MOJTABA YOUSEFI

mysfi@mit.edu | 617 817 6662 | Boston, MA
Will not need sponsorship for work. | ysfi.me | github.com/myousefi | /in/mysfi
SCAN ME

SUMMARY

Data Scientist/Software Engineer with over **6 years** of experience in *industry* and *academia*, specializing in *developing*, *integrating*, and *optimizing* data-driven solutions for **production environments**. Committed to fostering **collaboration** within cross-functional teams and members of the team at all levels. Thrives in *fast-paced*, *agile environments*, leveraging **strong communication skills** to bridge the gap between technical teams and business stakeholders. Advocates for a **rapid prototyping methodology** that embraces *iterative improvement*, *accelerated development cycles*, and *pivoting when justified and needed*.

SKILLS

Programming & Development Python, SQL, Julia, TypeScript, Go, Scala, Java, C++, Node.js, C#, Matlab *(Sorted by recency of professional experience)* | React, Vue, HTML, CSS, JavaScript

Machine Learning & Data Science Pandas, Numpy, Statsmodel, Plotly, Dash, Scikit-learn, Jax, Pytorch, Numba, Ray | Gurobi, Google-OR Tools

Cloud, DevOps & MLOps: OpenShift, AWS (EKS, ECS, Lambda, SQS, EC2, S3, CloudWatch)| Docker, Docker Swarm, Kubernetes, Terraform | Apache Kafka, Apache Airflow | MLflow, WandB

Data Management & Storage SQL (Postgres, SQLite, TimescaleDB, DuckDB), NoSQL (MongoDB, Redis, DynamoDB, Neo4j) | RedShift | Prometheus, Grafana, ELK Stack

EXPERIENCE

Graduate Research Assistant 8/2022 – Present — **MIT Transit Lab**

- *Analyzed* the *causal impact* of CTA Blue Line shutdown on Uber/Lyft ridership using **difference-in-differences design** and **Bayesian structural time-series models**, identifying the temporal and spatial patterns of affected origin/destination areas.
- *Conducted a cohort analysis to study* CTA passengers impacted by the Blue Line shutdown using **SQL** on **RedShift** databases, discovering **chrun rates** and alternative route choice patterns within the system.
- *Presented* and discussed the results to diverse audience of **100+ employees**, **C-Suit Executives**, and the **President of CTA**.
- *Modeled passenger demand* as a dynamic **stochastic Poisson process** utilizing Automated Fare Collection (AFC) data, employing **data imputation techniques** such as **k-Nearest Neighbors (kNN)** and **Multivariate Imputation by Chained Equations (MICE)** to handle missing data, effectively capturing realistic demand patterns and temporal variations while ensuring **data integrity** and **robustness** of the model.
- *Developed* a *microscopic agent-based simulation model* in **Python** adhering to **SOLID** and **OOD principles**, accurately representing vehicle movements, signal systems, and passenger demand, enabling evaluation of service planning scenarios and operational challenges.

Data Scientist Intern 5/2023 – 8/2023 — **Chicago Transit Authority**

- *Formulated MIP models* for rail service restoration with dynamic demand, leveraging **Julia**, **JuMP**, and **Gurobi** to reduce system-wide waiting time by 7%.
- *Optimized complex **SQL** queries* in **Redshift** to derive operational insights by combining data from SCADA (rail vehicle location system), AFC (automated fare collection), transit scheduling, and workforce planning tables, enabling data-driven decision making for scheduling, operations, and customer communication teams.
- *Spearheaded cross-functional discussions* between operations, service planning, and scheduling teams, identifying and presenting on critical operational bottlenecks.

Machine Learning Intern 5/2022 – 8/2022 — **Harvard Medical School**

- *Prototyped* a working *3D collision and ray-triangle intersection detection* pipeline in **Python**, reducing the processing times of each case using coarse-grained meshes in days to fine-grained meshes in minutes. This line of work was used in a startup spin-off that won the Harvard President's Innovation award (BONEPIXEL).

