**Data Analytics Report**

**Topic: When Planes Attack**

**Team Members: Gengzhe Li, Jiaying Ni, Pujiao Zhu, Shuqiao Li**

1. **Questions/Hypotheses**

According to reports, from 1990 to 2016, there are about 179,000 records of wildlife struck aircraft. Wildlife strikes from birds might cause severe damage to aircraft and related costs. Knowing the relationship between damage and strike factors can help people better understand bird-strikes and establish an aggressive protection mechanism.

Federal Aviation Administration Wildlife Strike Database mainly stores reported bird-strikes. Based on the database, we find records of different details related to strikes, such as aircraft, birds, situation and the like. Our primary goal is to find out the factors that can lead to the damage of strikes, and establish the model to predict the probability of damage that might occur in the future.

There are three major steps involved.

(1) The first is to study which factors can contribute to damage. As momentum is decided by mass and velocity, we make the hypothesis that damage is positively related to the speed and mass of both birds and aircraft. According to flying habits and migration characteristics of birds, we also assume that year, month, and sky situation are all significant to damage. Therefore, we select a dozen of the most useful independent variables – Type of engine, Aircraft Mass, Number of ENGS, Month, Year, Time, Height, Speed, Phase, Sky, Precipitation, Birds (Seen), Birds (Struck), and Size – and make assumptions that each variable is statistically significant to the dependent variable – damage.

(2) Notice there are many zeros in the column of “cost”, so we decide to discuss the damage and cost separately. At the second step, we further study which factors could impact the cost after damage. We make hypotheses that these variables above are statistically significant to the dependent variable – cost, and then adopt forward method to derive a linear relationship between cost and them and find an optimal subset, by comparing Cp.

(3) Lastly, we try another method, K-Nearest Neighbors to predict the outcome and compare with the result of logistics regression. In other words, verify our goal and hypotheses by different methods.

1. **Data Description**

The FAA Wildlife Strike Database contains records of reported wildlife strikes since 1990. Strike reporting is voluntary. Therefore, this database only represents the information FAA has received from airlines, airports, pilots, and other sources and the statistical bias is unavoidable. The data URL is: <https://wildlife.faa.gov/databaseSearch.aspx>

The dataset our team selected is the Strike Report of Maryland from Jan 1st, 2000 to Dec 1st, 2018, including more than 2,000 records and 93 original columns. As for most of the variables in the report are the professional technical specifications about the airplanes, we clean the data to keep the most valuable and analytical data columns and omit the incomplete records (with NAs), and then we get the dataset with 1409 records and 20 columns. We will choose different variables to fit in different analytical methods (Check R code part 0).

The basic information of the records includes airports (28 reported in MD), cost, month (1-12) and year (2000-2018). The other codes and explanations are attached as follows:

**Table 1. Explanation of Column Name and Code from FAA**



(Source:[*https://wildlife.faa.gov/databaseSearch.aspx*](https://wildlife.faa.gov/databaseSearch.aspx))

As a matter of fact, the real damage and cost that the wildlife strikes bring are much lower than we expected, so the output variable is whether the strikes cause damage (0-1), the main predictors include engine types and numbers, airplane weight, time (month, year & time of the day), height, speed, phase, sky & precipitation conditions, and birds’ numbers and sizes. And we plan to do logistics regression based on the hypothesis and dataset we have.

1. **Methodologies**

1. Logistics regression (Check R code part 1)

In order to solve our first question, we aim to find out variables that can cause the damage to the airplane. Since the response is qualitative, we choose to conduct a logistics regression between the response (damage or not) and a variety of predictors in our dataset.

After importing the dataset “Wildlife log”, we remove all n/a data and check what the variables are. Then we run logistic regression, in which the response is whether or not the airplane damaged (1= damaged, 0=not damaged), and the predictors are the type of engine of the airplane, the size of the airplane, the number of engines, incident month, year, time of the day, flying height, airplane speed, weather condition, how many birds seen, how many birds struck and the size of the bird. By interpreting the initial model, we are able to pick out predictors that have a statistically significant effect on the output and verify our hypothesis.

2. Estimation accuracy (Check R code part 2)

Base on the optimal logistics regression model we acquired between whether or not the airplane damaged and other predictors, we decide to evaluate the estimation accuracy of this model. In this way, we divide the dataset evenly into training and testing part and run logistic regression on the training data with our predictors. Then we predict the probability that whether or not the airplane damaged in the test data. Then we classify the prediction as “1” if the probability is > 50% and “0” if the probability is < 50%. For the last step, we check the accuracy of this prediction for the test data.

