Bilal Malik

Email | LinkedIn | GitHub | Medium | Washington D.C., USA

EDUCATION

Georgetown University *Washington D.C., USA*

M.S. in Data Science & Analytics *August 2025 – May 2027*

Ghulam Ishaq Khan Institute of Engineering Sciences and Technology *Topi, Pakistan*

B.S. in Computer Science *August 2020 – May 2024*

PROFESSIONAL EXPERIENCE

Jr. Data Scientist *August 2024 – May 2025*

AI Systems *Karachi, Pakistan*

- Conducted time-series analysis on legal topic frequency to detect seasonality and emerging regulatory issues, supporting monthly insights across 5k+ legal queries.
- Modelled 8k+ healthcare entity relationships in Azure Cosmos DB, building a graph-based knowledge structure to support clinical data exploration and entity linkage.
- Built PowerBI dashboards tracking active users, subscriptions, and growth, directly informing feature prioritization during weekly product reviews.

Data Analyst *January 2024 – June 2024*

Carvan Goods Transport *Karachi, Pakistan*

- Automated KPI dashboards for fleet utilization, order volume, and delivery performance, enabling near real- time operational visibility.
- Forecasted shipment demand using 48+ months of historical data to support demand planning and scheduling.
- Integrated and standardized multi-source operational datasets to improve reporting consistency.

Data Science Intern *June 2023 – August 2023*

United Bank Limited (UBL) *Karachi, Pakistan*

- Developed a computer vision-based cheque fraud detection model using supervised learning techniques, achieving 83% validation accuracy.
- Expanded training data from 4k to 14k images via data augmentation, improving model generalization.

PROJECTS

Hospital Readmission Prediction | *Python, XGBoost, Lasso, Random Forest, EHR Data*

- Predicted 30-day readmission and length of stay on 100k+ inpatient encounters from the Diabetes 130-US Hospitals dataset using Lasso, Random Forest, and XGBoost (AUC = 0.649, RMSE = 2.17 days).
- Applied L1 regularization and RFE for feature selection, and addressed class imbalance through model-specific weighting and threshold optimization.

Insaf.AI - Legal NLP Data Extraction System | *Python, NLP, OpenAI Agents SDK*

- Designed an NLP-driven data collection pipeline that extracts structured legal entities (complainant, accused, incident details) from free-text user inputs to support automated FIR (First Information Reports) drafting.
- Analyzed conversation flows and failure cases to improve data completeness and reduce invalid or missing fields.

CFPB Complaint Modeling & Classification | *Python, XGBoost, Regression, Clustering, Feature Engineering*

- Integrated CFPB complaints with ACS, postal, and holiday data; applied clustering and Lasso-regularized regression ($R^2 \approx 0.55$) to derive and validate socioeconomic features capturing community-level disparities..
- Engineered temporal and high-cardinality metadata features and trained XGBoost models achieving 0.77 multiclass accuracy, with probabilistic outputs enabling 0.93 F1 for binary triage of high-priority complaints.

SKILLS

- **Programming & Data Tools:** Python, SQL, PyTorch, Scikit-learn, XGBoost
- **Data Visualization & Reporting:** Streamlit, Dash, Dash Cytoscape, Matplotlib, Seaborn, PowerBI
- **Databases & Cloud:** MySQL, MongoDB, Azure Cosmos DB, Azure
- **APIs and Deployment:** Git, GitHub, Flask, FastAPI, REST APIs