Aakash Kunarapu

3302810912 — aakashkunarapu17@gmail.com — LinkedIn — Portfolio

Summary — Data Scientist with hands-on experience building and deploying Python-based data pipelines and machine learning models to enhance compliance monitoring and risk mitigation. Proven expertise in applying statistical analyses, including regression and hypothesis testing, to drive data-driven decision making. Successfully increased proactive issue resolution by 20 percent through effective model implementation and cross-team collaboration. Holds an M.S. in Computer Science and demonstrates strong proficiency in tools such as SQL, TensorFlow, and Scikit-learn.

Education

Kent State University **Aug 2023 - May 2025**
Master of Science in Computer Science *GPA: 3.8*
Kakatiya University **Aug 2019 - May 2022**
Bachelor of Computer Applications *GPA: 3.5*

Professional Experience

Genpact Hyderabad,India. **Jun 2022 – Jul 2023**
Data Scientist

- Trained a gradient-boosted-tree risk model in scikit-learn; 25 engineered signals and Bayesian hyper-parameter tuning lifted F1-score to 0.87, capturing 78% of high-risk cases and cutting manual review hours 25%.
- Designed daily ETL workflows (PySpark Redshift) orchestrated in Airflow, delivering subhourly, audit-ready tables to BI tools and downstream ML jobs.
- Re-wrote complex joins and window functions, slashing dashboard refresh latency 60% and reducing cloud warehouse cost by $4 thousand per month.
- Audited 400+ product integrations across nine regions, mining telemetry logs, transactional data, and API traces to surface systemic UX and compliance gaps.
- Launched a company-wide Tableau analytics layer, giving Product, Legal, and Engineering instant visibility into incident spikes, churn signals, and conversion funnels.
- Ran data-driven workshops with 120 external dev teams; model-guided remediation boosted first-login success 12% MoM and improved NPS scores by 8 pts.
- Authored SOPs for data validation, model versioning, and alert thresholds; framework adopted by a 15-member global analytics squad for all new ML deployments.
- Automated model retraining in Airflow; MLflow triggers when population-stability drift exceeds 5%, keeping AUC within two% of baseline for six straight quarters.
- Re-engineered Redshift tables with sort and distribution keys plus incremental materialised views, trimming query latency from one-hundred-eighty to forty-five seconds and saving four thousand dollars per month in compute.
- Deployed the risk scoring service as a Docker container on AWS ECS with a blue–green rollout, achieving zero downtime releases and a median inference latency of two-hundred milliseconds.

Genpact Hyderabad,India. **Jan 2022 – May 2022**
Data Science Intern

- Queried and cleaned 75 K daily event records (PostgreSQL + Pandas) to uncover data-quality defects and user-journey drop-offs; findings fed the senior team's weekly KPI review.
- Prototyped a 25-feature store (time-series drift, error frequency, behavioral entropy) that later powered a production scoring model.
- Built an interactive Tableau dashboard (login success, latency, regional heat-maps) that trimmed triage lead-time from 3 days to 1 day.
- Translated analytics into concise remediation playbooks, helping partner engineering teams close all open issues before quarterly OKR checkpoints.
- Authored PostgreSQL window-function queries to build weekly retention cohorts; analysis exposed a 14% user-drop-off in APAC, informing the product roadmap for Q3.
- Added Pandas/Pydantic checks to the ETL prototype, cutting downstream null-pointer exceptions 30% and raising pipeline reliability to 99.7%.
- Normalised high-volume telemetry tables and introduced surrogate keys, trimming complex join execution time 50% and cutting analysis runtime from two hours to fifty minutes.
- Built a parameterised SQL harness that replayed authentication edge cases on demand, reducing feature validation cycles from one day to ten minutes.
-

Projects

Skytrax Reviews Analysis for British Airways - (Data Science Virtual Internship) **Spring 2025**

- Analyzed three thousand nine hundred forty passenger reviews with a VADER-based sentiment and NLP pipeline, generating a data-driven view of customer perception across cabin classes and route lengths.
- Quantified sentiment split at fifty-six percent positive versus forty-one percent negative, isolating long-haul flights as the primary source of adverse feedback.
- Extracted and ranked topic clusters; cabin-crew professionalism and premium-economy comfort emerged as the top positive drivers, while seat ergonomics and baggage handling dominated the negative cohort.
- Delivered an insight brief to customer-experience leads that recommended crew service refreshers and seat-upgrade roadmaps, projecting a four-percent uplift in Net Promoter Score on transatlantic routes.

