



RAGHAV SHARMA

Contact: 571-274-1930 | [LinkedIn](#) | raghav.sharma.240796@gmail.com | [Portfolio](#) | [Github](#)

SKILLS

Programming Languages: Python, R, MySQL

Libraries: Pandas, Numpy, Matplotlib, Seaborn, Plotly, Altair, NLTK, Spacy, Sklearn, Keras, Tensorflow, Pytorch, Pickle

Machine Learning models: Regression, Naïve Bayes, KNN, Decision Trees, SVM, Random Forest, Adaboost, K-Means, Hierarchical clustering

Data Tools: MS Excel, AWS (EC2, S3), Azure BLOB storage, Amazon Redshift, Firebolt, Oracle SQL Developer

EDUCATION

Master of Science, Data Science and Analytics | Georgetown University

May 2024 | Washington, DC

Relevant Courses: Statistical Learning, Advanced Data Visualization, Deep Learning, Big Data and Cloud Computing, Natural Language Processing

Bachelor of Engineering, Electronics & Communication | Thapar Institute of Engineering and Technology

May 2019 | Patiala, India

Relevant Courses: Data Structures and Algorithms, Data Analytics, Machine Learning, Optimization

WORK EXPERIENCE

Graduate Research Assistant | Lab for Globalization and Shared Prosperity

May 2023 – May 2024 | Washington, DC

- Curated a comprehensive dataset of 40,000+ SEC 10-K filings spanning from 1993 to 2023 by web scraping and utilizing SEC APIs.
- Engineered a comprehensive ETL pipeline for labor risk forecasting in manufacturing firms, incorporating advanced NLP and supervised learning techniques, enhancing data accessibility and processing efficiency.
- Achieved 86% test accuracy through hyperparameter optimization across five classification algorithms, enhancing predictive reliability.

Data Engineer | Sigmoid Analytics

Aug 2021 – Jul 2022 | Remote (India)

- Facilitated seamless management of over 1 million rows and integration with Apache Superset for real-time analytics ensuring data modeling integrity by engineering a robust DB-API, Dialect, and Service for SQLAlchemy adapter customized for Firebolt cloud data warehouse.
- Integrated 'firebolt-python-SDK' with Prefect for executing complex SQL queries within Firebolt databases by orchestrating 2 ETL pipelines to optimize data processing and scalability.
- Attained improved software reliability by conducting over 100 use cases in smoke testing across firm projects, ensuring bug identification and resolution.

Software Engineer | HSBC Software Development India (HSDI)

Jul 2019 – Aug 2020 | Pune, India

- Achieved a notable 75% reduction in time for capturing transaction details by delivering Trade Transaction Posting utility API, a REST API utilizing layered infrastructure with Spring, Maven, and Java.
- Orchestrated Awaiting Approval Requests (AWPR) (REST API) using Clean Java MVC Architecture to automate transaction approvals within HSBC policy limits, reducing manual approval tasks for HSBC employees.
- Delivered Inbound Payments API leveraging JDBC, JSP, Servlets, and RESTFUL-Web Service, streamlining global internal transfers (HSBC to HSBC) and achieving a remarkable 90% reduction in transaction processing time.

PROJECTS

Reddit Analysis | Big Data, EC2, S3, Sagemaker, Azure, Spark, Python, NLP, ML [Project Website](#)

Sept 2023 – Dec 2023

- Analyzed 2M Taylor Swift subreddit comments and Spotify data using AWS EC2, S3, Sagemaker, and Microsoft Azure to analyze user engagement trends, time series analysis, and Taylor Swift's music's emotional nuances by creating comprehensive summary tables, visualizations, and machine learning models.
- Implemented a sentiment analysis model utilizing Spark NLP on Reddit comments, unveiling community sentiments and nuances, and designed a cyberbullying classifier, identifying 11% of comments as sexist, prompting discussions on fostering diversity and inclusion.
- Applied predictive modeling techniques to anticipate controversial comments and popularity, providing valuable insights into the reception of Taylor Swift's music on digital platforms.

Demarcation Line Detection | Tensorflow, Pytorch, UNET, Deep Learning, Computer Vision, Python [Project Website](#)

Sep 2023 – Dec 2023

- Directed automation of demarcation line identification in post-corneal cross-linking (CXL) surgery, managing a diverse dataset of 939 eye scans from 61 anonymized patients, sourced from Elza Institute in Switzerland and AUBMC in Lebanon.
- Deployed a foundational U-Net model for biomedical image segmentation, incorporating advanced practices like model checkpointing and learning rate scheduling.
- Leveraged Weights and Biases software for tracking and analysis of model progression, culminating in promising outcomes and a significant leap forward in computer vision applications for ophthalmology.

Soccer Match Summarization | Spacy, NLTK, Generative AI, NLP, Streamlit, Python [Project Website](#)

Mar 2023 – May 2023

- Orchestrated end-to-end data pipeline to collect 10K+ UEFA Champions League 2020–21 commentaries from goal.com that creates 15-minute intervals for each commentary, conducts text processing, data manipulation, and visualization of team formations for key insights.
- Implemented extractive summarization with Spacy and NLTK, alongside abstractive summarization utilizing GPT3, Pytorch, and Hugging Face Transformers library, and evaluated summarization quality via ROUGE scores for comprehensive assessment.
- Developed and deployed a real-time Streamlit application leveraging the data pipeline to encapsulate match highlights, offering insights into soccer matches listed on goal.com.

UFO Sightings Analysis | Data Visualization, Plotly, Altair, Highcharts, Python [Project Website](#)

Mar 2023 – May 2023

- Analyzed 80,000+ UFO sightings using the National UFO Reporting Center (NUFORC) dataset spanning from 1973 to 2022 to identify trends and patterns in UFO sightings across the United States.
- Uncovered a 1200% increase in sightings, highlighting significant temporal shifts and correlations with technological advancements and social media proliferation, delivering impactful observations into potential societal influences on reporting trends.
- Crafted 10+ interactive visualizations utilizing Plotly, Altair, and HighCharts showcasing temporal trends, geospatial distributions, weather impacts, and regional/shape patterns in UFO sightings.