Business Logos as Predictors of Industry Domain

Hyeonu (Eric) Kim
British Columbia Institute of Technology
Coquitlam, Britsh Columbia, Canada
A01494656
hkim505@my.bcit.ca

Jia Qi (Jacky) Chen
BCIT Master in Applied Computing
Vancouver, British Columbia, Canada
A01315278
jchen574@my.bcit.ca

Hsuan-Chen (Alex) Liu
British Columbia Institute of Technology
Burnaby, British Columbia, Canada
hliu@my.bcit.ca

Abstract

This project explores whether a company's logo can reveal its business domain. Using a dataset of over 10,000 company logos from Crunchbase, which includes attributes such as company name, category group, country, and employee count, we aim to investigate whether visual features in logos are predictive of the industry domain (e.g., Technology, Food Beverage, Finance). The study involves extracting image embeddings from logos using pre-trained convolutional neural networks (ResNet-18) and applying classification models such as Random Forest and Support Vector Machines to predict business categories. The expected outcome is to identify whether distinct visual patterns such as color palettes, shape composition, or texture correlate with certain industries. This work could inform logo design strategies, marketing analytics, and automated company classification systems. The broader impact lies in understanding how design aesthetics reflect corporate identity and sector differentiation in modern branding.

ACM Reference Format:
Hyeonu (Eric) Kim, Jia Qi (Jacky) Chen, and Hsuan-Chen (Alex) Liu. 2025. Business Logos as Predictors of Industry Domain. In . ACM, New York, NY, USA, 4 pages. https://doi.org/none

1 Introduction

Logos are among the most recognizable representations of corporate identity, often conveying the nature of a business through color, typography, and shape. In marketing and design theory, visual symbolism plays a critical role in communicating industry cues: tech companies often favor minimalist or blue-toned designs, while restaurants tend to use warmer colors and organic forms. This raises an intriguing question: Can a machine learning model learn these associations and accurately predict a company's business domain from its logo image?

Our research specifically addresses the following questions:

(1) Which logos are most suggestive of the business domain they represent?
(2) Can we predict the business domain from the logo image with meaningful accuracy?
(3) Do certain industries (e.g., technology, food services) exhibit distinctive visual patterns?

Stakeholders include brand analysts, marketing consultants, and AI researchers interested in visual classification and design perception. The scope includes supervised image classification using existing Crunchbase logos, focusing on data visualization and domain prediction accuracy, while excluding text-based analysis or company metadata beyond the logo itself.

2 Related Work

Prior marketing and design research has established that logo features, such as color, typography, and shape, carry symbolic meanings that influence brand perception and recognition. Henderson and Cote analyzed how design complexity, naturalness, and harmony affect perceived brand personality and consumer recall, suggesting that visual elements communicate key brand traits aligned with business identity [2]. Similarly, Labrecque and Milne demonstrated that color cues evoke domain-specific associations. For example, blue often signals professionalism and trust in technology and finance sectors, while red and orange communicate excitement or food-related contexts[4]. These studies provide a theoretical foundation that certain visual elements systematically correspond to business types, motivating the hypothesis that logos alone may reveal the firm's sector.

The computational study of logos as image data has gained traction with the availability of large-scale logo datasets and the rise of convolutional neural networks (CNNs). Bianco et al. showed that CNN-based feature extraction significantly improves performance in object and logo recognition compared to handcrafted features[1]. Transfer learning with pre-trained networks such as ResNet enables effective extraction of image embeddings from limited logo datasets[3]. These embeddings capture high-level visual characteristics (color distributions, texture, and geometry) that can be used as inputs for classical classifiers such as Support Vector Machines (SVMs) or Random Forests. However, most prior work focuses on brand identification rather than the link between logo design and the company's business domain, an unexplored dimension this study aims to address.

