JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ANTICIPATING COMPANY SUCCESS OR FAILURE USING MACHINE LEARNING MODEL

[1] Dr.Yazdani Hasan, [2] Archana Jain

[1] Associate Professor, [2] Professor

[1] , School of Engineering and Technology(CS)

[1] IIMT University, Meerut, India

ABSTRACT:

Both scholars and practitioners have always had difficulty predicting the success of a commercial initiative. However, because of businesses that compile data on other businesses, it is now possible to develop and evaluate predictive models based on a record number of real-world instances. In this study, we make use of data from Crunchbase, one of the biggest platforms for integrating company data. 223182 businesses made up our final training set. In order to anticipate a company's performance, this effort tries to build a predictive model based on machine learning. Similar attempts have been done a lot lately. Many of those tests, which frequently made use of information acquired from a number of sources, had encouraging findings. However, we discovered that they were frequently materially biased by their usage of data that revealed information that was a direct result of a business experiencing some kind of success (or failure). A typical illustration of the look-ahead bias is a strategy like this.

It produces incredibly positive test results, but any effort to use such a strategy in a situation that might occur in real life could have disastrous outcomes. We planned our experiments to stop any information from reaching the training set that was not accessible at the time of the choice. We contrasted three algorithms: gradient boosting classifier, logistic regression, and support vector machine. Despite the deliberate choice to restrict the amount of predictors In terms of accuracy, recall, and F1 scores, which were 57%, 34%, and 43% for the best model, respectively, we arrived at extremely encouraging findings. With the gradient boosting classifier, the best results were attained. The top three factors are the nation and location the company works in, the sector the firm is in, and we provide thorough information about the significance of various attributes. For many kinds of venture capital funds, our model may be used right away as a decision support system.

KEYWORDS:

Supervised learning , Support vector machine ,Logistic Regression.

1. INTRODUCTION

Entrepreneurs and investors have good cause to be happy when their company endeavour succeeds. A cash incentive is also closely related to it. Both entrepreneurs and investors are always seeking for resources, techniques, and guidance that will put them ahead of their rivals. It is debated if having certain inherent talents is a need for success as an entrepreneur or whether those skills may be learned (for example, through formal business school). Furthermore, it is highly challenging to gauge the importance of external elements like the sector in which a firm works, the neighbourhood in which its headquarters are situated, or the intensity of competition in a certain sector and its subsectors.

The question of what elements are essential for a business to prosper has long been the subject of inquiry. According to Stuart and Abetti (1987), a team's alignment with an entrepreneur's technical and commercial experience is essential for success. They also made other less insignificant discoveries, such as the fact that businesses in "more slowly growing or less dynamic markets" tend to succeed initially more easily. Another strategy was looking at how informal and formal information affected the performance of small and medium-sized businesses in Shanghai (Vaughan, 1999). Using variables defining the corporate environment, information utilisation, and market scenario, Vaughan (1999) constructed a linear mathematical model to investigate the quantitative link between information and success.

For scholars studying management theory, predicting corporate performance has been a fascinating task. They have examined how management tools and theories affect a company's success (Spyros Makridakis, 1996). The difficulties of forecasting the success of young enterprises and the various, shifting elements of their operating environment were also major topics of business study. Cooper also looked at the compatibility between the founder's personal objectives and the expansion of the business.

Venture capital (VC) firms make investments in start-ups and small businesses in the hopes that they will experience long-term growth and provide a sizable return on their cash. A startup is described as "a human institution designed to create a new product or service under conditions of extreme uncertainty" (Ries, 2011). Nine out of ten companies fail, according to industry statistics. More than 75% of businesses, including those with venture capital backing, fail or only manage a subpar life (Picken, 2017). The performance of VC funds could be enhanced by a model that predicts business success. While venture capital funds often offer favourable return on investment, the study (Harris et al., 2014) demonstrates that throughout the 2000s they underperformed when compared to the S&P 500 index. For venture capital funds, the objective is to identify companies with higher success rates.

2.SOFT COMPUTING MODELS FOR PREDICTING BUSINESS SUCCESS

For a very long time, company success has been predicted using machine learning techniques. Using data gathered from surveys of US enterprises, Lussier (1995) utilised logistic regression to predict a new firm's performance. Other studies looked into how to forecast the performance of Australian ICT businesses using macroeconomic parameters. Support vector machine (SVM), naive Bayes, and k-nearest neighbours (k-NN) algorithms were effectively applied by Tomy and Pardede (2018) (Tomy & Pardede, 2018). However, the datasets utilised in both publications were just 216 and 250 occurrences, respectively, making them quite tiny.

Researchers were able to integrate information about financing events in the models and expand the datasets to thousands of cases by using data from Crunchbase (Krishna et al., 2016). In order to forecast firm acquisition using Bayesian networks, Xiang et al. (2012) analysed Crunchbase data and factual elements from TechCrunch articles. Xiang et al (2012). Using logistic regression, SVM, and random forest algorithms, Bento (2018) predicted acquisitions or initial public offerings (IPOs) for firms with US addresses using Crunchbase. A gradient-boosted decision tree model for series prediction was created by Sharchilev et al. (2018). For businesses that have already received seed or angel finance, a funding in the following year. They enhanced the dataset from Crunchbase that was collected in monthly snapshots.

Additionally, investment behaviors modelled with graph approaches were predicted using Crunchbase data. Yuxian and Yuan discovered that it is feasible to anticipate whether investors are likely to invest by utilising several link predictors, such as the shortest path in the graph or the number of neighbors (Yuxian & Yuan, 2013). Hybrid intelligence techniques are another possibility for forecasting company success. By combining judgements made by a group of people with decisions made by machine learning algorithms utilizing hard data (team size, entrepreneurial experience), Dellermann et al. established a framework. Both professionals and non-experts would utilise their instinct, experience, and market knowledge to make a success prediction for a business. After that, the data would be combined to provide the categorization output (Dellermann et al., 2017).

3 OBJECTIVES AND CONTRIBUTION

Using supervised machine learning techniques, we examined the issue of predicting the success of commercial endeavours in this study. Numerous instances of the use of contemporary machine learning models for that specific aim can be found in the literature. Those studies frequently make use of many data sources, integrating numerical aspects with information taken from text data that has been Internet-crawled. Although this strategy is incredibly alluring, it might easily make it impossible to apply the findings. Aligning the data on the timeline properly is the major issue. Data that was acquired at a certain time in the past is frequently used in research.

Rarely would decision-makers wish to determine if a certain enterprise qualifies for financing at this point. Furthermore, adding data that was crawled at the time the experiment began to such a dataset introduces the bias mentioned in the preceding section. The major goal of this study was to undertake an experiment that would result in the creation of an information system that would be free from the aforementioned biases and could be used in the real world to forecast economic success.

The following are the primary contributions of this paper:

1. To the best of our knowledge, this is the first study that places a high emphasis on the application of its findings by minimizing the amount of biases added to the dataset. We succeeded by deliberately restricting the collection of predictors to knowledge available at the start of the company's activities.

2. By far the biggest dataset among comparable studies was employed for this research. There are 213 171 firms in our practice set.

3. We gathered information from the dataset and performed statistical analysis to generate insights that could be useful to founders, legislators, and investors. The main startup centres, secondary sites, and businesses from other industries are all included in the scope of the investigation. In terms of its significance, the offered analysis gains from the sample size. For the purpose of supporting our research and data selection, we used information acquired by crawling more than 700 000 websites.

4. One of the first such methods is the development of a goal variable that incorporates data on IPO, acquisition, and later funding rounds.

4 DATA COLLECTION AND METHODOLOGY

The Crunchbase database (www.crunchbase.com) was used to gather information for the study. Crunchbase is a website that provides access to business data about both private and public firms, founders or other key decision-makers, investors, and investment rounds (Crunchbase Inc., 2020). For this study, we requested permission to access the Crunchbase database. We were given access to the daily snapshots of the Crunchbase database after receiving a favourable answer to our request. On March 10, 2020, the information needed for the study and experiments was acquired. We will go into more detail about creating the dataset in this part so that machine learning models may be trained on it.

4.1. *DATASET FROM CRUNCHBASE*

Multiple tables that may be connected together by distinctive IDs make up the dataset offered by Crunchbase for research purposes.

Fig. 1 depicts the crunchbase tables' simplified entity-relationship diagram (ERD).

