# AI-Driven Company Intelligence Through Data-Driven Segmentation

SDS Datathon 2026 - Final Report

Team Minions
*Singapore Data Science Datathon 2026*

January 2026

**Abstract**

This paper presents a prototype system for deriving actionable business intelligence from company-level data through multi-dimensional clustering, lead scoring, and risk detection. We analyze 8,559 companies, engineering 15+ features across organizational structure, productivity, and data quality. Our approach is rigorously validated: we achieve a **Clustering Stability (ARI) of 0.94** via bootstrap analysis and validate our Lead Scoring model using a Gradient Boosting classifier with **93.11% accuracy**. Furthermore, t-tests confirm that high-priority leads have statistically higher unit efficiency ($p < 0.003$). Using K-Means ($k = 5$) and Isolation Forest, we identify 3,063 potential shell companies and distinct anomaly archetypes. Integration with Large Language Models (LLMs) and SHAP values provides transparent, interpretable explanations for every insight, bridging the gap between raw data and strategic decision-making.

## Contents

# 1 Introduction

## 1.1 Background and Motivation

The modern B2B marketplace is characterized by an overwhelming volume of company data, yet extracting actionable business intelligence remains a significant challenge (Chen et al., 2012). Decision-makers across sales, marketing, investment, and risk management functions need efficient tools to segment markets, identify high-value prospects, and detect potential risks (Provost and Fawcett, 2013).

Traditional approaches to company analysis often rely on manual review or simple filtering, which fails to capture the multi-dimensional nature of business entities. Machine learning techniques, particularly unsupervised clustering, offer promising solutions for discovering natural groupings within company data (Jain, 2010).

## 1.2 Project Objectives

This project develops a prototype system that transforms raw company-level data into interpretable business intelligence. By leveraging data analytics, machine learning techniques, and large language models (LLMs), our system generates data-grounded insights that help users understand how companies operate and compare with similar firms (Vaswani et al., 2017).

Our solution enables users to:

- Identify and group companies with similar characteristics or operating profiles
- Understand key differences and similarities within and across groups
- Highlight notable patterns, strengths, risks, or anomalies
- Demonstrate commercial value through actionable lead scoring and risk assessment
- Generate interpretable, data-grounded explanations using LLM integration

## 1.3 System Architecture

Figure 1 presents the end-to-end system architecture:

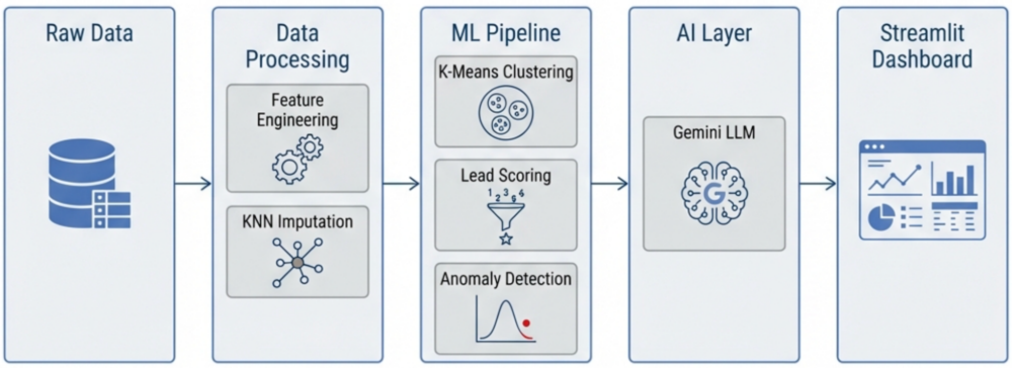


Figure 1: System Architecture: From raw data to actionable intelligence.

### 1.4 Commercial Value Proposition

The insights generated directly support multiple business functions:

- **Sales teams**: Prioritized lead lists based on quantitative scoring
- **Risk analysts**: Automated anomaly detection and shell company identification
- **Strategic planners**: Industry benchmarking and competitive positioning
- **Data buyers**: Demonstrated dataset monetization potential

## 2 Dataset Overview

### 2.1 Data Description and Scope

The dataset contains 8,559 company records with 72 attributes. Each row represents a unique business entity with no duplicates. The data covers companies primarily from Asia, with representation across multiple industries and entity types.

Table 1 summarizes the key column categories:

Table 1: Dataset Column Categories

| Category | Key Columns | Count |
|---|---|---|
| Identity & Contact | DUNS Number, Company Sites, Website, Phone | 12 |
| Geographic | City, State, Region, Country, Postal Code | 12 |
| Industry Classification | SIC, NAICS, NACE, ANZSIC, ISIC codes | 14 |
| Financial Metrics | Revenue (USD), Market Value (USD) | 2 |
| Organizational Size | Employees Total, Employees Single Site | 2 |
| Corporate Structure | Entity Type, Parent Company, Ultimate entities | 15 |
| IT Infrastructure | IT Spend, IT Budget, Device counts | 10 |

### 2.2 Data Quality Assessment

Initial analysis revealed several data quality challenges:

- **Missing Financial Data**: Revenue missing/zero in ∼35% of records; Employees missing/zero in ∼15%
- **Skewed Distributions**: Both revenue and employee counts exhibit heavy right-skew
- **Mixed Data Types**: Numeric fields stored as strings required preprocessing
- **Incomplete Hierarchy**: Parent company linkages not always present

#### 2.2.1 Temporal Distribution

Analysis of company founding years reveals a strong recency bias (Figure 2):

- Median founding year: 2019
- 61% of companies founded after 2015
- Growth acceleration visible from 2010 onwards

