Startup Success Rate Prediction

1st Harshitha Mohanraj Radhika
Department of Applied Data Science
San Jose State University
San Jose, United States
harshitha.mohanrajradhika@sjsu.edu

2nd Fnu Maria Poulose
Department of Applied Data Science
San Jose State University
San Jose, United States
mariapoulose@sjsu.edu

3rd Samiksha Talwekar
Department of Applied Data Science
San Jose State University
San Jose, United States
samikshatalwekar@sjsu.edu

***Abstract*—Startups are newly developed companies focusing on creating only one product and launching them in a market. Investors invest money in startups based on the company's potential and growth. Many kinds of research state the reasons for the success and failure of startups, which may or may not be accurate because there needs to be 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. In this paper,we performed a study to compare the performance of the machine-learning models like Logistic Regression, SVM, KNN, Random Forest, that can predict the success or failure of a company. The evaluation metrics like accuracy, Precision, recall, F1 score, Roc-Auc score, Cohen Kappa score were used to estimate the performance of the models .This model will be helpful for venture capitalists to invest in promising companies and also for people who are planning to establish a startup company or work for a startup can also use this model to understand the crucial features that result in the company's success.**

***Index Terms*—Startup, Success prediction, Random Forest, Logistic Regression**

I. Introduction

The known fact is that most startup companies fail, and very few are sustaining this competitive market and earning 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 whether the startup idea will survive. Many people in academic and research fields are interested in launching their idea in the real world. 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. With the increasing number of ideas, it is the need of the hour to build a system that will evaluate the ideas' success. A business success prediction model could help venture capitalist funds perform better. The primary goal is to use various classification algorithms to evaluate the business venture's success. With the help of a classifier, we can label the idea as successful or not. We are implementing Logistic Regression, Random Forest Regressor, and Support Vector Machine to solve the problem and reduce the effort and time of the investors. This model would motivate people who have ideas but need more acumen to understand the market. They can leverage this system and analyze whether the idea is worth pursuing by evaluating it against the market trends.

II. Significance to real world

Predicting a startup's success allows investors to identify companies with the potential for rapid growth, allowing them to stay one step ahead. Startups play a significant role in 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. Success prediction also helps audiences who want to implement their business idea and need guidance for estimating their idea's success, thus encouraging them to pursue the idea. There will be an increase in good ideas coming into the market, thus giving founders and investors the tools, methods, and advice that will give them a competitive advantage.

III. Literature survey

The author considers a machine learning approach to find low-risk venture capital investments by predicting the risk involved during the beginning period of the startup. The authors have considered many factors like IPO achieved in which stage of progress and making the prediction based on funding and the founder's data. In their previous research, they found that authors have used just IPO, so the authors of this paper wanted to predict the company's status after the IPO. They used a dataset of about one hundred thousand newly started companies for which they wanted to forecast the success in three years. The data was collected from Crunchbase, and there was bias in the survival rate of the companies. They had five target variables: acquired, the funding round, IPO, Closed, and no event. They calculated their corresponding frequency and ratio to understand the class distribution and then performed the prediction. They used models like SVM, DT, GTB, and RF, in which GTB had high accuracy. They also used some metrics like precision, Recall, and F1 score to evaluate the performance of the model[1]

The authors have investigated predicting the success of startup companies in their early stages of development. They defined 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 focused chiefly on mining structured data sources, such as startup ecosystem databases comprised of investors, incubators, and startups. Instead, they investigate the feasibility of using web-based open sources to predict startup success and model the task with a rich set of signals from such sources. They enrich structured data about the startup ecosystem, particularly with information from a business- and employment-oriented social networking service and the Internet. The authors train a robust machine learning pipeline encompassing multiple base models using gradient boosting using these signals. They demonstrated that using company mentions on the Web yields a significant performance boost compared to using structured data about the startup ecosystem. They also provide a thorough analysis of the obtained model, which allows us to gain insights into the types of valuable signals found on the Internet and the market mechanisms that underpin the funding process [2].