Finally, while prior studies on startup success prediction demonstrate strong performance using structured tabular data such as funding, size, and location, they exclude visual branding features altogether[5]. This creates a crucial research gap: whether a company's logo contains predictive information about its business domain and how design aesthetics encode organizational identity.

Our study addresses these gaps by systematically examining whether machine learning models can learn and generalize the associations between logo design and business domain. By combining pre-trained CNN embeddings (ResNet-18) with interpretable classification models such as Random Forest and SVM, this research aims to provide evidence on whether the corporate logo connects to the sector a company operates in. In doing so, it bridges the theoretical work on visual branding and the technical work on

Conference'17, Washington, DC, USA
2025. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/none

computer vision, offering a new perspective on how design reflects industry conventions and corporate strategy.

3 Dataset Description

The dataset originates from Crunchbase (https://www.crunchbase.com/), containing metadata and logo URLs for over one million companies. Due to computational constraints and long download times (approximately one hour per million logos), we restricted our scope to the top 10,000 companies by visibility and ranking. The version used is `top10k_logos.csv`, derived from the full dataset for efficient processing.

The original `logo_url` column in the Crunchbase data contained broken or outdated links, making it impossible to retrieve images directly. To address this issue, we developed a custom web scraping pipeline to re-collect logo image URLs from Crunchbase and external APIs (e.g., Clearbit and Crunchbase image endpoints). The new URLs were stored in an additional column named `new_logo_url`, which now provides working links to company logo images.

The dataset includes companies from a wide range of industries and countries between roughly 2007–2022. The sample represents diverse domains including technology, finance, healthcare, retail, and food services.

Some of the key variables that we used to analyze data are:

- `company_name`: Company identifier
- `category_groups_list`: Business domain(s) - primary prediction target
- `country`: Company location
- `employee_count`: Employee size range (converted to numeric midpoints for analysis)
- `new_logo_url`: Updated and functional image URL for feature extraction

Logo Generation. After examining several entries in the dataset, we observed that all logo URLs followed a consistent structure within the `images.crunchbase.com` domain. A Python script was developed to generate a new column, `new_logo_url`, by extracting the public ID from each original link and reconstructing a standardized URL using the Crunchbase image template with the transformation string `c_pad, h_160, w_160, f_auto, b_white, q_auto:eco,dpr_2`. This ensured that all images were uniformly processed to 160×160 pixels with consistent padding, background, and quality settings. Standardizing the logos facilitates reliable comparison of visual features such as color distribution, shape, and aspect ratio by eliminating variations in image resolution and format.

Missing Data and Outliers. We found no missing values in the key variables used for analysis, including company name, business domain, and logo image. As our dataset primarily consists of categorical variables, there were no numerical outliers present.

Employee Count Distribution. Table 1 summarizes the company size distribution. Most firms have fewer than 500 employees, with a long tail of large enterprises.

To better understand the data distribution, several visualizations were generated.

Table 1: Distribution of Companies by Employee Count Range

Employee Count Range	Number of Companies
1–10	687
11–50	2,742
51–100	1,548
101–250	2,093
251–500	1,042
501–1000	664
1001–5000	594
5001–10000	176
10000+	386

Category Distribution. Figure 1 shows the number of companies by business domain. The largest categories include *Software*, *Financial Services*, *Information Technology*, and *Internet Services*, indicating the dataset's strong representation of digital and tech-driven industries.

Geographical Distribution. Figure 2 displays the top 20 countries by company count. The United States overwhelmingly dominates with over 6,000 entries, followed by the United Kingdom, India, and Canada. This geographic skew mirrors Crunchbase's concentration in English-speaking and tech-focused economies.

Correlation Heatmap of key-variables. Figure 3 The correlation heatmap of the key variables shows that the columns exhibit only very weak relationships. This suggests that the ranking was determined almost entirely from the image of the logo itself.

