Sujay Nagesh Koujalgi

Machine Learning Engineer / Data Scientist , 814-996-8691 , koujalgisujay@gmail.com , github.com/sujayk96

Dynamic Engineer with strong foundation in Machine Learning, Artificial Intelligence, Data Analysis and building AI products. Eager to join an organization to empower data-driven decision making that drive business growth. Proficient in Python, GCP, SQL, Numpy, Pandas, GenAI and Pytorch with hope to learn from industry experts and support innovative edge technology.

Work Experience

Research Assistant - Explainable AI lab , Penn State University **August 2022 – Current**

- **AI dev/ integration**: Spearheaded advanced encoding research for war games with a 20,000 state space, securing over $1M in DOD grants and enhancing AI model explainability, while training Deep Learning models focused on RL and MCTS behavior.
- **Statistics / Explainable AI:** Pioneered research on novel methods for analyzing human prediction (of AI decision) task for gaming data in a higher granularity of output space. Proposed ways to reduce floor and ceiling effects in binary classification.
- **NLP:** Conducted NLP research on legal document classification and built a RoBERTa-based pipeline capable of classifying Terms of Service agreements with an F1 score of 0.74. Assisted with distributed-GPU training using Torchrun.
- **LLM**: Finetuned LLama-2 model on 15k Terms of Service documents and created an interface using streamlit to classify them.
- **CV:** Implemented a semi-supervised computer vision strategy employing OpenCV and ResNet models to classify 25,000 images with a confidence score of 75% utilizing Pytorch. This enlarged the dataset specifically designed for applications that classify extremism.
- **PicAlert:** Instrumental in the setting up and benchmarking of image classification algorithms for creating a personalized privacy dataset, which was accepted by the AAAI. Paper: https://ojs.aaai.org/index.php/ICWSM/article/view/19387.

Data Scientist , Merkle, Bangalore, India **October 2019 – July 2021**

- **Search and Ranking**: Engineered a data-driven POC using NLP techniques like Word2Vec, Glove to enhance in-site search for specialized categories, leading to its implementation by a top 3 U.S. home improvement retailer.
- **Revenue Prediction:** Developed a scalable tool using boosting algorithms and time series analysis for assessing revenue driver impacts on business KPIs. Successfully deployed and productionalized on cloud-based distributed ML platforms such as GCP and VertexAI.
- **A/B testing**: Led and mentored a cross-functional A/B testing team of 3 members, conducting an extensive series of over 100 tests to enhance online customer experience like ads optimization, recommendations, product placement, marketing campaigns and recommend data-informed business strategies. Leveraged Adobe Experience and GCP to analyze and implement data-driven improvements.
- **Data Visualization**: Integrated real-time Tableau visualization, optimizing data observability (PySpark, DataBricks). Collaborated with the Merchandising team to provide immediate revenue insights for agile inventory planning.

Software Engineer - Data , Wells Fargo EGS, Bangalore, India **August 2018 – September 2019**

- **Scripting:** Utilized Python and shell scripts to successfully migrate B2B web API integrations, enhancing large file transfers from third-party servers. Ensured zero data loss and established error-handling procedures, specifically addressing issues like timeouts.
- **ETL**: Engineered more than 50 ETL pipelines, ensuring data integrity for analytics driving Financial insights.

Machine Learning Intern, Oracle Financial Service Software **June 2017 - August 2017**

- **Unsupervised learning:** Streamlined customer segmentation model for credit card users using kmeans and presented optimal clusters.

Project Experience

- **Influencer Stats** (link): Automated playlist statistics monitoring for Influencers, reducing manual tracking time by 10% by building an end-to-end Machine Learning pipelining system with PySpark, Youtube's APIs, and real time data processing using Kafka.
- **Mental health Indicator** (link)**:** Collaborated with an NGO to identify and support Reddit users with depression symptoms. Led the development of an app using Reddit's APIs, Flask, and NLP models, identifying approximately 100 potential users. Enabled timely outreach and improved mental health trend understanding through d3.js visualizations.

Technical Skills

Languages and Tools: Python (NumPy, Pandas, Scikit-learn, Pytorch, Keras, TensorFlow, GeoPandas, NLTK, spaCy, label-studio), R, C++, SQL, Adobe Workspace, DataBricks, PySpark, Informatica, Kafka, R, Microsoft Project, Advanced Microsoft Excel, ETL, Java
Visualization Tools: Tableau, Python (Matplotlib, Seaborn, Plotly), d3.js, Prompt Engineering
Analytical techniques: Machine Learning (Linear Regression, KNN, SVM, Decision Trees, Clustering, Random Forests, Deep Learning), Reinforcement Learning, CV (object detection, segmentation), NLP (Word2Vec, TF-IDF, NER, Sentiment Analysis), Optimization, Principal Component Analysis, A/B Testing, Hypothesis Testing, Data Warehousing, Data Lakes, Stored Procedures, Data Warehouse
Cloud and DevOps: GCP, Distributed System, Bigquery, Vertex AI, Azure, Docker, Kubernetes, AWS, ML-Ops, Dev-Ops

Education

The Pennsylvania State University, University Park (Master's in Information Science and Tech , 3.95/4.0) **August 2021 – August 2023**

Ramaiah Institute of Technology, India (B.E. Computer Science and Engineering) **August 2014 – May 2018**