Predicting the success of tech startups in South Africa

Thabo Rachidi
School of Computer Science and Applied Mathematics
University of the Witwatersrand
South Africa
1632496@students.wit.ac.za

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page limits.

II. Related Work

A. Challenges and issues faced by startups

he reported statistics indicate a high failure rate of startups with 9 out of 10 startups failing [1]. Every entrepreneur wishes to see their innovative idea succeed. Some of the few factors that can increase their chances of success include traction, access to capital, management, skilled individuals, a viable product and marketing [1]. Deciding what constitutes as a successful startup or not is essential to the study. [2] states that a startup can be considered successful if its current state is either acquired or has issued an IPO[1]. [3] have similar ideas about what state a startup would need to be in, in order to be considered successful with the inclusion of having completed a Series B round of funding.

The impact of the challenges faced by startups during development can be presented using four holistic dimensions which are team(main driver of development), product(an innovative developed solution),financial and market(the needs of the final customers) [4]. [5] states that the model for evolution of product development for a startup consists of three phases, namely, startup(time between conception and first sale), stabilisation(from when first customer receives product until when the product is stable enough to be commissioned without any overhead on product development) and growth(which starts when the product is stable enough to be commissioned without any overhead on product development and ends when the market size, share and growth rate have been established).

[1]Initial Public Offering(IPO): Is when a private company offers its shares to the public.

Taking the above points into consideration we can list and categorise the most common challenges and issues faced by startups in the tables below.

TABLE I
Challenges/Issues most commonly affected by startups during the startup phase.

Startup phase	
Challenges/Issues	Dimension
Inexperienced developers	Team
Product isn't really a product	Product
Product has no owner	Product
No strategic plan for product development	Product
Unrecognised product platform	Product
Acquiring first paying customer	Market
Acquiring initial funding	Financial
Having entrepreneurial teams	Team
Defining a minimum viable product	Product

TABLE II
Challenges/Issues most commonly affected by startups during the stabilisation phase.

Stabilisation phase	
Challenges/Issues	Dimension
Founders won't let go	Team
Development team fails to gel	Team
Product is unreliable	Product
Requirements become unmanageable	Team
Product expectations are too high	Product
Service provision delays development	Team
Managing multiple tasks	Team
Targeting a niche market	Market
Focus and discipline	Team
Reaching break-even	Financial

B. Data collection

Previous literature relied on data collected from surveys which were limited in their use for prediction as they were only able to collect a few sample points. Machine learning algorithms were the preferred way over a more statistical approach, which would need a large amount of data to predict startup success [2]. A widely used resource for the collection of large scale data of startups that was common among literature,

TABLE III
CHALLENGES/ISSUES MOST COMMONLY AFFECTED BY STARTUPS DURING THE GROWTH PHASE.

Growth phase	
Challenges/Issues	Dimension
Skills Shortage	Team
Product pipeline is empty	Product
Platform creep delays development	Product
No process for product introduction	Market
Thriving in uncertainty	Market
Creating customer value	Market

such as [2] and [1] , is a website called Crunchbase[2]. Other sources of data that were used were Mattermark, Dealroom, Forbes and Techcrunch. Additionaly, [3] crawled websites using web scrapers to collect data about whether the startups' websites were still operational by testing their HTTP status code. Although [4] was able to get 5389 survey responds it does not compare to the amount of data [2] and [3] were able to collect which was 44 522 and 213 171 startups respectively.

A huge issue that was found is the amount of bias there was in the data collected. [3] had 87.8% of the collect data belonging to the failure class and 12.2% belonging to the success class.Thus the authors ,including [2], handled the class imbalance in the data by oversampling the minority sample and creating artificial data. This was achieved using Adaptive Synthetic Sampling Approach (ADASYN) which is based on the Synthetic Minority Over- sampling Technique (SMOTE) concepts.

The survival of a startup involves a number of key factors that may be huge determinants of their success. The table IV below lists the most common factors that were included as features in their data set.

