Sam Clastine Jesumuthu

Email: samclastine.jesumuthu@gmail.com | Mobile: +44 7831 505015 | LinkedIn: https://www.linkedin.com/in/sam-clastine/

PROFILE

I am a highly skilled Data Scientist with an MSc in Data Science from City, University of London. My academic and professional journey is distinguished by top grades and a proven track record in innovative machine learning projects. As a Research Assistant, I have pioneered the use of Large Language Models for data visualization, and my experience as a Software Developer Intern and Data Science Intern has honed my expertise in Data Science and full-stack development. With advanced skills in Python, R, several data science libraries, and web development frameworks, I am eager to contribute to projects that require deep technical knowledge and creative problem-solving abilities.

EDUCATION

City, University of London, UK **MSc Data Science** Sep 2022 - Oct 2023

- Grade: Distinction
- Completed extensive coursework covering core data science disciplines including Principles of Data Science, Visual Analytics, Machine Learning, Neural Computing, Big Data, Natural Language Processing, and Computer Vision.
- Gained practical experience applying data science techniques, performing statistical analysis, creating data visualizations, and building proficiency in Python and R programming languages.
- Had the opportunity to collaborate with industry experts for my Dissertation on research in the field of Visual Analytics through Large Language Model (LLM). This allowed hands-on exploration of cutting-edge data science topics.

Karpagam College of Engineering, IN **BEng Automobile** July 2019 - June 2022

- Grade: First Class Distinction
- Best Academic Performance Award During 2019 – 2022
- Got guidance and mentoring from Data Science and Analytics Centre during 2021-2022
- Event Coordinator for “Dhruva 19” a National Level Cultural Techno Festival
- Student Placement Coordinator

WORK EXPERIENCE

Research Assistant, City University of London October 2022 - Present

- Conducted extensive literature reviews on state-of-the-art language model applications for data visualization task.
- Spearheaded research on leveraging LLMs to generate Vegalite Specification with 100% Valid JSON.
- Implemented Python for data visualizations and compared methodologies like Zero Shot, RAG, and Chain-of-Thought prompting using LLMs (GPT-4 Turbo, Mistral 7B, Deepseek Coder 6.7B).
- Optimized model performance through prompt engineering techniques (zero-shot-CoT, few-shot-CoT, chain-of-thought, baseline prompting).
- Developed a custom GPT evaluator to score JSON outputs based on subjective criteria, enabling more comprehensive evaluation of generated visualizations.
- Experimented with quantized and non-quantized models to explore trade-offs between performance and accuracy.
- Enforced custom outputs and constraints to better align with desired visualization specifications.
- Utilized libraries like llama-cpp and vllm for faster inference and improved efficiency during model deployment.

Software Developer Intern, Halleyx Feb 2022 - Aug 2022

- Designed UI/UX wireframes to plan application layout and user flows.
- Developed full stack application with NoSQL databases, OpenCV, Tesseract OCR engine, and Vue.js frontend.
- Led core feature development of document text highlight extraction using OpenCV and Tesseract.
- Utilized Agile development methodologies to iteratively build, test, and improve the document highlight extractor throughout the project lifecycle.

Data Science Intern, Data Science Analytics Center March 2021 - Jan 2022

- Worked on machine learning techniques including regression, classification, clustering, and recommendation systems
- Developed object recognition in video system using YOLO v4 for real-time detection and tracking

- Built speech recognition capabilities (text-to-speech and speech-to-text) using RASA NLP
- Completed time series forecasting project analyzing and modelling seasonal air pollution data
- Identified trends and cycles in pollutant levels over time using clustering algorithms
- Compared temporal patterns of different pollutants through clustering time series data
- Developed strong analytical skills and end-to-end knowledge of ML workflows through working on diverse projects

ACADEMIC PROJECTS

Data visualization (visual information) mediated through language July 2023 - Sept 2023

- The main goal of the research is to help user to visualize and analyse datasets through natural language.
- Employed state-of-the-art language models GPT-3.5 and GPT-4 to interpret text prompts and output Vega-Lite JSON specifications for the visualizations.
- Optimized model performance through prompt engineering techniques including zero-shot-CoT, few-shot-CoT, chain-of-thought and baseline prompting.
- Achieved 96% accuracy on the nvUtterances benchmark dataset using GPT-4 with zero-shot prompting, outperforming other techniques.
- Built a custom JSON Comparator evaluation metric based on Jaccard similarity to numerically assess output quality against ground truth.
- Implemented an intuitive frontend UI with Vue.js and Flask that allows users to easily create visualizations powered by the backend large language models.