The organisations table contains data about businesses and investment funds. The table includes the company's name, headquarters address, number of workers, website, social media connections, email address, and phone number. The financial information that has been condensed includes the total money, the total number of fundraising rounds, the date of the most recent funding event, and the total number of exits from investments. Whether an organisation is operating, closed, purchased, or a public firm becoming public, Crunchbase maintains track of that information as well. Additionally, each organization's core function (business or investor) and the categories and subcategories that characterise it are outlined.

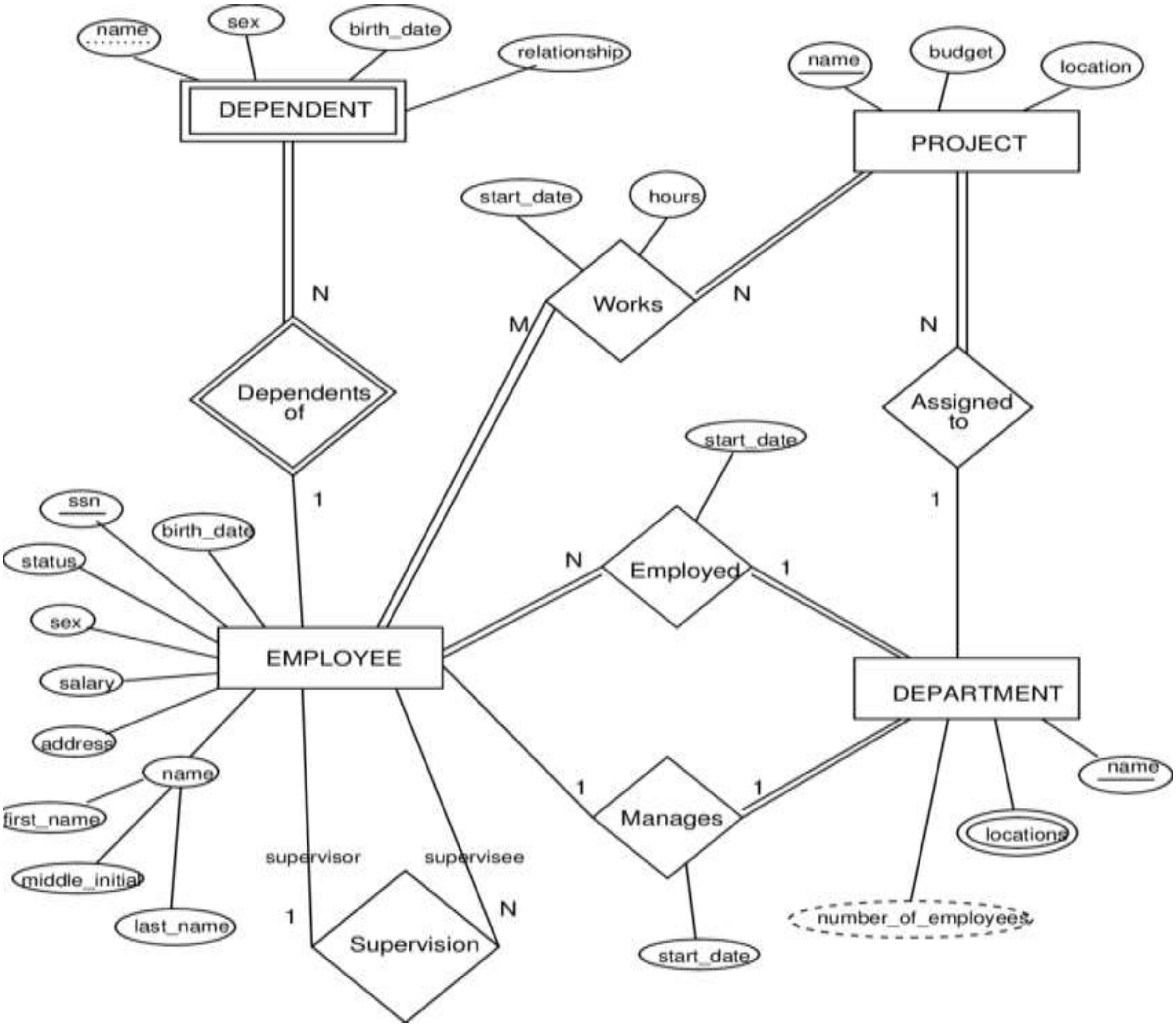


Fig. 1. Simplified ERD diagram of Crunchbase data.

In the appropriate tables, further details on fundraising occasions, acquisitions, IPOs, and investors are kept. Data on the dates of such events, the amount of cash raised, and the kind of investment (seed, angel funding, series A, B, C, etc.) are all included.

The persons table lists the people who are the organisations' founders, investors, or workers. The person's name, gender, residence, links to their social media accounts, employer, and position within the employer are all shown in the table. The degrees table contains details on each individual's schooling. Each item may include details about the degree's field of study, the institution where it was completed, and the dates of matriculation and graduation.

4.2 *COMPANIES' HOMEPAGE RESPONSE*

The URL for each company's home page is also included in the Crunchbase dataset. Investors and prospective consumers can find information about a firm on its site. An organization's failure to succeed on the market is unmistakably shown by an idle webpage. A target variable might be made using this information to determine if a firm is successful.

We prepared and carried out an experiment by crawling replies from the websites of the businesses listed in the organisations table that had the functioning status. Using Python's urllib package, the replies from the websites were gathered. The script that does the website crawl records the website's HTTP response code as well as any error messages that urllib raised while handling the request. CSV files are used to store the script's output. In the Crunchbase dataset, the identification of the organisation is used to index each entry.

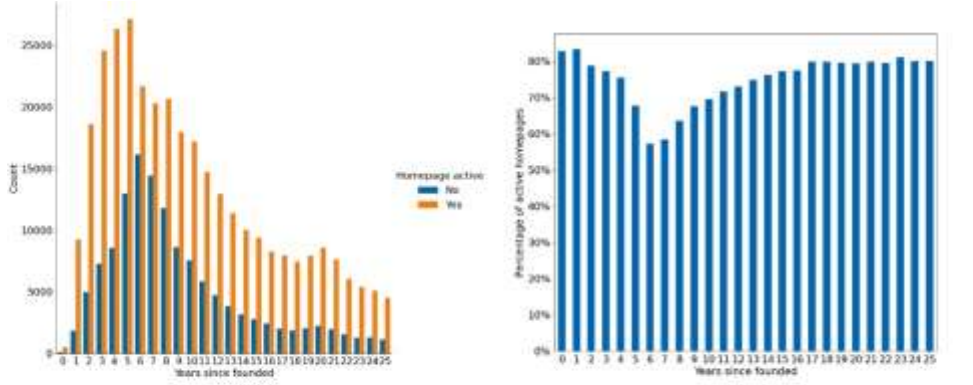
In the organisations table, we added a new column with a flag indicating whether the homepage was active or inactive based on the data gathered throughout the trial. In this column, organisations having HTTP response codes of 200 were given the number 1 (active), whereas all other organisations received the number 0 (inactive) (Fielding & Reschke, 2014). Figure 2(a) shows how homepage activity has changed throughout the years since the firm was established. The distributions are inversely correlated, with the highest number of inactive homepages for businesses occurring in the sixth year after establishment and the maximum number of active homepages occurring in the fifth.

The percentage of businesses having live and inactive homepages is depicted in Fig. 2(b). In the sixth year following a company's inception, we may see the minimal value. After then, the proportion of businesses with a live homepage increases. As seen in Fig. 2, we may deduce that it takes most businesses five years to verify their idea on the market. The proportion of businesses with dormant homepages is at its lowest six years after inception. Then it goes back to being steady at a value of more than 70%.

