Investigating the Predictors of Damaging Wildlife Strikes at Part 139 Airports

-Ross Dickinson, M.Sc.
-Dr. Flavio Mendonca

Embry-Riddle Aeronautical University

Agenda

01

Introduction

What is the goal of this research?

02

Data Collection

Data science approach

03

Methodology

Where can you access this dataset?

04

Data Exploration

Deep dive into our dataset…

05

Principal Predictors

Which predictors should we be concerned about?

06

Conclusion

Summary of our results and Q & A

01

Introduction

- Purpose of the research?
- How can we answer this question?

Purpose of the study

Two components:

- Can we identify the most pertinent conditions leading to damaging wildlife strikes;
- and if so, can we then predict if a strike is going to be damaging or non-damaging?

To answer our research questions, we defined a single hypotheses:

1. ML models can sufficiently predict damaging wildlife strikes because there are relations between the features and damaging wildlife strikes (H0).

Which states are high -risk?

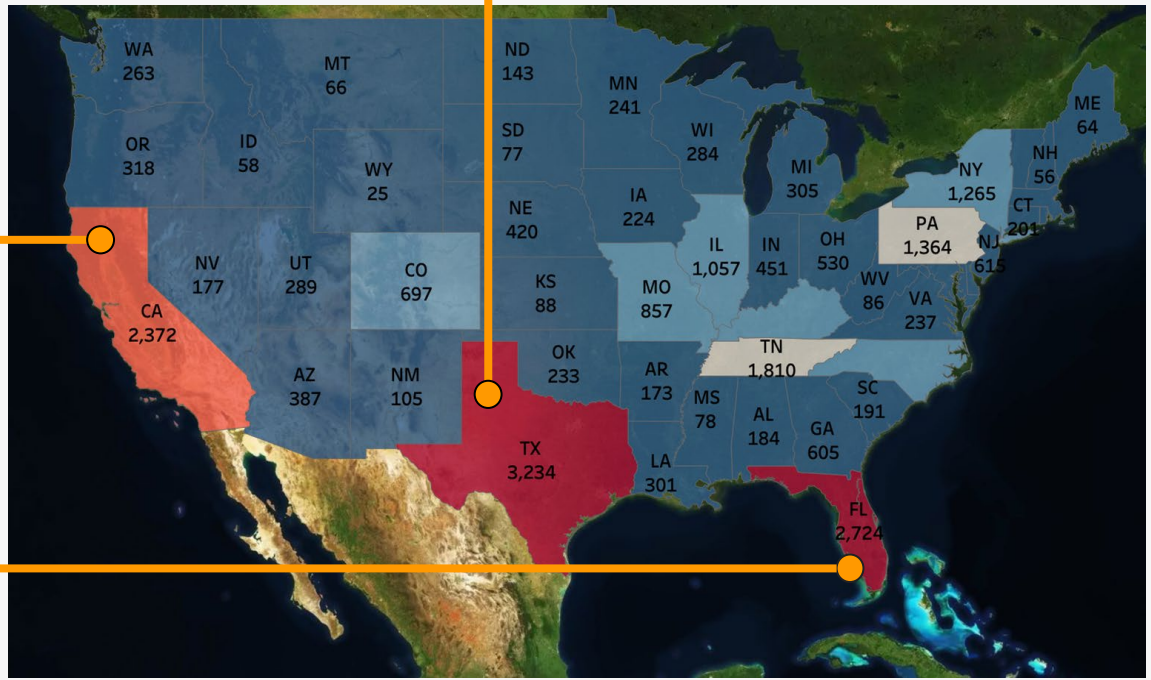
California

State:	CA
Nr. Cases:	2,372
Nr. Fatalities:	0
Nr Injuries:	
Cost of Repairs:	1,825,966
Distinct count of Airport:	31
Distinct count of Species:	162

