COMPARISON OF CLASSIFICATION ALGORITHMS TO PREDICT THE SUCCESS OR FAILURE OF A STARTUP

Harshitha Mohanraj Radhika, Fnu Maria Poulose, Samiksha Talwekar

Department of Applied Data Science, San Jose State University

Data 245: Machine Learning Technologies

Dr. Vishnu S. Pendyala

September 30, 2022

ABSTRACT:

Startups are newly developed companies focusing on creating only one product and launching them in a market. The investors for these companies are called venture capitalists who look for a company's potential. Based on the potential and company's growth, these investors invest the money in the startups. "90% of startups fail," says Forbes research, and there are many kinds of research that state the reasons for the success and failure of the startups, which may or may not be true because there is no proper proof for the numbers given. A startup's success or failure depends on various features like the technology they use, the company's location, the position of the company amongst the other companies, etc. The company's success is when the creator of the company gets a high amount when the company becomes privately or publicly traded. The company's failure is marked when they plan to shut down because of no funds or other reasons. We, as a team, intend to build a machine learning model that can predict the success or failure of a company based on the data. The project's primary goal is to compare the different model's accuracy using the evaluation metrics and present a better model that can predict the startups' success rate. This model will be helpful for venture capitalists to invest in promising companies. This model can also be useful for companies to understand what they lag and develop in that particularly. People planning to establish a startup company or work for a startup can also use this model to understand the important features that result in the company's success.

MOTIVATION FOR THE PROJECT:

It is a well-known fact that most start-up companies fail. There are very few which can sustain this competitive market and earn profit. Finding businesses that are more likely to succeed is a worthwhile challenge for venture capital funds. It is essential to analyze the trends and evaluate current requirements to understand the start-up idea will survive or not. There are lots of people in academic as well as research field interested in launching their idea in the real world. However, not every individual has the acumen of a business person.The socio-economic factors such as the industry in which the company operates, the area in which the headquarters is located, or the level of competition in a specific sector and its sub-sectors are also difficult to quantify. There are various investors who are ready to fund ideas but needs a proper evaluation for the investment. With the increasing number of ideas, it is the need of the hour to build a system which will evaluate the ideas' success. A business success prediction model could help venture capitalist funds perform better.

The scope of the course enables us to use various algorithms like classifier and regressor for evaluating the success of the business venture. With the help of classifier, we can label the idea as success or not. Similarly, using multiple regression , we can examine the relationship between various variables and the outcome. We are looking into implementing these algorithms to solve the problem. We are building a model in the hope to reduce the effort and time of the investors. The system would motivate people who have ideas but lacks the acumen to understand the market. They can leverage this system and analyse their idea is worth pursuing by evaluating against the market trends.

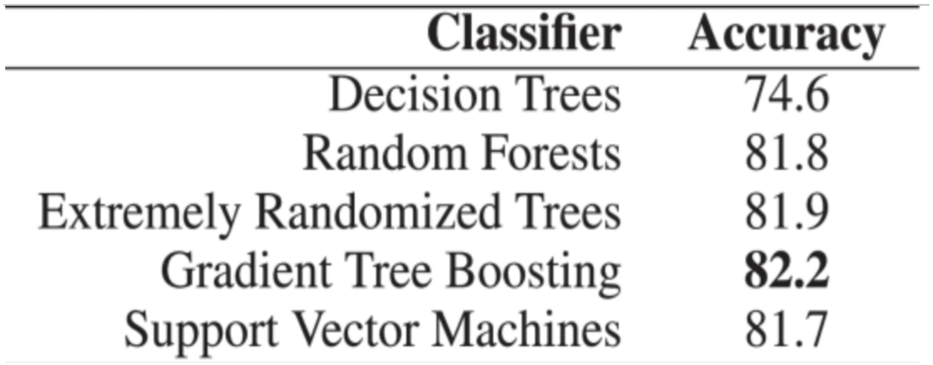
LITERATURE SURVEY:

Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments

This paper considers a machine learning approach to find venture capital investments with low risk. Machine learning algorithm help to predict the risk involved during its beginning itself. Here they have considered many factors like IPO achieved in which stage of progress. Based on funding and the founder's data, they are making the prediction. Here they have used various classifiers and calculated their corresponding frequency (Arroyo et al., 2019).

FIGURE 1

Comparison of accuracy of various classifiers



Classifier	Accuracy
Decision Trees	74.6
Random Forests	81.8
Extremely Randomized Trees	81.9
Gradient Tree Boosting	**82.2**
Support Vector Machines	81.7

Note: Accuracy of classifiers

Also, they have tried to differentiate between good or bad classes for each classifier. This will be a huge benefit for investors. For feature engineering, tree-based ensemble classifiers are used. They have trained classifiers again to get better features. Venture capital investors found this to be helpful for initial evaluation and to calculate performance (Arroyo et al., 2019).

