SOBHAN MOOSAVI

Email ◇ LinkedIn ◇ Google Scholar ◇ Github ◇ Kaggle

SUMMARY AND OBJECTIVE

Senior Machine Learning & Applied Scientist with 6+ years of experience developing and deploying scalable, interpretable ML systems across domains like autonomous driving, large-scale routing, and financial fraud detection. My work at Lyft and Zoox spans deep learning, generative modeling, and real-time decision systems — focused on building transparent, production-grade AI solutions that drive measurable impact in the real world.

WORK EXPERIENCE

- Senior Applied Scientist (Bellevue, WA) 2025 – 2026
  - Researching, prototyping, and designing a machine learning–based **card fraud detection system** integrated into FIS's existing framework to improve precision and recall in identifying fraudulent transactions. Developing a two-tier architecture comprising a **generic fraud model** and a **real-time residual model** that adapts to missed detections and evolving fraud patterns. Working with **multi-billion daily transaction volumes**, leveraging distributed training and inference frameworks to ensure scalability, efficiency, and high accuracy in production. Emphasizing **interpretability and debuggability** for a broad audience, ensuring model transparency and explainability while maintaining state-of-the-art accuracy.

Data Scientist (Machine Learning) at Zoox (an Amazon subsidiary), San Mateo, CA

- Senior Data/ML Scientist 2023 – 2025

  Contributed to the development and validation of Zoox's Collision Avoidance System (CAS), driving measurable safety and reliability improvements through data-driven validation and cross-functional collaboration. Leveraged state-of-the-art **deep learning–based** modeling approaches to enhance perception and decision-making pipelines.
  - **Safety Validation**: Defined safety metrics, led large-scale validation cycles, and delivered reproducible insights guiding CAS development and leadership decisions.
  - **Machine Learning Integration**: Enhanced kinematic and behavioral models (e.g., KEM) for multi-agent and dynamic scenarios; applied **multi-modal transformer foundational models** and deep learning techniques to improve dataset quality, scene representation, and risk estimation.
  - **Methodology Innovation**: Advanced validation frameworks and simulation methods to better capture real-world behavior while optimizing cost and data efficiency.
  - **Cross-Functional Leadership**: Represented CAS V&V in company-wide forums, aligning with autonomy, safety, and data science teams to improve communication and technical documentation.

Data Scientist (Machine Learning and Algorithm) at Lyft, San Francisco, CA

- Senior Data Scientist 2021 – 2023
  - Led science and engineering vision, strategy and roadmap on a team of 15+ engineers (ML, SWE, DE), data scientists, and product managers working on the design, implementation and experimentation (A/B testing) for in-house optimized and safe routing
  - Mentored junior DSs, SWEs, ML SWEs, and interns, enabling their skill growth and career advancement
  - Drove continuous improvement of products, internal (e.g. workflows) and external tools (e.g. map and navigation) by challenging the status quo and constantly seeking new ways to enhance quality through multidisciplinary collaborations
- Data Scientist 2020 – 2021
  - Designed, implemented, and improved a routing cost model that leveraged machine learning and optimization techniques to find optimal routes, and enhance routing quality and accuracy. This model was crucial in improving navigation experience (as in LyftMap), ETA and distance prediction, pricing, and dispatch. It was a significant departure from sub-optimal, potentially unsafe, or illegal routes, and proved to be a game-changer for the industry (e.g. resulted in improving route accuracy by 15% and driver compliance by 10%)
  - Led cross-functional efforts to establish the essential infrastructure required to effectively scale our new routing model. Conducted multiple online experiments to rigorously test each variation of the model prior to its launch
- Research Science Intern May 2018 – Aug 2018

- Developed a deep neural network model for ETA estimation based on spatiotemporal input data
- Co-authored a research paper on "estimating travel times by routeing-aware supervised learning"

General skills/tools: Python, Java (familiar), Tensorflow, PyTorch, Scikit-learn, SQL (Spark, PySpark), S3/AWS, Databricks.

EDUCATION

Ph.D. in Computer Science, Ohio State University 2014 – 2019

- Dissertation: Telematics and Contextual Data Analysis and Driving Risk Prediction
- Advisors: Prof. Rajiv Ramnath and Prof. Srinivasan Parthasarathy
- Research Areas: Data mining and Machine Learning

M.S. in Computer Software Engineering, University of Tehran, Tehran, Iran 2009 - 2012

B.S. in Computer Science, Shahid Beheshti University, Tehran, Iran 2005 - 2009

SELECTED PUBLICATIONS AND PATENT

- Context-aware driver risk prediction with telematics data, Elsevier Accident Analysis & Prevention 2023
- Predicting road constructions based on heterogeneous spatiotemporal data, ACM SIGSPATIAL 2022
- Locationtrails: A federated approach to learning location embeddings, IEEE/ACM ASONAM 2021
- Driving style representation in convolutional recurrent neural network model of driver identification, Preprint 2021
- System and method for analyzing vehicle data, US Patent 10740990 (2020)
- Short and Long-term Pattern Discovery Over Large-Scale Geo-Spatiotemporal Data, ACM SIGKDD 2019
- Accident risk prediction based on heterogeneous sparse data: New dataset and insights, ACM SIGSPATIAL 2019
- QDEE: question difficulty and expertise estimation in community question answering sites, AAAI ICWSM 2018
- Characterizing driving context from driver behavior, ACM SIGSPATIAL 2017

PUBLISHED DATASETS

- US Accidents (2016 - 2023), available on Kaggle (about 180K downloads, among the top 40 popular datasets)
- US Weather Events (2016 - 2022), available on Kaggle
- US Road Construction and Closures (2016 - 2021), available on Kaggle
- US Traffic Congestions (2016-2022), available on Kaggle
- CityTrek-14K: A large-scale dataset of 14,000 car trajectories, available on Kaggle