Predicting Avian Patterns

The Bird Brains

Lynn Jacobs

Sri Moukthika Aluri

Jeffrey Cris

Gayathri Gandham

Dec 8, 2025

Abstract

Bird presence patterns provide valuable insight into habitat conditions and ecological change, and large citizen-science datasets offer an opportunity to study these patterns at scale. In this project, we used data from the Cornell Lab of Ornithology's Project FeederWatch to explore whether the presence of mallards at backyard feeder sites can be predicted using site characteristics, habitat features, and observation effort. Our initial goal was to build a model capable of predicting the presence of all bird species in the dataset; however, due to the complexity and variability across species, we refined the scope to focus on a single species, mallard ducks, so that we could develop a clear and reliable modeling pipeline. After cleaning and preprocessing the data, we applied several supervised learning methods, including linear, tree-based, and distance-based models, and evaluated them using metrics appropriate for imbalanced classification. The results showed that tree-based models captured the relationships in the data most effectively. Overall, the project demonstrates that machine learning can successfully estimate mallard presence from feeder site information and highlights how citizen-science data can support ecological understanding and species-distribution modeling.

Table of Contents

CHAPTER 1

Introduction

1.1 Background

Understanding bird presence and distribution is an important component of ecological monitoring, habitat management, and conservation planning. Citizen-science platforms such as eBird and FeederWatch collect large volumes of bird observation data that, when combined with detailed site-level habitat information, provide opportunities to model species presence patterns with machine learning. Predictive models built using such data can help identify which environmental features most strongly influence the occurrence of a species, offering insights for habitat design, conservation strategies, and ecological forecasting.

This project investigates the presence of the mallard (mallar3), a widely distributed waterfowl species whose occurrence can be influenced by site characteristics, habitat structure, and human activity. By integrating observational records with environmental and site-level variables, the project applies modern data-driven methods to evaluate how habitat features contribute to predicting mallard presence.

1.2 Problem Statement

Although large amounts of observational bird data exist, ecological datasets are often fragmented across multiple tables and vary in structure, scale, and completeness. This makes it challenging to determine which habitat and environmental characteristics are most relevant for predicting the presence of specific species.

We asked a core question in this project: how accurately can mallard presence be predicted using a combination of observational bird data and site-level habitat features?

1.3 Aims and Objectives

Aim

To develop and evaluate machine learning models that predict the presence of the mallard (mallar3) using integrated observational and site habitat data.

Objectives

- To clean, preprocess, and merge observational and site-level datasets into a unified analytical dataset.
- To identify relevant ecological and habitat features that contribute to the prediction of mallard presence.

- To train and compare multiple machine learning models (Logistic Regression, Decision Tree, Random Forest).
- To evaluate model performance using accuracy, precision, recall, and related metrics.
- To interpret key predictors and assess the ecological relevance of the results.

1.4 Solution Approach

The project follows a structured data science methodology:

- **Data Preparation:** Standardizing column names, handling missing values, merging observations with site-level habitat data, and creating a binary presence variable for *mallar3*.
- **Feature Engineering:** Selecting meaningful predictors such as location, yard type, vegetation characteristics, feeding activity, and presence of humans/animals.
- **Model Development:** Implementing Logistic Regression, Decision Tree, and Random Forest classifiers to predict presence.
- **Model Evaluation:** Comparing performance across models using accuracy, recall, precision, F1-score, and ROC curve analysis.
- **Interpretation:** Identifying which habitat features most strongly influence mallard presence based on model outputs and feature importance.

1.4.1 Dataset Source

The dataset used in this project was obtained from **Project FeederWatch**, an established citizen-science program managed by the Cornell Lab of Ornithology. Project FeederWatch collects winter bird observations submitted by volunteers across North America through the platform available at *feederwatch.org*. Participants record bird species, counts, and observation effort at their home feeders or local sites, while also providing detailed information about habitat characteristics such as yard type, vegetation, presence of water sources, and supplemental feeding activity.

The data is structured into multiple tables, including an **observational dataset** containing species-level sighting records and a **site dataset** containing environmental and habitat variables associated with each location.

1.4.2 Dataset Integration

Observational data and site habitat data were merged based on location and project period identifiers. Both datasets required cleaning, standardization, and validation before integration.

1.4.3 Modeling Strategy

Models were trained using an 80–20 train-test split. Due to imbalance in the target variable, a resampler called SMOTE was used to balance the training data.

1.5 Summary of Contributions and Achievements

- Successfully merged two heterogeneous datasets into a single modeling dataset.
- Implemented three predictive models and demonstrated that **Random Forest outperformed the others**, capturing complex interactions among habitat features.

- Identified key predictors of mallard presence, including proximity to water, wooded areas, vegetation density, supplemental feeding, and human/animal activity.
- Highlighted methodological considerations such as class imbalance and feature selection in ecological modeling.

CHAPTER 2

Methodology

2.1 Data Cleaning

2.1.1 Overview of the Datasets

This study relied on two primary datasets collected between 2020 and 2024. The first dataset, the Observer Table, contains approximately one million observations along with metadata such as date, time, effort, and additional contextual details. The second dataset, the Site Description Table, consists of roughly thirty thousand records describing the habitat characteristics of each site, including yard type, vegetation, feeder availability, animal presence, and environmental attributes. Because both datasets contained a large number of columns, it is not practical to describe every variable in detail. Instead, a representative subset of important features is highlighted below to illustrate the types of attributes included in the analysis.

