Imperial College London

Department of Earth Science and Engineering

MSc in Applied Computational Science and Engineering

Independent Research Project

Investor Contact Data – EDA & Cleaning Summary

Unlocking Global Capital: AI-Powered Prediction, Access, and Verification of Investor Decision Makers

by

Daniel Bowman

Email: daniel.bowman24@imperial.ac.uk

GitHub username: acse-db1724

Repository: ese-ada-lovelace-2024/irp-db1724

Supervisors:

Antony Sommerfeld

Dr. Yves Plancherel

June 2024

Table of Contents

List of Tables

List of Figures

Abstract

Automating investor contact maintenance is critical for scalable, reliable capital raising in a sector plagued by high turnover and data decay. Existing commercial services like Hunter and ZoomInfo rely on paid lookups and often take days to return results. In contrast, our novel pipeline seamlessly integrates three components—a comprehensive offline template miner that uncovers every common formatting skeleton, a lightweight real-time classifier that instantly predicts the correct template for any new name–domain pair, and on-the-fly third-party deliverability scoring—into one end-to-end system. Performance will be evaluated on held-out contacts, reporting template coverage, prediction accuracy, API precision/recall, and sub-50 ms query latency. This hybrid approach not only outperforms standalone pattern-mining or machine-learning methods but also undercuts costly data-provider fees, democratizing access to fresh, validated investor email information.

1. Introduction

The purpose of this document is to summarize the key findings from Exploratory Data Analysis (EDA) performed on all available investor contact data. The focus is on the LP dataset, although EDA is done on both the GP and combined datasets as well to account for future stretch goals that will include these datasets. The primary aim was to assess data quality, identify patterns in missing or inconsistent values, and decide on the appropriate preprocessing tasks for downstream tasks such as email template prediction and pattern mining.

2. Data Overview

This section outlines the general structure and characteristics of the raw LP and GP datasets before cleaning or transformation. Understanding the baselines schema and field distribution was critical to informing the subsequent cleaning and imputation decisions.

2.1. Dataset Structure

The LP dataset has 155963 rows within it, where as the larger GP dataset has 267454. Data is originally provided in csv format and then added to a SQL database.

2.2. Field Types and Summary Stats

The dataset is entirely categorical (except for telephone number and row ID which was added when porting from csv to SQL). With that in mind the numeric statistics are not very informing.

	id	time_stamp
count	155963	155963
mean	77982	2025-06-19 12:21:47
min	1	2025-06-19 12:21:47
25%	38991.5	2025-06-19 12:21:47
50%	77982	2025-06-19 12:21:47
75%	116972.5	2025-06-19 12:21:47
max	155963	2025-06-19 12:21:47

std	45022.78

Table 1: Numeric Stats LP

And of course we see the same thing in the GP dataset.

	id	time_stamp
count	267454	267454
mean	133727.5	2025-06-19 12:24:24
min	1	2025-06-19 12:24:24
25%	66864.25	2025-06-19 12:24:24
50%	133727.5	2025-06-19 12:24:24
75%	200590.8	2025-06-19 12:24:24
max	267454	2025-06-19 12:24:24
std	77207.46	

Table 2: Numeric Stats GP

For categorical features, we see strong cardinality in investor and firm fields (which supports their use as joint primary identifiers in deduplication and modelling). In contrast title, firm_type only have a few unique values. Job_title is highly varied but not as cardinal as the investor and firm fields.

	count	unique	top	freq
investor	155963	141147	Michael Miller	22
firm_type	155963	51	Wealth Manager	47355
title	134364	18	Mr.	101648
firm	155963	24133	Jones Day Defined Benefit Master Trust	1769
alternative_name	10620	9771	Monty Cleworth	10
role	143815	288	Investment Team	72284
job_title	155873	51998	Partner	4873
asset_class	155963	258	GEN	29617

email	82811	79925	mmiller@crewcialpartners.com	12
tel	143838	57238	+1 212 761 4000	383
city	151260	4392	New York	12138
state	106990	528	NY	14415
country	155963	138	US	95650
zip_code	144722	12798	10022	1459
linkedin	118080	111892	www.linkedin.com/in/michael-miller-5a060523/	13
region	155926	7	North America	101374
address	155963	22607		1848
website	149414	20697	https://www.jonesday.com/en	1769
general_email	102980	14786	contact@weil.com	954
source_file	155963	1	LP Contact Data	15596 3

Table 3: Categorical Stats LP

See similar trends, scaled to the larger dataset, for GP.

	count	unique	top	freq
investor	267454	227167	Wei Wang	23
firm_type	239571	41	Private Equity Firm	122801
title	249022	18	Mr.	193967
firm	267454	40838	GC&H Investments	831
alternative_name	0	0		
role	0	0		
job_title	267191	65299	Partner	18582
asset_class	253688	573	PE	79427
email	170748	166555	peter.chapman@accretion.com.au	2
tel	203264	70087	+1 212 583 5000	257
city	231157	3531	New York	27720
state	131911	670	NY	29161

country	233516	161	US	108915
zip_code	206423	11755	10022	6174
linkedin	200604	193140	www.linkedin.com/in/bragiel/	8
region	239229	7	North America	125076
address	267454	38172		6964
website	254016	36801	http://www.cooley.com	831
general_email	213221	29223	irares@aresmgmt.com	708
source_file	267454	1	GP Contact Data	267454

Table 4: Categorical Stats GP

3. Missingness Analysis

This section evaluates the completeness of key fields in the various datasets, this helps guide the cleaning and imputation decisions.

