Sarah Vastani and JaDarius Jones

05/01/2024

SURVIVAL OF MIGRATION

Texas A&M University

STAT 654-600

Prepared For :
Dr. Guha

Contributions

Sarah Vastani- Introduction, Methodology, Modeling approach: Decision Tree, Code

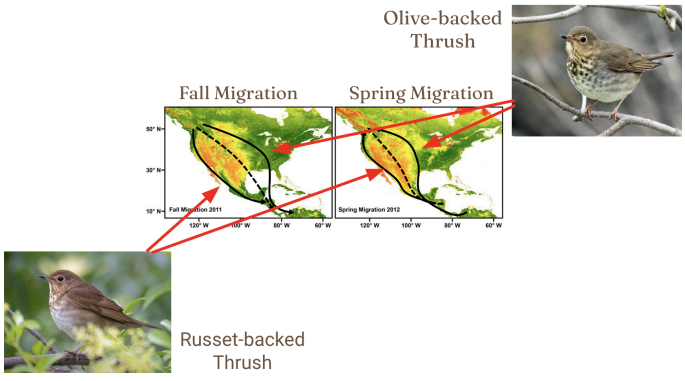
JaDarius Jones- Modeling approach: Multiple Logistic Regression Model, Probit Model, Conclusion, Discussion

Introduction

The Swainson's thrush (Catharus ustulatus) presents an intriguing subject for study due to its migratory behaviors and the genetic complexities associated with its hybridization. This avian species, distributed widely across North America, comprises two distinct subspecies distinguished primarily by the coloration of their feathers: the olive-backed and the russet-backed thrushes. To comprehensively address the question of which migratory and morphological traits influence the survival of these thrush populations, our study harnesses a multifaceted approach encompassing genetic analysis, telemetry tracking, and morphological assessments provided by the Delmore team.

Notably, these subspecies, the olive-backed and the russet-backed thrushes, follow different migratory paths during both fall and spring migrations, with the former favoring inland routes while the latter favors coastal routes. The interbreeding of these subspecies gives rise to hybrid individuals, particularly the first-generation hybrids, which inherit genetic material favoring both coastal and inland migration routes. Consequently, these hybrids often exhibit an intermediate migratory pattern, which is shown by the dashed line in the middle of the map in Figure 1.

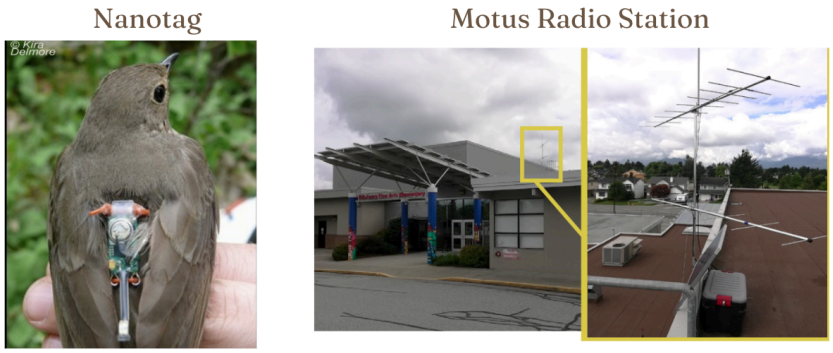
Figure 1



Later-generation hybrids, the offspring of two first-generation hybrids, further complicate the genetic landscape with a random assortment of alleles, leading to a spectrum of migratory behaviors. Thus, all the samples can be assigned with the feature ‘heterozygosity’. A value close to 0 would mean the bird has more pure DNA, and a value close to 1 would mean that it has more hybrid DNA. They can also be assigned with the feature ‘ancestry’. A value close to 0 would mean the bird has more coastal DNA, and a value close to 1 would mean that it has more inland DNA.

