Suraj Bilgi

AI/ML Engineer

• suraj.b@themailpad.com •+15512576126 • LinkedIn • GitHub • Portfolio

SUMMARY

- AI/ML Engineer with 4+ years of experience in designing, deploying, and scaling end-to-end machine learning solutions across NLP, Generative AI (LLMs), Computer Vision, and advanced analytics using deep learning and statistical modeling techniques.
- Expert in building supervised and unsupervised models using Linear/Logistic Regression, Naive Bayes, SVM, KNN, Decision Trees, K-means, Random Forest, Bagging, and Gradient Boosting for predictive analytics and pattern recognition.
- Well-versed in training and fine-tuning deep learning architectures including CNNs, RNNs, LSTMs, Transformers, and LLMs (GPT, BERT, LLaMA) using PyTorch, TensorFlow, and Keras for real-world NLP and vision-based applications.
- Advanced Python programming for feature engineering, data preprocessing, and visualization using Pandas, NumPy, Scikit-learn, SciPy, OpenCV, NLTK, Matplotlib, Seaborn, and Ggplot2 to support experimental workflows and insights.
- Skilled in ML infrastructure using AWS (EC2, S3, SageMaker, IAM, SQS, SNS), Google Cloud (Vertex AI), and Azure DevOps; experienced with CI/CD pipelines, Docker, Terraform, and MLOps practices to manage the full ML lifecycle.

SKILLS

Language/ IDE's: Python, MATLAB, Jupyter Notebook, Google Colab, VS Code, SSMS
Machine Learning: Linear Regression, Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), Naive Bayes, A/B Testing, Feature Engineering, Model Evaluation & Validation
Deep Learning: CNN, RNN, LSTM, NLP, Large Language Models (LLM), LangChain, Hugging Face Transformers (BERT, GPT-3)
Cloud/Visualizations: AWS (EC2, S3, Lambda, API Gateway, CloudWatch, CodeDeploy), GCP (Vertex AI, Google Cloud Storage, AI Platform Pipelines), Tableau, Power BI, Excel Advanced Analytics
Statistical Techniques: Hypothesis Testing, Data Visualization, Data Modelling, A/B Testing, Predictive Analytics, Statistical Inference
Packages and Frameworks: NumPy, Pandas, Matplotlib, Scikit-learn, Seaborn, TensorFlow, Keras, NLTK, XGBoost, PyTorch, LightGBM
Database and Tools: SQL Server, MySQL, PostgreSQL, Redis, Neo4j, Apache Airflow, MLflow
Certifications: Generative AI Engineering with LLMs Specialization, AWS Certified Solutions Architect – Associate

PROFESSIONAL EXPERIENCE

Tracker Groups LLC **May 2024 – Present | Albany, NY**
Machine Learning Engineer

- Engineered a high-fidelity 3D reconstruction pipeline for indoor house environments using LiDAR-derived point cloud data, integrating Open3D and MeshLab for surface reconstruction and semantic segmentation of architectural elements.
- Developed and containerized backend microservices for preprocessing, mesh generation, and rendering using Python, deployed via Docker and orchestrated on AWS ECS; optimized reconstruction throughput by 40%.
- Designed and deployed a FastAPI-based ML inference service to serve XGBoost models in real-time; integrated JWT authentication for secure access and leveraged AWS SQS to handle asynchronous prediction requests, reducing latency by 30% and enabling scalable model consumption in production.
- Executed development of a model dashboard using Python (Flask), integrated with MLflow and PostgreSQL to track runs, hyperparameters, and model versions. Enabled collaboration between data science and MLOps teams, improving reproducibility and traceability by 40%.
- Constructed and trained deep learning models using CNNs and RNNs to automate document classification and object detection across enterprise workflows, increasing operational throughput by 40% and enabling scalable AI adoption
- Architected and fine-tuned NLP models using BERT and GPT-3 via Hugging Face Transformers for sentiment analysis and intent detection, improving customer feedback classification accuracy by 20% and reducing escalation cycles
- Built custom TensorFlow pipelines to automate model training, hyperparameter tuning, and batch inference, reducing model training time by 60% and enhancing deployment consistency across scalable production workflows.

Ineuron Intelligence Private Limited **Jul 2022 – Aug 2023 | Bangalore, India**
AI/ML Developer

- Architected a CNN-based image classification pipeline using TensorFlow and OpenCV to detect diabetic retinopathy in high-resolution retinal scans, achieving 91% AUC and improving early-stage detection in nationwide trials.
- Crafted a scalable NLP workflow using spaCy and regex-based extractors to mine clinical symptoms from unstructured EHR text, leading to a 47% increase in structured patient data and enhancing downstream analytics.
- Formed a Random Forest-based risk prediction model to identify patients likely to develop adverse drug reactions during post-trial follow-ups, reducing clinical error rate by 19% over 6 months.
- Implemented a LLM-powered summarization module using LangChain with GPT-4 APIs to convert complex diagnostic reports into physician-ready insights, improving clinical efficiency across 3 departments.
- Created a statistical modeling dashboard using Python's statsmodels and seaborn to identify trends in patient response data across 8 hospitals, helping clinical staff fine-tune treatment procedures.
- Executed custom TensorFlow pipelines to automate model training, hyperparameter tuning, and batch inference, reducing training time by 60% and enhancing consistency across scalable production workflows.

Resolute AI Software Private Limited **Jan 2020 – Jul 2022 | Bangalore, India**
Machine Learning Engineer

- Formed real-time AI services using OpenCV and LSTMs to detect abnormal user behaviors in video streams, reducing manual monitoring needs by 55% in fraud detection and compliance scenarios
- Migrated legacy ML models to AWS SageMaker with Docker containers and implemented model endpoints using API Gateway and Lambda, reducing deployment time by 50% and enabling seamless scaling
- Led exploratory data analysis and implemented decision tree models using scikit-learn to identify customer churn indicators, enhancing prediction accuracy by 20% and informing targeted retention strategies across multiple business units.
- Established and deployed predictive models leveraging gradient boosting and ensemble methods, achieving 88% forecasting accuracy for sales trends; this enabled inventory planning and reduced stockouts by 15%, saving significant operational costs.
- Fabricated and automated scalable data pipelines for complex feature engineering, data cleaning, and transformation using Python and SQL, improving data processing speed by 35%, which enhanced model training efficiency across multiple projects.
- Applied advanced NLP techniques, including sentiment analysis and topic modeling, on large volumes of customer feedback data, generating insights that drove a 22% increase in customer satisfaction and informed product roadmap prioritization.

EDUCATION

Master of Science in Applied Artificial Intelligence
Stevens Institute of Technology, Hoboken, NJ