HIMANSHU GUPTA

Data Scientist | NLP | Machine Learning
Portfolio | LinkedIn | GitHub

PROFESSIONAL SUMMARY

AI & Machine Learning undergraduate specializing in Natural Language Processing and applied Data Science, with end-to-end experience from text/data preprocessing and feature engineering to predictive modeling and deployment. Built an NLP spam classifier (98.84% accuracy, 95.24% F1-score) and a customer satisfaction prediction system (79.2% accuracy, 0.804 F1-score, XGBoost + SHAP) during a 6-month AI/ML internship. Comfortable across the ML stack — supervised and unsupervised learning, deep learning (ANNs, CNNs), and NLP (TF-IDF, Word2Vec, n-grams) — and focused on turning models into decisions people can act on. Seeking a Data Scientist or NLP-focused role.

CORE SKILLS

- Programming & Databases: Python, SQL
- NLP: Text Preprocessing (Tokenization, Stopword Removal, Stemming), Bag-of-Words, N-Grams, TF-IDF, Word2Vec, NLTK
- Machine Learning: Supervised & Unsupervised Learning, Regression, Classification, Naive Bayes, Ensemble Learning, Feature Engineering, Hyperparameter Tuning (GridSearchCV), Model Evaluation, SMOTE, SHAP Explainability
- Deep Learning: ANNs, CNNs, Backpropagation, Transfer Learning (MobileNetV2)
- Data Analysis & Visualization: Excel, Power BI, DAX
- Libraries: Pandas, NumPy, Scikit-learn, TensorFlow, Keras, XGBoost
- Tools & Deployment: Git, GitHub, Streamlit

PROFESSIONAL EXPERIENCE

AI/ML Intern — Labmentix Pvt. Ltd. *Jul 2025 – Jan 2026*

- Built and deployed end-to-end machine learning and deep learning pipelines across structured and image datasets, covering preprocessing, feature engineering, model evaluation, and deployment.
- Engineered a customer satisfaction (CSAT) prediction system with XGBoost, achieving 79.2% accuracy and a 0.804 F1-score by applying SMOTE for class balancing and GridSearchCV for hyperparameter tuning.
- Increased model transparency and stakeholder trust by integrating SHAP explainability and shipping the model as an interactive Streamlit application for real-time prediction and scenario analysis.
- Designed, trained, and benchmarked CNN-based brain tumor classification models (custom architecture vs. MobileNetV2 transfer learning), selecting the top performer via comparative evaluation.

PROJECTS

Email/SMS Spam Classifier — NLP | GitHub

Python, NLTK, Scikit-learn, TF-IDF, Naive Bayes, Streamlit

- Built an end-to-end NLP pipeline to classify messages as spam or ham, including tokenization, stopword removal, and stemming for text preprocessing.
- Compared CountVectorizer and TF-IDF feature extraction across five models (Decision Tree, Random Forest, Gaussian/Bernoulli/Multinomial Naive Bayes), evaluating each on accuracy, precision, recall, and F1-score.
- Selected TF-IDF + Bernoulli Naive Bayes as the final model, achieving 98.84% accuracy and a 95.24% F1-score.
- Deployed the model as an interactive Streamlit web application for real-time message classification.

Customer Satisfaction (CSAT) Prediction | GitHub

Python, Scikit-learn, XGBoost, SHAP, Streamlit

- Built an end-to-end ML pipeline combining structured and textual features to flag dissatisfied customers before escalation.
- Benchmarked Logistic Regression, Random Forest, Decision Tree, and XGBoost with SMOTE and GridSearchCV tuning; XGBoost delivered the best precision/recall balance (0.804 F1, 79.2% accuracy).
- Applied SHAP for model interpretability, surfacing the key drivers of customer satisfaction for business stakeholders.
- Deployed a Streamlit app enabling real-time CSAT prediction and feature-impact scenario testing.

Customer Behaviour & Purchase Decision Analysis | GitHub

Power BI, SQL, DAX

- Analyzed customer purchasing behavior and revenue drivers across segments using SQL and Power BI, translating findings into business strategy recommendations.
- Identified high-income adult customers as the top-value segment, contributing 50%+ of total revenue, informing targeted marketing prioritization.
- Found non-loyal customers spent ₹27 more per purchase on average, surfacing an opportunity to redesign the loyalty program.
- Evaluated marketing channel ROI using DAX measures, showing ad engagement drove the highest revenue while discount campaigns underperformed.
- Built a multi-page interactive Power BI dashboard with drill-through filters and cross-page slicers to support data-driven revenue and retention decisions.

EDUCATION

B.Tech, Artificial Intelligence & Machine Learning

Vivekananda Institute of Professional Studies, New Delhi | 2023 – 2027