Predicting Saudi Startups Success Using Machine Learning

Completed research paper

Zahra Al Huraiz, Reda Alhamza, Khaled Halawani

King Fahd University for Petroleum and Minerals

Dhahran, 31261, Kingdom of Saudi Arabia

Supervised by: Dr. Mousa Al Bashrawi

26th Sep 2022

Introduction

A startup describes a business that are still in its formative phases and founded by one or more individuals to penetrate an existing market with distinctive goods or services. Majority of startups' first funding comes from a variety of different private types of finance. However, investors will not fund any business, startups must first meet the funding conditions. Beginning with the viability of the idea to be converted into a profitable business. Secondly, evaluating if the startup has a long-term growth potential or has already grown quickly. Finally, assessing if the startup appears positive signs to expand till it reaches the unicorn stage (unicorn is the startup that has a valuation equals to or more than $1B) (Embroker, 2022). On the other hand, Venture capital firms that fund the startups are considered private equity companies that act as a bridge between investors who wish to invest their money and startups that need the money to boost their business. Hence, they pave the way for startups to get funded.

This study has a good fit to be conducted in Saudi Arabia as the Kingdom is going through a radical transformation to support the promising 2030 vision, which focuses on releasing new prospects and unlocking new opportunities at an unprecedented pace to diversify Saudi Arabia's sources of revenue and relief the dependence on oil income. Therefore, it is important to use artificial intelligence tools to develop strategies that can make Saudi Arabia an attractive investment destination and so cease on empowering the investors and VCs to access investment opportunities as well as provide them with facilities, flexibility, and logistics services (GOV.SA, 2021). Aligning with this effort, the Saudi government represented in Public Investment Fund (PIF) established a company called "Jada Fund of Funds" intending to increase opportunities for small and medium-sized enterprises (SMEs) to have access to capital by investing in private equity funds and venture capital. Jada has invested nearly SAR 1,150 million in 14 investment funds. This contributes in creating an attractive investment environment for both entrepreneurs and venture capitalists (Fund, 2022). Additionally, the government has established its own Saudi Venture Capital (SVC) in 2018 to participate in developing the startup ecosystem by investing in funds and decreasing funding gaps for startups. (SMEs) (SVC, 2021). Overall, new businesses both large and small, are responsible for most of Saudi's business advancements, economic flexibility, and growth, as well as for creating new job opportunities (Vision, 2021, P.36).

Startup ecosystem has been characterized by a high degree of uncertainty and volatility, which makes it challenging for venture capital firms to conduct the due diligence in order to evaluate the current performance and future potential success for startups. Contrary, traditional business companies show more stability. Therefore, the forecast of future success of such traditional companies are easier than startups by referring to their previous financials, key performance indicators (KPIs), sales, and production statistics. Since none of these statistics are available for startups the likelihood of a startup company becoming successful is dependent on a variety of different factors and metrics. Hence, using a quantified model that permits a machine learning approach can help confront a different set of business risks and uncertainties to forecast the success of startups.

Using conventional ways of evaluating startups success may cause a great risk to venture capital and other investors in early-stage organizations, as over ninety percent of new businesses fail, while ten to twenty two percent fail during the first year of operation (Carter, 2021). Therefore, there has been an effort to develop machine learning models that are trained on the historical performance of over one million different organizations to determine which businesses have a greater chance of becoming successful (Ross, 2021). We address the same subject in this paper by suggesting that VC firms may be able to benefit from using machine learning to screen potential investments using publicly available information. Few studies suggest that an ensemble approach can predict the exit scenario of a startup with over four times the accuracy of a successful venture capitalist, confirming the strength of machine learning's models (Cemre, 2019; Krishna et al., 2016).

Our study aims to develop a machine learning prediction model geared towards the success of startups. This would enable venture capital firms to investigate and invest in businesses that have a high potential for success in their respective industries. Ultimately this brings the attention to our study question; how incorporating a machine learning approach in VCs can assist in the prediction of startups success?

