ie UNIVERSITY

PRIVATE EQUITY MATCHMAKER

A Predictive Tool for Screening Buyers of Spanish Companies

Madrid, July 2025

Capstone Project

Master in Business Analytics & Data Science

Group 2

Moayad Barayan
Ana Cortés Barquier
Camilla Perotti
Samir Barakat
Shadi Alfaraj
Vibhushan Balaji Neethi Mohan

Mentored By

Gustavo Martin Vela

Ethics Statement

This project was conducted in full compliance with ethical standards for academic and professional research. Proprietary client data provided by EY-Parthenon was used strictly for internal analysis and was not shared outside the project team. We adhered to responsible AI practices, ensuring transparency in model interpretation, safeguarding data privacy, and maintaining the confidentiality of all sensitive information throughout the development of the PE Matchmaker tool.

Table of Contents

1. Executive Summary

In a competitive Private Equity (PE) landscape, where deal origination speed and strategic alignment are more critical than ever, identifying the right buyer for a company remains an inherently complex and resource-intensive process. This report introduces a data-driven solution designed to streamline how analysts at EY Parthenon match acquisition targets with the most relevant PE buyers, combining machine learning intelligence with intuitive, natural language interaction.

Developed exclusively for the Spanish market, the solution leverages EY's proprietary data on PE firms, portfolio companies, and historical transactions to recommend the ten most likely acquirers for a given Spanish company. These recommendations are based on quantifiable behavioural patterns, including sector preferences, deal size history, and geographic focus, ensuring relevance and precision.

The tool is delivered through a user-friendly interface powered by a Large Language Model (LLM), allowing analysts to query potential acquirers simply by prompting the company of interest. By automating what is traditionally a manual screening process, EY gains a faster, smarter, and more consistent method of identifying aligned PE buyers, enhancing analyst productivity and supporting more effective outreach strategies.

This initiative represents a practical and scalable application of AI within M&A advisory, reinforcing EY's commitment to innovation and to delivering high-impact, data-informed solutions to its clients.

2. Industry Context and Strategic Challenge

Private Equity continues to evolve in response to shifting macroeconomic conditions and intensifying market dynamics. With interest rates elevated and access to credit more constrained, firms are under pressure to be more selective and efficient in how they deploy capital. At the same time, record levels of dry powder have created fierce competition for high-quality assets, heightening the importance of origination as a critical lever for value creation.

Against this backdrop, deal sourcing has become increasingly complex. Traditional approaches relying on personal networks, intermediaries, and manual screening are proving insufficient in a market where speed, precision, and sector expertise are paramount. PE firms are now required to assess a broader and more fragmented landscape of opportunities, often across multiple geographies and sectors, while also considering a growing number of strategic fit criteria.

This complexity has a direct impact on execution timelines and opportunity cost. The longer it takes to identify the right acquirer or investment partner, the greater the risk of losing the deal to a faster-moving competitor. For firms on the sell side, the challenge is equally pressing: finding the most strategically aligned buyer who can move quickly and close with conviction.

Our project is designed to address this exact challenge. By analyzing patterns in historical transaction data and firm behaviour, we aim to improve the accuracy and speed with which potential buyers are identified. The result is a more focused and data-driven origination process that reduces time to insight and strengthens the alignment between targets and potential investors.

3. Business Problem and Opportunity

Identifying the right PE buyer for a given company remains one of the most time-intensive and uncertain aspects of the deal process. Despite access to industry databases and internal CRM systems, much of the matching process is still driven by manual research, anecdotal knowledge, and network-based outreach. This fragmented approach often leads to inefficiencies, blind spots, and missed strategic fits.

For investment banks, M&A advisors, and even the PE firms themselves, the challenge is twofold: first, determining which firms are most likely to be interested in a specific asset, and second, prioritizing those with the highest probability to act. Without a structured method to connect target company profiles with PE firms' investment behaviour, teams are forced to cast wide nets contacting dozens, sometimes hundreds, of firms in the hope of securing interest. This results in prolonged timelines, significant resource allocation, and an overall dilution in the quality of engagement.

Beyond inefficiency, the current process often fails to capture the full potential of the market. Promising matches are overlooked due to lack of visibility into evolving investment themes, nuanced firm strategies, or subtle signals in transaction histories. The consequence is a deal flow that is not only slower but also less aligned with where real interest and strategic fit exist. This creates a clear opportunity for innovation. By leveraging data to uncover hidden patterns in PE firm behaviour such as sector preferences, deal size, geographic focus, and co-investment activity matching can become faster, more accurate, and significantly more effective. A machine learning driven approach transforms the process from reactive to proactive, enabling deal teams to identify top-fit acquirers earlier and with greater confidence.

