RINDHUJA TREESA JOHNSON

Baltimore, MD **(Willing to Relocate)** ◇ 914-746-5465 ◇ rindhuj1@umbc.edu ◇ LinkedIn ◇ GitHub ◇ Portfolio

PROFESSIONAL SUMMARY

Data Scientist and ML expert with a mission to augment human potential through ethical and scalable solutions. Specialized in model deployment, supervised and unsupervised learning, computer vision, deep learning. Proficient in Python, TensorFlow, PyTorch, SQL, and MLOps tools. Skilled in building scalable ML pipelines, optimizing performance, and deploying models on cloud platforms like AWS, Azure, and GCP

SKILLS

Tools	SQL, Python, C++, MS Power BI, Tableau, MS Excel, Streamlit, C#, .NET, HTML
Cloud & Big Data	Google Cloud Platform, Snowflake, AWS, Azure, Databricks, Spark, Hadoop
Database Management	MySQL, MS SQL Server, Azure Data Studio, PostgreSQL
IDE & Project Management	Docker, Jupyter NB, Visual Studio Code, GitHub, Jenkins, Google Colab, SharePoint
Python APIs/Lib	Pandas, NumPy, Matplotlib, Seaborn, StatsModels, Sci-Kit Learn, PySpark, TensorFlow, XGBoost, PyTorch, NLTK, Keras, LangChain, Transformers, HuggingFace, Flask
Expertise	ETL, Data Analysis, Data Visualization, Data Management, Machine Learning, Big Data, A/B Testing, MLOps, Time-series Analysis, Feature Engineering, Automated ML Pipelines, Predictive Analytics, Model Development, Deep Learning - CNN, RNN, CI/CD, Model Evaluation, Model Deployment, Generative AI, Natural Language Processing, Hypothesis-testing, Scalable AI, Distributed Training, Cloud Computing, Reinforcement Learning with Human Feedback, Supervised Fine-tuning

EXPERIENCE

Junior Data Scientist, Leveragai Inc. Dec 2023 - Present

- Designed and deployed ML models for image captioning and context-aware retrieval using Salesforce BLIP, MS GIT, and ViT-GPT2, increasing metadata generation efficiency by 30% and reducing manual intervention in image annotations.
- Developed and optimized LLM pipelines by implementing vector embeddings, Retrieval-Augmented Generation (RAG), and transformer-based models, enhancing image-to-text system performance and scalability across large datasets.
- Improved deep learning-based feature extraction and caption generation by fine-tuning multimodal models using PyTorch, Hugging Face Transformers, and TensorFlow for enhanced model accuracy.
- Conducted in-depth data analysis on large-scale image datasets, applying preprocessing, feature engineering, and dimensionality reduction techniques to improve model accuracy and efficiency.
- Leveraged statistical analysis and A/B testing to evaluate ML model performance and recommend data-driven optimizations for improved business insights and decision-making.
- Reduced model deployment time and ensured real-time inference by integrating vision-based models into Azure AI and AWS, implementing MLOps best practices for scalable and efficient model serving.
- Designed and automated CI/CD pipelines for ML model training and deployment using Docker, Kubernetes, and MLflow, ensuring version control and reproducibility.
- Decreased data processing time by 30% by building end-to-end data pipelines for image and video processing, incorporating preprocessing, augmentation, and transformation techniques with OpenCV, Pandas, and NumPy.

Data Scientist, Smart Ecosystems Inc. Jan 2024 - Present

- Led a high-priority Reinforcement Learning with Human Feedback (RLHF) project, conducting in-depth qualitative analysis to identify and mitigate key loss categories, resulting in improved AI model refinement and performance.
- Enhanced AI model adaptability and reasoning accuracy by leveraging advanced data preprocessing, feature engineering (Pandas, NumPy), and machine learning techniques (Scikit-Learn, TensorFlow, PyTorch), improving automation and decision-making capabilities.
- Refined AI-generated content quality by developing optimized response strategies for Image Generation models, contributing to Supervised Fine-Tuning (SFT) efforts to ensure more accurate, human-like outputs aligned with user expectations.
- Conducted in-depth data analysis to identify patterns and optimize AI model performance, applying NLP techniques such as LangChain, Transformers, and NLTK to enhance model reasoning and response generation.
- Accelerated data-driven decision-making by 20% through the design and deployment of interactive Power BI dashboards, integrating real-time data from multiple sources for cross-functional teams to gain actionable business insights.
- Spearheaded AI model quality control improvements, achieving 95% accuracy in content generation through rigorous testing, validation, and implementation of advanced QA frameworks over a 3-month period.
- Designed and implemented scalable ML pipelines to automate the training, fine-tuning, and deployment of reinforcement learning models, optimizing workflow efficiency and model performance in production environments.