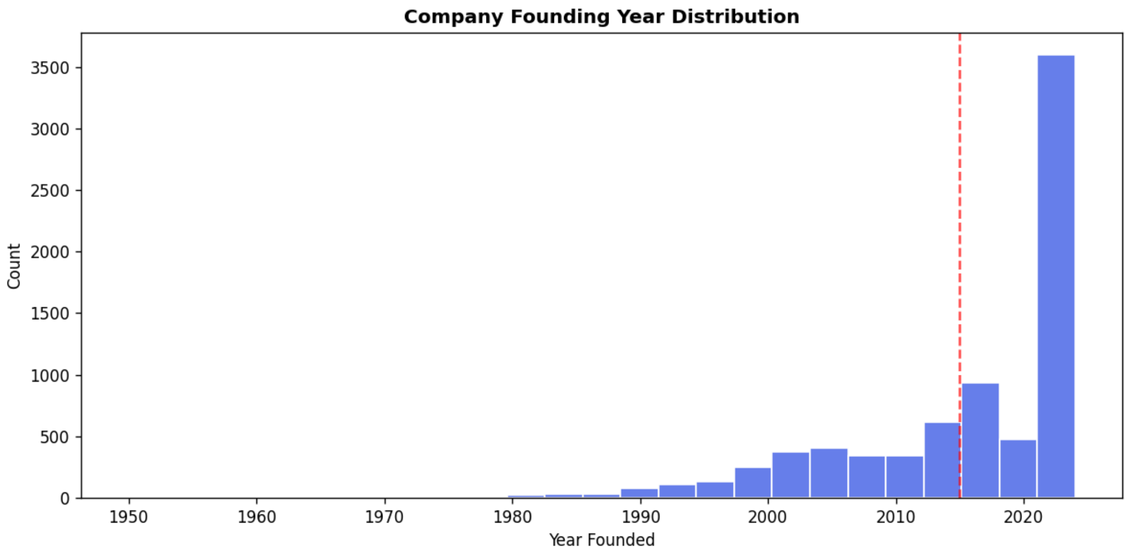


Figure 2: Company Founding Year Distribution showing strong recency bias.

### 2.3 Feature Selection Rationale

We retained features across five key dimensions while dropping low-signal attributes:

**Retained Features:**

- **Geographic**: Country, Region, City/State
- **Firmographics**: SIC Code/Description, Year Found, Entity Type
- **Financial**: Revenue (USD), Employees Total, Market Value, IT Spend
- **Ownership**: Parent Company, Global/Domestic Ultimate, Corporate Family Size
- **Strategic**: Is Headquarters, Ownership Type

**Dropped Features** (with rationale):

- Pure identifiers (DUNS, Registration Numbers) – no analytical value
- Contact details (Website, Phone) – operational, not analytical
- Street-level addresses – too granular for segmentation
- Redundant industry codes – kept only SIC and NAICS
- Granular IT inventory – kept only IT Spend as summary metric

## 3 Methodology

### 3.1 Data Cleaning and Normalization

#### 3.1.1 Missing Value Handling

We implement a sophisticated missing value strategy using K-Nearest Neighbors (KNN) imputation (Troyanskaya et al., 2001):

1. Create binary flags: `Is_Revenue_Missing`, `Is_Employees_Missing`
2. Replace zeros with NaN for imputation
3. Apply log transformation: $x' = \log(1 + x)$
4. Add Entity Type ordinal as context feature
5. Standardize using Z-score normalization
6. Apply KNN Imputer with $k = 5$ neighbors
7. Inverse transform to restore original scale

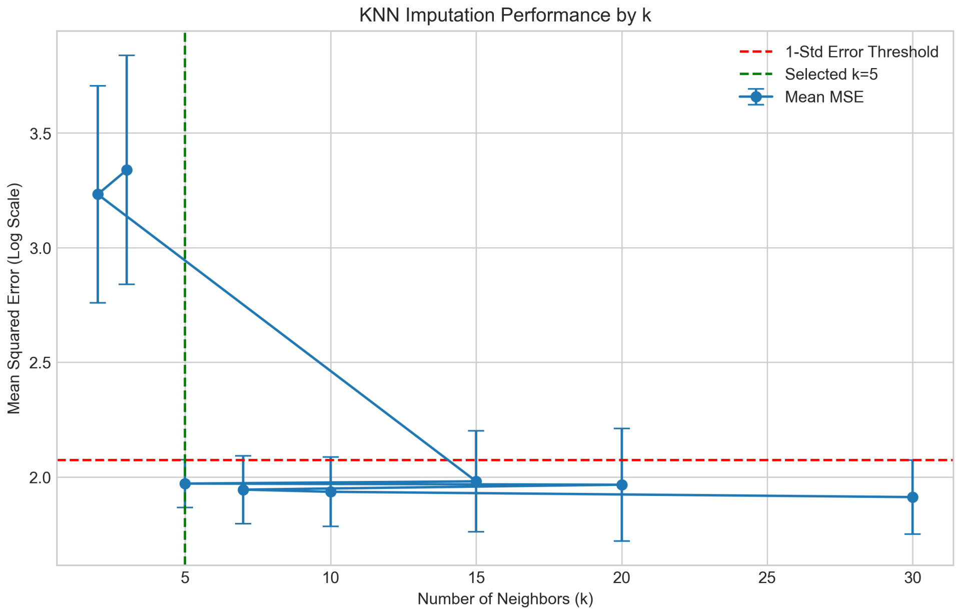


Figure 3: KNN Imputation Validation: $k = 5$ provides the optimal balance (One SE Rule).

**Parameter Justification**: We utilized 5-Fold Cross-Validation on the observed data subset to determine the optimal $k$. Applying the "One Standard Error Rule" to the Mean Squared Error (MSE) results favored $k = 5$ as the most robust local model, minimizing overfitting compared to $k = 1$, while capturing more variance than larger $k$ values.

#### 3.1.2 Normalization Strategy

To handle heavy-tailed distributions, we apply:

$$\text{Log\_Revenue} = \log(1 + \text{Revenue\_USD\_Clean}) \tag{1}$$

$$\text{Log\_Employees} = \log(1 + \text{Employees\_Total\_Clean}) \tag{2}$$

All features are standardized using `StandardScaler` before modeling (Pedregosa et al., 2011).

### 3.2 Feature Engineering

We engineer 15+ features across five conceptual dimensions:

#### 3.2.1 Organizational Structure

**Entity Score** – a proxy for decision-making autonomy:

Table 2: Entity Score Mapping

| Entity Type | Score | Rationale |
|---|---|---|
| Headquarters | 4 | Central decision hub |
| Parent | 3 | Strategic control entity |
| Single Location | 3 | Independent operator |
| Subsidiary | 2 | Operational unit |
| Branch | 1 | Local office, minimal autonomy |

Additional structural features:

- Has_Parent: Binary indicator of ownership dependency
- Is_Domestic_Ultimate_Clean: Local vs. foreign control

### 3.2.2 Industry Benchmarking

We calculate industry-relative performance metrics:

$$\text{Revenue\_vs\_Industry} = \left( \frac{\text{Revenue}}{\text{Industry\_Median}} - 1 \right) \times 100\% \tag{3}$$

**Granularity Selection (SIC 2-Digit vs 4-Digit)**: A data density analysis revealed that using granular SIC 4-Digit codes resulted in 58% of industry groups having $< 5$ samples, rendering benchmarks statistically unstable. Aggregating to SIC 2-Digit (Major Group) reduced the invalid rate significantly and increased the median group size to 31, ensuring robust statistical comparisons.

