Tanav Thanjavuru

tanavthan@gmail.com | (267) 467-1239 | Washington DC | https://www.linkedin.com/in/tanavt/ | https://github.com/tanav1

EDUCATION:

- **University of California, Berkeley** **May 2024 - Dec 2025**
  - **Master of Information and Data Science**
  - Relevant Coursework: Natural Language Processing, Deep Learning, Machine Learning Systems Engineering, Probability Theory, Applied Machine Learning, Data Engineering
- **The Pennsylvania State University** **May 2019 - Aug 2023**
  - **B.S. in Applied Data Science with Minor in Security & Risk Analysis**
  - Cybersecurity Application Focus, Information Sciences & Technology Certificate; National Security Agency (NSA) Letter of Recognition
  - Relevant Coursework: Data Structures and Algorithms, Neural Networks, Machine Learning, Databases, Programming for Big Data, Probability Theory, Natural Language Processing, Network Analytics

PROFESSIONAL EXPERIENCE:

Booz Allen Hamilton | Staff Data Scientist, Senior Consultant Washington DC | **Jul 2023 - Present**

Technology Stack: Python (Transformers, Tensorflow, Keras, PyTorch, Streamlit, NLTK, PySpark), Qlik, SQL, Amazon S3 Storage, Databricks

- Data Scientist supporting the Advanced Data Analytics Platform (ADVANA) for the Chief Digital and Artificial Intelligence Office (CDAO) under the Office of the Under Secretary of Defense (OSD), focusing on data science and machine learning needs.
- Led team of data scientists to develop a **Retrieval-Augmented Generation (RAG) pipeline,** enabling efficient and secure processing of sensitive PDF documents, improving data extraction and insight-generation capabilities.
- Utilized **VGGNet** to achieve 89% accuracy in detecting **deep-fake AI images**, incorporating **GradCAM** for model explainability and integrating with a custom Streamlit front-end. Also explored additional models including ElasticNet, ResNet-50, and transfer learning approaches for performance comparison.
- Implemented **T5-small and GPT-2** models on Databricks for text summarization and comparison tasks, enhancing the ability to distinguish between source truth and non-source truth content, thereby improving the efficiency and reliability of content verification processes by 18%.
- Assisting the FDA in launching a Gen AI challenge by leveraging domain expertise and technical ML/AI knowledge to effectively execute specific use cases.
- Building a Q&A chatbot for client databases, exploring models like **Mistral-7b, Dolly-12b, and Phi-2** to generate SQL queries for efficient data retrieval from databases, enhancing user interaction and data accessibility.
- Led sentiment analysis and named entity recognition on 8,000+ social media comments for the DoD client office, employing Python, and NLTK Vader Model to extract and analyze data, resulting in a 14% improvement in sentiment analysis accuracy.
- Developed a tool leveraging **LLaMA 405b** to evaluate outputs generated by large language models, providing robust assessments to minimize **hallucinations** and enhance response accuracy by analyzing content relevance, factuality, and coherence. Migrated solution to **Palantir Aritifical Intelligence Platform (AIP)** for production.
- Conducted on-site technical qualitative interviews with DoD clients to identify and extract key ML/AI use cases, enhancing project alignment and innovation.
- Engineered a solution utilizing **Google Gemini 1.5 Pro API,** fuzzy string matching, **PDFMiner**, and Streamlit UI to extract and summarize key points from 300-page unstructured government request for proposal (RFP) PDFs.

Lithia Motors Inc. | Data Science Analyst Medford, OR | **Jul 2021 - Apr 2023**

Technology Stack: Snowflake, Neo4J, Python, Microsoft SQL Server, Cypher, Azure Data Factory, SQL, Databricks, Git

- Developed a **TensorFlow-based** neural network to predict the time a vehicle would remain unsold at a dealership prior to its resale decreasing inventory holding costs and improving sales efficiency.
- Built and executed Database Comparison algorithm with **Python** which compares objects in different Snowflake database environments and SQL DDL scripts using **string matching and text comparison** to assist in Snowflake database migration and validation saving Enterprise Data Services team over $5000.
- Developed machine learning text comparison algorithm using **Natural Language Processing (NLP)** to compare vehicle trims extracted from different databases and predict correct models which optimized current workflow practices and accuracy.
- Designed and implemented an image enrichment algorithm to improve image quality on Lithia's e-commerce portal using Azure Databricks, Azure Container, AutoML Computer Vision, and parallel processing.
- Aided in migration and validation of database scripts from **Microsoft Azure Synapse Analytics** to **Snowflake** data warehouse.
- Wrangled and preprocessed large datasets to run through different machine learning models and language processing algorithms.
- Designed and built dealership profiling knowledge graphs using **Neo4J graph database** to provide insights into dealer's customer demographics, sales characteristics, and service loyalty.

Apoio Clinica | Machine Learning Engineer State College, PA | **Mar 2021 - Nov 2021**

Technology Stack: Python, R, Azure Cloud Services, IBM Cloud Pak for Data, IBM Natural Language Understanding (NLU), Git

'Apoio Clinica' is an IBM-sponsored project that built a software for mental health professionals using artificial intelligence to provide faster diagnostics, treatment, and identify patients who may be hard to treat by aggregating initial self-reported data and tracking patients' progress over time. **Runner up in Microsoft 2022 Imagine Cup Epic Challenge. Runner up in Nittany Artificial Intelligence Challenge.**

- Built machine learning model using **Python** and **Azure Form Recognizer** that analyzes patient in-take form data then extracts variables that are flagged as important risk factors which could affect patient health.
- Wrangled and cleaned JSON text data extracted from Azure Form Recognizer to be used in test models to provide meaningful patient insights.
- Worked with IBM Watson **Natural Language Understanding (NLU)** service to detect patient sentiment, risk factors, and behavioral patterns that would be flagged for review.

SKILLS AND CERTIFICATIONS:

- **Certifications:** Azure Data Scientist Associate (Microsoft), Tensorflow Developer (Google), Azure Fundamentals (Microsoft), Foundry & AIP Builder Foundations (Palantir), Generative AI Fundamentals (Databricks)
- **Programming Languages:** Python [TensorFlow, PyTorch, Keras, NumPy, Pandas, Sckit-learn, Matplotlib, PySpark, Seaborn, Streamlit, NLTK, Keras, Statsmodels], SQL, R, Cypher, JavaScript, HTML/CSS
- **Tools/Frameworks:** Django, React, Tableau, RStudio, Git, PowerBI, Databricks, Docker, Wireshark, Apache Spark, Linux, EC2 Computing, Docker, Kubernetes, Palantir Foundry + AIP
- **Databases:** MySQL, MongoDB, Microsoft SQL Server, Redis, Neo4J, Snowflake, Databricks Data Warehouse, PostgreSQL