sustainability

MDPI

Article

The Dynamics of the Profit Margin in a Component Maintenance, Repair, and Overhaul (MRO) within the Aviation Industry: An Analytical Approach Using Gradient Boosting, Variable Clustering, and the Gini Index

Nur Şahver Uslu [1,*] and Ali Hakan Büyüklü [2]

[1] Graduate School of Science and Engineering, Yıldız Technical University, Davutpasa Campus, 34220 Istanbul, Turkey

[2] Department of Statistics, Faculty of Arts & Science, Yıldız Technical University, Davutpasa Campus, 34220 Istanbul, Turkey; hbuyuklu@yildiz.edu.tr

[*] Correspondence: tnursahver@gmail.com

Abstract: This study focuses on the dynamics of the profit margin within the aviation MRO industry, using operational data from a small and medium-sized enterprise (SME) MRO company between 2013 and 2021. Especially in SME MROs, profit margin analysis provides an advantage in competing with the large companies that dominate the industry. Therefore, the operational data were prepared for analysis to identify the variables related to the profit margin. This study's data cleaning and transformation processes can serve as a guideline for similarly sized companies. The research aims to address the complex relationships among the factors influencing profit margins in this industry. The objective is to utilise these factors in making strategic decisions to increase the profit margin of an SME MRO company. Applying gradient boosting algorithms as the analytical framework should allow identifying the correct relationships between the profit margin and input variables according to time for the SME MRO company. Another important aspect of this study is to increase the accuracy of the gradient boosting model by utilising the interactive grouping methodology. The variable selection was performed by using the Gini indexes of the variables using interactive grouping as a criterion in selecting the variables to be included in the model. After the data cleaning, transformation, and selection, the input variables for the gradient boosting model were Part Description, Parts Billed Current (part cost), Labour Billed Current (labour cost), Diff Shipping Entry (turnaround time (TAT)), Diff Quote Entry (time to quotation (TTQ)), Manager, Department, and Status. In this study, the profitability model indicates that the SME MRO company should initially focus on part numbers and the departments, secondly on standardisation of and expertise in preferred workshop units, and lastly, on highly qualified and effective technical department leaders and increasing labour. The aviation industry emerges as a sector that requires such analytical studies. It is hoped that the study will serve as a foundational work for SME MRO companies in the aviation industry.

Keywords: aviation; aircraft; aircraft components; maintenance, repair, and overhaul; MRO; data science; gradient boosting; cluster analysis; interactive grouping; data analytics; Gini index; profit

check for updates

Citation: Uslu, N.Ş.; Büyüklü, A.H. The Dynamics of the Profit Margin in a Component Maintenance, Repair, and Overhaul (MRO) within the Aviation Industry: An Analytical Approach Using Gradient Boosting, Variable Clustering, and the Gini Index. *Sustainability* **2024**, *16*, 6470. https://doi.org/10.3390/su16156470

Academic Editor: Yichuan Zhao

Received: 18 June 2024
Revised: 21 July 2024
Accepted: 23 July 2024
Published: 29 July 2024

1. Introduction

Maintenance involves the processes used to rehabilitate and restore the performance of a system or device, ensuring it returns to its expected operating levels after deteriorating over time [1]. The main objective of aircraft maintenance is to ensure the safety and availability of the aircraft for flight operations by minimising operational disruptions due to system failures. This goal is economically achieved best through the optimal use of maintenance resources [2]. In aviation, this is specifically known as the maintenance, repair, and overhaul (MRO) industry. The MRO industry encompasses a spectrum of entities, ranging from industry giants such as Airbus and Boeing to relatively smaller MRO firms that provide maintenance and repair services to ensure aircraft operate safely and efficiently.

The MRO environment is highly sensitive to macro factors such as the global economy, geopolitical events, technical issues, environmental factors, and pandemics. Although driven by post-pandemic challenges and geopolitical tensions, significant headwinds remain for airlines, MROs, and original equipment manufacturers (OEMs) [3]. The MRO industry is expected to get back on track by 2024, with the global aviation MRO market making considerable progress in recovering from the COVID-19 pandemic. According to Oliver Wyman, the size of the commercial aircraft MRO market reached USD 101 billion in 2023, amounting to 98% of the 2019 pre-COVID peak in real dollars. MRO spending is expected to set a new record of USD 104 billion in 2024. In the coming years, the MRO industry is predicted to expand at an annual rate of 1.8% through 2034, reaching USD 124 billion [4].

The aircraft MRO market is segmented by MRO type into airframes, engines, components, and line maintenance. The market consists of four main MRO categories: engine MRO at 46% component MRO at 20% airframe MRO at 20% and line MRO at 14% [5]. The safety and availability of aircraft depend on comprehensive maintenance activities across these segments [6]. Thousands of MRO companies operate worldwide, including manufacturers, authorised MROs, and independent MROs, all adhering to defined procedures monitored by aviation authorities such as the Federal Aviation Administration (FAA) and the European Union Aviation Safety Agency (EASA) [7]. To ensure aircraft safety by economically performing maintenance, various methods in the MRO industry are used, including reactive, proactive, and aggressive maintenance strategies. Industry trends increasingly favour preventive and predictive maintenance, both part of the proactive approach.

Preventive maintenance involves performing failure-finding tasks at regular intervals to detect deterioration and extend the system's remaining useful life. On the other hand, predictive maintenance uses predictive algorithms and real-time condition monitoring to estimate failure times and detect system degradation [8]. Data analytics play a critical role in predictive and preventive maintenance, guiding MRO companies in becoming competitive and differentiating themselves in the market [9].

Analytical methods in MRO are generally categorised into descriptive, predictive, and prescriptive analytics [10]. Various techniques are applied in aviation, including Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and k-means clustering for flight data [9], machine learning (ML) for maintenance data [11], system dynamics for airline operations [6], and Bayesian networks for predictive maintenance [12].

In the MRO market, OEMs hold an advantage due to their control over repair procedures, data, and materials. However, demand and capacity constraints also sustain small and medium-sized enterprise (SME) MRO companies [13]. These smaller firms, however, often lack the data science resources and expertise of larger MRO players and OEMs [11]. This study focuses on how robust data analysis affects the business models of an SME MRO, rather than the general importance of data for maintenance methods previously examined in the literature.

Oliver Wyman identifies the elements that disrupt MRO companies' business models the most, including cost management, labour shortages, and economic factors [4]. These pressures necessitate that SME MRO companies emphasise value propositions to differentiate themselves in this competitive industry.

Medium-sized MRO companies see much data science activity at the larger MRO players and OEMs. The amount of data and applications are growing fast. However, the smaller companies lack the resources and expertise [11]. This study investigates the dynamics of profit margins within the aviation MRO industry, utilising operational data from an SME MRO company collected between 2013 and 2021. Analysing the profit margins of the SME MRO company offers a competitive edge over the larger companies that dominate the industry. By focusing on these dynamics, the study aims to provide strategic insights to help the SME MRO company enhance its profitability and sustainability in a highly competitive market.

The study was conducted using a structured dataset comprising 19,621 observations from the SME MRO company for the training and validation datasets. The analysis involved several steps: data understanding, data cleaning, new data derivation, data transformation, variable grouping, variable selection, variable clustering, and variable reduction [14]. Gradient boosting algorithms were employed, followed by validation, to analyse profit margins. The study utilises 19,621 operational data entries after all data preparation steps, split into training and validation datasets, to discern the relationship between profit margin and various operational variables, optimising processes and prioritising decisions to maximise profitability. The interactive grouping methodology was employed to select significant variables and enhance the model's accuracy and efficacy. After variable grouping, the Gini index was applied to determine which variables should be included in the gradient boosting model. Furthermore, variable clustering was conducted to analyse which variables in the gradient boosting model clustered together, facilitating the interpretation of the input variables for strategic decision-making. The misclassification rate and RASE statistics were utilised to interpret and compare two different gradient boosting models from a validation perspective.

The primary contribution of this study is that the analysis is based on real data from the SME MRO company. By applying interactive grouping methods to the data, the study facilitates the use of complex datasets in decision-making processes, thereby enhancing the accuracy of the gradient boosting algorithm. The dynamics of the profit margin can potentially be utilised by the SME MRO company when making strategic decisions to increase its profitability.

2. Materials and Methods

In order to summarise the steps of the analyses conducted in Section 3, Figure 1 was prepared.

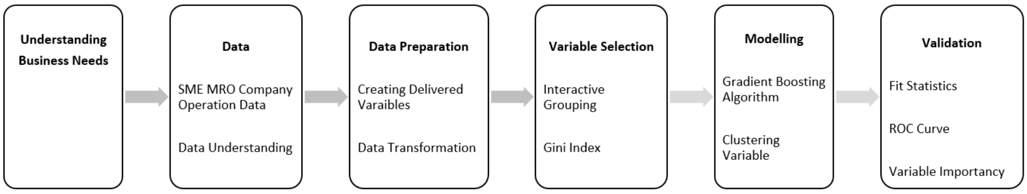

Figure 1. Flow chart of the study.

Figure 1 illustrates the steps taken from the raw data to obtaining the results.

Under the "Understanding Business Needs" step, the process begins with discussions about and understanding the business needs with business stakeholders and obtaining the data from the SME MRO company.

Under the "Data" step, this process continues with understanding the nature of the data obtained for the business needs and performing data cleaning.

In the "Data Preparation" step, variables that could impact the profit margin, but are not present in the raw data were derived and transformed.

In the "Variable Selection" step, data were grouped to have a monotonic relationship with the target variable, explaining the profit margin in the best way. Variable selection was performed using the Gini indexes of these grouped variables.

The data were divided into training and validation datasets, and a gradient boosting model was run to compare the performance of models built with standardised variables and grouped variables. The accuracy of the models was also compared using fit statistics on the training and validation datasets.

The study utilised variable clustering to examine which variables related to profit margin were grouped distinctively. Considering these clusters, strategic recommendations for the SME MRO company are shared in Section 4.

The details of the steps performed are provided in Sections 2 and 3.

2.1. Data and Data Preparation

In this study, the analysis focused on identifying the operational variables that influence the profit margin of an MRO company, London, UK. Due to the challenges faced in simulating this type of operational data, it was decided to work with real-world data. This dataset consists of structured data including nominal, categorical, and numerical variables derived from an MRO company, based on real-time operational data entries [14].

Within the scope of the Understanding Business Needs and Data step in Figure 1, initially, there were 83 variables in the raw operational data. Each variable was analysed regarding the data meaning, data availability, missing data ratio, and data reliability. Following the initial data-understanding phase, the decision was made to proceed with 30 variables out of the initial 83 for the rest of the data cleaning process.

In the research conducted using operational raw data comprising 30 variables belonging to the SME MRO company, 21,708 observations covering the period from 2013 to 2022 were included. Following the data cleaning and transformation procedures mentioned below consistent with industry standards, the analysis was continued with 19,621 observations and 30 variables. The data cleaning stages applied are included under the Data step in Figure 1 and are listed below.

Remove Billed Profits less than or equal to 0 (remaining 19,780). Upon examination of the data meeting these criteria, it was determined that these repairs had not occurred.

Remove entries containing the case study MRO company name as the customer name (remaining 19,766). The records are maintained for internal transactions within the company; therefore, these observations do not hold any significance in explaining the profit margin.

Remove entries where the Last Shipment Date equals zero (remaining 19,755). It is expected that these numerical data are not less than 0, and upon examining the transaction details, erroneous records have been identified.

Remove entries where the Need Date equals "missing" (remaining 19,744). Errors have been identified in observations that meet this criterion, and they have been excluded from the analysis.

Remove entries where the Current Labour Billed exceeds 40,000 (remaining 19,622). The labour costs cannot be at or above this amount, as like the company's business scope and due to the identified error, it was decided to exclude the model from consideration

Replace labour hours greater than 11,000 with "missing" (remaining 19,621). The labour hours cannot be at or above this amount, and due to the identified error, it was decided to exclude the model from consideration.

Table 1 contains the data (long list of variables) that are likely to enter the gradient boosting model based on the result of the initial data analysis mentioned above. From the remaining 30 variables after the initial data cleaning, data containing the same information and unique numbers were extracted at this stage. Variables that may have meaning for the model are included in the long list of variables. The derived variables in Table 1 are the data created for the model, but not presented in the raw operational dataset.

