SAKSHI GUPTA

Email: sakshi.k425@gmail.com **Phone:** (980) 210-9830 **Portfolio:** https://sakshigupta.me
GitHub: @sakshi1989 **Tableau:** @sakshi.gupta7992 **LinkedIn:** @sakshigupta89

SKILLS

Large Language Models	NLP	Machine Learning	Deep Learning
Statistical Analysis	Data Visualization	Agile Methodologies	Version Control

Technical Skills

Python	SQL	T-SQL	SAP ABAP

Software, Services & Packages

Python	Langchain	Pytorch	Scikit-Learn
PySpark	SQL Databases	Snowflake	Microsoft Azure
Tableau	SAP ERP	Microsoft Excel	SAS Enterprise Miner

EDUCATION

University of North Carolina at Charlotte **(Jan 2022 - Dec 2023)**

- Master of Science in Data Science and Business Analysis (4.0 GPA)
- Secretary of the School of Data Science Student Council
- Instructional Assistant for Database Systems for Data Scientists

Arya College of Engineering and IT **(Aug 2007 - May 2011)**

- Bachelor of Technology in Computer Science (Honors)

PROFESSIONAL EXPERIENCE

Data Scientist

Siemens Energy (Charlotte, NC) **(Apr 2024 - Present)**

Leading the project to enhance multi-skilling and knowledge sharing among energy engineers through the development of a LLM-driven Retrieval-Augmented Generation (RAG) application, enabling efficient information retrieval for design tasks and accelerating project timelines. Separately, contributing to the development of an AI-powered Product Change Management workflow using Microsoft Azure AI Search Intelligence.

Accomplishments

- Developed a novel document post-parsing strategy that significantly improves the accuracy of information extraction from internal documents (PDF, Microsoft Office formats). It addresses the following limitations of standard parsing services (Unstructured.io and Azure AI Document Intelligence).
  a. Accurately identifying heading levels.
  b. Associating captions with figures and tables.
  c. Avoided fragmentation and text misidentification by restructuring complete figures.
  d. Accurately parsing formulas as LaTeX.
  e. Correcting overall document layout.
- Created advanced chunking strategy that retains complete context for LLM processing by incorporating title/subtitle/heading hierarchy with each chunk, and integrating context from figures and captions.
- Developed custom LCEL (LangChain Expression Language) chains employing map-reduce technique with a custom threshold implementation to improve query precision. These chains effectively integrate information from diverse document sources, enabling more accurate responses.

- Crafted prompts and utilized few-shot prompt engineering techniques to guide the LLM in extracting answers from context documents and effectively responding to user queries.
- Prepared experiments to evaluate various LLM models for the use case considering response accuracy, response time and relevance.
- Streamlined development workflows and enhanced team collaboration through Git usage.

Data Science Intern

Sia Partners (Charlotte, NC) **(Jun 2023 - Aug 2023)**

Specialized in analyzing unstructured safeguard data to streamline audit expenses. My role entailed extracting and analyzing data from various sources using PrestoSQL to identify the factors driving safeguard audit costs. Additionally, I supported the company's Large Language Model (LLM) initiative and contributed to its project refinement.

Accomplishments

- Devised KPIs that allowed the team to track the completion status in each safeguard Google Sheet, enabling accurate and transparent progress tracking.
- Engineered advanced formulas in Google Sheets to streamline multiple safeguard sheets.
- Crafted comprehensive dashboards and visualizations to provide insights and recommendations to stakeholders on Privacy Risk Assessment responses.

Data Analyst Intern

Caring.com (Charlotte, NC) **(Jan 2023 - May 2023)**

Focused on analyzing senior care partners' data to optimize business operations. I utilized Snowflake SQL for exploratory data analysis and employed Looker for visualizations. A crucial part of my role was developing approaches to classify partners at risk of churn and clustering them into tiers, enhancing the efficiency of lead distribution and partner management.

