Binit Kumar Jha

+91 7016153673 binitjha2000@gmail.com linkedIn : Binit Kumar Jha Github :Binitjha2000

Education

BITS Pilani Expected July 2025
M.Tech in Artificial Intelligence and Machine Learning *Pilani, India*

Sir M. Visvesvaraya Institute of Technology May 2021
B.E in Electronics and Communication (CGPA: 7.82 / 10) *Bangalore, India*

Experience

Cognizant Technology Solutions Sep 2021 – Present
Associate *Bengaluru, India*

- Orchestrated AI agents via Google Agentspace to proactively detect and prioritize billing discrepancies .
- Developed automated pipelines on GCP, utilizing BigQuery and Dataflow, to integrate diverse billing, usage, and contract data for AI-driven anomaly detection and revenue leakage prioritization.
- Designed and implemented a multi-layered AI/ML detection framework, integrating rule-based systems with unsupervised anomaly detection for proactive identification of billing discrepancies.
- Applied Time Series Analysis for pattern recognition and utilized supervised classification models (Logistic Regression, SVM, Gradient Boosting) to categorize and prioritize detected anomalies.
- Designed and implemented Agentic AI workflows within Agentspace for intelligent dispute communication analysis and automated invoice line item matching solutions.
- Project Leadership: Led project lifecycle, ensuring handover with documentation/training.
- Architected Azure infrastructure for a Retrieval-Augmented Generation chatbot, leveraging Azure OpenAI and Azure Cognitive Search.
- Engineered the integration and maintenance of Azure Blob Storage as the chatbot's knowledge repository, optimizing it for efficient indexing and semantic retrieval via Azure Cognitive Search.
- Built Python pipelines for RAG, integrating Azure OpenAI embeddings, Azure Cognitive Search vector capabilities, and contextual data retrieval.
- Designed/implemented scalable GCP infrastructure (Terraform) for 50+ ETL pipelines supporting ML datasets.
- Real-time ML Data (Kafka/BigQuery): Integrated Kafka streams into BigQuery (Pandas) for real-time ML data.
- Automated Data Pipelines (GCP/BQ): Developed automated ETL pipelines (GCP, Terraform, BigQuery) for scalable ML data processing (CI/CD).
- Automation for Efficiency (Python): Achieved 30-40% manual effort reduction via Python automation for various workflows.

The Sparks Foundation May 2020 – Jul 2020
Data Analyst *Remote*

- Developed a predictive model to forecast sales for a retail store using historical sales data.
- Performed data cleaning and preprocessing to handle missing values and outliers.
- Utilized regression techniques and time series analysis to predict future sales trends.
- Presented findings and recommendations through interactive visualizations and reports.

Projects

AgentAI Support Chatbot | *NLP, RAGs , GENAI*

- Developed an intent-driven chatbot (Python, JSON, NumPy) to automate task execution by accurately understanding user requests.
- Engineered a Retrieval-Augmented Generation (RAG) system using LangChain, Hugging Face Transformers and FAISS to provide precise and contextually relevant responses.
- Developed a semantic knowledge base using Hugging Face embeddings and a FAISS vector store for accurate retrieval.
- Leveraged the **google/flan-t5-large** language model via Hugging Face Transformers for generating accurate and coherent responses, carefully tuning parameters like temperature and beam search to optimize output quality.
- Designed and implemented dynamic context filtering of retrieved documents based on query relevance .

LoRA Fine-tuning for DistilBERT on SST-2 | *Hugging Face Transformers, PEFT*

- Implemented LoRA fine-tuning for DistilBERT on SST-2, dramatically reducing trainable parameters while maintaining performance.
- Achieved comparable accuracy to full fine-tuning with 1.8% of total parameters trained, enabling efficient training on consumer hardware.
- Applied LoRA adapters across DistilBERT's Transformer layers, including attention and feed-forward matrices, for comprehensive model adaptation.
- Developed a custom callback for real-time monitoring of model predictions during the fine-tuning process.
- Utilized Hugging Face Transformers and PEFT for streamlined model and LoRA configuration, alongside mixed-precision training.

Real-time Fraud Detection System | *scikit-learn, AutoKeras, MLflow*

- Developed a real-time fraud detection system using Python and Scikit-learn for model training .
- Implemented an MLflow tracking system to manage experiments, model versions, and deployments .
- Utilized a publicly available fraud detection dataset (IEEE-CIS Fraud Detection) for model training and evaluation, processing and analyzing the data efficiently .

Predictive Maintenance for Industrial Machinery | *Python,TensorFlow , Pandas, NumPy*

- Utilized Long Short-Term Memory (LSTM) networks for analyzing and predicting time-series data related to machinery performance.
- Designed and implemented a LSTM model tailored to predict failures based on historical sensor data from NASA's Turbofan Engine Degradation Dataset.
- Focused on feature engineering, model training, and evaluation to accurately predict potential machinery failures.

Deep Reinforcement Learning for Autonomous Spaceship Navigation | *CNN, TensorFlow, OpenAI Gym*

- Developed a Deep Q-Network (DQN) agent with Convolutional Neural Networks (CNNs) for navigating a spaceship through a simulated asteroid field environment using OpenAI Gym.
- Implemented and rigorously tested the DQN agent, focusing on optimizing its learning strategy for survival and collision avoidance.
- Demonstrated expertise in deep reinforcement learning principles and real-time decision-making in a challenging simulated environment.

Technical Skills

Languages: Python, SQL, Linux
Tools/Technologies: TensorFlow, scikit-learn, Pandas, NumPy, NLTK, spaCy, LangChain, Hugging Face Transformers, FAISS(Vector DB), MLflow, AutoKeras, Docker, Git/GitHub
Concepts: Machine Learning (Regression, Classification, Clustering), Deep Learning (CNNs, RNNs, LSTMs, Transformers), Reinforcement Learning (DQN), Generative AI, Natural Language Processing , Data Analysis, Data Visualization , Feature Engineering, Model Evaluation, Hyperparameter Tuning, MLOps, DevOps
Cloud Platforms: Google Cloud Platform(GCP), Azure Databricks
Certification: GCP Certified Machine Learning Professional

Social Engagements

Volunteer: Akshay Patra Foundation
Technical Coordinator: Lead of all events in Tech fest and member of Sponsorship Team for Kalanjali 2020
Blog: Medium blog
Sports: Intra College Chess Champion