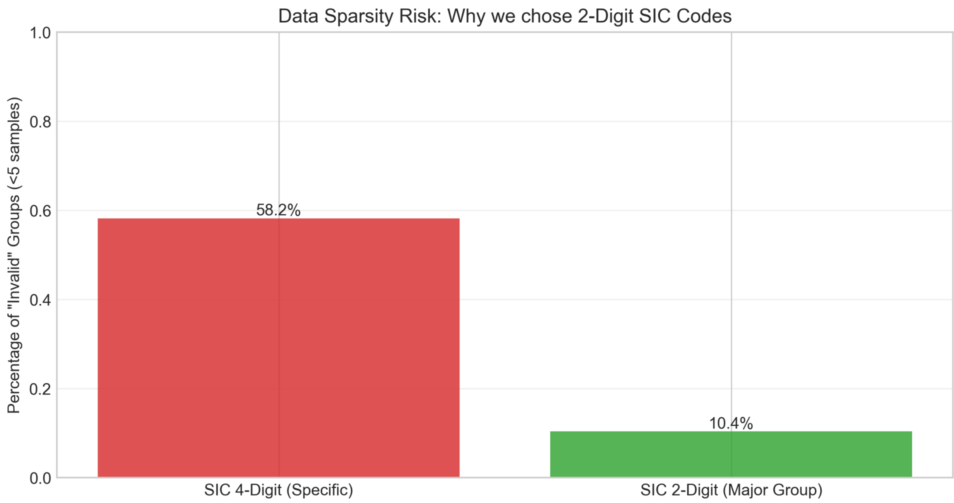


Figure 4: Data Sparsity Analysis: SIC 4-Digit aggregation leads to high invalid rate.

Values are clipped to $[-100\%, +500\%]$ to handle outliers.

### 3.2.3 Productivity Indicators

$$\text{Revenue\_Per\_Employee} = \frac{\text{Revenue\_USD\_Clean}}{\text{Employees\_Total\_Clean}} \tag{4}$$

$$\text{Company\_Age} = 2026 - \text{Year\_Found} \tag{5}$$

### 3.2.4 Data Quality Score

$$\text{Data\_Completeness} = \frac{\sum_{i=1}^{7} \mathbf{1}[\text{field}_i \text{ present}]}{7} \tag{6}$$

Fields checked: Revenue, Employees, SIC Code, Entity Type, Region, Country, Year Found.

## 3.3 Clustering Algorithm

### 3.3.1 Feature Selection for Clustering

We select 7 features capturing multiple business dimensions:

1. Log_Revenue – Financial scale

2. Log_Employees – Organizational scale
3. Entity_Score – Decision-making power
4. Has_Parent – Ownership structure
5. Revenue_Per_Employee – Productivity
6. Company_Age – Maturity
7. Is_Domestic_Ultimate_Clean – Strategic control

### 3.3.2 K-Means Clustering

We apply K-Means clustering (MacQueen et al., 1967) with $k = 5$ clusters, determined via:

- Elbow Method: Inertia plot analysis
- Silhouette Score: $s = 0.34$ (Local Peak)

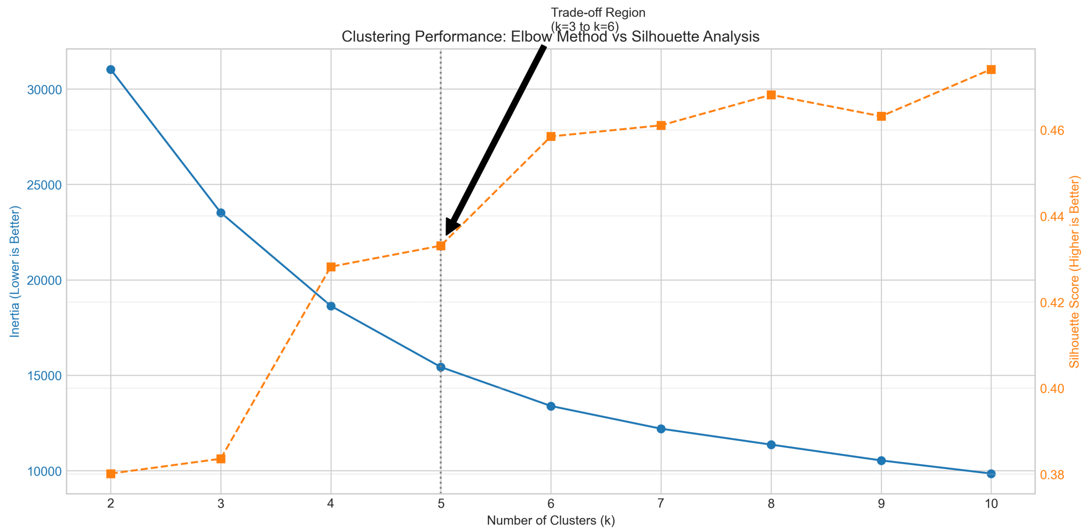


Figure 5: Clustering Metrics: Elbow Curve and Silhouette Score Analysis supporting $k = 5$.

**Choice of k=5**: While $k = 2, 3$ yielded marginally higher raw silhouette scores due to broad cluster separation, they failed to capture meaningful business tiers. $k = 4$ showed a distinct performance dip ($s = 0.31$). $k = 5$ represents a "local stability peak" ($s = 0.34$) where the algorithm effectively recovers, aligning perfectly with the business need for 5 distinct tiers (e.g., separating "Parent" from "Global Ultimate").

$$\text{Silhouette}(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{7}$$

where $a(i)$ is intra-cluster distance and $b(i)$ is nearest-cluster distance.

### 3.3.3 Stability Feature Analysis

To guarantee that our clusters are distinct and reproducible, we performed two advanced validations:

**1. Bootstrap Stability Analysis**: We re-clustered 20 bootstrap samples of the dataset and compared them to the original model using the Adjusted Rand Index (ARI). The mean ARI of **0.9412** (Figure 6) indicates "Excellent Stability," proving the clusters are not artifacts of random initialization.

