Preliminary report of the Final Paper

for the CEU MSc in Business Analytics program

Gábor Horváth

June 2017

CEU CENTRAL EUROPEAN UNIVERSITY

1 Introduction

The structure of the document follows the Cross Industry Standard Process for Data Mining (CRISP-DM) process model, which is a non-proprietary, documented, and freely available data mining model (Shearer 2000). Whenever the model sections can be matched to (and can fulfill) the requirements stated by CEU for the Final Paper I'm using the appropriate section identified by the CRIPS-DM model. Please keep in mind that the model supports the full end-to-end process of a data mining project, but the project does not require the use of all the model elements.

2 Business Understanding

2.1 Determine Business Objectives

2.1.1 Business Objectives

There are two main objectives what the project is aiming to complete.

1. Create a statistical analysis to identify those reasons (based on the data available), which are determining the the risk of an animal strike for an airport.
2. Create a prediction model, which can be used to predict the risk of an animal strike for a given flight.

The result of the statistical analysis could be used in the completion of the model building and evaluation the recommended order of the completion is the order of the objectives stated above.

2.1.2 Business Success Criteria

- Identification of features determining the risk potential of an airport
- Working model for animal strike prediction

2.2 Assess Situation

2.2.1 Inventory of resources

- Flight Data
- Animal Strike Data
- R
- Buckets

2.2.2 Requirements, Assumptions, and Constraints

- Additional requirements:
  - No additional requirements identified on top of the requirements already stated in this document.
- Assumptions
  - No initial assumptions made.
- Constraints
  - No initial hard constraints identified.

2.2.3 Risks and Contingencies

- Risks
  - No initial risks identified
- Contingencies
  - No initial contingencies identified

2.2.4 Terminology

The project is using different terminologies from the different domains. The terms/definitions used will not be marked or explained in details, if based on the context the reader can easily identify the domain of the particular term. In case there are uncertainties about a term (and it's not explained in the paper), the following sources can be used for the definitions:

- Aviation:
  - Aviation Terms / Directory: http://www.aviation-terms.com/index2.php
  - Aviation Glossary: http://www.aerofiles.com/glossary.html
  - Aviation Glossaries: https://www.flightsimaviation.com/_glossaries.html?s=aviation_terms
- Data Mining
  - Data Mining Glossary: http://www.thearling.com/glossary.htm
  - Data Mining - Terminologies: https://www.tutorialspoint.com/data_mining/dm_terminologies.htm
  - Data Mining and Predictive Analytics Glossary: http://www.kdnuggets.com/2015/06/data-mining-predictive-analytics-glossary.html
- Data Science / Big Data
  - Data Science Glossary: http://www.datascienceglossary.org/
  - Analytics and Big Data Glossary: http://data-informed.com/glossary-of-big-data-terms/
  - Data Science Glossary: http://www.kdnuggets.com/2015/09/data-science-glossary.html

2.2.5 Costs and Benefits

This is a one-man project, no significant cost is expected. Main benefit is to put to and almost end-to-end scenario the topics covered during the courses and discovering bits and bolts of the techniques for creating the project.

2.3 Determine Data Mining Goals

2.3.1 Data Mining Goals

- Understand, Analyse, Clean and Merge the source data correctly
- Create the required attributes
- Generate the required records (if applicable)

2.3.2 Data Mining Success Criteria

- Identification of featured determining the risk potential of an airport
- Working model for animal strike prediction

2.4 Produce Project Plan

2.4.1 Project Plan

The project is managed in an agile way, where all the tasks, requirements, issues, solutions, and ideas are kept in a project at buckets.