For firms operating in a competitive M&A environment, this is more than an operational improvement it represents a strategic edge. Faster outreach to the right buyers shortens execution cycles, enhances credibility with clients, and maximizes the potential for successful outcomes. In this context, data-driven matching is not just a technological upgrade, but a business-critical advantage.

4. Our Solution: The PE Matchmaker Tool

To address the challenge introduced in earlier sections, namely, identifying the most suitable PE firms for specific market opportunities, a PE Matchmaker tool was developed. This platform integrates machine learning (ML) and large language models (LLMs) to provide EY consultants with a seamless and intuitive user experience.

Application Capabilities and User Journey

The core user experience of the PE Matchmaker platform unfolds across the following steps:

a) **Input Phase**

The EY user inputs the target company.

b) **ML-Driven Scoring Engine**

In the backend, a supervised machine learning model is triggered. Trained on historical transaction data and investment behaviour, the model predicts the likelihood that a given PE firm would express interest in the target opportunity. The algorithm considers structured features such as:

- Sector alignment
- EBITDA range
- Geographic focus (total number of Spanish companies)

c) **Result Interpretation and RAG-Enhanced Explanation**

The model returns a ranked list of suitable PE firms along with a relevance score. To ensure transparency and build user trust, a connected LLM interface generates tailored justifications for each recommendation. The language model is enhanced via Retrieval-Augmented Generation (RAG), allowing it to dynamically incorporate internal EY knowledge – such as specific sectors, number of divested Spanish companies - during runtime. This ensures explanations reflect both public data and confidential institutional insight.

d) **Interactive Q&A and Strategic Advisory Support**

Upon receiving the list of recommended PE firms, users can engage with a conversational AI assistant embedded within the tool. This chatbot is powered by the same LLM and allows for natural language queries (e.g., “What sectors is this PE focusing on?”). It acts as a strategic sparring partner, assisting EY professionals in deepening their understanding of the results and enriching client conversations in M&A contexts.

Technology Stack

To ensure a scalable, robust, and business-ready solution, the following technologies were selected across the machine learning, language model, and interface layers of the PE Matchmaker platform:

- **Machine Learning**
  Python was used as the core programming language for developing the supervised CatBoost machine learning model. The model was trained on historical transaction data to capture patterns in PE behaviour, incorporating libraries such as Pandas, Scikit-learn for feature engineering, classification, and evaluation. Further information can be found in the Data Science Pipeline section.

  Large Language Model (LLM)
  The application leverages LangChain to orchestrate LLM-based workflows and integrates the Gemini 2.5 Pro model as the reasoning engine. Gemini 2.5 Pro was selected because its advanced reasoning capabilities are essential for generating accurate and insightful justifications for the ML model's outputs. Its superior ability to synthesize information from large RAG contexts ensures that these justifications are robustly grounded in the provided data. This combination of deep understanding and coherent explanation is critical for delivering trustworthy and actionable insights in high-stakes M&A scenarios.

- **Web Application and Deployment**
  The user interface was built and deployed using Streamlit. Streamlit was selected for its agility in developing interactive data applications and its capability to operate as a lightweight standalone web service, independent of EY's core IT and data infrastructure. This ensures ease of deployment while maintaining flexibility for future integration into enterprise environments.

5. Data Science Pipeline

Data Sources

We used three proprietary EY datasets:

- **PE Firms:** Profile data including sector focus, geography, and ticket size.
- **Portfolio Companies:** Historical investments with company metadata.
- **Transactions:** Deal-level records including value, timing, and parties.

These sources collectively enabled profiling PE behaviour and linking past deals to firm characteristics.

Data Cleaning

A robust cleaning process was applied to unify and prepare the datasets for modelling. This included:

- **Deduplication:** to eliminate overlapping records across data sources
- **Standardization:** of sector, geography, and revenue fields to ensure consistent categorization
- **Missing value handling:** using targeted imputation methods for categorical and numerical fields
- **Normalization:** of financial figures (e.g., revenue bands) to align deal comparability
- **Relational integrity checks:** ensuring consistent IDs across PE firms, portfolio companies, and transactions

This step was critical to reduce noise, improve model inputs, and preserve analytical integrity. Further technical details are documented in the Jupyter Notebook.

Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase focused on understanding and characterizing the two key datasets used in the project: Private Equity Dataset *(Cleaned_PE)* and Spanish Company Dataset *(OutputCheckCombined)*. The objective was to identify relevant patterns and insights to inform the development of the recommendation model.

Private Equity Dataset

This dataset contains information on PE firms, including their sector focus, investment size preferences (ticket_EBITDA), and their Spanish portfolio.

- **Ticket_EBITDA**
  The distribution of ticket_EBITDA revealed a concentration of PE firms targeting companies with EBITDA values in the mid-market range (€10M–€30M), reflecting the typical size of transactions in the Spanish market. However, the field is inconsistently populated, containing a mix of numeric, categorical, and free-text values, which suggests a need for standardization before use in recommendation logic.
- **Total_Spanish_Company**
  Analysis showed a wide variance in the number of Spanish companies in PE portfolios. While a few PEs had extensive Spanish holdings (10+ companies), a significant portion had 1-3, suggesting these firms are either newer entrants to the Spanish market or have sector/geographic preferences elsewhere or a strategic focus on limited but high-impact deals.
- **Sectors**
  Sectoral analysis indicated that most PEs maintain diversified portfolios across multiple industries, with notable concentrations in Technology, Healthcare, and Consumer Goods. This diversity underscores the importance of sector matching when recommending suitable PEs for specific opportunities.
- **Geographic focus**
  The Top_Geographies variable confirms that Spain is the top target country, followed by the USA, UK, and France. Notably, 90 out of 169 firms have a physical office in Spain, and 34% of them are headquartered in Spain. This underlines the relevance of local presence in active portfolio management and deal sourcing.

Spanish Companies Dataset

This dataset includes financial and sectoral information on Spanish industrial companies.

- **Sectors**
  Companies were well distributed across sectors, with dominant representation in Manufacturing, Retail, and Technology. Subsector analysis provided more granularity, enabling precise alignment with PE sector focus areas. This distribution complements the sector diversity observed in the *Cleaned_PE* dataset, enhancing the effectiveness of sector-matching algorithms.
- **EBITDA**
  Most firms exhibit EBITDA levels that align with stable operational performance, making them attractive candidates for private equity investment and value creation initiatives. However, a subset of companies displays highly disproportionate or negative EBITDA figures, which may reflect financial distress, atypical business models, or residual data inconsistencies despite prior cleaning. These anomalies highlight the need for robust outlier handling and financial health filters to ensure that recommendation models prioritize sustainable and scalable targets.
- **Ownership structure**
  The ownership structure of the companies analyzed is highly concentrated in a few categories. Private Owned firms dominate the dataset, followed by Subsidiaries and Family-Owned businesses, together accounting for most observations. Private Equity ownership, while less frequent, represents a notable segment, reflecting active institutional participation in the market. A long tail of other ownership types such as Cooperatives, Foundations, and Religious Organizations appears with significantly lower frequencies.
- **Regional Concentration**
  The regional distribution of companies is heavily concentrated in Cataluña, Madrid, and Andalucía, which together account for the largest share of firms in the dataset. These three regions represent key economic hubs in Spain, offering significant business density and investment opportunities. Secondary clusters are observed in Comunidad Valenciana, Galicia, and Murcia, while the remaining regions exhibit a much lower concentration of companies.

Machine Learning

To support EY in identifying the most suitable PE firms for specific market opportunities, we developed a supervised Machine Learning (ML) model. The model is designed to analyze key characteristics of both PEs and companies, providing ranked recommendations with clear reasoning to enhance business development efforts.

The core objective of the ML model is to automate and scale the PE recommendation process. It enables consultants to quickly pinpoint PEs whose investment theses best align with a given opportunity, reducing manual effort and accelerating go-to-market strategies.

The model was trained on historical data from PE investments and Spanish industrial companies, leveraging features that reflect real-world investment criteria:

- **EBITDA Range Fit**

  The company's EBITDA was evaluated against each PE's typical investment size (ticket_EBITDA). A numeric alignment score was derived, favouring PEs whose target deal sizes matched the opportunity's financial profile. This feature consistently ranked among the top three predictors across all tested models.
- **Geographic Presence**

  The presence of a PE's portfolio companies in Spain was incorporated as a feature to capture their familiarity with and commitment to the local market. This allowed the model to prioritize PEs with demonstrated interest in Spanish investments.
- **Sector Matching**

  Using Natural Language Processing (NLP), we transformed the sector descriptions from both PEs and companies into dense vector representations via pre-trained SentenceTransformer embeddings (MiniLM). Cosine similarity was then used to calculate a sector similarity score, capturing nuanced relationships between sector focus areas.

