Predicting Startup Status

William Fang

15/06/2020

Contents

1 Introduction

Kaggle makes available a dataset "StartUp Investments (Crunchbase)", which provides the status of a company (acquired, operating, or closed), investment series, and other data. The aim of this report is to develop a model that can predict the status of a startup, based on the features available in this dataset.

In the modern technology landscape * Startups

- Dataset available from Kaggle "StartUp Investments (Crunchbase)" (see https://www.kaggle.com/arindam235/startup-investments-crunchbase/)

Crunchbase is an online platform for @@@@ Section that describes the dataset and variables, and summarizes the goal of the project and key steps that were performed.

@@@@ Problem, can startup status (acquired, operating, closed) be predicted?

1.1 Variables

The dataset contains 39 variables:

```
 [1] "permalink"          "name"               "homepage_url"
 [4] "category_list"      "market"             "funding_total_usd"
 [7] "status"             "country_code"       "state_code"
[10] "region"             "city"               "funding_rounds"
[13] "founded_at"         "founded_month"      "founded_quarter"
[16] "founded_year"       "first_funding_at"   "last_funding_at"
[19] "seed"               "venture"            "equity_crowdfunding"
[22] "undisclosed"        "convertible_note"   "debt_financing"
[25] "angel"              "grant"              "private_equity"
[28] "post_ipo_equity"    "post_ipo_debt"      "secondary_market"
```

```
[31] "product_crowdfunding" "round_A"              "round_B"
[34] "round_C"              "round_D"              "round_E"
[37] "round_F"              "round_G"              "round_H"
```

1.1.1 Market and Categories

The `market` variable is string representing the main market the startup is targeting. `category_list` contains one or more categories the startup belongs to. Each category is separated by a |, and there is no specific ordering to the list.

Table 1: Sample category_list Valuse

x
|Entertainment|Politics|Social Media|News|
|Games|
|Publishing|Education|
|Electronics|Guides|Coffee|Restaurants|Music|iPhone|Apps|Mobile|iOS|E-Commerce|
|Tourism|Entertainment|Games|
|Software|

1.1.2 Name

Startup name, there may be some sort of predictive power here. For example, a catchy or memorable startup name might attract more attention and lead to more investment, consumer interest etc.

1.1.3 Location

There are four variables that describe location. `country_code` which is a 3 character string, `state_code` a 2 character string for companies within the US, `region` and `city` which are strings.

1.1.4 Status

The focus of this report is to create a model for predicting status. The variable `status` has the following values:

Table 2: Status Values

x
acquired
operating
NA
closed

There are companies with a missing status value, these will be removed as they can't be used.

1.1.5 Date

founded_at founded_month

" "funding_total_usd" [8] "" "" "funding_rounds"
[15] "founded_quarter" "founded_year" "first_funding_at" "last_funding_at" "seed" "venture" "equity_crowdfunding" [22] "undisclosed" "convertible_note" "debt_financing" "angel" "grant" "pri-

vate_equity" "post_ipo_equity"
[29] "post_ipo_debt" "secondary_market" "product_crowdfunding" "round_A" "round_B" "round_C" "round_D"
[36] "round_E" "round_F" "round_G" "round_H"

1.1.6 Ignored Variables

The following variables will be ignored, as they are very unlikely to have any predictive power. `permalink`, which is a hyperlink from the Techcrunch data, and `homepage_url`, which is the company's web page.

2 Analysis

@@@@ Turn status, country, state etc into factors. Dates from strings to datetime.

@@@@ What to use for measuring accuracy/loss function?

