



Optimizing headcount decisions in the apparel industry: Leveraging Predictive Analytics and Machine Learning

**A dissertation submitted for the Degree of Master of
Business Analytics**

A.G.N.K. Ariyasena

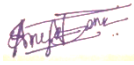
University of Colombo School of Computing

2024

Declaration

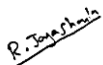
| |
|---|
| Name of the student: A.G.N.K. Ariyasena |
| Registration number: 2020/BA/003 |
| Name of the Degree Programme: Master of Business Analytics |
| Project/Thesis title: Optimizing headcount decisions in the apparel industry: Leveraging Predictive Analytics and Machine Learning |

1. The project/thesis is my original work and has not been submitted previously for a degree at this or any other University/Institute. To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.
2. I understand what plagiarism is, the various types of plagiarism, how to avoid it, what my resources are, who can help me if I am unsure about a research or plagiarism issue, as well as what the consequences are at University of Colombo School of Computing (UCSC) for plagiarism.
3. I understand that ignorance is not an excuse for plagiarism and that I am responsible for clarifying, asking questions and utilizing all available resources in order to educate myself and prevent myself from plagiarizing.
4. I am also aware of the dangers of using online plagiarism checkers and sites that offer essays for sale. I understand that if I use these resources, I am solely responsible for the consequences of my actions.
5. I assure that any work I submit with my name on it will reflect my own ideas and effort. I will properly cite all material that is not my own.
6. I understand that there is no acceptable excuse for committing plagiarism and that doing so is a violation of the Student Code of Conduct.

| | |
|---|------------------------------|
| Signature of the Student | Date (DD/MM/YYYY) |
|  | 28/09/2024 |

Certified by Supervisor(s)

This is to certify that this project/thesis is based on the work of the above-mentioned student under my/our supervision. The thesis has been prepared according to the format stipulated and is of an acceptable standard.

| | Supervisor 1 | Supervisor 2 | Supervisor 3 |
|------------------|---|---------------------|---------------------|
| Name | Mr. R.J. Amaraweera | | |
| Signature |  | | |
| Date | 28/09/2024 | | |

ACKNOWLEDGEMENT

I express my sincere appreciation to my research supervisor, Mr. R.J. Amaraweera, Lecturer at the University of Colombo School of Computing, for his invaluable guidance, advice, and unwavering support throughout this research endeavour.

I am also grateful to all the lecturers who participated in my proposal defence and interim presentations, offering their insightful advice and perspectives. Special thanks are extended to Mr. R.J. Amaraweera and Ms. Nimali Hettiarachchi, the final year Business Analytics research coordinators, for their assistance and support.

I extend my gratitude to Emjay International and Penguin Sportswear (Pvt) Ltd. for providing the dataset essential for conducting my research.

My heartfelt thanks go to my friends for sharing their knowledge and supporting me throughout this journey. Additionally, I am thankful to my family for their unwavering support and assistance during the research process.

Lastly, I want to express my appreciation to everyone who contributed to the successful completion of my research.

ABSTRACT

Labour management, especially in labour-intensive sectors like apparel manufacturing, is crucial for maintaining operational efficiency and managing costs effectively. The COVID-19 pandemic and ensuing economic challenges have worsened these concerns, prompting the need for innovative solutions. In light of devising a solid mechanism to facilitate headcount-related decisions, this research focuses on Enjay Penguin, a prominent garment manufacturer in Sri Lanka, which sought to optimize its warehouse department's headcount management amidst changing demands and digital transformation initiatives. A data-driven approach was adopted, leveraging machine learning algorithms to predict optimal worker and staff cadre requirements. Through rigorous experimentation and performance evaluation, Gradient Boosting Regressor emerged as the most effective model, which together with the Suffering Index offering precise predictions and insights into headcount reduction or reallocation potential. Key features influencing cadre prediction were identified, aiding informed decision-making. The research successfully achieved its objectives, providing a robust framework for headcount management and cost reduction. The findings underscore the importance of aligning labour resources with production targets and demonstrate the value of advanced technologies in enhancing operational efficiency.

Keywords: Labour management, apparel industry, headcount optimization, machine learning, Gradient Boosting Regressor, cost reduction, data-driven decision-making.

TABLE OF CONTENTS

| | |
|--|------|
| ACKNOWLEDGEMENT..... | ii |
| ABSTRACT | iii |
| LIST OF FIGURES | vii |
| LIST OF TABLES..... | viii |
| LIST OF ACRONYMS | ix |
| CHAPTER 1 INTRODUCTION | 1 |
| 1.1 Motivation | 1 |
| 1.2 Statement of the problem..... | 2 |
| 1.3 Research Aims and Objectives | 5 |
| 1.3.1 Aim | 5 |
| 1.3.2 Objectives | 5 |
| 1.4 Scope | 6 |
| 1.5 Structure of the Thesis | 6 |
| CHAPTER 2 LITERATURE REVIEW | 7 |
| 2.1 Headcount Management and Business Performance..... | 7 |
| 2.2 Predicting Headcount | 9 |
| 2.2.1 Early Approaches and Intuition..... | 11 |
| 2.2.2 Spreadsheet-Based Models..... | 11 |
| 2.2.3 Statistical Methods | 11 |
| 2.2.4 Machine Learning and Advanced Analytics | 11 |
| 2.2.5 Predictive Analytics Platforms | 11 |
| 2.2.6 Real-Time Data and AI-Powered Forecasting..... | 12 |
| 2.2.7 Cloud-Based Solutions and Scalability | 12 |
| 2.3 Headcount Prediction and Demand Conditions..... | 12 |
| 2.4 Headcount Optimization..... | 14 |
| 2.4.1 Evolution of Headcount Optimization..... | 14 |
| 2.4.2 Methods of Headcount Optimization | 14 |
| 2.4.3 Challenges in Headcount Optimization..... | 15 |
| 2.5 Predicting and Optimizing Headcount in Labour Intensive Manufacturing Industries | 16 |
| 2.6 Critical Analysis of Related Work..... | 17 |
| 2.7 Drawbacks and Limitations of Previous Research Work..... | 21 |

| | | |
|------------|--|----|
| CHAPTER 3 | METHODOLOGY | 23 |
| 3.1 | Research Questions..... | 23 |
| 3.2 | Data Collection | 23 |
| 3.3 | Data Pre-processing | 25 |
| 3.3.1 | Data Reduction | 26 |
| 3.3.2 | Data Cleaning and handling missing values..... | 26 |
| 3.3.3 | Data Transformation..... | 26 |
| 3.3.4 | Feature Selection | 26 |
| 3.4 | Machine Learning Models Selection..... | 27 |
| 3.4.1 | Model Architecture..... | 27 |
| 3.4.2 | Assumptions | 29 |
| 3.4.3 | Model Selection..... | 30 |
| 3.4.4 | Prediction Experiments..... | 30 |
| 3.4.5 | Evaluation..... | 31 |
| 3.5 | Justification for the Methodology Selected..... | 32 |
| CHAPTER 4 | EVALUATION AND RESULTS | 34 |
| 4.1 | Implementation..... | 34 |
| 4.1.1 | Data Preprocessing | 34 |
| 4.1.2 | Data Exploration..... | 36 |
| 4.1.3 | Feature Selection and Model Selection | 38 |
| 4.1.4 | Model Building and User Interface Development..... | 38 |
| 4.2 | Results | 40 |
| 4.2.1 | Experiment 1: Prediction of headcount with original data | 40 |
| 4.2.2 | Experiment 2: Prediction of headcount with Principal Component Analysis .. | 40 |
| 4.2.3 | Experiment 3: Prediction of headcount with Permutation Feature Importance | 41 |
| 4.3 | Evaluation..... | 42 |
| 4.3.1 | Algorithm Performance | 42 |
| 4.3.2 | Model Evaluation | 45 |
| CHAPTER 5 | CONCLUSION AND FUTURE WORK | 47 |
| 5.1 | Conclusion..... | 47 |
| 5.2 | Future Work..... | 49 |
| APPENDICES | | I |
| | Appendix A: Screen printing of codes – Model Training and selection | I |
| | A.1 Importing useful libraries and loading data | I |
| | A.2 Preprocessing | I |

| | |
|--|-----|
| A.3 Training model..... | III |
| A.4 Prediction and Decision support..... | IV |
| A.5 Model Evaluation..... | V |
| Appendix B: Screen printing of codes – Final Model | V |
| Appendix C: Consent Letter | X |
| REFERENCES | XI |

LIST OF FIGURES

| | |
|--|----|
| Figure 1.1: Yearly Headcount Vs Sales Analysis..... | 3 |
| Figure 1.2: Monthly Warehouse Indirect Cadre Vs Sales | 4 |
| Figure 3.1: Research Methodology | 23 |
| Figure 3.2: Proposed Architecture..... | 28 |
| Figure 3.3: Scikit-learn Machine Learning Algorithm Cheat Sheet..... | 33 |
| Figure 4.1: Correlation Matrix Heatmap | 37 |
| Figure 4.2: Proposed interface for Required Cadre Prediction | 39 |
| Figure 4.3: Mean Cross-Validation Score at Different k-values | 45 |

LIST OF TABLES

| | |
|---|----|
| Table 2-1:Summary of Critical Analysis of Related Research Work | 19 |
| Table 3-1: Interpretation of Typical MAPE Values | 31 |
| Table 4-1:Descriptive Statistics of the dataset | 36 |
| Table 4-2: Attributes with High Correlation | 37 |
| Table 4-3: Comparison of results from Experiment 1 | 40 |
| Table 4-4:Comparison of results from Experiment 2 | 41 |
| Table 4-5:Comparison of results from Experiment 3 | 41 |
| Table 4-6: Feature Importance Values - Permutation feature importance | 42 |
| Table 4-7: Algorithm Performance based on R^2 Score | 42 |
| Table 4-8:Algorithm Performance based on MSE | 43 |
| Table 4-9:Algorithm Performance based on MAE..... | 43 |
| Table 4-10:Algorithm Performance based on MAPE | 44 |

LIST OF ACRONYMS

| | |
|-----------|----------------------------------|
| AI | Artificial Intelligence |
| CMS | Cellular Manufacturing Systems |
| CSV | Comma Separated Values |
| CV | Cross-Validation |
| Dtype | Data Type |
| EMJ | Emjay International (Pvt) Ltd |
| ERP | Enterprise Resource Planning |
| FastReact | FastReact Planning Tool |
| FTE | Full-Time Equivalent |
| GRN | Good Receiving Note |
| IFS | Industrial and Financial Systems |
| IT | Information Technology |
| KGL | Karandagolla |
| KPI | Key Performance Indicators |
| MAE | Mean Absolute Error |
| MAPE | Mean Absolute Percentage Error |
| Max | Maximum |
| Min | Minimum |
| ML | Machine Learning |
| MSE | Mean Squared Error |
| OptQuest | OptQuest optimizer |
| PAN | Panvila |
| PCA | Principal Component Analysis |
| PSO | Particle Swarm Optimization |
| PSW | Penguin Sportswear (Pvt) Ltd |
| Qty | Quantity |
| R^2 | Coefficient of Determination |
| RM | Raw Material |
| RQ | Research Question |
| RVU | Relative Value Unit |
| SA | Simulated Annealing |
| SAH | Standard Achieved Hours |
| SI | Suffering Index |
| SIMUL8 | SIMUL8 Simulator |
| STD | Standard |
| TAL | Talawinna |
| TLD | Theldeniya |
| TT | Transport Task |
| Vs | Versus |

CHAPTER 1

INTRODUCTION

The apparel industry stands as one of the most labour-intensive sectors globally. Consequently, high labour costs have become commonplace within this industry (Rodrigo & Ratnayake, 2021). However, amid the economic downturn witnessed in various regions, a significant decline in order placements has been observed, directly impacting the revenue forecasts of apparel companies (Li, 2022). Despite the revenue decrease, labour costs persist, prompting apparel companies to seek cost-cutting measures wherever possible (Rodrigo & Ratnayake, 2021).

Technological advancements have enabled many companies to lessen their reliance on direct labour (Sjödín et al., 2018). Various tools and machines available in the market have facilitated apparel production. However, the number of indirect staff remains stagnant despite technological availability (Sjödín et al., 2018). This stagnation is attributed to companies' inability to determine the optimal headcount levels that drive targets and identify potential areas for staff reduction without affecting objectives (Trevor & Nyberg, 2008).

Consequently, management in these apparel companies resorts to forecasting required headcount levels based on intuition, posing challenges in personnel management and achieving Key Performance Indicators (KPIs). Thus, the necessity for a data-driven model arises which is capable of predicting optimal headcount levels throughout the company. Such a model would facilitate the identification of areas for staff reduction without compromising goal achievement (Goubko & Mishin, 2009).

1.1 Motivation

Identifying the ideal headcount stands as a critical concern for apparel companies due to their high reliance on labour. However, amidst the European market recession, global companies are witnessing a significant decline in order placements. This market downturn has left the industry grappling with strategies to meet established targets. While pinpointing factors directly impacting sales should be a priority for management, companies lack comprehensive insights into company-wide determinants beyond basic production parameters.

Moreover, despite the revenue stream dilution, the cost structure remains unchanged, prompting companies to restructure their expenses. Given that labour costs represent a substantial portion, achieving an optimal headcount mix becomes pivotal to ensure efficient utilization of human

capital and mitigate excessive hiring and training expenses due to labour shortages. However, preceding the COVID-19 pandemic, many companies extensively expanded their workforce with aspirations of growth (Ardiyono, 2022). Nonetheless, the current decline in sales fails to align with this decision, resulting in soaring labour costs. The absence of a proper mechanism to determine the optimal headcount exacerbates the situation.

Emjay Internation & Penguin Sportswear (Pvt) Ltd (referred to as Emjay Penguin Group hereafter) stands as a prominent apparel manufacturer in Sri Lanka, striving to meticulously plan its headcount to control labour costs. With the goal of achieving \$100 million in sales by 2025, they are actively formulating strategies. However, similar to other companies in the industry, Emjay Penguin Group grapples with achieving sales targets while optimizing the headcount mix. The company relies on management's intuition and experience to propose various headcount reduction strategies.

Management remains apprehensive about excessively downsizing the workforce, fearing disruptions to company operations. However, the absence of readily available tools for predicting optimal headcount levels leaves them in a state of uncertainty. This spurred the researcher to explore potential models or tools to aid the apparel industry in headcount planning.

1.2 Statement of the problem

Emjay Penguin Group comprises two primary entities: Penguin Sportswear (Pvt) Ltd and Emjay International (Pvt) Ltd. Penguin Sportswear (Pvt) Ltd was established in 1990, while Emjay International (Pvt) Ltd commenced its operations in 2005 (Emjay International, 2023). The consolidation of these two companies occurred in 2010, resulting in the operation of a single entity (Emjay International, 2023), with production activities spanning four factories located in Kandy and Kurunegala. Over time, the company has expanded its presence in numerous international markets by servicing renowned brands such as Guess, Hugo Boss, Helly Hansen, George, Tesco (F&F), New Look, and others. With a daily production capacity exceeding 20,000 pieces, the company employs over 4000 individuals nationwide (Emjay International, 2023).

The apparel industry is characterized by market demand volatility, a challenge faced by Emjay Penguin Group and many other apparel manufacturers worldwide. Consequently, escalating costs, not fully offset by generated income, pose a continual concern. Given the labour-intensive nature of the industry, salary expenses constitute a significant portion of the cost structure,

making it a prominent component.

The company's workforce can be broadly categorized into two groups: Direct and Indirect. The Direct cadre encompasses shop floor machine operators, directly involved in production activities, while the Indirect cadre includes all other employee categories such as managers, executives, office staff, and workers. While an increase in the Direct cadre to meet demand is essential due to their direct association with production, the headcount within the Indirect cadre has experienced significant fluctuations over the years, often not directly correlated with demand or sales targets. Figure 1.1 below illustrates the fluctuations in the Direct and Indirect cadre against sales targets within the company.