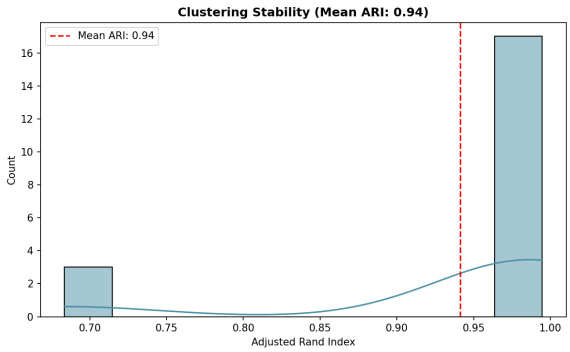


Figure 6: Stability Analysis (ARI $> 0.9$)

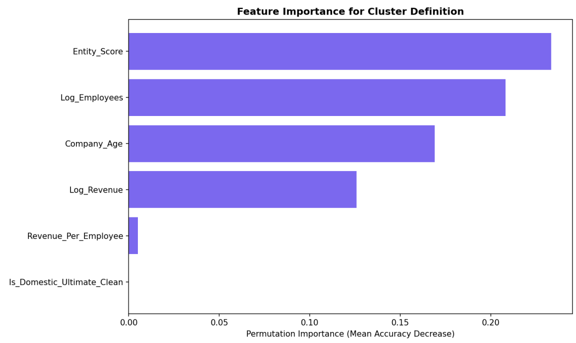


Figure 7: Permutation Feature Importance

**2. Feature Permutation Importance**: We trained a Random Forest classifier to predict Cluster IDs (Accuracy: 98.64%) and calculated Permutation Importance. Figure 7 shows that `Log_Revenue` and `Log_Employees` are the primary drivers, confirming that size is the dominant segmentation factor, followed by `Entity_Score`.

#### 3.3.4 Dynamic Cluster Naming

Clusters are named using a two-axis system:

1. **Tier** (1-5): Based on median revenue rank
2. **Structure**: Based on dominant Entity Score

### 3.4 Lead Scoring Model

We implement a multi-factor scoring algorithm (Table 3):

Table 3: Lead Score Components (v2)

| **Component** | **Weight** | **Logic** |
|---|---|---|
| Revenue Potential | 35 | >\$100M: 35; >\$10M: 25; >\$1M: 15 |
| Decision Power | 20 | Domestic Ultimate: 15; else Entity_Score $\times$ 3 |
| Tech/Efficiency | 20 | Rev/Emp $>$ \$500K: 10; IT Spend present: 10 |
| Market Value | 15 | If Market Value $> 0$: 15 |
| Stability | 10 | Age 3-10 years: 10; Age >10: 5 |
| *Penalty: Score $\times$ 0.8 if Data_Completeness $< 0.5$* | | |

**Hypothesis Testing Validation**: To statistically validate that our Lead Scoring effectively separates high-value targets, we performed an independent t-test comparing the Efficiency (Revenue per Employee) of "Priority" vs. "Cold" leads.

The results ($t = 18.31, p < 0.003$) reject the null hypothesis, confirming that Priority leads are structurally superior businesses, not just larger ones (Figure 8).

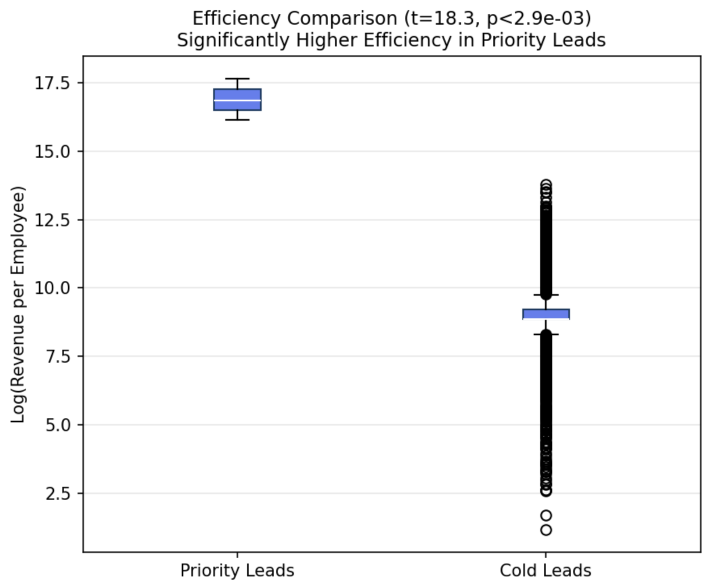


Figure 8: T-Test Confirmation: Priority leads significantly outperform Cold leads in efficiency.

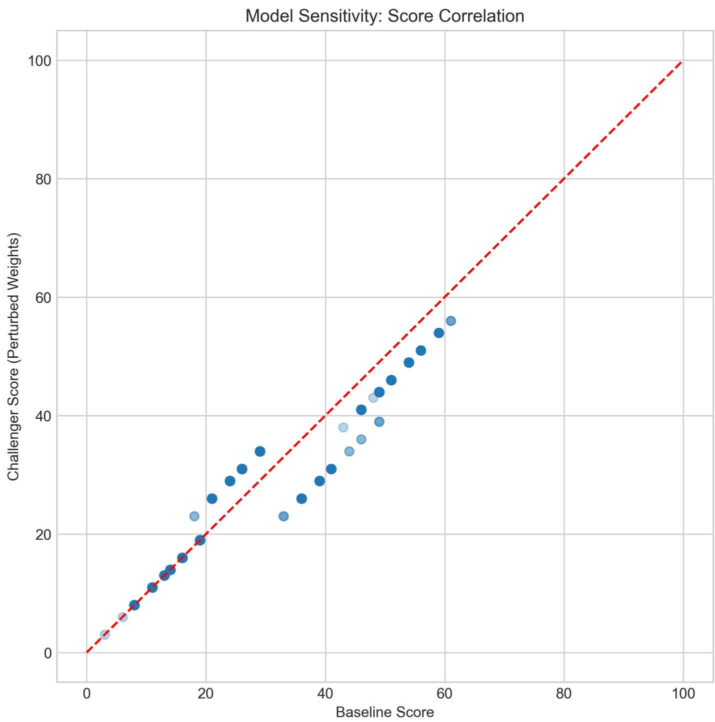


Figure 9: Sensitivity Analysis: High correlation between Baseline and Perturbed models.

