Anomaly Detection Exercise

Premise

You are providing a basic anomaly detection analysis to the security team to help identify suspicious authentication attempts and other anomalous authentication activities. The security team is noticing a change in failed log in attempts so they have asked you to analyze (synthetic) authentication data to better understand what insights there may be around suspicious activities.

Expectations and Deliverables

This exercise is expected to take between 1-2 days. We expect an initial analysis pass over this dataset, go to the depth you feel is appropriate. Please submit an R or Jupyter notebook with the code you used to answer the questions along with comments on your approach and what each code block is doing. Additionally, please also submit a pdf or html version of the notebook which include the output from each block (again this isn't expected to be ultra polished go to the detail you feel is appropriate)

Schedule: 1-2 days from time of receipt
Submission Materials:
- Notebook code file (ipynb, rmarkdown)
- Separate non-code file including code output (pdf, html)

We will be evaluating:

- The code that you use for data manipulation and basic modeling tasks (clarity and effectiveness).
- Your understanding of the relevant modeling and statistical concepts.
- Your ability to create understandable and appropriate visualizations or data summaries when needed.
- Your ability to think about how to improve and enable future analysis.

Simple modeling and analysis is sufficient - no need to over-engineer.

Dataset

You have been provided 4 data files

- `orgs.csv` - Contains info on organizations

- `users.csv` - Contains info on users
- `devices.csv` - Contains info on devices
- `auth.csv` - Contains authentication events

This is an entirely artificial dataset, so don't try to reason too hard about the motivation behind the users.

An authentication represents a single user attempting to authenticate to an application,

Data Descriptions

Device Data (`devices.csv`)

- `device_id`: device ID (PK)
- `type`: the type of device (mobile, desktop, laptop, etc.)
- `os`: operating system of the device (ios, windows, linux, etc.)

Organization Data (`org.csv`)

- `organization_id`: organization ID (PK)
- `type`: type of organization (corporation, llc, nonprofit, etc.)
- `start_date`: date the organization joined

User Data (`users.csv`)

- `user_id`: User ID (PK)
- `organization_id`: organization ID (FK)
- `start_date`: date the organization joined

Authentication Data (`auth.csv`)

- `time`: timestamp of authentication event
- `id`: authentication event ID (PK)
- `device_id`: Device ID (FK)
- `user_id`: User ID (FK)
- `result`: authentication attempt result { success : the user could access the service, failure : unable to authenticate for any reason }
- `method`: method of authentication (push, fingerprint, sms, etc.)
- `country`: country associated with the authentication attempt

Questions

Q1: Data Cleaning

Load the datasets and prepare the data for analysis. Document any assumptions, modifications, etc.

The attached datasets include the columns documented above. Please load the datasets, evaluate whether the data matches expectations, and prepare the data for further analysis. If you remove any rows or otherwise modify the data, please explain your reasoning.

Q2: Anomaly Detection Analysis

The security team has noticed a significant change in failed log in attempts. A sample of (synthetic) data has been collected and provided to the analytics team and your task is to provide an initial investigation into the data.

- Use the authentication data to analyze and visualize anomalies in authentication attempts. Consider factors like user org, devices, locations etc.
- In a short paragraph (5 sentences or less), briefly summarize your findings as though reporting to a product manager.

Q3: Data Improvements

How would you improve this dataset to make it more robust or provide additional insights? What remaining questions do you have about the data?