2.4.2 Initial Assessment of Tools and Techniques

- Programming language:
  - R: https://www.r-project.org/
- IDE for the programming language:
  - RStudio: https://www.rstudio.com/
- Documentation is created using:
  - knitr: https://yihui.name/knitr/
  - MiKTeX: https://miktex.org/
  - ReporteRs: https://cran.r-project.org/web/packages/ReporteRs/index.html
- Data visualization:

    - ggplot2: http://ggplot2.org/
- Data manipulation:
    - access2csv: https://github.com/AccelerationNet/access2csv
    - dtplyr: https://cran.r-project.org/web/packages/dtplyr/index.html
- Project plan / task management:
    - Buckets: https://www.buckets.co/
- Source code repository:
    - GitHub: https://github.com/

Note: The list above do not contain the list of all the tools and packages used to create the project, but the full list will be provided in the source code.

3 Data Understanding

3.1 Collect Initial Data

3.1.1 Initial Data Collection Report

There have been two data sources acquired in the initial phase of the project. These sources are the following:

3.1.1.1 Federal Aviation Administration

- Data source: Wildlife Strike Database
- The FAA provides the database as a compressed Microsoft Access file.
- The database version used is Version 2016.4-P (as of 24-10-2016).
- The database contains 180,177 Strike Reports from 1-1-1990 through 30-4-2016.
- The compressed file size is 44,730,852 bytes.
- The uncompressed Microsoft Access database file size is 193,495,040 bytes.
- The extracted tables are:
  - STRIKE_REPORTS (1990-1999) - 30082 rows - CSV size is 21,523,668 bytes
  - STRIKE_REPORTS (2000-2009) - 69960 rows - CSV size is 51,833,820 bytes.
  - STRIKE_REPORTS (2010-Current) - 70577 rows - CSV size is 53,973,874 bytes.
  - STRIKE_REPORTS_BASH (1990-Current).csv - 8046 rows - CSV size is 5,412,394 bytes.

3.1.1.2 United States Department of Transportation

- Data source: Bureau of Transportation Statistics
- The BTS provides the database as separate compressed CSV files. One file contains data of one month.
- The datestamp of the first CSV file available is 1-1-1987.
- The datestamp of the first data available is 1-10-1987.
- The datestamp of the last data acquired from BTS in the project is 31-12-2016.
- The number of files is 360.
  - Compressed size of the files is 6,196,385,360 bytes.
  - Uncompressed size of the files is 71,146,030,010 bytes.
- The download speed of the public access to these files seems to be limited, which needs to be taken into account in case of reproducing the results.

3.2 Describe Data

3.2.1 Data Description Report

The two main data sources have the following column explanations, which is attached to the downloaded files as well, by the data provider agencies.

3.2.1.1 Animal strike data

Column name	Explanation of Column Name and Codes
INDEX NR	Individual record number
OPID	Airline operator code
OPERATOR	A three letter International Civil Aviation Organization code for aircraft operators. (BUS = business, PVT = private aircraft other than business, GOV = government aircraft, MIL - military aircraft.)
ATYPE	Aircraft

