Vedant Sahai

814-769-0201 vedantsahai18@gmail.com linkedin.com/vedantsahai18 github.com/vedantsahai18

Experience

Julep AI Inc. Oct 2024 – Present
Machine Learning Engineer *New York City, NY*

- Increased User Engagement by 15% within a month of launch by developing the ”Browser Use” feature, which enables automation of browser tasks through AI-driven workflows.
- Optimized the workflow performance to accelerate executions from hours to under 45 min for 2.5k runs by migrating the database from CozoDB to TimescaleDB.
- Accelerated Julep stars from 900 to 4.4k by creating 10+ detailed, templates for Agentic Workflows, improving developer engagement.
- Enhanced workflow capabilities by 20% through the development of an integration service that enables interaction between Agents and external tools.

JP Morgan Chase & Co. Jun 2023 – Aug 2023
AIML Summer Associate *Wilmington, DE*

- Improved the Transaction Detection Rate by 100-150 basis point by implementing the Online ML XGBoost algorithm for the Transaction Risk model (TRS).
- Attained 90% accuracy by programming a TabNet-based Deep Neural Network as a challenger for the TRS XGBoost model.
- Ensured 95% code coverage by implementing a PyTest-based testing framework for the TRS Feature Engineering codebase.

Plexflo Oct 2021 – Jul 2022
ML Engineer *Mumbai, India*

- Developed Evidence, a Meter Data Management & Analytics (MDMS) software with a latency of less than 90ms, by leveraging ITRON, Sensus Xylem, and Siemens data streams, powered by AWS, Apache Flink, and a custom ML model.
- Achieved an F1 score of 85% for a non-intrusive load monitoring model by leveraging a Variational Autoencoder.
- Enhanced rooftop solar assessment accuracy by 30%, using Mask R-CNN and Geo-Spatial Image Processing techniques.

Sync Energy AI Jul 2020 – Sept 2021
ML Research Intern *Mumbai, India*

- Optimized the power outage extraction, resulting in a 40% faster response, by deploying an AWS Lambda-Python-REST API.
- Improved accuracy to 83% in estimating the locations of utility poles from Google Street View images by employing Mask R-CNN and Image Processing.
- Boosted research capabilities by generating Neo4J-based knowledge graphs from research papers on wildfires and their effects.

Technical Skills

Languages: Python, C, JavaScript, Java, LaTeX
ML Frameworks: PyTorch, TensorFlow, RASA, Scikit, XGBoost, HugginFace, NLTK, Spacy, LlamaIndex, Langchain, Autogen
Databases: MySQL, PostgreSQL, MongoDB, Timescale
Cloud: Amazon Web Services [Ec2, S3, Lambda, API Gateway, Sagemaker, IAM]
Technologies: React.JS, Flask, FastAPI, PyTest, GitHub Actions, Elasticsearch, Git, Docker, PySpark, Grafana, Sphinx, Temporal

Education

Pennsylvania State University May 2024
Master of Science in Computer Science & Engineering ***(GPA: 3.75 / 4.00)*** *University Park, PA*

- **Teaching Assistant:** CMPSC 132: Programming and Computation II: Data Structures
- **Relevant Coursework:** Computer Vision II, Data Structures & Algorithms, Natural Language Processing, Introduction to Deep Learning, Machine Learning and Algorithmic AI, Vision & Language, Computer Security

University of Mumbai May 2021
Bachelor of Engineering in Computer Engineering ***(CGPA: 9.58 / 10.00)*** *Mumbai, India*

Projects

Datacertus May 2022
Reacts. JS, AWS APIs, Python, Docker, AWS ELB, AWS Lambda, DynamoDB, Scikit, RASA, PyTorch, Keras, Scikit-Learn

- Trained a BERT summarization model to summarize updates on disaster information with 75% ROGUE-L.
- Reduced update frequency by over 50% by developing a pipeline using Lambda, Selenium, and the BERT model to track natural calamities.
- Made data processing 10% faster by integrating data pipelines using Lambda, S3, and API Gateway-based trigger events.

Conversational AI for Secure Healthcare Assistance May 2021
Docker, Node.JS, MongoDB, RASA, Blockchain, Python, Express.JS, jQuery, AJAX, Vault, Nginx

- Handled over 10,000 EHR records concurrently by architecting a Docker-NodeJS-MongoDB-Vault-based software.
- Attained the intent prediction confidence up to 95% by developing and integrating a RASA-powered therapy chatbot.
- Integrated the BigchainDB + IPFS blockchain database with the RASA for behavior analysis and secure medical record storage.

Publication

Vedant S., Jason D., Mayank S., Mahendra M., Dhananjay K. (2021) Leveraging Deep Learning and IoT for Monitoring COVID19 Safety Guidelines Within College Campus. In: Garg D., Wong K., Sarangapani J., Gupta S.K. (eds) Advanced Computing. IACC 2020. Communications in Computer and Information Science, vol 1367. Springer, Singapore. https:doi.org10.1007978-981-16-0401-0_3