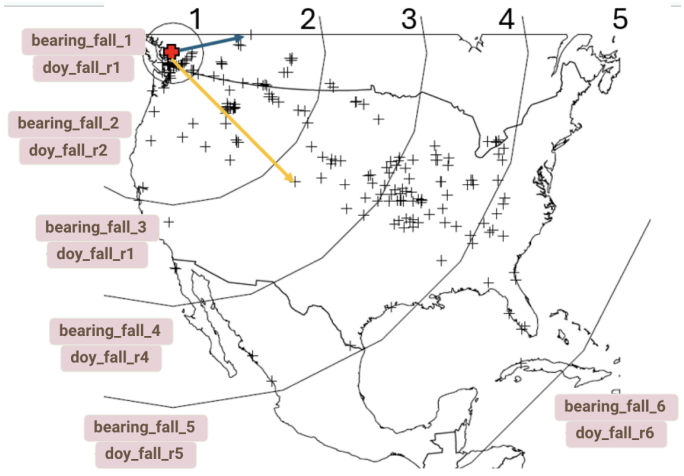
Additionally, these birds are fitted with nanotag, which are radio transmitter devices that are attached to them (Figure 2). These devices emit a unique signal, which can be detected by a motus radio station. These radio stations are placed across North America to help document the bird’s location throughout the year. Each time a bird is detected passing a certain latitude, a 1 is placed in the feature corresponding to the latitude and season of the year. Their migration starts in the fall at 40 degrees latitude, around Vancouver Canada. Thus, if there is a 1 in ‘t1_fall40’, it would mean the bird didn't die at the start of the migration. As they travel south, they reach 25 degrees latitude. If they have survived until this point in migration, there will be a 1 in variable ‘t2_fall25’. Traveling even more south, near the tip of Florida, is where their fall migration ends. Thus, if there is a 1 in ‘t3_winter00’, that means they survived the fall migration and the beginning of the spring migration. They then travel back north, and if they reach 25 degrees latitude, there will be a 1 for the variable ‘t4_spring25’. Lastly, if there is a 1 in the feature ‘t5_spring40’, that means they survived both the fall and spring migrations and made it back to the breeding grounds.

Figure 2



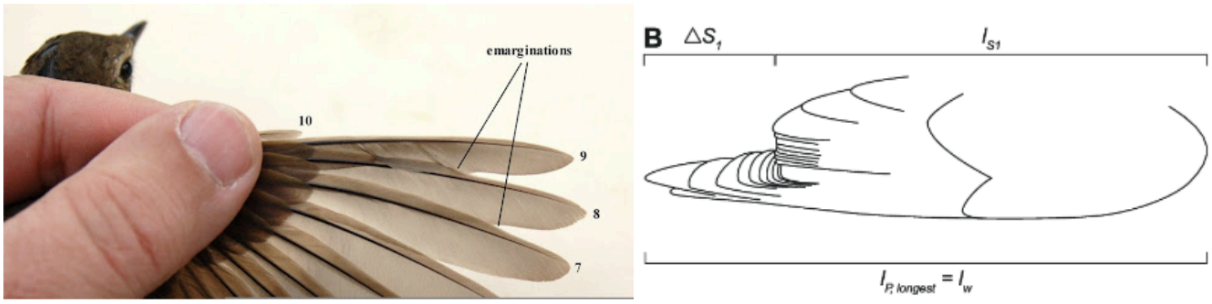
In addition to collecting survival based on latitude, there is data collected for bearings. All of the birds in our sample were released at the red point shown in Figure 3. The black crosses are where the birds were detected during migration. To describe where they are going quantitatively, six rings were generated around the release site. Each bearing would be the angle measured between the red release site and the black crosses inside each of the rings. For example, the blue arrow shows how the value for 'bearing_fall_2' was collected, while the yellow arrow indicates a value for 'bearing_fall_3'. The day of the year that each bearing was taken is also recorded as doy. For example, if the doy is 1, that would mean that the corresponding bearing was taken on January 1st.

Figure 3



Various morphological data can be obtained from the bird, which can also serve as predictors for the survival of the migration. This includes the lengths of certain feathers, like ‘p7’, ‘p8’, ‘p9’, and ‘p10’ (Figure 4). Additionally, ‘distal’ is measured from the tip of ‘p10’ to the longest feather. Wing measurements include delta S1, which is recorded as ‘kipps’, LS1 which is recorded as ‘carpal’, and LW which is recorded as ‘wing.cord’. These are all measured in millimeters.

