Hack Zika 2017

Prescriptive Analytics to Curb Zika

Cone Nose Assassins
Robert Ferrari
Jennifer Gronek
Cliff Lezark
Linggih Saputro

Statement of Problem

Hillsborough County sprays regularly to control mosquito populations and prevent the spread of disease, especially Zika fever. At the moment, Hillsborough County relies on the expertise of one man, Ron Montgomery, to determine which zones to spray when.

We propose to replace Ron's expertise with a more stringent, scientific method.

Suggested Solutions

Various solutions may be used to determine which zones are at greatest risk for vector-borne disease, Zika in particular. We particularly suggest:

- Statistical Methods
  - Logistic Regression
- Machine Learning
  - Decision Trees
  - Random Forests

Explanatory Variables

Regardless of method used, we consider the same relationship.

The potential presence of Zika may be considered the result of weather trends, existing Zika cases, the presence of *Aedes aegypti* and *Aedes albopictus*.

The need to spray a certain zone is also determined by population density and the presence of immunodeficient populations.

Logistic Regression

The presence of Zika is considered a binary dependent variable – i.e., either present or not.

We estimate the log odds of the presence of Zika based on our explanatory variables as such:

$$logit(p) = \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

Decision Trees

- Decision trees consist of nodes nested in branches connected to other nodes
- Nodes consist of conditional decision rules that dictate the observation path
- Advantages
  - Simple to understand and modify
  - Handles more than binary outcomes
- Disadvantages
  - Can be very cluttered with large variable sets
  - Tend to overfit training sets

Decision Trees

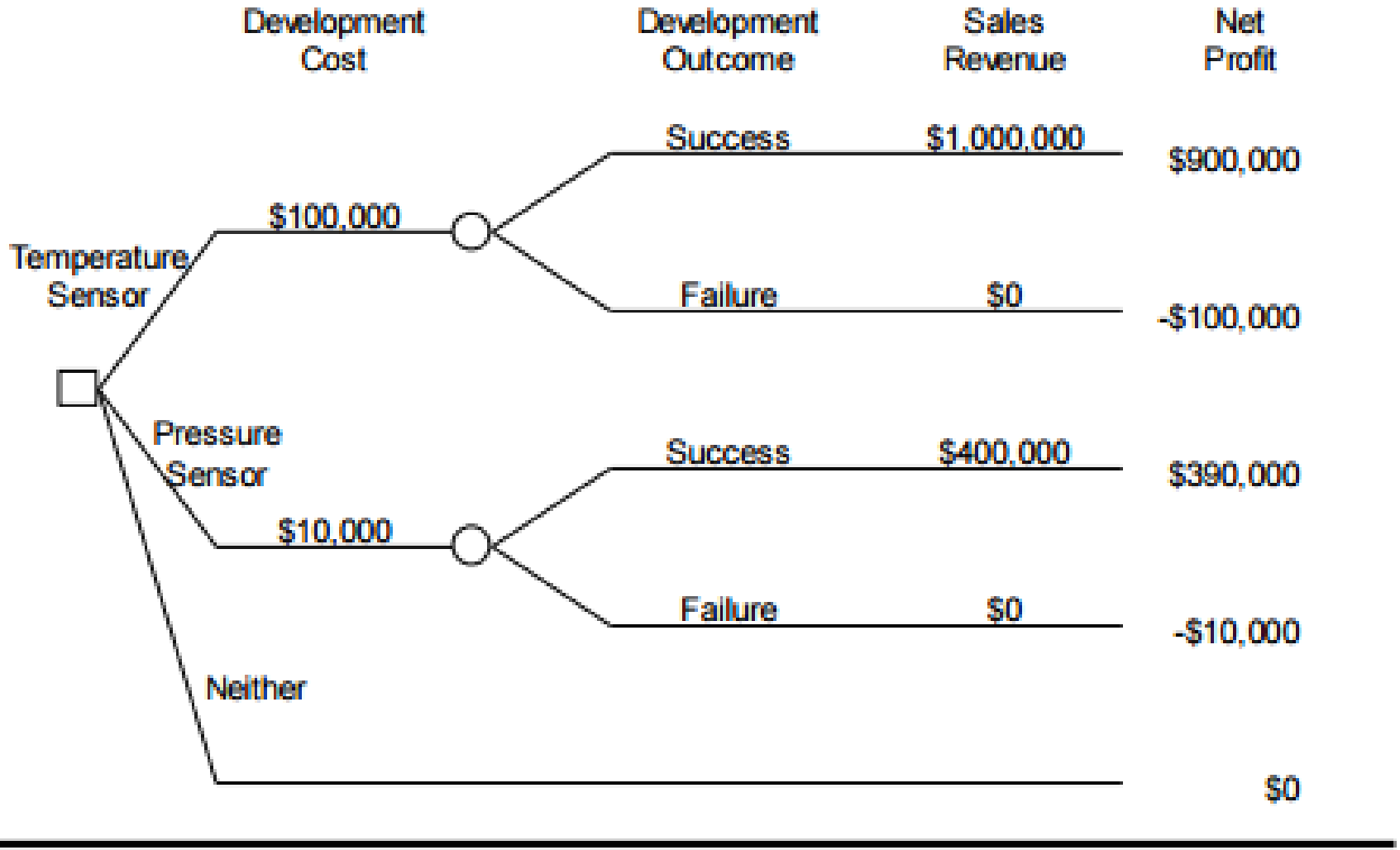

Figure 1.1 *Special Instrument Products decision*

Random Forests

- Random forests are used to refine complex decision trees and reduce variability in test sets
- Random forest are formed from bootstrap aggregating. Data samples are randomly selected with replacement to form the training trees and then the predictions are averaged.
- Usually a subset of variables are eliminated, so the random forest tree is simplified
- Advantages
  - Generally more effective than decision trees
  - Effective for all magnitudes of data
- Disadvantages
  - Large computing power

