Sandeep Bandi

Hyderabad, Telangana | +91 94944 20777 | bandisandeep1423@gmail.com | LinkedIn | GitHub | Website

Summary

Results-driven AI/ML Engineer with expertise in delivering end-to-end Machine Learning and Generative AI solutions across social media, healthcare, and insurance domains. Skilled in building scalable data pipelines and production-ready models using **Python**, **TensorFlow**, **FastAPI**, PyTorch, **LangChain**, and **Hugging Face**. Experienced in **MLOps** practices including **CI/CD**, model tracking, and cloud deployment **(GCP, AWS)**. Passionate about solving real-world problems through human-centric and performance-optimized AI systems.

Technical Skills

Programming Languages : Python, R, SQL, Flask
Data Science & Analytics: Data Preprocessing, Feature Engineering, Statistical Modeling, Probability Theory
Machine Learning & Deep Learning: scikit-learn, TensorFlow, Keras, PyTorch, ANN, CNN, RNN
Natural Language Processing (NLP): Transformers, Hugging Face, BERT, GPT
MLOps & CI/CD: MLflow, DVC, GitHub Actions, Jenkins, Apache Airflow
Cloud & Computing: AWS (EC2, S3, Lambda), GCP (BigQuery), Distributed Systems Concepts
Databases & Big Data: MySQL, PostgreSQL, SQL Server, MongoDB, AWS Glue
Data Visualization & BI Tools: Power BI, Matplotlib, Seaborn, Excel
Development Tools: Git, GitHub, Docker, Kubernetes, Jupyter Notebook, VS Code
LLM Tools & Frameworks: LangChain, OpenAI APIs, Hugging Face Models
Methodologies & Concepts: Agile, Data Warehousing, Business Analytics

Experience

AI/ML Engineer June 2024 – September 2024
BeWorld Web Services *Hyderabad, India*

- Designed and implemented an end-to-end personalized feed ranking algorithm to enhance user engagement and content relevance on a social media platform.
- Extracted and cleaned structured data from **MongoDB**, including user interactions, post metadata, and engagement patterns, to generate actionable features such as engagement scores, post behaviors, and content metadata.
- Developed and trained supervised **ML models** (Logistic Regression, XGBoost) for predicting user interest scores, improving feed content relevance.
- Integrated hybrid recommendation techniques combining **collaborative filtering** (user behavior) and **content-based filtering** (post metadata) to provide personalized recommendations.
- Built a feed ranking system that prioritized followed users' posts first, followed by interest-based recommendations, driving increased user retention and engagement.
- Delivered Post IDs and relevance scores via API integration to the backend, enabling real-time ranking of posts in the platform.
- Deployed the solution as a **FastAPI** microservice on **AWS EC2**, ensuring low-latency predictions and scalability.
- Implemented **CI/CD pipelines** using Docker and Jenkins for automated model training, testing, and deployment, ensuring smooth iteration and reproducibility.
- Collaborated closely with cross-functional teams (product, backend, and frontend) to seamlessly deploy and integrate the solution into the platform.

ML Engineer January 2022 – February 2024
Tata Consultancy Services *Hyderabad, India*

- Collaborated with M&G Prudential client to deliver **AI/ML-driven** solutions across insurance and investment platforms, focusing on churn prediction, customer segmentation, and policy risk assessment.
- Analyzed and modeled structured datasets, including customer profiles, policy histories, behavioral indicators, and fraud risk flags sourced from enterprise SQL databases.
- Designed and implemented **customer segmentation** models using KMeans clustering, enabling targeted retention strategies and improving campaign performance.

- Built and optimized predictive **models** using XGBoost, Logistic Regression, and Random Forest, achieving an F1 score of 0.84 for churn prediction and reducing customer loss by ∼15%.
- Developed a real-time fraud risk scoring model with an AUC of 0.88, streamlining risk assessments and cutting down manual review efforts by 25%.
- Translated complex ML outputs into actionable business insights using SHAP for model explainability and Matplotlib for stakeholder-friendly visualizations.
- Automated ML pipelines for data processing, model training, and evaluation using **MLflow**, **Docker**, **Airflow**, and Jenkins, ensuring seamless **CI/CD workflows**.
- Deployed models via FastAPI microservices, enabling real-time predictions and easy integration into existing customer management systems.

Project Intern November 2021 – February 2022

iNeuron.ai *Remote*

- Developed a machine learning system for income price prediction through an end-to-end ML pipeline.
- Executed core stages including **data ingestion**, **EDA**, **transformation**, **model training**, **prediction**, and **deployment**.
- Built and containerized the solution using **Python** and **Docker**; deployed it on AWS for real-time inference.
- Gained hands-on experience with VS Code, GitHub version control, and cloud-based model delivery.

Education

Lovely Professional University Punjab, India

Bachelors of Technology in Computer Science Engineering *2016 - 2020*

Projects

Hotel Reservation Cancellation Prediction | *Python, MLflow, Jenkins, Docker, Flask* GitHub | Live App

- Developed an end-to-end machine learning solution to predict hotel reservation cancellations, enabling proactive customer engagement and revenue optimization.
- Designed and implemented the entire ML pipeline, from data ingestion (GCP) and preprocessing to model training and deployment.
- Conducted feature engineering, data versioning (DVC), and model versioning for reproducible experiments.
- Implemented experiment tracking (MLflow, CometML) and automated model deployment using CI/CD (Jenkins, Docker, GCP Run).
- Built a Flask-based web application with an interactive UI (HTML, CSS) for real-time predictions.

Hybrid Anime Recommendation System | *Python, Scikit-learn, Pandas, Flask, Streamlit* GitHub | Live App

- Developed a hybrid recommendation system combining collaborative filtering and content-based filtering techniques to provide personalized anime suggestions.
- Utilized user-item interaction matrix and anime metadata to generate tailored recommendations.
- Designed a modular and scalable ML pipeline ensuring clean code practices and easy experimentation.
- Implemented a Flask-based backend and integrated with a Streamlit front-end for deployment-ready application.

Adult Census Income | *Python, Flask, scikit-learn, Pandas, MongoDB, AWS, S3, SQL* GitHub | Live App

- Designed and implemented a binary classification system to predict whether individuals earn more than $50K annually, leveraging Logistic Regression for the best performance with an accuracy of 85%. Developed a modular workflow comprising Data Ingestion, Data Transformation, and Model Training pipelines.
- Built a user-friendly web application for real-time predictions, incorporating a preprocessing pipeline for handling numeric and categorical variables and integrating Logistic Regression models for seamless deployment.

Gemstone Price Prediction | *Python, LightGBM, XGBoost, scikit-learn* GitHub | Live App

- Built and deployed a predictive model leveraging supervised learning techniques to estimate gemstone prices, optimizing performance using cross-validation and advanced regression algorithms.
- Analyzed and visualized key factors influencing gemstone prices using tools like Python, Pandas, and Matplotlib, enabling clear interpretation of market trends and model outputs.

Certifications

Full Stack Data Science Certificate | iNeuron Certificate Link

Complete ML, DL, NLP Bootcamp Certification | Udemy Certificate Link

Python | Kaggle Certificate Link