KUMAR BAIBHAV

Email: baibhav06june@gmail.com
Phone: (716) 253-5029
Linkedin/kumar-baibhav
Github/kumar-baibhav

EDUCATION

University at Buffalo, SUNY — **Buffalo, NY**
Master of Professional Studies in Data Science (MPS-DS) | GPA – 4.0/4.0 — Dec 2024
Selected Coursework – Machine Learning, Data Mining, Probability & Statistics, Database Management Systems

Manipal Institute of Technology, MAHE — **Manipal, India**
Bachelor of Technology in Civil Engineering | GPA - 8.57/10 — Jun 2022

SKILLS

Languages: Python, R, PostgreSQL, Redis, Linux
Libraries and Frameworks: NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn, Statsmodels, Tensorflow, PyTorch, Langchain
Data Science Techniques: Machine Learning, Time Series Forecasting, NLP, Experimental Design, ETL, Data Visualization
Tools: AWS(S3, Redshift, Kinesis), MLFlow, Databricks, Git, PowerBI, PySpark, Airflow, Agile, CI/CD, REST APIs

EXPERIENCE

Spillbox — **San Jose, CA**
Software Engineer AI/ML — Feb 2025 - Present

- Developed a GPT-4o-mini powered AI assistant with a RAG pipeline using OpenAI embeddings and FAISS, integrated into the website to automate technical FAQ retrieval, reducing response time to under 5 seconds.
- Fine-tuned the model on internal support material to enhance contextual accuracy by 30%, reducing customer queries by 20% and lowering manual support costs by 80%.
- Deployed the assistant on AWS EC2 using FastAPI and Gunicorn, reducing latency by 3 seconds and enabling concurrent support for ~50 internal and external users.
- Led the end-to-end lifecycle of the assistant and collaborated with R&D to automate the assistant's training process, reducing manual update efforts by 25% and ensuring consistent support quality.

Baldwin Richardson Foods (Internship) — **Buffalo, NY**
Data Scientist — Aug 2024 - Dec 2024

- Built time series forecasting models (ARIMA, Prophet) by using 14 years of sales data, achieving 90% accuracy across 20+ SKUs, reducing stockouts by 15%, streamlining inventory planning.
- Analyzed the effects of inflation, holiday schedules, and stock prices on demand patterns, generating insights that informed a 12-month procurement strategy for the supply chain team.
- Led model development efforts in a 4-member team and delivered data-backed product mix recommendations to senior leadership, contributing ~$800K in revenue opportunity and 8.5% projected sales growth.

StatSkew — **Remote**
Data Science Intern — Mar 2023 - May 2023

- Automated product data scraping using Python, Selenium, and BeautifulSoup, reducing manual collection time by 20%.
- Built classification models (Logistic Regression, Random Forest, XGBoost) to predict likelihood of health insurance purchase among car insurance customers.
- Increased the model's ability to identify likely buyers from 60% to 85% by addressing class imbalance with SMOTE, resulting in more accurate targeting and lead prioritization.

VRC Constructions — **Manipal, India**
Research Data Analyst — Jan 2022 - Jun 2022

- Analyzed 1,000+ procurement and on-site logs using Python(Pandas) to uncover weather and equipment-related delays, reducing project lead time by 36.4% and saving $18K through better scheduling.
- Performed ABC inventory analysis and built an Excel dashboard to flag surplus and critical materials, identifying $31K in unused stock and reducing waste by 3.3%.

PROJECTS

Bike Rental Demand Prediction | *Python, FastAPI, Azure, Docker*

- Performed EDA and hypothesis testing on rental bike data to identify demand trends by time of day, season, and weekday, contributing to a 25% increase in demand planning accuracy.
- Developed and deployed a real-time prediction pipeline using AWS Kinesis and FastAPI to stream weather data into a hyperparameter-tuned Random Forest model, achieving 90% prediction accuracy.
- Delivered interactive demand forecasts through a Streamlit dashboard, enabling timely operational insights and contributing to a 30% increase in bike reallocation efficiency.

Cardiovascular Risk Prediction | *Python, FastAPI, Docker, AWS, Streamlit*

- Developed a cardiovascular risk prediction model by experimenting with Random Forest, GBM, AdaBoost, and CatBoost, improving early detection of high risk patients by 60% using threshold tuning and class imbalance handling.
- Automated deployment using Docker, AWS ECR, and EC2 with CI/CD integration, and built a real-time Streamlit UI to enable clinical risk assessment for healthcare professionals.

Fraud Detection in Credit Card Transactions | *PySpark, Databricks*

- Improved fraud detection on 1M+ transactions by addressing severe class imbalance and applying threshold tuning with Random Forest, enhancing identification of fraudulent transactions by over 40% and overall model accuracy by 18%.
- Discovered high-risk fraud patterns by engineering geo-distance anomalies, flagged merchants/categories, identified high risk transaction hours and user-level behavior signals to significantly enhance detection accuracy.