# Karthik Murumulla

**+1 (716)808-4366 | karthikmurumulla20@gmail.com | Linkedin**
**Houston, TX.**

## SUMMARY

Ingenious Machine Learning Engineer with a Masters Degree in Engineering Data Science possessing strong foundation in linear algebra, statistics, probability theory. Extensively contributed towards the ML community taking up projects and building products in various domains including Supervised Learning, predictive modeling, Clustering, Reinforcement Learning, Deep Learning, CV, NLP, GANs, Generative AI, LLMs, RAGs, Diffusion Models, BigData Analytics on Apache Spark, Kafka, Snowflake. AWS Certified Developer Associate with strong skills in designing efficient ETL solutions and scalable software development and CI/CD on cloud. Professional Engineer who can Bridge the gap between complex data and clear action with exceptional collaboration skills and adept at interpreting intricate datasets and translating them into actionable insights.

## EDUCATION

**Master of Science in Engineering Data Science** | University at Buffalo | Buffalo, NY | GPA: 3.4/4.00 | **December 2023**

**Coursework:** Probability and Statistics, Statistical Data Mining, Machine Learning, Reinforcement Learning, Data Modelling and Query Languages.

## TECHNICAL SKILLS

**Languages**: Python, R, Scala, Java, SQL, Linux, Bash.
**Machine Learning Skills** : Regression analysis, Classification, Clustering, Supervised Learning, Unsupervised Learning, Reinforcement Learning, NLTK NLP, Scikit Learn, Encoder-Decoder models, Transformers, Diffusion Models.
**Databases and tools**: RDBMS SQLite, MySQL, PostgreSQL, NoSQL & Big Data- Hadoop, Kafka, Hive, Apache Spark, Spark Streaming.
**Cloud**: AWS (EC2, S3, Kinesis, DynamoDB, RDS, SNS, SageMaker, EMR, BedRock), Azure, Terraform, Snowflake.

## WORK EXPERIENCE

**Machine Learning Engineer | Quills.ai, Houston, TX** **Feb 2024 – Present**

- **Constructed a structured dataset containing 350,000 records** pipelining huge chunks of raw data from various sources. Pioneered the development of a GenAI tool called Quilla.ai allowing users to generate SQL queries with natural language prompts. Extensively worked with GPT-3.5, BERT, Flan T5 (base, small) to fine-tune transformers and underlying neural networks in LLMs.
- **Achieved a 72% improvement in model performance by employing clustering for data segmentation**, LoRA, soft prompts as a part of PEFT and refined prompts in accordance with anomalies in generated text**.**
- **Spearheaded the development of personalized language models** that revolutionized user experience by leveraging cutting-edge reinforcement learning with human feedback. Leveraged AWS Kinesis for real-time data ingestion and SageMaker for efficient model orchestration. Ensured seamless production rollouts with a blue/green strategy for zero downtime and instant rollback capabilities.

**Machine Learning Intern | Eitacies Inc. Santa Clara, CA** **June 2023 – Dec 2023**

- **Engineered high-velocity clickstream data pipelines** processing around 20 terabytes of data per leveraging Azure Event Hub for real-time data ingestion at 5 million events per second.
- **Developed backend un-supervised machine learning and Deep Learning algorithms** to leverage insights from collected data and enhance user experience by personalization, predictive search, anomaly detection, and promotion placement and dynamic pricing.
- Achieved 40% growth in customer engagement with the website and 24% tangential growth in overall sales of the company.

**Data Scientist | TCS Bangalore India** **Jul 2021 – Jul 2022**

- **Implemented end-to-end ETL processes using Apache Airflow** to extract, transform, and load data from diverse sources like relational databases, flat files, APIs. Designed and developed efficient, scalable ETL pipelines that ensure smooth data movement and processing. Leveraged unsupervised learning techniques like K-Means clustering, to automatically group system failures based on categories, enabling faster identification of recurring issues.
- **Enhanced Service Maintenance by reducing service resolution time by 40%** by providing real-time insights into service cluster behavior through intuitive dashboards. Facilitated proactive maintenance and troubleshooting, improving overall service uptime by 30%.
- **Fostered a collaborative environment as a key team player** with testing, code reviews, debugging, also effectively communicating complex data concepts to non-technical stakeholders, business and working cross-functionally to translate analytical findings into strategic initiatives.

## PROJECTS

**Customer Segmentation and Recommendation Engine** [ Deep Learning, TensorFlow, Apache Spark, Kafka]

- **Scalable Customer Segmentation with Machine Learning:** Engineered a customer segmentation system leveraging Apache Spark for large-scale data processing, resulting in categorization of distinct customer groups based on their purchasing behavior and preferences. Employed deep learning frameworks such as TensorFlow, PyTorch to deliver actionable insights from real-time streaming customer datasets.
- **Real-Time Recommendation Engine for Boosted Engagement :** Spearheaded the development of a hyper-personalized recommendation engine with a 92% click-through accuracy powered by Spark Streaming and Kafka for real-time analysis of customer behavior and purchase patterns. Witnessed a 45% increase in overall user lifetime value and a 3x boost in conversions for targeted product categories.

**Reinforcement Learning with Large Language Models (LLMs) for Conversational Agents** [RL, Hugging Face, LLM, DQN, A2C]

- **Pioneered a hybrid approach integrating Reinforcement Learning (RL)** with Large Language Models (LLMs) like GPT-3 for enhanced conversational agent capabilities. Leveraged RL algorithms like DQN, A2C and policy gradients to train the agent to select optimal responses generated by the LLM, improving dialogue fluency and user engagement.
- **Developed state-action-reward functions** to optimize replay buffer utilization and fine-tuned Large Language Models (LLMs) through RL-based feedback. Benchmarked a **63% increase in ROUGE scores**, signifying significantly improved precision in LLM in-context response generation.

**Real-Time Content Moderation AI system for Twitter** [ETL, Kafka, NLP, BERT, LLMs, AWS, DynamoDB, ElastiCache, Lambda]

- **Spearheaded the development of scalable ETL pipelines,** leveraging Kafka for real-time data ingestion from Twitter API. Engineered efficient data processing workflows to extract and preprocess text data for opinion mining on Twitter and its cross-platform societal impact analysis.
- **Pioneered the development of a robust content moderation system** utilizing pre-trained language models such as BERT within TensorFlow, ensuring the identification and flagging of potentially harmful content. Enhanced processing efficiency through optimized performance techniques like in-memory caching on AWS ElastiCache for real-time analysis of user-generated content.

## CERTIFICATIONS AND PUBLICATIONS:

- **AWS Developer Associate**
- "**Surface Roughness Prediction using Machine Learning Algorithms while Turning under Different Lubrication Conditions**".
 Citation: A Varun et al 2021 J. Phys.: Conf. Ser. 2070 012243.