Ujan Dasgupta

✉ ujan@cmi.ac.in   in ujan-dasgupta
ujandasgupta   ☎ +91-8240717719

Profile

A dedicated Data Science student with a solid foundation in statistics, seeking to leverage my analytical prowess and academic background to excel in the role of Data Scientist. Equipped with a passion for uncovering actionable insights from data, I am committed to driving data-driven decision-making and delivering impactful solutions to real-world problems.

Education

2022 - 2024 (ongoing)
Tamil Nadu, India

M.Sc. Data Science

Chennai Mathematical Institute
SGPA (after semester 2): 8.19

2019 - 2022
West Bengal, India

B.Sc. Statistics (Honours)

Presidency University Kolkata
CGPA: 7.26

Internships

May 2023 - August 2023

Data Scientist Intern - Synergy Marine Group
Tools used: Python, pandas, GPT-4, MongoDB, Amazon S3, Pinecone, KNIME

Automated Email Analysis and Classification using Python, GPT-4, MongoDB, Amazon S3, and Pinecone on the KNIME Platform

– Created a searchable email database for domain experts to search relevant emails.

Domain Specific Document Parser for chatbot

– Led the development of a cutting-edge chatbot project at Synergy Marine Group, focusing on optimizing technical support and information accessibility.

– Spearheaded the standardization of PDF manual parsing for ship equipment, leveraging Adobe PDF Extract API, PDF miner, and Google Vision to extract and convert diverse content into embedded vectors.

– Implemented OCR techniques for image text extraction and fine-tuned token sizes to enhance search capabilities, resulting in a highly efficient chatbot system. Integrated Pinecone and MongoDB for vector and metadata storage, significantly improving information retrieval and user experience.

Projects

July 2023

Building a Hybrid Search System with SPLADE
Tools used: Python, Pandas, MongoDB, Pinecone, PyTorch
– Refined pipelines for embedding creation of short texts by incorporating a hybrid vector system of transformer based and token based embeddings in order to increase retrieval accuracy.
– Used SPLADE (Scalable Probabilistic Latent Analysis for Distance Estimation).

June 2023

Building an n-gram language model and word vectors using a subset of the English-wiki corpus
Tools used: Python, NLTK, scikit-learn
– Demonstrated corpus cleaning, tokenization, and verification of empirical laws like Zipf's law.
– Illustrated the creation of a 4-gram language model, including next word prediction and sentence generation, with conditional options based on POS tags.
– Constructed word vectors using Co-occurence Analogue to Lexical Semantics (CoALS), a method leveraging co-occurrence statistics. It also identifies words with similar meanings.

Jan 2023 - Feb 2023

Comparative Analysis of Decision trees, Naive Bayes and Ensemble models
Tools used: Python, scikit-learn, matplotlib, pandas
– Compared performance of these models on different datasets for both regression and classification.
– Performed feature engineering and hyperparameter optimization to get the best results possible.
– Carried out hyperparameter tuning for both models and cost complexity pruning for decision tree to improve predictive metrics and generalisability further.

Aug 2022 - Oct 2022
Link

Financial Literacy Among 18-35 year olds
Tools used: Python, NumPy, Matplotlib, pandas
– Conducted a survey to collect responses to assess financial literacy.
– Performed extensive data cleaning to remove invalid responses.
– Performed tests for Cronbach Alpha to check for internal consistency among the responses and used clustering techniques to group individuals with similar literacy score and the level of financial literacy.

Oct 2022 - Dec 2022
Link

Visualization Project (Dashboard Creation)
Tools used: R, ggplot, tidyverse, R-Shiny, shinyjs, shinydashboard, tidyquant, forecast | Chennai Mathematical Institute
– Created visualisations to better understand the change in honey production over the years in the US. [Report]
– Built a dashboard web app with `R Shiny` for the above. [Dashboard]

August 2023 - December 2023

Understanding Deep Fakes
Tools used: Python, OpenCV, MTCNN, Autoencoder, tensorflow
Guide: Prof. Sourish Das | Chennai Mathematical Institute
– Created deep-fakes of US Presidents Donald Trump and Joe Biden using autoencoder models

Technical Skills

Languages	`Python`,`R`, `MySQL`, `C`,`Java`, LaTeX
ML/AI	NumPy, Pandas, Matplotlib, Scikit-Learn, ggplot2, spaCy, RShiny
Misc	KNIME, MongoDB, Pinecone, AWS Textract, Tableau, Power BI

Key Courses Taken

Postgraduate
Machine Learning and ML theory, Deep Learning and Advanced ML, Statistics and Visualisation with `R`, Regression techniques, NLP, RDBMS and SQL, Python and Data Structures, Design and Analysis of algorithms

Undergraduate
Probability, Real Analysis, Linear Algebra, Sampling Distributions, Statistical Inference, Multivariate Analysis, Linear Models, Design of Experiments, Time Series Analysis, Econometrics, Survival Analysis and Bio-statistics, Stochastic Processes, Operation Research, `R`, `C`

Others
Financial Modelling using Python, Programming with Julia, FinTech: Foundations, Payments, and Regulations

Positions of Responsibility

CMI
2023
Teacher's Assistant
Serving as a TA to Prof. Rajeeva Karandikar.

IIT Madras
2023
Teacher's Assistant
Served as a TA for evaluating paper of Statistical Computing.

Presidency University
Kolkata
2022
Logistics Team
Was a member of the Logistic Team of Milieu '22 at Presidency University