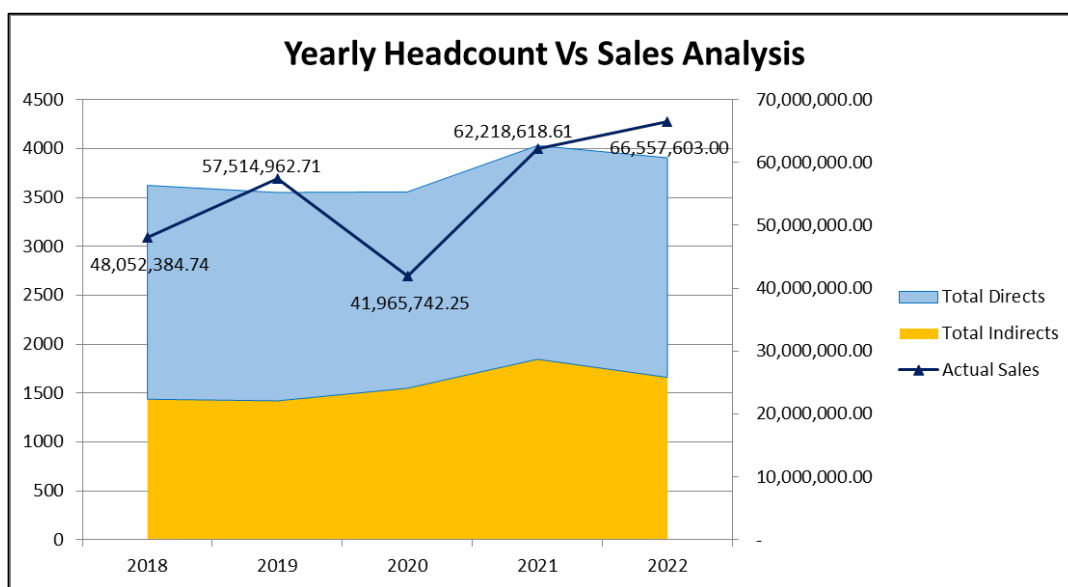


Figure 1.1: Yearly Headcount Vs Sales Analysis

Source: Emjay Penguin Internal Reports

The direct cadre requirements are straightforwardly determined by converting sales figures into Standard Achieved Hours (SAH) and subsequently calculating the number of running machines needed to achieve such SAH. However, as mentioned earlier, the correlation between the indirect cadre and sales figures is not direct. Consequently, decision-makers tend to establish headcount requirements based on department-specific Key Performance Indicators (KPIs).

Particularly in departments like the warehouse, there is a significant involvement of the indirect cadre due to the substantial physical labour required. Hence, it becomes crucial to identify the optimal headcount to ensure that employees are neither idle nor overburdened. The analysis of warehouse indirect cadre in Emjay Penguin Group relative to total sales from April 2018 to March 2022 is depicted in figure 1.2 below.

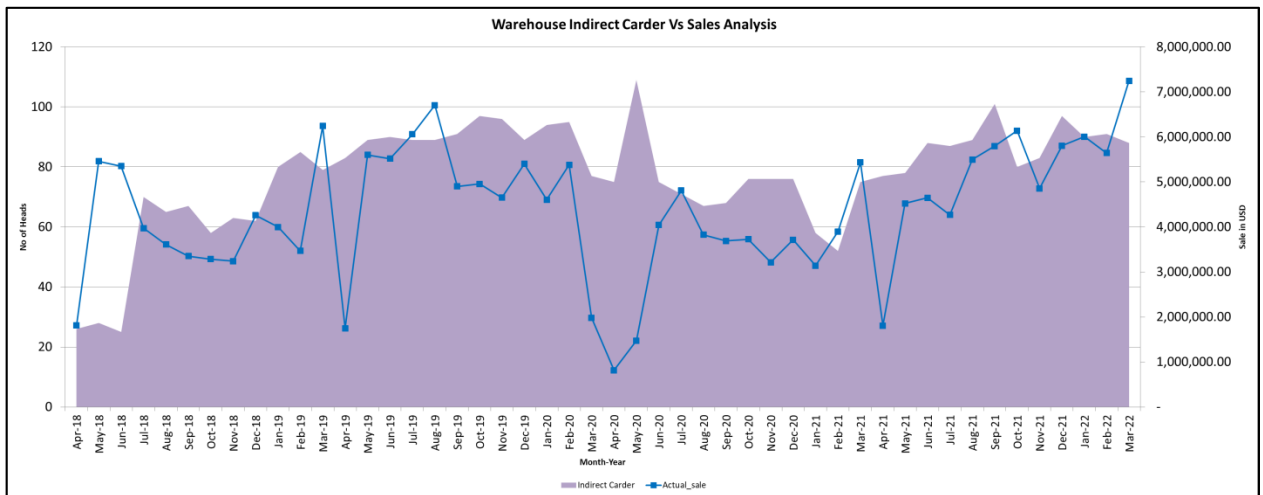


Figure 1.2: Monthly Warehouse Indirect Cadre Vs Sales
Source: Emjay Penguin Internal Reports

The fluctuation in headcount has inflicted substantial costs on the company, including salaries paid for underutilized resources and expenses incurred during rushed recruitment drives to address shortages. Notably, the costs incurred for worker and staff categories are particularly significant compared to other employment categories. Consequently, the company is eager to identify headcount requirements in a systematic manner supported by robust evidence to justify the incurred costs.

As the company embraces digital transformation, its aspiration for data-driven decision-making has prompted the researcher to explore technologically advanced solutions to address this issue through this paper.

Therefore, the importance of resolving the aforementioned problem can be summarized as follows:

- Proposing new paradigms for staffing costs of the company to ensure a balanced cost structure aligned with income sources.
- Facilitating optimized workforce planning across the organization to anticipate and fulfill staffing needs in different departments, locations, or timeframes promptly, ensuring adequate staffing with requisite skills to meet business demands.
- Providing insights into enhanced recruitment and retention strategies by analyzing historical data on employee turnover, hiring rates, and other relevant factors to identify patterns and trends.
- Improving performance management by identifying potential performance issues or areas requiring additional support.
- Supporting strategic decision-making through data driven models to enable expansion

into new markets, changes in business strategy, or mergers and acquisitions regarding staffing needs that align with their long-term objectives.

- Supporting strategic decision-making through data-driven models to facilitate decisions regarding expansion into new markets, changes in business strategy, or mergers and acquisitions, ensuring alignment with long-term objectives regarding staffing needs.

1.3 Research Aims and Objectives

Aims and objectives of carrying out this research are listed below.

1.3.1 Aim

- Main aim of this research is to provide insights on reducing indirect labour cost to ensure a secured profit margin.

In addition to the main aim, sub aims to be achieved at the completion of the research are listed below.

- To provide insight into the department specific features that affect the sales targets directly.
- To help management identify the worker and staff cadre mix of a department to achieve a potential sales/production target.
- To enable the senior managers to make decisions about headcount reduction/reallocation based on data.

1.3.2 Objectives

- Identify department specific significant features that affect headcount prediction.
- Predict the optimum no of Required worker cadre to achieve a given sale/production quantity
- Predict the optimum no of Required Staff cadre to achieve a given sale/production quantity
- Uncover the headcount reduction /reallocation potential

1.4 Scope

This research is mainly focusing to develop a model to predict the headcount which drives the achievement of a given sales/production target while identifying the potential areas for the headcount reduction.

Due to the time constraint as well the vastness of the domain, the scope is limited to the apparel industry focusing on one company in Sri Lanka in particular. Narrowing down the area of consideration further, the research will focus on the Warehouse department's worker and staff cadre, as the number of heads employed in these indirect categories in this department is significant compared to other supportive departments as a result of high labour-intensive tasks involved. Moreover, the availability of technological solutions such as warehouse management systems motivated to focus on this department, since it has a great potential of headcount reduction through automation. However, this model will be designed in a way that it could be naturally extended further to any department, covering the overall company.

1.5 Structure of the Thesis

Following are the main chapters of the dissertation.

Chapter I: Introduction

This will provide the basic idea of research overview, research objectives and expected methodology on how to carry out the research.

Chapter II: Literature Review

The Literature Review is based on related work under some subtopics such as the work based on the objective, attributes, and performance.

Chapter III: Research Methodology

The research design is presented, with an overview and in-depth description, followed by a justification of the methods utilized.

Chapter IV: Results and Evaluation

The research's findings and evaluation are presented, in which the findings were assessed from a variety of angles.

Chapter V: Conclusion and Future Research

Discusses the research challenges, goals, and objectives as well as the limitations and implications for the future.

CHAPTER 2

LITERATURE REVIEW

This chapter will bring about a detailed review of the literature considering the main concepts addressed through this research paper. Moreover, related or similar work done by past researchers on the subject matter is critically analyzed to gain insights on modelling the method to carry out this research.

Accordingly, for this research, combination of three domains will be considered as explained follows.

Human Resources: The first domain is human resources, which deals with managing people within an organization. In this research, ways to optimize headcount reduction will be explored, which is a key concern for HR professionals. Understanding HR practices, policies, and procedures will equally be essential for this research.

Sales: The second domain is sales, which involves the process of selling products or services and achieving sales targets. In this research, it will need to be ensured that the headcount reduction does not negatively impact sales targets. It will be assumed that production quantity equals sales quantity in order to make sure production capacity is met while achieving sales targets.

Machine Learning: The third and most important domain is machine learning, which is a subset of artificial intelligence that involves using algorithms and statistical models to enable computers to learn from data and make predictions. This research will involve using machine learning and predictive analysis to optimize headcount reduction without affecting sales targets, so a solid understanding of machine learning principles and techniques will be important.

2.1 Headcount Management and Business Performance

In today's dynamic and competitive business landscape, effective headcount management stands out as a fundamental pillar that profoundly influences the performance and success of companies across various industries (Wright, et al., 2003). The strategic allocation and optimization of human resources have emerged as crucial factors in achieving organizational goals, maintaining financial health, and fostering innovation (Jamal & Saif, 2011).

In manufacturing and production sectors, headcount management plays a pivotal role in streamlining operations, reducing costs, and enhancing productivity (Wright, et al., 2005). When considering operational efficiency, proper headcount management ensures that the workforce is aligned with production demands (Sjödén, et al., 2018). Overstaffing can lead to

excessive labour costs and underutilization of resources, while understaffing can result in delays, compromised product quality, and decreased output (Jamal & Saif, 2011). Maintaining an optimal workforce size helps streamline production processes, reduce waste, and enhance overall operational efficiency. Similarly, maintaining consistent product quality is crucial for manufacturing businesses (April, et al., 2006). Overburdened staff can lead to errors and decreased product quality. Proper headcount management ensures that employees have sufficient time to focus on their tasks, reducing the likelihood of defects and enhancing the overall quality of manufactured goods (Hertz, et al., 2010).

Labour costs often constitute a significant portion of a manufacturing company's expenses. Effective headcount management allows companies to control labour-related expenditures (Hertz, et al., 2010). By matching workforce levels to production needs, manufacturers can minimize unnecessary labour expenses and allocate resources more efficiently (Jamal & Saif, 2011). Further, better flexibility and adaptability is provided to the manufacturing industry through proper headcount management. The manufacturing industry is subject to market fluctuations, changing customer preferences, and evolving technologies (Sjödin, et al., 2018). A well-managed headcount enables manufacturers to adapt to these changes more swiftly. When the workforce is appropriately sized, manufacturers can pivot production lines, introduce new products, and scale operations without facing unnecessary disruptions (Sjödin, et al., 2018).

In service-oriented industries like hospitality and retail, customer satisfaction is paramount. Headcount management directly impacts the quality of service provided (Baier, et al., 2012). Insufficient staffing can lead to longer wait times, reduced attention to customers, and overall dissatisfaction (Baier, et al., 2012). A properly managed headcount ensures that customers receive prompt and personalized service, leading to positive experiences and repeat business. Employees are often the face of the business, in service industries (Kawas, et al., 2013). Proper staffing levels prevent employee burnout and stress caused by excessive workloads. When employees are adequately supported, they can provide better service, maintain a positive attitude, and contribute to a healthier work environment (Jamal & Saif, 2011). Going hand-in-hand with employee wellbeing, Service businesses rely heavily on the direct interaction between employees and customers. Well-managed staffing levels ensure that every customer interaction is meaningful and contributes to revenue generation (Kawas, et al., 2013). From upselling to cross-selling, a properly staffed team can capitalize on sales opportunities and boost profitability.

Another inherent characteristic of the service industry is the fluctuating demand patterns. Effective headcount management allows businesses to adjust their workforce according to peak and off-peak periods. This agility ensures that service levels remain consistent regardless of customer demand variations (Baier, et al., 2012).

The healthcare and pharmaceutical industries require precise headcount management to ensure quality patient care, maintain compliance with medical regulations, and drive medical advancements. In healthcare, having the right number of healthcare professionals, administrators, and support staff is essential to deliver optimal patient outcomes while managing costs (Olya, et al., 2022). This very reason has led many researchers to revolve their studies around this industry due to this criticality.

The technology industry thrives on innovation and rapid development cycles. Effective headcount management allows tech companies to scale their teams appropriately as projects expand or evolve. Avoiding workforce bloat ensures that companies can invest resources into research and development, attract top talent, and maintain a flexible structure conducive to embracing disruptive technologies (Sjödín, et al., 2018).

The energy and utilities sectors, often dealing with capital-intensive projects, require careful headcount management to optimize resource allocation (Chien, et al., 2010). Skilled engineers, technicians, and support staff must be allocated efficiently to ensure smooth operations, minimize downtime, and meet environmental and safety standards (Chien, et al., 2008).

2.2 Predicting Headcount

Headcount prediction, also known as workforce forecasting or staff demand forecasting, involves using historical data and statistical techniques to anticipate future workforce needs (Olya, et al., 2022). This is a critical aspect of human resource management that helps organizations plan their staffing levels efficiently.

Headcount prediction plays a vital role in strategic workforce planning (Zhao, et al., 2022). It allows organizations to align their staffing levels with business goals and market demands. Accurate predictions enable companies to avoid the costs and disruptions associated with sudden shortages or surpluses of personnel (Chien, et al., 2008). Moreover, it facilitates better allocation of resources, improves employee engagement, and enhances overall organizational performance (Chien, et al., 2008).

Past literature emphasizes the significance of using data-driven approaches for accurate headcount prediction. Historical data on employee turnover, hiring rates, promotions, retirements, and other relevant factors are crucial inputs for predictive models (Rodrigo &

Ratnayake, 2021). Advanced statistical techniques, such as time series analysis, regression, and machine learning algorithms, have been employed to capture underlying patterns and trends in workforce dynamics (Mundschenk & Drexler, 2007). The literature underscores the need to consider the appropriate time horizon and granularity when predicting headcount. Short-term predictions are useful for immediate staffing decisions, such as dealing with seasonal demand fluctuations, while long-term predictions support strategic workforce planning and talent development initiatives (Hertz, et al., 2010).

Effective headcount prediction accounts for external factors that influence workforce demand. These factors can include changes in market conditions, economic trends, industry growth rates, technological advancements, and regulatory changes (Pac, et al., 2009). Incorporating external variables into prediction models enhances their accuracy and reliability (Pac, et al., 2009).

Some studies have highlighted the importance of moving beyond generic headcount prediction to skill-based or role-based predictions (Drexler & Mundschenk, 2008). Different roles within an organization require varying skill sets and qualifications. Predicting the demand for specific skills or roles allows organizations to address talent shortages in critical areas (Drexler & Mundschenk, 2008).

Literature suggests that static prediction models may not fully capture the dynamic nature of workforce fluctuations. Dynamic models that can adapt to changing conditions, feedback loops, and unexpected events are more resilient and accurate in anticipating headcount needs (Olya, et al., 2022). The quality of input data significantly impacts the accuracy of headcount prediction models. Ensuring data accuracy, consistency, and completeness through proper preprocessing is crucial (Gröger, et al., 2012). Inaccurate or incomplete data can lead to unreliable predictions and suboptimal staffing decisions (Gröger, et al., 2012).

To ensure the reliability of predictions, past literature highlights the importance of model validation and refinement. Employing techniques like cross-validation, sensitivity analysis, and back testing helps assess the performance of prediction models against actual outcomes, leading to continuous improvement (Rodrigo & Ratnayake, 2021). Effective headcount prediction involves collaboration between human resources, finance, operations, and other relevant departments (Jamal & Saif, 2011). Open communication and sharing of insights gained from prediction models facilitate informed decision-making and alignment with overall business strategies (Jamal & Saif, 2011).

The evolution of headcount prediction stands as a testament to the relentless progress of data analytics, technology, and business strategies. From rudimentary methods to sophisticated

predictive models, the journey reflects the transformation of organizations' ability to anticipate and manage their workforce needs.

2.2.1 Early Approaches and Intuition

In the early stages of workforce management, headcount prediction relied heavily on gut feelings, anecdotal evidence, and historical patterns. Organizations made staffing decisions based on subjective estimations of workload and employee turnover (Trevor & Nyberg, 2008). While these methods were often intuitive, they lacked precision and struggled to accommodate the complexities of modern business environments (Trevor & Nyberg, 2008).

