CA683 Data Analytics and Data Mining

Predicting if a startup will be successful using data modeling techniques.

Group 42

Eoin Clayton - 20216428
Conor Reilly - 20216376

Contents:

Abstract

Question: Is it possible to predict if a startup will be successful based on the current company funding data?

This report details the following process of applying the CRISP-DM approach to prepare and preprocess data to answer the above question using data analytic and machine learning and processes. These processes include logistic regression, and random forest method.

The following report is based on the following dataset provided by Kaggle that takes financial attributes and information on startups from Crunchbase, a popular company analytics provider: https://www.kaggle.com/arindam235/startup-investments-crunchbase.

1. Introduction

Startup funding is more popular now than ever before with Venture capital funding in 2020 rising from 14% from 2019 [1] despite a global pandemic. More recently, Venture funding in January 2021 hit at an all-time monthly high of $39.9 billion, an analysis of Crunchbase data shows.[2]. The environment in which startups grow and develop is traditionally very complex, so that there are numerous intrinsic and extrinsic variables to be taken into consideration in building a prediction model.

In this case, the report will detail the steps in predicting how likely a startup is to survive or fail depending on certain factors such as market type, total rounds of funding, total amount in funding and which country has the best record for producing successful startups. Also noting if there are any common trends among those that are successful.

With data mining, modelling and machine learning techniques currently at the forefront of modern technology, the need for application of these processes towards company funding is something that is currently underutilized. Previous data analysis towards startup success has primarily been based off of survey analysis and personal opinions [3].

This report will look to provide a clear and concise method of applying the CRISP-DM method for data preprocessing, preparation and modelling under the following headings:

Business Understanding, to provide an understanding of the implications for researching and answering the question of startup successfulness.

Data Understanding and Quality, this step will identify, collect, and analyze the data sets that can help you accomplish the project goals.

Data Preparation, to determine which data will be used and document reasons for inclusion/exclusion.

Modelling, in this case using random forest calculations to model this cleaned data using python and sklearn.

Evaluation, to analyze the results from this data modeling and suggest future improvements. [4]

2. Related Work

This section will overview related work and research into specific aspects prior to the data mining phase. Discussed topics include data mining method, data modeling methods and related aspects research.

2.1 Data Mining Model Research

The major methods to consider for data mining are primarily KDD, CRISP DM and SEMMA. Each of these provides certain positives and negatives that can be taken into consideration. [5]

KDD	SEMMA	CRISP-DM
Pre KDD	----------	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	----------	Deployment

At first sight, it is evident that CRISP-DM is more complete than SEMMA. Another clear benefit of using CRISP-DM is the first initial step of applying business understanding towards. For this data mining research, the application of business understanding is an important part of assessing startup based predictions. Given that this research is likely to be used by venture capitalists or large hedge funds, the application of business needs and expected outcomes would likely be a crucial step to include [6].

The CRISP-DM model is also highly iterative and in general will not require many new requirements at each stage of the process. This will allow for new data to enter the model and be prepared, modeled and evaluated with ease.[7] This factor is very important when dealing with financial and startup data as companies current standings can change rapidly with new investments, closures and acquisitions occurring regularly in this industry.[8]

2.2 Related Aspects Research

2.2.1 Research positive and negatives

This research is set to provide a positive outcome in data modelling of startup based. Risk management through machine learning is still in its earliest stages of development, and it's already proving to be a potent tool. [9] This research will provide a risk assessment data modelling for startups in earlier phases looking to achieve success and large companies such as hedge funds who are looking to back startups with good foundations.

Negatives from this research may be that startups are naturally unpredictable, with more startups failing than being successful [10]. Another factor to consider is that not all data is reliable within this dataset as crunchbase can be publicly edited, predictions will likely give rise to anomalies.[11]

2.2.2 Ethical Research

This data is publicly available within a Kaggle dataset [12]. This dataset is originally from Crunchbase, a financial data website which can be publicly edited. Most company information is publicly available, however, there may be cases of companies within the dataset that have been incorrectly credited or private financial data included, although currently there is no way to track which data is private vs public within the dataset.

