Prediction of Startup Success Using Machine Learning

Smit Patel

***Abstract*— In the dynamic world of startups, predicting success remains a challenge. As an investor, this makes it difficult to deploy capital to a startup with a high probability of success. Leveraging machine learning has the potential to offer insights into factors that contribute to startup success and to provide a predictive model that can help investors. We used Neural Network and Gradient Boosting classification algorithms to predict startup success. We were able to achieve accuracy of 96% for Neural Network and 100% for predicting startup success through Gradient Boosting classification. We concluded that investors should include machine learning based prediction models in their analysis along with other methods they already used today to make investment decisions.**

I. INTRODUCTION

Startups play a pivotal role in driving innovation, economic growth, and job creation in the modern world. They also contribute significantly to the diversification of industries and generation of new employment opportunities. In essence, startups are not just businesses; they are catalysts that stimulate progress and prosperity in the global economy. And when startups succeed, they create amazing returns for their founders and investors. However, only 1 out of 12 startups succeed [1]. That makes it very difficult for investors to determine which startup will succeed and which one they should fund. In this paper, we are using publicly available structured dataset [5] and supervised machine learning algorithms [6] to predict startup success.

II. BACKGROUND

There are other studies that have tried to answer similar questions. For example, Harvard Business Review [2] found 4 factors that predict startup success. This mostly focused on founders and how certain founder types are able to achieve success. Another MIT study [3] focused on the action the founder took. For example, any founders that registered in Delaware and protected their firms through patents and trademarks were more likely to receive funding and ultimately succeed as a company. Another interesting study [4] looked at the relationship between startups (either business relationships or employee moving between companies) and how it translates into startup success. Informed with these studies, we would like to bridge the gap between traditional predictors of startup success and modern data driven approaches.

III. METHOD

Dataset

We started with a publicly available dataset on startup. The original dataset had 54294 rows worth of data and after removing empty and duplicate rows we are left with 45850 rows of clean data. The dataset had 23 columns describing various attributes of a startup. Some of the interesting attributes are Status, Market, Total Funding, Founding Year, Founding Country, and Seed Funding. We explore these attributes in our data analysis stage to figure out which one should be used in our supervised machine learning and which one should not be.

Status

The companies were either operating, acquired or closed in the dataset. We considered operating and acquired as a successful outcome for this research.

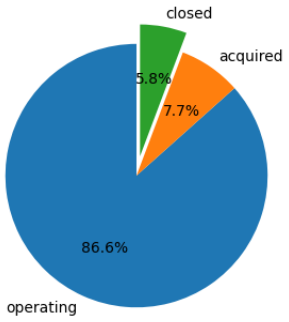

Figure 1: Company breakdown by status

Market/Industries

There were a lot of unique categories in the dataset for Industries. For our data exploration we only looked at the top 10 categories to see where most of the companies are from. Software has the largest number of startups and Advertising has the smallest number of startups in the top 10 categories.

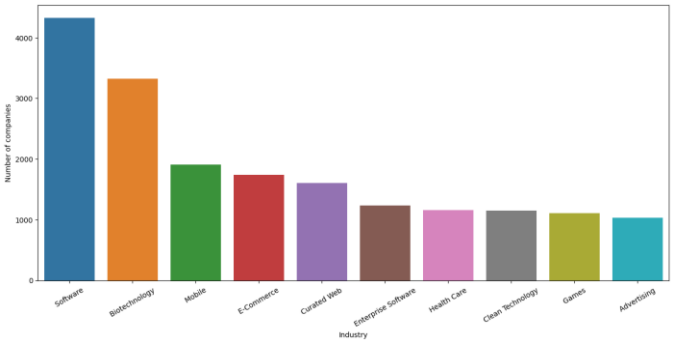

Figure 2: Group by Industries

Country of Origin

Another interesting category is country of origin and from the dataset it looks like it's skewed towards the U.S.A.

