Robert Andrews

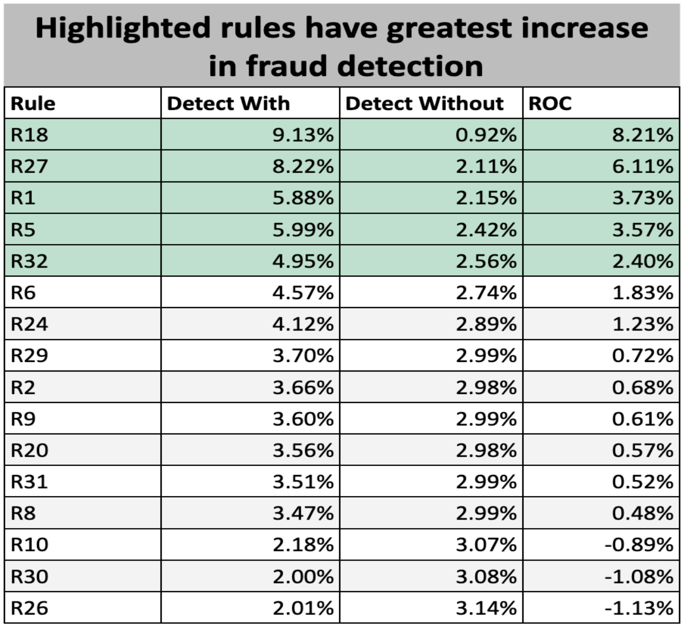
Memo Regarding RBC’s Fraud Prevention Technique Efforts

Daniel:

I hope this finds you well. I am writing to communicate the results of my research to develop a model that meets your fraud prevention objectives: reduce losses, enhance customer experience, lower costs, and minimize overall risk. I have created a model whose test results appear to meet these goals. To be clear, “fraud” refers to first-order fraud in this memo; however, I am more than willing to investigate solutions to our third-order fraud issue in the future – I digress.

RBC’s fraud-related business problem is two pronged. First, it is a major cost driver as it increases bad debt and strains existing fraud-prevention resources. We currently spend $11 million per year in fraud prevention data storage. [1] It is also costly to deploy fraud prevention techniques. Second, fraud potentially drives away customers in good standing through inaccurate fraud flagging, which can harm customers’ reputations and cause inconvenience associated with the absolvement of culpability. [1] It is imperative that we improve fraud prevention techniques to stop fraud on the front end while also protecting the customer experience. The model I created will help achieve this by leveraging our best Social Network Analysis (SNA) rules.

I investigated the relative performance of the sixteen SNA rules provided to determine which ones were effective determinants of fraud. Once the best rules were identified, they were incorporated into different models for testing. I was surprised to discover that most of the rules were not particularly useful.



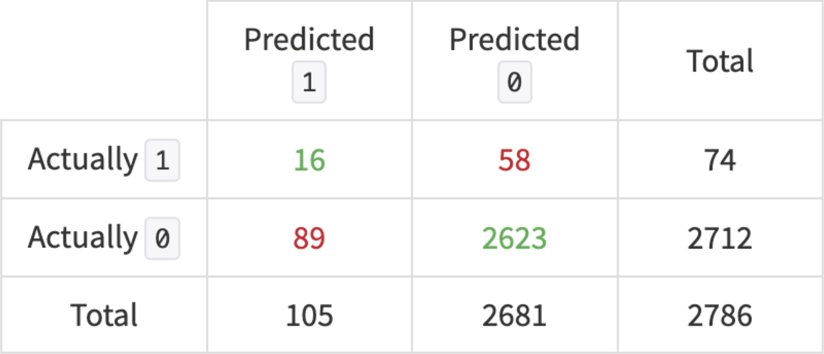
Highlighted rules have greatest increase in fraud detection

Rule	Detect With	Detect Without	ROC
R18	9.13%	0.92%	8.21%
R27	8.22%	2.11%	6.11%
R1	5.88%	2.15%	3.73%
R5	5.99%	2.42%	3.57%
R32	4.95%	2.56%	2.40%
R6	4.57%	2.74%	1.83%
R24	4.12%	2.89%	1.23%
R29	3.70%	2.99%	0.72%
R2	3.66%	2.98%	0.68%
R9	3.60%	2.99%	0.61%
R20	3.56%	2.98%	0.57%
R31	3.51%	2.99%	0.52%
R8	3.47%	2.99%	0.48%
R10	2.18%	3.07%	-0.89%
R30	2.00%	3.08%	-1.08%
R26	2.01%	3.14%	-1.13%

Rule effectiveness in order of greatest improvement in fraud detection when deployed

Exploratory data analysis (EDA) revealed that rules 1, 5, 18, 27, and 32 best predicted fraud detection. These were incorporated into the candidate models. The chart above, particularly the the “ROC” column, shows the net increase (or decrease) in fraud detection attributed to each rule. This also offered evidence to support discarding the other eleven rules since they do not appear to be useful in improving fraud detection. Rules 10, 26, and 30 even saw a net decrease in fraud detection when deployed.

I experimented with three models: logistic regression, light gradient boost, and random forest. I leaned into our fraud prevention goals to decide which model would be best. The key objective is to prevent fraud while enhancing the customer experience, principally by not targeting customers with false accusations. This translates to striking an appropriate balance between the true positive rate and false positive rate. Random Forest produced the best results based on these metrics. Although random forest is not the most easily explainable model in a business context, our goals clearly require the most accurate method rather than the most easily understandable. To not load you down with dense data science terminology, a random forest is essentially a group of smaller models whose outputs are combined into one final result. Rather than rely on the results of just one model, a random forest incorporates the best elements of multiple models – producing robust results.



	Predicted 1	Predicted 0	Total
Actually 1	16	58	74
Actually 0	89	2623	2712
Total	105	2681	2786

Confusion matrix of random forest model

In testing, the random forest model I generated reduced the ratio of false positive rates to true positive rates from RBC's current ratio of 89:1 to 5.5:1 – even besting your goal of 10:1 [1]. You can see this by taking the proportion of the top left number (green) and the bottom left number (black) in the chart above. While the 5.5:1 ratio probably will not hold when deployed to new data, it does strongly suggest we would major improvement over the status quo, even within range of your 10:1 target. This will strike the much-needed balance of rooting out fraud while preserving the customer experience.

Given the strong results of this model and of previous analysis, I recommend that we discard all SNA rules but 1, 5, 18, 27, and 32. Further, I recommend that we deploy the random forest model to predict fraudulent transactions. Additionally, we should modify our existing data pipeline to ensure that new data is procured and fed into the model so that its results continue to improve. While it is expensive to store the data we have, I am confident that the reduction in rule deployment and the cost savings from a drastically improved fraud ratio will make this a worthwhile investment until our data engineering team can discover new ways to handle our fraudulent transaction data.

Please do not hesitate to contact me if you have any questions.

Sincerely,
Robert

[1] RBC: SOCIAL NETWORK ANALYSIS. 2017, Richard Ivey School of Business Foundation