3. Linear regression (Check R code part 3)

In order to find out the linear relationship between damage cost and other variables, we narrow down the dataset to a smaller one in which we only include observations with specific costs. After importing the dataset “Wildlife cost”, we run forward stepwise selection that allows models with up to 15 predictors to find the best subset of variables. Then we plot Cp as a function of how many predictors are in the model and use this information to pick the best model with the minimize Cp. The reason why we choose forward stepwise selection is that the method is less complex than best subset selection and easier to fit than backward stepwise selection.

Once we have the best subset with the method, we can check the variable names in this model and their coefficients. According to the output, we can utilize this model to predict the cost of new observations.

4. K-Nearest Neighbors (Check R code part 4)

To solve our last question, we choose K-Nearest Neighbors, the non-parametric method, to interpret the dataset. Even though we have conducted linear and logistics regression to analyze the dataset, the true relationship between inputs and response may be different from these assumed models. We consider KNN as a necessary step in our analysis to deliver more flexible predictions.

Before we import the dataset, we convert our variables into 0-1 variables and name as a new dataset “Wildlife KNN”. Then we import the data and divide the dataset evenly into training and testing part and run KNN with K = 7 to obtain predictions that whether or not the airplane damaged (1 = damaged, 0 = not damaged) for the test data. After checking the accuracy of the prediction with the real data, we change K to 8, 9, 10, and 11 and repeat the KNN steps to select the optimal K value and compare the accuracy of KNN method and the result of logistics regression in our first question.

5. Cluster (Check R code part 5)

We regroup the data according to the A\_Type (aircraft models) into 30 models, and then count the number of damage records and average the airplane weights. And we use K-means and Hierarchical methods to find 5 clusters by each way and plot the charts.

1. **Results and Conclusion**

According to the Logistics Regression Method, we found out that Aircraft Mass, how many birds were seen, the size of the birds, incident month, incident year, and airplane speed are statistically significant to the dependent variable - damage. This result is out of our expectation since we believe when considering the damage, the “Time” - light conditions (dusk, day, dawn, or night) and the “Sky” - type of cloud cover (no cloud, some cloud, overcast) should be statistically significant to the dependent variable - damage. After our discussion, we interpret the reason that Sky and Time are not statistically significant factors may under the external factors, for example, birds are striking the airplane because of the high speed of the airplanes (high speeds prevent most birds from escaping from flying aircraft and cause damages), but not the light condition. In addition, the powerful airflow from jet air intakes often draws flying birds into the engine, causing bird strikes and damages, it is also not the reason of light/time conditions. There is also a possibility that we are processing the data only limited in Maryland, the sample size is not large enough, which is impossible to fully summarize the situation in the United States. As the factors of sky and light, they could possibly affect another district in the United States.

Through the Estimation Accuracy Method, we check the accuracy of this prediction for the test data and compute the percentage of time we were correct, which is 92.69%.

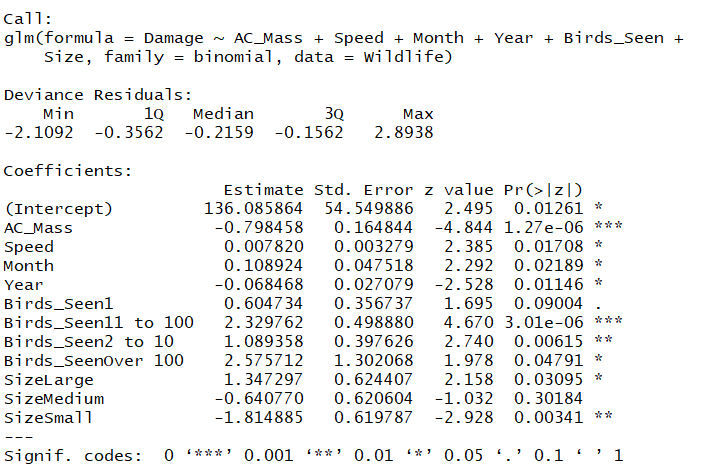
In order to find out the linear relationship between damage cost and other variables, we narrow down the dataset to a smaller one in which we only include observations with specific costs. We finalize the best model with 8 predictors, they are TimeDawn, TimeDusk, Phase Climb, Phase, Take-off-run, Sky No Cloud, Sky Overcast, Size Small, and Size Large.