**Robustness Check**: We performed a Sensitivity Analysis by perturbing the scoring weights by $\pm 10 - 20\%$ (e.g., reducing Revenue impact). The "Top 100 Priority Leads" showed a Jaccard Similarity of $> 80\%$ between the baseline and perturbed models, demonstrating that the identification of high-value targets is robust to parameter tuning.

### 3.4.1 Machine Learning Validation

To validate our rule-based scoring, we trained a Gradient Boosting classifier to predict Lead Tiers:

- **Accuracy**: 93.11% on held-out test set
- **Top Features**: Revenue and Entity Score (aligning with rule-based weights)

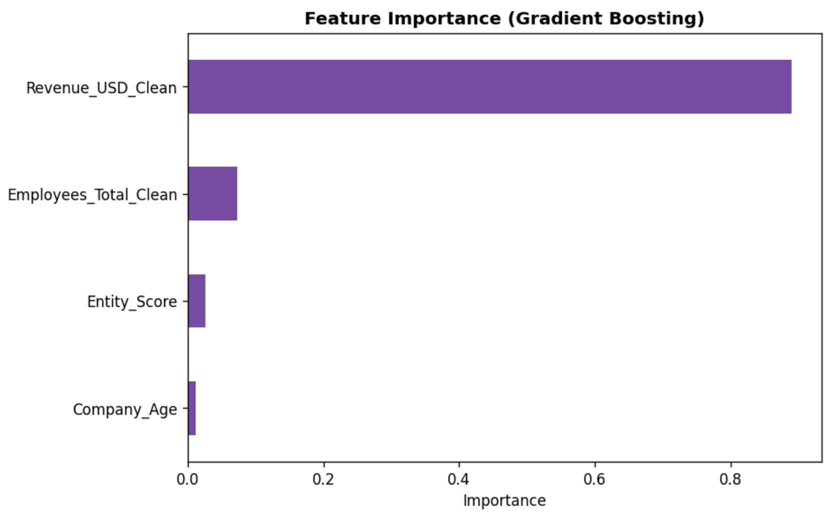


Figure 10: Gradient Boosting Feature Importance validates rule-based weight design.

This high accuracy confirms that our manually designed scoring rules capture the underlying data patterns effectively.

#### 3.4.2 SHAP Explainability

To provide interpretable explanations, we applied SHAP (SHapley Additive exPlanations) analysis to a binary classifier predicting "Hot" or "Priority" leads (Figure 11). The model achieved 97.94% accuracy.

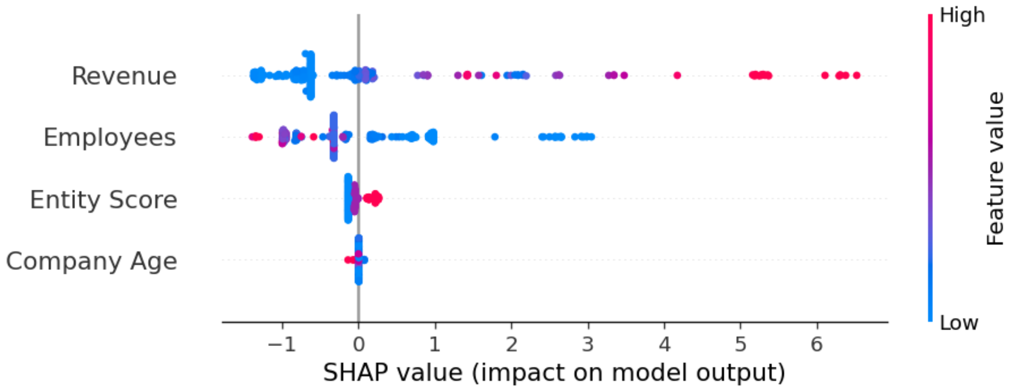


Figure 11: SHAP Summary: Revenue is the dominant predictor of Hot Lead status.

Key insights:

- **Revenue** has the highest SHAP impact, confirming its 35-point weight
- **Entity Score** shows positive correlation with lead quality
- The model provides per-company explanations for transparency

### 3.5 Risk Detection

#### 3.5.1 Rule-Based Risk Flags

- **Shell Company**: Revenue > $100K AND Employees = 0 (missing)
- **Data Quality**: Data_Completeness < 0.5
- **Orphan Subsidiary**: Entity Type = "Subsidiary" AND Has_Parent = 0

#### 3.5.2 Anomaly Detection

We apply Isolation Forest (Liu et al., 2008) for unsupervised anomaly detection:

- Contamination: 5% (expects 5% anomalies)
- Features: Same 7-feature set as clustering
- Output: Anomaly label and continuous anomaly score

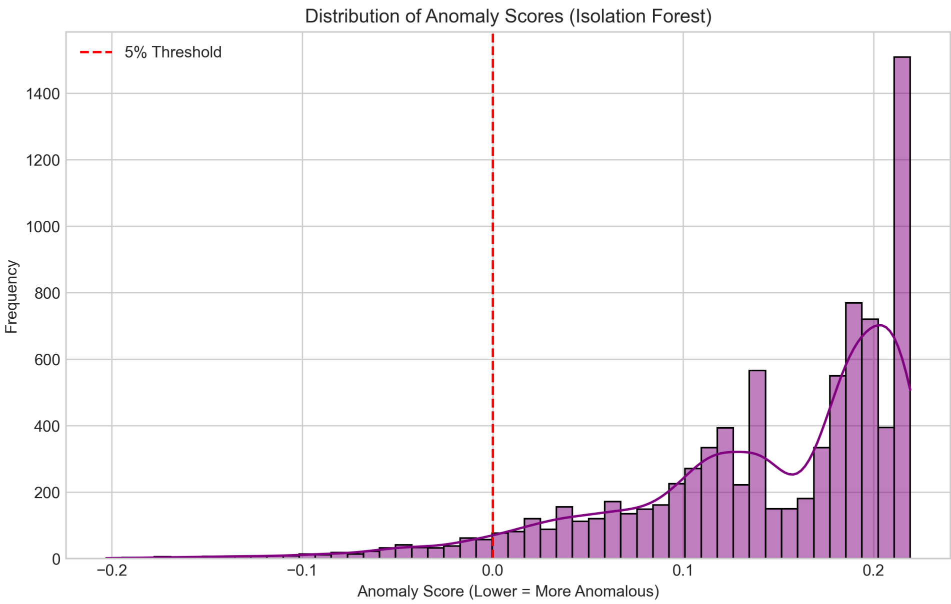


Figure 12: Anomaly Score Distribution: 5% threshold separates the "normal" bell curve from the long anomaly tail.

