VIVEK GIRI

Machine Learning & Data Science

Phone: 9064447153 | Email: vivekgiri556@gmail.com | GitHub: github.com/vivek7557 | Kaggle: kaggle.com/vivek7557 | Location: Bangalore

---

Professional Summary

KYC Analyst with strong real-world experience in compliance and fraud detection, now transitioning into Machine Learning and Generative AI. While working full-time in the fintech domain, I've actively built endtoend ML solutions across NLP, recommender systems, and deep learning. My hands-on projects involve LLMs, LangChain, and Hugging Face, highlighting my commitment to continuous learning and practical application of AI technologies. Skilled in blending business insight with data science to solve real problems and create future-ready AI tools.

Core Skills

- Programming & Data: Python, Pandas, NumPy, MySQL, PostgreSQL
- Machine Learning: Linear/Logistic Regression, Decision Trees, Random Forests, SVM, Recommender Systems, NLTK
- Deep Learning: Neural Networks, Keras, TensorFlow
- NLP: Text Classification, TF-IDF, CountVectorizer, Pipelines
- Visualization: Matplotlib, Seaborn
- Tools: Jupyter Notebook, Anaconda, PyCharm, VS Code
- Generative AI Exposure: Hugging Face, LangChain, Prompt Engineering, RAG (RetrievalAugmented Generation)
- LLMs: GPT-4, GPT-3.5, BERT, LLaMA, Claude, Gemini, DeepSeek

---

Professional Experience

KYC Analyst – Khatabook | April 2024 – Present

- Verified customer identities using internal tools, enhancing compliance accuracy
- Analyzed case trends and optimized workflows using SQL and Excel
- Maintained high data integrity across verification pipelines
- Automated data quality checks, reducing manual verification time by 30%
- Developed internal dashboards for real-time case status using Google Sheets & Excel
- Collaborated with engineering teams to improve ID verification APIs
- Flagged fraudulent applications through pattern recognition and data analysis

CRM Executive – Home Credit | July 2023 – April 2024

- Communicated new service offerings to customers and resolved inquiries effectively

- Adapted to regular updates in products and services
- Influenced customers to purchase services using consultative selling
- Consistently exceeded 100% of sales goals while handling high call volumes
- Maintained accurate transaction logs and customer databases

Data Entry QA Operator | April 2023 – July 2023

- Verified and edited data to ensure compliance with internal standards
- Routed and tracked clinical and administrative documentation
- Facilitated document flow and supported compliance reviews
- Ensured confidentiality of records per policy and regulation- Maintained logs and tracking systems for audit readiness

Projects

OpenAI-Powered Data Explorer with Gradio & SQLAlchemy

- Built an AI-powered **Natural Language Data Agent using Gradio + OpenAI**, enabling users to convert plain English queries into SQL and Pandas code with **95% accuracy.**
- Integrated multi-database support (SQLite, MySQL, PostgreSQL) with automated schema extraction, reducing manual query writing time by **80%.**
- Implemented CSV upload and dynamic dataset exploration, allowing non-technical users to interact with their own data, boosting usability by **70%** in testing.
- Designed an AI-driven query engine capable of handling JOINs, aggregations, and Pandas method chaining, achieving consistent execution success across **200+ test queries.**
- Deployed the tool with a shareable Gradio interface, enabling real-time data analysis and making it accessible for both local and external users

End-to-End Fraud Detection System with 99.99% ROC-AUC

- Engineered a high-performance fraud detection system on a 6M+ transaction dataset, leveraging LogisticRegression, Random Forest, and XGBoost to maximize predictive accuracy (**ROC-AUC: 0.9999, F1-score: 0.8924**)
- Built end-to-end ML pipelines using ColumnTransformer & Pipeline for seamless preprocessing,transformation, and modeling
- Tackled severe class imbalance with cost-sensitive learning and SMOTE, boosting rare fraud detection rates- Delivered data-driven insights through rigorous evaluation: ROC-AUC, PR-AUC, F1, precision, recall, and confusion matrix
- Optimized decision-making via threshold tuning and best-model saving for real-world deploymentreadiness

Credit Risk Scoring Model – Advanced ML

- Model Development & Deployment → Built an advanced credit risk scoring model

leveraging Logistic Regression, Random Forest, XGBoost, and LightGBM, achieving ROC-AUC 0.91 and 87% accuracy on test

- data.Exploratory Data Analysis (EDA) → Analyzed 10,000+ customer financial & behavioral records, uncovering patterns in credit utilization, transaction activity, and churn behavior.
- Feature Engineering → Created 20+ derived features (transaction ratios, utilization trends, spending patterns) to strengthen model predictive power and reduce noise.
- Imbalanced Data Handling → Applied SMOTE & cost-sensitive learning techniques to address class imbalance, improving recall of high-risk customers by 22%.
- Model Optimization → Conducted cross-validation & hyperparameter tuning, enhancing precision and minimizing false positives in customer risk classification.
- Business Insights & Recommendations → Identified high-risk customer segments and proposed data-driven strategies for churn reduction & targeted retention campaigns.

Uber Ride Booking Prediction | Machine Learning Project

- Performed **data preprocessing & feature engineering** (encoding, scaling, imputation, and handling missing values) to prepare large-scale ride data for ML.
- Conducted **advanced EDA** (heatmaps, boxplots, violin plots) uncovering insights on **ride distance, payment method, turnaround time, and ratings** affecting booking status.
- Built and compared **7 ML models (Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, SVM, KNN, XGBoost)** achieving **~100% model accuracy** with Logistic Regression & XGBoost.
- Developed a **robust ML pipeline** ensuring **scalability, reproducibility, and production readiness** for financial/transportation data analysis.

AI Brochure Generator

- Designed AI assistant to generate company brochures using GPT and structured prompts
- Accepted company inputs via a form and output formatted PDFs
- Integrated LangChain for dynamic prompt chaining and content generation
- Output generation time: 1.2s with 95% relevance rate
- Validated brochure formats with manual and automated testing

---

Certifications

- LLM Engineering: Master AI, Large Language Models & Agents (Udemy)
- Applied Data Science with Python – University of Michigan (Coursera)
- The Complete SQL Bootcamp – Jose Portilla (Udemy)
- Python for Data Science and ML – Jose Portilla (Udemy)
- The Ultimate MySQL Bootcamp – Colt Steele (Udemy)

Education

Bachelor of Business Administration (BBA) – BB College, West Bengal