Random Forests

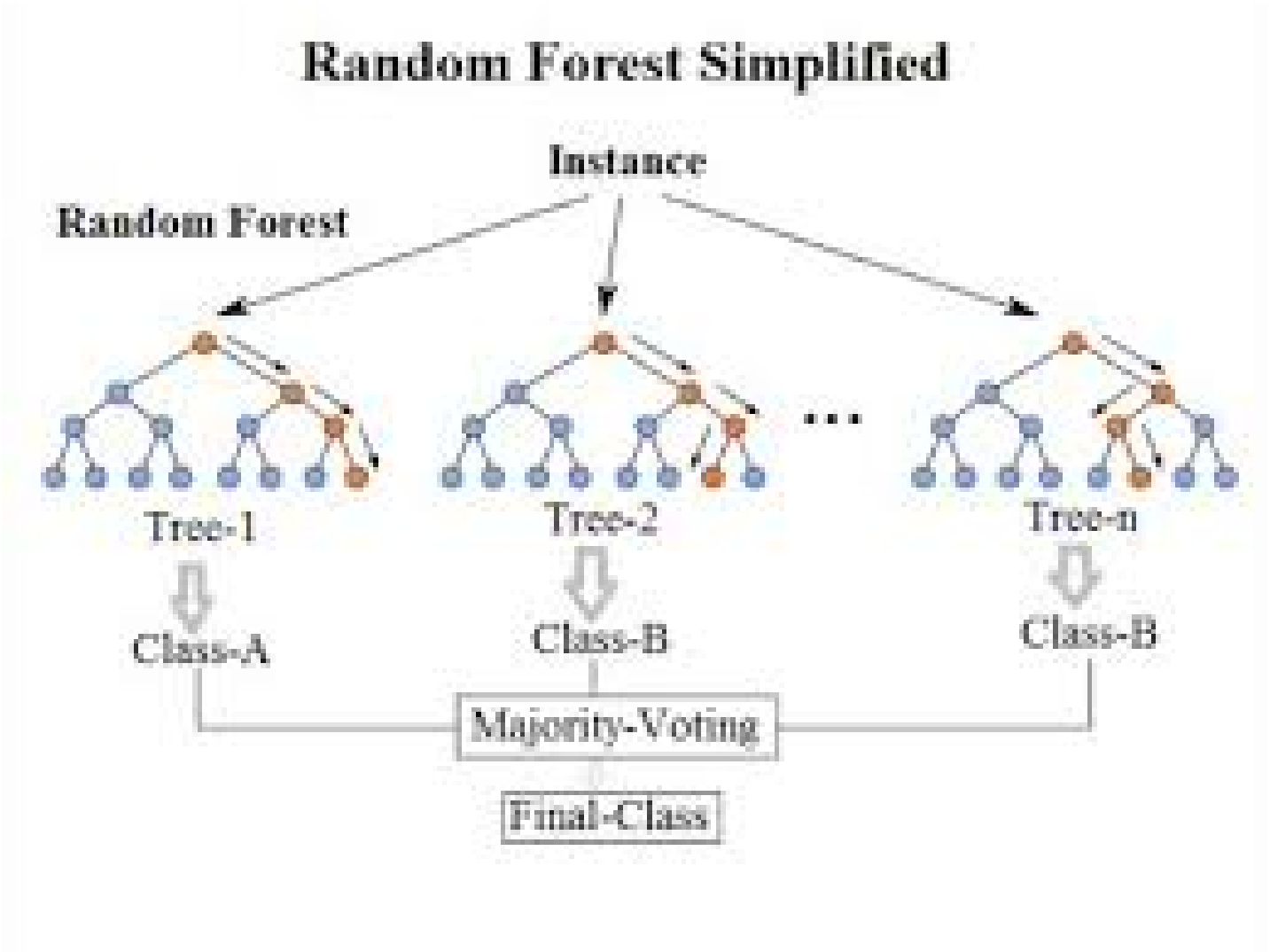

Figure 2:

Moving Forward

This project presents the unique problem of identifying trends in both space and time. Unfortunately, the spatial and tabular data are not easily related in the existing dataset.

The data as it is available to us poses several issues with our proposed solutions.

- There are no global relationships between findings and their location
- Inconsistent header names
- Inconsistent writing style and contents
- Inconsistent sampling periods
- Large number of NULL values

Moving Forward

We propose a collaboration between Cone Nose Assassins and the GIS experts at Hillsborough County to address data integrity and database design issues and implement the concepts presented here.

Relationships between tables need to be defined and implemented.

Data entry forms need to be codified and consistent.