Accomplishments

- Achieved monthly time savings of 2 hours for the Sales Director by designing an automated report on partner associations and streamlining its distribution using Google Apps Script.
- Analyzed partner activity patterns to predict the probability of partners churning at different time intervals.
- Effectively segmented Homecare Partners into tiers for personalized engagement and support using clustering and core statistical methods on the data from Snowflake.
- Designed charts to showcase insights on both the Snowflake dashboard and Google Sheets.

Data Science Associate Intern

Azimuth GRC India **(Jun 2022 - Aug 2022)**

Specialized in developing and managing data pipelines using Microsoft Azure Data Factory. I played a crucial role in automating and enhancing data validation strategies, utilizing T-SQL to create stored procedures.

Accomplishments

- Gained experience in Azure Data Factory (ADF) pipelines and data validation strategies through learning on the go.
- Automated the data quality check process using stored procedures running as an ADF pipeline step, improving the accuracy and efficiency of the data validation process.
- Enabled communication and transparency in the data validation process by creating a Logic Apps step in the pipeline that notifies clients of data issues via email.

Other Experiences

Senior Consultant, HCL Technologies, India **(May 2017 – Feb 2021)**

Associate, Cognizant, India **(Sep 2014 – Apr 2017)**

Application Developer, Fujitsu Consulting, India **(Aug 2011 – Sep 2014)**

- Gained proficiency as an SAP ABAP developer, specializing in reports, form design, system enhancements, and integrations while expertly handling multiple SAP modules.
- Mentored colleagues for smooth integration, demonstrating leadership and fostering their professional growth, with a focus on team collaboration and effective communication.
- Led automation initiatives, streamlined processes like bulk company code updates, and developed a reusable report-generation framework, significantly improving team efficiency.

Data Science/Machine Learning Projects (Full Portfolio)

LLM-based Chatbot using LangChain (Launch App)

- Developed an app using the LangChain framework and GPT-3.5 Turbo model, specializing in providing answers related to Bank of America and Wells Fargo's 2022 annual reports.
- Implemented parsing of PDF reports, converting data into vectors for indexing and information retrieval.
- Engineered precise prompts to help the Large Language Model (LLM) decide whether to use the conversation history in context, enhancing the response accuracy and relevance.
- Launched a user-friendly Streamlit application interface for easy access and interaction with Chatbot.

Meme Generation using Deep Learning (Launch App)

- Implemented the model using Pytorch and encoder-decoder architecture.
- Used ResNet50 and InceptionV3 for image encoding and pre-trained GloVe embeddings on the Twitter dataset for text embeddings.
- Employed text pre-processing, including breaking down hashtags into meaningful tokens, fuzzy matching to optimally align input tokens with GloVe dictionary, n-gram language detection, and Google Translate to translate non-English captions.
- Applied Beam Search with Top-K to select appropriate meme output.
- Evaluated the model using BLEU and BERT scores, ensuring quantifiable benchmarks for assessing the quality and relevance of the generated memes.

Customer Default Prediction – Kaggle AMEX Dataset (View Source)

- Collaborated with a team of four to predict the likelihood of credit card payment defaults.
- Performed Exploratory Data Analysis to identify data irregularities and address imbalances, strategically selecting appropriate resolution techniques.
- Utilized Logistic Regression with VIF-based variable elimination, PCA for feature transformation to address multicollinearity, and the incorporation of SHAP values for model simplification.
- Performed classification including RandomForest, XGBoost, and LightGBM, with hyper-parameter tuning using GridSearchCV, gaining a 5% accuracy boost.

Airline Dashboard (View Dashboard)

- Designed an impactful Tableau Dashboard showcasing comprehensive flight information, utilizing data sourced from the Snowflake Marketplace.
- The dashboard helps answer the following questions:
  - Are there any flights that will take them to their desired destination?
  - Will their desired in-flight service be available on the flight they choose?
  - What is the flight frequency between the source and destination airports, and how can this data inform credit card companies' airline partnerships for loyalty programs?