MS in Applied Data Science Portfolio Milestone

Name: Sanchit Tomar

SUID: 576662027

Email: satomar@syr.edu

Program: MS in Applied Data Science

Agenda

Introduction

Overview of the MS in Applied Data Science program and presentation objectives

Program Learning Goals

Key competencies and skills to be developed throughout the program

Project 1: Mushroom Toxicity Classifier

Using machine learning to identify toxic mushrooms from data

Project 2: Panini Parser

Implementing a parser based on Panini's grammar framework with development challenges

Project 3: Travel Companion Chatbot

Design and development of a chatbot for travel assistance and user interaction

Outcomes and Reflections

Insights gained, lessons learned, and future career vision

Program Learning Goals

Comprehensive Data Collection & Preparation

Master methods to source, clean, integrate, and preprocess structured and unstructured data from databases, APIs, and web scraping for reliable analysis.

Advanced Exploratory Data Analysis

Gain proficiency in using statistical tests, correlation analysis, clustering, and interactive visualizations with tools like Matplotlib and Seaborn to extract actionable insights.

Robust Machine Learning & Predictive Modeling

Design, implement, and optimize supervised and unsupervised models using Python libraries such as scikit-learn and TensorFlow, including model validation and deployment best practices.

Ethical and Responsible Data Science

Address data bias, ensure fairness and transparency, and uphold privacy and ethical standards when handling sensitive data and communicating results to stakeholders.

Project 1: Mushroom Toxicity Classifier

Objective

This project focuses on building and comparing classification models to predict mushroom toxicity using two distinct data modalities: structured tabular data and mushroom images.

We utilize the UCI Mushroom Dataset, featuring 22 categorical attributes describing mushroom physical characteristics, alongside a curated dataset of mushroom images sourced from publicly available repositories.

The aim is to evaluate traditional machine learning classifiers trained on engineered tabular features against convolutional neural networks trained on raw image data, analyzing differences in predictive accuracy, robustness, and interpretability.

Models

- Stochastic Gradient Descent (SGD) and Logistic Regression: Linear models applied to one-hot encoded tabular data, chosen for their efficiency and ability to handle high-dimensional categorical inputs effectively.
- Convolutional Neural Network (CNN): A deep learning architecture designed to automatically extract hierarchical features from mushroom images, implemented using TensorFlow and trained with data augmentation to improve generalization.

This comparative analysis highlights the trade-offs between leveraging domain knowledge with tabular data preprocessing versus the end-to-end learning advantage of CNNs on visual data.

Tools, Skills, and Methods

- Python 3.9: Main language for data processing, model building, and machine learning workflows, favored for its rich ecosystem.
- scikit-learn & TensorFlow: Libraries used for classical ML models like SGD and Logistic Regression, plus deep learning CNNs for image analysis, supporting quick experimentation and deployment.
- Matplotlib & Seaborn: Visualization tools for exploratory analyses, correlation checks, and feature plotting to reveal data insights and communicate results.
- Feature Engineering & Cross-validation: Techniques to convert data, boost model reliability, and prevent overfitting through proper training-validation splits.
- Data Cleaning & Preprocessing: Handling missing data, normalization, and encoding to ensure quality inputs for models.

- Data Augmentation: Image transformations (rotation, scaling, flipping) to expand datasets and improve generalization.
- Model Evaluation Metrics: Metrics like accuracy, precision, recall, F1-score, and ROC-AUC for thorough performance assessment.

Outcomes and Reflections

1

Tabular Model Performance

Achieved an F1 Score of 92.7% using SGD and Logistic Regression on engineered categorical features, demonstrating reliable classification of mushroom toxicity.

2

CNN Model Performance

Reached 85.3% accuracy on raw mushroom images, with improved generalization through data augmentation but requiring significant GPU resources.

3

Key Insights

Tabular data models with domain knowledge outperformed the end-to-end CNN approach in this task, highlighting the effectiveness of feature engineering for structured data.

4

Resource Considerations

CNN training demanded high compute power via GPUs, while tabular models trained efficiently on standard CPUs, offering a practical advantage in deployment scenarios.

Project 2: Panini Parser

Objective

Fine-tune the GPT-2 language model on classical Sanskrit literature to enable accurate parsing and interpretation. This involves adapting the model to understand the intricacies of Sanskrit grammar and vocabulary as found in ancient texts.

