

# Project Report: AI-Powered Lead Prioritization Engine

**Candidate:** Kartik Dinesh Sangle

**RE:** Caprae Capital Intern Interview Pre-Work

## 1. Problem Analysis & Opportunity

The provided reference application, [saasquatchleads.com](https://saasquatchleads.com), successfully identifies potential leads based on broad categories. However, its primary limitation lies in its output: a high-volume, unsorted list of prospects. This creates a significant operational bottleneck, forcing a founder or sales team to manually sift through hundreds of entries to find a handful of high-potential targets. This process is time-consuming and inefficient, directly opposing the goal of leveraging technology to create scalable growth.

The opportunity was not merely to replicate the scraping functionality but to solve the underlying business problem: **how to transform raw data into actionable intelligence**. My goal was to build a tool that doesn't just generate a list, but delivers a prioritized action plan, providing true "horsepower" to a sales team.

## 2. Architectural Approach: Learning-to-Rank (LTR)

To address this challenge, I chose to build a full-stack application centered around a **Learning-to-Rank (LTR) machine learning model**. I deliberately avoided a simple, rule-based scoring system in favor of this more sophisticated, data-driven approach for three key reasons:

1. **Scalability:** A rule-based system is brittle and requires constant manual adjustment. An LTR model, powered by Gradient Boosting (XGBoost), can learn complex, non-linear relationships from data, making it more robust and adaptable over time.
2. **Holistic Analysis:** The model evaluates leads based on a multi-dimensional "fit" rather than a simple checklist. It can learn, for instance, that high revenue is most valuable when paired with a mid-range employee count, a nuance that simple rules often miss.
3. **Future-Proofing:** The architecture is designed to be retrained on a portfolio company's actual historical sales data (e.g., characteristics of won vs. lost deals). This would create a self-improving, company-specific asset that gets smarter over time, directly embodying the Caprae mission of using technology to unlock long-term value.

The technical workflow is as follows:

- **Data Acquisition:** A user's query is translated into a targeted search via the **Google Custom Search API**.
- **Concurrent Scraping:** The system uses a multi-threaded process to scrape company websites in real-time using **BeautifulSoup**, extracting key firmographic features from unstructured HTML.
- **Real-Time Inference:** The scraped and processed data is fed into the pre-trained **XGBRanker model**, which predicts a relevance score for each lead. The results are then sorted and displayed.

### 3. Model Selection & Justification

- **Model:** xgboost.XGBRanker
- **Objective:** rank:pairwise

I selected XGBoost for its industry-proven performance, speed, and its native support for ranking objectives. The rank:pairwise objective trains the model to learn the optimal ordering of a list by comparing pairs of documents (leads), which is more effective for ranking tasks than simple classification or regression.

To enable this within the challenge's constraints, I developed a full training pipeline where the model is trained once on a synthetic dataset generated by a "teacher" function. This demonstrates the end-to-end workflow of an LTR system, from data generation to real-time prediction, showcasing a production-ready mindset.

### 4. Business Impact & Alignment with Caprae's Vision

This project is a direct reflection of a founder-operator mindset: it's a tool built not for the sake of technology, but to solve a core business problem and create a competitive advantage.

- **For a Founder:** It saves dozens of hours of manual labor, allowing them to focus on building relationships and closing deals rather than on prospecting.
- **For Caprae:** It serves as a blueprint for a scalable, data-driven sales process that can be implemented across portfolio companies. It is a tangible example of how practical AI can be leveraged post-acquisition to immediately enhance operational efficiency and drive top-line revenue growth.

By moving beyond simple scraping to intelligent ranking, this tool demonstrates a commitment to creating systems that deliver not just data, but decisive, actionable insights.