Matin Zarei

Mobile : 07427673358 **E-mail** : matinzarei.zm@gmail.com **Location:** London **website**

PROFILE

Data analyst with hands-on experience in Python, SQL, Databricks, and machine learning, passionate about applying advanced analytics to real-world problems. I bridge the gap between data and business by designing pipelines, building models, and translating results into actionable insights. With proven success in end-to-end projects — from defining KPIs to delivering stakeholder-ready solutions — I'm expanding my expertise in machine learning and data science to drive even greater business impact. Curious, self-driven, and always evolving.

I aim to create actionable value, not just analysis. Website: matin-zarei.com

KEY SKILLS

- **Programming & Data Tools:** Python,pyspark, SQL, Git, Web Scraping
- **Data Analysis & Visualization:** Pandas, NumPy, Matplotlib, Seaborn, Power BI, AI/BI
- **Machine Learning:** scikit-learn, TensorFlow | Classification, Regression, Clustering, Tree-based Models
- **NLP:** spaCy, Hugging Face (Sentence-BERT) | Vectorization & Embedding Models for Text Classification, Semantic Similarity, Clustering
- **Statistics:** Hypothesis Testing, A/B Testing, Statistical Inference, Experimental Design, Probability Modeling
- **Big Data & Cloud:** Databricks (Delta Lake, MLflow, Spark SQL), PySpark | Building ETL pipelines
- **Development Tools:** Power BI, Power Apps, Power Automate, Excel, Docker, Jupyter, VS Code
- **Soft Skills:** Stakeholder Communication, Storytelling, Business Acumen, Cross-functional Collaboration

EMPLOYMENT HISTORY

June 2023 –Now, Thames Water (Data Analyst)

- Designed and implemented a SCADA signal name standardization system using **Sentence-BERT embeddings**, **DBSCAN**, and **Transformer-based seq2seq models** in Azure Databricks, improving signal clarity and enabling automated processing.
- Led an **alarm cycle SLA analysis** project across millions of records to uncover delay patterns in alarm handling, classify alarm states, and support KPI monitoring across teams.
- Applied NLP clustering (Sentence-BERT + KMeans) on operator logs to identify recurring operational issues and enhance incident response analysis.
- Automated data collection pipelines with **Power Automate**, integrating data from SharePoint, emails, databases, and CSVs to streamline daily reporting.
- Delivered the **Total Alarm Report**, producing actionable insights that reduced waste and clean water alarms.
- Collaborated with engineers, control room analysts, and performance leads to align data insights with operational improvements.
- Completed advanced SCADA training, strengthening my domain knowledge in telemetry and control systems.

2019 – 2021 Rohamweb (Data scientist , Analyst)

- Conducted A/B testing and funnel analysis for digital marketing campaigns, **improving bounce rate and session engagement.**
- Collaborated with content, web design, and finance teams to apply insights and guide decisions
- Prepared and transformed datasets for customer segmentation and recommendation models in Python, enabling **improvements in click-through rates** and session duration.
- Collected and integrated data through large-scale web scraping (Scrapy)and SQL databases (MySQL)
- Partnered with data engineers to maintain scalable pipelines in SQL and Python, ensuring high-quality, consistent data for client deliverables.

2017 – 2019 Rohamweb / Nikan (Web Developer)

- Developed RESTful APIs and full-stack web applications (frontend & backend), designed MySQL databases, implemented SEO tracking using Google Analytics, and performed Python-based web scraping for data collection.

EDUCATION

2021–2022 MSc Data Science & Analytics, **Brunel University** London **,**
Key modules: Data Visualisation, Quantitative Analysis, Machine Learning, HPC
Dissertation: Gender Classification on Twitter using 20+ ML models
BSc Computer Science (IT), Pishtazan University

TECHNICAL PROJECT WORK

SCADA Signal Name Standardization *(Thames Water – Databricks + NLP)*

- Developed a semi-automated signal name standardization pipeline using Sentence-BERT embeddings to detect high-similarity pairs for human review.
- Designed custom parsing logic to extract structured components (e.g., Region, Asset Levels, Signal Type) from raw SCADA names, improving naming consistency and enabling scalable analytics across telemetry systems.

Performance Compliance Reporting System *(Thames Water – Big Data & KPI Reporting)*

- **Architected end-to-end reporting solution:** Designed and built a centralized analytics platform to track service performance against internal targets, translating raw logs into actionable KPIs.
- **Built robust ETL pipelines:** Used Python and SQL to extract, transform, and load data from multiple systems.
- **Developed dashboard with role-based access:** Created visualizations in Power BI, implementing user-level access controls so that managers, and front-line staff each see only their relevant metrics.
- **Implemented KPI-driven alerting:** Defined key performance indicators and threshold rules; automated email and in-app notifications to both users and managers when performance dipped below targets.
- **Automated reports & notifications:** Scheduled daily and weekly summaries, with escalation workflows for critical breaches, ensuring stakeholders received updates 8 times a day.
- **Drove performance improvements:** Partnered with operations leads to refine metrics and workflows—boosted on-time delivery from 90% to 94% within six months.

Implementing a recommendation system for a movie streaming start-up

- Working with a team of four to make and improve recommendations. improving click-through rate by 15%.
- Making an integrated database (APIs, Database, web scraping).
- Preparing data for use by an ML engineer, Dealing with missing data and false information (Python).

Clean Alarm Response Power-BI project

- Designed and developed a comprehensive Clean Alarm Response Power BI report, consolidating alarm data and key performance metrics to provide real-time insights and drive informed decision-making.
- Facilitated frequent requirement-gathering sessions and collaborative workshops with managers and shift controllers to align dashboard functionality with operational needs, streamlining communication and improving incident response processes.

Gender Classification on Twitter *(Dissertation Project)*

- Built a full ML pipeline using **TF-IDF, Word2Vec, GloVe**, and **Text Bleaching**, evaluating 20+ models (RF, SVM, LR, NB) across datasets.
- Data pre-processing containing removing: words that carry minimal weight in text using NLTK package.
- Built a full ML pipeline using **TF-IDF, Word2Vec, GloVe**, and **Text Bleaching**, evaluating 20+ models (RF, SVM, LR, NB) across datasets.
- Implementing machine learning models including **RF**, **SVM**, **LR** and **NB**.
- Cross-validation is used to evaluate and compare best performance models and to test on new datasets.

Learning & Certifications

- **Certificates:** Google Data Analytics (Coursera), AWS Cloud Practitioner Essentials, Hugging Face NLP Specialization, Databricks Certified Data Engineer Associate
- **Continuous Learning:** Active Kaggle participant (applied ML on real-world datasets), regular reader of data science & machine learning books to expand knowledge