Predicting Dog Trainability to Mitigate Bite Incidents in New York City

Penggao Gu
GitHub Repository

December 15, 2024

1 Introduction

1.1 Problem Statement

Dog bite incidents are a significant concern in areas like New York City, with potentially severe physical and emotional consequences for victims. According to a 2020 study published in *Injury Epidemiology*, over 4.5 million dog bites occur annually in the United States, with children and elderly individuals disproportionately affected. In New York City, dog bite incidents often lead to emergency department visits and pose a challenge to public safety [1]. Understanding the trainability of dogs involved in these incidents can aid in identifying high-risk animals and designing targeted training programs. Trainability, as a continuous variable, reflects a dog's ability to respond to training efforts and is crucial for mitigating future bite incidents and enhancing public safety.

1.2 Dataset Description

The dataset used in this project combines information from two reliable sources:

- **National Electronic Injury Surveillance System (NEISS)**: Provides comprehensive records of dog bite incidents reported by emergency departments. Key details include dog characteristics (e.g., breed, size, and age), incident context (e.g., location and time), and victim demographics.
- **American Kennel Club (AKC)**: Evaluates dog traits, including trainability, using breed-based criteria and expert observations. These evaluations provide a general measure of trainability, making them suitable as a target variable for this analysis.

The combined dataset includes 12 variables describing dog bite incidents. Key columns include `Breed`, `Gender`, `SpayNeuter`, `Season`, and `Age_Range`, which provide information about the dogs and the context of the incidents. Additional columns like `group` categorize dogs further, while `trainability_value` serves as the target variable, representing a continuous score of a dog's trainability. Missing values are present in `Age_Range`, requiring careful preprocessing to ensure data quality.

1.3 Previous Work

Extensive research has been conducted to understand the factors contributing to dog bite incidents. Studies published in journals such as the *Journal of Veterinary Behavior* and *Injury Epidemiology* have highlighted the overrepresentation of certain breeds in bite statistics, including Bull Terriers and German Shepherds, often associated with higher bite incidents due to their physical strength and behavioral tendencies [2, 3]. Seasonal changes and crowded urban settings have also been identified as key environmental factors influencing dog behavior. These studies provided a foundational understanding of the interplay between breed-specific traits, environmental conditions, and behavior, underscoring the importance of further exploration into trainability as a predictor of behavior.

2 Exploratory Data Analysis (EDA)

2.1 Overview of the Dataset

The dataset comprises 13,299 records of dog bite incidents in New York City, with 12 features describing attributes of the dogs, incident context, and the target variable `trainability_value`. Each feature provides specific insights:

- **Breed**: Categorizes dogs by breed, such as Bull Terrier and Pit Bull, which appear frequently.
- **Gender**: Binary classification of dogs as male or female.
- **SpayNeuter**: Indicates whether the dog has been spayed or neutered.
- **Borough**: Geographic identifiers for where the incidents occurred.
- **Season**: Categorical variable representing the time of year the incident occurred (e.g., Summer, Winter).
- **Age_Range**: Represents dog age grouped into `Puppy/Young`, `Adult`, and `Senior`, but includes missing values that require imputation.
- **Group**: Categorical identifiers grouping dog types; the `Group` feature was later dropped due to collinearity.
- **trainability_value**: A continuous float variable ranging from 0 to 1, representing the trainability score.

2.2 Key Findings

2.2.1 Breed Distribution

The dataset reveals a highly disproportionate distribution of dog breeds. Breeds such as Pit Bulls and Bull Terriers appear far more frequently, accounting for a significant portion of bite incidents.

