SHRUTI HOUJI

812.778.4724 • Louisville, KY, 40223 • sghouji@iu.edu • LinkedIn • GitHub • Portfolio

EDUCATION

Indiana University, Bloomington 08/2022 – 05/2024
Master of Science – **Data Science** (GPA - 3.770)
(Coursework - Statistics, Applied Machine Learning, Cloud Computing, Computer Vision, Social media mining, Time-series analysis, Elements of Artificial Intelligence, Data Visualization, Advance Database Technologies)

Pune University 06/2016 – 04/2020
Bachelor of Engineering – **Computer Engineering** (GPA - 4.0)

WORK EXPERIENCE

Project 990 - *Remote* 06/2024 – Present
Data Scientist

- Performed **EDA** to analyze organizational counts, identify financial discrepancies, and examine grant, asset distributions, and financial metric correlations, enhancing data quality by 29% on U.S Charities and Non-Profits dataset using **Python.**
- Utilized **Tableau** to create geographic heat maps of educational philanthropy, identifying areas with high dropout rates, enabled targeted resource allocation to underserved regions, improving geographical coverage across U.S. states by 35%.
- Created a robust text pipeline using **MySQL**, **Python** to **ETL** 270,000 records from the database, enhancing data throughput.

Danfoss Power Solutions - *Cleveland, OH* 05/2023 – 08/2023
Data Analyst Intern

- Analyzed maintenance request data in **Power BI**, **R** with clustered column and time-series charts using seasonal decomposition to identify trends, prioritize tasks, optimize resource allocation, and reduce maintenance task completion time by 57%.
- Examined labor data using **seaborn** radar, scatter plots, uncovering workload imbalances, improving task distribution by 21%.
- Integrated diverse data sources into data warehouses, built **ETL pipelines** in **Python** using techniques like data cleaning, normalization, and incremental loading, and used **SQL databases** for data modeling and processing to streamline workflow.
- Automated status of order reports for the production planning team using **Excel VBA scripts** and **macros** to merge, format, rename, separate, and sort columns, reducing daily report creation time by 30 minutes each morning.
- Crafted 7 KPIs per assembly line using **DAX**, **Power BI**, and **Python** to track productivity targets, enabling informed decision-making on employee scheduling to ensure coverage of required productivity hours and enhance overall efficiency.

Cognizant Technology Solutions - *Pune, India* 12/2019 – 05/2022
Programmer / Data Analyst

- Applied **k-means clustering** and **DBSCAN** to analyze policyholder details, revealing causes for unclaimed policies like age groups, address changes, deaths, missing information, and non-payments, leading to a 33% reduction in policy lapse rate.
- Enhanced extracted handwritten data by 41% using **Power BI**, ensuring detailed data profiling, standardization, and validation.
- Employed **SQL** and **Oracle** databases to query and filter data, boosting filtering efficiency by S48% for criteria-specific policies.
- Boosted business performance by enhancing **OCR with Python** to identify unclaimed LIC policies by 64% from handwritten documents and optimize data processing, leveraging **.NET**, **Git**, **Azure**, and **agile** methodologies within the **SDLC** framework.
- Coordinated **Jira** task updates with a 10-person cross-functional team, demonstrating strong organizational, problem-solving, interpersonal, and communication skills, while leveraging **HTML** and **JS** webpages to improve customer service satisfaction.

PROJECTS

Skin Lesion Classification for Melanoma Detection

- Spearheaded an end-to-end skin cancer detection system leveraging ML XGBoost algorithm with 74% accuracy.
- Leveraged AWS services (S3, SageMaker, Lambda, IAM) to streamline analysis, achieving 50% faster processing.

Datawage Navigator

- Structured a robust MySQL database design, allowing users to access job listings in specific locations within milliseconds.
- Integrated CRUD operations and dashboard page using Flask to get real-time updates with a 28% decrease in data latency.

Energy analytics and prediction in the USA

- Implemented time series models such as ARIMA, VAR, and LSTM for forecasting optimal energy resources for the upcoming decade by applying data-driven recommendations.

Real-time Intrusion Detection Systems for IOT Networks using ML

- Built a dataset using an IoT node with Tshark scripts to capture 1 GB of instantaneous network logs, used data mining technique to develop an IDS using a Random Forest model, and integrated websockets for real-time attack detection notifications.

SKILLS

- **Programming Languages**: C++, Java, Python, R, DAX
- **Data Analysis Tools**: Microsoft Power BI, MS Excel VBA, Tableau, Looker, Matplotlib, QlikView, Seaborn, ggplot
- **Databases and Web Technologies**: MySQL, NoSQL, PostgreSQL, Neo4j, BigQuery, Snowflake, HTML5, JavaScript
- **Cloud Platforms**: AWS (S3, EC2, Lambda, SageMaker), Google Cloud Platform (GCP), Microsoft Azure
- **Statistical Technologies/Tools**: Regression, Hypothesis testing
- **Misc**: Data Structures, NLP, Spark, PyTorch, Git, Flask, Pandas, NumPy, Scikit-learn, Hadoop MapReduce, A/B Testing