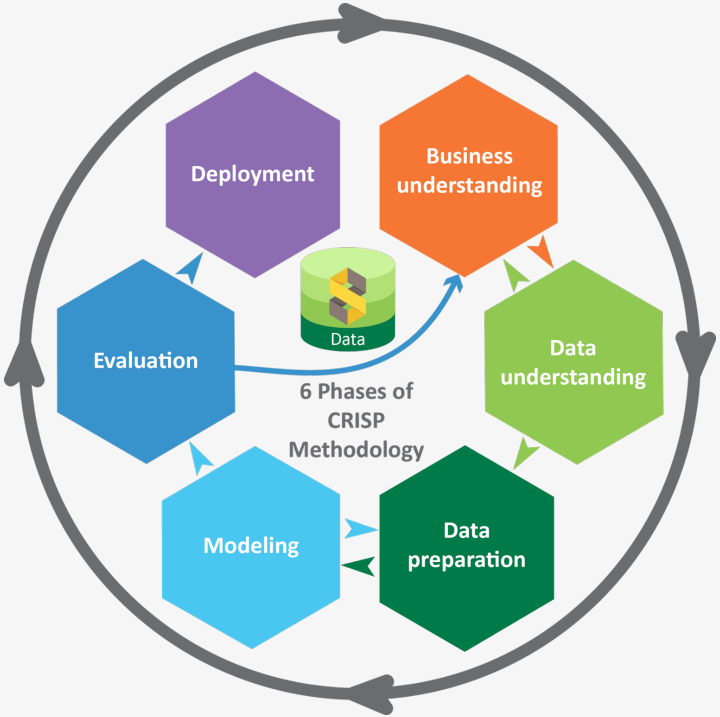
Florida

State:	FL
Nr. Cases:	2,724
Nr. Fatalities:	0
Nr Injuries:	4
Cost of Repairs:	22,951,657
Distinct count of Airport:	27
Distinct count of Species:	164

State:	TX
Nr. Cases:	3,234
Nr. Fatalities:	0
Nr Injuries:	
Cost of Repairs:	2,398,504
Distinct count of Airport:	30
Distinct count of Species:	155



How can we answer this question?



Reference 1

Cross-Industry Standard Process for Data Mining (**CRISP-DM**):

- 1. Business development
- 2. Data understanding
- 3. Data preparation
- 4. Modeling
- 5. Evaluation
- 6. Deployment

Data visualization:

- Data distributions (histograms, box-plots)
- Scatter plots, tree maps

Statistical methods:

- Chi-Squared Test
- Analysis of Variance (ANOVA)

Machine Learning

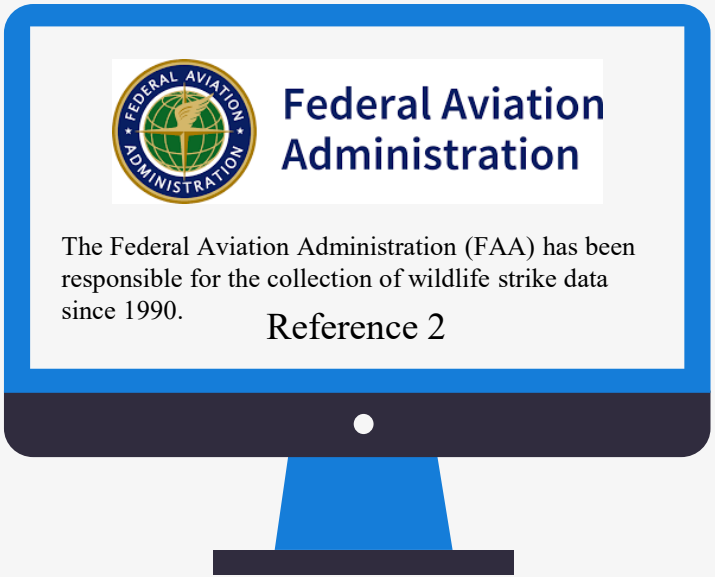
- Binary classification predictive model
- Ensemble methods

02

Data Collection

- Where is the data from?
- What does is show?
- Dealing with scarcity in the dataset

Where is the data from?



- The FAA provides a platform for pilots, airlines, and aviation professionals to record wildlife strikes on a voluntary basis
- Collected data from **January 2012 to January 2022**
- Contains 137K records
- Filtered to retrieve Part 139 Airport data only - 106K records in this category

What does it show?

