Proceedings of ICFTBA 2024 Workshop: Finance's Role in the Just Transition
DOI: 10.54254/2754-1169/143/2024.GA18965

Estimate Company's Profit Using Least Squared Method and Random Forest Model

Bingkun Zhang[1,a,*]

[1]*Faculty of Computer Science, Hong Kong Baptist University, Hong Kong, China*
a. 22258531@life.hkbu.edu.hk
**corresponding author*

Abstract: Profit plays a significant role in company operations. It reflects operational efficiency and market competitiveness of a company. Companies not only rely on profit for their survival but also use it as an important parameter for decision-making. Moreover, Profitable companies are better positioned to fulfill social responsibilities. Therefore, companies consider profit to be one of the most important parameters. Another parameter that significantly impacts companies is cost. Costs play a unique role in the formulation of a company's business strategy, financial health analysis, and other areas, closely related to each enterprise's competitive advantage. Therefore, reducing costs and increasing profits have become the goals pursued by every profit-oriented enterprise. Finding the relationship or pattern between these two important pieces of data becomes even more significant. By establishing a model, the company can use the known patterns and existing cost data to estimate profits, thereby creating plans that better align with its future development. This paper aims to explore the relationship between cost and profit by categorizing total cost into three classes: R&D costs, marketing costs, and administrative costs. A sample size of 1,000 is used in this research to build ordinary least squares and random forest models to analyze the relationships between these attributes.

Keywords: Ordinary least square, random forest, cost, profit.

1. Introduction

Profit is not only a residual calculated after accounting for liabilities and other negative values, but it also serves as a key metric for profit-seeking companies [1]. Consequently, predicting profit is a significant aspect of performance management for these companies.

On the other hand, cost is deeply connected to profitability and has a significant impact on net profit [2-4]. More specifically, total cost can be separated into marketing cost, R&D cost, and administration cost. In terms of marketing cost, it positively influences companies' future performance. Researchers believe that marketing expenditure is directly related to profitability and may also impact future income [5]. Regarding R&D cost, from the investment process to technology innovation, R&D plays significant but varied roles [6, 7]. Lastly, concerning administration cost, studies have shown that it does not always have a strong connection with profit, but a correlation still exists [8, 9]. For most organizations, cost is a measurement that is relatively easy to calculate and helps stakeholders—both inside and outside the company-- understand the company's state.

DOI: 10.54254/2754-1169/143/2024.GA18965

This paper aims to provide a profit estimation method not only through simple linear regression, but also discovered relations between cost independents. Firstly, this research collects a suitable data set which contains required parameters and estimate the data. Secondly, it use ordinary least square method to build a model and make a prediction. Then, it finds out the relation between parameters according to the reflection after using OLS model to predict data. Lastly, this group of data is utilized to train random forest model. The prediction results is adjusted in order to discover the ability of trained models.

2. Data and Methods

2.1. Data

The data being used is from Kaggle, updated in 2022. In this dataset, 1,000 entries related to companies' performance are selected, including R&D expenditure, administration costs, marketing expenditure, and profit. As mentioned, three categories of costs are considered as independent variables, while profit is the dependent variable. The mean value of profit is 119,546.16. There are no missing data points.

2.2. Methods

2.2.1. Ordinary Least Square Regression

Least squares regression is a useful method for analyzing the relationship between attributes, as it assumes a linear relationship between variables.When analyzing a dataset, the presence of noise and outliers is unavoidable. The ordinary least squares method (OLS) provides a solution to this issue. This formula illustrates the concept of minimizing the gap between the actual values and the estimated values. There exists a linear function that minimizes the difference in this formula, and this linear estimation function represents the best-fitting line for the relationship.

$$L = \sum_{i=1}^{n}(y_i - f(x))^2 \quad (1)$$

The sum of the squared discrepancies between the actual target values (y_i) and the outputs predicted by the model ($f(x)$) for each of the n samples is the total loss, or L. The formula inside is summed over the index i from 1 to n, where n is the total number of samples, as shown by the summation symbol (Σ). And the true target or label value for the i-th sample is denoted by y_i. There exists a linear function that minimizes the difference in this formula, and this linear estimation function represents the best-fitting line for the relationship.

