Assessment 3: Data Analysis Report

Upcoming Startups Success Rate Prediction

Figure 1: Startup Journey (Thomas Kings Private Limited, n.d.)

Group 2

Rohit Shrestha (c3502400)

Pradeep Regmi (c3501889)

Bikram Ghimire (c3502820)

Executive summary

This Project focuses on predicting the success of startups by analysing various key factors from the startups funded by Y Combinator. Y Combinator is a startup accelerator company .The dataset includes information of those companies funded by Y combinator . It includes the information about founders , the industries they belong to , the school they attended , and other relevant features. The main objective is to predict whether a startup will be acquired or fail based on historical data.

Four Machine learning models, Logistic Regression, Support Vector Classifier (SVC), Random Forest, and Gaussian Naive Bayes were evaluated using metrics like accuracy , precision , recall and f1 score. Random Forest and Gaussian Naive Bayes showed the most balanced performance among all , effectively predicting the success of startups.

According to the study, the chances of success are influenced by founders' education, industry and past experience.The experiment also showed that using machine learning to forecast startup results could be useful in making investment decisions for investors. To increase the accuracy , future improvements can be done by using deeper models and larger dataset with more information.

Introduction

Startups are a crucial part of the global economy. They open new markets , and give room for new ideas and innovation. Startup ventures face dynamic environments which often have unexpected outcomes.It makes it very difficult to predict whether the startup will succeed or fail. This project focuses on analysing the key factors influencing the success of startups and predicting if it will become successful or not. Y Combinator, one of the most well-known startup accelerators globally, has a dataset of all the companies funded by it, which has key factors such as education of the founders , industry they are in , region and company history. For this, machine learning algorithms for classification are used. The effectiveness is tested on the dataset for every model and performance metrics are compared to determine the best algorithm for startup success prediction.

Dataset

The dataset for the project is extracted from the official site of “YCombinator All Funded Companies” dataset (Korovkina, n.d.). The dataset contains different CSV files which govern information of founders, regions, companies, industries, schools and badges.
The dataset on the 8 CSV files are:

1. companies.csv: It include data related to information of the company
2. industries.csv: It includes information about the industry in which the company is operating
3. schools.csv: It includes the data in which schools the founders have attended.
4. badges.csv: It contain data that represent the badges awarded to the companies
5. regions.csv: It represent the geographical data of the companies
6. tags.csv: It include the tags associated with the companies
7. prior_companies.csv: It contains the data of the details on the prior companies the founder has worked at.
8. founders.csv: It includes the founders, which include their name and other attributes.

The data have its own columns to represent the data, here some of the data are missing which are handled during the preprocessing stage by dropping or filling it with null values as needed.

Dataset Analysis

The dataset includes various features with both categorical and numerical variables. To understand the relationships and find patterns , several analyses were made. A correlation matrix was computed for numerical features. This helped to reveal how factors like team size, region and company history are related to the success of a startup.

1) Top Schools : Startups founded by individuals from top universities like Harvard, Stanford tend to have very high success rates .
2) Industry Influence : Certain industries such as software and biotechnology were found to be more likely to succeed
3) Regional Factors: Regional factors are also found to play a significant role in startups success. Companies from regions like Silicon Valley show higher success rates.
4) Prior Company Experience: Founders with experience in prior companies which are successful have a higher probability of success in their current centuries.

Visualisations, including a heatmap of correlations and various box plots, highlight these relationships and offer a deeper understanding of the dataset's structure and key variables.

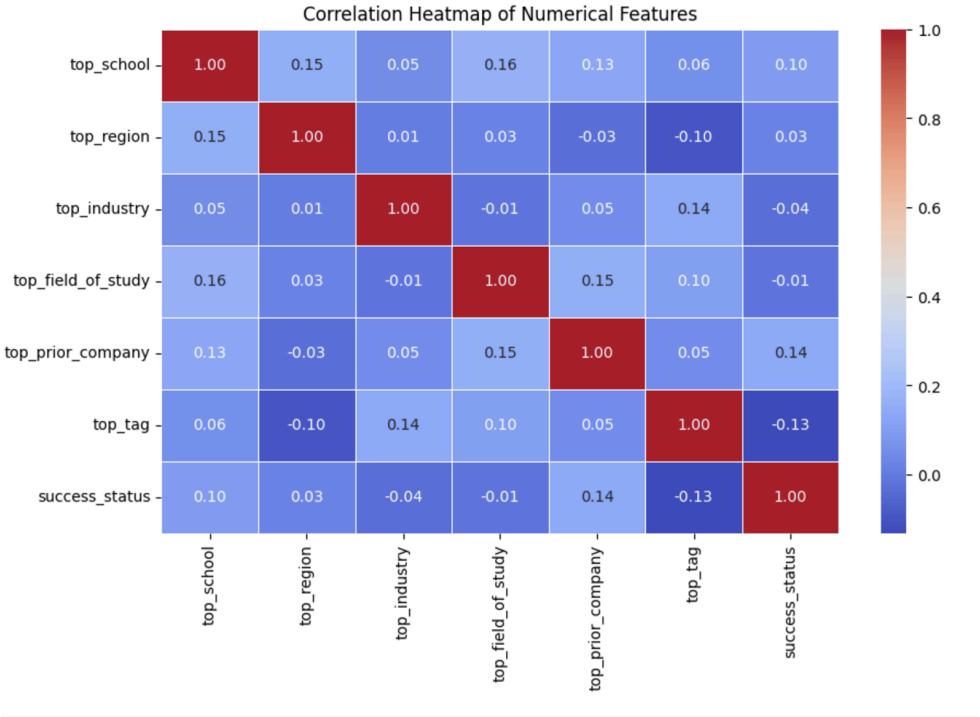

Fig. Correlation Heatmap for all numerical columns

The Heatmap shows the correlation between key features used in predicting the startup success. It shows that most correlations are weak, as indicated by the blue colour. It tells that the individual features are relatively independent of each other and do not strongly affect the success status . We can see that the highest positive correlation with success status is from the top prior company with a score of 0.14. There are mild correlations between features like top school and field of study , from which we can tell that the graduates from top schools often have related fields of study. Overall , the heatmap suggests that while no single feature is a strong predictor of success , their combined analysis might still give us valuable insights.

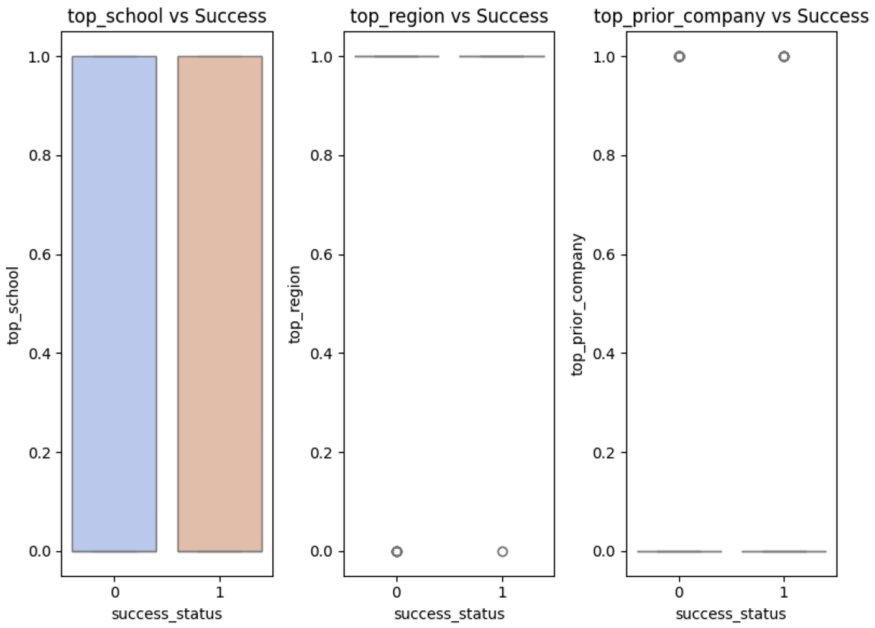

Fig. Box plot for top school, top region and top prior companies on success status

The box plot shows the relationship between three features namely top school ,top region and top prior company and the success status of startups. For top schools , the proportion of founders from top schools seems similar for both successful and unsuccessful startups , suggesting that there is limited impact on success. The top region plot shows no difference , meaning that the region feature does not differ significantly between success categories. The top prior company plot shows sparse data points , meaning low representation of founders from top companies and potentially weak influence on the success status. Overall these plots also suggest that none of these features alone can strongly differentiate between successful and unsuccessful startups.

Data Modelling and Results

After carefully cleaning up the data, separation of the possible columns and removing all empty values, we split up the data into a 70/30 ratio for training and test purposes using train_test_split from sklearn. The models SVC and logistic regression require the standardisation of the data hence we used StandardScalar for preprocessing.
We then for all the separated clean data trained four different models with the training dataset and validated using the test dataset. The models used are:

1. Logistic Regression
2. SVC
3. Random Forest
4. GaussianNB

	precision	recall	f1-score	support
0	0.80	0.99	0.88	87
1	0.50	0.04	0.08	23
accuracy			0.79	110
macro avg	0.65	0.52	0.48	110
weighted avg	0.73	0.79	0.71	110

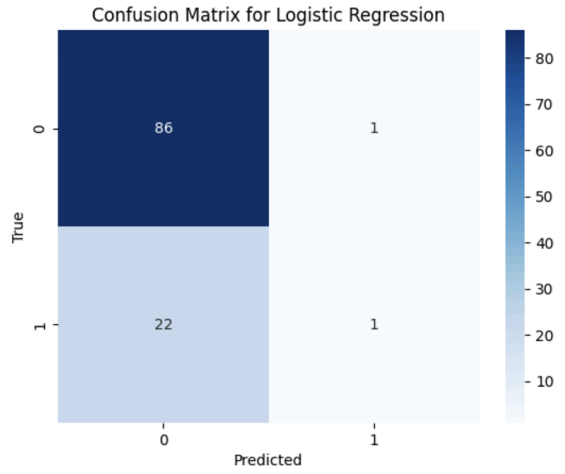

Fig: Logistic Regression

	precision	recall	f1-score	support
0	0.83	0.97	0.89	87
1	0.67	0.26	0.38	23
accuracy			0.82	110
macro avg	0.75	0.61	0.63	110
weighted avg	0.80	0.82	0.79	110

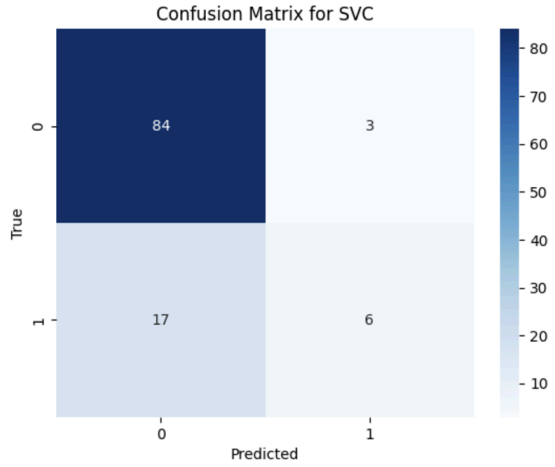

Fig: SVC

The results were categorised for each on the basis of precision, accuracy, recall and f1 score. The computed results from the model were then used to create the confusion matrix for each of the models.

	precision	recall	f1-score	support
0	0.84	0.95	0.89	87
1	0.64	0.30	0.41	23
accuracy			0.82	110
macro avg	0.74	0.63	0.65	110
weighted avg	0.80	0.82	0.79	110

