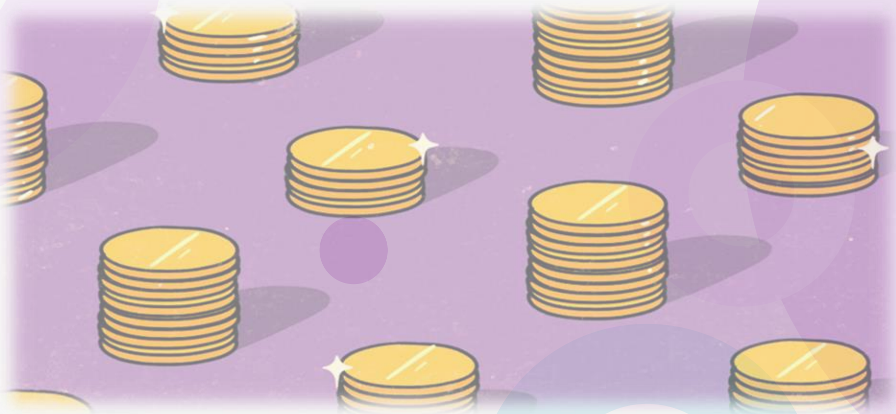
Investment Data Analysis

- Minija Kannath



Contents

- ★ Introduction.
- ★ Methodology.
- ★ Process workflow.
- ★ Results
- ★ Conclusion
- ★ Future opportunities
- ★ Appendix

Introduction

I am a data analyst working for an asset management Company

Representing Organization

TIAA
GLOBAL ASSET MANAGEMENT

Target Audience

The CEO and board of directors of TIAA

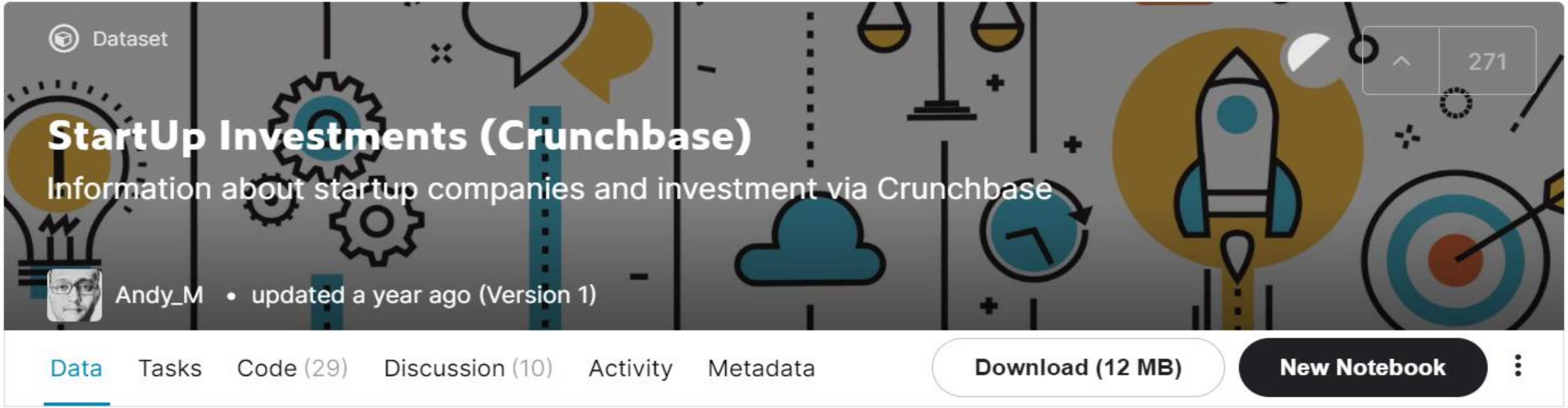
Business Objective

TIAA wants to make investments in a few companies. The CEO of TIAA wants to understand the global trends in investments and to build a prediction model to classify the status of the companies so that she can take the investment decisions effectively

Business Constraints

- TIAA wants to invest between 6 to 16 million USD per round of investment .
- TIAA wants to invest only in English-speaking countries.

