Applied Data Science Research Project

Group G

Index

01. COMPANY

	count	unique	top	freq	Data Type
COMPANY_ID	9971	9971	drift	1	object
COMPANY_NAME	9971	9971	Drift	1	object
CATEGORY	9948	8695	Health Care, Medical, Medical Device	27	object
LOCATION	9938	1682	San Francisco, California, United States	789	object
FOUNDED_ON	9971	1237	2014	1201	object
EXITED_ON	1481	1060	Jun 20, 2018	7	object
CLOSED_ON	431	188	2020	83	object

Variable	Null Count
COMPANY_ID	0
COMPANY_NAME	0
CATEGORY	23
LOCATION	33
FOUNDED_ON	0
EXITED_ON	8490
CLOSED_ON	9540

02. INVESTMENT

	count	unique	top	freq	Data Type
COMPANY_ID	19789	8648	suncayr	23	object
FUNDING_TYPE	19789	27	Seed	6939	object
ANNOUNCED_DATE	19789	3100	Jan 1, 2016	137	object
INVESTMENT_STAGE	13502	4	Seed	8446	object
MONEY_RAISED	14770	3659	$1,000,000	437	object

Variable	Null Count
COMPANY_ID	0
FUNDING_TYPE	0
ANNOUNCED_DATE	0
INVESTMENT_STAGE	6287
MONEY_RAISED	5019
Year	0

03. ACQUISITION

	count	unique	top	freq	Data Type
ACQUIRER_ID	950	619	animoca-brands-corporation	11	object
ACQUIREE_ID	950	948	tableapp	2	object
ANNOUNCED_DATE	950	741	Oct 16, 2018	4	object
ACQUISITION_TYPE	877	4	Acquisition	821	object
PRICE	118	104	$2,000,000	4	object

Variable	Null Count
ACQUIRER_ID	0
ACQUIREE_ID	0
ANNOUNCED_DATE	0
ACQUISITION_TYPE	73
PRICE	832

04. EMPLOYEE

	count	unique	top	freq	Data Type
COMPANY_IDS	20515	9736	oculus-vr	40	object
JOB_TITLES	20515	6232	Founder	1014	object
ATTENDED_SCHOOLS	7628	3321	Stanford University	130	object

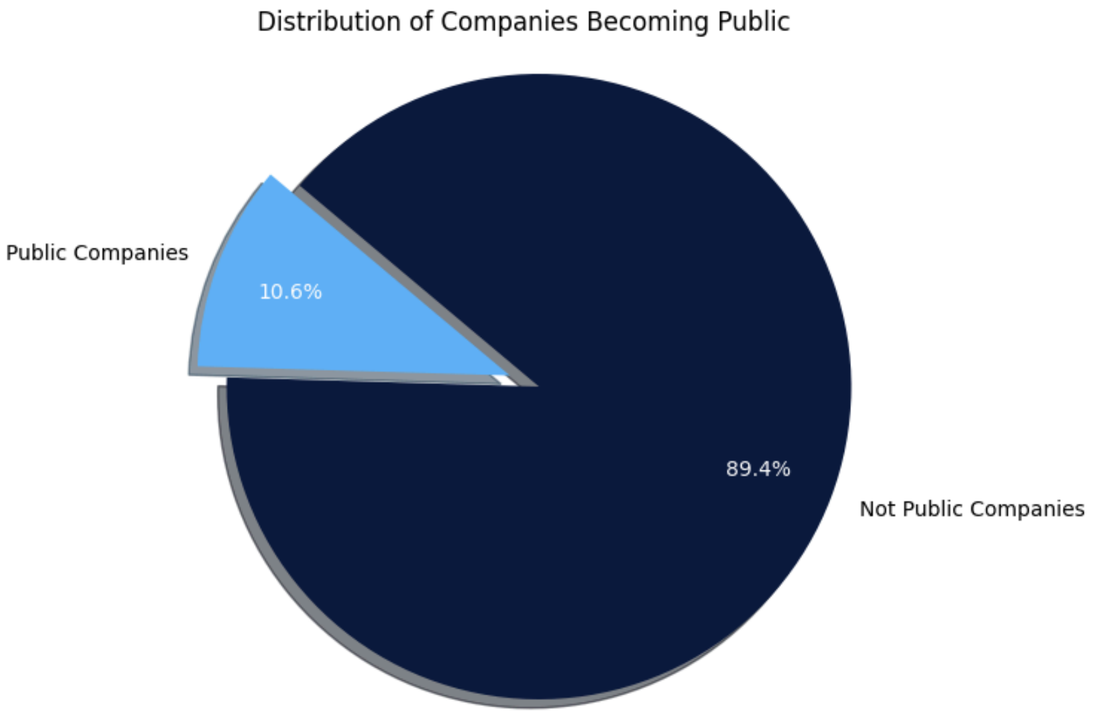
Variable	Null Count
COMPANY_IDS	0
JOB_TITLES	0
ATTENDED_SCHOOLS	12887

05. NEWS

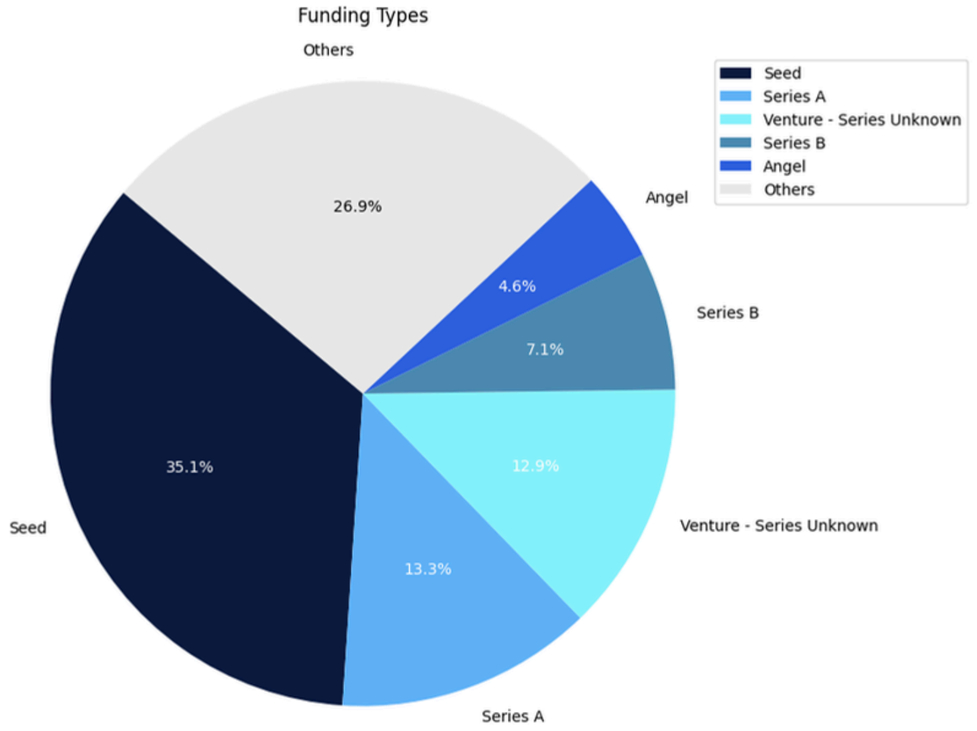
	count	unique	top	freq
COMPANY_ID	25497	5273	cresco-labs	41
NEWS_DATE	25497	3472	Sep 13, 2021	39
NEWS_VENUE	24186	4133	TechCrunch	1279
NEWS_TITLE	25465	24950	Alkami Announces Acquisition of Digital Account Opening and Loan Origination Provider MK Decision	24

Variable	Null Count	Data Type
COMPANY_ID	0	object
NEWS_DATE	0	object
NEWS_VENUE	1311	object
NEWS_TITLE	32	object

0.1. COMPANY

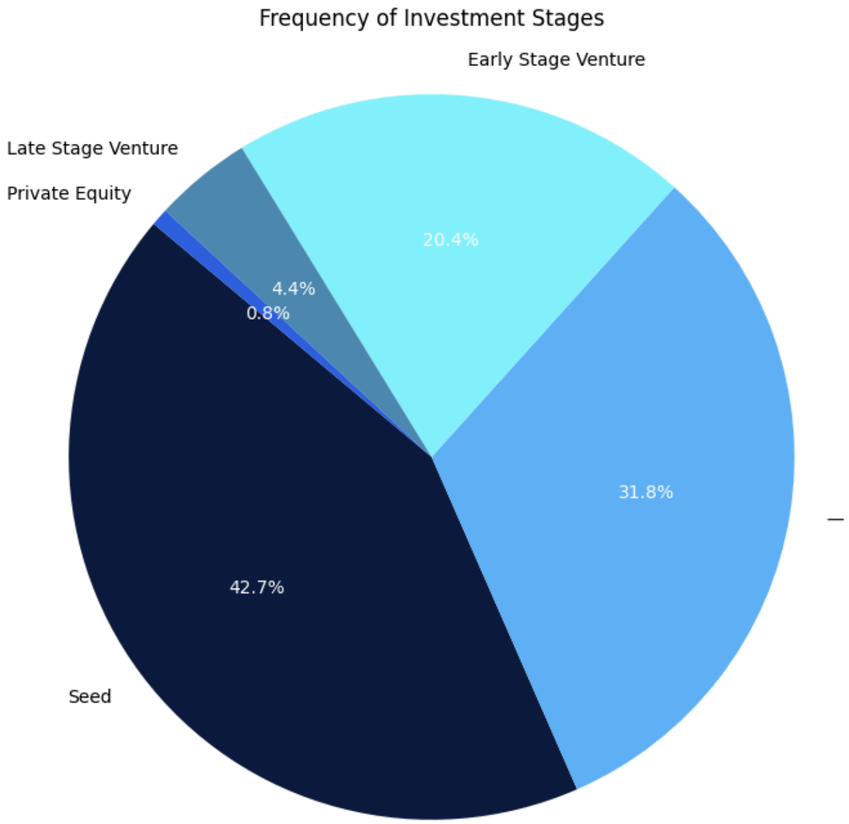


0.2. INVESTMENT

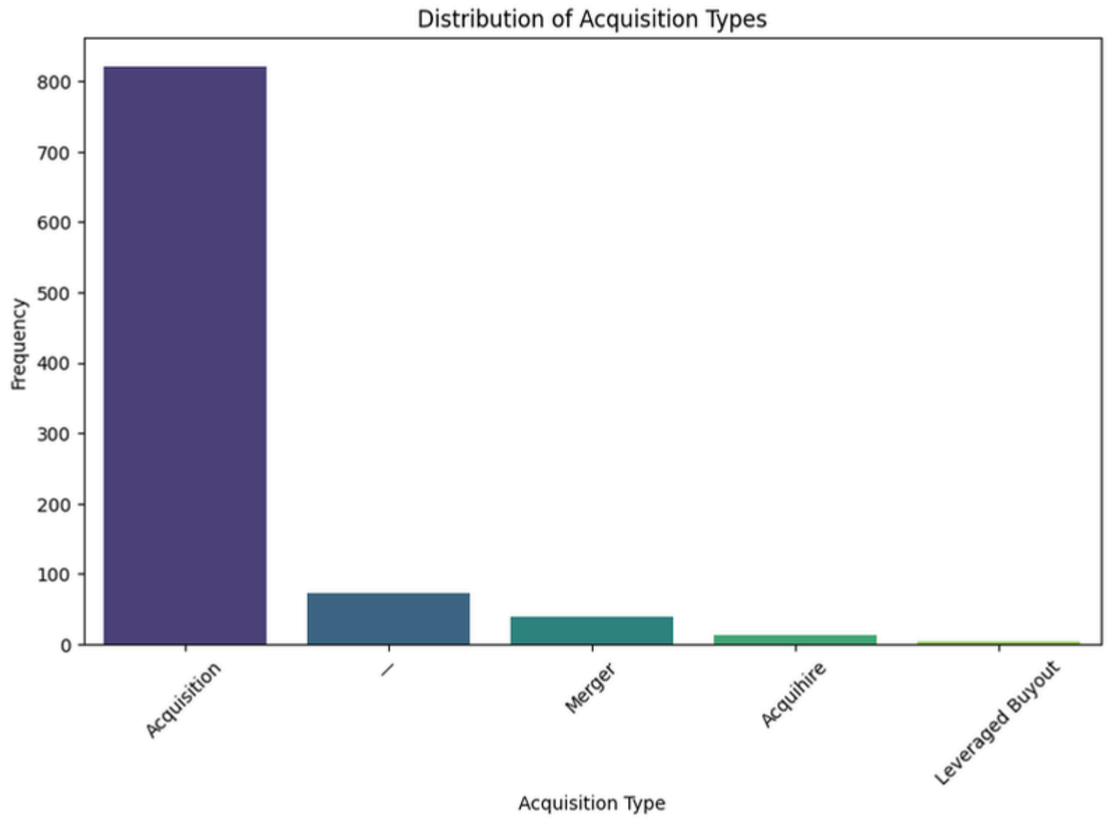


Data Pre-Analysis

0.2. INVESTMENT

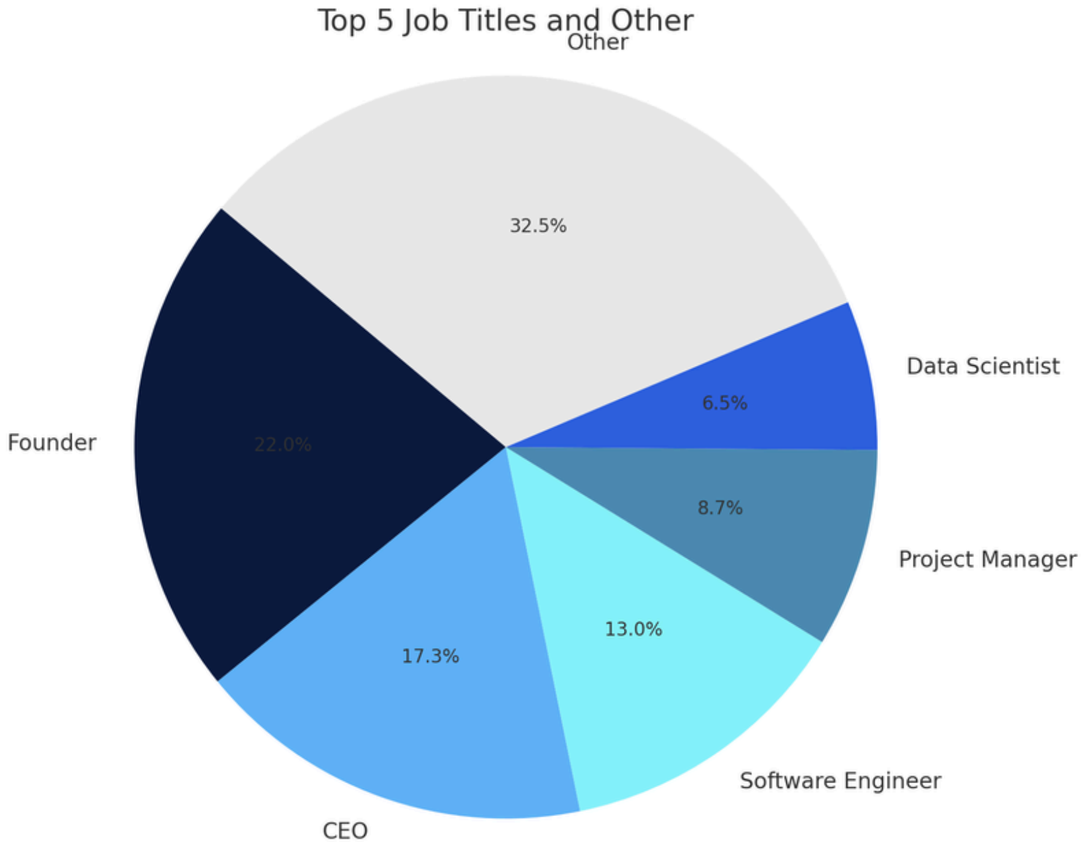


0.3. ACQUISITION

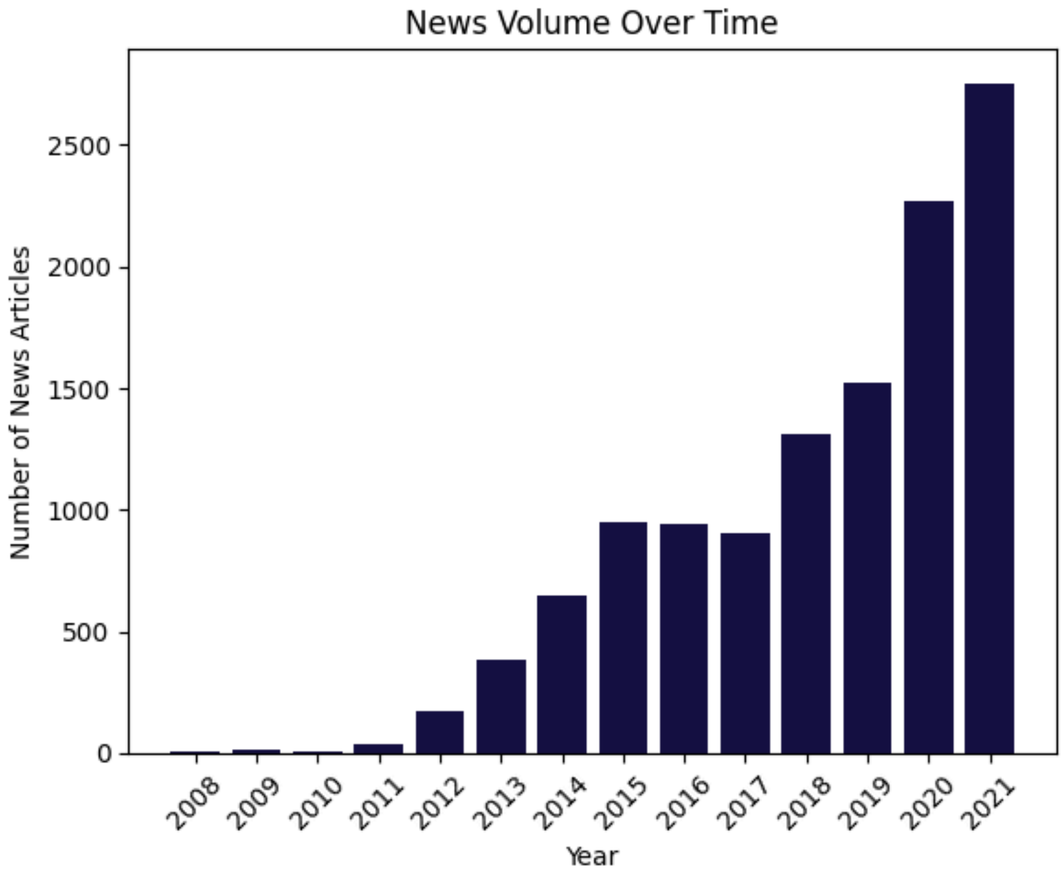


Data Pre-Analysis

0.4. EMPLOYEE



0.5. NEWS



Data Pre-Analysis

Research Question

Which factors Influence a Company's Chance to Secure Non-Seed Investments*?

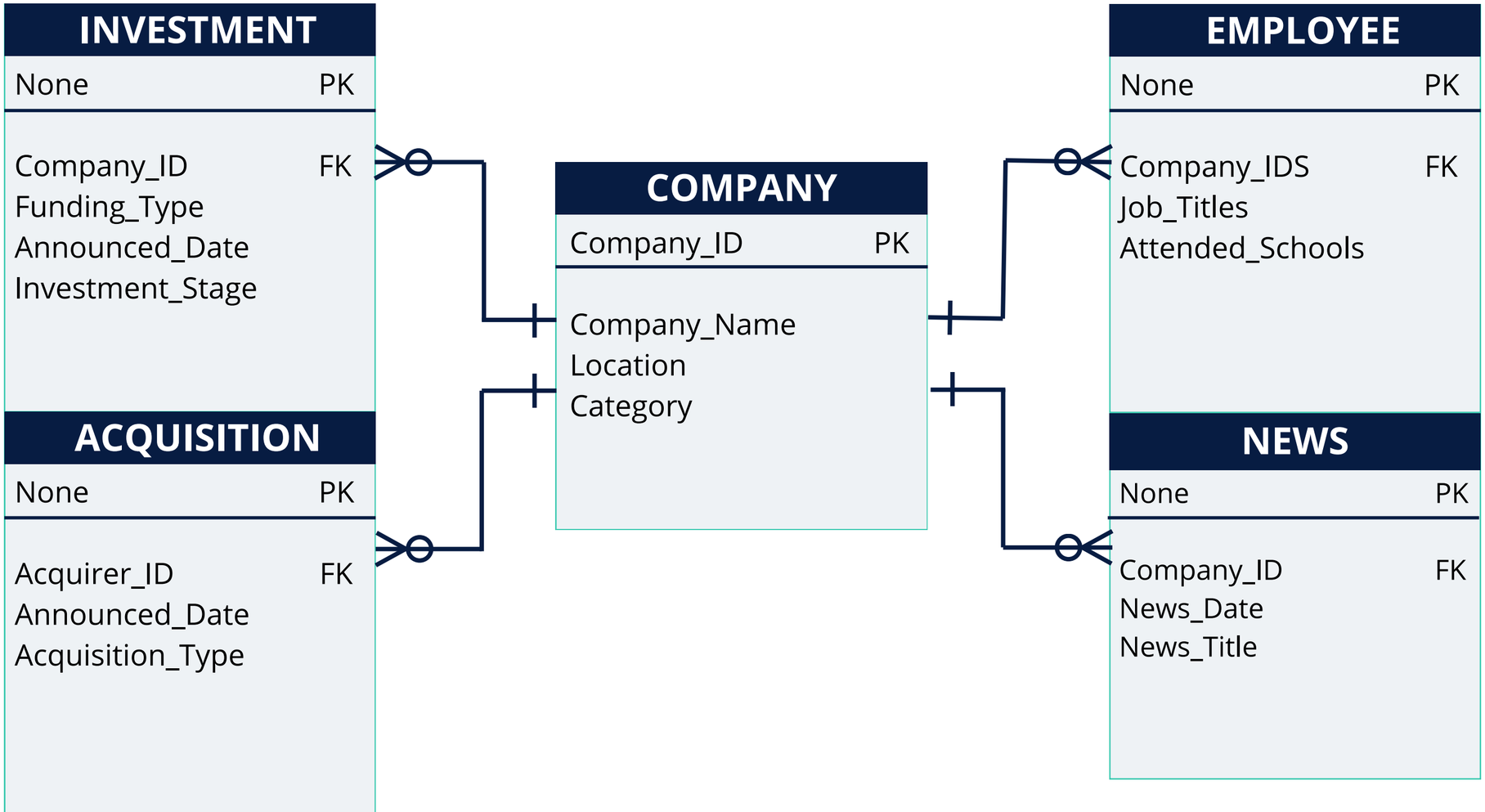
**Early Stage Venture, Late Stage Venture and Private Equity*

OBJECTIVE

WE IDENTIFY FACTORS THAT INFLUENCE STARTUPS TO SECURE NON-SEED INVESTMENTS

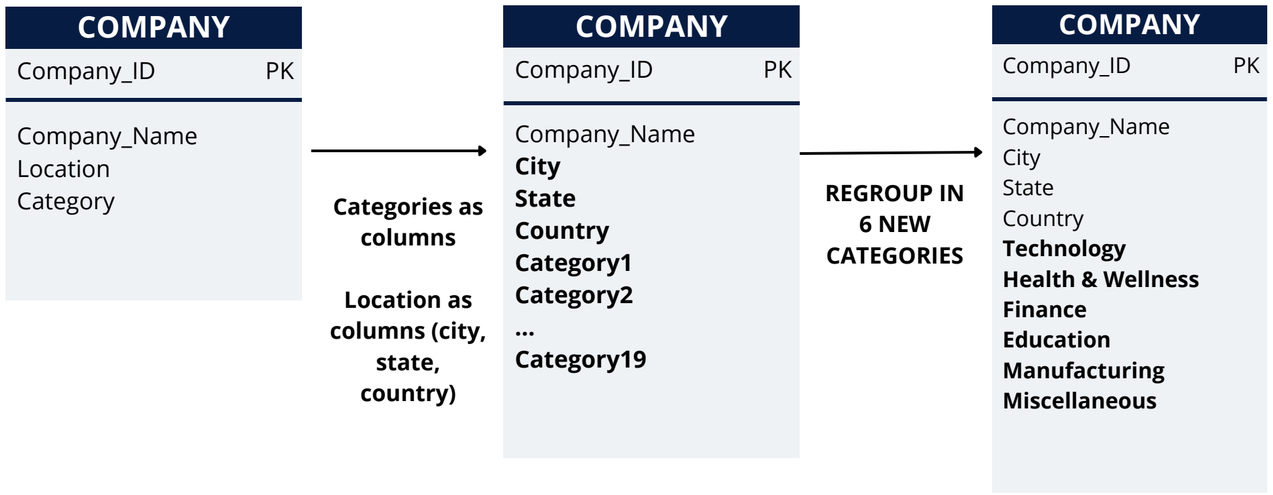
WE BUILD A MODEL TO ESTIMATE IF A STARTUP WILL SECURE NON-SEED INVESTMENT

The Dataset


INVESTMENT
None PK
Company_ID FK
Funding_Type
Announced_Date
Investment_Stage
ACQUISITION
None PK
Acquirer_ID FK
Announced_Date
Acquisition_Type
COMPANY
Company_ID PK
Company_Name
Location
Category
EMPLOYEE
None PK
Company_IDS FK
Job_Titles
Attended_Schools
NEWS
None PK
Company_ID FK
News_Date
News_Title

Cleaning &Transformation

01. COMPANY



CATEGORY

- Split the 'CATEGORY' column into separate columns, each containing a single category
- Categorized each category into broader industries using a predefined mapping dictionary

LOCATION

- Split the 'Location' column into separate columns: “City”, “State” and “Country”

DATE

- Converted only-years values to the first day of the year
- Standardized dates into the format “YYYY - MM - DD

Cleaning & Transformation

02. INVESTMENT

INVESTMENT	
COMPANY_ID	PK
Announced_Date Investment_Stage	

→

INVESTMENT	
COMPANY_ID	PK
Announced_Date Investment_Stage **ADVANCED_STAGE**	

Announced Date

- Change the Data Type to DateTime in order to make Date calculations.

ADVANCED STAGE

- We classify every row as “No” (Investment_Stage = “seed”) and “Yes” in the remaining cases. As they where many investment stage for each company, if the company had advanced through Seed, we kept the value that had the earliest “Announced Date”. In the case it had only reached “Seed”, we stay with the oldest date.

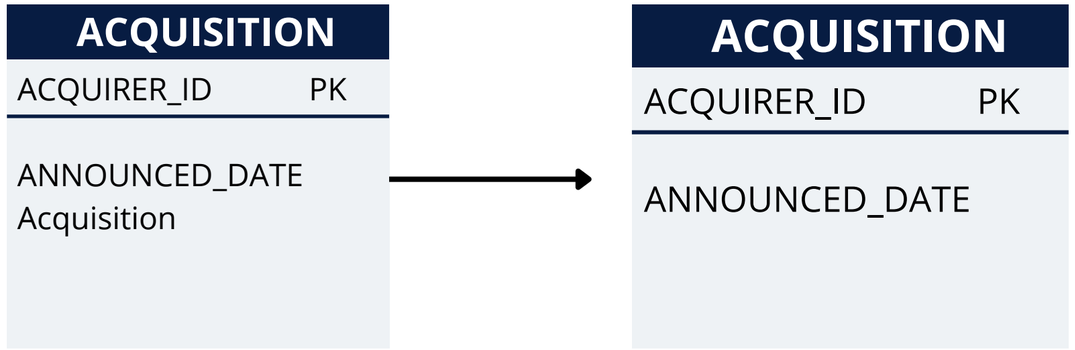
	A	B	C	D
1	COMPANY_ID	ANNOUNCED_DATE	INVESTMENT_STAGE	ADVANCED_STAGE
2	fly-now-pay-later	May 5, 2021	Early Stage Venture	Yes
3	fly-now-pay-later	May 18, 2020	—	Yes
4	fly-now-pay-later	May 18, 2020	Early Stage Venture	Yes
5	fly-now-pay-later	Sep 12, 2016	—	Yes
6	fly-now-pay-later	Jun 1, 2015	Seed	No
7	fly-now-pay-later	May 1, 2018	—	Yes
8	animoca-brands-corpor	May 13, 2021	—	Yes
9	animoca-brands-corpor	Nov 15, 2011	Early Stage Venture	Yes

→

	A	B	C	D
1	COMPANY_ID	ANNOUNCED_DATE	INVESTMENT_STAGE	ADVANCED_STAGE
2	%C3%81eron	2014-08-01 00:00:00	Early Stage Venture	Yes
3	-the-one-of-them-inc-	2011-10-03 00:00:00	Early Stage Venture	Yes
4	1-page	2012-04-01 00:00:00	Seed	No
5	10alike	2013-01-01 00:00:00	Seed	No
6	10k-2	2015-09-04 00:00:00	Seed	No
7	10sec	2013-10-31 00:00:00	Seed	No
8	133t-db8f	2016-03-11 00:00:00	Seed	No
9	16lab-inc-	2014-11-26 00:00:00	Seed	No

Cleaning & Transformation

03.ACQUISITION



Only 136 companies had an acquisition before reaching an Advanced Stage for the first time
(1.4%)

Announced Date

- Change the Data Type to DateTime in order to make Date calculations.

Acquirers

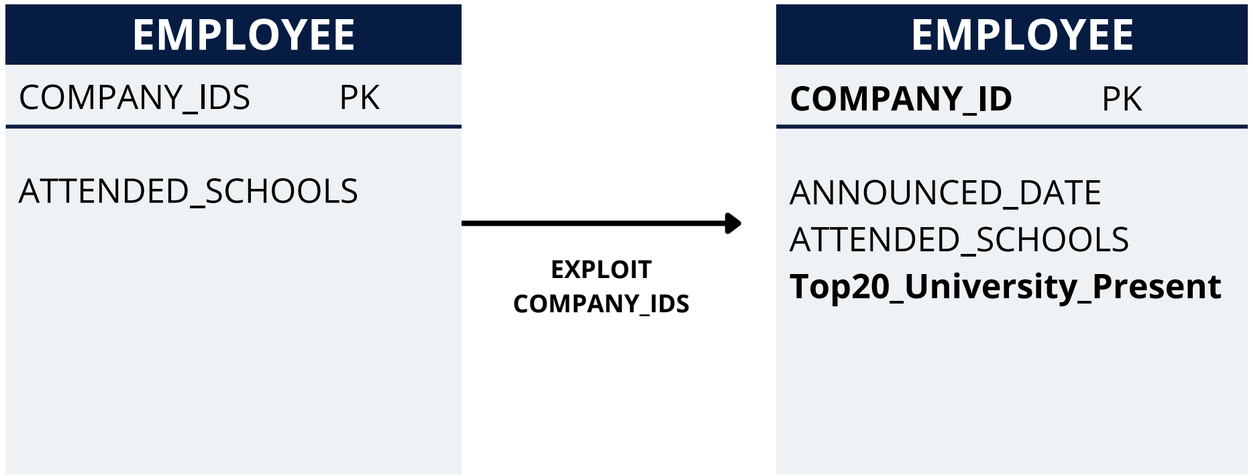
- Filtered Acquirers to include only companies that had made an acquisition prior to receiving their first investment at an ADVANCED STAGE.
- Resulted to be an insignificant portion of the total sample, thus not considered to build the model.

JOIN between ACQUISITION & INVESTMENT

- It was necessary to perform a Join between both tables to carry out the comparison.

Cleaning & Transformation

04.EMPLOYEE



ATTENDED_SCHOOLS

- Exploit “COMPANY_IDS” to have “COMPANY_ID” per row.

Top20_University_Present

- Creation of a list of the top 20 universities with all their possible variations, then each row was searched to see if one of these universities was mentioned. Finally, they were marked with "Yes" in the cases where they appeared and "No" in the remaining cases.

```
Top20_Universities = [
    "University of Oxford", "Oxford University", "MIT",
    "Stanford University", "Stanford Graduate School of Business", "Stanford",
    "Massachusetts Institute of Technology (MIT)", "Massachusetts Institute of Technology - MIT", "MIT",
    "Harvard University", "Harvard", "Cambridge College",
    "University of Cambridge", "Princeton University", "Caltech",
    "California Institute of Technology", "California Institute of Technology (Caltech)", "Imperial College London",
    "University of California, Berkeley", "University of California, Berkeley (UCB)", "UC Berkeley",
    "Yale University", "Yale", "ETH Zurich", "Tsinghua University",
    "Tsing Hua", "The University of Chicago", "University of Chicago",
    "Peking University", "Johns Hopkins University", "Johns Hopkins",
    "University of Pennsylvania", "Columbia University", "Columbia Business School",
    "University of California, Los Angeles (UCLA)", "UCLA", "University of California",
    "National University of Singapore (NUS)", "National University of Singapore", "Cornell University"
]
```

Cleaning & Transformation

05.NEWS

NEWS	
COMPANY_ID	PK
NEWS_DATE	
NEWS_TITLE	

→

NEWS	
COMPANY_ID	PK
NEWS_DATE	
NEWS_TITLE	
Classification	

NEWS_DATE

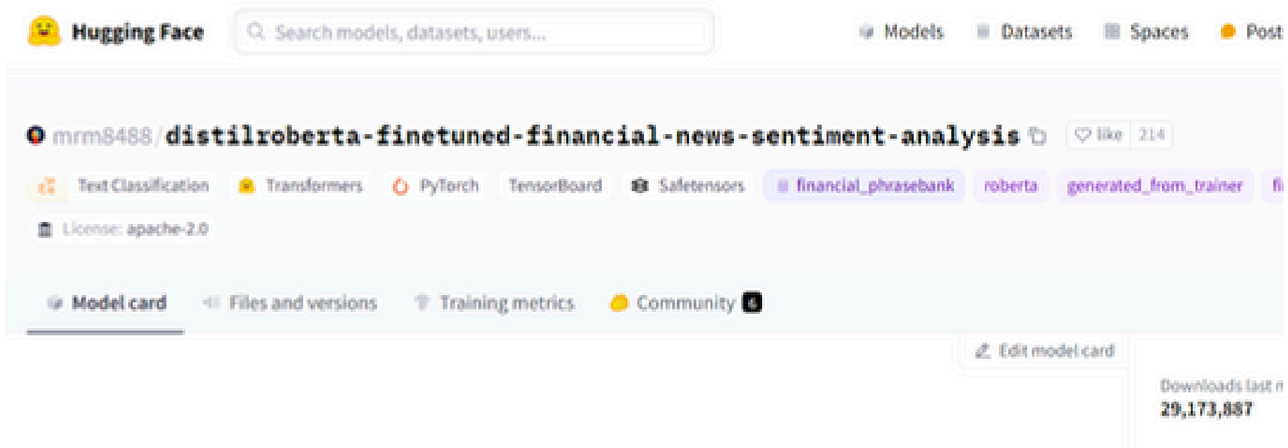
- Change the Data Type to DateTime in order to make Date calculations.

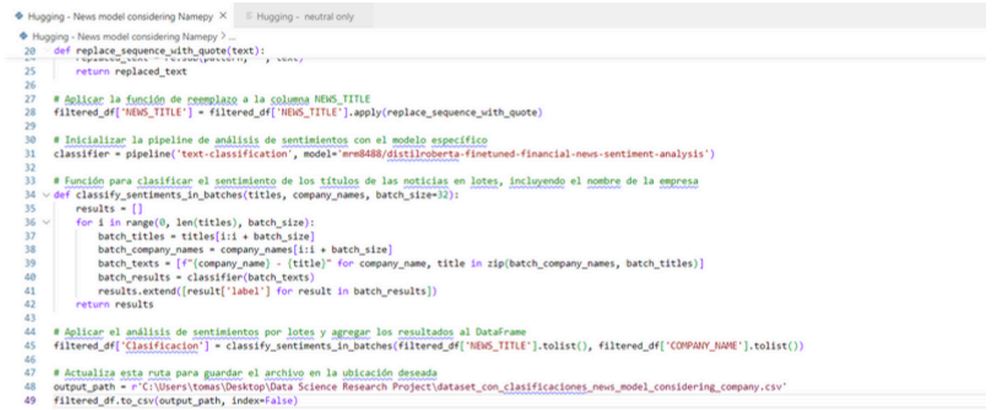
NEWS_TITLE

- We retained only the news headlines that included the company's name.

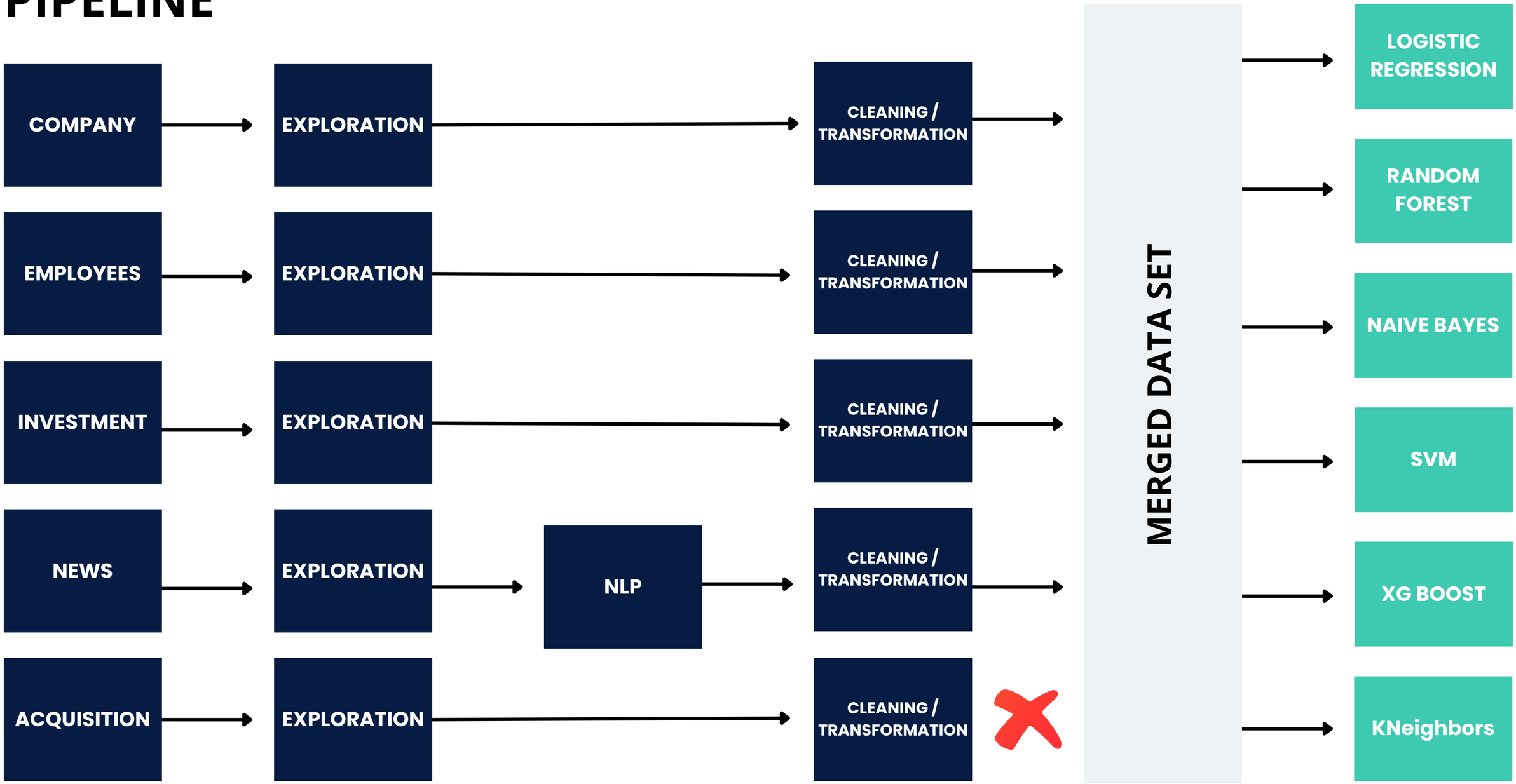
Classification

- Using API of model mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis, which specializes in the sentiment analysis of financial news, we managed to classify the news into "Positive," "Negative," and "Neutral" based on the impact on the company.



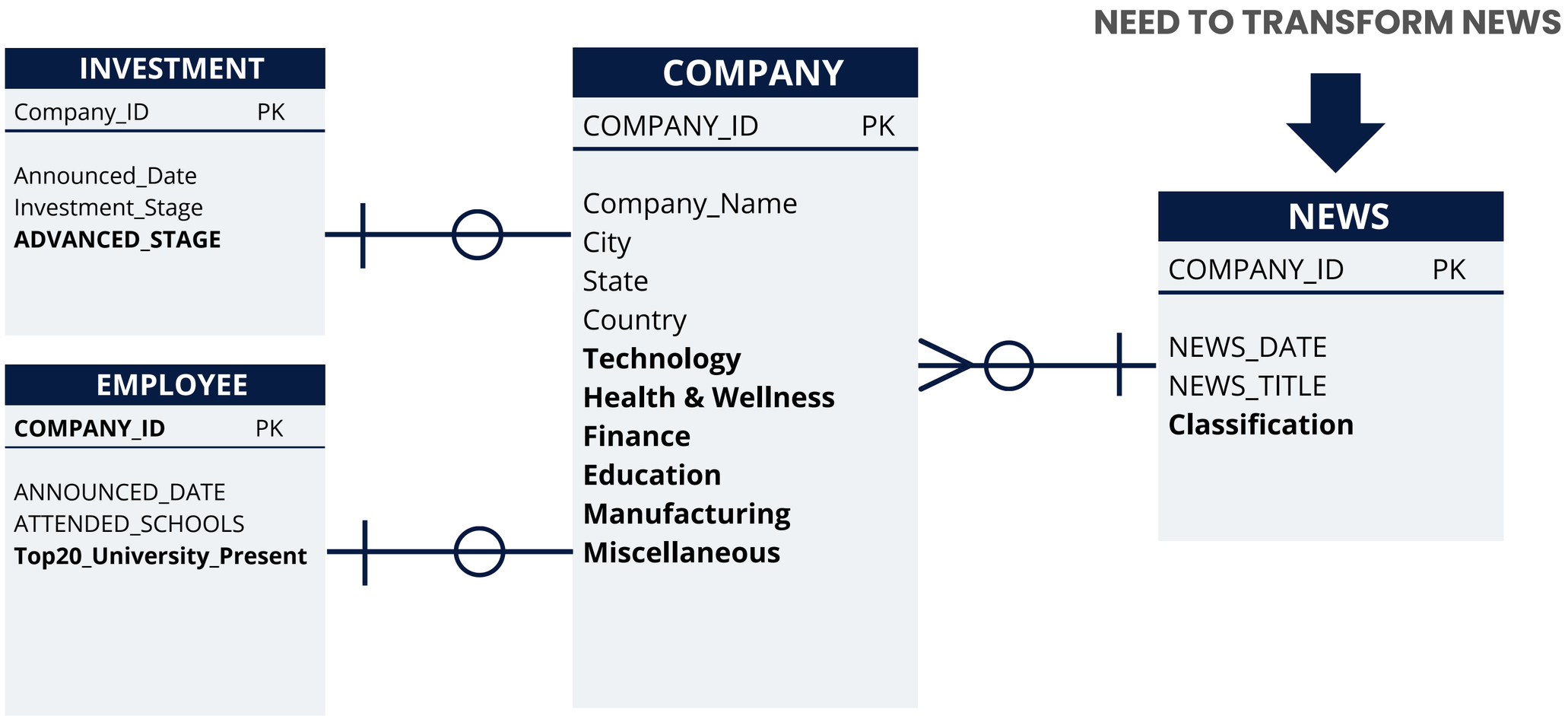


PIPELINE


COMPANY
EMPLOYEES
INVESTMENT
NEWS
ACQUISITION
EXPLORATION
EXPLORATION
EXPLORATION
EXPLORATION
EXPLORATION
NLP
CLEANING / TRANSFORMATION
CLEANING / TRANSFORMATION
CLEANING / TRANSFORMATION
CLEANING / TRANSFORMATION
CLEANING / TRANSFORMATION
MERGED DATA SET
LOGISTIC REGRESSION
RANDOM FOREST
NAIVE BAYES
SVM
XG BOOST
KNeighbors

Cleaning & Transformation

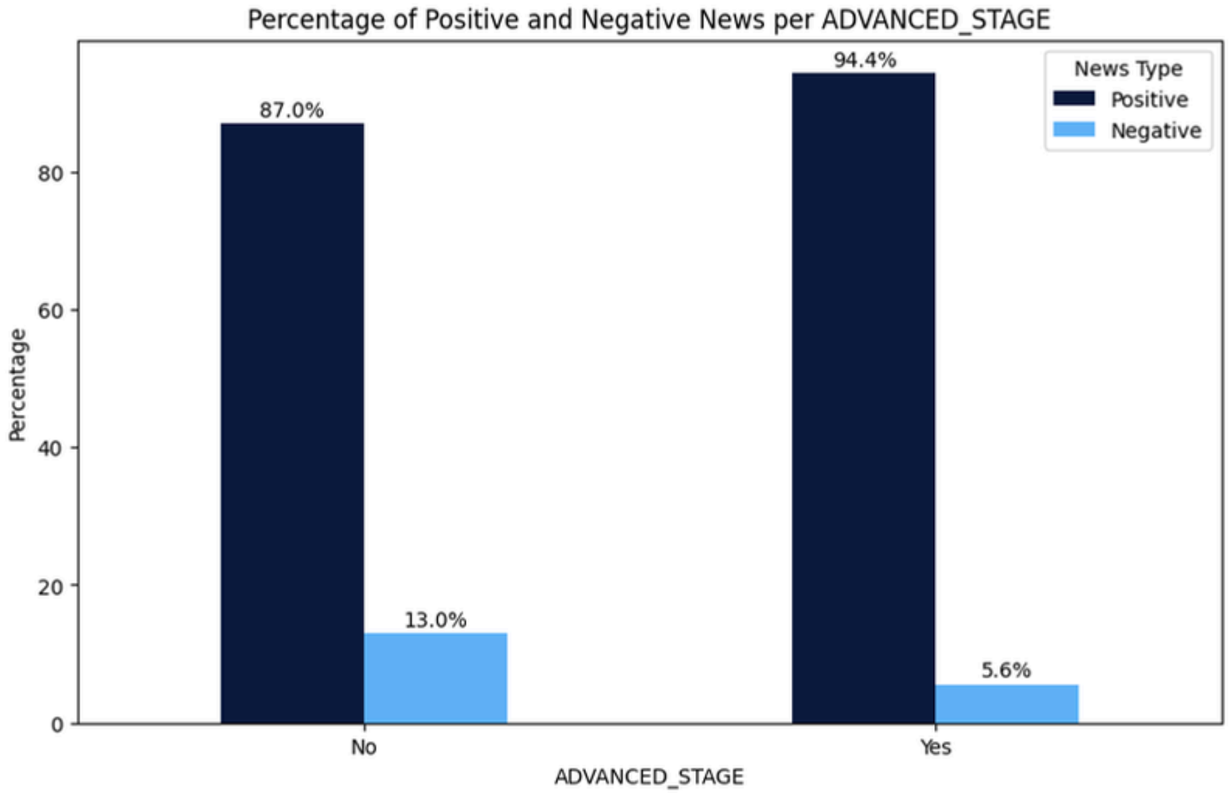
The Merge



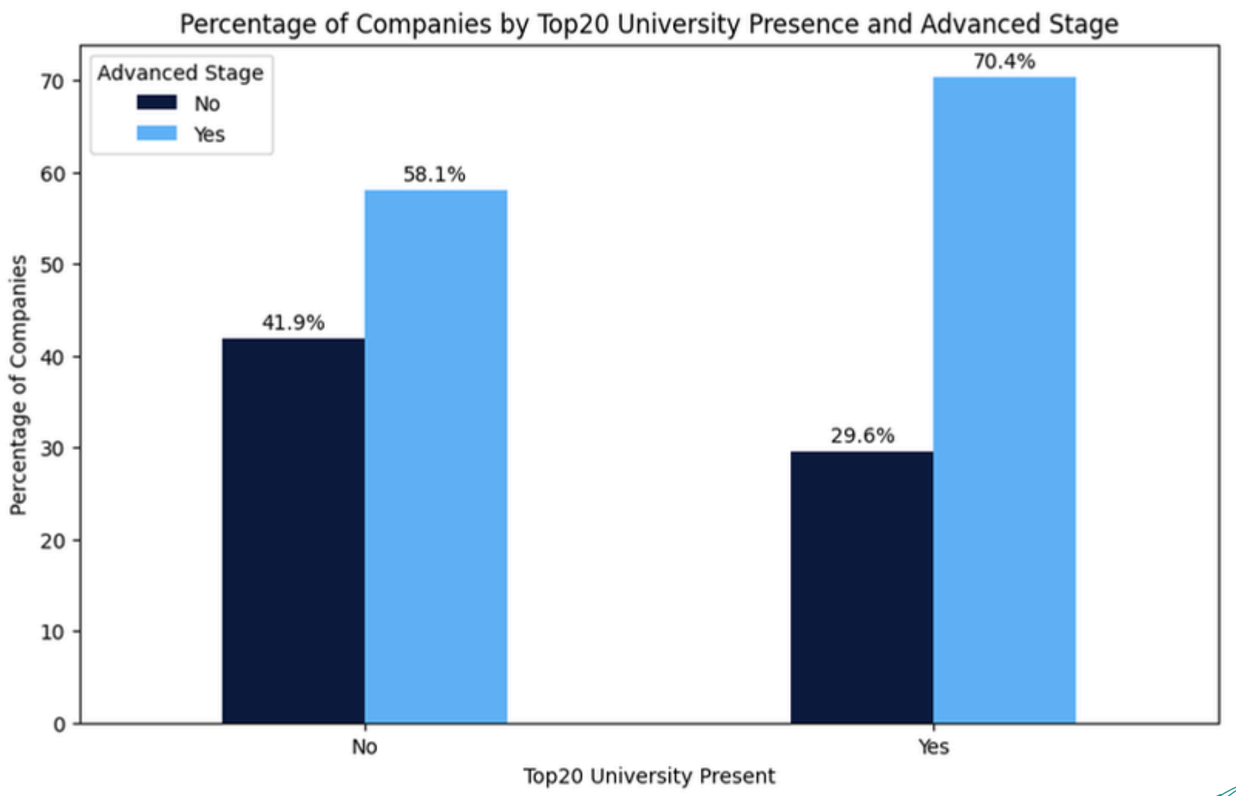
Merged Dataset

COMPANY	
Company_ID	**PK**
ADVANCED_STAGE	
Positive	
Negative	
Neutral	
city	
state	
country	
Top20_University_Present	
Technology	
Health & Wellness	
Finance	
Education	
Manufacturing	
Miscellaneous	

Impact of News on ADVANCED_STAGE

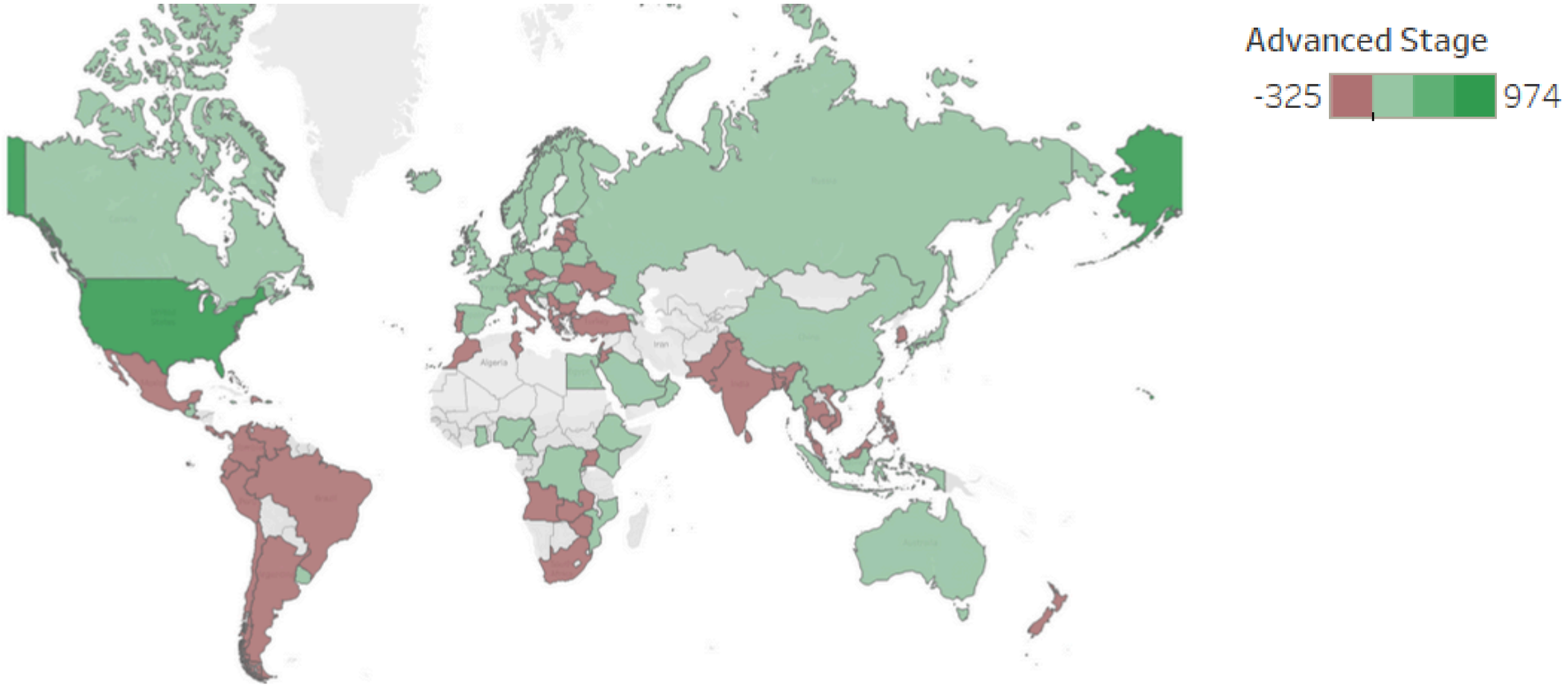


Impact of Employee Education on ADVANCED_STAGE



Data Analysis

LATE-STAGE INVESTMENT SUCCESS VS. FAILURE BY COUNTRY



Data Analysis

VARIABLES

Thirteen variables were considered for the analysis, three of which are numeric and the rest are text.

DEPENDENT VARIABLES

Advanced_Stage

INDEPENDENT VARIABLES

NUMERICAL

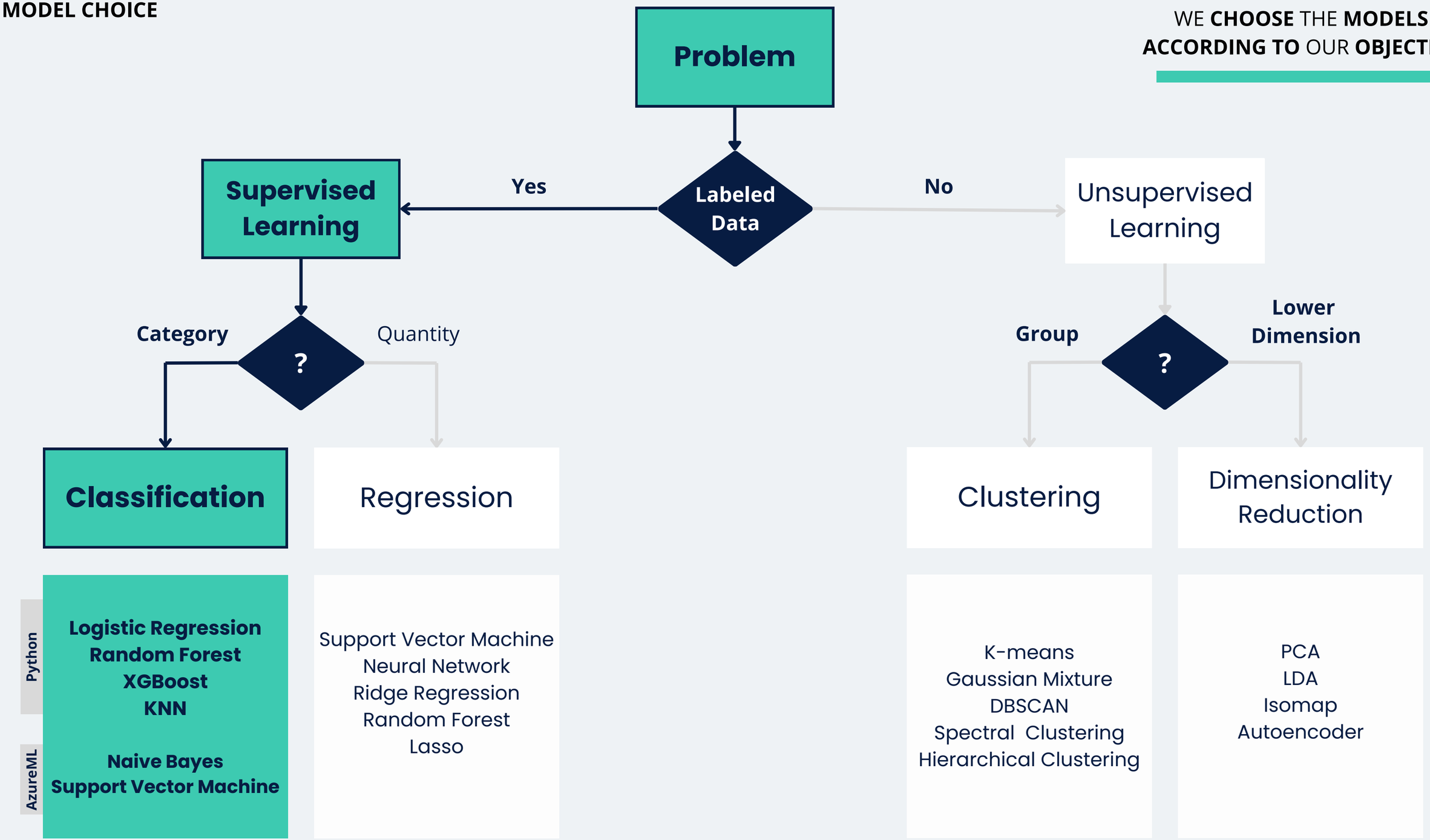
Positive
Negative
Neutral

TEXT

City
State
Country
Top20_Univesity_Present
Technology
Health & Wellness
Finance
Education
Manufacturing
Miscellaneous

MODEL CHOICE

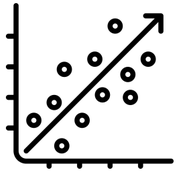
WE **CHOOSE** THE **MODELS**
ACCORDING TO OUR **OBJECTIVE**



MODEL EVALUTATION

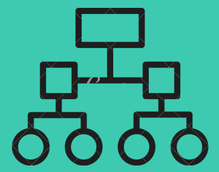
WE **CHOOSE** THE **RIGHT PERFORMANCE METRICS, GIVEN** OUR **OBJECTIVE** & **MODEL**

REGRESSION



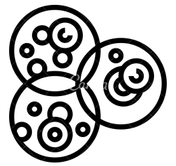
MAE
MSE
RMSE
R-Squared

CLASSIFICATION



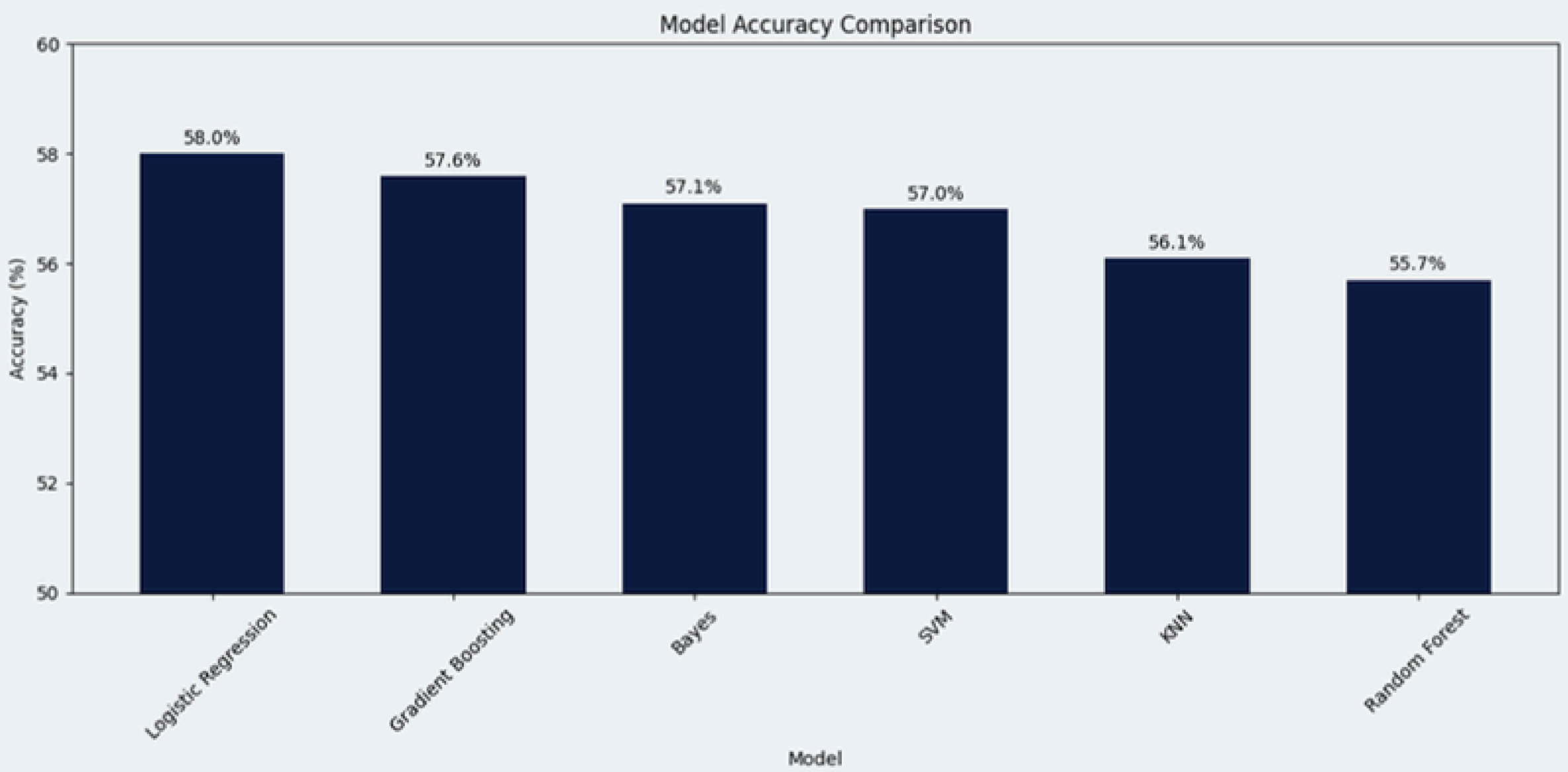
Accuracy
Precision & **Recall**
F-Score
AUC-ROC
Confusion Matrix
Gini Coefficient

CLUSTERING

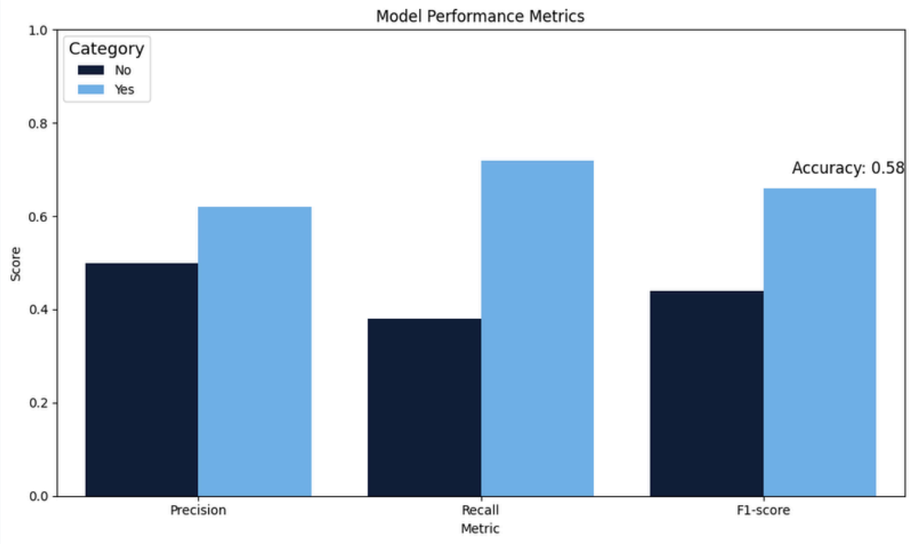


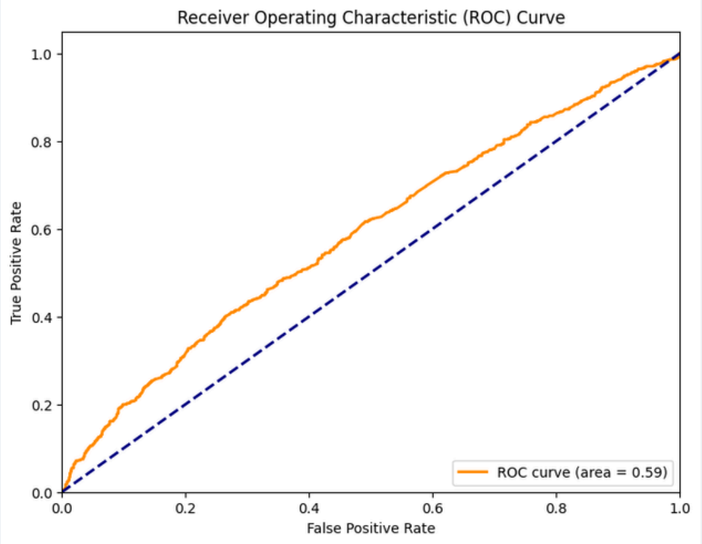
Rand Index
Mutual Information
Silhouette Score

