

# VINAYAK SHRISHAIL ANADINNI

+1 (312) 610-2915 | [vinayak14anadinni@gmail.com](mailto:vinayak14anadinni@gmail.com) | [www.linkedin.com/in/vinayaksa99](http://www.linkedin.com/in/vinayaksa99) | [github.com/vinayak99](https://github.com/vinayak99) | [Portfolio](#) | U.S.A (Authorized to work)

## PROFESSIONAL SUMMARY

---

Data Science professional with 3+ years automating analytics workflows across financial services and CRM operations, currently completing an MS in Data Science at Illinois Institute of Technology. Specializes in eliminating manual bottlenecks through SQL-driven ETL pipelines, data quality frameworks, and BI dashboard automation reducing reporting cycles and scaling decision-making for cross-functional stakeholders. Applies MLOps and GenAI tooling as production-grade accelerators to operationalize experimentation, KPI design, and end-to-end data delivery at scale.

## TECHNICAL SKILLS

---

- **Programming Languages:** Python (Pandas, NumPy, scikit-learn, PySpark), SQL (CTEs, window functions, stored procedures, views, triggers), R
- **Database Platforms:** MS SQL Server, PostgreSQL, BigQuery, Snowflake, SQLite, Databricks, schema design, query optimization, indexing strategies
- **Data Modeling & Transformation:** Star Schema, Snowflake Schema, Denormalization, Data Warehouse design, Normalized forms (1NF–3NF), ETL/ELT pipeline architecture, dbt (data build tool), data contracts
- **Data Quality & Governance:** Great Expectations, Evidently AI, Monte Carlo, metadata management, data lineage, business definitions, HIPAA/PII compliance, data access governance, anomaly detection, drift monitoring, Population Stability Index (PSI)
- **ETL & Orchestration:** Apache Airflow, Google Composer, SSIS, Apache Spark (PySpark), Apache Kafka, Prefect, FastAPI, Streamlit
- **BI & Reporting Tools:** Power BI, Tableau, Google Looker Studio, SSRS, Excel (pivot tables, Power Query, advanced analytics), Power Automate
- **Statistics & Experimentation:** Hypothesis Testing, A/B Testing, Experimental Design, Regression Analysis, Classification, Clustering, Feature Engineering, Error Analysis, Bias-Variance Tradeoffs, Agile project planning, cross-functional sprint delivery
- **Cloud & DevOps:** AWS (S3, EC2, SageMaker, Lambda, Redshift, CloudWatch), GCP (BigQuery, Vertex AI, Google Composer), Terraform, Docker, Kubernetes, CI/CD, GitHub Actions, MLflow, DVC, Git
- **NLP & GenAI:** Embeddings, Transformers, RAG, LangChain, LlamaIndex, Sentence Transformers, Vector Databases (Pinecone, ChromaDB, FAISS), Arize Phoenix, RAGAS, Prompt Engineering

## PROFESSIONAL EXPERIENCE

---

### *Data Scientist* Illinois Institute of Technology, Chicago, Illinois, USA

Jan 2026 – Present

- Spearheaded an end-to-end agentic AI market intelligence pipeline, achieving a 98% reduction in competitor research-and-draft cycle time (6+ hours to under 5 minutes per campaign), by architecting a BrightData SERP + Playwright + BeautifulSoup scraping system that autonomously extracts engaging keywords and content patterns from competitor blog posts, product pages, and landing campaigns to generate platform-native marketing emails.
- Surpassed manual content scouting benchmarks by 80%, processing 500+ competitor pages per query run, by engineering a multi-signal scoring system that combines keyword frequency, topic clustering, and formatting density to surface high-engagement themes and detect competitor content patterns at scale.
- Doubled the volume of actionable competitor signals per analysis run, as measured by engagement keyword recall and messaging shift detection rate, by implementing a semantic analysis engine using sentence-transformers (all-MiniLM-L6-v2) with cosine similarity thresholding to perform trend analysis and sentiment classification across competitor blog and campaign content.
- Achieved under 10-second end-to-end query response time across 1,536-dimensional embeddings spanning competitor datasets, by developing a weighted relevance scoring algorithm persisted in a WAL-mode SQLite store with upsert deduplication for zero-recompute retrieval, delivered via a Streamlit dashboard and FastAPI backend serving keyword recommendations and AI-generated content strategy reports.

