HACKATHON

IE7275 Spring 2024– Data Mining

Instructions for Hackathon's submissions

1. Teams are allowed to submit only once. Any revised files or submitting any of the tasks post 5:00 PM EST shall be disqualified. **There will be no editing allowed post submission.**
2. The format allowed for the submission file is in ".ipynb" file for and Dashboard with publicly accessible link for Task 2. One of the team members can submit the submissions via email on the following address: **ie7275datamining@gmail.com**
3. A reference template for the submissions will be provided during the Hackathon. All teams should only use the template provided for Task 1.
4. All teams need to mention their Team name and their members name in the "ipynb" file as per the instructions mentioned in the template provided.
5. Teams with similar code/pattern/ having plagiarism shall be disqualified.

The teams need to submit both tasks (Task 1 and Task 2) before 5:00 PM EST. If teams successfully submit Task 1 before 2:30 PM EST and Task 2 before 4:30 PM EST, they get 5 bonus points for each task. (Note: Bonus points are limited to hackathon only)

Attendance for the Hackathon will be awarded. The lowest grade of the homework will be replaced by 100 (e.g., homework 1,2,3 is 100,90,80 points, respectively. Homework 3 (80 points) will be 100).

Dataset: https://bit.ly/3KXg4y0
U.S. Offense Type by Agency 2012.xlsx

Use this link if the above link doesn't work:
https://northeastern-my.sharepoint.com/:x:/g/personal/mehta_pre_northeastern_edu/EWrDpLBQQbdBm1Q_wngmjP4B2guBC4G1ZZur76ntfaz8UQ?e=yJSveM&wdLOR=c3B05C541-83B6-5C4D-A3C7-AB72D18258A7

MetaData: https://bit.ly/37pN6bl

Consider "Total Offenses" as sum of Crimes Against Persons, Crimes Against Property, and Crimes Against Society

Data description: The FBI collects these data through the UCR Program's NIBRS. This table provides the number of offenses distributed by offense type.

This table counts offense types using the following rules:

- Offenses – if the offense type in Data Element 6 (UCR Offense Code) is:

- Crime Against Person: count one for each victim, i.e., Victim Segment.
- Crime Against Property: count one for each unique offense type.
- Crime Against Society: count one for each unique offense type.

- The data used in creating this table were from law enforcement agencies submitting 12 months of NIBRS data to the FBI UCR Program for 2012 and were converted and published in CIUS, 2012. Upon meeting both criteria, the agencies' actual NIBRS submissions were used in this table.

Task #1:

Template for Task #1: https://bit.ly/3xuGTFK

Data cleaning & Exploratory Data Analysis: **20 Points**

a. Top 5 states with highest number of *Assault offenses* registered in all the agency types except cities. (Show pivot of sub-types of Assault offenses)
b. Which category of crimes were most registered in universities?
c. Compare offenses at Michigan State University with offenses at all other universities.
d. Which provinces have state agencies with lowest number of digital offenses registered (Credit Card/Automated Teller Machine Fraud, Wire Fraud)
e. Which category of agency type and their respective agency names have the highest number of offenses registered per million people?
f. Geospatial visualization of the data showing the offence type with highest number of offences in that province (Hint: for each state find the offense type with highest number of offenses and create a geoplot)

Suggestion:
Select the best type of visualization (along with title, data labels, etc.) in order to show the insights for the above questions.
Mention your interpretation along with the visualization
Feel free to share other insights as well by searching other datasets apart from the one provided
Only use the provided template for this task

Data modelling: **40 Points**

a. **Data: X**: Population, Drug/Narcotic Offenses, Drug/Narcotic Violations, Drug Equipment Violations, Theft from Building, Theft from Coin-operated Machine, Theft from Motor Vehicle, Theft of Motor Vehicle Parts or Accessories. **Y**: Total number of offenses

Task: Build a linear regression model that fits best to the above data. You can use feature engineering to derive new features from columns in X.

Output: Can linear regression model fit the above data well? What type of linear regression model performs best? Justify your findings.

b. Build a statistical model of your choice from data available predicting total number of offenses.

Evaluate performance of the model. Show and explain the findings.

Github project for ref: Link

Task #2: **40 Points**

- Upload the same FBI UCR dataset (cleaned) in the data warehouse (GCP Big Query/AWS Redshift) and connect it with Self Service Business Intelligence tool (Tableau Recommended)
- Create an interactive dashboard

Follow Visual best practices: Link

Make sure you're the visualization is self-explanatory and easy to comprehend for the viewer.

Proper title, footnotes, data source link must be included

Deliverables: Share public accessible link of the dashboard with ie7275datamining@gmail.com

Leaderboard: Link

Here you need to update your model evaluation parameters for **Task #1 b** – Data models