2022 - 2023 H5N1 Bird Flu Modeling and Prediction in the United States

Weilin Cheng
University of California, Davis
wncheng@ucdavis.edu

Hengyuan Liu
University of California, Davis
hyliu@ucdavis.edu

Kathy Mo
University of California, Davis
kamo@ucdavis.edu

Sida Tian
University of Michigan
startian@umich.edu

Li Yuan
University of Michigan
leeyuan@umich.edu

June 23, 2023

Abstract

This report presents an analysis of the likelihood of H5N1 outbreaks in different counties of the United States in March 2023 using logistic regression, ridge regression, lasso regression, ridge & lasso regression models. The models were trained using historical data from January 2022 to March 2023, and the accuracy of the models in predicting H5N1 outbreaks in March 2023 is about 99.348%. The ridge & lasso regression model performed the best among the four models, with an AUC of 0.7959950. The map generated based on the ridge & lasso regression model indicated that counties in the north and west were at a higher risk of having H5N1 outbreaks in March 2023, which matched the actual result. The report concludes that there are limitations to the models, including the consideration of only a limited set of factors affecting the spread of the virus and the use of historical data. Future work could incorporate additional data sources and use more sophisticated machine learning techniques to improve the accuracy of the models. The report also proposes some possible remedies to help control the spread of H5N1.

1 Introduction

Poultry, such as chicken, turkey, goose, duck, and others, are staple foods on our dinner tables. According to the United States Department of Agriculture (USDA) in 2022, each person had access to 68.1 pounds of chicken for consumption in 2021. This indicates that chicken is the most popular meat in the United States, and per capita egg consumption has increased by 15% in the past 20 years (UEP, 2021). However, like humans, poultry can also be infected with viruses, and in the context of the COVID-19 pandemic, we are reminded of how a small virus can have a significant impact on our lives. Bird flu caused by the H5N1 virus is one such example, and highly pathogenic avian influenza (HPAI) A(H5) viruses have been detected since January 2022 in U.S. wild aquatic birds, commercial poultry, and backyard or hobbyist flocks (CDC, 2023).

The H5N1 virus can have severe effects, and its outbreak has already caused economic, ecological, and environmental consequences with long-term effects. For example, the price of a dozen large Grade A eggs has more than doubled in 2022 in the United States (Iacurci, 2023). Moreover, sometimes grocery stores run out of eggs due to the virus, making it difficult for millions of people in the U.S. to maintain their usual levels of egg and poultry consumption. The virus has affected more than 58.7 million poultry in 47 states and about 7,098 wild birds in 50 states and 1,022 counties (CDC, 2023), posing incalculable risks to our ecological environment and the poultry industry.

Avian influenza is not a new occurrence, and its effects on humans have been long-lasting and severe since its discovery in the 1880s. The H1N1 virus of avian influenza, for instance, caused 50 million deaths in 1918 (CDC, 2019), and the H5N1 virus has infected 868 people and caused 457 deaths since 2003, according to the World Health Organization (WHO) in 2018. Therefore, this virus not only affects people's food consumption but also their health.

Given the economic, ecological, environmental, and health effects of avian influenza, we aim to perform analyses on its cases to provide predictions and suggestions for reducing its negative impacts. We will use mathematical and statistical methods to determine which counties are more likely to be infected by the H5N1 virus and should thus implement more countermeasures. We will also conduct visualizations and analyses based on the datasets provided by various authoritative organizations and institutions such as the CDC, USDA, U.S. Census Bureau, and the Bureau of Labor Statistics.

The objective of this report is to develop and evaluate machine learning models to predict the outbreak of the H5N1 virus in the United States in the future. Our report will focus on analyzing data from past outbreaks to build models that can accurately predict the likelihood of future outbreaks in different regions of the country. By identifying high-risk areas and providing actionable insights, we hope to contribute to efforts to mitigate the impact of the H5N1 virus and protect public health.

2 Data Description and Visualization

To develop a predictive model for identifying counties that might be at risk of H5N1 infection in the future, we need to understand the structure and content of our data. Our approach involves merging four datasets to create a single, curated dataset that contains information on reported H5N1 cases in each county during a specific month, from January 2022 to March 2023.

The cleaned dataset will be used to train our classification model to predict which counties might be likely to experience H5N1 infection in the upcoming month. By analyzing patterns in the data, we can identify key variables that are correlated with an increased risk of infection, such as location, temperature, and flock type. This information will help us develop targeted interventions

and public health strategies to mitigate the spread of H5N1 in high-risk areas.

2.1 United States Counties Database

This public dataset is provided by Pareto Software, LLC, who builds it from the ground up using authoritative sources such as the U.S. Census Bureau and the Bureau of Labor Statistics. It contains all 3,143 county names, their Federal Information Processing Standard (FIPS) codes, longitudes, and latitudes with respect to the 51 states, including Washington D.C., in the United States in 2023.

We made some changes to this dataset for future convenience. Specifically, we changed the state full names to their abbreviations. This dataset makes it possible to generate a detailed geographical report of H5N1 cases in each county in the United States by matching observations in the latter datasets provided by the CDC.

There are 3,143 observations (counties) and 5 variables after the modification, as shown in Table 1 below.

Table 1: First 6 Observations of United States Counties Database

FIPS Code	**State**	**County**	**Latitude**	**Longitude**
6037	CA	Los Angeles County	34.3209	-118.2247
17031	IL	Cook County	41.8401	-87.8168
48201	TX	Harris County	29.8578	-95.3936
4013	AZ	Maricopa County	33.3490	-112.4915
6073	CA	San Diego County	33.0343	-116.7350
6059	CA	Orange County	33.7031	-117.7609

2.2 H5N1 Bird Flu Detections across the United States (Backyard and Commercial)

The second public dataset is about H5N1 bird flu outbreaks involving commercial poultry facilities, backyard poultry, and hobbyist bird flocks by county in the United States.

We made some modifications to this dataset for future convenience. Since the original dataset includes all records with detection dates, which is too specific, we split the date, year, and month in each observation for future data cleaning and analysis. We also excluded cases that occurred after March 31st, 2023, as the later data is not complete enough.

Moreover, we generalized flock types into `Poultry` and `Non-Poultry`. Originally, there were 15 commercial flock types in addition to `Poultry` and `Non-Poultry`. However, as defined by the World Organization for Animal Health (WOAH) on March 8th, 2022, poultry includes "all birds reared or kept in captivity for the production of any commercial animal products or for breeding for this purpose, fighting cocks used for any purpose, and all birds used for restocking supplies of game or for breeding for this purpose, until they are released from captivity." Therefore, we categorized all these 15 commercial flock types as poultry.

After the modification, there are 817 observations and 6 variables in this dataset. It will play a magnificent role in our future analyses.

Table 2: First and Last 5 Outbreaks in the United States till Mar. 31st 2023 (Backyard and Commercial)

State	County	Year	Month	Day	Type	Cases
Indiana	Dubois County	2022	02	08	Poultry	29000
Virginia	Fauquier County	2022	02	12	Non-Poultry	90
Kentucky	Fulton County	2022	02	12	Poultry	231400
Kentucky	Webster County	2022	02	15	Poultry	53300
Indiana	Dubois County	2022	02	16	Poultry	26600
...	...	...	...	...	...	...
Michigan	Lapeer County	2023	03	23	Poultry	950
Colorado	Arapahoe County	2023	03	24	Non-Poultry	10
Kansas	Ellsworth County	2023	03	24	Non-Poultry	50
Colorado	Yuma County	2023	03	28	Poultry	310
Oregon	Umatilla County	2023	03	30	Non-Poultry	50

Table 2 shows the first and last 5 H5N1 backyard and commercial outbreaks in the United States till March 31st 2023. We can see that the first outbreak happened on February 8th, 2022 in Dubois, Indiana with 29,000 cases and its outbreak type was `Poultry`. The last outbreak happened on March 30th, 2023 in Umatilla, Oregon with 50 cases and its outbreak type was `Non-Poultry`.

