MAHESH YAGANDLA

+1 (346) 977-6876 | ymgmahesh@gmail.com | linkedin.com/in/ymgmahesh

PROFESSIONAL SUMMARY

- AI/ML Engineer with over Eight years of experience in NLP, Machine Learning, and MLOps. Skilled in buildingLLM-powered chatbots with RAG, Conversational AI for real-time responses.
- Expertise in NLP, including intent recognition and prompt engineering. Proficient in ML model development using TensorFlow, PyTorch, and Scikit-Learn, with experience in recommendation systems and
- predictive analytics. Strong MLOps background, implementing CI/CD, Kubernetes, cloud deployments (AWS, GCP), and model monitoring (Prometheus, LangSmith) for scalable AI solutions.
- Skilled in designing, fine-tuning, and deploying large language models for diverse applications such as natural language understanding, text summarization, sentiment analysis, and conversational AI, leveraging these models to enhance automation and decision-making processes.
- Proficient in building and optimizing machine learning models, with hands-on experience in feature engineering, model optimization, and hyperparameter tuning.
- Expert in data extraction, transformation, and processing using Python, SQL, Pandas, and NumPy, enabling robust data pipelines and analysis.
- Extensive experience in fine-tuning LLMs for specific tasks, optimizing their performance for targeted applications.
- Strong expertise in deploying and managing LLMs on cloud platforms, ensuring scalable and efficient model performance.
- Applied NLP techniques for sentiment analysis, text mining, and extracting actionable insights from unstructured data.
- Integrated Generative AI models and Large Language Models (LLMs) to enhance natural language processing capabilities.
- Developed and optimized deep learning models using TensorFlow, Keras, and PyTorch, driving innovation in predictive analytics.
- Implemented MLOps practices to ensure seamless model deployment and maintenance.
- Applied advanced statistical modeling and machine learning techniques, including regression, classification, and clustering, to analyze large datasets, leveraging Python libraries like Scikit-learn and TensorFlow for model building and optimization.
- Worked with structured and non-structured databases, including SQL and NoSQL systems, for data extraction, transformation, and storage in cloud environments like AWS and GCP.
- Developed and optimized big data pipelines on Databricks and Apache Spark, improving data processing efficiency and reducing computational time for large-scale machine learning projects.
- Managed cloud-based machine learning workflows, using services like AWS S3, EC2, and Lambda, ensuring scalability and cost-effectiveness of deployed solutions.
- Worked in a Linux OS environment, leveraging shell scripting for automation and system management, ensuring efficient deployment of machine learning models and pipelines.
- Strong advocate for the ethical use of AI and data, ensuring adherence to data privacy regulations and promoting responsible AI practices.

PROFESSIONAL EXPERIENCE

Truist **Charlotte, NC, USA**

Generative AI Engineer *August 2022 - Present*

- Designed and deployed an end-to-end AI-powered virtual assistant pipeline integrating structured/unstructured data, vector databases (Milvus), and large language models (LLMs).
- Built scalable ingestion pipelines for both structured data (e.g., customer profiles, order history) and unstructured data (e.g., manuals, FAQs), feeding into a SQL database and vector DB respectively.
- Implemented Retrieval-Augmented Generation (RAG) using a combination of text retriever microservices, Milvus vector DB, and reranking/embedding NIMs to improve context-aware responses.
- Built NL-to-SQL AI assistants using LangChain and LangGraph, enabling seamless natural language querying of databases with automated error handling.

- Fine-tuned Llama 2 using LoRA and QLoRA, optimizing model efficiency for domain-specific applications while reducing latency.
- Implemented hybrid search techniques combining dense vector embeddings with keyword-based retrieval to improve response accuracy and relevance.
- Designed agentic AI architectures using LangGraph and CrewAI, enabling multi-agent coordination for complex decision-making and workflow automation.
- Integrated CrewAI to develop collaborative AI agents, allowing distributed task execution and enhanced reasoning capabilities.
- Integrated function calling into AI agents, allowing dynamic interaction with external APIs for automated task execution.
- Developed and optimized prompt engineering strategies, including few-shot, zero-shot, chain-of-thought, ReAct, and self-reflection prompting, to improve AI response accuracy and contextual awareness.
- Developed interactive generative AI applications using Streamlit and React, enhancing user engagement with real-time AI-powered responses.
- Designed observability and debugging pipelines using LangSmith, improving model performance through detailed tracing and evaluation.
- Ensured system integrity by implementing guardrails to prevent prompt injection and adversarial attacks in AI applications.
- Implemented AI governance frameworks to ensure compliance with financial regulations and standards, reducing risk and enhancing the reliability of automated financial decision-making systems.
- Optimized large-scale document processing through chunking and indexing, enabling faster information retrieval in AI applications.
- Designed graph-based AI solutions using Neo4j, enabling advanced relationship-based querying and knowledge graph integration.
- Integrated human oversight into AI workflows, ensuring ethical decision-making and high-quality outputs.
- Built scalable ML pipelines using TensorFlow, PyTorch, and Hugging Face, deploying AI models in real-world production environments.
- Applied MLOps best practices, including CI/CD, model monitoring, and observability, leveraging Docker, Kubernetes, Prometheus, and Grafana.

