Classifying Cats: Exploring ML Models

Cat-pstone Project 2
Springboard Data Science Career Track Program

The goal

- Catering to BuddyTheCat's growing demand, we've created a binary classification system to identify cats as "suckers" or "non-suckers."
- Helps match individuals with feline companions that exhibit this adorable behavior.

Cat-pstone 1 findings

- Cat-pstone 1 used a logistic regression model with balanced class weights
- Breed and excessive grooming were highly predictive

Positive Predictive

Turkish Van, Angora

Saint Birman

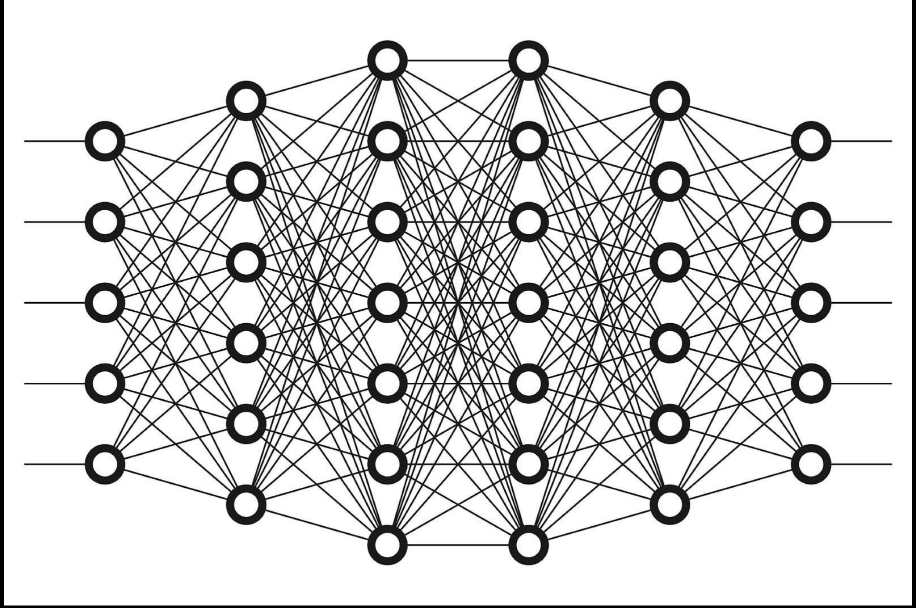
Balinese, Oriental L/S,
Seychellois L/S,
Siamese

Is deep learning really better than a more simple supervised ML model?

WHAT IS DEEP LEARNING?

WHAT IS DEEP LEARNING

- Subfield of machine learning
- Neural networks with multiple layers
- Automatic learning and hierarchical data representation extraction



WHY DEEP LEARNING

WHY DEEP LEARNING

PROS

- Captures more complex hierarchical patterns
- Can learn directly without extensive feature engineering
- Can handle non-linear relationships

CONS

- Requires more computational resources
- Less efficient
- Less interpretability

PREPPING THE DATA

PREPPING THE DATA: Imputing without Leakage

Cat-pstone 1: Imputation of missing data used entire dataset, which could result in minor data leakage.

Cat-pstone 2: Imputed missing data exclusively within the training set to maintain data integrity.

PREPPING THE DATA: Feature selection

Breed Group (One-hot encoded)

Neuter Status

(Excessive) Grooming

Shyness Novel

Shyness Strangers

Aggression Owners

Aggression Cats

Behavior Problem

Aggression Strangers

Outdoors

Contact People

Age

Activity Level

Other Cats

Gender

PREPPING THE DATA: Feature selection

Breed Group (One-hot encoded)