2.3 H5N1 Bird Flu Detections across the United States (Wild Birds)

This public dataset contains information about detections of highly pathogenic avian influenza (HPAI) A(H5) viruses in wild birds by county in the United States.

We have also made some modifications to this dataset for future convenience. Similar to the dataset in section 2.2, we have changed the format of dates and excluded cases that occurred after March 31st, 2023. We have also changed the column names to match the previous dataset for future data cleaning and analysis.

After the modification, there are 2,752 observations and 6 variables in this dataset. It is another dataset that will play a magnificent role in our future analyses.

Note that there are two outbreak types: `Wild bird`, which means "an animal that has a phenotype unaffected by human selection and lives independently without requiring human supervision or control (WOAH, 2022)," and `Captive wild bird`, which means "an animal that has a phenotype not significantly affected by human selection but that is captive or otherwise lives under or requires human supervision or control (WOAH, 2022)."

Table 3: First and Last 5 Outbreaks in the United States till Mar. 31st 2023 (Wild Birds)

State	County	Year	Month	Day	Type	Cases
North Carolina	Hyde County	2022	01	12	Wild bird	2
South Carolina	Colleton County	2022	01	13	Wild bird	2
North Carolina	Hyde County	2022	01	16	Wild bird	2
North Carolina	Hyde County	2022	01	20	Wild bird	3
North Carolina	Pamlico County	2022	01	20	Wild bird	34
...	...	...	...	...	...	...
Alaska	Sitka County	2023	03	31	Wild bird	1
Maryland	Harford County	2023	03	31	Wild bird	1
Minnesota	Wright County	2023	03	31	Captive wild bird	1
Utah	Millard County	2023	03	31	Wild bird	1
Washington	Benton County	2023	03	31	Wild bird	1

Table 3 shows the first and last 5 H5N1 wild bird outbreaks in the United States till March 31st 2023. We can see that the first outbreak happened on January 12th, 2022 in Hyde, North Carolina with 2 cases and its outbreak type was `Wild bird`. The last outbreak happened on March 31st, 2023 in Benton, Washington with 1 case and its outbreak type was also `Wild bird`.

2.4 Monthly Average Temperature of each County across the United States

This dataset is combined from the public datasets provided by the National Centers for Environmental Information (NCEI), which provides the average temperature in Fahrenheit degrees (°F) for all counties, except those in the state of Hawaii, from January 2022 to March 2023, and Cedar Lake Ventures, Inc., which provides the average temperature in Fahrenheit degrees (°F) for all five counties in the state of Hawaii from January 2022 to March 2023.

We have changed the formats and column names of state, county, and month for future convenience. Additionally, we have corrected some mismatched county names in this dataset based on dataset in section 2.1. Moreover, since the average temperature data for Hawaii is not available in any offline format, we have manually filled in those values. We change the months January 2022 to March 2023 to month index from 1 to 15, which makes our future analyses easier.

After the modification, there are 47,160 observations and 4 variables in this dataset. This dataset provides an important factor in our analysis and prediction model.

Table 4: First 6 Observations of Average Temperature by County in °F across the United States

State	County	Month Index	Average Temperature
AL	autauga county	1	45.1
AL	baldwin county	1	50.1
AL	barbour county	1	45.4
AL	bibb county	1	43.2
AL	blount county	1	41.6
AL	bullock county	1	44.9

2.5 Monthly H5N1 Cases by County from Jan. 2022 to Mar. 2023 in the United States

This cleaned data is combined and derived from the datasets in 2.1 - 2.4 mentioned above, which is the major dataset we will use in the rest of this report. There are 188,580 observations and 10 variables.

- `fips`: Each FIPS code represents a unique county in the United States, so it is a categorical variable with 3,143 unique values, each unique value has 60 entries.
- `state`: Abbreviation of each state in the United States, so it is a categorical variable with 51 unique values, each states has its own number of counties.
- `county`: The names of counties, independent cities, census areas, and same administrative level regions in the United States, so it is a categorical variable with 3,143 unique values with respect to states, each unique value has 60 entries (note that some sates have some counties with the same name).
- `lat`: The latitude of each county.
- `lng`: The longitude of each county.

- `month.index`: The order of month of avian influenza outbreak from 1 (January 2022) to 15 (March 2023). Each `month.index` has 12,572 entries. Every `month.index` has the same number of entries because the cleaned dataset contains all counties' H5N1 situations regardless of how many cases they have, if there is no case in a county, then the case number is just 0.
- `type`: The type of outbreak in a specific county and month, so it is a categorical variable with 4 unique values, including `poultry` (47,145 entries), `non-poultry` (47,145 entries), `wild bird` (47,145 entries), and `captive wild bird` (47,145 entries). Every type has the same number of entries because the cleaned dataset contains all counties' H5N1 situations regardless of how many cases they have, if there is no case in a county, then the case number is just 0.
- `avg.temp`: The average temperature in a specific county and month in Fahrenheit degree (°F).
- `cases`: The number of H5N1 cases detected in a specific county and month.
- `binary.case`: If the case of a type of outbreak in a specific county and month is 0, then it is marked as `uninfected` (185,907 entries). Otherwise, it is marked as `infected` (2,673 entries).

It is important to visualize the data to understand the patterns and trends that are present in the datasets before building models and doing analyses.

Figure 1: New H5N1 Cases each Month by County from Jan. 2022 to Mar. 2023

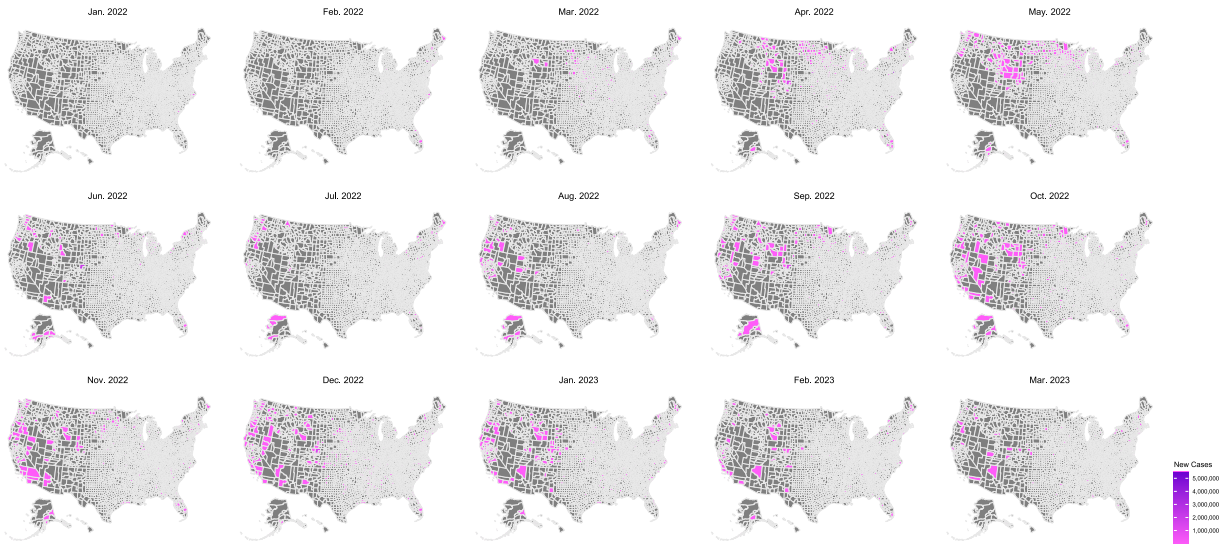


Figure 1 indicates that as the months went by, there were more new cases of the H5N1 virus. The majority of the new cases were in the west and midwest regions. There was a fluctuation of new cases in March, April, May, and from August to the end of the year of 2022. Based on the colors, new cases did not exceed 1,000,000 in March 2023. Moreover, monthly cases were decreasing since 2023.