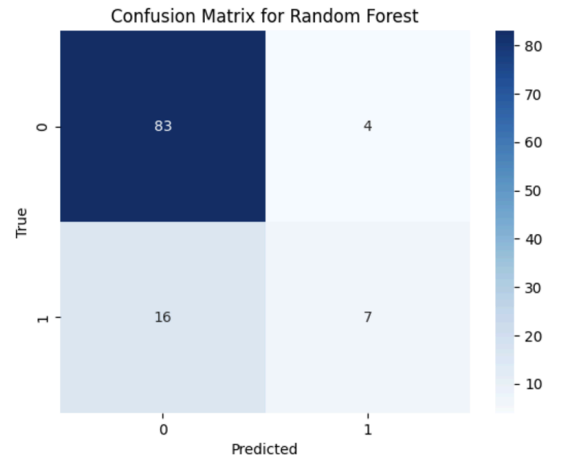

Fig. Random Forest

	precision	recall	f1-score	support
0	0.88	0.89	0.88	87
1	0.55	0.52	0.53	23
accuracy			0.81	110
macro avg	0.71	0.70	0.71	110
weighted avg	0.81	0.81	0.81	110

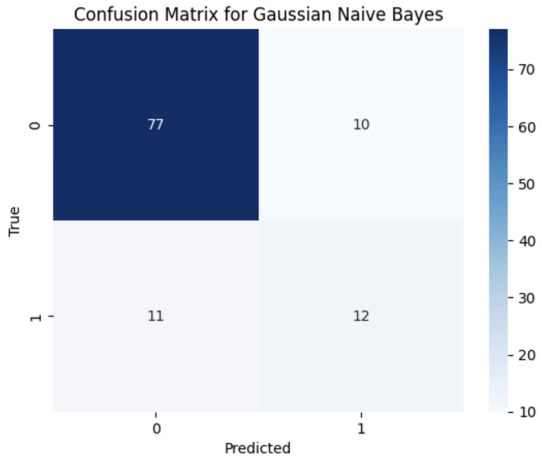

Fig. Gaussian NB

On comparing the results from the different models, there were 110 total data in the dataset, with 87 not successful startups and 23 successful startups. The Logistic regression model generated a precision of 80% for detecting not successful and 50% for detecting successful startups correctly. The SVC model performed slightly better than Logistic regression in the context of precision with 83% correct not successful predictions and 67% correct successful predictions. For Random Forest and GaussianNB, their performance was more balanced.
Overall the testing dataset was quite imbalanced with a heavy number of unsuccessful cases. However, the SVC and Random Forest performed better than the logistic regression. The GaussianNB showed the highest balanced performance with lower overall accuracy. Based on the weighted average F1 scores, GaussianNB and Random Forest show best results with balanced prediction.

Discussion & Conclusion

The models provide valuable insights into the success factors of startups. The Random Forest model outperformed others in terms of accuracy and balanced metrics, making it the most suitable model for this problem. The analysis shows that factors like founder education, industry, and region are strong predictors of startup success.
Future work could involve incorporating more granular features, using deeper models like neural networks, and analysing larger datasets to further refine predictions.
In conclusion, the project successfully demonstrated that machine learning can be a powerful tool in predicting startup success, which can be leveraged by investors and accelerators to guide their decisions.

References

Cohen, S., & Hochberg, Y. (2014). Accelerating startups: The seed accelerator phenomenon. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2418000

Marketeers. (2023, October 3). What is Y Combinator? - Marketeers. Medium. https://medium.com/@markeeters/what-is-y-combinator-1de598e83cd4

Shane, S., & Stuart, T. (2002). Organisational endowments and the performance of university startups. Management Science, 48(1), 154–170. https://doi.org/10.1287/mnsc.48.1.154.14280

Sussan, F., & Acs, Z. J. (2017). The role of big data in startup ecosystems. Small Business Economics, 49(2), 279–290. https://doi.org/10.1007/s11187-017-9874-9

Y Combinator Companies Directory. (n.d.). Y Combinator. Retrieved October 26, 2024, from https://www.ycombinator.com/companies

Thomas Kings Private Limited. (n.d.). Startup journey: From idea to IPO in India and beyond. Retrieved November 16,2024, from https://thomaskingsprivatelimited.com/startup-journey-from-idea-to-ipo-in-india-and-beyond/

Korovkina, S. (n.d.). *YCombinator All Funded Companies Dataset (Version 3)* [Data set]. Kaggle. https://www.kaggle.com/api/v1/datasets/download/sashakorovkina/ycombinator-all-funded-companies-dataset?dataset_version_number=3