Neuter Status

(Excessive) Grooming

Shyness Novel

Shyness Strangers

Aggression Owners

Aggression Cats

Behavior Problem

Aggression Strangers

Outdoors

Contact People

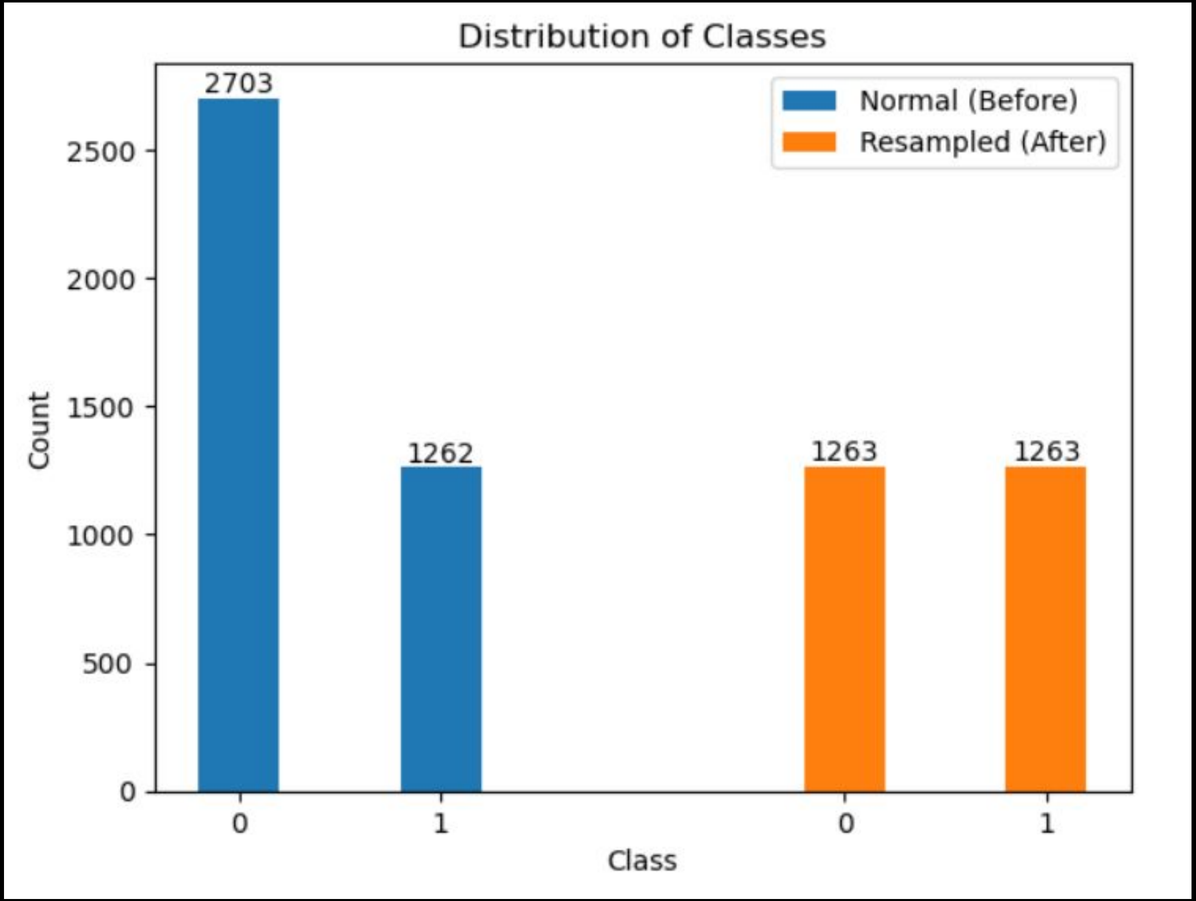
Age

Activity Level

Other Cats

Gender

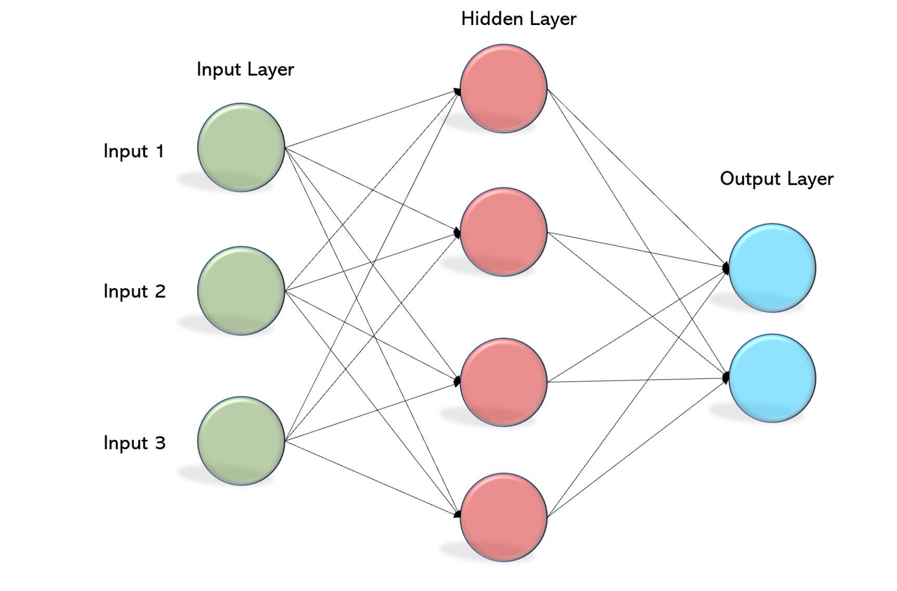
PREPPING THE DATA: Resampling



DEEP LEARNING MODEL SELECTION

DEEP LEARNING MODEL SELECTION: MLP

- Multilayer Perceptron (MLP)