Table 5: Most and Least 5 Monthly Cases by County in the United States from Jan. 2022 to Mar. 2023

FIPS code	State	County	Month Index	Type	Cases
19021	IA	buena vista county	3	poultry	5486700
19143	IA	osceola county	3	poultry	5011700
42071	PA	lancaster county	4	poultry	3782700
39039	OH	defiance county	9	poultry	3748500
55055	WI	jefferson county	3	poultry	2750700
...	...	...	...	...	...
56039	WY	teton county	10	wild bird	1
56039	WY	teton county	6	wild bird	1
56039	WY	teton county	9	wild bird	1
56043	WY	washakie county	13	wild bird	1
56043	WY	washakie county	14	wild bird	1

Table 5 shows the most and least 5 monthly cases by county in the United States till March 31st 2023. We can see that the county that has the most monthly cases was Buena Vista, Iowa with 5,486,700 in March 2022. Its outbreak type is `poultry`. The 5 counties that has the least monthly cases are all in Wyoming with only 1 `wild bird` case each.

Figure 2: Cumulative Cases each Month by County from Jan. 2022 to Mar. 2023

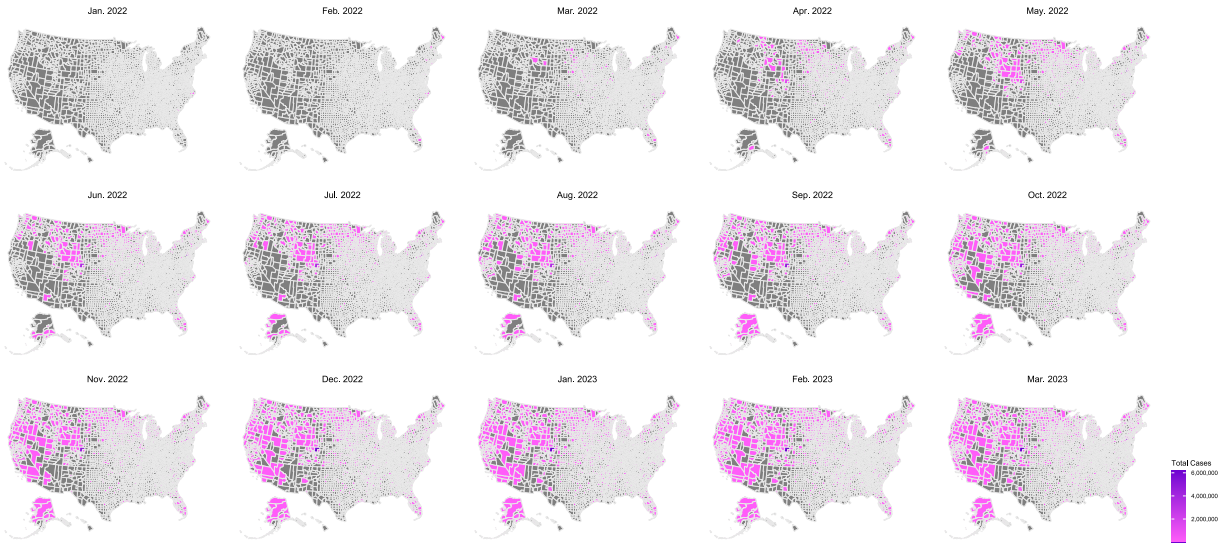


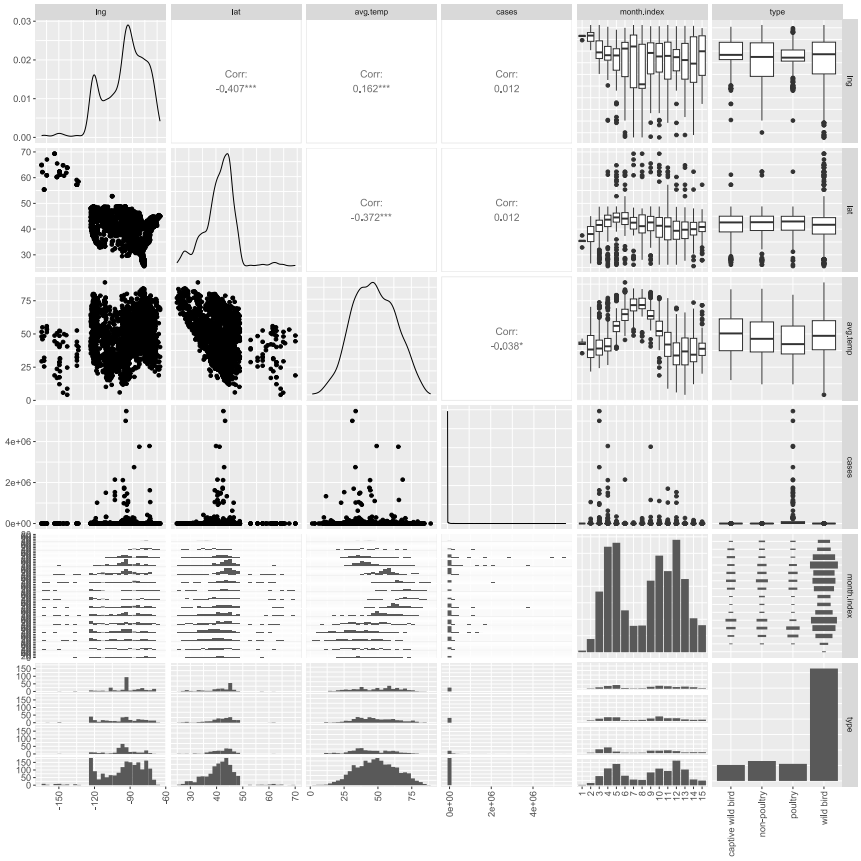
Figure 2 above shows the cumulative cases of the H5N1 virus from January 2022 to March 2023. By the time it was March 2023, most of the United States had cases of this virus. However, most of the cumulative cases are fewer than 2,000,000.

Table 6: Most and Least 5 Cumulative Cases by County in the United States till March 31st 2023

FIPS code	State	County	Cases
8123	CO	weld county	6188790
19021	IA	buena vista county	5606301
19143	IA	osceola county	5011700
42071	PA	lancaster county	3855188
39039	OH	defiance county	3748500
...	...	...	...
55127	WI	walworth county	1
55135	WI	waupaca county	1
56003	WY	big horn county	1
56037	WY	sweetwater county	1
56043	WY	washakie county	1

Table 6 shows the most and least 5 cumulative cases by county in the United States till March 31st 2023. We can see that the county has the most cumulative cases is Weld, Colorado with 6,188,790 cases. Notice that Buena Vista and Osceola in Iowa also have a lot of cumulative cases, 5,606,301 and 5,011,700 respectively. The counties have the least 5 cumulative cases are all in Wisconsin and Wyoming with only 1 case each.

Figure 3: Scatterplot Matrix



In figure 3, the x-axis represents the variables of the columns and the y-axis represents the variables of the rows.

