Sitanshu Kushwaha Data Engineer | ML Engineer | Software Engineer

sitanshu.kushwaha@nyu.edu +1 (929) 643-4480 sitanshu.tech Brooklyn, NY
linkedin.com/in/sitanshukushwaha/ github.com/Sitanshuk

SKILLS

Big Data (Apache Spark, Kafka, Databricks, Azure Data Factory, AWS, GCP, Data Lake, ETL, NoSQL),
Machine Learning (Scikit Learn, Tensorflow, NLP, Neural Networks, Deep Learning),
Data Analytics (SQL, Pandas, Numpy, Matplotlib, Seaborn, Web Scraping, Tableau),
Languages (Python, R, C, C++, JAVA.), **Tools** (Git, Docker, Databricks MLOps)

EDUCATION

New York University, MS in Computer Science Sep 2023 – May 2025 | New York
Design and Analysis of Algorithms I, Big Data, Fundamentals of Data Science

Mumbai University, BE in Computer Engineering Aug 2016 – Nov 2020 | Mumbai
9.32/10 CGPA (Rank: Top 10)
Big Data Analytics, Machine Learning, Data warehousing & Mining, Artificial Intelligence & soft computing, Digital Signal & Image Processing, Cloud Computing, NLP, Distributed Computing, Software Engineering

PROFESSIONAL EXPERIENCE

Data Engineer, LTIMindtree Jan 2021 – Jun 2023 | Mumbai
Technical Lead, Visioncare MFF Data Engineering team - Johnson and Johnson
- Spearheaded **optimization** efforts in Databricks Spark code, resulting in a **30% reduction** in **execution time** for 50% of Transformation Jobs, significantly improving data timeliness.
- Championed the adoption of **event-based triggers** for ETL pipelines, leveraging **Azure Data Factory**, to handle **Big Data** from multiple sources.
- This strategic switch enhanced efficiency, **reduced costs by 25%**, and **mitigated unforeseen outages** by **40%.**

Machine Learning Engineer, Oniria Creations Mar 2020 – Dec 2020 | Remote, Poland
- Developed a high-accuracy CNN TensorFlow model to precisely identify Pet service provider websites from Bing search results, achieving an impressive **92% accuracy** rate.
- Automated the extraction of valuable CRM Data from service providers' websites using **NLP** and ML techniques, leveraging the powerful **BERT language model**.

Intern Machine learning Engineer, Toflo Fintech Consulting Dec 2019 – Mar 2020 | Mumbai
- Engineered a sophisticated **Recommender System** for their e-commerce platform, leveraging **Big Data analytics.**
- **Increased click-through rate by 33%** for products displayed on a recommended section of the page.

Python Developer Intern, Innolearn Solutions pvt. ltd. Dec 2018 – Jun 2019 | Mumbai
- Implemented a real-time web-scraping pipeline for BSE corporate announcements, achieving **95% accuracy** in classifying them into 50 categories using **ML**, surpassing the previous keyword matching approach.
- Designed a **user-friendly Website in Django** to interact with the extracted data and **hosted it on Heroku**.

PROJECTS

Subjective Answer Evaluation using Machine Learning, (NLP, Django, TensorFlow)
- Utilized state-of-the-art **NLP** techniques, including **BERT**, **USE**, and **Word2Vec** language models, to assess students' subjective answers by measuring **semantic similarity** against the teacher's answer.
- Published a **research paper** in International Journal for Scientific Research and Development

Analysis of scanned prescriptions, (Computer vision, OpenCV, TensorFlow)
- Utilized **computer vision** techniques like EAST for watermark removal, enhancing **text detection**, and **CNN** models for **text recognition** from scanned prescriptions.

Classification of Punjabi BBC Articles, (NLP, Keras)
- Developed a generic **Punjabi language model** using a public corpus for effective processing and understanding of Punjabi text, alongside a **sentiment classifier** with **87% accuracy** for categorizing news articles as political or non-political.

Geolocation Data Preprocessing and Clustering, (Unsupervised learning, Geospatial Data Analysis)
- Processed and cleansed **geolocation data**, ensuring data quality and reliability, and applied advanced **clustering** algorithms, including **K-Means** and **DBSCAN**, to analyze **proximity** and **density** patterns.
- Evaluated and **compared** clustering models' **performance**, providing valuable insights for optimal selection, and **visualized** geolocation **data** interactively to uncover and **analyze spatial patterns**.