Hyma Roshini Gompa Data Scientist

+1 (940) 843-6984 | hymaroshini7@gmail.com | LinkedIn | GitHub

SUMMARY

Data Scientist with 4+ years of experience delivering end-to-end predictive analytics and machine learning solutions for high-volume consumer platforms. Expert in building scalable data pipelines, developing and deploying ML models (Random Forest, XGBoost, LightGBM), and translating complex data into actionable business insights through customer segmentation, retention strategies, and KPI-driven dashboards. Proven track record of driving measurable impact on churn reduction, revenue optimization.

SKILLS

Data Engineering & Cloud Analytics: Scalable Data Pipelines, ETL/ELT, Data Modeling, Feature Engineering, Data Quality Frameworks, Data Validation & Anomaly Detection, Data Governance, Cloud Data Warehousing (Snowflake)
Programming & Distributed Processing: Python, SQL (Advanced Joins, Window Functions, Query Optimization), Apache Spark
Machine Learning & Advanced Analytics: Predictive Modeling, Churn & Retention Analytics, Customer Segmentation, Logistic Regression, Random Forest, XGBoost, Hyperparameter Tuning, ROC-AUC, Precision/Recall, Resampling Techniques
MLOps & Deployment: Production ML Pipelines, Automated Model Scoring & Monitoring, Containerization (Docker), Orchestration
Generative AI & LLMs: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), LLM Pipeline Development
BI & Visualization: Executive Dashboards, KPI Frameworks, Power BI, Tableau, Trend & Root-Cause Analysis
Stakeholder & Program Management: Cross-Functional Collaboration, Requirements Gathering, Metric Definition, Executive Communication, Analytics-to-Business Translation

EXPERIENCE

Data Scientist | Enigma Technologies **May 2025 – Present | USA**

- Own a flagship end-to-end predictive analytics initiative for proactive customer retention, analyzing large-scale behavioural and transactional data using SQL, Apache Spark, and Snowflake to identify churn risk and optimize engagement strategies.
- Design and maintain scalable, production-grade data pipelines using Python, SQL, and Spark, performing data validation, feature engineering, and data quality checks to support reliable model training and scoring.
- Design and implement advanced ML pipelines using Logistic Regression, Random Forest, and XGBoost, incorporating feature engineering, hyperparameter tuning, and resampling techniques to improve model performance metrics like precision & ROC-AUC.
- Deploy automated model scoring and monitoring workflows, leveraging Docker and Kubernetes-aligned environments to enable near real-time churn risk visibility and faster, data-driven intervention for product and marketing teams.
- Apply Large Language Models (LLMs) by implementing Retrieval-Augmented Generation (RAG) pipelines using LangChain to enhance insight retrieval from internal documents, model outputs, and customer data.
- Collaborate cross-functionally and deliver executive-ready insights, translating ML and LLM outputs into customer segmentation, personalized engagement strategies, and KPI-driven dashboards using Power BI and Tableau.

Data & Cloud Analytics | Trulogik **Aug 2024 – May 2025 | USA**

- Led end-to-end analytics delivery by acting as the primary liaison between business stakeholders and engineering teams, translating complex reporting requirements into scalable dashboards and automated insights, cutting reporting cycle time by 35%.
- Designed and owned executive-ready Power BI and Tableau dashboards, tracking critical KPIs and performance metrics while conducting deep trend and root-cause analysis to drive data-backed leadership decisions.
- Built and automated backend analytics workflows using SQL and Python, streamlining data preparation and validation processes and reducing reporting errors by 20% across multiple datasets.
- Implemented proactive data quality monitoring and anomaly detection, ensuring high data accuracy, reliability, and on-time delivery of insights for operational and strategic reporting.
- Standardized reporting logic, data models, and documentation, enabling repeatable analytics processes and aligning cross-functional teams around consistent, trusted business metrics.

Data Engineer Technical Program Analyst | Josh Technology Group **Apr 2021 - Jul 2023 | India**

- Designed and automated scalable data pipelines integrating large multi-source datasets using SQL and Python, enabling reliable KPI reporting and operational analytics while reducing manual data preparation effort by 30%.
- Developed and optimized complex SQL transformations and analytical queries for recurring and ad hoc reporting, performing advanced trend, variance, and root-cause analysis to surface anomalies and insights for leadership review.
- Built and maintained production-grade Power BI and Tableau dashboards to visualize KPIs, trends, and data quality indicators, supporting executive and operational decision-making across teams.
- Implemented robust data quality frameworks including duplicate detection, missing-data analysis, and validation workflows, cutting data discrepancies by 25–30% and improving trust in enterprise reporting.
- Partnered with cross-functional stakeholders to gather requirements, define standardized metrics, and clarify data structures, ensuring consistent, repeatable analytics outputs aligned with business objectives.
- Documented end-to-end data flows, transformation logic, and validation rules, strengthening reproducibility, governance, and compliance across multiple analytics initiatives.

EDUCATION

Master of Science in Data Science — **Aug 2023 – May 2025**
University of Maryland, Baltimore County — Baltimore, MD
Bachelor of Science in Computer Science — **Aug 2018 – May 2022**
Lovely Professional University — Punjab, India

CERTIFICATIONS

- Microsoft Power BI
- Microsoft Azure Fundamentals
- SQL (Intermediate–Advanced)
- Python for Data Analytics

PROJECTS

AI-Powered Legal Document Summarization (NLP)

- Built a transformer-based legal document summarization platform using Legal-BERT, LED-16384, and RAG, automating multi-format document processing and reducing review time by 70%.
- Implemented vector-based semantic search and a Streamlit UI to enable clause-level retrieval and interactive summarization.
- Evaluated and optimized model performance using ROUGE, BLEU, and BERTScore, ensuring reliable summarization quality.

Real-Time Voting System

- Designed and implemented a real-time voting platform capable of processing 500K+ events per minute using Kafka, Spark Streaming, and PostgreSQL.
- Optimized data pipeline throughput by 45% via advanced batching strategies, stream tuning, and efficient event processing.
- Developed a user-facing Streamlit interface for live vote tracking, analytics visualization, and operational monitoring.
- Ensured scalability and reliability of the pipeline for high-velocity, concurrent event streams.

End-to-End Azure Data Engineering Pipeline

- Designed and implemented a Medallion Architecture pipeline (Bronze → Silver → Gold) for e-commerce data, enabling scalable, reliable, and business-ready analytics.
- Ingested, transformed, and enriched data from HTTP endpoints and SQL databases using ADF and Databricks (PySpark), optimizing performance for large-scale datasets.
- Delivered actionable insights via dashboards in Power BI and Tableau while implementing production-grade workflows with automation, monitoring, and error handling.