Face Mask Detection on Video [Report] August 2023 - September 2023

- Developed a video-based face mask detection system to identify individuals without proper face masks in public spaces.
- Utilized models including HOG + MLP, baseline CNN, and ResNet-50 for face mask classification, achieving a test accuracy of 96% with ResNet-50.

Classifying Machine Reading Comprehension using SQUAD [Report] June 2023 - August 2023

- The goal of this study is to predict the passage and the question is answerable or not, this helps MRC models to verify or evaluate the correct answers.
- We had used Logistic regression and state-of-the-art models such as BERT, ROBERTA-Base, and ROBERTA-XLM-Base for feature extraction and training.
- The Best model was ROBERTA with Test Accuracy of 55% and F1 score of 35% for unanswerable questions.

Fake (Deceptive) Review Detection [Report] May 2023 - July 2023

- Spearheaded an in-depth study on fake review detection by assessing linguistic and structural characteristics of online reviews, using supervised learning algorithms such as Logistic Regression and Random Forest.
- Implemented extensive data preprocessing and feature engineering techniques including tokenization, lemmatization, and TF-IDF vectorization, enhancing the predictive accuracy of the models.
- Developed and refined machine learning models to distinguish deceptive reviews, achieving an accuracy of up to 80%. This involved rigorous testing and validation using advanced analytics to optimize performance.
- Uncovered significant trends in verified purchase data, revealing that 82% deceptive reviews are from unverified users, indicating key difference between deceptive and truthful reviews in terms of verification status.

PERSONAL PROJECTS

AIVA – AI for Visual Analytics [aiva.samclastine.me] May 2024 - Present

- Developed an AI-driven visual analytics application enabling users to visualize and analyze datasets through natural language. The project utilized a comprehensive tech stack for optimal performance and scalability. On the frontend, Vue.js was chosen for building interactive and dynamic user interfaces due to its flexibility and efficiency. For the backend, Flask was used to handle server-side logic and API endpoints, integrated with GPT-3.5 to interpret text prompts and generate visualizations. Flask's simplicity facilitated rapid development and deployment.
- LangChain is used to integrate Retrieval-Augmented Generation (RAG) technique and for creating chains for processing documents, prompt and memory, allowing for more sophisticated handling of queries and improving the accuracy and relevance of visualizations. Docker was used to containerize the backend, ensuring a consistent and reliable environment across different stages of development, testing, and deployment, thus streamlining the CI/CD process.
- AWS services played a crucial role in the deployment and operation of the application. AWS Amplify was used for deployment, providing a robust and scalable solution with its CI/CD pipeline for automatic builds and updates. Static assets and dataset files were stored in S3, leveraging its high durability and scalability. DynamoDB managed fast and scalable database operations, essential for performance. API Gateway secured and managed API endpoints, ensuring efficient communication between

frontend and backend. Additional backend processing and compute power were provided by EC2 instances, allowing for handling large data volumes and complex computations.

- The integration of these technologies resulted in a sophisticated application demonstrating modern web development and machine learning techniques, ensuring scalability, efficiency, and reliability.

KEY SKILLS

Programming Language: C, JavaScript, Python, R; **Database**: SQL, MongoDB (serverless); **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Scikit Learn, LightGBM, Scipy, Statsmodels, NLTK, Spacy, TextBlob, OpenCV, Pillow, Tesseract, Plotly, ggplot; Pytorch, Tensorflow, Pyspark, Darts, Transformers, skorch, skimage, langchain, llama-cpp, openAI, VLLM
Web Frameworks: Flask, Django, Vue.js, React.js, React-Native;
Cloud Platforms: AWS (EC2, Lambda, SageMaker, S3, CloudWatch), GCP, IBM Watson
Other: Node.js, HTML, CSS/SCSS, MATLAB, HDFS, Apache Spark, MATLAB, Linux, Tableau, Qlik Sense, Machine Learning Algorithms, Deep Learning Algorithms, GitHub, Docker, Anaconda
Spoken Languages: Tamil (Native), English (Fluent), Hindi (Fluent)

CERTIFICATION

- Qlik Sense Business Analyst Qualification Feb2020
- Deep Learning Specialization by Coursera
- Self-Driving cars Specialization by Coursera
- Advanced Machine Learning and Signal Processing by Coursera