Saransh Gupta

A career driven professional, targeting opportunities with a reputed organization to leverage experience and diverse skill set to lead and drive strategic initiatives, in Data Science and Machine Learning domain and contribute towards organizational goals.

saransh.official.iitkgp@gmail.com | +91- 9530277421 | LinkedIn Profile

Executive Profile

- Dedicated professional offering over 2 years of experience in developing machine learning models and transforming data science prototypes into production grade solutions.
- Assessed strategies and validate modifications in machine learning models to enhance NLP systems continually, ensuring consistent improvement.
- Understanding of the concept of data science - advanced analytics, predictive modeler, machine learning algorithm in multiple technical and functional domains.
- Expertise in conducting full lifecycle analysis including data gathering and cleansing, deep dive advanced statistical analysis/modeling and recommendations to optimize performance.
- Skilled in leveraging Python, PyTorch, Transformers, BERT, scikit-learn, and TensorFlow to develop advanced machine learning models and optimize NLP systems.
- A focused individual with a zeal to learn and adapt to new technologies quickly; capabilities in managing critical situation.

Education & Credentials

- Indian Institute of Technology Kharagpur (2017 – 2022)
  B. Tech + M. Tech in Engineering Product Design, Industrial and Systems Engineering
  Grade: 8.09 / 10

Publications

- "ClotCatcher: A Novel Natural Language Model to Accurately Adjudicate Venous Thromboembolism from Radiology Reports" BMC Medical Informatics and Decision Making doi: 10.1186/s12911-023-02369-z
- S. Gupta et al. "Integrative Network Modeling Highlights the Crucial Roles of Rho-GDI Signaling Pathway in the Progression of Non-Small Cell Lung Cancer" in IEEE - JBHI, 2022, doi: 10.1109/JBHI.2022.3190038
- Entity-aware Question-Answer Extraction for Shopping Guidance, Amazon Machine Learning Conference

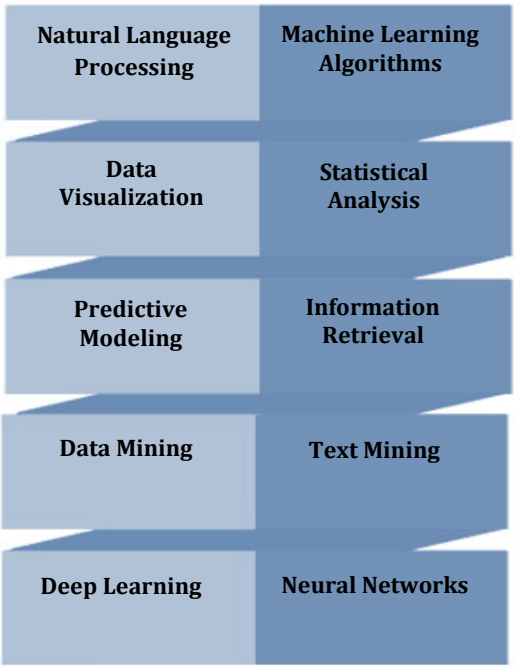
Achievements

- Conferred Blue-Award at the American Express for impactful contribution to the organization in Jan 2023
- Received scholarship of 248 USD for Harvard College Project for Asian International Relations conference - 2022
- Featured as one of the Top 30 Undergraduate Achievers of IIT Kharagpur in the UG Achievers Directory 2020
- Awarded scholarship of 2200€ by The A*Midex Foundation of Aix-Marseille University, France, Feb 2020
- Selected among Top 5 percent out of all for the summer fellowship at Institute of Science Technology Austria
- Featured in the ISE Newsletter Autumn-2020 under Department Spotlight of ISE fights COVID- 19, 2020

