Who is suspicious?

1. ABSTRACT

Criminal profiling is used as a technique used to identify the perpetrator of a crime by identifying the personality and behavioral characteristics of the offender based upon an analysis of the crime committed. The sensational and dramatic elements of profiling is often portrayed in movies, television shows, and books but how does it fare as an investigative tool against crime in real life?

Well, more often than not, there has been doubts within the crime-fighting community about the viability of profiling as an investigative tool. Why? This is because there are views hat a criminal profile only gives a broad indication of the type of person who may have committed the crime. It does not indicate a specific individual who happens to fit the profile. The profiler is therefore unable to say whether it is more probable than not that a specific offender did, in fact, commit the crime. Nevertheless, criminal profiling has proven to be helpful in solving some criminal cases.

Definitely though, more research has to be done before criminal profiling can truly become a useful part of the criminal investigation process. This is a fascinating topic and hence, through our project, we aim to use a range of data visualisation techniques to develop an R Shiny application to maximise the insights obtained from studying the different aspects of an individual's daily life e.g.credit card transactions and email communications in order to successfully profile an individual. The analytics and design choices made in the development of the application and initial findings and future work are discussed.

Keywords— Profiling, spending patterns, Analysis of credit card transations, Interactive Data Visualisation, network graphs, R Shiny

2. 1. INTRODUCTION

The goal of the annual Institute of Electrical and Electronics Engineers (IEEE) Visual Analytics Science and Technology (VAST) Challenge is to advance the field of visual analytics through competition. The 2021 IEEE VAST Challenge brings back a classic challenge from 2014 to see how approaches and techniques have developed since the original release of the challenge. The background of the challenge is as below:

In January, 2014, the leaders of GAStech are celebrating their new-found fortune as a result of the initial public offering of their very successful company. In the midst of this celebration, several employees of GAStech go missing. An organization known as the Protectors of Kronos (POK) is suspected in the disappearance, but things may not be what they seem. It appears that certain employees of GAStech may be involved in the disappearance.

Aligning with our intent to create an application that aims to aid the profiling process, our application will serve as an interactive tool for users to visually investigate and identify which employee profile is indicative of suspicious behaviour. This paper documents our approach to designing and developing the interactive application targeted at crime investigators. This introduction is followed in Section 2 by an explanation of our motivation and objectives. Section 3 details the data used and methodology selected. Section 4 provides a visual overview of the final product. Section 5 concludes the report and offers ideas for further development.

3. 2. MOTIVATION & OBJECTIVES

This project was motivated by our findings that there were suspicious activities within GAStech itself which were worth investigating. This project aims to create a data analytics applications to visualize these suspicious activities and relationships for users to judge, who exactly are the suspicious people in GASTech. On a larger scale, this could assist in profiling which can be used for generic crime investigations.

Through our application, users can start to decipher for themselves what defines suspicious behaviour and how can we make use of everyday data to raise red flags on suspicious behaviours. This project aims to understand better the individuals and organizations that are involved in this situation. We do this by exploring the following:

- Conducting exploration data analysis and inferential data analysis on
  - Credit card expenditure data

  - Email headers data
  - Employee categories

- Delivering a R-Shiny app that achieve the following through an interactive user interface design:
  - Identifying any anomalous or suspicious behavior.
  - Identifying formal (work-related) or informal (non-work related) relationships.

  - Discover any associations based on common interest given in the data.

  - Discover relationships between CC expenditure, email headers and employee records.
  - Decide who are the suspicious GASTech employees * Obtaining a holistic profile on these suspicious employees.

4. 3. METHODOLOGY

Our methodology is as below:

- Data preparation using dplyr and other R packages.
- Analysis of VAST 21 data set with background research using some of the following methods:
  - Exploratory Data Analysis (EDA) methods in R.
  - Inferential Analysis methods in R.
  - Network Analysis in R.
- Creating a R shiny dashboard showing our findings/insights and conclusions:
  - R Markdown development for functionality checks
  - R-Shiny app development for user interactivity

4.1 3.1 Data

We'll be using data sets from Mini-Challenge 1 [MC1] and Mini-Challenge 2 [MC2].