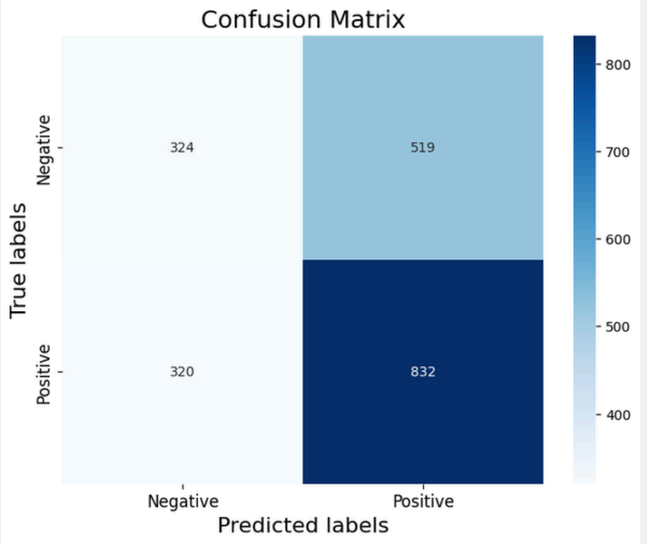
WE COMPARE ACCURACIES AND CHOOSE THE BEST MODEL



LOGISTIC REGRESSION - PERFORMANCE







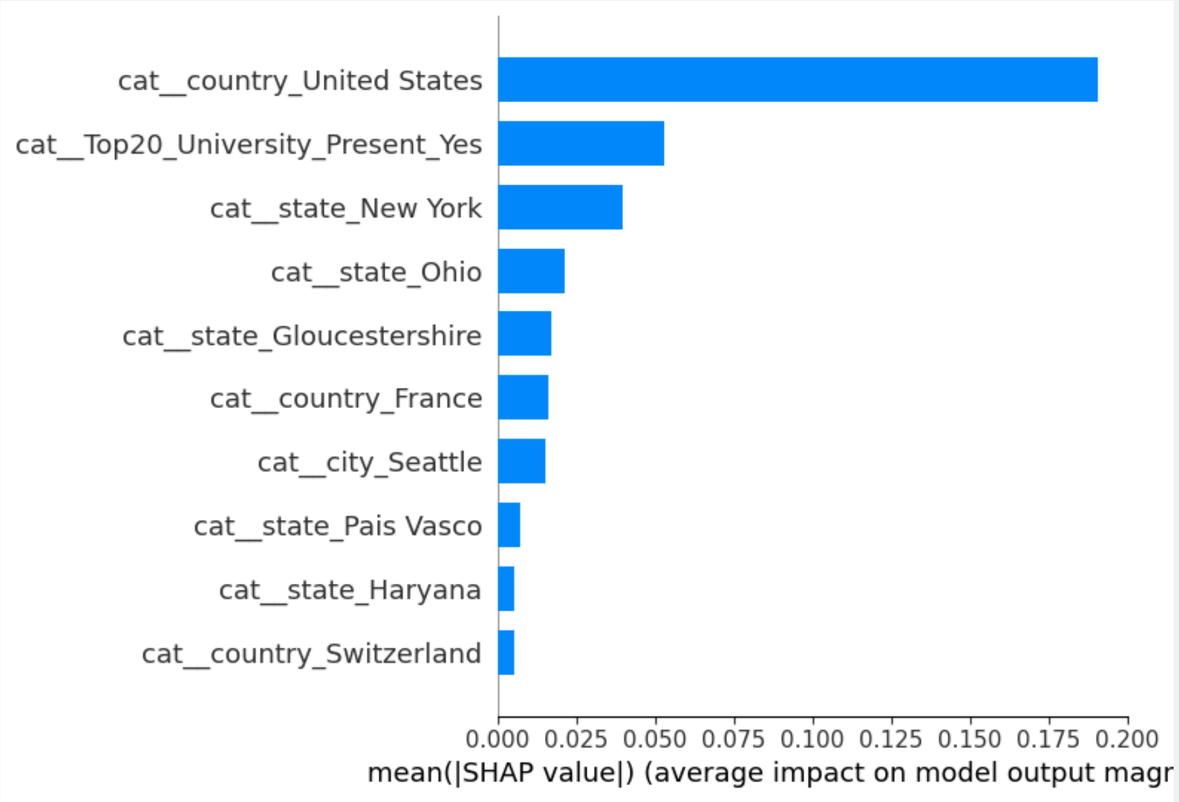
As we can see, the **model is better at predicting success for positives than for negatives.**

The ROC curve shows an AUC of 0.59, suggesting the **model can** reasonably **differentiate between classes.**

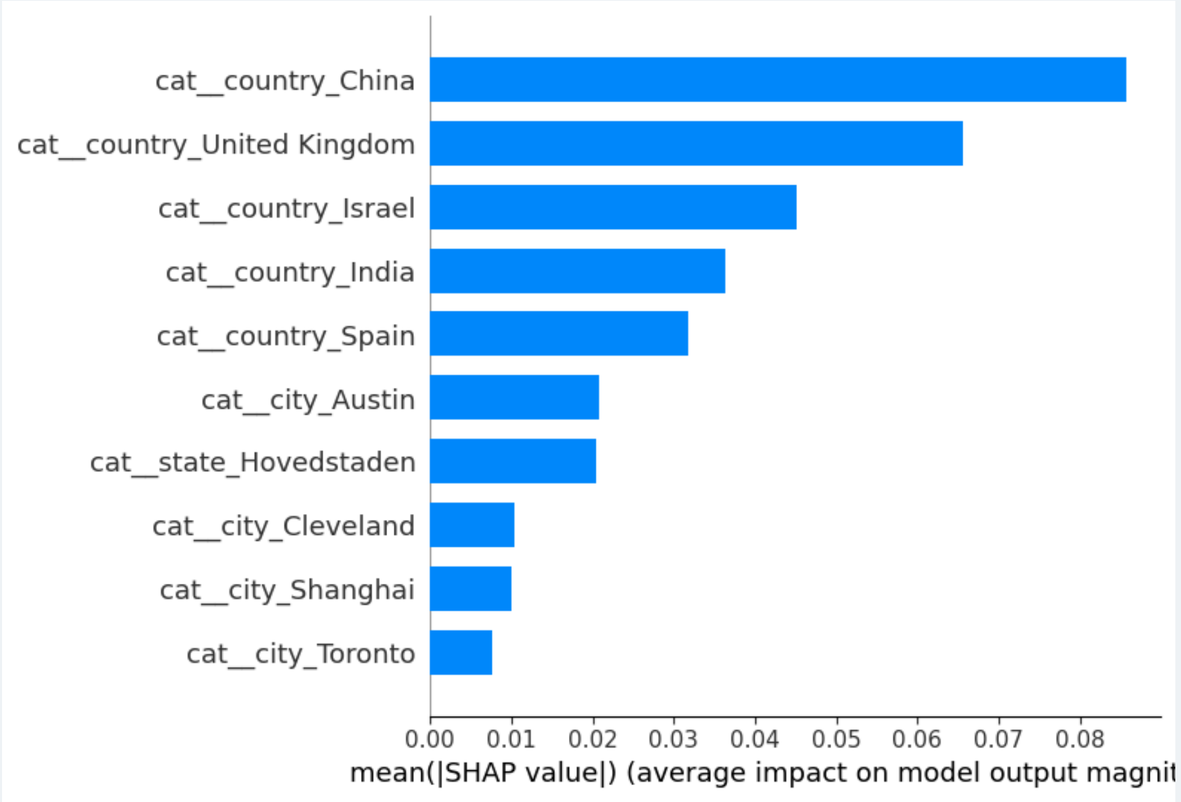
It predicted 'Negative' 324 times and 'Positive' 832 times. However, there were 519 **false positives** and 320 **false negatives**.

LOGISTIC REGRESSION - RESULTS

Top 10 Positive SHAP Values



Top 10 Negative SHAP Values



We calculate the **mean absolute SHAP** - a measure of the average impact of a feature on the model's output magnitude. “**Top Positive SHAP Values**” indicate features that significantly **increase the model's predicted probability of the positive class.** "**Top Negative SHAP Values**" denote **features** that substantially **lower it**.

Conclusions & Key Takeaways

Positive Impact Factors

According to SHAP, the variables that most positively affect the dependent variable are **being in the United States, having employees from the top 20 universities, and being in New York state.**

Negative Impact Factors

The variables that most negatively affect are being in China, Israel, or the UK.

Positive News Correlates with Late-Stage Investment

We identified that **companies reaching** a **late-stage investment** generally **have a higher ratio of positive news** than those that do not.

Improvements

Some possible improvements to the model could include: **incorporating external data** that affects startups, **performing sentiment analysis** based on the **entire article** and including more data.

This could improve the accuracy of our most performant model, which was 58%.