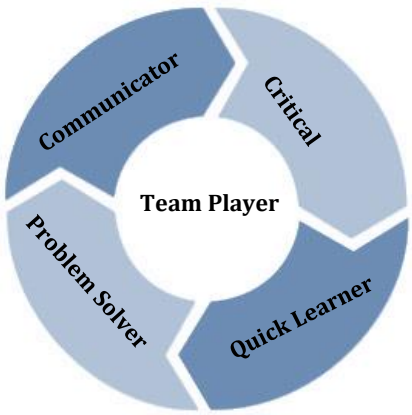
Links

- GitHub
- Google Scholar
- Website

Core Competencies



Soft Skills



Technical Skills

- Python
- PyTorch
- Transformers
- BERT
- Transfer Learning
- scikit-learn
- TensorFlow

Professional Experience

American Express | Engineer-III | Aug 2022 – Present

- Developing and implementing machine learning algorithms to enhance predictive analytics capabilities for optimizing business processes.
- Collaborating with cross-functional teams to gather & analyze large datasets, extracting insights to drive data-driven decision-making.
- Designing and maintaining scalable data pipelines, ensuring the efficient flow and transformation of data for various analytical purposes.
- Mapping client's requirements, performing system analysis and finalization of technical / functional specifications and high level design documents for the project.
- Coordinating with users for system study, requirements gathering, analysis & testing of applications & managing implementation of the same.
- Documenting business requirements, preparing process related documentation and coordinating for user acceptance testing.

Amazon Development Centre India | Applied Scientist – Intern | Jan 2022 - June 2022

- Developed machine learning models for data analysis and pattern recognition.
- Implemented and optimized algorithms to enhance the efficiency of existing systems.
- Collaborated with cross-functional teams to integrate machine learning solutions into product development.
- Conducted experiments and A/B testing to evaluate model performance and iterate on improvements.
- Utilized programming languages such as Python and frameworks like TensorFlow for model development.
- Assisted in the collection, preprocessing, and cleaning of data for training and validation purposes.

ZS Associates| Data Science Associate – Intern | Jan 2021 – June 2021

- Conducted exploratory data analysis on large datasets using Python and data visualization libraries to extract meaningful insights.
- Assisted in the development of machine learning models for predictive analytics, contributing to improvement in accuracy.
- Collaborated with cross-functional teams to clean and preprocess raw data, ensuring data quality and reliability for analysis.
- Created and maintained data pipelines, automating data retrieval and cleaning processes, reducing manual workload.
- Applied statistical techniques to analyze A/B test results, providing actionable recommendations for optimizing product features.
- Contributed to the creation of data-driven reports and presentations for stakeholders, summarizing key findings and insights.

Internships

Sapio Analytics | Data Analyst - Intern | April 2020 - June 2020

Project: Building a data-driven Decision Support System for the prediction of COVID-19 at the hyper-local level

- Fabricated SEIRD model using Migration, Lockdown conditions, reduced RMSE up to 4.8% by tuning parameters.
- The Govt. Of Telangana employed the application to mitigate COVID-19 while maximizing the economic activities.

Tools and Software: Python, MySQL, scikit-learn, SciPy, AWS Server, NumPy, pandas, Plotly, joblib

Research Experience

Emory University | Volunteer Researcher, Atlanta, GA, USA (Remote) | Jul 2022 – Aug 2023

Project: Predict the type of Venous thromboembolism (VTE), from the medical diagnosis and clinical Impressions

- Reduced manual adjudication of dataset by 20 times using pegasus paraphrasing model on sample dataset
- Achieved F1 score of 0.97 in predicting the type of VTE on test dataset by fine-tuning a Bio- BERT model
- Improved F1 score on test dataset by 20 percent by deploying paraphrasing and Bio-BERT finetuning pipeline

Tools and Software: Python, PyTorch, Transfer Learning, pegasus model, BERT

Osaka University | Research Assistant, Ibaraki, Osaka, Japan (Remote) | Jan 2020 - Dec 2020

Project: Predict Non-Small Cell Lung Cancer (NSCLC) using Machine Learning, identify potential drug targets