For the KNN Method, we run KNN with K = 7, 8, 9, 10, and 11. Then we compare the accuracy of each of the result under KNN method, the accuracy when K = 9 is 90.91%, which is the best compared to K = 7, K = 8, K = 10, and K = 11.

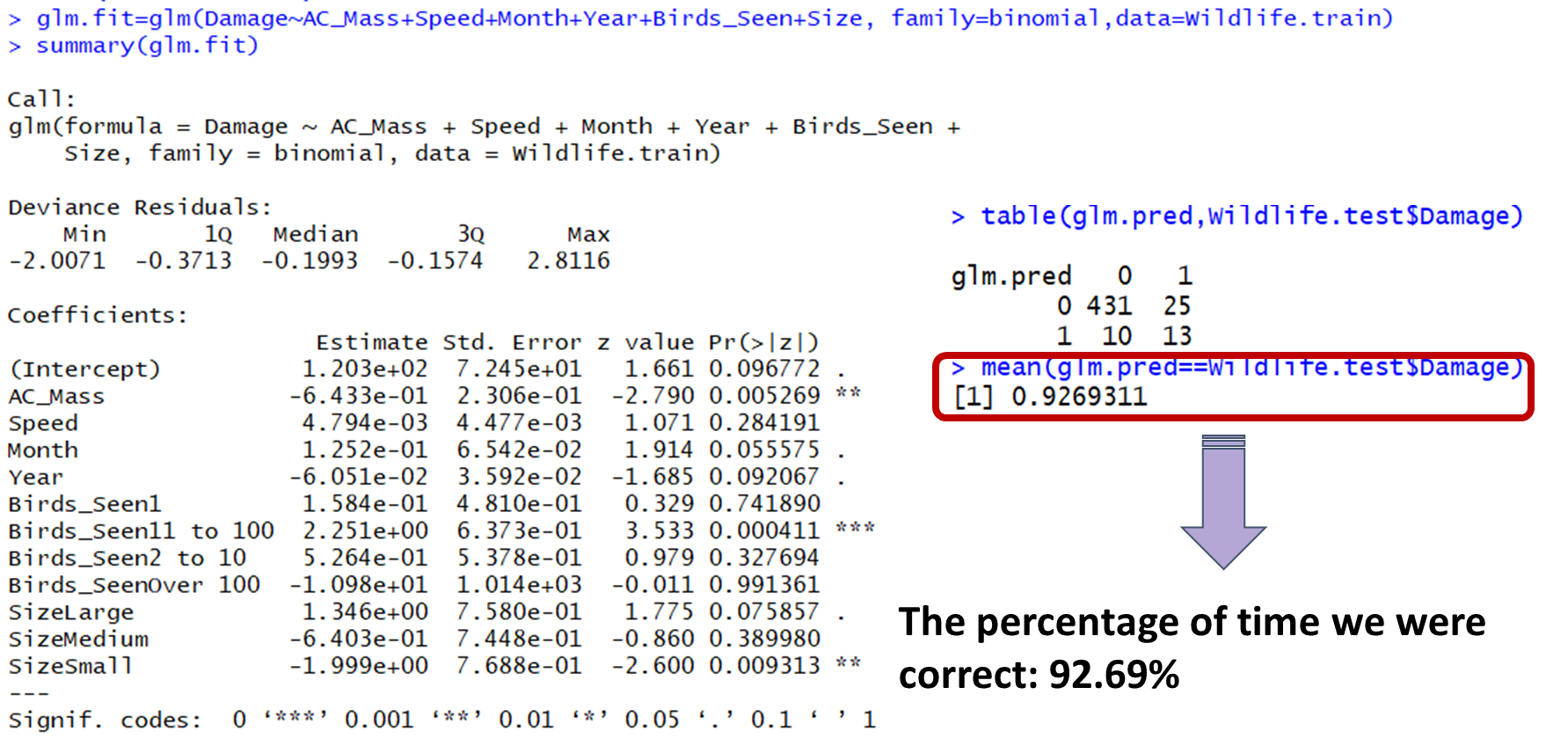
At last, we need to also consider the voluntary report mechanism. It is insufficient because, under the voluntary report mechanism, the data in the database is much less than the actual situation. Plus, self-report will certainly exist a margin of error, for example, people consider the size of the birds may be biased. Another problem with self-reporting is that in order to avoid trouble, any damage that is not obvious or does not directly related to maintenance costs, the data will not reflect. However, those may cause potential damage to the aircraft related to depreciation and future maintenance.

