Final Report: Wool-Sucking Cat Adoption Predictive Modeling

1. Introduction

This report summarizes the findings and conclusions of a data science project (Cat-pstone 1 and 2) aimed at developing a predictive model for distinguishing wool-sucking cats from non-suckers to enhance the cat adoption process. Cat-pstone 1 initially utilized a logistic regression model but later, in Cat-pstone 2, we explored the potential of incorporating deep learning techniques, specifically a multilayer perceptron (MLP) algorithm. The objective was to assess the feasibility and effectiveness of deep learning in improving predictive accuracy. This report outlines the approach, key findings, and recommendations based on the project's outcomes.

2. Approach

2.1 Data Preprocessing

During the initial modeling epoch, imputation of missing data was performed using the entire dataset, potentially leading to minor data leakage. Cat-pstone 2 addressed this issue by imputing missing data exclusively within the training set, ensuring data integrity and eliminating bias caused by the test set.

2.2 Feature Selection

To streamline the modeling process, several predictors were excluded, resulting in a reduced feature set of 25 predictors, 19 of which were binary breed group membership indicators. Factors were dropped based on low monotonic correlation with the target variable (a binarized set of suckers and non-suckers), low predictive power observed in the logistic regression model, and concerns about the validity of certain factors. For instance, the `Behavior_problem` factor was excluded due to doubts surrounding the content validity of the original questionnaire, as some respondents might interpret wool-sucking as a behavioral problem itself, while others may associate behavioral problems with aggression or destruction.

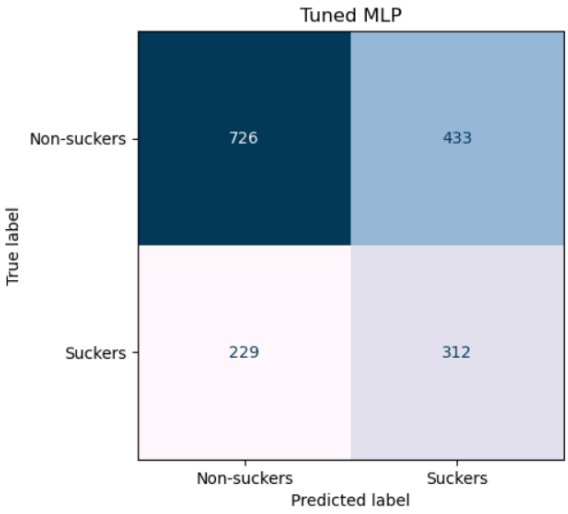
2.3 Model Selection

Given the limitations of a small dataset, a basic multilayer perceptron (MLP) algorithm was chosen as the deep learning approach for the project. The MLP model employs multiple layers of calculations to generate predictions, effectively categorizing samples as non-suckers or suckers based on the class with the highest vote.

3. Findings

3.1 Deep Learning vs. Logistic Regression

Deep learning algorithms excel at capturing complex hierarchical relationships and patterns that may elude traditional machine learning models. However, this complexity comes at the cost of efficiency. In this project, the MLP model was significantly slower than the logistic regression model, both during training and prediction stages, with training times nearly 80 times slower and prediction times over 130 times slower.

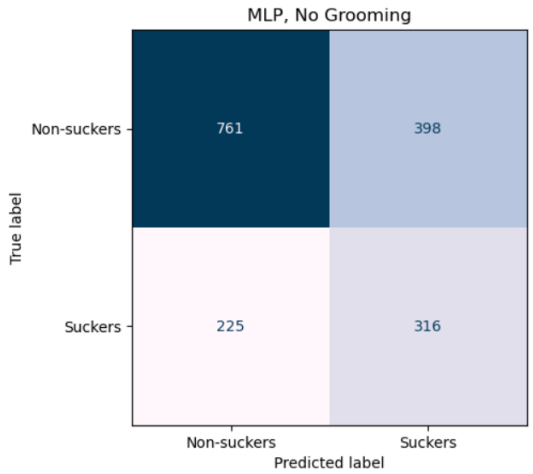


3.2 Class Imbalance

The logistic regression model, when trained with balanced class weights, did not require additional steps such as resampling the training data to address the issue of class imbalance in the target feature like the MLP model required. Although this step was not factored into the training time, it should be considered when comparing the performance of different models.

3.3 Grooming Feature

Surprisingly, the MLP model performed similarly to the logistic regression model, even without including the highly predictive grooming feature identified in Cat-pstone 1. When a similar experiment was conducted with logistic regression, using the same hyperparameters as Cat-pstone 1, the logistic regression model outperformed the one trained with grooming included.



4. Conclusion and Recommendations

After thorough experimentation and analysis, it is our recommendation that employing a deep learning model, such as the MLP algorithm explored in this project, would be impractical in this context. Despite the stakeholders' enthusiasm and the buzz surrounding deep learning, the MLP model exhibited performance comparable to the logistic regression model, while being significantly slower in training and prediction. Moreover, it required more preprocessing steps during training.

To enhance the cat adoption process, we suggest leveraging the well-performing logistic regression model with balanced class weights. Additional research and exploration can focus on refining the model's feature set. Further investigation into the grooming feature's impact on predictive accuracy and its inclusion/exclusion from the model is also recommended.

Ultimately, this project highlights the importance of evaluating various modeling approaches while considering their efficiency, interpretability, and practicality in real-world applications.

5. Acknowledgments

We extend our gratitude to the stakeholders and project team members for their valuable input, collaboration, and support throughout the duration of this data science project.