Feature Name	**Description**
yard_type_landsca	Landscape-type yard
yard_type_woods	Yard dominated by woods
hab_dcid_woods	Deciduous woods habitat
hab_mixed_woods	Mixed woods habitat
hab_water_fresh	Freshwater habitat presence
hab_residential	Residential habitat type
evgr_trees_atleast	Evergreen trees score
evgr_shrbs_atleast	Evergreen shrubs score
dcid_trees_atleast	Deciduous trees score
dcid_shrbs_atleast	Deciduous shrubs score
brsh_piles_atleast	Brush pile availability
bird_baths_atleast	Bird bath availability

nearby_feeders	Presence of feeders nearby
squirrels	Squirrel activity (binary)
dogs	Dog activity (binary)
humans	Human activity (binary)
housing_density	Ordinal residential density (1–4)
fed_yr_round	Year-round feeding indicator
count_area_size_sq_m_atleast	Observation area size score
supp_food	Supplemental food availability

The combination of these features provided a comprehensive representation of both environmental and anthropogenic factors.

2.1.2 Cleaning the Observer Table

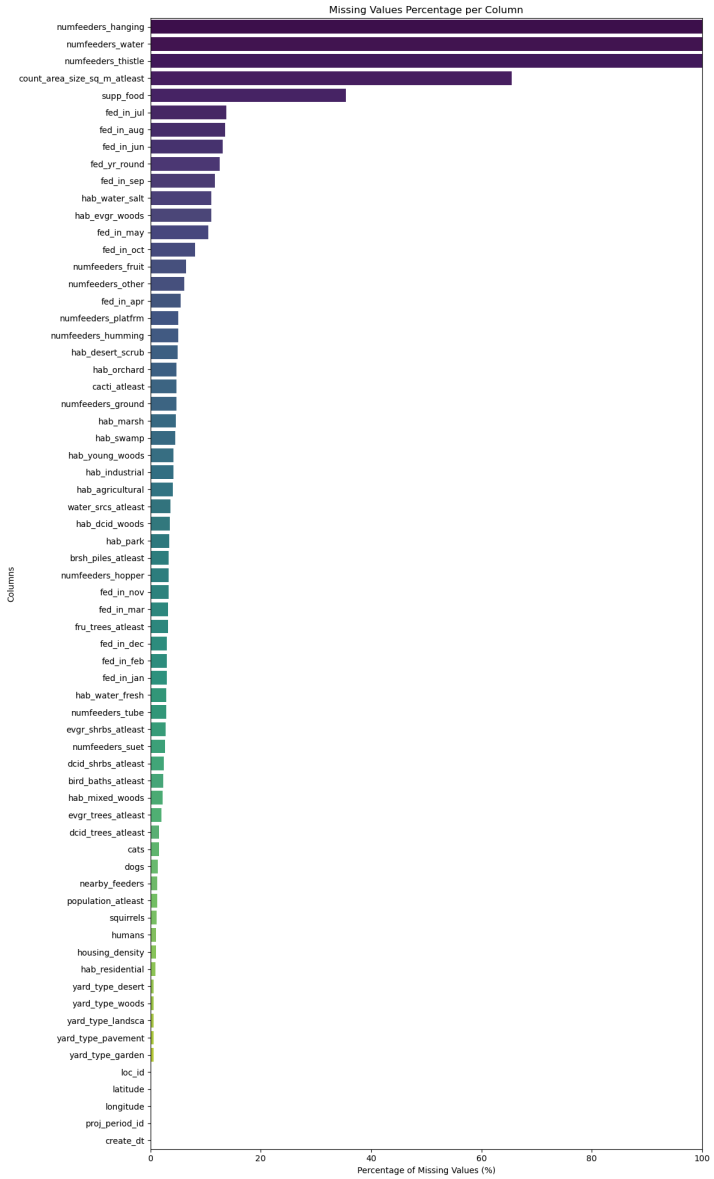
The first step in the cleaning process involved addressing data quality issues within the Observer Table. Initial inspection of missing values revealed several concerns regarding data reliability. Any rows for which both the **VALID** and **REVIEWED** flags were set to zero were removed, as these entries represented observations that raised concerns but were never checked by reviewers. Keeping such records would risk introducing noise and potentially diminishing model performance.

Following this, columns with excessively high missingness were evaluated. Those with more than 70% missing values, specifically *PLUS_CODE* and *alt_full_spp_code*, were removed because they contained insufficient information to support reliable analysis. Additional administrative fields such as *data_entry_method*, *create_dt*, *VALID*, and *REVIEWED* were also removed since they do not contribute to the model's predictive ability.

Some variables with moderate missingness were retained and imputed appropriately. For example, *SNOW_DEP_ATLEAST*, a numeric variable representing snow depth, exhibited a distribution heavily skewed toward zero. Because of this skewness, median imputation was applied to preserve central tendency without being influenced by outliers. After completing these steps, the Observer Table was left with a more consistent and analytically useful structure.

2.1.3 Cleaning the Site Description Table

The Site Description Table required a different cleaning approach due to its temporal structure and the variety of categorical, ordinal, and count-based fields. Since the objective of this study was limited to the years **2020–2024**, all records from earlier years were removed. This filtering step significantly reduced irrelevant variation and ensured temporal consistency.



As with the Observer Table, variables with greater than 70% missingness were removed. These included *numfeeders_thistle*, *numfeeders_water*, and *numfeeders_hanging*. Administrative fields such as *create_dt* were also discarded due to their lack of analytical relevance.

A structured imputation strategy was then applied based on the underlying data type of each feature. Binary indicators, such as those describing habitat presence (*hab_evgr_woods*, *yard_type_pavement*) or animal activity (*squirrels*, *cats*), had low missingness and were imputed using the mode, which best reflects the dominant category. Ordinal variables such as *evgr_trees_atleast*, *dcid_shrbs_atleast*, and *bird_baths_atleast*, which use a discrete scale (0, 1, 3, 4, 11), were imputed using the median to preserve category order without distorting distributional structure. Similarly, *housing_density*, an ordinal variable with values ranging from 1 to 4, was imputed using the median for the same reasons.

