**Nida Madina Khan**

**+92 333 9255528 |** [**nmkhan096@gmail.com**](mailto:nmkhan096@gmail.com) **| linkedIn.com/in/nmkhan096 | Karachi, Pakistan**

**Work Experience**

**Securiti, Karachi, Pakistan** **11/2024 – Current**

*Data Scientist*

• Developed AI and Machine Learning solutions to process sensitive data securely, managing the full ML lifecycle from data collection and preparation to model training, evaluation, and production-readiness.

• Built a CNN-based model to detect the header row location within semi-structured tabular files leveraging sentence embeddings and cosine similarity, achieving **96% accuracy, <250ms inference latency, and a broader context window** compared to the previous system.

• Integrated **PyTorch** models into a **Java**-based production pipeline for performance testing, collaborating closely with engineering teams to ensure seamless deployment.

• **Fine-tuned LLaMA 3 using PEFT (LoRA)** to classify personal data (PD) types of columns in structured tables, achieving ~**92% accuracy** across 15+ PD types, and enabling classification of newly added PD types during inference.

• Iteratively improved model accuracy through **prompt-formatted instructions** and data augmentation techniques, utilizing AI copilots like **Claude Code** to automate data transformations.

**Afiniti Software Solutions Ltd., Remote** **02/2020 – 10/2024**

*Data Scientist II, Artificial Intelligence Team 05/2022 – 10/2024*

*Data Scientist, Artificial Intelligence Team 05/2021 – 05/2022*

*Junior Data Scientist, Artificial Intelligence Team 02/2020 – 05/2021*

• Optimized call center interactions using machine learning techniques like Decision Trees and performance metrics to estimate customer lifetime value and agent impact, generating over **$1 million in revenue per month** for SKY UK.

• Conducted extensive grid search experiments comprising of **2000+ models** to find optimal hyper parameters.

• Improved the efficiency of the ML pipeline by implementing strategies to **reduce grid search time by 50%**, and adding features to streamline and standardize the process of model logging.

• Designed interactive dashboards using **R** and **Python**, resulting in streamlined tracking and visualization of KPIs.

• Conducted real-time production monitoring using **MySQL**, leading to improved data quality for decision-making.

• Used **gradient-boosted decision trees** (e.g., **LightGBM**) for predictive revenue forecasting and customer churn models.

• **Trained 10+ members** across three clients on the ML pipeline, R programming and client business, contributing to an increase in project efficiency.

**Projects**

**• End-to-end PySpark data pipeline** to process large-scale taxi trip data. Used Spark local mode for prototyping and deployed the final pipeline on a single-node Dataproc cluster on GCP. Ingested Parquet-formatted data from GCS, standardized schemas, and wrote monthly aggregated metrics to BigQuery for reporting and analysis.

• **Machine Learning Pipeline Automation**: Deployed an automated, end-to-end machine learning pipeline in R, encompassing data preprocessing, model training, validation, and deployment phases, greatly enhancing productivity.

**Skills**

**Programming:** Python, R, Java, SQL, Bash

**Data Engineering & MLOps:** Spark, Docker, MLflow, Git, GCP

**Machine Learning & Modeling:** Probability & Statistics, Deep Learning, LLM fine-tuning, Feature Engineering, Model Evaluation, Data Visualization (ggplot, Streamlit, Tableau)

**Languages:** English (C2), Urdu (mother tongue)

**Education**

**National University of Sciences and Technology, Islamabad, Pakistan 2015 – 2019**

*B.E. Electrical Engineering*

• **Cumulative GPA: 3.83/4.00**

• Courses: Machine Learning, Probability & Statistics, Data Structures & Algorithms, Object-Oriented Programming