Figure 4



Other morphological data collected include ‘tarsus.length’, which is the leg bone length, and the ‘tail.length’, both of which are also in millimeters. We also have the sex binary data; a 0 would indicate a female and a 1 would indicate a male. The fat score is measured through observation. A score of 0 indicates a lack of visible fat on the bird. The score ranges up to 5, which represents the highest level of visible fat.

Through meticulous data preparation, including imputation techniques to address missing values and feature selection to streamline the model, we refine our dataset for predictive modeling. Leveraging techniques including logistic regression, probit regression, and decision tree analysis, we aim to construct predictive models capable of discerning the factors influencing migration survival.

In this paper, we present a comprehensive exploration of the Swainson's thrush migratory dynamics, encompassing genetic, environmental, and morphological factors. By elucidating the intricate interplay between genetics and environmental cues in shaping migratory behaviors, our study contributes to a deeper understanding of avian ecology and evolutionary processes.

Methodology

Our study employs a combination of genetic analysis, telemetry tracking, and morphological assessments to investigate the migratory dynamics of Swainson's thrush populations. The methodology can be divided into the following key components:

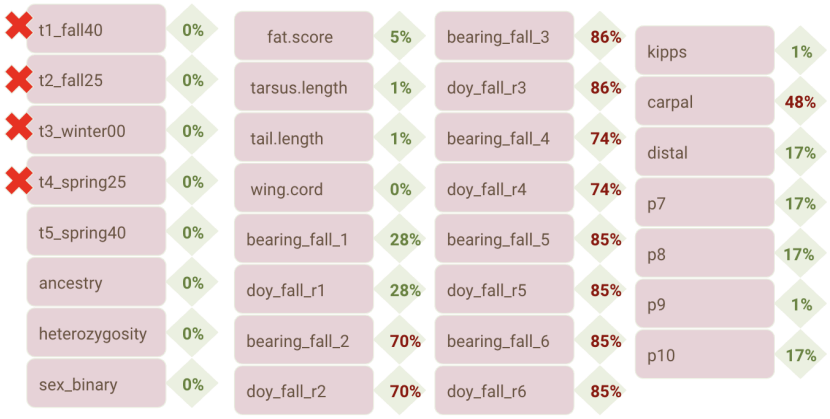
Data Collection:

The dataset was obtained thanks to the Delmore team.

Data Preparation:

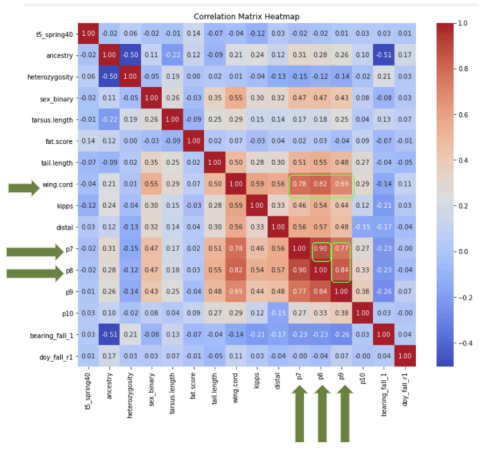
Our goal with this project is to predict the variable ‘t5_spring40’, which tells us if the bird survived both fall and spring migration and made it back to the breeding grounds. Thus, we will remove columns ‘t1_fall40’, ‘t2_fall25’, ‘t3_winter00’, and ‘t4_spring25’, since that would be redundant information. In other words, if there is a 1 in ‘t5_spring40’, then it is implied that there would also be a 1 in ‘t1_fall40’, ‘t2_fall25’, ‘t3_winter00’, and ‘t4_spring25’. Because if they survived the whole migration, then they also survived the intermediate checkpoints of the migration. We also have some columns that contain a significant amount of null values. So we removed any column that had around 50% or more of the data missing, which included ‘carpal’, bearings 2 through 6, and doys 2 through 6. We also removed a sample which contained null for 10 features.

Figure 6: Null values shown as a percentage and column removal visualization



Feature selection techniques are employed to identify relevant predictors for migration survival, with redundant or irrelevant variables removed from the dataset. Looking at the correlations, we found that ‘p7’, ‘p8’, ‘p9’, and wing cord are highly correlated to each other with correlation values of 0.77 and above, so we decided to remove ‘p7’ and ‘p8’ since they contain more null values than ‘p9’ and wing cord (Figure 7).