- Flexible and adaptable, even with small datasets
- Multilayer design allows for hierarchical learning



LOGISTIC REGRESSION VS MLP

(spoiler: they perform *very* similarly)

LOGISTIC REGRESSION VS MLP: Speed

Compared to Logistic Regression, MLP is:

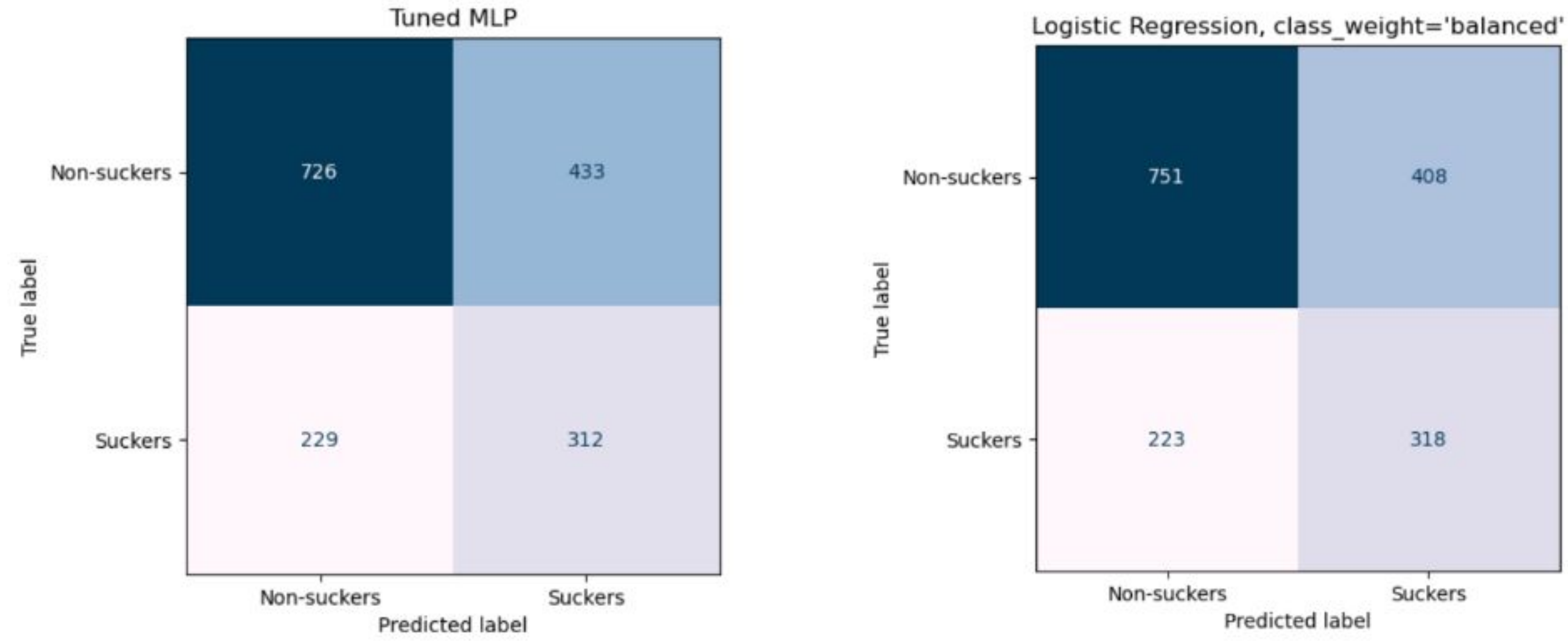
- 80x slower for training (not including resampling)
- 130x slower for predicting

LOGISTIC REGRESSION VS MLP: Grooming

How does MLP perform without the Grooming feature?

