DSCI-549: Introduction to Computational Thinking and Data Science

USC Viterbi School of Engineering

Homework 8

Goals: The purpose of this homework is to learn to analyze real data science scenarios and to describe data collection, pre-processing, and analysis steps using technical data science terms. (100 points).

Assignment: Complete this assignment in a Word/PDF document, with sections that correspond to the requirements below. Submit the homework on D2L.

1. Propose a data science project based on one of the scenarios listed below. You are free to choose to complete this assignment about any of the scenarios.
    a. Based on the scenario, describe the data science project in your own words.
    b. Describe in technical and data science terms (underline each of them):
        i. The data that you would use for your project. You may use all the data described in the scenario or only use some of the data. However, you can only use the data described in the scenario.
        ii. The data pre-processing steps that you anticipate will be required based on the description from the scenario.
        iii. The kinds of analysis that you would propose to do in order to achieve the goals of your project. **Note**: when considering the data pre-processing and analysis needed, take into account the challenges and costs described in the scenario.
    c. Sketch a high-level workflow for the data pre-processing and analysis steps. Describe the workflow have sketched.
    d. Enumerate and describe the outputs that you believe would be generated.

IMPORTANT NOTES

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in the Student Handbook https://policy.usc.edu/studenthandbook/. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, https://policy.usc.edu/research-and-scholarship-misconduct/.

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the American Language Institute http://dornsife.usc.edu/ali, which sponsors courses and workshops specifically for international graduate students.

For more information, see the class syllabus and the USC web site.

Product Problem

- Context
  - You will interact with a company that has an online shopping site which would like to start to push products to customers as they browse the site.
- Data
  - There is data about the on-line purchases of customers for the last 5 years.
  - There is profile data for some customers who are repeat customers: address, credit card, shipping preferences.
  - There is data about customers who recommended products to their friends in order to get a discount.
- Cost
  - When you push random products, 10% of customers do not like what is pushed to them and they leave the site.
- Challenges
  - When you have a sale, many repeat customers buy many more items than usual.
  - There is no profile data for many customers that pay through a third-party service.

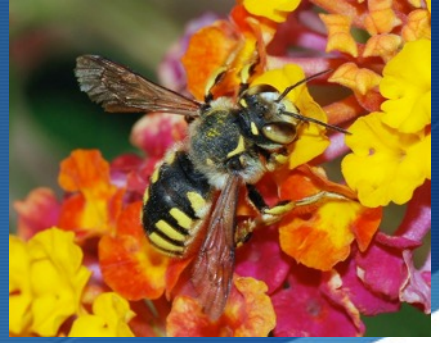
https://commons.wikimedia.org/wiki/File:Shopping_cart_icon.svg

Factory Maintenance

- Context
  - A factory has numerous complex machinery. It wants to build a system to predict when an equipment may fail to perform proactive maintenance and minimizing disruptions and costly repairs.
- Data
  - Real-time and historical data from sensors embedded in the machinery, e.g., temperature, pressure, vibration, acoustic emissions, motor current, and fluid levels.
  - Maintenance logs, e.g., date, type of maintenance (e.g., repair, replacement), the reason, and the parts involved.
  - Information about each equipment, e.g., make, model, manufacturing date, operational parameters, expected lifespan.
  - Temperature and humidity in the factory.
- Cost
  - An inaccurate prediction could lead to unnecessary maintenance, wasting resources.
- Challenges
  - The sensor data contain a large number of features. Sensor readings are noisy or may contain anomalies that are not indicative of actual failures.
  - Equipment failures are rare compared to normal operation. The dataset contains far fewer instances of failure compared to normal operation.
  - Different types of equipment may fail in different ways, and the patterns leading to failure might vary significantly.
  - Understanding why a model predicts a failure is crucial for maintenance teams to take appropriate action. A “black box” prediction might not be sufficient.

https://commons.wikimedia.org/wiki/File:A380_Assembly_line.jpg

Bee Problem

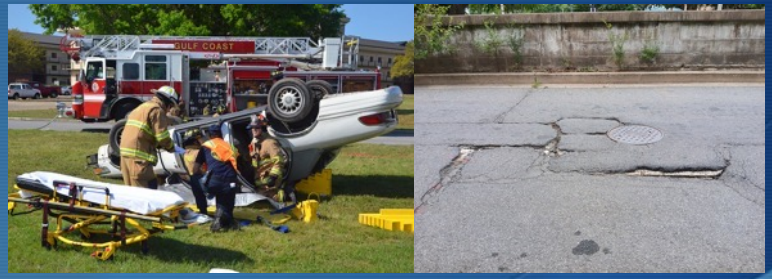


- Context
  - You will interact with the government of an island who would like to investigate how to reduce the bees so tourism can thrive again.
- Data
  - There is data available about the weekly water levels of all rivers and ponds for 30 years.
  - There are medical reports of bee bites and pollen allergies for the last 20 years.
  - There is climate data and population data, including rainfall and temperatures as well as pollen levels.
  - There is a lot of data about population, pollution, pesticide use, and bird populations (bee predators).