From the remaining 30 variables after the initial data cleaning, data containing the same information and unique numbers were extracted at this stage. Variables that may be meaningful for the model are included in the long list of variables in Table 1. Additionally, the derived variables in Table 1 were created for the model, but are not included in the raw operational dataset.

The decisions regarding the inclusion of variables from Table 1 in the model, within the scope of the Data step in Figure 1, are outlined below. As a result of these decisions, the final variables selected from the long list of variables in Table 1 have been transitioned to Table 2, which contains the short list of variables included in the model.

The profit margin, referred to as the Billed Margin Calculated, was calculated manually and then standardised in Table 2, despite the availability of the billed margin original variable in Table 1. It was observed that, although few, some variables yielded incorrect results in the automatic calculation of the billed margin original variable. Therefore, the billed margin original variable was not used in the model development. The Billed Profit and Total Billed Current variables from Table 1 were used to derive the target variable, standardised Billed Margin Calculated in Table 2.

Table 1. Long list of variables.

Variable	Definition
Billed margin original	Profit rate from the raw data
Billed Profit	Profit amount
Billing Date Last	Billing Date
Customer Name	Customer name
Entry Date Last	Entry Date of the part (date)
Entry Date Num	Entry Date of the part (dummy)
Invoice Number	Unique invoice number
Labour Hours	Labour hours
Last Ship Date	Shipping date of the part (date)
Last Ship Date Num	Shipping date of the part (Derived to Numeric)
Part Number	Part number
Quote Date last	Quote Date of the part (date)
Quote Date Num	Quote Date of the part (Derived to Numeric)
Sales Invoice	Sales invoice
Total Billed Current	Total bill amount
WO Number	Work order number
WO Type	Work order type
Work Type Description	Work type description
Labour Billed Current	The selling price of labour
Parts Billed Current	The selling price of parts
Department	The department responsible for the part
Part Description	Part description
Status	Status of part

Table 2. Short list of variables.

Variable	Definition
Billed Margin Calculated (Target)	Profit margin (derived)
Diff Quote Entry	The number of days between the arrival of the part and the quotation (derived)
Diff Shipping Entry	The number of days between the arrival of the part and the shipping (derived)
Labour Billed Current	The selling price of labour
Parts Billed Current	The selling price of parts
Department	The department responsible for the part
Part Description	Part description
Status	Status of part

The Last Ship Date and the Billing Date Last in Table 1 contained the same information; hence, the Billing Date could not be included in the shortlist in Table 2. As will be explained in the next paragraphs, the Last Ship Date was used to derive other variables, Diff Quote Entry and Diff Shipping Entry in Table 2, and was not directly included in the model.

Customer Name, Invoice Number, Part Number, and Work Order (WO) Number from Table 1 were not used in the model. They were only used for joining and similar data preparation and control steps.

To differentiate between pre- and post-pandemic transactions, a dummy variable named Entry Date Num was created from the Entry Date Last from Table 1. The Entry Date Num variable is not included in Table 2, which lists the variables used in the model, because its Gini index calculated against the target variable was below the cutoff value of 15.

It was decided to include the Labour Billed Current variable from Table 2 in the model as it has fewer missing values instead of the Labour Hours variable from Table 1.

The Quote Date Last variable from Table 1 was not used directly in the model; however, it was used to create the standardised Diff Quote Entry variable included in Table 2.

Since there is only one category left in the WO Type variable in Table 1 after the data cleaning process mentioned above (remaining observations 19,621), WO Type was not included among the variables to be included in the model.

The Labour Billed Current, Parts Billed Current, Department, Part Description, and Status variables from Table 1 were included in the model and placed in Table 2, the short list of variables, because they had a Gini index above 15 after interactive grouping.

The Work Type Description variable from Table 1 did not make it to the short list of variables and is not included in Table 2 because its Gini index against the target variable was below 15.

Within the scope of the "Data Preparation" step in Figure 1, following discussions with industry professionals, derived variables including Billed Margin Calculated, Diff Quote Entry, and Diff Shipping Entry were included in the final model in Table 1 containing variables available in the dataset. The derived variable Billed Margin Calculated represents the profit margin achieved for the aircraft part. Diff Quote Entry is a derived variable that calculates the period between the Entry Date Last (the date the part enters the system) and the Quote Date Last (the date on which the quote for the part is generated), expressed in numeric format. Diff Shipping Entry is another derived variable, also expressed numerically, that calculates the time between the Entry Date Last (the date the part was recorded in the system) and the Last Ship Date (the date the part was shipped).

As a result of the decisions mentioned above regarding the data in Table 1, the list of variables to be included in the model was obtained and is presented in Table 2. Within the scope of the "Variable Selection" step in Figure 1, based on the Gini index, the theory mentioned in Section 2.2 was used as the criterion for variable selection. Table 2 contains the short list of variables that we decided to enter into the model after variable selection.

The mean and median values of the numerical variables included in the model, which are listed in Table 2, are provided in Table 3.

Table 3. Descriptive statistics of numerical variables.

Variable	Mean	Median
Billed_Margin_Calculated	55.71	55.25
Diff_Quote_Entry	19.38	8.00
Diff_Shipping_Entry	58.64	31.00
Labour_Billed_Current	383.39	246.00
Parts_Billed_Current	2433.43	204.49

In Table 3, the mean and median values of Diff Quote Entry, Diff Shipping Entry, Labour Billed Current, and Parts Billed Current suggest that the majority of the data are clustered around low values, while a few extremely high values create a right-skewed distribution.

The descriptive statistics for the categorical variables included in the model and listed in Table 2 are provided in Table 4. The aim is to manage the outlier problem by applying Interactive Grouping to the data. In Section 3.3, the event rates of the data after grouping are detailed comprehensively.

Table 4 illustrates the number of levels, mode value, and mode percentage value for each categorical variable.

Table 4. Descriptive statistics for categorical variables.

Variable	Number of Levels	Mode Percentage	Mode
Department	8	48.25	Electromechanical
Manager	17	11.75	Manager1
Part_Description	1670	19.80	Pinion
Status	5	51.65	Ship

Before running the model, the numerical variables were standardised, and subsequently, all variables were grouped. Variable selection involved the application of the Gini index [15]. Within the scope of the "Modelling" step in Figure 1, the gradient boosting model was run with training and validation datasets in a ratio of 70% and 30% respectively. The "Validation" step was tested as outlined in Figure 1.

Python and the SAS software were used for all data cleaning, preprocessing, and model applications.

2.2. Interactive Grouping

The interactive grouping utilises models of non-linear functions of multiple modes of continuous distributions. It computes initial bins through quantiles and allows interactive splitting and combining of these bins. This is used to create bins, buckets, or classes for all input variables, encompassing both categorical and interval variables. Bins are established to decrease the number of unique levels and potentially enhance the predictive capability of each input. The selection of robust characteristics is based on the Gini statistic, and the chosen characteristics are grouped according to business considerations. This analysis aids in structuring the data to reflect trends in profit margin ranking rather than modelling idiosyncrasies, which could result in overfitting. Grouping attributes in some instances also enhances predictive capability.

In the modelling process, interactive grouping plays a crucial role in minimising the effects of outliers, as defined in Table 3. This method effectively reduces the adverse impacts of outliers and significantly enhances model performance by grouping data. The Gini index is widely recognised as a key selection criterion for assessing the uniformity of event rates across different categories for manually grouped variables in interactive grouping and was employed in the process of selecting variables [16]. Its selection is attributed to the Gini index's remarkable ability to measure distributional equality, reduce bias, and provide ease of comparison and use [17].

The Gini index is expressed as a percentage. The Gini coefficient ranges from 0 to 1, so the Gini index ranges from 0 to 100. The Gini index measures the uniformity or non-uniformity of a distribution. The lower the Gini index, the more uniformly distributed a variable is. In the context of the profit margin model applying gradient boosting, the Gini index is used to assess how evenly the event rates are distributed across the attributes of a characteristic. Utilising the Gini index for computing feature importance also aided in removing inefficient variables from the dataset [18].

First, sort the data in descending order based on the proportion of events associated with each attribute. Suppose a characteristic has m attributes. Then, the sorted attributes are represented as groups 1, 2, . . ., m. Each group corresponds to an attribute, with group 1 having the highest proportion of events.

Then, for each of these sorted groups, calculate the number of events (n_i^{event}) and non-events ($n_i^{\text{non-event}}$) in group i. Then, compute the Gini index.

The Gini index calculation is performed as in Equation (1):

$$\text{Gini Index} = \left(1 - \frac{2\sum_{i=2}^{m}\left(n_i^{\text{event}} \times \left(\sum_{j=1}^{i-1} n_j^{\text{non-event}}\right)\right) + \sum_{k=1}^{m}\left(n_k^{\text{event}} \times n_k^{\text{non-event}}\right)}{N^{\text{event}} \times N^{\text{non-event}}}\right) \times 100 \tag{1}$$

Here, event N^{event} and non-event $N^{\text{non-event}}$ are the total numbers of events and non-events in the data, respectively.

2.3. Cluster Analysis

Cluster analysis consists of grouping objects (or variables) that are similar to each other. Here, similarity is the concept that should be defined before starting the procedure. Usually, a distance measure is used so that similarity becomes spatial proximity [19].

Cluster analysis is an important tool for "unsupervised" learning—the problem of finding groups in data without the help of a response variable. A major challenge in cluster analysis is to estimate the optimal number of "clusters" [20].

Variable clustering can be performed to reduce the number of variables. The rule dictates selecting the variable with the minimum $1 - R^2_{ratio}$ as the cluster representative. This $1 - R^2_{ratio}$ is given by Equation (2).

$$1 - R^2_{ratio} = \frac{1 - R^2_{own}}{1 - R^2_{nearest}} \tag{2}$$

Intuitively, cluster the representative to be as closely correlated to its cluster $1 - R^2_{own}$ to 1 and as uncorrelated to the nearest cluster $1 - R^2_{nearest}$ to 0. Therefore, the optimal representative of a cluster is a variable where $1 - R^2_{ratio}$ tends to zero [21].

Even though variable clustering is used as a variable reduction methodology, here it was applied to understand the similarity of variables well to utilise some strategic decisions for the SME MRO company. The method visually presents how input variables were grouped. As a result, suggestions are made in Sections 4 and 5 based on both these clusters and individual variables. These clusters were instrumental in informing the strategic decisions intended for the company.

2.4. Gradient Boosting Algorithm

Gradient boosting of trees produces competitive, highly robust, interpretable procedures for regression and classification, especially appropriate for mining less than clean data [22].

Gradient boosting is a sophisticated boosting technique that iteratively resamples the analysis dataset multiple times, generating results that aggregate into a weighted average of the resampled datasets. Tree boosting involves the construction of a sequence of decision trees that collectively form a single predictive model. Each tree in the sequence is fit to the residuals of the predictions produced by the preceding trees, with these residuals defined based on the derivative of a loss function.

Each iteration of the algorithm uses the data to grow a decision tree, and the accuracy of the resulting tree is then evaluated. Subsequent samples are adjusted to address previously identified inaccuracies. Since each successive sample is weighted according to the classification accuracy of previous models, this approach is sometimes referred to as stochastic gradient boosting. Boosting can be applied to binary, nominal, and interval targets.

The Stochastic Gradient Boosting Loss optimisation aims to reduce the losses generated during the training process [23]. The residual for interval targets is defined using a squared error loss function. To compute the residual for an interval target using squared error loss, one subtracts the predicted value from the target value. For binary targets, the residual is defined using the negative binomial log-likelihood loss function, also known as logistic loss.

Similar to decision trees, boosting does not make assumptions about the data distribution. For interval inputs, the model relies solely on the ranks of the values. For interval targets, the impact of extreme value theory depends on the chosen loss function. Boosting tends to be less prone to overfitting compared to a single decision tree. When a decision tree fits the data reasonably well, typically, boosting enhances the model fit.

The total loss over the entire dataset is then defined as the sum of the individual losses of overall samples in Equation (3) [23]:

$$L_{\text{total}} = \sum_{i} L(f(x_i, \theta), y_i) \tag{3}$$

where (x_i) is the input and (y_i) is the target value. It uses a decision tree with parameters θ to predict the output $\hat{y}_i$ for input (x_i). The output can be any function of the parameters and the input, represented as $\hat{y}_i$ = f (xi, θ).

