Anurag Hambir

(812) 606-3617 | hambiranurag@gmail.com | linkedin.com/in/ah10 | anuraghambir.github.io

EDUCATION

Indiana University Bloomington, Master of Science in Data Science; GPA: 3.9/4.0 — Bloomington, IN
Courses: Machine Learning, Artificial Intelligence, Graph Analytics, Cloud Computing — *May 2023*

Savitribai Phule Pune University, Bachelor of Engineering in Computer Engineering; GPA: 3.7/4.0 — Pune, India
Courses: Data Structures, Big Data Analytics, Data Management, Software Engineering, Business Intelligence — *June 2018*

SKILLS

Languages: Python, Java, R, SQL, NoSQL
Databases: SQL Server, PostgreSQL, MongoDB, Hive, Elasticsearch, BigQuery
Machine Learning: Scikit-learn, SpaCy, NLTK, CoreNLP, XGBoost, Pandas, Numpy, TensorFlow, PyTorch, Keras
Tools/Frameworks: AWS, GCP, Snowflake, Hadoop, Spark, Git, REST API, Airflow, DBT, Kibana, JIRA, Docker, Tableau, Flask, Microsoft Office Tools
Certifications: Data Analysis using PySpark

WORK EXPERIENCE

Data Scientist, Credit One Bank, Las Vegas, NV — *November 2023 – Present*

- Leveraged Snowflake and SQL to develop **200+** KPIs, including customer acquisition, revenue growth, payment trends, and card sales. Delivered actionable insights through Tableau dashboards to the CEO and key stakeholders, driving strategic data-driven decision-making
- Transformed Python code into PySpark for segmenting customers using KMeans clustering to send out targeted promotional offers on a **6.5**M dataset, achieving a significant reduction in processing time from **4** hours to under **5** minutes
- Leading a team of **4** to design data pipelines for data migration from legacy systems to a hybrid cloud, enhancing data availability and enabling seamless downstream access
- Developed robust ELT workflows including data cleaning and preprocessing with PySpark and Hive, data transformation using SQL in Snowflake and orchestration using Airflow

Data Scientist, Social Science Consulting LLC, Bloomington, IN — *August 2023 – November 2023*

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

Data Scientist, ProMazo Inc., Bloomington, IN — *June 2022 – December 2022*

- Developed a customer retention model using Python for a Fortune 500 banking and insurance client, improving the F1 score by **70%** with an XGBoost model optimized through a genetic algorithm, resulting in projected savings of **$269,000** from retained customers
- Enhanced the Attrition model by engineering 50 new features using SQL and Snowflake, achieving a **5%** improvement in accuracy
- 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*

- Engineered a scalable recommendation system with Python, leveraging profile-based user segmentation and Elasticsearch, contributing to a **1M+** increase in OttPlay application downloads
- Implemented comprehensive data-driven and A/B testing strategies to evaluate the performance of the recommendation engine, resulting in a **15%** increase in click-through rates
- Collaborated with cross-functional teams, employing statistical modeling and data analysis, to extract insights from user preferences and improve content recommendations, resulting in a **20%** increase in average time spent on the platform
- Developed Tableau visualizations for competitor ad data analysis, generating **10%** more leads for new client acquisition
- Designed and implemented a Python ETL pipeline for extracting data from over **100** 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 infrastructure, reducing processing costs by **20%** by eliminating the usage of on-premises servers

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

- Implemented a machine learning pipeline using Python for sentiment analysis on clients' email data, employing Stanford's CoreNLP library and SpaCy, and delivered weekly client satisfaction Tableau 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%**