Barclays Hackathon 2015

UCT's team: J Combrink, R Nhapi, D Rance, T Wolf-Piggott, T Phaweni, Q Dube, A Scarcella, G Dlamini

Introduction

Barclays' hackathon is a 24-hour event, in which teams from around the world 'gather' and tackle a set of real-world problems from Barclays. It held a high level of international competition, with a total of 54 team entries. We, as UCT's student team, faced off against 53 professional teams within the field of data-analytics. The event was due to run 12:00pm Thursday 8 October - 12.30pm Friday 9 October.

Topic Outline

12:15pm We were presented with five topics, of which as a team we had an hour to select which we felt we could break within 24 hours. The problems could be generally divided into Forcasting problems; Classification problems; and (one) HR problem. Our selected topic was **Collection data analysis**, and it involved predicting the success, as well as solving for the optimal days on which to hit a transactional bank account with a NAEDO (Non-Authorised Electronic Debit Order).

Begin the attack

We download a virtual machine on which to access the protected data. Four of us log on, and it seems we are unable to stay connected for very long, continuously getting kicked off. It took us approximately 10 minutes to realize we were kicking each other off; as a VM can only be used by a single machine simultaneously.

The team dispersed into several groups, some working on the final app, some considering how to most efficiently code predictors for the data, and someone performing regressions on the 43 or so variables. Logistic Regression was used along side an exploratory data analysis, looking for any obvious univariate relationships.

- All save one of the variables were categorical
- Almost none of the variables had reasonable correlations with the output, except for a select few, which were *perfectly* correlated with the desired output variable (the binary variable predicting whether the NAEDO hit would be successful). After a detailed analysis into the nature of the data, we realized the variables were indeed too good to be true; the perfectly correlated variables were *ex post facto* related to the success of a variable.

A Decision Tree and its sister, a Random Forest, were found to be simple a quick to implement learning algorithms which complemented well the observations made during the analysis, this was further extended to boosting algorithms and, separately a Neural Network was fitted.

Summary of results

6am We had a trained Boosting Model (which was great):

Table: Confusion matrix for GBM

		True status code		
		1	0	Total
Predicted status code	1	0.76	0.13	0.89
	0	0.03	0.08	0.11
	Total	0.79	0.21	1

Neural network (which was less than useless, classifying *every* NAEDO to a 'success' regardless of data input):

Table: Confusion matrix for neural network

		True status code		
		1	0	Total
Predicted status code	1	0.79	0.21	1
	0	0	0	0
	Total	0.79	0.21	1

We had implemented Bayesian techniques to describe the distribution of the best days on which to hit an account (initiate the NAEDO) for an individual, given his/her profile.

$$P(D|success) = \frac{P(success|D) \times P(D)}{m(D)}$$

$$m(D) = \sum_{D \in \{1,2,\ldots 31\}} P(success|D) \times P(D)$$

Along-side an app designed with *Shiny* to manage and present data, we had successfully answered all components of the *Collections Data Challenge*. Our final presentation was done by skype, condensing the cumulative results within a three-minute skype call.

As the only student (seven MSc. Statistics and one BSc. (Hons) Mathematics) and only Cape Town team, we are proud to announce that we were third distinguished team; placed second in South Africa (to Revenue Science) in this highly competitive international competition (The Global winners are the Bohemians from Czech republic).