 **U of W Data Mining Cup**

**Case: Winter 2013**

***Background***

With more than 50 billion made each year in the North America, credit card transactions are increasingly leveraged by fraudsters as a tool to obtain goods, services and cash without paying. The intensity of transaction fraud has soared in the past several years, with losses across the financial industry in the billions of dollars. Not only does fraud damage the profitability of financial institutions and threaten the security of consumers, it also greatly impacts the stability of the financial system as a whole.

To combat fraudsters, the financial services industry makes tremendous investments in IT, data collection, data mining, and analytical techniques to develop strategies that protect customers against fraud. In many fraud prevention programs, statistical models are built to estimate the probability of fraud for each transaction. Model predictions are used in conjunction with other inputs to define strategies that determine which transactions are approved or declined due to the high likelihood of fraud.

Of course, minimizing fraud loss is not the only objective as an incorrect rejection of non-fraudulent transactions negatively impacts customer relationships and can cause customers to attrite or to decrease use of their card. The goals of fraud strategy therefore should balance potential losses with a great customer experience (e.g. we cannot decline all transactions).

***Scenario***

Once upon a time there was a young and ambitious analyst, who was exploring, with hungry eyes, the vast world of business opportunities. Our protagonist decided to start a credit card company, since he or she was very analytical in nature and had a keen business mind.

Within a few short years, the company grew to become an industry leader for credit cards thanks to its cutting edge underwriting, driven by its powerful statistical models.

Guess what? **You are the young and ambitious analyst** and you are confronted with a real business problem.

You find yourself perusing your latest financials, thoroughly analyzing each cash flow. Everything seems to be in order, until you notice something odd in the fraud section. You drop and shatter your coffee cup in the most dramatic way possible when you see the massive red figures. It seems fraud losses have quintupled since two years ago and you start sweating at the thought of facing the board of directors with this news.

You immediately dive into your data to find the cause. After days of diligent research, you find the likely culprit: you haven’t updated your fraud defenses in years and attacks have been getting more and more sophisticated. Fraudsters are not much unlike electrons in that they look for the path of least resistance: if you fall behind on your defenses, most of the fraud in the market will shift towards you.

Beaming with pride, you run off to your board of director with your discovery. The board tasks you with building a transaction fraud detection model as well as a strategy to go along with it (which transactions do we approve and which ones do we decline as suspected fraud).

***Call to Action***

Your team has to design a transaction fraud prevention strategy based on historical data. In particular, your team needs to maximize the portfolio value based on your proposed strategy. Furthermore, the board has hired several other analytics firms to develop competing models and strategies. If any of them beat you, you will be replaced.

The data management team has prepared a dataset for your analysis. It includes detailed information for each transaction, the spending profiles of customers, the credit line information, and an indicator of whether or not the transaction was fraud. It has around 168K observations and 354 variables (full data dictionary to be provided to all registered teams). The dependant (target) variable is called “FRAUD\_IND” and is a binary indicator, flagging fraudulent transactions.

After consulting with various teams, you also get an assessment of the cost matrix for your fraud prevention strategy, which includes the cost for wrongly rejecting a non-fraudulent transaction (false positive), and the cost for wrongly approving a fraudulent transaction (false negative). There is no cost for approving a non fraudulent transaction (true negative) or rejecting a fraudulent transaction (true positive).

The board has asked you to conduct statistical/business analyses to answer the following questions:

*What is the probability of fraud for each transaction?*

*Which transactions should we decline to minimize cost?*

***Cost Matrix***

There are two categories of cost for your strategy:

*False Negative Cost*: cost for approving a fraudulent transaction (“Fraud Loss”)

*False Positive Cost:* cost for declining a non-fraudulent transaction (“Poor customer Experience Loss”)

The total cost of your strategy for a given dataset is the sum of the false negative cost and the false positive cost.

For simplicity, we assume that approving a fraudulent transaction costs the bank the total amount of that transaction.

There are two potential sources of false positive cost. Due to the unpleasant experience of getting declined, customers may suspend their card usage temporarily or they may close their account permanently.

The attrition cost for false-positives is calculated by multiplying the attrition probability and the estimated account value, given in the table below. The suspended purchase volume cost is calculated by multiplying the suspended purchase assumption and the net interchange rate. The four metrics differ greatly in low credit line and high credit line accounts.

|  |  |  |
| --- | --- | --- |
|  | Credit Limit<=2K | Credit Limit>2K |
| Estimated attrition probability | 0.02 | 0.05 |
| Estimated account value | $75 | $275 |
| Total suspended purchase | Current Purchase | 4\*Current Purchase |
| Net interchange rate | 0.0075 | 0.005 |

You are asked to develop an optimal strategy by minimizing total cost, not just fraud loss.