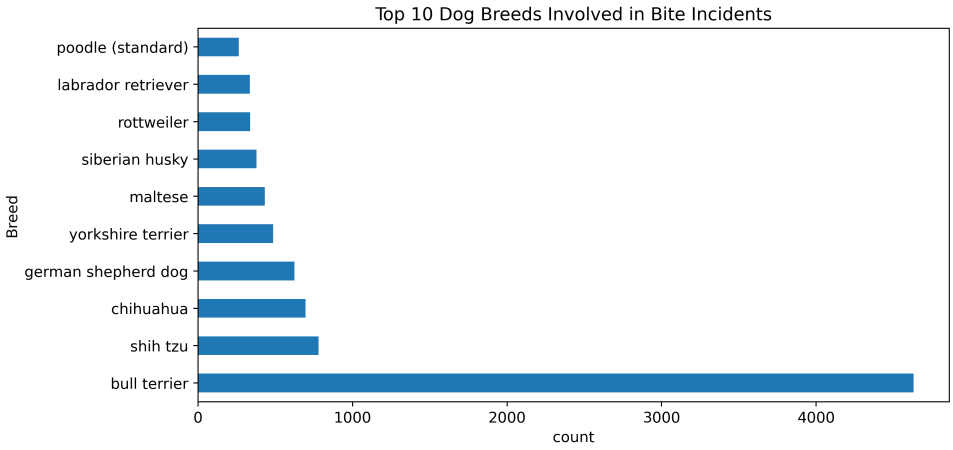

Figure 1: Distribution of Dog Breeds in the Dataset

2.2.2 Trainability Value Distribution

The target variable, `trainability_value`, displays a bimodal distribution with peaks around specific values, potentially indicating systematic patterns in scoring or evaluation biases.

- **Insight**: This bimodal trend highlights the need to investigate whether external factors, such as breed or incident severity, influence the scoring process.

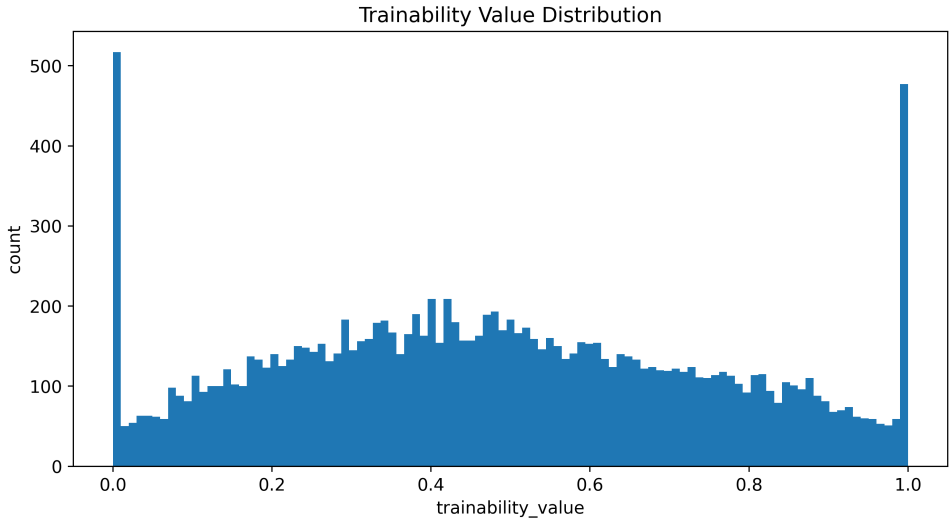

Figure 2: Distribution of Trainability Values

2.2.3 Seasonal Trends

Dog bite incidents fluctuate across seasons, with a notable spike observed in summer months. This trend aligns with increased outdoor activity and higher exposure to unfamiliar environments or individuals.

- **Insight**: Seasonality is an important feature that may influence dog behavior and incident frequency.