Social Media Analyst, Redwood Algorithms Jan 2021 - Dec 2022

- Drove a 200% increase in customer engagement by leading data-driven ad campaigns for five business clients, utilizing predictive analytics, customer segmentation, and ML models to optimize targeting and personalize marketing strategies.
- Improved lead generation and conversion rates by analyzing large-scale data sets from Google Analytics, Meta Ads, and Google Ads, applying A/B testing, time-series analysis, and clustering techniques to identify high-impact elements.
- Enhanced marketing ROI by implementing SQL-based data pipelines to automate performance tracking, integrating data from Google Analytics, Meta Ads, and internal databases, and reducing manual reporting time by 40%.
- Utilized data analysis and machine learning to enhance supply chain optimization, forecasting demand and reducing inventory costs by 15% over 6 months, leveraging tools like Python (Pandas, Scikit-Learn) and Tableau for insights.
- Strengthened strategic decision-making by designing interactive Power BI dashboards, visualizing key digital marketing metrics such as click-through rates, engagement trends, and ad spend efficiency, enabling cross-functional teams to adapt campaigns in real-time.

- Optimized campaign performance prediction models using Python (Pandas, NumPy, Scikit-Learn), machine learning algorithms, and NLP techniques, uncovering actionable insights that increased ad relevancy and customer retention.
- Facilitated data-driven decision-making by documenting and presenting campaign performance insights via Microsoft SharePoint, collaborating with stakeholders to align marketing efforts with data-backed strategies.

PROJECTS

CLV Prediction for Insurance Companies with Python, MS Power BI, AWS, SQL, Streamlit, and LangChain

- Achieved a 91% accurate CLV prediction model using Random Forest, implementing rigorous statistical hypothesis testing and regression analysis to optimize customer segmentation and personalized marketing strategies.
- Developed an interactive QnA Interface using LangChain and Google's Gemini LLM API, translating natural language queries into SQL for seamless non-technical access to complex data insights.
- Integrated AWS S3 storage with Python API access, establishing a scalable, cloud-based data pipeline to streamline CLV analytics and predictive modeling.
- Designed and deployed an interactive Power BI dashboard visualizing CLV for different insurance policies and premium tiers, enabling data-driven decision-making for targeted customer engagement strategies.
- Hosted a web-based CLV prediction platform, providing businesses with real-time forecasting tools to optimize customer retention and revenue growth.

Breast Cancer Prediction with Computer Vision using PyTorch

- Developed and deployed a deep learning model leveraging Convolutional Neural Networks (CNNs) with TensorFlow and Keras to classify histopathological images for breast cancer detection, aiding in early diagnosis and treatment planning.
- Engineered a robust data pipeline for preprocessing medical images, including resizing, normalization, augmentation like rotation, flipping, contrast adjustment, and grayscale conversion, ensuring improved model generalization and performance.
- Trained the CNN model on high-resolution histopathological image datasets, fine-tuning hyperparameters (batch size, learning rate, dropout regularization) to optimize classification accuracy and reduce false negatives.
- Achieved a left-skewed probability distribution in model predictions, indicating a bias toward detecting malignant cases, prompting further model refinement through class-balancing techniques, weighted loss functions, and ensemble learning to improve sensitivity and specificity.
- Evaluated model performance using precision, recall, F1-score, and AUC-ROC metrics, analyzing classification thresholds to ensure a high recall rate, which is critical for minimizing missed cancerous cases.
- Deployed the trained model in a scalable cloud-based environment (Google Cloud/AWS) using Flask/Streamlit for a user-friendly interface, allowing medical professionals to upload images and receive real-time probability scores for malignancy.

Automatic Traffic Light Management Application with AI sensors using C# in .NET framework

- Engineered an AI-powered traffic management system using .NET, C#, and MS SQL Server, achieving a 25% reduction in commute times, a 40% decrease in wait times, and a 20% drop in fuel consumption through adaptive signal optimization.
- Developed and deployed a scalable SQL Server database via Azure Data Studio, enabling real-time data storage, retrieval, and analytics, enhancing traffic signal responsiveness and congestion prediction.
- Implemented real-time traffic data collection using GPIO-connected sensors, facilitating dynamic traffic flow adjustments based on vehicle density and pedestrian activity.
- Leveraged Docker and DotNetEnv for containerized deployment, ensuring secure, modular, and scalable system implementation across diverse urban traffic environments.

Steam Review Analysis on Big Data using Apache Spark and Hadoop in Python

- Led a team of three data scientists in analyzing an 8GB dataset with over 7 million records using HDFS and PySpark, identifying a 40% surge in user reviews on the Steam platform in November 2021, providing key insights into user engagement trends.
- Implemented extensive data cleaning, preprocessing, and feature engineering using SparkSQL and PySpark, leveraging Spark DataFrames and RDDs for scalable processing and real-time interactive visual analysis.
- Developed a personalized game recommendation model utilizing the Alternating Least Squares (ALS) algorithm from Spark MLlib, enabling tailored recommendations for new players based on their gaming history, improving recommendation accuracy and user retention.
- Optimized the model's hyperparameters (rank, regularization, iterations) to enhance prediction accuracy, employing cross-validation and evaluation metrics such as Root Mean Square Error (RMSE) to fine-tune recommendations.

Portfolio Website Using HTML and CSS

- Escalated public visibility by 300% by developing a personal website, launched with Jekyll on GitHub that showcases career journey
- Drafted the website using HTML for content and CSS for styling and formatting

EDUCATION

University of Maryland Baltimore County (UMBC) — Baltimore, MD
Master of Professional Studies in Data Science — May 2024
GPA: 4.0/4.0

Pondicherry University — Puducherry, India
Master of Science in Physics — May 2022
GPA: 3.61/4.0 Valedictorian