Customer Churn Analysis using Data Mining Techniques **Spring 2024**

- Cleaned and normalised the IBM telecom churn dataset of 7,043 customers in Python, then profiled tenure, contract type, and monthly charges to guide targeted feature engineering.
- Balanced the target with SMOTE plus random undersampling, created 15 behavioural and billing features, and tuned Random Forest and Decision Tree models to 93% and 92% accuracy while lifting minority recall 21%.
- Used SHAP to rank churn drivers, confirming month-to-month contracts, premium add-ons, and high support demand as the most influential factors.

Marvel Universe Character Network Analysis **Fall 2023**

- Modelled a 6 426-node, 167 219-edge Marvel social graph from open "hero–comic" data, normalised and de-duplicated records, loaded them into NetworkX, and used degree, closeness, and betweenness centrality to pinpoint Iron Man, Captain America, and Spider-Man as the dominant influencers while flagging bridge nodes whose removal would splinter the network.
- Applied Greedy Modularity plus Girvan–Newman community detection to reveal 14 cohesive hero factions, exposing intra-team affinity patterns that mirror real-world social clustering and informing broader graph-theoretic research.
- Built interactive matplotlib and Plotly dashboards that visualise degree distributions, community boundaries, bottlenecks, and core–periphery splits, then distilled the insights into a report advocating bridge monitoring and core-node reinforcement for resilient network design.

Ensemble Model for Sepsis Detection **Spring 2024**

- Built an early-warning sepsis model on the PhysioNet 2019 ICU dataset (40,336 patients, 41 hourly vitals) by cleaning raw CSV files with Python, Pandas, and NumPy and scaling features with min-max normalisation for a consistent modelling base.
- Balanced the severe class skew with a two-stage pipeline—random undersampling then SMOTETomek oversampling—equalising class counts at 138 099 each and boosting minority recall 22 percentage points in cross-validation.
- Trained a Keras LSTM for temporal patterns and a scikit-learn Gradient Boosting classifier for static signals, integrating their outputs in an ensemble that reached 85.2 percent accuracy and 0.93 AUC up to six hours before clinical detection.

Interactive Mutual-Fund Performance Dashboard (D3.js) **Spring 2025**

- Cleaned and normalised the a1-mutualfunds.csv dataset in JavaScript and D3, parsing category labels, one-, three-, five-, and ten-year returns, yield, expense ratio, and manager tenure to create a tidy data frame ready for visual analysis.
- Engineered an interactive dashboard with five linked views — pie, bar, scatter, stacked bar, and parallel-coordinates — each built in D3.js and enhanced with tooltips, hover animations, and a dropdown axis filter, enabling multi-dimensional exploration of fund performance and cost structures.
- Optimised the D3.js dashboard for any screen size and surfaced three findings: International Stocks and Large Value lead fund share, three-year returns outstrip shorter and longer horizons, and short- to mid-term returns move in tandem, all documented for future investment analysis.

Skills

Programming: Python, Java.

Databases: SQL, PostgreSQL, MongoDB.

Big Data: Apache Spark.

Cloud: AWS (EC2, S3, SageMaker).

Statistics: Descriptive & inferential stats, hypothesis testing, regression, PCA, probability.

Machine Learning: Linear / Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, SVM, K-Means, NLP, feature engineering, model tuning.

Visualization: Tableau, Power BI, Gephi.

Tools: Anaconda, Jupyter, Git, Agile, OOP.

Certification

- Python for Data Science and Machine Learning.
- Introduction to prompt engineering for Generative AI.
- Big Data Analytics with Hadoop and Apache Spark.