Rohail Taimour

Python/R Software Engineer | Contract Data Scientist | Statistician | AI/Machine Learning Specialist

1040, Brussels, Belgium
rohail.taimour@gmail.com | +32 489 83 64 76 | Linkedin | Github

Education

MSc in Statistics - KU Leuven, Leuven, Belgium (2014-2016)

Cum Laude Graduation, Master's thesis on continuous optimization of production processes in **MATLAB**

BSc (Hons.) in Accounting and Finance - Lahore University of Management Sciences (LUMS), Lahore, Pakistan (2010-2014)

Graduated with Distinction (3.6/4.0)

Technical competencies

Programming:	Python and R (5+ years each)
Containerization:	Docker
OS:	Windows, MacOS, Linux (Redhat/Ubuntu)
IDE:	Pycharm, VScode, Rstudio, jupyter notebooks, Azure databricks
CI/CD Tools:	Git, GitHub/Azure Pipelines, GitHub CLI tools and Git flow branching
Cloud platforms:	AWS, Azure
Databases:	Postgresql, SQLite3, Neo4j, sqlalchemy
Documentation:	Pandoc, Markdown, sphinx

Summary

Experienced data scientist with an expertise in working on data products that have a machine learning focus. I am able to play a dynamic role between data scientist/machine learning engineer and data engineer given the needs of the project. I am passionate about adopting development best practices wherever possible and am comfortable working in an ambiguous and fast changing environment, laser focused on delivering user requirements in an agile manner.

Freelance projects (Oct 2022-present)

Automated SQL Script Generation for Cross-Platform Data Migration

Python Software Engineer, Illumina, Mechelen, Belgium July-Aug 2023

- Designed and implemented an ORM Mapper for dynamic ingestion of various file formats, automating the SQL script generation process for data migration.
- Implemented the solution as a Python package encapsulating the entire data migration logic within a Docker entrypoint for portability and ease of deployment.
- Conducted comprehensive testing of generated SQL scripts using mock PostgreSQL database tables, ensuring script accuracy and reliability.
- Parameterized key inputs allowing for seamless deployment across multiple environments (Development, Integration, Production).

Configurable Bioinformatics Pipeline for Event-Triggered Secondary Analysis of Sequencing Data Using Python and Docker

Python Software Engineer and Data Pipeline Architect, Illumina, Mechelen, Belgium April 2023 - Present

- Developed a configurable multi-stage pipeline for secondary analysis, implemented in Python and deployed as a Docker entrypoint.
- Designed and implemented an event-driven system that actively monitored for new sequencing data, triggering corresponding analyses upon detection of state changes, thereby ensuring timely and efficient processing of data.
- Implemented unit testing using pytest and implemented fail-safe mechanisms for robust error handling.
- Optimized pipeline performance by implementing an SQLite database for tracking previously launched analyses, enabling the pipeline to function as a daemon with persistent memory.
- Provided guidance on automation strategies, leveraging CLI tools and API calls to enhance interoperability between Illumina platforms ICA and Basespace.

Design and implement information retrieval methods using Natural language processing (NLP)

Machine Learning Engineer, IT Supply Quality, GSK Belgium Oct 2022-Feb 2023

- Improved performance of information retrieval by 20% on unseen test data using a custom named entity recognition (NER) from Spacy
- Performed POC's on improve model performance using rule based techniques as well as NER and annotated data to train custom NER
- Added text preprocessing features to the NLP pipeline such as spacy tokenization, Part of speech (POS) tagging, better handling of non-english emails, breaking emails into sentences, etc

Personal details

- **Nationality:** Belgian, Pakistani
- **Languages:** English (fluent/bilingual), Urdu (Native), French (B1)
- **Mobility:** Driving Licence available, flexible for hybrid setup with up to 4 days on site
- **Availability:** Immediately
- **Hobbies:** Drumming and percussion instruments, Bouldering/Climbing, productivity, Squash, reading

Data science projects at IT AI team, UCB Pharmaceutical (2016-Oct 2022)

Yield optimization for batch and continuous production processes using Machine Learning in Python

Lead Data Scientist, Supply and Manufacturing, UCB Switzerland/Belgium Aug 2020-Oct 2022

- Production setting proposed by model directly led to an increased throughput of 20%, turning in a recurring 1.5 million euro in annual cost savings
- Analyze time series data collected from equipment sensors and visually summarize golden batch insights
- Created (bayesian) and tree-based regression models to quantify impact of process changes and predict batch performance
- Performed a thorough model validation and hyperparameter tuning exercise before recommending model insights be tested in a live production environment
- Supported delivery of workshops demystifying the process of conducting AI projects and machine learning to process experts

Optimizing Resource Efficiency and Customer Engagement through Channel-Specific Promotional Responsiveness

AI/ML engineer, Lead Data Scientist, Go to Market/Commerical EU5, US and Japan, UCB June 2019-June 2021

- Developed a Python package that abstracts the complexities of the data science workflow, enabling configurable deployments across diverse scenarios such as different countries and disease areas.
- Enhanced the package to seamlessly wrap over scikit-learn, thereby simplifying key data science tasks from preprocessing to model training and tuning.
- Incorporated MLflow into the package for robust artifact management, allowing for the tracking of model versions, data inputs, and predictions.
- Created customer segmentation models and proposed optimal resource allocation based on customer responsiveness to different marketing channels.
- Investigated adaptations to data science methodology for country/product specificities for maximum reusability. Delivered as many as ten different use cases as lead data scientist for different products and countries
- Supported data engineers in the creation of features using pyspark and validated ingested data using data visualization methods and discussions with subject matter experts

Scientific influencer (KOLs) identification, ranking and profiling using network analytics and Neo4j

Data scientist/Product owner, Drug Development, Commercial, Medical affairs, UCB 2018-2019

- Created custom Neo4j databases by ingesting additional data sources to quantify influence. Delivered tailored KOL ranking/profiling reports and presentations to meet stakeholder requirements.
- Made network visualizations using networkx, Cytoscape and performed custom analysis.
- Support improvements in the intake of customer requests to reduce time to deliver reports from days to hours.

Developed an automated forecasting workflow of claims data from US healthcare system

Lead Data Scientist, US Finance and claims, UCB 2017-2018

- Created modular R packages to extend the functionality of Facebook's prophet package with an end to end workflow for ingesting, forecasting and reporting to analyze forecasting results
- Achieved forecasting accuracy of **> 90%** across the different use cases
- Prototyped different time series forecasting methods to flexibly model multiple time series models and performed hyperparameter tuning and validation on a batch compute machine
- Applied anomaly detection methods to account for outlying behavior in time series automatically

Hands-on workshop to demystify Artificial intelligence and Machine Learning

Data science instructor, IT departments US, EU, UCB May- June 2017

- Created a **R shiny** application to create an engaging way for participants to learn about typical AI use cases
- Delivered the workshop to over **100 people** in four different venues and received great feedback on level of engagement

Personal projects

- Developed webscraper in python (Beautiful soup, Selenium) to compare apartments based on price, area, etc
- Created a predictive model for whether movie will achieve an imdb rating > 7.5 using best practices for data science projects