The study contributes by, first, using the power of machine learning which allows venture capital firms to adjust to the ever-shifting realities of the market and to improve the startups evaluation process. machine learning can help venture capital firms to acquire a deeper comprehension regarding the necessities that each startup should have to be subject for investment Second, the developed model can elevate venture capital firms' abilities and contribute to minimizing the amount of time spent on the assessment of the deal flow, reducing the amount of risk associated with investments, and increasing the accuracy of management decision making. Overall, the study could help to build a predictive classification model for Saudi startups' success.

Subsequent sections in the research begin with defining the study problem in the introduction section then reviewing related works in the literature review section, followed by the study methodology used and analysis conducted. Finally, the research concludes with discussion and insights to present recommendations to industry and future outlets for the research community.

Literature Review

According to Arroyo et al. (2019), venture capitals invest in startups to seize investment opportunities despite the high difficulty in deciding which startups to fund due to their unstable nature and lack of robust evaluation tools in VCs. Ries (2011) defines startup as "a human institution designed to create a new product or service under conditions of extreme uncertainty". Moreover, Embroker (2022) elaborates further on the motives for VCs to bet on high risk startup, as it usually leads to high return on investment specially if the startup succeeds in reaching at least $1B evaluation and became a unicorn. Currently, more than 600 unicorn startups exist around the world and are valued at about $2 trillion. Therefore, Unicorn startup covers the expenses for the fund with a remarkable revenue to both VC and investors providing the fund (Embroker, 2022).

There has been a plenty of studies that have use machine learning techniques to predict startups success, for example, Bento (2017) predicted startups success by analyzing 495,798 startups retrieved from Crunchbase database and represent different countries. The size of the dataset helped to tackle the issue of using a small sample and avoid the usage of financial metrics that are hard to obtain for startup's early stage. In his study, he applied several machine learning, but selected random forest algorithm to implement the model as it achieved a true-positive rate of 94.1%, a false-positive rate of 7.8%, and a precision of 92.2%, which indicates a better performance than logistic regression and the other tested model algorithms. Da Silva (2016) developed a prediction model for the success and failure of 50 Portuguese startups by choosing fourteen variables that include founder personality, capital, some characteristics of the startup, and external factors. He selected logistic regression to implement the model since it achieved an overall accuracy of 82%, sensitivity of 84.85%, and specificity of 76.47%. It's important to mention that the characteristics of the founder(s) and external factors are highly significant variables in predicting the success of Portuguese startups, while marketing expertise has a negative correlation with their success (Da Silva, 2016).

In the same vein, Dellermann et al. (2017) predicted the success probability at an early stage of startups that are looking for series A funding. In their approach, they used a hybrid intelligence method by combining humans' capabilities and machine learning algorithms for binary classification, while using dataset from Crunchbase, Mattermark, and Dealroom. They . With a dataset of of 1,500 startups, they applied multiple algorithms, for instance logistic regression, support vector machine, random forest, and artificial neural network and claimed this hybrid methodology provides much better results than using one of the approaches. Cemre (2019) implemented a predictive model to classify a startup as successful (operating startups) or failure (closed start-ups). He compared six models such as random forest, recursive partitioning tree, logistic regression, conditional inference tree, and XGBoost by analyzing Crunchbase detest of 44,522 startups from different areas around the world and selected Extreme Gradient Boosting (XGBoost) as the best-performing model, with an accuracy of 94.45%, a sensitivity of 97.53%, and a specificity of 88.28%.

Recently, Ross et al. (2021) built prediction models to classify whether a startup has an initial public offering (IPO), mergers and acquisitions, or fail through employing three machine learning algorithms. Using Crunchbase dataset of 942,605 startups from all over the world, they used XGBoost, random forest, and k-nearest neighbor as predictive models for classifying startups and assisting in making better investment decisions. One more study by Żbikowski et al. (2021) suggested a data-driven approach for startups success prediction by using a main set of independent variables of demographic, geographical, and basic startups information. Using Crunchbase dataset of 213,171 startups, they claimed the model reproducibility for real world scenarios relying on the available data at the time of making a decision, however, they selected XGBoost as the best model, with an accuracy of 85%, a precision of 57%, and a recall of 34% for the current scenario.