Software Engineer/Data Scientist 1/2019 – 8/2021 — **Yas Group**

- *Designed, developed, and deployed* a *high-throughput streaming data pipeline* on **OpenShift** processing various limit order book data from 5+ different sources using **Apache Kafka** for real-time data ingestion, applied streaming *feature engineering techniques*, and persisted processed data into **Apache Cassandra** for further analysis, basically enabling in-house data capabilities.
- *Deployed* **CatBoost/Statsmodel** regression models on **Kafka** for real-time inference, reducing prediction latency by an order of magnitude, and enabling 1000+ predictions per second.
- *Packaged* complex feature engineering, modeling, and testing logic developed by Data Scientists into a unified **Python** library with **JIT compilation** where applicable using **Numba**, and **CI/CD** pipelines through **Gitlab Actions**.
- *Architected and deployed* a *real-time database monitoring service* using **Jaeger** for distributed tracing, **Prometheus** for metrics collection, and **Grafana** for visualization, enabling the optimization of database read/write performance, and ensuring seamless data access across multiple microservices.
- *Developed* multiple internal tools for the data science team, including project templates using **cookiecutter** with **static code analysis**, **linting**, **formatting** and **CI/CD** for experiment tracking and access to model registry. Reducing project setup time, and enforcing best practices.
- *Spearheaded **Python** and **Linux** coffee/brown bag sessions* for 15+ data science team members, fostering knowledge sharing and collaborative culture.
- *Designed, developed and deployed* an on-premise infrastructure using the **ELK stack** (**Elasticsearch**, **Logstash**, **Kibana**) to monitor **model drift** and **residual distribution** in real-time for deployed machine learning models.
- *Developed and trained* advanced **gradient-boosted decision tree regression** and **statistical time series models** on 1000+ financial time-series data using **CatBoost**, **Ray**, and **Dask**, leveraging feature engineering and hyperparameter tuning to identify trading strategies with Sharpe ratio greater than 2.
- *Worked* with a *cross-functional* team of software engineers, DevOps, data scientists, and financial experts.

Data Scientist 9/2018 – 1/2019 — **Tosan Ofogh**

- *Developed* a *backtesting engine* for retrospective **A/B testing** of trading strategies, incorporating risk management and performance metrics, allowing for hyper-optimization of 75+ strategies.
- *Developed* a data pipeline that leveraged web scraping with **Beautiful Soup** to exploit a bug in the **Tehran Stock Exchange (TSE)** website, uniquely identifying major shareholders. Ingested data into a **Neo4j graph database** and matched it with annual reports to monitor the activities of market makers, resulting in the identification of trading opportunities with Sharpe ratios exceeding 3.
- *Worked* with a *cross-functional team* of economists, software engineers, and domain experts in analyzing financial reports.

SAMPLE PROJECTS WITH SOURCE CODE AVAILABLE

MIT RailSim Python, Sphinx, SQL, Pandas, Dash Plotly — **Link**
MIT RailSim is an urban heavy-rail operations simulation model developed at the MIT Transit Lab, built upon decades of research.

Kaggle LLM Prompt Recovery Transformers, Datasets, Streamlit, t-SNE, Spectral Clustering, Peft, Spacy, Slurm, Bash — **Link**
My scripts and models for the LLM Prompt Recovery Competition on Kaggle

EDUCATION

Master of Science May 2024 — **Northeastern University**
Electrical and Computer Engineering: Computer Vision, Machine Learning, and Algorithms
Relevant Coursework: Data Mining for Engineering, Applied Probabilities & Stochastic Processes, Machine Learning & Pattern Recognition, Advanced Reinforcement Learning, Advanced Machine Learning

Bachelor of Science in Engineering August 2018 — **Sharif University of Technology**
Civil and Environmental Engineering, Minor in Computer Science
Relevant Coursework: Statistcs and Applications, Linear Algebra, Operations Research I & II, Data Structures, Design of Algorithms, Relational Database Management Systems