Among the numerical variables, latitude and longitude have the highest correlation. The correlation of -0.407 indicates that these two variables have a moderate negative relationship with each other. When the latitude is between 35 and 45, there is a surge in cases. Meanwhile, when longitude is between -120 and -70, cases increase dramatically. They indicate the locations in the United States that have a larger number of cases.

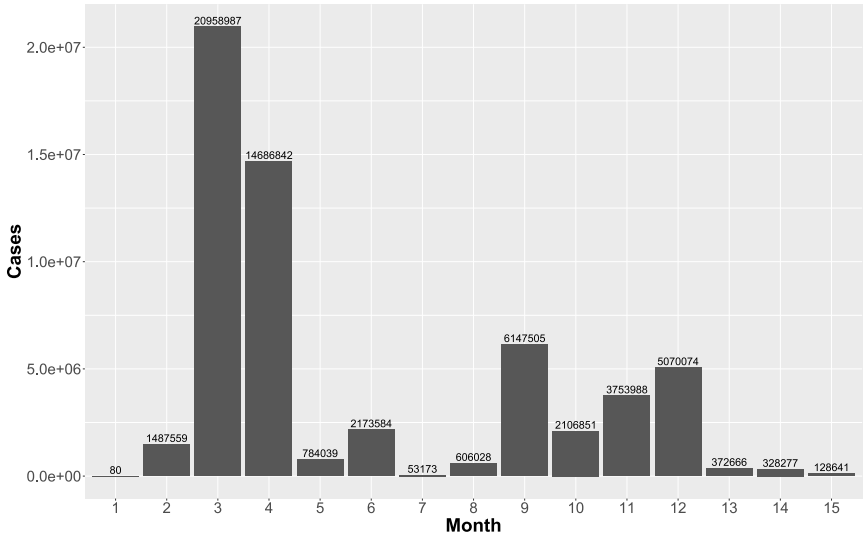
In addition, we can notice that the correlation between average temperature and latitude is -0.372, which makes sense, as the latitude increases (close to the north pole) the average temperature decreases.

The type variable shows that most of the H5N1 cases in the dataset are wild birds.

Furthermore, the `cases` variable indicates that there are not a lot of cases in each outbreak, yet there are many outliers. Possible reasons for this are that although wild birds make up most of the dataset, poultry are in large groups while wild birds are not. Since viruses spread more easily through close contact, most of the cases are poultry.

Another interesting phenomenon is shown in the scatterplot as x-axis is `avg.temp` y-axis is `cases`. Seems the distribution of cases against average temperature looks like a normal distribution. Moreover, there seems to have larger cases when the average temperature is between 30 and 60 degrees Fahrenheit.

Figure 4: Bar Chart of New H5N1 Cases Each Month

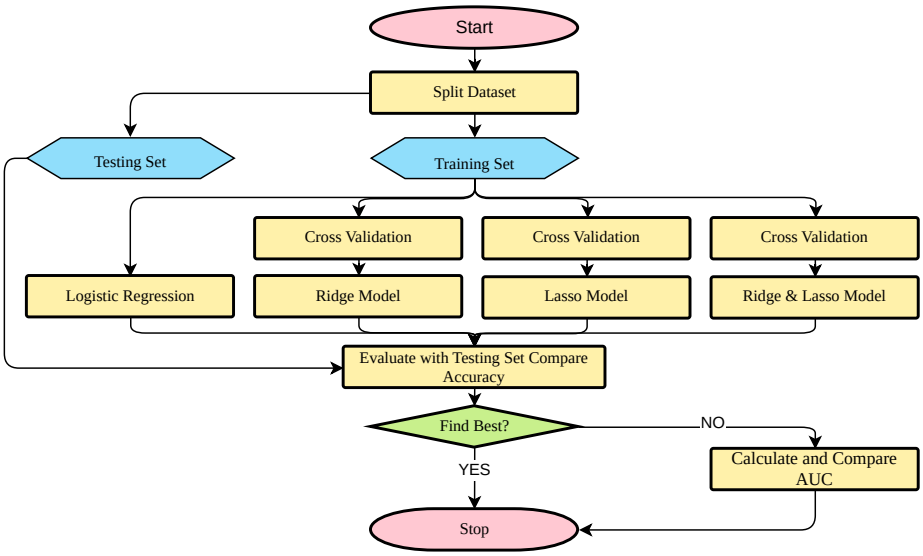


The bar chart as figure 4 shown above indicates the H5N1 cases happened each month. As we can see, March 2022 has the most cases, recall that Buena Vista county, Iowa had the highest monthly cases in table 5. Since 2023, the H5N1 cases became less and less.

3 Methods

Figure 5 below is a flowchart that shows our general approaches and steps to build, evaluate, and select a model that performs the best in predicting which counties may have H5N1 outbreak(s) in the upcoming month.

Figure 5: Flowchart of Steps and Approaches



We use the outbreaks from January 2022 to February 2023 as our training set, which has 176,008 observations. Moreover, we let the outbreaks happened in March 2023 as the testing set, which has 12,572 observations. The testing set will tell us how well our model performs on predicting which county will have H5N1 cases.

Our model is

$$\begin{aligned} Y =& \beta_0 + \beta_1 X_{\text{lat}} + \beta_2 X_{\text{lng}} + \beta_3 X_{\text{month.index}} + \beta_4 X_{\text{type(non-poultry)}} \\ &+ \beta_5 X_{\text{type(poultry)}} + \beta_6 X_{\text{type(wild bird)}} + \beta_7 X_{\text{avg.temp}} + \beta_8 X_{\text{lat * lng}}, \end{aligned} \tag{1}$$

where β_0 is the intercept of the model, β_1 to β_8 are the coefficients of explanatory variables X_{lat} to $X_{\text{lat * lng}}$.

Moreover, we need to use the sigmoid function

$$p(\boldsymbol{X}) = \frac{1}{1 + e^{-\boldsymbol{X}\boldsymbol{\beta}}},$$

where

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{1\ \text{lat}} & X_{1\ \text{lng}} & \cdots & X_{1\ \text{lat * lng}} \\ 1 & X_{2\ \text{lat}} & X_{2\ \text{lng}} & \cdots & X_{2\ \text{lat * lng}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{176008\ \text{lat}} & X_{176008\ \text{lng}} & \cdots & X_{176008\ \text{lat * lng}} \end{bmatrix}_{176008 \times 9}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_8 \end{bmatrix}_{9 \times 1}.$$

The sigmoid function guarantees that the predicted probability is in the range $(0, 1)$ and hence allows us to obtain a sensible prediction.

Because this model determines whether a county will have H5N1 case(s) based on each outbreak type, so it is a binary classifier.

Since the result is either ‘infected‘ or ‘uninfected‘, so for each observation, its distribution is a Bernoulli Distribution

$$\text{Bern}(p).$$

Since we have 176,008 observations in the training set, the distribution of Y should be a Binomial Distribution

$$\text{Binomial}(176008,\ p).$$

p is the probability of a county, based on each outbreak type, to have H5N1 case(s), which will be obtained by performing the model with the sigmoid function described above.

Now let us start building models to predict potential H5N1 outbreak(s) in the future.

3.1 Logistic Regression Model

3.1.1 Modeling

We decided to use logistic regression to model our data with the response variable as to whether a specific county will get infected by avian influenza.

The log-likelihood function for logistic regression is

$$\ell(\beta) = \sum_{i=1}^{n} Y_i \log(p(X_i)) + (1 - Y_i) \log(1 - p(X_i)),\ n = 176,008.$$

In order to estimate the parameters of the logistic regression model, we will apply the method of Maximum Likelihood Estimate (MLE), which solves the objective function

$$\hat{\beta} = \arg\max_{\beta} \Big[\sum_{i=1}^{n} Y_i \log(p(X_i)) + (1 - Y_i) \log(1 - p(X_i)) \Big],\ n = 176,008.$$

By performing the logistic regression with the MLE method, we get our estimated coefficients, rounded to five decimals, as table 7 shown below.

