

Harshit Kashyap

Data Scientist

kshp95@gmail.com | [Portfolio](#) | [LinkedIn](#)

PROFESSIONAL SUMMARY

- Data Scientist with 6+ years across healthcare, financial services, and analytics, specializing in machine learning, statistical modeling, and production ML systems.
- Own the end-to-end ML lifecycle — feature engineering, model validation, MLOps, CI/CD, monitoring — on cloud-based data workflows processing 2M+ records per month.
- Proficient in Python, SQL, R, and Snowflake, with hands-on production exposure to Docker, Git, GCP, and deployment orchestration.
- Build applied AI projects spanning LLM-powered RAG systems, agentic workflows, and MCP tool integration alongside core statistical/ML modeling (XGBoost, Random Forest, K-Means, deep learning, NLP).

PROFESSIONAL EXPERIENCE

Claritev | New York, NY

08/2023 – 01/2026

Data Scientist

- Streamlined large-scale claims analytics for ProPricer, refining cost-management and repricing models across 2M+ monthly claims.
- Engineered an LLM-driven text-to-SQL interface over healthcare claims data — designing prompt logic, schema alignment, and query guardrails for plain-English analyst access — cutting turnaround from ~3 hours to under 10 minutes.
- Built cumulative appeal-rate-vs-savings analyses across pricing thresholds, quantifying the tradeoff between repricing aggressiveness, savings, and provider appeal risk.
- Used Bayesian posterior updates on weekly claim outcomes to quantify uncertainty in savings and provider appeal rates, then recalibrated claim-risk thresholds as evidence accumulated, increasing identified savings by approximately 12%.
- Developed risk-scoring and disease-rate prediction models using contextual features, producing calibrated probability outputs across a population of 1M+ members to support care-management prioritization and population-health planning.
- Executed production ML pipelines for new client populations, validating feature calculations and predictions, supporting up to \$22M in potential annual savings per client.
- Monitored and benchmarked model performance across client populations, driving targeted improvements and a ~10% lift in overall savings performance.
- Automated Python analytics workflows (~30% faster processing) and optimized SQL pipelines (~50% faster feature retrieval).
- Applied MLOps practices across production workflows — Git version control, CI/CD, model monitoring, Docker-based deployment, scheduled jobs, API testing, and production handover.
- Operated CERES, an XGBoost-based benefit-utilization platform, on its Snowflake-based production environment, supporting policy/benefit-metrics pipelines and QA.
- Partnered with clinical and engineering stakeholders to translate model outputs into actionable population segments; documented model assumptions and validation methodology, reducing QA review cycles 40%.

Connecticut Information Technology Institute (UConn) | Stamford, CT

08/2021 – 12/2022

Student Administrative Specialist III (Data Science)

- Engineered a Selenium-based web-scraping pipeline to extract and structure client data for 2,000+ accounts.
- Designed predictive time-series models to forecast client-interaction and quiz-completion trends, and built automated client-response workflows.
- Built and deployed a Random Forest binary classifier on Azure ML.
- Ran A/B tests across landing-page designs and quiz-question variants to optimize completion rates and engagement.
- Delivered Power BI and Streamlit dashboards alongside SSIS/SSRS-based ETL reporting, partnering cross-functionally with the web team to translate analytics findings into design decisions.

American Century Investments | Kansas City, MO

05/2022 – 07/2022

IT Intern (Data Science)

- Built a BERT-based NLP sentiment-scoring pipeline over GDELT news-event data, powering a Tableau dashboard that tracked controversy sentiment and article-volume intensity across major portfolio holdings.
- Built an S3-to-PostgreSQL ETL pipeline for MSCI ESG controversy assessments, flags, and source articles used in Tableau reporting.
- Partnered cross-functionally on KPI analysis via REST APIs; ran ETL across SSIS/SSRS and MongoDB (via the Hadoop connector) plus ANOVA testing in Python.
- Benchmarked classical ML baselines — decision tree, Naïve Bayes, logistic/linear regression — against a BERT-based model for ESG sentiment classification, selecting BERT for higher accuracy.

