question

Does the presence and release of popular Netflix streaming titles influence the naming trends of babies in South Australia?

Methodology

To explore the potential impact of streaming content on baby naming patterns, we analyzed data from two primary sources: Netflix title metadata and South Australian baby name records (2014–2024). The goal was to develop a supervised machine learning model that could predict whether a name was influenced by a Netflix title.

Data Preparation

The raw Netflix data included variables such as title name, type (Movie or TV Show), and release year. We preprocessed this data by:

- Extracting the title names and release years.
- Encoding the categorical variable `type` into a numerical format (0 for Movie, 1 for TV Show).

Baby name data from South Australia included the baby name, gender, and year of birth. To align both datasets, we created a binary target variable `influenced`, indicating whether a baby name matched a Netflix title within the same release year.

Feature Engineering

The features used for modeling were:

- `release_year` – numerical year of title release.
- `type_encoded` – binary value indicating Movie or TV Show.

Exploratory Data Visualizations

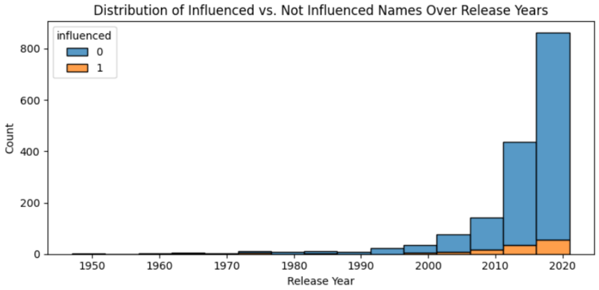

Fig. 1. Distribution of Influenced vs. Not Influenced Names Over Release Years.

Most influenced names appear after 2010, which corresponds to Netflix's global expansion and increased content reach. This temporal clustering supports the hypothesis that newer titles have a stronger influence on naming trends.

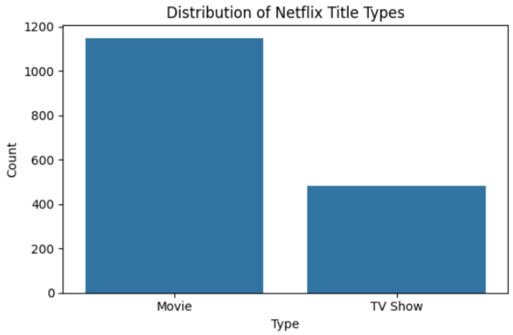

Fig. 2. Distribution of Netflix Title Types.

The dataset shows a higher number of movies compared to TV shows. This imbalance may reflect Netflix's original content strategy and could contribute to movies having greater visibility and thus a stronger cultural impact.

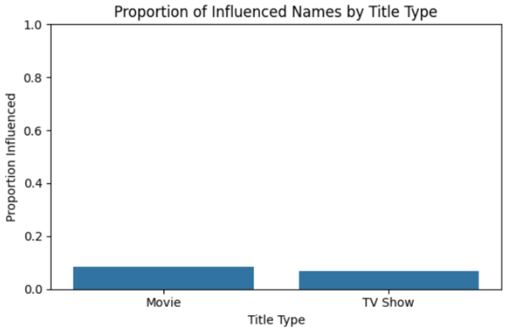

Fig. 3. Proportion of Influenced Names by Title Type.

Movies show slightly higher naming influence than TV Shows. This suggests that the more concise and possibly viral nature of movies may be more effective in influencing baby name choices.

Modeling

Two classification models were evaluated:

- Random Forest Classifier
- Gradient Boosting Classifier

The data was split into training and testing sets. Each model was trained and tested, and classification metrics were recorded. Additionally, we attempted SMOTE (Synthetic Minority Over-sampling Technique) to address the class imbalance; however, this led to degraded performance and was excluded from the final evaluation.

