Adhiraj Banerjee

Data Professional

linkedin.com/in/adhiraj-banerjee
github.com/adhirajbane13
adhirajbane13@gmail.com
+447438720064
UK

Summary

Data Professional with 2.5+ years of experience in data analysis, automation, and predictive analytics using Python, SQL, Power BI, and Azure Databricks. Skilled in ETL pipelines, machine learning, and financial modelling, with a strong focus on optimising workflows and extracting actionable insights. Adept at building interactive dashboards and high-impact reports to support data-driven decision-making.

Skills Summary

- **Programming & Scripting**: Python, SQL (PostgreSQL, MySQL), PySpark
- **Data Science & ML**: Scikit-learn, TensorFlow, NLP (NLTK), Statistical Analysis
- **Data Handling**: Pandas, NumPy, SciPy, MongoDB, Relational Databases
- **Big Data & Cloud**: Azure Databricks
- **Visualisation & Reporting**: Power BI, Tableau, Matplotlib, Microsoft PowerPoint
- **Productivity Tools**: Advanced Microsoft Excel

Work Experience

- **University of Sheffield** — Sheffield, UK
  Data Scientist — *February 2024 – September 2024*
  - Engineered a hybrid financial model using **Transformer-based forecasting** and the **Markowitz model** in **Azure Databricks**, enhancing returns by **2%** and reducing data prep time by **40%** with automated ETL pipelines.
  - Optimized data management using **Spark SQL** and **Apache Hive**, creating **Parquet tables** for efficient portfolio weight retrieval; built an encoder-only **Transformer model**, achieving an **RMSE of 0.877** in market trend predictions.
  - Conducted a **400-day risk-return trade-off analysis** with advanced statistical techniques, proving that **Transformer-enhanced portfolios** delivered **1.4x higher returns** with lower risk profiles than traditional methods.
  - Designed and presented interactive **Power BI dashboards** and **Matplotlib visualisations** for senior stakeholders, showcasing portfolio insights like weight allocations, cumulative returns, and volatility.

- **Cambium Networks** — Bengaluru, India
  Data Analyst — *June 2022 - July 2023*
  - **Automated RF test data extraction processes**, implementing Python scripts, SCPI commands, and advanced SQL techniques that **improved data accuracy by 40%** and **reduced manual effort by 35%**.
  - **Designed and deployed automated data pipelines**, enhancing **data consistency by 30%** and **minimizing reporting errors by 45%**; **slashed report generation time from 8 hours to 2 hours**.
  - **Developed and customized Excel Pivot Charts**, significantly enhancing stakeholder understanding of key **RF performance KPIs**, leading to **quicker and more informed decision-making processes**.
  - **Streamlined RF test workflows**, achieving a reduction in data acquisition time from 8 to 1.5–3 hours, thereby **boosting the agility of product development teams**.

- **CSIR-CMERI** — Durgapur, India
  ML Research Engineer — *June 2021 - October 2021*
  - Constructed a **clustering-based Gaussian Mixture Model (GMM)** speaker recognition system using **scikit-learn** and **statistical packages such as NumPy and SciPy**, securing **96%** accuracy in noisy environments across **10 speakers**.
  - Implemented **Mel-Frequency Cepstral Coefficients (MFCC)** for feature extraction with **librosa** and **NumPy**, strengthening model robustness and reducing classification errors by **20%**.
  - Optimised **Expectation-Maximisation (EM) training** within the GMM framework using **SciPy**, improving speaker classification under varying noise levels.
  - Conducted **noise variance analysis** using **SciPy's Wiener filter**, ensuring reliable speaker recognition performance in challenging acoustic conditions.

- **Tata Consultancy Services** — Kolkata, India
  NLP-ML Engineer(Internship) — *May 2020 - July 2020*
  - **Engineered** a supervised machine learning pipeline for **grammatical error detection**, leveraging **custom SpaCy-based tokenizers** and **LinearSVC classification**. Optimized model accuracy from **78% to 85%** through advanced feature engineering.
  - **Developed scalable NLP preprocessing pipelines**, incorporating **text normalization, lemmatization, and stopword removal**, achieving a **10% reduction in false positives** in classification.
  - **Designed and implemented** an end-to-end **NLP pipeline**, integrating **data cleaning, tokenization, feature extraction, and classification models**, improving computational efficiency by **30%**.

