Report

Introduction

Why was the project undertaken?

We were interested in the subject matter. There have been news reports of airline failures and we were curious about the chances of having an accident while in flight. We first discovered a bird collision dataset while browsing Kaggle.com.

What was the research question, the tested hypothesis or the purpose of the research?

Are birds more likely to collide with commercial or military aircraft during a twelve-month calendar?
This report will allow one to predict the month and the location of collisions between aircraft and birds among a set of airports. The question for our data will be “What is the expected
number of collisions that a given type of plane will collide with a bird in a certain location during a certain month?”
The project will predict the time, location, and type of aircraft that will have a collision with a bird. The project will use three predictors in the dataset: Month of the incident, plane model,and location of incident. The imported columns that will represent this are labeled “Incident Month”, “Aircraft Model”, and “Airport ID”. These three variables will create a graph that will show the monthly likelihood of a collision occurring and enable more interpretation of the incidents and aircraft types.

Selection of Data

What is the source of the dataset?

The data that was used by the Federal Aviation Administration had all of the features that we were looking for. We found it when we were searching in Kaggle for our research question.

Characteristics of data?

The team will use a dataset from Kaggle that includes 174104 rows and 66 columns of Federal Aviation Administration public data. There are columns that show airport locations, types of aircraft, general time of collision, and other aircraft specifics. Columns at the end of the row are related to the specific plane location that was damaged by the collision. Of the 66 columns, much of the data in the set was not necessary due to our choice of what would be predicted.

Any munging or feature engineering?

The data that was found was not immediately ready for accurate analysis and contained some irregular values that do not fit into the model presented. Preprocessing

of the dataset was necessary to visualize predictions without extraneous information. First, the value "ZZZZ" is a special code which is used when no ICAO code exists for the airport. There is also a corresponding airport name column that reflects this fact by using the value "UNKNOWN". 18570 of such rows were excluded with preprocessing.

#Christal and Kyle's data exploration

We originally wanted to use linear regression to analyze relationships between a few categorical columns, but this proved to be a significant challenge when it came to scaling and normalizing our dataframe. About 75 percent of the columns were not integer based and needed to be converted to numeric. This was tricky however, in part because some columns contained string values and there was often no clearly useful numerical form to convert them to. We ended up having to forgo changing the columns to numeric and instead created a slice of the original dataframe containing numeric values.

However this was not the only deciding factor in choosing our second set of columns. It was important to consider whether correlations exist within the data. Correlations exist between "Distance", "Speed", and "Height"; they are also likely to correlate with other columns because of the fact that some types of planes tend to fly higher, faster, or on longer trips than other planes. Specifically, military planes tend to fly at higher altitudes than commercial planes (AircraftCompare.com). There may also exist a distinction between the frequencies of collisions at different altitudes and the types of birds involved.

We chose to utilize linear regression in order to predict the height at which a collision will occur based on the speed of the plane and distance of the trip. We selected the columns "Speed" and "Distance" as predictors because we found them to be correlated with "Height". There ended up being a positive correlation between the distance traveled and speed of the plane.. However, since our dataframe only contains values for "Height" of up to 31,300 feet, we were not able to prove there was correlation between military and commercial planes hitting the birds using "Height" as the target variable.

Instead, we made a copy of the dataframe, dropping any rows which had an NA value in any of the three relevant columns. We specified 'Height' as the target, with "Distance" and "Speed" as the predictors, then split the data into 75% training data and 25% test data. We created the linear regression object and fit the model to the training data.

After inspecting the coefficients to confirm their importance, we plugged the test data into the model and saved the predictions. We then plotted the test data against the predictions as a scatter plot with a line plot superimposed on it.

The r-squared value we got was 0.7 when using our training set. Which is above the average of 0.6. The RMSE of 1330.78 we achieved was also good, as the best value for our dataset is between the values of 0-1000. These results were conclusive with respect to our testing data fitting well to the model.

We also ran a variation of this experiment, adding 'Unique Operator Type Value' as a predictor. Based on the coefficients and R^2 value

Methods

What materials/APIs/tools were used or who was included in answering the research question?
Google Colab
Github
Youtube (demo video)
Kaggle.com
FAA data
No persons were interviewed

Results

What answer was found to the research question; what did the study find?
In the course of the research, the data showed that more birds are hit during the period between July and August.

Was the tested hypothesis true?
Yes, the month of July and August have the highest correlation of monthly incidents between Military and Commercial aircraft. The Commercial aircraft have more overall collision incidents.

Any visualizations?
Yes. Describe or summarize each one in the notebook.
Regression scatterplot:
(characteristics)... actual test values of 'Height' on the y-axis. Predicted values of 'Height' for test data on the x-axis. The more accurate the predictions, the closer the dots are to the line.
Positive correlation.
Test data is 80 percent accurate

Discussion

What might the answer imply and why does it matter?
The chance of hitting a bird in the month of July is substantially higher than three months before or after.
How does it fit in with what other researchers have found?
Each researcher on this team found new methods of estimating the data. Overall common correlations existed across both models, but one was stronger than the other.
What are the perspectives for future research?
Future research in aviation collisions will reveal more data that can be transformed into information.
Survey about the tools investigated for this assignment.
kNN: slide 1
The data from the Federal Aviation Administration was classified into a kNN model. The question being asked was "Are birds more likely to collide with commercial or military aircraft during a twelve-month calendar?" The classification processing done used to predict the month that a bird would hit an aircraft. The first step was to build a visualization based on commercial

and military frequency. The data was sliced for the following predictors: 'Record ID', 'Unique Operator Value', 'Operator', 'Incident Month', and 'Unique Airport Value'. Each of these were numeric values. The Seaborn library was imported into the script to create a violin plot that measured the sample data slice. A distinction between the number of commercial aircraft and the number of military aircraft was measured across one calendar year.

kNN: slide 2

kNN Data Model training:

The preprocessed data was divided into thirty five percent test data vs sixty five percent training data. A kNN classification model was trained using the data split for these categorical numeric pairs. The data was scaled before testing. Early results of the scaled data showed low accuracy on the training model. Different features were attempted to find a more accurate combination to test.

To find the month of the collision, the training and test model used predictors called, 'Unique Operator Type Value', and 'Unique Airport Value', which were numeric representations of military or commercial aircraft and airport location. Early baseline accuracy suggested fourteen percent. Testing was even lower at ten percent.

Summary

Most important findings:

Aircraft that fly in the summer months have a much higher chance of hitting the birds. July through August is the peak of the damage.

Do not forget to include team members' names and **references/bibliography** and a **Github link**.

https://www.kaggle.com/faa/wildlife-strikes