# Khushbu Kakdiya

216-340-3545 ⋄ khushbukakadiya5025@gmail.com ⋄ linkedin.com/in/khushbukakdiya/
⋄ https://khushbukakdiya.github.io/resume ⋄ New York, U S A

## Skills

**Languages:** C/C++, Python, R, SQL, HTML, CSS
**Cloud Platforms & Services:** AWS (EC2, S3, Lambda)
**Tools & Frameworks:** Flask, TensorFlow-Keras, PyTorch, Seaborn, Matplotlib, SciPy, Tableau, PowerBI

## Education

**Master's of Computer Science** *Aug 2022 - May 2024*
*Cleveland State University, Cleveland, Ohio* ***GPA:*** *3.3/4.0*

**Bachelor's of Engineering in Computer Engineering** *Aug 2019 - May 2020*
*Gujarat Technological University (GTU)* ***GPA:*** *3.6/4.0*
**Relevant Coursework:** Python, Artificial Intelligence, Machine Learning

## Experience

**Data Scientist** *Oct 2024 - Present*
*Fanalyze, CA, USA*

- Built scalable ETL pipelines to ingest NBA data from web, GitHub & Kaggle; automated ingestion & model workflows using Python.
- Trained ML models (Logistic Regression, Random Forest, SVM) achieving 65.3% accuracy; integrated BERT for NLP on unstructured data.
- Used AWS/GCP for scalable storage and training; improved model reliability via backtesting (XGBoost, CatBoost, Ridge).

**Data Engineer-AI Solution** *Aug 2020 – July 2022*
*Softvan, GJ, INDIA*

- Built AI-powered UPSC/GPSC interview assistant using AWS and Alexa for real-time voice-based interviews.
- Processed audio, video & text inputs with TensorFlow, OpenCV & NLP for response analysis, scoring, and non-verbal cue evaluation.
- Ensured scalability via modular pipelines and robust data validation.

## Projects

**Weather Streaming Pipeline** *GitHub*

- Developed a real-time weather data pipeline using Apache Kafka, Spark Structured Streaming, and PostgreSQL.
- Containerized the solution with Docker and orchestrated components via docker-compose for seamless deployment.
- Streamed and processed live weather data from APIs, enabling continuous ingestion, transformation, and storage.

**Predicting NBA Games** *GitHub*

- Built predictive models for NBA game outcomes using 10 seasons of team-level stats and key metrics (eFG%, TOV%, ORB%, FTr).
- Engineered aggregated features over rolling windows (10–30 games) and applied PCA to reduce dimensionality.
- Developed data pipelines with custom web scrapers and delivered insights to support fantasy sports and betting analysis.

**Insurance Regression** *GitHub*

- Predicted individual insurance costs using regression models on demographic and health-related features from a Kaggle dataset.
- Conducted EDA, feature engineering, and categorical encoding; identified key drivers like smoking status and BMI.
- Built and fine-tuned models (Linear, Ridge, Lasso, Random Forest), achieving best performance with Random Forest via GridSearchCV.

## Certifications

- Earned the Data Analysis with Python Developer Certification by freeCodeCamp after completing 300 hours of hands-on projects involving data wrangling, statistical analysis, and visualization using Pandas, NumPy, Matplotlib, and Seaborn.
- Participated in a 3-day National Workshop on AWS Cloud, gaining practical exposure to AWS services, cloud architecture, and deployment strategies.
- Completed a Power BI & Excel Dashboard Course with 10 hours of guided training focused on building interactive dashboards and mastering data visualization techniques.
- Received an International Experience Program Participation Certificate for completing a global academic and cultural exchange initiative, enhancing cross-cultural communication and global collaboration skills.