Data Collection



This data is collected from Data world and Kaggle. It is originally from Crunchbase database. Which is a collection of information about startup companies and investments . This data set consists of more than 40,000 records and 16 columns

Data Collection

funding_total_usd

Total funds received

Funding round

The rounds of funding that start-ups go through to raise capital.

Last_funded_year

The last year the company received the fund

Funding Type

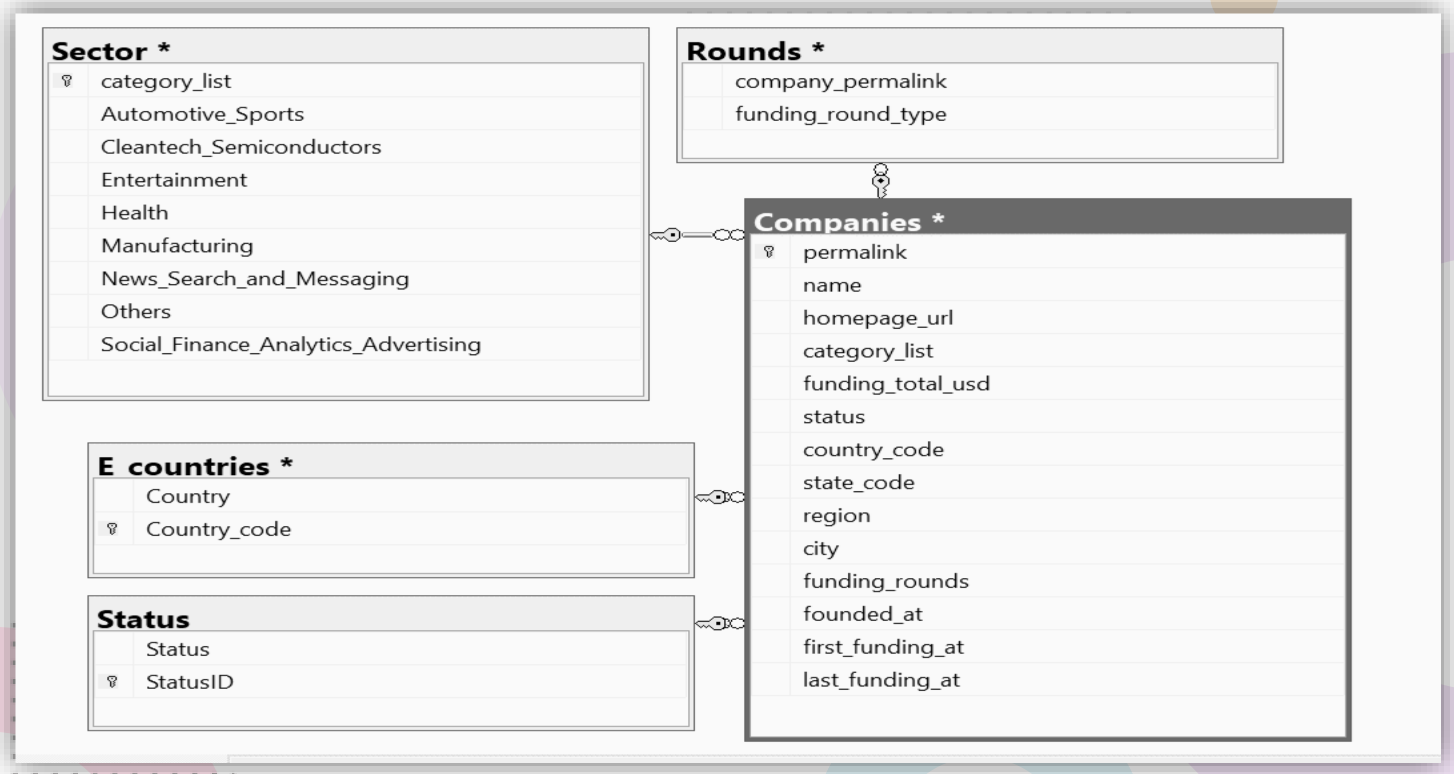
The type of funding company received

Founded_at

The year company was founded

Database schema

Creating Schema in SQL:



Methodology

Model

Baseline Model: Naive Bayes
Alternative Model: Logistic Regression
Random Forest, Decision tree

Metrics

Classification Report(precision,recall,f1-score)
Confusion Matrix

Tools

Process workflow

★ Data Cleaning/Data Analysis
★ Data Preparation/Feature Engineering
★ Data Transformation
★ ML Model Training and Evaluation

Data cleaning

- For the business objectives the column homepage_url and name is not used. These columns are dropped.
- State_code, region and city are highly correlated with country_code.So these columns are dropped.

category_list	funding_total_usd	status	country_code	funding_rounds	founded_at	first_funding_at	last_funding_at	funding_round_type
Curated Web	NaN	0	USA	1	2010-01-01	2014-07-24	2014-07-24	venture
Marketplaces	41250.0	0	HKG	1	None	2014-07-01	2014-07-01	undisclosed
Finance Technology	2000000.0	0	CHN	1	2007-01-01	2008-03-19	2008-03-19	venture
Clean Technology	762851.0	0	CAN	2	1997-01-01	2009-09-11	2009-12-21	venture
Clean Technology	762851.0	0	CAN	2	1997-01-01	2009-09-11	2009-12-21	seed

Data cleaning

summary statistics of funding_total_usd:

```
count     9.589700e+04
mean      3.520118e+07
std       2.827248e+08
min       1.000000e+00
25%       8.250000e+05
50%       5.000000e+06
75%       2.385000e+07
max       3.007950e+10
Name: funding_total_usd, dtype: float64
```

Summary of the missing values (column-wise) and fraction of NaNs:

```
permalink            0.00
name                 0.00
category_list        0.00
funding_total_usd    0.00
status               0.00
country_code         4.42
funding_rounds       0.00
founded_at           0.00
first_funding_at     0.06
last_funding_at      0.00
funding_round_type   0.00
dtype: float64
```

Data cleaning

Column wise analysis of country code column:

Displaying frequencies of each category

```
USA     62627
GBR      5073
CAN      2642
CHN      2153
IND      1663
         ...
MKD         1
MNE         1
QAT         1
PSE         1
ZWE         1
Name: country_code, Length: 134, dtype: int64
```

Removing the rest of the missing values from country_code and first funding

```
permalink              0
name                   0
category_list          0
funding_total_usd      0
status                 0
country_code           0
funding_rounds         0
founded_at             0
first_funding_at       0
last_funding_at        0
funding_round_type     0
dtype: int64
```

Data Preparation/Feature Engineering

Dimensionality Reduction:

Any funding type having less than 1000 data points are tagged as "others" funding type.

Hot encoding:

After that hot encoding is used to transform categorical feature “funding type into binary form of representation

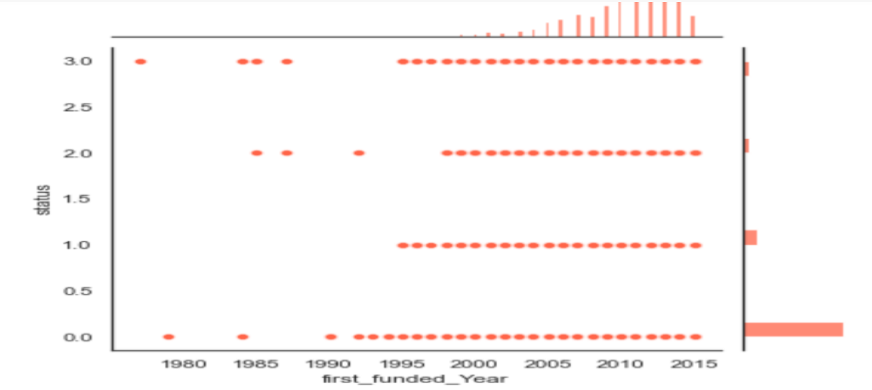
Label encoding:

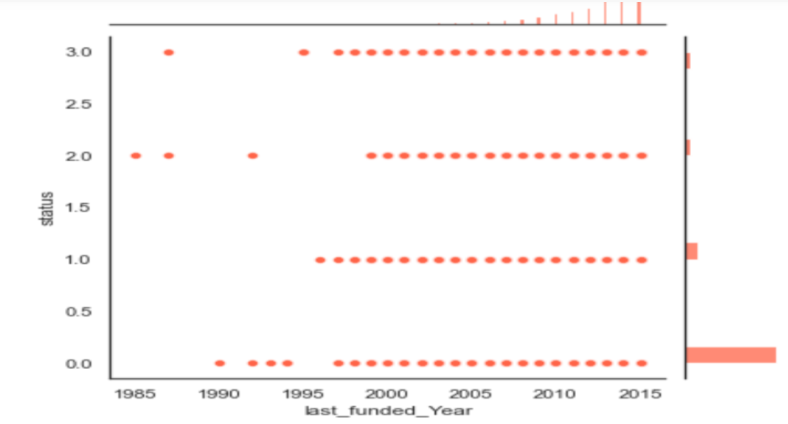
Label encoding is used to transform the categorical feature ‘country_code’ into numerical form of representation

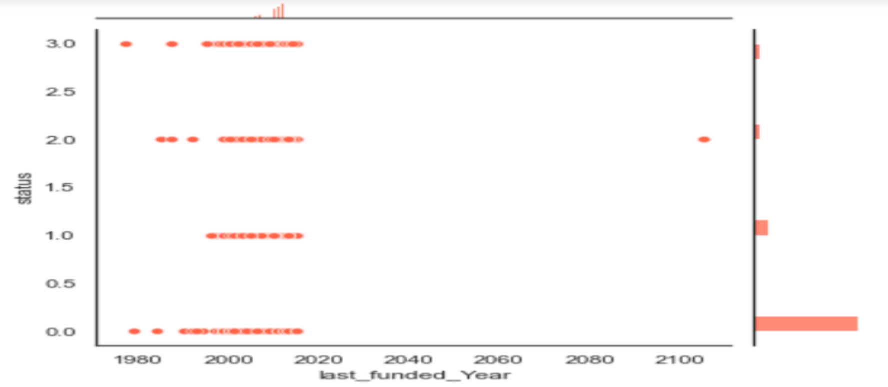
country_code
CHN
CAN
CAN
USA
USA

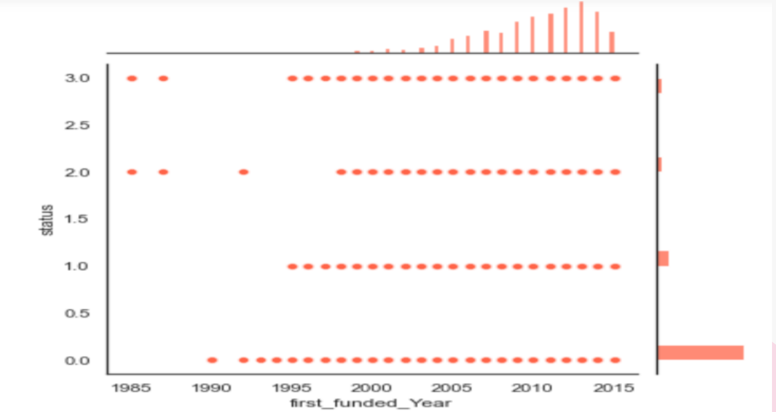
country_code
20
17
17
118
118

Data Preparation/Outlier removal



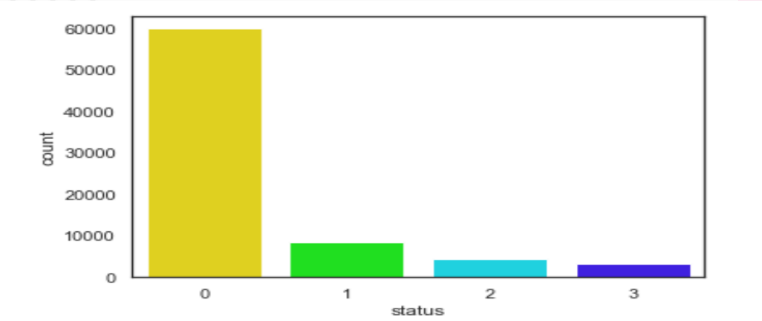






Data Analysis

Explore target value distribution:
It is an imbalanced data set



To impose the constraint:
investment amount should be between 6 and 16 million USD.

We will choose the funding type such that the average investment amount falls in this range.

```
funding_round_type
private_equity           56063148.0
venture                  15015744.5
debt_financing           10028500.0
others                    7795780.0
convertible_note          1258032.0
angel                     1000000.0
grant                     1000000.0
seed                       847708.0
non_equity_assistance      800000.0
equity_crowdfunding        195649.0
Name: funding_total_usd, dtype: float64
```

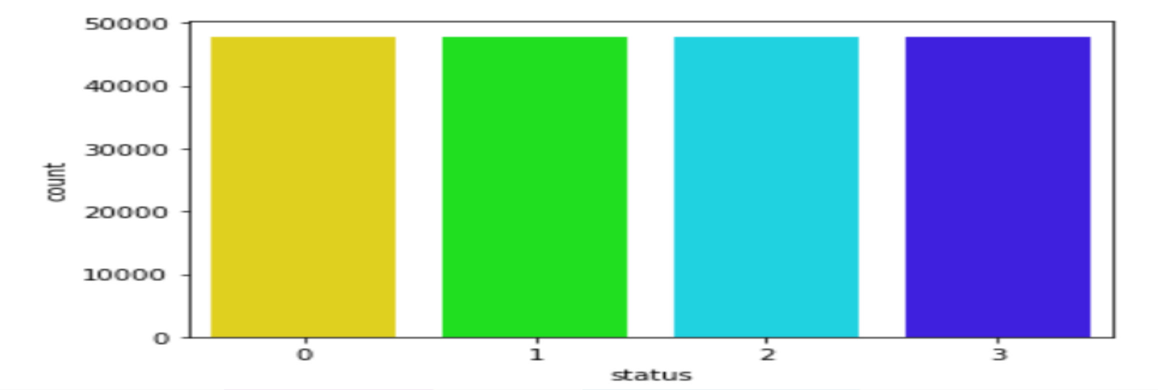
Data Transformation

Split the data into training and testing datasets:
test_size = 0.2

Data Normalization:
Min Max Scaler: Naive Bayes

Robust Scaler: Logistic regression, Decision Tree & Random Forest