2.2.2 Spreadsheet-Based Models

As technology advanced, businesses transitioned to using spreadsheets for basic headcount forecasting (Sjödín, et al., 2018). These models involved manually inputting historical data and applying simple mathematical formulas (Sjödín, et al., 2018). While a step forward from intuition, they were limited in their ability to handle dynamic factors and make accurate predictions in the face of changing market conditions (Pac, et al., 2009).

2.2.3 Statistical Methods

With the rise of computational power, statistical methods like time series analysis and regression came into play. These techniques allowed organizations to analyze historical data more systematically, considering trends, seasonality, and external variables (Sjödín, et al., 2018). While offering improved accuracy, they still struggled to incorporate the intricate interdependencies within complex business ecosystems.

2.2.4 Machine Learning and Advanced Analytics

The advent of machine learning marked a significant turning point in headcount prediction. Businesses began leveraging algorithms capable of learning from vast datasets, identifying patterns, and making increasingly accurate forecasts (Gröger, et al., 2012). Machine learning models, such as decision trees, random forests, and neural networks, could factor in a multitude of variables and account for nonlinear relationships, enabling organizations to achieve higher prediction precision (Rodrigo & Ratnayake, 2021).

2.2.5 Predictive Analytics Platforms

The evolution of headcount prediction culminated in the development of predictive analytics platforms. These platforms integrated a range of advanced algorithms and tools, offering user-friendly interfaces for organizations to model their workforce needs (Rodrigo & Ratnayake, 2021). With features like data visualization, scenario analysis, and real-time updates, these platforms empowered decision-makers to make informed choices based on data-driven insights (Gröger, et al., 2012).

2.2.6 Real-Time Data and AI-Powered Forecasting

In recent years, the integration of real-time data streams and artificial intelligence (AI) has revolutionized headcount prediction. AI-powered algorithms can analyze vast amounts of data from various sources in real time, allowing organizations to respond rapidly to changing conditions (April, et al., 2006). These systems can also adapt and improve over time, continuously enhancing prediction accuracy.

2.2.7 Cloud-Based Solutions and Scalability

Cloud computing has further accelerated the evolution of headcount prediction by providing scalable and cost-effective solutions. Organizations can now leverage cloud-based platforms to handle large datasets, perform complex calculations, and deploy predictive models without the need for extensive on-premises infrastructure (April, et al., 2006).

The journey of headcount prediction evolution showcases the remarkable progression from subjective estimations to data-driven insights powered by advanced analytics and AI. Businesses across industries have reaped the benefits of increased accuracy, better resource allocation, and improved operational efficiency (Gröger, et al., 2012). As technology continues to advance, the future of headcount prediction holds promises of even greater precision and adaptability, enabling organizations to thrive in an increasingly dynamic and competitive global landscape (Gröger, et al., 2012).

Headcount prediction is a multidimensional and dynamic process that draws on data, statistical techniques, and strategic insights to forecast future workforce needs accurately (Olya, et al., 2022). Past literature emphasizes the importance of data-driven approaches, the incorporation of external factors, dynamic modeling, validation, and collaboration across departments (Pac, et al., 2009). Accurate headcount prediction enables organizations to optimize staffing levels, enhance resource allocation, and align workforce strategies with business objectives (Mundschenk & and Drexler, 2007).

2.3 Headcount Prediction and Demand Conditions

Predicting headcount requirements under varying demand conditions is a vital component of workforce management, impacting operational efficiency and cost-effectiveness. Forecasting strategies differ in certain and uncertain demand contexts. Thus, understanding the symbiotic relationship between headcount and demand is foundational as this relationship holds unique implications for manufacturing and service industries.

Demand conditions can vary widely based on factors like seasonality, economic cycles, market trends, and business expansion or contraction (Olya, et al., 2022). Accurate headcount

prediction is essential to ensure that an organization's workforce aligns with these dynamic demand conditions. Another major aspect to be considered under headcount prediction is the way of predicting under various demand conditions. In situations of relatively certain demand conditions, historical data analysis becomes a reliable tool for predicting headcount (Chien, et al., 2008). Time series analysis, regression models, and moving averages are commonly employed. These approaches identify trends, seasonality, and cyclic patterns to forecast staffing needs accurately (Chien, et al., 2010).

However, uncertain demand conditions pose a greater challenge. Traditional forecasting models may prove insufficient due to abrupt market changes, economic shifts, or unforeseen events. Thus, agile approaches are essential to accommodate fluctuations (Olya, et al., 2022). Scenario Analysis is one way to solve this issue. Developing multiple staffing scenarios based on different demand projections enables organizations to be prepared for a range of outcomes (Baier, et al., 2012). This flexibility helps balance workforce needs amidst uncertainty.

Advanced analytics and machine learning models excel at handling complex and uncertain data. These models can incorporate a wide array of variables, including economic indicators, social trends, and external factors, to provide nuanced predictions under volatile conditions (Pac, et al., 2009).

Given the volatility of modern markets, the ability to adjust workforce levels quickly is critical. Businesses must be prepared to redeploy resources or train staff to tackle changing demands effectively (Hertz, et al., 2010). Leveraging technologies such as Cloud-based platforms, real-time data integration, and predictive analytics tools enhances prediction accuracy and strategic decision-making (April, et al., 2006). In uncertain demand scenarios, businesses might explore flexible work arrangements, such as part-time or temporary employment, to manage staffing needs without compromising employee well-being (Hertz, et al., 2010). Regardless of demand certainty, prediction models must be regularly reviewed and updated to account for changing market dynamics, emerging trends, and unforeseen events.

Workload analysis is crucial for predicting headcount needs. By assessing the volume and complexity of tasks, organizations can determine the appropriate staffing levels needed to meet demand (Olya, et al., 2022). Workload analysis helps avoid overstaffing during lulls and understaffing during peak periods. In addition to predicting headcount numbers, skill mapping and role-based prediction are gaining traction (Drexel & Mundschenk, 2008). By identifying the skills and competencies required for different roles, organizations can ensure that they have the right talent in place to address diverse demand scenarios (Drexel & Mundschenk, 2008).

Hence, predicting headcount under various demand conditions is a multifaceted process that requires a combination of historical analysis, advanced analytics, scenario planning, and dynamic adjustments. Past literature emphasizes the importance of data-driven approaches, skill mapping, collaboration, and the integration of demand forecasting with workforce planning. Accurate predictions empower organizations to optimize staffing levels, ensure operational efficiency, and align their workforce strategies with the ever-changing demands of the business environment.

2.4 Headcount Optimization

Headcount optimization is a strategic imperative that transcends industries, influencing operational efficiency, cost-effectiveness, and overall business performance.

2.4.1 Evolution of Headcount Optimization

Over the years, headcount optimization strategies have evolved significantly, transitioning from simplistic approaches to data-driven and technology-enabled methodologies.

Early research focused on simple cost-cutting measures like downsizing and layoffs during economic downturns (Lu & Sturt, 2022). However, these practices often resulted in talent loss, decreased morale, and compromised productivity (Kawas, et al., 2013). With the advent of advanced analytics, researchers explored data-driven approaches. These involved analyzing historical workforce data, performance metrics, and market trends to inform headcount decisions (Kawas, et al., 2013). This shift towards informed decision-making laid the foundation for optimizing headcount based on actual business needs. In recent years, emerging technologies like machine learning and artificial intelligence have revolutionized headcount optimization. Advanced algorithms can process vast datasets, identifying intricate patterns, and suggesting optimal staffing levels to meet dynamic demand conditions (Kawas, et al., 2013).

2.4.2 Methods of Headcount Optimization

Researchers have devised various methodologies to optimize headcount, catering to different industries and organizational complexities. Statistical methods such as linear regression and time series analysis are such methods to identify correlations between historical data and headcount requirements (Kawas, et al., 2013). These models considered factors like seasonality, trends, and external variables. Workload-based optimization involves assessing the volume and complexity of tasks performed by each employee (Olya, et al., 2022). Researchers established benchmarks to allocate headcount according to workload, preventing overburdening of employees (Baier, et al., 2012). Researchers recognized the importance of scenario analysis, allowing organizations to simulate different workforce allocation scenarios. By considering

best-case, worst-case, and baseline scenarios, businesses can prepare for uncertain demand fluctuations (Ayough & Khorshidvand, 2019). Recent studies leveraged machine learning algorithms like decision trees, random forests, and neural networks. These models excel at handling vast and complex datasets, offering highly accurate predictions and optimal staffing recommendations (Baier, et al., 2012).

2.4.3 Challenges in Headcount Optimization

Researchers have highlighted several challenges in headcount optimization that need to be addressed for effective implementation. Accurate optimization relies on high-quality data (Lu & Sturt, 2022). Inaccurate or incomplete data can lead to flawed predictions and suboptimal headcount decisions (Lu & Sturt, 2022). Businesses operate in dynamic environments with ever-changing market conditions (Baier, et al., 2012). Incorporating real-time data and adapting optimization models to changing circumstances is essential. Balancing workforce optimization with employee well-being and organizational culture is a challenge (Peterson, et al., 2011). Drastic staff reductions can lead to reduced morale and talent loss. Different industries have unique requirements and challenges (Rodrigo & Ratnayake, 2021). For example, the service industry deals with customer-centric fluctuations (Baier, et al., 2012), while manufacturing contends with production cycles (Ayough & Khorshidvand, 2019).

Past research offers numerous real-world applications showcasing the effectiveness of headcount optimization strategies. Retailers have employed predictive analytics to optimize staffing during peak shopping seasons; ensuring customers receive exceptional service without excessive labour costs (Kawas, et al., 2013). Hospitals have utilized workload-based optimization to allocate nursing staff effectively, improving patient care and employee satisfaction (Olya, et al., 2022). Manufacturers have leveraged data-driven approaches to align workforce levels with production demands, preventing overstaffing or understaffing on assembly lines (Chien, et al., 2010). Technology companies have adopted machine learning algorithms to forecast workforce needs based on research cycles and technological advancements (Chien, et al., 2010) (Ayough & Khorshidvand, 2019).

Headcount optimization, informed by past research, has evolved into a complex and data-driven practice that requires a nuanced understanding of industry dynamics, technology integration, and employee considerations. Researchers have moved beyond simple cost-cutting approaches to harness the power of analytics and advanced algorithms for informed decision-making. The journey of headcount optimization continues to unfold, marked by a commitment to aligning workforce needs with business objectives, maximizing productivity, and fostering a harmonious work environment.

2.5 Predicting and Optimizing Headcount in Labour Intensive Manufacturing Industries

Labour-intensive manufacturing industries, characterized by heavy reliance on human labour, face intricate challenges in predicting and optimizing headcount to maintain operational efficiency (Baier, et al., 2012). While there is a substantial body of research on workforce management, several gaps exist in understanding the nuances of headcount prediction and optimization specific to labour-intensive manufacturing (Baier, et al., 2012).

A significant gap exists in research that tailors headcount prediction and optimization models specifically to the unique dynamics of labour-intensive manufacturing. Many existing studies focus on general workforce management strategies that may not effectively address the intricacies of industries like textiles, apparel, and assembly-line manufacturing (Ayough & Khorshidvand, 2019).

Labour-intensive industries often grapple with volatile and unpredictable demand patterns due to fashion trends, market shifts, and economic conditions (Lu & Sturt, 2022). The literature lacks comprehensive studies that develop models capable of accurately predicting headcount needs amidst such erratic demand fluctuations.

In sectors like apparel manufacturing, demand can vary greatly based on seasons and trends (Rodrigo & Ratnayake, 2021). However, there is limited research that delves into the development of robust forecasting models capable of accommodating seasonality, allowing companies to optimize headcount during peak and off-peak periods effectively (Mundschenk & and Drex1, 2007).

Headcount optimization is closely linked to supply chain dynamics, including raw material availability, production schedules, and distribution processes. Current literature often overlooks the integration of these factors into headcount prediction models, leaving a gap in understanding how to synchronize labour needs with broader manufacturing operations (Zhao, et al., 2022).

High turnover rates are a persistent challenge in labour-intensive manufacturing industries. Research addressing the impact of turnover on headcount prediction and optimization is scarce (Rodrigo & Ratnayake, 2021). There's a need to explore strategies for managing headcount fluctuations caused by employee attrition and recruitment efforts.

In the realm of labour-intensive manufacturing, the gaps in literature regarding headcount prediction and optimization pose substantial challenges. To bridge these gaps, future research should focus on developing industry-specific models that account for volatile demand, incorporate supply chain dynamics, address employee turnover, integrate sustainability goals,

and leverage emerging technologies. By addressing these gaps, researchers can contribute to more effective and tailored strategies for workforce management in labour-intensive manufacturing industries.

2.6 Critical Analysis of Related Work

Olya, et al., (2022) in their paper named “An integrated deep learning and stochastic optimization approach for resource management in team-based healthcare systems” address two critical healthcare team-based resource planning challenges through a combination of machine learning and stochastic optimization.

The first challenge pertains to measuring and predicting the required workload of patients. The second challenge focuses on resource planning and optimization within healthcare teams, including patient allocation, need satisfaction, and cost minimization (Olya, et al., 2022). The study introduces a novel integrated model that provides a systematic solution to predict healthcare providers' workload and balance their workloads, even when the required workload is unknown (Olya, et al., 2022). This model comprises predictive and prescriptive where the predictive phase involves using a deep multi-task learning approach to predict the workload for different patient types while the prediction serves as input for the prescriptive phase, which assigns patients to healthcare teams, determines the required team count, and balances team workloads (Olya, et al., 2022).

The research findings suggest that utilizing multi-task learning on represented data yields better predictions compared to conventional methods (Olya, et al., 2022).. Additionally, the proposed stochastic optimization model for resource planning indicates that considering randomness and stochastic variables in team-based resource allocation significantly reduces the total cost of healthcare operations (Olya, et al., 2022).

However, this study considers workload calculated using Relative Value Unit (RVU) of teams created and inputting this predicted data to arrive at the optimum team mix. Due to consideration of service provision nature and subjective to healthcare industry, employing the same model as it is in another industry might be questionable. Even though this could be alleviated to an extent due to the use of a deep learning model, employing multi-tasking model would have to be tested in other industrial contexts.

With the hope of providing a way to derive an optimum organizational hierarchy, M. Goubko and S. Mishin have proposed a normative model of optimal hierarchy design which drives higher organizational performance (Goubko & Mishin, 2009). As these researchers highlight, the effects of features such as employees' cost and effort, size of the firm, monitoring costs and

etc. on revenue have been considered in this model (Goubko & Mishin, 2009) to make it better than prior models which hadn't taken these features in to account. As per the results of the model, span of control, headcount, efforts distribution and wage differentials are identified as optimal hierarchy attributes, showing them as a function of exogenous parameters (Goubko & Mishin, 2009).

Further, M. Goubko and S. Mishin elaborate how these attributes permit the analysis of the impact the environment parameters have on a firm's size, financial results, employees' wages, and hierarchy shape. Moreover, the model provides answers for manager subordinate ratio of a company, cost of maintenance control system, ways in which the management expenses are increased with the expansion of a company and if requires a radical restructuring of the control system as well the changes required in an organizational structure in order to cater new management technologies, production modernization and standardization as well as environment changes.

However, this model is subject to imperfect and asymmetric information such as considering only efforts of employee for wage consideration (Goubko & Mishin, 2009). This suggests requiring advanced techniques for the model development in order to compensate for the imperfect information. In addition, the symmetric organization hierarchy is obtained assuming a common plan for all the employees. This might not be the scenario in a practical instance, since employ specific workload and/or different technology dependencies can affect the results (Goubko & Mishin, 2009). Thus, even if this model seems to provide satisfying results theoretically, practicality could be questionable.

Emphasizing the benefit of high accuracy levels as well as the less time taken for processing, April, J., Better, M., Glover, F., et al elaborates on a simulation optimizing model for headcount management decisions which employs a meta-heuristic approach (April, et al., 2006). The research explains how this model can be used in real life scenarios by illustrating two case studies where one is focusing on selecting the best headcount level for a Personal Claims Process at an Insurance Company (April, et al., 2006).

The selection of this optimum mix is done through SIMUL8 simulator and OptQuest for SIMUL8 optimizer (April, et al., 2006) through the iteration of results over and over again to identify the solution which gives highest throughput of insurance claim requests with the least cycle time of claim processing. The authors highlight that using a meta-heuristic approach along with advanced statistical analysis techniques can bring more accurate solutions using the least

computer time while ensuring the precision and clarity of solutions as well as the quality of process performance (April, et al., 2006).

However, the researchers do not elaborate on the reasons for using the above specific commercial simulation optimizing model nor do they explain the algorithms they have used for feature selection. Thus, it remains questionable if this model can be used for any kind of business scenario, especially in the manufacturing industries where processes are more complex compared to insurance policy handling.