A lot of data had to be removed during the cleaning phases such as data points that had missing values. Removing some features is a highly discussed topic in literature mainly because of its benefits [2]. It helps in reducing the computation time and complexity of the machine learning models.It is also essential that data from a certain date not be included when collecting the data, as the company would not be recognised as a startup anymore.

C. Review of classification techniques

There are two types of prediction models being used in literature; statistical models and intelligent models(machine learning) [2]. Early studies in literature focused on using statistical modelling to make predictions. These statistical models were used to make accurate predictions in the context of financial decisions [2]. Thus these early studies did not put any focus on the ability and experience of the development team. Machine learning models became more popular during the recent few decades. Both methods essentially allow us to learn from data, but machine learning algorithms are not bound by rule-based programming.

[2]www.crunchbase.com

TABLE IV
COMMON FEATURES AND A DESCRIPTION.

Features	Description	Type
Founding date	When the startup was founded i.e. start date.	Numeric
Seed funding	Initial funds raised by the startup.	Numeric
Time for seed funding	The months it took the startup to raise the seed funds	Numeric
Rounds of Funding	The number of funding rounds the startup raised	Numeric
Last funding	Years that have passed between last funding and current year.	Numeric
Valuation	The amount of funds the startup raises at each round of funding	Numeric
Defunct date	Date when the startup closed down (only applies to failed companies)	Numeric
Severity factors	Factors responsible for the startups failure/success	Categorical
Months Active	Total number of months the startup has been active in market	Numeric
Market value	The current market value of the startup	Numeric
Total funds	The total amount of funds it has received to date (Seed funds + Venture funds)	Numeric
Burn rate	It is calculated as total funds/months active	Numeric
Knowledge support	Type of support the startup received e.g. incubator, accelerator etc.	Categorical
Web analytics	Number of website visits, duration of visit, bounce rate etc.	Numeric
Social media analytics	Number of followers,posts and sentiment analysis of tweets	Numeric
Revenue Model	The model used to make money e.g. commisions, advertising etc.	Categorical
Number of Employees	The number of team members or people employed at the startup.	Numeric
Contact details	Yes/No indicating if there is an email address or contact number.	Boolean

Machine learning is an application of Artificial Intelligence (AI) and is based on the theory that computers can learn and perform certain tasks without being explicitly programmed to do so. Machine learning was born out of a subject called pattern recognition, thus they learn to predict from previous computations. Many of the machine learning algorithms have been proved to outperform statistical models [2]. In this research, machine learning algorithms are used to predict the chance of success of a startup.

[6] state that artificial intelligence machines usually fail in 2 kinds of situations; firstly when interpreting information that cannot be quantified and secondly when making predictions in situations of extreme uncertainty. While humans intuition is still the best when it comes to predictions in the situations previously listed. A single persons decision can be highly biased as people rely on heuristics like mental shortcuts to make decisions. Hence the idea of a collective intelligence which is essentially the "wisdom of crowds". A Hybrid Intelligence method can be used to combine collective

intelligence and machines learning. The limiting factor of the Hybrid Intelligence method is that interviews and group focus workshops would be needed and thus may not allow the large data required for effective machine learning models.

The table V below shows the more common machine learning algorithms that were used in various literature.

TABLE V
MACHINE LEARNING ALGORITHMS.

Model	Reference
Random Forest	[1], [2], [3], [6]
Naive Bayes	[1], [6]
Logistic Regression	[1], [2], [6]
Extreme Gradient Boosting	[2], [3]
Support Vector Machine	[3] , [6]

Implementation of machine learning algorithms involves randomly splitting the data into a training and testing data sets. The training data is the subset of the data that will be used to create the prediction models. [6] chose to use a 10-fold cross validation approach to learning by splitting the training data into 10 mutually exclusive and approximately equally sized subsets. Its main aim is to reduce the bias that comes with randomly sampling the training data. Thus each model was trained using 10 subsets and tested using 1 subset.

During [1] experiments with their chosen models the WEKA toolkit was used for classification, analysis and modeling. Different milestones were created so that different models could be create labelled M0 to M9. Initially 30 different classification machine learning algorithms were used but only the top 6 were selected. Experiments were then conducted with the 6 algorithms on each of the models. Leave-One-Out Cross Validation (LOOCV) was used for evaluation which in the case of [1] is equivalent to a 11000-fold cross validation.