The optimisation algorithm focuses on estimating the values of the parameters θ that minimise this total loss. This is typically performed using gradient descent, which updates the parameters θ in the opposite direction of the gradient of the total loss concerning the parameters in Equation (4) [23]:

$$\theta_{\text{new}} = \theta_{\text{old}} - \alpha \nabla_{\theta} L_{\text{total}} \tag{4}$$

Here, α is the learning rate, which controls the size of the parameter update, and $\alpha\nabla_{\theta}$L total is the gradient of the total loss concerning the parameters θ [23].

3. Results

The study utilised operational data from the SME MRO company, collected from 2013 to 2022. After discussions with sector professionals, derived variables were created to examine the profit margin. Interactive grouping was applied for variable selection. Among the grouped variables, those with a Gini index greater than 15 were chosen for the model. The final dataset for model development included the variables listed in Table 2 and contained 19,621 observations following the data cleaning process. Numerical variables were standardised. The methodology involved using interactive grouping for variable selection and gradient boosting for modelling. Validation analysis was conducted by dividing the data into training and validation datasets with a 70:30 ratio, respectively. Model fit statistics were compared between the gradient boosting models with standardised variables and grouped variables. The results indicated that gradient boosting performed better with the grouped variables. The following results were obtained using Figure 1.

3.1. Understanding Business Needs and Data

Initially, 83 variables in the raw operational data were analysed for the data meaning, availability, missing data ratio, and reliability. After this initial data and business understanding phase, 30 variables were selected for the subsequent data cleaning process. Data cleaning procedures, as detailed in Section 2.1 were applied to the raw dataset of 21,708 observations. After these procedures, the dataset was reduced to 19,621 observations.

3.2. Data Preparation

As was discussed with industry professionals, some variables in the raw data alone might not be sufficient to explain the profit margin. Therefore, it was decided to create some derived variables as detailed in Section 2.1. Additionally, numerical variables were standardised for analysis.

3.3. Variable Selection—Interactive Grouping

Interactive grouping was used to comprehensively analyse the relationship between the variables in the model and the profit margin. The continuous profit margin variable was transformed into a binary variable in the interactive grouping. Considering the company's average profit margin of 0.56, a cutoff of the profit margin higher than 0.5 was used when creating a new binary target variable for the interactive grouping.

All interval input variables were treated as such. Groups were generated by ranking quantities so that each group contained approximately the same frequency. The initial number of groups was set to four.

After initial groups were obtained in the interactive grouping, final groups were created with manual adjustments. Since the groups obtained after manual correction better explain the relationship with the profit margin, these groups were also used in the gradient boosting model.

Sections 3.3.1–3.3.8 contain the Gini indexes of the groups and variables obtained from the result of the interactive grouping for each variable included in the gradient boosting model. The Gini index was used to measure the relationship of each variable with the profit margin variable, and variables with a Gini index below 15 were not used in the gradient boosting model.

3.3.1. Parts Description

Part Description refers to the name and definition of the component, based on its location, shape, or function, which is approved for installation on type-certified aircraft. Part Description helps us understand the unique part numbers that are coded with numbers and letters by the OEM. Initially, four groups were created for the categorical variable Part Description. The Part Description variable provides information about the name of the aircraft part and covers 1670 different part numbers that the MRO company can repair and overhaul for airline and aviation companies. The MRO capabilities for these parts are extensive. To express the relationship between the binary profit margin variable (where a profit margin greater than 0.5 is labelled as 1, while that which less than or equal to 0.5 is labelled as 0) and the categorical variable Part Description in a linearly monotonic decreasing manner, parts with similar event rates among the 1670 parts were grouped. This grouping was performed manually. Table 5 shows the adjusted groups.

Table 5. Classification of Part Description into groups.

Group	Event Count	Non-Event Count	Group Event Rate	Gini Coefficient	Gini Ordering
1	1563	90	0.946	65.246	1
2	442	167	0.726	65.246	1
3	319	182	0.637	65.246	1
4	621	558	0.527	65.246	1
5	495	776	0.389	65.246	1
6	671	1371	0.329	65.246	1
7	558	1719	0.245	65.246	1
8	342	3861	0.081	65.246	1

In Table 5, for the parts included in group 1, the number of events representing parts with maintenance or repair completed and a profit margin value above 0.50 is 1563, while the number of parts with a profit margin value of 0.50 or below is only 90. This implies that parts in group 1, with a group event rate of 0.946, are the most profitable parts.

In group 2, there are 442 parts with profit margin values above 0.50 and 167 parts with profit margin values of 0.50 and below, resulting in a group event rate of 0.726.

In group 3, there are 319 parts with profit margin values above 0.50 and 182 parts with profit margin values of 0.50 and below, yielding a group event rate of 0.637.

In group 4, there are 621 parts with profit margin values above 0.50 and 558 parts with profit margin values of 0.50 and below, with a group event rate of 0.527.

In group 5, there are 495 parts with profit margin values above 0.50 and 776 parts with profit margin values of 0.50 and below, resulting in a group event rate of 0.389.

In group 6, there are 671 parts with profit margin values above 0.50 and 1371 parts with profit margin values of 0.50 and below, leading to a group event rate of 0.329.

In group 7, there are 558 parts with profit margin values above 0.50 and 1719 parts with profit margin values of 0.50 and below, yielding a group event rate of 0.245.

In group 8, there are 342 parts with profit margin values above 0.50 and 3861 parts with profit margin values of 0.50 and below, resulting in a group event rate of 0.081.

During the initial phase of interactive grouping, separate groups were created for all parts, and the Gini index for these groups was calculated as 69.169. Manual grouping was performed to reduce the number of groups from 1500 to 8, which led to a slight decrease in the Gini index to 65.246 in Table 5. A high Gini index indicates a strong relationship between the profit margin and the Part Description variable. While this manual grouping slightly reduced the Gini index, the interactive grouped Part Description variable can now be monotonically associated with the event rate, i.e., whether the profit margin is higher or lower than 0.50. This approach provides a better explanation of the target variable, namely the profit margin. Therefore, it can be stated that there is a strong relationship between the grouped part types and the profit margin.

3.3.2. Parts Billed Current

The Parts Billed Current variable, which contains information about part cost, is a numerical variable. This variable represents material costs consumed and used during the repair process. These direct costs are recorded in real-time to calculate actual costs at any time to monitor and track the repair process. The relationship between profit margin and part cost was analysed by categorising standardised part costs. Intervals were determined for significant groups of the numerical variable for the binary target variable, created as 0 and 1 for profit margin values above 0.50 or 0.50 and below. Table 6 shows the adjusted groups.

Table 6. Classification of Parts Billed Current into groups.

Group	Event Count	Non-Event Count	Group Event Rate	Gini Coefficient	Gini Ordering
1	3847	2986	0.563	49.050	2
2	1001	2491	0.287	49.050	2
3	163	3247	0.048	49.050	2
4	0	0	0	49.050	2

Initially, the Part Billed Current variable was divided into five groups as follows:

Parts Billed Current < -0.36, $-0.36 \leq$ Parts Billed Current < -0.33, $-0.33 \leq$ Parts Billed Current < -0.08, $-0.08 \leq$ Parts Billed Current, and missing values. The Gini index for this variable was calculated as 50.222 initially. Although the Gini index was high, there was no monotonic relationship with the binary profit margin variable defined as the target variable. Therefore, some modifications were made manually to the initial groups.

In Table 6, the group where missing values were classified was labelled as group 4. Since there were no missing values in this variable, observations in this group were not analysed. Groups 2 and 3 were merged into group 1 due to the low number of events in group 3. Within the combined group 1, there were observations with Part Billed Current < −0.33, out of which 3847 were an event (i.e., profit margin above 0.50) and 2986 were a non-event (i.e., profit margin at or below 0.50). The group event rate for group 1 was 0.563. It can be inferred that observations in group 1, where part costs are low, have good profit margins.

Observations in group 2, with Parts Billed Current between −0.33 and −0.08, consisted of 1001 event count observations (i.e., profit margin above 0.50) and 2491 non-event count observations (i.e., profit margin at or below 0.50). The group event rate for Group 2 was 0.287. It can be inferred that observations in Group 2, where part costs are relatively higher compared to group 1, have lower profit margins.

Observations in group 3, with Part Billed Current greater than −0.08, consisted of 163 event count observations (i.e., profit margin above 0.50) and 3247 non-event count observations (i.e., profit margin at or below 0.50). The group event rate for group 3 was

0.048. It can be inferred that observations in group 3, where part costs are relatively higher compared to group 1 and group 2, have the lowest profit margins.

The Gini index for these groups is calculated as 49.050 in Table 6. Although slightly lower compared to the initial Gini index from the initial groups, this Gini index indicates that the interactive grouped Parts Billed Current variable can be monotonically associated with the event rate, i.e., whether the profit margin is higher or lower than 0.50. This approach provides a better explanation of the target variable, that material cost increases reduce profit margins.

3.3.3. Labour Billed Current

The Labour Billed Current variable represents direct labour costs, calculated based on direct labour hours booked during the rectification of the part. This variable is recorded in real time to track total labour costs during the repair process. The Labour Billed Current variable reflects the company's labour efficiency and direct labour performance, which are within its control; on the other hand, Parts Billed Current is an external cost driver influenced by market fluctuations in material prices depending on location and purchase timing. Table 7 shows the adjusted groups for Labour Billed Current.

Table 7. Classification of Labour Billed Current into groups.

Group	Event Count	Non-Event Count	Group Event Rate	Gini Coefficient	Gini Ordering
1	3757	3133	0.545	42.664	3
2	858	2570	0.25	42.664	3
3	396	3021	0.116	42.664	3
4	0	0	0	42.664	3

Initially, the Labour Billed Current variable was divided into five groups as follows:

Labour Billed Current < −0.82, −0.82 ≤ Labour Billed Current < −0.29, −0.29 ≤ Labour Billed Current < 0.34, 0.34 ≤ Labour Billed Current, and missing values. The Gini index for this variable was calculated as 42.664. Although the Gini index was high, there was no monotonic relationship with the binary profit margin variable defined as the target variable. The adjustments made manually to the initial groups are given below:

In Table 7, the group where missing values were classified was designated as group 4. Since there were no missing values in this variable, observations in this group were not analysed. Groups 2 and 3 were merged under group 1 due to the low number of events in group 2. Within the combined group 1's scope, there were observations with Labour Billed Current < −0.29, out of which 3757 were events (i.e., profit margin above 0.50) and 3133 were non-events (i.e., profit margin at or below 0.50). The group event rate for group 1 was 0.545. It can be inferred that observations in group 1, where labour costs are low, have the best profit margins.

Observations falling into group 2, with Labour Billed Current between −0.29 and 0.34, consisted of 858 event count observations (i.e., profit margin above 0.50) and 2570 non-event count observations (i.e., profit margin at or below 0.50). The group event rate for group 2 was 0.250. It can be inferred that observations in group 2, where labour costs are relatively higher compared to group 1, have lower profit margins.

Observations falling into group 3, with Labour Billed Current greater than 0.34, consisted of 396 event count observations (i.e., profit margin above 0.50) and 3021 non-event count observations (i.e., profit margin at or below 0.50). The group event rate for group 3, 0.116, suggests that observations in group 3 demonstrate a notable pattern in event occurrence, where labour costs are relatively higher compared to group 1 and group 2, which have the lowest profit margins.

The Gini index remains at 42.664 in Table 7 for the interactive grouped Labour Billed Current variable, which can be monotonically associated with the event rate (i.e., whether the profit margin is higher or lower than 0.50). It provides an opportunity to explain the target variable more effectively, for which the increase in labour costs reduces the profit margin.

3.3.4. Manager

The Manager variable stands for the leader or mechanic of the specific technical department in the MRO company. The company has sub-departments, which are experts based on part types such as hydraulic, mechanic, pneumatic, electrical, and avionic. This variable also monitors and measures the department's performance or specific part contribution to the company, as well as the expertise of the technical staff. The Manager variable provides information about the managers' identity related to the operation and encompasses 58 managers. To express the relationship between the Profit Margin binary variable (where a profit margin higher than 0.5 is labelled as 1 and <=0 is labelled as 0) and the categorical variable of Manager in a linearly monotonic decreasing manner, variables with similar event rates among the 58 managers were grouped. This grouping was performed manually. Observations with a missing Manager variable were grouped separately. Table 8 shows the adjusted groups.

