Roshil Shrestha Bana

ML Engineer Associate

✉ roshil.bana@fusemachines.com | Linkedin | GitHub

Machine Learning Engineer Associate with 4+ years of experience and a Bachelor's in Computer Science and Information Technology. My expertise lies in delivering end-to-end machine learning solutions, particularly in Object Detection (YOLO), Text Recognition, and Natural Language Processing (NLP). I possess a strong foundation in Generative AI, Knowledge Graph construction, and leveraging Large Language Models (LLMs), with hands-on experience utilizing frameworks like LangChain to build sophisticated AI applications. I am proficient in major cloud platforms, including GCP, AWS, and Azure, and well versed in essential ML tools such as Python, TensorFlow, Keras, and Scikit-learn, Google ADK, Neo4j . My analytical skills extend to both text and object data, enabling me to develop and deploy high performing and accurate models. I thrive in cross-functional teams, bringing a collaborative, adaptable, and continuous-learning mindset with strong communication skills, to deliver impactful results. Strong experience translating complex data and business requirements into scalable ML solutions, with a focus on model performance, explainability, production deployment, and measurable business impact.

PROJECT EXPERIENCE

Early VIP & High-Potential VIP Player 360 | Entain

March 18 – August 28, 2025

- Built an end-to-end **Player 360 ML and Graph Analytics platform** for Early VIP and High-Potential VIP detection across **3, 7, 14, and 21-day behavioral windows**, enabling proactive CRM targeting.
- Designed scalable **Google BigQuery-based Player 360 pipelines** integrating transactional, behavioral, derived, product-mix, and graph data across player identity and behavioral entities including **IP Address, Device, Session, Affiliate, Tracker, Country, and Deal**.

- Engineered advanced **behavioral, monetary, temporal, engagement, and cross-product features** covering deposits, stakes, active days, sessions, NGR, product mix, and player activity patterns.
- Developed and optimized **XGBoost, Homogeneous GNN, and Heterogeneous GNN models**, applying feature selection, class-imbalance handling, threshold optimization, and **PR-AUC-focused evaluation** for early VIP prediction.
- Engineered graph intelligence features including **connectivity, neighborhood statistics, propagation signals, behavioral clustering, and shared-entity patterns**, with relationship filtering to reduce graph noise and improve signal quality.
- Applied **SHAP-driven feature selection and explainability** to identify high-value predictors, eliminate redundant/low-value features, and improve model interpretability.
- Built **leakage-safe, time-aware training and inference pipelines** with frozen feature schemas, reproducible preprocessing artifacts, model checkpoints, validation checks, and training/inference consistency controls.
- Integrated a **Gemini-powered explainability layer** that uses the top SHAP features generated during inference to produce human-readable explanations for why a player was classified as **Early VIP or High-Potential VIP**.
- **Technologies:** Python, Google BigQuery, SQL, Pandas, NumPy, XGBoost, Scikit-learn, PyTorch, PyTorch Geometric, GAT, Homogeneous & Heterogeneous GNNs, SHAP, Explainable AI, PR-AUC Optimization, GPU Computing.

Demand Forecasting System Sanmina

- Built an end-to-end API-driven demand forecasting pipeline enabling item-level forecasts using either BigQuery enterprise data or client-uploaded datasets.
- Developed a modular FastAPI service supporting both automated cloud-based inference and flexible file-based forecasting workflows.
- Designed scalable preprocessing pipelines to transform demand, MDS, capacity, WIP, inventory, CTB, and calendar data into clean, continuous ITEM × WEEK time series.
- Integrated hierarchical reconciliation to align item-level weekly forecasts with quarterly business demand plans.
- Deployed MLForecast (XGBoost) models for large-scale item-level forecasting with configurable forecast start dates and dynamic horizons.
- Built an automated report generation engine producing structured, client-ready Excel reports combining forecasts with capacity, inventory, and supply pipeline insights.
- Enabled fully automated enterprise inference via secure BigQuery integration, removing the need for manual data handling.
- Designed a production-grade ML architecture with clearly separated modules for data ingestion, preprocessing, forecasting, reconciliation, and reporting.
- **Technologies:** Python, FastAPI, Pandas, BigQuery, XGBoost, MLForecast, Time-Series Forecasting, Hierarchical Forecasting

Persistent Cross-Session Memory for AI Agents (Proof of Concept)