Web-based Startup Success Prediction

We investigate the problem of predicting the success of startup companies in their early stages of development. We define the task as predicting whether a company that has already received initial (seed or angel) funding will attract a second round of investment within a specified time frame. Previous work on this task has mostly focused on mining structured data sources, such as startup ecosystem databases comprised of investors, incubators, and startups. Instead, we investigate the feasibility of using web-based open sources to predict startup success and model the task with a very rich set of signals from such sources. We enrich structured data about the startup ecosystem, in particular, with information from a business- and employment-oriented social networking service and internet.

We train a robust machine learning pipeline encompassing multiple base models using gradient boosting using these signals. We demonstrate that using company mentions on the Web yields a significant performance boost when compared to only using structured data about the startup ecosystem. We also provide a thorough analysis of the obtained model, which allows us to gain insights into both the types of useful signals that can be found on the Internet and the market mechanisms that underpin the funding process (Ozorin et al., 2018).

CapitalVX: A machine learning model for startup selection and exit prediction

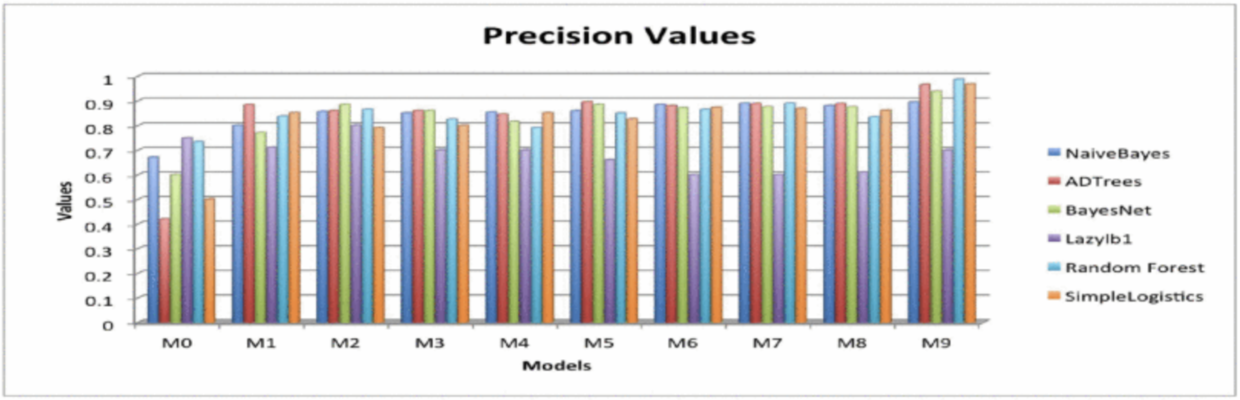
Here the aim is to predict whether a startup can successfully exit from IPO. Using machine learning algorithms prediction is done and accuracy is calculated based on various models. The startup can plan to make the investments according to the prediction results. Feature engineering is done and later on, this is also included to balance the model. In this paper, they have used models like Multilayer Perceptron (MLP), Random Forest, XGBoost, Ensemble performance, and K-Nearest Neighbors. After calculating the accuracy of various models, it is between 80 - 90 %. This plays an important role as many startups fail to successfully exit from IPO. Machine learning models help to predict and can make the investment decision-making process easy (Ross et al., 2021).

Predicting the Outcome of Startups: Less Failure, More Success

Startups can fail due to multiple reasons. In this paper, the prediction is done whether a startup will succeed or fail based on various factors like funding. The goal is to create a predictive model. Here they have used many classification techniques like Lazy lb1, Random Forest, Naive Bayes, Adtree, Bayesian Network, and Simplelogistic. Data preprocessing is done to clean the data. Classifiers are applied to the cleaned data for the prediction. 88.9 % is one of the top precision results from prediction algorithms for which a startup will succeed or not. From the results, Random forest and Simple Logistics has the highest accuracy (Krishna et al., 2016).

FIGURE 2

Comparison of precision of various classification models



Note: Precision of classification models.

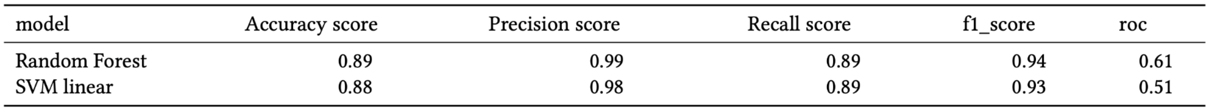
Prediction of the Success of Startup Companies Based on Support Vector Machine and Random Forest

There can be a lot of economic development from startups. For this, we need to predict the success of startups. In this paper, they are predicting the success of startups. Models used here are Random Forest and Support Vector Machine. Here they have considered various features.