3.1. Column Level Missingness

A horizontal bar chart was used to illustrate missing rates in various columns.

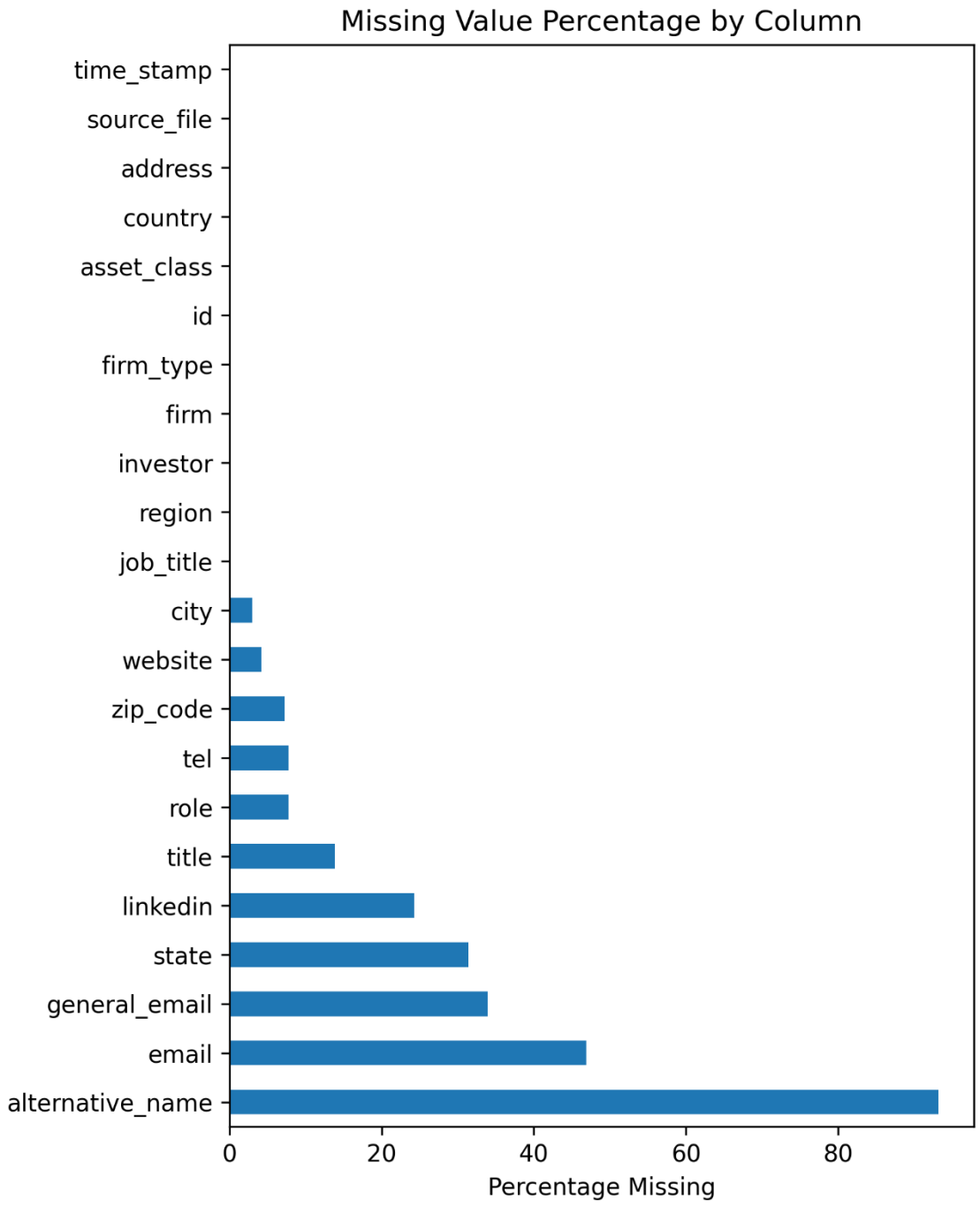

Figure 1: Missingness Bar Chart LP

The email field has a lot of missing values, whereas website, and LinkedIn seem to be less sparse. This will motivate downstream imputation as the domain can be inferred manually from the website there might even be opportunity to do manual recovery via LinkedIn although that will be a big task. Alternatively, if

pattern mining finds that certain firms use the same template every time, missing emails could be programmatically recovered for those firms.