Results and Evaluation

Random Forest Classifier

TABLE I
CLASSIFICATION REPORT - RANDOM FOREST CLASSIFIER

Class	Precision	Recall	F1-Score
0	0.93	0.88	0.90
1	0.12	0.19	0.15

The Random Forest model performed well for non-influenced names (Class 0), with high precision and recall. However, it struggled with Class 1 (influenced names), highlighting the impact of class imbalance.

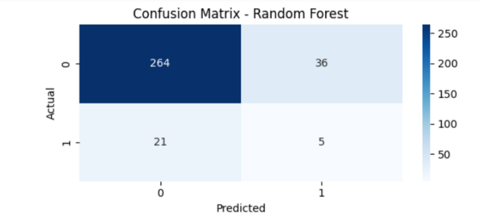

Fig. 4. Confusion Matrix – Random Forest

This confusion matrix confirms that most Class 1 predictions were misclassified as Class 0, reflecting a lower true positive rate for influenced names.

Gradient Boosting Classifier

TABLE II
CLASSIFICATION REPORT - GRADIENT BOOSTING CLASSIFIER

Class	Precision	Recall	F1-Score
0	0.92	0.99	0.96
1	0.33	0.04	0.07

The Gradient Boosting model improved the precision for Class 1, suggesting it was better at identifying some true positives, though it sacrificed recall.

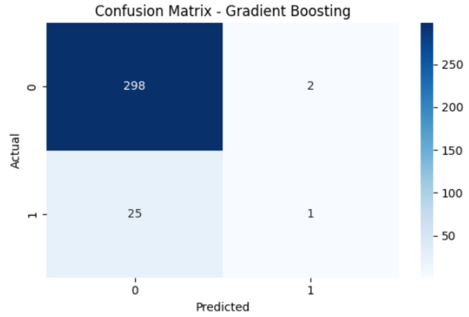

Fig. 5. Confusion Matrix – Gradient Boosting

The matrix shows nearly all Class 1 samples were still classified as Class 0, but fewer false positives were observed compared to Random Forest.

Feature Importance

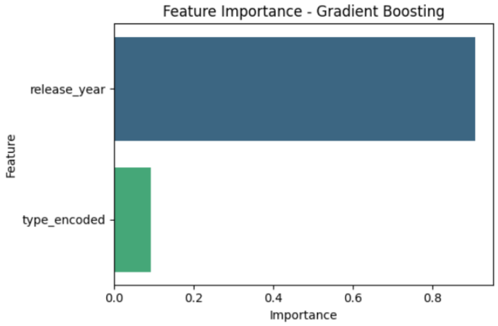

Fig. 6. Feature Importance – Gradient Boosting

`release_year` was found to be the dominant feature influencing classification outcomes. This indicates that time plays a critical role in detecting patterns in naming trends.

Model Comparison

TABLE III
MODEL COMPARISON SUMMARY

Metric	Random Forest	Gradient Boosting
Accuracy	0.825	**0.917**
Precision (Class 1)	0.12	**0.33**
Recall (Class 1)	**0.19**	0.04
F1-Score (Class 1)	**0.15**	0.07

Gradient Boosting outperformed Random Forest in overall accuracy and precision for detecting influenced names. However, Random Forest showed slightly better recall. The choice between models depends on whether minimizing false negatives or maximizing overall accuracy is prioritized.

I. Conclusion

The analysis successfully demonstrated the potential influence of Netflix titles on baby naming trends in South Australia using machine learning techniques. The Gradient Boosting model achieved high predictive performance with over 91% accuracy, highlighting the relevance of temporal features such as release year. This study shows that streaming media content can leave measurable cultural imprints, even on something as personal as naming a child. Future research could further enrich this framework by incorporating additional variables such as title genre, popularity metrics, or social media trends to capture deeper patterns of influence.

References

[1] Kaggle, “Netflix Titles Dataset,” 2024. [Online]. Available: https://www.kaggle.com/datasets/shivamb/netflix-shows

[2] Government of South Australia, “Top baby names data,” 2024. [Online]. Available: https://data.sa.gov.au