Solution Abstract Problem 1: SmartFly

Cindy Lamm

January 16, 2015

Explain your methodology including approach, assumptions, software and algorithms used, testing and validation techniques applied, model selection criteria, and total time spent.

1 Approach

1.0.1 Get to know the data

As first step I do a basic structure and content analysis of the given data files on the command line in order to prepare loading the data into any tool/program[1]. The main goal is to get a first impression of data quality: Do all lines of the csv file have the same number of fields? Are there any quotes within the fields that might need escaping?

As second step I do an exploratory data analysis on the historic as well as the prediction data set. I analyse both sets to avoid potential issues arising when the prediction data is slightly (or strongly) different in structure or content from the historic data that will be used for training a model.

I use the open source statistics software R[2] for the exploratory data analysis since it has all necessary analysis functions and plotting tools already implemented and the knitr package[3] allows for reproducible analysis.

2 First basic model

In order to setup the infrastructure from beginning to end I estimate a first basic model with all necessary validation techniques. This allows me to then

[1] see `problem1_smartfly/src/00_file_structure_analysis/file_structure_analysis.sh`

[2] `http://www.r-project.org/`

[3] `http://yihui.name/knitr/`

quickly iterate over other models/algorithms with each iteration producing the required output file of ordered flight IDs.

I thus perform the following steps always bearing in mind to achieve the "minimal viable solution"[4] (meaning the least effort to get to the next working increment):

- prototype fast by using a minor fraction of the original historic data for the first iterations
- prepare data for modeling: convert one input variable and the target variable to the necessary format
- estimate model

[4]Eric Ries, Lean Startup