Caprae Capital Submission - Lead Scoring & Deduplication Tool

Tool Name: caprae_leadgen_scraping

Overview:

I chose the 'Quality First' approach to improve the most critical phase in the lead generation funneldata validation, cleaning, and enrichment. Rather than replicating scraping logic, my tool focuses on ensuring that leads are reliable, accurate, and prioritized for outreach.

Approach:

The app is built using Python and Streamlit, offering a lightweight web interface that allows users to upload CSV files of leads, automatically validate them, score their potential, and export cleaned results.

Model Selection:

A rule-based heuristic model was applied instead of a machine learning model. Lead quality is calculated based on:

- Email validity (regex validation)
- Free domain detection (e.g., gmail.com)
- Revenue tiering (parsed from 'Cr' to int)
- Employee strength
- Missing website or phone fields

Each factor contributes positively or negatively to a total lead score, which is then categorized into High, Medium, or Low priority leads.

Data Preprocessing:

Removed duplicates based on Company and Email.

Cleaned missing values and whitespace

Added Google-based LinkedIn search URL enrichment

Parsed numeric fields and normalized formats

Performance Evaluation:

Accurate filtering of invalid and free emails

Successfully prioritized leads based on real business indicators

Deduplication logic removed redundant contacts while preserving one unique row. Users can export both cleaned data and removed duplicates for auditability

Business Rationale:

Caprae seeks better leads, not just more leads. This tool ensures clean, enriched, and prioritized leads ready for sales engagement aligning directly with Caprae's value proposition. The interactive UI and CSV-based workflow makes it deployable into any sales funnel with minimal effort.

Submitted by: Sakshi Gupta