Table 8. Classification of Manager into groups.

Group	Event Count	Non-Event Count	Group Event Rate	Gini Coefficient	Gini Ordering
1	750	195	0.794	33.178	4
2	1530	1944	0.440	33.178	4
3	804	1233	0.395	33.178	4
4	1325	2669	0.332	33.178	4
5	415	1356	0.234	33.178	4
6	187	1327	0.124	33.178	4

In Table 8, for the managers included in group 1, the number of event counts have a profit margin value above 0.50, which is 750 while the number of managers with a profit margin value of 0.50 or below is only 195. This implies that managers in group 1, with a group event rate of 0.794, are led to the highest profit margin operations.

In group 2 and 1530 managers with profit margin values above 0.50 and 1944 managers with profit margin values of 0.50 and below. The group event rate for group 2 is 0.44.

In group 3, the observations with missing values in the Manager variable are classified. There are 804 managers with profit margin values above 0.50 and 1233 managers with profit margin values of 0.50 and below. The group event rate for group 3 is 0.395.

In group 4 are 1325 managers with profit margin values above 0.50 and 2669 managers with profit margin values of 0.50 and below. The group event rate for group 4 is 0.332.

In group 5 are 415 managers with profit margin values above 0.50 and 1356 managers with profit margin values of 0.50 and below. The group event rate for group 5 is 0.234.

In group 6, there are 187 managers with profit margin values above 0.50 and 1327 managers with profit margin values of 0.50 and below. The group event rate for group 6 is 0.124.

The calculated Gini index considering these groups is notably high at 33.178 in Table 8. A high Gini index can be interpreted as a sign of a strong relationship between the Profit Margin and the Manager variables. Additionally, during the initial phase of Interactive Grouping, separate groups were created for all managers, and the Gini index for these groups was calculated as 34.677. Manual grouping was performed to transition from 59 groups to 6 groups, and this manual grouping slightly decreased the Gini index to 33.178. A marginally lower Gini index, but grouped Manager variable that can be monotonically associated with the event rate, i.e., whether the profit margin is higher than 0.50, will provide a more comprehensive understanding of the Profit Margin. By interpreting this Gini index, the Manager variable has a relationship with profit margin.

3.3.5. Diff Shipping Entry

This variable represents the difference between the Shipment Date and Entry Date, known as turnaround time (TAT) in the industry. TAT measures the duration a part or component is in process at the MRO company site, serving as a critical indicator of operational efficiency. Table 9 shows the adjusted groups.

Table 9. Classification of Diff Shipping Entry into groups.

Group	Event Count	Non-Event Count	Group Event Rate	Gini Coefficient	Gini Ordering
1	1842	1396	0.569	30.367	5
2	1418	2097	0.403	30.367	5
3	953	2587	0.269	30.367	5
4	798	2644	0.232	30.367	5
5	0	0	0	30.367	5

Initially, the Diff Shipping Entry variable was divided into five groups as follows: Diff Shipping Entry < -0.39, $-0.39 \leq$ Diff Shipping Entry < -0.25, $-0.25 \leq$ Diff Shipping Entry < -0.01, $-0.01 \leq$ Diff Shipping Entry, and missing values. The Gini index for this variable was calculated as 30.367. Although the Gini index was high, there was no monotonic relationship with the binary profit margin variable defined as the target variable. Therefore, some adjustments were made manually to the initial groups.

In Table 9, the group where missing values were classified is designated as group 5. Since there were no missing values in this variable, observations in this group were not analysed. Within group 1, there were observations with Diff Shipping Entry less than −0.39, out of which 1842 were events (i.e., profit margin above 0.50) and 1396 were non-events (i.e., profit margin at or below 0.50). The group event rate for group 1 was 0.569. It can be included that observations in group 1, where Diff Shipping Entry values are the lowest, have the best profit margins.

Observations falling into group 2, with Diff Shipping Entry between −0.39 and −0.25, consisted of 1418 event count observations (i.e., profit margin above 0.50) and 2097 non-event count observations (i.e., profit margin at or below 0.50). The group event rate for group 2 was 0.403. It can be concluded that observations in group 2, where Diff Shipping Entry values are relatively higher compared to group 1, have lower profit margins.

Observations falling into group 3, with Diff Shipping Entry between −0.25 and −0.01, consisted of 953 event count observations (i.e., profit margin above 0.50) and 2587 non-event count observations (i.e., profit margin at or below 0.50). The group event rate for group 3 was 0.269. It can be inferred that observations in group 3, where Diff Shipping Entry values are relatively higher compared to group 1 and group 2, have lower profit margins.

Observations falling into group 4, with Diff Shipping Entry greater than −0.01, consisted of 798 event count observations (i.e., profit margin above 0.50) and 2644 non-event count observations (i.e., profit margin at or below 0.50). The group event rate for group 4 was 0.232. It can be inferred that observations in group 4, where Diff Shipping Entry values are relatively higher compared to group 1, group 2, and group 3, have the lowest profit margins.

After the manual grouping adjustment, the Gini index remained the same as the initial interactive grouping at 30.367 in Table 9. The interactive grouped Diff Shipping Entry variable that can be monotonically associated with the event rate—whether the profit margin is higher or lower than 0.50—provides an opportunity to explain the profit margin more effectively. As Diff Shipping Entry increases, the profit margin decreases. The longer TAT decreases more in profitability, for which this data grouping is aligned with the negative effect of TAT on profit.

3.3.6. Department

The Department variable provides information about the department responsible for completing the operation and encompasses different departments. Four groups were created for the categorical variable of Department initially. To express the relationship between the binary variable of profit margin (where a profit margin higher than 0.5 is labelled as 1 and $\leq$0.5 is labelled as 0) and Department in a linearly monotonic decreasing manner, variables with similar event rates among the 11 categories, including the missing category, were grouped. This grouping was performed manually. Table 10 shows the adjusted groups.

Table 10. Classification of Department into groups.

Group	Event Count	Non-Event Count	Group Event Rate	Gini Coefficient	Gini Ordering
1	2452	2594	0.486	23.390	6
2	1171	2089	0.359	23.390	6
3	1388	4041	0.256	23.390	6

In Table 10, for the departments included in group 1, the number of event counts (with a profit margin value above 0.50) is 2452, while the number of non-event counts (with a profit margin value of 0.50 or below) is 2594. This implies that departments in group 1, with a group event rate of 0.486, are the most profitable among the other groups.

In group 2, there are 1171 event count observations with profit margin values above 0.50 and 2089 non-event count observations with profit margin values of 0.50 and below. The group event rate for group 2 is 0.359. Group 2 contains departments that generate lower profit margins compared to group 1.

In group 3, there are 1388 event count observations with profit margin values above 0.50 and 4,041 non-event count observations with profit margin values of 0.50 and below. The group event rate for group 3 is 0.256. Group 3 contains departments with the lowest profit margins compared to group 1 and group 2.

In the initial phase of interactive grouping, separate groups were created for all departments, and the Gini index for these groups was calculated as 25.080. Manual grouping was performed to transition from 11 groups to 3 groups, which reduced the Gini index slightly to 23.390, as shown in Table 10. Despite the marginal decrease in the Gini index, the interactively grouped Department variable, which can be monotonically associated with the event rate (i.e., whether the profit margin is higher or lower than 0.50), offers a more comprehensive explanation for the profit margin. Interpreting this Gini index reveals an association between the Department variable and the profit margin. It can be inferred that certain department groups have higher profit margins compared to others.

3.3.7. Diff Quote Entry

This variable represents the time elapsed between the date a quotation is sent to the customer and the work order for the part or component is generated. This metric measures the time taken to provide a quote to the customer, an important performance indicator for both the customer and the company, called time to quote (TTQ). Table 11 shows the adjusted groups.

Initially, the Diff Quote Entry variable was divided into five groups as follows: Diff Quote Entry < -0.25, $-0.25 \leq$ Diff Quote Entry < -0.18, $-0.18 \leq$ Diff Quote Entry < -0.07, $-0.07 \leq$ Diff Quote Entry, and missing values. The Gini index for this variable was calculated as 23.193. The Gini index was high, and there was a monotonic decrease with the binary profit margin variable defined as the target variable. Some modifications were made manually for the missing group in the initial groups.

Table 11. Classification of Diff Quote Entry into groups.

Group	Event Count	Non-Event Count	Group Event Rate	Gini Coefficient	Gini Ordering
1	1533	1301	0.541	23.193	7
2	1769	2964	0.374	23.193	7
3	879	2023	0.303	23.193	7
4	830	2436	0.254	23.193	7
5	0	0	0.000	23.193	7

In Table 11, the group where missing values were classified was designated as group 5. Since there were no missing values in this variable, observations in this group were not analysed. Within group 1's scope, there were observations with Diff Quote Entry < −0.25, out of which 1533 were events (i.e., profit margin above 0.50) and 1301 were non-events (i.e., profit margin at or below 0.50). The group event rate for group 1 was 0.541. It can be inferred that observations in group 1, where Diff Quote Entry is low, have good profit margins.

Observations falling into group 2, with Diff Quote Entry between −0.25 and −0.18, consisted of 1769 event count observations (i.e., profit margin above 0.50) and 2964 non-event count observations (i.e., profit margin at or below 0.50). The group event rate for group 2 was 0.374. It can be inferred that observations in group 2, where Diff Quote Entry is relatively higher compared to group 1, have lower profit margins.

Observations falling into group 3, with Diff Quote Entry between −0.18 and −0.07, consisted of 879 event count observations (i.e., profit margin above 0.50) and 2023 non-event count observations (i.e., profit margin at or below 0.50). The group event rate for group 3 was 0.303. It can be inferred that observations in group 3, where Diff Quote Entry is relatively higher compared to group 1 and group 2, have lower profit margins.

Observations falling into group 4, with Diff Quote Entry greater than −0.07, consisted of 830 event count observations (i.e., profit margin above 0.50) and 2436 non-event count observations (i.e., profit margin at or below 0.50). The group event rate for group 4 was 0.254. It can be inferred that observations in group 4, where Diff Quote Entry is the highest compared to group 1, group 2, and group 3, have the lowest profit margins.

The Gini index remains unchanged at 23.193 in Table 11. The interactive grouping of the Diff Quote Entry variable initially shows a monotonically associated trend with the event rate (i.e., whether the profit margin is higher than or equal to 0.50 or lower). This provides an opportunity for a clearer explanation of the target variable, which is the profit margin. There have been no changes to the groupings. The profit margin decreases as the Diff Quote Entry increases, indicating the negative impact of quotation time on profitability.

3.3.8. Status

This variable indicates the status of a customer's order at a specific time while the part is under the repair process at the MRO company from entry to shipment date. Table 12 shows the adjusted groups.

Table 12. Classification of Status into groups.

Group	Event Count	Non-Event Count	Group Event Rate	Gini Coefficient	Gini Ordering
1	1345	1467	0.478	19.876	8
2	2999	4762	0.386	19.876	8
3	667	2495	0.211	19.876	8
4	0	0	0.000	19.876	8

In Table 12, group 1 consists of parts with a status indicating that they have been shipped. The number of event counts, where the profit margin is above 0.50, is 1345, while the number of non-event counts, where the profit margin is 0.50 or below, is 1467. This suggests that parts in the shipped status included in group 1, with a group event rate of 0.478, are the most profitable among all status groups.

In group 2, which includes parts with a status of work completed, there are 2999 event counts (profit margin above 0.50) and 4762 non-event counts (profit margin at or below 0.50). The group event rate for group 2 is 0.386.

In group 3, which includes the rest of the work completed and shipped parts, there are 667 event counts (profit margin above 0.50) and 2495 non-event counts (profit margin at or below 0.50). The group event rate for group 3 is 0.211.

The calculated Gini index considering these groups is 19.876. An acceptable Gini index can be interpreted, which indicates a relationship between the Profit Margin and the Status variables. Additionally, during the initial phase of interactive grouping, separate groups were created for all statuses, and the Gini index for these groups was calculated as 20.865. Manual grouping was performed to transition from 11 groups to 4 groups, and this manual grouping slightly decreases the Gini index to 19.876 in Table 12.

