# Nikita Kumari

nikitak.nk11@gmail.com | Portfolio | LinkedIn | GitHub | U.S. Permanent Resident

## Education

**Arizona State University, Tempe, AZ** Aug 2024 – May 2026

*M.S. Data Science, Analytics and Engineering (Computing and Decision Analytics)*

Relevant Coursework: Machine Learning, Big Data Analytics, Data Mining, Optimization, Advanced Databases, Applied Statistics, Linear Algebra

**Arizona State University, Tempe, AZ** May 2024

*B.S. Biological Sciences (Biomedical)*

## Technical Skills

| | |
|---|---|
| **Languages & Libraries** | Python, R, SQL, Pandas, NumPy, Scikit-learn, PyTorch, XGBoost |
| **AI & Machine Learning** | NLP, Generative AI, LLMs, RAG, LangChain, HuggingFace |
| **Platforms & Data** | Databricks, PySpark, Docker, Tableau, ChromaDB, Git |
| **Methods & Analysis** | A/B Testing, Exploratory Data Analysis (EDA), Statistical Analysis |
| **Certifications** | AWS AI Practitioner, AWS Cloud Practitioner, Google Data Analytics |

## Experience

**Databook Analytics** Aug 2025 – Present

*Machine Learning Engineering Intern* *Remote*

- Architecting an end-to-end ML pipeline for an AI-powered educational platform, including components for experimentation and production to enhance user learning experiences.
- Developing the core AI engine using Python, text-to-speech (TTS), and NLP to generate personalized educational content and create an interactive, data-driven learning environment.

**School of Life Sciences, ASU** Feb 2022 – May 2024

*Data Research Analyst* *Tempe, AZ*

- Derived critical research insights by leading the development of a structured dataset from over 500 unstructured academic sources, engineering a robust data pipeline to solve a previously ambiguous research problem.
- Enhanced data quality and enabled advanced statistical analysis by developing Python scripts to systematically process, clean, and validate large research datasets.
- Drove **data-driven decision-making** by creating interactive Tableau dashboards that translated complex findings into actionable strategic recommendations for a cross-functional team.

**School of Molecular Sciences, ASU** Aug 2022 – Dec 2023

*Teaching Assistant (General Chemistry)* *Tempe, AZ*

- Demonstrated excellent communication and presentation skills by simplifying complex concepts for over **75 students** weekly, fostering an environment of collaborative learning.

## Projects

**Large-Scale Fraud Detection System** | *Python, PySpark, Scikit-learn, XGBoost*

- Led a full-cycle fraud detection project, using PySpark for initial analysis and baseline modeling on 280k+ transactions, and ultimately boosting recall from **18% to 82%** with a final stacked model.
- Achieved a robust Matthews Correlation Coefficient (MCC) of 0.83 by engineering novel time-series features and developing a stacked classifier (RandomForest, XGBoost) to overcome severe class imbalance.
- Drove product decision-making by explaining the model's logic to stakeholders using SHAP, turning complex data into clear and actionable recommendations.

**Natural Language-to-SQL** | *Python, ChromaDB, HuggingFace Transformers, LLMs*

- Architected a full-stack NL2SQL copilot and iteratively improved its execution accuracy from a 25% baseline to 87% on the industry-standard Spider dataset.
- Engineered a multi-stage query pipeline that boosted accuracy by **+42 point**s by implementing schema awareness with vector search and sophisticated prompt engineering.
- Drove the final performance leap to **87%** by fine-tuning a T5-based model on the Spider dataset, specializing its SQL generation capabilities for complex, cross-domain queries.

**Medical AI Diagnostic Agent** | *Python, RAG, Llama 3, ChromaDB, LangChain*

- Architected a privacy-first RAG agent for medical diagnosis, achieving **68% Top-1 accuracy** by validating performance against a custom built clinical test suite.
- Engineered a **100% local AI stack** for complete data privacy by orchestrating open-source tools, including Llama 3, LangChain, and a ChromaDB vector store.
- Implemented a custom evaluation framework that identified retrieval failures as the key performance bottleneck, enabling targeted optimizations to the RAG pipeline..

**Personalized Promotion System** | *Python, Reinforcement Learning, Streamlit*

- Engineered a promotional personalization system using a Contextual Multi-Armed Bandit, boosting conversion rates by **132%** compared to a traditional A/B test control group in a simulated environment.
- Developed the LinUCB (Linear Upper Confidence Bound) bandit algorithm from scratch in Python, enabling the system to learn user preferences in real-time by dynamically balancing exploration and exploitation.
- Built and deployed an interactive Streamlit dashboard to visualize the A/B test performance and the bandit's learned strategy, effectively communicating the model's business value to stakeholders.