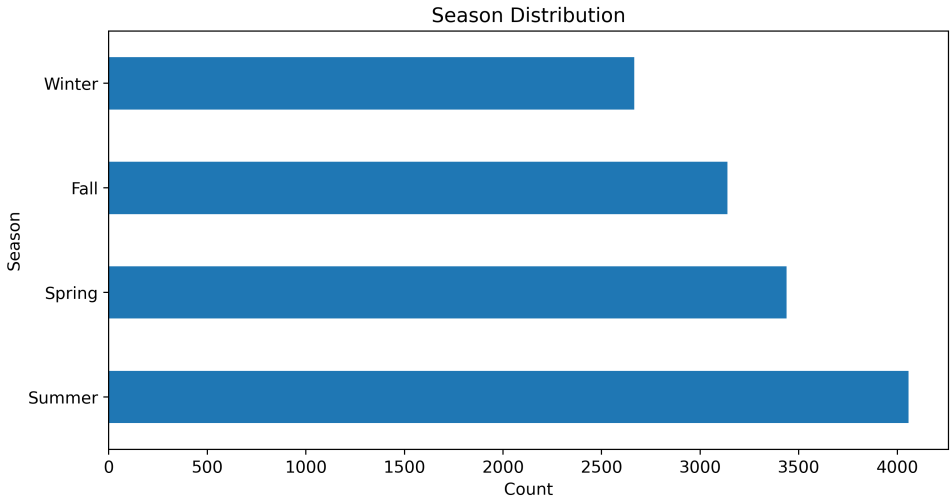

Figure 3: Seasonal Trends in Dog Bite Incidents

2.2.4 Demographic Analysis

Gender: The dataset shows an approximately even split between male and female dogs, suggesting no significant gender imbalance in bite incidents.

Spay/Neuter Status: Unneutered dogs are overrepresented in bite incidents, supporting findings in behavioral research that link intact dogs to higher aggression.

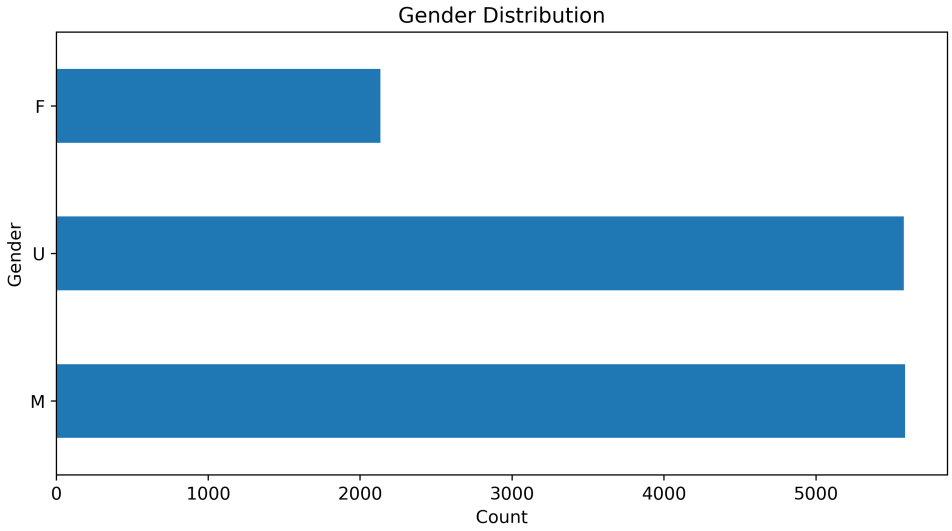

Figure 4: Gender Trends in Dog Bite Incidents

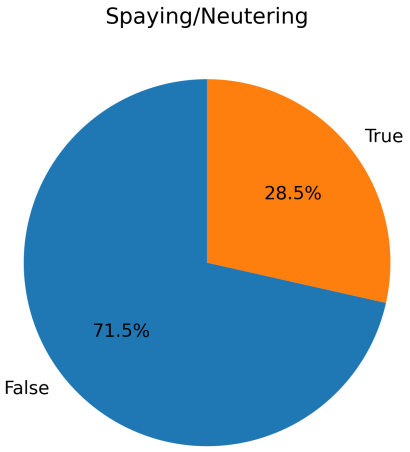

Figure 5: Analysis of Spay/Neuter Status

3 Methods

3.1 Splitting Strategy

The dataset was divided into three subsets:

- **Training Set (60%)**: Used for model training.
- **Validation Set (20%)**: Used for hyperparameter tuning.
- **Testing Set (20%)**: Used for final evaluation of model performance.