Column name	Explanation of Column Name and Codes
AMA	International Civil Aviation Organization code for Aircraft Make
AMO	International Civil Aviation Organization code for Aircraft Model
EMA	Engine Make Code (see Engine Codes tab below)
EMO	Engine Model Code (see Engine Codes tab below)
AC_CLASS	Type of aircraft (see Aircraft Type tab below)
AC_MASS	1 = 2,250 kg or less: 2 = ,2251-5700 kg: 3 = 5,701-27,000 kg: 4 = 27,001-272,000 kg: 5 = above 272,000 kg
NUM_ENGS	Number of engines
TYPE_ENG	Type of power A = reciprocating engine (piston): B = Turbojet: C = Turboprop: D = Turbofan: E = None (glider): F = Turboshaft (helicopter): Y = Other
ENG_1_POS	Where engine # 1 is mounted on aircraft (see Engine Position tab below)
ENG_2_POS	Where engine # 2 is mounted on aircraft (see Engine Position tab below)
ENG_3_POS	Where engine # 3 is mounted on aircraft (see Engine Position tab below)
ENG_4_POS	Where engine # 4 is mounted on aircraft (see Engine Position tab below)
REG	Aircraft registration
FLT	Flight number
REMAINS_COLLECTED	Indicates if bird or wildlife remains were found and collected
REMAINS_SENT	Indicates if remains were sent to the Smithsonian Institution for identification
INCIDENT_DATE	Date strike occurred
INCIDENT_MONTH	Month strike occurred
INCIDENT_YEAR	Year strike occurred
TIME_OF_DAY	Light conditions
TIME	Hour and minute in local time
AIRPORT_ID	International Civil Aviation Organization airport identifier for location of strike whether it was on or off airport
AIRPORT	Name of airport
STATE	State
FAAREGION	FAA Region where airport is located
ENROUTE	If strike did not occur on approach, climb, landing roll, taxi or take-off, aircraft was enroute. This shows location.
RUNWAY	Runway
LOCATION	Various information about aircraft location if enroute or airport where strike evidence was found. Some locations show the two airports for the flight departure and arrival if pilot was unaware of the strike.
HEIGHT	Feet Above Ground Level
SPEED	Knots (indicated air speed)
DISTANCE	Miles from airport
PHASE_OF_FLT	Phase of flight during which strike occurred
DAMAGE	
Blank	Unknown
M = minor	When the aircraft can be rendered airworthy by simple repairs or replacements and an extensive inspection is not necessary.
M? = uncertain level	The aircraft was damaged, but details as to the extent of the damage are lacking.
S = substantial	When the aircraft incurs damage or structural failure which adversely affects the structure strength, performance or flight characteristics of the aircraft and which would normally require major repair or replacement of the affected component.
D = Destroyed	When the damage sustained makes it inadvisable to restore the aircraft to an airworthy condition.

Column name	Explanation of Column Name and Codes
STR_RAD	Struck radome
DAM_RAD	Damaged radome
STR_WINDSHLD	Struck windshield
DAM_WINDSHLD	Damaged windshield
STR_NOSE	Struck nose
DAM_NOSE	Damaged nose
STR_ENG1	Struck Engine 1
DAM_ENG1	Damaged Engine 1
STR_ENG2	Struck Engine 2
DAM_ENG2	Damaged Engine 2
STR_ENG3	Struck Engine 3
DAM_ENG3	Damaged Engine 3
STR_ENG4	Struck Engine 4
DAM_ENG4	Damaged Engine 4
INGESTED	Engine ingested the bird/ animal
STR_PROP	Struck Propeller
DAM_PROP	Damaged Propeller
STR_WING_ROT	Struck Wing or Rotor
DAM_WING_ROT	Damaged Wing or Rotor
STR_FUSE	Struck Fuselage
DAM_FUSE	Damaged Fuselage
STR_LG	Struck Landing Gear
DAM_LG	Damaged Landing Gear
STR_TAIL	Struck Tail
DAM_TAIL	Damaged Tail
STR_LGHTS	Struck Lights
DAM_LGHTS	Damaged Lights
STR_OTHER	Struck Other than parts shown above
DAM_OTHER	Damaged Other than parts shown above
OTHER_SPECIFY	What part was struck other than those listed above
EFFECT	Effect on flight
EFFECT_OTHER	Effect on flight other than those listed on the form
SKY	Type of cloud cover, if any
PRECIP	Precipitation
SPECIES_ID	International Civil Aviation Organization code for type of bird or other wildlife
SPECIES	Common name for bird or other wildlife
BIRDS_SEEN	Number of birds/wildlife seen by pilot
BIRDS_STRUCK	Number of birds/wildlife struck
SIZE	Size of bird as reported by pilot is a relative scale. Entry should reflect the perceived size as opposed to a scientifically determined value. If more than one species was struck, larger bird is entered.
WARNED	Pilot warned of birds/wildlife
COMMENTS	As entered by database manager. Can include name of aircraft owner, types of reports received, updates, etc.
REMARKS	Most of remarks are from the form but some are data entry notes and are usually in parentheses.
AOS	Time aircraft was out of service in hours. If unknown, it is blank.
COST_REPAIRS	Estimated cost of repairs of replacement in dollars (USD)
COST_OTHER	Estimated other costs, other than those in previous field in dollars (USD). May include loss of revenue, hotel expenses due to flight cancellation, costs of fuel dumped, etc.