The random forest classifier helps to get a clear picture of the importance of various features. From the results, both models have almost the same prediction accuracy (Jinze Li, 2020).

FIGURE 3

Comparison of two models



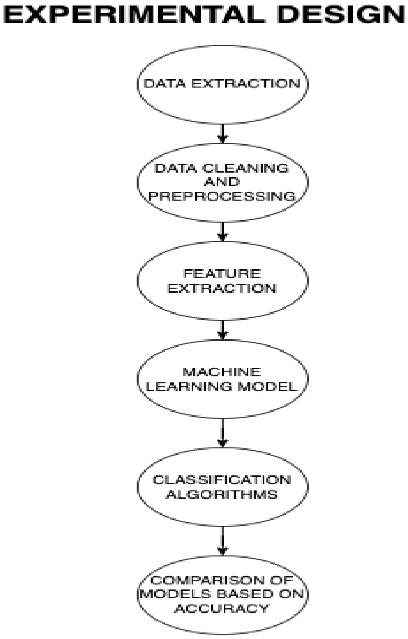
model	Accuracy score	Precision score	Recall score	f1_score	roc
Random Forest	0.89	0.99	0.89	0.94	0.61
SVM linear	0.88	0.98	0.89	0.93	0.51

Note: Accuracy of Random forest and SVM model.

METHODOLOGY:

FIGURE 4

PROCESS WORKFLOW DIAGRAM

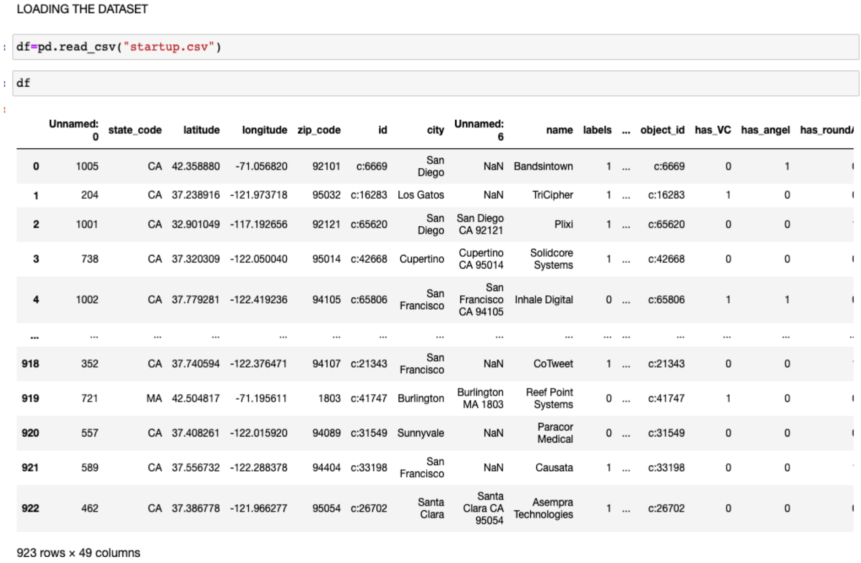

EXPERIMENTAL DESIGN
DATA EXTRACTION
DATA CLEANING AND PREPROCESSING
FEATURE EXTRACTION
MACHINE LEARNING MODEL
CLASSIFICATION ALGORITHMS
COMPARISON OF MODELS BASED ON ACCURACY

Note. Flow chart of the project

As part of our project, predicting the success or failure of the startups, we plan to use a CSV dataset from Kaggle that contains various columns, including the location details, type of industry, name of the companies, about the venture capitalists, is it in the top 500 companies or not, etc. The dataset has 923 rows and 49 columns.

FIGURE 5

Loading the dataset and viewing the number of rows and columns.



LOADING THE DATASET

```
df=pd.read_csv("startup.csv")
```

```
df
```

	Unnamed: 0	state_code	latitude	longitude	zip_code	id	city	Unnamed: 6	name	labels	...	object_id	has_VC	has_angel	has_roundA
0	1005	CA	42.358880	-71.056820	92101	c:6669	San Diego	NaN	Bandsintown	1	...	c:6669	0	1	
1	204	CA	37.238916	-121.973718	95032	c:16283	Los Gatos	NaN	TriCipher	1	...	c:16283	1	0	
2	1001	CA	32.901049	-117.192656	92121	c:65620	San Diego	San Diego CA 92121	Plixi	1	...	c:65620	0	0	
3	738	CA	37.320309	-122.050040	95014	c:42668	Cupertino	Cupertino CA 95014	Solidcore Systems	1	...	c:42668	0	0	
4	1002	CA	37.779281	-122.419236	94105	c:65806	San Francisco	San Francisco CA 94105	Inhale Digital	0	...	c:65806	1	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
918	352	CA	37.740594	-122.376471	94107	c:21343	San Francisco	NaN	CoTweet	1	...	c:21343	0	0	
919	721	MA	42.504817	-71.195611	1803	c:41747	Burlington	Burlington MA 1803	Reef Point Systems	0	...	c:41747	1	0	
920	557	CA	37.408261	-122.015920	94089	c:31549	Sunnyvale	NaN	Paracor Medical	0	...	c:31549	0	0	
921	589	CA	37.556732	-122.288378	94404	c:33198	San Francisco	NaN	Causata	1	...	c:33198	0	0	
922	462	CA	37.386778	-121.966277	95054	c:26702	Santa Clara	Santa Clara CA 95054	Asempra Technologies	1	...	c:26702	0	0	

