Kashmira Golatkar

+1 812-543-3872 | kgolatka@gmail.com | LinkedIn | LinkedIn | Portfolio | Bloomington, IN

Education

Indiana University Bloomington, Indiana, USA **Aug 2022 - May 2024**
Master of Science in Data Science
Coursework: Applied Database Technologies, Data Mining, Artificial Intelligence, Statistics, Data Visualization
Graduate Teaching Assistant: Big Data and Cloud, Applied Machine Learning

University of Mumbai, India **Aug 2016 - Oct 2020**
Bachelor of Engineering in Computer Science
Coursework: Software Engineering, Natural Language Processing, Data Warehouse, Data Structures and Algorithms

Work Experience

Central Indiana Corporate Partnership **Aug 2023 - Dec 2023**
Healthcare Data Analyst | Life Sciences and Clinical trials Indianapolis

- Led integration of large, complex life sciences data of 20M+ records from diverse sources, to identify **clinical** trials gaps in Indiana.
- Developed a clinical genomics dataset by utilizing **Python** for **data mapping** and **feature engineering**, revealing specific sponsor trends and identifying a 20% gap in urban-rural clinical trial participation.
- Conveyed analytical findings to the stakeholders through interactive **Tableau** dashboard, advocating enhanced regional strategies.

Indiana University Bloomington **Aug 2023 - Dec 2023**
Research Assistant | Finance and Social Sector Bloomington

- Directed **quantitative analysis** on 5M+ IRS tax records, generating features for network graphs for uncovering complex relationships in the philanthropy sector from **XML** files using Python, **NLP**, **Neo4j Graph DB** and Cypher Query Language.
- Built recommendation models accelerating optimal funding opportunity identification by 25% for 180K+ grantees and foundations.

Accenture **Jan 2021 - Jul 2022**
Data Engineer Associate | Customer Meter Integrations Mumbai

- Architected **IBM DataStage** jobs ingesting 10GB+ data from **OLTP** to **data marts**, automating 80% of manual workload.
- Designed **ETL** pipelines using SCD1/SCD2 and Change Data Capture techniques for efficient **data warehousing and analytics.**
- Optimized performance by 60% through parallel ingestion and reduced query execution time by 40% through strategic partitioning.
- Conducted root cause analysis using **Oracle SQL Developer** to resolve data discrepancies, boosting accuracy by 70%.
- Leveraged **Azure Boards** for work item tracking, delivering the completion of 12+ user stories, and ensuring accountability and transparency within the cross-functional teams.

Key Projects

ETL Automation in the Retail Domain using Apache Airflow **Git**

- Orchestrated **Airflow** DAG, ingesting 10M+ transactions into a star schema model in **BigQuery**, leveraging Astro CLIs.
- Designed **DBT** reporting models enhancing BI analytics querying through a structured semantic layer.
- Dockerized and automated Airflow **DAG** for CI/CD pipelines, enabling task monitoring and management at scale.

Demand Forecasting and Optimization in the Supply Chain Domain using Python and ML models **Git**

- Engineered 200+ features, using lagged variables, **hypothesis** testing and moving averages, for store sales forecasting.
- Optimized **LightGBM** model hyperparameters through randomized search CV, boosting performance on 10M+ records.
- Fine-tuned using early stopping and reduced to 89 features, achieving 12% SMAPE for improved **inventory planning**.

Customer Behavior in VR Gaming Data using Snowflake and AWS for Transformation **Git**

- Consolidated semi-structured user data of the gaming industry from **AWS** S3 to capture 3 key KPIs involving usage patterns, engagement levels, and churn risk factors by automating pipelines using **Snowflake**'s tasks, stream and pipes.
- Performed Lookup, Joins **transformations** on 8M+ records, facilitating data modeling for analytics consumption.

Predictive modeling for anticipating flight delays using PySpark **Git**

- Built a PySpark **ML pipeline** on 4M+ records using a high performance IU Cloud Engine - Jetstream2.
- Employed data wrangling to conduct EDA on 60 features, aimed at identifying **correlations** among delay factors.
- Refined **Random Forest** model with GridSearchCV achieving 85% prediction rate; visualized results with statistical plots.

Technical Skills

Programming Languages: Python, SQL, T-SQL, R-Studio, Java, C++, JavaScript, Node.js, SAS
Big Data and Cloud Services: GCP, AWS, IBM DataStage, Airflow, Docker, PySpark, Snowflake
Database systems: Microsoft SQL Server, Oracle Database, PostgreSQL, SSIS, NoSQL, MongoDB, Neo4j Graph Database
Visualization Tools: Tableau, Microsoft Power BI, AWS Quicksight, Google Data Studio, IBM Cognos, DBT

Leadership Experience and Recognition

Led **code review** sessions for 150+ students to enhance their understanding in the **Big Data and Cloud** course.

Recognized for consistently contributing efforts to the integration team at **Accenture** across 5 **OPCOs** on a daily basis.

Orchestrated a 24-hour **Hackathon** for 50+ participating teams at **Atharva Group of Institutes**.