Context: Excessive grooming emerged as one of the most influential predictors in Cat-pstone 1.

LOGISTIC REGRESSION VS MLP: Grooming



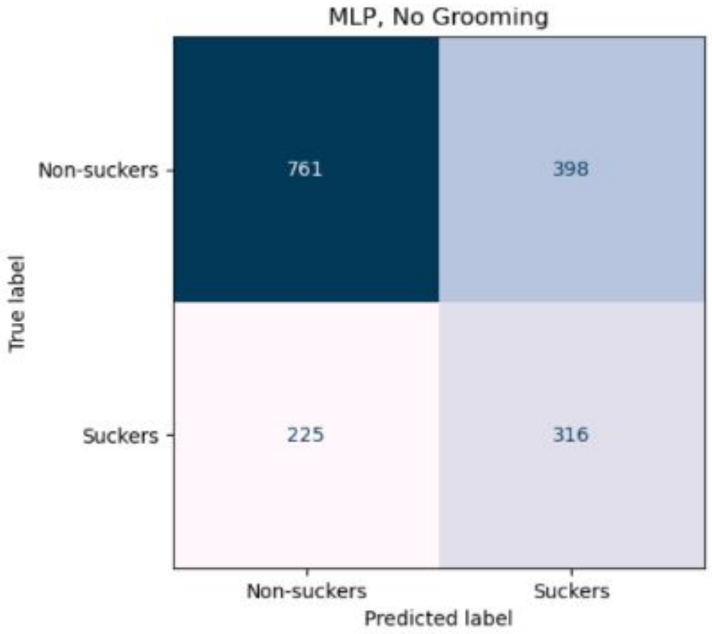
Tuned MLP

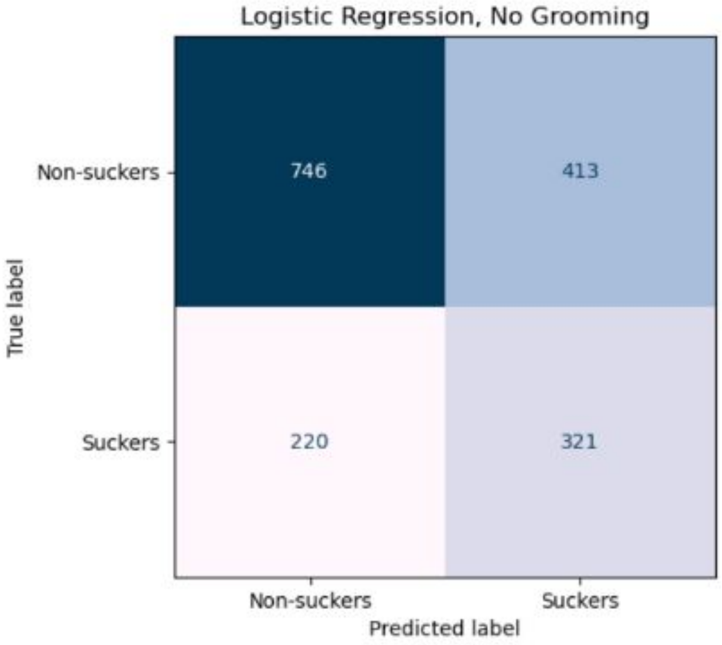
True label \ Predicted label	Non-suckers	Suckers
Non-suckers	726	433
Suckers	229	312

Logistic Regression, class_weight='balanced'

True label \ Predicted label	Non-suckers	Suckers
Non-suckers	751	408
Suckers	223	318

LOGISTIC REGRESSION VS MLP: No Grooming





RECAP AND RECOMMENDATIONS

RECAP AND RECOMMENDATIONS

Deep learning (MLP) is impractical for this use-case; logistic regression is preferred due to its efficiency.

Logistic regression is faster in data preparation, model fitting, and prediction.

Refine the model's feature set through further research and and analyze the grooming feature's influence.