Yash Krishna Bheke

yashbheke2000 | yash-bheke | 🌐 Portfolio | ✉ yash.bheke2000@gmail.com | ☎ +1 5132765130

Professional Experience

Graduate Assistant (IS 8034: Big Data Integration) Sept 2025 - Present
University of Cincinnati Cincinnati, Ohio

- Improved processing efficiency by 31% for 20+ graduate students by optimizing 15 legacy data pipeline modules in Databricks platform using PySpark and SQL to enable faster model inference.
- Integrated AWS RDS and S3 with databricks, automated data ingestion and implemented scalable Unity Catalog solutions for reproducible ML experiments and collaborative model development.

Software Developer (Data Analytics and Machine Learning) Sept 2022 - Jul 2024
Accelya Mumbai, India

- Accelerated model inference time by 97% (from 60+ min to < 2 min) by enhancing the predictive analytics pipeline using XGBoost and feature driven dimensionality reduction techniques.
- Reduced operational costs by $1.2M annually by developing and deploying ensemble classification models (Random Forest, Gradient Boosting) to predict flight delays with 92% precision, which served 5k+ daily predictions to optimize flight scheduling and airline resource allocation.
- Improved user engagement by 32% and reduced financial reporting errors by 16% by designing an A/B testing framework combining Bayesian inference, multi-armed bandit algorithms and hypothesis testing (t-tests, chi-square) to identify 9 statistically significant process improvements ($p < 0.05$).
- Eliminated 600+ hours of monthly manual work by automating data pre-processing workflows using Python/Bash, which streamlined data cleaning, validation and EDA across 500+ GB datasets.

Projects

Airbnb Pricing Tool (Tech Stack: Python, Scikit-learn, SHAP, Google Colab)

- Preprocessed Airbnb listings dataset through text parsing (bathroom extraction), currency normalization, outlier removal, and engineered revenue-signal features for optimum pricing calculation.
- Optimized a Random Forest regressor (RandomizedSearchCV) to achieve R^2=0.58, MAE $49.69, RMSE $103.83; outperformed baseline by 35.9%. Identified top price drivers through permutation importance and SHAP to allow hosts to refine listings and access 90% prediction confidence bands.

Mood Disorder Predictor (Tech Stack: Python, SHAP, Matplotlib, Seaborn, Jupyter Notebook)

- Developed a privacy focused mood disorder prediction pipeline, achieving 96% accuracy and perfect recall on minority at risk classes via cross-validation enabling identification of vulnerable individuals.
- Visualized clinical drivers (Sleep, Exhaustion, Euphoria) for clinicians for early targeted intervention.

Education

University of Cincinnati, Carl H. Lindner College of Business August 2024 – May 2026
Master of Science, Information Systems, 3.97 GPA
Courses: Gen AI, Statistical Computing, Datamining for BI, Data Analysis, AI ML, Data Visualization
University of Mumbai August 2018 – May 2022
Bachelor of Engineering, Electronics and Telecommunications, 3.66 GPA

Technical Skills

- **Languages:** Python (Numpy, Pandas, TensorFlow, Keras, Spark), SQL, R, Bash, C++
- **Data Science & ML:** NLP, LLMs, PyTorch, RAG, A/B Testing, Time Series Analysis
- **Databases & Big Data:** Oracle, MySQL, PostgreSQL, Snowflake, Databricks, BigQuery
- **Data Engineering:** ETL Pipelines, Batch & real-time Processing, Feature Engineering
- **Data Visualization & Analysis:** Tableau, PowerBI, Looker, Excel, Zoho, Matplotlib, Seaborn
- **DevOps, API and Deployment :** Docker, Git, GitHub, CI/CD, Rest API, FastAPI, Flask