Table 7: Estimated Coefficients Generated by Logistic Regression Model

Coefficient	**Estimation**
$\hat{\beta}_0$	18.21486
$\hat{\beta}_1$	-0.2842
$\hat{\beta}_2$	0.0804
$\hat{\beta}_3$	-0.05675
$\hat{\beta}_4$	-0.20253
$\hat{\beta}_5$	-0.05453
$\hat{\beta}_6$	-1.99719
$\hat{\beta}_7$	-0.00434
$\hat{\beta}_8$	-0.00172

As a result, our predicted model is (coefficients rounded to three decimal places)

$$\begin{aligned}\hat{Y} =& 18.215 - 0.284 X_{\text{lat}} + 0.08 X_{\text{lng}} - 0.057 X_{\text{month.index}} - 0.203 X_{\text{type(non-poultry)}} \\ & - 0.055 X_{\text{type(poultry)}} - 1.997 X_{\text{type(wild bird)}} - 0.004 X_{\text{avg.temp}} - 0.002 X_{\text{lat * lng}}.\end{aligned} \tag{2}$$

3.1.2 Interpretation

The coefficient for `lat` (-0.2842) indicates that a one-unit increase in latitude is associated with a negative change in probability that `lat` multiple by 0.7526161, holding all other predictor variables constant.

The coefficient for `lng` (0.0804) indicates that a one-unit increase in longitude is associated with a positive change in probability that `lng` multiplies by 1.0837205, holding all other predictor variables constant.

The coefficient for `month.index` (-0.05675) indicates that a one-unit increase in `month.index` (which represents the month of the year) is associated with a negative change in probability that month multiplies by 0.9448302, holding all other predictor variables constant.

The coefficient for each `type` of outbreak (`poultry`, `non-poultry`, `wild bird`) represents the difference in log odds of the outcome compared to the reference category (in this case, wild bird). The coefficient for `non-poultry` (-0.20253) indicates that `non-poultry` animals are 0.816662 times less likely than the `wild bird`, holding all other predictor variables constant.

The coefficient for `avg.temp` (-0.00434) indicates that a one-unit increase in average temperature is associated with a negative change in the log odds of the outcome by 0.9956694, holding all other predictor variables constant.

The coefficient for the interaction term `lat * lng` (-0.00172) indicates that the effect of latitude on the log odds of the outcome depends on the value of longitude. Specifically, a one-unit increase in latitude is associated with a negative change in the log odds of the outcome by 0.9982815 units for each one-unit increase in longitude.

Overall, this logistic regression model can be used to predict the probability of the binary outcome based on the values of the predictor variables included in the model. The estimated coefficients can also be used to interpret the effects of each predictor variable on the log odds of the outcome, holding all other predictor variables constant.

3.2 Ridge Regression Model for Classification

3.2.1 Modeling

In this section, we will try the ridge regression model for classification. Ridge regression attempt to solve the objective function

$$\hat{\beta}^{\text{ridge}} = \arg\max_{\beta} \Big[\ell(\beta) + \lambda \sum_{i=1}^{n} \beta_i^2 \Big], \ n = 176,008.$$

where $\ell(\beta)$ is the loss function of the original logistic regression model, $\lambda \sum_{i=1}^{n} \beta_i^2$ is the penalty term. λ is called the penalty parameter and $\lambda \in [0, \infty)$.

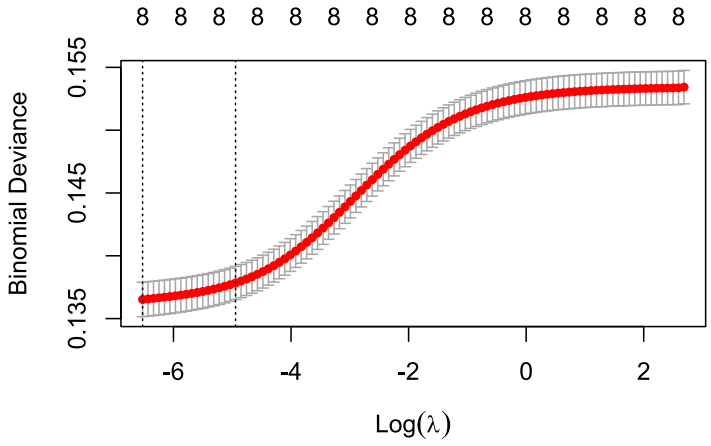
Since we need to find λ to do the ridge regression model, we first use a 10-fold cross validation to find out the λ that yields the smallest deviance

$$CV_{(10)} = \frac{1}{10} \sum_{i=1}^{n} d_i^2 = \frac{1}{10} \sum_{i=1}^{n} 2 \left[Y_i \log\left(\frac{Y_i}{p(X)} \right) + (1 - Y_i) \log\left(\frac{1 - Y_i}{1 - p(X)} \right) \right], \ n = 176,008.$$

Figure 6 below shows different deviances when using different λ's. The dashline on the left side indicates the λ (0.0014659) producing the smallest deviance (0.1365223).

The dashline on the right side of the first one indicates the largest λ (0.0071283) at which the deviance is within one standard error of the smallest deviance.

Figure 6: 10-Fold Cross Validation for Proper λ in Ridge Regression Model



After performing the 10-fold cross validation and getting the value of λ we want, we can start building the ridge regression model for classification, which gives us the estimated coefficients as shown below in table 8.

Table 8: Estimated Coefficients Generated by Ridge Regression Model

Coefficient	**Estimation**
$\hat{\beta}_0$	9.09331
$\hat{\beta}_1$	-0.08512
$\hat{\beta}_2$	0.0063
$\hat{\beta}_3$	-0.04842
$\hat{\beta}_4$	0.0434
$\hat{\beta}_5$	0.16787
$\hat{\beta}_6$	-1.69266
$\hat{\beta}_7$	-0.00105
$\hat{\beta}_8$	-6×10^{-5}

As a result, our predicted model becomes (coefficients rounded to three decimal places)

$$\begin{aligned}\hat{Y} =& 9.093 - 0.085X_{\text{lat}} + 0.006X_{\text{lng}} - 0.048X_{\text{month.index}} + 0.043X_{\text{type(non-poultry)}} \\ &+ 0.168X_{\text{type(poultry)}} - 1.693X_{\text{type(wild bird)}} - 0.001X_{\text{avg.temp}} - 6 \times 10^{-5}X_{\text{lat * lng}}.\end{aligned} \tag{3}$$

3.2.2 Interpretation

The coefficient for `lat` (-0.08512) indicates that a one-unit increase in latitude is associated with a negative change in probability that `lat` multiple by 0.9184021, holding all other predictor variables constant.

The coefficient for `lng` (0.0063) indicates that a one-unit increase in longitude is associated with a positive change in probability that `lng` multiplies by 1.0063199, holding all other predictor variables constant.

The coefficient for `month.index` (-0.04842) indicates that a one-unit increase in `month.index` (which represents the month of the year) is associated with a negative change in probability that month multiplies by 0.9527336, holding all other predictor variables constant.

The coefficient for each `type` of outbreak (`poultry`, `non-poultry`, `wild bird`) represents the difference in log odds of the outcome compared to the reference category (in this case, wild bird). The coefficient for `non-poultry` (0.0434) indicates that non-poultry animals are 1.0443556 times more likely than the wild birds, holding all other predictor variables constant.