PROJECTS

- **NHS Accident and Emergency (A&E) Performance Analysis**:
  An interactive Power BI dashboard with automated ETL for NHS A&E trends, featuring trust-level insights, KPIs, and dynamic drill-through storytelling.
  ***Technologies**:* Python, Pandas, PowerBI(Interactive Dashboard,DAX querying,Data Storytelling), ETL, BeautifulSoup, PostgreSQL, Docker, Windows Task Scheduler[*GitHub*]
- **Tender Intelligence Assistant – GenAI-Powered Tender/RFP Q&A Platform**:
  A GenAI tool that answers natural language questions from tender PDFs with structured, context-aware responses.
  ***Technologies**:* Streamlit, Python, OpenAI GPT-4 & text-embedding-3-large, FAISS(Facebook AI for Semantic Search), PDFMiner, PDFPlumber[*GitHub*]
- **Advanced TensorFlow Deep Learning-Based Spam Detection for Email Classification**:
  A deep learning-based email spam classifier using TensorFlow, Keras, and NLP techniques.
  ***Technologies**:* TensorFlow, Keras, Scikit-learn, NLTK, Pandas, NumPy, Matplotlib, Seaborn[*GitHub*]
- **UK Regional Salary & Working Hours Analysis 2024**:
  A Tableau dashboard analyzing 2024 UK regional salary disparities and working hours.
  ***Technologies**:* Tableau, Python (Pandas, Jupyter), Microsoft Excel[*GitHub*]
- **Financial Portfolio Optimization: Integrating Transformers with the Markowitz Model**:
  A Transformer-enhanced Markowitz portfolio model on Azure Databricks, boosting small-cap returns with reduced risk.
  ***Technologies**:* Python, TensorFlow, Pandas, NumPy, Matplotlib, Yahoo Finance API, PowerBI, Deep Learning[*GitHub*]
- **Gaussian Mixture Model-based Speaker Recognition**:
  GMM-based speaker recognition using MFCCs for robust identification across varied acoustic conditions.
  ***Technologies**:* Machine Learning, Voice Signal Processing, MATLAB, NumPy, SciPy, librosa, scikit-learn, Matplotlib, pickle(for model serialisation)[*GitHub*]

EDUCATION

- **The University of Sheffield** — Sheffield, UK
  Masters of Science - Data Analytics; Grade: Distinction — *September 2023 - November 2024*
  - **Modules:** Data Science with Python, Scalable Machine Learning (including PySpark), Machine Learning and Adaptive Intelligence, NLP, Parallel Computing with GPU, Text Processing, Professional Issues.
  - **Academic Project: AI + IoT for Smart Homes via Free LLMs** – Fine-tuned **Vicuna LLM** for **smart home automation**, enhancing command execution accuracy from **85% to 92%**. Developed a **specialized dialogue dataset** to improve response precision, simulated **rule-based vs. LLM-enhanced interactions**, reducing execution errors by **30%**, and conducted a **comparative analysis** showing a **25% improvement** in adaptability to dynamic smart home environments.

- **Indian Institute of Engineering Science and Technology** — Shibpur, West Bengal, India
  Bachelors of Technology - Electronics Engineering; Grade: 1st Class(Honors) — *August 2018 - May 2022*
  - **Modules:** Signals and Systems, Digital Signal Processing, Wireless and Mobile Communications, Communication Systems, Digital Image Processing & Computer Vision.
  - **Academic Project: Pulse-Oximeter using Arduino and MAX30100 Pulse Sensor** – The project focused on creating a pulse oximeter using an **Arduino Uno board** and **MAX30100 pulse sensor**, motivated by the COVID-19 pandemic's demand for **healthcare monitoring technologies**. It highlights the application of embedded systems and sensor integration to address urgent healthcare needs.