CAHIERS MAGELLANES-NS
Volume 06 Issue 2
2024

ISSN:1624-1940
DOI 10.6084/m9.figshare.263651
http://magellanes.com/

STARTUP PROFIT RATE PREDICTION SYSTEM USING MACHINE LEARNING ALGORITHMS

[1] Vanshika, [2] Srishti Mathur, [3] Vandita Singh, [4] Srishti Vashisht
[1234] Department of CSE, MIET, Meerut

ABSTRACT:
Startups are dynamic entities that contribute significantly to economic growth and innovation. Predicting their profit rate is crucial for investors, stakeholders, and entrepreneurs to make informed decisions. We introduce a new method employing machine learning algorithms to forecast startup profit rates. The study investigates various factors such as funding, market trends, and operational metrics to develop accurate prediction models. The findings showcase the effectiveness of machine learning methods in forecasting startup profit rates, offering valuable insights for stakeholders in the startup ecosystem, including investors and entrepreneurs.

KEYWORDS – Startup profitability prediction, Machine learning algorithms, Predictive analytics, Feature selection, Data preprocessing, Regression analysis.

1. INTRODUCTION

Startups represent a dynamic and vital component of the modern economy, driving innovation, job creation, and economic growth. However, the high failure rates associated with startups pose significant challenges for investors seeking to allocate capital effectively. Making informed investment decisions in the startup landscape requires a comprehensive understanding of various factors influencing their success or failure. Among these factors, predicting the profitability of a startup remains a crucial yet elusive task. The goal of this project is to create a predictive modeling system that utilizes machine learning algorithms to forecast startup profit rates. By analyzing historical data on startup attributes, market conditions, and outcomes, this system aims to generate reliable predictions of profitability, enabling investors to make data-driven investment decisions.

The project aims to create a predictive modeling system that utilizes machine learning algorithms to forecast startup profit rates by analyzing historical data on startup attributes, market dynamics, and outcomes. This system aims to provide reliable predictions of profitability.

1.1 LITERATURE REVIEW

Machine learning algorithms are crucial in addressing classification challenges, illuminating their strengths, weaknesses, and practical applications. In their study on predicting house prices using ML algorithms, Satish, G.N. and Raghavendran [1] utilized Linear and multiple regression to analyze variable relationships. They also incorporated techniques such as lasso regression and gradient boosting to enhance model optimization and accuracy.

ISSN:1624-1940
DOI 10.6084/m9.figshare.263651
http://magellanes.com/

In these papers [4-6] research aims to reduce error, and probability and enhance future performance estimation. While the results are promising, there is room for improvement by considering additional parameters in the predictive models. They used programmed prediction methods, such as Artificial Neural Network (ANN) and Random Forest; there has been improved efficiency in stock price prediction.

Čeh, M.,Kilibarda [11] used the dataset that included various explanatory variables such as structural characteristics of apartments, age, environmental factors, and neighborhood information, organized in a Geographic Information Systems (GIS) format. Results consistently showed that Random Forest outperformed traditional models, indicating its potential for accurate apartment price prediction.

The literature on predicting startup success has increasingly recognized the vital role startups play in economic dynamism, innovation, and competition. In these [13-15], machine learning methods have gained traction for startup success prediction due to their ability to handle large-scale data and complex patterns. These models aim to provide reproducible and quantified predictions, offering an objective approach to evaluating startup success.

2. PROPOSED METHODOLOGY

The outlined methodology comprises the following steps:-

Data Gathering: Gather a comprehensive dataset containing attributes of startups such as market size, funding rounds, team composition, industry sector, geographical location, and profitability metrics. Publicly available datasets from platforms like Kaggle, UCI Machine Learning Repository, and government databases are also explored to supplement the dataset with additional variables and observations, enhancing its breadth and depth.

Data Pre-processing: - Data pre-processing for a startup profit rate prediction system involves several essential steps. Initially, relevant data sources are collected, including financial records and market trends. The data is then cleaned to handle missing values, errors, and inconsistencies, ensuring its quality. Categorical data is encoded into numerical format, and imbalanced datasets are addressed

using techniques like over-sampling or under-sampling. The dataset is divided into training, validation, and testing sets to develop and assess the model. Finally, the pre-processed data is ready for training predictive models to forecast startup profit rates accurately.

Selecting a Model:- Selecting a model for a startup profit rate prediction system involves understanding the problem and dataset characteristics, choosing candidate models such as linear regression, decision trees, support vector regression, and Random Forest, and evaluating performance using metrics like mean squared error or classification accuracy. Hyperparameter tuning and ensemble methods like bagging or boosting can enhance model performance.