Since this is a regression task, random splitting was employed to ensure a fair distribution of data while preserving the continuous nature of the target variable, `trainability_value`.

During model training, K-Fold cross-validation was applied to further assess model robustness and mitigate the impact of random splits on performance. This approach minimizes overfitting and ensures a robust foundation for model evaluation.

3.2 Data Preprocessing

Data preprocessing was implemented using a pipeline to streamline transformations and ensure reproducibility. The following steps were included:

1. **Feature Selection**: The `group` variable was removed due to high collinearity with `breed`, which provides more granular information.
2. **Handling Missing Values**: Missing values in categorical variables (e.g., `age_group`) were imputed with a placeholder value (`"None"`). This ensured no data was lost during preprocessing and distinguished missing values from valid categories. However, this choice may influence the model if missing values are not uniformly distributed.
3. **Feature Encoding**:
   - Ordinal encoding was applied to `age_group`.
   - One-hot encoding was applied to other categorical variables.

3.3 Machine Learning Pipeline

The machine learning pipeline was designed to evaluate and compare multiple models to predict the target variable (`trainability_value`). The pipeline included the following steps:

1. **Baseline Comparison**: The mean of the target variable (`trainability.value`) was used as a baseline for evaluation. This simple approach serves as a benchmark to ensure that more complex models offer meaningful improvements. No hyperparameters were involved.
2. **Ridge Regression**: A regularized linear model designed to reduce overfitting. The best hyperparameter was found to be `alpha = 1.0`, which controls the strength of the L2 regularization to balance bias and variance.
3. **Lasso Regression**: Lasso Regression performs feature selection by shrinking coefficients of irrelevant features to zero, improving interpretability. The best hyperparameter was determined as `alpha = 0.001`, which minimizes overfitting while maintaining predictive performance.
4. **Random Forest**: A tree-based ensemble method that captures complex feature interactions. The optimal hyperparameters were:
   - `max_depth = 30` (maximum depth of the trees), and
   - `max_features = 0.5` (fraction of features considered at each split).
5. **XGBoost**: A powerful gradient boosting algorithm known for its high predictive power and efficiency. The best hyperparameters include:
   - `max_depth = 10` (maximum depth of the trees),
   - `n_estimators = 300` (number of boosting rounds),
   - `reg_alpha = 0.001` (L1 regularization term), and
   - `reg_lambda = 100.0` (L2 regularization term).
6. **Support Vector Regression (SVR)**: A model effective for high-dimensional data. The optimized hyperparameters were:
   - `C = 100.0` (regularization parameter), and
   - `gamma = 0.001` (kernel coefficient for the RBF kernel).

The details of the models and their hyperparameters are summarized in Table 1.

3.4 Best Hyperparameters Summary

The following table summarizes the best hyperparameters obtained for each machine learning model. These hyperparameters were selected based on grid search or validation performance.

Table 1: Summary of Machine Learning Models and Hyperparameters

Model	Description	Hyperparameters
Baseline Comparison	Mean of the target variable (`trainability.value`).	None
Ridge Regression	Regularized linear model to reduce overfitting.	`alpha`: {0.001, 0.01, 0.1, 1, 10, 100}
Lasso Regression	Performs feature selection by shrinking coefficients to zero.	`alpha`: {0.001, 0.01, 0.1, 1, 10, 100}
Random Forest	Captures complex feature interactions.	`max_depth`: {1, 3, 10, 30, 100} `max_features`: {0.5, 0.75, 1.0}
XGBoost	Gradient boosting for enhanced predictive power.	`reg_alpha`: {0, 0.001, 0.01, 0.1, 1} `reg_lambda`: {0, 0.01, 0.1, 1, 10, 100} `max_depth`: {1, 10, 100, 1000} `n_estimators`: {100, 200, 300}
Support Vector Regression	Handles high-dimensional data effectively.	`C`: {0.1, 1, 10, 100, 1000, 10000, 100000, 1000000} `gamma`: {0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1}