The coefficient for `avg.temp` (-0.00105) indicates that a one-unit increase in average temperature is associated with a negative change in the log odds of the outcome by 0.9989506, holding all other predictor variables constant.

The coefficient for the interaction term `lat * lng` (-6×10^{-5}) indicates that the effect of latitude on the log odds of the outcome depends on the value of longitude. Specifically, a one-unit increase in latitude is associated with a negative change in the log odds of the outcome by 0.99994 units for each one-unit increase in longitude.

3.3 Lasso Regression Model for Classification

3.3.1 Modeling

In this section, we will try the lasso regression model for classification. Lasso regression attempt to solve the objective function

$$\hat{\beta}^{\text{lasso}} = \arg\max_{\beta} \Big[\ell(\beta) + \lambda \sum_{i=1}^{n} |\beta_i| \Big], \ n = 176,008.$$

where $\ell(\beta)$ is the loss function of the original logistic regression model, $\lambda \sum_{i=1}^{n} |\beta_i|$ is the penalty term. λ is called the penalty parameter and $\lambda \in [0, \infty)$.

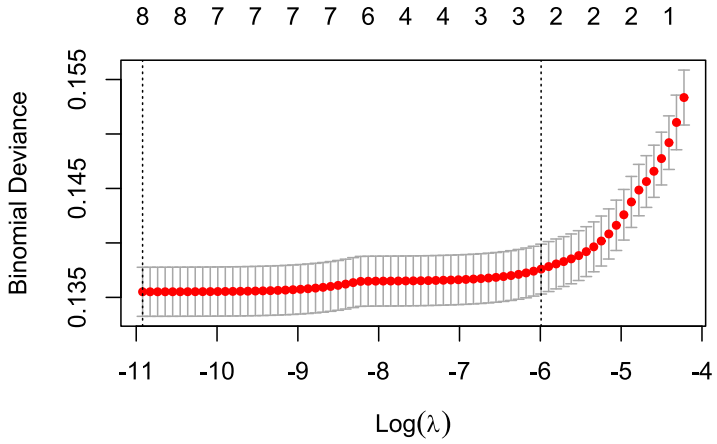
Since we need to find λ to do the ridge regression model, we first use a 10-fold cross validation to find out the λ that yields the smallest deviance

$$CV_{(10)} = \frac{1}{10} \sum_{i=1}^{n} d_i^2 = \frac{1}{10} \sum_{i=1}^{n} 2 \left[Y_i \log\left(\frac{Y_i}{p(X)} \right) + (1 - Y_i) \log\left(\frac{1 - Y_i}{1 - p(X)} \right) \right], \ n = 176,008.$$

Figure 10 below shows different deviances when using different λ's. The dashline on the left side indicates the λ (1.807284×10^{-5}) producing the smallest deviance (0.1355127).

The dashline on the right side of the first one indicates the largest λ (0.0025029) at which the deviance is within one standard error of the smallest deviance.

Figure 7: Cross Validation for Proper λ in Lasso Regression Model



After performing the 10-fold cross validation and getting the value of λ we want, we can start building the lasso regression model for classification, which gives us the estimated coefficients as shown below in table 9.

Table 9: Estimated Coefficients Generated by Lasso Regression Model

Coefficient	**Estimation**
$\hat{\beta}_0$	17.38428
$\hat{\beta}_1$	-0.2659
$\hat{\beta}_2$	0.07337
$\hat{\beta}_3$	-0.05613
$\hat{\beta}_4$	-0.18213
$\hat{\beta}_5$	-0.03353
$\hat{\beta}_6$	-1.97957
$\hat{\beta}_7$	-0.00408
$\hat{\beta}_8$	-0.00157

As a result, our predicted model becomes (coefficients rounded to three decimal places)

$$\begin{aligned}\hat{Y} =& 17.384 - 0.266X_{\text{lat}} + 0.073X_{\text{lng}} - 0.056X_{\text{month.index}} - 0.182X_{\text{type(non-poultry)}} \\ & - 0.034X_{\text{type(poultry)}} - 1.98X_{\text{type(wild bird)}} - 0.004X_{\text{avg.temp}} - 0.002X_{\text{lat * lng}}.\end{aligned} \tag{4}$$

3.3.2 Interpretation

The coefficient for `lat` (-0.2659) indicates that a one-unit increase in latitude is associated with a negative change in probability that `lat` multiple by 0.7665158, holding all other predictor variables constant.

The coefficient for `lng` (0.07337) indicates that a one-unit increase in longitude is associated with a positive change in probability that `lng` multiplies by 1.0761286, holding all other predictor variables constant.

The coefficient for `month.index` (-0.05613) indicates that a one-unit increase in `month.index` (which represents the month of the year) is associated with a negative change in probability that month multiplies by 0.9454162, holding all other predictor variables constant.

The coefficient for each `type` of outbreak (`poultry`, `non-poultry`, `wild bird`) represents the difference in log odds of the outcome compared to the reference category (in this case, `wild bird`). The coefficient for `non-poultry` (-0.18213) indicates that `non-poultry` animals are 0.833493 times less likely than the `wild birds`, holding all other predictor variables constant.

The coefficient for `avg.temp` (-0.00408) indicates that a one-unit increase in average temperature is associated with a negative change in the log odds of the outcome by 0.9959283, holding all other predictor variables constant.

The coefficient for the interaction term `lat * lng` (-0.00157) indicates that the effect of latitude on the log odds of the outcome depends on the value of longitude. Specifically, a one-unit increase in latitude is associated with a negative change in the log odds of the outcome by 0.9984312 units for each one-unit increase in longitude.

3.4 Ridge & Lasso Regression Model for Classification

3.4.1 Modeling

In this section, we will try to use the combination of ridge and lasso regression model for classification. The Ridge & Lasso regression attempt to solve the objective function

$$\hat{\beta}^{\text{ridge \& lasso}} = \arg\max_{\beta}\Big[\ell(\beta) + \lambda\Big(\frac{1}{2}\sum_{i=1}^{n}|\beta_i| + \frac{1}{2}\sum_{i=1}^{n}\beta_i^2\Big)\Big],\ n = 176,008.$$

where $\ell(\beta)$ is the loss function of the original logistic regression model, $\lambda\Big(\frac{1}{2}\sum_{i=1}^{n}|\beta_i| + \frac{1}{2}\sum_{i=1}^{n}\beta_i^2\Big)$ is the penalty term. λ is called the penalty parameter and $\lambda \in [0, \infty)$.

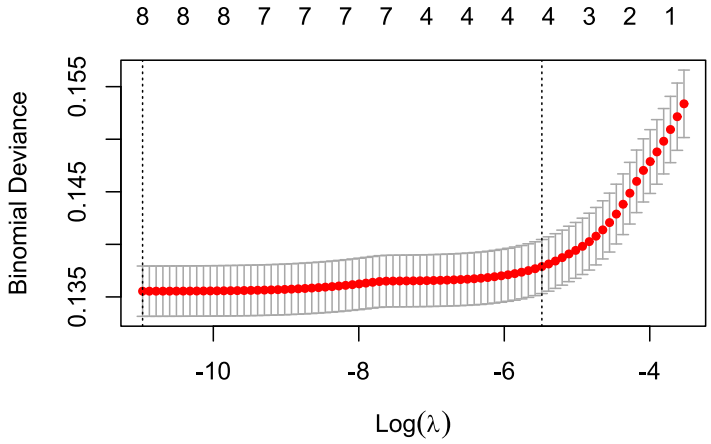
Since we need to find λ to do the ridge & lasso regression model, we first use a 10-fold cross validation to find out the λ that yields the smallest deviance

$$CV_{(10)} = \frac{1}{10}\sum_{i=1}^{n} d_i^2 = \frac{1}{10}\sum_{i=1}^{n} 2\left[Y_i \log\left(\frac{Y_i}{p(X)}\right) + (1 - Y_i)\log\left(\frac{1 - Y_i}{1 - p(X)}\right)\right],\ n = 176,008.$$

Figure 11 below shows different deviances when using different λ's. The dashline on the left side indicates the λ (1.7172127×10^{-5}) producing the smallest deviance (0.1355458).

