

Mostafa Varzaneh

Phone: +1 (865) 582-5589

Email: mostafagv@gmail.com

LinkedIn: <https://www.linkedin.com/in/mostafa-gv>

Summary

I have a strong foundation in Machine Learning, Large Language Models (LLMs), Generative AI, Data Analysis, and Statistical Methods. I have a proven track record of designing, implementing, and deploying comprehensive ML pipelines into production across various industries, including energy, online advertising, supply chain, and customer service chatbots and voice agents.

Additionally, I excel in expanding and leading machine learning teams and effectively communicating complex data and algorithms to non-technical audiences, including sales and marketing professionals.

Area of Expertise

- Machine Learning
- Generative AI
- Deep Learning
- Full Stack ML Pipelines
- Large language models (LLM)
- Natural Language Processing
- Ads Ranking
- Speech Recognition
- Reinforcement Learning

Professional Experience

IBM

Manager of Machine Learning

Dec 2021 – Now

- Expanded the machine learning team of WatsonX Orders (8 new hires) and helped building the data science team (10 new hires) also as the interim manager, hired 5 machine learning engineers for IBM Concert.
- While being hands on, lead multiple teams of developers across different groups to:
 - Develop and deploy a voice-based customer service conversational agent that served over 100 million clients (WatsonX Orders).
 - Implement a distillation pipeline with multiple large language models (LLMs) to train an edge deployable Natural Language Understanding (NLU) and intent detection model with less than 7 million parameters (WatsonX Orders).
 - Develop a pipeline with combination of Function calling LLMs and Retrieval-Augmented Generation (RAG) techniques to analyze the security of the codebase and application with a chat interface (early version of IBM Concert).
 - Implement a LLM distillation and reinforcement learning pipeline to make WatsonX Orders' English-only chat model, bilingual (Spanish and English with code switch in middle of sentence) while keeping the final model size small.
 - Build multiple PyTorch-based ASR (Automatic Speech Recognition) pipelines employing conformer-transducer, jasper CTC and RNNT in collaboration with IBM research.
 - Created a data selection pipeline to tag audio files for human transcription and injection into the training pipeline. This pipeline incorporates multiple models, including Speech Activity Detection and an ensemble of ASR models such as Wave2Vec and Whisper.
 - Employed shallow fusion and a language model to enhance the performance of the ASR system.

McD Techlabs (Acquired by IBM)

Staff Machine Learning Engineer (NLP)

Aug 2020 – Nov 2021

- **Team Lead:**

- Quick Eval: Developed an evaluation pipeline for predicting the performance of the voice agent (Speech Recognition, Natural Language Understanding models).
- Refactored and enhanced NLU pipeline.
- Compared the state-of-the-art models including LSTM, Encoder/Decoder, and Transformers in Natural Language Understanding.
- Scaled the training pipeline using containerization and Kubeflow.

Adara

Data Science Lead

Mar. 2018 – July 2020

- **Project Lead:**

- Developed and deployed a Discretionary Model for predicting the next purchase of a user with an estimated purchase time frame.
- Lead a team of Data Scientists and Engineers to develop a Recommender System, MicroService, and APIs for Audience Recommender engine for Display and Search advertising platforms.
- Auto-ML: Managed a team to develop an automated modeling pipeline using a variety of Google Cloud tools like GKE (Kubernetes), Cloud Composer, BigQuery ... for click/conversion prediction in real-time bidding consuming data from hundreds of data partners.

- **Developer:**

- Bid Optimization: Prototyped a bid optimization model using Reinforcement Learning and Bayesian Optimization at audience level for Google Ads search, Meta Search, and DV360.
- Ads Bidding Model: Created a bidding system for Google Ads on the audience and keyword level.
- Developed model for incremental/lift insertion orders which broke the sales record of the company.

LCG Consulting

Data Scientist

Mar. 2017 – Mar. 2018

- Used machine learning to predict the bidding behavior of market participants.
- Developed an optimization model to find the best location for the installation of batteries in a power network.
- Creating and updating network databases using Python and SQL.
- Utilizing LSTM and Deep Learning for Nodal Price prediction.

Niagara Bottling *Supply Chain Analyst Intern*

Jun. 2016 – Aug. 2016

Azad University *Lecturer*

Jun. 2009 – Jun. 2011

Customer Centric Group *Project Manager*

Mar. 2007 – Mar. 2009

Software Proficiencies

Machine Learning: PyTorch, Scikit-learn, TensorFlow, Keras, LangChain

Programming: Python, C++, Matlab

Database and utility: Different flavors of SQL and NoSQL, RestAPI, Docker, Beam, Airflow, Argo ...

Education

University of Tennessee-Knoxville

Doctor of Philosophy, Industrial and Information Engineering

Minor in Statistic

Amirkabir University of Technology, Tehran, IR

BSc and MSc, Electrical Engineering