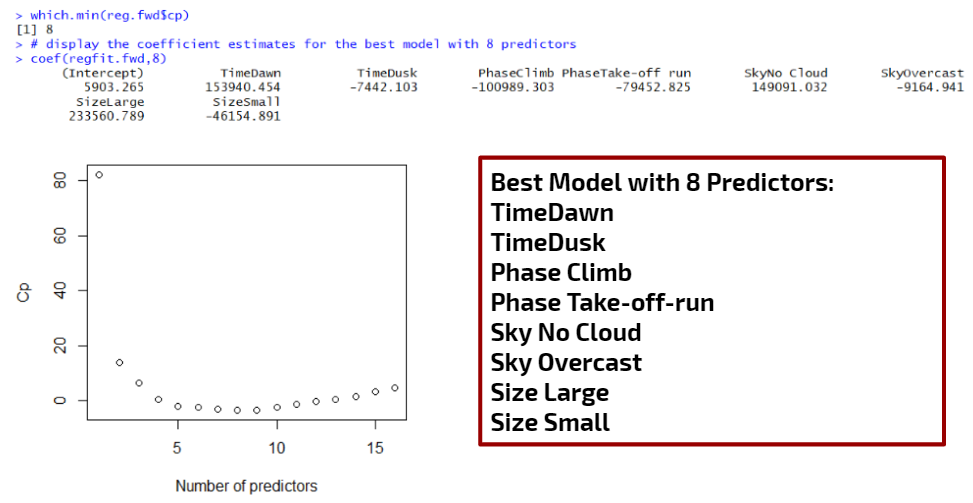
1. **Appendix**

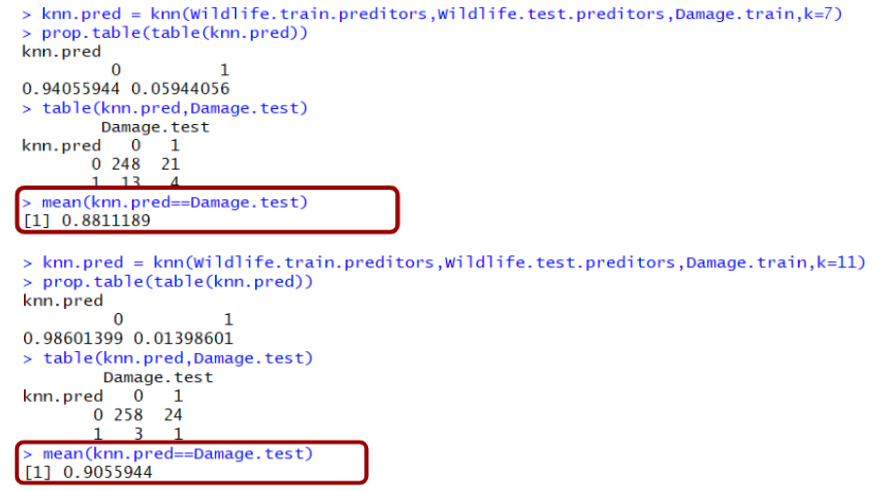
[1] Federal Aviation Administration. (2018). Retrieved from <https://wildlife.faa.gov/databaseSearch.aspx>

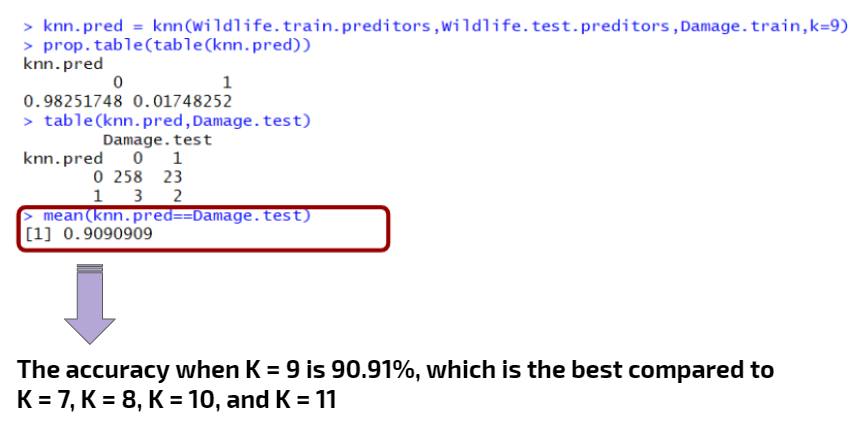
[2] R Results of Logistics Regression

[3] R Results of Estimation Accuracy



**[4] R Results of Linear Regression

**[5] R Results of K-Nearnest Neighbours

**