These trends largely hold true for the GP dataset as well.

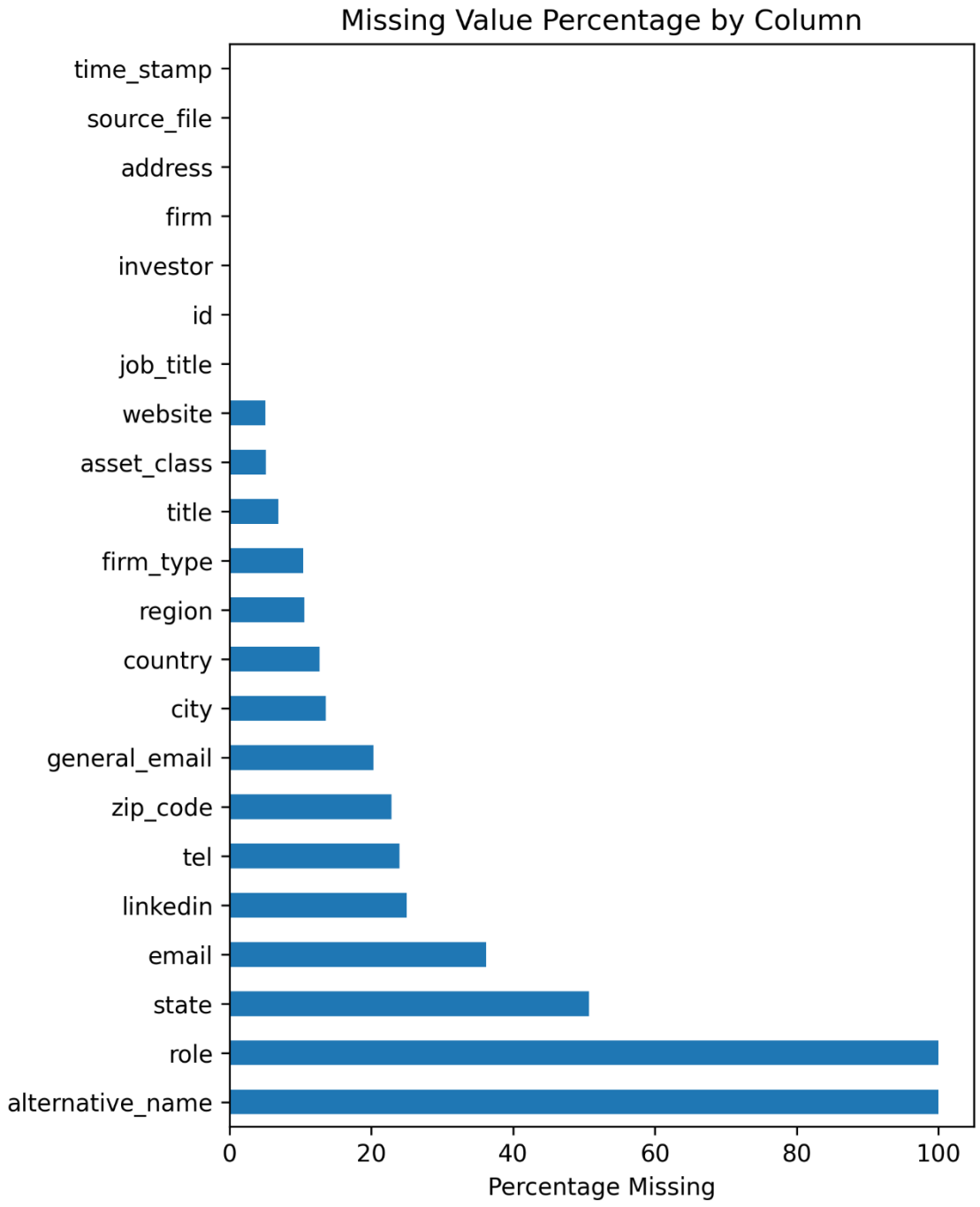

Figure 2: Missingness Bar Chart GP

3.2. Missingness Correlation

A heatmap was used to try and deduce any corelation between missing fields.

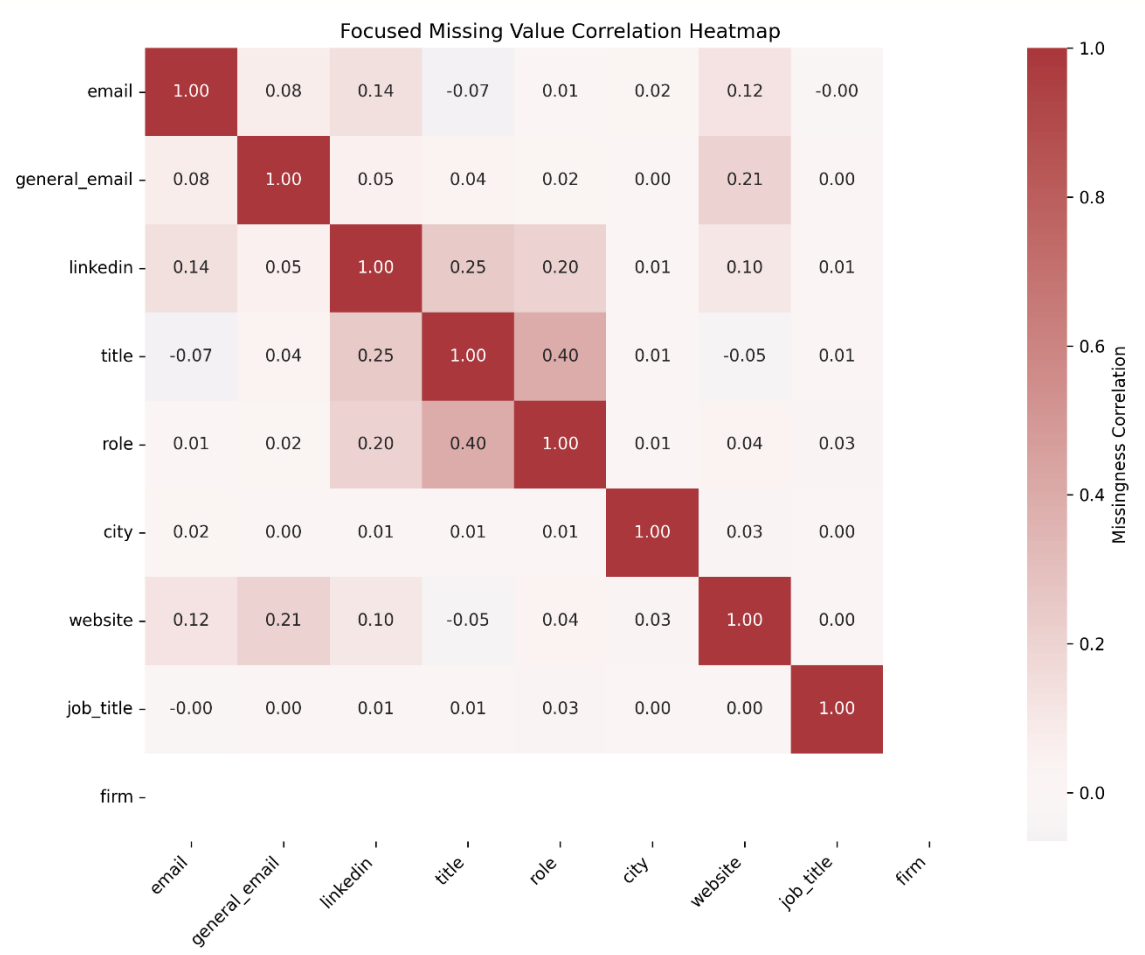

Figure 3: Missingness Correlation LP

This shows a weak corelation between missing key fields, indicating that they are independent in most cases. No strong evidence of structured or cascading data loss patterns, which is not helpful for informing imputation strategies. The analysis suggests that we cannot rely on other missing fields and must instead attempt imputation techniques based on our pattern mining findings.

The same holds true for the GP dataset.