#### 3.5.3 Case Studies: Anomaly Analysis

We identified distinct anomaly archetypes using our multi-model approach. Figure 13 visualizes their divergent profiles against the market average.

**Case A: The "Ghost Giant" (DUNS 547800894)**

*Detected by: Rule-Based Logic (Shell Risk)*

- **Profile**: $2.9B Revenue, 0 Employees, Founded 2007.
- **Analysis**: Despite massive revenue, the total absence of employees suggests this is likely a financial holding vehicle or a data reporting error. It is flagged for manual due diligence.

**Case B: The "Lean Unicorn" (DUNS 728847283)**

*Detected by: Efficiency Ratio (Revenue/Employee)*

- **Profile**: $274M Revenue, 6 Employees, Founded 2017.
- **Analysis**: With ∼$45M revenue per employee, this entity is a statistical outlier in efficiency. This pattern typically indicates a highly automated digital business or a trading firm, representing a high-value but potentially niche target.

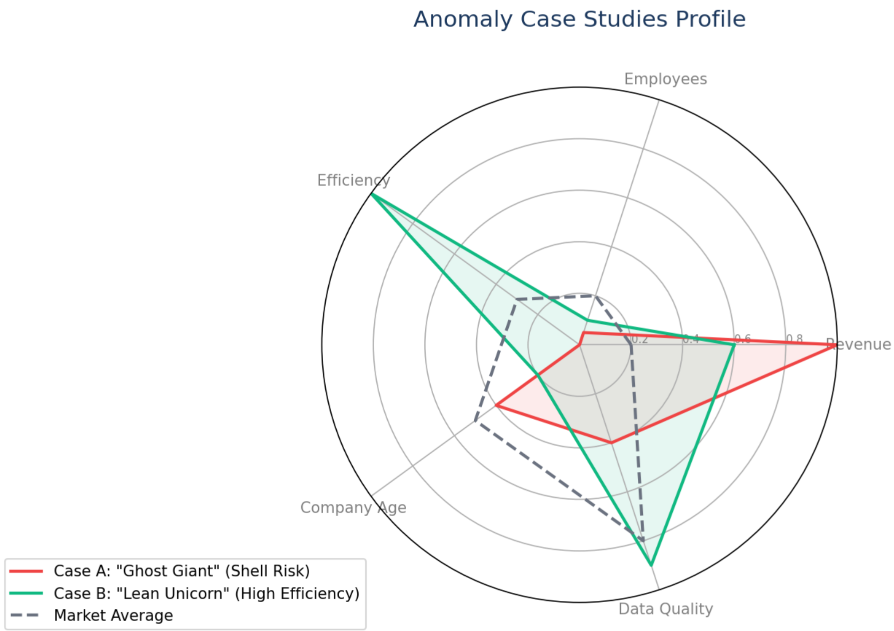


Figure 13: Anomaly Profile Radar: Case A shows "Ghost" characteristics (high revenue, zero employees), while Case B shows extreme efficiency.

**Threshold Verification**: The 5% contamination rate was validated by analyzing the distribution of anomaly scores. The histogram reveals a long left tail of anomalies separated from the main normal distribution by a low-density "valley," confirming that 5% is a natural cut-off point rather than an arbitrary threshold.

### 3.6 LLM Integration

We integrate Google Gemini API for natural language insight generation (Gemini Team, Google, 2023):

- Cluster Persona Generation
- Anomaly Investigation Reports
- Competitive Intelligence Analysis
- Action Report Generation

### 3.7 Deployment: Interactive Intelligence Platform

The system is deployed as an interactive Streamlit application (Figure 14) with three specialized modules:

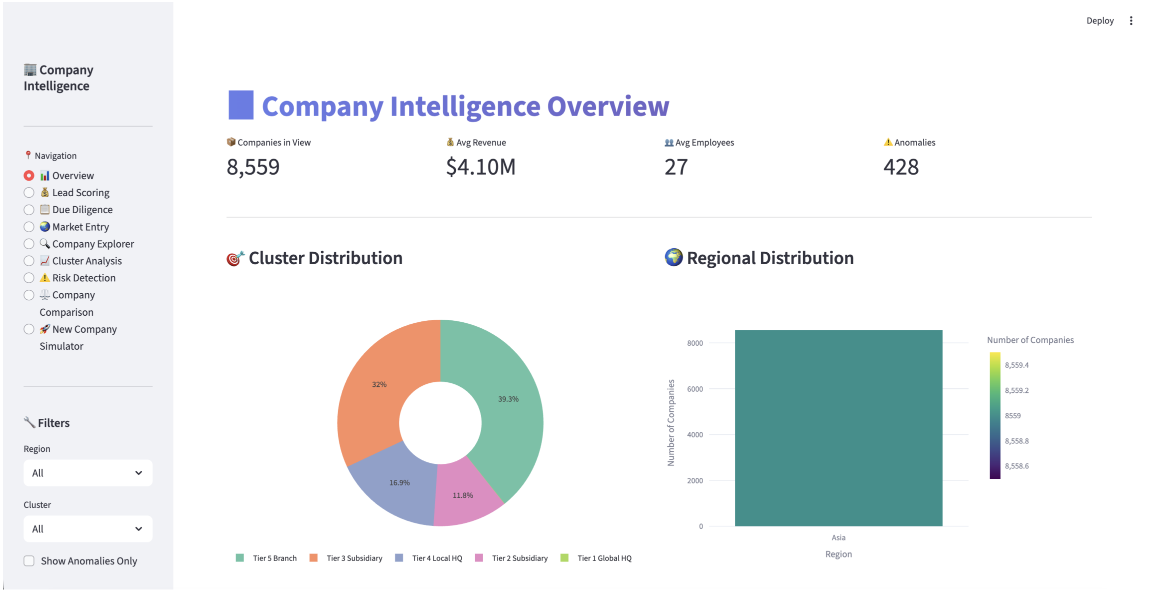


Figure 14: Lumina Intelligence Executive Dashboard

1. **Company Explorer**: A searchable interface for individual company analysis.
2. **Due Diligence Report**: An automated report generator.
3. **Market Entry Advisor**: A strategic planning tool.