4.3 STATISTICAL EVALUATION

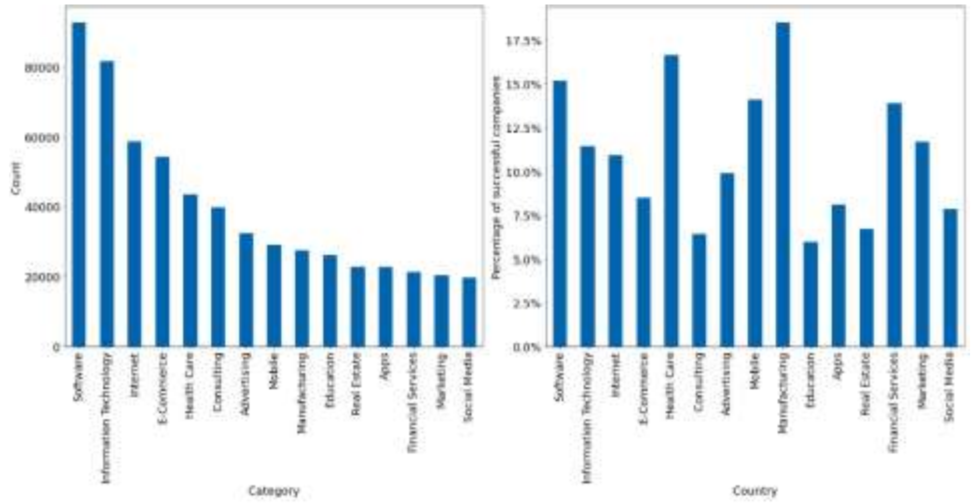
4.3.1 Most popular startup industries

The top fifteen markets in which the analysed businesses do business are depicted in Fig. 3. Information technology makes up four of the top five most popular categories. Due to their "high flexibility, adaptability, and inclination to implementing innovational products and business processes," small and medium-sized firms (SMEs) constitute the fundamental component of the Internet economy (Sukhodolov et al., 2018).The Fourth Industrial Revolution also heavily relies on these sectors of the economy. The automation of production and services, made possible by digital transformation, is poised to revolutionise the globe. Companies like Uber, AirBnB, and Amazon disrupted marketplaces and saw market growth at a never-before-seen rate in the previous ten years (Schwab & World Economic Forum, 2016). Entrepreneurs attempting to replicate these businesses' success could find inspiration from their example.



(a) Active and inactive homepage distribution (b) Ratio of companies with active homepage

Fig. 2. Active homepage distribution versus companies' age.



(a) Number of companies per category (b) Percentage of successful companies per category

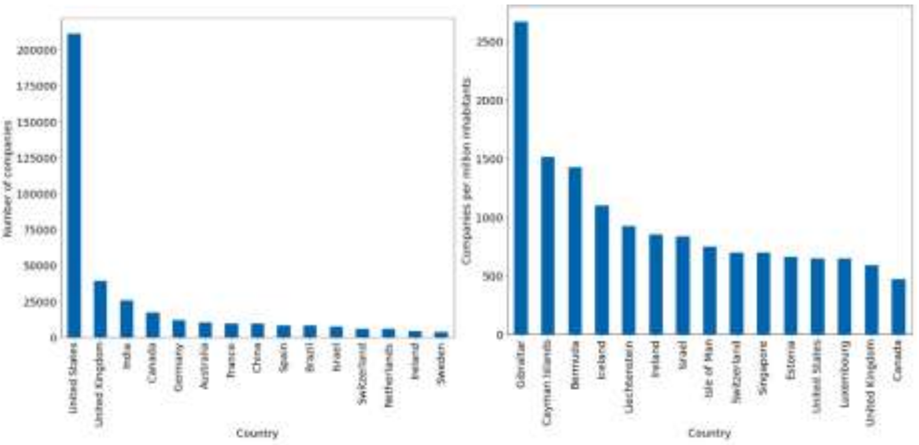
Fig. 3. Most popular industry categories among analyzed companies.

4.3.2 *Top startup hubs*

According to Fig. 4(a), the dataset's country with the most startups is the United States. According to CB Insight (2018), Silicon Valley, New York, Boston, and Los Angeles make up the majority of the world's top startup clusters. Startup hubs are situated in areas that constitute an innovation cluster because they fit the criteria. Universities, businesses, established firms, and sizable private funding sources are the constituents of an innovation cluster (Engel, 2014). However, if we look at the number of startups per million people, we may see an almost entirely distinct group of nations, as illustrated in Fig. 4(b). Some of the best nations have tax havens and advantageous fiscal systems, like Gibraltar, Bermuda, and the Cayman Islands. For tax-saving purposes, businesses decide to domicile their financial centres in these nations (Palan et al., 2010).

Israel and Estonia are among the nations with the highest startup rates per million people, as well. Considering how much public services have been digitalized, the former is known as the Digital Republic. Most administrative duties may be completed by Estonian citizens using their identity card with an electronic chip. The fact that programming is taught beginning in first grade may be another factor in the high rate of startups per capita (Gat, 2018).

Israel, recognised as the Startup Nation, is also renowned for having an economy that is focused on technology. It is notable for having a significant number of businesses that were formed by former members of the military's cybersecurity unit 8200 (Fraiberg, 2017). Another well-known startup hotspot is Tel Aviv. Compared to Silicon Valley, London, or New York, it does not have as many companies functioning there, but the investments provide high return. 14% of exits between 2012 and 2017 were valued over $100M. It is comparable only to Silicon Valley (CB Insight, 2018).



(a) Number of companies

(b) Number of companies per million inhabitants (2018)

Fig. 4. Countries with most companies in the dataset.

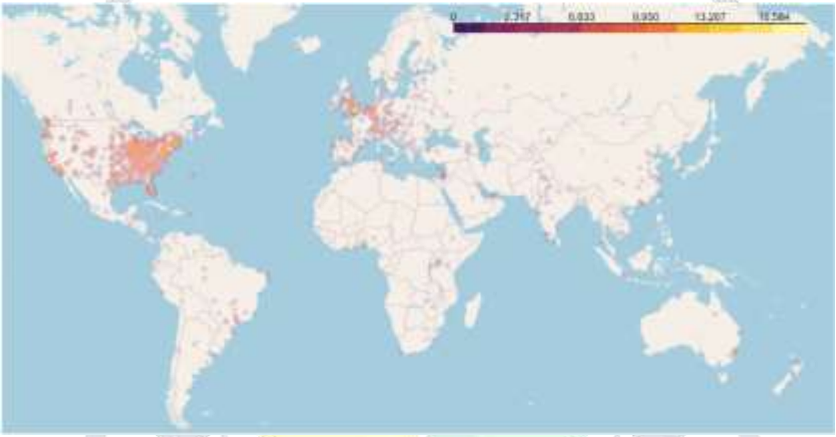


Fig. 5. Distribution and size of startup hubs in cities around the world. Map by © OpenStreetMap contributors under ODbL.

The dispersion of businesses across global cities is seen in Fig. 5. The size of the bubble depicts how many businesses are present in the city. We can see that the United States has a very high number of startup clusters. London is the continent's most well-known startup hotspot. Only in this city do functioning startups outnumber hubs in the US in terms of number. The majority of the Asian enterprises under analysis are located in major Asian cities including Singapore, Hong Kong, Beijing, Tokyo, and Seoul. We can observe the existence of startups functioning from smaller cities in Europe and the US. It is likely that Asian businesses searching for outside finance or planning a worldwide expansion are included in the dataset. According to this theory, having a presence on Crunchbase would help them draw in new clients and investors. We do not have access to the Crunchbase user base analysis that would be required to confirm this notion, though.

4.4 FEATURE ENGINEERING

4.4.1 Data selection

Organisations created between 1066 (HM Treasury) and 2020 are included in the dataset from Crunchbase. Numerous additional well-known organisations are listed as firms in Crunchbase and have company biographies. In keeping with earlier studies (Bento, 2018; Krishna et al., 2016; Xiang et al., 2012), we chose to only include the newest businesses that satisfy the criteria of a startup. Companies created between 1995 and 2015 were included in the subgroup we chose. We choose to exclude the businesses established in 2015 Considering that they are only getting started, and 2020 were excluded from the sample. We would not be able to appropriately classify those businesses with ground truth information. According to 4.2 in Fig. 2's examination of homepage answers, there are less businesses with active status and inactive homepages after five years. The identification of profitable businesses is therefore achievable. The majority of the pioneering businesses in the software and Internet sectors were established before 1995. The proportion of American households with computers increased from 22% in 1993 to 51% in 2000, which contributed to the late 1990s Internet boom (Ryan & Lewis, 2017). New Internet-based business models were introduced as a result (Litan & Rivlin, 2001). The Internet, a brand-new and quickly developing technology at the time, was the foundation for many new businesses that were established to offer their services and create their goods around it. Many of them added the.com prefix to their names, giving rise to the term "dot-com companies." Between January 1995 and February 2000, the NASDAQ Composite Index increased 450% as a result of excessive speculative activity in these businesses (MacroTrends LLC, 2020).

Table 1

List of final features in the dataset.

