Shivam Choudhary

Data Scientist – Fractal Analytics

Shivamchoudhary2014@gmail.com

+91 9062330628

https://github.com/shiv-choudhary

Hyderabad

https://linkedin.com/in/choudhary-shivam

https://shiv-choudhary.github.io/portfolio

Summary and Experience:

➢ Data Scientist specialized in Natural Language Processing (**NLP**) with ~5 years of experience in developing Data Science solutions, Machine Learning pipelines, Big Data Pipelines, Cloud Migration, and API Development.

➢ Extensive domain experience in healthcare & life sciences, insurance, biopharmaceuticals, and media & entertainment.

➢ Expert in deriving insights from unstructured data using Large Language Models (**LLMs**), Deep Learning (**DL**) Machine Learning (**ML**) & **Statistics** techniques, particularly in **Product Management**.

➢ Proficient in designing data-driven solutions utilizing advanced technologies of unsupervised learning i,e. Topic Modelling using BERTopic, Transformer models other Foundation Models. Generative AI Framework such as Langchain and Llamaindex

➢ Skilled in text classification, text-to-text generation, text summarization, conversational AI and Retrieval-Augmented Generation (RAG).

➢ Skilled in both agile and waterfall methodologies for project delivery.

➢ Awarded INSTA Award for implementing a cost-effective data pipeline to process millions of unstructured data (PDF).

➢ Received Impact Award for contributing to the design of an ETL pipeline integrated with ML models, scaling inference to ~40 million records.

➢ Master of Computer Application (**MCA**) from the University of Mysore.

➢ Bachelor of Computer Application (**BCA**) from the West Bengal University of Technology.

Tools & Skillset:

1. OS: Linux, Win
2. Language: Python
3. Database: MySQL, MongoDB, Snowflake
4. Cloud: AWS, Azure, GCP
5. Framework: Flask, FastAPI, GraphQL, PyUnit, PySpark
6. Tools: Airflow, RabbitMQ

Certifications:

➢ Microsoft Certified Azure Data Scientist (DP:100)

➢ Microsoft Certified Data Fundamental (DP-900)

➢ Microsoft AI Fundamental (AI-900)

➢ Microsoft Azure Fundamental (AZ-900)

➢ Microsoft Certified AI-Engineer (AI-100)

➢ Microsoft Certified Python Programmer (MTA 98-381)

➢ AWS Cloud Practitioner

Project Experience:

Fractal Analytics – *(July 2023 – till date)*

1. Hardware Device Insights
- ✓ **Description:** Processed a large volume of chat(~40k) and call transcripts(~70k) to derive insights on customer satisfaction rate, repeat call rate, and SLA adherence for the Customer Satisfaction Index. Automated the process, achieving an ~85% automation rate, to efficiently generate these insights. These KPIs helped the hardware/device stakeholders to drive strategic decisions and make better customer experiences.
- ✓ **Domain**: Entertainment & Media
- ✓ **Role Played**: Data Scientist
- ✓ **Contribution**:
    - i. Leveraged Unsupervised techniques like Clustering, HDBSCAN, UMAP and BERTopic.
    - ii. Used LLM and Statistical techniques to support business-driven decisions
- ✓ **Technology Used**: Python, LLM(Gemini Flash 1.5), BERTopic, Data warehouse, Topic Modelling, SQL

2. Advance Asia Analytics Platform
- ✓ **Description:** Designed, implemented & orchestrated AWS Cloud native solution for three use cases: generation of customer 360 data/view, smoker propensity use case and Top Advisor analytics. Combining three processed ~90 million data over the cloud. Conversion of SAS code to Python equivalent.
- ✓ **Domain**: Insurance
- ✓ **Role Played**: Data Scientist
- ✓ **Technology Used**: Python, AWS Sagemaker, Lambda, Step Function, DL

Carelon Global Solutions – *(Sept 2022 – July 2023)*

1. Care Management
- ✓ **Description**: Engineering & Orchestration of big data pipeline and ML models enabling Care Management applications for model inference. Developed a patient outreach analytics automation that will optimize the expenses of the organization and resource management.
- ✓ **Domain**: Healthcare
- ✓ **Role Played**: AI Engineer
- ✓ **Contribution**:
    - o Created a pipeline to get data from Hive tables and mongoDB to perform the pre-processing feature using **Pyspark**.
    - o Plugged **XGBoost-based** AutoML model in **BigData** Pipeline which will infer ~40 Million datapoints.
    - o Implemented **SHAP** to understand local/global feature interpretation and feature importance while real-time inferencing.
    - o **Technology Used**: Airflow, Snowpark, MongoDB, PySpark, Hadoop, Yarn, Responsible AI

Deloitte USI – *(Jan 2021- Sept 2022)*

1. Data Migration and Analytics Platform
- ✓ **Description**: Migration applications of Big Data Stack to Snowflake and exposed via restful service.
- ✓ **Domain**: Healthcare
- ✓ **Role Played**: AI Engineer
- ✓ **Contribution**:
    - o Migrated on-premises big data stack to snowflake leveraging AWS data services (AWS glue, step functions)
    - o Analyzing the big data stack and suggesting similar cost-cutting options to clients to optimize the cost of deliverables.
    - o Handling the project deliverables from offshore, plan Scrum planning to avoid the risk of delivery on time.
- ✓ **Technology Used**: Snowflake, Spark, Scala, AWS

2. **Data Migration and Application Modernization**
   - ✓ **Description**: Application migration from Mainframe Technology to Digital Cloud Service (AWS) in the form of scalable microservices.
   - ✓ **Domain**: Healthcare
   - ✓ **Role Played**: Data Science cum Data Engineer
   - ✓ **Contribution**:
     - o Identified the layer of data and designed separate microservice for all the data access layer
     - o Deployed designed microservices on EKS.
     - o Written 100+ transformation rules to achieve graphQL user requests.
     - o Integrated restful services with graphQL framework to reduce user requests' latency by 60%.
   - ✓ **Technology Used**: Python, AWS, MongoDB, Flask-restful, graphQL, pytest, Docker, Kubernetes

Infosys Ltd - *(Sept 2019- Dec 2020)*

1. **Data Remediation Tool**
   - ✓ **Description**: An in-house ETL workbench tool customization for the client, digitization of legacy paper recipes for modern MES system to be migrated and implemented.
   - ✓ **Domain**: Biopharmaceuticals
   - ✓ **Role Played**: Data Science Associate
   - ✓ **Contribution**:
     - o Designed a data pipeline that will extract data from PDFs and word files and dump it into PostgreSQL for data processing and data wrangling
     - o Implemented ML model to generate keywords for building block suggestion and Template recommendation (BoW, TF-IDF, rake-NLTK)