Count-based variables describing feeder availability (e.g., *numfeeders_suet*, *numfeeders_ground*) were also imputed using the median because their distributions were right-skewed with many zero values. In some special cases, such as *population_at_least*, the variable contained typically large values but only

minimal missingness, so median imputation was again the appropriate method. Variables with moderate missingness, such as *supp_food* at approximately 35%, were retained and imputed using the mode. Even variables with higher missingness, such as *count_area_size_sq_m_atleast* at roughly 65%, were retained because their missing rates fell below the 70% threshold, and were imputed using the median. After completing these imputation procedures, the Site Description Table became consistent, complete, and suitable for merging.

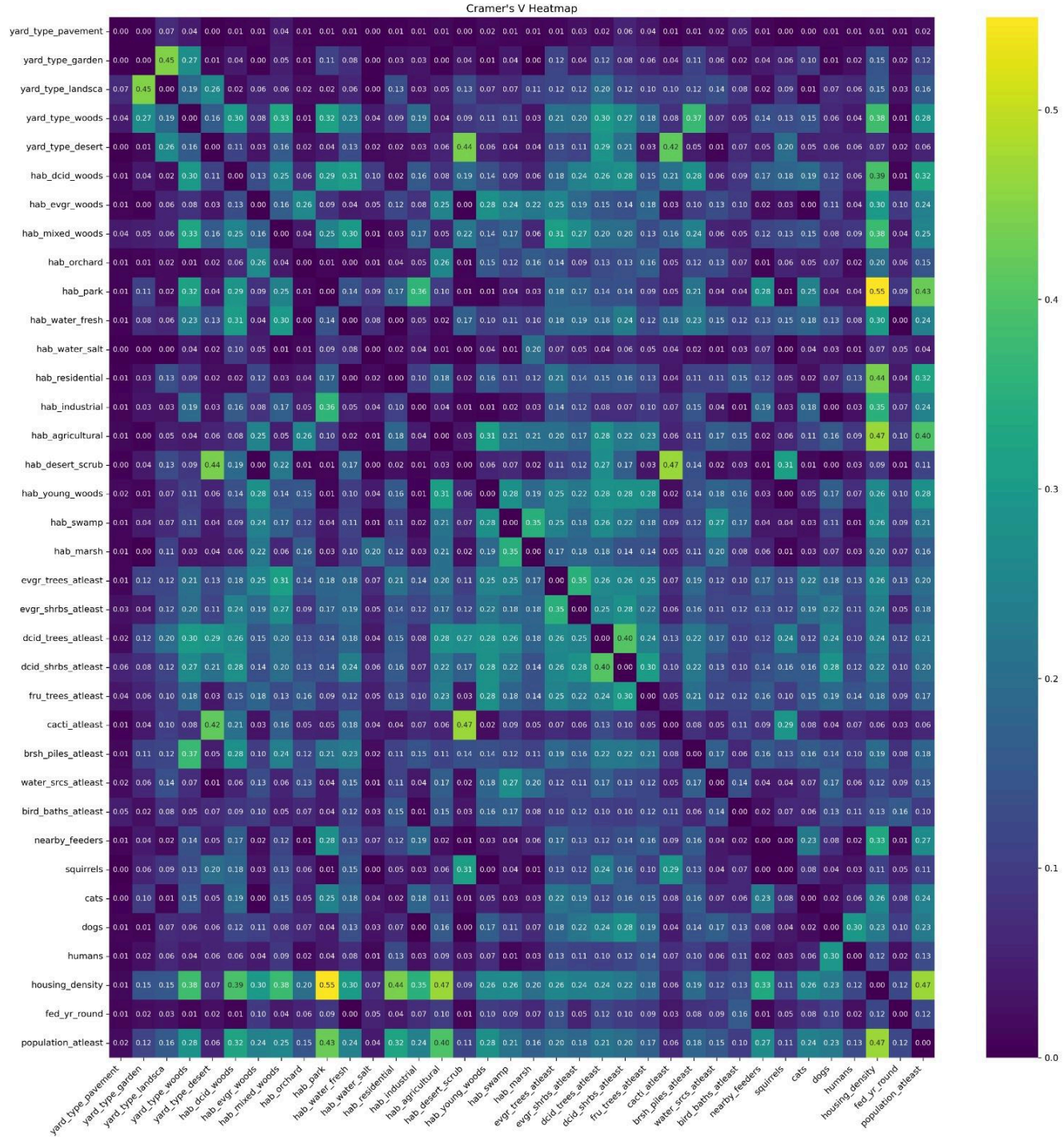
2.1.4 Merging Datasets and Final Imputation

Once both tables were individually cleaned, they were merged to form a unified dataset for modeling. The merge was performed using a set of identifiers—**location_id**, **latitude**, **longitude**, and **proj_period_id**—to ensure that each observer record was matched to the correct site information. Following the merge, a small amount of missingness was introduced, primarily due to non-overlapping site attributes or observer entries.

