MACHINE LEARNING & PREDICTIVE ANALYTICS

sbharadwaj@uchicago.edu | anishgera@uchicago.edu | stevieb@uchicago.edu

Objective

The course project is designed to apply your learnings relating to machine learning and deep learning skills by sourcing data in raw format, engineering features and preparing the data for model development, evaluation, and deployment. The focus is to make sure that the solution is developed referencing the CRISP-DM or TDSP) while making recommendations to the client/management based on findings using visualizations.

PROJECT

The goal behind the project is to 'put it all together' by developing a coherent, concise, and realistic analysis in the form of a presentation to the client/management. The project will provide you with the opportunity to apply your knowledge and understanding of creating an automated end to end ML pipeline that facilitates the data collection, feature engineering, model development and visualization from identifying a dataset/problem statement to providing recommendations to your client.

The project should contain the following and be written for the intended executive audience in mind:

- Executive Summary
- Problem Statement/Research objective(s)
- Exploratory Data Analysis
  - Data analysis, visualizations, data mining techniques
- Data Preparation/Feature Engineering
  - Handling of features, assumptions, and tests
- Methodology and various tools used in the process
  - Evaluation of analytical or transactional data stores for the use cases
  - At-least 4 ML/DL Models implemented and evaluated
- Lessons Learned
  - scope for improvement
  - Assumptions
  - Feature Engineering
  - Data Preparation
  - Model Metrics
- Recommendations
  - Next Steps
  - Methods that can be used
  - Datasets that can add value to the existing analysis

- References
    - Literature review, URLs

PROJECT TIMELINES

- Week 2: Form project teams, research and socialize project ideas
- Week 4: Define scope and finalize project data sources and datasets
- Week 6: Feature Engineering, Exploratory Data Analysis
- Week 8: Design and Develop Models, evaluate key metrics
- Week 10: Presentation (findings, recommedations, learnings) and submit artifacts

DATA

Students have the flexibility to can use any public dataset. The following URLs can also be used to refer for additional datasets

- Enron emails dataset (https://www.cs.cmu.edu/~./enron/)
- https://pushshift.io/kavanaugh-twitter-dataset/
- https://toolbox.google.com/datasetsearch/
- https://data.cityofchicago.org/
- https://opendata.cityofnewyork.us/
- https://data.gov.in/catalogs/
- https://github.com/awesomedata/awesome-public-datasets/
- https://www.springboard.com/blog/free-public-data-sets-data-science-project/
- IRI Dataset
    - NDA and data dictionary available at Modules > final Project > datasets > IRI
    - NDA signed by every member of the team and sent to gguevara@uchicago.edu
    - Once permission granted, login to midway. Data located at /project/databases/IRIData

SUBMISSIONS

- Students will work in teams of 2 to 4 people.
- Single submission per team.
- Following artifacts to be submitted as a single submission per team in canvas:
    - GitHub Repository comprising of dataset, scripts, analysis, visualizations
    - All scripts file(Python/R) containing all analysis, model build & deploy
    - Visualization Dashboards/Reports – Notebook, Tableau/ PowerBI, etc.
    - Final Presentation slides (as PPT)

GRADING RUBRIC

The final project accounts for 40% of your overall grade, and project grade will be determined on:

- Executive Summary & Problem Statement - 10%
- Data Analysis, Feature Engineering & Data Preparation - 25%
- Model selection, design, and evaluation – 25 %
- Proposed Solution and reasoning for the choosing the model/s for deployment – 20%
- Presentation along with findings and insights - 20%