The author established a model that can predict the success of startups by analyzing the features for the success of those companies. In this paper, they predict the success of startups. The author has used the Kaggle dataset of 22000 startups, and the models used here are Random Forest and Support Vector Machine. Here they have considered various features like state code, venture, crowdfunding, and more. The random forest classifier helps to get a clear picture of the importance of various features. From the results, they have concluded that both SVM and Random Forest Regressor have almost the same prediction accuracy.SVM has performed slightly well because it can handle multi-dimensional data. The Random Forest Regressor had the best precision, F1 score, and roc[3].

The author aims to predict whether a startup can successfully exit from IPO and if the venture can gain additional funding. The authors have considered using machine learning algorithms to predict, and accuracy is calculated for various models. The target variable here is successful and unsuccessful, and the main goal is to classify the companies that fall under these classes. The data was taken from Crunchbase, and the dataset was converted to an RDBMS to represent the data. Feature engineering and selection were made using the SQL query; the raw and created features were compiled for modeling. The author used various algorithms for models like Multilayer Perceptron (MLP), Random Forest, XGBoost, Ensemble performance, and K-Nearest Neighbors. The accuracy of various models was between 80 - 90. The author concludes that by using this model, and the startup can plan to make investments according to the predicted results. This plays a vital role as many startups need help to exit successfully from IPO[4].

The author states that startups can fail due to multiple reasons. The prediction was made if a startup would succeed or fail based on various factors like funding and various stages of the startup. The data was taken from Crunchbase and Tech Crunch, and they also used some data mining techniques to preprocess the data with other optimization and validation techniques. Data preprocessing was done to clean the data. They aim to create a predictive model using the Random forest, ADTrees, Bayesian Networks, Lazy lb1, Naive Bayes, and Simplelogistic. They used metrics like accuracy, precision, Recall, and AUC to evaluate the model's performance. Classifiers were applied to the cleaned data for the prediction. 88.9 percentage is one of the top precision results from prediction algorithms for which a startup will succeed or not. The results show that the Random forest and Simple Logistics have the highest accuracy[5].

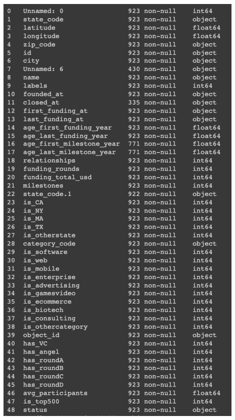
IV. Project management methods

An agile method was followed for the project, where we managed the tasks required to complete the project. We had regular meetings in teams and via zoom to brainstorm the tasks to be done in the upcoming week. We used Trello, a project management tool using CRISP-DM, to split the project into 6 phases. We created sub-tasks under each phase, assigned each team member tasks, and worked accordingly. We maintained a board where we moved tasks from to To-Do to done once completed. If there is any need to review tasks, they were moved to the In review card of the agile board.The code was developed using the pair programming were we used Google collab to work at the same time.

V. Data Preparation

A. Data Collection

The dataset for the project was taken from Kaggle, which we are using to build the model to predict the startup's success. The dataset has 923 rows and 49 columns, which impacts the startup's success or failure prediction. The dataset contains data about each startup, like the startup's name, the startup domain, statecode, details about funding, whether it has VC, has an angel, and is in the top 500. The target variable here is status, which has two classes acquired and closed.Figure 1 shows the features or the columns that are present in the dataset



