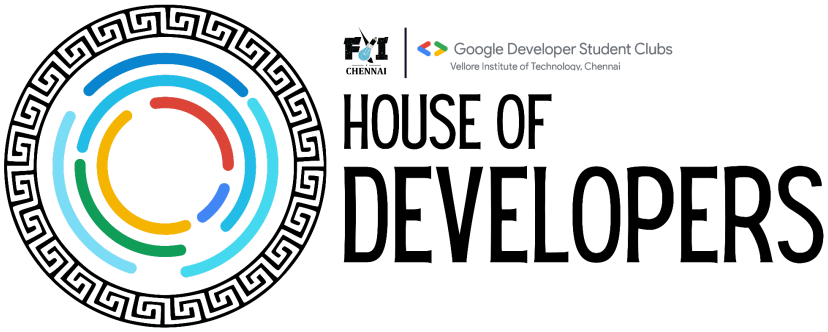

Hackathon - Data Science Track

Data Science Problem Statement

Ms Oprah Windfury, an employee of a government agency in charge of the machines in the office, realized that they had been behaving strangely for the past couple of days. They were frequently crashing and receiving unusual error messages, but most importantly, almost all the computers carried several failed logins, hence displaying compromised conditions pertinent to the workings of the machines and devices. Worried that sensitive information may be compromised, she resorted to venerated specialists who told her what she feared most, that a menacing malware had infected them, the Peepeepoopoo. Malware, from a previous incident she knew, was a well-run, well-funded sector devoted to getting around established security measures. Once the malware has infected a machine, it can harm consumers and businesses in several ways. The specialists advised her to seek the cybercrime department's intel to acquire more perspective. The cybercrime department informed her of the increasing cases of malware infection.

The members of the cyber crime department have dedicated their lives to equipping cyber security to the vulnerable commons, legitimate businesses, and governments that needed their help. For the later part of their career, they spent their time ethically collecting data on systems they encountered, hoping it to help their cause. Each row in this dataset corresponds to a machine, uniquely identified by a machine_id. They recorded essential features such as machine version, access to the root shell, and region. Ms Oprah Windfury had decided to approach the data science community to

postulate a solution to avoid vital and sensitive information of national importance being given away. When such a cause presented itself to the data science community you are involved in, you, a pompous melioristic member, decided to work on this problem. An ideal way for a data science practitioner to aid Ms Oprah would be to help automate the identification of such outlier machines using the data made by cybersecurity specialists. The team that can make accurate predictions is to be rewarded with riches and honour.

Judging Criteria

Stage 1 – Understanding the problem and defining the project
Stage 2 – Exploratory and Statistical Data Analysis
Stage 3 – Defining the solution space containing all possible solutions
Stage 4 – Architecture and workflow diagram of the proposed solution
Stage 5 – Partial Implementation of the proposed solution
Stage 6 – Final completed Implementation of the proposed solution
Stage 7 – Presentation of achieved results

Stages 1-4 will have 20 points for each respective stage. Stage 5 will have 10 points, Stage 6 will have 60 and Stage 7 will have 20 points.

Rounds (2)

Round 1
(80 Points)

The round 1 submissions will be judged and scored on the basis of Stage 1, Stage 2, Stage 3 and Stage 4 of the 7-Stage Process. The submissions are to be made in the form of a write-up and submitted as a pdf. You are expected to summarise the problem and identify the features and target variable in stage 1. In stage 2, you are required to perform a detailed exploratory and statistical data analysis. Stage 3 involves you making a detailed literature survey of all the possible solutions. Stage 4 requires you to create a detailed workflow diagram of your final proposed model(s) and methodologies and explain them with a write-up.