923 rows × 49 columns

Since there are many columns, we may need to perform feature engineering to extract the features or factors related to predicting the success of the startups. The dataset will be classified into training and testing datasets.

Here the target variable is the company's status, which tells whether the company has been acquired or closed. The training dataset is where we fit our model, and the test data is used to evaluate the model. We aim to use three classification algorithms to find the accuracy of the three models and compare them to find the best model. We plan to use Linear Regression, Random Forest, and Support Vector Classifier algorithms to build the model.

DELIVERABLES

Deliverable	Description	Due date
Project proposal	A report containing the project topic, abstract, the motivation behind the project, the literature survey, and the models used.	**30 SEP 2022**
Intermediate status Report	A report containing works that would have been completed and things that have to be done later	**OCT 28 2022**
Project Presentations	A presentation explaining the model built by the team and the accuracy to predict the success or failure of the startups	**DEC 2 2022**

Project Report and Code	The report about the model developed, the codes of the project and the presentation slides	**DEC 2 2022**

MILESTONE

On a long term we like to write a technical paper describing the models that we used and the metrics to evaluate the model and further develop the model and research on how to increase the accuracy by using various other algorithms so we can submit this a journal or a thesis paper.

TEAM MEMBERS AND THEIR ROLES:

We are a team of 3, Harshitha Mohanraj Radhika, Fnu Maria Poulose, and Samiksha Talwekar where everyone is new to the domain. We planned to work together as a team for all the data extraction, cleaning preprocess, and feature engineering and also develop a model individually and show the accuracy to compare the three models and find which model outperformed well.

RELEVANCE TO THE COURSE:

This problem will be solved using a Supervised Machine Learning approach, which will involve training a model based on the history of startups that have been acquired or closed. To implement this system, we would be using machine learning algorithms like classifiers and linear regression. These algorithms will give the output as “success” or “failure” on the basis of a certain idea. The dataset obtained from the internet has various features. We will use feature engineering included in the course, and segregate data on the basis of the most relevant features to the outcome.

TECHNICAL DIFFICULTY

One of the technical limitations is an interactive front-end application that can be useful for investors as well as students who are planning to start their startups. Also, we are unable to provide an interactive map showing the demographics. Like to show which states in the United States are best for a startup. Based on prediction using machine learning algorithms.

NOVELTY.

We found this idea to be unique and valuable for many people, especially students who are planning to start their businesses. Knowing the risk involved at the early stage of business development is always better. Machine learning paves the way for this problem. As there are many classifiers that help to know the accuracy of the models. If we can predict the success of startups, thereby reducing the risk involved for investors. Our dataset consists of startup companies in the US until 2012. Also, there are 49 unique columns in which the majority can be considered features. We found this might be helpful for investors in the United States who are planning to start their businesses. A startup has a high chance of failing due to various reasons.

IMPACT

Predicting a startup's success allows investors to identify companies with th**e** potential for rapid growth, allowing them to stay one step ahead of the competition. Startups are critical to economic growth. They bring new ideas, stimulate innovation, and create jobs, thereby moving the economy forward. Startups have grown at an exponential rate in recent years. The aim to build this system is to reduce the time and effort required by venture capitalists.

It also targets audiences who want to implement their business idea and needs guidance for estimating the success of their idea. Thus encouraging them to pursue the idea. This will lead to an increase in good ideas coming into the market. Thus giving both founders and investors the tools, methods, and advice that will give them a competitive advantage.

For us, it will give an insight into the course, which contains various algorithms and concepts. It will help us implement any algorithm for real case scenarios. This system will build our technical skills and acumen as well.

REFERENCE

Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019b). 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

Das, S., Sciro, D., & Raza, H. (2021, November). 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

Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P., & de Rijke, M. (2018, October 17). Web-based Startup Success Prediction. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management.* https://doi.org/10.1145/3269206.3272011

Agrawal, A., & Choudhary, A. (2016, December). 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

Prediction of the Success of Startup Companies Based on Support Vector Machine and Random Forset. (2020, November 6). *2020 2nd International Workshop on Artificial Intelligence and Education.* https://doi.org/10.1145/3447490.3447492