Takeda **Exton, PA, USA**
AI/ML Engineer *January 2021 - July 2022*

- Designed and implemented a robust document extraction pipeline using Google Cloud Platform, incorporating Google Cloud Vision API for accurate OCR capabilities and Google Cloud Natural Language API for advanced text analysis.
- Developed and fine-tuned supervised ML models including Random Forest, XGBoost, and Logistic Regression for improving model accuracy
- Performed in-depth exploratory data analysis (EDA) and feature engineering using Pandas, NumPy, and Scikit-learn, leading to significant performance gains.
- Employed Grid Search, Randomized Search, and Cross-Validation techniques to optimize hyperparameters and mitigate overfitting.
- Evaluated model performance using ROC-AUC, Precision-Recall, F1-score, and confusion matrix, and communicated results to stakeholders.
- Implemented data preprocessing pipelines using Scikit-learn Pipeline API, standardizing workflows and reducing data leakage risk.
- Packaged ML training and inference code into reusable modules and tested using pytest and MLflow for experiment tracking and version control.
- Architected and deployed an end-to-end MLOps pipeline on Google Cloud Platform (GCP) using Vertex AI, Cloud Run, and Big Query.
- Built automated CI/CD pipelines with Cloud Build, Artifact Registry, and Terraform, enabling reproducible and production-ready ML deployments.
- Designed data ingestion workflows using Cloud Pub/Sub, Dataflow (Apache Beam), and Cloud Storage to handle both real-time and batch processing.

- Used Vertex AI Pipelines and Kubeflow Pipelines to orchestrate training, validation, and deployment steps for scalable model lifecycle management.
- Set up continuous monitoring using Vertex AI Model Monitoring and Cloud Logging, with automated retraining triggers based on drift detection.
- Managed experiment tracking and model versioning using Vertex AI Experiments, Model Registry, and Tensor Board, ensuring reproducibility and auditability.
- Containerized ML services using Docker and deployed scalable inference endpoints on Cloud Run and Kubernetes Engine (GKE).

Axalta **Philadelphia, PA, USA**
Data Scientist *November 2019 - December 2020*

- Designed and automated ETL pipelines in a Hadoop and Kafka ecosystem, integrating data from IoT devices, SQL databases, and real-time streams for seamless data ingestion and processing.
- Developed custom Python scripts, utilizing Kafka for data streaming and Hadoop for distributed data processing to clean, transform, and aggregate data for real-time analytics.
- Leveraged Spark for exploratory data analysis (EDA), identifying trends and anomalies in manufacturing processes to provide actionable insights for process optimization.
- Built and deployed predictive models using scikit-learn and Spark MLlib, enabling real-time equipment failure prediction and optimized maintenance schedules, significantly improving operational efficiency.
- Used Run-Length Encoding (RLE) to optimize the storage and processing of segmented image data, leveraging Spark for efficient handling of high-volume data streams.
- Evaluated model performance using cross-validation and fine-tuned models to ensure accuracy and reliability, utilizing Spark for scalable, distributed computations.
- Visualized data insights using Matplotlib and Seaborn, and integrated ELK for real-time monitoring, enabling stakeholders to track key metrics through dynamic dashboards.
- Documented data science workflows in Jupyter Notebooks, ensuring project reproducibility, with version control and monitoring supported by GitHub and ELK for real-time data tracking.
- Delivered real-time insights via Power BI and Tableau dashboards, integrated with Elasticsearch for instant updates and enhanced reporting capabilities.
- Preprocessed and cleaned large datasets using Pandas, NumPy, and Spark, ensuring data quality for predictive models, while maintaining data flow integrity through Kafka.
- Deployed and monitored machine learning models using Kafka and ELK, ensuring continuous performance and adaptability to evolving production conditions.
- Optimized data pipelines with Kafka for real-time streaming and Spark for large-scale processing, reducing latency and improving overall system efficiency.
- Implemented CI/CD pipelines for streamlined development, testing, and deployment of AI models, leveraging Kafka for message-driven architectures and ensuring smooth integration into production environments.