The Gini index is slightly lower, but with an interactively grouped Status variable demonstrating a monotonically increasing association with the event rate (i.e., whether the profit margin is higher or lower than 0.50), there is a better opportunity to explain the profit margin. Interpreting this Gini index indicates a clear relationship between the Status variable and the profit margin. Consequently, the profit margin tends to increase as the completion level of customer orders rises.

3.3.9. Summary of Interactive Grouping Results

The interactive grouping results explained above for individual variables are collectively included in Figure 2 for all variables.

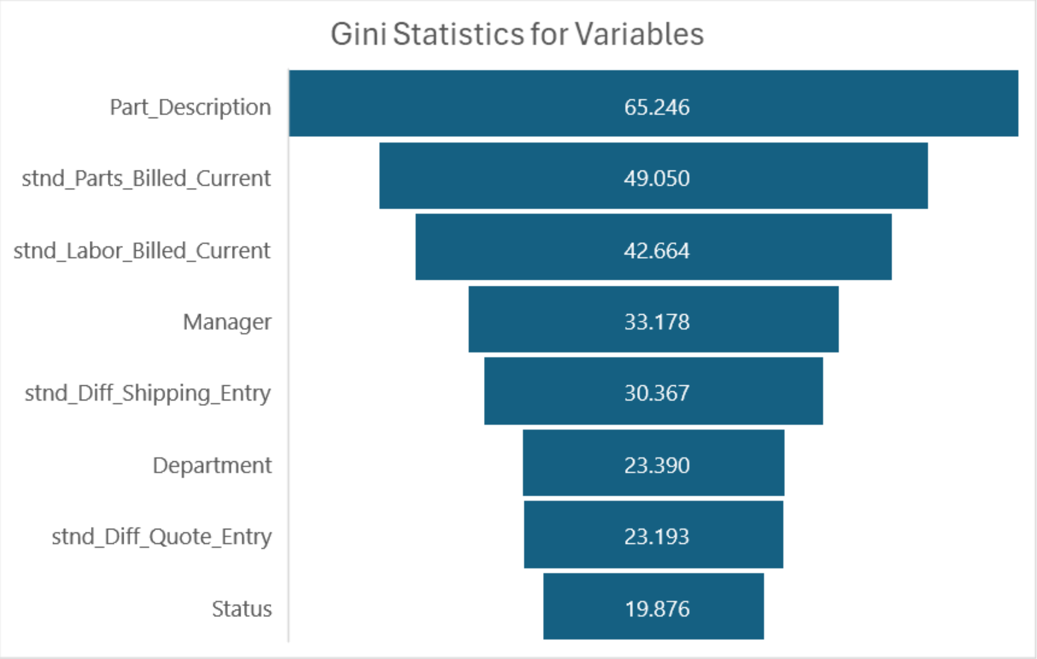

Figure 2. Gini index for grouped variables.

Figure 2 shows the Gini indexes explaining the relationships of the grouped variables included in the model with the target variable. As seen in Figure 2, the variable that explains profit margin best is Part Description, while the variable that explains it least is Status. Before interactive grouping, there was no observed correlation between standardised input variables and target variables.

Following the manual adjustments made to the variable groups, a monotonic trend was observed, indicating that the profit margin decreased as the variable group number increased. The targeted monotonic trend between the input and target variable is illustrated in Figure 3.

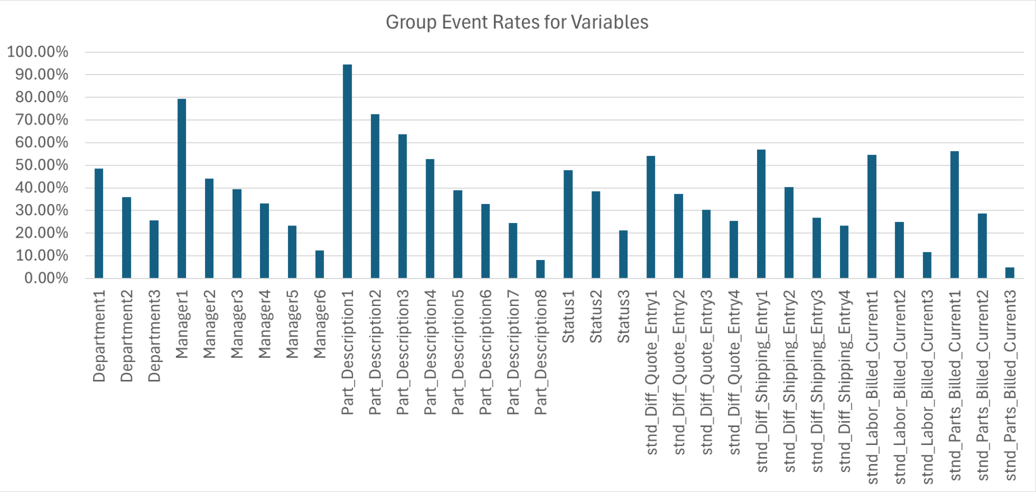

Figure 3. Comparison of event rates across different groups for variables.

In Figure 3, the numbers at the end of the variables on the x-axis indicate the groups. As the group numbers progress from 1 to n, it can be said that the rate, that is a profit margin above 0.5, decreases.

After performing manual adjustments on variable groups, a monotonic trend was observed, indicating that the profit margin decreases as the number of variable groups increases. The targeted monotonic trend between the input and target variable is illustrated in Figure 3.

To show how successful the Part Description variable is in explaining the target variable by manual grouping, the relationship of the 1670 parts in the Part Description variable with the event rate before manual adjustment is shown in Figure 4.

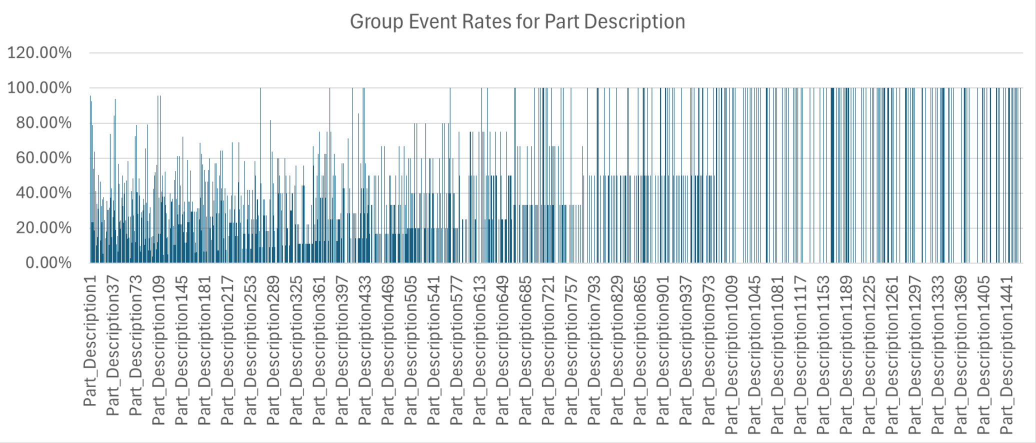

Figure 4. Comparison of group event rates for Part Description.

Figure 4 indicates that establishing a relationship between the 1670 parts and the profit margin was challenging before manual grouping.

Furthermore, during the data preparation phase, to differentiate between pre- and post-pandemic transactions, a dummy variable named Entry Date Num was created. The relationship of this variable with the target variable was limited, and therefore, it was not included in the model. There was no difference in the profit margin variable before and after the pandemic.

3.4. Variable Clustering

The model was initially run by standardised input variables and a continuous target variable to show the contribution of applying interactive grouping transformation to the variables. Then, it was run with the newly grouped variables and binary target variables. Before executing this algorithm, variable clustering was performed to measure the distances between the input variables. The results are presented in Figure 5.

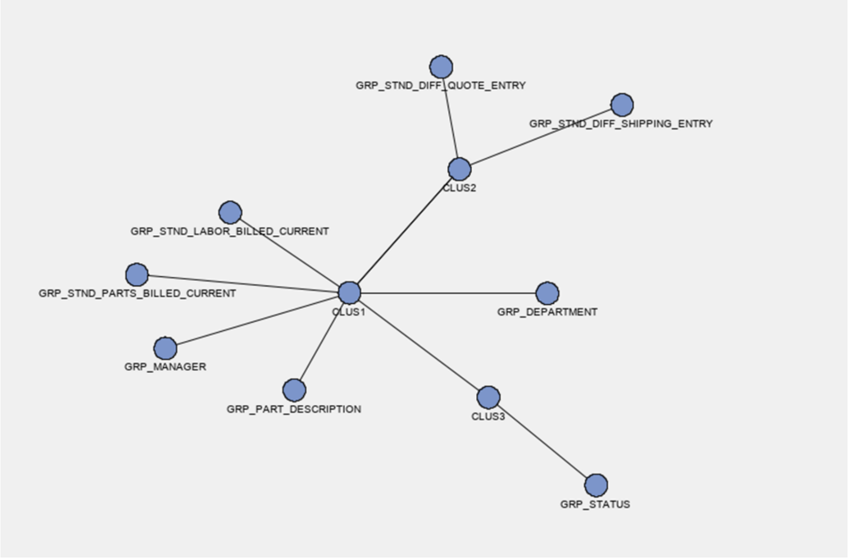
GRP_STND_DIFF_QUOTE_ENTRY
GRP_STND_DIFF_SHIPPING_ENTRY
CLUS2
GRP_STND_LABOR_BILLED_CURRENT
GRP_STND_PARTS_BILLED_CURRENT
CLUS1
GRP_DEPARTMENT
GRP_MANAGER
GRP_PART_DESCRIPTION
CLUS3
GRP_STATUS

Figure 5. Variable clustering diagram.

In Figure 5, squares represent the individual terms and circles represent the cluster centres. The names of the variables that were grouped show obvious similarity relationships. Since the variables with similar factual attributes were clustered together [21], this can be interpreted and analysed visually in Figure 5. Variable clustering provides groups of variables where variables in a group are similar to other variables in the same group and as dissimilar as possible to variables in another group. In this study, the variable clustering method was not used as a variable reduction technique. Instead, it was used to make the data more understandable by organising the variables into clusters. The variables within these clusters were used to provide insights into the strategic decisions shared in the Discussion and Conclusion sections.

When the diagram is examined, it is seen that Diff Quote Entry and Diff Shipping Entry are in Clus2. This is a cluster that explains the concepts of TAT and TTQ, which stands for the operational process [24]. Bidding time and TAT describe similar information in Clus2. It can be concluded that TAT and TTQ differ from other variables. The Parts Billed Current, Manager, Part Description, Labour Billed Current, and Department variables are grouped as Clus1. One may conclude that Clus1 is a cluster containing information related to the part. The Status variable is located in Clus3 alone. The status of part information is disclosed in a separate cluster. The dendrogram of the variables is shown in Figure 6.

Figure 6 visualises the points grouped and the similarity of group members and shows that the variables of TAT and TTQ are distinguished from the other variables.

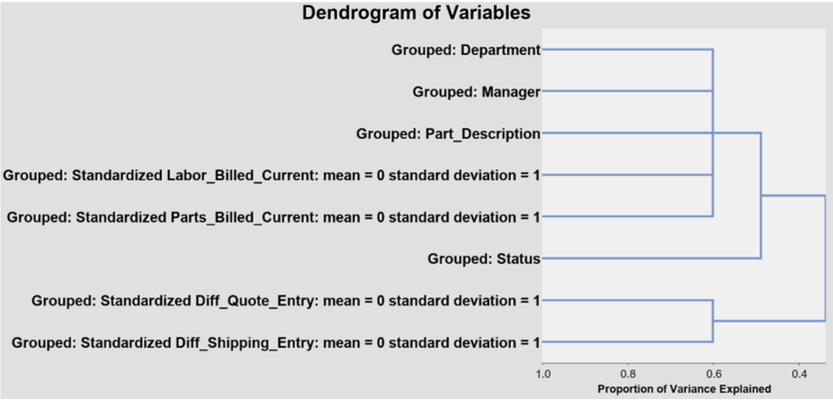

Figure 6. Dendrogram of variables.

3.5. Modelling and Validation—Gradient Boosting

In this study, the steps from data cleaning to validation are discussed to illustrate how SME MRO companies can use analytical methods to inform their profit margin-enhancing decisions. Two separate gradient boosting algorithms were applied, one for grouped variables and another for standardised variables, and their results were compared. The choice of gradient boosting was based on its demonstrated high performance in classification problems [22,25]. Consequently, the study did not include comparisons with other ML techniques.

During the course of this research, hyperparameter tuning was conducted manually using a trial-and-error approach [26]. Initially, a range of values for key hyperparameters was selected. Each combination of these hyperparameters was iteratively tested by training the model on the training dataset and evaluating its performance on the validation dataset. Adjustments were made to the hyperparameters based on the performance outcomes. This iterative process was repeated until an optimal set of hyperparameters was identified, providing the best balance between training and validation performance metrics. The tuned hyperparameters for Boost79 and Boost86 are discussed in Sections 3.5.1 and 3.5.2, respectively. Optimisation methods could also have been used in the selection of the hyperparameters.

Furthermore, the dataset is limited to a specific SME MRO company and spans from 2013 to 2021. This temporal and contextual limitation may affect the generalizability of the results to other contexts or time periods.

Within the scope of the Modelling step in Figure 1, two different gradient boosting algorithms were run, and their results are compared in this section. The Boost79 model built with standardised variables has an RASE value of 0.637 in Section 3.5.1 for the validation dataset, which is a very high error value. In the Boost86 model developed with grouped variables, the 0.164 misclassification rate and 0.821 Gini value in Section 3.5.2 show that a classification model could be established with this data transformation with a lower error rate. A scatter plot was drawn to analyse the compatibility of variable importance values in the models for the training and validation datasets. It can be inferred that the scatter plot for the Boost79 model, Section 3.5.1, shows less fit in the training and validation datasets than the scatter plot for Boost86 model, Section 3.5.2. It has been determined that transforming continuous and categorical data into meaningful groups explains the profit margin well in the Boost86 model compared to the Boost79 model. Interactive grouping was used for this purpose, and a model with higher accuracy was obtained.

The confusion matrices for the Boost86 model, which demonstrated better prediction performance, are presented in Tables 13 and 14 for the training and validation datasets, respectively.

Table 13. Confusion matrix for the training dataset.

	Actual Positive (1)	Actual Negative (0)	Total
Predicted Positive (1)	3825	834	4659
Predictive Negative (0)	1202	7872	9074
Total	5027	8706	13,733

The confusion matrix for the model's performance for the training dataset is shown in Table 13 and provides a breakdown of actual positive and negative cases versus predicted positive and negative cases. Here is a concise summary of the matrix:

Actual positive (1): 3,825 instances correctly predicted as positive.
Actual negative (0): 834 instances incorrectly predicted as positive.
Actual positive (1): 1202 instances incorrectly predicted as negative.
Actual negative (0): 7872 instances correctly predicted as negative.

Table 14. Confusion matrix for the validation datasets.

	Actual Positive (1)	Actual Negative (0)	Total
Predicted Positive (1)	1589	397	1986
Predictive Negative (0)	567	3335	3902
Total	2156	3732	5888

The confusion matrix for the model's performance for the validation dataset is shown in Table 14 and provides a breakdown of actual positive and negative cases versus predicted positive and negative cases. Here is a concise summary of the matrix:

Actual positive (1): 1589 instances correctly predicted as positive.
Actual negative (0): 397 instances incorrectly predicted as positive.
Actual positive (1): 567 instances incorrectly predicted as negative.
Actual negative (0): 3335 instances correctly predicted as negative.

Tables 15 and 16 below present key performance metrics for the model on the training dataset. These metrics include precision, recall, and F1 score, which are crucial for evaluating the effectiveness of a classification model.

Precision measures the accuracy of positive predictions. A high precision indicates a low false positive rate. Recall measures the ability of the model to identify all relevant instances. A high recall indicates a low false-negative rate. The F1 score is the harmonic mean of precision and recall, providing a balance between the two. This metric is useful when a balance between precision and recall is desired.

Table 15. Performance metrics for the model on the training dataset.

Metric	Value
Precision	0.821
Recall	0.761
F1 Score	0.790

In Table 15, a precision value of 0.821 indicates that approximately 82% of the instances classified as positive are indeed positive. This high precision suggests the model makes few false positive errors. A recall value of 0.761 implies that the model successfully identifies 76% of all actual positive instances. While this is a decent score, it indicates that there is still room for improvement in detecting positive instances. An F1 score of 0.790, which balances precision and recall, indicates an approximate value of 0.79. This score suggests that the model maintains a good balance between precision and recall, making it a reliable classifier for the given task.

Table 16. Performance metrics for the model on the validation dataset.

Metric	Value
Precision	0.800
Recall	0.737
F1_Score	0.767

In Table 16, a precision value of indicates that approximately 80% of the instances classified as positive are indeed positive. This high precision suggests that the model makes few false positive errors. A recall value of 0.737 implies that the model successfully identifies 73.7% of all actual positive instances. While this is a decent score, it indicates that there is still room for improvement in detecting positive instances. The F1 score, which balances precision and recall, is approximately 0.767. This score suggests that the model maintains a good balance between precision and recall, making it a reliable classifier for the given task.

The presented performance metrics indicate that the model performs well, with high precision and a reasonable balance between precision and recall. The consistency of these metrics between the training and validation datasets further validates the model's robustness and generalizability. However, there is potential for further improvement, particularly in enhancing recall to ensure more positive instances are correctly identified.

The outputs of the Boost79 and Boost86 models will be examined in subsequent Sections 3.5.1 and 3.5.2, respectively.

3.5.1. Gradient Boosting for Standardised Variables versus Continuous Target Variable

The gradient boosting model was run multiple times using different hyperparameter tuning strategies. Parameter values for models other than the best model are not provided. To construct the Boost79 model gradient boosting model, a training proportion of 70% was utilised. The square loss function was exclusively employed as the splitting rule. The base learner was designed with a maximum of three branches and a depth of five. The number of iterations and the shrinkage parameter were set to 40 and 0.1, respectively.

Table 17 presents the fit statistics for the model results when the profit margin variable is used as a numerical variable, utilising ungrouped standardised variables.

Table 17. Fit statistics for the training dataset and validation dataset: continuous target variable.

Model (Continuous Target Variable)	Train RASE	Validation RASE	Train ASE	Validation ASE
Boost79	0.593	0.637	0.405	0.351
Boost83	0.609	0.640	0.410	0.372
Boost84	0.621	0.645	0.416	0.386
Boost82	0.747	0.776	0.603	0.558

Table 17 shows that the Average-Squared Error (ASE) and Root-Average-Squared Error (RASE) values are very high and indicate that the model does not fit the training data well. The Boost79 model built with standardised variables has an RASE value of 0.637.

Table 18 presents the importance of the training and validation datasets for the Boost79 model.

Table 18. Variable importance.

Variable Name	Training Importance	Validation Importance
Parts Billed Current	1.00	1.00
Labour Billed Current	0.681	0.603
Diff Quote Entry	0.335	0.294
Diff Shipping Entry	0.311	0.252
Part Description	0.278	0.297
Status	0.277	0.272
Manager	0.149	0.155
Department	0.108	0.095

When looking at the order of importance of the variables in Table 18 obtained for standardised variables before grouping, the order is Parts Billed Current, Labour Billed Current, Diff Quote Entry, Diff Shipping Entry, Part Description, Status, Manager, and Department.

Figure 7 presents the scatter plot of variable importance for the training and validation datasets. Variable importance uses the reduction in the sum of squares from splitting a node, summing over all nodes in the split-based approach to calculate the importance [22].

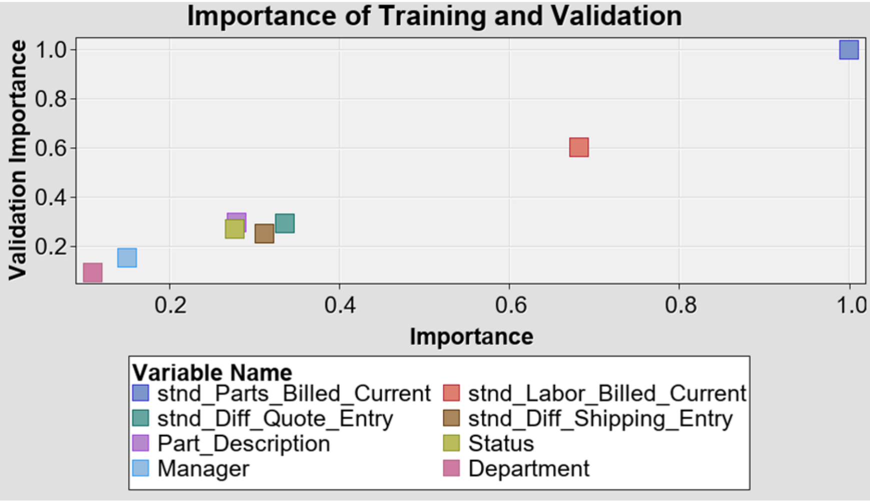

Figure 7. Scatter plot for importance of Boost79.

In Figure 7, the variable importance values of the Boost79 model are not similar in the training and validation datasets compared to the Boost86 model, which will be shown in the following subsection.

3.5.2. Gradient Boosting for Grouped Variables versus Binary Target Variable

Under Section 3.1, all variables were grouped, and those with a Gini index greater than 15 were selected for gradient boosting modelling. For variable grouping, the target variable, profit margin, was also converted from a numerical variable to a binary variable. The models were constructed in this sub-section using the transformed variables.

In each model, different hyperparameter tuning strategies were applied, and the model was run multiple times to achieve the best model. To construct the Boost86 model, a training proportion of 70% was utilised. The square loss function was exclusively employed as the splitting rule. The base learner was designed with a maximum of three branches and a depth of five. The number of iterations and the shrinkage parameter were set to 40 and 0.1, respectively. Table 19 presents a subset of results from the gradient boosting models applied to interactively grouped variables derived from the binary-converted profit margin variable.

Table 19. Fit statistics for the training dataset and validation dataset: binary target variable.

Model (Binary Target Variable)	Training MCR	Val MCR	Train ROC	Val ROC	Train Gini	Val Gini
Boost86	0.148	0.164	0.925	0.911	0.849	0.821
Boost63	0.154	0.167	0.919	0.909	0.839	0.818
Boost64	0.156	0.168	0.916	0.908	0.833	0.815
Boost65	0.183	0.186	0.893	0.889	0.785	0.779

Table 19 shows that the misclassification rate, ROC index, and Gini value yield similar results for training and validation datasets for different models. By an examination of Table 19, the validation misclassification rate was considered the primary criterion for selecting the best model. The misclassification rate represents the percentage of incorrectly predicted outcomes. In the Boost86 model, which was developed using grouped variables, the misclassification rate is 0.164. This transformation of the target variable and input data using interactive grouping has positively impacted the model's accuracy.

When examining the ROC curves for the training and validation datasets, it is observed that the model referred to as the Boost86 model made the best predictions for both datasets.

Figure 8 presents the ROC Curves based on the predictions of different models.

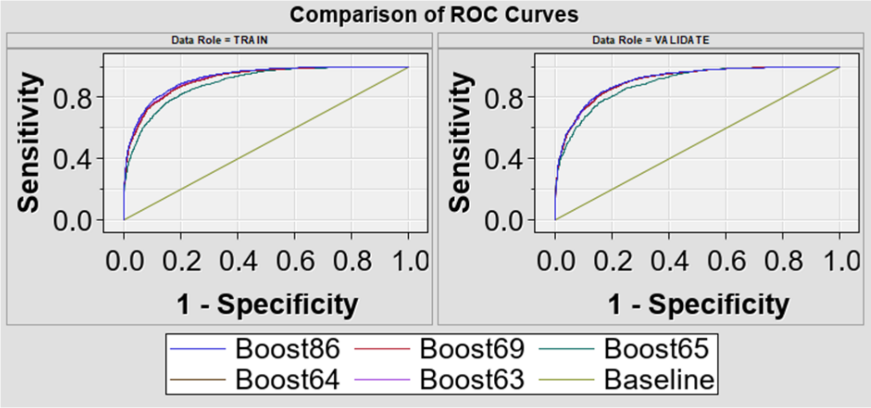

Figure 8. ROC Curves for different gradient boosting models.