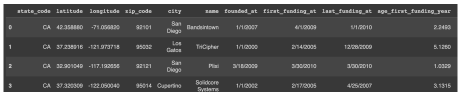
```
0   Unnamed: 0                 923 non-null    int64
1   state_code                 923 non-null    object
2   latitude                   923 non-null    float64
3   longitude                  923 non-null    float64
4   zip_code                   923 non-null    object
5   id                         923 non-null    object
6   city                       923 non-null    object
7   Unnamed: 6                 430 non-null    object
8   name                       923 non-null    object
9   labels                     923 non-null    int64
10  founded_at                 923 non-null    object
11  closed_at                  335 non-null    object
12  first_funding_at           923 non-null    object
13  last_funding_at            923 non-null    object
14  age_first_funding_year     923 non-null    float64
15  age_last_funding_year      923 non-null    float64
16  age_first_milestone_year   771 non-null    float64
17  age_last_milestone_year    771 non-null    float64
18  relationships              923 non-null    int64
19  funding_rounds             923 non-null    int64
20  funding_total_usd          923 non-null    int64
21  milestones                 923 non-null    int64
22  state_code.1               922 non-null    object
23  is_CA                      923 non-null    int64
24  is_NY                      923 non-null    int64
25  is_MA                      923 non-null    int64
26  is_TX                      923 non-null    int64
27  is_otherstate              923 non-null    int64
28  category_code              923 non-null    object
29  is_software                923 non-null    int64
30  is_web                     923 non-null    int64
31  is_mobile                  923 non-null    int64
32  is_enterprise              923 non-null    int64
33  is_advertising             923 non-null    int64
34  is_gamesvideo              923 non-null    int64
35  is_ecommerce               923 non-null    int64
36  is_biotech                 923 non-null    int64
37  is_consulting              923 non-null    int64
38  is_othercategory           923 non-null    int64
39  object_id                  923 non-null    object
40  has_VC                     923 non-null    int64
41  has_angel                  923 non-null    int64
42  has_roundA                 923 non-null    int64
43  has_roundB                 923 non-null    int64
44  has_roundC                 923 non-null    int64
45  has_roundD                 923 non-null    int64
46  avg_participants           923 non-null    float64
47  is_top500                  923 non-null    int64
48  status                     923 non-null    object
```

Fig. 1. Features of the dataset

B. Data Cleaning

The duplicate values would be a result of a wrong data entry. These values can impact the accuracy of the model. Removing all the duplicate values in the dataset is essential as they do not contribute to model generation. When we explored the dataset, there were no duplicate values.The next step in data cleaning is to remove the null values

A few columns were not included in the factors for predicting the startup's success because they are irrelevant data. For example, some columns like id and object id were just numerical values that were not used as a part of modeling. Unnamed columns were identified during analysis as trivial, which had many null values and were irrelevant. Closed at has more than half of the values as null . We dropped all these columns to hold onto the essential features for modeling, as shown in Figure 2.



	state_code	latitude	longitude	zip_code	city	name	founded_at	first_funding_at	last_funding_at	age_first_funding_year
0	CA	42.358880	-71.056820	92101	San Diego	Bandsintown	1/1/2007	4/1/2009	1/1/2010	2.2493
1	CA	37.238916	-121.973718	95032	Los Gatos	TriCipher	1/1/2000	2/14/2005	12/28/2009	5.1260
2	CA	32.901049	-117.192656	92121	San Diego	Plixi	3/18/2009	3/30/2010	3/30/2010	1.0329
3	CA	37.320309	-122.050040	95014	Cupertino	Solidcore Systems	1/1/2002	2/17/2005	4/25/2007	3.1315

Fig. 2. Sample dataset after dropping irrelevant columns

A few columns have NaN values, which were not required during the computation. Hence, to carry out any operations, we convert it into a numeric value such as 0 or the corresponding column's median value. The null values have been replaced or filled with the above two methods.The column lastyearfundings, have negative values and cause error in the model generation.The negative values are identified and dropped, so it does not impact the model's accuracy.

C. Exploratory Data Analysis

Data exploration is an essential task where we find the patterns in the data and also helps us to find the outliers. To explore the data, we use some data visualization libraries like Plotly, seaborn, and matplotlib.

Verifying the data distribution in the target variable is the most important task to check if the data in the column is balanced. If the data is imbalanced, we need to perform SMOTE, which will undersample or oversample according to our dataset.The distribution of the data in the status column in shown in Figure 3.