To ensure that models do not overfit to a validation set [3] performed cross validation on the training set. Cross validation is recommended during hyperparameter tuning to reduce selection bias and overfitting. The data was then preprocessed using feature scaling such as min-max normalisation for logistic regression and standardisation for Support Vector Machine. The advantages of feature scaling are that we can improve the performance of the algorithms and reduce the time it takes the algorithms to converge. The final models were chosen after hyperparameter tuning. These are the models used on the entire training set and then tested on the test set.

D. Analysis of results

[6] used a balanced performance measure for binary classification called Matthews correlation coefficient (MCC). It is a relevant measure for biased data as the data they collected contained a large amount of successful startups with a low amount of failed startups. They then compared the performance of a logarithmic regression algorithm, as their baseline, with each machine learning algorithm, a crowd prediction and then a weighting algorithm using a two-way analysis of variance (ANOVA).

[1], [2], and [3] used a combination of the following performance metrics to analyse the results of their final models: area under ROC curve (AUC), accuracy and precision and recall. The table VI below shows a comparison of the top performing machine learning algorithms and the relevant metrics.

TABLE VI
COMPARISON OF THE PERFORMANCE OF THE MACHINE LEARNING MODELS

Author	Model	AUC	Accuracy	Precision	Recall
[1]	Random Forest	96%	-	97%	97%
[1]	Logistic regression	98%	-	95%	94%
[1]	ADtrees	98%	-	95%	93%
[2]	Full logistic regression	85%	77%	-	-
[2]	Reduced logistic regression	85%	77%	-	-
[2]	Random Forest	92%	94%	-	-
[2]	Extreme gradient boosting	93%	94%	-	-
[3]	Logistic regression	-	86%	67%	21%
[3]	Support Vector Machine	-	84%	49%	31%
[3]	Extreme gradient boosting	-	85%	57%	34%

To get the area under ROC curve (AUC) we can take the integral of a ROC curve between 0 and 1. The ROC curve plots the true positive rate vs the false positive rate at all classification thresholds. High recall means that the algorithm was able to return more relevant results while high precision means it was able to return substantially more relevant than irrelevant results.

[2] also included Type I error, Type II error and confusion matrices when evaluating the performance of their models. Type I error occurs when a model misclassifies a successful startup as failed(false negative) and Type II error is when a failed company is misclassified as successful. A discussion of variable importance was included as it can be used to explain the model to achieve a better understanding of what its doing.

III. METHODOLOGY

A. Data

The data used to build the models was taken from Kaggle[3]. It contains roughly 49438 datapoints with 39 features which are permalink, name, homepage url, category list , market, total funding(USD), status, country code, state code, region, city, funding rounds, founded date, founded month, founded quarter, founded year , first funding date, last funding date, seed, venture, equity crowdfunding, undisclosed, convertible

[3]Startup investments crunchbase dataset https://www.kaggle.com/arindam235/startup-investments-crunchbase

note, debt financing, angel, grant, private equity, post IPO equity, post IPO debt, secondary market, product crowdfunding, round A, round B, round C , round D, round E, round F, round G and round H.

1) Preprocessing: Missing and any duplicated data was identified in the dataset so that they could be analysed further. Firstly the following columns were removed, as they were not useful when trying to predict if a startup will be successful or not. The columns were permalink, homepage url, category list, state code, region, founded date, founded month, founded quarter, first funding date, last funding date and city. The startups with no name were removed as they could not be identified without the company name. Startups that had a missing status were also removed from the dataset. The missing values for market and country code were replaced with other and the missing founded year was replaced with the mean founded date of the dataset which was 2007.

Finally a target column was created using the features of the dataset. The target was created as follows:

$$y = \begin{cases} 1 & (\text{status} == \text{acquired OR post IPO equity} > 0) \\ & \text{OR}(\text{status} == \text{operatingANDround b} > 0) \\ 0 & otherwise \end{cases}$$

[3]

An IPO means that the company has issued a publicly available valuation, which indicates a level of success. An acquisition is also another indication of success as it usually means that investors get a return on investment. Startups that are still operating and have managed to receive a round B of funding are also marked as successful. Although we can not be sure that they will be successful in the future, a round B of funding usually indicates high investor trust in the start [3].