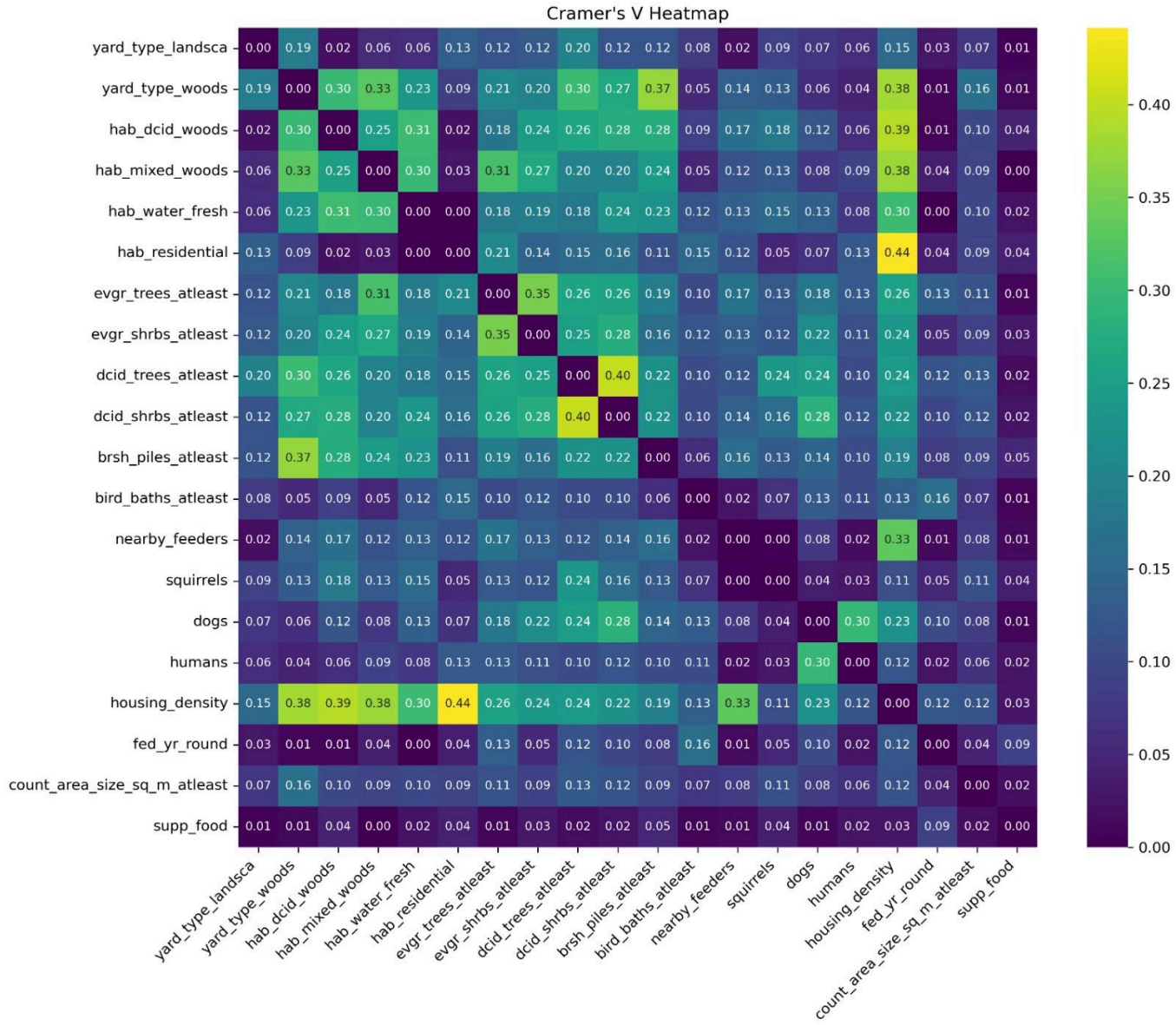
Because these newly missing values accounted for 20% of any given column, they were addressed using a final round of imputation: numeric variables were filled using the median, and categorical variables using the mode. This ensured that the merged dataset was fully consistent and ready for the subsequent stages of feature selection and modeling.

2.2 Feature Selection

To reduce dimensionality and keep only the most informative predictors, we performed a combination of correlation analysis and mutual information–based feature ranking. We started by computing **Cramér's V correlation matrix** for all categorical and binary features to identify any pairs that were highly correlated.

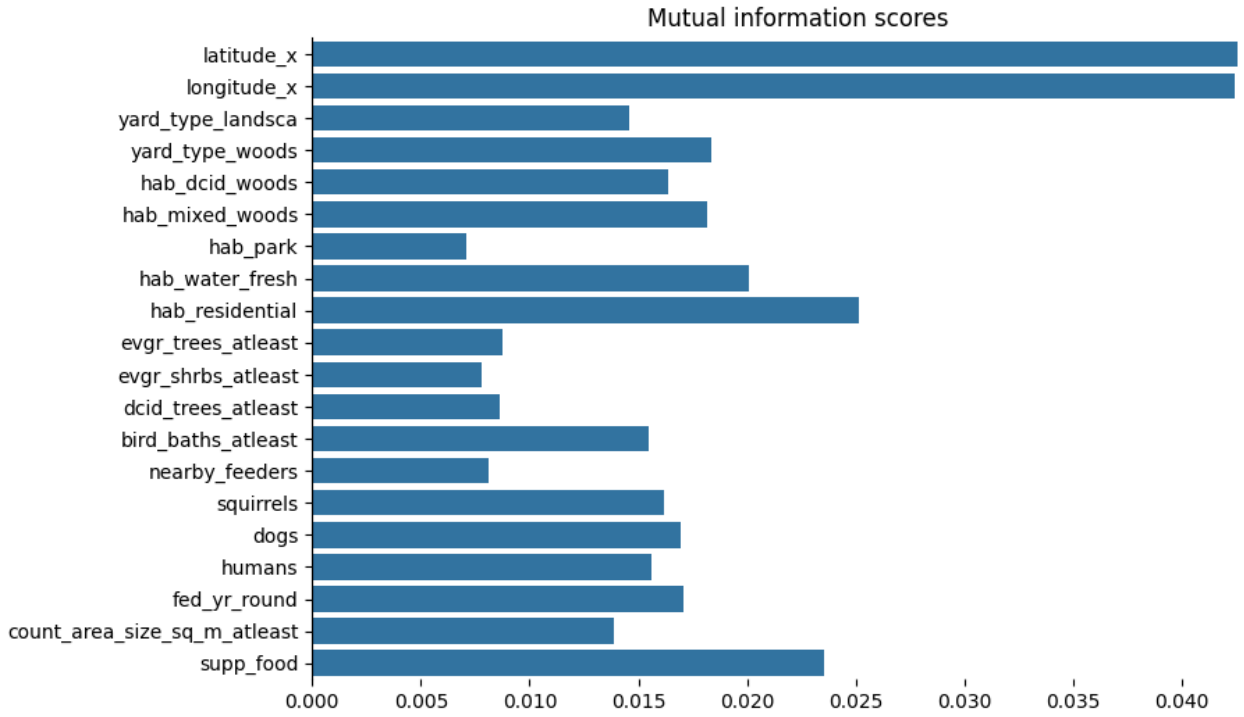


According to the full matrix , none of the feature pairs exceeded a correlation of 0.85, which suggested that most variables carried unique information and did not create strong multicollinearity issues. Next, we plotted Cramér's V again for the narrowed-down set of features to confirm that the remaining features did not have problematic overlap.



