Akshay Kulkarni

akshayavinashkulkarni@gmail.com
857-869-0944 | Boston, MA

Github: github.com/Akshay-A-Kulkarni
Website: akshaykulkarni.netlify.com
LinkedIn: linkedin.com/in/-akshaykulkarni

EDUCATION

Northeastern University **Sep 2018 - May 2020**
Master of Science, Data Science Boston, MA.
Khoury College of Computer and Information Sciences - GPA 3.715/4.0
Relevant Courses: Algorithms, DBMS, Information Retrieval, Large Scale Parallel Data Processing, Supervised & Unsupervised ML

Birla Institute of Technology & Science, Pilani – Dubai **Aug 2012 – Aug 2016**
B.E. [with honors] in Electronics and Communications Engineering Dubai, U.A.E.

EXPERIENCE

– **Northeastern: Khoury College of Computer Science** **Jan 2020 – Present**
Research Assistant – Vitek Lab [MSstatsQC-ML] Boston, MA.

- Collaborating with an international group of researchers to leverage statistical sampling, factorial design and simulation in conjunction with Tree-based & Neural network models to enhance individual, longitudinal/time-series & complex mass spectrometry proteomics experiments.
- Developing tools using statistical process control & ML methods to detect anomalies or classify sub-optimal runs of instruments/sensors.
- Implementing efficient parallelized models (Ensembles, XGBoost etc.) for MSstatsQC-ML toolkit to be included in the open-source distribution of the framework in the form of an R/Bioconductor package.
- Building a web application to enable users to upload, test and visualize their data to facilitate optimal analysis and maintenance of their workflow.

– **Northeastern: Khoury College of Computer Science** **Jan 2019 – Dec 2019**
Teaching Assistant – Database Design + Data Collection, Integration and Analysis Boston, MA.

- Designed coursework and assignments. Evaluated submissions and managed teams of students in designing and implementing of projects.
- Guided & tutored students in class or during office hours to help them reinforce learning concepts such as tidying, storing, analyzing data and employing Data Mining & Machine Learning techniques in R as well as designing databases in MySQL

– **Predikly** **Feb 2017 - May 2018**
Junior Data Analyst Pune, India.

- Performed data cleaning, visualizations & contributed in development of dashboards & predictive analytics solutions with client-sourced data.
- Worked as part of the sentiment analysis team, assisting in mining and tokenization for large corpora of texts ranging from customer support incidents, product reviews, social media feeds, news articles, documents, and other sources.

– **Zio Technologies** **Aug 2015 - Feb 2016**
Data/Sales Insights Intern Dubai, U.A.E.

- Worked with the analytics team to analyze the sales of various AV components (Security, Digital signage) in different parts of the Gulf region.
- Directly worked with the managers to come up with strategies and take decisions to help maximize profit from the analysis.

SKILLS

Languages: Python, R, Java, SQL, Scala, Bash, HTML, CSS

Frameworks & Platforms: Flask, Django, Spark, MapReduce, Apache Beam, Lucene, Docker, AWS, GCP

VCS & CI/CD: Git, DVC, GitHub, GitLab, GitHub-Actions,

Libraries & Tools: NumPy, Pandas, Sklearn, H2O, TensorFlow, Keras, PyTorch, OpenCV, spaCy, RShiny, Tableau, PowerBI, MySQL

PROJECTS

Topic Modelling and Clustering on Complex COVID-19 Twitter Networks- [Python, Google BigQuery, BEAM, spaCy, GenSim]

- Implemented a streaming data ingestion ETL pipeline with Apache Beam to filter, process & store tweets from Twitter API to a BigQuery table.
- Generated network-graphs from streamed tweets to employ modularity-based graph clustering as well as other unsupervised clustering algorithms.
- Created an NLP pipeline with pretrained language models to tokenize, lemmatize and do specialized handling associated with social media texts.
- Employed Latent Dirichlet Allocation and other topic modelling algorithms to extract and visualize various topics surrounding COVID-19.

Distributed Matrix Factorization for Recommender Systems - [Scala, Apache Spark, AWS - EC2/Elastic MapReduce]

- Implemented a scalable parallelized version of the ALS algorithm to process user-item interaction data containing ~100 million ratings.
- Used the resultant user & item latent factored matrices to optimize large scale dimensionality reduction and computation of unseen interactions
- Deployed the algorithm on varying sizes of clusters on AWS. Achieved near linear speedup and scaleup while processing and generating over 8+ billion records to predict missing ratings for a user-item pairs.

Classification of Radiographs using Convolutional Neural Networks - [Python, FastAI, Keras, Google Compute Engine & Cloud Storage]

- Built CNN architectures to classify high resolution radiographs from the NIH Chest X-ray Dataset containing 112,120 patient records into one of the 10 possible conditions (Cardiomegaly, Effusion, Pneumothorax etc.)
- Reduced computational overhead by training on pre-trained CNN architectures such as ResNet50 & InceptionV3 in Pytorch-FastAI & Keras.
- Employed techniques such as one-cycle-policy and cyclic momentum to tune the network optimally and allow for significantly faster convergence while training models in deep learning cloud VM instances. Achieved ~80% validation accuracy on binary classification across all labels.

Decision Support/Multi-Objective Optimization & Visualization Tool - [Python, DASH, SQL, Heroku]

- Deployed a multi-database tool with React front-end and back-end developed in Python to allow users to perform decision support analysis in presence of trade-offs between two or more conflicting objectives. Used in Finance, Process-Optimization & other BI applications.
- Implemented functionality to display an interactive data table as well as calculate & highlight the Non-Dominated/Best set of solutions to users.
- Incorporated visualization methods to plot an interactive Pareto Curve for the selected features or objectives for intuitive analysis.