

RAVINDRA KUMAR YADAV

Mob: 9466301856 | **Email:** yadavravi.it@gmail.com | **LinkedIn:** [linkedin.com/in/ravindra-kumar-yadav](https://www.linkedin.com/in/ravindra-kumar-yadav)

Portfolio: ravindrph.blogspot.com | **GitHub:** github.com/ravindra-yadav

Data Scientist | Machine Learning Engineer | Credit Risk Modelling | Generative AI Engineer

Professional Summary

Data Scientist with 5+ years of experience in Machine Learning, Credit Risk Modelling, Predictive Analytics, SQL-based ETL and Business Intelligence across FinTech (NBFC), EdTech and Digital Platforms. Currently building and deploying production-grade credit risk models at an NBFC, owning the full lifecycle from data extraction and feature engineering to REST API serving with FastAPI and cloud deployment on Google Cloud Platform (GCP), backed by PostgreSQL and MongoDB. Experienced in end-to-end ML pipelines, recommendation systems, fraud detection and KPI reporting using Python, SQL, Scikit-Learn, XGBoost, FastAPI and Streamlit. Upskilled in Generative AI, Deep Learning, NLP, LLMs, Retrieval-Augmented Generation (RAG), Agentic AI, LangChain and Vector Databases. Strong foundation in statistics, model explainability and data storytelling, with a focus on solving business problems through data-driven decision making.

Core Technical Skills

Programming: Python, SQL, PySpark, Pandas, NumPy, Matplotlib, Seaborn, PyTest, Streamlit

Machine Learning & Libraries: Regression, Classification, Credit Risk / PD Modelling, Time Series Forecasting, Feature Engineering, Scikit-learn, XGBoost, Optuna, Imbalanced-learn, Statsmodels

Credit Risk Analytics: Probability of Default (PD) Modelling, Weight of Evidence (WoE), Information Value (IV), Gini / KS Statistic, AUC-ROC, Class Imbalance Handling (SMOTE, Undersampling), Model Monitoring

Deep Learning and NLP: Text Classification, Text Preprocessing, Tokenization, Named Entity Recognition (NER), Semantic Similarity, Information Retrieval, TF-IDF, Cosine Similarity, NLTK, SpaCy, PyTorch

Generative AI & LLMs: Large Language Models (LLMs), LangChain, RAG (Retrieval-Augmented Generation), Agentic AI, Prompt Engineering, Vector Database (ChromaDB)

Backend, MLOps & Deployment: FastAPI, REST APIs, Pydantic, Docker, MLflow, CI/CD, Google Cloud Platform (GCP), AWS SageMaker, AWS Cloud

Databases: PostgreSQL, MongoDB, ChromaDB (Vector DB)

Data & BI: Power BI, ETL Pipelines, Data Visualization, KPI Reporting

Statistics: Hypothesis Testing, A/B Testing, Probability, Regression Analysis, VIF Analysis, Multicollinearity Resolution

Professional Experience

Data Scientist | (FinTech NBFC)

[Noida, UP]

End-to-End Credit Risk Scoring System (Production)

- Built an end-to-end credit risk (Probability of Default) model to predict loan defaults and support credit underwriting decisions on 91.6% + loan applications, achieving an AUC-ROC of 89% / Gini of 94%.
- Engineered features from applicant demographics, bureau data, income and repayment history; applied WoE/IV and VIF analysis for feature selection and multicollinearity resolution, and handled class imbalance to prioritise recall on defaulters.
- Developed high-performance REST APIs using **FastAPI** with Pydantic request validation to serve real-time credit scores to the loan origination system, with response latency under 0.39 ms.
- Containerised the model service with Docker and deployed it on **Google Cloud Platform (GCP) Vertex AI** enabling scalable, low-latency inference in production.
- Designed the data layer using **PostgreSQL** for structured loan, customer and repayment data (complex SQL, CTEs, window functions) and **MongoDB** for semi-structured application data, prediction logs and model audit trails.

Tech Stack: Python, SQL, XGBoost, Scikit-learn, FastAPI, Docker, GCP, PostgreSQL, MongoDB

SME | EdTech & Digital Platforms, Consulting & Freelance

NLP-Based Personalized Content Recommendation System & AI Personalization Engine

- Developed a content recommendation engine using NLP techniques (text preprocessing, TF-IDF vectorization, cosine similarity) for personalized content suggestions.
- Improved user engagement and click-through rate (CTR) by 17% and reduced platform complaint tickets by 15% through real-time inference and model monitoring workflows.

- Executed the complete ML lifecycle including data preprocessing, feature engineering, model training, evaluation and deployment.

Enterprise ETL & Analytics Platform

- Built, optimized and maintained large-scale ETL pipelines using Python and complex SQL (advanced CTEs) to process high-volume datasets for ML applications.
- Reduced dashboard data load times by 40% and cross-functional data discovery time by 25%, improving real-time decision-making for technical and non-technical stakeholders.
- Automated data ingestion, transformation, validation and reporting workflows for business analytics.

Fraud Detection & User Behavior Analytics

- Analyzed user interaction patterns to detect fake profiles, inappropriate content uploads and policy violations using SQL, Python and Power BI.
- Identified anomalous activities and operational risks through exploratory data analysis and pattern recognition, improving platform trust and operational efficiency.
- Conducted deep-dive fraud analysis, academic integrity auditing and interaction behavior tracking.

Student Performance Analytics

- Analysed learner engagement and performance metrics to identify key success factors.
- Developed analytical frameworks and automated recommendation mechanisms that improved student performance scores by 15% and overall user acceptance by 10%.

Consulting & Freelance Analytics Projects

- Delivered 100+ end-to-end analytics and machine learning projects across multiple business domains, maintaining an average client satisfaction rate of 95%.
- Built scalable, production-ready ML systems for real-time data processing and low-latency inference, reducing data processing latency by 20%.
- Managed the complete project lifecycle from requirement gathering to deployment and documentation.

Machine Learning & AI Engineering Projects

Predictive Health Insurance Premium Model | Regression | Python, XGBoost, Scikit-learn, MLflow, Streamlit

- Built a premium prediction model on 50,000 records achieving R^2 of 98.19%, exceeding the client target of 97%; EDA identified top premium drivers: insurance plan (0.83), age (0.77) and normalized risk score (0.52).
- Resolved multicollinearity between income-related features ($r=0.91$); tuned hyperparameters with GridSearchCV and RandomizedSearchCV (3-fold CV), reducing RMSE from ₹2,247 to ₹1,130.
- Implemented experiment tracking and model monitoring using MLflow for performance optimization and reproducibility.

End-to-End Credit Risk Classification | Classification | XGBoost, SQL, Optuna, Imblearn, Streamlit

- Developed and deployed a credit risk model on 50,000+ records to predict loan defaults, achieving AUC-ROC of 98.4% and minority-class Recall of 97%.
- Used VIF analysis, Weight of Evidence (WoE) and Information Value (IV) for feature selection, identifying Credit Utilization as the primary risk driver (IV: 2.35).
- Addressed heavy class imbalance (91/9) using RandomUnderSampler and SMOTETomek, prioritizing recall to minimize lending risk.
- Architected a modular Streamlit application with a decoupled backend for real-time inference, including automated data cleaning and MinMaxScaler transformations.

Education & Certifications

- **Professional Certification:** Data Science & Generative AI Advanced Program | Codebasics.com — Math and Statistics for AI, SQL for Data Science, Advanced Python, Machine Learning, Deep Learning, NLP, Generative AI and Agentic AI applications.
- **B.Tech in Information Technology** | Graduation Year: 2015 (Score: 69%)