Revisiting the CoIL Challenge 2000

DATA 621

Fall 2019

Critical Thinking Group #3

Corey Arnouts, Adam Douglas, Michael Silva

The Challenge

- Held from March to May 2000
- A data mining competition organized by the the Computational Intelligence and Learning Cluster
- 147 participants registered, 43 solutions submitted, and 2 winners chosen

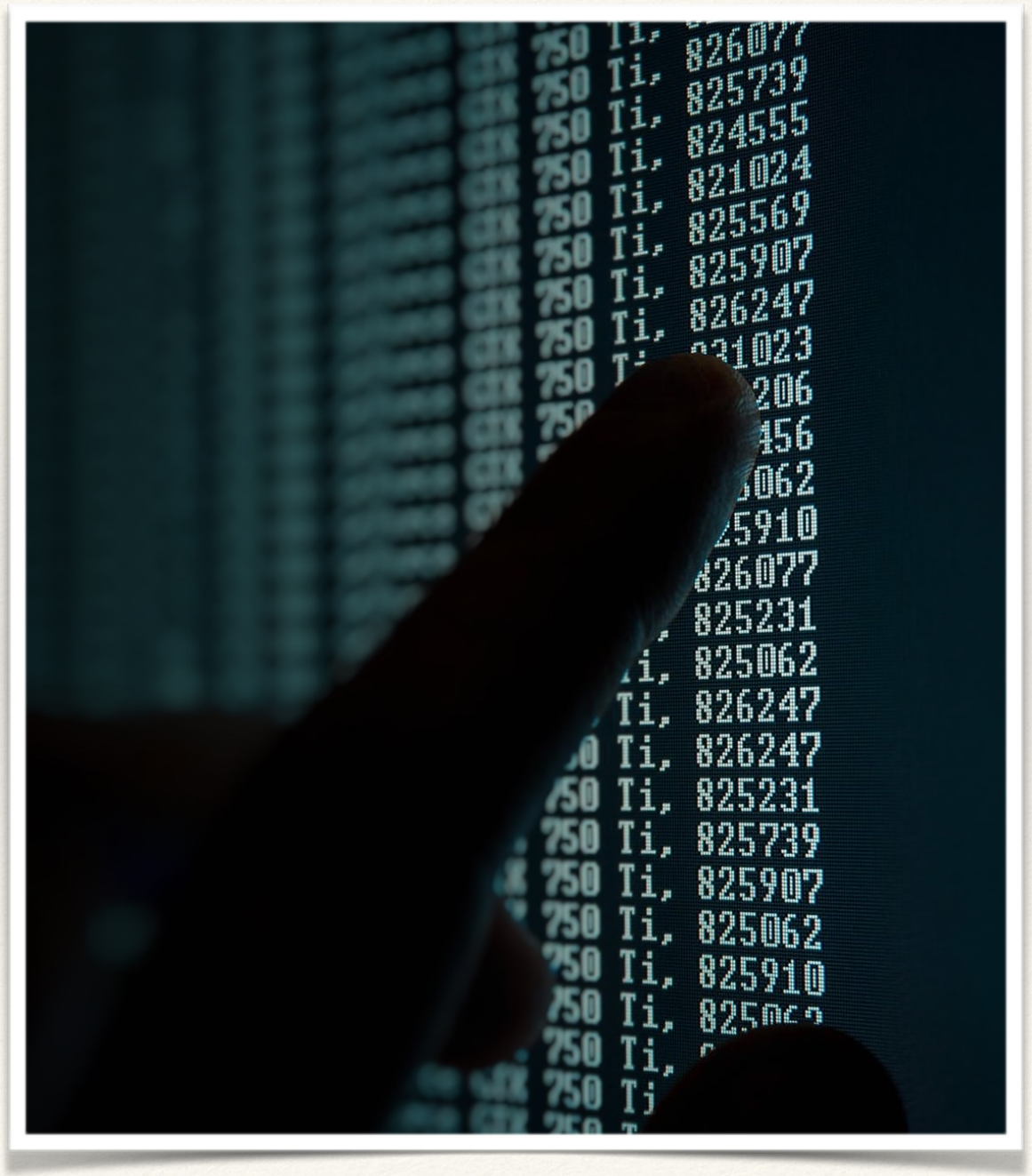
The Goal

Given a dataset with actual and potential customers:

- **Goal #1** - Predict who would be interested in purchasing a caravan insurance policy
- **Goal #2** - Describe actual and potential customers

The Data

- 5,822 observations (customers and potential customers)
- 86 potential predictor variables
- 1 outcome variable
- 4,000 observation data set used for scoring



Our Approach

- Literature review revealed that simpler models performed much better in the original challenge
- Due to the numerous variables, a decision tree algorithm was used to select the ones most likely to be relevant
- The unbalanced data set required us to use oversampling techniques

Our Approach (continued)

- Three (3) logistic regression models trained using variables identified by the decision tree
- Each model was repeatedly retrained, tested, and it's specificity measured

Outcome

- Our preferred model only contained 3 explanatory variables (2 of which were derived)
- We were able to correctly predict 165 of 238 customers correctly
- The best submission from the original challenge only identified 121 correctly

Works Cited

- http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/problem.html
- http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/
- http://liacs.leidenuniv.nl/~puttenpwhvander/library/2000synergy3.pdf

Questions?