2.2.3 Research Limitations

Limitations from the research and application of data mining principles may include that the dataset may be inaccurate as Crunchbase is publicly editable.[11]

It is also worth noting that startups are by nature unpredictable and can change rapidly depending on many different outside factors such as economic conditions or leadership within the company. [13].

The limitations towards the dataset is that there are currently only around 50,000 individual companies listed. At this time, there is no way to tell what specifics were applied to compile this list of companies other than the financial data currently present within the dataset.

3. Data Mining Methodology

3.1 Business Understanding

At this stage, the following assessments were made. The research conducted will use data mining techniques to determine if financial data can be used to predict a startups success rate. The proposed research would be conducted using Crunchbase data and technologies used would be Python and libraries such as Pandas for data manipulation and Sklearn to apply machine learning models to the data.

3.2 Data Understanding and Quality

Initial data understanding period gave an interesting insight to the many different attributes that it takes to define a startup. The dataset contained around 50,000 rows with 39 columns of unique attributes all relating to each individual company.

```
[10] #Total rows and columns within initial data
     df.shape

     (49438, 39)
```

Fig 1. Dataset shape

The dataset had many missing attributes depending on the size of the company and rounds of funding. The next step would be to determine which columns were necessary to keep or remove.

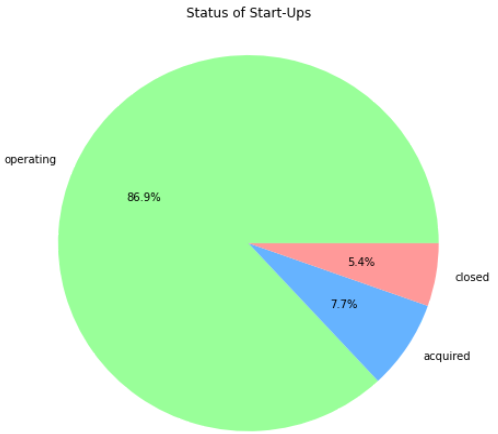
Another factor to consider was that many attributes were in a different format. For example, company_name, market, country were all listed as strings whereas total_funding_usd would need to be converted to an integer.

```
[7] df.dtypes

    name                 object
    market               object
    funding_total_usd    object
    status               object
    country_code         object
    funding_rounds        int64
    dtype: object
```

Fig 2. Dataset types

The following areas were represented as the most important in terms of company success: Startup market, country of origin, total funding in USD and number of funding rounds. This gave an overall coverage of the companies current standings and enough data that an accurate prediction model could be made on whether a company was operating, acquired or closed.



3.3. Data cleaning and Preparation

Many steps were taken in the data cleaning and data preparation phase. The first step in cleaning the dataset is to remove all columns that will not be used for the final prediction. The dataset was then left with 6 columns; Name, Market, Total funding in USD, Status, Country Code and number of funding rounds.

	name	market	funding_total_usd	status	country_code	funding_rounds
0	#waywire	News	1,750,000	acquired	USA	1
1	&TV Communications	Games	4,000,000	operating	USA	2
2	Rock' Your Paper	Publishing	40,000	operating	EST	1
3	(In)Touch Network	Electronics	1,500,000	operating	GBR	1
4	-R- Ranch and Mine	Tourism	60,000	operating	USA	2
...	...	...	...	...	...	...
49433	Zzish	Education	320,000	operating	GBR	1
49434	ZZNode Science and Technology	Enterprise Software	3,174,602	operating	CHN	1
49435	Zzzzapp Wireless ltd.	Web Development	97,398	operating	HRV	5
49436	[a]list games	Games	9,300,000	operating	NaN	1
49437	[x+1]	Enterprise Software	71,000,000	operating	USA	4