Forecasting Profit:- Utilizing one of the previously mentioned models,profit prediction can be conducted by inputting the relevant variables.These factors encompass expenditures allocated for various purposes like research and development (R&D), marketing, and administrative functions.

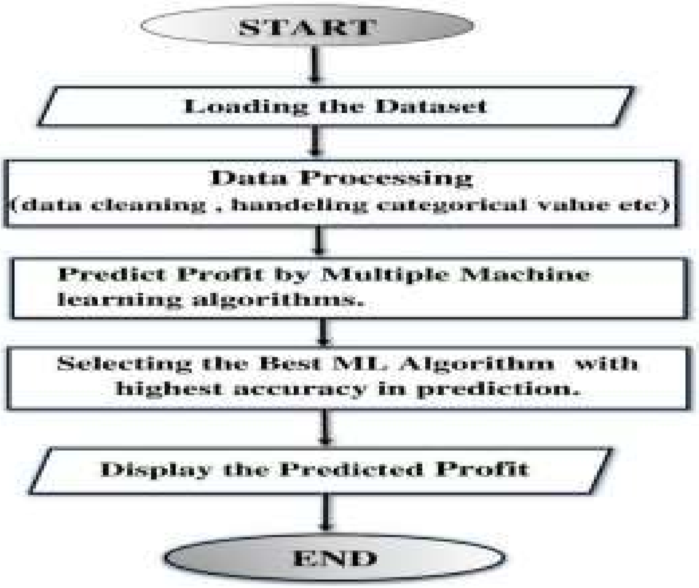

START
Loading the Dataset
Data Processing
(data cleaning , handeling categorical value etc)
Predict Profit by Multiple Machine learning algorithms.
Selecting the Best ML Algorithm with highest accuracy in prediction.
Display the Predicted Profit
END

Fig. 1 Flowchart of Methodology

Multiple Linear Regression: - It functions as supervised machine learning method primarily utilized for prediction and forecasting objectives. It comes into play when there are multiple independent variables to consider. In this context, the focus is on predicting the profit (the target variable) based on various dependent variables such as Administration Spend, Marketing Spend, and R&D Spend. The model is trained using multiple linear regression, resulting in an R-squared score of 0.9282. The regression equation, with the dependent variable as Y and k data points, is expressed as $Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + ... + \beta kXk$, where X1, X2, X3,...,Xk represent independent variables and β1, β2, β3,...,βk denote their corresponding coefficients.

K-Nearest Neighbors (KNN): - For the startup profit prediction system, the KNearest Neighbors (KNN) algorithm can be adapted to predict profit rates based on the characteristics of startups. Instead of classifying or regressing based on features like in traditional applications, KNN can be utilized to identify similar startups in the dataset based on their attributes such as R&D spending, marketing budget, and administrative costs. To implement KNN for profit prediction, the system would first need to define a distance metric to measure the similarity between startups. This could involve calculating Euclidean distance or other similarity measures between feature vectors representing each startup's attributes.

Support Vector Regression (SVR): - It is an algorithm applied in startup profit prediction systems to forecast profit rates. It selects a hyperplane within a high-dimensional feature space to optimally fit the data points, Ensuring the widest margin between the hyperplane and the nearest data points is a central focus of SVR. It can handle linear and nonlinear relationships between variables using different kernel functions. After preprocessing the data and training the SVR model, it can predict profit rates for new startups based on their feature values. SVR offers the advantage of capturing complex relationships but requires careful selection of hyperparameters to avoid overfitting.

Random Forest: - It is a versatile algorithm commonly utilized for regression purposes,including predicting profit rates in startup profit prediction systems. It belongs to the ensemble learning family, which combines multiple individual models to enhance predictive precision and generalization.

It offers several advantages for startup profit prediction:-

Robustness to Overfitting: By aggregating forecasts from numerous trees,it mitigates the chances of overfitting in contrast to a solitary decision tree.

Ensemble Learning Benefits: Leveraging ensemble learning, it combines multiple decision trees to achieve superior prediction performance and generalization ability.

Scalability: It can handle large datasets with numerous features effectively, making it suitable for real-world applications in startup environments.