Overall, the above-mentioned prior studies have been conducted in Portuguese, the United States, Germany, and other developed countries and hence lacked the developing-country context. Therefore, we have been motivated to conduct this study in Saudi Arabia, first to address this gap context, and second to provide better insights for VCs to make good investment decisions when selecting which startup to fund through harnessing the power of machine learning.

Method

We obtained a sample of Saudi startups from Crunchbase database with 21 features (variables) and 979 records (startups) where Crunchbase is a community-based structured database where the users provide their companies' data. Moreover, the dataset gets further processed by Crunch base's in-house data experts to ensure accurate data for their customers. According to our domain knowledge and data availability of Saudi startups, we selected 17 features (or variables) that can serve our objective of this study. Table 1 below shows the selected features and provides a definition for each to give a better understanding of the nature of those features.

No	Feature	Description
1	Organization Name	Company commercial name.
2	Estimated Revenue Range	Estimated revenue, which represents the range of return to each investor and how successful the business can be.
3	Total Funding Amount	Total amount raised across all funding rounds. Can give a hint if compared with the overall revenue.
4	Headquarters Location	Where the organization is headquartered.
5	Founded Date	Date the organization was founded.

6	Number of Founders	Total number of founders.
7	Funding Stage	The stage in which the startups is operating (Early Stage, Mid Stage, Late Stage).
8	Number of Funding Rounds	The number of times in which startup received fund from investors.
9	Number of Investors	Total number of investment firms and individual investors who invested in the startup.
10	Industry Groups	Superset of industries (e.g., IT, Financial Services).
11	Number of Employees	Total number of employees, more employees will induce more cost.
12	Number of Apps	Total numbers of apps a given publisher have consolidated between iTunes and google play, as detected by Apptopia.
13	Visit Duration	Average time spent by users on a website, per visit in seconds. Includes both desktop and mobile web.
14	Bounce Rate	The percentage of visitors to the site who navigate away after viewing only one page. Includes both desktop and mobile web. Includes both desktop and mobile apps.
15	Monthly Visits	Total (non-unique) visits to site for the last month; includes desktop and mobile App.
16	Monthly Visits Growth	Percent change in total visits to the site from the previous month. Includes both desktop and mobile web.
17	Exit status	Whether the organization got acquired or went to IPO
	Table 1. Features Descriptions	

Prior to applying different algorithms, the obtained data need to be processed through cleaning and preprocessing steps to have better structured data and avoid biased prediction. The target variable (Exit status) had 796 observations labeled as “not exited” and 183 observations labeled as “exited”, representing only 23% of the overall startups. First, since the nature of CrunchBase’s data are crowd-sourced, many missing information had been encountered. Therefore, the encountered missing values were treated according to its nature (categorical / numerical). For example, the categorical missing values encountered were replaced with a value of 1 backed by the domain knowledge and for numerical missing values encountered were replaced by the median. Second, we detected the existence of outliers using the z-score method and so we treated this by replacing the outlier values with the median for each numerical variable. Third, we removed a number of variables such as description, industry, funding status, and download last 30 days due to irregularities and redundancy.

In addition, the founded startups before 2014 were not considered in this study, stating the fact if the company age is more than five years old, it’s not considered a startup. However, in our data sample, we extended the age period to eight years due to the lack of data on Saudi startups (Sanchez, 2021). ANOVA was applied for continuous variables and Chi-square for categorical variables resulted in recommendation to eliminate insignificant features as well as we considered other important features suggested by HALA Ventures (2020). Therefore, we ended up with 15 variables as the final selected features. The final features list includes all those mentioned in Table 1 above except for , “Number of Apps” and “Organization Name”.