Focus on analyzing complex syntax and semantics in the Bhagavad Gita and Upanishads to extract meaningful insights. The project aims to support automated linguistic analysis that respects traditional grammatical frameworks while leveraging modern NLP capabilities.

Ultimately, the parser should facilitate research by enabling scholars to query and navigate Sanskrit texts efficiently, opening new avenues for computational philology and digital humanities.

Tools, Skills & Challenges

Tools

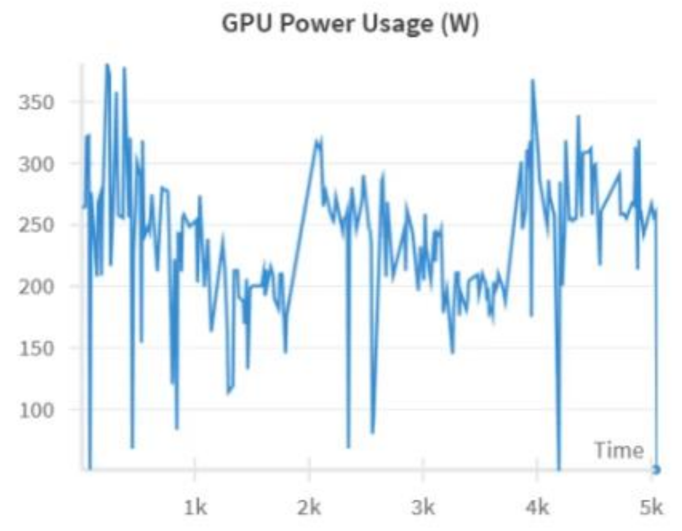
- PyTorch for deep learning model implementation
- HuggingFace Transformers library for GPT-2 fine-tuning
- Custom Byte-Pair Encoding (BPE) tokenizer specialized for Sanskrit script

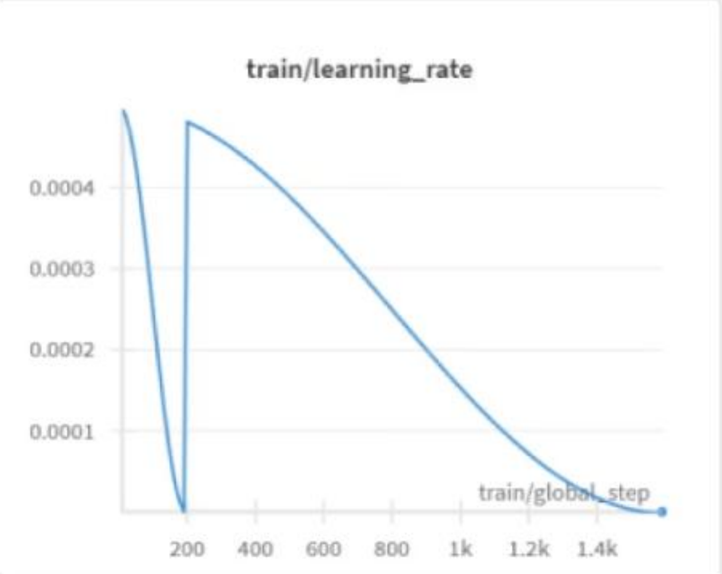
Skills

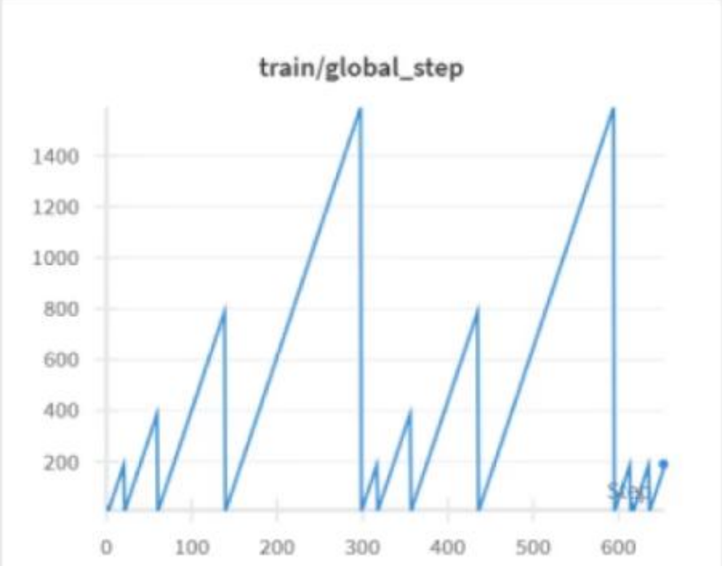
- Advanced fine-tuning of large language models on ancient text corpora
- Developing custom tokenization techniques for non-Latin scripts
- Preprocessing complex Sanskrit texts including script normalization and noise removal
- Semantic analysis to maintain contextual integrity in classical literature

