ILLINI DATATHON 2021

FEBRUARY 20TH - 27TH

LOCATION: VIRTUAL

Sponsors:

synchrony

P&G

Sandia National Laboratories

CITADEL | CITADEL | Securities

OPEN TO ALL STUDENTS

DATATHON ILLINI

How to be successful this Datathon...

1. Join the Discord! Link: https://discord.gg/XAY9MtpB
2. Check your email and our website frequently :
   https://illinidata.wixsite.com/illinidatathon
3. Follow the Submission Directions!
4. Use all of your resources (Office hours, workshops, etc.)

SCHEDULE

MONDAY 02/22

8:30am - 11am → Synchrony office hours

10am - 12pm → P&G Workshop on Desegregation of Data + Q&A session

2:30pm - 3:30 pm → Sandia Validation Workshop

6pm - 7pm → Community event!

Tuesday 02/23

11am - 1pm → P&G Office Hours

1pm - 2pm → Synchrony workshop on Voice data

4:30pm - 7pm → Synchrony Office Hours

SCHEDULE

Wednesday 02/24

10am - 12pm → P&G Office hours

12pm - 2pm → Citadel Office Hours

6pm - 7pm → Community event!

Thursday 02/25

1pm - 3pm → Sandia Office Hours

SWAG PICKUP (information found in email)

SCHEDULE

Friday 02/26

1pm - 3pm → Sandia Office Hours

4pm - 6pm → Stats Club Office Hours (for submission help only)

6pm - 7pm → Community event!

11:59pm → SUBMIT YOUR WORK!!

SWAG PICKUP (information found in email)

Saturday 02/27

4pm - 5pm → Closing Ceremony

SWAG PICKUP (information found in email)

Deliverables

1. Code -- A submission guide can be found on our website, the github link will be assigned to your team (found in your email)
2. Video -- Up to 7 minute video must be uploaded, going over your analysis and findings

Nervous about submitting? Don't be!

Even if you are not confident about your model or code, submit your work! We will be grading based on innovative approach and presentations, not *only* model accuracy!

CRITERIA

All deliverables will be taken into consideration during judging. The following criteria will be on a 5-point scale and will be weighted equally.

ACCURACY

- Based on Classification Error Rate
- Takes into account possibility of skewed data

METHOD

- Model fit and techniques
- Creative, innovative approach

PRESENTATION

- Engaging presentation
- Clear, thoughtful explanations during Q&A
- Supportive visualizations
- Lessons learned

APPLICABILITY

- Solution addresses given business problem
- Proposed solution can be implemented in similar real life situations

A word from our sponsors:

A word from our sponsors:

A word from our sponsors:

synchrony

Problem set #1

Abstract: As virtual assistants become more widely used by consumers to simplify tasks and corporations to increase customer service efficiency, some individuals are getting left behind. Many voice assistants are designed to understand non-accented, dictionary language, which leaves out a significant portion of people. Leveraging publicly available data sets, deliver a model that can create equity in this technology for underserved populations.

Problem set #1

Notes:

Underserved populations in the US may include minorities, immigrants, English-Second-Language (ESL) households, and low income communities.

Other populations negatively impacted may include individuals with speech impediments, learning disabilities, or other.

Assume we are solving for virtual assistants in an English speaking environment.

Why: It is critical that we identify opportunities to improve technology by removing implicit bias and breaking down barriers for underserved populations.

Data Sets: There are troves of voice data (one site I found had over 2tb). I can share starting points with the students, but this leaves the door open for them to approach it from a way that they feel connected.

A word from our sponsors:

Problem set #2

This Data Challenge is for enthusiastic students to conduct pioneering work in the consumer IoT space using their analytical and data science skills. The participating students will have to develop machine learning algorithms for processing telemetric bath tissue consumption data in order to gain knowledge on consumer needs and habits. The raw data is neither classified nor labelled, but hides the natural and inherent groupings to be uncovered through feature engineering and unsupervised machine learning. The insights revealed will fuel the business strategy to better serve consumers.

Problem set #2

Problem to Solve: identify bathroom task types, e.g., #1 and #2 “go-events”, using telemetric bath tissue consumption data collected from smart devices

Data: sheet pulls by household, including timestamp, number of sheets, and household characteristics

Scope: the project may proceed in the following steps.

1.**Task Definition**: determine a time limit, e.g., 30 seconds, to combine sheet pulls into a task

2.**Feature Engineering**: once tasks are defined, task level features need to be created for clustering analysis, e.g., number of sheets used, number of sheet pulls, sheets per pull, max sheet per pull, time between first and last pulls, average time between pulls, time of day, etc.

3.**Unsupervised Machine Learning (ML)**: clustering algorithms can be used to discover patterns, identify and separate distinct task types

4.**Task Type Profiling**: use available data for profiling and showing the characteristics of each task type

- **Deliverables:**

1.Rules for feature engineering

2.ML algorithms

3.Profiling results

Rules

1. Submission by 11:59 pm on FRIDAY 02/26
2. All submissions must have all deliverables or they won't be counted (code + video)
3. We are looking for clarity and ability to answer both business and data science questions not for flashiness of submissions. Work on what counts.
4. Teams can only leverage any publicly available data source for generating input features.
5. Teams can leverage any tools and/or libraries for this task but must provide the links for those tools in their code deliverable.
6. Teams are not allowed to ask anyone outside of mentors or their own team for help. Please do your own work because this is a learning opportunity first and foremost.
7. Be courteous to your peers as you work.
8. Have fun and get to know your fellow competitors and mentors.

R Resources

Useful R Libraries

- Data loading – DBI, odbc, RMySQL, xlsx, haven
- Data manipulation – dplyr, tidyr, stringr
- Visualizations – ggplot2, rgl, htmlwidgets
- Modeling – car, mgcv, lme4/nlme, randomForest, vc, glmnet, caret
- Spatial data – sp, maptools, maps, ggmap, leaflet, R-ArcGIS bridge
- https://support.rstudio.com/hc/en-us/articles/201057987-Quick-list-of-useful-R-packages

R Resources

- https://www.r-project.org/
- https://rstudio.com/
- https://www.codecademy.com/learn/learn-r
- https://www.coursera.org/learn/r-programming
- https://www.datacamp.com/
- https://www.lynda.com/R-training-tutorials/1570-0.html
- https://www.reddit.com/r/Rlanguage/
- https://stackoverflow.com/questions/tagged/r

Python Resources

Useful Python Libraries

- https://pypi.org/project/googlemaps/
- https://pypi.org/project/geopy/
- https://pro.arcgis.com/en/pro-app/arcpy/get-started/what-is-arcpy-.htm
- https://geopandas.org/
- https://gdal.org/
- https://numpy.org/
- https://pandas.pydata.org/
- https://matplotlib.org/
- https://scikit-learn.org/stable/
- https://www.reportlab.com/

Python Resources

- https://www.learnpython.org/
- https://www.codecademy.com/learn/learn-python-3
- https://www.python.org/about/gettingstarted/
- https://docs.python-guide.org/intro/learning/

THANK YOU AND GOOD LUCK!!