Available online @ https://jjem.jnnce.ac.in
https: www.doi.org/10.37314/JJEM.SP0279
Indexed in International Scientific Indiexing (ISI)
Impact factor: 1.395 for 2021-22
Published on: 08 December 2024

Startup Success Prediction Using Machine Learning Algorithms

Harish B. G[1], Achuth Kumar R[2], Ranjith Kumar C K[3]

UBDT College of Engineering, Affiliated to Visvesvaraya Technological University, Belagavi, Davangere-577 004

[1]harishbg@ubdtce.org, [2]acuthkumar11@gmail.com, [3]ranjithkumarckranjith@gmail.com

Abstract

The importance of startups for a dynamic, innovative and competitive economy has already been acknowledged in the scientific and business literature.. The startup ecosystem is characterized by high levels of uncertainty and volatility, which makes the process of evaluating company success through information analysis and interpretation laborious and computationally demanding. This prediction dilemma highlights the necessity for a quantitative methodology that should allow for an unbiased, fact-based method of predicting startup success. The information utilized in this analysis was obtained from crunchbase.com, an online investment platform. The oversampling technique, ADASYN, has been used to pre-process the data for sampling bias and imbalance. Four distinct models are employed in order to forecast the success of a startup. The best models chosen are the ensemble approaches, random forest, and extreme gradient boosting, with test set prediction accuracy of 94.1% and 94.5%, respectively, using goodness-of-fit metrics that are applicable to each model situation.

Keywords: *Decision tree, Random forest , K-Nearest Neighbors (KNN)*

1. Introduction

The fast expansion of new businesses has emerged as a major catalyst for creativity and economic progress on a global scale. Although they offer promising potential returns, startups encounter considerable challenges, including difficulty in obtaining funding, establishing a market presence, or maintaining business activities. As a result, forecasting the triumph of startups has attracted significant interest from entrepreneurs, investors, and scholars alike. Precise predictions offer important guidance for making investment choices, allocating resources, and developing strategic plans.

In this research, we investigate the use of different machine learning algorithms for forecasting the success of startups. Utilizing data from Crunchbase, a thorough platform for business data on private and public firms, our goal is to construct predictive models that can pinpoint the crucial factors that influence startup success. Our main focus is on three commonly-used classification algorithms:

Decision Tree, Random Forest, and K-Nearest Neighbors (KNN).

Decision Trees provide a clear and easy-to-understand method for classification, dividing data according to the importance of features and hierarchical decision rules. Random Forests, a technique that combines multiple Decision Trees, boost prediction accuracy and reliability by addressing overfitting and variance issues. K-Nearest Neighbors (KNN), a non-parametric approach, categorizes instances based on the majority class among their nearest neighbors, making it a straightforward yet powerful algorithm for diverse applications.

To tackle the inherent imbalance in startup success and failure data, we utilize Adaptive Synthetic Sampling (ADASYN), which generates synthetic samples to equalize the class distribution. This preprocessing step guarantees that our models are trained on a more balanced dataset, leading to enhanced predictive performance.

Our study enhances the field of business analytics by offering a comparative analysis of various machines learning algorithms concerning the prediction of startup success. The knowledge obtained from this research can help entrepreneurs comprehend the key

elements that impact the success of their ventures and assist investors in making well-informed decisions.

2. Materials and Methodology

2.1 Materials(Dataset)

Dataset revolves around the placement season of India. Where it has various factors on candidates getting hired such as work experience, exam percentage etc., Finally it contains the status of recruitment and remuneration details. I have used the historical dataset. The dataset given by the source is fairly accurate and it goes back several years. The dataset has 215 rows and 15 columns.

2.2 Methodology

2.2.1 Input Parameters

The input parameters for the startup success prediction project include: state_code, latitude, longitude, city, name, labels, has_VC, has_angel,has_roundA,has_roundB,has_roundC, has_roundD, avg_participants, is_top500, and status.

2.2.2 Data Collection

The data utilized in this research is sourced from Crunchbase, a well-known platform that offers in-depth insights into both private and public companies, including startups. Crunchbase serves as a valuable tool for gathering thorough data on startups, encompassing different facets of their activities, financing, and expansion.

2.2.3 Data preprocessing

Data Cleaning:
Address missing values by imputing them with appropriate statistics (e.g., mean, median, mode) or removing records with excessive missing data. Eliminate duplicates to maintain data integrity.

Feature Selection:
Identify and choose relevant features that contribute to predicting startup success. Transform categorical variables into numerical values using techniques such as one-hot encoding or label encoding.

Handling Imbalanced Data:
Utilize ADASYN (Adaptive Synthetic Sampling) to balance the class distribution between successful and unsuccessful startups

2.2.4 Model Development

Decision Tree:
Decision tree displays predictions from a succession of feature-based splits using a flowchart that looks like a tree structure. To make a decision it starts with the root node and concludes with the leaf nodes. The algorithm is divided into two classes: the main class, which includes the algorithm, and a helper class, which defines a node.