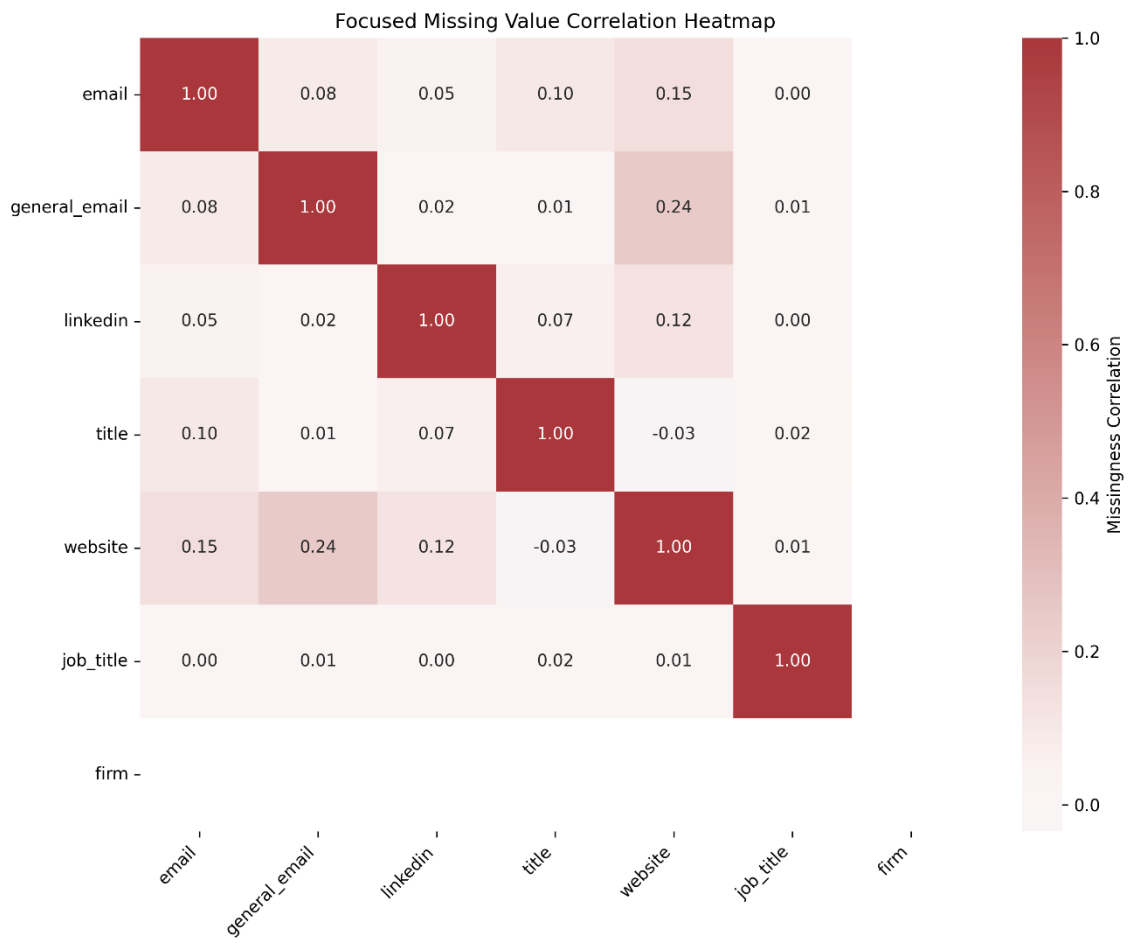

Figure 4: Missingness Correlation GP

4. Duplicates

This section identifies both exact and near duplicates in both datasets to ensure uniqueness and reduce redundancy in the training data.

4.1. Definite Duplicates

Entries watching matching investor name and either email or LinkedIn details were flagged as duplicates. Those duplicates were then group by email and pruned, keeping one ‘best’ row that was decided based on completeness. A second pass was done but instead grouped by LinkedIn. This resulted in 6642 dropped rows in the LP dataset and 11868 in the GP dataset.

4.2. Ambiguous Duplicates

Rows with the same investor but varying in other fields were flagged but not immediately removed as they could represent career moves or different investors with the same name.

investor	email	firm
aaron ammerman	aammerman@alphaky.com	alpha financial partners
aaron ammerman		alpha financial partners
aaron brodt	abrodt@at-pw.com	ashton thomas private wealth
aaron brodt		amplify financial
aaron brodt		ashton thomas private wealth
aaron cohen		gyl financial synergies
aaron cohen		city national bank investment management
aaron foster	aaron.foster@compoundplanning.com	compound planning
aaron foster		rbc dominion securities
aaron jones	amjones@ft.newyorklife.com	eagle strategies
aaron jones		nicholas wealth management
aaron king	aaron.king@stblaw.com	retirement plan for non legal employees of stb
aaron king		lgt wealth management
aaron kowal		creative planning
aaron kowal		creative planning
aaron lim	aaronlim@azalea.com.sg	azalea investment management
aaron lim		roths investment bank

Table 5: Ambiguous Duplicate LP

And for the GP dataset.

investor	email	firm
aadil chitalwala	aadil@peakventures.in	peak sustainability ventures
aadil chitalwala		peak venture partners
aakanksha sharma	aakanksha@venturehighway.vc	general catalyst india
aakanksha sharma	asharma@generalcatalyst.com	general catalyst

aakash butala	aakash.butala@rothschildandco.com	five arrows
aakash butala		rothschild merchant banking
aakash desai	adesai@rvcapital.com	rv capital management
aakash desai		360 one asset
aakash desai		360 one
aakash jain	aakash.jain@arvog.com	arvog
aakash jain	aakashjain@venture.university	venture university venture partners
aakash kumar	aakash@caspian.in	caspian
aakash kumar	aakash@matrixpartners.in	z47
aakash kumar	aakash@z47.com	z47
aakash patel	aakash@bluewolfcapital.com	blue wolf capital partners
aakash patel	apatel@riverside.ac	riverside acceleration capital
aakash shah	aakash@peakventures.in	peak venture partners
aakash shah	ashah@centeroakpartners.com	centeroak partners
aaron brown	aaron.brown@nfiindustries.com	nfi ventures, llc

Table 6: Ambiguous Duplicate GP

a) Career Move Detection

Career moves were detected by identifying investor-role-job title grouping where the number of records matched the number of distinct firms. These entries were retained and merged back into the clean dataset as valid, the idea being that despite it possibly representing a duplicate, the syntactic structures of the same name used across different firm/domains will provide valuable training data to our classifier.