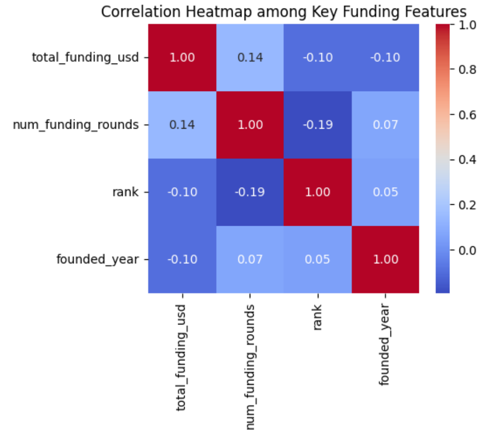

Figure 3: Correlation heatmap of key-variables

Reducing the dataset from over one million records to the top 10,000 companies introduces a potential sampling bias. The ranking metric on Crunchbase tends to favor well-known, well-funded, and predominantly North-American technology firms. As a result, the dataset may under-represent smaller startups, non-English-speaking regions, and traditional industries with limited digital presence. This concentration could influence model learning by reinforcing visual and categorical patterns common in the tech sector,

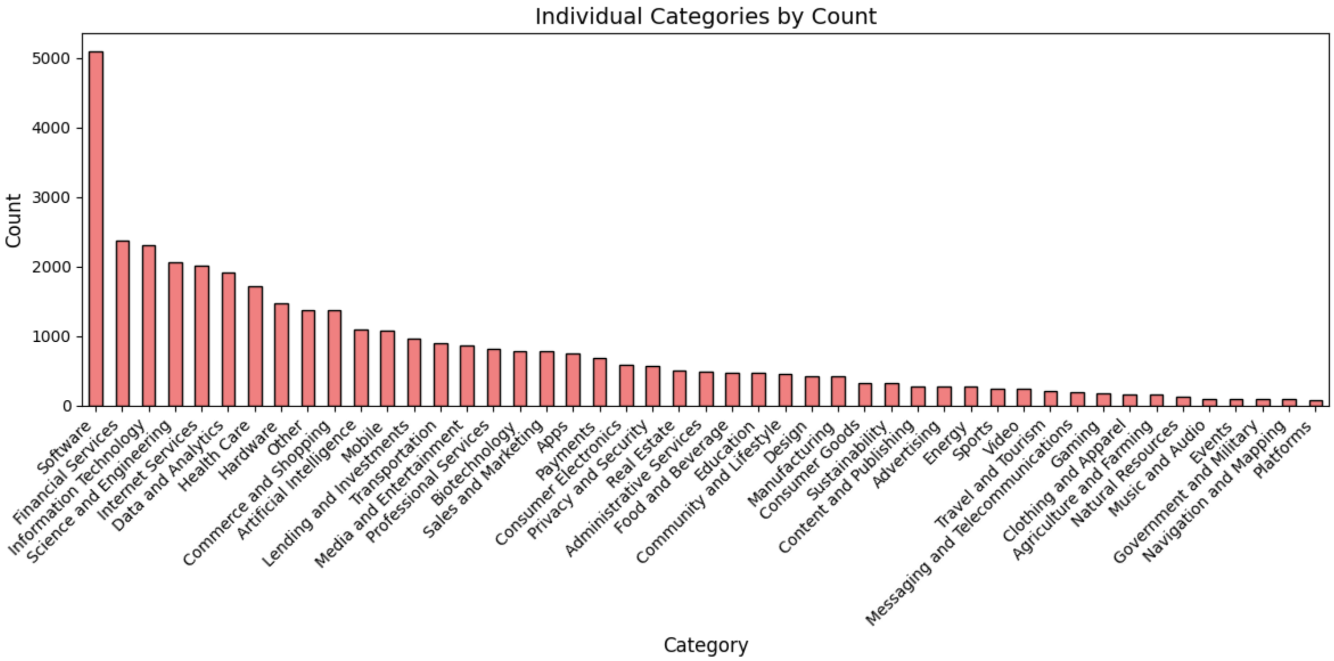

Figure 1: Distribution of Companies by Business Category

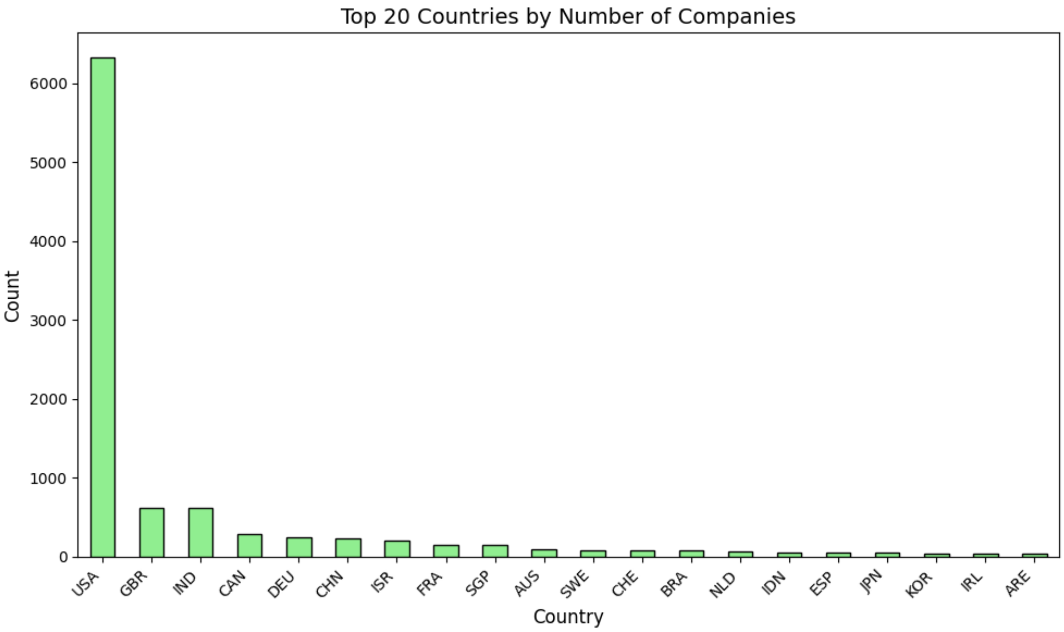

Figure 2: Top 20 Countries by Number of Companies

thereby reducing generalizability to the global business population. Future work could address this limitation by incorporating stratified or randomly sampled subsets from the full dataset to ensure more balanced coverage across countries and sectors.

