

Umar Khan

uraza.khan2000@gmail.com | +1(647)-454-7907 | www.linkedin.com/in/umarkhan | theumarkhan.com

SKILLS

Python, R, **SQL**, **Databricks**, **PySpark (Spark)**, **Tableau**, **Power BI**, **Azure (Azure Data Factory, Synapse)**, Statistics, **Airflow**, **LLMs**, **GenAI**, SAS, Scikit-Learn, Pandas, Scala, Tensorflow, NLP, **Docker**, **Azure DevOps**, A/B Testing, **Agile**, **Waterfall**, Hadoop, **Git**, Jira, Java, C/C++, **CI/CD**

EDUCATION

University of York, United Kingdom Sep '23 – Jun '25
Masters of Science, Computer Science and Data Analytics

University of Toronto Sep '18 – Dec '22
Honors Bachelors of Science, Statistics Specialist (Data Science and Machine Learning)

PROFESSIONAL EXPERIENCE

Data Scientist II Dec '25 - Current
TD Bank – Enterprise Digital Analytics

- Drove a projected **2–3% lift** in customer survey response rates by applying **ensemble based time-series forecasting** to optimize survey **consideration scores** and targeting strategy using **Databricks, PySpark, SQL** and **Power BI**.
- Collaborated with **governance** team to create migration plan for over **10+ reports** from **Tableau** to **Power BI** using **Azure Synapse Analytics, Databricks**.
- Partnered **closely** with cross-functional stakeholders across Marketing and **Customer Analytics** to perform **exploratory data analysis** on customer **appointment and contact data**, surfacing behavioral insights that informed **NPS** propensity modeling across **100+ data sources** using **PySpark, Databricks, SQL**.
- Migrated a static **PDF-based executive report** to an interactive **Power BI** dashboard, enabling **self-serve, repeatable reporting** for senior leadership and eliminating manual report generation.
- Orchestrated and automated **10+ recurring data processes** with **workflow orchestration** in **Databricks** and **Azure Data Factory**, implementing **data validation checks** and reducing weekly manual effort by **45%**.
- Applied **MLOps** practices to monitored a production **customer propensity model** for data drift, segment misalignment and accuracy decay using **Power BI** dashboards and **SQL/PySpark** checks on a monthly cycle catching shifting customer behavior that would have misrouted **15%** of targeted offers.

Senior Business Intelligence Analyst Jul '24 – Dec '25
TD Bank - Wealth

- Developed an **automated ETL pipeline in Databricks** using **PySpark, SQL**, and **Tableau** Dashboard encompassing **9 metrics** for the **DI / Retail** teams to coach **1000+** retail colleagues and analyze branch performance saving managers over **160 hours** a month in manual report generation.
- **Collaborated** with broader team to develop a metric curation framework for TD Wealth, utilizing **Databricks, PySpark, SQL** and **data modelling and Data Warehousing** techniques to encompass nearly **80** metrics, resulting in a **60%** reduction in query complexity through query and workflow optimization.
- **Coached** and **project-managed** onboarding for **2 new** analysts, enhancing accuracy and structure of Wealth reporting dashboards; **improved reporting efficiency** and deepened their knowledge of key business metrics; administered **Azure DevOps** repository access and permissions for the team and business stakeholders.
- Created and **administered Unity Catalog** catalogs and schemas in **Databricks** for **10+ internal teams** migrating to **Azure**, establishing **governed access** controls across the platform.
- Ingested data into **Azure Data Lake Storage (ADLS) curated zones** for 10+ teams, structuring and validating **100+ tables** to support **downstream** reporting and analytics.
- Conducted **valuation** assessment for a project using **regression analysis**, resulting in the transfer of nearly **3,000 leads** to a different team and raising an additional **\$30 million** in opportunity value.

Business Intelligence & Reporting Analyst III

Sep '22 – Jul '24

TD Bank - Wealth

- **Streamlined** operations by automating diverse **ETL** pipelines through **Pyspark, Powershell, Hadoop**, and **SQL** reclaiming over 10 hours a week from manual processing.
- **Conducted** in-depth data collection using diverse **big data** to set up ETL workflows for various business partners using **PySpark, SQL, Hadoop** ensuring extremely clean and well documented scripts as well as writing optimized code using **normalization, multi-threading**, etc.
- **Migrated** over **50** on-prem processes to **automated workflows** in **Azure** using **Databricks, Synapse Analytics, ADF**.

Data Science and Advanced Analytics Intern

Sep '21 – Apr '22

TD Bank - Canadian Personal Banking, Wealth

- **Analyzed** customer and merchant data using **K-Means Clustering** and **Market Basket Analysis** to find trends throughout the pandemic and present insights to business partners to increase brand loyalty.
- Built a **data pipeline** for the **MyTD** app to personalize notifications for customers based on their behavioral pattern.

SELECTED PROJECTS

PreBuildIQ – Data Pipeline & Platform Engineering (Founder)

- Designed and built ELT data pipelines using **Python (FastAPI)** and **PostgreSQL/PostGIS** to integrate and model data from **43 municipal sources**, spanning **37,900+ bylaw sections** and **36,000+ decision records**.
- Built automated **web scraping and ingestion workflows** (Playwright, deployed on Azure VMs) to keep municipal planning and zoning data current, feeding a **pgvector-based retrieval layer** for downstream analytics and search.
- Maintain a test suite of **227 backend and 149+ frontend tests** to enforce data quality and prevent regressions across the production platform.
- Implemented a **CI/CD pipeline** in **Azure DevOps** for a production SaaS application, gating merges on a **227+ backend / 149+ frontend** automated test suite and **TypeScript validation**, with an **approval-gated deployment** stage to **Fly.io**.
- Applied **A/B testing** principles to cold outreach and LinkedIn content, **iterating** on content across sequential waves and tracking engagement rates to inform positioning changes for **100+ leads**.

Performance Analysis of Proposed Enhanced Spatio-Temporal Graph Convolutional Network for Urban Traffic Speed Optimization – Research for Master's Thesis

- Designed and implemented an enhanced **Spatio-Temporal Graph Convolutional Network (STGCN)** with **dynamic graph adaptation**, attention-modulated propagation, and **multi-modal IoT data fusion (PyTorch, PyTorch Geometric)** for urban traffic prediction; benchmarked against **CNN/LSTM** baselines across 3 seeds, achieving a **45.6% MAE reduction** vs. LSTM and the lowest RMSE overall (14.97), demonstrating superior error stability over CNN and LSTM baselines.
- **Skills: Graph Neural Networks, Convolutional Neural Networks, Simulations.**

Analysis of Different Fraud Detection Algorithms

- Analyzed fraudulent financial transaction data using various algorithms; **Local Outlier Factor** and **Isolation Forest** using **SkLearn** to achieve 29% accuracy of predicting **fraud** transactions.
- **Skills: Python, Pandas, Scikit Learn, Decision Trees, Local Outlier Factor, Isolation Forest**

CERTIFICATIONS

- **Tensorflow** Developer, **IBM Data Scientist**, **Azure Data Fundamentals**, **Azure Fundamentals**, **Azure Data Engineer (IP)**