Feature name	Description	Type
category_list	List of organizations' subcategories	nominal
category_groups_list	List of organizations' categories	nominal
gender	Founder's gender	nominal
is_completed	Founder has completed a degree	boolean
has_multiple_degrees	Founder has more than one degree	boolean
region_org_size	Rank of region in number of startups	categorical
city_org_size	Rank of city in number of startups	categorical
years_between_graduation_and_founding		Number of
years between founder's graduation and company's foundation		interval
years_of_studying	Number of years between founder's matriculation and	
graduation	interval	

The dot-com boom burst in 2000, which was one of the factors contributing to the US recession in 2001 (Kraay & Ventura, 2005). To cover the businesses that were active during the dot-com bubble, we decided to drop the analysis's constraint to 1995.

A similar phenomena to that seen when gauging the performance of hedge funds may arise when forecasting corporate success (Amin & Kat, 2003). Companies that have existed in the past but failed to survive are most likely not included in our dataset. The oldest samples tend to favour businesses that have endured adversity (perhaps through numerous economic downturns). The percentage of profitable businesses should be significantly lower in the actual distribution than it is in the Crunchbase data. This will cause trained models' sensitivity to be overestimated. Since there isn't a publicly accessible dataset that would collect information on failed company initiatives, it is difficult to assess the size of this effect. It is conceivable, though, that this prejudice gradually diminished after Crunchbase became well-known. A "master database for companies" is what the business aspires to be. Even with the assumption of nearly complete coverage of all organisations that have ever been, we still need to deal with the issue of outdated data. In example, this is a fairly typical scenario for businesses that no longer operate but nevertheless maintain a "active" status in the database. The next parts, particularly 4.4.3, go into further information about this topic.

4.4.2 Dataset creation

The organisations, persons, degrees, and funding_rounds tables' contents were utilised to build the actual dataset that was put to use in the experiment. The other tables were not utilised either because they included text data (people_description and organization_descriptions tables) or because they could have introduced a look-ahead bias (investments, investors, ipos, and funds tables). Usually, the individuals who fit the specifications of the people and the members of the organisations are those specific persons and members. Utilising identifiers, the chosen tables were connected. The tables for individuals and degrees were combined, and the merged table retained the initial degree in chronological order. Following that, the resultant table was combined with the organisations table. If there were numerous persons working for one organisation, the founder or executive officer was chosen. The funding_rounds table's target variable was made. The funding_rounds table's data was not utilised in the dataset's features.

State code (organisations table), subject, institution name, degree type, city (people table), region (people table), and country code (people table) were the characteristics that were eliminated from the dataset after analysing value distribution and calculating the missing values ratio.

Three categories—nominal, interval, and binary—represent the chosen characteristics. Building an appropriate prediction model could be challenging since many nominal features have numerous distinct values. They need to go through extra transformation processes before encoding.

Table 1 displays the complete feature list prior to encoding. In comparison to earlier research that utilised Crunchbase data, the number of characteristics that have been chosen is purposefully lower. Previous works contained knowledge that was acquired towards the end of the company's existence. As an illustration, consider details on financing occasions or Venture Capital (VC) support (Bento, 2018; Krishna et al., 2016; Xiang et al., 2012). Only the information that would have been known at the start of the business' operations will be preserved, we determined.

4.4.3 Target variable

To define the goal variable, it is necessary to define what business venture success entails. It takes more than just surviving in the market to call a business successful. Startups frequently rely on outside capital to expand their businesses, which is also an indication that investors are optimistic about the future of the business. Rounds of business finance are typical. Engaged investors grow more institutionalised, and each succeeding round is often greater than the one before it. Reaching a particular milestone for the firm implies receiving each funding series. In some of the earlier efforts (Sharchilev et al., 2018; Spiegel et al., 2016), reaching a certain financing milestone served as a success indicator.

We made the decision to use the conclusion of series B as the benchmark for defining the company's success. A startup has often demonstrated that it has a reliable user base that generates profit when it successfully completes a series B financing. Among the firms analysed, the median amount raised in series B is $11.2M, compared to $5M raised in series A. Receiving Series B signifies a firm has successfully navigated both rounds of investment fund selection, which is a powerful predictor of success.

Another sign of the company's market worth is that it was acquired. The acquiring party seizes control of the acquired company's concept, goods, and personnel. For some startups, being bought could be viewed as a growth strategy. For instance, research suggests that university spinoffs have a higher acquisition probability than they have a public offering. This may be due to academics' tendency to place a greater emphasis on the technical aspects of a product than on business expansion. A more established business may be able to meet these demands while developing expertise. For both sides, the purchase may be viewed as a "win-win" approach (Bonardo et al., 2010*)*.

The goal variable is binary, with the positive class including only successful firms and the negative class including all other enterprises. The criteria for the positive class are designed to clearly identify instances of achievement. By launching an IPO, a firm has completed the necessary preparations, which include publishing a prospectus with the publicly disclosed value. The purchase of the business is another surefire sign of success since it allows the founders to recoup their investment in starting the business. We also consider businesses who have obtained series B investment and are still in business to be successful. Although it is impossible to predict whether they will succeed or fail in the future, getting series B investment is a big step forward for the business and a vote of confidence from the investors. The likelihood of the businesses in the positive class doing well in the market is very high. The following defines the target variable:

$$y = \begin{cases} 1, & \text{if } x_{status} = \text{"acquired"} \vee x_{status} = \text{"ipo"} \\ & \vee (x_{status} = \text{"operating"} \wedge x_{investment\_type} = \text{"series\_b"}) \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

In the fifteen nations with the most organisations in the dataset, shown in Fig. 6, the proportion of cases falling into the positive class. With more than 30% of its enterprises falling into the favourable category, China stands out. However, among the examined nations, Brazil, Spain, and India have the lowest percentages of successful businesses.

4.4.5 *Binning values based on frequency*

The information includes a wide range of the startup companies' operating cities and areas. There are, correspondingly, 21159 and 1755 distinct values in those columns. For the number of unique values to be reduced to a processable level, the cutoff value would need to be quite high. The frequency distribution of the values, on the other hand, indicates that classifying them into tiers might inform the model while simplifying it. Five equal-width bins were created for the values based on how frequently they appeared in the dataset. As a result, the top bins were sparsely inhabited, but it also distinguished the top startup centres from the others. Because of its distinct position among startup clusters, California, for instance, is in a different bucket than New York or England.

It was also attempted to use the equal-depth binning method, however the results did not produce the needed separation between the top hubs and the remainder. We think that employing equal-depth binning would lead to the information being lost.

On the basis of the previously discussed binning, the city and region characteristics were replaced by ordinal attributes of city size and region size.

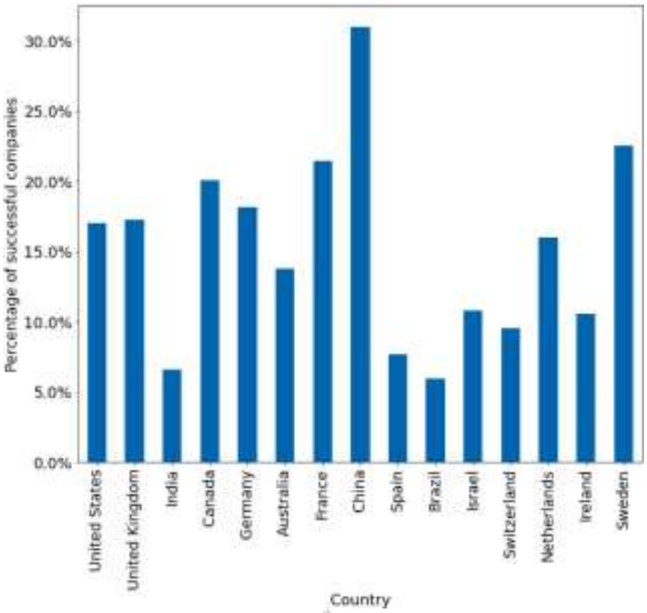


Fig. 6. Percentage of instances in positive class in countries with most instances in the dataset.

4.4.6 Transforming dates into time ranges

Giving the model dates in an explicit manner could introduce bias. A corporation has a greater probability of succeeding on the market, for instance, the longer it has been in business. When predicting success, a model would favour older organisations if the dataset's date of firm creation were included. Our first investigations showed that dates should be converted into relative time ranges to prevent this situation.

The number of years that passed between the founder's graduation and the company's founding was used in place of the founding year of the business. The updated function makes the professional background of the company's founder, who founded it, plain information. The number of years the founder spent attending the institution was used in place of the dates of his matriculation and graduation. The values lacking is filled with 0.

4.4.7 Encoding attributes

The final set of characteristics should be encoded before being used as a dataset for model training once the dataset's number of unique values has been decreased. One-hot encoding was used to change the gender, category, and category group characteristics. Using ordinal encoding with missing data labelled separately, region size, city size, country code, and boolean flags were changed.