Similarly, Ayough, A., & Khorshidvand, B. compare how effective meta-heuristic models in designing a Cellular Manufacturing Systems (CMS) for allocating workforce for manufacturing cells when there are uncertain demand conditions (Ayough & Khorshidvand, 2019). The methods compared are Simulated Annealing (SA) and Particle Swarm Optimization (PSO) algorithms which resulted in spotlighting PSO algorithm to be more satisfactory in the subject matter (Ayough & Khorshidvand, 2019). The issue of previous studies which assumed certain demand conditions for cell manufacturing has been addressed by this study since dynamic production times and uncertain demands are well considered in this model (Ayough & Khorshidvand, 2019).

However, this study does not necessarily consider the supply chain aspects of the manufacturing industry leaving a gap for further improvement through the introduction of more features. In addition, this model only considers product, its allocations to cells and workforce leading to assume some important aspects such as processing time and workload are not affecting factors.

The summary of the above critical analysis could be presented as follows in table 2.1:

Table 2-1: Summary of Critical Analysis of Related Research Work

| Authors | Aim of the Study | Methodology Used | Outcomes | Limitations |
|---------------------|--|---|--|--|
| Olya, et al. (2022) | Address healthcare team-based resource planning challenges using machine learning and stochastic optimization. | Integrated model with predictive and prescriptive phases. | - Better predictions of healthcare provider workload using multi-task learning. - Reduced total cost of healthcare operations through stochastic | - Reliance on Relative Value Unit (RVU) for workload calculation may limit applicability to other industries. - Multi-tasking model's applicability to |

| Authors | Aim of the Study | Methodology Used | Outcomes | Limitations |
|------------------------|--|--|--|--|
| | | | optimization in resource planning. | other contexts needs testing. |
| Goubko & Mishin (2009) | Propose a normative model of optimal hierarchy design to enhance organizational performance. | Normative model analysis considering various features' effects on revenue. | - Identification of optimal hierarchy attributes. - Analysis of environmental parameter impact on firm size, financial results, etc. | - Imperfect and asymmetric information might affect model accuracy. - Assumption of symmetric organization hierarchy may not align with practical scenarios. |

| Authors | Aim of the Study | Methodology Used | Outcomes | Limitations |
|------------------------------|---|--|---|---|
| April, et al. (2006) | Present a simulation optimizing model for headcount management decisions using a meta-heuristic approach. | Employ SIMUL8 simulator and OptQuest for SIMUL8 optimizer. | - More accurate solutions with less processing time. - Illustration of model's applicability through case studies. | - Lack of explanation for the selection of specific commercial simulation optimizing model and algorithms used for feature selection. - Uncertain suitability of the model for complex manufacturing scenarios. |
| Ayough & Khorshidvand (2019) | Compare meta-heuristic models for designing Cellular Manufacturing Systems (CMS) under uncertain demand conditions. | Comparison of Simulated Annealing (SA) and Particle Swarm Optimization (PSO) algorithms. | - Highlight PSO algorithm as more satisfactory for CMS design. - Addressing the issue of uncertain demand conditions in cell manufacturing. | - Lack of consideration for supply chain aspects in the manufacturing industry. - Limited focus on product allocation and workforce, neglecting factors like processing time and workload. |

2.7 Drawbacks and Limitations of Previous Research Work

The limitations and drawbacks identified during the critical analysis can be listed as follows.

- Variables used to predict the headcount being less applicable to other industries than on which it is used
- Less concern on applicability of the used model, i.e. Multi-tasking models, to other industries
- Not considering the impact of imperfect and asymmetric information on model accuracy

- Assumptions employed, i.e. symmetric organization hierarchy, not being aligned with practical scenarios.
- Lack of explanation for the selection of models employed for feature selection as well as prediction.
- Suitability of the model being uncertain for complex scenarios than that of ones used for modelling.
- Lack of consideration for main aspects in the domain, i.e. supply chain aspects in the manufacturing industry.

Considering the limitations outlined in the critical analysis above, the researcher intends to develop a more industry-specific and easily scalable model. This model aims to identify determinants even when they are initially unknown. Consequently, the gap prevailing in the literature regarding the availability of a headcount model for the apparel industry would be addressed upon the completion of this research.

CHAPTER 3

METHODOLOGY

This chapter outlines the planned methodology and approaches for conducting the research. It begins with discussing the study's design and proceeds to detail the steps involved in data collection and validation, feature selection for the model, the choice of machine learning algorithms for modeling, and the subsequent evaluation before and after deployment. The methodology is visually represented in figure 3.1 below, followed by a comprehensive explanation of each step.

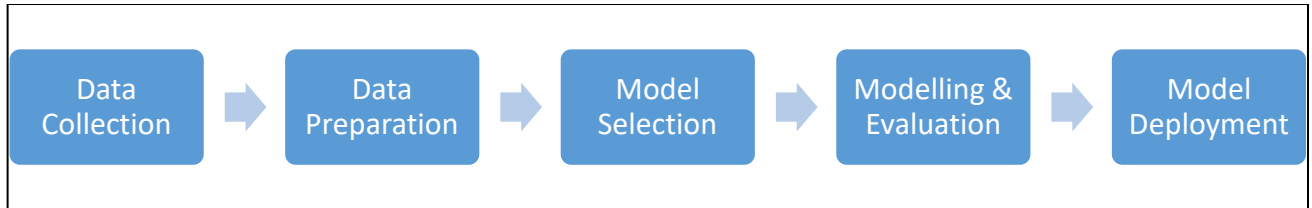


Figure 3.1: Research Methodology

3.1 Research Questions

The proposed model will be utilized to tackle the following two research questions:

RQ1: What are the primary features influencing the prediction of worker and staff cadre?

RQ2: Which machine learning model is the most appropriate for predicting the worker and staff cadre based on the identified features?

3.2 Data Collection

The data essential for this research is sourced from various databases within the Emjay-Penguin Group. Specifically, it will be extracted from the company's Enterprise Resource Planning (ERP) system and planning tool. The primary ERP system utilized is IFS, from which data encompassing invoice details, sales records, organizational structure, and associated headcount information will be retrieved. Additionally, customer-specific order details and their production planning are sourced from the FastReact planning tool. Furthermore, time-study data is obtained from manual Excel reports utilized within the company.

A five-year dataset forms the basis of this research, with a summary presented in the table below. Upon data collection, the initial dataset comprises raw attributes that require preprocessing and feature selection to enhance their suitability for analysis that can be detailed as follows in table 3.1.

Table 3.1: List of Available Features

| Attribute | Explanation | Values | Data Type |
|---------------------------------|---|-------------------------------------|----------------------|
| Applied Date | Date of the transaction | | Date - DD/MM/YYYY |
| Factory | Production factory | KGL, PAN, TAL, TLD | Char |
| Production Qty | Quantity produced on transaction date | | Integer |
| Part Type | Raw Material Part Type | Fabric, Trims | Char |
| Transaction Type | Tasks carried out in the department | Arrival, GRN , TT, Issuing | Char |
| Transaction Qty | Quantity processed under each transaction type on each Applied Date | | Float |
| Per Day Requirement Type | Categorization of daily requirement of Raw material transactions | Inhousing , TT, Issuing | Char |
| Per Day Requirement | Factory wise Average per day requirement of each raw material transaction | | Float |
| Current Worker Cadre | Existing No of employees on the applied date to provide physical effort in tasks such as loading, unloading, etc. No of Heads calculated based on Man Hours | | Float |
| Current Staff Cadre | No of existing employees on the applied date to carry out system data entering and other documentation related tasks. No of Heads calculated based on Man Hours | | Float |

| | | | |
|------------------------------|--|--|-------|
| Required Worker Cadre | Required number of employees on the applied date to provide physical effort in tasks such as loading, unloading, etc | | Float |
| Required Staff Cadre | Required number of employees on the applied date to carry out system data entering and other documentation related tasks | | Float |

The target variables of this headcount prediction model are "Required Worker Cadre" and "Required Staff Cadre." It's important to note that "headcount" in this context doesn't directly refer to the actual number of employees needed in the department. Instead, it represents the Full-Time Equivalent (FTE) of employees, which quantifies a head as the number of hours worked by a full-time employee over a specific time frame (Johansson, 2022).

In this research, FTE is defined as the number of man-hours worked by a full-time employee in a day, which equals eight hours. Therefore, if a predicted target value is 1.5 for an instance, it implies that the equivalent of one full-time employee's eight-hours plus an additional half of man hours of another full-time employee would be required.

The consent of the company was taken prior to the data extraction and a formal letter has been attached herewith in the appendix to the report.

3.3 Data Pre-processing

Data pre-processing, also referred to as enhancing data informativeness and significance, entails the transformation of data from one form to another that holds greater value and relevance (GeeksforGeeks, 2021). Leveraging machine learning methodologies, mathematical modeling, and statistical insights, this process can often be automated. The resultant output can manifest in various formats such as graphs, charts, tables, images, and more, contingent upon the specific task at hand and the requisites of the computational system (GeeksforGeeks, 2021).

The data utilized in this study is structured in a tabular format. Consequently, Microsoft Excel serves as the primary tool for preliminary data processing tasks such as conversion to Comma Separated Values (CSV) format to facilitate readability within Python. Furthermore, any non-numeric data, including symbols, undergoes transformation into numerical equivalents (e.g.,

dashes replaced by zeros), thereby ensuring uniformity and compatibility for subsequent analyses.

3.3.1 Data Reduction

Given the datasets' extensive attribute sets, correct selection of features is pivotal to yield optimal analytical outcomes. Accordingly, less significant attributes are identified and omitted to prioritize the utilization of the most pertinent features for predictive modeling.

3.3.2 Data Cleaning and handling missing values

Incomplete data stems from errors in data gathering or corruption during storage. Given that many machine learning algorithms are intolerant of missing values, strategies must be employed to address such instances to ensure the accuracy of analytical results.

For the purpose of this research, tuples containing multiple missing values are disregarded if their insignificance relative to the dataset is established. With the dataset comprising over 20,000 records, a threshold of up to 1% of tuples is deemed acceptable for exclusion due to missing value concerns. Beyond this threshold, alternative approaches are adopted. Categorical variables undergo imputation using the mode of the respective field, while numerical variables are imputed on a case-by-case basis.

3.3.3 Data Transformation

Data undergoes transformation into formats compatible with machine learning algorithms, necessitating various preprocessing steps. Data pivoting, mapping, encoding, normalization, and integration represent pivotal examples of such transformations. These processes are integral for ensuring that raw data is structured, standardized, and optimized to facilitate efficient model training and predictive accuracy.

3.3.4 Feature Selection

After data preprocessing, the next step involves selecting significant features to emphasize those most influential in prediction. This feature selection process aims to pinpoint the factors that drive changes in the target variable, facilitating deeper analysis.

Various statistical and machine learning techniques are employed for feature selection. For example, the Pearson correlation matrix, Gini index, Principal Component Analysis can be highlighted as means used for this purpose. Regardless of the specific approach, the overarching goal remains consistent: to identify and prioritize the key attributes that impact the target variable.

3.4 Machine Learning Models Selection

After preparing the dataset with machine-compatible structures through various preprocessing steps and conducting feature selection, the next task is to determine the most suitable machine learning model that yields accurate predictions based on the dataset. Initially, several machine learning models will be selected and trained. Subsequently, a comparative analysis will be conducted to ascertain the most appropriate model for addressing the problem at hand.

3.4.1 Model Architecture

The modeling process adheres to a specific architecture comprises two main stages:

1. Training: In this stage, the dataset is allocated for training the model to forecast the necessary worker and staff cadre within the company.
2. Testing: Following training, the dataset is reserved for testing the trained model's performance in predicting the required worker and staff cadre. It's crucial to ensure that the data preparation procedures applied to the training dataset are replicated identically in the test dataset to maintain prediction consistency.

At both training and testing stages, cross-validation would be carried out to validate the model. Rather than depending on the simple split of dataset into training and test based on 80:20 method, cross-validation would be selected since the full dataset would be considered for training and validation dividing it to several number of samples pre-defined (Pandian, 2023). Past researchers such as Olya, et al., (2022) has discussed about running the model considering different sample sets of the same dataset to articulate the model validation encouraging to utilize similar instance for the research.

For this research endeavor, a similar architecture will be adhered while incorporating additional evaluation metrics to ascertain prediction performance as shown in the figure 3.2 below.

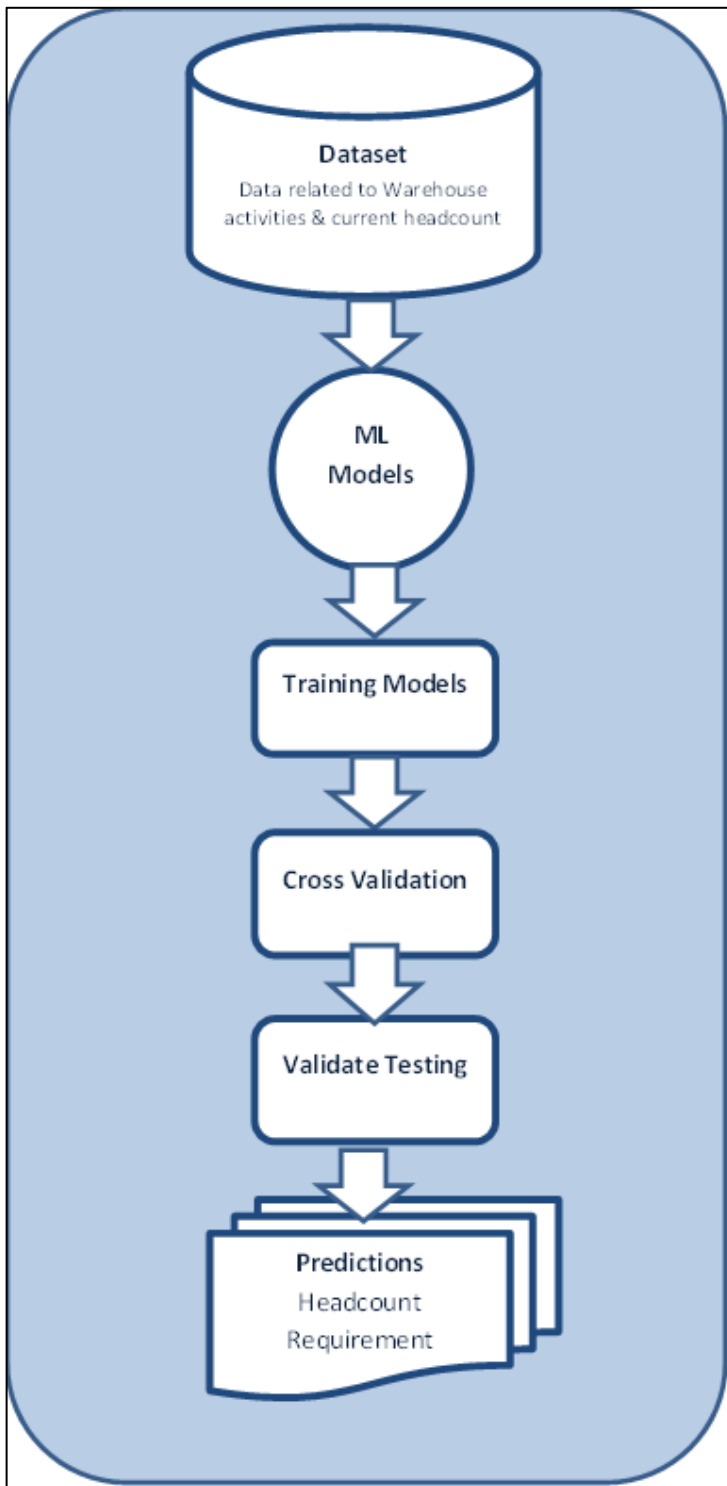


Figure 3.2:Proposed Architecture

As highlighted in the above architecture, a predictive analysis is conducted to determine the necessary workforce and staff levels for a given sales or production target.

3.4.2 Assumptions

Following assumptions were established during the development of the model to remove any potential ambiguity.

Assumption 1: Sales quantity of a day is equal to the Production quantity.

Despite potential variations in sales quantity in real-world scenarios, for the sake of simplifying the model, it is assumed that all items produced on a specific transaction date will be sold. Additionally, confirmation of this assumption was obtained due to the direct impact of warehouse tasks on production.

Assumption 2: Consider the capacity of employees as a constant value.

The maximum capacity of each employee is utilized, assuming they work at full capacity. Additionally, employee capacity is assumed to be unaffected by their performance, the quality of work carried out, or the required workload. The total available capacity for an employee is calculated using headcount and employee availability during a specific transaction period.