In Figure 8, the blue line in the graph represents the Boost86 model. It can be seen that the area under the ROC Curve for both the training and validation datasets is greater than that of the other models. It is observed that the model referred to as the Boost86 model makes the best predictions for both datasets among the other models.

Figure 9 presents the misclassification rate versus the number of iterations for the Boost86 model.

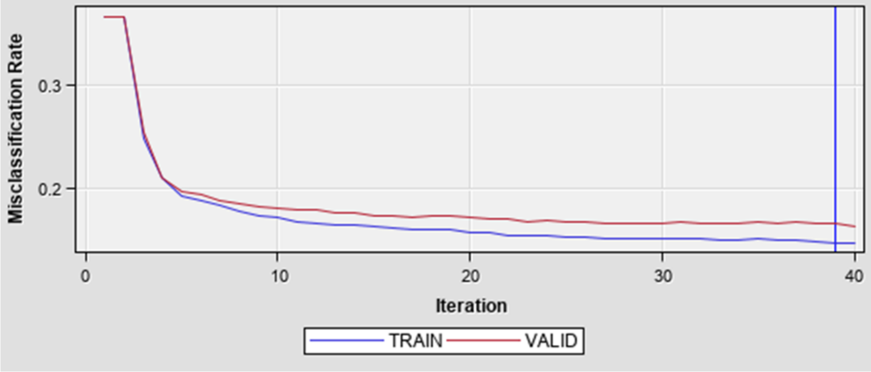

Figure 9. Misclassification rate vs. iteration number.

In Figure 9, a lower misclassification rate means better model accuracy. In addition, when the figure is examined, it is observed that, when the number of iterations approaches 40, the difference between the training and validation datasets becomes stable.

Figure 10 presents the misclassification rates for the target categories (1: the profit rate is higher than 0.5, 0: the profit rate is lower than 0) across different gradient boosting models.

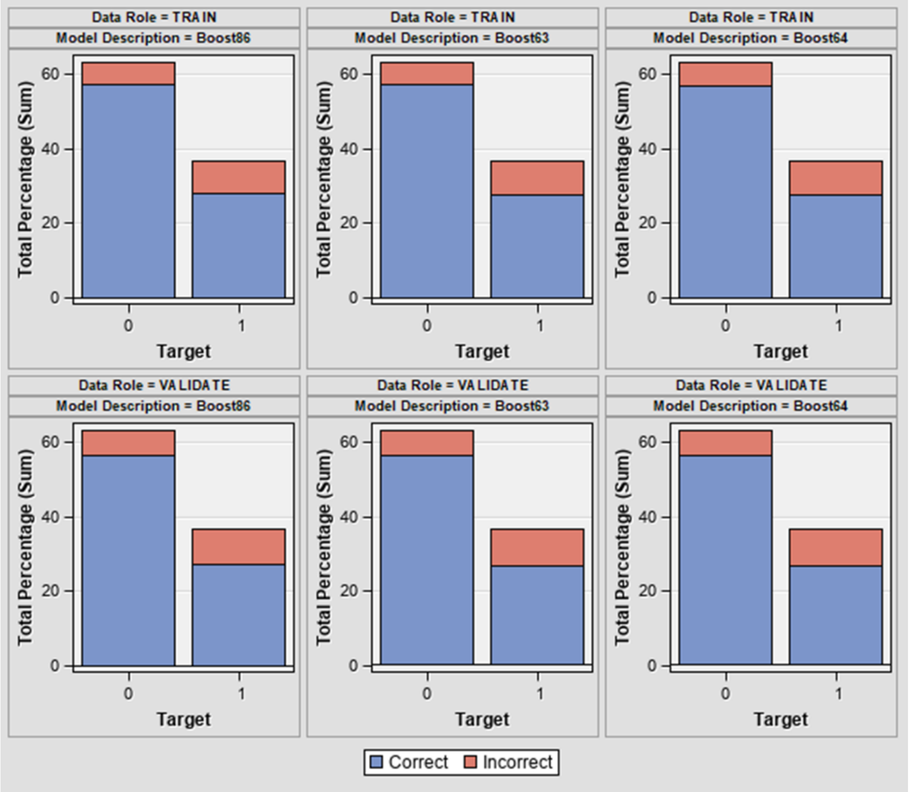

Figure 10. Misclassification rates across different gradient boosting models.

In Figure 10, the top row presents the misclassification rates for the training dataset, while the bottom row shows the rates for the validation dataset, analysed based on the binary target variable. It is a graphical representation of the MCR values in Table 19. Although the error rates are close compared to the other models, Boost63 and Boost64, the Boost86 model demonstrates fewer errors in predicting a profit margin below 0.5 for the training (6.07%) and validation (6.74%) datasets. Furthermore, the Boost86 model demonstrates fewer errors in predicting a profit margin above 0.5 for the training (8.75%) and validation (9.63%) datasets.

Table 20 presents the importance of training and validation datasets for the Boost86 model.

Table 20. Variable importance.

Variable Name	Train Importance	Validation Importance
GRP Part Description	1.00	1.00
GRP Parts Billed Current	0.542	0.655
GRP Labour Billed Current	0.493	0.499
GRP Manager	0.402	0.377
GRP Status	0.335	0.309
GRP Diff Shipping Entry	0.219	0.140
GRP Diff Quote Entry	0.197	0.107
GRP Department	0.192	0.122

In Table 20, the variable importance values of the Boost86 model are very close to each other in the training and validation datasets. This is one of the positive criteria for model accuracy. The order of importance of the grouped variables entering the model is Part Description, Parts Billed Current, Labour Billed Current, Manager, Status, Diff Shipping Entry, Diff Quote Entry, and Department.

Figure 11 presents the scatter plot of variable importance for the training and validation datasets.

Figure 11 illustrates the scatter plot of variable importance for the training and validation datasets.

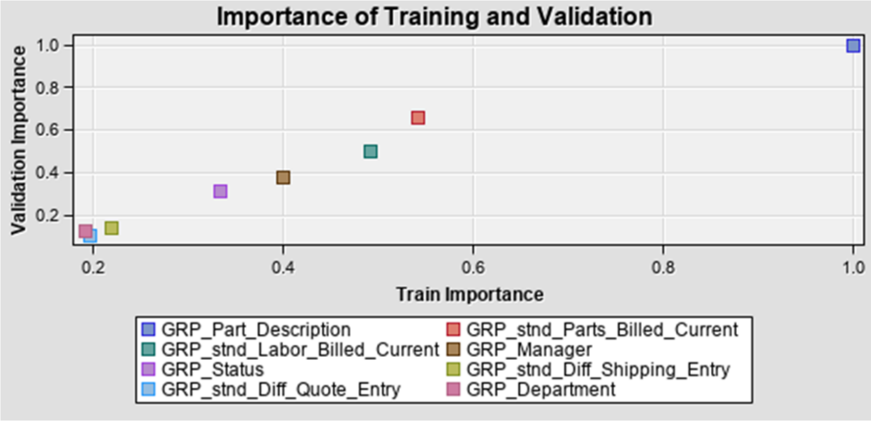

Figure 11. Scatter plot for importance of Boost86 model.

The scatter plot in Figure 11 reveals that the variable importance values for the training and validation data are closely aligned, improving model accuracy and reliability [22].

4. Discussion

Data analytics certainly becomes more critical for MRO companies to use their data for better planning and to be more effective, predictable, and profitable despite data analytics being new in this industry. The data available to MRO companies, including technical information, aircraft performance data, malfunction details, and reasons for failures, are largely held by aircraft manufacturers, OEMs, and airlines, and standardisation of these data is still a work in progress.

Real-world data were used in the study. As the raw operational data were not suitable for modelling, the study carried out the processes involved in data preparation (data cleaning, data preparation, creating derived variables, data transformation, data selection, etc.). These challenging processes were discussed in detail in Section 2.1. Variables were examined to retain those with high data quality, while errors and inconsistencies were eliminated from the dataset. Additionally, after consultations with professionals, several derived variables believed to explain the profit margins were created.

Furthermore, due to the difficulty in accessing such real-world data, the results were specific to the component maintenance segment of the SME MRO company. This study utilises data spanning from 2013 to 2021. Unique factors such as the company's organisational structure, market position, and specific operational challenges have significantly influenced the findings. Therefore, while the findings provide valuable insights for this company, caution should be exercised when generalising these results to other settings.

An interactive grouping method was employed for variable transformation. Interactive grouping also addressed the issue of outliers in the continuous variables, ensuring that the influence of extreme values was minimised, leading to more robust and reliable models.

As detailed in Section 3.5, hyperparameter tuning was performed manually [26], without the use of optimisation methods. Numerous models were developed for various hyperparameter combinations, targeting similar and minimal misclassification rates in the training and validation datasets.

The selection of gradient boosting for this study was driven by its proven efficacy in classification tasks [22,25]. Consequently, comparisons with other ML techniques were not included in this research.

To balance the trade-off between the complexity and interpretability of the gradient boosting model, variable importance values are discussed in Sections 3.5.1 and 3.5.2.

Two gradient boosting algorithms, Boost79 and Boost86, were compared in the study, with the details explained in Section 3.5. The Boost86 model, utilising grouped variables, demonstrated better performance with a lower misclassification rate and higher Gini value compared to the Boost79 model, which uses standardised variables. In addition, with variable clustering, clusters were created at the level of variables entering the model, and variables within the same cluster and separate clusters were analysed Figure 5. Variable clusters will guide strategic decisions by allowing a better understanding of the data (see Figure 6).

This study is based on the output of the gradient boosting algorithm, showing that profitability is not only related to direct variables such as Parts Billed Current and Labour Billed Current, but also to indirect variables like TAT, TTQ, Department, Manager, and Status. In the standard perspective, profitability is calculated mathematically by subtracting direct costs (labour and materials) from the price, and companies focus on increasing prices and reducing costs in this narrow philosophy. However, supply chain challenges and labour shortages, particularly after COVID-19, are prompting MRO companies, especially SME MROs, to explore alternative areas where they can devise unique strategies to boost profitability and differentiate themselves from their competitors.

The study draws attention to the Part Number variable, which refers to the unique and specific part code. There are thousands of parts installed in aircraft systems, and each aircraft model and type uses different parts, requiring different maintenance procedures, special equipment, tools, and materials, and specific workforce qualifications. Having many part numbers in the MRO capability list requires managing different tools, equipment, machinery, materials, and labour, which is a huge effort and cost burden on the management. It proves that finding the optimum and the best part number is the most important variable in the profitability model, where the best model output is shown in Table 20. Interactive grouping by the Gini index confirms that 1563 parts in group 1, shown in Table 5, provide a significant profit of over 0.5. Thus, the profitability model developed through gradient boosting gives important strategic business outcomes: MROs should focus on defining part number capability lists and prioritise the part numbers they choose to service.

The model confirms that direct variables related to profit margins, such as Parts Billed Current and Labour Billed Current, significantly affect profitability, as in Table 20. Certainly, as long as you reduce your material costs and provide more standard work, that is a higher number of the same parts, this will reduce your direct labour costs and contribute positively to profitability.

The most crucial results of this study are the identification and verification of hidden factors affecting profitability:

Manager: Each department provides repair services for its specific parts, so the department and its leader could play a critical role in the profitability; we addressed this factor based on the models in the study [27]. The model outputs indicate the leader's performance, such as planning, managing the process, finishing the tasks, and releasing the parts into service, is related to profitability. The Boost86 model with a validation ROC of 0.911 shows the manager's strong impact with a validation importance of 0.377, while the Boost79 model with a validation RASE of 0.637 indicates a lower impact of the Manager variable with a validation importance of 0.155. Per the analysis, an effective department leader, herein the manager, creates a distinctive effect on profitability, which is another business strategy the MRO company should consider.

Repair Status: The status of parts during repair impacts profitability. MROs issue a release-to-service certificate and invoice upon completion of the repair process, enabling profits upon shipment to the customer. In the Boost79 model, for the standardised variables, the Status variable has a 0.272 validation importance (see Table 18 and Figure 7) and is in the last three rows. On the other hand, the Boost86 model, which uses an interactive grouping of variables, ranks the variable higher with a validation importance of 0.309 (see Table 20 and Figure 11). Completing work promptly leads to higher turnover and effective time management, which maximises employee productivity and positively impacts business management. Incomplete tasks and prolonged work-in-progress statuses in the shop system result in reduced profitability and adversely affect overall performance.