Insights

- **Ridge Regression** achieved optimal performance with a regularization strength of `alpha` = 1.0, balancing bias and variance.
- **Lasso Regression**, which performs feature selection, favored a smaller regularization strength (`alpha` = 0.001).
- **Random Forest** and **XGBoost** both captured complex relationships effectively. XGBoost required higher regularization (`reg_alpha`, `reg_lambda`) to prevent overfitting.
- **SVR** (Support Vector Regression) worked best with a higher margin tolerance (`C` = 100.0) and a small kernel parameter (`gamma` = 0.001).

Table 2: Best Hyperparameters for Machine Learning Models

Model	Best Hyperparameters
Ridge Regression	`alpha` = 1.0
Lasso Regression	`alpha` = 0.001
Random Forest	`max_depth` = 30, `max_features` = 0.5
XGBoost	`max_depth` = 10, `n_estimators` = 300, `reg_alpha` = 0.001, `reg_lambda` = 100.0
SVR	`C` = 100.0, `gamma` = 0.001

The hyperparameters reflect the need for balancing model complexity, regularization, and fitting capacity across different algorithms.

3.5 Evaluation Metric

Mean Squared Error (MSE) was used as the primary metric for evaluating model performance. MSE penalizes larger errors more than smaller ones, making it suitable for regression tasks where precision is critical. Metrics were averaged across K-Fold splits to ensure robustness.

3.6 Uncertainty Analysis

To measure uncertainty, models were trained and evaluated across five random splits of the dataset using different random seeds. For each split, 5-Fold cross-validation was applied during training to ensure robust evaluation and tuning, and the variability in MSE across folds and random states was recorded to assess model robustness. For non-deterministic models such as Random Forest and XGBoost, results were averaged over multiple runs to account for randomness in tree construction and boosting iterations. This approach provided a comprehensive view of model stability and generalization.

4 Results

4.1 Test Score Comparison

We evaluated multiple models against a baseline Mean Squared Error (MSE) of 0.07478, which was calculated using the mean of the target variable as predictions. The following observations summarize model performance:

- **Baseline Comparison**: All models significantly outperformed the baseline MSE of 0.07478.
- **Best Performing Models**: Ridge Regression achieved the lowest average Test MSE (0.02905), closely followed by Support Vector Regression (SVR) with an MSE of 0.02930.
- **Model Stability**: Ridge Regression and SVR exhibited the lowest variance across random states, highlighting their robustness.
- **Final Model Selection**: Ridge Regression was selected as the final model due to its consistent performance, low variance, and interpretability.

Global feature importance was assessed using three complementary methods: **Permutation Importance**, **SHAP Values**, and **Ridge Coefficients**. These methods consistently identified Breed and Season as dominant predictors.

4.1.1 Model Performance and Uncertainties

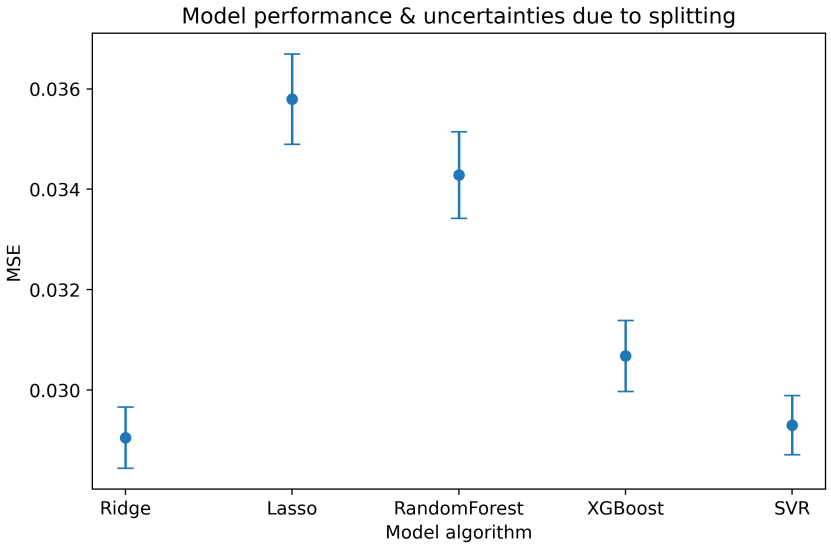

Figure 6: Model performance and uncertainties due to splitting.