- Integrated PostgreSQL with Google ADK's session framework to enable persistent, queryable memory storage for AI agents, allowing context-aware responses across multiple user sessions.
- Transformed agents from stateless chatbots into contextual assistants capable of recalling past interactions, user preferences, and instructions, enhancing perceived intelligence and engagement.
- Developed structured datasets and sample conversation questions to evaluate agent reasoning, memory retention, and multi-session capabilities, covering temporal reasoning, open-domain knowledge, adversarial scenarios, multi-hop, and single-hop QA.
- Built a mock Demand Forecasting Agent simulating real-world user sessions, demonstrating retrieval and utilization of past interactions, validating context continuity and memory effectiveness in realistic scenarios.
- Tested and validated the memory system for robustness, including knowledge graph creation with and without ontology, session-specific isolation of UserID, Engine Name, and Project Name, and long-term memory consistency across sessions.
- Enhanced user experience by delivering personalized, consistent, and context-aware AI interactions, increasing engagement, retention, and overall satisfaction.
- **Technologies Used:** PostgreSQL, Google ADK, Python, LLMs, Knowledge Graphs(Neo4j), Prompt Engineering

Image Generation with Gemini (Nano Banana) – Fashion Mood Boards (Proof of Concept)

September 24 – October 6, 2025

- Designed and implemented a fashion-focused image generation pipeline using Google Gemini (Nano Banana) to create high-quality, editorial-style fashion visuals from textual prompts and reference images.
- Enabled rapid prototyping of fashion concepts without physical photoshoots, significantly reducing design iteration time and production costs.
- Built automated workflows to generate fashion mood boards and editorial compositions by combining multiple reference images with user prompts for fast concept visualization.
- Implemented dynamic visual variation generation (poses, expressions, camera angles, environments, lighting, and accessories) while preserving core fashion design consistency.
- Accelerated creative ideation for fashion collections by enabling fast experimentation and reducing manual design effort for designers and marketing teams.
- Supported multi-image inputs to generate collages and mood boards, enhancing flexibility for creative exploration and marketing visualization.
- Demonstrated prompt-free image generation using reference images alone, enabling fully automated workflows for fashion design and branding use cases.

- **Technologies Used:** Google Gemini (Nano Banana), Python, Multimodal LLMs, Image Generation, Prompt Engineering

Sales Copilot (Proof of Concept)

- Developed an intelligent sales copilot system leveraging HubSpot APIs to streamline deal management, CRM interaction, and sales pipeline optimization.
- Created autonomous agents to perform tasks like deal creation and information retrieval from HubSpot in real-time using CrewAI flows.
- Explored and integrated HubSpot's API endpoints for managing deals, companies, CRM data, pipelines, and deal stages, enabling seamless data retrieval and automation.
- Constructed dynamic knowledge graphs using Neo4j to represent interlinked CRM entities (deals stages, deals, companies, notes, contacts, emails), enhancing contextual understanding and supporting intelligent querying.
- Implemented a question-answering system powered by ChatGoogleGenerativeAI to enable natural language querying over CRM and deal-related data.
- Built a recommendation engine leveraging historical deals, notes, and email interactions to suggest actionable insights and provide summaries of deals and all associated activities.
- Enhanced decision-making by surfacing relevant patterns and next-step recommendations based on previous sales activity.
- Applied prompt engineering techniques to refine question-answering performance and improve knowledge graph-based query resolution.
- **Technologies Used:** Python, HubSpot API, Neo4j, ChatGoogleGenerativeAI, LangChain, google-adk, LLMs, Graphiti.

Project Management System Agent (Proof of Concept)

- Designed and implemented a robust recommendation system that intelligently integrates and analyzes data from multiple Monday.com boards, significantly enhancing decision making through cross board analysis and providing holistic project summaries.
- Created dynamic sample email templates for project-related communications, standardizing team interactions and improving overall productivity.
- Refined the logic for multi-board data combination to accurately merge project metrics and generate cohesive insights across all teams.
- Focused on the logic for combining data from multiple Monday.com boards, ensuring tables and insights are merged accurately.

- Enhanced plot quality through prompt tuning and integrated Graphiti for scaling and visualizing Knowledge Graphs.
- Tested and validated the video feature module, ensuring that the system could successfully process and confirm visual inputs from project-related clips.
- Explored modern tools like CopilotKit and CoAgents Starter CrewAI Flows, including full codebase review and local environment setup to assess applicability for future system expansions.
- **Technologies Used:** Python, OpenAI, Retrieval Augmented Generation (RAG), CopilotKit, CoAgents Starter CrewAI Flows, Git , REST APIs, Graphiti.

