Shruti Houji

812-778-4724 • Austin, TX 78660 • sghouji@iu.edu • LinkedIn • GitHub • Portfolio

SUMMARY

- **Experienced data analyst and data science professional with 7+ years of work experience in exploratory data analysis, predictive modeling, data visualizations, implementing data models, data pipelines, and cloud platforms.**
- **Proficient in Python, SQL, R, Tableau, ETL pipeline, BigQuery, Spark and ad hoc data analysis to derive insights.**

WORK EXPERIENCE

PayPal 10/2024 – Present

Data Science Analyst

- Executed SQL-based live data analysis on **200 million Venmo's Bad Tag Dataset** in **BigQuery and Tableau**, cutting risk response time by **86%** and enabling proactive **detection of spikes** in Bad, Dormant, and noSuccess accounts across tiers.
- Built an ETL pipeline with **Databricks and Spark** to analyze **6-month account transition trends** bad to historical bad, identify risks, and improve transition rates by **65%**.
- **Improved** BigQuery performance by designing partitioned and clustered tables, enabling faster data retrieval and reducing costs for billion-record datasets using **Python**.

Project 990 - *Remote* 06/2024 –09/2024

Data Scientist & Analyst

- Conducted extensive **Exploratory Data Analysis** on **200+ GB IRS** tax data, extracting insights on philanthropic donations across sectors, identifying financial discrepancies, metric correlations, trends, improving overall data quality using **Python**.
- Automated the **ETL process** by developing a robust text pipeline using **AWS Glue, SQL**, and **Python**, processing **270,000** records efficiently, which significantly enhanced data throughput and improved overall processing efficiency
- Fine-tuned the **RoBERTa model** using **AWS SageMaker** with DeepSpeed to enhance processing speed and employed Keras Tuner for hyperparameter optimization, resulting in an **85% accuracy** in extracting real-time key phrase-relevance scoring.
- Integrated **Tableau with AWS Redshift** to analyze and visualize data, creating **heat maps** that identified high dropout regions, enabling targeted resource allocation and improving coverage across U.S. states by **35%.**

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

Data Analyst

- Leveraged **Power BI and R** to analyze maintenance request data using clustered column and time-series charts with **seasonal decomposition**, identifying **trends** that optimized resource allocation and reduced maintenance task completion time by **60%.**
- Examined labor data using **seaborn** radar, scatter plots, uncovering workload imbalances, improving task distribution by **21%.**
- Streamlined workflows by integrating **50+ GB** of data into **Snowflake** and building **ETL pipelines in Python**, utilizing data cleaning, normalization, and **SQL** for modeling, which significantly improved data processing efficiency.
- Automated the status of order reporting process using **Excel VBA script** and **macros**, reducing daily report creation time by **30 minutes** each morning and improving the report generation process to merge, format, rename, separate, and sort columns.
- Addressed productivity issues by crafting daily, weekly, monthly **KPIs** for 7 assembly lines using **DAX and Power BI**, comparing productivity targets with actual employee productivity, resulting in a 70% improvement in productivity efficiency.

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

Data Analyst

- Collected and integrated customer data for insurance documents using **Azure Data Factory**, creating and managing **5+ data pipelines** to streamline data flow using **Agile** methodologies.
- Boosted filtering efficiency by **48%** for criteria-specific policies by employing optimized SQL queries and leveraging **Azure SQL Database** within **Azure Synapse Analytics**, improving data querying and filtering performance.
- Applied **statistical modeling** with logistic regression using Python to predict policy lapse or not lapse with a precision of **82%**.
- Leveraged **Azure Databricks** to integrate **k-means clustering and DBSCAN**, analyzing policyholder details to reveal causes for unclaimed policies, resulting in a **32% reduction** in policy lapse rate.
- Created visualizations using **Power BI** integrated with **Azure Synapse Analytics** and **HDInsight** to track classification metrics and compare unprocessed versus processed documents, improving **OCR efficiency by 30%**.
- Co-ordinated **Jira** task updates with a **10-person** multi-functional team, demonstrating strong problem-solving, interpersonal, and communication skills, while leveraging **Looker**, **HTML** & **JavaScript** webpages to improve customer service satisfaction.

SKILLS

- **Programming Languages**: C++, Python, R, DAX
- **Data Analysis Tools**: Microsoft Power BI, MS Excel VBA, Tableau, Looker, Matplotlib, QlikView, Seaborn, ggplot, SPSS
- **Databases & Web Technologies**: SQL, NoSQL, PostgreSQL, MongoDB, BigQuery, Snowflake, HTML5, JavaScript, CSS
- **Cloud Platforms**: AWS, Google Cloud Platform (GCP), **Microsoft Certified** - (Azure Data Fundamentals)
- **Competencies**: Keras, Tensorflow, Numpy, Pandas, Scikit-learn, Hadoop, Regression Analysis, Hypothesis testing, A/B Testing, Key Performance Indicators, Mathematics, Github, multivariate testing, Spark, SAS, generative AI

PROJECTS

Skin Lesion Classification for Melanoma Detection ***AWS(S3, SageMaker, Lambda, IAM), XGBoost, Flask***

- Spearheaded an end-to-end skin cancer detection system using the XGBoost algorithm with 74% accuracy, leveraging AWS.

University Admission prediction for Big Data ***Big Data Analytics, Hadoop, PySpark, Linear Regression***

- Capitalized on diverse data sources for Big Data Analysis of admission of students in the university using PySpark and Hadoop to transform large datasets and linear regression model for prediction, achieving an impressive R2 score of 0.80.

EDUCATION

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