4.1.2 Permutation Importance

Permutation importance measures the increase in MSE when each feature is shuffled. The results revealed the following top features:

- **Breed**: The most influential feature, increasing the MSE to approximately 0.11 when perturbed.
- **Season**: Particularly `Season_Summer`, which showed moderate importance with an MSE increase of approximately 0.04.
- **Age_Range** and **Borough**: Exhibited moderate importance, but much lower compared to Breed and Season.
- Features like **Spay/Neuter** and **Gender** had negligible impacts on the model's predictions.

4.1.3 SHAP Values (Global)

SHAP values provided a detailed breakdown of each feature's contribution at the global level:

- **Breed_Bull Terrier**: The most significant negative impact on trainability scores.
- **Season_Summer**: Consistently lowered trainability predictions.
- **Breed_Yorkshire Terrier**: Another impactful feature, though with a smaller magnitude.
- Features such as **Spay/Neuter** and **Gender** were consistently low in importance.

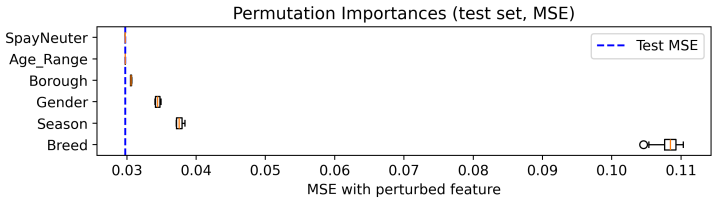

Figure 7: Global Permutation Importance of Features

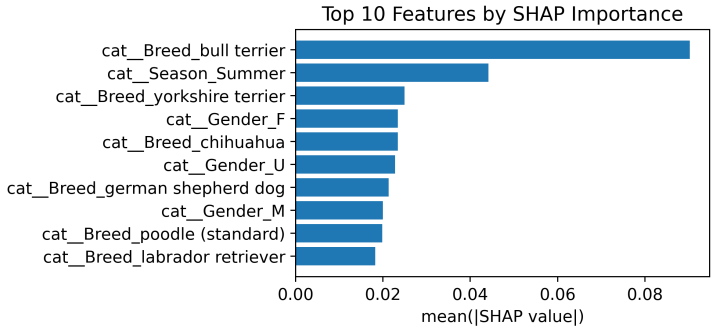

Figure 8: SHAP Values for Global Feature Importance

4.1.4 Ridge Coefficients

Ridge coefficients confirmed the dominance of Breed and Season when features were standardized (mean = 0, std = 1):

- **Breed_Bull Terrier**: -0.07399 (largest negative effect).
- **Breed_German Shepherd Dog**: 0.07135 (positive effect).
- **Season_Summer**: -0.04633 (moderate negative effect).
- Other breeds, such as **Yorkshire Terrier** and **Labrador Retriever**, also showed significant contributions.
- Low-coefficient features like **Spay/Neuter** and **Gender** corroborated their minimal impact.

