Monika Mittal

Evanston, IL , USA • 224-432-8020 •Email • Inspire • LinkedIn • GitHub • Website

Bringing a rich blend of research experience from CERN and advanced skills in data science, I am a seasoned professional with over 5 years of experience in machine learning and big data analytics. Recognized with awards like the CMS Achievement Award and a contributor to key scholarly publications, my career is marked by a commitment to excellence and innovation. Eager to apply my analytical acumen in data science role, I aim to harness data insights for strategic business growth. I am authorized to work in the USA (EAD holder).

SKILLS

Data Science & AI: Machine Learning, Big Data Analytics, Computer Vision AI, Large Language Models.
Programming: Proficient in Python, SQL, C++, PySpark, Shell Scripting, GitHub, CD/CI, Docker.
Visualization & Cloud Technologies: Skilled in Matplotlib, Seaborn, Tableau, AWS Sagemaker, ROOT.
Additional Tools: LATEX, Keynote, Microsoft Excel.
Soft Skills: Analytical thinking, problem-solving, clear communication.

EXPERIENCE

Data Science Consultant (Contractual) **EULER RESEARCH LLC, IL, USA** Sept 2023 - Present

- Spearheaded the creation of an end-to-end document extraction and summarization pipeline, integrating advanced computer vision and NLP techniques, along with API implementation.
- Trained and fine tuned deep learning model to detect table boundaries in documents with 90% accuracy.
- Integrated optical character recognition and large language models for efficient extraction and summarization of tabular data.

Data Science Postdoctoral researcher **CMS Exeriment CERN, Switzerland** Sept 2020 - Aug 2023

- Engineered Python classes to handle pre-processing and cleaning of 10 pb (~10^{15} records) of proton-proton collision data and implemented selection to identify records.
- Innovate key features, crafted and deployed a Graph Neural Network (GNN) model using PyTorch for discerning Vector Boson Scattering (VBS) from non-VBS events, yielding a 10% sensitivity enhancement.
- A 10-fold reduction in processing time obtained through utilizing NumPy arrays for vectorization.
- Efficient parallel processing of massive dataset through deployment of GitHub's CI/CD to a Docker image.
- Extracted quantitative results with statistical hypothesis testing and a enhance in exclusion limits on dark matter production cross-section by 20% through implementation of category using maximum likelihood.
- Designed an automatic tailored Python-based algorithm and customized dashboard visualization using matplotlib to pinpoint the source of noise for the comprehensive testing of electronic sensors, resulting in an impressive 80% reduction. Presented results at ICHEP 2022 in Bologna, Italy.
- Co-ordinated and managed data-taking operations of CMS muon detectors, ON-CALL to carry out detector-related issues during commissioning and crucial initial phase.

Data Science Postdoctoral researcher **ATLAS Experiment CERN, Switzerland** Apr 2017 - Aug 2020

- Lead the team of researchers to strategize and execute data analysis for imbalanced dataset, successfully identified findings, lead to publications in prestigious scientific journals including Nature, PRL, and PRD.
- Developed and optimized a ML-based classification algorithm to detect and classify VBS events resulting in accuracy surpassing 90% and improving efficiency by 50% relative to existing methods.
- Modeling of event process through simulations were used to develop a groundbreaking approach that correlates interference and quantifies uncertainty due to interference at around 6-7%.
- Curated and prepared simulation and real data to analyze the correlation between hit rates in the ATLAS Muon Spectrometer and LHC luminosity. Developed a linear regression model to enhance the understanding of detector behavior as a function of operational conditions. The model successfully predicts detector behavior for new operating conditions, aligning well with operational data.
- Formulated mathematical modelling of noise in cavern and forecasted to generate synthetic data.
- Expertly leveraged SQL for intricate data extraction and processing, managing voluminous and structured datasets encompassing diverse attributes.

EDUCATION

Ph.D. in Computational Physics

Punjab University, Chandigarh & CERN, Switzerland Jan 2011 - Feb 2016

Dissertation: A novel approach to Vector Boson Fusion (VBF) search analysis

- Proposed the conceptualization of the data analysis, employing BDT to enhance the efficiency by 5% .
- Employed the test statistic to successfully reject alternate hypotheses and results published in EPJC.
- First Indian recipient awarded with the CMS Achievement Award, CERN. Leveraged data mining techniques and time series analysis to monitor health and mapping of newly deployed sensors.

PROJECTS:

Univariate and Multivariate Time Series Forecasting : Aug,2023 - Present

- Time series forecasting using regression, ARIMA family models to predict the stock prices.
- Enhanced the metric (MSE) using SOTA deep learning methods for time series forecasting LSTM and transformers.
- Envisioned a streaming pipeline to make real-time predictions of stocks data.
    - Saved and access dataset via Cassandra cluster deployed on local machine.

Credit Card Fraud Mitigation Strategies using ML on AWS sagemaker using RAPIDS April 2023

- Designed and implemented a comprehensive suite of fraud mitigation strategies, including anomaly detection algorithms and risk scoring models, resulting in minimised fraud losses and increased efficiency.
- Trained machine learning-based system to detect and prevent credit card fraud, achieving an exceptional 85.9% F1 score with a precision and recall rate of 88% and 84%.

Sentiment analysis model for Twitter Data Oct 2022

- Devised and executed a data pipeline using Apache Cassandra, processing and storing a dataset of ~4000 tweets.
- Used NLTK to pre-process and train machine learning models, including SVC, Naive Bayes, and regression techniques to analyze the sentiment of tweets with an accuracy of 81%, extracting valuable insights from the data.

Statistical Analysis of COVID-19 Outbreak in Indian States Aug 2020

- Utilized descriptive statistics to calculate the percentage of infected individuals per state and the deceased percentage, providing insights into the severity of the outbreak in different areas.
- Estimated the reproduction number (RO) for various states in India, offering valuable insights into the transmission dynamics of the virus. Maximum RO value measured to be 1.79 for Gujarat.

COURSES/CERTIFICATIONs

- Time series analysis and Forecasting Jan 2024
- Python For Finance : Investment Fundamentals and Data Analytics Dec 2023
- JPMorgan Chase & Co. Quantitative Research Virtual Experience Program on Forage Oct 2023
    - Completed a simulation focused on quantitative research methods
    - Analyzed a book of loans to estimate a customer's probability of default
    - Used dynamic programming to convert FICO scores into categorical data to predict defaults
- Hands-on Machine Learning with AWS and NVIDIA July 2023
- Databases and SQL for Data Science by IBM Feb 2021
- Neural Networks and Deep learning by DeepLearning.ai Mar 2021
- Statistics for Data Science by IBM Mar 2021
- Applied Data science with python specialisation by University of Michigan Aug 2020