SUMEDH SHAH

(412) 954-8546 **sumedhns@andrew.cmu.edu** **https://www.linkedin.com/in/sumedhs/**

EDUCATION

Carnegie Mellon University, Pittsburgh

Master of Information Systems Management - Business Intelligence and Data Analytics **December 2022**
GPA- 3.96/4.00
Teaching Assistant- Unstructured Data Analytics, NoSQL Database Management

Maharashtra Institute of Technology, India

Bachelor of Engineering in Electronics and Telecommunication Engineering **June 2019**
GPA- 7.79/10

SKILLS

- ***Languages***: Java, Python (w/ framework – Pyspark), SQL, Bash (Linux), Scala
- ***Database and Big Data***: SQL (Oracle, MySQL, Postgres), Hadoop (w/ Hive and Apache Spark), ETL Modeling, Apache Solr, MapReduce, PyTorch, Tableau, Spacy, Redis, MongoDB, and Neo4j database
- ***Cloud***: AWS Certified Solutions Architect – Associate, Microsoft Azure, Azure Data Factory, Databricks

WORK EXPERIENCE

PPG Industries, Pittsburgh

AI/ML Intern **May 2022- August 2022**

- **[azure pipelines]** Designed and managed data flows and pipelines in Azure Data Factory for ingestion from relational databases, REST APIs, Delta Lake, Sharepoint and other cloud file storage systems
- **[automation]** Automated the manual process of ingesting Design of Experiment data from workbooks in Sharepoint using PySpark scripts written in Databricks, thereby reducing ingestion and curation time by 90%

Modak Analytics LLP, India

Data Engineer **September 2019- July 2021**

- **[data ingestion**] Developed PySpark code that minimized ingestion time by 60% to consume terabytes of data from structured and unstructured sources daily using automated bot workflows into the client data lake as Hive tables, leading to convenient downstream analysis and data transformation (using SQL) applications
- **[data curation]** Built complex Spark SQL queries to curate pharmaceutical data on Dataframe collection, these were utilized in data analysis and modelling for accelerating the timeline of the drug discovery process by 50%
- **[pipeline automation and visualization]** Conducted the compilation of Python workflows using Subprocess and OS modules which executed Scala scripts to index curated data into Apache Solr and Neo4j graph database for utilization by pharmaceutical SMEs; automated workflows would be triggered daily to bring in updated data
- **[communication]** Collaborated with clients and took the initiative to plan strategies to enhance the company's native platform and helped develop creative business requirements for a clinical trials project
- **[leadership]** Spearheaded a team of 5 members who collectively created an efficient data pipeline in PySpark for a client use case that involved crawling, ingestion and transformation of terabytes of clinical trials data, which resulted in the reduction of original processing time by 50%

ACADEMIC PROJECTS

Predicting Order Returns (https://github.com/Sumedh1197/Predicting-Order-Returns) **March 2022- May 2022**

- Predicting whether an online order will be returned to an online retailer as a step to increase net margins and improve supply chain & inventory management; according to research this is soon expected to be a trillion-dollar problem
- Initially performing data preparation and feature engineering for downstream modelling and finally predicting the orders which would be returned using Machine Learning classification algorithms such as Decision Tree, Random Forest, Logistic Regression, Naive Bayes, and Perceptron
- Improvement of performance metrics mainly Recall since False Negatives, in this case, are a bigger concern for a retailer

[Python, Jupyter Notebook, Scikit-Learn, Numpy, Plotly, Seaborn]

EDA on H&M Transactions Data (https://github.com/Sumedh1197/EDA_Project) **January 2022- March 2022**

- Manufactured call-to-action (CTA's) for H&M regarding production of category, colour, and pricing using EDA. Associating the best channel of sales for the above actionable items. Presented the output in a time series format with CTAs for each quarter
- Sample CTA Quarter 1 - Jan-March - H&M should focus on selling lower-priced products in the Ladieswear category in Black, Blue and Pink colour and sell them through online channels

[Python, Jupyter Notebook, Altair, Pandas, Scipy, Numpy]

Other Projects: Travel Planner GUI, Recommendation System (https://github.com/Sumedh1197)

COMMUNITY SERVICE AND LEADERSHIP EXPERIENCE

- **Live Life Love Life Charity Foundation- Fundraising Coordinator** **October 2017- October 2019**
  Organized events to spread cancer awareness and raised approximately $40,000 from donations to provide medical assistance for the underprivileged to fight breast cancer
- **Thermax Foundation** **June 2017- August 2017**
  Focused on educating underprivileged children in mathematics and science for entrance exams