Section: Research Paper

Profit Prediction Model using Machine learning Algorithms

Riya Verma, Satyam Gupta, Swati Bhadana ,Ritika Rajput ,Vaishali Deshwal, Vijay Kumar

riya.verma.cs.2019@miet.ac.in ,satyam.rajesh.cs.2019@miet.ac.in,swati.bhadana.cs.2019@miet.ac.in, ritika.rajput.it.2019@miet.ac.in

Meerut Institute of Engineering and Technology, Meerut

Abstract: - In today's world, people are more attracted towards starting their own new start-up but there is cut throat competition in the market in terms of product, profit, etc. To survive in the market and earn something high, different strategies are followed and several opinions and points are considered. The model aims at predicting the profit of any start up based on the values of certain parameters such as R&D Spend, Marketing Spend, Administration Spend and State in which start-up is located. Our model will be helpful in providing the profit amount to the users beforehand. Prediction of profit in advance, will be helpful for users to adopt and examine different strategies to enhance it. Random Forest and Multiple Linear Regression are the two machine algorithms that are implemented to provide the result with higher accuracy.

Keywords- Profit, Random Forest Regression, Prediction, Multiple Linear Regression, Startup

Introduction: - Start-up Companies plays an indispensable role in uplifting the economy by providing employment to experienced and young professionals. Start-up companies are generally set up with high costs and limited revenue and require capital from a variety of sources such as venture capitalists. Therefore, these reasons make start-up companies an important target of analysis. It becomes very difficult for the start-up companies to operate and tackle the problems in a very highly competitive environment. Profit earned by the start up company is an important factor in determining its ability to survive and enhance its business.

Profit of any start-up company largely depends on the structure of money distribution for different causes. Evaluation of profit will not only provide an overview of the gain but also helpful in developing future strategies and to reorder the money distribution for different parameters. Since Machine Learning possess the ability to solve complex problems, therefore machine learning is implemented in our model to make prediction of profit. Using appropriate dataset and machine learning algorithms, it is possible to predict the profit. The dataset and algorithm used in the model plays an indispensable role to determine the accuracy of result obtained.

This Profit Prediction Model aims at predicting the profit of the start-up based on the values of Administration Spend, R&D Spend, Marketing Spend and Location using different machine learning algorithms. The dataset consists of the data of 1000 start-ups. The dataset consists of the values of four parameters which are R&D Spend, Marketing Spend, Administration Spend and State in which the start-up is located. The selected dataset is then fed for further processing. Data Pre-processing including data cleaning and outlier removal is

done. The pre-processed data is then trained and tested. Out of the complete dataset, 30% of the dataset is used for testing purpose and the remaining 70% precent of the dataset is used for training.

Multiple Linear and Random Forest Regression are the two algorithms implemented on the dataset after training. From the results obtained by Multiple linear and Random Forest regression algorithm, it is found that Random Forest regression provides more accurate results as compared to multiple linear regression. Since people are very fascinated towards setting up their own business, this model would be helpful in providing them an overview about the profit that can be earned based on the money spent on R & D, Administration, Marketing, and location of the start-up. This Model will be helpful for the people willing to set up their new business firms and for the investors as well. It will be helpful for the owners of the start-up to analyse the links between the success parameters and the criteria.

Specifications of the Dataset used for the prediction-

Field Name	Value
Number of Rows	1000
Number of Columns	4
Number of Categorical Values	1
Number of numeric values	3
Number of target variables	1
Type of Problem	Classification

Table.1 Attributes of Dataset

Table.1 represents the attributes of the dataset of one thousand startups. There are different numeric values of spend on different areas involved such as marketing, administration and Research and Development. The dataset also includes the area in which the startup is situated.

	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
5	131876.90	99814.71	362861.36	New York	156991.12
6	134615.46	147198.87	127716.82	California	156122.51
7	130298.13	145530.06	323876.68	Florida	155752.60
8	120542.52	148718.95	311613.29	New York	152211.77
9	123334.88	108679.17	304981.62	California	149759.96

Fig.1 Dataset Representation

Figure.1 represents the actual state of the dataset used in this model to predict profit of different startup firms.