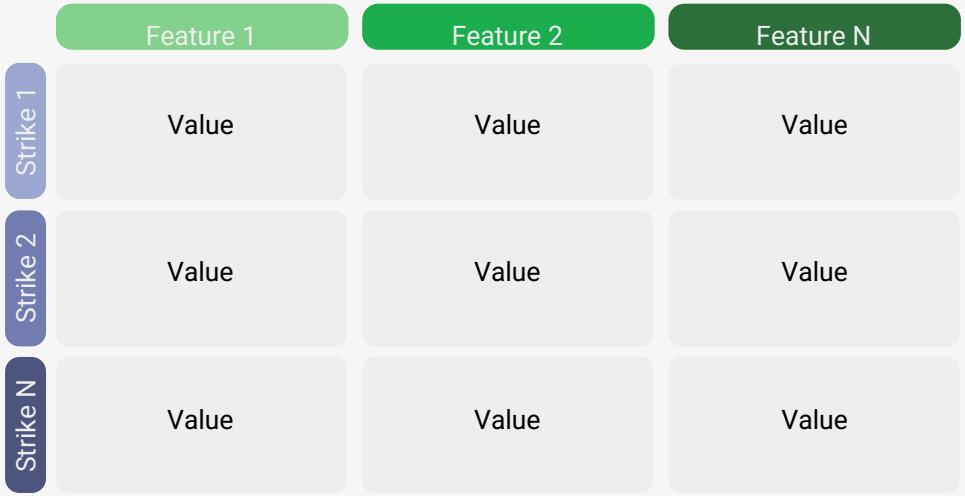
- This database contained 100 data features and 137K records of which 106K records were associated with Part 139 Airports
- We removed all records that did not have a damage level entry
- Leaving **55K records** to train and test our machine learning models

Scarce Dataset

A major concern from a data science perspective is the volume of missing data values. In total, there were **63,653 missing data entries**.

Feature	Number of Missing Values
Time of Day	5276
Phase of Flight	344
Height	10279
Speed	26766
Sky	11957
Size	0
Damage Level	9031

Dataset Layout



Deep dive into the features

Feature	Entry Values	Units	Report Category
Damage Level	None, Unknown, Minor, Severe Destroyed	-	Damage Information
Phase of Flight	Approach, Landing Roll, Climb, Take-Off, Departure, Descent, Arrival, Local, Taxi, Parked	-	Aircraft Information
Height	Continuous	ft.	Aircraft Information
Airspeed	Continuous	Knots (IAS)	Aircraft Information
Time of Day	Day, Dusk, Night, Dawn	-	Incident Date and Time
Sky Condition	No Cloud, Some Cloud, Overcast	-	Environment Condition

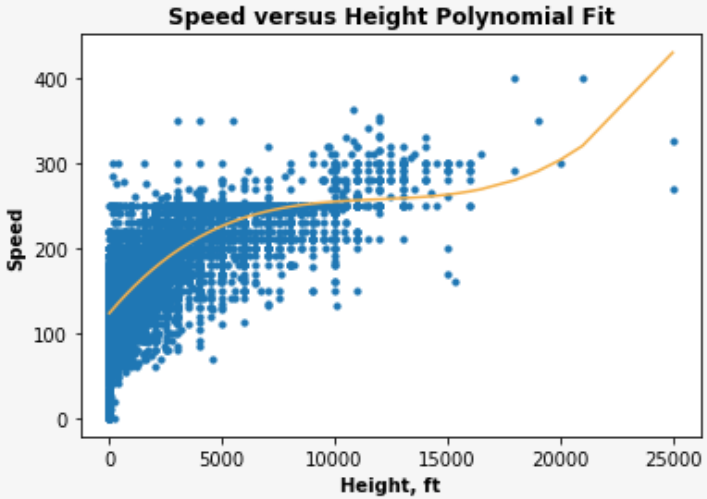
Damage level was encoded using this schema:

- Damaging = {‘Minor’, ‘Severe’, ‘Destroyed’}
- Non-damaging = {‘Unknown’, ‘None’}

Solving the problem of missing values

What approaches did use to predict missing values?

