Gaurav Nandode

gsnandode@gmail.com ❖ LinkedIn ❖ Website ❖ (+91) 8172961868

Data Scientist with 9 years of experience developing data-driven algorithms, mentoring junior data scientists and leading data science initiatives. Proven track record in optimizing logistics, econometric modeling, and sales forecasting for global companies.

WORK EXPERIENCE

Coditas **July 2022 – Present**

Lead Data Scientist

Spearhead data science initiatives as an individual contributor and mentor junior data scientists, fostering their professional growth.

Scalable Product Review Insights Engine

- Developed a production-grade RAG application for e-commerce product intelligence, ingesting Google Shopping data and reviews via SerpAPI, transforming them into an embedded corpus with SentenceTransformers, and indexing in ChromaDB for low-latency retrieval.
- Built a Streamlit interface that enables grounded Q&A using Gemini 1.5 Flash, constraining responses to retrieved context to reduce hallucinations.
- Implemented session caching, de-duplication, and rate-limit safeguards to ensure reliability and responsive UX. Delivered actionable, review-driven insights across up to 40 products per query.

Efficient News Classification with LoRA Fine-Tuning

- Fine-tuned a DistilBERT model on the AG News dataset using LoRA for efficient and scalable text classification, achieving robust accuracy and F1 scores.
- Automated the workflow for data preprocessing, training, evaluation, model saving/loading, and inference, including comparison with zero-shot classification.
- Integrated the solution with Hugging Face pipelines for easy deployment and real-world text classification tasks.

Ship Fuel Optimization using Mathematical Programming and Econometrics

- Formulated and solved a Linear Programming (LP) model to optimize bunker fuel planning for a bulk liquid logistics fleet, minimizing fuel costs while adhering to operational constraints.
- Integrated behavioral analytics and econometric modeling to capture real-world fuel consumption patterns based on ship movement, port conditions, and historical refueling behavior.
- Designed SQL-based logic to enhance Remainder-on-Board accuracy at ports.

Stress Prediction Model Using Physiological Signals from Wearable Devices

- Built a semi-supervised stress prediction pipeline leveraging multi-sensor physiological data (BVP, EDA, skin temperature) from wearable devices. Implemented a masked autoencoder for self-supervised pretraining on unlabeled biosignal windows to learn robust temporal representations. Fine-tuned the encoder on limited labeled data for stress classification using tree-based models.
- Extracted domain-relevant features including heart rate variability (RMSSD, SDNN), respiration rate, and galvanic skin response metrics. Employed Leave-One-Subject-Out Cross-Validation (LOSOCV) to ensure subject-independent generalization.
- Additionally, derived strain and recovery scores, computed sleep debt, and incorporated circadian alignment features from time-windowed sensor data..

AI-Powered Healthcare Logistics Optimization System

- Developed an integrated AI system for healthcare logistics, combining real-time fleet assignment with geospatial demand forecasting using tree-based models. Enabled precise, location-based predictions, automated retraining, and dynamic van allocation to optimize resource utilization and service delivery.

Customer Prediction from Leads

- Improved lead conversion for a shipping client by developing a random forest classifier model to predict customer from a lead and deploying it on Azure ML Studio.

Sales Forecasting

- Predicted sales forecasts for a container shipping leader using tree-based regression models, improving sales planning.

Profit Margin Estimation
- Developed real-time gross margin estimations for an e-commerce leader in children's clothing.
- Built Looker dashboards for ML predictions, streamlining decision-making processes. Segmented customers using Recency-Frequency-Monetary scores, enhancing marketing strategies.

AppZen **June 2021 – July 2022**

Senior Data Scientist
- Automated invoice processing, extracting key entities using a Random Forest Classifier.
- Extended multilingual support by training a Neural Network model with multilingual embeddings, enhancing scalability.

Talentica **November 2018 – June 2021**

Data Scientist
- Enhanced crash log analysis by implementing MinHash and Locality Sensitive Hashing, improving issue detection accuracy by 30%.
- Developed methods to extract key indicators from crash logs, including critical log lines, key function calls, register values, action keywords, and crash keywords using advanced Regex techniques.
- Created unsupervised clustering pipelines to organize unclassified logs into actionable categories, leveraging features like Word2Vec, TFIDF, and Jaccard Similarity.
- Built a custom K-means clustering algorithm optimized with Jaccard scores to improve clustering efficiency.
- Reduced data processing times by 25% through optimization of PySpark pipelines, streamlining data transformation and analytics workflows.
- Designed and implemented end-to-end architecture for data-driven products, ensuring robust testing and efficient deployment pipelines.
- Contributed to log analysis and categorization strategies, enabling faster debugging and better insights for development teams.

IBM **August 2016 – November 2018**

Data Specialist

Cognitive Signature Capture and Verification
- Implemented thresholding and segmentation techniques to properly form segments over signatures.
- Worked on extraction of text and signatures within table format by detection of lines using Hough Lines.
- Extracted crucial features from an ensemble of HOG (Histogram of Gradients) and LBP (Local Binary Patterns) features for signature and text segments using PCA (Principle Component Analysis)
- Trained a SVM (Support Vector Machine) classifier on crucial features
- Worked on NER (Named Entity Recognition) and Class Recognition for the authorised signatories

C3 - Cognitive Content Collator
- Played a crucial role in implementing an algorithm for image segmentation.
- Trained a SVM classifier to classify image and text segments using HOG features.
- Tried and tested Google and IBM Speech-to-Text conversion tools for the audio chunking part.
- Worked on silence detection to detect pauses in the video and speaker label identification.
- Extracted POS (Parts of Speech) tags from text and script chunks in videos.

Auto Insurance Damage Assessment
- Damage extent classifier based on the confidence score of 3 classes low, moderate and heavy damage.
- Built a separate classifier for damaged part to tell which part of the car is damaged.
- Deployed demo in NodeJS environment on IBM Bluemix cloud

EDUCATION

IIT Kanpur **July 2014 - August 2016**

M. Tech., Electrical Engineering

Thesis - Unbounded PML-based Forward Modelling of a 2D GPR Tomography Problem
Optimization and Implementation of a finite-element-method (FEM) based formulation for a Helmholtz-equation modeled forward problem of GPR (Ground Penetrating Radar) tomography using an unbounded function as a perfectly matched layer (Link for thesis)

VNIT Nagpur **July 2009 - July 2013**

B. Tech., Electronics and Communication Engineering

PATENT

- Log analysis in Vector Space - US11138059B2 https://patents.google.com/patent/US11138059B2/en

SKILLS

- **Programming Languages & Tools**: Python, SQL, TensorFlow, Scikit-Learn
- **Machine Learning & AI**: Random Forest, SVM, Neural Networks, Tree-Based Models, LLMs, Retrieval-Augmented Generation (RAG)
- **Data Engineering**: PySpark, Azure ML Studio, Looker
- **Specializations**: Image Processing, Computer Vision, NLP
- **Tools**: **Cloud Platforms**: Azure, AWS, GCP, Snowflake

INTERESTS

- Weightlifting, hiking, and adventure sports, which demonstrate discipline and problem-solving abilities.