We removed a couple of features—such as *housing_density* and *dcid_shrbs_atleast*—because they showed both low individual relevance and relatively higher redundancy compared to the other variables.

After that, we computed **Mutual Information (MI) scores** for each feature with respect to the target variable (mallard presence).



The MI plot helped us rank variables based on how much useful signal they contributed. MI is especially helpful here because it can capture non-linear relationships that simple correlation measures may miss.

Using the combination of MI scores, ecological interpretability, and the effect on model performance, we selected a final set of **20 features**. These features provided a good balance between keeping enough information for the model to learn meaningful patterns and avoiding unnecessary noise or redundancy. This final feature set was used for training all machine learning models in the study.

CHAPTER 3

Results

This section presents the predictive performance of all four machine learning models that we used to estimate the presence of mallards at feeder sites: Logistic Regression, Random Forest, K- Nearest Neighbours and Decision Tree. Because the presence of mallards is rare in the dataset (only 1221 presence instances out of 88714 total observations), metrics such as recall, precision, F1- score and ROC AUC are more informative than accuracy alone.

3.1 Logistic Regression

Logistic Regression serves as a baseline linear model. While it achieved a moderate ROC AUC of 0.7479, its performance was strongly affected by severe class imbalance. The model predicted the majority class (“absent”) well, but struggled to distinguish true presence cases accurately. Precision for the presence class was extremely low (0.03), indicating many false positives.

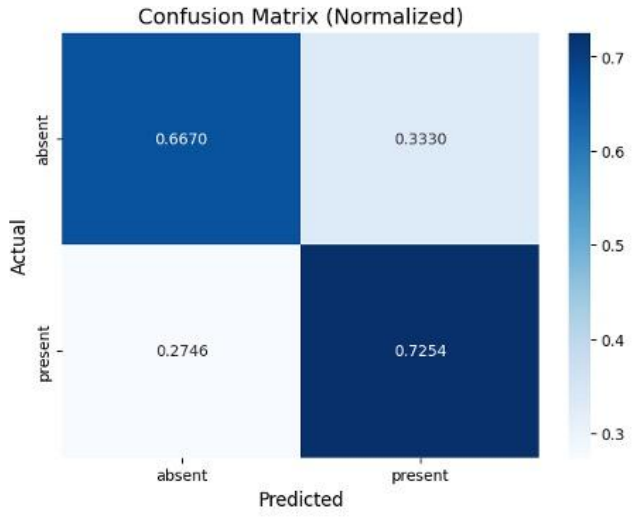
```
--- Logistic Regression ---
Accuracy: 0.667760048924223
ROC AUC: 0.7478981551717926
              precision    recall  f1-score   support

           0       0.99      0.67      0.80     17743
           1       0.03      0.73      0.06       244

    accuracy                           0.67     17987
   macro avg       0.51      0.70      0.43     17987
weighted avg       0.98      0.67      0.79     17987

Confusion Matrix:

                 Predicted: absent  Predicted: present
Actual: absent               11834                5909
Actual: present                 67                 177
```

Key Metrics of Logistic Regression Model:

- Accuracy : 0.67
- ROC AUC : 0.7479
- Precision for presence : 0.03
- Precision for absence : 0.99
- Recall (presence) : 0.73
- F1 - score (presence) : 0.06

Confusion matrix:
Below is the confusion matrix that has been plotted for this model.

- True Absent correctly predicted : 0.6670
- True Present correctly predicted : 0.7254



Overall although Logistic Regression captured some presence cases, it incorrectly predicted presence for thousands of absent sites, making it unsuitable for prediction.

3.2 Decision Tree

The Decision Tree model demonstrated stronger performance than Logistic Regression. The model achieved a ROC AUC of 0.9399 and correctly identified 83% of mallard presence instances. While

precision remained modest (0.26), it offered a more balanced trade-off between false positives and false negatives.

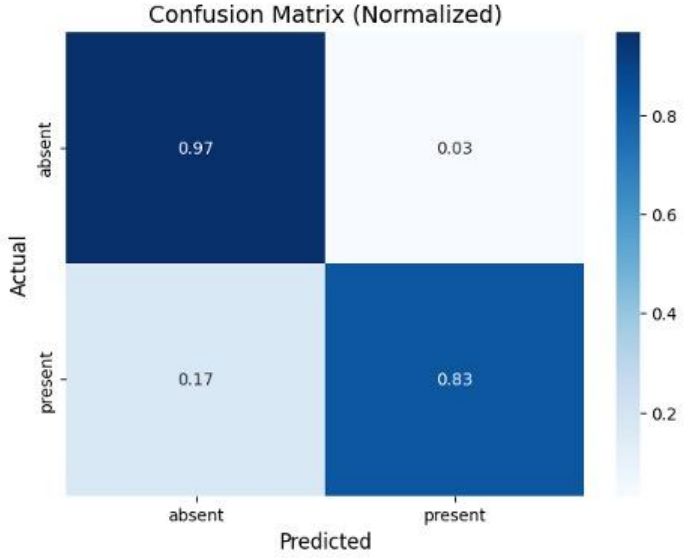
```
--- Decision Tree ---
Accuracy: 0.9660866181130817
ROC AUC: 0.9398544842898099
              precision    recall  f1-score   support

           0       1.00      0.97      0.98     17743
           1       0.26      0.83      0.40       244

    accuracy                           0.97     17987
   macro avg       0.63      0.90      0.69     17987
weighted avg       0.99      0.97      0.97     17987

Confusion Matrix:
 [[17175   568]
 [   42   202]]


Confusion Matrix:

                 Predicted: absent  Predicted: present
Actual: absent               17175                 568
Actual: present                 42                 202
```

Key Metrics of Decision Tree Model:

- Accuracy : 0.97
- ROC AUC : 0.9399
- Precision for presence : 0.26
- Recall (presence) : 0.83
- F1 - score (presence) : 0.40

Confusion matrix:
Below is the confusion matrix that has been plotted for this model.

