Lathish Balaji Baskaran

| Glassboro, NJ | lathishb510@gmail.com | (856) 462-2845 | Github |

SUMMARY

Data Scientist with expertise in Machine Learning, AI, and Data Engineering. Skilled in building AI Agents, ML models, multimodal AI applications, and BI dashboards. Proficient in TensorFlow, PyTorch, and LangChain, with a focus on data processing, predictive modeling, and delivering insights through Tableau. Enthusiastic about AI, with a strong focus on advancing skills in deep learning and predictive analytics.

WORK EXPERIENCE

Data Scientist, Rowan University **May 2024 - Present**

Job Description:

- Designed and deployed an AI-driven validation tool to analyze higher ed survey data for accuracy, leveraging Azure OpenAI models to generate concise discrepancy reports and automate validation, reducing manual effort by 70%.
- Developed a web application to automate survey data collection using AI agents that assign documents to officials, track submissions, and send scheduled reminders, streamlining workflows and improving efficiency. Utilized Git for version control and collaborative development.
- Engineered RAG-based NLP chatbots with department-specific knowledge bases for IT, integrating them into the institution's help page to enhance self-service for students and faculty, reducing support center calls from 150 to 80 per week.
- Designed complex SQL queries to extract tables and views for machine learning analysis, AI application development, and ad-hoc reporting, ensuring data accuracy and usability.
- Performed predictive analysis on higher education data using multiple ML models to uncover patterns, simulate scenarios, and identify optimal strategies, leading to a 5-point increase in university ranking.
- Built dynamic BI dashboards and reports, enabling both technical and non-technical stakeholders to visualize ranking impact across different scenarios for strategic decision-making.
- Integrated APIs across multiple platforms, including the Informatica Business Terms API, using Python with Tableau and TabPy. Enabling real-time data exchange, optimized dashboard performance, and reduced latency while cutting manual work by 80%.
- Developed a complex R script to parse and restructure ETL job data from the Veera platform, ensuring seamless integration with Informatica Data Governance by aligning data formats to required standards.
- Built a Django application using OpenAI API and Llama Parse to parse and process 750 resumes, extracting candidate work experience for surveys and institutional research. Automated data extraction and standardization, improving accuracy, efficiency, and reducing manual effort.

Graduate Research Assistant, Rowan University **March 2023 – May 2024**

Job Description:

- Developed a machine learning pipeline leveraging PCA for dimensionality reduction, identifying 8 key variables influencing Carnegie rankings of research universities, highlighting drivers of institutional performance and areas for growth.
- Enhanced model architecture by adding dense encoding layers with ReLU activation and dropout regularization, reducing model overfitting by 25% and improving feature extraction for accurate classification of university performance.
- Performed extensive data cleaning and preprocessing to address inconsistencies, missing values, and outliers in large datasets, enabling the machine learning model to uncover meaningful patterns that provided actionable insights, shaped strategic decisions, and guided future research initiatives.
- Deployed the machine learning model as a Flask-based web app with an interactive Plotly dashboard, containerized using Docker for scalable, cross-environment deployment, enabling real-time ranking predictions and scenario analysis, improving decision-making.

Junior Data Analyst, Vinayak Communication **Jan 2022 - Dec 2022**

Job Description:

- Analyzed large sales datasets using Pandas and Excel, identifying trends and performance patterns for a wholesale and retail electronics company.
- Utilized PyTorch based LSTM model to analyze sales data, identifying high-revenue regions and key sales drivers by capturing seasonality and market trends.
- Developed interactive Power BI dashboards and data visualizations using Matplotlib and Seaborn, transforming complex datasets into clear, actionable insights that enhanced data-driven decision-making.
- Analyzed historical sales data to uncover revenue trends, optimize pricing strategies, and enhance supply chain efficiency, leading to more accurate demand forecasting and cost reductions.

SKILLS

Programming Languages: Python, SQL, R, JavaScript, Go, MATLAB
RDBMS: Oracle, MySQL, Mongo DB, PostgreSQL
Big Data Technologies: PySpark, SparkSQL, Hadoop
Data Visualization: Tableau, Cognos, Power BI
Machine Learning/Python libraries: TensorFlow, PyTorch, Keras, Hugging Face, Scikit-learn, Pandas, NumPy, Docling
Cloud Technologies: Microsoft Azure (Azure AI Foundry, Azure OpenAI, Azure Machine Learning), Amazon Web Services (AWS IAM, S3, EC2, SageMaker), Google Cloud Platform (Vertex AI Studio, Auto ML, Vision AI)
Web Development: Docker, Django, Flask, Streamlit
ETL Tools: Veera Construct, Azure Data Factory, SAS, Tableau Prep Builder

PROJECTS

Multimodal AI Application for Speech and Hearing Disabilities

- Developed a multimodal AI system that processes user prompts through a chat completion model with Retrieval-Augmented Generation (RAG) for knowledge retrieval. Utilized NLTK for text preprocessing, including tokenization and stop word removal, ensuring clean input for Transformer-based models like GPT, which generated context-aware responses before conversion into video or speech.
- Fine-tuned the Text-to-Video model using a dataset of 500+ sign language video clips, featuring sign language experts signing individual words with corresponding captions, ensuring accurate representation of the knowledge base. The Text-to-Speech model converted responses into speech for improved accessibility.
- Built an interactive user interface with real-time customization options, enabling users to adjust video speed, language preferences, and accessibility settings, ensuring a personalized experience.

Natural Language Database Query Chatbot

- Designed a natural language chatbot using LangChain and LLM-based chat completion models, enabling users to input queries in natural language, which the chatbot translates into structured SQL queries for seamless data retrieval.
- Processed and stored structured data in Oracle/MySQL, designing optimized schemas for efficient querying and retrieval.
- Implemented query optimization techniques such as indexing, normalization, and query caching, improving database performance and reducing response time.
- Ensured accurate SQL query generation by leveraging few-shot prompting, query templates, and context-aware query refinement, reducing errors and improving query reliability.
- Developed an interactive chatbot interface using Streamlit, enabling users to input natural language queries and retrieve insights dynamically from the SQL database.

Generative AI-Enhanced Tableau Dashboard

- Developed a Generative AI-powered Tableau extension to provide real-time explanations for business terms, enhancing data comprehension for users.
- Built a JavaScript-based Tableau extension that detects hover events and triggers NLP-based AI-generated explanations, enhancing interactive BI dashboards.
- Integrated pre-trained AI models via an API to dynamically generate contextual explanations in Tableau.
- Automated the delivery of concise, real-time insights, enhancing self-service analytics for business users.

Traffic Violations Analysis

- Utilized Databricks and PySpark to manage and preprocess large-scale traffic violations datasets, leveraging distributed computing for efficient data handling.
- Conducted exploratory data analysis (EDA) with Pandas and NumPy, uncovering patterns and key factors influencing traffic violations.
- Trained and optimized a Random Forest model using Spark MLlib, tuning hyperparameters such as tree depth, number of estimators, and feature selection to improve prediction accuracy.
- Developed interactive Tableau dashboards to visualize traffic trends, model predictions, and violation patterns, making complex data more interpretable and actionable.

EDUCATION

ROWAN UNIVERSITY, Glassboro, NJ **Jan 2023 - Dec 2024**

Master of Science in Data Science

Related Courses: Deep Learning, Data Mining, Data Warehousing, Visual analytics, Applied Multivariate

CERTIFICATIONS

Azure AI Engineer – Issued by Microsoft
AWS AI Practitioner – Issued by AWS