49438 rows × 6 columns

Fig 4. Dataset Example

The next step was to get rid of all duplicates and any null or NaN values in the status column as the status of the company is the value we will be predicting. As this was the classification for the data model, rows with this information would be considered too crucial to ignore.

```
#Then we will check for duplicate rows

print('Duplicated entries:',df.duplicated().sum())
df[df.duplicated()].isna().mean()
df.dropna(how='all',inplace=True)
```

```
Duplicated entries: 0
```

The status column has three values; operating, acquired and closed. As the data mode to be used is Logistic Regression which uses binary classification, the data needed to be narrowed down in values. The initial decision that for a startup to be successful its status should be either operating or acquired. To satisfy this, a decision was made to change all acquired values to operating. This would later change to only acquired and closed being considered.

In regards to outliers, there were several outliers that appeared especially in terms of total rounds of funding. For example, outliers in the figure below show that although a majority of companies secured between 1 to 3 rounds of funding, there were examples where companies have secured 18 rounds. This is an unexpected value but as this is financial data, there can be expected anomalies. A decision was made to keep these companies within the dataset, even though this may skew the results, it is an example of how financing can differ in different companies.

status	count	mean	std	min	25%	50%	75%	max
acquired	3692.0	2.013814	1.398832	1.0	1.0	2.0	3.0	15.0
closed	2603.0	1.434114	0.965478	1.0	1.0	1.0	2.0	11.0
operating	41829.0	1.689522	1.302072	1.0	1.0	1.0	2.0	18.0

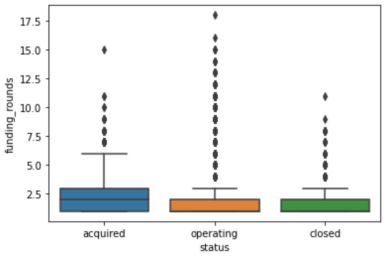


Fig 5. Funding round outliers

4. Modelling

The two types of modelling processes that would best represent this data were logistic regression and random forest. These would be implemented using python and sklearn.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. [14] In this case, operating or closed.

Random forest is a machine learning algorithm that can be used for a variety of tasks. A random forest model is made up of a large number of small decision trees, which each produce their own predictions. [15]

4.1 Logistic Regression

The goal of analysis was to determine if the startup would be operating or closed given a number of relevant columns of data. As logistic regression was being used, the first step was to convert the status to binary. 0 was assigned to operating and 1 to closed.

status		status_binary	
operating	18884	0	18884
closed	990	1	990

There were 18884 total companies still operating and only 990 companies closed.

Although the training model was extremely accurate, the problem was that because of the large number of companies still operating the data became skewed to assuming that all companies that were tested are successful.

```
logreg = LogisticRegressionCV(max_iter=10000)
logreg.fit(X_train,y_train)

print('The training model accuracy: {:.4}'.format(logreg.score(X_train,y_train)))
print('The test model accuracy: {:.4}'.format(logreg.score(X_test,y_test)))
```

```
The training model accuracy: 0.9463
The test model accuracy: 0.9492
```

This large difference in operating vs closed companies later became an issue while we were attempting to predict if future startups would be successful.

```
# What % are Operating and closed

num_operating = np.count_nonzero(results == 0)
print("Number operating: " + str(num_operating))

num_closed = np.count_nonzero(results == 1)
print("Number Closed: " + str(num_closed))
```

```
Number operating: 7301
Number Closed: 0
```

The next step was to reevaluate the objective of our program. After looking carefully at our dataset we determined that the best way forward is to predict if the status is either acquired or closed. This would allow for useful results and predictions to be generated, rather than having highly skewed predictions.

4.2 Random Forest Algorithm

Using random forest, the data was also split into testing and training data before the model was fit using the predefined sklearn features from the RandomForestClassfier.

```
[89] # Random Forest

     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state = 10)
```

```
#Testing the impact of increasing desicion trees used to predict y.
n = 10
while n <= 200:
  model = RandomForestClassifier(n_estimators=n)
  model.fit(X_train, y_train)
  print('{} - train score: {:.3f} | test score: {:.3f}'.format(n,model.score(X_train,y_train),m
  n = n+10
```

```
10 - train score: 0.985 | test score: 0.946
20 - train score: 0.991 | test score: 0.946
30 - train score: 0.996 | test score: 0.946
40 - train score: 0.997 | test score: 0.946
50 - train score: 0.999 | test score: 0.946
60 - train score: 0.999 | test score: 0.946
70 - train score: 0.999 | test score: 0.946
80 - train score: 1.000 | test score: 0.946
90 - train score: 1.000 | test score: 0.946
100 - train score: 1.000 | test score: 0.946
```

Again after the first evaluation, it was clear the data was skewed towards predicting companies to be operating rather than closed the model which would eventually mean that the CRISP-MD model would need to be reassessed for our final evaluations

5. Evaluation

After remodeling the data process the training and testing model were less accurate in predicting the status of any new startup. Although we believe it is a more accurate representation of the data and the unpredictability of creating a startup. Logistic regression and random forest were once again used to evaluate the data.

5.1 Logistic Regression Evaluation

```
logreg = LogisticRegressionCV(max_iter=10000)
logreg.fit(X_train,y_train)

print('The training model accuracy: {:.4}'.format(logreg.score(X_train,y_train)))
print('The test model accuracy: {:.4}'.format(logreg.score(X_test,y_test)))
```

```
The training model accuracy: 0.8072
The test model accuracy: 0.6848
```

The new logistic regression model has a test accuracy of 68%, this resulted in the model predicting that 69% of the startups that were tested will be acquired and 31% to be closed.

The next method of evaluation that was implemented is the confusion matrix.
This matrix shows the number of true and false positive and negative predictions.
In the matrix below it is seen that the model predicted 1000 true positive, 229 false positive, 429 false negative and 439 true negative values.

```
from sklearn.metrics import confusion_matrix

conf_mat = confusion_matrix(y_test, prediction)
print(conf_mat)
```

```
[[1000  229]
 [ 426  423]]
```

5.2 Random Forest Algorithm Evaluation

```
# Random Forest

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state = 10)
```

```
#Testing the impact of increasing desicion trees used to predict y.
n = 10
while n <= 200:
  model = RandomForestClassifier(n_estimators=n)
  model.fit(X_train, y_train)
  print('{} - train score: {:.3f} | test score: {:.3f}'.format(n,model.score(X_train,y_train),model.score(X_test,y_test)))
  n = n+10
```

```
10 - train score: 0.967 | test score: 0.667
20 - train score: 0.987 | test score: 0.662
30 - train score: 0.995 | test score: 0.676
40 - train score: 0.997 | test score: 0.671
50 - train score: 0.999 | test score: 0.679
60 - train score: 0.999 | test score: 0.677
70 - train score: 1.000 | test score: 0.676
80 - train score: 1.000 | test score: 0.675
90 - train score: 1.000 | test score: 0.682
100 - train score: 1.000 | test score: 0.673
110 - train score: 1.000 | test score: 0.675
120 - train score: 1.000 | test score: 0.674
130 - train score: 1.000 | test score: 0.672
140 - train score: 1.000 | test score: 0.668
150 - train score: 1.000 | test score: 0.680
160 - train score: 1.000 | test score: 0.674
170 - train score: 1.000 | test score: 0.675
180 - train score: 1.000 | test score: 0.675
190 - train score: 1.000 | test score: 0.675
200 - train score: 1.000 | test score: 0.679
```

Like the logistic regression model the random forest model was also less accurate after remodeling. The new model received a test accuracy of 67%. This model predicted 67% of companies to be acquired and 53% to be closed.

The confusion matrix results were similar to that seen in logistic regression.
983 were predicted as true positives, 218 as false positives, 461 as false negatives and 416 as true negatives.

```
[37] from sklearn.metrics import confusion_matrix

     conf_mat = confusion_matrix(y_test, pred_results)
     print(conf_mat)

     [[983 218]
      [461 416]]
```

There are a number of reasons for the false predictions. The model was predicted using the market, the location and the amount of funding the startup received. The model may have been influenced by the startups that are in the same market but a different location to another startup as some locations require more funding than others to become successful.

