Anshuman Singh

Gurugram, India
anshumanr434@gmail.com • linkedin.com/in/anshumansingh2023

> B.Tech (CS, AI/ML) applying for an Intern role. I am interested in automated interpretability, i.e.
> Turning manual analysis into repeatable pipelines that form and test structured hypotheses about model internals.
> Experience with feature attribution (SHAP/Grad-CAM), simple probing/classifiers, calibration, and scripting
> Evaluation/benchmarking loops. Python and PyTorch;
> I value rigorous experiments, reproducible code, and clear notes.

Education

SGT University, Gurugram: B.Tech, Computer Science (AI & ML Specialization)
2023 - 2027 • Current CGPA: 9.4/10.0
Relevant Coursework: Machine Learning, NLP, Data Engineering, Statistics, Linear Algebra, Algorithms

Experience

Machine Learning Team Lead | CodeSangam Hackathon 2024 - Gurugram
3rd Place, Healthcare Track • Feb 2024

- Collected Ayurveda symptom and remedy data from online sources and PDFs; organized in spreadsheets and JSON.
- Generated synthetic examples for underrepresented classes to balance the training set.
- Outlined an image data plan for dermatology (sampling and deduplication); focused on text for the final prototype.
- Built a Flask prototype (scikit-learn) with simple notes on labeling and evaluation.
- Led a 4-member team through data collection - labeling - training - deployment in 48 hours.

Data Science Team Lead | AIFusion Hackathon 2024 - Gurugram
Special Recognition • Sep-Oct 2024

- Created synthetic tabular datasets for campaign benchmarking; defined column names and what each feature meant.
- Cross-checked data from different sources for consistency; flagged obvious issues.
- Made simple charts and slides showing trends; kept notes on what we tried.

Projects

StreaxBot - Reddit Topic Quiz Platform (Recent, Oct 2024)
Tech: React, Node.js, Express, Firestore, Google Gemini API, Devvit SDK

- Built a daily quiz bot for a Reddit hackathon; users pick topics, bot scrapes Fextralife/Wikipedia, generates questions via Gemma-3n.
- Implemented real-time deduplication to prevent re-scraping and duplicate questions across daily updates.
- Stored scraped data in Firestore; wrote rules to update datasets auto-daily while filtering out already-used content.
- Pipeline: user input - web scraping - data cleaning - question generation - presentation; handled edge cases.
- Designed guardrail prompts and ran spot-check sampling on generated questions to correct low-confidence items.

LocalLibrary - Book Metadata Scraper
Tech: Python, Playwright, aiohttp, asyncio

- First real scraping project attempt; built pipeline to collect book metadata from Amazon, Anna's Archive, and other sites (for learning).
- Implemented rate limiting, user-agent rotation, and session management to avoid network overload during scraping.
- Cleaned duplicates, normalized ISBN/title fields; saved to JSON for downstream use.
- Later found a simpler API-based approach via LibGen, but this taught me scraping fundamentals and ethical considerations.

- Maintained a simple data dictionary and source log; documented dedup and normalization rules for reproducibility.

Optimizing Personal Loan Campaigns with Machine Learning
Tech: Python, scikit-learn, LightGBM, XGBoost, Random Forest, SHAP

- Internship project analyzing 10,000+ customer records to predict loan acceptance for a banking client (TechnoHacks Edutech, Summer 2024).
- Performed EDA: explored income distributions, family size, education, pin-code; middle-income groups had higher acceptance rates.
- Engineered polynomial interaction features (income x family size, pin-code x income) to capture non-linear relationships.
- Built *predictive ensembles* (XGBoost/LightGBM/Random Forest) with feature selection; used SHAP to explain drivers.
- Final model: 92.8% accuracy; delivered clear decision rules to reduce targeting costs while maintaining acceptance.
- Wrote a short report with methodology and results; documented assumptions and evaluation split for transparency.

Sentiment Modeling with Synthetic Augmentation
Tech: PyTorch, Hugging Face Transformers, LoRA, ADASYN

- Curated an initial corpus and hand-labeled multi-class sentiment; kept a brief labeling note for consistency.
- Cleaned text and created a stratified train/validation split to keep class balance stable across splits.
- Fine-tuned RoBERTa-base with LoRA adapters (Transformer-based *feature extraction*) on a local GPU.
- Handled class imbalance with ADASYN and synthetic examples for minority classes; macro F1 improved by ~12%.
- Ran error analysis (confusion matrix, common error buckets) and did manual spot checks to refine labels and prompts.
- Produced small, reusable artifacts: dataset CSV/JSON, preprocessing script, training/eval notebook, and a brief model card-style summary.

Skills

Python | NumPy | Matplotlib | scikit-learn | PyTorch | Hugging Face Transformers | LoRA fine-tuning | RoBERTa-base | Predictive ensembles | XGBoost/LightGBM | Stratified splits | Git | Linux

Data: collection (web/APIs), labeling, deduplication, leakage checks
Evaluation: accuracy, macro F1, calibration
Docs: concise notes and clear reports

Independent Learning & Experimentation

- Reading and summarizing papers to improve data/evaluation practices.
- Testing small models locally and keeping simple experiment logs.
- Writing quick benchmarking scripts to compare approaches on small datasets.

Achievements

- **9.44 CGPA** across 4 semesters (AI specialization)
- Top 3 winner in multiple AI/ML hackathons (200+ participants)
- Built and labeled datasets for NLP and healthcare projects from scratch
- Comfortable working independently on data tasks - organized, detail-oriented, keeps clear notes