Assumption 3: Number of working hours per day is constant.

The number of hours available for a normal working day is set to be a standard of eight hours. Each employee is assumed to have the same eight hours for carrying out their daily tasks regardless of their employment category.

Assumption 4: Only one fabric type is used to produce one garment.

One garment will always comprise of only one fabrication such as Single Jersey and workload of an employee will not vary based on the type of fabrication.

Assumption 5: Required worker and staff headcount is correlated with the attributes of Per day production quantity.

Assumption 6: The required workload and per day requirement to be achieved through this workload by each employee is known. Thus, every employee's effort is an input of the model.

Assumption 7: Identical employees have the same efficiency. Thus, similar employees provide identical effort in terms of quality and productivity.

Assumption 8: Identical providers do collaborate. It is presumed that the assignment of a task to an employee is one to many allowing the same employee to carry out several tasks pertaining the quantity produced on a day.

Assumption 9: A cost will be incurred when employees are required to work more than their available capacity.

3.4.3 Model Selection

Considering the nature of the dataset and the aforementioned assumptions, a selection process would be conducted to determine suitable machine learning models. Given that the dataset focuses on predicting a continuous variable—specifically, the required headcount of workers and staff based on various input factors—the problem is classified as a regression task. Furthermore, since the goal is to predict two target variables, any regression model capable of facilitating multi-output prediction will be utilized.

In addition, a suffering index would be calculated to provide more insight into the decision-making process in terms of workforce allocation. In allocating the workforce to various units, fairness is a significant concern, regardless of whether the workforce is adequate or insufficient. This becomes especially crucial during periods of workforce shortages, as there is a desire to distribute the workload shortage evenly among units. The suffering index (SI) quantifies the extent of the workforce shortage by standardizing the current workforce level (Chien, et al., 2008). Even though there are many factors that could be considered for this calculation, As Chien, et al. (2008) devises, for the purpose of this study, suffering index can be calculated as follows:

$$\text{Suffering Index} = \frac{\text{Required Labour}}{\text{Available Labour}}$$

For the purpose of this research, the formula would be customized as follows:

$$\text{Suffering Index} = \frac{\text{Required Cadre}}{\text{Current Cadre}}$$

Consequently, the suffering index would be computed separately for each cadre type, offering a clearer insight into areas of workforce shortage or excess staffing.

3.4.4 Prediction Experiments

To achieve the research objectives, the following experiments will be conducted to determine the most appropriate prediction algorithm among the regression models identified in the model selection phase.

- **Prediction of headcount with original data**

Without carrying out feature selections, the original dataset would be used with all the attributes available to train and obtain headcount results.

- **Prediction of headcount with features selected with Principal Component Analysis (PCA) method**

Features would be selected through PCA to be used in headcount prediction.

- **Prediction of headcount with features selected with Permutation Feature Importance method**

Features would be selected through Permutation Feature Importance approach to be used in headcount prediction.

Each experiment will be executed for all the algorithms identified in the model selection section to determine which feature selection method yields superior performance across each algorithm used.

3.4.5 Evaluation

The outcomes of these experiments will be assessed based on the algorithmic performances of each model. In contrast to classification models, where metrics like Recall, Precision, or F1 Score are commonly used for evaluation, regression models require different measurements for meaningful assessment (Chicco et al., 2021). Hence, metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Coefficient of Determination (R-squared Score) will be utilized (Chicco et al., 2021).

As elaborated in table 3.2 below, lower values approaching zero indicate better performance for MSE and MAE, while a positive R-squared value closer to one signifies good performance (Chicco et al., 2021). Furthermore, MAPE values are interpreted as follows, as highlighted by Chicco et al. (2021) citing Lewis (1982):

Table 3-1: Interpretation of Typical MAPE Values

| MAPE | Interpretation |
|-------|-----------------|
| <10 | Highly Accurate |
| 10-20 | Good |
| 20-50 | Reasonable |
| 50< | Inaccurate |

Source: Lewis (1982)

For this research, all the aforementioned evaluation metrics would be employed to assess the accuracy of the algorithms used, aiming to compare and select the most suitable algorithm for the model.

The model evaluation will be conducted through k-fold cross-validation, inspired by the utilization of these metrics in literature, which has provided valuable insights into headcount planning as in Olya, et al., (2022).

3.5 Justification for the Methodology Selected

As noted in the Model Selection section, this study addresses a regression problem falling within the domain of supervised learning techniques. The rationale behind this choice stems from the availability of data labels for all instances within the dataset and the continuous nature of the target variable.

In Chapter 2, the Literature Review highlighted the utilization of various algorithms for predicting workload and headcount across diverse industries, ranging from healthcare services to semiconductor manufacturing. Among these methods, Multi-Task Learning and Linear Regression emerged with the most favorable outcomes in instances discussed during the Critical Analysis of Related Work. Additionally, Decision Tree algorithms, Random Forest, and Lasso Regression were employed for comparison purposes, offering insights into the selection of more suitable algorithms.

According to the Scikit Learn Machine Learning Cheat Sheet, datasets with under 100,000 samples can utilize algorithms such as Elastic Net and Ridge Regression (scikit-learn.org, n.d.) as shown below in the figure 3.3 below.

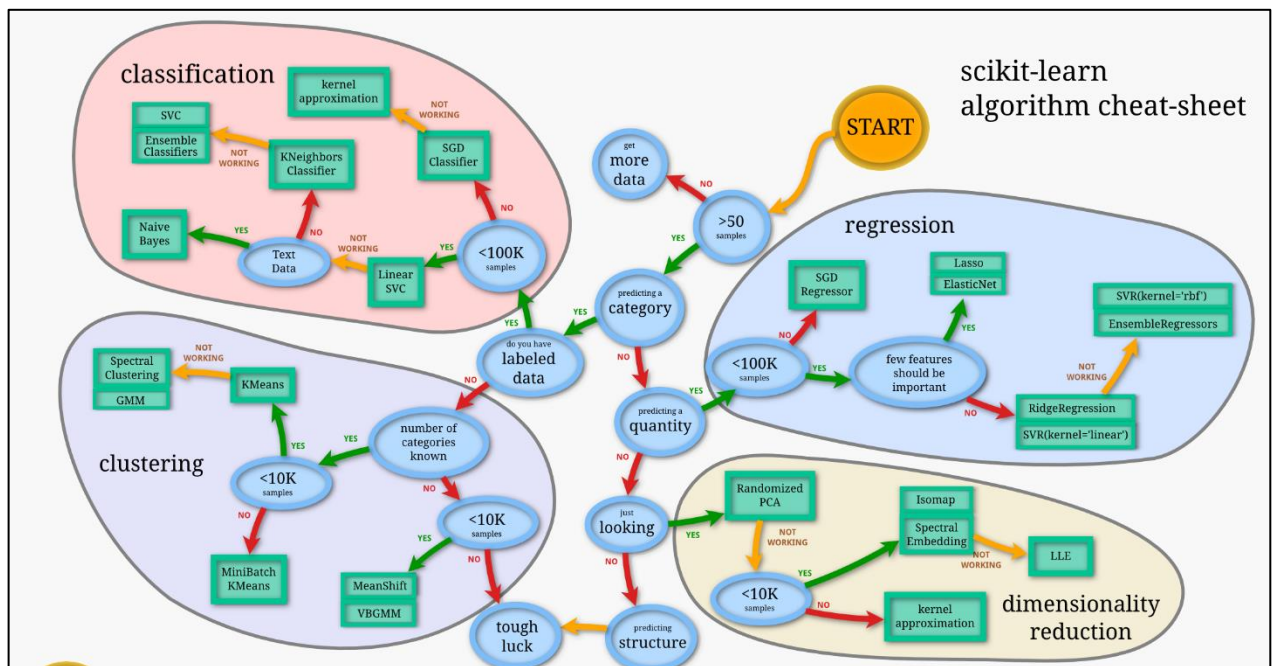


Figure 3.3: Scikit-learn Machine Learning Algorithm Cheat Sheet

The Gradient Boosting algorithm, categorized under supervised learning techniques, has garnered attention for its capability to surpass many other regression models (scikit-learn.org, n.d.). This is attributed to its methodology of combining a collection of weak individual models to generate a more precise final model (scikit-learn.org, n.d.). In line with this, the following algorithms were selected for comparison to identify the best algorithm for the model.

- Linear Regression
- Random Forest
- Decision Tree
- Lasso Regression
- Ridge Regression
- Gradient Boosting

Each of these algorithms were trained using Multi output regressor as there are two target variables which need to be predicted at once.

CHAPTER 4

EVALUATION AND RESULTS

This chapter portrays the execution of the case study by applying the methodology detailed in the preceding chapter. It encompasses the presentation and discussion of the results obtained from this implementation. Subsequently, the outcomes achieved from each devised experiment will be interpreted, accompanied by evaluation measures to validate the selection and implementation of the final model.

4.1 Implementation

The implementation of the research would be presented in line with the methodology to improve the understandability.

4.1.1 Data Preprocessing

As outlined in section 3.2, Data Processing, the dataset was transformed into Comma Separated Values (CSV) format to enhance its machine readability. Following the loading of the converted file, several preprocessing steps were undertaken to optimize the model's performance.

Data Reduction

Certain transaction data related to unissued raw materials, finished goods, and reversals of goods received and transferred were omitted from the analysis. The focus was narrowed to solely include the receipt, transfers between factories, and issuances of raw materials, with consideration given to the initial recording of production quantities. Additionally, the attribute 'Applied Date' was removed from the dataset due to its repetitive nature and lack of significance compared to other variables.

Data Cleaning and Missing Values

As the data extraction was done directly from the ERP system of Emjay Penguin Group, the presence of missing values was not directly identifiable. Upon investigation, it was observed that days without specific transaction types did not generate any data entries, resulting in naturally handled missing values at the data collection stage. Descriptive analysis, detailed later in this section, validates this observation. Furthermore, entries with factory codes 'EMI' or 'PSW' were removed from the dataset, as these locations represent transactions related to the group's head office rather than warehouse locations.

Data Transformation

To prepare the data for mining and machine learning algorithms, data transformation techniques were employed. Categorical variables were encoded into numerical format using data mapping, as most machine learning algorithms cannot directly process categorical variables.

To facilitate processing, values in attributes such as 'Factory,' 'RM Part Type,' 'Transaction Type,' and 'Per Day Requirement Type' were encoded from string to numeric type. This encoding involved a transcoding procedure to convert the "n" values of a class into numeric variables ranging from 0 to 1 where 1 is arrived at from $(n - 1)$.

For instance, the binary variable "RM Part Type" underwent the following transformation:

- The value "Fabric" was assigned to 0,
- The value "Trims" was assigned to 1.

Subsequently, a function was created for the 'RM Part Type' column to generate a new column with encoded part type values, enabling compatibility with machine learning algorithms. Hence, a new attribute labeled "Part Type" was introduced, featuring numerical values.

The "RM Factory" variable, comprising four values, underwent a similar transformation as shown below:

- "KGL" was mapped to 0,
- "PAN" was mapped to 1,
- "TAL" was mapped to 2,
- "TLD" was mapped to 3.

Following this approach, other categorical variables, namely 'RM Transaction Type' and 'RM Per Day Requirement Type,' were mapped to numerical values ranging from 0 to $(n - 1)$. Subsequently, new attributes were created, featuring the mapped numerical values for these columns.

Standardization

To mitigate potential issues arising from outliers, which could adversely affect the model's performance, data standardization was performed (Prabhu, 2020). This process encompassed standardizing all numerical attributes by subtracting the mean from attribute values and scaling them to unit variance. This was achieved utilizing the 'StandardScaler' function from the 'sklearn.preprocessing' library. Each attribute is scaled by computing the relevant statistics on the samples within the training dataset. Subsequently, the 'transform' function was employed to retain the calculated mean and standard deviation for application to future data.

Python codes related to the preprocessing are attached as an appendix to the report under Appendix A.

4.1.2 Data Exploration

Descriptive statistics of the entire dataset were constructed to understand the characteristics of all variables utilized in the headcount prediction model. This includes values such as count, minimum and maximum (min/max), unique count, mean, and standard deviation, quartiles, data type and availability of missing values as presented in table 4.1 below.

Table 4-1: Descriptive Statistics of the dataset

| Statistic | Production Qty | Transaction Qty Per Day | Per Day Requirement | Current Worker Carder | Current Staff Carder |
|-----------|----------------|-------------------------|---------------------|-----------------------|----------------------|
| count | 35800 | 35800 | 35800 | 35800 | 35800 |
| unique | 227 | 1784 | 6 | 23 | 4 |
| mean | 51103.44637 | 159237.9552 | 101020.8882 | 2.860697 | 0.664804 |
| std | 28469.71321 | 237746.3667 | 81455.3166 | 1.814158 | 0.230027 |
| min | 0 | 2146.792 | 20640 | 1 | 0.416667 |
| 25% | 33842 | 13257.386 | 21849.16 | 1.529412 | 0.416667 |
| 50% | 47492 | 46863.58 | 124883.954 | 2.352941 | 0.5 |
| 75% | 64820 | 223751.75 | 199456.36 | 3.529412 | 0.833333 |
| max | 186067 | 2071545.6 | 199456.36 | 8.823529 | 1 |
| dtype | int64 | float64 | float64 | float64 | float64 |
| missing | 0 | 0 | 0 | 0 | 0 |

Furthermore, as illustrated in figure 4.1 below, a correlation matrix heatmap was employed to assess the strength of relationships among numerical variables. This graphical representation showcases correlations between all attributes within the dataset. Certain variables exhibited high correlations, while others displayed weaker correlations. To ensure consistency, encoded categorical attributes were utilized for this correlation analysis.

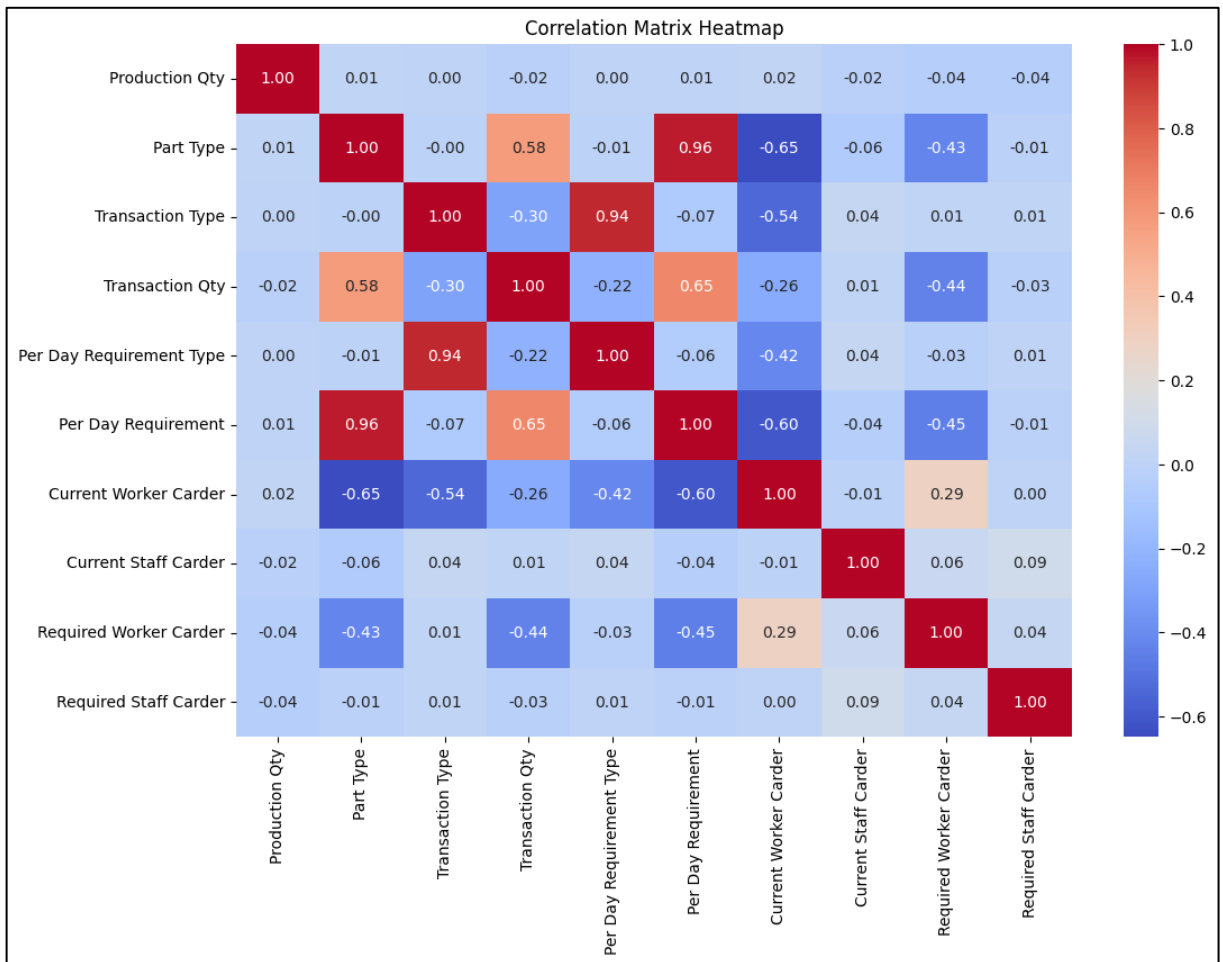


Figure 4.1: Correlation Matrix Heatmap

According to the analysis, features in the following table 4.2 demonstrated significant correlations with each other, based on the heatmap results. Positive correlations ranged from 0.6 to 1, while inverse correlations ranged from -1 to -0.6.

