Shikhar Johri

(646) 749-6893 | sj3327@columbia.edu | Linkedin: @shikhar-johri | GitHub: @johrilab

WORK EXPERIENCE

Barclays New York, US

Data Scientist | Python, SQL, AWS, PySpark, SAS Studio, LLM *Jun 2024 - Present*

- Developed a metadata tool and Rule Engine, automating validation for audit compliance, reducing errors, and improving reporting.
- Streamlined Monthly Business Review Analysis, reducing report processing time by a week for Consumer Bank Analysts.

Tata Consultancy Services Karnataka, IN

Data Scientist | Python, Computer Vision, Pytorch, Tensorflow, PyQt, VTK *Sep 2021 - Jul 2023*

- Automated blood sample reporting with YOLOv4, reducing the time by 20x and saving **Roche** $400K annually.
- Integrated Meta Detectron2 for catheter inspection, enhancing precision and efficiency for **J&J**'s medical devices.
- Created a tool to convert 2D CT scans to 3D models, enhancing medical diagnostics and implant development.

Data Engineer | Python, Spark, SQL, AWS, NLP, Recommender System *Nov 2020 - Sep 2021*

- Designed and deployed an AWS data lake from scratch, improved UI/UX via A/B testing, and led R&D on OCR-based report summarization and employee recommender systems.

PROJECTS

MTA Commute Pal Project @Columbia University 🔗 *Oct 2023*

- Analyzed **geospatial data** from NYC's MTA turnstile database to predict subway traffic and optimize commute routes.
- Built interactive dashboards to visualize congestion patterns and weather correlation, improving commute time estimation.

Metadata Analysis and Rule Engine for Automated Data Validation @Barclays *Oct 2024*

- Developed LLaMA (via Ollama API) to automate metadata analysis and ensure audit compliance for Payments and Credit Data.
- Deployed a Rule Engine for automated data validation, reducing errors and improving reporting timelines by 30%.
- Leveraged NLP to detect metadata inconsistencies, enhancing data accuracy and compliance with governance policies.

Oasis Data Lake - Genentech @Tata Consultancy Services *Sep 2021*

- Led a team to deploy a 40TB AWS data lake with ETL pipelines using PySpark and SQL, centralizing data for Genentech.
- Improved UI/UX through **A/B testing** and click pattern analysis in AEM, enhancing user engagement.
- Developed a **geospatial sales analysis utility**, identifying high-performing regions and driving $2M+ in revenue within the Q1.

AI-Driven Price Prediction and Trading Model @Columbia University 🔗 *Apr 2024*

- Devised and backtested an AI trading model to predict IWM ETF stock price using historical and macroeconomic data.
- Achieved 29.48% ROI during a 7-day live trading simulation, leveraging advanced backtesting techniques.

Explainable AI for Transaction Fraud Detection @Columbia University 🔗 *Dec 2023*

- Developed interpretable ML solution for fraud detection using Decision Trees, Random Forest, and SHAP.
- Conducted comprehensive financial risk analysis, identifying fraud indicators and feature engineering with AutoML.

GenAIJudge @Harvard Hackathon 🔗 *Jan 2024*

- Fine-tuned GPT-4 for scalable evaluation of 10K+ proposals on the circular economy using dynamic scoring rubrics.
- Integrated 3D topic clustering with Tensorboard for intuitive visualization, optimizing the proposal selection process.

RealFlow Assist *Feb 2024*

- Developed a real-time conversation and de-escalation assistant powered by Google's Gemini LLM.
- Leveraged prompt engineering techniques for accurate speech-to-text conversion and dynamic response generation.

PUBLICATIONS

- A novel ML-based analytical framework for auto-detection of COVID-19, IMA, 2021 (doi.org/10.1002/ima.22613)
- Serum and CSF Cytokine biomarkers for diagnosis of Multiple Sclerosis, MOI, 2020 (doi.org/10.1155/2020/2727042)
- TENSOPIT - Tensor Structured Bloom Filter: data caching using real-time space **forecasting** (under review)
- Image Super-Resolution using GAN: 2-stage Semantic Information GAN to improve resolution (under review)

SKILLS

- **Programming:** Python (Adv.), SQL (Adv.), C/C++ (Int.), PySpark (Int.), R/D3 (Beg.), Shell/Bash (Int.)
- **Frameworks:** Pandas, Numpy, NLTK, Scikit-Learn (Adv.), PyTorch, TensorFlow, Keras, OpenCV (Int.), Open-AI
- **Databases:** MS SQL Server, MySQL, PostgreSQL, Oracle RDBMS
- **Tools:** AWS (Glue, S3, Athena, EC2, SageMaker, EMR, Airflow, Redshift), GCP, Kubernetes, Looker, Tableau, Git, AEM

EDUCATION

Columbia University New York, US

*M.S. in Data Science, **GPA**- 3.7/4.0*

- Honors: R.G.S. Scholarship ($120K); TA of AML course; Graduate Assistant at NE Big Data Hub; Data Science Student Council