Nazih Kalo | 510-506-1332 | nazihkalo@gmail.com | Linkedin |Github | nazihkalo.com

PROFESSIONAL EXPERIENCE

CyberConnect|Head of Data | San Francisco, CA *May 2022 – Present*

- Built all data pipelines, including indexing & decoding on-chain data across 3 chains using Airflow/Spark/dbt and scraping off-chain sources like Github/Twitter/Discord
- Developed **nft & wallet recommendation engines, providing recommendations to >2M addresses,** leveraging wallet trading/minting history to power follow/content suggestions
- Maintained all internal/external dashboards (incl. dune, internal), retention/growth insights, & analytics for partners on link3.to
- Built 4 NextJs/TS demo dApps leveraging our on-chain contracts and wrote accompanying technical walkthrough articles

Scale AI | Product/Data Analyst & Data Engineer| San Francisco, CA *September 2020 – May 2022*

- **Built & maintained data pipelines for** the company's largest data extraction/scraping project, **scraping 12M+ products from ~5000 ecommerce sites**. Extracted data was parsed, categorized/normalized to fit into customers' taxonomy.
- Developed internal Payout Optimizer to dynamically adjust payout functions to hit target rates; **reduced pay variance by ~50% and led to $90k savings/month**
- Deployed **self-hosted data cataloging tool** (Amundsen), improving data discovery across the company & significantly reducing analytics team onboarding time. Extracted & **linked Snowflake, dbt, BigQuery, Tableau, & Salesforce metadata.**
- **Reduced LiDAR labeling time 34%** through 1) optimizing ML pre-labels in product, 2) developing a new labeling pipeline (isolating 2D/3D labeling stages). New 2D labeling pipeline reduced computer spec requirement & increased labor pool.

Hive AI | Product Analyst| San Francisco, CA *June 2020 – September 2020*

- Product lead for company's new **ML based text-moderation product**; scope included dataset management, model training/deployment, post-training optimization, and monitoring/maintenance of SLAs
- Collaborated with the ML team to develop a human-assisted/in-the-loop model auditing system to identify model deficiencies and error patterns in production data. **Improved model F-1 score by 24%** with minimal additional training data.

Apple Inc. | Operations Internship | Cupertino, CA *January 2018 – December 2018*

- Built data pipelines integrating internal & vendor data to reduce spend forecasts latency from 168->24hrs
- Managed data for $50M budget for iPhone XR dev builds and identified $1M fraudulent invoices through my analysis.

EDUCATION

The University of Chicago| MSc Data Science | Chicago, IL | *GPA: 4.00/4.00* *June 2020*

- Relevant Coursework: Advanced ML, Deep Learning, NLP, Big Data, Data Engineering
- Awards: Facebook Hackathon 2019 – WebBuilder ChatBot - 1st Place Prize

University of California, Berkeley | B.A Economics | Berkeley, CA | *GPA: 3.85/4.00* *December 2017*

- Certification: Certificate in Entrepreneurship & Technology | UC Berkeley, IEOR Department

PROJECTS

Crypto Token Developer Analysis App *January 2022*

- App for analyzing a token/coin's Github contribution & engagement history (stars, commits, lines of code,etc.)
- Created graph network of crypto tokens' github repos, with edges representing github contributor/developer overlaps

IBM Supply Chain Risk Advisor – Capstone Project *October 2019 – June 2020*

- Built web-scraping pipeline to extract entities/information from news articles and populate Neo4J graph of 1200 IBM suppliers
- Trained semi-supervised model on over 50M articles, labeled using heuristic functions, to predict supply chain disruptions

Scientific Research Paper Knowledge Base *Jan 2020 – March 2020*

- Implemented a knowledge graph connecting 175M scientific research papers; using papers as entities and citations as edges
- Preprocessed 280GB+ of text; extracting relevant features for the graph and citation predicting models.
- Developed semantic search function to discover new papers & found relevant features for improving papers' citation count

LEADERSHIP

Scale Labeling Cost Initiative – Spearheaded various operations/engineering initiatives to **reduce company labeling cost by 18% or $500k/quarter**; led multiple experiments with tasker incentives, labeling pipelines & hosting cost infrastructure.

DevOps Team Lead for IBM Project – Led team of 5 data engineers/scientists, delivering a software solution hosted entirely on GCP with >99% uptime for APIs & <2% error rate from predictive models

SKILLS AND LANGUAGES

Technical: NLP, Experimental Design, GLMs, ML Models (Parametric/Non-parametric/NN), RDBMS/NoSQL, Time Series

Languages/Platforms: Python, ML Libs (MLflow, scikit, keras/Torch), SQL, AWS/GCP, Neo4j, Spark, Docker, Airflow/Dagster, Web3 Tools (Brownie, Infura, Remix, Dune Wizard), RDBMS (postgres, mysql, timescaleDb)