ARKA CHOWDHURY

• (716)-986-2817 • arka.chowdhury31@gmail.com • https://www.linkedin.com/in/arka-chowdhury • 1021 Cheryl Dr, Iselin, NJ, 08830

Data Scientist with 7 years of experience in developing, designing, and deploying Artificial Intelligence Applications. Highly skilled in Generative AI, NLP, Auto-ML, Forecasting, Statistical and Optimization software implementation. Proven ability to lead cross-functional teams and deliver robust AI solutions.

EXPERIENCE

- **Senior Data Scientist**　　　　　**GEP Worldwide, Clark, NJ**　　　　　**03/2021 - Current**
Python, Spacy, Open AI, Llama, Azure Blob, PySpark, Docker, Kubernetes, VectorDB, Databricks, Mlflow
Generative AI & NLP
  - Led the creation of an advanced chatbot for Supply Chain Analytics that leverages large language models (LLMs) to transform user-input text into SQL queries, facilitating data-driven decision-making. This solution significantly streamlined the query process, reducing time-to-insight by 30%.
  - Built a RAG Agent to detect user intent and retrieve data from a vector database (Qdrant DB) using chunking techniques, significantly enhancing data retrieval efficiency and accuracy for various P2P services.
  - Engineered an Anonymization API that detects and masks personally identifiable information in contract documents using Spacy and Faker for minimal processing time before LLM processing. This API acts as an orchestration layer for GenAI model selection and parameter tuning, enhancing data security and compliance with regulations.
  - Spearheaded the development of an automated system for generating real-time summaries of supplier profiles, directly integrated into the UI. This system used LLMs to summarize and rate suppliers based on responses in questionnaire forms, effectively aiding in risk assessment and quick decision-making for procurement teams.
  - Designed a robust Invoice Extraction Engine utilizing advanced OCR technologies (Azure OCR, Form Recognizer) to process PDFs in 12 languages with 92% accuracy. Developed additional features such as logo detection, address validation, amount conversion, and rubber banding, ensuring precise selection and validation of invoice components.

  Auto-ML Platform
  - Led the development and deployment of Minerva Studio, a cloud-agnostic Auto-ML Platform with MLflow and Fast API, integrating ML algorithms for classification and implementing automatic feature selection and anomaly detection.
  - Configured Docker and Kubernetes to optimize GPU usage and ensure smooth application deployment.
  - Coordinated with DevOps to deploy and test pipelines in multiple environments, achieving over 90% code coverage in SonarCloud. Trained and mentored team members on key concepts like MLOps, CI/CD pipelines, and API testing, ensuring efficient task completion.

  Demand Planning and Forecasting Analysis
  - Developed and deployed a user-friendly web portal for demand planning, integrating forecasting models and providing real-time analytics to clients.
  - Analyzed part booking demand history and applied statistical segmentation(ABC Analysis) techniques to recommend stocking policies, significantly reducing lead times and improving inventory management.
  - Developed an ensembling model consisting of time series (ARIMA, SARIMA, Prophet) and Machine Learning (Random Forest, XGBoost, LSTM) algorithms to forecast short-term and long-term demand.
  - Converted Python code to PySpark in Databricks for parallelization, creating a pipeline that reads data from ADL and writes forecasts and validation metrics back to ADL.
  - Supervised IT, Engineering and consulting teams, to ensure smooth project delivery and client satisfaction.

