

Hamza Mohd Zubair

Data Scientist & AI Engineer

hamzamohdzubair@gmail.com +91 76009 53379 hzubair.com
in/hamzamohdzubair hamzamohdzubair

Data scientist with **9+ years** of deploying ML in production across **consumer healthcare**, **retail**, **BFSI** and **QSR**. Productized demand forecasting, supervised learning, reinforcement learning, deep learning, NLP, computer vision, Markov models and **LLM** pipelines. Also builds independent AI products and speaks on AI.

EXPERIENCE

2026 EPAM Systems — Jan 2026 – Present

Senior Data Scientist

Client: **Opella** (Sanofi Consumer Healthcare). Weekly e-commerce sell-out forecasting for Amazon and other retailers in 8+ markets, on Snowflake, SageMaker Pipelines and MLflow with Chronos-2, Prophet, MFLES and Ridge models.

- **Product discontinuation lifecycle** in the production forecaster: phased-out products fade to zero, are capped by retailer stock and hand volume over to their successor.
- **Model vs. covariate study** over 27 rolling backtest origins: showed model choice is worth 6–9 WAPE points while external data adds under 1, and ensembling gains about 2.
- **Plackett-Burman factorial screen** of feature groups in 16 parallel SageMaker runs, replacing one-feature-at-a-time testing; cut screening time by **63%** (246 → 91 min) and SHAP cost by about 60%.
- Redesigned **cannibalization detection** to remove false positives from zero-inflated sales; found by reproducing runs that production had silently switched to the wrong Chronos checkpoint.
- Audited the 9 forecast-exclusion rules (12.7% of recent volume excluded, several rules dead) and proposed a three-tier redesign; ranked cross-market data-quality issues by revenue impact.
- Self-initiated internal tools: **nx**, a Snowflake data-access and run-reproduction library used in 40+ notebooks; **benchpress**, a FastAPI app for benchmarking forecast runs; and **ops**, a Rust CLI tying Jira, GitHub and SageMaker test runs together.

2024 Turing.com — Aug 2024 – Apr 2025

Senior Data Scientist

- Part of the team training **Gemini** with supervised fine-tuning (SFT).
- **Complex chart production tool** (sole designer and developer): LLM agent pipelines that cut chart generation time by **93%**, from 45 min to 3 min, at about ₹0.4 per chart.
- Built **RAG** with optimized inference and caching, and production agentic workflows in **LangChain** and **DSPy** with observability and evaluation built in.

2018 Cartesian Consulting — Aug 2018 – Aug 2024 · 3 promotions

Senior Lead Data Scientist — Apr 2022 – Aug 2024

- **Campaign prioritizer** (solo): contextual bandits (Thompson sampling, UCB) that pick each customer's campaign; deployed for 2 major clients (pharma retail, brokerage), lifting click-through and revenue.
- **Automated MLOps platform** (team of 2): configure, deploy and maintain XGBoost / LightGBM models with automated feature engineering, tuning, calibration and retraining. Used by nearly all Solus clients.

Lead Data Scientist — Jul 2020 – Mar 2022

- **SMS composer** (solo): generative NLP module and UI that breaks a message down by purpose, tone and personalization fields and writes multiple variants from a prompt.
- **Multi-touch attribution** (solo): Markov-model attribution adapted from digital to CRM and direct marketing.

Senior Data Scientist — Apr 2019 – Jun 2020

- **Open rate predictor** (led a team of 3): NLP pipeline that predicts email open rates from subject lines, plus a semantically controlled LSTM that writes subject lines. Deployed for a major Indian bank.

Senior Analyst – AI Aug 2018 – Mar 2019

- **Product segmentation** (solo): LDA topic modelling to auto-categorize products for targeted marketing.

2017 **Untrodden Labs** Jan 2017 – Aug 2018

Founding Employee & AI Lead

- Bootstrapped startup building an Alexa-like desktop robot with voice commands, vision, face tracking and an animated emotive display.
- Led a team of 5 building speech intent classification (SVM, random forests), novelty detection (one-class SVM), information extraction (grammar and dependency parsing, SyntaxNet) and live object detection over 90 classes (YOLO, MobileNet SSD, ResNet v2).

PRODUCTS

Redeo · Android, on Google Play play.google.com

Spaced-repetition flashcards where you write the answer from memory, an on-device model scores your recall and FSRS schedules the next review. Kotlin, on-device ML, Room.

dontwatch.in · Web app dontwatch.in

Learn a university lecture series by answering questions linked to the exact minute of each lecture; AI grades answers against what the lecturer said. Cloudflare Workers, LLM grading.

EDUCATION

MS, Computer Science (ML/AI) · Woolf University 2024 – 2026

Certification, Reinforcement Learning · IISc Bangalore 2019 – 2020

MSc, Cognitive Science · IIT Gandhinagar 2013 – 2015

BTech, Bioinformatics · VIT Vellore 2009 – 2013

SKILLS

Languages	Python, Rust, Kotlin, SQL, R
ML	PyTorch, TensorFlow, Keras, scikit-learn, XGBoost, LightGBM, SHAP, Vowpal Wabbit
Forecasting	Chronos-2, Prophet, MFLES, rolling backtests, factorial experiments
Data & MLOps	pandas, NumPy, SciPy, Snowflake, AWS SageMaker, MLflow, FastAPI, GitHub Actions
LLM	LangChain, DSPy, RAG pipelines
Areas	LLMs, agentic AI, NLP, bandits & RL, computer vision, forecasting, attribution

AWARDS

Employee of the Quarter · Cartesian Consulting Q1 2020

Innovation Award, for work in NLP · Cartesian Consulting 2020

Merit-Based Scholarship · IIT Gandhinagar 2013 – 2015

Research Fellowship (€1,000) · ISCTE-IUL, Lisbon May – Jul 2014

TALKS

State of AI: Facts vs Hype · MINE Nov 2025

AI: Between Myth, Mind and Morality · BIFT, Bangalore Nov 2025

AI 101 · The Success Institute, Bangalore Oct 2025

Sustainable AI · St. Claret College, Bangalore Oct 2025

HackerTech hackathon, judge & panelist · VIT University, Vellore Dec 2019

AI for Startups · Nexus, US Embassy, New Delhi Aug 2017