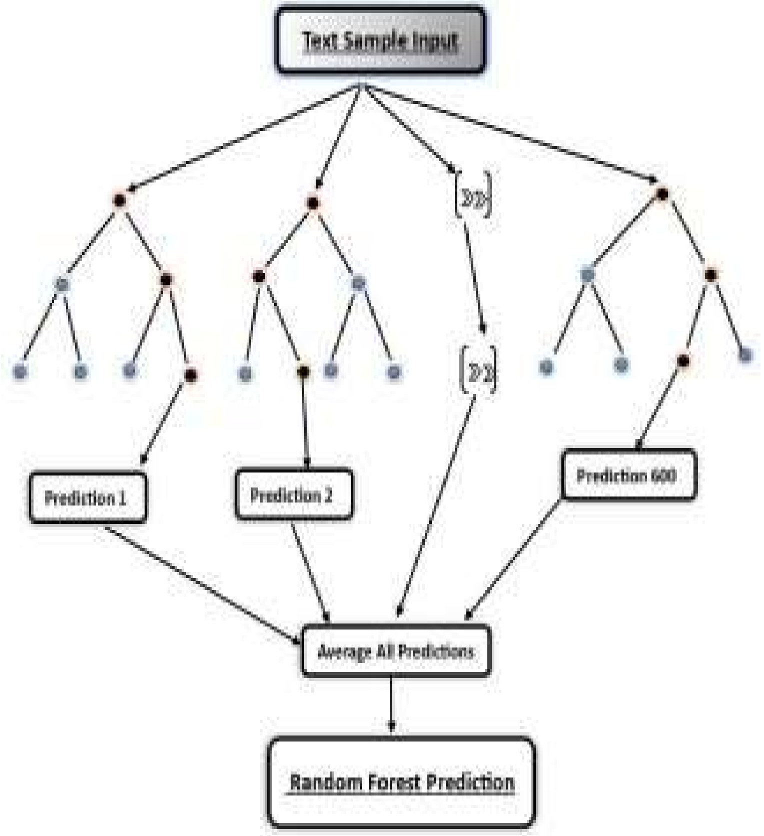
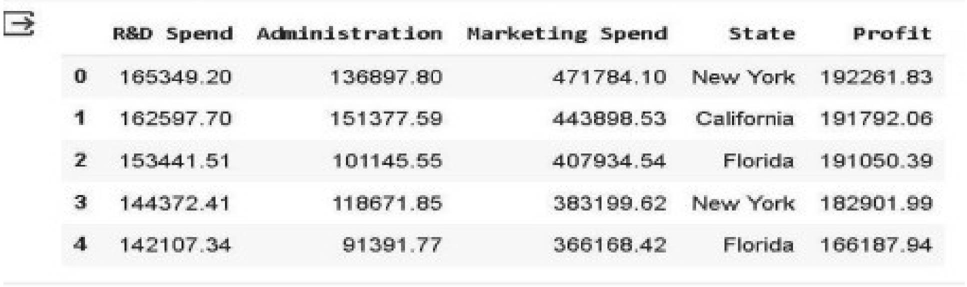


Fig. 2 Working of Random Forest regression

3. EXPERIMENTAL RESULT ANALYSIS

Description of the data set used: - The dataset incorporates key variables such as research and development (R&D) spending, marketing spending, administration expenses, and geographical location (state). These factors offer insights into how startups allocate resources, execute operational strategies, and navigate regional economic dynamics, ultimately impacting their profitability.