After finishing the data preprocessing and cleaning stage, we applied data partitioning to prepare the data for the next stage of applying different models and algorithms. To do so, we converted all categorical

variables to dummies, then we standardized and scaled the data, finally we were able to partition the dataset into a 60% training set and 40% validation set. Our target class (exit class or successful) in the training dataset was represented only 3.32% as shown in figure 1. Therefore, oversampling was performed on the training data to overcome the class imbalance in the target variable and to have a better model prediction accuracy.

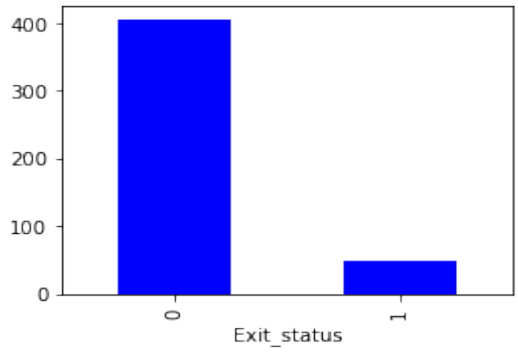

Figure 1. Target Variable Distribution

ML Algorithms Used

We encoded the target variable as class "0" represents not successful (survival) startup and "1" represents a successful startup. We employed the following machine learning algorithms represented in Table 2 in our study since they had been proved to be reliable to predict startup success in previous reviewed research.

No	Algorithms	Description
1	Naïve Bayes	A data-driven machine learning algorithm based on the Bayes theorem that finds the conditional probability to assign a class to a new record. Moreover, it requires the variable to be categorical, while the numerical variable to be binned. It makes no assumptions about the data and the variables are independent of each other (Gandhi, 2018).
2	Logistic regression	A type of linear regression in which the outcome variable is binary and usually utilized as a starting point for classification problems. Additionally, it's one of the most common algorithms in machine learning, known for its simplicity and low computational time to train and implement. It is less prone to overfitting due to its low variance, making it ideal for binary classification with a clear distinction of the classes (Bento, 2017; Cemre, 2019).

3	K-nearest neighbor (KNN)	A simple supervised machine learning technique that works by calculating the distances between a query and all of instances in the dataset by specifying K to the nearest query. Then, it votes for the most frequent label for the categorical target variable. It can be easily implemented, however, the larger the data gets, the algorithm gets noticeably slower (Harrison, 2019).
4	Support vector machine (SVM)	A supervised learning method that can be used for classification as well as regression. The concept is that the algorithm tries to discover the best hyperplane based on the labelled data in the training set to classify new records. The hyperplane is a simple line in two dimensions. Typically, this learning algorithm attempts to learn the most frequent characteristics of a class. The SVM operates in the opposite direction compared with other classification algorithms though it is the most alike examples found between classes (Zoltan, 2022).
5	Decision tree (DT)	It creates a training model that can predict the target variable value by learning simple rules of decision inferred from training data. It uses a tree-like graph to show predictions arising from a series of splits based on features. The DT is simple to understand, interpret, visualize, and effectively handle numerical and categorical data. However, a common flaw in decision trees is overfitting (Shmueli et al., 2022).
6	Random forest	It is based on the bagging algorithm. It creates as many trees as possible on the subset of the data and combines the output of all the trees. In this way, it reduces the overfitting problems in decision trees and reduces the variance and therefore improves the accuracy. However, a forest is less interpretable than a single decision tree (Shmueli et al., 2022).
7	Extreme Gradient Boost (XGBoost)	A decision-tree-based that uses a gradient boosting framework. XGBoost generally can't predict the future very well, but it is well suited for classification tasks, especially those related to real-world business problems. It can give high accuracy and prevent overfitting by making use of more trees (Morde, 2019).
8	Neural networks (NN)	NNs are algorithms that can be used to perform nonlinear statistical modeling. It's composed of many nodes and layers including the hidden layer, where most of the work is being conducted. The strength of NT is that it requires less formal statistical training, the ability to detect complex nonlinear relationships between independent and dependent variables. On the other hand, disadvantages include its "black box" nature, greater computational burden, and overfitting (Shmueli et al., 2022).

