Duke Datathon 2019

Valassis, a leader in marketing technology and consumer engagement, is providing three datasets for Duke Datathon 2019, which contain various user and customer actions. Using information about customer interests and actions, Valassis hopes to understand what shoppers are most likely to engage with their marketing technology.

Your team's goal is to analyze and/or visualize any (or all) of the data in a creative and insightful way. Pretend you're a team of data scientists at Valassis, and you've been tasked with exploring a few datasets over twelve hours to create as much value as possible for the company. Section III details potential questions and prompts for you to explore, but you're free to formulate and pursue any questions or visualizations you think might be interesting! Feel free to ask a mentor for help, and attend a workshop to learn some data science and engineering skills and techniques.

The dataset can be found at dukeml.org/dataset, and the password to the file will be provided at 9:30am on Saturday, November 2, 2019 via email and Slack. Please download the dataset as soon as possible to prevent potential network issues on Duke's wifi!

Instructions for submission: Submissions will be due at 9pm on Saturday, November 2, 2019 at dukeml.org/submit, and projects will be judged on design, creativity, technicality, and presentation. The top five teams will be selected to

present and will be awarded prizes! In Section I, you'll find more details about the competition.

I. Datathon Schedule

Schedule: Saturday Nov 2 (The Edge)

9:00am Registration and Breakfast
9:30am Dataset Released!
10-11am Workshop: Exploratory Analysis in Python & R*
11-12pm Workshop:Intro to Keras/Tensorflow
12-1pm Workshop: Intro to Causal Inference (Advanced)*
12:30pm Lunch
2:00 pm Soccer & Frisbee
4-5pm Workshop: Appian Presents Smart Web Apps Using ML*
5:00 pm Painting
6:30pm Dinner
9:00pm Submission Deadline

* Workshops are a great way to learn a new statistical technique or data science tool! All workshops will be held in Old Chem 116.

Make sure to take advantage of mentors walking around and through Slack. You are encouraged to discuss your work with other teams, mentors, and students, and can use online and offline resources. However, a maximum of four individuals (the members of your team) should make large, meaningful contributions to your submission in fairness to all teams participating in the competition. **Teams must submit the following materials by the 9pm deadline.** It is recommended that teams work continuously from the beginning on deliverables rather than finish it all within the last hour. You should begin working on the deliverables at least three hours before the deadline.

Schedule: Sunday Nov 3 (Gross Hall)

9:00am Finalists announced
12:00pm Lunch (Fosters)
1:00pm Keynote by Dr. Fayyad
2:00pm Final Presentations
3:00pm Deliberations
3:30pm Winners announced

II. Components of Submission:

A. Written Report

Teams must write a report that describes the steps taken to answer their proposed question or prompt. There is no set format for how the report should be written, but example sections of the report can include, but are not limited to, the following:

- Introduction: What question are you answering with the data, and why is it important?
- Data Engineering Process: How did you clean and prepare the data, and what data did you use?
- Analysis: What analytical techniques did you use, and why?
- Findings: What did you discover (include visualizations)?
- Conclusion: What can a layperson at Valassis conclude from your team's work?

At minimum, the report must include the question being answered, findings and visualizations, and a conclusion. Your report may not be longer than 4 pages in length. You may have an optional appendix (maximum 2 pages), however, this may not be read by the reviewers. From the report, it should be clear as to how you approached your analysis. Do not include any identifying information about the members of your group in the submission. Please submit your written report as a .pdf document.

B. Programs

All programs written during the competition will need to be submitted. The programs can be messy, uncommented, in multiple files, etc., and will not be judged on their quality. You must submit your code as a .zip file. Do not include any identifying information about the members of your group in the submission.

C. Submission

In order for judges to properly evaluate a team's performance, it is required that teams submit both written report and the programs used for the analysis. Submit your work at dukeml.org/submit by the deadline.

Ensure that all materials are submitted by 9pm. *Unfortunately, in fairness to all teams participating in the competition, we cannot offer any extensions to the deadline.* A group of judges will review the submissions and select the top finalists. The top 5 finalists will be selected to present their work at 2pm the following day, and a panelist of professors at Duke will rank the finalists. Additionally, 3 teams will be selected for honorable mention. Honorable mention teams will not present. We'll

distribute over $2,000 in prizes, as well as some additional surprise awards provided by sponsors, amongst the top 5 teams, and wrap up the event!

D. [FINALISTS ONLY] Slide Deck Presentation

The top 5 teams will be required to submit and present a slide deck presentation (up to 5 slides). The goal of the presentation is to guide judges on how you utilized the available data to answer the question you came up with. The presentation should include meaningful visualizations, text, video, and/or other relevant multimedia content. It is up to your discretion as to what kind of material you would like to put in the presentation, but the analytical process, findings, and conclusion should be clear. In general, the content in the presentation should be a condensed version of the written report. Finalists will be given further instructions on Sunday morning, before they present at 2:00pm on Sunday.

III. Dataset Information

Refer to https://github.com/DukeUndergraduateML/datathon for information about the dataset!