LITERATURE REVIEW

Manasi Chhibber [1] introduced Start-up Profit Predictor using Machine Learning which makes use of Random Forest regressor model to predict the profit of start-up companies based on the values of Administration Spend, R&D Spend, Marketing Spend and the State in which the start-up is located. The accuracy of their model turned out to be 96.91%. Mr. Kanhaiya Pandey [2] used Deep Neural Networks (DNN) to predict the profit of newly established companies and start-ups based on the values of R&D Spend, Marketing Spend, Administration Spend and State where the firm is present. In this model, seventy percent of the data is used for training purpose and rest thirty percent data is used for testing purpose.

Ajit Kumar Pasayat [4] developed an empirical approach to determine essential features for predicting the success of start-up. In the first part, an empirical machine learning approach is applied on ten datasets to determine critical features. The features obtained from the first part were validated using feature selection techniques. Hence the accuracy of the model turned out to be 90 percent.Ünal [5] introduced "A machine learning approach towards startup success prediction. Reproducible methods for startup success prediction have been presented using machine learning algorithms. Ensemble methods, Random Forest and Extreme gradient boosting are implemented. The accuracy of the ensemble methods turned out to be 94.1% and 94.5% and 92.22% and 92.91% in case of Random Forest and Extreme Gradient Boosting.

Pingwen Xue [6] designed a machine learning model using Improved Support Vector Machine (SVM) to forecast the net profit of enterprises. Different performance metrics such as Mean Squared Error, Root Mean Squared Error and Mean Absolute Percentage Error has been evaluated to make better analysis.

PROPOSED METHODOLOGY

Proposed Methodology consists of the following steps: -

1.Data Gathering: - This is the first step involved in the process of making profit prediction. The dataset used in this model is taken from the website of Kaggle. It consists of the data values of thousand start-up companies [7][8].

2.Data Pre-processing: - Data Pre-processing is done after loading the dataset. This step involves optimizing the dataset by removing the outliers and unnecessary details from the dataset to provide most accurate results. Since location does not has any continuous relationship with the value of profit therefore the location column has been dropped from the dataset for further steps. The entire dataset is split in the ratio of 3:7 for testing and training the model respectively [9][10].

3.Selecting a Model: - Random Forest and Multiple Linear Regression are the machine learning algorithms implemented in this model to predict the profit.

Random Forest Regression is majorly used to compute a variety of prediction problems where company needs a prediction of continuous value such as prediction of future prices, comparison of performances etc.

Multiple Linear Regression is usually used when there is a continuous dependent variable and two or greater than two independent variables. Similarly, for this model we have profit as a dependent variable and rest other parameters Marketing Spend, Administration Spend and R&D Spend as the independent variables [11][12].

4.Prediction of Profit: - One of the above-mentioned models can be used to make prediction of profit by providing the values of the required variables. These parameters include money spent for different causes such as R&D, Marketing and Administration purpose.

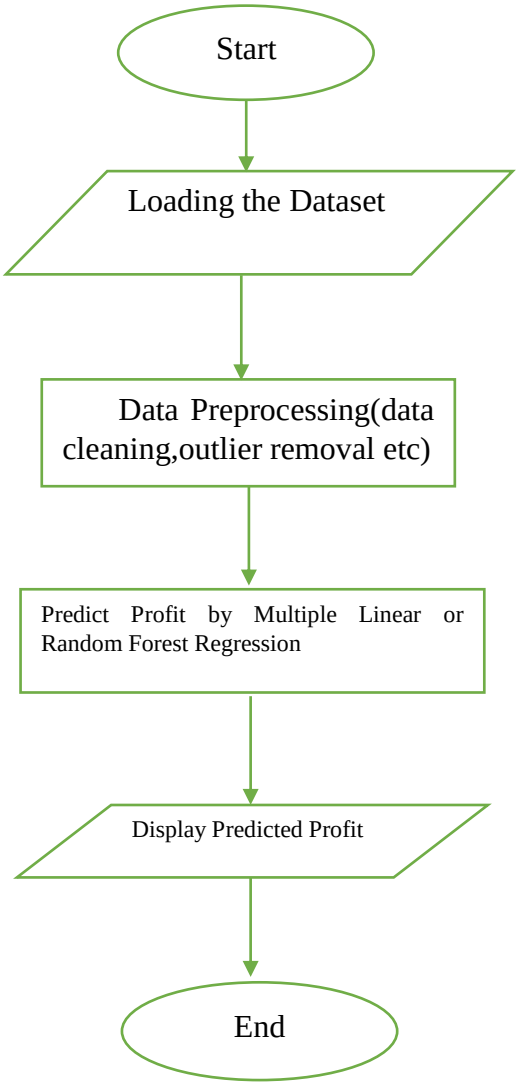

Fig.2 Flowchart of Methodology

Fig.2 shows the flowchart of this model. Different steps used in the methodology are represented in the above flowchart.

Multiple Linear Regression: - It is a supervised machine learning algorithm that is mostly used for the prediction or forecasting. Multiple linear regression algorithm is used when more than one independent variable is present. In this Model, the target variable is profit and rest all are the dependent variables, i.e., Administration Spend, Marketing Spend, R&D Spend

etc. After Data Pre-Processing the model is trained and fitted in the multiple linear regression model. The value of R2 score turns out to be 0.9282 [13][14].

Let Y be the dependent variable such that there are k data points

$Y=\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \ldots\ldots + \beta_kX_k$, where X_1, X_2, $X_3\ldots\ldots X_k$ represents various independent variables and β_1, β_2, β_3,…..,β_k represents the coefficients of these variables and k represents the number of parameters.

Random Forest Regression: - It is a supervised machine learning algorithm that implements ensemble learning for regression. In this method, different decision trees are built and the average of the results obtained from all the trees is returned as the answer. The model is trained with the number of estimators equal to ten. The R2 Score of Random Forest Regression turned out to be 0.989 which is far better than that of the Multiple linear regression algorithm. The basic working of Random forest regression is displayed by the figure below: [15][16] [17][18].

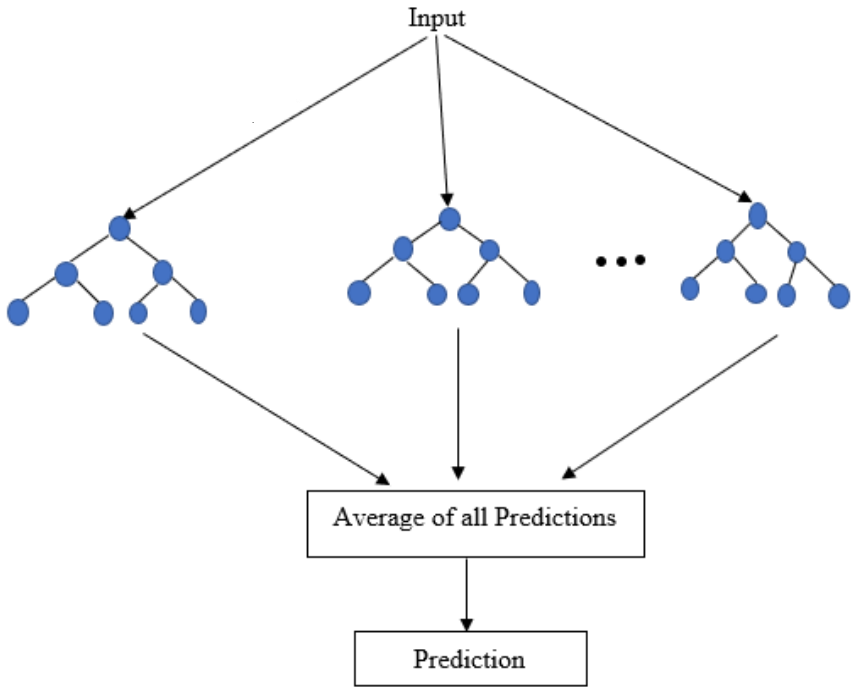

Fig.3 Working of Random forest regression

Fig.3 displays the working of Random forest algorithm in machine learning. From the input dataset, multiple decision trees are built as per user requirement. Then the mean of the results obtained from all decision trees is returned as the final answer. In our model, this machine learning algorithm produces the average of the results obtained from ten decision trees [19] [20].

