Rajasvi Vinayak Sharma

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

EDUCATION

University of California - San Diego | California, US Sep. 2021 – Jun. 2023
M.S. in Electrical & Computer Engg. ***(Major: Machine Learning & Data Science)*** GPA: **3.66 / 4.0**

- **Coursework**: Statistical Learning, Linear Algebra, ML: Learning Algorithms, Python Programming for Data Analysis, Probability & Statistics for Data Science

Indian Institute of Technology (Banaras Hindu University), Varanasi | IN Jul. 2014 – May 2018
B.Tech. in Electronics Engineering GPA: **8.81 / 10**

SKILLS

Languages : Python, R, SQL, Java, C++
Big Data : Apache Flink, PySpark, Hadoop, Map Reduce, Kafka, Redis, Hbase, HDFS, Yarn
Machine Learning : PyTorch, Pandas, Numpy, scikit-learn, Spacy, TensorFlow, Spark NLP
Technologies/Dev Tools : React, Node.js, MongoDB, Alteryx, Tableau, Git, Kubernetes, VS Code, REST API

EXPERIENCE

Goldman Sachs
Jun. 2018 – Aug. 2021
Analyst | Core Machine Learning - Search Engineering Team Bengaluru, IN

ML pipeline : Cross-language infrastructure ML pipeline for real-time Big Data Analytics | PySpark, Flink

- Developed Python based PySpark ML signals pipeline integrated with Apache Flink, processing **>10 million emails per day** and providing real-time predictions a rate of **>24k data points per min**
- Solved Apache Flink's inability to use Python's ML libraries using PySpark streaming & Redis cache
- Deployed trained scikit-learn ML models for spam classifier, entity recognition use-cases over PySpark realtime ML pipeline.

Entity recognition : Salutation, Disclaimer, Signature (SDS) block extraction | Spacy, SparkNLP, Scikit-learn

- Developed Conditional Random Field(CRF) model using Spacy, sklearn-crfsuite, and Spark NLP achieving **85.7% accuracy** to identify SDS blocks and scaled up to extract contact entities embedded from **>8 million emails/day**
- Enriched Goldman's knowledge graph using extracted entities, improving graph surveillances for external Bloomberg contacts

Search Engineering : Conversation Stitching Model | Apache Flink, Java, Redis

- Built data engineering infrastructure, REST endpoints, in Java aggregate real-time stream of daily Bloomberg trader conversation snapshots (**>4 million per day**) into common chatroom bins, for stitching messages into a single merged conversation view
- Developed chatroom based indexing algorithm which **reduced indices size by 40% and search latency by 30%**

Machine Learning : Trader chat analysis for predicting geographic location | Pandas, PySpark, Python

- Extracted semantic & temporal information from Bloomberg trader conversations **(>6 million per day)** for resolving external traders geographic location and determine possible jurisdiction violations
- Performed feature engineering using PySpark & Spacy followed by tuning XGBoost, LGBM models achieving **78% precision**

Samsung R&D Institute
May. 2017 – Jul. 2017
Summer Intern | Bixby AI Team Noida, IN

- Developed offline image-classification android app, integrating custom-build optimized CNN models using Tensorflow
- Final model achieved accuracy of **82% and occupied mere 7kb** on phone with offline prediction capability

PROJECTS

Adverse Food Events Analysis | Pandas, Plotly, Numpy Sep. 2021 – Dec. 2021

- Detailed Exploratory Data Analysis of Adverse Food Events reports (2004-2020) gathered from FDA site, identifying causes of serious outcomes based on factors like age, symptoms, gender & food category. [code]
- Identified key brands & potential outcomes to help users beware of potential health risks before purchasing a product.

Web Development : Online-game player rating tracker | MongoDB, React, Node.js Feb. 2021 – Apr. 2021

- Developed webapp based on MERN Stack, utilising Brawlhalla game's API to track ratings of players; deployed independent frontend - backend server on Netlify & Heroku platform [site]

Business Intelligence : Front-to-Back Data Modelling & Analytics | Alteryx, Tableau Aug. 2018 – Mar. 2019

- Built dimensional data models, handling **>1M trades/day**, using SQL, Python, ElasticSearch APIs & Alteryx by transforming trade-level data from multiple OLTP sources into a unified OLAP data warehouse
- Created visualisation layers in Tableau to surface Key Performance Indicators and provide tracking across the trade life-cycle