Note: For this case, revenue generated for each account is assumed to be independent from the fraud strategy proposed, thus the strategy that minimizes costs will also maximize profit.

***Statistical Package and Methods***

You can apply any statistical software package (SAS, SPSS, R, etc) for the modeling portion of the competition. Any statistical method can be used to build your model, such as logistic regression, linear regression, decision trees, clustering, SVM, neural network, etc…

In addition, you are told that one requirement for your work is to understand the features that are used to predict the fraudulent transactions (why did you select those variables and how do they relate to or shed light on fraud?).

***Strategy Assessment***

A random holdout sample (hereinafter referred as “validation sample”) will be used to apply & assess the fraud prevention rules you proposed. The total cost of your proposed strategy will be calculated as a major criterion to assess the results you submit.

***Data Access***

Data will be made available for download online at http://capitalonecampus.ca/.

***To Submit Results***

Each team (with a maximum of four members) needs to submit three files for a qualifying entry: your approval decisions for the validation sample, your code and your slideshow presentation

1. The **dataset** should:
   1. be in a simple file format (.csv, .dat, .txt, etc...)
   2. only contain two columns:   
      ENCRYPTION\_KEY – ID variable from validation sample  
      FRAUD\_DECLINE – A binary indicator for your decision, 0 meaning you approve the transaction and 1 meaning you decline for suspected fraud
   3. contain all transactions from the validation sample
2. Your **code** should :
   1. run in a common statistical language (SAS, R, C, etc...)
   2. be able to reproduce your results
3. Your **slideshow presentation** should:
   1. be a business proposal that includes a description of your approach, method and strategy
   2. be a maximum 15 slides not including appendix
   3. have at least one slide describing your approach and strategy
   4. have at least one slide describing the statistical methods used and the steps in the model-building process
   5. have at least one slide on the interpretation of the statistical, business insights, and long term strategy
   6. have all non-essential material in an Appendix section at the end of the slideshow

We will ask each team to submit their entry via email to waterloo@capitalone.com by 11:59:59 pm EST Monday, March 25th, 2013.

***Judging***

There are two rounds of judging.

*Round One*

All entries received by the submission date will be reviewed by a panel of Capital One associates who will select the top five (5) teams as finalists. Decisions will be made based on the following criteria:

* Performance of the proposed fraud prevention strategy, evaluated on the validation sample using the cost matrix
* Articulation of strategy in slideshow
* Evidence of analytical rigour and creativity

The finalists will be invited to the final round of judging. Teams will be notified of Round One judging results by 5:00:00 pm EST Tuesday, March 26th, 2013.

*Finals*

The five (5) finalist teams will be invited to give a 20-minute presentation on their business strategy proposal to a panel of leaders from Capital One who will be on-campus Thursday, March 28th, 2013. Presentations will be judged on the following criteria:

* Clarity and organization of thought
* Overall presentation skills
* Demonstrated analytical ability

The winning team will be selected by the judging panel based on the overall persuasiveness of the proposal. The winning team members will:

* Be **immortalized** by having their Team name engraved on the coveted Capital One Data Mining Cup, on display in the MathSoc display case.
* Be treated to **dinner with Capital One leaders** where you can relax and network to your heart’s content.

Also, if you’ve chosen to participate in the Capital One recruiting process, as an added bonus, you’ll skip the candidate pre-screening stage**.** For other prizes, see the short rules below.

It’s not easy running a credit business. The key is to manage your time and to play into the strengths of your team. The work can be split effectively between team members, but don’t approach each part independently. Analytics and business strategy are inextricably linked, so be sure to cross-pollinate.

**Thank you for your interest, and good luck!**

**Short Rules**

No purchase necessary. Contest is open to all students who have reached the age of eighteen (18), who are residents of Canada (excluding Quebec) and who are currently enrolled at the University of Waterloo (U of W). Register a team of between one and four members by 7:59:59 p.m. on March 22, 2013. Complete the U of W Data Mining Cup Case as a team and submit your entry by 11:59:59 p.m. March 25, 2013 for your chance to win! First place prize is a $25 gift card for each Team member, dinner with Capital One leaders and the honour of having your team’s name engraved on the Capital One Data Mining Cup! First-place prize value is approximately $75 CAD per person. Second- and third-place team members will also each receive a $25 gift card and an invitation to dinner, total value is approximately $75 CAD per member. Fourth- and fifth-place team members receive a $25 gift card. Chances of winning are based on the number and quality of the submissions. Visit [http://capitalonecampus.ca/] to register and for full contest details.