The dashline on the right side of the first one indicates the largest λ (0.0041559) at which the deviance is within one standard error of the smallest deviance.

Figure 8: Cross Validation for Proper λ in Ridge & Lasso Regression Model



After performing the 10-fold cross validation and getting the value of λ we want, we can start building the ridge & lasso regression model for classification, which gives us the estimated coefficients as shown below in table 10.

Table 10: Estimated Coefficients Generated by Ridge & Lasso Regression Model

Coefficient	**Estimation**
$\hat{\beta}_0$	17.34946
$\hat{\beta}_1$	-0.26485
$\hat{\beta}_2$	0.07299
$\hat{\beta}_3$	-0.05629
$\hat{\beta}_4$	-0.19021
$\hat{\beta}_5$	-0.04208
$\hat{\beta}_6$	-1.98563
$\hat{\beta}_7$	-0.00411
$\hat{\beta}_8$	-0.00156

As a result, our predicted model becomes (coefficients rounded to three decimal places)

$$\begin{aligned}\hat{Y} =& 17.349 - 0.265X_{\text{lat}} + 0.073X_{\text{lng}} - 0.056X_{\text{month.index}} - 0.19X_{\text{type(non-poultry)}} \\ & - 0.042X_{\text{type(poultry)}} - 1.986X_{\text{type(wild bird)}} - 0.004X_{\text{avg.temp}} - 0.002X_{\text{lat * lng}}.\end{aligned} \tag{5}$$

3.4.2 Interpretation

The coefficient for `lat` (-0.26485) indicates that a one-unit increase in latitude is associated with a negative change in probability that `lat` multiple by 0.767321, holding all other predictor variables constant.

The coefficient for `lng` (0.07299) indicates that a one-unit increase in longitude is associated with a positive change in probability that `lng` multiplies by 1.0757198, holding all other predictor variables constant.

The coefficient for `month.index` (-0.05629) indicates that a one-unit increase in `month.index` (which represents the month of the year) is associated with a negative change in probability that month multiplies by 0.945265, holding all other predictor variables constant.

The coefficient for each `type` of outbreak (`poultry`, `non-poultry`, `wild bird`) represents the difference in log odds of the outcome compared to the reference category (in this case, wild bird). The coefficient for `non-poultry` (-0.19021) indicates that `non-poultry` animals are 0.8267855 times less likely than the `wild birds`, holding all other predictor variables constant.

The coefficient for `avg.temp` (-0.00411) indicates that a one-unit increase in average temperature is associated with a negative change in the log odds of the outcome by 0.9958984, holding all other predictor variables constant.

The coefficient for the interaction term `lat * lng` (-0.00156) indicates that the effect of latitude on the log odds of the outcome depends on the value of longitude. Specifically, a one-unit increase in latitude is associated with a negative change in the log odds of the outcome by 0.9984412 units for each one-unit increase in longitude.

4 Results

Now we have four models to determine whether a county will have a specific type of outbreak in the future, which are logistics regression model, ridge regression model for classification, lasso regression model for classification, and ridge & lasso regression model for classification. One of the best ways to see how these model performs is to use the testing set to see the accuracy of their predictions.

We apply the testing set, which are the cases happened in March 2023, into our four models and find out that all the models produce the exact same result with the threshold probability of `infected` being 0.5 (i.e. when the predicted probability is larger than 0.5, it is classified as `infected`, otherwise `uninfected`).

These models indicate that all counties are uninfected. The confusion matrix is shown as table 11 below, which represents the counts of all combination of values between the predicted label and the true label.

Table 11: Confusion Matrix of the Test Set

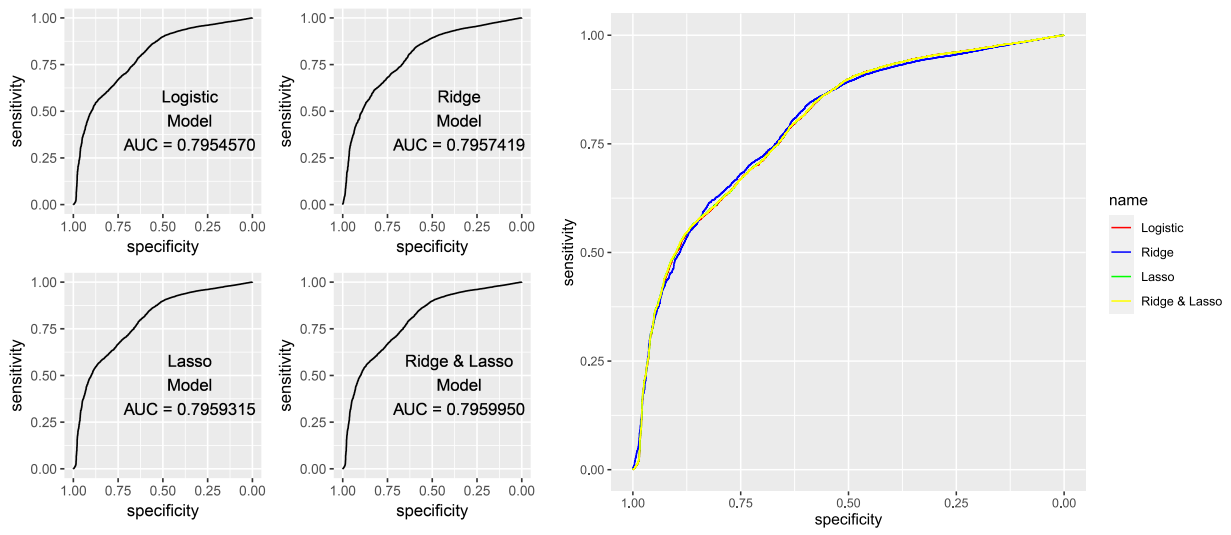
Predicted / True	**Infected**	**Uninfected**
Uninfected	82	12490

By looking at the confusion matrix, we get error rate equals to $\frac{82}{12572} = 0.00652$, and the accuracy equals to $1 - \frac{82}{12572} = 0.99348$, which means that most of the cases, 99.348%, are correctly classified.

Although the accuracy is very high, but that is not what we really want because all these models classify all the 3,143 counties as `uninfected` in March 2023, whereas there were 82 counties had outbreaks, which cannot bring us any useful information and may bring risks to public health.

In order to know which model performs better, we use Receiver Operating Characteristic (ROC) curve, which tests the goodness of fit, and compare the Area Under the Curve (AUC). The range of AUC is $(0, 1)$, where higher AUC means the classifier is better. Figure 9 below shows the ROC curves and AUC of logistic regression, ridge regression, lasso regression, and ridge & lasso regression models for classification.

