SOHAM BHAGWAT

Bloomington, IN ◇ sobhagwa@iu.edu ◇ LinkedIn ◇ Github ◇ Website ◇ 812-327-4070

EDUCATION

Indiana University, Bloomington, IN — Expected May 2024
MS in Data Science

Savitribai Phule Pune University, Pune, India — May 2022
Bachelor of Engineering, Computer Engineering (Honors in ML)

SKILLS

Programming Languages	Python, R, Java
Databases	MySQL, PostgreSQL
Tools	Git, Jenkins, Ansible, Excel
Libraries	Pandas, NumPy, PyTorch, Scikit-learn, scipy, dplyr
Visualization tools/libraries	Tableau, Matplotlib, seaborn, ggplot2
Machine Learning	Regression, Classification, Clustering, Graph Neural Networks, Decision Trees
Cloud	AWS: S3, RDS, VPC, Beanstalk, Sagemaker

EXPERIENCE

Indiana University-Bloomington — Aug 2023 - Present
Graduate Associate Instructor: **Data Analysis & Modeling** (Spring'24) — *Bloomington, IN*
Statistical Analysis for Effective Decision Making (Fall'23)

- Led weekly instruction for 40+ graduate students on statistical techniques, including advanced regression analysis.
- Instructed students on complex concepts **(Chow tests, Hausman test, chi-squared, nested models)** during office hours, delivering practical insights.
- Demonstrated hands-on application of complex concepts like **predictive modeling, hypothesis testing**, and **data visualization** using ggplot2 in R.

PTC Inc. — Aug 2021 - June 2022
DevOps Engineer Intern — *Pune, India*

- Automated deployment of newer windchill builds with Python and Ansible, saving **2+** hrs daily.
- Optimized and maintained Windchill CI/CD pipelines, achieving a **31%** reduction in test execution time for 100k+ tests.
- Automated and optimized Oracle DB deployment on Azure VM, cutting pipeline failures by **20%**, and saving 1+ hour daily for 40k Windchill regression tests.

PROJECTS

Amazon Products Recommendation using GNNs:

- Compared the performance of various Recommender System using models like traditional Collaborative Filtering, NGCF and Graph Sage on Amazon Products Dataset with more than **500k** data points.
- The hit-rate for NGCF and GraphSage was **0.817** and **0.846** respectively.

Asteroid Deep Water Impact:

- Conducted intricate scientific visualizations using Paraview on a **6TB** NASA Deep Water Impact dataset, revealing the probable impact of the asteroid on ocean surfaces.
- Designed a batch processing pipeline using Python to seamlessly transfer unstructured VTK grid files from the source silo to Paraview software.Checkout Video1 Video2.

Scientific Publications: Classification and Link Prediction

- Classified and predicted links among scientific publication networks in seven classes using the CORA dataset.
- Achieved a **93%** accuracy in node classification using Logistic Regression and an **89%** accuracy in link prediction.
- Enhanced classification accuracy by 2% through embedding dimension adjustment from 50 to 120.

Soccer Club Reddit Metrics Dashboard:

- Built a real-time ETL pipeline utilizing Mage AI and PRAW to scrape Reddit metrics, for leading soccer clubs.
- Automated the data scraping of more than 50k data points and leveraged Google BigQuery for robust storage.
- Created an engaging and intuitive dashboard interface using Looker Studio with daily updates.

Property Listings in India:

- Designed and implemented a database application with an interactive user dashboard, utilizing a cloud-hosted SQLite database featuring a dataset of over **150k** properties across the top 10 metro cities in India.
- Leveraged SQL indexing techniques to optimize querying, leading to a notable 750ms reduction in latency per retrieval.