

Arth Talati

Philadelphia, PA | P: 267206-8721 | atalati@seas.upenn.edu | arthtalati.github.io

EDUCATION

University of Pennsylvania Master of Science in Data Science

Philadelphia, PA
May 2021

- Cumulative GPA: 3.78/4.0
- Best Project Design Award for Football Freak project – Databases and Info Sci Course
- Relevant Coursework: Big Data Analytics, Machine Learning, Internet & Web Systems, Computational Linguistics, Databases, Statistics for Data Science, Data Science for Public Policy

SKILLS

Programming Languages: Python, Java, SQL, NoSQL, R

Big Data & Machine Learning: Spark, Pandas, NumPy, Hadoop, TensorFlow, Pytorch, PostgreSQL, Mongo DB, Neo4j, MATLAB

Data Science Technologies: ETL, Database Management, Data science pipeline (mining, cleansing, wrangling, visualization, modeling, interpretation), Statistics, Experimental design, Hypothesis testing, Git, A/B testing, Airflow, Jira

EXPERIENCE

School of Engineering & Applied Sciences

Philadelphia, PA

Teaching Assistant

Aug 2020 – Dec 2020

- Collaborated with instructor to design HW assignments, answer 300+ student questions and mentored 8 project groups.
- Hosted HW datasets in S3 buckets and developed a guide to spin up an EMR cluster on AWS for students to work with Spark ML.

CSPACE

Boston, MA

Data Science Intern

May 2020 – Aug 2020

- Derived a quantitative approach to cluster companies based on levels of customer centricity as measured with CSPACE survey features.
- Performed feedback sentiment analysis and K-means clustering to segment companies (300+) into cohorts w.r.t their public opinion.
- Modelled correlation of financial performance of companies with their perceived image in different consumer segments.
- Improved the built inhouse analytics capabilities to enhance consulting framework.

Penn Data Science Group

Philadelphia, PA

Research Analyst

Jan 2020 – May 2020

- Assisted the Procurement Department to identify high-risk transactions as part of Purchasing Services High Risk Project.
- Designed anomaly detection algo using Local Outlier Factor and Isolation Forests algos for dataset of 1.8 million Oracle Financial invoices.

Penn Medicine

Philadelphia, PA

Research Intern (Anaesthesiology and Critical Care Department)

Nov 2019 – Jan 2020

- Analyzed EEG signal data from anesthetized subjects, to know more about their responsivity and side-effects of drugs.
- Extracted quantifiable information from raw data using MATLAB after doing preliminary signal processing and spectrum analysis.

SELECT PROJECTS

Football Freak

- Created a soccer app using data scraped from [sofifa.com](https://www.sofifa.com) hosted on AWS Lambda serverless instance. Designed a relational model for the dataset in 3NF format deployed on AWS RDS instance with appropriate indexes and an optimal query plan to decrease the query execution time by 10x.

SEPTA On-Time Performance Analysis

- Customized TWINT API to crawl tweets from @SEPTA Twitter handle and validated the claim by performing regression between actual delays and the ones claimed in tweets. Scrapped the weather data and streamlined Spark pipeline to train a random forest regressor model to predict the delays in arrival time.

Cook-Assist

- Scraped BBC-GoodFoods website for 13000+ food recipes using Chromium browser and BeautifulSoup package. Developed a R Shiny app with Neo4j DB to query the graph database to assist a cook in exploring new recipes and optimizing user time and effort spent while cooking.

Deep Learning for Authorship Identification

- Executed multi-class classification for 50 news article authors using LSTM, Bi-LSTM, GRU (RNNs) neural networks, improved the classification accuracy by 20% over the baseline LSTM model using SVM and 40 different Stylometry features.

US Flight Delay Analysis

- Analysed the dataset of 5.8 Million flights delays and cancellations by US DOT to find the effects of delays on domestic flight operation.
- Developed an interactive dashboard (R Shiny) to demonstrate the key functionality of comparing different airlines departure/ arrival timings for given airport based on the day-of-week, time-of-day, taxi times and other features.

FPGA based Machine Learning for Cardiac Diseases Detection | Senior year Thesis-Project

- Designed an ASIC (Application Specific IC) prototype to be embedded on handheld devices to detect heart diseases by ECG machine learning algorithm (SVM and Naive Bayes).
- Modelled using extracted features to give disease predictions with 93% accuracy of a person having Premature Ventricular Contraction.