To ensure that the dataset could be used by a classification model the categorical features needed to be converted to a numerical representation. This can be done using two methods a one-hot encoder or a label encoder. The one-hot encoder encoders each categorical feature as a one-hot numerical array. This greatly increases the dimension of the dataset and be costly during training of the classification models. A label encoder encoders the categorical feature with a value between 0 and the number of unique values of the features minus one. Both Encoding techniques were used on the dataset so that their performances could be monitored. Thus the preprocessing process results in a total of 49437 datapoints with 896 features for the one-hot encoder dataset and 29 features for the label encoder. The dataset is imbalanced with 83% of the datasets being label as not successful and 17% being labeled as success.

B. Models

1) Logistic Regression: A baseline Logistic Regressing Model was used as a model to predict if a startup is successful or not. The first Logistic regression model was created using scikit-learn's[4] parameters for the model which was using a L2 penalty term, and uses liblinear solver. Another model using 10-fold cross validation was also created using the same parameters as the baseline model.

2) Naive Bayes: A Gaussian Naive Bayes and a Complement Naive Bayes model were used for predicting. The Gaussian Naive Bayes classifier assumes that the likelihood of the features is Gaussian. The Complement Naive Bayes Model implements the complement naive bayes algorithm which is an adapted form of the multinomial naive bayes that is particularly suitable for imbalanced datatsets. The complement naive bayes algorithm works by using statistics from the complement of each class to compute the model's weights. Both models were trained using 10-fold cross validation with the default parameters set by scikit-learn.

3) Support Vector Machines: The main idea behind Support Vector Machines (SVM) is to find a maximum marginal hyperplane efficiently divides the dataset into classes. The SVM algorithm uses a kernel to transform the input data space in the required form. There are different types of kernels that could be used to this such as a linear kernel, polynomial kernel and a radius basis function kernel. The SVM model was trained using 10-fold cross validation and parameter tuning using a grid search technique. The best parameters that were used are : regularization parameter(C) of 100, gamma was 1, and the kernel was the radius basis function kernel.

For all of the models used on the dataset a random state of 21 was used to ensure reproducibility of the models.

C. Evaluation

To evalaute the performance of the models the accuracy and confusion matrics were used.

IV. Results

V. Conclusion

References

[1] A. Krishna, A. Agrawal, and A. Choudhary, "Predicting the outcome of startups: Less failure, more success," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 12 2016.

[2] C. Ünal, "Searching for unicorn: A machine learning approach towards startup success predicition," Masters, Huboldt-Universität zu Berlin, 07 2019.

[3] K. Źbikowski and P. Antosiuk, "A machine learning, bias-free approach for predicting business success using crunchbase data," *Information Processing and Management*, 2021.

[4] C. Giardino, S. S. Bajwa, X. Wang, and P. Abrahamsson, "Key challenges in early-stage software startups," in *Agile Processing Software Engineering and Extreme Programming*, vol. 212, 05 2015, pp. 52–63.

[5] M. Crowne, "Why software product startups fail and what to do about it. evolution of software product development in startup companies," in *IEEE International Engineering Management Conference*, vol. 1. IEEE, 2002, pp. 338–34.

[6] D. Dellermann, N. Lipusch, P. Ebel, K. M. Popp, and J. M. Leimeister, "Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method," in *International Conference on Information Systems (ICIS)*, 12 2017.

[7] G. G. Kingdon and J. Knight, "Unemployment in south africa: The nature of the beast," *World Development*, vol. 32, no. 3, pp. 391–408, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0305750X03002407

[8] J. J. Chengalroyen, "A computational model to predict the organisational performance of startups in south african incubators," Masters, University of the Witwatersrand, 2018.

[4]Scikit-learn is a machine learning library for Python https://scikit-learn.org/stable/index.html