Introduction

Welcome! Analyze This 2014 is the third edition of a first-of-its-kind Pan-IIT data analytics competition by American Express. Through this game, you will get a firsthand experience of the various facets of the exciting field of Data Sciences.
By the end of this 9 day nerve-wracking, nail-biting, roller coaster ride we are sure you would agree that Data analytics is as addictive as gaming.
Gear up and Game On!!!

The sections below have details on the

1. Problem Statement
2. Data for analysis
3. Clues and Milestones
4. Tips on data analysis
5. Popular data analysis techniques

Background

The increasing crimes in the Gametris islands had left the citizens terrified yet the police were clueless. One day, acting on a tip-off the police conducted a daring raid at a mafia premises. One of the items seized was a Flash Drive.

Commissioner: Sir, I am really proud of my team for the successful raid. All the criminals who were involved in the crimes in the city of Ebetor for the last 3 months are behind bars. Thanks to the Flash drive.

Chief: That's great news! What all does that Flash Drive have?

Commissioner: Sir, the mafia operates like a recruiting agency. Once they plan a crime, local criminals come forward to execute it. But the mafia is very shrewd. Based on past history they carefully handpick the criminals who they think would execute the crime successfully. I must say, their criteria are foolproof. They always manage to execute their crimes without leaving a trace. This Flash drive had the list of those criminals who executed the crimes in Ebetor.

Chief: That's stunning! I am glad that we could solve the mystery for Ebetor. But, what concerns me more now are the other parts of the GAMETRIS ISLANDS – Sevestra & Kaldar. We know the same mafia must be operating in those cities as well. The pattern looks similar! We have to stop this somehow!

Commissioner: Sir, the flash drive does have information on Sevestra and Kalder. We have data on the past history of all the criminals. We don't know who the mafia has picked but we could

simply arrest all of them before September – That's what the informant mentioned as the planned execution time! There are 20,000 of them

Chief: But we don't have the resources to chase all of them…We have to find a better way to do this and do it fast

Commissioner: How do we choose whom to arrest? Whom do we let go? We can't go wrong. If we let go of even one criminal, it would mean a huge loss for us. Besides, there are some criminals who could be convinced to be police informers – spies in the mafia…

Chief: Rightly said, in fact, if we end up arresting a former criminal now leading a peaceful life, there will be a lot of outrage in the press. We need to be very careful

Commissioner: Sir, based on the data in the flash drive, we could try identifying the criminals of other cities – they would have the same criteria to select the criminals… but how do we handle the data? We have no experience in that.

Chief: I think we are going to need some external help. It is high time the Gametris Islands are freed of the ever increasing crimes. I am going to make a few phone calls to arrange for the same.

Problem Statement

The mafia has 2 crimes planned in the cities of Sevestra and Kalder of the Gametris Islands and criminals offered to execute these crimes. The mafia heads selected criminals for each crime based on their past history. Also, unknown to the mafia, some criminals are potential spies. Given the right incentive from the police, they could become police informants

As Guardians of Gametris, based on the data in the Flash drive captured during the raid, you have to:

1. Identify which criminals were selected for a crime.
2. Identify the potential spies. (Bonus points for every correct identification of a potential spy)

Remember,

1. The criteria for selecting criminals for a crime are the same irrespective of which city the crime will be executed.
2. A criminal from a given city cannot offer to execute a crime in another city. Gang members guard their territory very fiercely.

3. **Note that the "acceptance rate" of the mafia i.e. the % of criminals chosen for a crime remains similar across cities.**

Data

The file "Analyze This Data.zip" contains 4 spreadsheets.

1. **Training_Dataset.csv**: This data is for the city of Ebetor. The police have already arrested the criminals. The data has information on:
    a. Past history of all the criminals from Ebetor who offered to execute a particular crime. (One criminal can offer to execute only one crime)
    b. Whether they were selected by the mafia heads to execute the crime
    c. Based on police records, whether the criminals were spies or not
2. **Leaderboard_Dataset.csv**: This data is for the city of Kalder. It has information on the past history of all the criminals from Kalder who had offered to execute a particular crime. (One criminal can offer to execute only one crime)**. (Remember, some of these are potential spies)**
3. **Final_Dataset.csv**: This is the data for the city of Sevestra which the police will target next based on the information you provide. This data also has information on the past history of all the criminals from Sevestra who had offered to execute a particular crime. **(Remember, some of these are potential spies)**
4. **Data_Dictionary.xlsx**: This sheet will give you the descriptions of all the variables contained in the 3 datasets above.

