

AI-Driven Lead Generation: Caprae Capital Challenge

Overview:

Caprae Capital seeks to identify high-quality acquisition targets by leveraging AI-readiness scores based on growth, scale, and maturity. This project creates a machine learning pipeline that clusters companies and ranks them using a composite lead score.

Dataset:

- Source: Kaggle's 7M+ Company Dataset
- Features: year_founded, employee size, growth estimate, domain, industry, location

Approach:

1. Data Cleaning & Feature Engineering:

- Removed nulls, calculated company_age
- Normalized employee size and growth metrics
- Constructed lead_score as a weighted metric combining size, growth, and maturity

2. Clustering:

- Used StandardScaler + PCA for dimensionality reduction
- Applied KMeans to segment businesses
- Identified cluster 0 as the most AI-ready group

3. Lead Extraction:

- Selected top 5% by lead_score
- Final filter: (cluster == 0) & (lead_score > 0.03)
- Output: CSV of top leads with domain, location, and LinkedIn

Why It Matters:

This pipeline aligns with Caprae's SaaS/MaaS strategy by filtering for scalable, growing, and mature companies likely to benefit from AI transformation. It prioritizes actionability and quality over volume.