Oversampling the train data set using SMOTE:



Machine Learning Model

MODEL

Baseline Model: Naive Bayes

Alternative Model: Logistic Regression ,Decision tree, Random Forest

RFE method for Feature selection:

Since the stability of RFE depends on type of model .I ran RFE with all four models

```
[ True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True]
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
```

Machine Learning Model Training & Evaluation

Baseline Model: Naïve Bayes

Hyperparameter: (fit_prior= True, alpha =1)

```
The accuracy is: 40.9%
```

```
Average Precision score: 0.4893249261525952
Average Recall score    : 0.4535846189878576
Average F1 score        : 0.4575571882881284
```

Alternative Model: Logistic Regression

Hyperparameter: (solver='lbfgs', max_iter=1000)

```
The accuracy is: 59.0%
```

```
Average Precision score: 0.4775182941289873
Average Recall score    : 0.30434246236796l
Average F1 score        : 0.311773754066795 2
```

Machine Learning Model Training & Evaluation

Alternative Model: Decision Tree

Hyperparameter: (min_samples_split’= 3, min_samples_leaf:=9, max_depth= 5)

```
The accuracy is: 88.3%
```

```
Average Precision score: 0.8944955809884085
Average Recall score    : 0.8925268817204302
Average F1 score        : 0.8927414709604884
```