All data used are publicly available company information. Logos represent corporate entities rather than individuals, minimizing privacy concerns. The dataset is used strictly for academic and non-commercial purposes, in compliance with open data and fair-use principles.

References

[1] S. Bianco, M. Buzzelli, D. Mazzini, and R. Schettini. 2017. Deep learning for logo recognition. *Neurocomputing* 245 (July 2017), 23–30. doi:10.1016/j.neucom.2017.03.051

[2] P. W. Henderson and J. A. Cote. 1998. Guidelines for Selecting or Modifying Logos. *Journal of Marketing* 62, 2 (April 1998), 14–30. doi:10.1177/002224299806200202

[3] L. Hsairi. 2024. Deep Learning to Predict Start-Up Business Success. *International Journal of Advanced Computer Science and Applications (IJACSA)* 15, 3 (March 2024). doi:10.14569/IJACSA.2024.0150336

[4] L. I. Labrecque and G. R. Milne. 2013. To be or not to be different: Exploration of norms and benefits of color differentiation in the marketplace. *Marketing Letters* 24, 2 (June 2013), 165–176. doi:10.1007/s11002-012-9210-5

[5] K. Żbikowski and P. Antosiuk. 2021. A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing & Management* 58, 4 (July 2021), 102555. doi:10.1016/j.ipm.2021.102555