Project Plan of the Final Paper

for the CEU MSc in Business Analytics program

Gábor Horváth

2017

CEU | CENTRAL
EUROPEAN
UNIVERSITY

1 High level description

The goal of the project is to show - creating a risk evaluation of wildlife strikes of flights in the US - the techniques, methods, interpretations and understanding of the data analytic. The project is based on the Cross Industry Standard Process for Data Mining (CRISP-DM) process model, which is widely used worldwide for various scientific and business related data analytic projects. The use of the CRISP-DM process model will enable to for the project to cover all those areas (i.e. Business Understanding, Data understanding, Modelling, Evaluation, etc.), which are crucial of managing and delivering a successful data analytic project.

2 Resource needs

2.1 Training requirements

No additional organized / official training requirements are required above the trainings received during the courses in the program. There are tools and techniques used to fulfill the project which have not been described in the program at CEU, but there are several useful user manuals available on the webpages of the tool's creators, which would enable the use of these tools and resources for any student who have been part of the program.

2.2 Tools & resources used

Fulfilling the completion need for the project the following tools are planned to be used:

- 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.

2.3 Data sources

The project will use the following data provided by multiple US government agencies:

- Federal Aviation Administration: Wildlife Strike Database
- United States Department of Transportation: Bureau of Transportation Statistics

Note: In case data enrichment would be required for the successful risk modelling, additional data sources might be used as well. These possible additional data sources will be listed in the Final Paper.

2.4 Restrictions

Restrictions apply as per the restrictions set by the tools, data providers and owners of additional resources used. No additional restrictions have been identified and set regarding the use of the results of this project.

2.5 Contributors

Student: Gábor Horváth
Mentor: Gergely Daróczi