Alternative Model: Random Forest

Hyperparameter: (n_estimators=10, min_samples_split= 5, max_depth=28)

```
The accuracy is: 90.0%
```

```
Average Precision score: 0.9513638553682598
Average Recall score    : 0.949911436337028
Average F1 score        : 0.9500340786278793
```

Finding the best model and Hyperparameter tuning

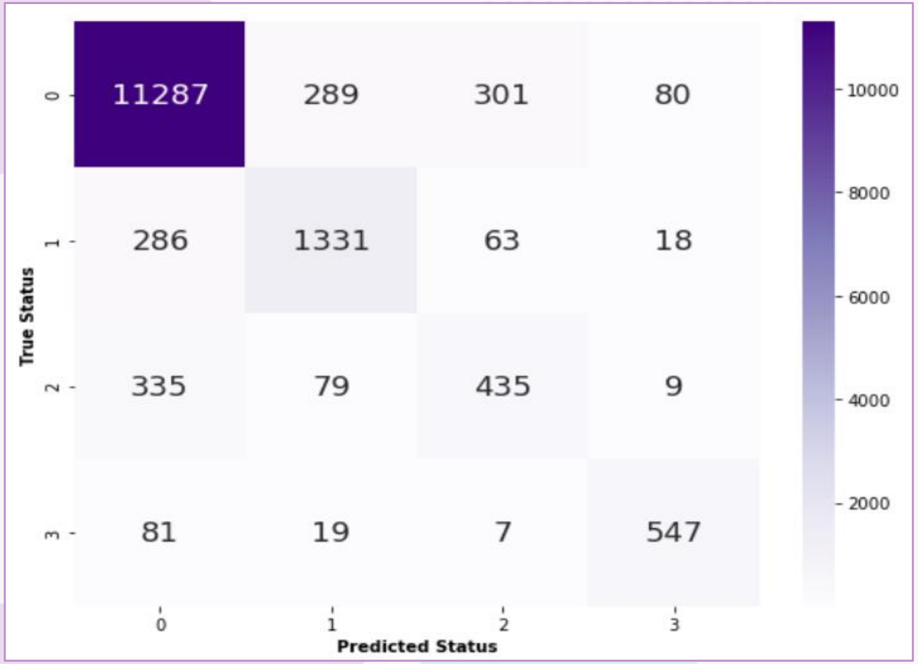
	model	best_score	best_params
0	Random_forest	0.912020	{'n_estimators': 225, 'min_samples_split': 2, 'max_depth': 28}
1	DecisionTreeClassifier	0.859221	{'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 15}
2	Multinomial Naive Bayes	0.788027	{'fit_prior': True, 'alpha': 1}

	model	best_score	best_params
0	logistic_regression	0.792958	{'C': 10}

Results

Confusion Matrix for the Most accurate model Random Forest

open
acquired
closed
ipo



open
acquired
closed
ipo

Model deployment using Flask

Company Status Prediction

Funding_total_usd
Country: --Please choose an option--
Funding_rounds
YearFounded:
FirstFundedYear:
LastFundedYear:

Select funding type

Angel: No
ConvertibleNote: No
DebtFinancing: No
EquityCrowdfunding: No
Grant: No
NonEquityAssistance: No
PrivateEquity: No
Seed: No
Venture: No

