Om Hitesh Vyas

Data Scientist, New York, NY | omvyas288@gmail.com | linkedin.com/in/om-hitesh-vyas-97340a204 | github.com/omvyas77

Experience

PNC, New York, NY | **Data Scientist** 09/2026 to present

- Ran exploratory analysis in Python and pandas on borrower demographics, credit history, repayment behavior, income and loan attributes, identifying 20+ predictive risk patterns for default modeling.
- Built SQL feature pipelines that turn repayment history, credit utilization, delinquency frequency, debt-to-income ratios and loan characteristics into 30+ model-ready predictors for borrower-risk segmentation.
- Used SHAP to quantify global and per-prediction feature contributions, surfacing the main drivers of default risk.

Synchrony Financial, Illinois | **Data Science Intern** 05/2025 to 05/2026

- Built and validated consumer credit-risk and loss-forecasting models (PD, EAD, expected loss) for large lending portfolios, benchmarking logistic regression, XGBoost and LightGBM with out-of-time backtesting and ROC-AUC and KS checks.
- Engineered 100+ macroeconomic, credit bureau, behavioral and account-level features from 18+ years of data, stabilizing forecasts.
- Designed a data-quality framework using Isolation Forest and statistical outlier detection (z-score, IQR) that monitors 9 critical KPIs across 2B+ records; adopted as the team's standard pre-deployment check.
- Built SQL, Hive and PySpark pipelines on AWS EMR and S3 (10 to 20 nodes) that process 900M+ customer records per snapshot.
- Moved Excel risk workflows to PySpark (hours to minutes) and generated 460+ dashboards from one parameterized notebook.

Johnson Controls-Hitachi Air Conditioning, India | **Data Scientist** 01/2024 to 07/2024

- Rebuilt retrieval for a customer-facing HVAC support assistant with LangChain, FAISS and OpenAI, raising Precision@3 from 68% to 74%, confirmed with a paired per-query significance test ($p = 0.03$) before going live.
- Ran the confirmatory 50/50 A/B test (4.2K visitors per arm, generator and prompt held constant): self-service resolution rose from 40.0% to 42.4% ($p \approx 0.025$), with latency and groundedness guardrails set to roll back automatically.
- Fine-tuned Mistral-7B with LoRA (PyTorch, Hugging Face) for domain-specific generation, gated by an offline evaluation harness.

Accenture, India | **Data Analyst** 08/2022 to 12/2023

- Analyzed customer transaction, basket, spending and channel data in Python and pandas, identifying 20+ behavioral patterns.
- Engineered 15+ RFM-based predictors in SQL and found 6 customer segments with K-Means, StandardScaler and PCA.
- Validated the segments with SciPy hypothesis tests and built logistic regression propensity models to prioritize retention.

Projects

FinHelm: question answering over bank SEC filings | Python, FastAPI, FAISS, MLflow, Docker, OpenTelemetry GitHub

- Built a QA system over ~43K chunks of SEC filings, Fed statements and CFPB complaints with hybrid BM25 and dense retrieval, cross-encoder reranking and a gated query planner (+0.19 recall on multi-part questions); 74% evidence recall at $0.03 a question.
- Made refusal a primary metric on a 202-question test set (21 hand-written unanswerable; audited and fixed 2 bad labels): 95% correct refusal at a 12% false-refusal rate. Error analysis found a wrong-filing-year failure that citation checks miss.
- Shipped a CI-gated evaluation harness that blocks metric regressions and checks reported figures against a frozen run.

Gies Program Health: enrollment and grading analytics | Python, dbt, DuckDB, Tableau, Streamlit, GitHub Actions GitHub

- Built a tested pipeline over 7 public sources (IPEDS, enrollment reports, course grades) with dbt on DuckDB, enforced column contracts and scheduled refreshes that publish only when every test passes; served in Tableau and Streamlit.
- Split a 21.7-point rise in A-range grades (2015 to 2025) into within-course and course-mix effects (17.5 points within courses), and backtested headcount forecasts against naive and drift baselines.

Credit-risk early signals in consumer complaints | Python, NLP, LightGBM, SHAP GitHub

- Built a rare-event NLP classifier for credit-risk signals in CFPB complaint narratives, using TF-IDF, metadata and domain-lexicon features and a time-based split to prevent leakage.
- Benchmarked logistic regression against LightGBM, MiniLM embeddings and a fine-tuned transformer on PR-AUC, with cost-sensitive thresholds, calibration, SHAP and domain-shift tests; chose the interpretable logistic model.

Education

University of Illinois Urbana-Champaign | M.S., Information Management (Data Science Pathway) | GPA 3.85/4.0 May 2026
Nirma University, Ahmedabad, India | B.Tech., Computer Science 2024
Tableau Desktop Certification, Tableau

Skills

Data and modeling	Python, SQL, pandas, PySpark, Kafka, dbt, DuckDB, AWS EMR, scikit-learn, XGBoost, LightGBM, PyTorch, SHAP
Retrieval and MLOps	RAG, LangChain, FAISS, BM25, reranking, LoRA, Hugging Face, MLflow, Docker, FastAPI, GitHub Actions

Publications

F-Transformer: A Federated Transformer for Efficient and Privacy-Preserving Sequence Generation, *Scientific Reports*, 2026
BookNest: An Ensemble-Based Book Recommendation System, IC4S 2024 proceedings (Lecture Notes in Networks and Systems)