ARSHIYA ANSARI

DATA-DRIVEN · CODE-BASED

☎ (571)-236-6177 | ✉ arshiyaxansari@gmail.com | ⌂ arshiyaansari.github.io | arshiyaansari | arshiya-ansari

EDUCATION

University of Virginia
Charlottesville, VA

Master's in Data Science — *Jun 2020 - May 2021*

Cumulative MS GPA: 3.85

University of Virginia
Charlottesville, VA

Bachelor's in Computer Science, Bachelor's in Statistics — *Aug 2016 - May 2020*

Cumulative BA GPA: 3.83

EXPERIENCE

Apple
Sunnyvale, CA

Machine Learning Engineer — *Jul 2021 - NOW*

- Owner of the ground truth data set generation pipeline (using PySpark, Kubeflow, and PrestoSQL) that provides versioned and validated data sets to the entire autonomous stack for machine learning model training and evaluation
- Collaborated on a daily basis with stakeholders and consumers on new data and ETL features, attributes, schemas, and data set designs
- Designed and developed infrastructure for efficient data set creation with new extensive validation techniques to aid data quality initiatives. Reduced production defects by 80%, decreased time for overall data set generation by 4x.
- Led the effort to integrate a new data lake API into existing pipelines to increase accessibility to data from every point of the ETL process. Introduction of the data lake increased data lineage tracking by 100% and reduced data quality issues by 50%.
- Created executive dashboards hosted in Tableau to monitor volume and quality of data starting at the point of data collection from external and internal vendors to any point of data processing in the ETL pipeline
- Working with Flask and PostgreSQL to host a data corrections API for stakeholders to self-report data quality issues

Apple
Sunnyvale, CA

AI/ML Intern: Autonomous Systems — *Jun 2020 - Sept 2020*

- Introduced new functionalities to improve performance to existing ETL processes and SQL queries for ground truth data set generation
- Migrated data set generation from a heavy, deeply-nested data schema to a light-weight alternative
- Developed infrastructure for scalable and maintainable data set generation using Presto SQL, Apache Spark, and Kubeflow
- Added new features to the ground truth data sets for more robust and efficient machine learning model training and evaluation

Capital One
McLean, VA

Software Engineering Intern — *Jun 2019 - Aug 2019*

- Implemented a live time-series forecasting algorithm utilizing deep neural networks for Capital One's internal data ETL engine
- Constructed an anomaly detection program for the performance metrics of the internal ETL engine via an automated Python script
- Established an anomaly alert functionality to indicate real-time performance issues to the ETL engine software development team

CGI Federal
Fairfax, VA

Technical Consultant Intern — *Jun 2018 - Aug 2018*

- Corresponded with clients at the Center for Medicare and Medicaid Services for data analysis requests on healthcare insurance information
- Identified and reconciled 20 different bugs in the database by querying and analyzing data with Oracle SQL Developer
- Devised 5 new report templates in Cognos using Oracle SQL Developer to aid in easier delivery of data to clients

PROJECTS

Capstone Research: Scale-Invariant Factors for Brain-Computer Interfaces
Charlottesville, VA

MSDS Candidate/Researcher — *Aug 2020 - May 2021*

- Developed a specialized LSTM network that models after the scale-invariant patterns of the brain's time cells and quantify how well it can identify actions from EEG data. The goal is to improve the prediction accuracy of current BCI (Brain-Computer Interface) technology on EEG data by implementing neural features such as time cells as artificial agents.

COURSES

Computer Science	Software Development, Machine Learning, Databases, Cloud Computing, Artificial Intelligence, Big Data, Algorithms
Statistics	Data Science, A/B Testing, Regressional Data Analysis, Probability, Linear Algebra, Bayesian Statistics, Data Visualization

SKILLS

Programming	Python, Java, Presto SQL, PostgreSQL, Apache Spark, Kubeflow, AWS, R, SAS, C++, Assembly, Kubernetes
Analytic Tools	Tableau, Mathematica, MatLab, Cognos, Microsoft Excel, LaTex
Languages	English, Hindi, Urdu, Telugu
General	Photography, public speaking, mentoring, writing, film production
Activities	Women in Computing Sciences, Society of Women Engineers, Hack Cville (Program Coordinator), Madison Volunteer House