RESULTS

The machine learning algorithms used in this profit prediction model on the chosen dataset provides different outcomes. A comparative analysis of the results obtained by the different machine learning algorithms is shown by the graphs below.

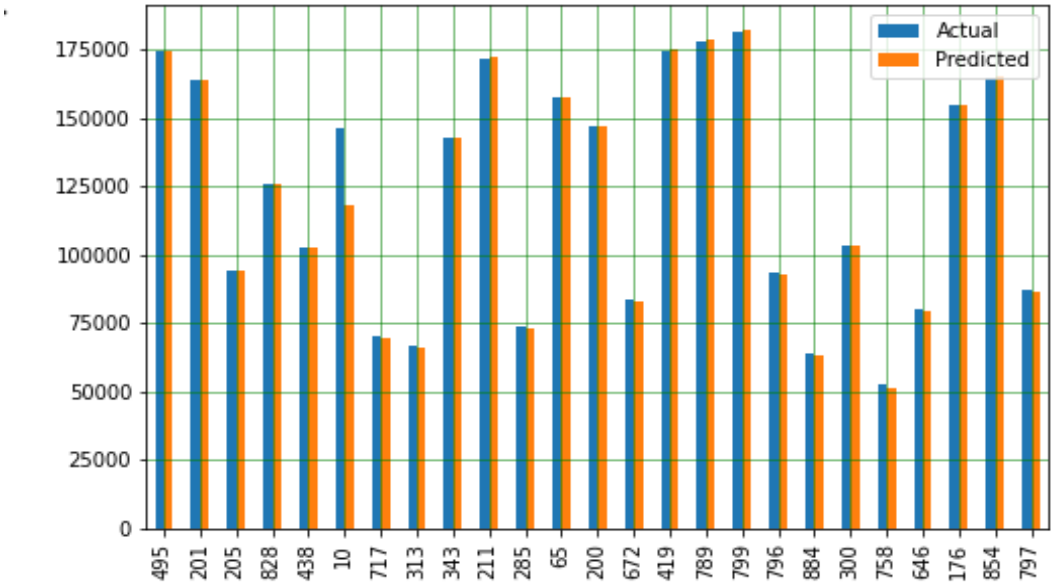

Fig.4 Multiple linear regression outcome

Figure.4 represents the results obtained by multiple linear regression. The blue line represents the actual profit and orange line represents the predicted profit. X axis shows row number and Y axis represents the profit.

	Actual	Predicted
927	60650.40749	60409.392180
226	111422.65230	111664.275523
977	89558.77320	89592.559439
254	164375.90290	165120.897404
449	72735.21323	72609.075514
421	71376.03566	71236.977592
941	148975.59230	149574.192180
383	136632.79060	137114.060389
979	60065.21791	59818.639728
11	144259.40000	103392.269483

Fig.5

Figure 5 represents the comparative analysis of the actual profit value and the predicted profit value obtained by multiple linear regression algorithm.

	Actual	Predicted
927	60650.40749	60409.392180
226	111422.65230	111664.275523
977	89558.77320	89592.559439
254	164375.90290	165120.897404
449	72735.21323	72609.075514
421	71376.03566	71236.977592
941	148975.59230	149574.192180
383	136632.79060	137114.060389
979	60065.21791	59818.639728
11	144259.40000	103392.269483

Fig.6

Figure 6. represents the comparative analysis of the actual profit value and the predicted profit value obtained by random forest regression algorithm.

