R – Ralfs

A- Andrei

A(slide 1): Hello and welcome to group’s number 2 presentation for this semester’s exam. In the following 50 minutes, we are going to talk about (slide 2) who we are, what we did and how we did it.

A(slide 3): So, who are we? My name is Birta Andrei and he is Ralfs Zangis. (play animation) Together we united in order to fight against the forces of evil, and make a project.

A(slide 4): But what is the project about? Well, just as the name might suggest, within the pales of this project we wanted to try and reduce, if not fully stop any and all accidents (play animation) that might happen in the future, regarding aircrafts and wildlife within the territory of the United States of America.

A(slide 5): Why this and not some other subject? Because, although in the period of 2009-2017 from all 142.921.417 flights in the USA, there have been 86.877 collisions with wildlife, giving us a 0.06% chance of collision; Not to mention the 30 people that died and 265 injuries and distrust that such an event will cause in the airlines customers.

A(slide 6): And considering there are around 5000 aircrafts in the sky, at any given moment, and an average of 87.000 flights per day, we thought: “how can we make sure that those will not partake into any accident?”; So this project came to life.

A(slide 7): How are doing that? Well, we couldn’t do anything, before learning some more about the case. So we thought of some data that might help us, found the data we though is going to help us from a reliable source, we acquired it, we cured it, analyzed it, visualized it and in the end, created a prediction model.

A(slide 8): But how exactly is all of that reducing the collision rate? Our entire project, is reducing the collision rate through multiple ways:

* by educating the uninformed about the matter, through the descriptive and diagnostic analysis’s we did;
* by offering a prediction model, which uses both supervised and unsupervised machine learning techniques, in order to predict whether a flight is going to or not partake into a collision with any wildlife.
* by offering some guidelines and examples of successfully proven ways of reducing such collisions from happening, used by multiple airports in the USA.

R(slide 9): The prediction model works by using a combination of Unsupervised and Supervised learning techniques. K-means, is our unsupervised learning technique of choice, and is used in order to further refine the data, by grouping it in 4 different clusters. (play animation) As you can, these are the flights from 2015 in USA, grouped in the 4 clusters using K-means. The reason why we decided to further separate the data, is probably best explained using a metaphor, and that is: it would make little to no sense for a doctor to look at female biology, if the presenting patient is a male. Using the same, school of thought, it would make little to no sense for our model to look at any group of flights, other than the one that the flight in question, is part of.

R(slide 10): Currently, the model predicts with a 99.55% accuracy, and it does so by using a supervised learning technique, that being a decision tree.

R(open .pdf file): This is a plotted image of our current decision tree. As you can see, there are a bunch of branches, in fact, the image alone, is so big, that it took my laptop around 30 minutes to plot it.

A(slide 11): Now you may ask yourself: how did we manage to achieve this? We did it by following a development framework called: Kanban; we did it by investing around 500 hours of working time (combined), most of which went into researching the matter, as we knew little to nothing about it, before dwelling into it, and another big chunk of time went into cleaning the data as we had to deal with NA values and misspellings and what not.

R(slide 12): In fact, none of this would have been possible without cleaning the data, and the way we did that, was by using python. First, we identified where the missing values are, by checking if there is something written there, or if it has a specific word in it, like “unknown”; then we filled all those empty slots with “NAN” – the typical values that python uses in order to signal an empty cell; then filled those NAN slots, with meaningful data, either by using the average of that attribute’s values, or most common or middle value. After that, we made sure that all attributes are loaded correctly, in their specific format; so data would be “dateTime” not integer or something else. We normalized the data, in the sense of making sure all values within an attribute are similar, so there wont be an airport name in the “date time” attribute, in this step we also took care of all misspells. Who knew there are 15 different ways of spelling the word “Dove”… In the end, we saved the new, clean dataset, into a .csv file and used it for further processing.

A(slide 13): But that’s about it, regarding our presentation, QNA time?