# SUDHANSHU MANOHAR

*Machine Learning Engineer*

## EDUCATION

Indian Institute of Information Technology Vadodara (IIITV)

*Bachelor of Technology in Computer Science and Engineering* — ***2018 - 2022***

## EXPERIENCE

### Cognam

***JAN 2022 - JUL 2023***

*Machine Learning Engineer*

- **Data Ingestion**: Automated the development of scalable data pipelines using AWS Lambda, Apache Airflow, and AWS Glue to streamline the ingestion of structured and unstructured data from multiple sources.
- **Data Transformation**: Cleaned, processed, and transformed raw data into a usable format using AWS Glue and Python libraries like Pandas, ensuring high data quality and consistency.
- **Model Training**: Developed and implemented a classification model for fraud detection using PyTorch, leveraging historical data features to enhance model accuracy and predictive performance. Established a continuous training pipeline that automates model retraining based on new data availability and performance degradation, ensuring adaptability to changing data patterns.
- **Model Deployment**: Deployed and monitored machine learning models using FastAPI and Docker, implementing CI/CD pipelines with GitHub Actions for seamless model versioning and rollback capabilities. Utilized monitoring tools to track performance in production environments.
- **Monitoring**: Implemented monitoring solutions to track model performance and data drift, ensuring models remain accurate and effective over time with tools like AWS CloudWatch.
- **Orchestration**: Managed end-to-end orchestration of data pipelines and model workflows using Apache Airflow, ensuring smooth execution of tasks and reliable data flow.

*API Developer*

- **Webhook Integration**: Led the development of a service to integrate Webhooks functionality with multiple external partners, streamlining data exchange and enhancing system interoperability.
- **Re-architected Legacy Systems**: Modernized Java-based systems with contemporary architectural frameworks, leading to significant improvements in throughput, scalability, and security.
- **Microservices Implementation**: Led the design and deployment of microservices architecture using Spring and RESTful practices, achieving a 50% increase in system scalability and a 40% reduction in maintenance costs.
- **Performance Optimization**: Developed and optimized high-performance Java backend applications and RESTful APIs, reducing response times by up to 50% and decreasing customer complaints by 60%.
- **Cross-functional Collaboration**: Partnered with front-end teams to integrate UI components with JavaServer Faces and RESTful web services, resulting in enhanced user experience and seamless functionality.
- **API Enhancement**: Improved API performance and documentation, leading to a 20% increase in developer productivity, a 10% reduction in API errors, and a 35% improvement in response times through the implementation of new caching systems.

## INFORMATION


LinkedIn

sudhanshu.m3939@gmail.com

https://github.com

+91-9953618442


## SKILLS

### DATA ENGINEERING SKILLS

- **DATA PIPELINES & ORCHESTRATION:** APACHE AIRFLOW, AWS GLUE, AWS LAMBDA
- **DATABASE SYSTEMS:** MYSQL, POSTGRESQL, MONGODB
- **VERSION CONTROL & CI/CD**: GIT, GITHUB ACTIONS, DOCKER, DVC
- **BIG DATA PROCESSING**: APACHE SPARK, HADOOP, KAFKA
- **ETL TOOLS:** PANDAS, PYSPARK, POLARS, NUMPY

### DATA SCIENCE SKILLS

- **PROGRAMMING LANGUAGES:** PYTHON, JAVA, SQL
- **MACHINE LEARNING LIBRARIES**: SCIKIT-LEARN, PYTORCH
- **REINFORCEMENT LEARNING:** PYTORCH RL, GYM, TORCHRL
- **COMPUTER VISION:** OPENCV, VIT, TORCH VISION, PILLOW
- **STATISTICAL ANALYSIS:** SCIPY, STATSMODELS
- **DATA VISUALIZATION:** MATPLOTLIB, SEABORN, PLOTLY
- **NLP TECHNIQUES**: NLTK, HUGGING FACE TRANSFORMERS
- **EXPERIMENT TRACKING:** MLFLOW

### ML ENGINEERING SKILLS

- **MODEL DEPLOYMENT:** DOCKER, FASTAPI, STREAMLIT, FLASK
- **MONITORING & LOGGING:** AWS CLOUDWATCH, GRAFANA
- **MODEL COMPRESSION & OPTIMIZATION:** QUANTIZATION, PRUNING, LOW-RANK ADAPTATION
- **CLOUD PLATFORMS:** AWS SAGEMAKER, AWS (S3, EC2, RDS), GOOGLE CLOUD