Table 4-2: Attributes with High Correlation

| Correlated Attributes | Correlation Type | Correlation Coefficient |
|--|------------------|-------------------------|
| Part Type Vs Current Worker Cadre | Negative | 0.65 |
| Part Type Vs Per Day Requirement | Positive | 0.96 |
| Transaction Type Vs Per Day Requirement Type | Positive | 0.94 |
| Per day Requirement Vs Current Worker Cadre | Negative | 0.60 |
| Per day Requirement Vs Transaction Qty | Positive | 0.65 |

Python codes related to the data exploration are attached as an appendix to the report under Appendix A.

4.1.3 Feature Selection and Model Selection

After preprocessing, feature selection aimed to pinpoint the most significant features for improving the model's performance. Given the inherent connection between feature selection and model selection, these steps were undertaken concurrently.

Three experiments detailed in the Prediction Experiment section of the methodology chapter were conducted. Initially, the selected algorithms were trained using all the attributes in the original dataset without any feature selection, and the model's performance was recorded. Subsequently, feature selection was initiated through Principal Component Analysis (PCA) in the second experiment. Model training then ensued using the PCA-selected features, followed by a comparison of model performance. In the final experiment, features were chosen via the Permutation Feature Selection method, and the results were utilized to train the algorithms for model selection through performance evaluation.

4.1.4 Model Building and User Interface Development

Based on the outcomes of each conducted experiment, the optimal model was determined, considering the algorithm with the highest performance. Model training entailed utilizing k-fold cross validation method for model validation.

The outcomes from each experiment were compared to determine the most effective feature selection method, yielding superior performance results under each trained algorithm, thereby guiding the selection of the optimal model. The evaluation metrics employed in each experiment remained consistent, encompassing Mean Squared Error (MSE), Mean Absolute Error (MAE), Coefficient of Determination (R-squared Score), and Mean Absolute Percentage Error (MAPE), as outlined in the Evaluation section of the Methodology chapter.

Subsequently, the chosen algorithm was employed to construct the model aimed at predicting the required worker and staff cadre within the warehouse department. This model was further expanded to compute the Suffering Index based on predicted and current cadre values, thereby indicating any shortage or excess in headcount for each prediction instance. The performance of the finalized model was evaluated using k-fold cross-validation method as detailed in the Model Architecture section of the Methodology chapter. The evaluated finalized model was then saved to the local disk using the 'Pickle' Python object.

Addressing one of the objectives outlined in the Objectives section of the Introduction chapter, a user-friendly model was developed to facilitate cadre requirement prediction by inputting new

values. Data input was anticipated to be streamlined through a user interface featuring prominent features identified during the best model selection process. The following prototype shown in figure 4.2 below serves as the ultimate output, facilitating the aforementioned requirement and aiding Human Resource professionals and Business Process Professionals in discerning headcount requirements necessary for achieving set targets at any given time, while providing additional insights for headcount reduction or reallocation decisions.

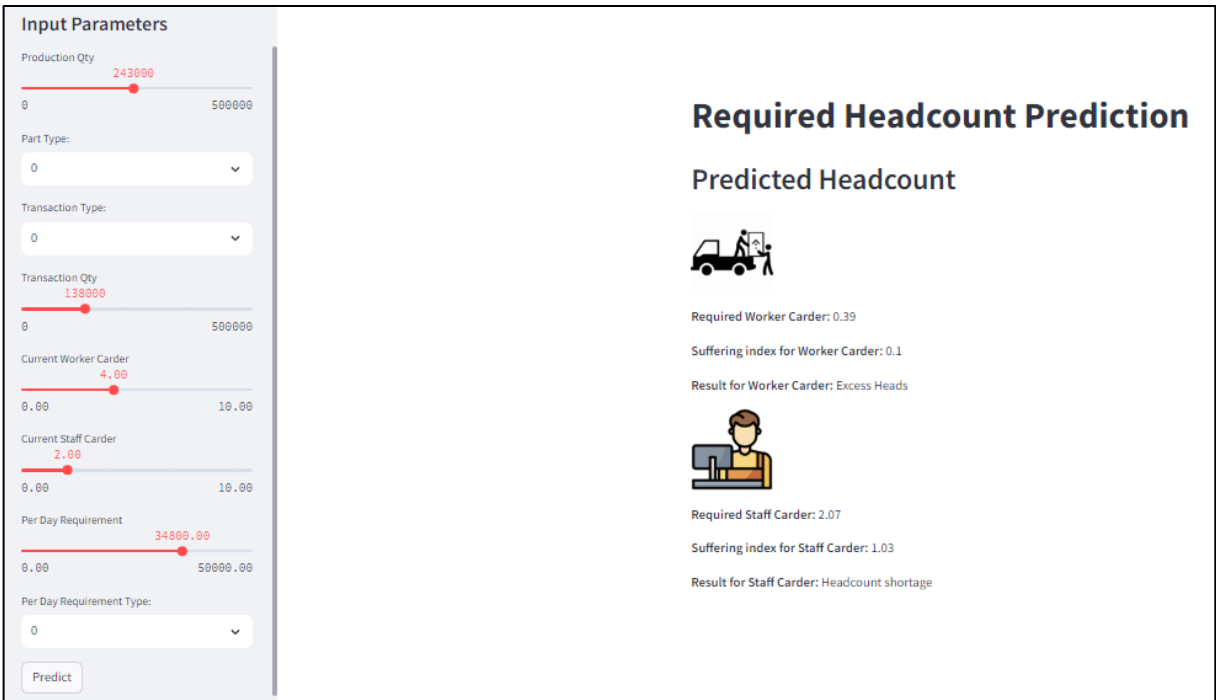


Figure 4.2: Proposed interface for Required Cadre Prediction

The interface was developed using the "Streamlit" library, which simplifies web app development (Streamlit, 2024). Predictions are generated based on the best model saved from this research, which focuses on predicting the required worker and staff cadre.

Attributes such as Part Type, Transaction Type, and Per Day Requirement Type were encoded to improve machine readability. However, the original categorical values are also displayed in the interface for user convenience, as users may be more accustomed to these values. For example, Part type values are presented as Fabric and Trims in the interface, despite being encoded as 0 and 1, respectively, to enhance user-friendliness.

After users input new data values for each parameter in the interface, predictions are generated to provide the required worker and staff cadre values separately. Additionally, the suffering index interpretation is provided, indicating whether there is a shortage or excess of headcount

considering both current and predicted values. This assists users in making data-driven decisions regarding headcount reduction or reallocation.

4.2 Results

The presentation of results aims to fulfill all research aims and objectives. Each experiment evaluated algorithm performance using R2 score, MSE, MAE, and MAPE, with detailed raw results as follows:

4.2.1 Experiment 1: Prediction of headcount with original data

Results shown in table 4.3 indicate Gradient Boosting as the top performer among algorithms, boasting the highest R2 score of 0.265, and the lowest MSE and MAE of 23.359 and 3.481, respectively. A low MAPE value of 32.792% underscores the prediction's fair accuracy compared to other algorithms.

Conversely, the Decision Tree regressor yielded a negative R2 score, indicating its inadequacy for the predictive modeling at hand. While Lasso and Ridge Regressors demonstrated similar outcomes, the Linear regressor showcased relatively better results, with the Random Forest regressor exhibiting the poorest performance among suitable algorithms.

Table 4-3: Comparison of results from Experiment 1

| Metric | Decision Tree | Gradient Boosting | Lasso Regression | Linear Regression | Random Forest | Ridge Regression |
|---------------------|---------------|-------------------|------------------|-------------------|---------------|------------------|
| R-squared Score | -0.761 | 0.265 | 0.262 | 0.214 | 0.177 | 0.262 |
| Mean Squared Error | 55.946 | 23.359 | 23.441 | 24.958 | 26.153 | 23.442 |
| Mean Absolute Error | 5.435 | 3.481 | 3.501 | 3.706 | 3.777 | 3.501 |
| MAPE % | 51.203 | 32.794 | 32.982 | 34.915 | 35.580 | 32.982 |

4.2.2 Experiment 2: Prediction of headcount with Principal Component Analysis

Results shown in table 4.4 indicate Gradient Boosting as the top performer again, achieving the highest R2 score of 0.288, along with the lowest MSE and MAE of 22.632 and 3.362, respectively. A low MAPE value of 31.673% reiterates the prediction's fair accuracy.

Conversely, the Decision Tree regressor exhibited a negative R2 score, indicating its unsuitability for the predictive modeling task. Both Linear and Lasso regressors displayed

comparatively better results, with Random Forest and Ridge regressors emerging as the weakest performers among suitable algorithms.

Table 4-4: Comparison of results from Experiment 2

| Metric | Decision Tree | Gradient Boosting | Lasso Regression | Linear Regression | Random Forest | Ridge Regression |
|---------------------|---------------|-------------------|------------------|-------------------|---------------|------------------|
| R-squared Score | -1.110 | 0.288 | 0.191 | 0.236 | 0.070 | 0.079 |
| Mean Squared Error | 67.043 | 22.632 | 25.704 | 24.259 | 29.562 | 29.254 |
| Mean Absolute Error | 5.500 | 3.362 | 3.757 | 3.559 | 3.928 | 4.035 |
| MAPE % | 51.818 | 31.673 | 35.395 | 33.532 | 37.003 | 38.014 |

4.2.3 Experiment 3: Prediction of headcount with Permutation Feature Importance

Results shown in figure 4.5 reveal Gradient Boosting as the standout performer, possessing the highest R2 score of 0.888, along with the lowest MSE and MAE of 2.530 and 0.6, respectively. A low MAPE value of 19.591% further confirms the prediction's good accuracy.

Notably, all models trained under permutation feature selection displayed suitability for prediction, with no negative R2 scores observed for any algorithm, unlike previous feature selection methods. Moreover, Random Forest, Decision Tree, Linear, and Ridge regressors demonstrated relatively better results, with Lasso regressors emerging as the weakest performer among algorithms.

Table 4-5: Comparison of results from Experiment 3

| Metric | Decision Tree | Gradient Boosting | Lasso Regression | Linear Regression | Random Forest | Ridge Regression |
|---------------------|---------------|-------------------|------------------|-------------------|---------------|------------------|
| R-squared Score | 0.811 | 0.888 | 0.106 | 0.257 | 0.874 | 0.257 |
| Mean Squared Error | 4.259 | 2.530 | 20.123 | 16.720 | 2.843 | 16.720 |
| Mean Absolute Error | 0.693 | 0.600 | 2.408 | 2.084 | 0.602 | 2.084 |
| MAPE % | 22.634 | 19.591 | 78.624 | 68.049 | 21.303 | 68.045 |

The features selected under this experiment were 'Production Qty,' 'Part Type,' 'Transaction Type,' 'Transaction Qty,' 'Per Day Requirement Type,' 'Per Day Requirement,' 'Current Worker Cadre,' and 'Current Staff Cadre' with importance values being greater than the threshold of

50%. Below table 4.6 illustrates the Importance value of each feature which guided the above feature selection.

Table 4-6: Feature Importance Values - Permutation feature importance

| Feature | Importance Value in % |
|--------------------------|-----------------------|
| Factory | 38% |
| Production Qty | 51% |
| Part Type | 51% |
| Transaction Type | 71% |
| Transaction Qty | 91% |
| Per Day Requirement Type | 66% |
| Per Day Requirement | 85% |
| Current Worker Carder | 68% |
| Current Staff Carder | 54% |

4.3 Evaluation

The evaluation of algorithms was conducted based on the results obtained from each experiment, along with the evaluation of the final model using the Cross-validation method.

4.3.1 Algorithm Performance

The results of each performance metric obtained for each algorithm under each experiment have been compared separately to provide insights for model selection, focusing on identifying the most suitable feature selection method to enhance overall model performance.

R-squared score

Coefficient of determination, if simply put, R-squared score articulates how well the target variable can be interpreted through the outlined input variables (Chicco, et al., 2021). As per the results obtained and showcased in table 4.7 below, Gradient Boosting algorithm consistently achieved the highest R-squared scores across all feature selection methods, with the permutation feature importance method yielding a notably higher value of 0.888.

Table 4-7: Algorithm Performance based on R^2 Score

| | Decision Tree | Gradient Boosting | Lasso Regression | Linear Regression | Random Forest | Ridge Regression |
|------------------------|---------------|-------------------|------------------|-------------------|---------------|------------------|
| Original Dataset | -0.761 | 0.265 | 0.262 | 0.214 | 0.177 | 0.262 |
| PCA | -1.110 | 0.288 | 0.191 | 0.236 | 0.070 | 0.079 |
| Permutation Importance | 0.811 | 0.888 | 0.106 | 0.257 | 0.874 | 0.257 |

Even the Decision Tree regressor showed improved results under permutation feature importance, whereas it was unsuitable for the predictive task under other methods.

Mean Squared Error

As per the results shown in below table 4.8, Mean Squared Error obtained under original dataset features and features obtained through PCA ranges from 20 to 67 whereas permutation feature importance indicates it to be within a range of 2 to 20 across all algorithms. Gradient Boosting consistently demonstrated the lowest MSE values across all feature selection methods, with a minimum MSE of 2.530 achieved under the permutation feature importance method.

Table 4-8: Algorithm Performance based on MSE

| | Decision Tree | Gradient Boosting | Lasso Regression | Linear Regression | Random Forest | Ridge Regression |
|-------------------------------|----------------------|--------------------------|-------------------------|--------------------------|----------------------|-------------------------|
| Original Dataset | 55.946 | 23.359 | 23.441 | 24.958 | 26.153 | 23.442 |
| PCA | 67.043 | 22.632 | 25.704 | 24.259 | 29.562 | 29.254 |
| Permutation Importance | 4.259 | 2.530 | 20.123 | 16.720 | 2.843 | 16.720 |

In contrast, Decision Tree and Random Forest Regressors exhibited comparatively higher error values under original dataset and PCA feature selection methods.

Mean Absolute Error

As shown in the table 4.9 below, out of all results obtained for MAE for the algorithms (shown below in the table), Gradient Boosting regressor consistently attained the lowest MAE values across all feature selection methods, with the permutation feature importance method yielding a MAE of 0.6, the lowest among all results.

Table 4-9: Algorithm Performance based on MAE

| | Decision Tree | Gradient Boosting | Lasso Regression | Linear Regression | Random Forest | Ridge Regression |
|-------------------------------|----------------------|--------------------------|-------------------------|--------------------------|----------------------|-------------------------|
| Original Dataset | 5.435 | 3.481 | 3.501 | 3.706 | 3.777 | 3.501 |
| PCA | 5.500 | 3.362 | 3.757 | 3.559 | 3.928 | 4.035 |
| Permutation Importance | 0.693 | 0.600 | 2.408 | 2.084 | 0.602 | 2.084 |

While Decision Tree and Random Forest regressors showed competitive performance under permutation importance, Gradient Boosting maintained a slight edge.

Mean Absolute Percentage Error (MAPE)

Based on the MAE calculated above, the percentage error has been obtained under MAPE and shown in the table 4.10 below. Accordingly, Gradient Boosting regressor consistently achieved better percentage error values compared to other algorithms. As detailed in the Evaluation section of the methodology chapter, all the algorithms have marked fair accuracy to inaccurate levels in predicting under all the feature selection instances while Gradient Boosting regressor has been able to secure its place in good accuracy level with a MAPE value of 19.591% falling under 10%-20%.