Random Forest:
Random Forests train several trees by utilizing a random sample of the statistics. This makes the model more robust and less overfitting. The final class of an instance is assigned by outputting the class that is the mode of individual tree outputs, which can create robust and accurate classification and manage a huge number of input variables

K-Nearest Neighbors (KNN):
Establish a K-Nearest Neighbors classifier, which categorizes a startup based on the predominant class among its closest neighbors. Fine-tune the number of neighbors (k) to determine the optimal value for precise forecasts.

2.2.5 Model Evaluation

2.2.6 Train-Test Split:
Split the dataset into training and testing sets (e.g., 80% training, 20% testing) to evaluate the performance of the models.

3. Results:

There are many options available for assessing model performance and choosing a metric by which to evaluate the three models that were put into use. An overall evaluation of perfor- mance at different threshold levels Since every metric has pros and cons.After appropriate data preparation for each model, and running the respective models, the accuracy that has been achieved for each is depicted in Figure . All models had an accuracy of over 96.6%. The logistic regression model gave the highest accuracy of around 100%. The lowest is the K-Nearest Neighbors model with the accuracy being a little over 90%. Decision tree and random forest classifiers per-formed similarly, with accuracies over 100%.

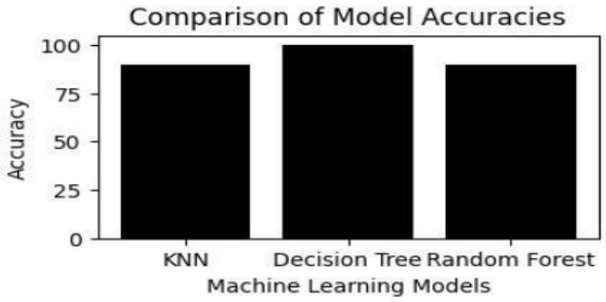

Fig 1. Comparative study of different machine learning models

Sr. No.	Model name	Accuracy (in %)
1	Decision Tree	100
2	K-Nearest Neighbors	90
3	Logistic Regression	100

Table 1 Accuracy of models used

4. Conclusion

This research takes a close look at predicting startup success. The volume of research on startup success has made it clear that additional study is required. The literature currently in publication focuses on forecasting established firm success rates. But there are big differences between corporate and startup success predic- tion, therefore the models that are currently in use are not useful for predicting startup suc cess. Because processing large amounts of data requires a lot of energy and time, actors in the startup ecosystem stand to gain a lot by using a quantitative approach when making decisions in such a high-risk environment.

We have constructed models for early-stage company success/failure prediction using a variety of machine learning algorithms. For the corresponding models, precision accuracy of 100%,90%,100% has been attained. We can categorically state that any early-stage startup can utilize our prediction models (at every milestone) to anticipate their fate, given the quality of the predictions. Our data also leads us to the conclusion that there is a high correla-tion between the above mentioned fea-tures(TABLE 1) and being a successful startup company. Because getting funds based on the idea does not lead to a successful company there should be people in the core-committee that have general and business-specific knowledge.

5. Future Scope

By utilizing deep learning techniques like neural networks, current methodologies can be advanced to capture more complicated patterns in the data. Boosting and stacking are examples of ensemble approaches that integrate many models to enhance overall predicting performance. Another crucial area is feature engineering, where the accuracy of the model may be improved by incorporating extra data like economic indicators, social media metrics, or the backgrounds of the founders. To further enhance data representativeness and diversity, additional datasets from other sources can be gathered and synthetic data generation techniques applied.

References

1. Afolabi, Ibukun, Ifunaya, T. Cordelia, Ojo, Funmilayo G., Moses, Chinonye. "A Model for Business Success Prediction using Machine Learning Algorithms." Journal of Physics: Conference Series, vol. 1299, no. 1, 2019, pp.012050. DOI: 10.1088/1742-6596/1299/1/012050.

2. Kim, Jongwoo, Kim, Hongil, Geum, Youngjung. "How to succeed in the market? Predicting startup success using a machine learning approach." Technological Forecasting and Social Change, vol. 193, August 2023, pp. 122614.DOI: 10.1016/j.techfore.2023.122614.

3. Misra, Ajai, Jat, Dharm Singh. "Machine Intelligence for Predicting New Start-ups Success: A Survey." International Conference on Data Science, Machine Learning and Artificial Intelligence, September 2021. DOI: 10.1145/3484824.3484919.

4. Sharchilev B., Roizner M., Rumyantsev A., Ozornin D., Serdyukov P., de Rijke M. Web-based Startup Success Prediction. Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 2018,2283–2291 DOI: 10.1145/3269206.3272011