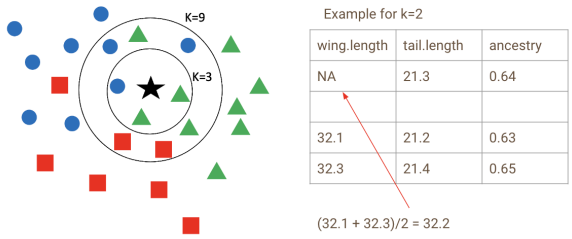
Figure 7



In our research, we utilized K-nearest neighbors (KNN) imputation to address missing values within our high-dimensional dataset. This technique was chosen for its ability to account for the intricate relationships among all variables present in the data, rendering it particularly

suitable for datasets characterized by complex interdependencies. Upon encountering a missing value for a specific variable (e.g., wing length), the KNN imputation process commenced by identifying the "k" closest neighbors to the observation with the missing value. Proximity was determined based on the values of other non-missing variables (e.g., tail length and ancestry) (figure 8). Subsequently, the missing value was imputed by computing the average of the values of the variable of interest (e.g., wing length) among the identified nearest neighbors. Leveraging the characteristics of similar observations in the dataset, KNN imputation facilitated the generation of a robust estimate for the missing value, effectively incorporating information from neighboring data points.

Figure 8



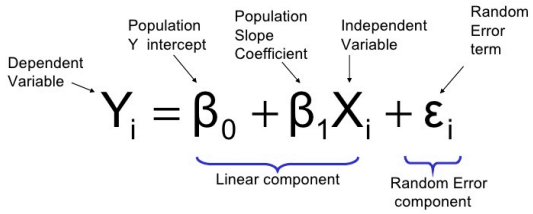
Modeling Approach:

We employ multiple regression techniques, including logistic regression, probit regression, and decision tree analysis, to construct predictive models for migration survival.

The first model we fit was the multiple logistic regression model which is an extension of logistic regression that involves more than one independent variable. In multiple logistic regression, the goal is to model the relationship between a binary dependent variable, ‘t5_sping40’, and two or more independent variables: ‘ancestry’, ‘heterozygosity’, ‘sex_binary’, ‘tarsus_length’, ‘fat_score’, ‘tail_length’, ‘wing_cord’, ‘kipps’, ‘p9’, ‘p10’, ‘bearing_fall_1’, ‘doy_fall_1’, and ‘distal’. Using the sigmoid function as a linking function, the model predicts

the probability of the dependent variable being in one category (e.g., 1 or 0) based on the values of the independent variables. Logistic regression is well-suited for modeling binary outcomes--in our case, whether a bird survives migration or not. Since we are studying the survivability of Swainson's thrushes, which can be categorized as either surviving or not surviving, logistic regression is an appropriate modeling technique. Additionally, logistic regression provides easily interpretable results. The coefficients obtained from the logistic regression model represent the log-odds of the outcome variable being in one category (e.g., survivability). This makes it straightforward to interpret the effects of predictor variables on the likelihood of survival for Swainson's thrushes.

Multiple logistic regression model equation



The next model we tested was the probit regression model. Similar to the logistic model, in probit regression, the link function used is the cumulative distribution function of the standard normal distribution, also known as the probit function. This model is often accompanied by logistic regression and can produce similar results. While the interpretation may be slightly less direct compared to logistic regression coefficients, probit regressions can provide an alternative explanation that can inform the relationship between predictor variables and the probability of survival for Swainson's thrushes.

Probit regression model equation

$$p_i = \Phi(\beta_o + \beta_1 X_1^i + \beta_2 X_2^i + \beta_3 X_3^i + \cdots + \beta_N X_N^i)$$

The final model we tested was a decision tree. A decision tree is a binary tree that recursively splits the dataset until we are left with pure leaf nodes. There are two kinds of nodes, decision nodes and leaf nodes. The decision node contains information to split the data, and the leaf node helps us to decide the class of a new data point. Using a decision tree model can be a suitable approach for analyzing the migratory dynamics of Swainson's thrush populations for several reasons, particularly non-linearity. Migration survival may be influenced by complex, non-linear interactions between genetic, environmental, and morphological factors. Decision trees can capture these non-linear relationships effectively, allowing for more accurate modeling of the migratory process. In our case, BayesSearchCV is utilized for hyperparameter optimization in decision tree modeling, allowing for the efficient selection of optimal model parameters based on performance metrics. How this works at first, it starts by randomly guessing values for the parameters without any real strategy, just to see how the model performs. After making these guesses, it looks at which guesses worked better and which ones didn't. Based on which settings seem to work well, it tries some variations of those. It keeps repeating this process, each time getting a bit smarter about which guesses to make, and finally identifying the best set of parameters that gives the highest performance for the model.

