****

**Week 6**

**Capstone Project**

ALY6140 Analytics Systems Technology

Instructor: Professor Zhi He

**Aircraft Bird Strike 2018-2021 Prediction**

By,

Sanjay Lokula

Shashan Mukesh Pande

May 20th, 2021

**Introduction**

In this project, we have taken the data set from the Federal Aviation Administration website. That data set is targeting the bird strike occurrences and the repercussion faced by the airport authority or the pilot flying the aircraft. Here, we will be predicting the prevention of damages due to bird strike, as well as, creating a warning signal that will help pilots and the authority to avoid this damage. The data set is the recent one and is obtained from the FAA Website. There are 44706 Rows and 99 Columns in the data set, which are dated from 1st May 2018 to 1st May 2021.

Within these 4 years, we will have to articulate a model that defines the highest accuracy of a model to predict damages due to bird strike. Some of the attributes in the tables contains the damages caused to which parts of the aircraft? how much did the damage cost? where did the strike occurred? which part of the country, the aircraft was flying to? did the damage occur, what kind of bird species was it? what was its size? how far was the aircraft from the airport? what was its speed? what was the altitude?

Considering these attributes, the data set was prepared, of course There are ample missing values to look at as all 99 attributes are not filled by the administrator. Let us proceed with the steps that we will be taking in the next headings.

**Proposed Methods**

We will be cleaning the data, removing the indenting and the spaces from the data, and then applying three models and then comparing those four models with the reference model that we have taken. We will be predicting the damages caused due to bird strike, which will act as a control system for the airport authority and the pilot. let's see how well we can predict the damage is and with what accuracy.

**Exploratory Data Analysis**

The very first thing that we do is understand the shape of the data, how big is the data set, what are the data types for each attribute- which were different because we had maximum object data types and few integer data types and few floats. The parts of the aircraft that were damaged were either two or false, hence they had a Boolean data type. After understanding the data with the help of the summary function, we will start evaluating the data with the ‘count’ function.

Table

Description automatically generated Table

Description automatically generated Table

Description automatically generated

With the first figure, we identified that Texas had the maximum number of bird strikes, this is also because Texas has the maximum number of airports, and one of the busiest in the United States. In the second figure, we see that Boeing 737-700, which has the highest bird strikes. This is also because most of the plane that we have, are all versions of Boeing 737. In the third figure, we see that usually, all small size birds are the most struck ones. All these values are identified not keeping ‘unknown’ into consideration.

Table

Description automatically generated Table

Description automatically generated

In the chart above, we see that the small size bird strikes are 31042, and the bird strikes have been majorly occurred between the months of July to October.

Chart, box and whisker chart

Description automatically generated

In the figure above, we see that maximum bird strikes have happened either during the landing roll, en-route or descend. We are also able to see the altitude of the flight, and how it has impacted the bird strike.

Chart, scatter chart

Description automatically generated

In the figure above, we see that bird strikes mostly happen either during the takeoff or landing. Majorly it is when the altitude of the aircraft is low, and the distance of the airstrip to the aircraft is also low. Hence, its either take off or landing.

Chart, histogram

Description automatically generated

In the figure above, we articulate that most of the strikes has happened when the aircraft is that a speed of 120 knots to 150 knots. Hence, this tells us that the aircrafts are at lower speed when the strike happens.

Chart

Description automatically generated

In the above figure, we see that there are different engine types. We have this in the data index, A = Piston Engine, B = Turbojet, C = Turboprop, D = Turbofan, E = None (glider), F = Turboshaft (helicopter). The chart shows Turbofan is heavily hit because major commercial flights have Turbofan Engines.

**Experiments- Modeling and Results**



**Cleaning:** The very first thing that we do is remove the index column as it is not required, this is the very first step of cleaning the data. Adding to this, we have also removed all the double spaces and single spaces from some of the records that is that are there in the attributes. The next step is defining the time of the day. As we, have hour stamps given in the data sets, we have defined what time of the day defined what hour. This is as given below:

Graphical user interface, text, application

Description automatically generated

In order to undermine the statistical significance of an independent variable, we are dropping one of the independent variables to get rid of the multicollinearity. Hence, in the first figure, we see that there are three co-dependent variables like speed, height, and distance. Hence, we have removed distanced, to get rid of the multicollinearity. We can clearly see the difference, in the chart below.

Chart

Description automatically generated Chart

Description automatically generated

**Predictive Models**

A part of predictive models will be finding the accurate model to predict the damage is caused by aircrafts, and to classify damaging and non-damaging bird strikes with the models that we have. Will be classifying on the basis of four models; KNN, Random Forest, Logistic Regression and Decision Tree.

Before getting into the model, we will create a reference model with the help of null model. Null model is the accuracy of the model if we choose the most frequent classes or the outcomes. It can be determined by the number of times the model is right to the number of protections that we have.