The model was developed using ensemble methods, including CatBoost, to achieve high predictive performance while handling categorical and numerical data efficiently.

Key Outcomes

- **Prioritized Recommendations**: For any given company, the model produces a ranked list of PEs most likely to consider the opportunity, complete with alignment scores for transparency.
- **Business Insights**: The model highlights areas of alignment and mismatch (size, geography, sector), enabling EY consultants to craft tailored pitches for each PE.
- **Scalability and Adaptability**: Designed to scale across sectors and geographies, the approach supports EY's global operations and future expansions.

Supervised Models

The supervised modelling framework utilized four classification algorithms trained on a balanced dataset derived from simulated positive matches and strategic sampling:

- CatBoost Classifier (final selected model)
- XGBoost
- Random Forest

Each model was trained on the same feature set: ebitda_in_range, Total Spanish Companies, and sector_similarity_score.

Model Performance Summary

Model	Accuracy	F1 Score	ROC AUC
CatBoost	**0.88**	**0.89**	**0.85**
Random Forest	0.87	0.87	0.87
XGBoost	0.81	0.82	0.87

CatBoost was selected as the final model based on:

- Superior predictive performance across all classification metrics.
- Robust handling of categorical features and imbalanced datasets.
- Interpretability, providing clear variable importance scores without requiring feature scaling or one-hot encoding.

Supplementary plots (confusion matrix, ROC curve, Precision–Recall curve) confirmed the model's ability to reliably discriminate between likely and unlikely matches across various thresholds.

Unsupervised Models

In parallel, several unsupervised models were implemented to explore latent match potential without labelled targets:

Unsupervised Model	Precision@5
HDBSCAN (sector + geo features)	0.0012
Autoencoder + HDBSCAN	0.0000
Refined Autoencoder + HDBSCAN	0.0000
Cosine Similarity Ranker	0.0000

Despite achieving a high Silhouette Score (0.978) in HDBSCAN, the resulting clusters exhibited poor overlap with known match labels. Precision@5 scores were near zero, suggesting that high-density areas in latent space did not correlate with real acquisition decisions.

The cosine similarity ranker, based on raw sector and geography embeddings, failed to produce useful differentiation. This is likely due to semantic redundancy across firms and the narrow expressive range of company descriptions.

Consequently, unsupervised approaches were deemed unsuitable for capturing complex buyer behaviour in this context, reinforcing the need for label-driven learning and carefully engineered features.

6. Business Impact and Value for EY

The **PE Matchmaker** tool introduces a strategic leap in EY's Private Equity origination capabilities by embedding machine learning into the deal sourcing process. Traditionally, identifying the right buyer involves a high-effort, low-yield screening process that relies heavily on manual analyst work and extensive partner networks. Our solution offers a faster, smarter, and more targeted approach that not only saves time but also enhances the probability of deal success.

Time Saved in Target-Purchaser Screening

In the current approach, consultants typically engage with 15–20 PE firms to generate 5–6 discovery calls, resulting in 3 formal proposals and 1–2 viable outcomes. Each PE firm outreach involves reviewing the firm's fit, preparing tailored communication, coordinating availability, and logging updates—activities that, on average, amount to 1 hour of effort per contact.

With PE Matchmaker, the initial longlist is algorithmically reduced from ~50 firms to the top 5 most probable acquirers in seconds, based on historical transaction patterns and firm-level investment theses. This narrows the active outreach pool to just 10 high-relevance targets, saving approximately 10 hours per project.

➔ **Time Savings per Deal Cycle:**
10 fewer outreach efforts × 1 hour = **10 hours saved**
➔ **Cost Savings per Deal Cycle:**
10 hours × €120/hour (average consultant cost) = **€1,200 saved**

Scalable Efficiency and Cost Reduction

When applied across a broader client base, the tool's impact scales significantly:

➔ **Annualized Value (100 projects/year)**
100 projects × €1,200 = **€120,000 in consultant cost savings**
100 projects × 10 hours = **1,000 hours of work freed for higher-value tasks**

This operational leverage allows EY teams to reallocate resources toward deeper analysis, stronger relationship-building, and faster time-to-close—accelerating the firm's overall deal throughput.

Enhanced Hit Rate in PE Outreach

By aligning outreach with firms most likely to acquire a given company, match relevance improves, thereby maintaining or increasing proposal and win rates despite a smaller outreach pool. The optimized funnel results in:

- ~10 PE firms contacted
- ~4–5 calls secured
- 2–3 proposals generated
- 1–2 deals closed

This preserves business outcomes while reducing upfront effort by 50%.