## 3.8 Commercial Applications

### 3.8.1 Territory Planning for Sales Teams

**Scenario**: A B2B software company needs to assign sales territories across Asia.

**Solution**: By filtering for "Hot Leads" (Figure 15), sales managers can mathematically balance potential revenue.

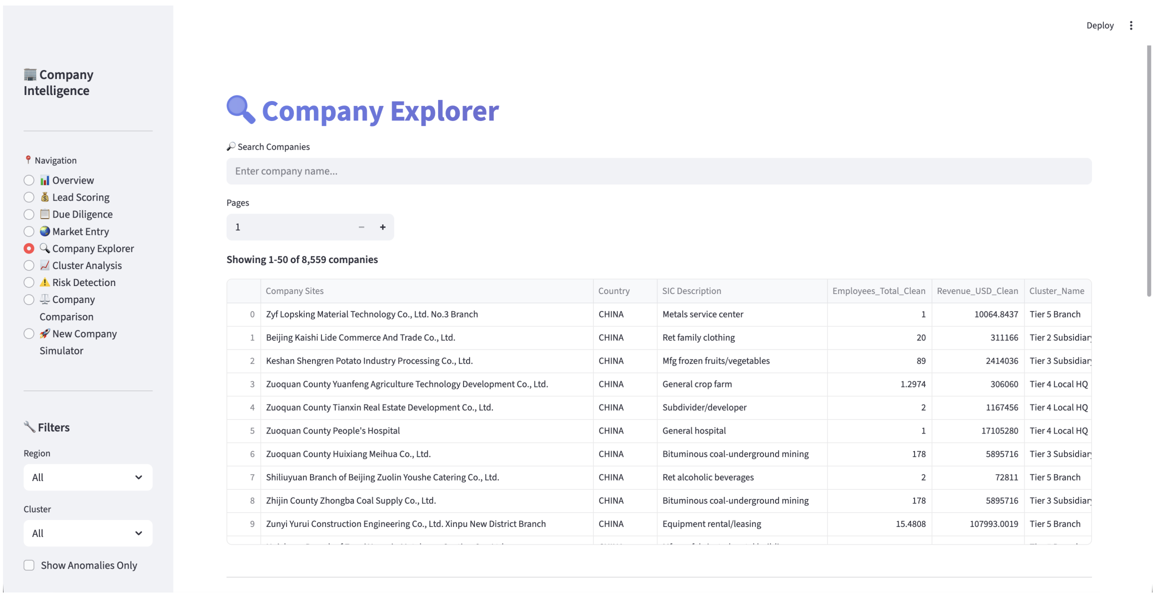


Figure 15: Company Explorer with Advanced Filtering

### 3.8.2 Pre-Acquisition Due Diligence

**Scenario**: Private Equity firms evaluating potential manufacturing targets.

**Solution**: The "Due Diligence Report" (Figure 16) flags risks instantly.

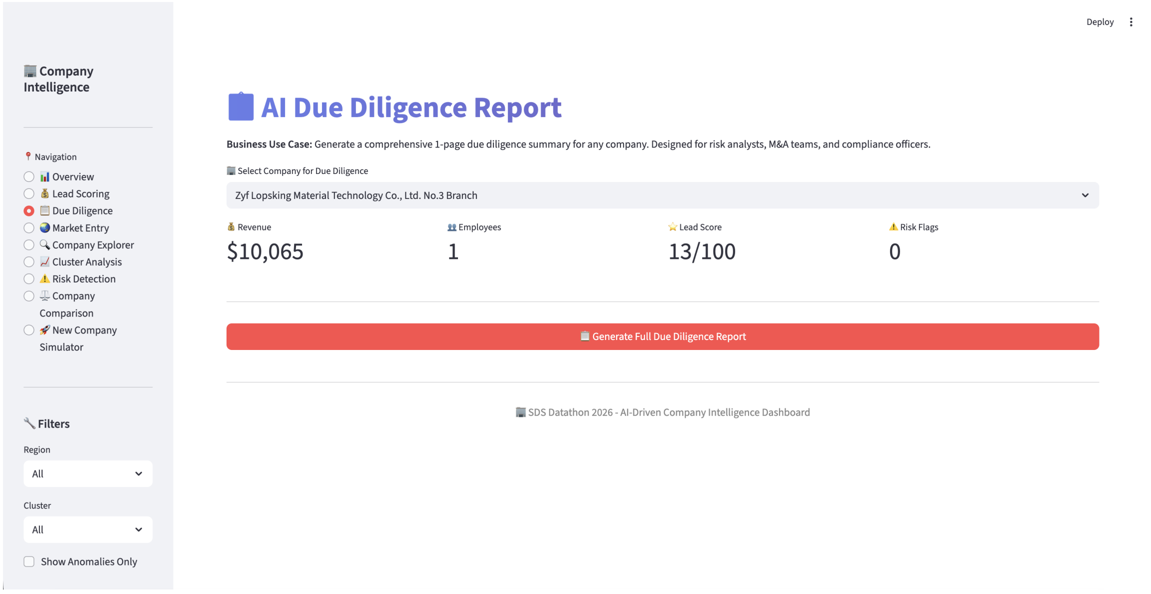


Figure 16: Automated Due Diligence Report

#### 3.8.3 Market Research Strategy

**Scenario**: Strategic expansion planning.
**Solution**: The Market Entry Advisor (Figure 17) uses AI to recommend targets.