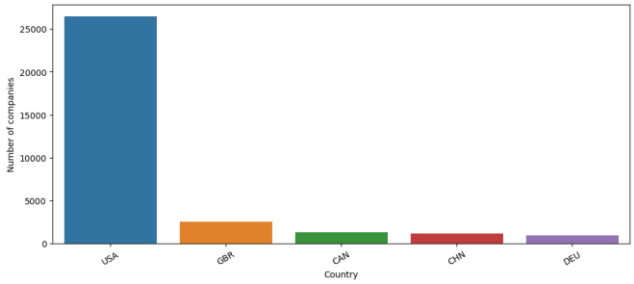

Figure 3: Group by Country of Origin

Founding Year

We also looked at the founding year to see if it signifies anything interesting and we found that the number of startups that are being founded have been consistently increasing. We also found that our dataset seems to stop in 2014 and number of startups may have peaked in 2012 (however this requires further research)

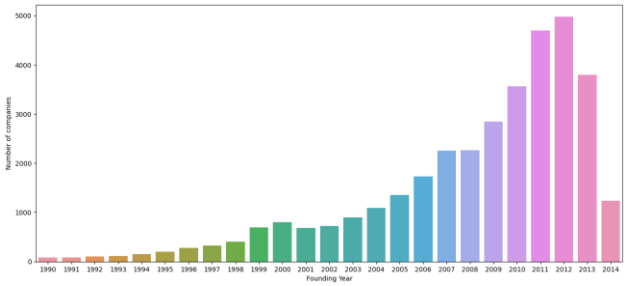

Figure 4: Group by Founding Year

Total Funding

We looked at the total funding to figure out if it has any impact on startup success or not. We had to perform more data cleaning to that column as it was in string format and contained commas and spaces. We also had to remove outliers as they were skewing the distribution and would impact on our machine learning model. After these steps, here is how the distribution of funding looks like with average funding of 3.8 million USD:

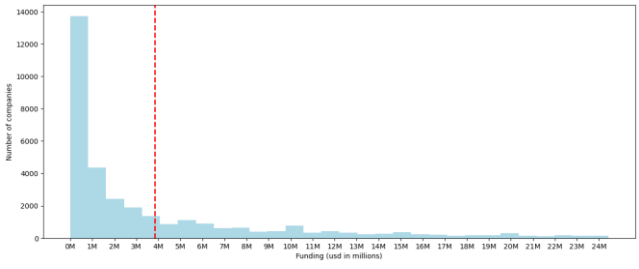

Figure 5: Total funding

Seed Funding

We wanted to explore, if seed funding was an indicator of success or not and lot of the companies in our dataset didn't receive any Seed funding

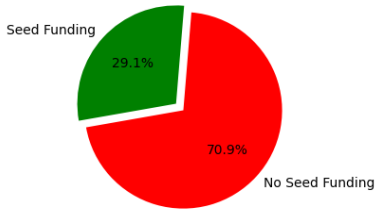

Figure 6: Funding at seed stage

Models

We were looking to predict the likelihood of a startup succeeding in their endeavors. This is fundamentally a classification problem [7]. We chose to explore two different algorithms for this paper, Artificial Neural Network [8] and Gradient Boosting [9]. For Artificial Neural Network, we chose sklearn's MLPClassifier and for Gradient Boosting, we chose sklearn's GradientBoostingClassifier algorithms.

MLPClassifier

A multi-layer Perceptron classifier which optimizes the log-loss function using the backpropagation algorithm. It has hidden layers, and each layer has activation functions which introduce non-linearity. The architecture can be deep, allowing for high model complexity. Depending on depth and data size, training can be slow for large networks. It requires preprocessing to handle missing data via imputation. It is considered a black-box model and doesn't offer great interpretability. Our model encompassed an architectural design defined by an input layer of 25 nodes, a subsequent hidden layer of 50 nodes, and a final output layer. The ReLU activation function was employed for the input and hidden layers, while the Sigmoid function was reserved for the output layer. The model's training was orchestrated with the Adam optimizer, using the binary cross-entropy loss function. Spanning 50 epochs, with a batch size of 10 and a validation split of 20%, the training process fine-tuned the model's weights

GradientBoostingClassifier

An ensemble method that builds an additive model in a forward stage-wise fashion. It constructs regression trees and fits them to the negative gradient of the loss function. It sequentially builds shallow trees and each tree tries to correct the errors of its predecessor. Depending on data size, training can be time-consuming since trees are built sequentially. It requires preprocessing to handle missing data via imputation. This specific model requires imputation but there are other boosting variants that can handle missing data natively. It offers a level of interpretability by visualizing individual trees, feature importance, and plots. Our model was tailored with hyperparameters including a learning rate of 0.1 and total estimators of 100.