Table 2 – Metrics Definition

Analysis and Findings

Several machine learning algorithms were used to compare different models' performance and select the highest-performing model. We employed confusion matrix for evaluating the developed models since it is commonly used for measuring the performance for any classification model by showing the metrics of overall accuracy, sensitivity, and specificity. Table 3 below defines the three metrics used to compare the performance of the different predictive models.

Metrics	**Description**
Accuracy	Used to measure the fraction of predictions our model correctly classified by measuring the ratio of sum of true positive and true negatives out of all the predictions made.
Sensitivity	Used to measure the model accuracy to correctly predicted positive class by measuring the count of true positives in a correct manner out of all the actual positive values.
Specificity	Used to represent the correctly classified negative values out of all by measuring the number of correct negative predictions divided by the total number of negatives

Table 3 – Metrics Definition

In our study, we applied eight machine learning algorithms on a sample of 451 Saudi startups to predict the successful startups. The best four models out of the eight were XGBoost (the best performing model), followed by decision tree, random forest, and lastly logistic regression.

As per Table 4 below, *XGBoost* model's evaluation scores had been the highest compared to the other models, with an accuracy of 86.19%, a sensitivity of 53.33%, and a specificity of 89.16%, while no overfitting issue was found. *Decision tree* model with the maximum depth equals to 4, resulted in almost the same evaluation metrics as the XGBoost and no overfitting issue noted. Whereas *random forest* showed a lower evaluation accuracy of 83.98%, a sensitivity of 46.67%, and a specificity of 87.35%, and encountered an overfitting issue. Lastly, *logistic regression* showed the lowest evaluation scores among the top performing models with an accuracy of 72.38%, a sensitivity of 46.67%, and a specificity of 74.70% and had an overfitting issue.

	Random Forest	**Naive Bays**	**Logistic Regression**	**K-Nearest Neighbor**	**Support Vector Machine**	**Decision Tree**	**XG Boost**	**Neural Network**
Accuracy Training	0.9160	0.6744	0.8067	0.9559	0.8151	0.9118	0.9013	0.8950
Accuracy Validation	0.8398	0.3425	0.7238	0.7956	0.6796	0.8619	0.8619	0.6685
Sensitivity	0.4667	0.6000	0.4667	0.2667	0.6667	0.5333	0.5333	0.5333
Specificity	0.8735	0.3193	0.7470	0.8434	0.6807	0.8916	0.8916	0.6807

Table 4 – Model Performance

In addition to the above analysis, we derived the feature importance by using the random forest to highlight most contributing features to the predictive model as shown in figure 2. It appeared that the estimated revenue range was the most contributing feature to the model, followed by founded date, and then number of founders, while both total funding amount and number of employees had the same contribution to the model. On the other hand, features such as sector size and visit duration each had lower than 5% contribution percentage to predict the Saudi startups success.