Evaluation and Validation:

The performance of each model is evaluated using appropriate metrics, such as accuracy, AIC, and ROC-AUC. Cross-validation techniques, such as k-fold cross-validation, are employed to validate model performance and mitigate overfitting.

Results

Multiple Logistic Regression Model

The logistic regression analysis aimed to assess the relationship between various predictor variables and the survivability of Swainson's thrushes during migration. The model revealed several coefficients that represent the estimated change in the log-odds of survivability associated with a one-unit change in each predictor variable, holding other variables constant.

Among the predictor variables in Figure 9, only fat score and ‘kipps’ demonstrated statistically significant associations with survivability at the alpha level of 0.05. Specifically, a one-unit increase in fat score was associated with a 0.531 increase in the log-odds of survivability ($p = 0.0186$), while a one-unit increase in ‘kipps’ was associated with a -0.202 decrease in the log-odds of survivability ($p = 0.0476$).

The remaining predictor variables, including ‘ancestry’, ‘heterozygosity’, ‘sex_binary’, ‘tarsus_length’, ‘tail_length’, ‘wing_cord’, ‘p9’, ‘p10’, ‘bearing_fall_1’, ‘doy_fall_r1’, and ‘distal’, did not show statistically significant associations with survivability based on $p > 0.05$.

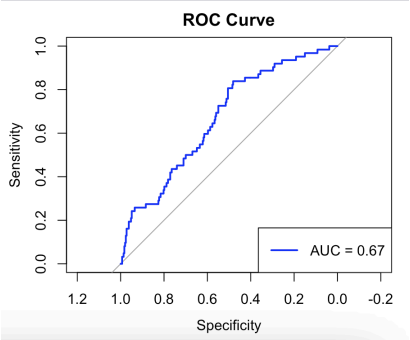
The model's goodness of fit was assessed using the null deviance and residual deviance. The null deviance, representing the deviance of the model with no predictors, was 328.86 on 354 degrees of freedom. The residual deviance, representing the deviance of the model with predictors included, was 309.27 on 353 degrees of freedom. The difference between the null and residual deviance provides a measure of how well the model fits the data. Additionally, the Akaike Information Criterion (AIC) value of 337.27 indicates the relative quality of the model, with lower AIC values suggesting better fit, and an accuracy of 0.825352.

Figure 9

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.310e+00  9.549e+00   0.451   0.6518
ancestry        1.776e-02  1.204e+00   0.015   0.9882
heterozygosity  1.360e+00  9.801e-01   1.388   0.1651
sex_binary      1.528e-01  4.283e-01   0.357   0.7212
tarsus_length  -2.994e-02  1.415e-01  -0.212   0.8324
fat_score       5.317e-01  2.258e-01   2.354   0.0186 *
tail_length    -1.304e-01  7.721e-02  -1.688   0.0913 .
wing_cord      -8.927e-02  1.037e-01  -0.861   0.3894
kipps          -2.021e-01  1.020e-01  -1.981   0.0476 *
p9              1.150e-01  7.710e-02   1.492   0.1357
p10             1.140e-01  8.370e-02   1.362   0.1731
bearing_fall_1  1.606e-01  1.344e-01   1.195   0.2321
doy_fall_r1     1.621e-03  5.659e-03   0.286   0.7745
distal         -2.847e-05  2.799e-02  -0.001   0.9992