Select Sector

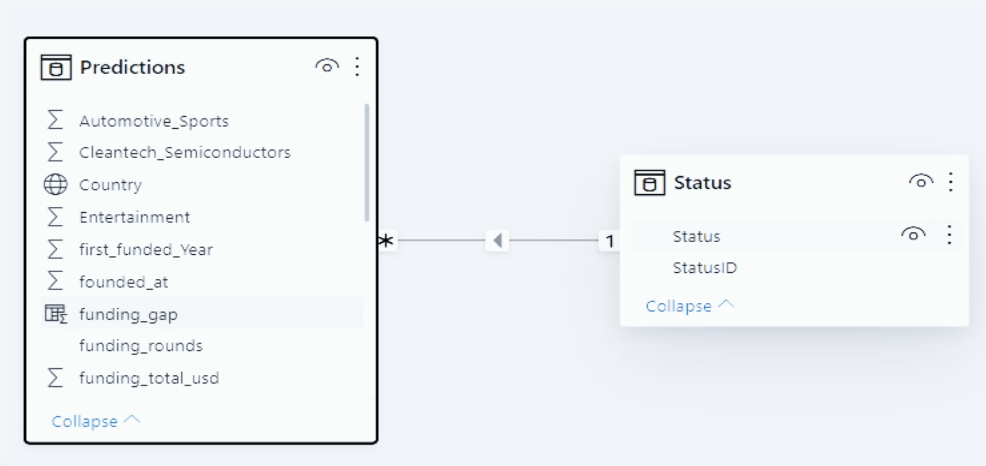
The status is: Operating

Deploying Flask App to Heroku

Link to my app: https://companystatusprediction.herokuapp.com

Predictive analysis

- ★ TIAA business constraints are applied to the test data
- ★ Predictions are obtained from this constraint dataset
- ★ The inputs and Predictions are inserted into SQL database as Predictions table.
- ★ Predictions table is connected to power BI .
- ★ New columns and measures are created using DAX formula and IF function.
- ★ Different charts are plotted using power BI.



Findings

Potential success rate of investment is high in South Africa (27%) followed by Ireland, Singapore, India, United states and United Kingdom.

Potential success rate of investment is high in Health sector (19%) followed by manufacture, Cleantech semiconductors.

South Africa has only one sector with potential success rate for investment. (Entertainment)
Australia and Singapore are the two countries with a high success rate in health sector.

Manufacturing sector investment is more successful in United kingdom followed by Canada.

Cleantech semiconductors sector investment is predicted as more successful in Ireland and Singapore.

The years of operation of the companies have a positive correlation on investment success.

Conclusion

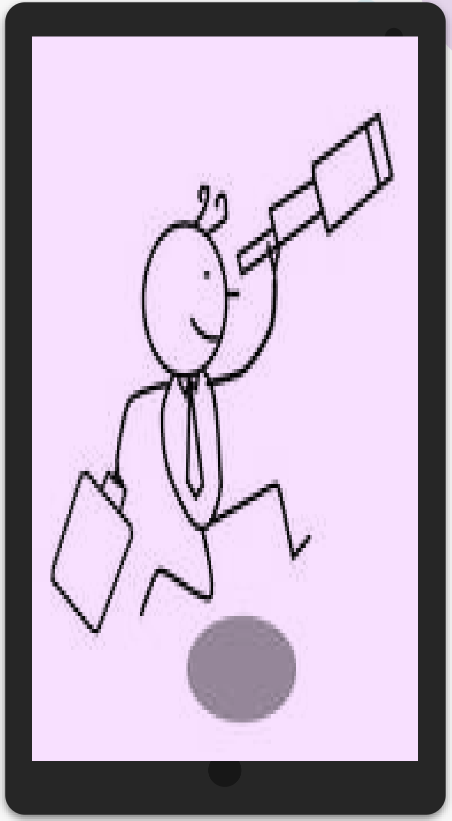
Sector	South Africa	Ireland	Singapore	India
Best	Entertainment	Automotive sport	Entertainment	Social Finance Analytics Advertising
Second Best		Cleantech Semiconductors	Health	Entertainment
Third Best		Entertainment	Automotive sport	Manufacturing

Future Opportunities

In future :

I will try to get my data from CrunchBase API so that I can update it if I want to.

I will use Principal Component Analysis (PCA) and try to improve the classification performance to build a more accurate model



Thank **you!**

https://data.world/fiftin/crunchbase-2015
https://www.kaggle.com/arindam235/startup-investments-crunchbase
https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population