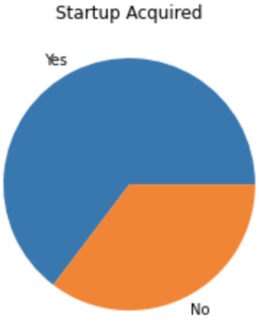


Fig. 3. Data Distribution

Correlation matrix is represented as heat map to find the correlation between the data.The strength of the relationship between the numerical values is the result of the heat maps.The correlation matrix for our dataset is shown in figure 4. Analyzing the outliers is also vital to make the model stand out while using various evaluation metrics. We used a dist plot to view the outliers in the dataset. Figure 5 shows the outliers analysis performed to check the outliers or noisy data in the dataset.

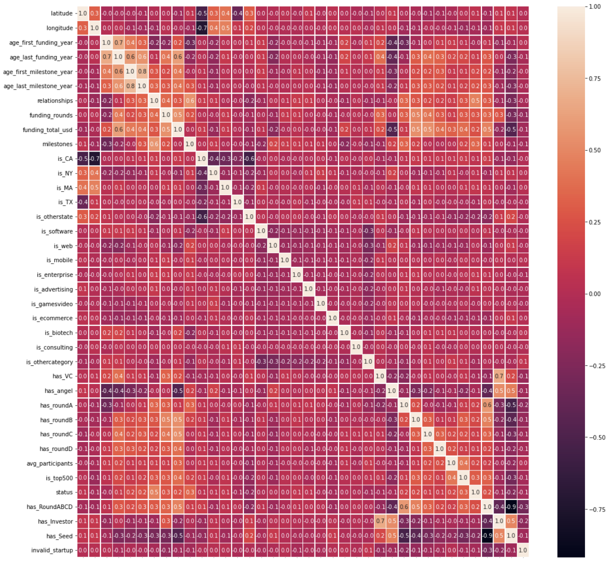


Fig. 4. Correlation Matrix

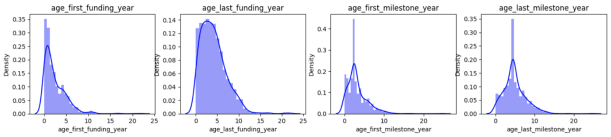


Fig. 5. Dist plot for outlier analysis

D. Feature Engineering

A machine learning technique called feature engineering uses data to generate variables not present in the training set. It can generate new features for supervised and unsupervised learning to streamline and accelerate data transformations while improving model accuracy. With machine learning models, feature engineering is necessary. A lousy feature will affect the model regardless of the architecture or the data.

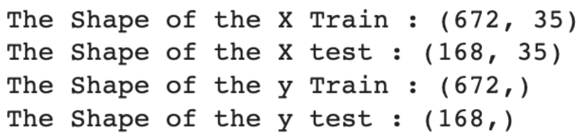
- To understand the credibility of each startup, we created "has investor" columns which let us know if any funding ever supported the startup.
- To analyze whether the startup company has its funding on its own or through any investors. We created a column "has seed."
- We analyzed rows by creating a new column as "is invalid startup" to maintain relevant data and avoid unnecessary computation. When the value of this column is 1, then the startup is invalid.

E. Scaling the Data

Feature scaling is a method for uniformly distributing the independent features in the data over a predetermined range. It is done as part of the prepossessing of the data to deal with highly variable magnitudes, values, or units. Without feature scaling, a machine learning algorithm would often prioritize larger values over smaller ones, regardless of the unit of measurement. We can standardize the dataset through the scikit learn object, StandardScaler.The scaler was applied to the entire dataset and then used to alter the dataset by individually standardizing each column. We can see that the values are centered around 0.0 and include both positive and negative values. The mean value in each column was a value of 0.0 if it exists.