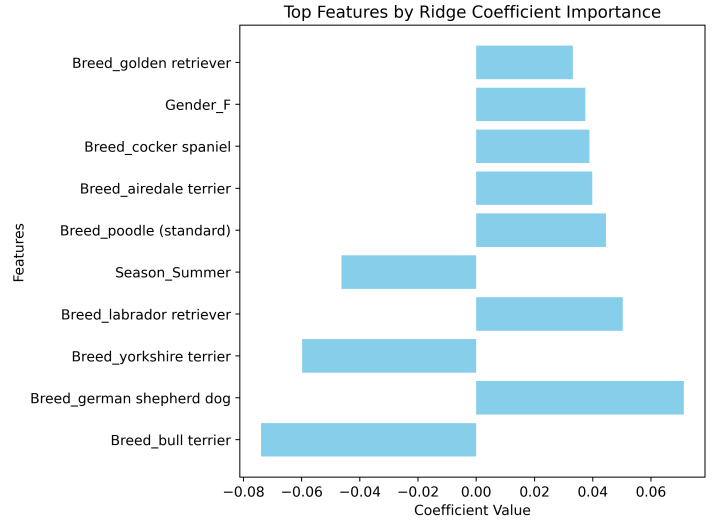

Figure 9: Ridge Regression Coefficients

4.2 Local Feature Importance

Local feature importance was analyzed using SHAP values for individual predictions, revealing contextual interactions between features.

4.2.1 Example 1: High Trainability Prediction (Index 80)

- **Positive Contributions**: Gender_Female, Breed_Portuguese Water Dog.
- **Negative Contributions**: The absence of Breed_Bull Terrier and `Season_Summer` mitigated their usual negative impact.

Figure 10: Ridge Regression Coefficients

4.2.2 Example 2: Unexpected Low Trainability Prediction (Index 1)

- **Positive Contributions**: Breed_Chihuahua and Gender_Undefined provided minor positive influence.
- **Negative Contributions**: Breed_Bull Terrier and `Season_Summer` significantly reduced the predicted trainability score.

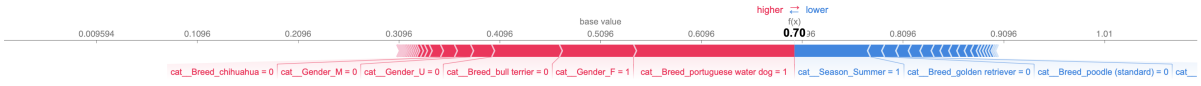

Figure 11: Ridge Regression Coefficients

These examples demonstrate that even globally low-importance features, such as Gender, can occasionally impact specific predictions, highlighting the importance of contextual analysis.

4.3 Interesting Findings

While `Breed_Bull Terrier` consistently reduced trainability predictions, its effect could be offset under favorable conditions (e.g., certain genders or seasonal contexts).

Environmental factors like `Season_Summer` had stronger-than-expected influence, suggesting actionable interventions to manage behavioral outcomes.

These results provide clear and actionable insights for improving trainability, such as focusing on breed-specific strategies and addressing seasonal effects.

4.4 Insights from Best Performing Model: Ridge Regression

The scatter plot of true vs. predicted values for the Ridge regression model provides key insights into its performance:

- **Overall Trends:** Most points align reasonably well along the diagonal, suggesting that the Ridge regression model effectively captures the overall trends in the data. This demonstrates that regularization has mitigated overfitting and improved generalization.
- **Deviations and Clustering:** Despite the strong alignment, there are noticeable deviations and clustering patterns. These may be due to the categorical nature of the training features, which can introduce systematic biases or limit the model's flexibility in capturing more complex relationships.
- **Performance Implications:** The alignment along the diagonal indicates the model's robustness, while the deviations highlight opportunities for further refinement. Incorporating additional feature engineering or transforming categorical features could help address the observed clustering.

The Ridge regression model thus serves as a reliable baseline, balancing bias and variance effectively while providing interpretable predictions.