Patient Monitoring and Pose Estimation System – Tampa General Hospital (Proof of Concept)

- Developed an AI-powered patient monitoring system to track patient movements and postures in real-time, enhancing healthcare efficiency and safety
- Implemented YOLOX for patient detection and RTMO (Real-Time Multi-Person Pose Estimation) to accurately track patient postures and movements in real-time.
- Leveraged Florence, a foundation model, to generate context-aware prompts, facilitating automated reporting and intelligent alerts for healthcare professionals.
- Designed and deployed a robust tracking system capable of detecting patient movements across different hospital environments, ensuring enhanced patient care and safety.
- Integrated ViTPose, a vision transformer-based pose estimation model, to enhance tracking accuracy, ensuring reliable monitoring of patient positioning and mobility.
- Optimized post-processing pipelines to improve system efficiency and minimize false detections, ensuring seamless integration with hospital workflows.
- Led the annotation process using Florence for prompt generation, streamlining dataset creation for improved AI model accuracy.
- **Technologies Used:** Python, Florence, YOLOX, RTMO, ViTPose, OpenCV, PyTorch

Document Verification and Classification System (Proof of Concept)

- Developed a document verification and classification application using Streamlit, integrating NLP models for dispute code prediction and entity extraction.
- Implemented a fine-tuned RoBERTa model for sequence classification, achieving accurate dispute code predictions based on user comments.
- Utilized a zero-shot classification pipeline with BART to classify various document types, enhancing the application's ability to handle diverse inputs.
- Integrated GLiNER for Named Entity Recognition (NER) to efficiently extract relevant entities.
- Created a user-friendly interface allowing users to upload PDF documents and input queries, streamlining the document verification process

- Enabled text extraction from PDFs using PyMuPDF, facilitating comprehensive document analysis.
- Built a Streamlit-based application for document verification and classification, incorporating advanced NLP models for dispute code prediction and entity extraction.
- **Technologies Used**: Python, Hugging Face Transformers (RoBERTa, BART), GLiNER, PyTorch, Streamlit

Object detection and tracking Darkhorse Soccer Analytics

- With the advent of ML and AI, object detection and tracking have become very popular. We created a system to detect and track the soccer ball, players, and jersey numbers of the players in short and long video clips.
- Collaborate with team members to make a system that detects and tracks the soccer ball, players, and detects jersey numbers in real game videos. We trained the model using images from real life or frames from actual game videos, then made a system to find and track soccer balls, players, and jersey numbers.
- Led an 8-member annotation team, ensuring high-quality labeled datasets using tools like Roboflow and CVAT, significantly enhancing model performance.
- Designed and implemented a ball detection system leveraging SAHI (Slicing Aided Hyper Inference) with 6-slice and 8-slice configurations to optimize detection in high-resolution imagery.
- Utilized state-of-the-art (SOTA) models such as RTDETR, YOLOv8, and YOLOX for detecting and tracking balls.
- Used algorithms like Kalman filters to track the ball's movement.
- Achieved real-time performance at 30 FPS, balancing accuracy and scalability for sports analytics.
- Designed robust tracking pipelines with optimized post-processing strategies for small object tracking, improving reliability and system efficiency.
- **Technologies Used:**

  Roboflow, Git, Python, AWS, OpenCV (cv2), Microsoft Azure, RTDETR, YOLOv8, and YOLOX, AWS

CERTIFICATIONS

- Machine Learning
- Deep Learning
- Microdegree in Artificial Intelligence

- Microdegree in Computer Vision
- Information Security Awareness Training
- Features and Boundaries
- Introduction to Computer Vision and Image Processing
- Object Tracking and Motion Detection with Computer Vision
- Advanced Computer Vision with TensorFlow