- Use existing data to estimate the missing values
- Relationships between the features can be used to estimate missing value
- Heavily imbalance distribution (97% non-damaging and 3% damaging)

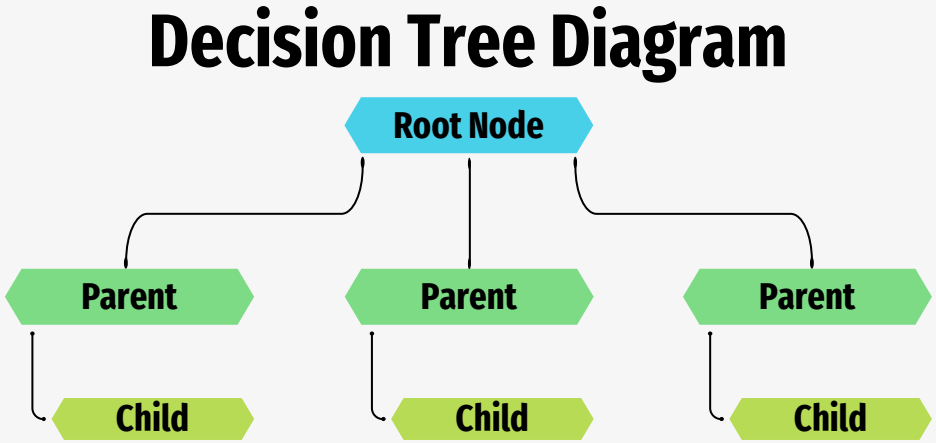


03 Methodology

Deep dive into Ensemble Methods

Structure of ensemble methods:

- Built from simple learners - decision trees
- Decision trees are formed using a random number of features
- Splits are made using information gain - entropy
- A random sample of the data is distributed to each tree
- Thousands of trees make individual predictions
- The outputs are aggregated to give a final output
- Helps to reduce bias in the model

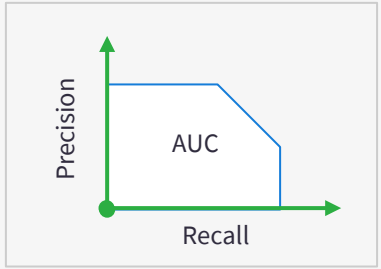


Evaluating classification models

	Positive	Negative
Existing Markets	TP	FP
New Markets	FN	TN

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

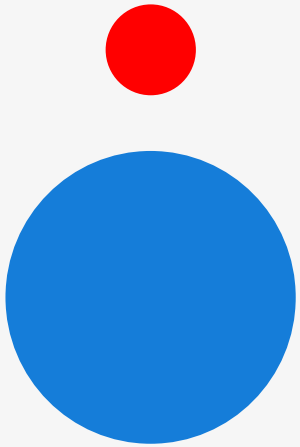


04

Results

Class Imbalance

Damage Level for all **55K records**

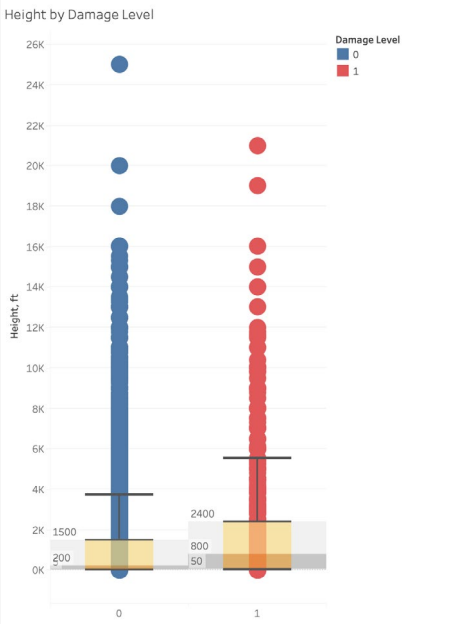


3% Damaging

97% Non-Damaging

Techniques to overcome class imbalance: **SMOTE** and **undersampling of the majority class**

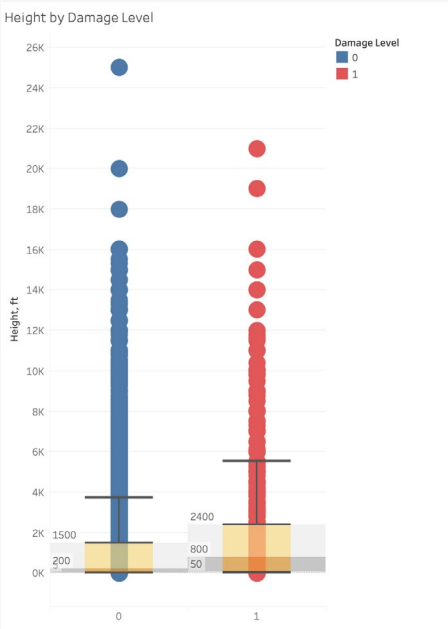
Speed



Speed (knots)