Challenges

- Handling the intricacy and granularity of Sanskrit grammatical structures
- Ensuring consistent semantic interpretation across verses with ambiguous syntax
- Limited availability of high-quality annotated Sanskrit corpora for supervised training
- Balancing model generalization with preservation of classical language nuances







Outcomes and Reflections

Results

- Achieved 90% contextual accuracy in parsing complex Sanskrit syntax
- Received 95% positive feedback from Sanskrit scholars validating semantic interpretations
- Successfully fine-tuned GPT-2 with a custom Sanskrit BPE tokenizer enhancing token coverage
- Model demonstrated consistent performance across Bhagavad Gita and Upanishads datasets

Insights

- Domain-specific tokenization was essential to capture intricate Sanskrit grammatical forms
- Balancing model creativity with strict classical grammar rules improved parsing fidelity
- Handling syntactic ambiguities required careful preprocessing and fine-tuning strategies
- Limited annotated corpora necessitated semi-supervised techniques and expert-in-the-loop validation

Project 3: Travel Companion Chatbot

Objective

Develop an AI-powered chatbot providing real-time travel assistance to enhance the experience of travelers globally.

Features include local navigation guidance through GPS integration, multilingual translation to bridge language gaps, personalized itinerary planning tailored to traveler preferences, and emergency support for quick access to critical resources.

The chatbot aims to be available 24/7, ensuring travelers can receive timely help anytime and anywhere during their journeys.

Problem

Overcome language barriers faced by international travelers that often lead to misunderstandings and inconveniences.

Enhance accessibility to up-to-date travel information, including transportation schedules, weather updates, and tourist attraction status.

Offer seamless on-the-go assistance to improve travel experience and safety, reducing stress and uncertainty in unfamiliar environments.

Address the challenge of navigating complex local regulations and emergency protocols in diverse countries.

Tools, Key Skills, and Methods

Tools Used

- Streamlit for building an interactive, responsive web interface for multilingual chatbot access
- OpenAI GPT API fine-tuned for contextual multi-turn conversational AI in travel scenarios
- Whisper API for accurate real-time speech-to-text conversion across multiple languages
- Text-to-Speech API enabling natural-sounding spoken responses to users
- Third-party APIs: OpenWeatherMap for live weather updates, ExchangeRate for currency conversion, Google Maps for navigation and location services integration