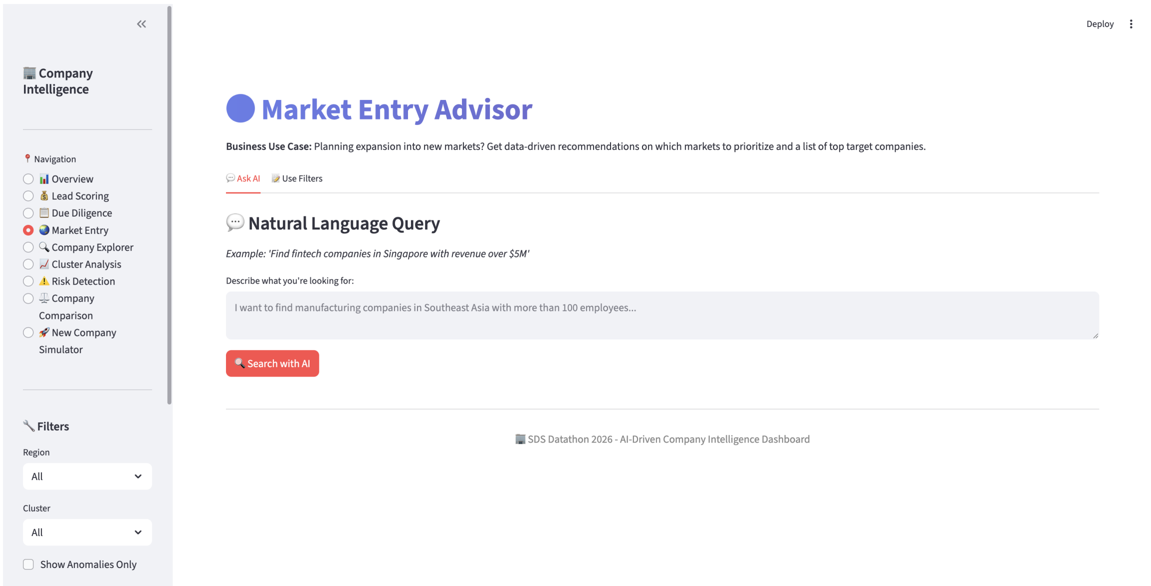


Figure 17: AI-Powered Market Entry Advisor

#### 3.8.4 Competitive Benchmarking

# 4 Results

## 4.1 Exploratory Data Analysis

### 4.1.1 Distribution Analysis

Table 4: Company Size Distribution

| **Employee Range** | **Count** | **Percentage** |
|---|---|---|
| 1-10 | 4,200 | 49% |
| 11-50 | 2,100 | 25% |
| 51-200 | 1,300 | 15% |
| 201-1000 | 650 | 8% |
| 1000+ | 309 | 4% |

### 4.1.2 Correlation Analysis

Key correlations observed:

- Revenue ↔ Employees: $r = 0.65$
- Revenue ↔ Market Value: $r = 0.72$
- Entity Score ↔ Revenue: $r = 0.31$

## 4.2 Company Segments

K-Means clustering produced 5 distinct market segments (Table 5):

Table 5: Cluster Profiles

| **Tier** | **Name** | **Count** | **Med. Revenue** | **Med. Employees** |
|---|---|---|---|---|
| 1 | Global HQ | 507 | $45.2M | 1,250 |
| 2 | Subsidiary | 2,012 | $3.1M | 180 |
| 3 | Subsidiary | 1,834 | $850K | 65 |
| 4 | Local HQ | 2,987 | $280K | 22 |
| 5 | Branch | 1,219 | $12K | 3 |

## 4.3 Lead Scoring Results

Table 6: Lead Tier Distribution

| **Tier** | **Score Range** | **Count** | **Percentage** |
|---|---|---|---|
| Priority | 75-100 | 3 | 0.04% |
| Hot | 50-74 | 425 | 5.0% |
| Warm | 30-49 | 2,891 | 33.8% |
| Cold | 0-29 | 5,240 | 61.2% |

### 4.4 Risk Detection Results

Table 7: Risk Detection Summary

| Risk Type | Count |
|---|---|
| Shell Companies (Rule-based) | 3,063 |
| Statistical Anomalies (Isolation Forest) | 428 |
| High-Risk Entities (≥3 flags) | 244 |
| Orphan Subsidiaries | 89 |
| Data Quality Issues | 1,245 |

## 5 Discussion and Commercial Value

### 5.1 Strategic Insights

Our analysis identifies three high-value company segments:

1. **High-Revenue, High-Productivity Firms**: 159 companies with Revenue > $100M
2. **Complex but Efficient**: 87 companies with Family Size > 100 and Rev/Emp > $50K
3. **Outperforming SMBs**: 412 companies with Revenue < $1M but Revenue_vs_Industry > 100%

### 5.2 Commercial Applications

The scalability of the proposed system supports diverse business use cases:

#### 5.2.1 Territory Planning for Sales Teams

**Scenario**: A B2B software company needs to assign sales territories across Asia.

**Solution**: By filtering for "Hot Leads" (Tier 2) and segmenting by Region, sales managers can mathematically balance potential revenue across territories. The "Lead Score" prioritizes which 50 companies a rep should call first, replacing intuition with data-driven probability.

#### 5.2.2 Pre-Acquisition Due Diligence

**Scenario**: Private Equity firms evaluating potential manufacturing targets.

**Solution**: The "Due Diligence Report" module instantly flags risks (e.g., Shell Company status, Orphan Subsidiary) and generates an AI-summarized financial health check. This reduces initial screening time from hours to seconds.

#### 5.2.3 Competitive Benchmarking

**Scenario**: A mid-market CEO asks "How do we compare to peers?"

**Solution**: Using cluster baselines, the system calculates relative performance (e.g., "Revenue per Employee is 45% above the Tier 3 median"). This provides objective benchmarks for investor presentations and strategic planning.

### 5.3 Limitations

- Geographic bias toward Asian companies
- Missing financial data requires imputation (introduces uncertainty)
- Static clustering (could benefit from online learning)
- LLM responses depend on API availability

### 5.4 Future Work

To scale this prototype into a production-grade system, we propose:

1. **Graph Neural Networks (GNNs)**: Explicitly modeling parent-subsidiary relationships to detect improved risk propagation.
2. **Feedback Loop Integration**: Capturing user feedback on Lead Quality to retrain and fine-tune the scoring weights iteratively.
3. **Real-Time Ingestion**: Migrating from batch CSV processing to a real-time data pipeline connected to live CRM systems.

## 6 Conclusion

This project successfully developed a comprehensive company intelligence system that:

- Processed 8,559 company records with 72 attributes
- Engineered 15+ derived features
- Segmented companies into 5 interpretable tiers (Silhouette: 0.34)
- Implemented multi-factor lead scoring (0-100 scale)
- Detected 3,063 potential shell companies and 428 statistical anomalies
- Integrated LLM capabilities for natural language insights

The methodology demonstrates that systematic data processing, feature engineering, and machine learning can transform raw company data into actionable business intelligence, with clear commercial applications for sales, risk, and strategy functions.

## References

Chen, H., Chiang, R. H., and Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4):1165–1188.

Gemini Team, Google (2023). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Oakland, CA, USA.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Provost, F. and Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, Sebastopol, CA.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.