0 - 58		4.2% of 333
59 - 115		4.3% of 4,063
116 - 174		4.5% of 15,734
232+		5.0% of 1,834

Height

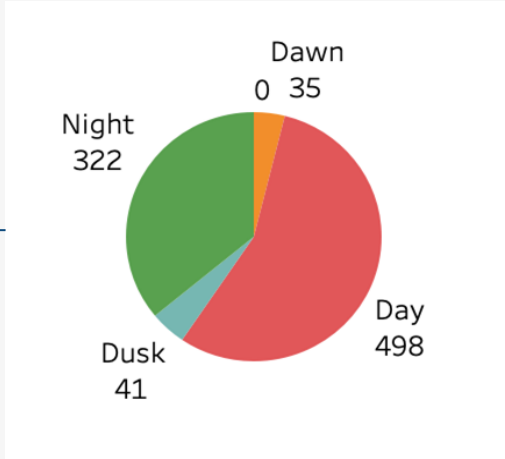


Height (ft)

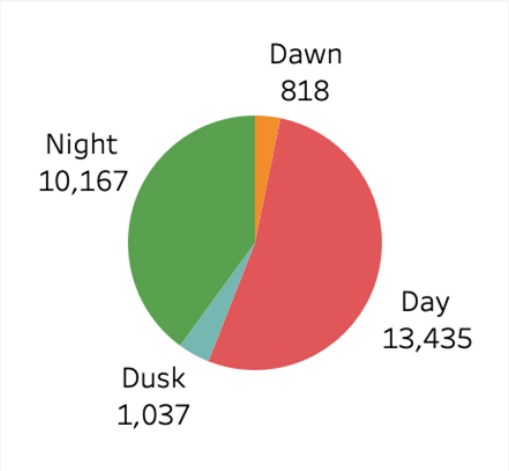
0 - 2K		3.0% of 20,265
2K - 4K		4.7% of 3,218
4K - 6K		1.9% of 1,446
6K+		2.2% of 1834

