Rajasvi Vinayak Sharma

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

EDUCATION

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

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

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

SKILLS

Languages : Python, Java, C++, R, SQL
Big Data : Apache Flink, PySpark, Redis, Hadoop, MapReduce, Hbase, HDFS, Yarn
Machine Learning : scikit-learn, crfsuite, NLTK, Spacy, TensorFlow, Spark NLP

EXPERIENCE

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

ML : Cross-language infrastructure for ML pipelines processing real-time Big Data stream

- Developed PySpark streaming pipeline integrated with Java-based Apache Flink realtime big data pipeline, processing **>10 million emails per day** at a rate of **>24k data points per min**
- Solved Apache Flink's native side-input limitation to utilise Python's ML libraries using Redis cache with PySpark streaming
- Scaled up ML model's inference stage for various trained sci-kit models over streaming Big Data

Big Data Engineering for Search Engine : Conversation Stitching Model

- Built a scalable infrastructure to 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%**

Entity recognition : Sequential Models identifying Salutation, Disclaimer, Signature (SDS) entity blocks

- Developed Conditional Random Field(CRF) + NER hybrid model using sklearn-crfsuite, Spacy, 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
- Built an in-house annotator modular Streamlit webapp to gather NLP multi-class token-sentence training dataset, which **replaced Goldman's best outsourced annotator**

Data Engineering : Front-to-Back Data Model for Securities & Derivatives

- 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 KPIs and provide tracking across the trade life-cycle

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

- Developed image-classification models, optimised CNN architectures using Tensorflow mainly for creating portable models to be integrated with image-classification app.
- Custom built model achieved accuracy of **82% and occupied mere 7kb** on phone with offline prediction capability

Indian Institute of Space Science & Technology (IIST)
May. 2016 – July 2016
Summer Intern | High-Performance Computing Lab Trivandrum, IN

- Studied ML theory & implemented algorithms specifically ensembles such as Random Forest, AdaBoost etc. from scratch
- Performed comparative performance analysis and verified findings based on research paper Fernandez-Delgado et al., with 10 classifier families across 15 UCI repository datasets for classification problems

PROJECTS

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

- Detailed EDA 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.

Trader chat analysis for predicting location | PySpark, scikit-learn, streamlit Jan. 2021 – Apr. 2021

- Extracted semantic & temporal information from Goldman's trader conversations **(>6 million per day)** to build a ML model resolving external traders geographic location with **78% precision** and determine possible jurisdiction violations