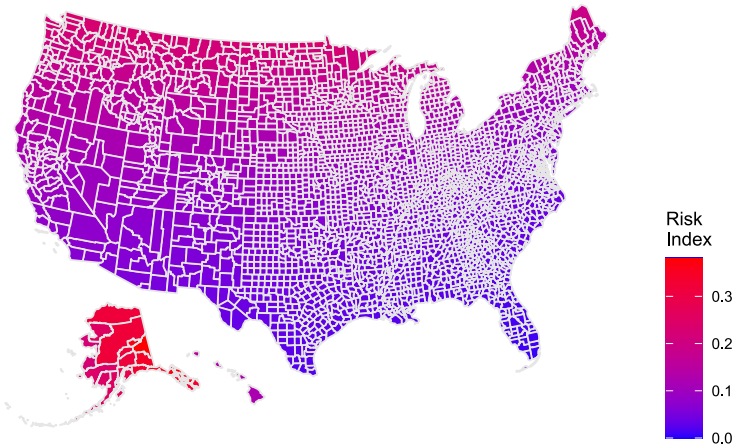
Figure 9: ROC Curves of All Four Models



We can see the all ROC curves look smooth and goes to the left corner, meaning that our classifier works good. By obtaining the value of AUC's, which are 0.7954570 for logistic regression model, 0.7957419 for ridge regression model, 0.7959315 for lasso regression model, and 0.7959950 for ridge & lasso regression model, indicating that most counties are correctly classified as `infected` and `uninfected`. Our detection of H5N1 looks good.

Moreover, since the AUC for the ridge & lasso regression model is the highest, although the difference is not very large, we still consider the ridge & lasso regression model for classification works the best among all the four models. Figure 10 is a map, which plots out the sum of predicted probabilities of all four possible H5N1 outbreak types (`poultry`, `non-poultry`, `wild bird`, and `captive wild bird`) of all counties in March 2023 calculated by the lasso regression model. Note that redder color means higher risk to have an outbreak. We can interpret the sum of probability of four types of a county as its risk index, with range $[0, 4]$, of having an outbreak.

Figure 10: Map of Risk Index Generated by Ridge & Lasso Regression Model



As we can see from figure 10, seem the counties in the south are less likely to have outbreaks

of H5N1 in March, but those counties in the north and west are more risky. This actually makes sense if we compare with figure 1, specifically the new cases in March 2023.

5 Conclusion and Suggestion

In conclusion, we have developed four models, logistic regression, ridge regression, lasso regression, and ridge & lasso regression, to predict the likelihood of H5N1 outbreaks in different counties of the United States. The models were trained using historical data from January 2022 to February 2023, and the results showed that all models had a high accuracy in predicting the occurrence of H5N1 outbreaks in March 2023, which is about 99.348%. However, the models failed to predict the occurrence of 82 H5N1 outbreaks in March 2023, if we set the threshold of infected to be 0.5, which highlights a limitation in our current models.

Our analyses show that the ridge & lasso regression model performed the best among the four models, with an AUC of 0.7959950. The map generated based on the ridge & lasso regression model indicated that counties in the north and west were at a higher risk of having H5N1 outbreaks in March 2023, which matches the actual result.

Despite the high accuracy of our models, there are still some limitations that need to be addressed. One limitation is that our models only considered the effects of temperature, outbreak types, time, and density on the likelihood of H5N1 outbreaks, but there may be other factors that also affect the spread of the virus, such as migration patterns of birds, human movement, egg production, breeding size, and so on. In addition, the models were based on historical data, and the emergence of new H5N1 strains may lead to changes in the spread of the virus that are not accounted for in our models.

To improve our models, we could incorporate additional data sources such as bird migration patterns, human travel data, egg production, breeding size, and so on. We could also use more sophisticated machine learning techniques such as convolutional neural network (CNN) and long short-term memory networks (LSTM) to capture more complex relationships between different variables. We suggest the USDA and CDC to make more detailed data collections, including these factors described above.

In terms of controlling the spread of H5N1, there are several measures that can be taken.

Implement strict biosecurity measures in poultry farms: Clean and disinfect regularly, limit farm access, and separate sick birds. Maintain cleanliness, limit property access, and handle dead birds properly. Buy healthy birds from reputable sources, keep them separate, sanitize equipment, and report sick birds promptly (Canadian Food Inspection Agency, 2013).

Practice good hygiene habits for chicken farmers: Wear masks and work clothes, clean and disinfect clothes, wash hands after contact with dirt, and use gloves when handling chicken farm manure. Minimize contact during an epidemic and use protective gear when handling poultry (Canadian Food Inspection Agency, 2012).

Establish surveillance programs: Monitor virus spread in wild birds and poultry farms to detect outbreaks early and prevent further transmission. Stay vigilant, observe signs of illness or distress, and report sick birds promptly to appropriate authorities (CDC, 2022).

Launch public education campaigns: Raise awareness about H5N1 risks and educate people on preventive measures. Enhancing knowledge can help prevent virus transmission.

Coordinate national and international efforts: Collaborate to track and contain the virus by sharing information and resources across countries. Joint efforts are crucial for effective control.

By implementing these measures, we can help control the spread of H5N1 and prevent future outbreaks.

6 Reference

Canadian Food Inspection Agency (2013, March 10). Government of Canada. Canadian Food Inspection Agency. https://inspection.canada.ca/animal-health/terrestrial-animals/biosecurity/tools/checklist/eng/1362944949857/1362945111651

Canadian Food Inspection Agency (2012, August 12). Government of Canada. Canadian Food Inspection Agency. https://inspection.canada.ca/animal-health/terrestrial-animals/biosecurity/veterinarians/eng/1344822097335/1344822345700

Centers for Disease Control and Prevention (CDC) (2022, October 31). Prevention and antiviral treatment of bird flu viruses in people. Centers for Disease Control and Prevention. https://www.cdc.gov/flu/avianflu/prevention.htm

Centers for Disease Control and Prevention (CDC) (2023, February 22). H5N1 bird flu detections across the United States (backyard and commercial). Centers for Disease Control and Prevention. Retrieved February 27, 2023, from https://www.cdc.gov/flu/avianflu/data-map-commercial.html

Centers for Disease Control and Prevention (CDC) (2023, February 15). H5N1 Bird Flu Detections across the United States (Wild Birds). Centers for Disease Control and Prevention. Retrieved February 27, 2023, from https://www.cdc.gov/flu/avianflu/data-map-wild-birds.html

Centers for Disease Control and Prevention (CDC) (2019, March 20). 1918 pandemic (H1N1 virus). Centers for Disease Control and Prevention. https://www.cdc.gov/flu/pandemic-resources/1918-pandemic-h1n1.html

Iacurci, G. (2023, February 8). Wholesale egg prices have 'collapsed.' why consumers may soon see relief. CNBC. Retrieved February 27, 2023, from https://www.cnbc.com/2023/02/07/wholesale-egg-prices-have-collapsed-from-record-highs-in-december.html

National Centers for Environmental Information (NCEI) (2023). County mapping: Climate at a glance. County Mapping — Climate at a Glance — National Centers for Environmental Information (NCEI). https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/mapping

Pareto Software, LLC (2023). United States Counties Database. simplemaps. https://simplemaps.com/data/us-counties

United States Department of Agriculture (USDA) (2022). Per capita availability of chicken higher than that of beef since 2010. USDA ERS - Chart Detail. Retrieved February 27, 2023, from https://www.ers.usda.gov/data-products/chart-gallery/gallery/chart-detail/?chartId=58312

United Egg Producers (UEP) (2021, March 10). Facts & stats. United Egg Producers. Retrieved February 27, 2023, from https://unitedegg.com/facts-stats/

World Health Organization (WHO) (2018, November 13). Influenza (avian and other zoonotic). World Health Organization. Retrieved February 27, 2023, from https://www.who.int/news-room/fact-sheets/detail/influenza-(avian-and-other-zoonotic)

World Organisation for Animal Health (WOAH) (2022, May 6). Health Standards Glossary WOAH. World Organisation for Animal Health. Retrieved February 27, 2023, from https://www.woah.org/fileadmin/Home/eng/Healthstandards/tahc/current/glossaire.pdf