Shipment Time: In the Boost79 model, the Diff Shipping Entry variable has a validation importance of 0.252, while it decreases to 0.140 in the Boost86 model. Both the Diff Shipping Entry and Diff Quote Entry variables are clustered separately, complementing each other and making similar contributions to the model shown in Figure 6. The model shows that faster jobs yield more profits, but longer jobs are less profitable; here, in this study, a shorter delivery date generates a profit margin above 0.50 with an event rate of 0.569 and longer dates reduce profitability with an event rate of 0.232, Table 9. This variable continues to be essential for customers and MRO companies, as it assists in managing labour and material resources by optimising time management.

Quote Time: This variable measures the time from entry to the shop to inspection completion, called TTQ. After inspection, the MRO determines the materials and labour needed for repair. These factors contribute to a quote price, factoring in the profit margin, which is then submitted to the customer for approval. The Diff Quote Entry variable is in the same cluster as the Diff Shipping Entry. Shorter quote times result in a higher event rate (0.541) with profit margins above 0.50, while longer quote times have a negative impact with a lower event rate (0.254) (see Table 11). Despite a reduced validation impact in the Boost86 model compared to the Boost79 model, the Diff Quote Entry variable continues to be important for the profit margin model. If the parts are quickly inspected, in a short time, the MRO company can calculate a quoted price and identify the projected profit margin for the part number.

Shop unit (Department): This variable also aligns with Part Number and Manager in the same cluster (see Figures 5 and 6). In this analysis, the Electro-Mechanical Department produces a higher event rate of 0.486 and an above 0.50 profit marginthan the other shop units, with the Electronics Department and Fuel Systems Department having a 0.359 event rate and the Hydraulic-Pneumatic Department a 0.256 event rate (see Table 10). Therefore, it is evident that some shop units have become more standardised and have more specific processes, which lead to a more profitable business. The Department variable carries the least influence compared to the other variables in both the improved model Boost86 model, with a validation importance of 0.122, and the standardised variables Boost79 model, with a validation importance of 0.095. It is essential to ensure that the department for specific part numbers has qualified manpower under the right unit leader, as this plays a critical role in contributing to the profitability model and achieving business targets.

When variable clustering was performed for the above variables that were significant according to the gradient boosting algorithm, three basic clusters were created. Clus1 contains the variables Part Description, Parts Billed Current (part cost), Labour Billed Current (labour cost), Department, and Manager (Figure 5). It can be stated that Clus1 primarily comprises information about the part. Updating the capability list with high-profit margin parts and combining high-profit margin departments with high-performance managers can increase profitability. Furthermore, changes in other variables in Clus1 can positively affect the labour cost of the part in Clus1

The second cluster, Clus2, includes TAT and TTQ. The fact is that Clus2 contains data for only two variables, which can be interpreted as meaning that the processes are independent of the Part Description, Parts Billed Current, Labour Billed Current, Department, and Manager variables in Clus1. It could be recommended to make general improvements to the processes, regardless of the part.

5. Conclusions

This study focuses on building a solid and sustainable profit margin model for the SME MRO company and addresses which variables are key to success in its future business.

The results of the models provide an authentic approach to MRO business management. Profitability is not only affected by direct, numerical variables such as labour and material costs, but also by qualitative variables such as Part Number, Manager, Department, and Status at significant levels. The study shows that the Part Number, Manager, and Department variables are in the same cluster as the Labour Billed Current and Parts Billed Current variables based on variable clustering. This provides a strategic overview to the MRO company as focusing on specific part numbers of certain departments with strong department leadership will bring sustainable success in profitability.

Another interesting view identified is that operational variables TAT and TTQ, which are also key performance metrics for the MRO company's success in the presence of its customers, are separated from the other variables for different clusters. Focusing on those variables also brings an advantage to the company as extra room for the profitability margin. In cases where competition is harsh, that option will bring strategic leverage to the business.

While the study offers valuable insights, it is important to acknowledge its limitations. First, the data used are from a single SME MRO company and are specific to the component maintenance segment, which may limit the generalizability of our findings to other segments or companies within the aviation industry. Additionally, the data span a specific period.

Future research should consider incorporating data from multiple SME MRO companies and different geographic regions. Furthermore, exploring alternative analytical techniques and machine learning algorithms may provide further insights. Investigating other segments within the aviation industry could also help in identifying unique factors influencing profit margins across different operational areas. The models can also be applied to other MRO segments such as engine, airframe, and line maintenance. Thus, the gradient boosting model for SME MROs developed in this study can be transformed into a more precise and robust MRO business forecasting model. This approach will enable us to assess the inclusiveness of the model for the entire MRO industry.

It is expected that future research will be able to conduct cross-studies using different data to produce similar results. Since this is the first study of its kind, its accuracy is expected to be verified through such cross-studies. This is a pioneering study; if there had been an opportunity to access similar operational data from different companies, the study could have been more comprehensive.

By addressing these limitations and pursuing the suggested research directions, future studies can build on these findings and contribute to a more comprehensive understanding of profit margin dynamics in the SME MRO industry.

Author Contributions: Conceptualisation, N.Ş.U.; methodology, N.Ş.U.; software, N.Ş.U.; validation, N.Ş.U. and A.H.B.; formal analysis, N.Ş.U.; investigation, N.Ş.U.; resources, N.Ş.U. and A.H.B.; data curation N.Ş.U.; writing—original draft preparation, N.Ş.U.; writing—review and editing, N.Ş.U. and A.H.B.; supervision, A.H.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MRO	maintenance, repair, and overhaul
WO	work order
ASE	Average-Squared Error
RASE	Root-Average-Squared Error
MCR	misclassification rate
OEMs	original equipment manufacturers
SMEs	small and medium-sized enterprises
TAT	turnaround time
TTQ	time to quote
Val	Validation
ROC	Receiver Operating Characteristic
ML	machine learning

References

1. Deac, V.; Carstea, G.; Bagu, C.; Parvu, F. The Modern Approach to Industrial Maintenance Management. *Inform. Econ.* **2010**, *14*, 133–144.
2. Sheng, J.; Prescott, D. A coloured Petri net framework for modelling aircraft fleet maintenance. *Reliab. Eng. Syst. Saf.* **2019**, *189*, 67–88. [CrossRef]
3. Berger, J.M. MRO Industry Forecast & Trends. In Proceedings of the IATA Maintenance Cost Conference, Geneva, Switzerland, 5 October 2022.
4. Oliverwyman.com. *MRO-Survey-2024-Aviation-MRO-Growth-Amid-Turbulence*; Oliver Wyman: New York, NY, USA, 2024. Available online: https://www.oliverwyman.com/our-expertise/insights/2024/apr/mro-survey-2024-aviation-mro-grows-amid-rising-costs-supply-chain-woes.html (accessed on 26 May 2024).
5. Statista.com. *Global Civil Air Transport MRO Market Share by Segment*; Statista: New York, NY, USA, 2024. Available online: https://www.statista.com/statistics/387977/global-civil-air-transport-mro-market-segmentation/ (accessed on 25 May 2024).
6. Tokgoz, A.; Bulkan, S.; Zaim, S.; Delen, D.; Torlak, N.G. Modeling airline MRO operations using a systems dynamics approach A case study of Turkish Airlines. *J. Qual. Maint. Eng.* **2018**, *24*, 280–310. [CrossRef]
7. Florio, F.D. *Airworthiness an Introduction to Aircraft Certification and Operations*, 3rd ed.; Butterworth-Heinemann: Oxford, UK, 2016; pp. 7–36.
8. Ulu, O.F. Data Analytics Methods Used for the Issues of Civil Aviation Maintenance Repair and Overhaul Industry—A Literature Review. Master's Thesis, Marmara University, Istanbul, Turkey, 2022.
9. Apostolidis, A.; Pelt, M.; Konstantinos, P.S. Aviation Data Analytics in MRO Operations: Prospects and Pitfalls. In Proceedings of the 2020 Annual Reliability and Maintainability Symposium, RAMS, Palm Springs, CA, USA, 27–30 January 2020; Abstract Number 9153694.
10. Sharda, R.; Dursun, D.; Turban, E. *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*, 4th ed.; Pearson Limited: London, UK, 2017; pp. 79–395.
11. Pelt, M.; Konstantinos, S.; Apostolidis, A. Data analytics case studies in the maintenance, repair and overhaul (MRO) industry. In Proceedings of the 9th EASN International Conference on "Innovation on Aviation and Space", MATEC Web Conference, Athens, Greece, 3–6 September 2019; Abstract Number 04005.
12. Dinis, D.; Barbosa-Póvoa, A.; Teixeira, Â.P. Valuing data in aircraft maintenance through big data analytics: A probabilistic approach for capacity planning using Bayesian networks. *Comput. Ind. Eng.* **2019**, *128*, 920–936. [CrossRef]
13. Available online: https://www.mordorintelligence.com/industry-reports/global-aircraft-maintenance-repair-and-overhaul-market-industry (accessed on 24 May 2024).
14. Nakhal, A.A.J.; Patriarca, R.; Di Gravio, G.; Antonioni, G.; PaltrinieriDinis, N. Investigating occupational and operational industrial safety data through Business Intelligence and Machine Learning. *J. Loss Prev. Process Ind.* **2021**, *73*, 0950–4230. [CrossRef]
15. Cortés-Ibañez, F.O.; Nagaraj, S.B.; Cornelissen, L.; Navis, G.J.; van der Vegt, B.; Sidorenkov, G.; de Bock, G.H. Prediction of Incident Cancers in the Lifelines Population-Based Cohort. *Cancers* **2021**, *13*, 2133. [CrossRef] [PubMed]
16. Sanche, R.; Lonergan, K. *Industrial Engineering and Operations Management*; Springer: Cham, Switzerland, 2022.
17. Strobl, C.; Boulesteix, A.L.; Augustin, T. Unbiased split selection for classification trees based on the Gini Index. *Comput. Stat. Data Anal.* **2007**, *52*, 483–501. [CrossRef]
18. Kuyoro, A.O.; Nagaraj, S.B.; Ogunyolu, O.A.; Ayanwola, T.G.; Ayankoya, F.Y.; van der Vegt, B.; Sidorenkov, G.; de Bock, G.H. Prediction of Incident Cancers in the Lifelines Population-Based Cohort. *Ingénierie Syst. d'Inf.* **2022**, *27*, 815–821. [CrossRef]
19. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a dataset via the gap analysis. *J. R. Statist. Soc. B* **2001**, *63*, 411–423. [CrossRef]

20. Lletí, R.; Ortiz, M.C.; Sarabia, L.A. Sánchez, M.Z. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Anal. Chim. Acta* **2004**, *515*, 87–100. [CrossRef]
21. Sanche, R.; Lonergan, K. *Variable Reduction for Predictive Modelling with Clustering*; Casualty Actuarial Society Forum: Arlington, VA, USA, 2006.
22. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *5*, 1189–1232. [CrossRef]
23. Jawalkar, A.P.; Swetcha, P.; Manasvi, N.; Sreekala P.; Aishwarya, S.; Bhavani, P.K.D.; Anjani, P. Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting. *J. Eng. Appl. Sci.* **2023**, 70–122. [CrossRef]
24. Esposito, M.; Lazoi, M.; Margarito, A.; Lorenzo, Q. Innovating the Maintenance Repair and Overhaul Phase through Digitalization. *Aerospase* **2019**, *6*, 53. [CrossRef]
25. Guelman, L. Gradient boosting trees for auto insurance loss cost modelling and prediction. *Expert Syst. Appl.* **2012**, *39*, 3659–3667. [CrossRef]
26. Paul, D.; Goswami, A.K.; Chetri, R.L.; Roy, R.; Sen, P. Bayesian Optimization-Based Gradient Boosting Method of Fault Detection in Oil-Immersed Transformer and Reactors. *IEEE Trans. Ind. Appl.* **2022**, *58*, 1910–1919. [CrossRef]
27. Pereira, D.P.; Gomes, I.L.R.; Melicio, R.; Mendes, V.M.F. Planning of Aircraft Fleet Maintenance Teams. *Aerospase* **2021**, *8*, 140. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.