	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

Table 1 Description of the data set

ISSN:1624-1940
DOI 10.6084/m9.figshare.263651
http://magellanes.com/

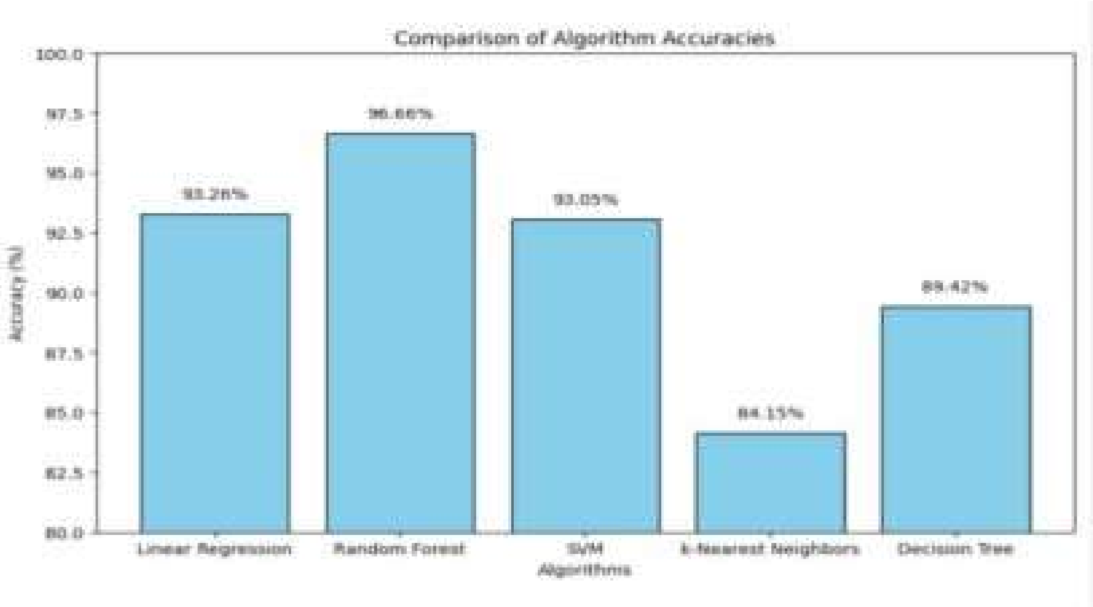


Fig. 3 Comparison of algorithms accuracies

Random Forest regression outcome:

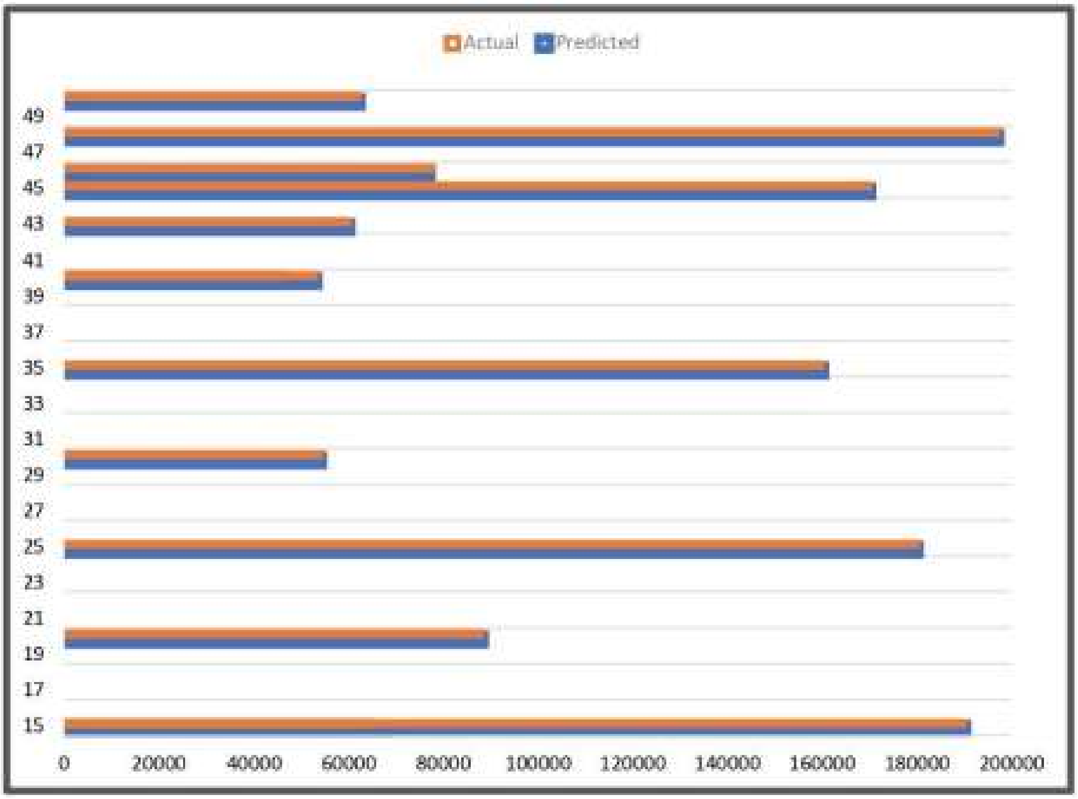
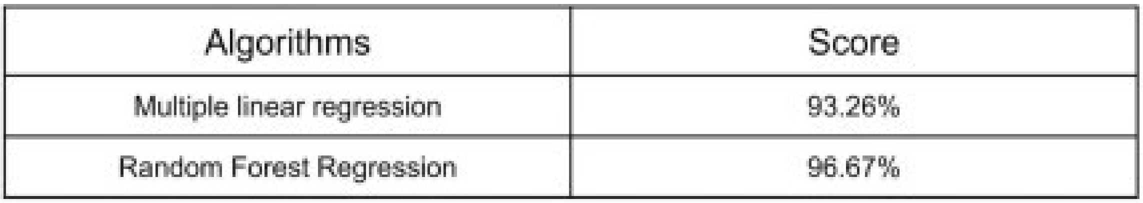


Fig. 4 This graph displays the actual profit values and predicted profit values across different data points.