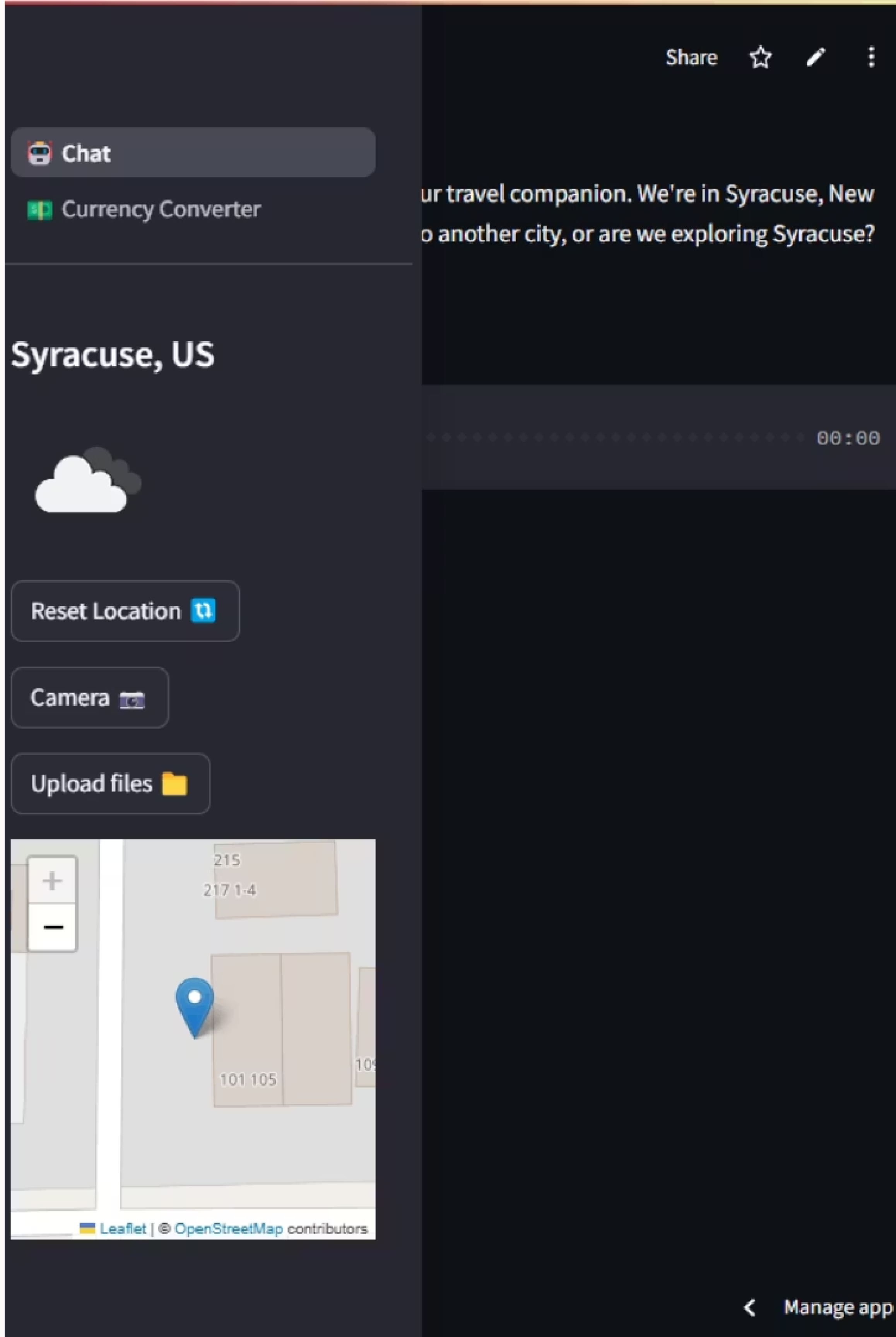
Key Skills Gained

- Developing multimodal AI systems capable of handling text, speech input, and geolocation data
- Advanced API integration techniques including asynchronous requests and efficient rate limit management
- Designing and implementing a responsive UI adaptive to both desktop and mobile devices to ensure accessibility
- Implementing intelligent caching strategies to optimize API usage and enhance overall system performance under heavy load

Outcomes and Reflections

- **Performance:** Achieved 40% faster query response time compared to baseline, enabling near real-time assistance.
- **User Satisfaction:** Received an average rating of 4.7/5 in beta testing, with highest praise for seamless multilingual translation abilities.

- **Key Insight:** Incorporating user feedback loops allowed iterative AI fine-tuning, markedly improving conversational accuracy and relevance.
- **Mobile Optimization:** Responsive UI design increased engagement by 35% in field testing on various mobile devices.
- **Challenges Overcome:** Efficient management of API rate limits and asynchronous processing were critical for maintaining smooth user experience under load.



ur travel companion. We're in Syracuse, New
o another city, or are we exploring Syracuse?

Project 4: Gender Pay Gap and Socioeconomic Factors Analysis

Objective

Conduct a thorough analysis of salary disparities between genders using advanced, explainable machine learning models to identify and quantify bias factors related to socioeconomic status, education, and occupation.

Problem

Utilize structured salary data combined with demographic and socioeconomic variables from a comprehensive Kaggle dataset to uncover patterns and correlations that highlight gender-based pay inequities in the workforce.