**Skills:** Python, BrightData SERP, Playwright, BeautifulSoup, Sentence-Transformers (all-MiniLM-L6-v2), Cosine Similarity, Vector Embeddings, SQLite, RAG, FastAPI, Streamlit, LLM Orchestration, Topic Clustering, Sentiment Analysis, NLP

### *Graduate Teaching Assistant*, Illinois Institute of Technology Chicago, Illinois, USA

Jan 2026 – Present

- Orchestrated academic support for 100+ students across Coursera-delivered database courses, triaging and resolving 150+ technical tickets per semester via Salesforce and Asana workflows, reducing average student resolution time by 45% through systematic categorization and priority-based routing.
- Applied and taught data modeling best practices including normalized schema design (1NF through 3NF), indexing strategies, stored procedure architecture, and query optimization across 30+ database projects, documenting specifications and design decisions that improved project delivery rates by 35%.
- Mentored 25–30 students on end-to-end project tech-stack selection guiding ORM choices, partitioning strategies, and multi-table join optimization, replacing ad-hoc advising with structured review cycles and reusable documentation that reduced unresolved blockers to zero.

**Skills:** Database Systems, SQL, Stored Procedures, Schema Design, Data Modeling (3NF), Salesforce, Asana, Technical Documentation

## **Data Scientist Labelmaster, Chicago, Illinois, USA**

Jan 2025 – May 2025

- Achieved enrichment of 440K CRM records from 52% to 89% completeness, enabling 30% improvement in account executive targeting accuracy for 50+ Chicago-based sales users, by executing XGBoost imputation and Sentence-BERT feature engineering to replace a manual audit process with an automated data quality scoring pipeline.
- Architected a holdout-based A/B testing framework with 95% statistical confidence gates, reducing post-deployment rollback incidents to zero across all pipeline rollouts, by designing and enforcing data accuracy checkpoints and validation protocols that protected 50+ active sales users from underperforming model versions before production release.
- Improved lead-scoring model precision by 15% across 270K CRM records, reducing inference time 2x and memory usage by 40%, by constructing GPU-batched Sentence-BERT pipelines stored in a vector index for real-time similarity retrieval.
- Eliminated 560 hours per year of manual data review, achieving 0.89 F1 score with drift detection and model monitoring for production retraining and reducing team workload by 62%, by tuning XGBoost classifiers with Optuna and SMOTE to handle class imbalance at scale.

**Skills:** Python, XGBoost, Sentence-BERT, Optuna, SMOTE, Scikit-learn, Vector Index, GPU Optimization, A/B Testing, MLOps, Drift Detection, Model Monitoring, Feature Engineering, NLP, CRM Analytics, Statistical Inference

## **Data Scientist, Tata Consultancy Services (TCS), India**

Aug 2021 – Dec 2023

- Delivered a credit risk segmentation report for a Canadian regional Equifax portfolio, by training an XGBoost classifier on masked historical credit records engineering 8 features (payment velocity, utilization ratio, delinquency recency, balance trajectory), surfacing three-tier risk classifications and default probability scores in a Looker Studio executive dashboard.
- Accelerated cross-functional discrepancy detection by 30% and reduced reporting cycle time by 20% across Finance, Marketing, BI, and Sales teams, by productionizing the MigrationWatch dashboard suite in Power BI and Looker Studio with unified data definitions, lineage documentation, and domain-specific business rules as a governed data contract layer.
- Designed and owned the MF-GCP Validation Engine, a column-level data quality framework in Python, SQL, and BigQuery verifying 1,200+ field mappings across Equifax Mainframe-to-GCP migration batches, enforcing metadata standards, PII compliance with Virtru encryption, and iterating the correction feedback loop from 60% to 99.6% match rate, eliminating 50+ monthly finance escalations.
- Operating within an Agile delivery model, cut validation runtime by 67%, cloud compute costs by 35%, and manual intervention by 80% across 1TB+ daily Equifax files, by re-engineering 50+ Python validation scripts into partitioned and clustered BigQuery-native SQL with modular stored procedure architecture and automated anomaly alerting.
- Compressed quarterly compliance audits from 5 days to 2 days and accelerated data quality degradation detection by 30%, by architecting a production drift monitoring pipeline using Apache Airflow and Google Composer that surfaced Population Stability Index alerts and metadata lineage signals into Looker Studio dashboards for 20+ executive stakeholders.

**Skills:** Python, SQL (stored procedures, partitioned/clustered tables), BigQuery, MS SQL Server, XGBoost, Apache Airflow, Google Composer, GCP, Looker Studio, Power BI, Virtru, Data Quality Frameworks, Metadata Management, Data Lineage, PII Compliance, ETL Pipelines, Schema Validation, Population Stability Index, Customer Segmentation, Credit Risk Modeling, Agile Delivery

## **PROJECTS**

---

### **Hallucination Hunter (RAG, LangChain/LlamaIndex, Arize Phoenix, MLflow)**

- Reduced RAG hallucination rate by 35% on a 500-document Wikipedia corpus by implementing RAGAS evaluation metrics (faithfulness, answer relevancy, context precision) with Llama 3 as judge via Groq API, achieving groundedness scores above 0.85 across 200+ test queries and validating retrieval accuracy using FAISS vector search with all-MiniLM-L6-v2 embeddings.
- Built an automated CI/CD evaluation pipeline using GitHub Actions and Arize Phoenix OSS that blocked merges when RAGAS faithfulness scores dropped below threshold, preventing 3 regression incidents across 6 deployment cycles and maintaining retrieval accuracy above 90% in LangChain RAG workflows tracked via MLflow.

### **Customer Churn Prediction (XGBoost, LightGBM, AWS SageMaker, SHAP)**

- Deployed XGBoost + LightGBM ensemble on AWS SageMaker achieving 85 AUC-ROC, identifying the top 3 churn drivers via SHAP analysis (tenure, product usage frequency, support ticket volume) and influencing retention budget reallocation that reduced at-risk customer base by 12% in high-value segments.
- Engineered an MLflow experiment tracking system with automated retraining triggers maintaining model accuracy above 82% across 6 months of production drift, processing 50K predictions/month via a FastAPI endpoint with zero downtime.

### **Text-to-SQL Analytics Agent (LangChain, LlamaIndex, ChromaDB, DVC, MLflow, Docker, AWS)**

- Built a LangChain-powered Text-to-SQL agent using ChromaDB for schema-aware vector retrieval and a self-correction feedback loop to resolve ambiguous multi-table joins across 6 database schemas, reducing analyst query turnaround from 3+ hours to under 10 seconds for standard queries and eliminating manual SQL writing for non-technical users.
- Achieved 84% exact-match accuracy on the Spider benchmark by implementing MLflow-tracked prompt versioning across 40+ iterations, containerized the full pipeline with Docker and DVC for reproducible evaluation, and deployed via FastAPI on AWS Lambda with CloudWatch monitoring and S3-backed artifact storage.

## **Education**

---

- Master of Data Science, Illinois Institute of Technology, Chicago — GPA: 3.72 / 4.0 Jan 2024 – May 2026
- Bachelor of Engineering, Siddaganga Institute of Technology, VTU University — GPA: 3.4 / 4.0 Aug 2017 – Aug 2021