5.EXPERIMENTS

5.1. SUPERVISED LEARNING

The task of forecasting business success is reduced to binary classification by defining the target variable's firm success as having two potential values: 1 (successful) or 0 (unsuccessful). A supervised learning strategy can be used to resolve this issue. The dataset is divided into the target variable (vector y) and a subset of characteristics (matrix X) in supervised learning. Models are trained to attempt to predict the value of y given a collection of x characteristics. Comparison is made between the anticipated value y and the actual value y.To increase the number of accurate predictions, the model's parameters are modified iteratively..

Multiple techniques may be used to tackle the binary classification problem because machine learning has extensively studied it. With the help of well-known Python machine learning libraries like scikit-learn and XGBoost, we choose to use three straightforward models whose implementations are readily available. The models in question were XGBoost, SVM, and logistic regression. SVM and logistic regression are well-known techniques that have been applied to several prior studies using Crunchbase data (Bento, 2018; Krishna et al., 2016; Xiang et al., 2012). Due to its success in Kaggle contests, the decision tree and boosting-based XGBoost algorithm has lately become more well-known (Kaggle Inc., 2019).

Table 2

Results of 10-fold cross validation on the training set. Pre ision, recall, and F1 score are reported for the positi e class.

Algorithm	Accuracy	Precision	Recall	F1
Logistic regression	0.86	0.70	0.21	0.33
SVM (RBF kernel)	0.87	0.86	0.20	0.32
XGBoost	0.86	0.90	0.17	0.28

5.2TRAIN/TEST SPLIT

After randomization, the dataset was divided into stratified train and test subsets (Sechidis et al., 2011). The stratified split ensures that the distribution of the target variable's classes in the training and test subsets is the same. As a result, the test set serves as a representative sample of the dataset for the target variable. The test set's size is set at 10,000 and cannot be changed. In the dataset, it represents about 5% of cases. The test set was used to provide the model's performance's final findings. Instead of the common 80-20 split, we employed a test set with a fixed size to deliver more examples during the training phase.

To prevent models from overfitting to the validation set, we used cross-validation on the training set instead of a separate validation test. To combat selection bias and overfitting, cross-validation is advised while tweaking hyperparameters (Cawley & Talbot, 2010). The distribution of the target variable was the same in a subset used for validation and the remainder of the training set utilised in the fold in each cross-validation fold.

5.3 EXPERIMENT SETUP

Model selection experimentation is shown in Fig. 7. As previously mentioned, the dataset was divided into training and test sets. In tests employing SVM (standardisation) and logistic regression (minmax normalisation), data was preprocessed using feature scaling techniques. The use of feature scaling with these algorithms enhances their efficiency and decreases the time needed for an algorithm to converge. Given that the technique employs lasso and ridge regularisation (see 5.5.1 for details and results), logistic regression requires scaling of inputs (Hastie et al., 2013a). Feature values were mapped to the [0, 1] range using minmax normalisation. Data standardisation was done prior to doing SVM training. The distribution of each characteristic is altered by standardisation to have a zero-mean and unit-variance. The models were decided upon during the hyperparameter tuning procedure, which is thoroughly explained in Section 5.5. The top-performing models were then trained across the board and put to the test on the test set. The test set, which did not serve as a baseline for the models' performance throughout the model selection process, yielded the findings that were reported.

5.4 INITIAL RESULTS

The XGBoost and scikit-learn packages' default set of settings were utilised in the initial trials. Tenfold cross-validation was used to train the classifiers. The validation subset in each fold was used to quantify accuracy, precision, recall, and F1 score (the harmonic mean of precision and recall). The metrics in Table 2 are the averages of values obtained from 10-fold cross-validation for the positive class. Nearly 90% accuracy values were achieved by all of the classifiers. The great accuracy, also close to 90%, of the SVM and XGBoost algorithms stood out. The classifier with logistic regression has the highest F1 score. All three classifiers had extremely low recall metrics, nevertheless. Therefore, 70% of successful businesses were misclassified as failures by models. The algorithm should, in our opinion, Find more effective companies, not just ones that adhere to the prevailing trends.

5.5HYPERPARAMETER TUNING

The preliminary findings indicated that models might enhance recollection performance. Finding the ideal collection of model parameters is the goal of the hyperparameter tuning procedure. We utilised two techniques—grid search and randomised search—to identify and evaluate various sets of model parameters since searching the whole space of potential values for parameters would be computationally costly. For studies using both of those strategies, we chose parameter values. Due to the small number of parameters that needed to be adjusted, an exhaustive grid search was used for logistic regression and SVM. We may test all possible combinations of the chosen values. Due to the large number of parameters, we used a randomised search to fine-tune the XGBoost model's performance. There was a thorough search.

5-fold cross-validation was used in both techniques to assess how well various model modifications performed. According to the importance of the F1 score, the verified models were ordered. The performance of the classifiers is balanced by maximising the F1 score, which balances the trade-off between precision and recall.

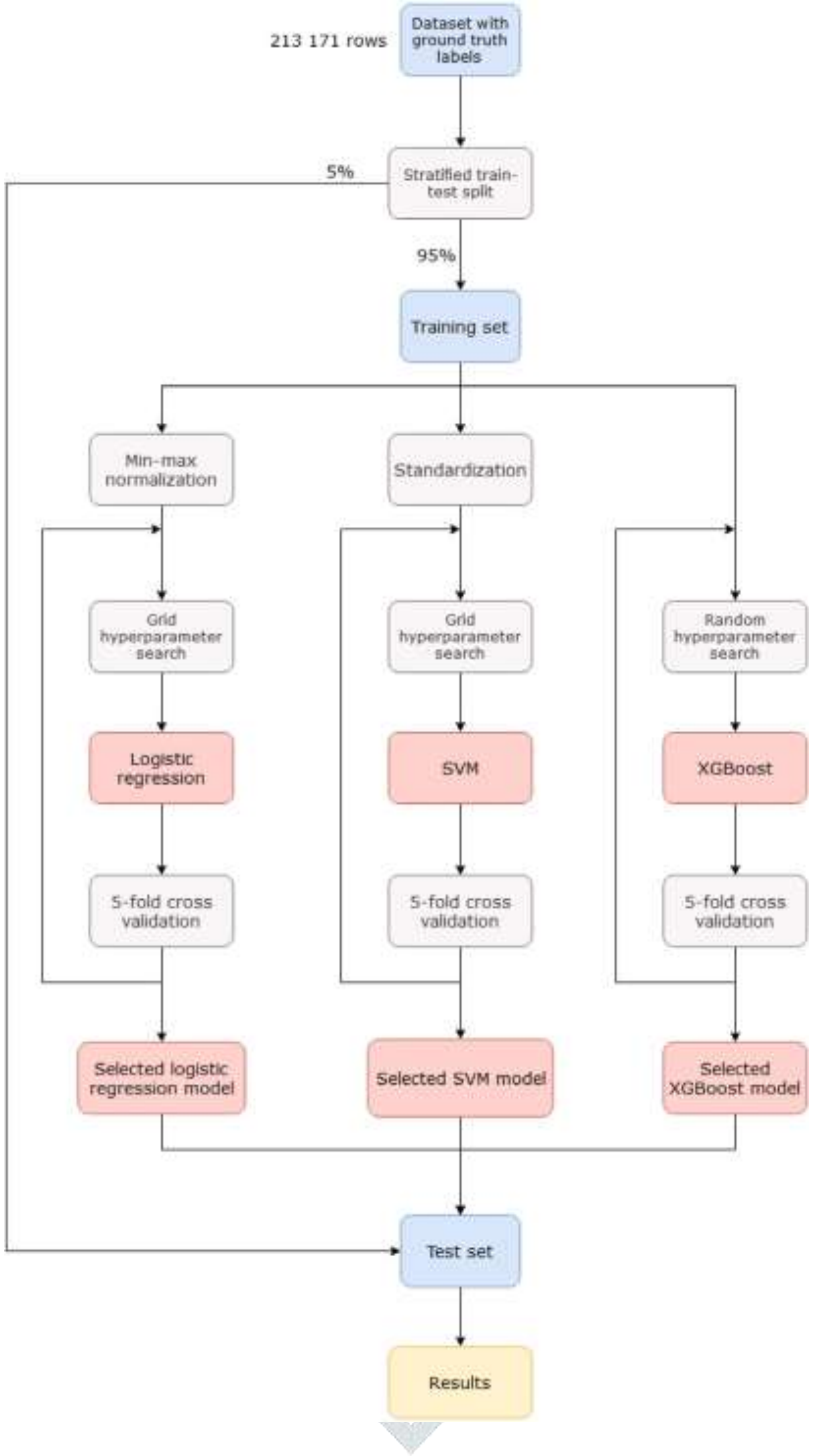
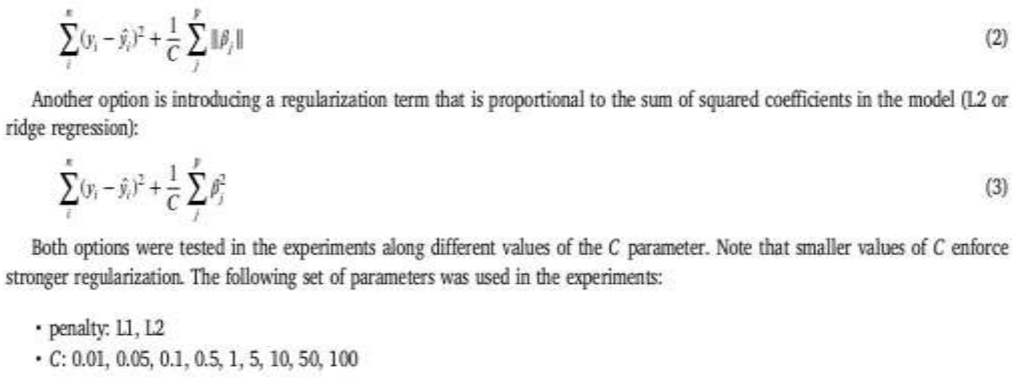