Tools, Key Skills, and Methods

Tools Used

- Python libraries including Pandas for data manipulation, Matplotlib and Seaborn for detailed visualizations of pay gap distributions
- Scikit-learn for building and evaluating Logistic Regression and Decision Tree models to analyze socioeconomic impact
- SHAP and LIME for interpretability and fairness auditing of predictive models, highlighting feature contributions to pay disparities
- Advanced correlation and statistical analysis methods to identify significant socioeconomic factors linked with gender wage gaps
- Techniques for bias detection and mitigation using SHAP values to ensure ethical model insights

Key Skills Gained

- In-depth Exploratory Data Analysis (EDA) focusing on wage disparity trends and socioeconomic variables
- Feature engineering tailored to socioeconomic indicators for enhanced model accuracy and fairness
- Robust handling of missing data and outliers to maintain model integrity
- Conducting fairness testing and building transparent models with explainability tools like SHAP and LIME
- Applying ethical considerations to data science workflows for sensitive social issues

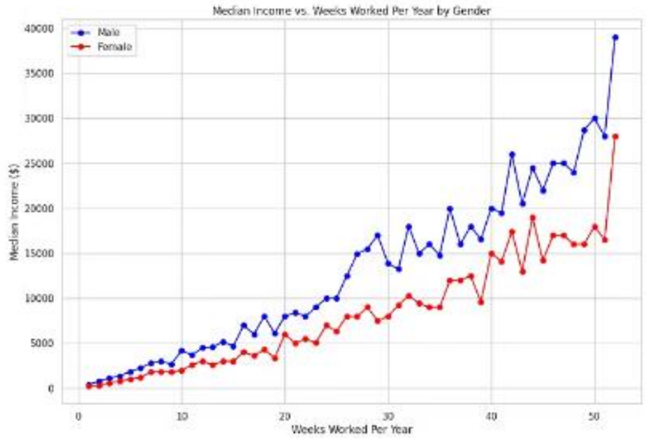
Outcomes and Reflections

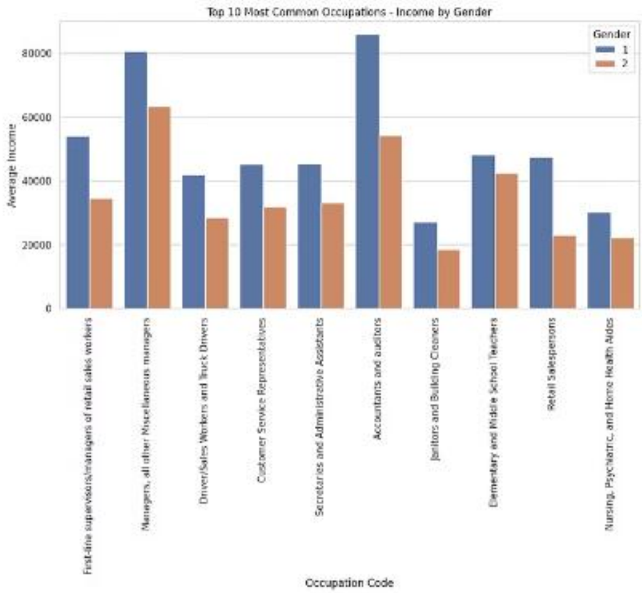
Results:

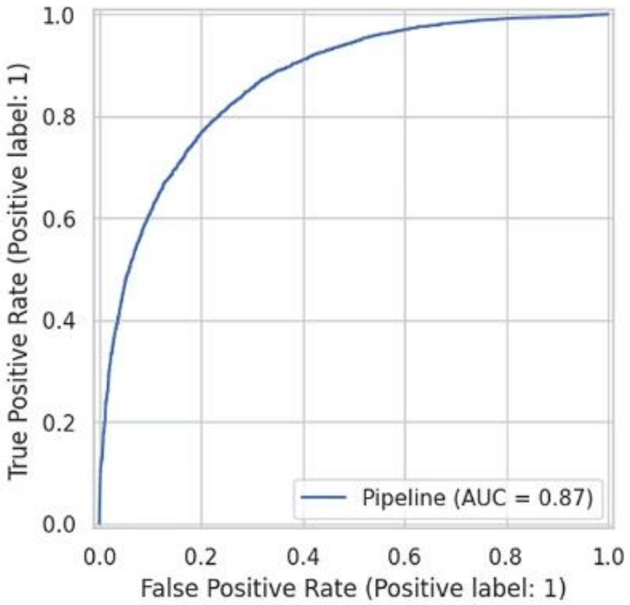
- Identified proxy features such as education level and job title that inadvertently contributed to gender bias in income predictions.
- Implemented bias mitigation strategies that improved model fairness metrics, increasing the demographic parity score by 15% and equalized odds by 12%.
- Enhanced model transparency through SHAP value analysis, enabling stakeholders to understand key drivers of pay disparities.