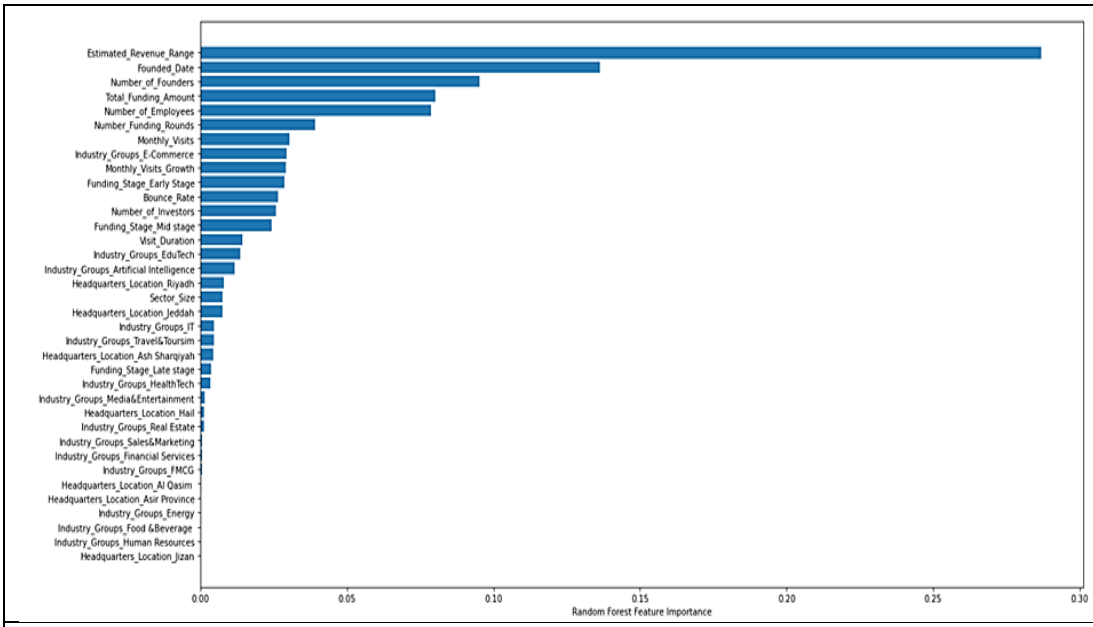

Figure 2. Features Importance

Lastly, choosing the best model was challenging due to the similar evaluation scores between XGBoost and decision tree. Although decision tree is simple and easy to interpret, it's usually arising with high variance, which means any small changes in the training data may cause large changes in the result (Glen, 2019). On the other hand, XGBoost depends on combined regression trees and uses each tree residual to learn and develop better prediction (Amazon, 2022). Thus, we select XGBoost as the optimal model to predict the startups success (Table 4).

Discussion and Insights

Based on the metrics evaluation scores provided in Table 4, XGBoost classification model is selected to classify startup success with an accuracy of 86%, and sensitivity of 53.33% on our final dataset of 451 Saudi startups. Prior studies that have employed machine learning applications to predict startups success are consistent with our findings; Cemre (2019) selects XGBoost algorithm in his study, as it accounts for accuracy of 94.45% and a sensitivity of 97.53%. Also, Żbikowski (2021) selected the XGBoost model as it shows an accuracy of 85% and a sensitivity of 34%. Hence, these studies can provide a reliable grounding to our study findings.

Machine learning models highlight how technology can impact industries that have historically been associated with a high-risk, and high-reward manner as it is the case for the venture capital industry. Venture capitalists frequently rely on the payoff from a single unicorn firm to meet the financial obligations that they have committed to their investors. Our selected model (i.e., XGBoost) has been able to make an accurate prediction of successful startups and hence demonstrate its high applicability to the venture capital industry. This means that venture capital firms would be able to adopt this specific prediction model and use it as a helping tool to make a decision on what investment should go with. On the other hand, we discovered that the model heavily depends on a number of variables which represent about 70% of the contribution to the model. These contributing variables are ranked from the highest to the lowest as follows:

1. Estimated Revenue Range
2. Founded Date
3. Number of Founders
4. Total Funding Amount
5. Number of Employees
6. Number of Funding Rounds

Accordingly, the VCs should highly pay attention to these contributing variables and consider their ranks when going through their analysis process. For example, VCs should give more weight to "estimated revenue range" during the startups analysis process and use it an anchor to expand on evaluating the other factors. Then, they should analyze the "founded date" and "number of founders" with their business backgrounds in line with considering the market size for the evaluated startups. However, it is important to indicate that this prediction model should be employed to get useful insights about the startups for better evaluation and cannot replace human intervention and expert judgement since there are other factors are not captured by the machine learning algorithms, for instance, the high/low potential of the startup considering market financials.

