Rajasvi Vinayak Sharma

+1(858)319-9642 | rvsharma@ucsd.edu | linkedin.com/in/rajasvi | rajasvi | rajasvi.github.io

EXPERIENCE

Signos
Apr 2023 – Present

Machine Learning Engineer | Python, Tensorflow, WandB — Remote, CA

- Identified & researched discrepancies in a training dataset of 1 million meals for post meal glucose prediction, enhancing ML model's generalizability & pinpointing noisy impacts on the test set. Implemented a strategy to reduce the training dataset by **10%** without compromising the overall MAPE (mean absolute percentage error).
- Conducted extensive statistical analysis on time series data, devising and implementing a heuristic algorithm for rapid increase alerts in Continuous Glucose Monitoring (CGM) readings. Successfully revamped the existing alerting algorithm, resulting in a **60%** increase in precision, an **80%** boost in recall, and a 10-minute reduction in average alert time, positively impacting over **10,000 users**.

Nvidia
Jun 2022 – Sep 2022

Data Scientist Intern | Python, PySpark, SQL, KubeFlow, MLFlow, CausalML — Santa Clara, CA

- Developed and deployed an end-to-end Time-series Anomaly Detection tool using Z-Score thresholding and Kubeflow pipelines to alert about malicious activities in 1000+ categories across 10M+ gaming sessions, reducing the alert response time from a few months to just 1 week.
- Improved in-place A/B test analysis by creating tool to identify significant sub-population using causal inference ML models like S-Learner, T-Learner, Double ML methods etc.
- Developed user engagement metrics using regression analysis to identify and target disengaged users, effectively tracking their local & absolute engagement while progressing in a game. Metrics used as a feature in a churn prediction model and in development of targeted marketing strategies for increasing customer retention.

Goldman Sachs
Jun 2018 – Aug 2021

Data Scientist, Analyst | Python, Java, PySpark, Apache Flink, SQL, Spacy, XGBoost — Bengaluru, IN

- Developed a scalable CRF model that accurately extracted Signature, Salutation, Disclaimer blocks, and related contact entities from over 8 million emails per day, achieving an impressive 85.7% accuracy rate, through persistent designing and testing.
- Engineered a PySpark-based ML feature extraction pipeline that efficiently processed >10M emails per day, delivering real-time predictions at a rate of >24K data points per minute, & successfully deployed ML models for entity recognition use-cases, by integrating it with an existing Apache Flink Big Data ETL pipeline.
- Built XGBoost & LGBM models for classifying jurisdiction violations among Bloomberg trader conversations (>6M daily interactions) by leveraging semantic & temporal information extracted from the data as features, achieving an impressive 78% precision.

EDUCATION

University of California - San Diego | California, US
Sep 2021 – Mar 2023

M.S. in Electrical & Computer Engineering ***(Major: Machine Learning & Data Science)*** — GPA: **3.91 / 4.0**

- **Coursework**: Recommender Systems & Web Mining, Deep Learning for Natural Language Understanding, Statistical Natural Language Processing (NLP), Big Data Mining & Spark Analytics, Statistical Machine Learning (ML), Linear Algebra

Indian Institute of Technology (Banaras Hindu University) Varanasi | India
Jul 2014 – May 2018

B.Tech. in Electronics Engineering — GPA: **8.81 / 10**

PROJECTS

Clickbait Spoiler Generation (SemEval'23 Shared task) [code] [report] | PyTorch, Hugging Face — Sep 2022 – Dec 2022

- Created a 2-stage process to generate Clickbait spoilers using text classification, question answering, and ranking models. Utilized DeBERTa, RoBERTa, and DistilBERT Transformer models, fine-tuning them to achieve optimal results. Evaluated the efficacy of the approach using metrics such as BLEU-4, exact match, and F-1 scores.
- Contributed to the Clickbait Challenge by surpassing the benchmark results through the implementation and enhancement of techniques outlined in the organizers' research paper, resulting in an outstanding 36.4 BLEU-4 score.

Neural Collaborative Filtering for Recommendation Systems [code] [report] | PyTorch — Jan 2022 – Apr 2022

- Built 3 Neural Collaborative Filtering models: Generalized Matrix Factorization (GMF), Multi-Layered Perceptron (MLP) & Neural Matrix Factorization (NeuMF) following architecture & key-metrics evaluation from original paper.
- Compared traditional recommendation algorithms like Matrix Factorization with state-of-art Neural collaborative filtering for Movie Recommendation task. Calculated nDCG and HR@10 scores to study the effect of latent hidden factors on performance.

SKILLS

Languages : Python, SQL, R

Big Data : Apache Flink, PySpark, SparkSQL, Hadoop, Map Reduce, Kafka, Redis, Hbase, HDFS

ML Frameworks/Libraries : Tensorflow, PyTorch, Hugging Face, Pandas, Numpy, scikit-learn, CausalML, DoWhy

MLOps/Dev Tools : Kubeflow, MLflow, Databricks, Weights & Bias