    Null deviance: 328.86  on 354  degrees of freedom
Residual deviance: 309.27  on 341  degrees of freedom
AIC: 337.27
```

Figure 9.2



When comparing the results of the logistic regression models in Figures 9 & 10, we observe some differences in the estimated coefficients and their significance levels across the predictor variables.

In the previous model (refer to Figure 9), the model includes additional predictor variables such as ‘ancestry’, ‘sex_binary’, ‘tarsus_length’, ‘wing_cord’, ‘bearing_fall_1’, ‘doy_fall_r1’, and ‘distal’. However, none of these variables showed statistically significant associations with survivability. Regarding model fit, p-values themselves do not directly contribute to AIC, however, they can indirectly influence model fit if they are used to guide model selection or parameter estimation. Thus, removing variables with a high p-value (indicating that it is not statistically significant) within the p-value threshold of 0.3 to 0.7 from the model would lead to a simpler model with a lower AIC value and improving model fit.

Moreover, fat score and ‘kipps’ were the only two variables that showed statistically significant associations with survivability at the alpha level of 0.05 across both models. Specifically, a one-unit increase in fat score was associated with a 0.512 increase in the log-odds of survivability ($p = 0.01695$), while a one-unit increase in ‘kipps’ was associated with a -0.237 decrease in the log-odds of survivability ($p = 0.00918$).

The goodness of fit measures, including the null deviance, residual deviance, and AIC, provide insights into the model's performance as well. While both models have similar null deviance values (328.86), the previous model shows slightly lower residual deviance (309.27) compared to the current model (311.21), indicating a better fit to the data. However, the AIC value for the current model (337.27) is slightly higher than that of the previous model (324.11), suggesting that the current model may have a better overall fit. Especially with its higher accuracy: 0.8225352. However, the AUC for the current model, which measures the discriminatory power of a classifier, was 0.65 and the previous model was .67, meaning that the ability of the model to distinguish between positive and negative cases was poor in both cases (refer to Figures 9.2 & 10.2).

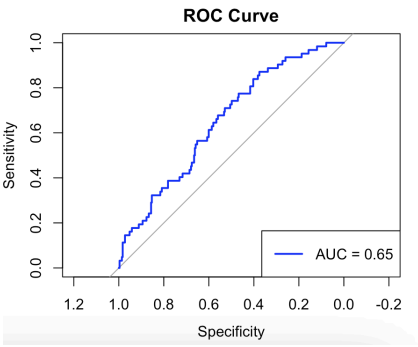
Figure 10

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.87298    4.72473   0.396  0.69179
heterozygosity 1.22318    0.79944   1.530  0.12600
fat_score      0.51202    0.21444   2.388  0.01695 *
tail_length   -0.12879    0.07096  -1.815  0.06954 .
kipps         -0.23783    0.09128  -2.605  0.00918 **
p9             0.05952    0.06305   0.944  0.34518
p10            0.11631    0.06708   1.734  0.08297 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 328.86  on 354  degrees of freedom
Residual deviance: 311.21  on 348  degrees of freedom
AIC: 324.11
```

Figure 10.2



Probit Regression Model

Among the predictor variables, the probit model (refer to Figure 11) also suggests that fat score and ‘kipps’ were found to have statistically significant associations with survivability at the alpha level of 0.05. Specifically, a one-unit increase in fat score was associated with a 0.303 increase in the standard deviation associated with survivability (p = 0.0235), while a one-unit increase in ‘kipps’ was associated with a -0.121 decrease in the standard deviation (p = 0.0358).

The remaining predictor variables, including ‘ancestry’, ‘heterozygosity’, ‘sex_binary’, ‘tarsus_length’, ‘tail_length’, ‘wing_cord’, ‘p9’, ‘p10’, ‘bearing_fall_1’, ‘doy_fall_r1’, and ‘distal’, did not show statistically significant associations with survivability (based on $p > 0.05$).

The null deviance was 328.86 on 354 degrees of freedom, while the residual deviance was 309.10 on 352 degrees of freedom with an AIC value of 337.1. The ROC-AUC result returned the same result of 0.67 (refer to Figure 12.2) with an accuracy of 0.8197183.