Time of day

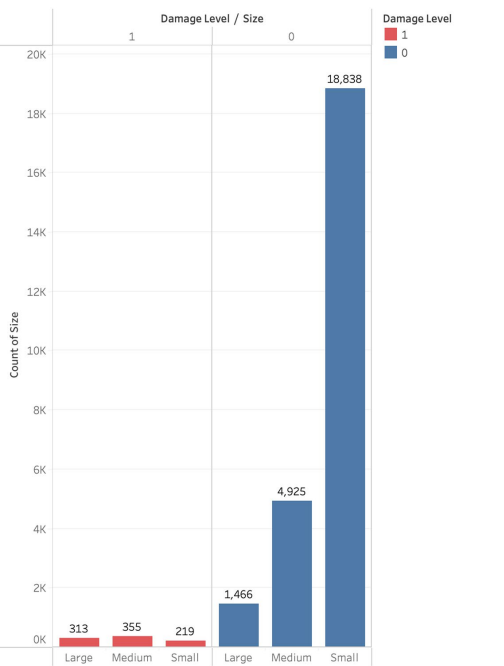
Damaging

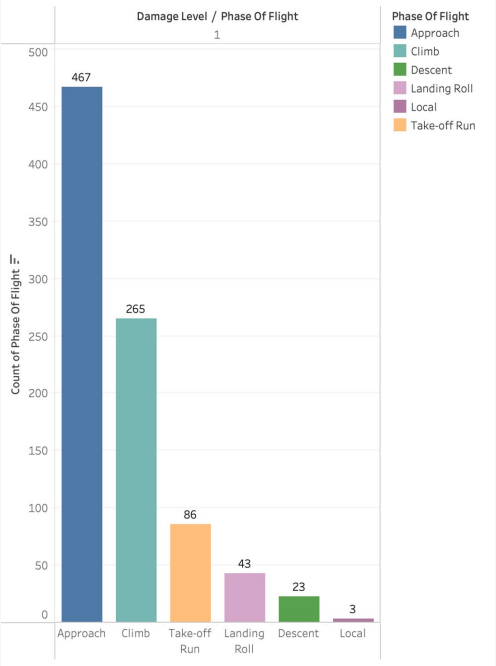


Non-Damaging



Phase of flight and size





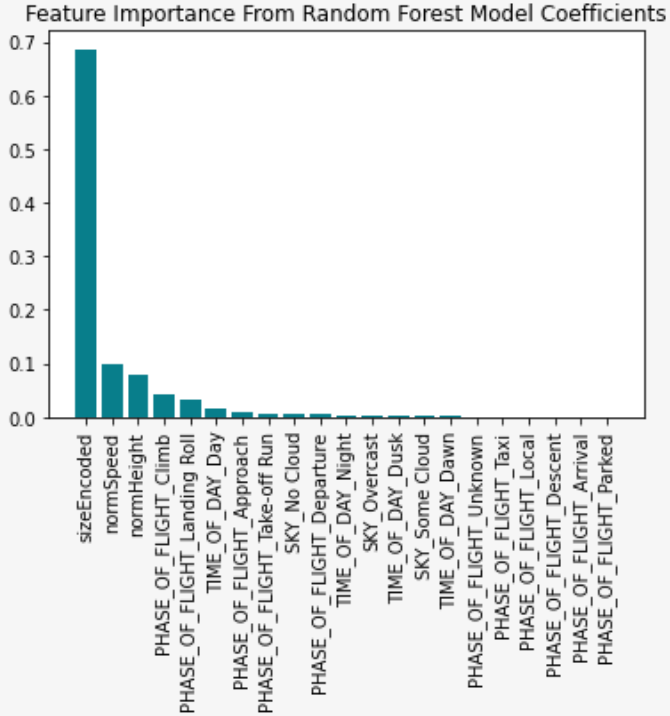
Measure of correlation

	Chi-Squared	ANOVA
Sky < 0.05	✓	-
Speed < 0.05	-	✓
Height < 0.05	-	✓
Phase of flight < 0.05	✓	-
Size < 0.05	✓	-
Time of day < 0.05	✗	-

Which model performed best?

Metric	Random Forest	XGBoost
Precision-Recall AUC	0.701	0.640
ROC AUC Score	0.885	0.860
Accuracy	0.950	0.990