- Extracted 412 essential genes out of 10,077 by applying Boruta Feature selection on gene expression dataset
- Obtained F-1 score of 1.0 on validation, 0.98 on test dataset by using the XGBoost model to predict NSCLC
- Predicted drug targets for the NSCLC by simulating a Bayesian Network Model on Rho-GDI signaling pathway
- Discovered methodology leads to an accurate treatment of the disease impacting 85% of the lung cancer

Tools and Software: Python, Pandas, NumPy, matplotlib, XGBoost, Bayesian Network, networkX

Competitions and Conferences

- The Harvard Project for Asian and International relations (HPAIR) – Hong Kong (SAR) - Aug 2023
- Annual Amazon Machine Learning Conference (AMLC) – Bengaluru, Karnataka - Aug 2022
- 23rd World Business Dialogue, Creation Lab at Evonik - Cologne, Germany - Jun 2022
- Amazon ML Summer School 2021: Offered PPI - Jul 2021
- International Conference on Human Interaction Emerging Technologies - Aug 2020
- Young Data Scientists annual meetup at Kaggle - days, Dubai World Trade Centre - Mar 2020
- Winner at Databuzz 2020 conducted by DoMS, IIT Madras - Jan 202

Personal Details

- **Date of Birth:** 20th Feb'1999
- **Languages Known:** English & Hindi
- **Permanent Address:** Rajasthan, India

REFER TO ANNEXURE FOR MAJOR PROJECTS

ANNEXURE

American Express
Tools and Software: Python, PyTorch, Pandas, NumPy, matplotlib, bash scripting, Linux

Project 1: Failure cause identification of applications on generated Incident for their automated resolve
- Implemented a Question-Answer based strategy on top of raw dataset to identify failure cause of applications
- Achieved F1 Score of 0.84 by fine tuning a pre-trained BERT based Question-Answering model

Project 2: Automation of various repetitive tasks to save the manual efforts
- Analyzed Incidents data, identified major issues in payment applications, recommended their automation
- Developed automatic PII data identification and encryption tool to improve the data security
- Reduced 12 business hours per month by automating the application availability report generation process
- Automated resolutions for certain repetitive Incidents saving on an average 2 business hours every day

Amazon Development Centre India
Tools and Software: Python, PyTorch, Transfer Learning, PAA, T5 Model, BERT, streamlit

Project: Generate Pre-Curated Question Bank (PCQB) by Question and Answer extraction from articles
- Developed a Transformers-based two-step model for Question Generation followed by the answer extraction
- Scrapped Texts, People Also Ask (PAA) questions and answers using queries related to the E- Commerce domain
- Achieved a Perplexity score of 82.3 on Question Generation by fine-tuning a T5 model on the PAA dataset
- Attained F-1 score of 0.79 on the answer extraction task by fine-tuning the T5-large model on the PAA dataset
- Deployed the two-step model pipeline on the streamlit-based demo web application that accepts user input

ZS Associates Jan 2021 – June 2021
Tools and Software: Python, PyTorch, Transfer Learning, Medline-Plus API, PubTator, Selenium, Snorkel

Project 1: Extract biomedical text dataset, identify entities, and classify if there exists a relation between entities
- Created a pipeline to extract texts from PubMed database, identifying entities using Selenium and PubTator
- Implemented Binary Classification rules, devised four labeling functions using bio-verbs, co- occurrence of entities
- Generated a training dataset utilizing the four labeling functions in Snorkel by applying the Weak Supervision
- Achieved F1 score of 0.88 on the test dataset in relation-classification by fine-tuning RoBERTa base model

Project 2: Identify the type of relationship between two entities if it exists from the result of the Project-1
- Created a new set of three labeling functions for relation-type identification by using the results of the project-1
- Attained F1 score of 0.83 on the test dataset using XGBoost Model followed by feature engineering