Recommendation

Manual market research and deal flow assessment which is done by VC's investment team may consume plenty of time and end up with unreliable decisions. By implementing our model, VCs would be able to gain more insights on the startups and the associated investment risk, which would be calculated by our model. Also, providing an easy-to-use API based on CrunchBase database, can result in reliable predictive or classification models that could be of use to the Saudi startups ecosystem including startups, venture Capital and other interested investors. Moreover, application of hyperparameters during the training process, will facilitate choosing the optimal parameters that correctly map the independent variables which should be included within the model (Nyuytiymbiy, 2020). Further, applying different algorithms (other than used in our paper) to the same data source may result in higher accuracy, sensitivity, and specificity.

According to the most recent findings from published research within the same topic, our research paper is thought to be one of the first studies that was done to create a Machine Learning model to predict the success of Saudi startups. The resulting outcome will assist VCs in making more precise selections, which eventually will participate in boosting the rate of return for investors. For that, the developed model is highly recommended to be used by underperformed venture capitals, by implying additional resources to invest in an even bigger number of enterprises. There is a possibility that Machine Learning will make the investment process more open and accessible to the public. After conducting qualitative research, such as getting to know the company's founders background and their startups in depth, investors will typically investigate the possible target companies before making any investment decisions. Together with a feature importance analysis that identifies the crucial features of the company which make it a successful investment or, on the other hand, raise red flags. Therefore, the contemporary venture capitalist will be able to look inside the black box, which enables them to make investment judgments that are both quicker and more reliable (Ross, 2021).

Limitations, Future Works, and Conclusion

Our study might suffer some limitations, which affect the overall accuracy of the model. We note that the study results are lower than the previous reviewed studies. This could be attributed to the fact that the

collected data sample obtained from Crunchbase has missing values in many variables. Also, the Crunchbase dataset is not being updated on a regular basis by the companies, which negatively could impact the quality of the data collected. In addition to this, the sample size of the cleaned dataset is deemed to be small, and there has been the issue of an imbalance distribution of target variables. Though we were able to resolve some of the mentioned issues however we can't claim that we get rid of the whole limitations.

We believe that these limitations can be considered to conduct better future research. First, future studies should avoid using data that involves many missing values to have better results in model prediction. Second, as an extension to our study, more variables need to be included e.g., the school background and job experience of each founder, which are believed would have a great impact to the overall prediction. Third, researchers should consult different venture capitals firms to understand their concerns in depth and tackle them through developing the machine learning models to achieve a higher payoff of the results.

In conclusion, predicting the startup success is a challenging task and should not only depend on the VCs expertise as it might lead to unsuccessful funding decisions imposing difficulties to overcome the resulted fund shortage. VCs can invest in machine learning applications to support their evaluation in making better informed decisions about identifying successful startups. Prior research has significantly lacked to provide how machine learning applications can help in predicting the success of new startups in Saudi Arabia. Therefore, this study attempts to fill this gap by developing different machine learning prediction models that can accurately classify the Saudi startups whether they will be successful. Overall, this study is one of the first studies that uses machine learning techniques to predict the success of Saudi startups. Hence, it can provide great insights to the Saudi Arabia venture capital industry by helping to lower the investment risk, elevate the evaluation process efficiency, along with saving the time and effort. It is highly recommended for venture capitalists to adopt machine learning models to increase their peering ability to extreme uncertainty and make faster and more reliable investment decisions.

References

Amazon. (2022). *How XGBoost Works*. Retrieved from aws.Amazon: https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html

Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment of Machine Learning Performance for decision support in Venture Capital Investments. *IEEE Access*, *7*, 124233–124243. https://doi.org/10.1109/access.2019.2938659

Bento, F. (2017). *Predicting Start-up Success with Machine Learning* (thesis).

Carter, T. (2021, Jan 3). The True Failure Rate of Small Businesses. *Entrepreneur*, p. 1.

Cemre, Krishna et al. Ü. (2019). *Searching for a unicorn: A Machine Learning approach towards startup success prediction* (thesis). Berlin.

