SHAHRUKH SOHAIL

srk.soh7@gmail.com | [redacted] | LinkedIn | GitHub | Website | Chicago, IL

SKILLS

Programming Languages: Python, SQL, NoSQL Databases, GIT, R, Java, Scala.
Frameworks: Pytorch, Tensorflow, Transformers, Keras, Pandas, NumPy, Scikit-learn, NLTK, OpenCV, Beautiful Soup.
Deployment: AWS, GCP, Docker, Kafka, Kubernetes, Fast API, Heroku, Flask, Django, Apache, Snowflake.
Statistics: ANOVA, Hypothesis Testing, A/B Testing, Statistical Analysis, Time Series Analysis, NLP, Text Mining.
Data Engineering: ETL, Hadoop, Spark, Airflow, Kafka, Cassandra, MongoDB, BigQuery, S3, EC2, Storm, Flink.
Data Analysis: Tableau, Power BI, R-Studio, dplyr, ggplot2, caret, tidyr, Shiny, Matplotlib, Plotly, Pillow, D3.js, Neo4j.
Machine Learning: BERT, GPT, Bloom, BigBird, T5, MobileNet, EfficientNet, Regression, Classification, XgBoost.
Certifications: Operations Analytics, Supply Chain Principles, Data Science Bootcamp, IBM Data Science Professional Certificate, Google Data Analytics Professional Certificate, Foundational Data, ML, and AI Tasks in Google Cloud

WORK EXPERIENCE

Pure Platform Chicago, IL
Machine Learning Intern Jan 2023 – May 2023

- Implemented an OCR-based solution to extract text data from 10TB of product images reducing manual effort by 75% and achieving a 95% accuracy rate in data extraction.
- Developed a Puppeteer-based E-Commerce Product Details Extractor in Node.js, reducing manual effort by 90% and achieving an impressive extraction accuracy rate of 98%.
- Built highly efficient and scalable Scrapy and Beautiful Soup pipelines for data extraction, resulting in enhanced overall productivity in data extraction workflows by 15%.
- Wrangled 10TB of trading data stored in hadoop distributed file system using scala to remodel and visualize16 previously inaccessible datasets to allow 500+ end clients to track risk management on their investments.
- Processed 25TB of unstructured data with ETL pipelines using Apache Airflow, automated data validation and monitoring, leading to a notable 20% enhancement in forecasting performance metrics.
- Fine-tuned BERT and GPT-3 models to predict product details from textual data, saving $200,000 in annual expenditure.

OMS Lahore, PK
Data Scientist Nov 2018 – April 2021

- Developed and deployed an interactive Tableau dashboard in production environment, saving 20 hours per week of manual reporting work and enabling data-driven strategy making for 2020-2021.
- Managed 500 GB data warehouse for the Information Systems department which included financial, project tracking and equipment stock data.
- Predicted future resource requirement using multivariate regression in order to reduce contractor spendings and saved $10k in costs.
- Worked closely with cross functional engineering teams to forecast demand and power production of a solar farm using regression and improved predictions by 40%.
- Led the design of a data intelligence tool for aggregating unstructured data from 7+ sources, enabling deep analysis.
- Lead a team of 6 in development of a recommendations engine which directly resulted in revenue growth by 15%.
- Designed a data pipeline architecture in a team of 5 for a new product, scaled from 0 to 100,000 daily active users.

EDUCATION

ILLINOIS INSTITUTE OF TECHNOLOGY Chicago, IL
Master of Science in Data Science Aug 2021 – May 2023
Minors: Computer Science and Statistics
Cumulative GPA: 3.92/4.0

GIK INSTITUTE Topi, PK
Bachelor of Science in Electronics Engineering Aug 2014 – June 2018

ADDITIONAL

Honors: Upsilon Pi Epsilon Recognition of Outstanding Talent in Computing at Illinois Tech
Languages: Fluent in English, Urdu; Conversational Proficiency in Hindi
Activities: Art Exploration, Photography, Swimming, Travel