Column name	Explanation of Column Name and Codes
COST_REPAIRS_INFL_ADJ	Costs adjusted for inflation
COST_OTHER_INFL_ADJ	Other cost adjusted for inflation
REPORTED_NAME	Name(s) of person(s) filing report
REPORTED_TITLE	Title(s) of person(s) filing report
REPORTED_DATE	Date report was written
SOURCE	Type of report. Note: for multiple types of reports this will be indicated as Multiple. See "Comments" field for details
PERSON	Only one selection allowed. For multiple reports, see field "Reported Title"
NR_INJURIES	Number of people injured
NR_FATALITIES	Number of human fatalities
LUPDATE	Last time record was updated
TRANSFER	Unused field at this time
INDICATED_DAMAGE	Indicates whether or not aircraft was damaged

3.2.1.2 Flight data

Column name	Explanation of Column Name and Codes
Year	Year
Quarter	Quarter (1-4)
Month	Month
DayofMonth	Day of Month
DayOfWeek	Day of Week
FlightDate	Flight Date (yyyymmdd)
UniqueCarrier	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years.
AirlineID	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.
Carrier	Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code.
TailNum	Tail Number
FlightNum	Flight Number
OriginAirportID	Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused.
OriginAirportSeqID	Origin Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time.
OriginCityMarketID	Origin Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market.
Origin	Origin Airport
OriginCityName	Origin Airport, City Name
OriginState	Origin Airport, State Code
OriginStateFips	Origin Airport, State Fips
OriginStateName	Origin Airport, State Name
OriginWac	Origin Airport, World Area Code

Column name	Explanation of Column Name and Codes
DestAirportID	Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused.
DestAirportSeqID	Destination Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time.
DestCityMarketID	Destination Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market.
Dest	Destination Airport
DestCityName	Destination Airport, City Name
DestState	Destination Airport, State Code
DestStateFips	Destination Airport, State Fips
DestStateName	Destination Airport, State Name
DestWac	Destination Airport, World Area Code
CRSDepTime	CRS Departure Time (local time: hhmm)
DepTime	Actual Departure Time (local time: hhmm)
DepDelay	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
DepDelayMinutes	Difference in minutes between scheduled and actual departure time. Early departures set to 0.
DepDel15	Departure Delay Indicator, 15 Minutes or More (1=Yes)
DepartureDelayGroups	Departure Delay intervals, every (15 minutes from <-15 to >180)
DepTimeBlk	CRS Departure Time Block, Hourly Intervals
TaxiOut	Taxi Out Time, in Minutes
WheelsOff	Wheels Off Time (local time: hhmm)
WheelsOn	Wheels On Time (local time: hhmm)
TaxiIn	Taxi In Time, in Minutes
CRSArrTime	CRS Arrival Time (local time: hhmm)
ArrTime	Actual Arrival Time (local time: hhmm)
ArrDelay	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
ArrDelayMinutes	Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.
ArrDel15	Arrival Delay Indicator, 15 Minutes or More (1=Yes)
ArrivalDelayGroups	Arrival Delay intervals, every (15-minutes from <-15 to >180)
ArrTimeBlk	CRS Arrival Time Block, Hourly Intervals
Cancelled	Cancelled Flight Indicator (1=Yes)
CancellationCode	Specifies The Reason For Cancellation
Diverted	Diverted Flight Indicator (1=Yes)
CRSElapsedTime	CRS Elapsed Time of Flight, in Minutes
ActualElapsedTime	Elapsed Time of Flight, in Minutes
AirTime	Flight Time, in Minutes
Flights	Number of Flights
Distance	Distance between airports (miles)
DistanceGroup	Distance Intervals, every 250 Miles, for Flight Segment
CarrierDelay	Carrier Delay, in Minutes
WeatherDelay	Weather Delay, in Minutes
NASDelay	National Air System Delay, in Minutes
SecurityDelay	Security Delay, in Minutes
LateAircraftDelay	Late Aircraft Delay, in Minutes
FirstDepTime	First Gate Departure Time at Origin Airport