Table 4-10: Algorithm Performance based on MAPE

| | Decision Tree | Gradient Boosting | Lasso Regression | Linear Regression | Random Forest | Ridge Regression |
|-------------------------------|----------------------|--------------------------|-------------------------|--------------------------|----------------------|-------------------------|
| Original Dataset | 51.203 | 32.794 | 32.982 | 34.915 | 35.580 | 32.982 |
| PCA | 51.818 | 31.673 | 35.395 | 33.532 | 37.003 | 38.014 |
| Permutation Importance | 22.634 | 19.591 | 78.624 | 68.049 | 21.303 | 68.045 |

However, Lasso, Linear, and Ridge regressors exhibited inaccurate predictions, with MAPE values exceeding 50% under permutation importance.

When considering all the evaluation results it is evident that Gradient Boosting regressor has performed well with the permutation feature importance method comparative to the other algorithms in consideration.

The reason for permutation feature importance method providing better results that stands out could be due to the fact that it uses individual features at a time to see how it affects the target variable and iterating the same by shuffling features to identify the most important features (Jensen, 2022). Moreover, in a situation where there is a possibility of providing misleading information by permutation feature importance method when the variables are highly correlated (Jensen, 2022), as the correlation matrix coming in Data exploration section under Implementation part of this chapter outlines, the correlation between the features and the target variable does not reflect to be strong, receipt of better performance under this method could be further confirmed.

Overall, Gradient Boosting regressor performed well with the permutation feature importance method compared to other algorithms considered. This method's effectiveness may be attributed to its approach of evaluating individual features' impact on the target variable by iteratively shuffling features to identify the most important ones (Jensen, 2022). Additionally, considering

the correlation matrix presented in the Data exploration section, which indicated weak correlations between features and the target variable, the superior performance of the permutation feature importance method further validates its effectiveness (Jensen, 2022).

4.3.2 Model Evaluation

After determining the Gradient Boosting regressor as the most suitable algorithm for the headcount prediction model, the final predictive model was developed and saved as described in the Implementation section of this chapter. To assess the final model, cross-validation with k-folds was employed, alongside a separate validation set containing attributes selected through the permutation feature importance method.

Negative Mean Squared Error was utilized to assess the performance of each fold, while the Mean Cross-Validation score was employed to indicate the average performance of the model across all folds (Shaikh, 2018). In order to identify the best cross-validation score, implementation was under several iterations with different k values at each time. Starting with five folds, the resulting values of Mean Cross-Validation score are presented in figure 4.3 below.

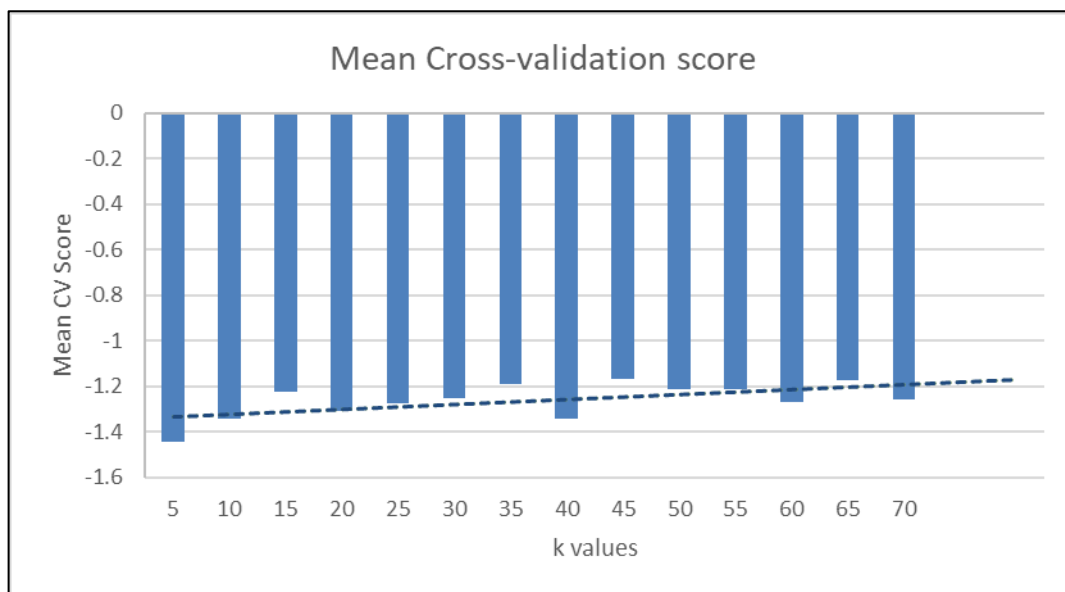


Figure 4.3: Mean Cross-Validation Score at Different k-values

The results indicate that when the k-values are increasing the Mean Cross-Validation score improves. The trendline shown in the figure above highlights this as the forecasted line heads further towards zero making it evident that the model is not overfitted. Hence, it provides more room for generalization of the model. Overall, the most common Mean Cross-Validation Score identifiable across different iterations lies around -1.26 suggesting a good

performance of the model as a whole, as it is not significantly distant from the zero threshold (Shaikh, 2018).

The codes and full results of the final model are presented in the Appendix to the report.

CHAPTER 5

CONCLUSION AND FUTURE WORK

The chapter encapsulates a summary of findings drawn from the attainment of the objectives and research questions outlined at the commencement of this study. Furthermore, it details the constraints encountered and suggests areas for enhancement to steer future research endeavors.

5.1 Conclusion

Labour constitutes a pivotal and costly resource globally, particularly in labour-intensive sectors where effective human resource management is paramount. The apparel industry, confronted with the repercussions of COVID-19 and widespread economic recession, has grappled with significant challenges in recent times.

Sri Lanka mirrors global trends, experiencing substantial revenue contractions amid unchanged cost structures, predominantly driven by labour expenses. Consequently, apparel manufacturers are endeavoring to strategize efficient headcount management, with direct labour management proving somewhat feasible due to its direct correlation with sales/production targets. However, managing indirect cadre poses a considerable challenge, often resulting in ad hoc decisions due to the absence of apparent links to set targets. The sensitivity surrounding human resources further complicates decisions pertaining to headcount reductions. Therefore, identifying a robust mechanism to ascertain the correct cadre requirements for indirect labour emerges as a pressing issue necessitating solutions.

Emjay Penguin, a leading garment manufacturer, embarked on a quest for innovative approaches to ascertain the indirect worker and staff cadre requirements for its warehouse department to meet set targets, aligning with its digital transformation journey. Consequently, this research sought to introduce a data-driven approach to headcount-related decision-making.

A model was developed, selecting the optimal algorithm from six popular regressors trained to provide multi outputs using a Multioutput regressor. Model selection hinged upon the outcomes of three experiments conducted based on the performance exhibited under different feature selection methods across each algorithm. Performance evaluation relied on R2 score, MSE, MAE, and MAPE, while the final model underwent assessment through cross-validation alongside the aforementioned evaluation matrices. Additionally, a suffering index was computed to furnish precise insights into headcount reduction or reallocation, revealing shortages or surpluses in current headcount.

This research was initiated to address two research questions and achieve specific aims and objectives. The research questions were addressed as follows:

RQ1: What are the primary features influencing the prediction of worker and staff cadre?

Three feature selection techniques were employed to identify the best method, with permutation feature importance emerging as superior, indicating 'Production Qty,' 'Part Type,' 'Transaction Type,' 'Transaction Qty,' 'Per Day Requirement Type,' 'Per Day Requirement,' 'Current Worker Cadre,' and 'Current Staff Cadre' as pivotal features for headcount prediction while being the most common features to be important under each feature selection method.

RQ2: Which machine learning model is the most appropriate for predicting the worker and staff cadre based on the identified features?

With the performance evaluation done for all the algorithms utilized for model selection using R^2 score, MSE, MAE and MAPE matrices, Gradient Boosting Regressor showcased the highest results for all matrices under all feature selection models while articulating the most prominent scores under permutation feature importance method accounting for R-squared score, MSE, MAE and MAPE values of 0.888, 2.530, 0.600 and 19.591% respectively.

Through comprehensive performance evaluation utilizing R-squared score, MSE, MAE, and MAPE matrices, the Gradient Boosting Regressor emerged as the optimal choice, exhibiting superior performance under all feature selection models, particularly excelling under permutation feature importance accounting for R-Squared score, MSE, MAE and MAPE values of 0.888, 2.530, 0.600 and 19.591% respectively.

Thus, it could be concluded that the research questions set forth in the inception of the research were addressed with solid answers.

In addition to the research questions, several objectives were outlined for accomplishment upon research completion:

- Identify department specific significant features that affect headcount prediction.
- Predict the optimum no of Required worker cadre to achieve a given sale/production quantity
- Predict the optimum no of Required Staff cadre to achieve a given sale/production quantity
- Uncover the headcount reduction /reallocation potential

While RQ1 addresses the identification of significant features, RQ2 answers the subsequent objectives by devising the final model for predicting worker and staff cadre requirements through Gradient Boosting regressor. The suffering index computed at the end of the model

unveils headcount reduction or reallocation potential considering predicted and current cadre values that clearly specifies under each instance whether the current cadre available is short to meet the production requirements or is in excess allowing room for reallocation and, thereby fulfilling the final objective.

In conclusion, the research has effectively achieved its aims and objectives through the developed headcount model, providing a mechanism to predict the required worker and staff cadre mix for the warehouse department. Additionally, it sheds light on significant features influencing cadre prediction and facilitates informed decisions regarding headcount reduction or reallocation. Ultimately, the research endeavors to achieve profitability by reducing the labour cost through informed headcount decisions made using the devised model.

In summary, the outcomes of this research underscore the importance of factoring in elements that connect outcomes with predetermined sales or production targets when determining headcount requirements within an organization, as this significantly influences the cost structure. Advanced technologies such as machine learning models, as demonstrated in this study, offer a means to predict these headcount requirements with greater precision and reliance on data, thereby providing valuable insights for decisions regarding headcount reduction or reallocation.

5.2 Future Work

In this section, potential avenues for future research endeavors are explored.

Given that this research delves into the realms of human resource management and machine learning, there exists an opportunity to expand this model to industries that encompass indirect staff not directly correlated with sales achievements. The model's flexibility allows for easy inclusion or exclusion of attributes, facilitating its extension to encompass multiple departments or an entire company. For instance, within the apparel industry, the model could incorporate factors like specific fabrication or trim type to predict warehouse department staff requirements, or extend to departments such as finance or IT, which lack direct links to sales targets but play supportive roles.

Moreover, future research could investigate headcount prediction in scenarios where workload data is unavailable. In this research, workload was inferred from employee transactions, but in situations where workload isn't provided, the model would need enhancements to predict workload employees for a particular period based on factors like cycle time.

Another area of interest is incorporating employee-specific behavioral factors into the model. The current model overlooks behavioral aspects such as work quality or employee skill levels, as well as cognitive factors like employee satisfaction or burnout. Addressing these factors could offer valuable insights and align with existing literature highlighting their significance (Olya, et al., 2022).

APPENDICES

Appendix A: Screen printing of codes – Model Training and selection

A.1 Importing useful libraries and loading data

#Import Libraries

```
#Importing Libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.ensemble import GradientBoostingRegressor, RandomForestRegressor
from sklearn.linear_model import LinearRegression, Lasso, Ridge
from sklearn.tree import DecisionTreeRegressor
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import cross_val_score
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from sklearn.multioutput import MultiOutputRegressor
from sklearn.inspection import permutation_importance
from sklearn.decomposition import PCA
```

#Load Dataset

```
# Load the dataset
data = pd.read_csv('/content/Transaction type wise full dataset.csv')
```

A.2 Preprocessing

#Drop irrelevant attribute columns

```
# Drop irrelevant columns
data.drop(['Applied Date'], axis=1, inplace=True)
```

#Encoding Categorical Variable into Numeric

```
# Get unique values for RM Part Type, RM Transaction Type, and RM Per Day Requirement Type
part_type_values = data['RM Part Type'].unique()
transaction_type_values = data['RM Transaction Type'].unique()
per_day_requirement_type_values = data['RM Per Day Requirement Type'].unique()

# Encoding variables
encoded_labels = {
    'Part Type': {'Fabric': 0, 'Trims': 1},
    'Transaction Type': {'Arrival': 0, 'GRN': 1, 'TT': 2, 'Issuing': 3},
    'Per Day Requirement Type': {'Inhousing': 0, 'TT': 1, 'Issuing': 2}
}

# Mapping of string labels to integer values
part_type_mapping = {label: value for value, label in enumerate(part_type_values)}
transaction_type_mapping = {label: value for value, label in enumerate(transaction_type_values)}
per_day_requirement_type_mapping = {label: value for value, label in enumerate(per_day_requirement_type_values)}
```

#Descriptive Statistics and Correlation Matix

```
# Descriptive Statistics
descriptive_stats = data.describe().transpose()

# Correlation Matrix Visualization
corr_matrix = data.corr()
plt.figure(figsize=(12, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix Heatmap')
plt.show()
```

#Split data into features and target

```
# Split data into features (X) and target variable (y)
X = data.drop(columns=['Required Worker Carder', 'Required Staff Carder'])
y = data[['Required Worker Carder', 'Required Staff Carder']]
```

#Standardization

```
# Standardize the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

A.3 Training model

#Model training and Cross-validation

```
# Define models
models = {
    "Decision Tree": MultiOutputRegressor(DecisionTreeRegressor()),
    "Gradient Boosting": MultiOutputRegressor(GradientBoostingRegressor()),
    "Random Forest": MultiOutputRegressor(RandomForestRegressor()),
    "Linear Regression": MultiOutputRegressor(LinearRegression()),
    "Lasso Regression": MultiOutputRegressor(Lasso()),
    "Ridge Regression": MultiOutputRegressor(Ridge())
}

# Train the models
for name, model in models.items():
    model.fit(X_scaled, y)
```

```
# Perform cross-validation
cv_scores = cross_val_score(model, X_scaled, y, cv=10, scoring='neg_mean_squared_error')

# Print cross-validation scores
print("Cross-Validation Scores (negative mean squared error):", cv_scores)
print("Mean Cross-Validation Score:", np.mean(cv_scores))
```

#Feature selection

```
# Calculate permutation feature importance
perm_importance = permutation_importance(model, X_scaled, y, n_repeats=10, random_state=42)

# Get feature importance scores and standard deviations
feature_importance = perm_importance.importances_mean
feature_importance_std = perm_importance.importances_std

# Select features based on permutation importance
threshold = np.mean(feature_importance) # Set a threshold
selected_features = X.columns[feature_importance > threshold]
print(selected_features)
```

#Define Suffering index

```
# Define the suffering index calculation function
def calculate_suffering_index(prediction, current_value):
    suffering_index = prediction / current_value
    return suffering_index
```


A.4 Prediction and Decision support

#Prediction

```
# Define the prediction function
def predict(Production_Qty, Part_Type, Transaction_Type, Transaction_Qty,
            Current_Worker_Carder, Current_Staff_Carder,
            Per_Day_Requirement, Per_Day_Requirement_Type, model_name):

    # Select the model based on the chosen name
    model = models[model_name]

    # Create a dataframe with the user inputs
    user_inputs = pd.DataFrame({
        'Production Qty': [Production_Qty],
        'Part Type': [Part_Type],
        'Transaction Type': [Transaction_Type],
        'Transaction Qty': [Transaction_Qty],
        'Current Worker Carder': [Current_Worker_Carder],
        'Current Staff Carder': [Current_Staff_Carder],
        'Per Day Requirement': [Per_Day_Requirement],
        'Per Day Requirement Type': [Per_Day_Requirement_Type]
    }, columns=feature_names) # Ensure the columns are in the same order as during training

    # Standardize the user data using the same scaler
    user_data_scaled = scaler.transform(user_inputs)

    # Make predictions
    prediction = model.predict(user_data_scaled)
```

#Calculate Suffering Index

```
# Calculate suffering indices
worker_suffering_index = calculate_suffering_index(prediction[0][0], Current_Worker_Carder)
staff_suffering_index = calculate_suffering_index(prediction[0][1], Current_Staff_Carder)

# Determine if there is a headcount shortage or excessive heads
worker_result = "Headcount shortage" if worker_suffering_index > 1 else "Excessive Heads"
staff_result = "Headcount shortage" if staff_suffering_index > 1 else "Excessive Heads"

print("Model:", model_name)
print("Predicted Required Worker Carder:", prediction[0][0])
print("Predicted Required Staff Carder:", prediction[0][1])
print("Suffering index for Worker Carder:", worker_suffering_index)
print("Suffering index for Staff Carder:", staff_suffering_index)
print("Result for Worker Carder:", worker_result)
print("Result for Staff Carder:", staff_result)
```

A.5 Model Evaluation

#Evaluating the models

```
# Dictionary to store evaluation metrics for each model
evaluation_results = {'Model': [], 'Mean Squared Error': [], 'R-squared Score': [], 'Mean Absolute Error': [], 'MAPE': []}
```