investor	role	count	unique_firms
david harris	investment team	7	7
michael smith	investment team	6	6
michael lee	investment team	6	6
brian johnson	investment team	6	6
john williams	investment team	5	5
brian kim	investment team	4	4
john lee	investment team	4	4
john kim	investment team	4	4

michael bradley	investment team	4	4
mike chen	investment team	4	4
jason chen	investment team	4	4
michael obrien	investment team	4	4
mark miller	investment team	4	4
scott wilson	investment team	4	4
william wang	investment team	4	4
john smith	investment team	4	4
jason smith	investment team	3	3
william scott	investment team	3	3
andrew park	investment team	3	3
scott anderson	investment team	3	3

Table 7: Career Move Stats LP

And for the GP dataset.

investor	job_title	count	unique_firms
investor relations	investor relations	10	10
david fischer	partner	3	3
general contact	general contact	3	3
michael chen	managing director	3	3
ed lascelles	partner	3	3
mark mccall	managing director	3	3
michael johnson	managing director	3	3
greg williams	managing director	3	3
alessandro benetton	founding managing partner	3	3
blake miller	vice president	3	3
john reed	managing director	3	3
milo werner	general partner	3	3
michael lee	managing director	3	3
omni team	data contact	3	3
edward lee	vice president	3	3
michael jones	managing director	3	3

david lee	partner	3	3
bo liu	partner	3	3
marnix roes	investment manager	3	3
james davis	managing partner	3	3

Table 8: Career Move Stats GP

5. Email-Firm-Domian Analysis

This section analyses the structural patterns between form names and email domains, this was done to get some early pre-pattern-mining insight on the quality of our data. The main areas of focus was reuse of email domains across different firms (domain ambiguity), the reverse (use of multiple domains within the same firm – subsidiaries) and superficial structural entropy in the local part of emails.

5.1. Domain Reuse Across Firms

Several domains use multiple firms. Cases such as ‘gmail.com’ or ‘yahoo.com’ are clearly individual investors or personal accounts, whereas ‘lpl.com’ must be some sort of shared infrastructure across various departments or subsidiaries.

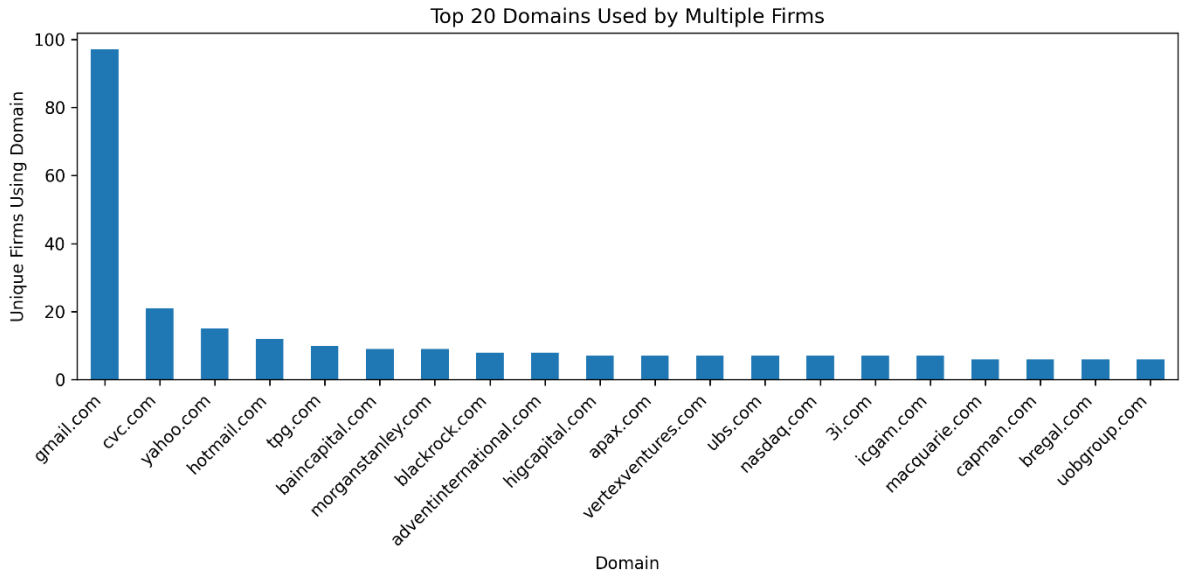

Figure 5: Domain Reuse LP

And for the GP dataset.