Insights:

- Effective ethical AI requires ongoing fairness auditing integrated throughout the model lifecycle, not just at deployment.
- SHAP and LIME provide crucial explainability that helps detect and mitigate hidden biases, making them essential tools for responsible machine learning.
- Socioeconomic factors such as education, occupation, and geographic location play significant roles and must be carefully considered to avoid reinforcing existing inequalities.







Strengths

Technical Strengths

- Expertise in fine-tuning transformer models like GPT-2 for specific NLP tasks
- Developing multimodal AI applications integrating text, audio, and image processing
- Conducting thorough model audits and fairness evaluations using SHAP and LIME explainability frameworks
- Advanced skills in exploratory data analysis and feature engineering for socioeconomic datasets

Professional Strengths

- Creating clear and comprehensive technical documentation paired with insightful visualizations
- Effectively communicating complex data science concepts to stakeholders and team members
- Strong project management skills, including coordinating cross-functional teams and meeting deadlines
- Applying ethical considerations and fairness principles in AI modeling and decision-making processes

Challenges

Technical Challenges

Overcoming API rate limits during peak user interaction in the Travel Companion Chatbot project, which caused intermittent service delays.

Addressing sudden drops in model accuracy after integration of new data sources, requiring rapid debugging and re-training.

Communication Challenges

Translating complex machine learning metrics like SHAP value interpretations and demographic parity scores into accessible insights for stakeholders without technical backgrounds.

Solutions Implemented

Implemented caching strategies and request batching to efficiently manage API call limits and reduce latency.

Adopted an iterative development and testing workflow to quickly identify and resolve model performance issues.

Enhanced presentation skills by practicing storytelling approaches that simplify technical findings for diverse audiences.

Growth & Continued Learning

Technical Growth

Evolved from theoretical knowledge to practical AI implementation by fine-tuning transformer models such as GPT-2 for natural language tasks.

Developed expertise in applying fairness and explainability frameworks like SHAP and LIME to ensure ethical AI deployment.

Built multimodal AI applications integrating text, audio, and image processing for complex real-world scenarios.

Professional Growth

Adopted user-centric design principles to enhance chatbot usability and accessibility, improving end-user satisfaction.

Implemented agile and iterative development methodologies to manage project timelines and quickly respond to emerging challenges.

Strengthened communication skills by translating complex model metrics into clear, actionable insights for non-technical stakeholders.

Continued Learning Plans

Pursue advanced study of Retrieval-Augmented Generation (RAG) techniques to enhance knowledge retrieval capabilities in AI systems.

Plan to achieve AWS Certified Machine Learning Specialty certification to validate cloud-based machine learning expertise.

Explore AI applications focused on improving accessibility and education, aiming to develop more inclusive technologies.

Career Goals

Immediate Career Goals

Pursue roles as a Machine Learning Engineer or Data Scientist specializing in natural language processing and ethical AI.

Target industries include Education Technology (EdTech) for adaptive learning platforms, and Healthcare focusing on AI-driven diagnostics and patient care.

Work Style Aspirations

Develop scalable, human-centered AI systems that prioritize user privacy and fairness across diverse populations.

Design transparent, explainable ML models using state-of-the-art interpretability frameworks to foster trust with stakeholders.

Skills to Leverage

Leverage deep expertise in natural language processing, fairness auditing techniques, and multimodal AI development integrating text, audio, and image data.

Utilize experience with transformer fine-tuning and explainability tools like SHAP and LIME to build ethical and performant models.

Long-Term Vision

Vision for the Future

Democratize access to AI in education and healthcare

Lead projects that integrate cultural preservation and AI

Build tools that make AI inclusive, ethical, and globally beneficial

Personal Mission

Become a bridge between technical excellence and societal good

Thank You

I deeply appreciate the opportunity to share my work and reflections.

Portfolio Link