Fig. 7. Experiment setup.

5.5.1 LOGISTIC REGRESSION

When the values of the coefficients are excessively high, the logistic regression method typically overfits the training data (Jurafsky & Martin, 2014). In order to reduce the algorithm's cost function, the coefficients are changed throughout each training step. Utilising the mean As a cost function in logistic regression, squared error (MSE) is used. Regularisation, which involves including a component in the cost function whose value is proportionate to the coefficient values, prevents models from overfitting. With L1 or Lasso regularisation, the regularisation term may be proportional to the total magnitude of all the coefficients. Assume that yi is the ground truth value for the i-th instance, yii is the prediction for the i-th instance, n is the number of occurrences, j is the value of the j-th coefficient, and p is the total number of coefficients.The C parameter determines how strong the regularisation will be.Consequently, the cost function's formula is equivalent to:

Both options were tested in the experiments along different values of the *C* parameter. Note that smaller values of *C* enforce stronger regularization. The following set of parameters was used in the experiments:



$$\sum_{i}^{n}(y_i - \hat{y}_i)^2 + \frac{1}{C}\sum_{j}^{p}\|\beta_j\| \tag{2}$$

Another option is introducing a regularization term that is proportional to the sum of squared coefficients in the model (L2 or ridge regression):

$$\sum_{i}^{n}(y_i - \hat{y}_i)^2 + \frac{1}{C}\sum_{j}^{p}\beta_j^2 \tag{3}$$

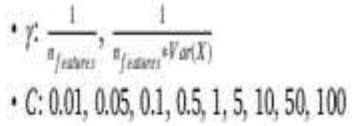
Both options were tested in the experiments along different values of the *C* parameter. Note that smaller values of *C* enforce stronger regularization. The following set of parameters was used in the experiments:

- penalty: L1, L2
- *C*: 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100

5.5.2 Support vector machine

Like logistic regression, SVM is a linear classifier. To divide classes into different groups, the algorithm seeks the best decision hyperplane. The regularisation of the decision function is controlled by the C parameter of the SVM algorithm. According to Pedregosa et al. (2011), its value is negatively correlated with the degree of regularisation. Lower C forces a simpler decision surface and a larger margin of safety for the classifier. This might cause a model to underfit the data. A model that has greater values for the C parameter has a tendency to accurately categorise all training data. Overfitting training data and poor test set generalisation might be a result of this (Cortes & Vapnik, 1995).

Using specialised functions referred to as kernels, SVM may be applied as a non-linear classifier (Smola & Schölkopf, 1998). The radial basis function kernel, utilised in this work, is one of them. The variable is in charge of regulating its behaviour. The model is compelled by high values of to build a decision surface based on examples that are near to one another. Low values of allow instances that are placed further away to have an impact on the decision surface (Pedregosa et al., 2011). Due to the computational difficulty of SVM's hyperparameter adjustment, we selected the following two values from the scikit-learn API:



- γ: $\frac{1}{n_{features}}$, $\frac{1}{n_{features}+Var(X)}$
- C: 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100

5.5.3 XGBoost

With CART serving as the fundamental model, we employed the XGBoost algorithm (Chen & Guestrin, 2016). We choose to adjust the ensemble's number of estimators. Since each tree only utilises a portion of the attributes, it makes sense that the larger the ensemble of trees, the more choices are examined. The colsample bytree parameter regulates the ratio of the number of features that are utilised to the total number of features. By adjusting a tree's maximum depth, we can also manage the number of levels that are permitted. The data is often overfit by deeper trees. The minimum child weight setting determines the minimum weight that must be added to the instance's weight in order to divide it. Higher values might stop a tree from growing sooner and avoid overfitting.

Contrary to other used algorithms, XGBoost provides a wider range of tuning options. Computational limitations prevented a complete grid search of all combinations. A few of the combinations may be tested using the randomised search approach. The best collection of parameters cannot be identified with certainty with this technique, but it is claimed to uncover models that may be on par with those discovered using grid search (Bergstra & Bengio, 2012). To determine the top-performing model, tests were conducted on 25 sets of the parameters listed below, which were randomly selected.

number of estimators: 100, 250, 500, 750, 1000

- maximum depth of a tree: 3, 5, 7, 10, 12, 15, 17, 20, 25

- minimum child weight: 1, 3, 5, 7

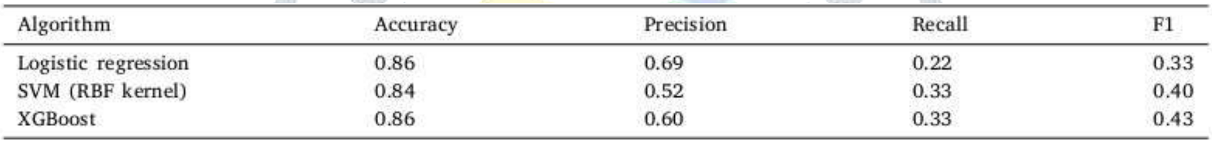
- gamma: 0.0, 0.1, 0.2, 0.3, 0.4

- colsample bytree: 0.3, 0.4, 0.5, 0.7

- learning rate: 0.05, 0.10, 0.15, 0.20, 0.25, 0.30

Out of all the evaluated combinations, the best performing set of parameters was discovered by a randomised search. The parameters of the chosen method were as follows: colsample bytree=0.5, gamma=0.1, learning rate=0.25, maximum depth=25, minimum child weight=5, number of estimators=750. According to the complete findings of the hyperparameter tweaking, performance might be further enhanced by increasing the number of estimators and tree depth. In a thorough grid search, the following set of parameters was examined using the model discovered in the random search procedure:

Table 3: Results of 5 fold cross validation on trainingset of selected models with highest f1 score after hyperparameter tuning,precision,recall and F1 score are reported for the positive class



Algorithm	Accuracy	Precision	Recall	F1
Logistic regression	0.86	0.69	0.22	0.33
SVM (RBF kernel)	0.84	0.52	0.33	0.40
XGBoost	0.86	0.60	0.33	0.43

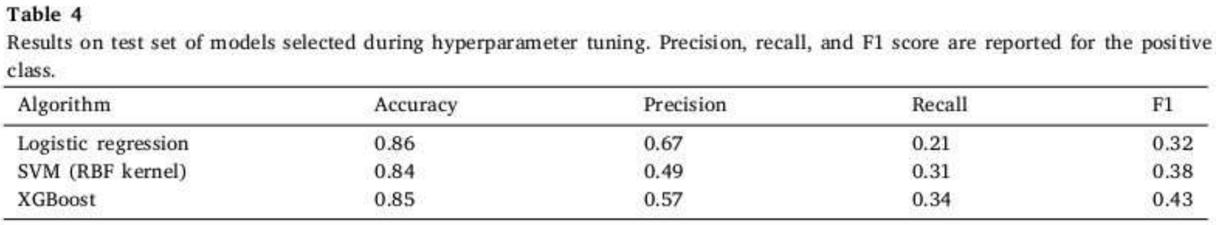


Table 4

Results on test set of models selected during hyperparameter tuning. Precision, recall, and F1 score are reported for the positive class.