## PROJECTS

### RAG Conversational Chatbot with PDF Search and Session Management

- **Conversational Chatbot**: Built a chatbot using the Retrieval-Augmented Generation (RAG) architecture to provide conversational responses with context-aware replies. Implemented chat history to maintain continuity of dialogue.
- **PDF Search with RAG**: Developed a PDF search feature using RAG, where document embeddings were created and stored in a vector database. The system retrieves relevant sections of the documents based on user queries and integrates them into chatbot responses.
- **Session Management**: Enabled users to create multiple, separate chat sessions, ensuring that each conversation is handled independently with its own chat history and context.
- **Technologies Used**: Leveraged libraries such as Hugging Face Transformers, Chroma for vector search, and integrated the chatbot into a frontend with Streamlit.

### YouTube Comments Search RAG

- **Data Pipeline**: Developed an end-to-end data pipeline using the YouTube Data API to extract comments from YouTube videos. Cleaned and preprocessed the text data using the NLTK library for tasks like tokenization, stop-word removal, and lemmatization. Stored the processed data in .parquet and CSV formats for efficient storage and retrieval.
- **Embeddings Creation**: Used the paraphrase-MiniLM-L6-v2 model from Sentence Transformers to generate embeddings for the comments, enabling semantic search functionality. Saved the embeddings in the Chroma vector database to enable fast and accurate query results.
- **API Development**: Built a RESTful API using FastAPI to allow users to pass a YouTube video URL and search query to retrieve relevant comments based on the search terms.
- **Deployment**: Dockerized the FastAPI application and deployed it on Google Cloud Run, ensuring scalability and ease of deployment.
- **Frontend**: Implemented a user-friendly frontend using Streamlit, where users can input a YouTube video URL and search query to see relevant comments in real-time.

### Fine-Tuning Mistral-7B GPTQ for Abstractive Text Summarization Using QLoRA

- **Summarization Expertise**: Fine-tuned the Mistral-7B GPTQ model using the QLoRA technique for abstractive text summarization, demonstrating proficiency in applying advanced NLP methods to generate concise and coherent summaries from extensive text inputs.
- **Efficient Model Training**: Implemented memory-efficient fine-tuning strategies, leveraging 4-bit quantization and low-rank adaptation to optimize the training of large language models, facilitating effective summarization on limited hardware resources.
- **Deployment and Integration**: Developed a real-time summarization application using Streamlit for the frontend and FastAPI for backend API services, showcasing the ability to integrate and deploy NLP models into production-ready environments for practical user interaction.

### Brain Tumor MRI Image Classification

- Employed PyTorch to create a brain tumor MRI image classification system through transfer learning, utilizing three deep learning architectures: AlexNet, VGG16, and VGG19.
- Refined the AlexNet, VGG16, and VGG19 models, achieving a remarkable 99% accuracy, with VGG19 emerging as a very proficient model.
- Carried out a thorough assessment utilizing various performance metrics including confusion matrices, precision, recall, and ROC curves.
- Utilized PyTorch's native features and external libraries to visualize model outcomes and performance metrics, enabling a comprehensive analysis.

## PAPER'S IMPLEMENTATION

- **Attention Is All You Need for English-Italian Translation:** Developed the Transformer architecture from scratch using PyTorch, focusing on self-attention mechanisms and multi-head attention to facilitate high-quality English-Italian translation.

- **An Image is Worth 16x16 Words:** Transformers for Image Recognition at Scale (ViT): Replicated the ViT paper using PyTorch for a FoodVision project, classifying images of pizza, steak, and sushi. Implemented the Vision Transformer model to analyze and recognize food items effectively.

## CERTIFICATES

- Machine Learning Specialization - DeepLearning.AI
- Supervised Machine Learning: Regression and Classification, Stanford Online
- Advanced Learning Algorithms - Stanford Online
- Unsupervised Learning, Recommenders, Reinforcement Learning - Stanford Online

## LANGUAGES

- ENGLISH
- HINDI
- JAPANESE

## INTEREST

- Competitive FPS Gaming
- Chess
- Tech News