- Challenges
  - There are two bee experts in the island, but they do not know anything about data science. What questions would you ask of them to help you figure out how to solve the problem?
  - There is a company that can release pesticides on crops
    - Pesticides cost $1,000 per square kilometer
    - Pesticides reduce the bee population for 3 months
  - There is an environmental group that claims that the bee population can be reduced naturally by planting crops that have no flowers (e.g., corn, wheat, etc.)

https://en.wikipedia.org/wiki/Bee#/media/File:Anthidium_February_2008-1.jpg

City Resources



- Context
  - A city wants to understand the demand for resources across neighborhoods, e.g., emergency services (police, fire, ambulance), sanitation services, and infrastructure maintenance.
- Data
  - Emergency service call logs, e.g., location, time, incident type, response time, resolution.
  - Date, location, request type, and status of non-emergency service requests, e.g., potholes, broken streetlights.
  - Infrastructure, e.g., location and condition of roads, traffic lights, sewer lines, and their scheduled maintenance and past repairs.
  - Neighborhoods Census data, e.g., population density, age, income, housing types.
  - Weather information and a calendar of events, e.g., festivals, concerts, sporting events.
- Cost
  - Inefficient resource allocation can lead to increased response times for emergencies, delays in addressing citizen requests, and higher maintenance costs
- Challenges
  - Combining data from different sources with different formats and levels of granularity can be complex
  - Service demands vary across different geographic areas and at different times of the day, week, and year.
  - Some events (e.g., major crimes, large fires) might be relatively rare.
  - Avoid bias and ensure equitable service delivery across all neighborhoods.
  - The analysis needs to generate interpretable insights that can be translated into resource deployment strategies.

https://commons.wikimedia.org/wiki/File:Emergency_service_personnel_remove_one_of_the_simulated_casualties_from_a_motor_vehicle_accident.jpg
https://commons.wikimedia.org/wiki/File:Nasty_pothole_in_the_middle_of_Virginia_Avenue.jpg

Fraud Problem

- Context
  - You will interact with a bank, interested in detecting fraudulent activity in credit card customers.
- Data
  - For each customer, there is detailed information from their card application about their address, salary and employment, and demographic.
  - For each customer, there is a record of all their transactions (date, charges, and vendor) for the last 4 years.
  - For 1% of customers, there is a flag that their credit card was reissued because of fraudulent use of their prior card.
  - You can buy additional data, like census data for any zip code.

- Cost
  - When a fraud goes undetected, the average loss to the company is $3K
  - Reissuing a customer card costs $50
  - When card is reissued and there is no fraud, 0.5% of customers cancel their card.
- Challenges
  - Volume of the data: there are 100M customers, with 30K transactions on average
  - Some credit cards were reissued but no fraud took place once investigated

https://pixabay.com/en/road-sign-attention-note-scam-464641/

Fitness App Problem

- Context
  - A fitness tracking company wants to provide personalized recommendations for their users. The company has wearable devices and a companion mobile app
- Data
  - Wearable devices data in the last two years, e.g., steps taken, heart rate, sleep duration, exercise, GPS data for outdoor activities.
  - Information collected during sign-up, e.g., age, gender, height, weight, fitness goals, dietary preferences, stress levels.
  - App logs, e.g., workout plans viewed, recipes saved, and engagement with past recommendations from the app.
  - Publicly available data on local gyms, healthy food restaurants, and air quality indices.
- Cost
  - If the recommendations are irrelevant or poorly aligned with a user's goals and preferences, they may decreased app usage or unsubscribe from premium features
- Challenges
  - Many users might not consistently track all aspects of their activity or may have incomplete profile information.
  - User goals and preferences can change over time.
  - Recommendations need to consider the user's current state, such as their activity level that day, their available time, and even the weather.

https://commons.wikimedia.org/wiki/File:Fitness_Tracking_Smartwatch.jpg

Disaster Relief

- Context
  - You will interact with a non-profit organization in a remote country who would like to understand where to send relief and in what form.
- Data
  - There is a lot of microblog data (e.g., Twitter) that has been made available to you, where people are posting issues with bridges, roads, and general access to remote locations.
  - Many hospitals are emailing hourly reports, with number of beds occupied and available, medical inventory status, and medical personnel.
- Cost
  - A number of data analytics team have contacted your headquarters to volunteer their time to help with data analysis and any data collection needed.
- Challenges
  - The remote country's government seems open to take your advice for what roads need repair, what hospitals need more personnel, etc, but will ask for detailed justifications of all your recommendations.

http://disaster-accident.com/81-rescue/001-rescue.html