ANALYTICS SYSTEMS ENGINEERING

shreenidhi.bharadwaj@northwestern.edu | ChristopherFiore2015@u.northwestern.edu

Objective

- End to end process of gathering, preparing data for ML Modeling.
- Collaborating with the team on business use case, data preparation and analysis
- Connecting to databases, analyzing the data and deriving valuable insights using Spark
- Build, train machine learning models and deploy them into production environment

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: Get started on the data preparation process & start on Machine Learning
- Week 8: Iterate on data loads, automated data pipelines, models & insights
- Week 10: Finalize, deploy models and upload artifacts for grading

Project

The goal behind the final project is to 'put it all together' by developing a coherent, concise, and realistic analysis in the form of a report and presentation to an executive audience (your client). The project will provide you with the opportunity to apply your knowledge and understanding of data collection, data preparation, storage in a relational or a non-relational database, analysis, modeling and visualization, by identifying datasets, analyzing it, and providing recommendations to your client.

The project report should contain the following sections and be written for the intended executive audience:

- Executive summary
- Research objective(s)
    - The problem to be solved and datasets you plan on using
- Data Ingestion, analysis and preparation
    - Gathering data and preparing it for storage and Analysis
- Methodology and various tools used in the process
    - Evaluation of analytical or transactional data stores for the use cases
    - Automation methodology for the End to End pipeline.
    - ETL/Streaming: Scripts, tools relating to data ingestion and transformations

    - Applicability of a relational database, graph database or Elasticsearch for the dataset and the corresponding data model
  - Deploying ML models to Production
    - Usage of containers and container orchestration
    - Publishing of a ML/DL model to production
  - Reporting
    - Delivering on insights via reports & dashboards.
  - Recommendations
    - Design considerations, model evaluations and platform choices
    - Corrective measures and scope for improvement
  - Lessons Learned
  - References

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/

Submissions

- Students will work in teams of 3 to 4 people.
- Single submission per team.
- Following artifacts to be submitted as a single submission per team in canvas:
  - All scripts file(SQL/Python/R) containing all analysis/modeling/queries
  - Visualization Dashboards/Reports – Tableau, Excel or PowerBI, etc. (raw files)
  - Final Presentation slides (as PPT)

Grading Rubric

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

- Business Use Case - 10%
  - Understanding the business problem and articulating projects goals
- Data Ingestion, Analysis & Preparation - 25%
  - Data Ingestion, cleaning, transforming to the target structure
- Data Modeling & End to End Platform Design – 25 %
- Tools / deployment – 20 %
- Evaluation of ML models & Insights - 20%

References

Below are some references to build the end to end pipeline

- https://aws.amazon.com/getting-started/tutorials/get-started-dlami/
- https://medium.com/merapar/pure-serverless-machine-learning-inference-with-aws-lambda-and-layers-979702d9ae49
- https://hackernoon.com/keras-with-gpu-on-amazon-ec2-a-step-by-step-instruction-4f90364e49ac
- https://aws.amazon.com/getting-started/tutorials/build-train-deploy-machine-learning-model-sagemaker/
- https://machine-learning-company.nl/deploy-machine-learning-model-rest-api-using-aws/
- https://towardsdatascience.com/ml-models-prototype-to-production-6bfe47973123