Cybersecurity Data Scientist Test

Job Description

We are looking for a Cybersecurity Data Scientist to join our Data Science team. You will be working on developing a real-time risk, anomaly and fraud detection platform, returning scores and signal intelligence about user behavior in less than 100 milliseconds of processing time to protect large FI and eCommerce customers from account hackers and fraud artists.

You will be working in our data science team under the product development division to build prototypes and new features. While a variety of fields can prepare an individual for this role (including software development, database management, and academic work) the key attribute is someone who possesses a "roll up your sleeves" attitude with an eagerness to learn and a desire to shoulder responsibility. We prefer mid-level career professionals (with 3-5 years of experience), however our ultimate goal is to find someone that is a good fit in our team. If you are willing to take on challenging tasks and have a passion for investigating massive sets of data then you are halfway there.

Job Duties and Responsibilities:

- Design, develop, and deliver AI/machine learning enabled solutions for behavioural authentication and fraud detection
- Build scalable, available, and supportable processes to collect, manipulate, present, and analyze large datasets in a production ready environment
- Articulate problem definition and work on all aspects of data including acquisition, exploration/visualization, feature engineering, experimentation with machine learning algorithms, deploying models
- Develop working prototypes of algorithms and evaluate and compare metrics based on the real-world data sets

- Data analytic responsibilities include pattern discovery, outlier detection, sample design, identification of appropriate analytic and statistical methodology, model development and documentation of process and results
- Oversee implementation of fraud detection and predictive models in production

Qualifications and Experience:

- 1-2 years of experience in a professional setting
- Degree in applied math, statistics, machine learning or computer science.
- Deep understanding of statistics and experience with machine learning algorithms/techniques.
- Passion for solving challenging analytical problems
- Ability to quickly qualitatively and quantitatively assess a problem
- Ability to work productively with team members, identify and resolve tough issues in a collaborative manner.
- Experience in applying machine learning techniques to real-world data.
- Strong understanding of Data Structures, Algorithms and Design Patterns
- Familiarity with scalable distributed data processing and visualization tools used in cloud based environments
- Experience with in Machine Learning, Predictive and Data Analytics and Data Analysis
- Experience working with programming languages, specifically, NodeJs, C, or Python, R, etc.

Bonus Assets:

- Experience with feature design and clustering
- Previous research experience in Machine Learning specifically anomaly detection
- Masters or higher degree in applied math, statistics, machine learning, computer science or relevant subject.
- Fraud modelling experience

Phone Interview Questions:

1. What is the difference between DOS attack and a buffer overflow attack?
2. To securely store a large file we decide to apply lossless data compression (e.g. zip) and encrypt the file using symmetric encryption such as AES algorithm. In your opinion should we encrypt the file and then compress the file or compress the file and then apply encryption? Please explain your answer
3. What are the three primary methods | technologies for user authentication?
4. What is the difference between firewall and intrusion detection system?
5. In your opinion what is the most severe security incident happened last year and why?
6. What is the difference between supervised learning and unsupervised learning? explain your answer using non technical terms
7. In few words explain the main idea behind Random Forest Algorithm?
8. What is overfitting and how we can avoid it?
9. What is logistic regression, and could you compare it to linear regression?
10. What is feature engineering and why it is important?

Written Test Questions

Programming Test

Please read the follows questions and provide your answers in the form of functions. Please test your work before submitting your answers. Please feel free to use any language of your choice (Python, Java, PHP, C/C++)

Question 1 - Given an array representing n points in a plane, write a function that that will find the closest pair of two points in the array. The function should return the distance between the closest pair of points

Each element of the array should contain:

1. name, e.g. "a"
2. x-coordinate, e.g. 1
3. y-coordinate, e.g. 10

Please indicate along side your answer what the performance of your approach will be in Big O Notation. If you break up the solution into multiple functions, please call your primary function "findClosest"

Question 2 - Given a HTTP access log, analyze it and calculate the percentage of failed requests in the log. The function should return an array of key-value pairs where:

1. the keys are the IP addresses
2. the values are the percentage of failed requests

Example: "127.0.0.1" => 0.25

It is OK to handle only 200 and 404 codes but you are free to support other codes too. If you break up the solution into multiple functions, please call your primary function "analyzeAccessLog"

Sample access log content:

192.168.2.20 - - [28/Jul/2006:10:27:10 -0300] "GET /try/ HTTP/1.0" 200 3395

127.0.0.1 - - [28/Jul/2006:10:22:04 -0300] "GET / HTTP/1.0" 200 2216

127.0.0.1 - - [28/Jul/2006:10:27:32 -0300] "GET /hidden/ HTTP/1.0" 404 7218

Data Analytics and Fraud Detection Test

In this written test, you are showing your ability to translate data and metrics into a convincing report. **This component is the most important part of the test** and will help set your investigative, analytical and report-writing skills out from other candidates.

This component will be evaluated on:

- Quality of your writing
- Report formating,
- Accuracy of your findings,
- Whether your graphs are persuasive, easy to understand and accurate

We have two MySQL tables: writtentest and writtentestflagged. The first table is the “writtentest” table, which contain data collected by our platform about purchase transactions. The second table, “writtentestflagged,” contains some events that have been flagged as fraudulent by a customer. Join these data sets together, analyze the data for more fraud and any interesting anomalies you can find, then prepare a report of 2-4 pages on this topic.

You may use non-SQL tools to help write this report (internally, we use Tableau and some Python, we also export CSVs for offline review). Please feel free to make use of any tools you are familiar with for graphing and reporting.

Suggested report format:

1. Overview
2. Provide metrics on the overall traffic for the week. Total account creation attempts, total Ips, accounts, devices
3. Identify the fraudulent attack and present specifics on it. IE Total accounts created, total ip addresses used, What user agent, browser, etc
4. Determine if there is more anomalous activity that hasn’t been flagged in the table, and present that information.

Your ‘audience’ would be a selection of non-technical managers and fraud analysts.