Anurag Hambir

812-606-3617 | anuraghambir@gmail.com | linkedin.com/in/ah10 | anuraghambir.github.io | San Francisco, CA

EDUCATION

Indiana University Bloomington — Bloomington, IN
Master of Science in Data Science — *May 2023*
Savitribai Phule Pune University — Pune, India
Bachelor of Engineering in Computer Engineering — *June 2018*

SKILLS

Languages: Python (Pandas, Numpy, Scikit-learn), Java, R, Shell Scripting
Databases: MySQL, PostgreSQL, MongoDB, Hive, Elasticsearch
Machine Learning: Recommendation System, Classification, Regression, Forecasting (LSTM, ARIMA), NLP (LLM), CV
Tools/Frameworks: Amazon Web Services (AWS), Google Cloud Platform (GCP), Snowflake, Hadoop, Spark, Airflow, GIT, REST API, Kibana, JIRA, Docker, Tableau
Certifications: Data Analysis using PySpark

WORK EXPERIENCE

Data Scientist Intern, Social Science Consulting LLC, Bloomington, IN, US — *August 2023 – Present*

- Employed PyTorch (Python) in conjunction with advanced large language models (BERT and RoBERTa) on a **300,000**-text data set to create precise contextual word embeddings for generating personalized recommendations
- Leading Graph Neural Network-based Recommendation system development (GraphSAGE, GATConv), enabling personalized foundation suggestions for grant recipients and promoting reciprocal recommendations

Data Scientist Intern, ProMazo Inc., Remote, US — *June 2022 – December 2022*

- Developed a customer retention model using Python for a Fortune 500 banking and insurance client, boosting the F-1 score by **70%** through the XGBoost model and evolutionary algorithm, resulting in projected savings of **$269,000** per retained customer
- Utilized SQL and Snowflake to add **50** new features to the Attrition model, resulting in a **5%** accuracy improvement
- Optimized LSTM hyperparameters via grid search and cross-validation, boosting churn rate prediction precision by **15%**

Data Scientist, HT Media Ltd, Pune, MH, India — *July 2020 – June 2021*

- Built a recommendation system using user profile-based segmentation and Elasticsearch database and deployed using Docker on AWS that contributed to the overall increase in the downloads of the **OttPlay application** by over **1 million**
- Designed and implemented a Python ETL pipeline for crawling movies and shows from OTT websites, efficiently managing data storage and retrieval through MongoDB, resulting in over **40%** reduction in data collection costs
- Migrated the data pipeline to AWS Cloud using AWS Lambda, S3, and DynamoDB, reducing processing costs by **20%** by eliminating the usage of on-premises servers
- Developed interactive Tableau visualizations that showcased competitor ad data on HT Media websites, enabling the sales team to identify market trends and generate **10%** more leads for new client acquisition
- Managed various data projects involving large unstructured data sets, while offering guidance and mentorship to a team of **four** engineers

Data Engineer, Persistent Systems, Pune, MH, India — *August 2018 – November 2019*

- Implemented a PySpark (Python) data pipeline for sentiments and entities analysis on clients' email data, employing Stanford's CoreNLP library and SpaCy, and delivered weekly client satisfaction reports to the CEO
- Processed and analyzed email data stored in Hadoop, through PySpark and stored the results in Hive tables
- Optimized the performance of the data pipeline by integrating Kafka, reducing the total processing time by **50%**

PROJECTS

Vehicle Detection and Counting in Images (Guided by Prof. David Crandall)

- Transfer learning of Computer vision models such as YOLOv7, RetinaNet, and Faster R-CNN models using TensorFlow and PyTorch on a custom image data set consisting of images of highway traffic
- Achieved Mean Average Precision score of **0.76** with YOLOv7, outperforming all the other models

Map Reduce operations on Google Cloud Platform (GCP) (Guided by Prof. Prateek Sharma)

- Built Map Reduce architecture for word count and inverted index problems using Google Cloud functions
- Created an API to invoke the operations and built a web UI to showcase the output