AI-LeadSquad: Smart Lead Generation Using AI

In the modern B2B sales landscape, lead generation needs to evolve beyond manual scraping and filtering. AI-LeadSquad is an AI-powered system designed to identify, enrich, and prioritize high-quality leads using machine learning and natural language processing. This tool is tailored for private equity use cases like Caprae Capital, where identifying strategic, AI-aligned acquisition or partnership targets is mission-critical.

Frame work used:-scikit-learn ,pandas, SerpAPI + requests Streamlit, joblib

Working Principle

1. The user inputs a keyword (e.g., 'AI SaaS Startups')
2. The system scrapes leads in real-time using SerpAPI or uses predefined mock data
3. Scraped leads are cleaned, deduplicated (via fuzzy matching), and validated (email regex)
4. Text descriptions are converted into numerical vectors using TF-IDF Vectorizer
5. SelectKBest (chi^2) is applied to retain the most informative features (top 1000)
6. A machine learning classifier — Logistic Regression or Multinomial Naive Bayes — predicts the lead's industry
7. Each lead is also scored using a custom keyword relevance function
8. The enriched leads are displayed via a Streamlit dashboard with CSV export functionality

Model Approach and Justification

Data Collection-Used a real-world structured dataset of 7M+ global companies. Filtered down to 100,000 rows using the top 10 industry categories and generated descriptions from available metadata (industry, location, size).

Data Preprocessing-Dropped rows with missing Industry or too-short Description. Normalized text (lowercase, punctuation removal). Converted text into TF-IDF vectors Applied SelectKBest (chi^2) to retain top 1000 relevant features.

Model Selection-Two models were evaluated and tuned using Pipeline and GridSearchCV with 3-fold cross-validation.:

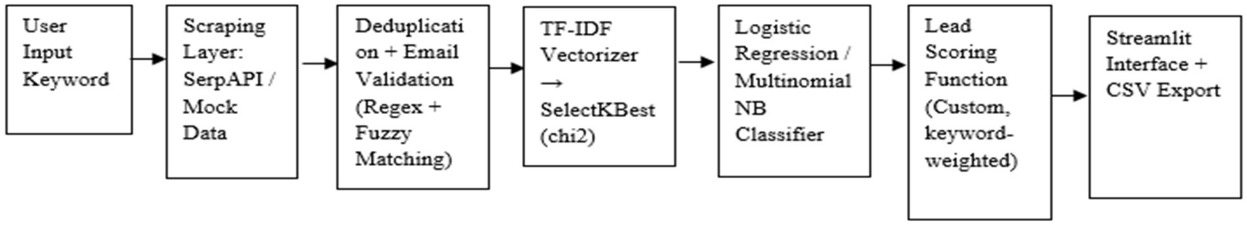
1. Logistic Regression (Deployed Model)- A linear model used for binary and multiclass classification. It learns weights for each feature to separate classes via a sigmoid function. Interpretable coefficients, Robust to overfitting (with regularization), and Works well with sparse data (TF-IDF)

Formula: $P(y|X) = 1 / (1 + e^{-(w \cdot x + b)})$

2. Multinomial Naive Bayes- A probabilistic classifier based on Bayes' Theorem that assumes feature independence, effective for document classification. Fast and efficient, Ideal for TF-IDF-like word count vectors and -Low resource consumption

Formula: $P(y|x_1, ..., x_n) \propto P(y) \times \prod P(x_i \mid y)$

Performance Evaluation-Both models achieved 100% accuracy, precision, recall, and F1-score on the cleaned and balanced dataset. Final deployed model: Logistic Regression, selected for its balance of speed, interpretability, and robustness.


User Input Keyword
Scraping Layer: SerpAPI / Mock Data
Deduplication + Email Validation (Regex + Fuzzy Matching)
TF-IDF Vectorizer → SelectKBest (chi2)
Logistic Regression / Multinomial NB Classifier
Lead Scoring Function (Custom, keyword-weighted)
Streamlit Interface + CSV Export

Conclusion

AI-LeadSquad demonstrates a full-stack ML pipeline that transforms unstructured web results into structured, scored, and classified B2B leads. This tool helps private equity firms like Caprae Capital identify operationally promising, AI-ready businesses faster. Future extensions include CRM integration, deeper embeddings (BERT), and real-time alert systems.