Column name	Explanation of Column Name and Codes
TotalAddGTime	Total Ground Time Away from Gate for Gate Return or Cancelled Flight
LongestAddGTime	Longest Time Away from Gate for Gate Return or Cancelled Flight
DivAirportLandings	Number of Diverted Airport Landings
DivReachedDest	Diverted Flight Reaching Scheduled Destination Indicator (1=Yes)
DivActualElapsedTime	Elapsed Time of Diverted Flight Reaching Scheduled Destination, in Minutes. The ActualElapsedTime column remains NULL for all diverted flights.
DivArrDelay	Difference in minutes between scheduled and actual arrival time for a diverted flight reaching scheduled destination. The ArrDelay column remains NULL for all diverted flights.
DivDistance	Distance between scheduled destination and final diverted airport (miles). Value will be 0 for diverted flight reaching scheduled destination.
Div1Airport	Diverted Airport Code1
Div1AirportID	Airport ID of Diverted Airport 1. Airport ID is a Unique Key for an Airport
Div1AirportSeqID	Airport Sequence ID of Diverted Airport 1. Unique Key for Time Specific Information for an Airport
Div1WheelsOn	Wheels On Time (local time: hhmm) at Diverted Airport Code1
Div1TotalGTime	Total Ground Time Away from Gate at Diverted Airport Code1
Div1LongestGTime	Longest Ground Time Away from Gate at Diverted Airport Code1
Div1WheelsOff	Wheels Off Time (local time: hhmm) at Diverted Airport Code1
Div1TailNum	Aircraft Tail Number for Diverted Airport Code1
Div2Airport	Diverted Airport Code2
Div2AirportID	Airport ID of Diverted Airport 2. Airport ID is a Unique Key for an Airport
Div2AirportSeqID	Airport Sequence ID of Diverted Airport 2. Unique Key for Time Specific Information for an Airport
Div2WheelsOn	Wheels On Time (local time: hhmm) at Diverted Airport Code2
Div2TotalGTime	Total Ground Time Away from Gate at Diverted Airport Code2
Div2LongestGTime	Longest Ground Time Away from Gate at Diverted Airport Code2
Div2WheelsOff	Wheels Off Time (local time: hhmm) at Diverted Airport Code2
Div2TailNum	Aircraft Tail Number for Diverted Airport Code2
Div3Airport	Diverted Airport Code3
Div3AirportID	Airport ID of Diverted Airport 3. Airport ID is a Unique Key for an Airport
Div3AirportSeqID	Airport Sequence ID of Diverted Airport 3. Unique Key for Time Specific Information for an Airport
Div3WheelsOn	Wheels On Time (local time: hhmm) at Diverted Airport Code3
Div3TotalGTime	Total Ground Time Away from Gate at Diverted Airport Code3
Div3LongestGTime	Longest Ground Time Away from Gate at Diverted Airport Code3
Div3WheelsOff	Wheels Off Time (local time: hhmm) at Diverted Airport Code3
Div3TailNum	Aircraft Tail Number for Diverted Airport Code3
Div4Airport	Diverted Airport Code4
Div4AirportID	Airport ID of Diverted Airport 4. Airport ID is a Unique Key for an Airport
Div4AirportSeqID	Airport Sequence ID of Diverted Airport 4. Unique Key for Time Specific Information for an Airport
Div4WheelsOn	Wheels On Time (local time: hhmm) at Diverted Airport Code4
Div4TotalGTime	Total Ground Time Away from Gate at Diverted Airport Code4
Div4LongestGTime	Longest Ground Time Away from Gate at Diverted Airport Code4
Div4WheelsOff	Wheels Off Time (local time: hhmm) at Diverted Airport Code4
Div4TailNum	Aircraft Tail Number for Diverted Airport Code4
Div5Airport	Diverted Airport Code5
Div5AirportID	Airport ID of Diverted Airport 5. Airport ID is a Unique Key for an Airport
Div5AirportSeqID	Airport Sequence ID of Diverted Airport 5. Unique Key for Time Specific Information for an Airport
Div5WheelsOn	Wheels On Time (local time: hhmm) at Diverted Airport Code5

