Sal_Minhas

Expert Machine Learning Engineer | Senior Data Scientist | AI & ML Specialist

salminhas32@gmail.com (929) 429-4441

New City, New York 10920 United States

Summary

With a decade of experience in Machine Learning, I offer deep expertise in Python, R, and Java. My skills encompass advanced algorithm development, data preprocessing, feature engineering, and model evaluation. I have worked across various domains, including computer vision, natural language processing, and recommender systems. Proficient in deploying models with Docker, Kubernetes, and AWS, I bring hands-on experience in production environments. My expertise extends to TensorFlow and PyTorch, where I leverage deep learning architectures to enhance model performance. I am skilled in SQL for effective data manipulation and have a solid grasp of distributed computing frameworks like Apache Spark. My achievements include developing NLP systems that significantly improved customer satisfaction and employing graph-based algorithms to enhance recommendation system diversity. My focus is on implementing scalable, efficient machine learning pipelines to deliver impactful solutions.

PROFESSIONAL EXPERIENCE

02/2022 – Present

Lead Data Scientist

47Billion

- Led the development and deployment of advanced AI/ML solutions, including Generative AI, Large Language Models (LLMs), and Retrieval-Augmented Generation (RAG) systems for enterprise clients in healthcare, finance, and telecom.
- Designed scalable AI architectures, delivering solutions like Executive Analytics Chatbots and Multi-Agent Support Bots using GPT-4, LangChain, LangGraph, and vector databases such as Pinecone and FAISS.
- Achieved measurable outcomes such as a 40–60% improvement in query handling efficiency and 20+ hours/week time savings for executive stakeholders through automation and intelligent insights.
- Built persistent, memory-enabled conversational agents using LLaMA 3.2, LangSmith, and deployed via Azure Kubernetes Service (AKS), resulting in enhanced customer engagement and streamlined service delivery.
- Implemented end-to-end MLOps pipelines on Google Cloud Platform (GCP) and Microsoft Azure, enabling automated training, deployment, and monitoring of ML models with cost-effective scalability.
- Developed transformer-based NLP models and predictive analytics workflows that generated actionable insights, supporting faster, more informed decision-making across client operations.
- Worked closely with cross-functional teams including data scientists, designers, and DevOps engineers to drive project delivery and client success in Agile environments.

05/2019 – 01/2022

Senior Data Scientist

Minute

- Led the development and deployment of advanced AI/ML solutions, including Generative AI, Large Language Models (LLMs), and Retrieval-Augmented Generation (RAG) systems for enterprise clients in healthcare, finance, and telecom.
- Designed scalable AI architectures, delivering solutions like Executive Analytics Chatbots and Multi-Agent Support Bots using GPT-4, LangChain, LangGraph, and vector databases such as Pinecone and FAISS.
- Achieved measurable outcomes such as a 40–60% improvement in query handling efficiency and 20+ hours/week time savings for executive stakeholders through automation and intelligent insights.

- Built persistent, memory-enabled conversational agents using LLaMA 3.2, LangSmith, and deployed via Azure Kubernetes Service (AKS), resulting in enhanced customer engagement and streamlined service delivery.
- Implemented end-to-end MLOps pipelines on Google Cloud Platform (GCP) and Microsoft Azure, enabling automated training, deployment, and monitoring of ML models with cost-effective scalability.
- Developed transformer-based NLP models and predictive analytics workflows that generated actionable insights, supporting faster, more informed decision-making across client operations.
- Worked closely with cross-functional teams including data scientists, designers, and DevOps engineers to drive project delivery and client success in Agile environments.

06/2016 – 04/2019 **Data Scientist**

HobbyDB

- Managed a team of four engineers to build a real-time event and dynamic topic detection pipeline on live tweets using NLP techniques across multiple moving time windows.
- Integrated Twitter Live Tweets API to stream and analyze real-time data, extracting named entities to identify and cluster similar events.
- Applied burst detection algorithms within defined time windows to surface trending events and measure audience engagement.
- Achieved 85% accuracy in the entity detection model for identifying bursting events.
- Utilized Dynamic LDA to uncover latent topics within each event, reaching 82% accuracy in topic classification.

Technical Skills

Machine Learning & AI
Proficient in Generative AI, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Neural Networks, and Time Series Forecasting.

Deep Learning
Strong expertise in Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Natural Language Processing (NLP).

Tools & Libraries
Hands-on experience with GPT-, LangChain, TensorFlow, PyTorch and scikit-learn.

Cloud Platforms
Skilled in deploying and managing ML solutions on Microsoft Azure and Google Cloud Platform (GCP).

MLOps & Deployment
Experience in building CI/CD pipelines and deploying ML applications using Flask and Streamlit.

Data Visualization & BI Tools
Proficient in Tableau and Power BI for insightful data visualization and reporting.

Programming Languages
Proficient in Python and R, with expertise in TensorFlow and Keras.

Big Data Technologies
Experience working with Apache Spark, Hadoop, and Kafka for scalable data processing and streaming.

CERTIFICATES

- Google professional Machine Learning
- Engineer Azure Data Science Associate
- Google Professional Data Engineer
- Google Data Analytics Professional
- GA Individual Qualification Certified
- MS Azure AI Fundamentals
- L2 Advanced Proficiency in KNIME Certified