Please note that the **Leader board data** submissions are restricted to only **10 submissions per day per team** and for the **Final dataset** you can submit **only one solution.** For further details, please refer to the submission guidelines document available at the link below:
http://www.axpindiacampus.com/AnalyzeThis/campusactivity/submission_guidelines.php

Clues and Milestones

During the game there would be **2 clues** released in order to help you solve the problem better.
These clues have the potential to give an extra boost to your solutions provided you can use them to your advantage. Check your mails regularly for information about the clues and keep an eye on the web site too…

We have defined **2 Milestones** on the scores calculated for the Leader Board Submissions.
The first 2 teams to cross a milestone will be awarded. The milestones will appear on the leader board on the site.
Milestone 1: **110,000**
Milestone 2 would be released during the game! Stay tuned!

Tips on Data Analysis

Following are some tips for the uninitiated on how you can approach this data analysis game.
Any exercise in the field of data analytics would start with understanding the data. So, start off by understanding the datasets and descriptions provided to you.
Once you are familiar with the data, try to answer these questions:
1. What all data do I have?
2. What all data is useful and what is junk?
3. How can I organize this data to solve my problem?
4. For the mafia, which variables would define a criminal's ability?
Then, try to build the variables on the training dataset, define dependent and independent variables and then start modeling on the Training Dataset. You need to match the mafia's choice of criminals.
Once you are satisfied with your model, use it on the Leaderboard Dataset (The data for the city of Kalder) and come up with your estimates of which criminals would have been picked. Follow the submission guidelines and upload your estimates. Your submission will be evaluated real time and you can compare how well you have estimated against other participants.
Keep fine tuning your estimates by trying to increase your leader board scores. Keep an eye on the clues to better your solution. Once satisfied, use the same logic to estimate the criminals that could have been picked for Sevestra (Final Dataset)
You can use any tool, write your own algorithms, and implement any predictive modeling/Data analysis methods you may want to. For your final submission, you will have to provide details of the techniques you have used.

Popular Data Analysis Techniques

1. **Regression:**

   Regression is a mathematical process used to find a function that closely fits a series of data. The analysis involves defining the function that minimizes the difference between the data point and the value predicted by the function. There are several different techniques, the most common being by the method of least squares.

   For example, say you wanted to find an equation that dictated a certain stock's performance. You could take the closing price of that stock for every day in the last year. You then would be trying to figure out what equation satisfies all those points. The equation could be used to try to predict future performance.

2. **Logistic Regression:**

   Say, you want to figure out whether the stock price for a certain day would go up or not. You would again have the closing price of that stock for every day in the last year. We can do this using Logistic Regression. It gives you the probability of stock price rising.

3 Support Vector Machine:

Imagine the previous scenario. In addition to closing price we have say some more indicators like volume traded as well, and we have a reason to believe that the price (as is often the case) is a complex function of these indicators. Then, to predict the upward or downward trends, SVM could be a better technique for the solution.

4 Neural Networks:

Again, referring to the previous example, let's say, that we have certain indicators which are themselves complex functions of several different variables, and suppose we want to use them for the final prediction. In such a scenario, neural networks may give a better solution.

A point to note, as we go down this hierarchy we might end up over fitting the data.

5. **Clustering algorithms :**

Clustering algorithms are used in search engines that try to group similar objects in one cluster and the dissimilar objects far from each other. It provides result for the searched data according to the nearest similar object which are clustered around the data to be searched.

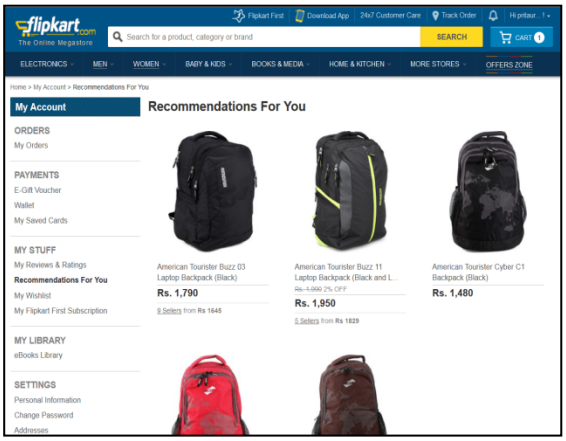
As an illustration, Google uses clustering algorithms to classify different contents as News by parsing though the matter and examining the keywords.

6. **Recommendation engines:**

Amazon/Flipkart/Netflix use collaborative filtering for recommendation.

In essence, the algorithm represents each customer as a vector of all items on sale. Each entry in the vector is positive if the customer bought or rated the item, negative if the customer disliked the item, or empty if the customer has not made his or her opinion known. Most of the entries are empty for most of the customers. The algorithm then creates its recommendations by calculating a similarity value between the current customer and everyone else.

Example –



7. **Naives Bavesian Text Classifier**:

   The best known use of Naives Bayesian classification is spam filtering. It is a probabilistic classifier based on Bayes' theorem.

   For example, Emails use Bayes' formula for calculating the probability of an email to be classified as a spam, given already existing spams. This can be done by calculating probabilities associated with each word of the text to be classified as a spam.