F. Split the data.

We created training and testing sets by dividing the data set by 75 and 25 percent. This entails taking a random sample without replacing roughly 75 percent of the rows, adding them to the training set, and adding the final 25 percent to your test set.



```
The Shape of the X Train : (672, 35)
The Shape of the X test : (168, 35)
The Shape of the y Train : (672,)
The Shape of the y test : (168,)
```

Fig. 6. Data split

VI. Modeling

Set of label times, historical samples of what we want to forecast, and features, target variables are used to train a model which helps to predict the label, are the byproducts of prediction and feature engineering. Modeling entails teaching a machine learning algorithm to infer labels from features, fine-tuning it for the needs of the business, and testing it with holdout data. A trained model that may be used for inference, or generating predictions on new data points, is the result of modeling.

a) Choosing algorithms: We have considered a few machine learning classifier algorithms like Support Vector Classifiers (SVC), Decision Trees (DT), KNN, and Random Forests (RF). We considered classifiers as our use case tells whether the startup company would be closed. Thus, we need to classify whether the particular startup would succeed. We did ensemble modeling of the best-performing classifier algorithms to achieve better accuracy.

b) SVC: The supervised machine learning technique, SVC, or Support Vector Classifier, is frequently used for classification problems. SVC separates the data into two classes by mapping the data points to a high-dimensional space and locating the best hyperplane. The model gave us 70 percent accuracy for the testing data.

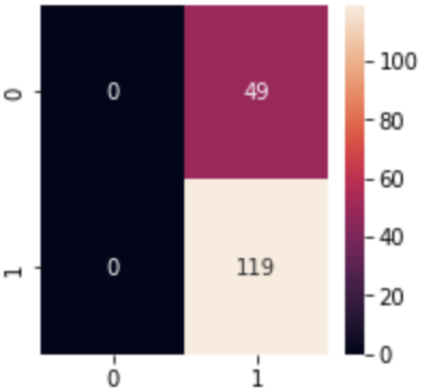


Fig. 7. Confusion Matrix for SVC

c) Random Forest Classifier: In decision trees, we rely on one node or the first tree, but the random forest takes the prediction from every tree. It bases its prediction of the final output on the majority votes of predictions. Random Forest is a classifier that contains many decision trees on various subsets of the dataset and takes the average of the trees to improve the predictive accuracy for that dataset. Random Forest Classifier gave the best accuracy among all the models. It gave us 87 percent accuracy for our datasets.

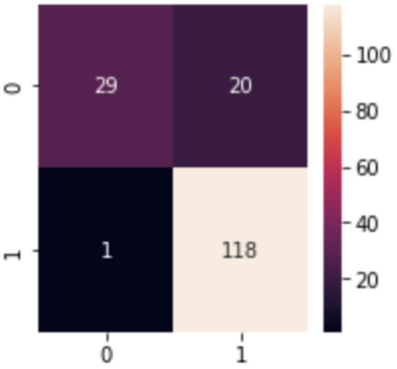


Fig. 8. Confusion Matrix for Random Forest Classifier

d) Logistic Regression: In a categorical dependent variable, the output is predicted via logistic regression. As a result, the result must be a discrete or categorical value. Rather of providing the exact values of 0 and 1, it provides the probabilistic values that fall between 0 and 1. It can be either Yes or No, 0 or 1, true or false, etc. It gave us 70 percent accuracy which was similar to SVC.

e) Decision Trees: Decision Trees are a sort of supervised machine learning in which the training data is continually segmented based on a particular parameter, with you describing the input and the associated output. Decision nodes and leaves are the two components that can be used to explain the tree. Since we had only two classes, decision trees performed better than SVC and Logistic Regression. It gave us accuracy of 7 percent.