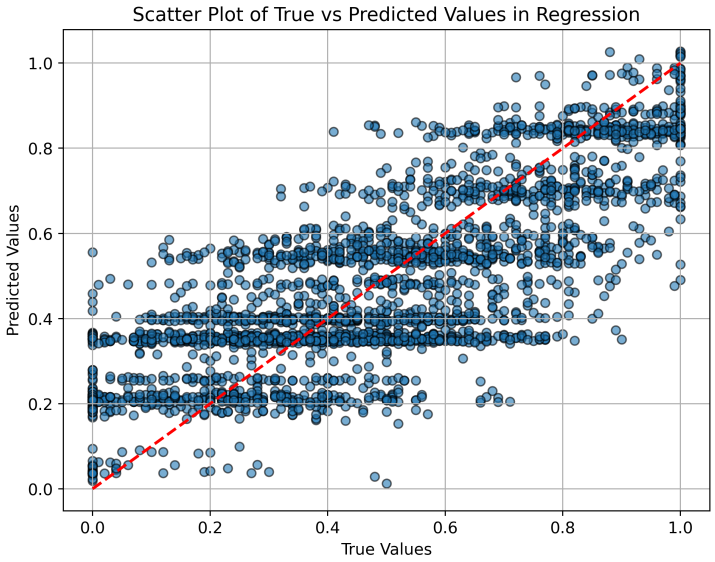

Figure 12: Ridge Regression Coefficients

5 Outlook

While the current model effectively predicts dog trainability and identifies key factors contributing to bite incidents, several improvements and opportunities for future work remain to enhance both the model's performance and interpretability.

5.1 Collaboration with Domain Experts

Incorporating insights from dog behavior experts, trainers, and veterinarians could add significant value to the model. Collaboration with professionals would enable qualitative data collection and additional feature engineering, including expert-defined traits such as:

- Temperament and behavioral triggers.
- Prior training exposure or intervention history.

These expert insights could provide a deeper understanding of breed-specific behaviors, which the current dataset may oversimplify.

5.2 Expanding Feature Dimensions

To improve model granularity and predictive performance, additional features can be considered:

- **Behavioral Data**: Incorporating dimensions such as aggression scores, bite severity ratings, and history of prior incidents.
- **Owner Data**: Adding ownership information, including training frequency, owner experience, and household environment, to explain unexpected predictions.
- **Geospatial and Environmental Factors**: Enhancing the dataset with weather data, park proximity, and urban density metrics to refine insights into location-based behavioral variations.

5.3 Improving Model Interpretability

While global and local SHAP analyses provided valuable insights, additional explainability tools could further clarify feature relationships and model behavior:

- **Partial Dependence Plots (PDPs)**: To illustrate how individual features impact predictions.
- **Individual Conditional Expectation (ICE) Plots**: To reveal feature-specific interactions for individual predictions.

These tools would aid stakeholders, such as policymakers and dog owners, in understanding how specific features influence trainability.

5.4 Future Data Collection

To strengthen predictive accuracy and expand model capabilities, future data collection efforts should focus on:

- **Training Program Data**: Information about dogs that underwent specific training regimens and their outcomes.
- **Longitudinal Data**: Tracking dogs over time to observe changes in trainability and behavior post-training or after spaying/neutering.
- **Incident Context**: Richer contextual data surrounding dog bites, such as the presence of children, loud noises, or unfamiliar individuals.

5.5 Conclusion

By integrating domain expertise, expanding feature dimensions, and improving feature interpretability, the model can evolve into a more robust and actionable tool for predicting trainability. These enhancements would not only improve predictive performance but also provide actionable insights for reducing dog bite incidents and promoting safer urban environments.

References

[1] Tuckel, P.; & Milczarski, W. The Changing Epidemiology of Dog Bite Injuries in the United States, 2005–2018. **2020**. *Injury Epidemiology.*

[2] Essig, G. F., Sheehan, C., Rikhi, S., Elmaraghy, C. A., & Christophel, J. J. (2019). Dog bite injuries to the face: Is there risk with breed ownership? A systematic review with meta-analysis. *International Journal of Pediatric Otorhinolaryngology*, 117, 182-188.

[3] Bini, J. K., Cohn, S. M., Acosta, S. M., McFarland, M. J., & Muir, M. T. (2011). Mortality, mauling, and maiming by vicious dogs. *Annals of Surgery*, 253(4), 791-797.