Text

Description automatically generated with medium confidence

With this, we get the accuracy as 90.23%. Hence, we are keeping 90.23% as a reference accuracy.

A picture containing text

Description automatically generated

**KNN- K-Nearest Neighbors** is a non-parametric classification method, where the object being classified or assigned to the class most common among its k nearest neighbors (k is a positive integer and of a smaller value). Here, when we articulate the code, we get the accuracy of 90.99%. Comparing it to a reference model, the accuracy of KNN is higher.

Text

Description automatically generated with medium confidence

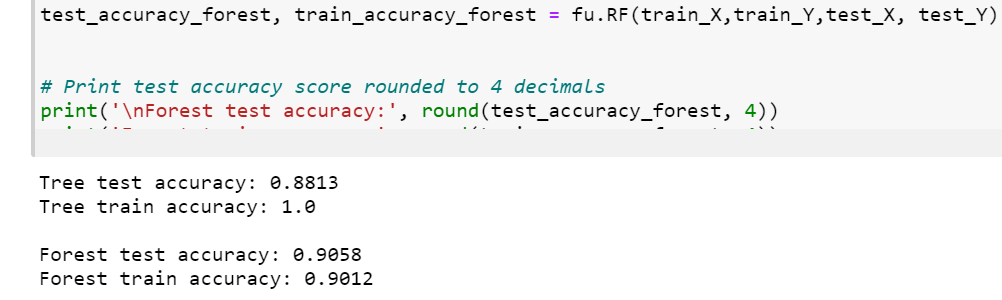
**Random forests** or random decision forests are an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the mode of the classes of the individual trees. Articulating the code, we get the accuracy of 91.83%, which is higher than KNN and the reference model. Hence, up till now, the highest accuracy half classification is with random forest classification model.

Text

Description automatically generated

Hyperparameter tuning of Random Forest model:

Hyperparameter n-estimators and max\_depth is tuned. Unfortunately, there has not been a great difference in the performance. The final accuracy with the Random Forest model is 91.83 and this is the best model so far.



**Decision tree** builds classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes. When we articulate, the code, we see that main cross validation score is 87.42 percent, and the accuracy of the model is 88.13%, which is low.

Text, letter

Description automatically generated

Hyperparameters, are parameters that directly control the behavior of the training algorithm and have a significant impact on the performance of the model that is being trained. In scikit-learn they are passed as arguments to the constructor of the estimator classes. Hence, we try to tune the model a bit with hyperparameters to get a higher accuracy. So, we will be changing the depth of the tree to four.

Text, letter

Description automatically generated

We see a significant impact under cross validation score as it has jumped to 90.52, and the overall accuracy of the model is 90.68. Next, we'll even try to further tune the data, by setting the number of features that are considered on a per-split basis to 20.

Text

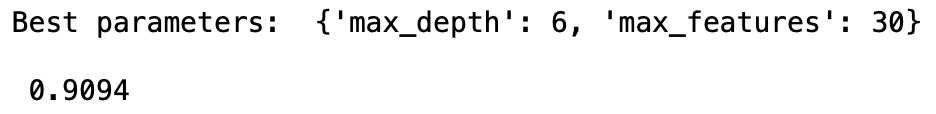
Description automatically generated

We get a cross validation score to 90.09% and the accuracy is declined to 90.53%. This means that there is a problem with the hyperparameter grid, that is the values of the maximum feature, and the maximum depth is not accurate. In order to do this, we define a function call GridSearchCV, which is a function that comes with model\_selection package, and that basically helps do loop predefined hyperparameters and fit the estimator on the training set.

A picture containing text

Description automatically generated

We have considered: 6,8,10,12 as maximum depth, and 20,25,30 as maximum features. We can get the accuracy of 90.94%, which is high, but, significantly low.



**Additional Model: Logistic Regression**

Logistic regression model is experimented to check the performance and this model didn’t perform better than the previous models. The solver used is “lbfgs” and max iterations are 1000. Logistic Regression gave a test accuracy of 91.93 that is the best model that we found.

Text

Description automatically generated with low confidence

**Conclusions**

Amongst the model’s test, Random Forest and LR outperforms and can be chosen for prediction, and a warning system for pilots. However, there is a lot of missing data, a filled data with all attributes might help us with a better model accuracy. Most strikes do not cause damages. A few have costs lives and monetary damage. Turbofan Engines are the most frequently struck and most bird strikes happen either during landing or take-off. Basically, in low altitude. Hence, pilots should be always informed of the bird situation before landing and taking-off. Most bird strikes have happened between July- September during the daytime, with smaller species of birds. Texas is the most bird struck state. Detection and countermeasures should be improved to prevent during daytime at landing and taking and for the designated months if resources are not available. Better understanding of bird migration routes and duration, bird control measure and airport traffic can surely improve the situation at large.

**References**

https://wildlife.faa.gov/