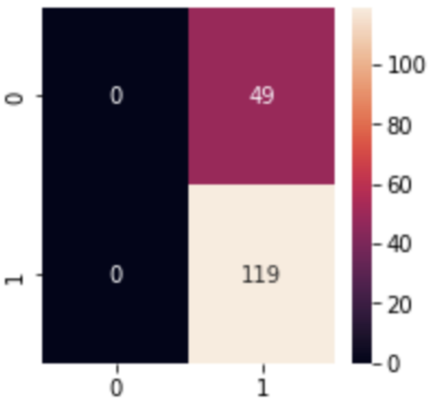


Fig. 9. Confusion Matrix for Logistic Regression

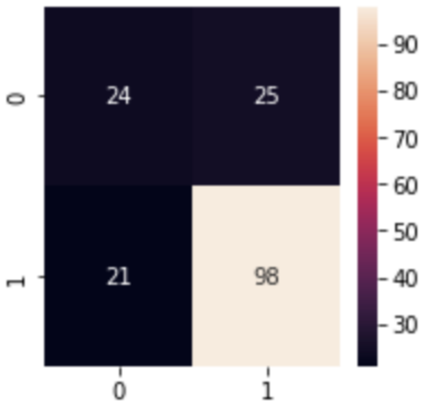


Fig. 10. Confusion Matrix for Decision Trees

f) Gradient Boosting: A machine learning method called gradient boosting is used, among other things, for classification and regression tasks. The model was used for prediction by combining the weak or individual decision trees. It performed better than SVC, logistic regression, and decision trees. The accuracy of the model was 78 percent.

g) KNN with neighbors = 2: To test how the dataset works with other algorithms, we implemented KNN with classes as two. The k-nearest neighbor's algorithm, also referred to as KNN or k-NN, is a supervised learning classifier that uses proximity to make classifications or predictions about the grouping of a single data point. It gave us the lowest accuracy of 64 percent.

h) Ensemble Modelling: Ensemble learning is a general meta-approach to machine learning that aims to improve predictive performance by combining the predictions from various models. Ensemble modeling often entails training each model on a distinct sample of the same training dataset while utilizing the same machine learning method, which is nearly invariably an unpruned decision tree. The ensemble member's forecasts have been merged using detailed statistics like voting or average.

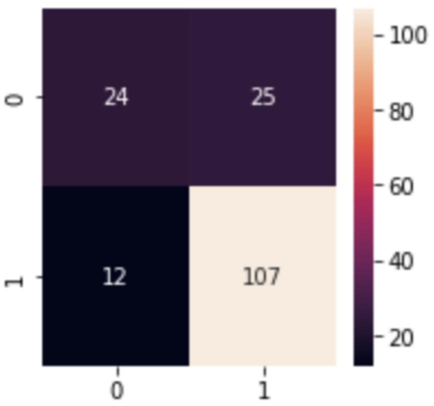


Fig. 11. Confusion Matrix for Gradient Boosting

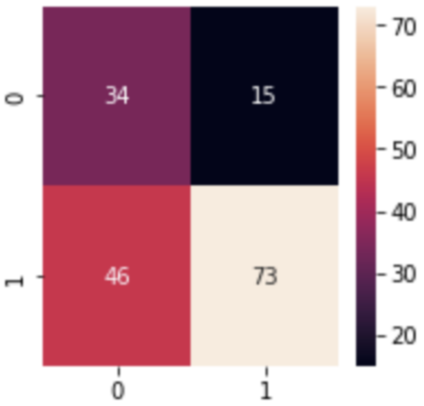


Fig. 12. Confusion Matrix for KNN

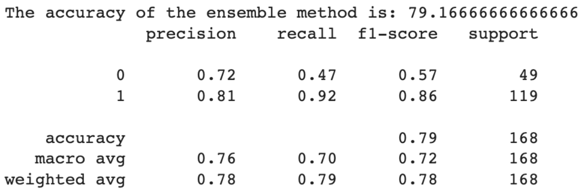
VII. Metrics for model evaluation

a) Recall Score: The machine learning model is more adept at recognizing both positive and negative samples the higher the recall score. Sensitivity or the true positive rate are other names for recall. A high recall score shows how well the model can locate examples of success.

b) ROC AUC: An indicator of performance for classification issues at different threshold levels is the AUC ROC curve. AUC stands for the level or measurement of separability, and ROC is a probability curve. It reveals how well the model can differentiate across classes.

c) Precision Score: The proportion of correctly predicted positive observations to all predicted positive observations is known as precision.

d) Cohen Kappa Score: The agreement between two raters who each assign N items to C mutually incompatible categories is measured by Cohen's kappa.To put it simply, Cohen's Kappa is a numerical indicator of reliability for two raters who are giving the same item the same rating, adjusted for the likelihood that the raters will agree by chance.

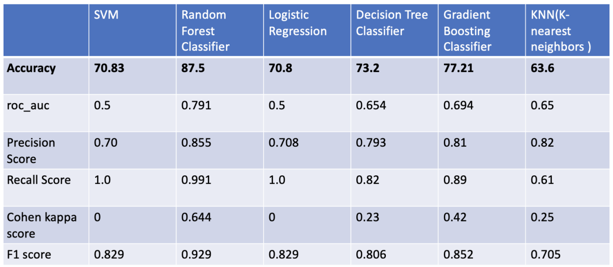


```
The accuracy of the ensemble method is: 79.16666666666666
              precision    recall  f1-score   support

           0       0.72      0.47      0.57        49
           1       0.81      0.92      0.86       119

    accuracy                           0.79       168
   macro avg       0.76      0.70      0.72       168
weighted avg       0.78      0.79      0.78       168
```

Fig. 13. Confusion Matrix for Ensemble Model



	SVM	Random Forest Classifier	Logistic Regression	Decision Tree Classifier	Gradient Boosting Classifier	KNN(K-nearest neighbors)
Accuracy	**70.83**	**87.5**	**70.8**	**73.2**	**77.21**	**63.6**
roc_auc	0.5	0.791	0.5	0.654	0.694	0.65
Precision Score	0.70	0.855	0.708	0.793	0.81	0.82
Recall Score	1.0	0.991	1.0	0.82	0.89	0.61
Cohen kappa score	0	0.644	0	0.23	0.42	0.25
F1 score	0.829	0.929	0.829	0.806	0.852	0.705

Fig. 14. Comparison of models using the metrics

VIII. Results

In our project, we implemented various classification algorithms. It helped us understand which algorithm would work best for this particular type of use case. To analyse further, we took help of various metrics and measures like Precision Score, Accuracy and ROC curve. We observed that Random Forest Classifier with the maximum accuracy of 86 percent. We choose to do ensemble modelling by picking the models and resulted into 80 percent accuracy.We can say that the features identified like amount of funding, location and current state are main factors in deciding the success of the start up.

IX. Conclusion

This project has demonstrated how machine learning can assist early-stage investors who are considering investments without any relevant quantitative data or track record in the baseline screening process. Our experiment in a practical environment shows that machine learning classifier can assist in boosting an investor's success rate.It's essential to understand some characteristics that could indicate a company will beat its competitors in order to lower risk and uncertainty while investing. It can't and shouldn't be the sole tool used to assess a firm, but it is unquestionably the first action an investor should do to determine whether or not to continue a dialogue. In our project, the logistic regression outperformed all the other classifiers. To balance the accuracy, we implemented ensemble modelling which improved our accuracy. The work that is described here can also be improved upon and some of its flaws can potentially be fixed. New variables could be taken into consideration in addition to those involved in classifier training, such as hyper parameter optimization or the use of sampling techniques for unbalanced classes.We can implement multi class classifier for better relevancy and accuracy.