- True Absent correctly predicted : 0.97
- True Present correctly predicted : 0.83



The decision tree provided strong recall for presence cases, making it useful for ecological monitoring scenarios where missing a real presence event is costly.

3.2.1 Feature Importance Analysis:

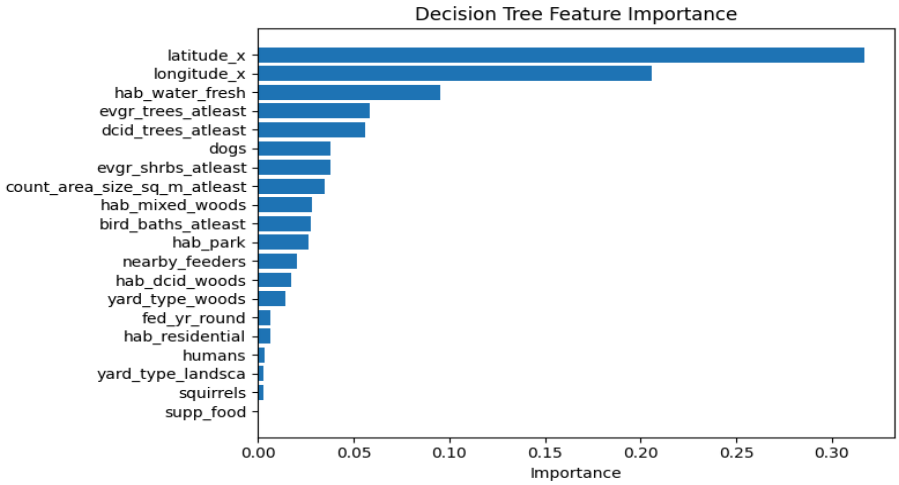
After training the decision tree model, we looked at the feature importance scores to understand which variables were most useful for predicting mallard presence. The results show that latitude, longitude, and hab_water_fresh were the top three features. This makes sense because mallard sightings depend a lot on seasonality and where the feeder site is located geographically.

Freshwater habitat also showed up as an important factor, which fits well with what we know about mallards since they prefer areas near ponds, lakes, and streams. Some vegetation-related variables, like evergreen and deciduous trees, had moderate importance, suggesting they may play a supporting role in the environment mallards use.

Human-related features such as dogs, nearby feeders, and bird baths had smaller importance values, and a few variables like squirrels and supplemental food contributed almost nothing to the model. Overall, the feature importance results suggest that time of year, geographic location, and access to freshwater are the strongest predictors of mallard presence in our dataset.

Index	Feature	Importance
0	latitude_x	0.316910
1	longitude_x	0.206127
7	hab_water_fresh	0.095386
9	evgr_trees_atleast	0.058425
11	dcid_trees_atleast	0.056119
15	dogs	0.037922
10	evgr_shrbs_atleast	0.037774
18	count_area_size_sq_m_atleast	0.034942
5	hab_mixed_woods	0.028057
12	bird_baths_atleast	0.027417
6	hab_park	0.026686
13	nearby_feeders	0.020543
4	hab_dcid_woods	0.017446
3	yard_type_woods	0.014485
17	fed_yr_round	0.006268

8	hab_residential	0.006262
16	humans	0.003532
2	yard_type_landsca	0.002995
14	squirrels	0.002703
19	supp_food	0.000000



3.3 k-Nearest Neighbours (kNN)

The kNN model delivered very high accuracy (0.99). However, accuracy is misleading in imbalanced datasets. kNN achieved strong precision for the presence class (0.60) but weak recall (0.38), meaning it frequently failed to detect true presence cases.

```
--- knn ---
Accuracy: 0.988102518485573
ROC AUC: 0.8440519835575887
              precision    recall  f1-score   support

           0       0.99      1.00      0.99     17743
           1       0.60      0.38      0.47       244

    accuracy                           0.99     17987
   macro avg       0.79      0.69      0.73     17987
weighted avg       0.99      0.99      0.99     17987

Confusion Matrix:
 [[17680    63]
 [  151    93]]


Confusion Matrix:

                 Predicted: absent  Predicted: present
Actual: absent               17680                  63
Actual: present                151                  93
```

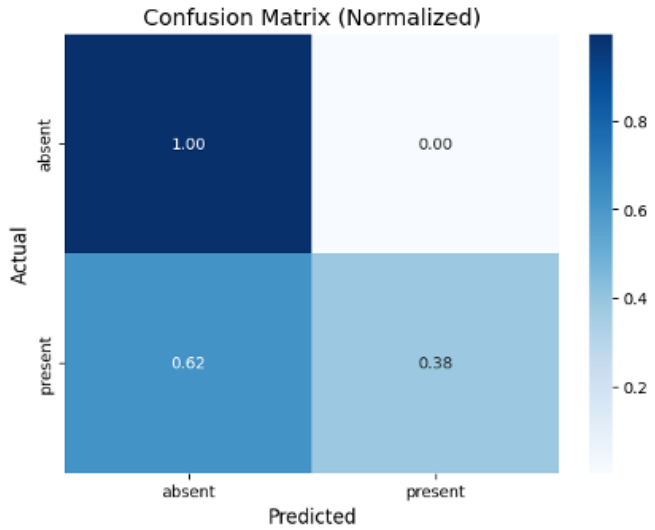
Key Metrics of kNN Model:

- Accuracy : 0.99
- ROC AUC : 0.8440

- Precision for presence : 0.60
- Recall (presence) : 0.38
- F1 - score (presence) : 0.47

Confusion matrix:

- True Absent correctly predicted : 1.0
- True Present correctly predicted : 0.38



kNN's performance shows that it only labels presence when very confident, resulting in many false negatives. This behaviour is undesirable for our prediction goals.