- **Lead Data Scientist**　　　　　**RSG Media Systems, New York City, NY**　　　　　**06/2020 - 02/2021**
Python, Apache Spark, PySpark, AWS S3, MySQL, Databricks, Mlflow
  - Designed a Machine Learning Model Pipeline for efficient and distributed Big Data ingestion and feature creation in Amazon S3 and Databricks (Apache Spark) which is used for ensembling models and to predict future audience.
  - Retrieved data from Data Lakes in AWS S3 storage; joined, validated, cleaned and structured the data as an input to the ML model. The insights from the data equipped the analysts with the trends and features of audience flow.
  - Developed the Pipeline of an ensembling model that uses multiple Machine Learning algorithms for prediction modeling and automatically selects the best learning algorithm to predict the future. This model ensures the selection of the best learning algorithm which has improved the prediction accuracy by 7% from the previous XG Boost model.
  - Implemented a cluster management system to optimize computation in Spark by managing the cluster type, size and workers. This system reduces the execution time by 30% and saves the cost of computation.

- **Data Scientist**　　　　　**RSG Media Systems, New York City, NY**　　　　**06/2018 - 06/2020**
  Python, CPLEX, FICO Xpress, Apache Spark, PySpark, AWS S3, MySQL, Spacy, R
  - Developed an XGBoost model on Apache Spark using PySpark to predict the audience viewing a network at any time. The accuracy of the predictions improved the program and break schedule with better delivery of promos and advertisements, saving 12% of the advertisement inventory.
  - Created and developed an optimization model using FICO Xpress solver to improve the existing break logs for the networks by shuffling the breaks. The objective was to deliver impressions for each break as per the advertisement deals (CPM) preventing any under or over delivery saving 1.2 million dollars equivalent of break inventory each year.
  - Developed a Lasso and Elastic-Net Regularized Generalized Linear Models using GLMNET in R to forecast the audience viewing a network based on the demography, location and frequency of viewing. This model was used to derive the relationship between reach and impressions considering the factors affecting the viewers.
  - Developed an NLP model to extract and organize media rights terms from contracts. This removed the need for analysts to enter terms manually, making the process fast and error-free.

- **Data Scientist Intern**　　　　**RSG Media Systems, New York City, NY**　　　　**05/2017 - 05/2018**
  - Designed and developed a scheduling model in Python to get the optimal promo schedule using CPLEX solver. It equipped campaign managers to schedule promo plans on multiple platforms (linear, digital) till the premiere date.

EDUCATION

Master of Science in Industrial Engineering - Operations Research Specialization, May 2018

University at Buffalo, The State University of New York, NY

Bachelor of Technology in Production Engineering, May 2015

Veer Surendra Sai University of Technology, Burla, India

TECHNICAL SKILLS

- **Programming Languages & Frameworks:** Python, R, Scala, PySpark, Fast API, Docker, Kubernetes, Rest API
- **Database Tools & Languages:** MySQL, PL/SQL, DB2, AWS S3, Azure Blob, Azure Pipeline, VectorDB, Elasticsearch, MongoDB
- **Tools and IDE:** Jupyter, Apache Spark, Minitab 17, Tableau, Power BI, MATLAB, GitHub, Visual Studio, PyCharm, Databricks, SonarCloud, TestRail, New Relic, JIRA, Confluence, Sreamlit
- **Mathematical Optimization Solvers:** CPLEX, Gurobi, FICO Xpress
- **Generative AI Models & AI Technology**: GPT-4o, GPT-4, GPT-3.5, Llama 3, Gemini, Vertex AI, Azure ML, Mlflow, Azure OCR, Form Recognizer, BERT, TensorFlow, PyTorch, Scikit-learn, Keras

PROJECTS

Enhancement of the University at Buffalo Bus (STAMPEDE) Schedule (2018)
Cleaned and transformed the bus GPS data to execute exploratory analysis in Tableau and list the issues in the current schedule. Collaborated with my advisor to create a new schedule that decreased the average waiting time by 12 minutes, lowered the peak hour load by 200 people and enhanced the routes reducing average travel duration by 8 minutes.

Traveling Salesman routing problem for UPS using nearest neighbor, MILP & Simulated Annealing (2017)
Plotted the depots and customers on a map using Folium and MapQuest in Python. Coded the algorithms and used Gurobi as a solver for MILP. Exhibited the optimum routes and costs for the algorithms in the map.