From [MC1], we will be using the email headers and employee records data. From [MC2], we will be extracting insights based on credit card transactions data.

We will also be joining the two data sets based on individuals to analyzing attributes by features across data sets.

4.2 3.2 Shiny Architecture

Development of the interactive tool was done on Shiny, and R packages were used to build interactive web apps. Shiny is widely adopted in the data analytics industry because:

- It has a framework that makes it user-friendly to collect input values from a web page, with R code written as output values back to the web page.
- The input values can be modified by the user at any time, through interaction with customisable widgets.
- The output values react to changes in input values, with the resulting outputs being reflected immediately.

The design of our Shiny web app is as follows:

1. Main layout - This will consist of two pain parts:
   - A top navigation bar that will help the user to navigate to the various modules that will be present in the blue area.
   - An input bar where user can decide for themselves who in GASTech is suspicious.

   The objective of this layout is when the user switch from one tab to another, based on the visualization and their own analysis, they are able to decide for themselves who they think are the most suspicious. Then at the end (in module 4), we would see the profile of the user's decision of their most suspected employees.

2. First module - EDA. On the side panel, we would include options for interactions. Then on the main panel, we will have histogram and data table to visualise the distribution of credit card spending amongst the GasTech employees.

   The filters applicable for interactions for the histogram include:
   - Employment type (Security, Executive, Administration...)
   - Expenditure type (Food, Retail, Company...)
   - Time category: Weekend vs Weekday

   Depending on our selection of the various filters, the Data Table will be automatically filtered to represent the data corresponding to the filters.

3. Second module -Inferential Analysis. On the side panel, we would include options for interactions. Then on the main panel, we will have violin plots to showcase the statistical analysis of credit card spending.

   The filters applicable for interactions for the violin plots include:
   - Employment Type (Security, Executive, Administration...)
   - Expenditure Type (Food, Retail, Company...)
   - Time Category: Day of week

   Appropriate statistical tests will be conducted as well. For the purposes of constructing the violin plots, outlier points will be excluded, i.e. credit card transaction which have price values that are outliers will not be included.

4. Third module – Network Graph.There will be similar filters to what we have for the first 2 modules, with the addition of Email Category (work related vs non-work related).

5. Fourth module - Parallel Plot.

   To facilitate exploration by users, we have created a user guide as well.

4.3 3.3 Analysis Techniques

4.3.1 3.3.1 Histogram

A histogram is a plot that illustrates the underlying frequency distribution of a set of continuous data. This allows the inspection of the data for its underlying distribution (e.g., normal distribution), outliers, skewness, etc

4.3.2 3.3.2 Violin

A violin plot depicts distributions of numeric data for one or more groups using density curves. The width of each curve corresponds with the approximate frequency of data points in each region. Densities are frequently accompanied by an overlaid chart type, such as box plot, to provide additional information.

4.3.3 3.3.3 Network Graph

Network graphs show interconnections between a set of entities. Each entity is represented by a Node (or vertice). Connections between nodes are represented through links (or edges). In the context of our shiny app, employees represent the nodes while the links represent the email correspondences.

4.3.4 3.3.4 Parallel Plot

Parallel Plots are a visualization method for the exploration of categorical data. They focus on the data frequencies instead of individual data points. The technique is built on the axis layout of parallel coordinates, with boxes representing the categories of data (e.g. gender, department, spending patterns etc.) and parallelograms between the axes showing the relations between categories. Our app enables the user to interactively remap the data to up to five levels of categorization.

5. 4. DESPCRIPTION OF PRODUCT & FINDINGS

6. 5. CONCLUSION & FURTHER IDEAS

This paper set out the development of a web application targeted at exploring how we could use visually motivated tools to facilitate the profiling process. The project was motivated by our realisation through the VAST 2021 project that there were suspicious behaviours demonstrated by some individuals which were worth investigating.

The application was developed using the Shiny architecture on R, supported with a range of statistical packages to provide users with a whole range of techniques to derive insights from the data. Definitely though, the range of techniques and ideas executed in this project is non-exhaustive and We suggest that the below can be explored in the future development of the app: