Vikas **Shukla**

LinkedIn | 716-903-2501 | Portfolio | vikasofficial927@gmail.com | GitHub

SUMMARY

Innovative and collaborative Data Scientist with 7+ years delivering large scale models. Detail oriented with a track record of structured execution. Experienced in cloud scale data solutions focused on market behavior analysis, insight generation, and decision support.

EXPERIENCE

Data Scientist **Capital One** *McLean, VA* **10/2024 - Current**

- Identified instability in high-volume card transaction signals that led to misclassification during peak shopping periods. Re-engineered the monitoring layer using XGBoost + segmented time series diagnostics, improving fraud risk prediction accuracy by 17% and reducing false positives by 22%, directly lowering customer friction and call-center escalations.
- Built consolidated behavioral dataset from 40M+ card transactions to replace scattered primary investigations. Introduced a repeatable insights pipeline that business teams used to size fraud exposure by segment, cutting manual analysis cycles by 60% and enabling faster go-to-market decisions for product and risk
- Designed and validated scoring framework combining ensemble models with business rules to surface early indicators of account compromise. Reduced average investigation time by 45 hours/month, freed analyst capacity, and improved time -to-action for fraud policy teams.
- Directed root cause analyses for Risk Management to isolate KPI fluctuations and enforce compliance with internal standards and US government.
- Led root cause analysis on metric swings for Risk Management, operationalized SQL diagnostics that caught data-quality issues before KPI impact in ~ 80% of cases and reduced variance in weekly metrics ~20%, documented fixes and aligned with internal standards and U.S. regulatory expectations.
- Partnered with engineering to productionize fraud-detection models, automate thresholds, and deploy explainability proxies for frontline teams. This increased model adoption rate to 98% and significantly improved stakeholder trust.
- Executed Risk Management and Audit testing using Test of Design and Test of Effectiveness for data pipeline controls. Mapped controls to risks and KRIs, reconciled populations, sampled and re-performed checks, and produced traceable audit evidence.

Data Scientist **Bayer** *St. Louis, MO* **07/2023 - 10/2024**

- Transformed aging analytics environment into clean and modern reporting system. Rebuilt RMarkdown workflows, containerized all analytical steps in Docker, and delivered platform that ran smoothly every time. Reporting failures dropped by 90%.
- Implemented scheduled job automation to generate updated reports on monthly basis, ensuring stakeholders receive most recent insights regularly.
- Built predictive models for complex trial and breeding data using Gradient Boosting Regressor and regularized Logistic model after detailed feature exploration. Tuned model behavior through repeated cross validation and precise threshold adjustments to balance precision and recall. Delivered a performance jump from 0.78 to 0.86 with 14% lift in F1 while cutting model iteration time from three days to one.
- Operationalized R Shiny application for breeding data visualization and model training. Added logistic regression with regularization and stratified cross validation, and exposed ROC and PR curves with simple threshold tuning. Improved AUC from 0.78 to 0.86, lifted F1 by 14 percent, and reduced iteration time from 3 days to 1.
- Produced complex plots and tables for large data analysis in LaTeX reports.

Software Engineer ML **Amazon** *Bellevue, WA* **06/2022 – 04/2023**

- Assembled notebook for Data Science team to securely explore, train, and deploy models, addressing need for streamlined model development and deployment processes, resulted in 40% reduction in model development time.
- Implemented comprehensive notification quality assessment framework, incorporating analysis of user settings, reachability metrics, and A/B testing results across push, SMS, and email channels. This initiative addressed challenge of balancing growth with user perception and resulted in 20% improvement in user engagement and retention rates for Alexa.
- Developed and implemented automated business logic for core marketing experiments, including A/B, Auto-Targeting, and Multivariate Testing resulted in higher conversion rates, increased ROI, and improved strategy effectiveness.
- Launched uplift-based targeting with guardrails to reduce notification fatigue, and cut total sends by ~22 percent with no loss in engagement, lowered SMS/email delivery failures by ~30 percent via reachability filters, and lifted incremental conversion by ~8 percent.

Data Scientist Intern **Teave Tech** *Buffalo, NY* **06/2021 – 12/2021**

- Conceptualized in Data transport, storage and cleansing techniques, microservice development, implementing Machine learning models and supporting deployment of a production SaaS platform.
- Developed containerized microservices for model scoring and feature extraction with Python and FastAPI. Set up CI and delivery with unit tests, with unit tests, logging, monitoring, and clear runbooks for production support.

Senior Research Aide | **University at Buffalo** | ***Buffalo, NY*** | **07/2021 – 12/2021**

- Conceptualized in Data transport, storage and cleansing techniques, microservice development, implementing Machine learning models and supporting deployment of a production SaaS platform.
- Developed containerized microservices for model scoring and feature extraction with Python and FastAPI. Set up CI and delivery with unit tests, with unit tests, logging, monitoring, and clear runbooks for production support.

Data Scientist | **SGS Tekniks** | ***Gurugram, IND*** | **11/2014 - 05/2018**

- Implemented Linear Regression model for SGS customer forecasts, boosting accuracy by 10% in AUC and reducing RMSE by 12% for order date prediction, while also decreasing RMSE by 5% for quantity prediction, enhancing decision-making and operational efficiency.
- Developed incident categorization models by integrating parametric and non-parametric methods (KNN), surpassing 95% accuracy. This approach streamlined incident management processes, reducing misclassifications and downtime, resulting in improved operational efficiency
- Consolidated vendors and created volume advantage, resulted in 12% overall reduction in marketing costs.
- Automated Python and SQL pipelines to ingest orders, tickets, and marketing data with schema and null checks, and scheduled daily model retrains. Improved data freshness from weekly to daily and saved about 25 analyst hours per month.
- Built experiment framework for marketing models with power analysis and CUPED variance reduction, and applied uplift modeling to target high impact segments. Delivered 7 percent lift in incremental conversion at flat spend and narrowed confidence intervals by about 25 percent.

EDUCATION

Master of Science | **Trine University** | ***Reston, VA*** | **02/2025**
- Information Studies

Master of Science | **University at Buffalo** | ***Buffalo, NY*** | **08/2020 - 02/2022**
- Data Science

GPA: 3.7 / 4

Bachelor of Technology | **SRM University** | ***Chennai, IND*** | **08/2010 - 05/2014**
- Electrical and Electronics Engineering

GPA: 8.6 / 10

PROJECTS

FRAUD DETECTION PROJECT | *Stack: Python, SQL*
- Engineered features (target encoding, transaction patterns), cleaned redundant attributes, and trained XGBoost classifier.
- Validated performance using decile analysis, Lorenz curve and Gini Coefficient for discrimination strength.
- Top two decile captures majority of fraudulent transaction, achieved high recall while minimizing false positives.

CANCER GENE DETECTION | *Stack: R, Python, SQL*
- Reduced dimension from more than 20k features to 500 features (99.7% reduction).
- Applied multiple clustering multiple algorithms on principle components in order to cluster patients.
- Validated clusters with original label present.

FORECAST BIKE RENTALS | *Stack: Python, SQL*
- Identified and plotted features gave high and low correlation with response variable. And anticipate total rentals per hour for last quarter of 2012 by using Gradient Boosting Algorithm.
- Estimated total rentals per hour for last quarter of 2012.

INTELLIQUERY ENGINE (LLM+RAG) | *Stack: Python, FastAPI, FAISS, SQL, React Native*
- Designed retrieval-based document query system using embeddings and FAISS vector search, enabling semantic search across 10K+ document chunks with average response time under 800 ms and reducing unsupported responses by ~60%.
- Implemented rule-based intent checks with lightweight classification to route finance and legal queries, improving query accuracy by ~45% while restricting responses to source documents and reducing manual validation effort by ~50%.
- Built and deployed scalable FastAPI services integrated with the CaseFit React Native application, delivering reliable low-latency query responses at production scale.

SKILLS

- Python | Pandas | Numpy | Matplotlib | scikit-learn | R | R Shiny | MATLAB | RS Connect| C++ | SQL | NoSQL | Git | Redshift
- Regression | Classification | Clustering | Statistics | Databases | Snowflake | Time Series | NLP | Deep Learning | LLM | RAG
- AWS | Azure | Cloud Computing | CI/CD | Quick Sight | Tableau | Sagemaker | Databricks | OOPs | SAS | Machine learning LifeCycle