IV. RESULTS AND DISCUSSION

MLPClassifier:

	Precision	recall	f1-score	support
0	0.00	0.00	0.00	108
1	0.97	1.00	0.98	3355
accuracy			0.97	3463
Macro avg	0.48	0.50	0.49	3463
Weighted avg	0.94	0.97	0.95	3463

Table 1: Classification report for MLPClassifier

	Predicted: No	Predicted: Yes
Actual: No	0	108
Actual: Yes	3	3352

Table 2: Confusion Matrix for MLPClassifier

GradientBoostingClassifier:

	Precision	recall	f1-score	support
0	1.00	1.00	1.00	152
1	1.00	1.00	1.00	4317
accuracy			1.00	4317
Macro avg	1.00	1.00	1.00	4469
Weighted avg	1.00	1.00	1.00	4469

Table 3: Classification report for GradientBoostingClassifer

	Predicted: No	Predicted: Yes
Actual: No	152	0
Actual: Yes	0	4317

Table 4: Confusion Matrix for GradientBoostingClassifer

The results of this research underline the efficacy of machine learning models in predicting startup success, specifically the Gradient Boosting Classifier, which showcased an impeccable accuracy. These findings coherently align with the experimental objective and address the initial research question, underscoring the transformative power of data-centric approaches in understanding startup dynamics. Tables and graphs were instrumental in providing a comprehensive view of these results, elucidating the nuances of the data. However, the perfect accuracy of the Gradient Boosting Classifier mandates a deeper analysis, raising questions about potential overfitting or biases. Such impeccable performance, while promising, necessitates skepticism and points towards potential sources of error. The dominance of the software market in the dataset hints at possible model biases. These results, while insightful, are not without limitations. The approach could benefit from a more diverse dataset or even the integration of ensemble methods. Additionally, potential modifications could encompass the exploration of more granular features or the application of advanced neural network architectures. Throughout the analysis, foundational principles presented in the background section served as guiding posts, ensuring that the research remained anchored in established theories while venturing into new territories.

V. CONCLUSION

The endeavor to employ machine learning models to predict startup success has elucidated some pivotal insights. The Gradient Boosting Classifier showcased remarkable accuracy. These results align with the objective question initially posited, reinforcing the potential of data-driven methodologies in deciphering the multifaceted landscape of startup success. Broadly, these outcomes accentuate the transformative power of machine learning in sectors beyond traditional technological realms, offering a data-centric lens to view business dynamics. Future research can potentially delve deeper into more granular features, like founder backgrounds or specific funding rounds, or even explore ensemble methods that combine various machine learning models. Such endeavors could not only refine the prediction accuracy but also provide more holistic insights into the myriad factors influencing startup trajectories.

REFERENCES

[1] Global Startup Ecosystem Report 2019 - https://startupgenome.com/reports/global-startup-ecosystem-report-2019

[2] 4 Factors that predict startup success - https://hbr.org/2016/05/4-factors-that-predict-startup-success-and-one-that-doesnt

[3] Passive versus active growth: evidence from founder choices and venture capital investment - https://www.nber.org/papers/w26073

[4] Predicting success in the worldwide startup network - https://www.nature.com/articles/s41598-019-57209-w#Sec3

[5] Startup Dataset - https://www.kaggle.com/datasets/arindam235/startup-investments-crunchbase

[6] Supervised Learning - https://en.wikipedia.org/wiki/Supervised_learning

[7] Classification Problem - https://en.wikipedia.org/wiki/Statistical_classification

[8] Artificial neural network - https://en.wikipedia.org/wiki/Artificial_neural_network

[9] Gradient Boosting - https://en.wikipedia.org/wiki/Gradient_boosting