Column name	Explanation of Column Name and Codes
Div5TotalGTime	Total Ground Time Away from Gate at Diverted Airport Code5
Div5LongestGTime	Longest Ground Time Away from Gate at Diverted Airport Code5
Div5WheelsOff	Wheels Off Time (local time: hhmm) at Diverted Airport Code5
Div5TailNum	Aircraft Tail Number for Diverted Airport Code5

3.3 Explore Data

3.3.1 Data Exploration Report

3.3.1.1 Animal strike data

3.3.1.2 Flight data

TODO

3.4 Verify Data Quality

3.4.1 Data Quality Report

TODO

4 Data Preparation

4.1 Data Set

4.1.1 Data Set Description

TODO

4.2 Select Data

4.2.1 Rationale for Inclusion / Exclusion

TODO

4.3 Clean Data

4.3.1 Data Cleaning Report

TODO

4.4 Construct Data

4.4.1 Derived Attributes

TODO

4.4.2 Generated Records

TODO

4.5 Integrate Data

4.5.1 Merged Data

TODO

4.6 Format Data

4.6.1 Reformatted Data

TODO

5 Modeling

TODO

5.1 Select Modeling Technique for Model 1

5.1.1 Modeling Technique

TODO

5.1.2 Modeling Assumptions

TODO

5.2 Generate Test Design for Model 1

5.2.1 Test Design

TODO

5.3 Build Model for Model 1

5.3.1 Parameter Settings

TODO

5.3.2 Models

TODO

5.3.3 Model Description

TODO

5.4 Assess Model for Model 1

5.4.1 Model Assessment

TODO

5.4.2 Revised Parameter Settings

TODO

5.5 Select Modeling Technique for Model 2

5.5.1 Modeling Technique

TODO

5.5.2 Modeling Assumptions

TODO

5.6 Generate Test Design for Model 2

5.6.1 Test Design

TODO

5.7 Build Model for Model 2

5.7.1 Parameter Settings

TODO

5.7.2 Models

TODO

5.7.3 Model Description

TODO

5.8 Assess Model for Model 2

5.8.1 Model Assessment

TODO

5.8.2 Revised Parameter Settings

TODO

6 Evaluation

6.1 Evaluate Results

6.1.1 Assessment of Data Mining Result with Business Success Criteria

TODO

6.1.2 Approved Models

TODO

6.2 Review Process

6.2.1 Review of Process

TODO

6.3 Determine Next Steps

6.3.1 List of Possible Actions

TODO

6.3.2 Decision

TODO

7 Contributors

Student: Gábor Horváth
Mentor: Gergely Daróczi

8 Environment

The following language, tool and library versions have been used to create the project:

R Studio version 1.0.143

R version 3.4.0 (2017-04-21) 72570

Package versions:

- RODBC version 1.3.15
- knitr version 1.15.1
- data.table version 1.10.4
- dplyr version 0.5.0
- dtplyr version 0.0.2
- ReporteRs version 0.8.8
- ReporteRsjars version 0.0.2
- installr version 0.19.0
- stringr version 1.2.0
- ggplot2 version 2.2.1
- yaml version 2.1.14

Base package versions:

- stats version 3.4.0
- graphics version 3.4.0
- grDevices version 3.4.0
- utils version 3.4.0
- datasets version 3.4.0
- methods version 3.4.0
- base version 3.4.0

```
MiKTeX Package Manager 2.9.6200 (MiKTeX 2.9.6210 64-bit)
Copyright (C) 2005-2016 Christian Schenk
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
```

Contents

References

Shearer, Colin. 2000. “The Crisp-Dm Model - the New Blueprint for Data Mining.” Journal of Data Warehousing 5 (4): 13–22.