Senior Data Analyst

- Extracted Azure data via SQL for EDA and PCA-based customer segmentation (K-Means/DBSCAN), cutting email campaign costs 20%.
- Constructed a Python/ETL pipeline cutting data-processing time 50%, enabling real-time metric analysis.
- Lifted landing-page conversion 45% through structured A/B testing.
- Designed Python/R/SAS predictive models on client behavior, improving Q1 retention 30%.
- Created Power BI dashboards across 15+ KPIs, improving executive visibility into operations 45%.
- Spearheaded an automated Excel-macro reporting system, saving \$78K per quarter in labor costs and cutting report time 29%.
- Engineered feature/imputation pipelines (scikit-learn, NumPy, pandas) and visualized customer behavior in R (ggplot2), pinpointing 3 key behavioral drivers behind operational adjustments.
- Applied XGBoost in R to model predicted probabilities and delivered results to the operations team for decision-making.

PROJECTS

CanadaRx: Canadian Drug Information RAG System | Qdrant, OpenAI, Streamlit — [Live Demo](#) | 02/2026 – 04/2026

- Built a citation-grounded RAG system over official Health Canada and Ontario drug sources, distilling 866 pages of PDFs and safety alerts into a 15.5K-record DIN catalogue and 1,327 traceable chunks, with hybrid retrieval (dense + BM25 + metadata filtering) on Qdrant.
- On a sealed 50-question holdout, achieved 87.5% Recall@3 on answerable queries, 100% citation validity, and 95% correct answer/refusal behavior; deployed on Streamlit and Qdrant Cloud with Docker packaging and automated tests.

Trading Strategy Text-to-SQL Agent (MCP) | OpenAI Agents SDK, DuckDB/MotherDuck — [Live Demo](#) | 04/2026 – 06/2026

- Engineered a DuckDB/MotherDuck warehouse consolidating 4.1M futures-market records with validated schemas and chronological train/val/test partitions.
- Built a 21-tool read-only MCP server and an OpenAI Agents SDK research assistant with guarded SQL, scope-refusal guardrails, and Streamlit/Altair charts; on a 50-case evaluation suite, 23/50 tasks passed end-to-end, with published failure analysis and zero write queries executed by design.
- Commodities Analysis (Python | Power BI) — [Link](#)
- Analytics Job Market Analysis (Python | Tableau) — [Link](#)
- Stroke Prediction (Machine Learning | Python) — [Link](#)

TECHNICAL SKILLS

Languages & Querying: Python, SQL, R

Machine Learning & Statistics: XGBoost, Random Forest, Logistic/Linear Regression, KNN, SVM, Naïve Bayes, K-Means Clustering, Neural Networks, Deep Learning, NLP, Decision Trees, A/B Testing, Statistical Analysis

AI/LLM Tooling: LLM Application Development (OpenAI API, Agents SDK), RAG, Qdrant, MCP (Model Context Protocol), Prompt Engineering, DuckDB/MotherDuck

Data, Cloud & Production: Snowflake, GCP, Docker, Git, Cron, REST APIs, MongoDB, Hadoop/Hive, Spark/PySpark, CI/CD & Model Monitoring

Visualization & BI: Tableau, Power BI, R-Shiny, Streamlit, MS SSIS/SSAS/SSRS, ETL, MATLAB, SPSS, SAS

EDUCATION

Master of Science, Business Analytics & Project Management (Data Science) — 12/2022

University of Connecticut, Stamford, CT

Bachelor of Technology, Mechanical & Automation Engineering — 06/2017

Guru Gobind Singh Indraprastha University, India

Publication: [Robotic Arm Aimed to Replace Cutting Processes](#)

CERTIFICATIONS

- [DeepLearning.AI — Generative AI with Large Language Models](#)
- [WorldQuant University — Applied Data Science II \(Machine Learning & Statistical Analysis\)](#)
- [WorldQuant University — Applied Data Science I \(Scientific Computing & Python\)](#)