Figure 11

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.180e+00  5.355e+00   0.407   0.6839
ancestry        5.768e-02  6.674e-01   0.086   0.9311
heterozygosity  7.388e-01  5.444e-01   1.357   0.1747
sex_binary      7.773e-02  2.393e-01   0.325   0.7453
tarsus_length  -1.406e-02  8.022e-02  -0.175   0.8609
fat_score       3.026e-01  1.336e-01   2.265   0.0235 *
tail_length    -7.074e-02  4.270e-02  -1.657   0.0975 .
wing_cord      -4.708e-02  5.802e-02  -0.811   0.4171
kipps          -1.207e-01  5.751e-02  -2.099   0.0358 *
p9              6.125e-02  4.324e-02   1.417   0.1566
p10             6.397e-02  4.707e-02   1.359   0.1741
bearing_fall_1  9.240e-02  7.487e-02   1.234   0.2171
doy_fall_r1     9.750e-04  3.164e-03   0.308   0.7580
distal          2.497e-06  1.571e-02   0.000   0.9999
---

    Null deviance: 328.86  on 354  degrees of freedom
Residual deviance: 309.10  on 341  degrees of freedom
AIC: 337.1
```

Figure 11.2

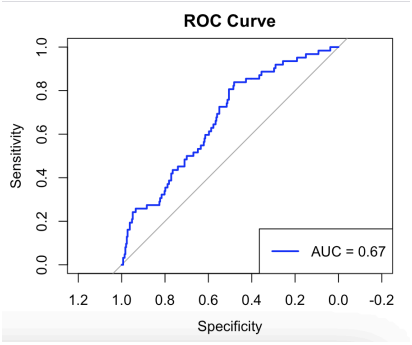


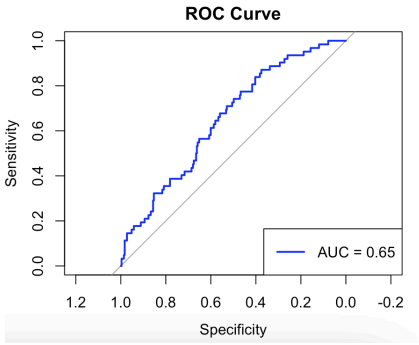
Figure 12

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.97493    2.56965   0.769   0.4422
heterozygosity  0.54334    0.43949   1.236   0.2163
fat_score       0.27571    0.12739   2.164   0.0304 *
tail_length    -0.04200    0.03543  -1.185   0.2359
kipps          -0.11053    0.04836  -2.286   0.0223 *
p9              0.05056    0.03328   1.519   0.1287
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 328.86  on 354  degrees of freedom
Residual deviance: 314.32  on 349  degrees of freedom
AIC: 326.32
```

Figure 12.2



After comparing the results of the two probit regression models, it reveals some differences in the estimated coefficients and their significance levels across the predictor variables as well.

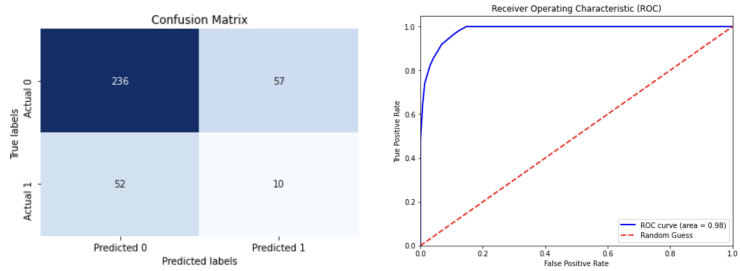
Comparatively, the previous model (Figure 11) includes additional predictor variables such as ‘ancestry’, ‘sex_binary’, ‘tarsus_length’, ‘wing_cord’, ‘bearing_fall_1’, ‘doy_fall_r1’, and ‘distal’. However, none of these variables showed statistically significant associations with survivability too, so the removal of predictors with p-value thresholds 0.4 to 0.7 was used to gauge the AIC fit. Although this threshold could be increased with this model, reducing too many variables can increase the AIC fit based on the AIC’s mathematical formulation. Also, the probit model returned the same effect on ROC-AUC.

While both models have similar null deviance values (328.86), the current model (Figure 12) shows a slightly lower residual deviance (314.32) compared to the previous model (309.10), indicating a better fit. Moreover, the AIC value for the current model (323.77) is slightly lower than that of the previous model (337.1) with a higher accuracy of 0.8225352, suggesting that the current model may have better overall fit.