Python is utilized in the following analysis steps. In the program, the 1000 lines of data are divided into two parts: the training set (train_set) and the testing set (test_set), comprising 80% and 20% of the data, respectively. The packages Pandas, Statsmodels, and Scikit-learn (Sklearn) are used to build the data frame, create the model, and test the results. In this research, only the training set is used to build the model, while the testing set is employed to evaluate the model's performance by examining the gap between predicted values and actual values.

2.2.2. Random Forest Regression

Random forest regression is composed of multiple decision trees. Each tree makes decisions based on the income data independently. This approach is intuitive and aligns well with human reasoning. After combining these trees, the random forest is created. When the forest encounters new samples, it uses all the individual trees to make decisions and provides a result based on their collective suggestions. The random forest method is resistant to overfitting. It addresses multicollinearity by

DOI: 10.54254/2754-1169/143/2024.GA18965

randomly selecting feature subsets and employing bootstrapping during the training of each tree. As a result, the predictions can be significantly more accurate.

2.3. Evaluation Metrics

Mean square error (MSE), mean absolute percentage error (MAPE), and R-squared (R^2) are useful metrics for assessing the quality of OLS models after prediction. Mean square error (MSE) measures the average squared difference between observed and predicted values. A smaller MSE indicates a better fit of the model to the data. Mean absolute percentage error (MAPE) measures the average absolute difference as a percentage of actual values, providing intuitive insights into the model's prediction errors. R-squared (R^2), known as the coefficient of determination, ranges from 0 to 1, where higher values indicate a greater explanatory power of the independent variables in predicting the dependent variable (y). Additionally, RMSE (Root Mean Square Error), which is the square root of MSE, is used to evaluate the results.

3. Analysis of Results

3.1. Data Visualization

In order to have a comprehensive understanding of the data distribution, a graph is generated to represent the relationship:

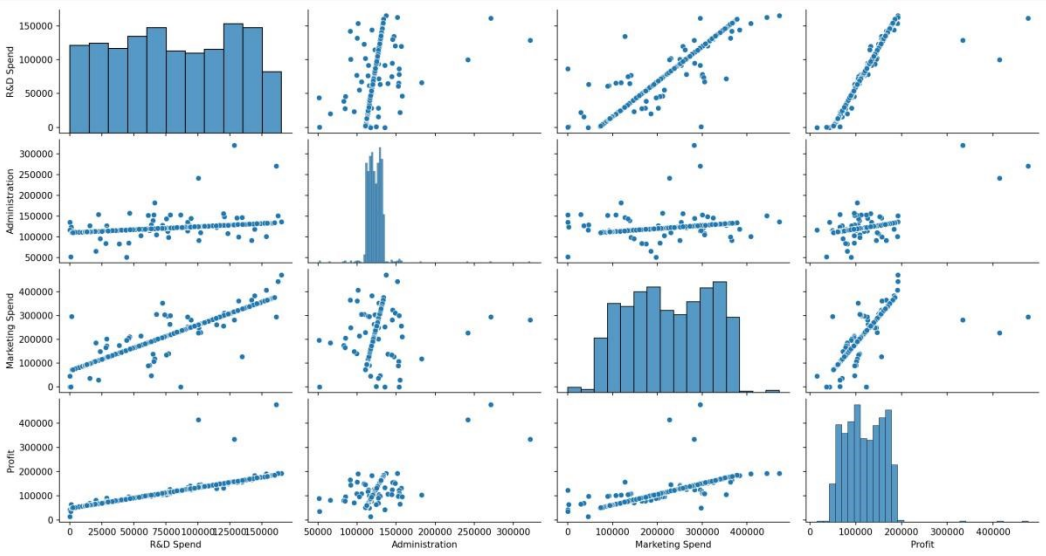

Figure 1: Data distribution of dataset(Photo credit: original)

In Figure 1, the R&D and marketing spend are correlated with profit. Due to the fewer discrete data points and the fact that a straight line in the graph is formed by the majority of the data points, therefore clear correlations and fewer anomalous values are shown. Administration is less correlated with profit and has more outliers in the graph. This aligns with the findings of Susetyo and Abdurohman [10].

The data exhibits strong continuity,which can better reflect and uncover the patterns of data changes, significantly improving the accuracy of model predictions. Additionally, this is beneficial for understanding the values of dependent variables under a wide range of independent variable values, providing decision-makers with more reliable evidence.

DOI: 10.54254/2754-1169/143/2024.GA18965

3.2. Analysis of model results

To analyze the dataset using the ordinary least squares (OLS) method, the estimation of R-squared, which represents the strength of correlation between parameters, is necessary to determine whether a linear relationship exists. All data points are selected to calculate the correlation.

Table 1: R-squared value of profits

RD_profit	Admin_profit	Marketing_profit
0.945245	0.74156	0.91727

From Table 1, it is evident that all three independent variables have high R-squared values when analyzed against profit. A high correlation coefficient indicates a strong linear relationship. Therefore, ordinary least squares regression can be applied to this dataset. The dataset is assumed to follow a linear function.

After training the model, 3 values are generated for testing it's effectiveness. First, 45,177.676 is the root mean square error (RMSE). The model may have considerable prediction mistakes in real-world applications because the RMSE value is comparatively large. As a result, it is necessary to further tune the model in order to improve its predictive capacity. The second number is 6.17% for the Mean Absolute Percentage Error (MAPE). As a relative indicator of prediction accuracy, MAPE shows that the more adaptable the model is to the data, the lower its value. Although there are still some prediction mistakes, the model's overall predict on performance is deemed satisfactory by the current MAPE rating. Lastly, 0.488935 is the coefficient of determination (R-squared). The model's fit appears to be mediocre, as evidenced by the current R-squared value, which shows that the model can only account for roughly 48.9% of the variation in the data, leaving nearly half of the variation in the data unaccounted for. This finding raises the possibility that the model's feature variable or structural selection is flawed.

If the results of Ordinary Least Squares (OLS) analysis are unsatisfactory, it may be due to multicollinearity. Strong correlations between independent variables in a linear regression model can skew estimates and make accurate estimation challenging. Additionally, one regression may be influenced by others, complicating the analysis. The Variance Inflation Factor (VIF) quantifies this issue by measuring the ratio of the variance of the estimated regression coefficients to the variance when the independent variables are not linearly related. It is widely used to assess multicollinearity.

Table 2: VIF value

Attributes	VIF
Const	170.089584
R&D	28.272465
Administration	1.656547
Marketing	25.622279

According to Table 2, the VIF values for "R&D Spend" and "Marketing Spend" are relatively high. The research from Blankley, W. also proved this idea [10]. After removing "Marketing Spend," the VIF values become acceptable, as shown in Table 3. This result indicates the presence of multicollinearity.

DOI: 10.54254/2754-1169/143/2024.GA18965

Table 3: VIF value after removed “Marketing”

Attributes	VIF
Const	119.435767
R&D	1.513384
Administration	1.513384

Multicollinearity influences OLS estimation in several ways. It leads to an increase in the standard error of the estimated regression coefficients and reduces the significance levels of these coefficients. Consequently, this issue may prevent the model from identifying important predictive variables effectively.

Since the OLS method does not provide a satisfactory explanation, this research utilizes random forest regression to mitigate the influence of multicollinearity. Python packages such as Pandas and Scikit-learn are employed in the training process. The independent variables remain the same as those used in the OLS model training. After analyze the model, this model is found that it performs well. First, the Root Mean Square Error (RMSE) is 4715.02171. The RMSE value is relatively low, indicating that the model has high accuracy in practical applications and can effectively capture data trends. Secondly, the Mean Absolute Percentage Error (MAPE) is 0.489231%. This shows that the model has a good prediction. Finally, the coefficient of determination (R-squared) is 0.987378. The current R-squared value is close to 1, meaning that this model can explain approximately 98.7% of the data variability. This result indicates that the constructed model can effectively capture the main features and trends.

3.3. Discussion

Both R&D spending and advertising spending are positively correlated with a company’s profitability [10]. An international comparison conducted in 17 countries, including several EU nations, found a significant correlation between R&D and advertising expenditures. This indicates a strong correlation between R&D spending and marketing spending. The results of the multicollinearity test within the dataset confirmed this finding.

Both the random forest model and the OLS model made predictions based on the independent variables. However, the OLS method is not particularly suitable in this case due to multicollinearity. In comparison to the random forest model, the performance of the OLS method is not satisfactory. Table 4 clearly illustrates the differences between the two methods. The random forest method demonstrates a better fit and greater explanatory power, resulting in more accurate predictions.

Table 4: Compare of two models

Model	RMSE	MAPE	R-squared
OLS	45177.676	6.17%	0.488935
Random Forest	4715.021711	0.489231%	0.987378

4. Conclusion

Throughout the paper, two models are developed to identify an effective method for predicting a company's profit based on its costs. Three parameters are defined that significantly influence the dependent variable. From the research, the following conclusions can be drawn: 1) Both OLS and random forest regression can be used to predict a company's profitability, but they exhibit different explanatory abilities in the model. 2) Due to the correlation between independent variables and the issue of multicollinearity, the predictions made by OLS may be affected. In contrast, the random

DOI: 10.54254/2754-1169/143/2024.GA18965

forest model provides more precise predictions. 3) The mean absolute percentage error is only 0.489231% using the model provided by the random forest method, indicating that this model is effective for predicting profitability based on costs.

Companies can use this model to forecast expected profits at different cost levels, thereby assisting in the formulation of the annual budget. Furthermore, companies can identify which costs have the greatest impact on profits and take action to innovate their cost structures. These two models can help stakeholders form better expectations about the company. Governments can use this model to adjust tax policies by predicting corporate profits or to formulate corresponding regulatory measures. Banks can monitor the financial status of enterprises to ensure debt security. Shareholders can assess the company's profit potential and return on investment to decide whether to increase their investments.

This paper still has areas for improvements. During training two models, the time series does not being considered, which ignored the time dependency of variables. In the future, this parameter will also be added as one significant variable to develop a more comprehensive model.

References

[1] *Domitilla, Magni. (2019). Theory of Profit. 1-264.*

[2] *Dunn, K. D., & Brooks, D. E. (1990). Profit Analysis: Beyond Yield Management. Cornell Hotel and Restaurant Administration Quarterly, 31(3), 80-90.*

[3] *Indrayani, I., Gani, A., Mursidah, M., & Yunina, Y. (2022). The Effect of Sales Production Costs, Total Debt and Working Capital on Net Profit of Manufacturing Companies Pharmaceutical Sub Sector. International Journal of Educational Review, Law And Social Sciences (IJERLAS).*

[4] *Hamdani, P., & Susianto, T. E, (2024). The Effect of Working Capital and Production Costs on Net Profit. Journal of Economic Management and Accounting. (2), 609-618.*

[5] *Markovitch, D. G., Huang, D. L., & Ye P. F.(2020) Marketing Intensity and Firm Performance: Contrasting the Insights Based on Actual Marketing Expenditure and Its SG&A Proxy. Journal of Business Research, (118), 223-239.*

[6] *Mubarok, F., Sultan, Z., Wibowo, M., & Wongsuwatt, S. (2023). Unlocking the Secrets of Profitability: Investigating the Role of Research and Development. Journal of Theory and Applied Management, (16), 356-367.*

[7] *Ravšelj, D., & Aristovnik, A. (2020). The Impact of R&D Expenditures on Corporate Performance: Evidence from Slovenian and World R&D Companies. Sustainability, 12(5), 1943.*

[8] *Jayathilaka, A. (2020). Operating Profit and Net Profit: Measurements of Profitability. OALib, (07), 1-11.*

[9] *Susetyo, D. P., & Abdurohman, M. R. (2024). Exploration of the Relationship Between Operating Costs and Net Profit: a Significance Test with Two Approaches–Revised Edition. Jurnal Ekonomi, Manajemen Dan Akuntansi, 2(2), 303-310.*

[10] *Blankley, W. (2007). Correlations Between Advertising and R&D Expenditures: Dealing with Important Intangibles. South African Journal of Science. 103. 94-98.*