3.4 Random Forest

Random Forest emerged as the best-performing model overall. By combining multiple decision trees through bagging, it generalized well and handled the class imbalance more efficiently. The Random Forest achieved a ROC AUC of 0.9433 and consistently identified presence cases with 83% recall while maintaining relatively low false-positive rates.

Key Metrics of Logistic Regression Model:

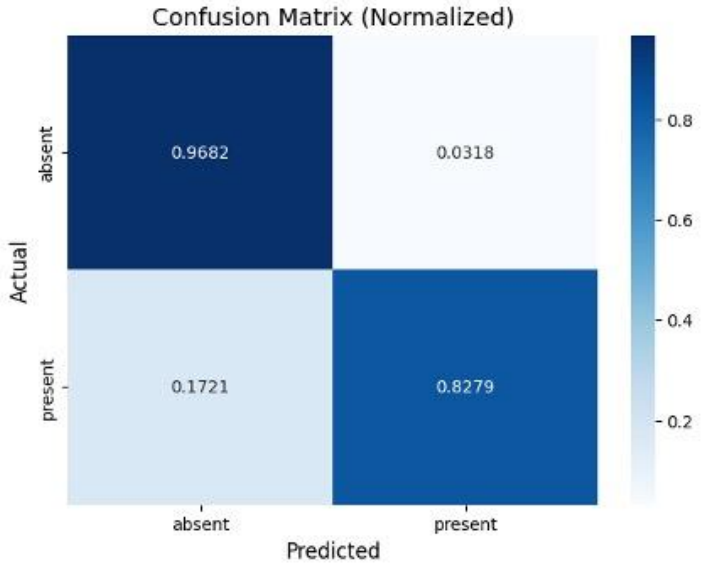
- Accuracy : 0.97
- ROC AUC : 0.9433
- Precision for presence : 0.26
- Recall (presence) : 0.83
- F1 - score (presence) : 0.40

```
--- Random Forest ---
Accuracy: 0.9662534052371157
ROC AUC: 0.9432991583843269
              precision    recall  f1-score   support

           0       1.00      0.97      0.98     17743
           1       0.26      0.83      0.40       244

    accuracy                           0.97     17987
   macro avg       0.63      0.90      0.69     17987
weighted avg       0.99      0.97      0.97     17987
```

```
Confusion Matrix:

                 Predicted: absent  Predicted: present
Actual: absent               17178                 565
Actual: present                 42                 202
```

Confusion matrix:

- True Absent correctly predicted : 0.9682
- True Present correctly predicted : 0.8279



Random Forest achieved the best balance between sensitivity and overall classification stability. Given the ecological significance of minimizing false negatives (i.e., missed mallard sightings), this model was selected as the most reliable for final interpretation.

All model scores at a glance:

Model	**Accuracy**	**ROC AUC**	**Precision (1)**	**Recall (1)**	**F1 (1)**
Logistic Regression	0.67	0.7479	0.03	0.73	0.06
Decision Tree	0.97	0.9399	0.26	0.83	0.40
Random Forest	0.97	0.9433	0.26	0.83	0.40
k-Nearest Neighbors	0.99	0.8440	0.60	0.38	0.47

CHAPTER 4

Discussion and Analysis

One of the main goals of our model is to help birdwatchers see a certain species they are interested in. Because of the interests of birdwatchers and the imbalanced nature of our data set when

predicting a single species, recall is a very important evaluation metric. Birdwatchers typically are willing to visit a place if there is a chance to see a species of interest, even if the chance is fairly low. Most birdwatchers would be more disappointed if an interesting species visited a site near them and they missed it than they would be if they visited and the species was not present on that day. Therefore, minimizing false negatives is more important than minimizing false positives for birdwatcher-oriented predictions, leading us to focus on evaluating recall values.

More broadly, to evaluate our four models, we looked at the combination of scores for accuracy, area under the ROC curve, precision, recall, and f1 scores. These scores changed as we made alterations to our modeling pipeline to improve model performance. Before we added SMOTE resampling to our pipeline, all four of our models had high accuracy scores but low recall scores, a poor result for our application. This was a predictable result of fitting models to highly imbalanced data, where the number of data points where a mallard was present is much higher than the number of the data points where a mallard was absent. Our pre-resampling logistic regression model actually classified every single point as absent. Before resampling, the other three models had recall scores between 0.55 and 0.6.

Applying SMOTE resampling to our training data vastly improved our final model performance. Our final logistic regression model had a decent recall score of 0.73, but its precision (0.03) and accuracy (0.67) were too low even for birdwatching prediction applications. Our k-nearest neighbors model had the best accuracy, precision, and f1 scores of our models, but its low recall value of 0.38 makes it a poor fit for birdwatching prediction. The k-nearest neighbors model might be a better fit for some ecological and conservation applications that value precision and f1 scores more highly than recall, but the model would likely have to be refined or tuned further to improve performance. For example, one application that might prioritize precision more highly would be a financially limited project that wanted to capture a specific species for bird banding. Due to budgetary constraints, they might only want to visit a low number of sites with a high likelihood of the species being present.