6. Conclusion

In conclusion, using the following CRISP-DM data mining model to apply business understanding, data quality and understanding, data preparation, and evaluation, resulted in a comprehensive review of how executing a model can be applied to a dataset involving startup financial data.

CRISP-DM was selected as it met the current requirements and business needs for a data mining research project of this nature.

evaluation differently? Also would the implication of different modelling practices such as clustering alter these results or give further unique insights?

The results from the logistic regression and random forest model were quite surprin insightful to understanding how test versus training data can affect a model.

Although there were difficulties with data preparation and a re-evaluation of attributes was required, this is a prime example of how CRISP-DM can be reiterated to suit business and evaluation needs.

Future research on this project could include the values that were left out of the dataset. For example, would including which each individual funding round or venture type have skewed the model in a different direction. Another option would be to attempt different data mining models such as clustering to see if that would give rise to more varied or detailed results.

7. GitHub Link

https://github.com/eoinclayton98/CA683-Data-Analytics-and-Data-Mining

REFERENCES

[1] Startup funding touches new records amid pandemic - Lizette Chapman, Jan. 17, 2021
https://www.seattletimes.com/business/startup-funding-touches-new-records-amid-pandemic/

[2] Monthly Recap January 2021: VC Funding, Just Shy Of $40B - Gené Teare, Feb 8, 2021
https://news.crunchbase.com/news/january-2021-vc-funding-report-record/

[3] Searching for a Unicorn: A Machine Learning Approach Towards Startup Success Prediction Master's Thesis submitted to Prof. Dr. Wolfgang Karl H¨ardle and Prof. Dr. Weining Wang
https://edoc.hu-berlin.de/handle/18452/21141

[4] What is CRISP DM? - Data Science Process Training
https://www.datascience-pm.com/crisp-dm-2/

[5] Data Science project management methodologies - Quantum
https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb

[6] KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW Ana Azevedo CEISE – ISCAP – IPP Rua Jaime Lopes de Amorim, s/n – 4465 S. M. Portugal Manuel Filipe Santos DSI - UM Campus de Azurém
https://www.researchgate.net/profile/Ana-Azevedo-48/publication/220969845_KDD_semma_and_CRISP-DM_A_parallel_overview/links/02bfe50cbb21f029f1000000/KDD-semma-and-CRISP-DM-A-parallel-overview.pdf

[7] CRISP-DM & Decision Management Creating a decision-centric, repeatable approach - James Taylor
http://www.decisionmanagementsolutions.com/wp-content/uploads/2018/07/CRISP-DM-and-Decision-Management-061718.pdf

[8] CRISPy Data Mining in Marketing Organizations
https://wyzoo.com/blog/crispy.html

[9] The Impact Of Data Science Analytics On Financial Institutions -Edwin Lisowski Jan 7 20
https://towardsdatascience.com/the-impact-of-data-science-analytics-on-financial-institutions-ee2d272427d1

[10] How Many Startups Fail and Why? By SEAN BRYANT Nov 9, 2020
https://www.investopedia.com/articles/personal-finance/040915/how-many-startups-fail-and-why.asp

[11] Where does Crunchbase get their data? - Crunchbase Product Team
https://support.crunchbase.com/hc/en-us/articles/360009616013-Where-does-Crunchbase-get-their-data-

[12] StartUp Investments (Crunchbase)
https://www.kaggle.com/arindam235/startup-investments-crunchbase.

[13] Economic influence on business activity
https://www.bbc.co.uk/bitesize/guides/zjjnnrd/revision/1

[14] Logistic Regression — Detailed Overview - Saishruthi Swaminathan
https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

[15] What is a Random Forest? Thomas Wood
https://deepai.org/machine-learning-glossary-and-terms/random-forest