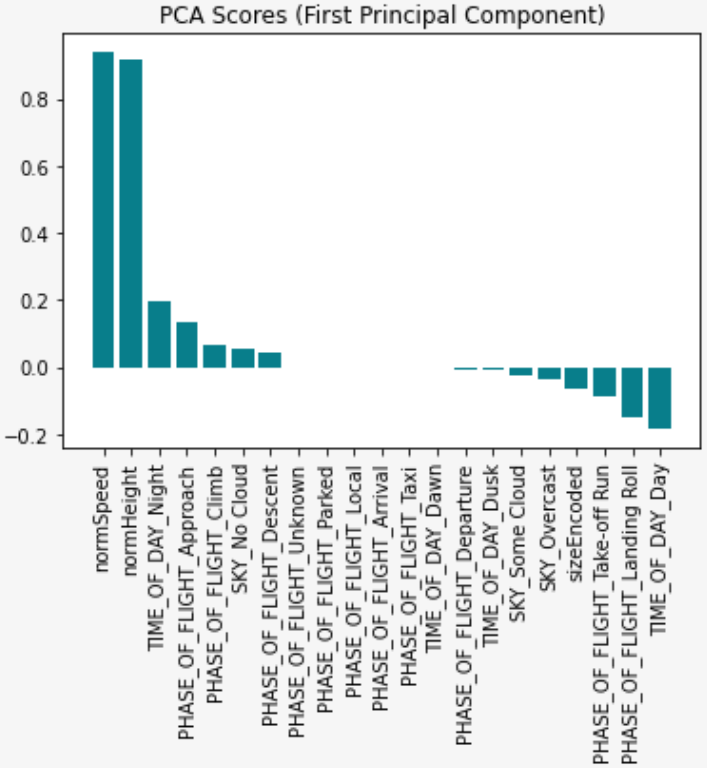
Feature Importance



High-Risk Indicators

1. Size
2. Speed
3. Height
4. Climb
5. Landing Roll

PCA Analysis



High-Risk Indicators

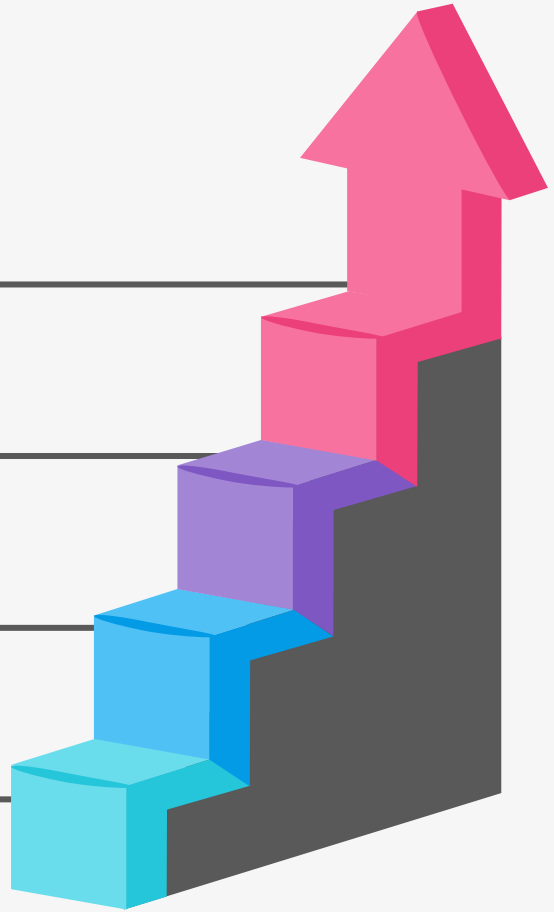
1. Speed
2. Height
3. Night
4. Approach
5. Climb

05

Conclusion

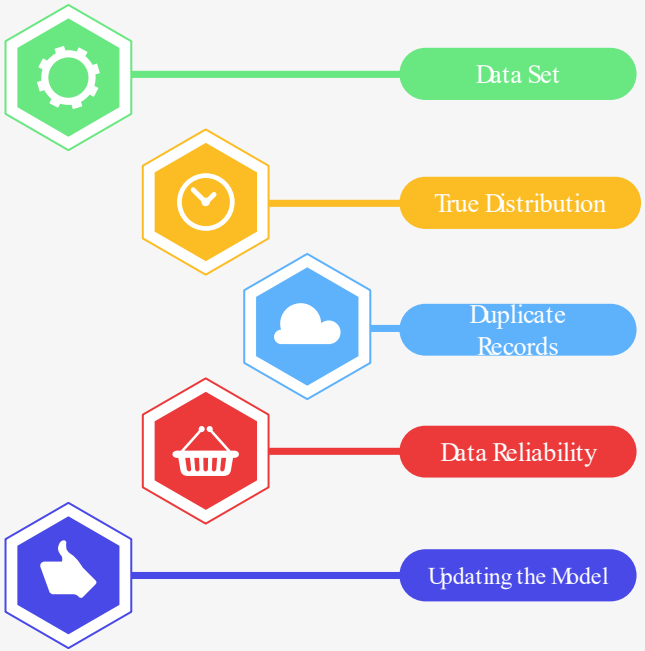
High-Risk Indicators

1. Wildlife Size: 4lb +
2. Speed between 116 - 174 knots (IAS)
3. Phase of flight: approach, climb, and landing roll
4. Time of day: day and night



What are the limitations?

Understanding the limitations of the data is paramount to make decisions on how to interpret the results of this research.



Data is voluntarily entered

Estimated only **20%** of all strikes are recorded

Was the same record input by numeral sources?

Validated by Dolbeer and Wright

Over time the true distribution may change over time

Do you have any questions?

dickinr3@my.erau.edu
Embry-Riddle Aeronautical University

References

[1] – Saliotech, CRISP-DM methodology. Accessed April 2022.
[2] – FAA Wildlife Strike Database, https://wildlife.faa.gov/home. Accessed March 2022.

Ensemble Method Types

XGBoost:

- Built in sequential format
- Uses stochastic gradient descent
- Each consecutive tree learns from the misclassification of the previous tree

Random Forest:

- Trees are distributed in a wide format
- Bootstrapping takes data with replacement