Dell Technologies **Round Rock, TX, USA**
Data Scientist *October 2018 - November 2019*

- Developed predictive models using Azure Machine Learning Studio and Python libraries such as Pandas, NumPy, and Scikit-learn for time-series analysis.
- Utilized Azure Data Factory and Azure Databricks for data cleaning, preprocessing, and ETL processes, ensuring data quality and accuracy.
- Leveraged Azure Synapse Analytics and Power BI for data visualization to identify trends and patterns in sales data.
- Applied machine learning algorithms, including Random Forest, XGBoost, and ARIMA, using Azure Machine Learning services to improve sales predictions and optimize sales strategies.
- Used Jupyter Notebooks within Azure Databricks and integrated with Microsoft Teams for effective communication and collaboration with sales teams.
- Created comprehensive sales reports and dashboards using Power BI and Azure Synapse Analytics to visualize key metrics and performance indicators.
- Designed and executed A/B tests using Azure Machine Learning and statistical techniques to evaluate the effectiveness of different sales strategies and campaigns.

- Performed customer segmentation analysis using K-means clustering and other unsupervised learning techniques within Azure Machine Learning to identify target markets and improve marketing efforts.
- Developed and maintained automated data pipelines using Azure Data Factory and Apache Airflow for seamless data integration and analysis.
- Delivered actionable insights using data analysis tools such as Azure SQL Database, Python, and Power BI to support decision-making processes.
- Evaluated sales performance against targets and benchmarks using Azure SQL Database and Python to identify areas for improvement.
- Conducted trend analysis using time-series decomposition and seasonal-trend decomposition (STL) techniques within Azure Machine Learning to identify emerging market trends and adapt sales strategies accordingly.
- Collaborated with various departments, including marketing and finance, using Azure DevOps and other project management tools like Asana and Trello to align sales strategies with overall business objectives.

DataReady Technology **Hyderabad, TG, India**
Data Analyst *August 2016 - May 2018*

- Collected, cleaned, and preprocessed data from various sources, ensuring accuracy and consistency.
- Performed exploratory data analysis (EDA) to uncover patterns, trends, and insights.
- Developed and implemented machine learning models and algorithms to solve complex business problems.
- Conducted feature engineering to enhance model accuracy and performance.
- Evaluated and validated models using metrics such as accuracy, precision, recall, and F1 score.
- Deployed machine learning models into production environments and integrated them with applications.
- Created data visualizations and reports using tools like Tableau, Power BI, Matplotlib, and Seaborn.
- Collaborated with cross-functional teams to drive data-driven decision-making processes.

SKILLS

- **Languages:** Python (NumPy, SciPy, Pandas, BeautifulSoup), R, Java, C++, JavaScript, SQL
- **Frameworks & Libraries:** TensorFlow, PyTorch, Keras, Scikit-learn, LangChain, LangGraph, CrewAI, LlamaIndex
- **Machine Learning:** Linear Regression, Logistic Regression, SVM, KNN, Random Forest, XGBoost, K-Means Clustering, Decision Trees, Naive Bayes, Neural Networks, Spacy, LLMs, NLP, PCA, A/B Testing, Probabilistic Networks, Statistics
- **MLOps & DevOps:** Docker, Kubernetes, Jenkins, Terraform, CI/CD Pipelines, Prometheus, Grafana, Model Monitoring, Drift Detection, Model Versioning, Git/GitHub
- **Cloud & Data Services:** AWS (SageMaker, Bedrock, Redshift, S3), GCP(BigQuery, VertexAI), Azure (OpenAI, Databricks), Snowflake, BigQuery, Kafka, Hadoop, Spark, ETL, MongoDB, PostgreSQL, Cassandra, Microsoft SQL Server
- **Visualization Tools:** Power BI, Tableau, Excel (Pivot Tables), PowerPoint

EDUCATION

J B Institute of Engineering and Technology
Bachelor's, Electrical and Electronics Engineering

CERTIFICATIONS

- **Microsoft certified:** Data Scientist Associate
- **Oracle Certified:** Generative AI Engineer
- Introduction to **LangGraph** By **Langchain**