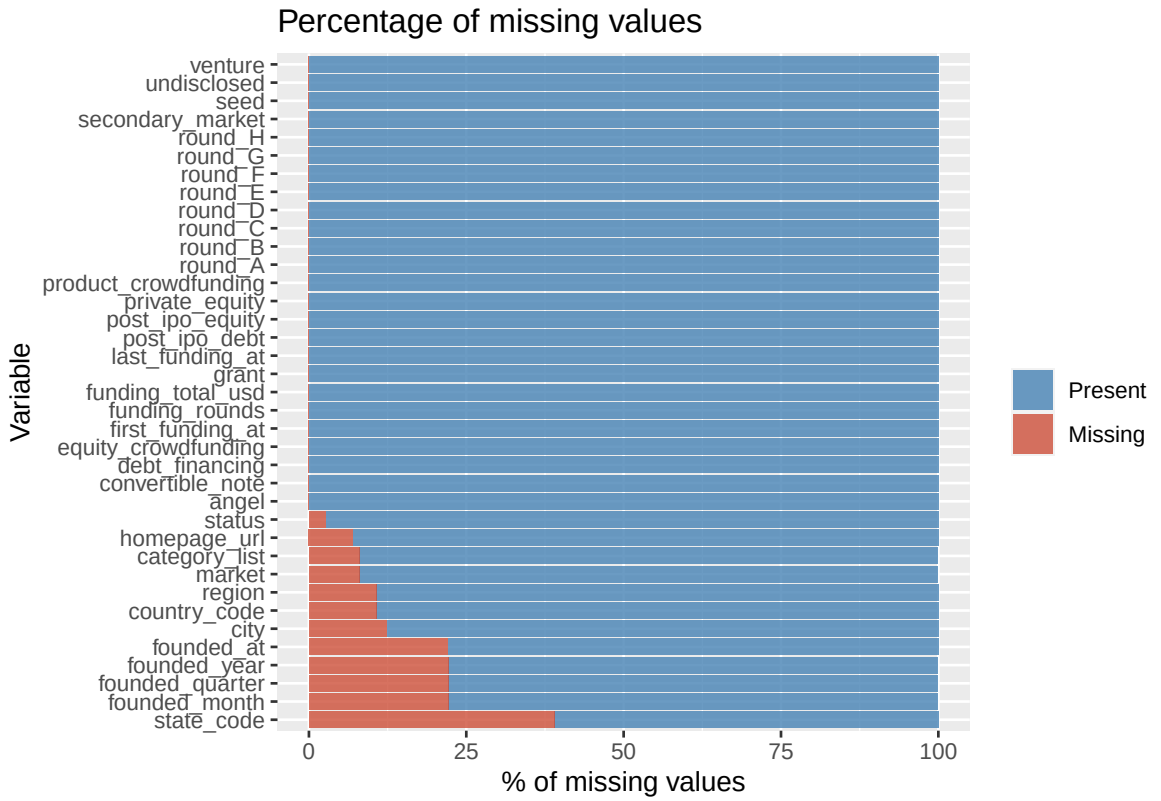
2.1 Missing Values

@@@@ Refer to (Laufer, n.d.)

Table 3: Variables With Missing Values Counts

variable	count
state_code	19172
founded_month	10909
founded_quarter	10909
founded_year	10909
founded_at	10838
city	6088
country_code	5247
region	5247
market	3959
category_list	3951
homepage_url	3441
status	1310
angel	1
convertible_note	1
debt_financing	1
equity_crowdfunding	1
first_funding_at	1
funding_rounds	1
funding_total_usd	1
grant	1
last_funding_at	1
post_ipo_debt	1
post_ipo_equity	1
private_equity	1
product_crowdfunding	1
round_A	1
round_B	1
round_C	1
round_D	1

variable	count
round_E	1
round_F	1
round_G	1
round_H	1
secondary_market	1
seed	1
undisclosed	1
venture	1



@@@@ section that explains the process and techniques used, including data cleaning, data exploration and visualization, any insights gained, and your modeling approach. At least two different models or algorithms must be used, with at least one being more advanced than simple linear regression for prediction problems.

3 Results

@@@@ A results section that presents the modeling results and discusses the model performance.

4 Conclusion

@@@@ A conclusion section that gives a brief summary of the report, its potential impact, its limitations, and future work.

Bibliography

Laufer, Jens. n.d. “Missing Value Visualization with Tidyverse in R.” Available at https://jenslaufer.com/data/analysis/visualize_missing_values_with_ggplot.html (2020/06/16).

“StartUp Investments (Crunchbase).” 2020. Available at https://www.kaggle.com/arindam235/startup-investments-crunchbase/ (2020/06/16).