References

[1] 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 ‘

[2] 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

[3] Jinze Li. 2021. Prediction of the Success of Startup Companies Based on Support Vector Machine and Random Forset. In 2020 2nd International Workshop on Artificial Intelligence and Education (WAIE 2020). Association for Computing Machinery, New York, NY, USA, 5–11. https://doi.org/10.1145/3447490.3447492

[4] 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

[5] A. Krishna, A. Agrawal and A. Choudhary, ”Predicting the Outcome of Startups: Less Failure, More Success,” 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016, pp. 798-805, doi: 10.1109/ICDMW.2016.0118.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

X. Appendix

A. Visualization

As recommended we as a team worked together to write the report .The report is our original work and for the references we used we have cited them in the reference section.The visualisations are also included in slides and report

B. Significance to the real world

Our project is really useful for the people who want to work in a startup and also the investors who are interested to provide fund for the growth of the companies.This is briefly explained in the section ii in report and also included in slides

C. Saving the model for quick demo

For saving and loading the model we used a pickle python package which is used to save the code for the model and when we want to reuse the model we can just all the pickle file were we stored the model.The .pkl file can be seen with the code

D. Code Walkthrough

During presentation we are planning to explain our code so that viewers can understand them better and get a hold on what we have done in our project

E. Report

Followed all these rubrics and the report is attached as both Latex and pdf format.The contents in the report are checked for format,language,plagiarism and grammar by using grammarly

F. Version Control Use of git/github

We are planning to attach the link of github with version control. The works which we had done as a team for this project is added in github link.

G. Discussion / Q and A

The questions and answer time will be used properly by the team to answer all the questions in the final presentation and also a slide is allocated at the last for discussion

H. Lessons learned

Analyzing the outliers helped us to show better accuracy and recall score. Model's performance is not just calculated with the accuracy ,but other metrics like recall, precision, Cohen kappa score plays a major role. Random Forest Regressor is a better classification algorithm for such kinds of dataset. When we need the data to be more accurate, we need to use SVM or Logistic Regression algorithms. Ensemble modelling gave us better accuracy than individual models.

I. Prospects of winning competition / publication

The projects main goal is to predict the success of the startups using the dataset.This model can be used by many startups to evaluate themselves and also help the investors to reduce the risk of losing money on a unsuccessful company.We have used multiple models and also compared the accuracy which is very unique about our project.We have used more than six metrics to evaluate the performance so that we can better understand which model is best with overall metrics.The chances are high for us to publish this paper in some reputed journal.

J. Innovation

We performed many data cleaning steps which gave us a better accuracy. Our visualisations and EDA gave us insights which helped us for modeling. Used more models to see which models outperformed well and also many evaluation metrics which helped us to choose the best model for the given dataset

K. Evaluation of performance

We used F1 score,recall,precision, accuracy,cohen kappa score as metrics to evaluate the performance of the model. We also found the roc score of the models we implemented to get the in depth understanding of the model.

L. Teamwork

The team has contributed equally for the project and report work.The coordination between the team helped to deliver the project and other deliverable at time

M. Technical difficulty

Interactive user interface visualization should have been implemented for users to easily make use of this prediction problem. But since our project is not focused on front end part. We were not able to do it. We were focused on models relevant to our coursework. May be other models might give a better prediction compared to the models we used. Volume of the dataset is very less, we felt we could have achieved better accuracy when there is more data. Due to time constraint, we were unable to implement hyper parameter tuning for all the models, but we tried to alter the parameters while doing modelling. Hyper parameter tuning might give a better accuracy for models.

N. Practiced pair programming

We collaborated and worked on projects.We followed virtual pair programming using the google collab and also added them in the report.We also used GitHub to do the pair programming where we added the code and report required and worked together

O. Practiced agile / scrum (1-week sprints)

We used Trello to manage our project which helps to schedule tasks for each members and also maintain a agile board.

P. Slides

We have created slides for all the topics covered in our project as per the rubrics.The slides are concise and contains the significant part of our project.

Q. LaTex Format

We have used Overleaf template. As per rubrics we have followed the IEEE LaTeX template.We generated the report with .tex format and also in a pdf format.

R. Creative presentation techniques

We have created slides with necessary screenshots, points and relevant images. Followed rubrics and included designer slides with respect to each topic.

S. Literature Survey

We have added the literature survey of the significant papers that are related to startup success prediction. Explained each paper with respect to title, goal, algorithms used and results. We have added this in slides and report as per rubrics.