Algorithm	Accuracy	Precision	Recall	F1
Logistic regression	0.86	0.67	0.21	0.32
SVM (RBF kernel)	0.84	0.49	0.31	0.38
XGBoost	0.85	0.57	0.34	0.43

5.4.1 Results of cross-validation on the training set

Based on how well the models performed on the validation subsets used in each fold, the models were chosen for cross-validation.

Table 3 displays the outcomes of the models with the greatest F1 scores.

The L2 penalty model with a C value of 100 produced the model with the greatest score in the logistic regression method. The outcomes resemble those of the model with default settings (see Table 2) rather closely. The F1 score did not much improve. The logistic regression model is straightforward and provides a starting point for more sophisticated models, but it does not give additional parameters to fine-tune its effectiveness.

The models that passed the cross-validation procedure were then put to the test using the test set of 10,000 representative samples. The test set results for the chosen models are displayed in Table 4 below. Comparing the performance to Table 3's cross-validation results, we only detect a very minor decline. The fact that the difference is so little illustrates how reliable the models are. When the models were validated and the hyperparameters were tuned, the training data was not overfit by the models.

The final results revealed an increase in recall at the expense of all tested models' accuracy declining. In the procedure, the logistic regression was least impacted. Similar performance is achieved with and without hyperparameter adjustment. Among the classifiers, the SVM algorithm's accuracy metric decreased the most while maintaining a recall that was comparable to XGBoost's. While keeping precision at 50%, the XGBoost algorithm was able to improve recall. It also outperformed other classifiers when the F1 score was taken into account. We were able to improve the classifiers' recall and F1 score through the hyperparameter tweaking method. The classifiers were able to identify more successful businesses among cases that had previously been incorrectly labelled as failures. Reduced accuracy was the price it demanded, which It indicates that the classifiers had a higher propensity to mistakenly label a failed business as successful.

5.5.4 Feature importance

We may examine the classification tree used by the XGBoost model as its base estimator so that we can better understand how the classifier makes decisions. It is feasible to plot a single decision tree, but it would be challenging to derive conclusions from the ensemble of 1250 trees. To visualise the variables that influenced the decision-making process, we may utilise the ensemble's aggregated data. The top-performing XGBoost model trained on the complete training set is shown in Fig. 8, and the top 10 features are shown there. A feature's F-score is determined by how frequently it appears in a tree.

The values of the characteristics that were used to make the categorization judgements cannot be ascertained from Fig. 8. This measure is utilised as an estimate of a feature's relevance since we know which features were most commonly employed to divide instances in leaves while constructing a tree. The size of the cities, the size of the regions, and the country code were among the most significant geographical characteristics. Governments work to create startup ecosystems in an effort to boost the number of profitable businesses in the market, as we shown in Section 4.3.2.

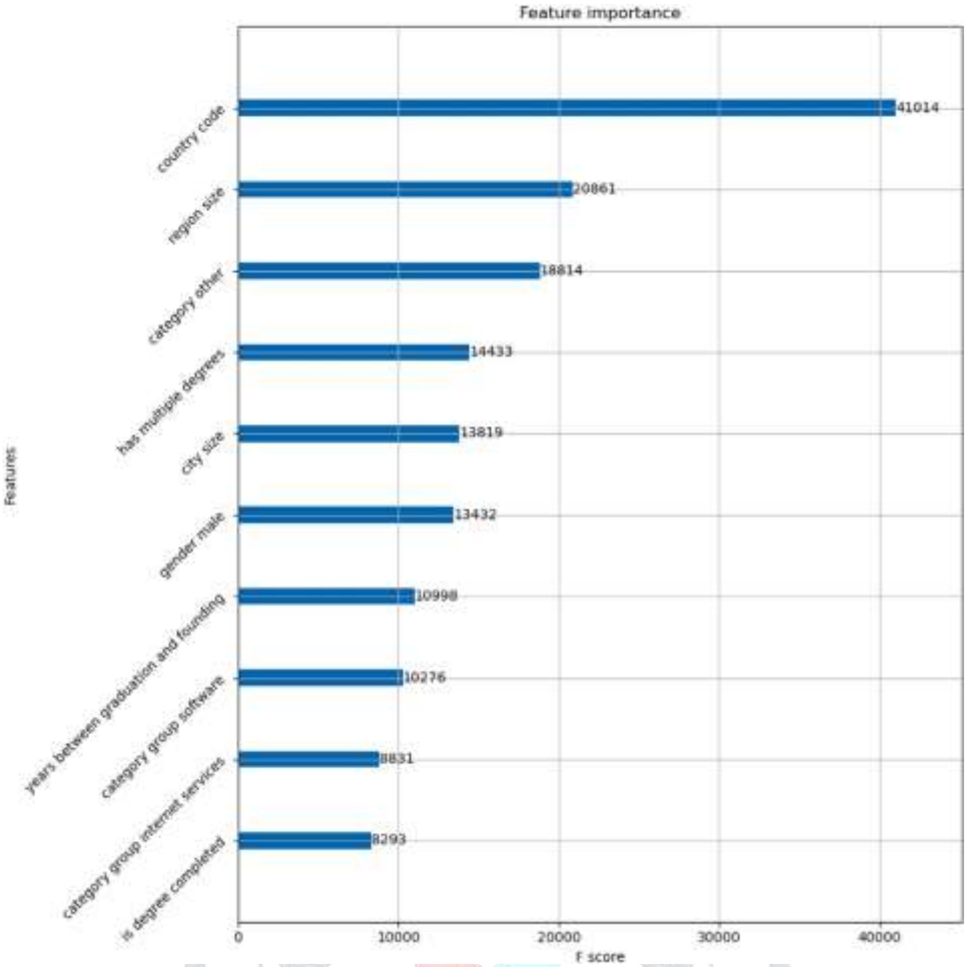


Fig. 8. 10 most important features in the selected XGBoost model. F-score is the number of times a feature appears in a tree.

6 CONCLUSIONS

Despite being a difficult endeavor, predicting company success is important to many public and private stakeholders who influence economics, decide where to invest money, and form businesses. It makes sense that as a firm develops, evaluates its product-market fit, and goes through the selection procedures for angel investors and venture capital funds, the work gets simpler. By limiting the range of attributes to regional, demographic, and fundamental company data, we suggested a machine learning technique for forecasting business success at the early stage. We did not utilize any information on external financing, even though it may have been accessible, unlike past studies. The key benefit of the study is that its findings can be replicated in real-world situations. We were able to do that by carefully limiting the dataset to just include variables that would be relevant at the moment of choice. We think that a decision-support system like this might be helpful in venture capital funds to uncover potentially profitable businesses that might otherwise go overlooked. For prediction purposes, the proposed technique, however, does not account for the most current data. It may be argued that businesses created recently differ greatly from older businesses in terms of the success-producing tendencies. This problem may be solved easily by either limiting the dataset to the more recent organizations or giving them greater weight throughout the learning process.

We demonstrated how a nation's fiscal policies, investment in innovation, and development of a system supporting the teaching of new technologies increases the number of startups per million people by analyzing chosen data from the Crunchbase database. Politicians that want to boost the influence of the new technologies sector on economies may find inspiration in Estonia's and Israel's experiences. We also demonstrated that the startup industries with the highest launch rates, such as software and internet services, do not always have the highest success rates. The odds of success are the highest among the most well-liked industries, yet starting a business in the healthcare or manufacturing sectors may demand more substantial upfront capital than in the software sector. But it's plausible that software companies have a lower success rate than businesses in other sectors because of a comparatively low entry barrier. Analyzing the rivalry in certain sectors is one more technique to approach this issue. In this way, our study backs up Stuart and Abetti's (1987) theory that slower-growing, less dynamic markets are better for starting a successful business.

We created machine learning models to forecast company success and evaluated the effectiveness of three different algorithms: logistic regression, SVM, and XGBoost. To the best of our knowledge, this is the first analysis of Crunchbase data using the XGBoost method, which has been well-known recently due to its impressive results in machine learning contests. By fine-tuning hyperparameters, we were able to improve the recall and F1 score over our original studies. Comparing the cross-validation and test set results demonstrates the trained models' robustness and generalizability. Out of the investigated methods, the XGBoost model achieved the best accuracy, recall, and F1 score values, which were 57%, 34%, and 43%, respectively, for the best model. The outcomes demonstrate the value of the XGBoost algorithm in upcoming commercial success prediction applications.