Competitive Differentiation through ML-Driven Prospecting

Adopting AI into the PE sourcing process positions EY as a **leader in innovation and client-centricity**. The PE Matchmaker tool delivers:

- **Speed**: reducing days of manual screening to seconds
- **Transparency**: justifiable, data-driven rationale for each suggested match
- **Precision**: tighter alignment between company profiles and PE investment behavior

This technology-enabled edge enhances EY's advisory credibility, particularly with clients expecting rapid and intelligent go-to-market strategies.

7. Limitations and Considerations

While the PE Matchmaker tool provides a valuable data-driven framework for PE targeting, several limitations must be acknowledged to contextualize its outputs and guide future improvements:

- **Data Gaps and Biases**
  The model is trained on historical investment data, which may contain omissions or inherent biases that affect the reliability of predictions.
- **Limited Training Data**
  The current dataset is relatively small, limiting the model's generalizability. Expanding the volume and diversity of ingested data would improve performance.
- **Market Volatility and Behavioral Shifts**
  The tool may not fully capture rapid changes in market conditions or shifts in PE firms' strategic focus, potentially impacting the accuracy of recommendations.
- **Model Confidence vs Analyst Judgment**
  While the LLM offers interpretability, its suggestions should be seen as decision support rather than standalone advice and always weighed against analyst expertise.
- **Unmeasured Relationship Dynamics**
  The model does not incorporate the qualitative strength of EY's existing relationships with specific PE firms, which can be a critical factor in real-world M&A engagements.

8. Next Steps and Roadmap

With the PE Matchmaker tool delivering strong initial value through ML-driven buyer recommendations, the next phase focuses on future-proofing and scaling the platform. This includes enhancing technical infrastructure, streamlining data operations, and unlocking broader market applicability, ultimately embedding the tool more deeply into EY's global PE advisory process.

Final Model Calibration and Deployment

The immediate priority is to finalize model calibration, validate interpretability with stakeholders, and stress-test recommendations against real transaction data. Once optimized, the model will be deployed via a conversational LLM interface in Streamlit, enabling analysts to receive high-confidence buyer matches in seconds through natural language prompts.

Real-Time Data Pipelines and Automation

To keep insights current, we will integrate the tool directly with EY's proprietary databases. Automated pipelines will ingest updates to firm profiles, transaction activity, and company metadata in real time. This ensures continuous learning and eliminates manual overhead, keeping the tool aligned with shifting market dynamics.

Building the EY Data Value Chain

A robust data value chain will enable structured scaling. We propose:

- Collecting and cleansing both structured and unstructured deal data
- Transforming and storing it in a central, queryable format
- Enriching model features and supporting retraining pipelines
- This infrastructure supports broader deployment and ensures repeatability across other use cases and regions.

Scalable Infrastructure and Cloud Architecture

Working with EY's technology teams, we recommend establishing a secure, cloud-native data platform using tools like Apache Airflow for orchestration and Snowflake or BigQuery for data warehousing. This architecture supports version control, automated ETL, and traceability, crucial for operationalizing AI responsibly at scale.

Market Expansion Opportunities

The architecture and learning models are inherently scalable, offering strong potential for geographic and strategic growth.

- **Geographic Scalability:**
  The matching logic, based on behavioural patterns like sector focus and deal size, can be retrained on local datasets to apply the tool in regions such as Portugal, Italy, or Latin America. This enables EY to deliver consistent value across deal teams globally.

- **Expansion to Other Buyer Types:**
  While the model is optimized for private PE firms, it can be extended to corporates, family offices, and strategic acquirers. These segments follow unique acquisition behaviours that the tool can learn and model to drive relevance.
- **Integration into Analyst Workflows:**
  Embedding the tool into EY's internal CRMs or M&A platforms will allow deal teams to trigger recommendations, track outcomes, and enrich firm profiles, all within their existing workflow. This ensures higher adoption and greater cross-functional value.

Future Enhancements

With a solid technical and operational foundation in place, additional innovations become possible:

- Enriching inputs with macroeconomic and sector-specific indicators
- Supporting multilingual interfaces for international adoption

9. GitHub Repository and Streamlit App

GitHub Repository

https://github.com/camillaperotti/mbd-corporateproject-EY/tree/main

Streamlit App

https://mbd-corporateproject-ey-npvtpf4mvfsmhj6hkwcdge.streamlit.app/