WORK EXPERIENCE

ML Engineer Associate, Fusemachines: **2023 November - Present**

- Built an end-to-end API forecasting pipeline using BigQuery or uploaded datasets, featuring time-series preprocessing, capacity-aware forecasting, hierarchical MLForecast (XGBoost) models, and automated Excel reporting with supply, inventory, and capacity insights.
- Developed an intelligent Sales Copilot system leveraging HubSpot APIs and Neo4j knowledge graphs to streamline deal management, optimize sales pipelines, and enable natural language querying and recommendation generation using ChatGoogleGenerativeAI, google-adk and LangChain.
- Designed and implemented a robust Project Management System that integrates and analyzes data from multiple Monday.com boards for enhanced decision-making and holistic project summaries, while exploring modern tools like CopilotKit and CoAgents Starter CrewAI Flows.
- Built a recommendation system leveraging historical deals and email notes to generate actionable sales insights, using LangChain and LLM-based prompt engineering.
- Developed an AI-powered Patient Monitoring and Pose Estimation System for Tampa General Hospital, utilizing YOLOX, RTMO, and ViTPose for real-time patient movement and posture tracking, and Florence for context-aware prompt generation and automated reporting.
- Built a Document Verification and Classification System using Streamlit, integrating fine-tuned RoBERTa for dispute code prediction, BART for zero-shot document classification, and GLiNER for Named Entity Recognition.
- Developed computer vision pipelines for soccer analytics to detect and track players, soccer balls, and jersey numbers using SAHI, RT-DETR, YOLOv8, and YOLOX; led an 8-member annotation team to support model development and dataset quality.
- Built an Early VIP & High-Potential VIP Player Detection Player360 platform using BigQuery, XGBoost, Homogeneous and Heterogeneous GNNs, SHAP, and Gemini-based explainability to identify high-value players across 3, 7, 14, and 21-day windows.

Ml Engineer Associate Trainee, Fusemachines

2023 Aug- 2023 November

- **Preliminary Courses**: Took & Completed Courses Named 'Anti-Harassment' and 'Introduction to Computer Vision and Image Processing'.
- **Joining the Computer Vision Upskilling**: Joined the upskilling project 'Computer Vision Upskilling' and built a decent foundation on concepts & techniques of Computer vision including libraries like python, OpenCV (cv2) and Matplotlib.
- **Soft Skill Development**: Built skills on self-learning, presentation skills, documentation skills & working with the team during 'Computer Vision Upskilling'. Learn to use AWS services like s3 bucket, EC2, lambda function and so on.
- **Hard Skill Development:**: Learned about and applied the principles and rules of Python Code Naming & Standardization, Contributed in Code Cleaning & Optimization making the Code base Efficient & Effective, Understood about various principles and tech-niques of Error Analysis.
- **Data annotation**: For fuse Extract and Darkhorse

SKILLS

Programming Languages: Proficient in Python, C/C++

Libraries & Frameworks: Experienced with NumPy, Pandas, scikit-learn, Matplotlib, Seaborn, Keras, TensorFlow, PyTorch, LangChain

Web Frameworks: Hands-on experience with Flask and FastAPI for building RESTful APIs

Databases: Proficient in using MySQL and MongoDB for data storage and retrieval, Neo4j for knowledge graph implementation

Cloud Platforms: AWS (EC2, S3), Microsoft Azure, Google Cloud Platform (GCP)

Version Control: Git – for source code management and collaboration

Machine Learning Techniques: Sentiment Analysis, Image Classification, Text Classification, Object Detection, Transfer Learning

Natural Language Processing (NLP): HuggingFace Transformers, Named Entity Recognition (NER), Text Summarization, NLTK, Spacy, BERT, GPT, HuggingFace, Langchain

Time-Series & Forecasting: MLForecast, XGBoost, Time-Series Feature Engineering, Hierarchical Forecasting

Computer Vision (CV) Tools: OpenCV, Pillow – for image processing and analysis

Model Deployment & MLOps: Docker containerization, API-based model serving, cloud deployment on AWS/GCP

Agentic AI Framework: LangChain, CrewAI, google-adk

Tools & Platforms: Jupyter Notebook, VS Code, Google Colab

TRAININGS

- AI For Everyone (Coursera)
- Structured Query Language
- Machine Learning Specialization by Andrew Ng
- Dive into Deep Learning
- Python Course Edureka
- Introduction to Deep Learning
- Introduction to Computer Vision and Image Processing
- Generative AI
- Kalman Filters
- AI Agents in LangGraph
- Understanding NLP and LLMs

EDUCATION

- **Bachelors of Science in Computer Science and Information Technology (BSc. CSIT)**
  Alliance for InnovativeManagement and Social Studies (AIMS) College Biratnagar, TU
  2023

- **Higher Secondary (+2) in Science**
  Shikshadeep Higher Secondary School of Science, HSEB
  2018