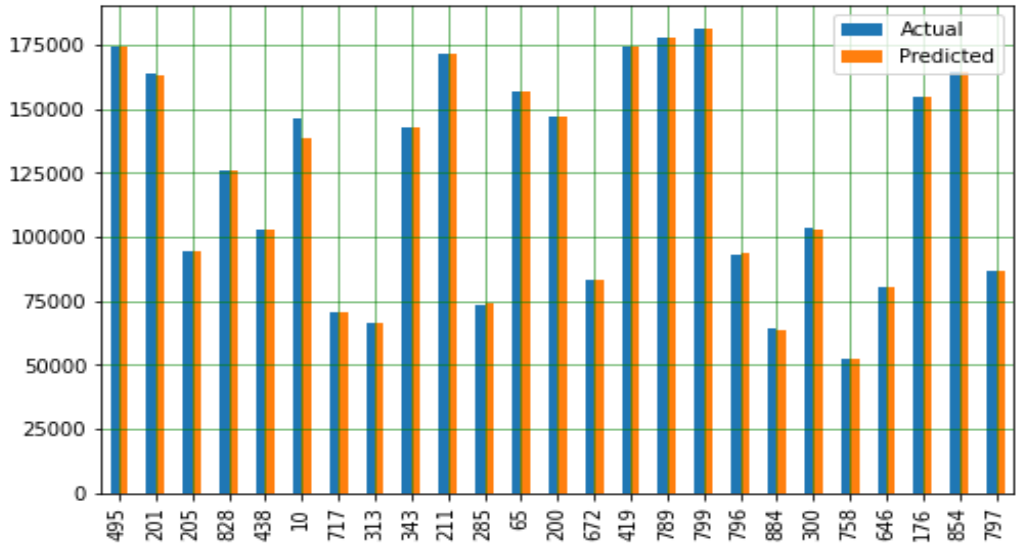

Fig.7 Random forest regression outcome

Figure.7 represents the results obtained by the implementation of random forest regression. X axis denote row number and Y axis represents the profit whereas blue line denotes the profit value and orange line represents the predicted profit value.

Algorithm	Mean Absolute Percentage Error	R2 Score
Multiple Linear Regression	0.021	0.9203
Random Forest Regression	0.0039	0.9899

Table.2 Performance Analysis

Table.2 **shows** the results obtained by the regression algorithms that are implemented in this model. Multiple Linear Regression and Random Forest Regression are the two machine learning algorithms that are used for the prediction of profit in this model. Thirty Percent of the dataset is used for testing and seventy percent of the dataset is used for training purpose. In case of Multiple Linear Regression, the value of R2 Score turns out to be 0.9203 and the value of mean absolute percentage error turns out to be 0.021. In case of Random Forest Regression, the value of R2 Score turns out to be 0.9899 and the value of mean absolute percentage error turns out to be 0.0039.

CONCLUSION AND FUTURE WORKS

Profit prediction model is able to provide an overview to the people willing to set up a startup or are already running a startup about the value of profit that could be achieved based on the values of money spent on different factors. This model will be helpful in deciding the money distribution for enhancement of profit. This model efficiently analyses the trends and the values of the different parameters present in the dataset to predict the profit more accurately. It is observed that Random Forest Regression works well as compared to Multiple linear regression algorithm for the prediction of the value of profit based on different parameters including Spend on Administration, Marketing and Research and Development(R&D). Random forest regression provides more accurate results as compared to multiple linear regression algorithm.

Future work can be including different other factors into consideration for the purpose of prediction of profit of any business or startup company. In addition, different other machine learning algorithms can also be applied to provide a comparative analysis of the performance of different algorithms on this dataset for the prediction of profit value.

REFERENCES

1.Durganjali, P., & Pujitha, M. V. (2019). "House Resale Price Prediction Using Classification Algorithms" .2019 International Conference on Smart Structures and Systems (ICSSS).