Da Silva, D. (2016). *Portuguese Startups: a success prediction model* (dissertation).

Dellermann, D., Lipusch, N., Ebel, P., Popp, K. M., & Leimeister, J. M. (2017). Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3159123

Embroker. (2022). Unicorn Startups by Industry and Lessons from the $1B+ Club. *Embroker, 1*.

Fund, P. I. (2022). *Public Investment Fund*. Retrieved from Pif.gov.sa: https://www.pif.gov.sa/en/Pages/VRP2021-2025.aspx

Gandhi, R. (2018, May 17). *Naive Bayes classifier*. Medium. Retrieved June 23, 2022, from https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

Glen, s. (2919). Decision Tree vs Random Forest vs Gradient Boosting Machines: Explained Simply. *TechTarget, 1*.

GOV.SA. (2021). *Investment*. Retrieved from The kingdom invest: https://www.vision2030.gov.sa/thekingdom/invest/

HALA. (2020). *HALA Ventures*. Retrieved from HALA Ventures: https://www.halavc.com/

Harrison, O. (2019, July 14). *Machine Learning basics with the K-nearest neighbors algorithm*. Medium. Retrieved April 2, 2022, from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761#:~:text=Summary-,The%20k%2Dnearest%20neighbors%20(KNN)%20algorithm%20is%20a%20simple,that%20data%20in%20use%20grows.

Krishna, A., Agrawal, A., & Choudhary, A. (2016). Predicting the outcome of startups: Less failure, more success. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. https://doi.org/10.1109/icdmw.2016.0118

KumarI, A. (2022, June 12). *Accuracy, precision, Recall & F1-Score - Python examples*. Data Analytics. Retrieved June 25, 2022, from https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/

Lee, A. (2013, November 2). *Welcome to the Unicorn Club: Learning from Billion-Dollar startups*. TechCrunch. Retrieved March 15, 2022, from https://techcrunch.com/2013/11/02/welcome-to-the-unicorn-club/

Morde, V. (2019, April 8). *XGBoost algorithm: Long may she reign!* Medium. Retrieved June 25, 2022, from https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Nyuytiymbiy, k. (2020). Parameters and hyperparameters in Machine Learning and Deep Learning . *Towards Data Science*, 1.

Ries, E. (2011). *The Lean Startup: How Today's entrepreneurs use continuous innovation to create radically successful businesses*. Currency.

Ross, G., Das, S., Sciro, D., & Raza, H. (2021). CapitalVX: A Machine Learning model for startup selection and exit prediction. *The Journal of Finance and Data Science, 7, 94–114. https://doi.org/10.1016/j.jfds.2021.04.001*

Sanchez, M. (2021, March 5). When is a startup no longer a startup?: *EU-Startups. EU. Retrieved June 25, 2022,* from https://www.eu-startups.com/2021/03/when-is-a-startup-no-longer-a-startup/

Saudi venture capital Company. SVC. (2021, October 24). Retrieved February 4, 2022, from https://svc.com.sa/

Shane, S. (2012). The importance of angel investing in financing the growth of entrepreneurial ventures. *Quarterly Journal of Finance*, *02*(02), 1250009. https://doi.org/10.1142/s2010139212500097

Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). *Data mining for Business Analytics: Concepts, techniques, and applications in Python*. John Wiley & Sons.

Vision, S. 2. (2021). *Public Investment Fund Program*. Retrieved from vision 2030: https://www.vision2030.gov.sa/

Żbikowski, K., & Antosiuk, P. (2021). A Machine Learning, bias-free approach for predicting business success using Crunchbase Data. *Information Processing & Management*, *58*(4), 102555. https://doi.org/10.1016/j.ipm.2021.102555

Zoltan, C. (2022, February 22). *SVM and kernel SVM*. Medium. Retrieved April 2, 2022, from https://towardsdatascience.com/svm-and-kernel-svm-fed02bef1200