Shubham Banerjee *Sr. Data Scientist*

Bangalore, India shubhambanerjee844@gmail.com 8617578490

https://github.com/ShubhamCoder007 https://www.linkedin.com/in/shubham-banerjee8

Professional Experience

2023/03 – present **Sr. Data Scientist,** *Deloitte & Touche LLP*

- Led the integration of GenAI technologies leveraging Langchain and LLMs such as GPT-4o into audit workflows, automating 80% of financial statement analysis tasks and reducing manual effort by 400+ hours annually.
- Pioneered the Doc-Chat project, allowing auditors to interact with financial statements for instant insights, reducing document review time by 50%.
- Engineered Metadata extraction project leveraging Generative AI to extract financial attributes from FS in multiple language, cutting manual efforts by 80%.
- Developed custom evaluation framework Auto-Eval for evaluation of Doc-Chat which was later adopted by other projects across the team.

2020/01 – 2022/12 **Data Scientist,** *Comviva*

- Designed and implemented predictive analytics models to forecast customer behaviors, including churn, offer recommendations, and service migration.
- Boosted marketing strategies via SARIMAX & LSTM forecasting insights, enhancing campaign performance; increased model accuracy by up to 30% through drift mitigation, retraining, and fine-tuning.
- Reduced campaign costs by 70% through targeted segmentation, directly contributing to a 15% revenue increase.

Education

2016 – 2020 **B.Tech Electronics and Computer Science,** *KIIT, CGPA*: *8.23*

Skills

Generative AI & NLP
Generative AI, LLMs, Fine-tuning, Langchain, Llamaindex, Agents, Prompt Engineering, RAG, Vector Database, NLP, Embeddings, BERT & Transformers

Probability and Statistics, Statistical testing

Computer Vision
Image processing, CNN, ResNets

Libraries & Frameworks
Sklearn, Numpy, Pandas, Matplotlib, Seaborn, Keras, TensorFlow, PyTorch, OpenCV, Statsmodel, HuggingFace

Data Science & ML Engineering
EDA, Data Visualization, Data Interpretation, Feature Engineering, Feature Selection, Hyperparameter Optimization

Machine Learning / Deep Learning
Bagging, boosting, SVM, GLM, ANN, CNN, RNN

Time Series Analysis
Time Series Analysis & Forecasting (Double/Triple Exponential Smoothing, ARIMA, SARIMA), Prophet

Projects

Doc-Chat
Objective: Design a conversational chatbot for querying financial statements with citations.
Solution:

- **Data Processing:** Designed table-aware semantic chunking to preserve tabular data which improved completeness score from 70% to 80%. Granular metadata such as content type, followed by multimodal embedding which combines text with numerical embeddings - weighted concatenation for richer representation.

- **Query Processing:** Performed query augmentation through decomposition, enrichment, and thresholding. Followed by hypothetical question generation and cluster assignment for better retrieval.
- **Contextual Query Completion:** Created contextually whole queries for dependent conversational queries by leveraging the chat-history component.
- **Retrieval Pipeline:** Implemented efficient priority/relevancy reranking where in, priority weights are assigned to every retrievals based on parameters such as relevancy score and frequency of occurence, and then performing minimization of cumulative priority scoring given the threshold for optimal retrieval.
- **Chunk Completion:** Fine-tuned BERT to predict continuation of candidate chunks, dynamically integrating relevant segments into context—increased completeness score from 80% to 87% and enhanced answer relevancy.
- **Response Generation:** Leveraged chat-history context, follow-up chunks, and citation sources to generate responses.
- **Chat-History Management:** Maintain the last 3 Q&A stacks in chat-history and identify follow-up questions.

Metrics:

- **Completeness:** 87% **Correctness:** 85% **Faithfulness:** 90%
- **Response Time:** 10 seconds (average time taken to generate a response)
- **Business impact:** Reduced FS Auditing time by 50%.

Get The Latest

Objective: To predict potential consumers of the GTL product and efficiently localize campaign targeting.
Solution:

- Resolved extremely imbalanced data problem (class A: class B – 1: 200); with advanced sampling techniques such as ratio-based smote up-sampling.
- Incorporated features to reduce the imbalanced nature and reduce noise by 31% for the models.
- Formulated techniques such as model-based imputation to tackle the issue of very high missing data.
- Achieved Overall Accuracy of over 85% and was able to correctly classify 75% Class A, and 85% Class B; where the given target was 80% overall accuracy, 60% Class A and 80% Class B.
- Boosted revenue by nearly 15% by increasing the number of consumers, while also decreasing campaign costs by 70% through targeted segmentation.

Certifications

Summer training in Machine Learninng using Python

Hands on Machine Learning and Deep Learning, *Udemy*

Time Series Data Analysis, *Udemy*