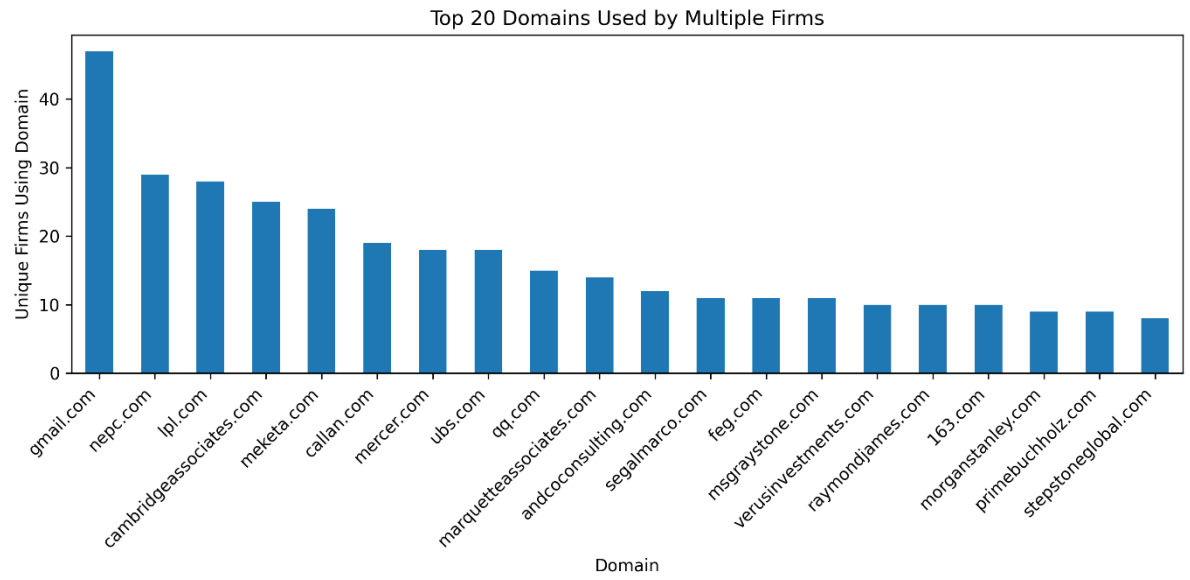

Figure 6: Domain Reuse GP

To account for this, the ‘is_shared_infra’ we introduced into the clean data set to help the model down weight reliance on domain alone during template prediction.

5.2. Firms Using Multiple Domains

Reversely, several domains are used by multiple firms. Som large firms use several different domain types, probably due to different departments or subsidiaries that all identify with their parent company but use different infrastructures for whatever reason.

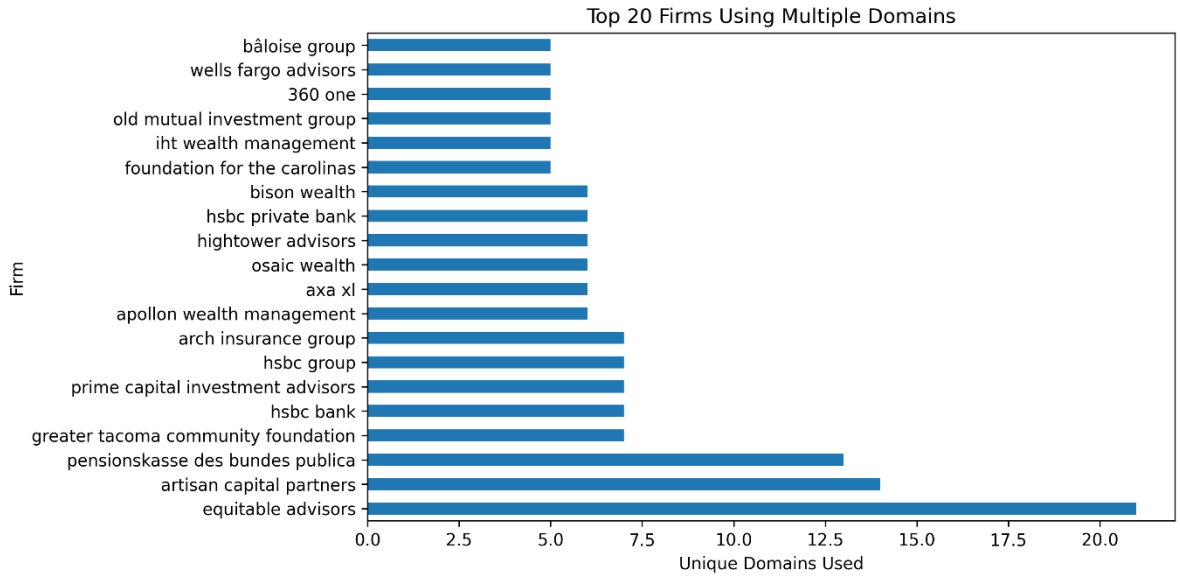

Figure 7: Firms Using Multiple Domains LP

And for the GP dataset.