Decision Tree Model

The results of the best hyperparameters were a ‘criterion’ of gini, ‘max_depth’ of 24, ‘min_samples_leaf’ of 19, and ‘min_samples_split’ of 8. This led to a decision tree with an accuracy score of 0.94 and an AUC of 0.98. According to this confusion matrix, the model is good at predicting instances where a bird won't survive, rather than those where it will (Figure 13). This is likely due to our unbalanced data, as 80% of our samples did not survive in the t5_spring40 column.

Figure 13



Discussion

Based on the results of the probit and logistic regression models, along with the findings from the decision tree analysis, we interpreted the outcomes in the context of our project goals.

The probit and logistic regression analyses revealed important predictors of survivability in Swainson's thrushes during migration. Both fat score and ‘kipps’ emerged as significant factors, indicating their potential importance in determining the likelihood of survival. These findings align with our project goals of identifying key factors influencing the survivability of thrush populations and providing insights for conservation efforts.

However, the decision tree analysis yielded promising results with high accuracy (0.94) and AUC (0.98). Despite the strong performance metrics, the confusion matrix highlights a potential limitation: the model's tendency to better predict instances where a bird won't survive rather than those where it will. This imbalance in predictions may be attributed to the skewed distribution of survival outcomes in our dataset, with a majority of samples indicating non-survival.

Comparing our results with existing literature or similar studies, we find support for the importance of factors such as fat score and ‘kipps’ in predicting survivability in bird populations during migration. However, the emphasis on the decision tree model in our study underscores the

value of machine learning techniques in analyzing complex ecological datasets and generating accurate predictions.

While the probit and logistic regression models provide valuable insights into specific predictors of survivability, the decision tree model offers a robust predictive tool with high accuracy. Nonetheless, it's essential to acknowledge the limitations of each approach, such as the imbalance in data distribution, and further validate our findings through ongoing research and collaboration within the scientific community.

Future research:

Moving forward, we recommend further exploration of additional predictors and their interactions to enhance predictive accuracy and robustness. Additionally, efforts to address data imbalance and incorporate more balanced datasets would improve model performance and generalizability. Furthermore, collaborative research endeavors and longitudinal studies can validate our findings and inform targeted conservation strategies aimed at safeguarding Swainson's thrush populations during migration. Ultimately, our project highlights the importance of interdisciplinary approaches and ongoing research efforts in advancing our knowledge of avian migration ecology and informing evidence-based conservation practices.

Conclusion

In summary, our project aimed to investigate the determinants of survivability in Swainson's thrushes during migration, utilizing logistic regression, probit regression, and decision tree analysis.

The main findings of our analysis reveal that fat score and ‘kipps’ are significant predictors of survivability, highlighting the importance of energy reserves and environmental cues in determining migration outcomes for thrush populations. Furthermore, while logistic and

probit regression models provided valuable insights into specific predictors, the decision tree model demonstrated superior predictive accuracy, although with a tendency to better predict non-survival instances due to data imbalance.

Our project significantly contributes to the field of avian ecology by facilitating key factors influencing migration survivability and showcasing the utility of machine learning techniques in ecological research. By integrating traditional statistical methods with advanced modeling approaches, we have expanded our understanding of the complex dynamics governing bird migration and provided practical insights for conservation efforts.

References

Barney, L. (n.d.). *PowerPoint presentation*.

matthewhalleymatthewhalleymatthewhalleymatthewhalley. (2017, February 8). *A closer look at the wing of Swainson's Thrush (Catharus ustulatus swainsoni) in First Basic plumage*. Matthew R. Halley. https://matthewhalley.wordpress.com/2017/02/08/a-closer-look-at-swainsons-thrush-catharus-ustulatus-swainsoni-in-first-basic-plumage/

Mixed genes mix up the migrations of hybrid birds. (2014, July 22). UBC News. https://news.ubc.ca/2014/07/22/mixed-genes-mix-up-the-migrations-of-hybrid-birds/

Swainson's thrush identification, all about birds, Cornell Lab of Ornithology. (n.d.). Retrieved May 1, 2024, from https://www.allaboutbirds.org/guide/Swainsons_Thrush/id

You are being redirected... (n.d.). Retrieved May 1, 2024, from https://www.surfbirds.com/community-blogs/northcoastdiaries/?p=852