CIBC Machine Intelligence Hackathon - Instructions

Background

Fraud and money-laundering are issues that have been faced by banks and insurance companies for decades. As companies have moved more on-line it is more difficult for banks and insurance companies to know their customers and the opportunity to commit fraud has increased. Technology investments in fraud detection and anti-money laundering capabilities to combat the rise in fraudulent activity have grown significantly. Finance industry companies have regulatory anti-money laundering requirements stipulated by the governments of their local countries. Many fintech companies offer sophisticated fraud detection and anti-money laundering capabilities using artificial intelligence.
This hackathon offers students the opportunity to try their hand at fraud detection.

Problem Statement

Teams will be provided an insurance claims data set and will be tasked with finding fraudulent claim records and medical providers. The claims data will contain 1 year of medical claims. The data is based on a sample of actual claims data from a US healthcare company. All the individual identifiers and medical provider identifiers have been altered to protect the privacy of the individuals.
This is an unsupervised learning task.
The claims data format is a comma-separated file with the following columns

1) Patient Family ID
2) Patient Family Member ID
3) Provider ID
4) Provider Type
5) State Code
6) Date of Service
7) Medical Procedure Code
8) Dollar Amount of Claim

The code descriptions are not important to the task and are left out to protect the privacy of the individuals.
Useful Definitions

1) Unique medical providers (i.e. doctors) are identified by a provider id.
2) The type of provider is identified by the provider type (i.e. medical specialty)
3) The state code identifies which state the provider practices in

4) Unique patient families are identified by a family id
5) A unique patient is identified by combination of family id and family member id.
6) A unique doctor visit is identified by a unique combination of columns 1,2,3,6 from the above list.

Instructions

The task is to return two csv files with the doctor visits and doctors ranked according to degree of outlying behavior.

File 1 – A CSV file Outlying providers containing all provider records sorted by Outlier Rank.
1) Provider Id
2) Outlier Rank (1 being the worst outlier)

File 2 – A CSV file Outlying visits containing the top 100 outlying visits for each provider type sorted based on outlier Rank first and then provide type
1) Family ID
2) Family Member ID
3) Provider ID
4) Date of Service
5) Outlier Rank (1 being the worst outlier)

File 3 - Students should also submit a powerpoint presentation describing the methodology that they used and a self-assessment of results

Although there is no correct answer in unsupervised learning, the above output will be compared to that achieved using Daisy's sophisticated artificial intelligence outlier detection method.

Evaluation

Teams will be judged on their explanation of how they evolved the solution, why they choose their approach, and why it is innovative.

FINAL SUBMISSIONS

Submissions Form
https://docs.google.com/forms/d/e/1FAIpQLSe0LIPE9Ln6YBQQiVS7sMR_ycjiGQTpUGujrsWpbYp1cDaBZw/viewform?usp=sf_link

Please submit only once.

Deadline 9:30am on Saturday, September 22, 2018.