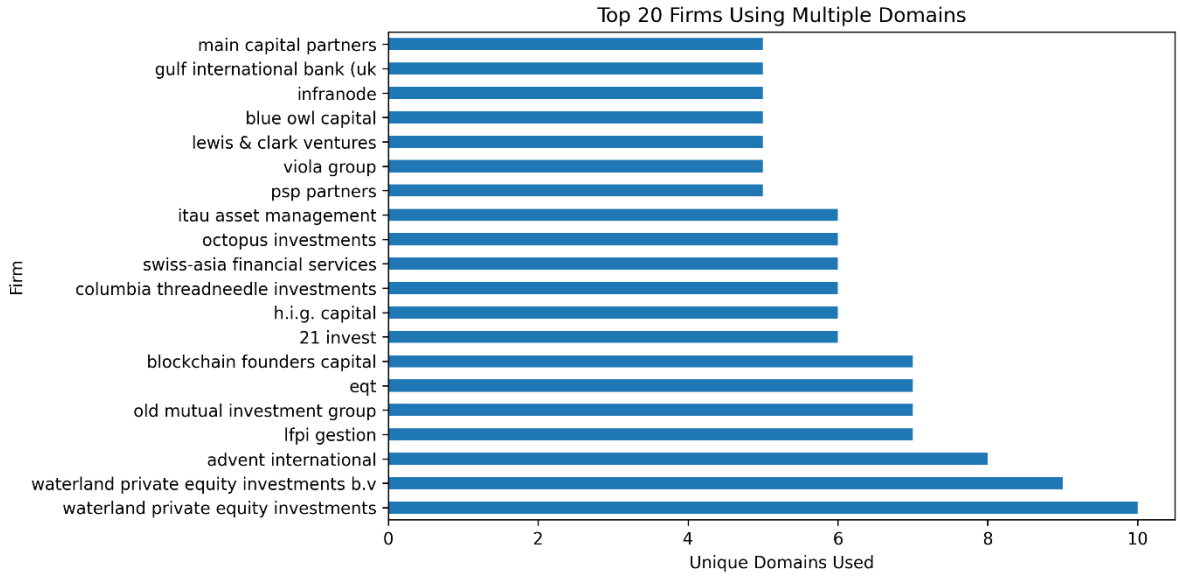

Figure 8: Firms Using Multiple Domains GP

To address this, ‘firm_is_multi_domain’ flag is included in the clean dataset to enable template logic to fall back to more granular rules or loosen confidence

thresholds in cases where multiple template structures are represented within the same firm (but with different domains).

5.3. Entropy of Email Local-Parts

Finally, a quick precursory look at variability of the local parts of email addresses. This was done to give us an idea of what how diverse the templates are – without doing full fledged pattern mining. Results generally point to high variability in the local parts and hint to multiple template structures used across the data set but we won't have a concrete idea until the next phase.

6. Cleaning Actions Taken

Based on the insights gained from EDA, we have taken the following steps for cleaning. For a more detailed understanding of the EDA journey and the points made, check the various EDA jupyter notebooks in the notebooks folder.

Step	**Description**
Dropped rows with missing or invalid emails.	Ensured all entries used in modelling contain valid email values with structurally correct format.
Removed emails with unmatched special characters.	Cleaned out entries with non-alphanumeric symbols not reflected in the investor's name.
Definite duplicate removal	Rows with matching email or LinkedIn were grouped and pruned, keeping the best representative based on non-null field score.
Career-move disambiguation	Ambiguous duplicates were retained if their investor, role, and job_title matched and were spread across distinct firms - interpreted as role transitions.

Domain structure flags	Two Boolean columns were added: - is_shared_infra: True if the domain is used by multiple firms. - firm_is_multi_domain: True if the firm uses multiple email domains.

Table 9: Cleaning Actions

These cleaning actions balance data quality and preserving legitimate variations, while enabling future predictive modelling stages to handle edge cases systematically.

7. Risks and Challenges

Despite the preprocessing steps taken and the significant improvement in data reliability that was gained as a result, several challenges and risks remain that may affect downstream modelling.

These risks inform the need for conservative fallback strategies, rigorous validation, and potentially a secondary model phase to catch misclassifications or refine prediction confidence thresholds.

7.1. Ambiguity and Duplicate Risk

Common names in the dataset have been pruned to some degree, but potential duplicate remain which could pollute our dataset. There is also the possibility that our preprocessing had erroneously removed valid data that could have been valuable in model training.

7.2. Email-Domain Edge Cases

Many domains are used across different companies, in some cases this is because of different departments using the same infrastructure and then quite possibly the same template structure. However, individual investors and other edge cases could lead to template confusion in our learning model. Similarly, firms that use multiple domains could lead to false generalizations when inferring email templates.

7.3. Template Learning Limitations

Some domains exhibit high entropy in email structures – at least from a quick precursory look – which could make deterministic template extraction difficult.

Also, field sparsity in certain metadata fields are too sparse to meaningfully assist with disambiguation or clustering at scale.

8. Next Steps

With a cleaned and profiled dataset in place, the following steps will incorporate our EDA findings on top of our existing project plan for the pattern mining phase.

8.1. Pattern Mining Phase

As already confirmed in the project plan, TRuleGrowth will be used to sequence mine and extract common email local part structures per domain. These will then be grouped by shared firm and domain (and maybe even job title) to isolate template consistency within different fields.

8.2. Template-Based Prediction and Recovery

Once template distribution has been computed, templates that are present consistently across various domains and firms will be applied to missing rows to attempt to recover missing data. If this does not prove fruitful, then fallback options will be considered (general email, manual recovery via LinkedIn).