2. Swain, S., Patel, P., & Nandi, S. (2017). "A multiple linear regression model for precipitation forecasting over Cuttack district, Odisha, India". 2017 2nd International Conference for Convergence in Technology (I2CT).

3. Kurniawati, N., Novita Nurmala Putri, D., & Kurnia Ningsih, Y. (2020). "Random Forest Regression for Predicting Metamaterial Antenna Parameters". 2020 2nd International Conference on Industrial Electrical and Electronics (ICIEE).

4. Rodrigues, N., Sequeira, N., Rodrigues, S., & Shrivastava, V. (2019). Cricket Squad Analysis Using Multiple Random Forest Regression. 2019 1st International Conference on Advances in Information Technology (ICAIT). doi:10.1109/icait47043.2019.8987367 .

5. Acharya, M. S., Armaan, A., & Antony, A. S. (2019). A Comparison of Regression Models for Prediction of Graduate Admissions. 2019 International Conference on Computational Intelligence in Data Science (ICCIDS). doi:10.1109/iccids.2019.8862140 .

6. Sun, W., & Zhao, W. (2010). Profitability evaluation of the power listed companies based on PSO-BP neural network model. 2010 2nd International Conference on Future Computer and Communication. doi:10.1109/icfcc.2010.5497824

7. Narayan, Vipul, et al. "To Implement a Web Page using Thread in Java." (2017).

8. Ulgen, T., & Poyrazoglu, G. (2020). Predictor Analysis for Electricity Price Forecasting by Multiple Linear Regression. 2020 International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM). doi:10.1109/speedam48782.2020.9161866

9. Narayan, Vipul, and A. K. Daniel. "FBCHS: Fuzzy Based Cluster Head Selection Protocol to Enhance Network Lifetime of WSN." ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal 11.3 (2022): 285-307.

10. Kavitha S, Varuna S, & Ramya R. (2016). A comparative analysis on linear regression and support vector regression. 2016 Online International Conference on Green Engineering and Technologies (IC-GET). doi:10.1109/get.2016.7916627

11. Narayan, Vipul, and A. K. Daniel. "Novel protocol for detection and optimization of overlapping coverage in wireless sensor networks." Int. J. Eng. Adv. Technol 8 (2019).

12. Mall, Pawan Kumar, et al. "FuzzyNet-Based Modelling Smart Traffic System in Smart Cities Using Deep Learning Models." Handbook of Research on Data-Driven Mathematical Modeling in Smart Cities. IGI Global, 2023. 76-95.

13. Faiz, Mohammad, et al. "IMPROVED HOMOMORPHIC ENCRYPTION FOR SECURITY IN CLOUD USING PARTICLE SWARM OPTIMIZATION." Journal of Pharmaceutical Negative Results (2023): 2996-3006.

14. Srivastava, Swapnita, et al. "An Ensemble Learning Approach For Chronic Kidney Disease Classification." Journal of Pharmaceutical Negative Results (2022): 2401-2409.

15. Mall, Pawan Kumar, et al. "Early Warning Signs Of Parkinson's Disease Prediction Using Machine Learning Technique." Journal of Pharmaceutical Negative Results (2023): 2607-2615.

16. Awasthi, Shashank, et al. "A Comparative Study of Various CAPTCHA Methods for Securing Web Pages." 2019 International Conference on Automation, Computational and Technology Management (ICACTM). IEEE, 2019.

17. Pentyala, S., Liu, M., & Dreyer, M. (2019). Multi-task networks with universe, group, and task feature learning. arXiv preprint arXiv:1907.01791.

18. Narayan, Vipul, et al. "Deep Learning Approaches for Human Gait Recognition: A Review." 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE, 2023.

19. Narayan, Vipul, et al. "FuzzyNet: Medical Image Classification based on GLCM Texture Feature." 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE, 2023.

20. Mall, Pawan Kumar, et al. "FuzzyNet-Based Modelling Smart Traffic System in Smart Cities Using Deep Learning Models." Handbook of Research on Data-Driven Mathematical Modeling in Smart Cities. IGI Global, 2023. 76-95.