Algorithms	Score
Multiple linear regression	93.26%
Random Forest Regression	96.67%

Table 2 Performance Analysis

The results from implementing two regression algorithms summarized as follows:

In a dataset split with 70% used for training and 30% for testing, the Multiple Linear Regression model achieved an R^2 score of 0.9326. This indicates that approximately 93.26% of the variance in profit can be explained by the independent variables included in the model.

On the other hand, the Random Forest Regression model achieved a higher R^2 score of 0.9667, indicating that approximately 96.67% of the variance in profit can be explained by the independent variables. These results suggest that both models perform well in predicting profit, with Random Forest Regression showing slightly higher predictive accuracy compared to Multiple Linear Regression.

4. CONCLUSION AND FUTURE WORKS

The development and implementation of the startup profit rate prediction system have shown promising results in aiding investors and entrepreneurs in making informed decisions.The experimentation conducted in this research has demonstrated the feasibility and effectiveness of such predictive models in the context of startup investments. The future scope of the startup profit rate prediction system lies in further refining its predictive capabilities by incorporating additional data sources, exploring advanced machine learning techniques, conducting longitudinal studies, integrating feedback mechanisms, and extending its application to other business domains beyond startups. These advancements will enhance the system's accuracy, adaptability, and relevance, thereby offering valuable decision support to investors and entrepreneurs in dynamic market environments.

5. REFERENCES

[1] Satish, G. N., Raghavendran, C. V., Rao, M. S., & Srinivasulu, C. (2019). House price prediction using machine learning. *Journal of Innovative Technology and Exploring Engineering*, *8*(9), 717-722.

[2] Thamarai, M., & Malarvizhi, S. P. (2020). House Price Prediction Modeling Using Machine Learning. *International Journal of Information Engineering & Electronic Business*, *12*(2).

[3] Rawool, A. G., Rogye, D. V., Rane, S. G., & Bharadi, V. A. (2021). House price prediction using machine learning. *Int. J. Res. Appl. Sci. Eng. Technol*, *9*, 686-692.

[4] Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. *Procedia computer science*, *167*, 599-606.

[5] Shakhla, S., Shah, B., Shah, N., Unadkat, V., & Kanani, P. (2018). Stock price trend prediction using multiple linear regression. *International Journal of Engineering Science Invention (IJESI)*, *7*(10), 29-33.

[6] Emioma, C. C., & Edeki, S. O. (2021). Stock price prediction using machine learning on least squares linear regression basis. In *Journal of Physics: Conference Series* (Vol. 1734, No. 1, p. 012058). IOP Publishing.

[7] Kanakam, R., Ramesh, D., Mohmmad, S., Shabana, S., & Prakash, T. C. (2022, May). Stock price prediction using multiple linear regression and support vector machine (regression). In *AIP Conference Proceedings* (Vol. 2418, No. 1). AIP Publishing.

[8] Polamuri, S. R., Srinivas, K., & Mohan, A. K. (2019). Stock market prices prediction using random forest and extra tree regression. *Int. J. Recent Technol. Eng*, *8*(1), 1224-1228

[9] Keerthan, J. S., Nagasai, Y., & Shaik, S. (2019). Machine Learning Algorithms for Oil Price Prediction. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *8*(8), 958-963.

[10] Jui, J. J., Imran Molla, M. M., Bari, B. S., Rashid, M., & Hasan, M. J. (2020). Flat price prediction using linear and random forest regression based on machine learning techniques. In *Embracing Industry 4.0: Selected Articles from MUCET 2019* (pp. 205-217). Springer Singapore.

[11] Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS international journal of geo information*, *7*(5), 168.

[12] Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P., & de Rijke, M. (2018, October). Web-based startup success prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 2283-2291).

[13] Ünal, C. (2019). *Searching for a unicorn: A machine learning approach towards startup success prediction* (Master's thesis, Humboldt-Universität zu Berlin).

[14] Bento, F. R. D. S. R. (2018). Predicting start-up success with machine learning. *Universidade Nova de Lisboa*.

[15] Krishna, A., Agrawal, A., & Choudhary, A. (2016, December). Predicting the outcome of startups: less failure, more success. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* (pp. 798-805). IEEE.