Our decision tree and random forest models had similar recall, precision and accuracy values, but the random forest model had a larger area under the ROC curve. The random forest model also had slightly higher precision than the decision tree, but they both round to 0.26. We concluded that the random forest model was the best model for birdwatcher-oriented predictions because of its high ROC AUC of 0.9433 and its recall value of 0.83. However, for some other application that highly values simplicity and interpretability, the decision tree model might be an acceptable substitute for the random forest model because of its similar recall score.

We considered tuning our random forest model further by altering the decision threshold. However, after graphing the recall and precision scores by decision threshold and considering the aims of our project, we concluded that the default threshold's recall of 0.83 and precision of 0.26 had a good balance for birdwatchers. With this precision, predictions of a mallard's presence will only be correct 26% of the time, leading birders to have to visit more places or visit more often to guarantee seeing the species. However, with 83% recall, birders will be unlikely to miss the chance to see a mallard if it visits a site near them. This type of recall-precision balance would be even more valuable for rarer bird species, but we might not be able to achieve this level of precision if we trained the model on a very rare species.

The group of twenty features that we selected for the model due to high mutual information scores have interesting implications from an ecological perspective. Predictably, latitude and longitude are individually highly predictive for the presence of mallards, reflecting how bird species tend to have defined geographic winter ranges. Residential habitat and supplemental food also had high mutual information scores with the target, although they did not have high feature importance in the decision tree model. This means that leaving out extra food probably increases the chances of seeing a mallard

substantially, if we assume causation. Residential areas must have a layout and set of ecological features that appeal to mallards. The residential habitat feature might be capturing a mixture of habitat characteristics that were not directly measured or were excluded from the model for having lower individual mutual information scores. Other features in our model include freshwater habitat, which might not be included as a feature for a non-aquatic bird species, other yard and habitat types, the presence of evergreen or deciduous trees, and the presence of other animals, specifically squirrels, dogs, and humans. It is particularly interesting that for mallard prediction we have multiple features related to the presence of trees, shrubs, and woods, with the presence of deciduous and evergreen trees being the fourth and fifth most important features for the decision tree. This may reflect that, when they are on shore, mallards may be able to find food in the form of plant matter more easily close to trees. Vegetation may also provide good nesting sites on shore.

Limitations

Since our model was trained on the FeederWatch site and observations data, any biases from that data will influence its real-world performance. If someone attempts to use the model to predict mallard presence for sites that are very different from the sites in our training set, model performance might be poor. For example, all of the observation sites were in the United States and Canada, so our model cannot make accurate predictions for sites in Europe, where mallards also live.

We trained our model to predict the binary presence or absence of a species. This approach obscures differences between sites that have low numbers of mallards on a single day and those that have high numbers of mallards. This could be addressed by fitting additional models to predict the number of mallards seen at a site on a certain day.

Our data was also limited to observations from November through March since the FeederWatch project is focused on the winter locations of migratory birds. A different data source might enable us to take year-round presence data into account.

CHAPTER 5

Conclusions

In this project, we used various site features to predict the presence of mallard ducks at particular locations on particular dates. We aimed to fit a model that could provide useful predictions for birdwatchers seeking the species. As a secondary goal, we wanted to find out which individual habitat features were highly predictive of species presence for ecology and conservation purposes. We found that, for mallards, a random forest model was the best fit for birdwatching-related aims. We also found a decision tree model that performed almost as well as the random forest model and was more easily interpretable.

Our findings can serve as a general promising proof of concept for using a random forest model to successfully predict the presence of bird species. Future random forest models to predict the presence of other bird species might have different selected features but might also be able to achieve good performance, as measured by recall and ROC-AUC scores. We also demonstrated that using SMOTE resampling can function well to help achieve good model performance when the presence/absence data set is highly imbalanced. This type of resampling might work well for other bird species prediction models, including those for rarer species than mallards. Additionally, our consideration of possible decision threshold adjustments led us to believe that adjusting the decision threshold could help models for other bird species to achieve a good balance of recall and precision.

We also were able to observe from our decision tree feature importances that factors like longitude, latitude, the presence of fresh water, and the presence of deciduous and evergreen trees were

the most important for predicting mallard presence. Ecologists and conservation professionals could make good use of these features and other predictive features in our model. They could use these features and features from similar models for other species to help protect the habitats of bird species, focusing on the most important habitat features.

Future Work

This project focused on predicting the presence of a single species—the mallard (*mallar3*)—using observational and habitat data. However, several avenues can further enhance and expand the predictive modeling approach:

- **Predicting Multiple Species:**
  Future work can extend the current framework to model the presence of multiple bird species simultaneously. This would allow for comparative analysis across species, highlight shared habitat drivers, and provide a broader ecological understanding of community-level patterns.
- **Species Grouping for Improved Accuracy:**
  Instead of modeling each species independently, species can be grouped based on habitat preference, feeding behavior, taxonomic similarity, or ecological traits. Group-based classification models may provide more stable predictions and reduce the sparsity problem caused by rare species with limited observations.
- **Advanced Modeling Techniques:**
  Incorporating ensemble learning, multi-label classification, or deep learning approaches could capture more complex interactions in the data. Models such as gradient boosting, random forest with tuned hyperparameters, or neural network–based architectures may further improve predictive performance.
- **Temporal and Spatial Extensions:**
  Additional features such as seasonal variation, migration timing, and spatial autocorrelation could enhance model depth. Integrating external data sources (e.g., climate data, land-cover maps) may also improve accuracy.

Overall, expanding the model to include more species and grouping strategies would create a more robust and generalizable ecological prediction system, opening the door to broader conservation.

References

- https://scikit-learn.org/stable/modules/classification_threshold.htmlclassification_threshold.html
- Mallard Life History, All About Birds, Cornell Lab of Ornithology
- Raw Dataset Downloads - Project FeederWatch
- https://www.frontiersin.org/journals/ecology-and-evolution/articles/10.3389/fevo.2021.619682/full
- python - The easiest way for getting feature names after running SelectKBest in Scikit Learn - Stack Overflow