Precision was greater than recall across the board for all models. This indicates that they were able to identify just the organizations that adhered to the most common success patterns. An analysis of feature significance revealed that the top-performing XGBoost model commonly employed a startup's location and mode of operation in sectors like software or Internet services while constructing decision trees.

By enhancing the dataset, further research may improve the models' recall. The performance of models might be enhanced by include more specific information on the founder's previous experience and the company's product or service. In order to do this, more sources of information besides Crunchbase would need to be included. Although we chose against using the subjective descriptions of individuals and businesses, the text data might be investigated as an alternative source of characteristics for the dataset. Taking periodic snapshots of the Crunchbase database is an alternative strategy. A dataset of this kind might offer potentially useful details regarding the mechanics of the business's expansion. It could be modelled with time-series methods. Our research, we feel, provides deeper understanding of the startup sector and offers reliable machine learning methods for forecasting commercial success.

REFERENCES:

[1] Patrick Bajari, Denis Nekipelov, Stephen P Ryan, and Miaoyu Yang. Machine learning methods for demand estimation. American Economic Review,105(5):481–85, 2015.

[2] Kris Johnson Ferreira, Bin Hong Alex Lee, and David Simchi-Levi. Analytics for an online retailer: Demand forecasting and price optimization. Manufacturing & Service Operations Management, 18(1):69–88, 2016.

[3] Ankur Jain, Manghat Nitish Menon, and Saurabh Chandra. Sales forecasting for retail chains, 2015.

[4] Grigorios Tsoumakas. A survey of machine learning techniques for food sales prediction. Artificial Intelligence Review, 52(1):441–447, 2019.

[5] Xiaogang Su, Xin Yan, and Chih-Ling Tsai. Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4(3):275–294, 2012.

[6] Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. Journal of the american statistical association, 83(404):1023–1032, 1988.

[7] Zheng Li, Xianfeng Ma, and Hongliang Xin. Feature engineering of machinelearning chemisorption models for catalyst design. Catalysis today, 280:232–238, 2017.

[8] Xinchuan Zeng and Tony R Martinez. Distribution-balanced stratified crossvalidation for accuracy estimation. Journal of Experimental & Theoretical Artificial Intelligence, 12(1):1–12, 2000.

[9] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 145–158. Springer, 2011.

[10] Chris Rygielski, Jyun-Cheng Wang, and David C Yen. Data mining techniques for customer relationship management. Technology in society, 24(4):483–502, 2002.

[11] Krzysztof J Cios, Witold Pedrycz, Roman W Swiniarski, and Lukasz Andrzej Kurgan. Data mining: a knowledge discovery approach. Springer Science & Business Media, 2007.

[12] Maike Krause-Traudes, Simon Scheider, Stefan Rüping, and Harald Meßner. Spatial data mining for retail sales forecasting. In 11th AGILE International Conference on Geographic Information Science, pages 1–11, 2008.

[13] Stephen Marsland. Machine learning: an algorithmic perspective. CRC press, 2015.

[14] ML documentation. https://www.mathworks.com/discovery/ machine-learning.html). Accessed: 2020-04-22.

[15] Ethem Alpaydin. Introduction to machine learning. MIT press, 2020.

[16] Arvin Wen Tsui, Yu-Hsiang Chuang, and Hao-Hua Chu. Unsupervised learning for solving rss hardware variance problem in wifi localization. Mobile Networks and Applications, 14(5):677–691, 2009.

[17] Bohdan M Pavlyshenko. Machine-learning models for sales time series forecasting. Data, 4(1):15, 2019.

[18] Taiwo Oladipupo Ayodele. Types of machine learning algorithms. New advances in machine learning, pages 19–48, 2010.

[19] Sanford Weisberg. Applied linear regression, volume 528. John Wiley & Sons, 2005.

[20] Gradient Boosting documentation. https://turi.com/learn/userguide/ supervised-learning/boosted_trees_regression.html). Accessed: 2020- 05-19.

[21] JN Hu, JJ Hu, HB Lin, XP Li, CL Jiang, XH Qiu, and WS Li. State-of-charge estimation for battery management system using optimized support vector machine for regression. Journal of Power Sources, 269:682–693, 2014.

[22] Wangchao Lou, Xiaoqing Wang, Fan Chen, Yixiao Chen, Bo Jiang, and Hua Zhang. Sequence based prediction of dna-binding proteins based on hybrid feature selection using random forest and gaussian naive bayes. PloS one, 9(1), 2014.

[23] İrem İşlek and Şule Gündüz Öğüdücü. A retail demand forecasting model based on data mining techniques. In 2015 IEEE 24th International Symposium on Industrial Electronics (ISIE), pages 55–60. IEEE, 2015.

[24] Takashi Tanizaki, Tomohiro Hoshino, Takeshi Shimmura, and Takeshi Takenaka. Demand forecasting in restaurants using machine learning and statistical analysis. Procedia CIRP, 79:679–683, 2019.

[25] Xu Ma, Yanshan Tian, Chu Luo, and Yuehui Zhang. Predicting future visitors of restaurants using big data. In 2018 International Conference on Machine Learning and Cybernetics (ICMLC), volume 1, pages 269–274. IEEE, 2018.

[26] Mikael Holmberg and Pontus Halldén. Machine learning for restaurant sales forecast, 2018. References 34

[27] I-Fei Chen and Chi-Jie Lu. Sales forecasting by combining clustering and machine-learning techniques for computer retailing. Neural Computing and Applications, 28(9):2633–2647, 2017.

[28] Malek Sarhani and Abdellatif El Afia. Intelligent system based support vector regression for supply chain demand forecasting. In 2014 Second World Conference on Complex Systems (WCCS), pages 79–83. IEEE, 2014.

[29] Jason Brownlee. Introduction to time series forecasting with python: how to prepare data and develop models to predict the future. Machine Learning Mastery, 2017.

[30] Python history. https://en.wikipedia.org/wiki/Python_(programming_

language). Accessed: 2020-04-29.

[31] Guido Van Rossum et al. Python programming language. In USENIX annual technical conference, volume 41, page 36, 2007.

[32] Travis E Oliphant. A guide to NumPy, volume 1. Trelgol Publishing USA, 2006.

[33] Wes McKinney. Pandas, python data analysis library. see http://pandas. pydata. org, 2015.

[34] Niyazi Ari and Makhamadsulton Ustazhanov. Matplotlib in python. In 2014 11th International Conference on Electronics, Computer and Computation (ICECCO), pages 1–6. IEEE, 2014.

[35] Raul Garreta and Guillermo Moncecchi. Learning scikit-learn: machine learning in python. Packt Publishing Ltd, 2013.

[36] Seaborn documentation. https://seaborn.pydata.org/introduction. html). Accessed: 2020-04-26.

[37] Chung-Jui Tu, Li-Yeh Chuang, Jun-Yang Chang, Cheng-Hong Yang, et al. Feature selection using pso-svm. International Journal of Computer Science, 2007.

[38] Tao Zhang, Tianqing Zhu, Ping Xiong, Huan Huo, Zahir Tari, and Wanlei Zhou. Correlated differential privacy: Feature selection in machine learning. IEEE Transactions on Industrial Informatics, 2019.

[39] Pearson documentation. https://en.wikipedia.org/wiki/Pearson_ correlation_coefficient). Accessed: 2020-04-25.

[40] Feature Importance documentation. https://machinelearningmastery. com/calculate-feature-importance-with-python/#:~:text=Feature% 20importance%20refers%20to%20a,feature%20when%20making%20a% 20prediction.). Accessed: 2020-06-06.

[41] Kedar Potdar, Taher S Pardawala, and Chinmay D Pai. A comparative study

of categorical variable encoding techniques for neural network classifiers. International journal of computer applications, 175(4):7–9, 2017.

[42] Cross validation documentation. https://towardsdatascience.com/ cross-validation-explained-evaluating-estimator-performance-e51e5430ff85). Accessed: 2020-04-28.

[43] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808, 2018.

[44] Accuracy documentation. https://medium.com/thalus-ai/ performance-metrics-for-classification-problems-in-machine-learning-part-i-b085 Accessed: 2020-05-01.

[45] scilearn max error. https://scikit-learn.org/stable/modules/model_ evaluation.html#max-error. Accessed: 2020-05-10