```
# Evaluate the model
mse = mean_squared_error(y, y_pred)
mae = mean_absolute_error(y, y_pred)
r2 = r2_score(y, y_pred)
mape = 100 * (mae / y.mean().mean())
```

```
# Store evaluation results
evaluation_results['Model'].append(name)
evaluation_results['R-squared Score'].append(r2)
evaluation_results['Mean Squared Error'].append(mse)
evaluation_results['Mean Absolute Error'].append(mae)
evaluation_results['MAPE'].append(mape)

# Convert evaluation_results dictionary to DataFrame
evaluation_df = pd.DataFrame(evaluation_results)

print(evaluation_df)
```

Appendix B: Screen printing of codes – Final Model

Only the different codes from the model selection part are shown.

#Import necessary libraries

```
from ipywidgets import interact, widgets
from sklearn.model_selection import cross_val_score
```

#Load validation dataset

```
# Load the dataset
data = pd.read_csv('/content/Transaction type wise validationset.csv')
```

#Cross Validation

```
# Train the Gradient Boosting model
model = MultiOutputRegressor(GradientBoostingRegressor())
model.fit(X_scaled, y) # Fit the model before using it

# Perform cross-validation
cv_scores = cross_val_score(model, X_scaled, y, cv=10, scoring='neg_mean_squared_error')

# Print cross-validation scores
print("Cross-Validation Scores (negative mean squared error):", cv_scores)
print("Mean Cross-Validation Score:", np.mean(cv_scores))
```

#Get the feature names

```
# Get the feature names
feature_names = X.columns.tolist()
print(feature_names)
```

#Create temporary interactive interface

```
# Define the interactive user interface
interact(predict,
          Production_Qty=widgets.IntSlider(min=0, max=500000, step=1000, value=5000),
          Part_Type=widgets.Dropdown(options=part_type_values, value=part_type_values[0], description='Part Type:'),
          Transaction_Type=widgets.Dropdown(options=transaction_type_values, value=transaction_type_values[0], description='Transaction Type:'),
          Transaction_Qty=widgets.IntSlider(min=0, max=500000, step=1000, value=5000),
          Current_Worker_Carder=widgets.FloatSlider(min=0, max=10, step=0.1, value=0),
          Current_Staff_Carder=widgets.FloatSlider(min=0, max=10, step=0.1, value=0),
          Per_Day_Requirement=widgets.FloatSlider(min=0, max=50000, step=100, value=1000),
          Per_Day_Requirement_Type=widgets.Dropdown(options=per_day_requirement_type_values, value=per_day_requirement_type_values[0], description='Per Day Requirement Type:'));
```

#Cross-validation results

-under 5folds

```
Cross-Validation Scores (negative mean squared error): [-0.21113101 -2.66695365 -0.28930383 -1.40847228 -2.64884483]
Mean Cross-Validation Score: -1.4449411220053585
```

-under 10folds

```
Cross-Validation Scores (negative mean squared error): [-0.2324232 -0.16644098 -0.31276824 -5.06294652 -0.11454711 -0.13095125
-2.20000664 -0.44066568 -1.92211247 -2.83804554]
Mean Cross-Validation Score: -1.3420907624929046
```

-under 15folds

```
Cross-Validation Scores (negative mean squared error): [-0.08450555 -0.46263004 -0.21522722 -0.36664222 -0.31840003 -7.20000736
-0.12623152 -0.52200006 -0.94826725 -1.31189966 -0.97014769 -0.16600175
-1.75253696 -0.61324185 -3.31568999]
Mean Cross-Validation Score: -1.224895276724535
```

-under 20folds

```
Cross-Validation Scores (negative mean squared error): [-0.07940199 -0.48939155 -0.06503549 -0.24539636 -0.52424664 -0.10806817
-0.39065589 -9.68253165 -0.12885564 -0.08957093 -0.17410342 -0.12342916
-3.14879319 -0.61275035 -0.67232403 -0.14250834 -0.31589704 -2.35701113
-4.49762802 -2.30909711]
Mean Cross-Validation Score: -1.30783480528857
```

-under 25folds

```
Cross-Validation Scores (negative mean squared error): [ -0.06277499 -0.0546404 -0.39191201 -0.32453419 -0.16672558
-0.39618908 -0.07737208 -0.08918853 -12.54204257 -0.0827168
-0.12931156 -0.08215069 -0.60765052 -0.13740819 -1.67320028
-1.89818667 -0.96589779 -0.96105773 -0.14932819 -0.16781593
-0.31082513 -3.85527275 -0.46463667 -5.82789989 -0.51663471]
Mean Cross-Validation Score: -1.27741491718785
```

-under 30folds

```
Cross-Validation Scores (negative mean squared error): [ -0.07893439 -0.06571669 -0.5904797 -0.06481145 -0.32530735
-0.14023016 -0.60483831 -0.06343953 -0.14028956 -0.57293353
-13.9886052 -0.07081932 -0.13135439 -0.0992439 -0.78802576
-0.18360952 -0.09860926 -2.12258043 -0.84010355 -2.72227036
-0.23502595 -0.91770182 -0.16704263 -0.14521245 -0.27086552
-3.61429597 -0.50425728 -0.35988109 -7.44171118 -0.20768164]
Mean Cross-Validation Score: -1.251862595157468
```

-under 35folds

```
Cross-Validation Scores (negative mean squared error): [ -0.08863106 -0.04044092 -0.0575472 -0.60016795 -0.08179186
-0.44090166 -0.07420337 -0.51113543 -0.10284398 -0.1236922
-0.05409125 -0.84231322 -16.47626158 -0.06446149 -0.13256103
-0.08311716 -0.07125621 -0.09998184 -0.17343326 -0.18170491
-2.17006495 -2.28666911 -0.43835422 -1.02109416 -1.03995371
-0.21250404 -0.11615973 -0.24613736 -0.2755062 -1.63537778
-2.9331744 -0.4774592 -0.76435445 -7.58841011 -0.19028805]
Mean Cross-Validation Score: -1.191315573699749
```

-under 40folds

```
Cross-Validation Scores (negative mean squared error): [ -0.09611178 -0.03552779 -0.07428862 -0.71405543 -0.05734431
-0.09164109 -0.35466843 -0.17700207 -0.43139452 -0.12270404
-0.07197775 -0.1220994 -0.7123808 -3.56241346 -18.61001795
-0.06344 -0.16959768 -0.06805972 -0.08739557 -1.29404798
-0.16647475 -0.1061122 -0.22096562 -2.60716095 -0.88109484
-2.25578328 -1.21834929 -0.2446134 -1.14750598 -0.16838036
-0.17027111 -0.22616627 -0.20932605 -0.81009274 -3.94752252
-0.21482383 -0.45887475 -8.60893861 -2.88883013 -0.25238328]
Mean Cross-Validation Score: -1.342995958980934
```

-under 45folds

```
Cross-Validation Scores (negative mean squared error): [ -0.05448405 -0.08980878 -0.08495109 -0.02300893 -0.74798202
-0.04820895 -0.08268049 -0.41391 -0.08326426 -0.46369179
-0.09029135 -0.05570673 -0.16456409 -0.03540543 -0.82837175
-0.66487004 -21.55051414 -0.03865659 -0.05344345 -0.18409661
-0.06486533 -0.06855107 -0.18199899 -0.20032671 -0.11020493
-0.24907869 -2.62205405 -0.59415777 -2.9696911 -0.12698359
-1.04617829 -0.45795272 -1.15008613 -0.17779128 -0.17216458
-0.14902928 -0.37589356 -0.12815982 -4.47117326 -0.53957706
-0.52071796 -0.35983554 -9.44538252 -0.26056148 -0.2606828 ]
Mean Cross-Validation Score: -1.1658002005185413
```

-under 50folds

```
Cross-Validation Scores (negative mean squared error): [ -0.06048779 -0.0831517 -0.08600251 -0.03184232 -0.73032299
-0.0391822 -0.08712666 -0.53189043 -0.05559198 -0.1883135
-0.49323122 -0.12637474 -0.0614708 -0.12105356 -0.05778174
-0.39138551 -0.62730972 -23.29413915 -0.12578746 -0.05544339
-0.1324128 -0.12710232 -0.06848755 -0.07165523 -0.04717173
-0.231574 -0.1338138 -0.2454809 -0.04250206 -3.42381995
-1.25068639 -2.71627247 -1.08199343 -0.27927479 -0.26976236
-1.36429222 -0.20780504 -0.10431772 -0.12094932 -0.30227773
-0.22107149 -0.11800386 -5.84780432 -0.47883147 -0.27308698
-0.49208272 -0.75772199 -11.78303791 -0.78911598 -0.31885568]
Mean Cross-Validation Score: -1.2109830717082495
```

-under 55folds

```
Cross-Validation Scores (negative mean squared error): [-5.44184729e-02 -8.91026931e-02 -4.27007000e-02 -8.90457956e-02
-2.10503679e-02 -1.17501822e+00 -3.73641634e-02 -8.13398311e-02
-5.72807540e-01 -6.52438655e-02 -3.15810752e-01 -5.52659928e-01
-1.45649825e-01 -5.75107875e-02 -6.02504264e-02 -2.17804140e-01
-6.71188712e-02 -9.59145722e-01 -1.71209268e+00 -2.54184100e+01
-4.49528370e-02 -5.51006794e-02 -2.04767530e-01 -8.68598308e-02
-6.10182361e-02 -8.05289156e-02 -6.58630776e-02 -1.21016916e-01
-2.12092464e-01 -5.62425553e-02 -2.86313309e-01 -3.40996448e+00
-9.51600652e-02 -1.70359182e+00 -3.53733237e+00 -1.20635755e+00
-1.93666629e-01 -2.74077788e-01 -1.71819756e+00 -2.06391483e-01
-1.67361859e-01 -1.38203779e-01 -1.36027025e-01 -3.65390861e-01
-1.77910413e-01 -2.19286749e-01 -2.68161500e+00 -4.82505646e+00
-2.07036178e-01 -5.70154729e-01 -2.05898935e-01 -5.67915108e+00
-5.43973636e+00 -7.66481741e-02 -3.85926652e-01]
Mean Cross-Validation Score: -1.2114444568623406
```

-under 60folds

```
Cross-Validation Scores (negative mean squared error): [-6.60591578e-02 -9.90563965e-02 -5.15223185e-02 -9.26788852e-02
-3.13186590e-02 -1.18781175e+00 -4.52109428e-02 -1.07710268e-01
-7.18318897e-02 -5.00044794e-01 -7.39392070e-02 -1.51147702e-01
-5.32239590e-01 -1.53206456e-01 -4.64287663e-02 -6.52814939e-02
-2.05092374e-01 -2.28928938e-02 -5.89866072e-02 -1.21987065e+00
-5.49527164e+00 -2.77654918e+01 -5.12241308e-02 -6.76801041e-02
-1.61657128e-01 -1.37243948e-01 -6.26242794e-02 -9.30719647e-02
-8.20865565e-02 -5.15061532e-02 -2.25475184e-01 -1.74755548e-01
-3.73705524e-02 -3.10878092e-01 -3.62786353e-02 -3.64915800e+00
-4.41013890e-01 -3.59544201e+00 -5.82705708e-02 -1.76231163e+00
-2.79524769e-01 -1.60592544e+00 -1.13095275e-01 -2.13352793e-01
-1.52145325e-01 -1.29441327e-01 -1.31140100e-01 -3.04050111e-01
-1.96672039e-01 -1.28543315e-01 -1.11185929e+00 -5.26859340e+00
-6.62855066e-01 -2.77415526e-01 -6.02060262e-01 -3.89471518e-01
-1.23791281e+01 -2.50549785e+00 -1.16514896e-01 -4.39436362e-01]
Mean Cross-Validation Score: -1.2674644241672224
```

-under 65folds

```
Cross-Validation Scores (negative mean squared error): [-6.50715482e-02 -7.68753301e-02 -4.19470789e-02 -8.99155011e-02
-2.55045259e-02 -2.43840031e-02 -1.54400171e+00 -4.38476346e-02
-9.23105414e-02 -6.36268848e-02 -5.64869977e-01 -3.72666310e-02
-1.04129117e-01 -1.16499335e+00 -5.20774049e-02 -1.63430997e-01
-5.08111819e-02 -1.05624091e-01 -1.70666782e-01 -1.73040871e-02
-7.70943182e-02 -1.20563554e+00 -2.85140256e-01 -3.09576688e+01
-1.20648558e-01 -5.87310344e-02 -5.11326605e-02 -1.67973593e-01
-1.27067013e-01 -6.61647991e-02 -8.66172380e-02 -8.62272368e-02
-4.41155201e-02 -1.53760362e-01 -1.64244389e-01 -1.63329038e-01
-3.36879718e-02 -3.43428344e-01 -3.66899304e+00 -6.88704467e-01
-5.74558039e-01 -4.62855701e+00 -5.74362507e-02 -1.43039481e+00
-2.97276273e-01 -1.74966926e-01 -1.79376249e+00 -2.15830218e-01
-1.28876832e-01 -1.23944557e-01 -1.45063432e-01 -1.48515612e-01
-3.35374140e-01 -1.28899562e-01 -7.33608476e-01 -5.77640417e+00
-7.12352616e-01 -2.37408335e-01 -6.38597647e-01 -2.03584400e-01
-8.50037940e-01 -1.28571045e+01 -4.47637486e-01 -4.10238223e-02
-4.74869215e-01]
Mean Cross-Validation Score: -1.1723862657428388
```

-under 70folds

```
Cross-Validation Scores (negative mean squared error): [-6.50715482e-02 -7.68753301e-02 -4.19470789e-02 -8.99155011e-02
-2.55045259e-02 -2.43840031e-02 -1.54400171e+00 -5.42582233e-02
-9.23105414e-02 -6.36268848e-02 -5.64869977e-01 -3.72666310e-02
-1.03380727e-01 -1.23845313e+00 -5.20774049e-02 -1.68783237e-01
-5.08111819e-02 -9.14463856e-02 -1.70666782e-01 -1.73040871e-02
-7.70943182e-02 -1.38875148e+00 -6.50960847e-02 -3.48197191e+01
-1.39900287e-01 -3.24707989e-02 -8.82111150e-02 -4.08365455e-02
-2.22397868e-01 -7.24744146e-02 -7.38879358e-02 -7.66043306e-02
-9.13140028e-02 -1.67532777e-01 -6.70080336e-02 -2.54398607e-01
-1.69810250e-01 -7.21722500e-02 -4.92357602e-02 -3.49500835e-01
-4.26265701e-02 -4.41531790e+00 -8.06578046e-02 -4.36624761e+00
-6.45260897e-01 -5.95605219e-02 -1.65758837e+00 -3.98668498e-01
-1.82711042e-01 -1.95456212e+00 -1.28498157e-01 -3.11766461e-01
-1.51208577e-01 -7.82023424e-02 -1.55885785e-01 -3.00583720e-01
-2.45996880e-01 -1.76883533e-01 -2.20237829e-01 -2.47928172e+00
-5.06934503e+00 -7.47594504e-01 -3.66632564e-01 -5.86843258e-01
-2.29118913e-01 -1.25211416e+00 -1.78872914e+01 -6.44649453e-01
-4.10238223e-02 -4.74869215e-01]
Mean Cross-Validation Score: -1.2605957192860728
```

#Final model predictions using validation set and evaluation results

Production...

Part Type:

Transactio...

Transactio...

Current_W...

Current_St...

Per_Day_...

Per Day R...

Predicted Required Worker Carder: 1.3030321378167602
Predicted Required Staff Carder: 0.8804676465661836
Suffering index for Worker Carder: 0.32575803445419005
Suffering index for Staff Carder: 0.4402338232830918
Result for Worker Carder: Excess Heads
Result for Staff Carder: Excess Heads
Mean Squared Error: 0.17021386739343464
Mean Absolute Error: 0.22275784874364213
R-squared: 0.9235459590254929
Mean Absolute Error Percentage (MAPE): 10.796558772418349

#Saving the best model

```
import pickle
```

```
# Save the model to a file using pickle
with open('final_model.pkl', 'wb') as f:
    pickle.dump(model, f)
```


Appendix C: Consent Letter



30th March 2023

Post Graduate Division,
University of Colombo School of Computing

Dear Sir/Madam,

Consent to Use Company Data for the Research Project

We are pleased to provide our consent to Ms. Nilusha Ariyasena, a Master of Business Analytics student at the University of Colombo - School of Computing, to use the data from our organization, Emjay International & Penguin Sportswear (Pvt) Ltd, for her research project.

We understand that her research project aims to analyze the key factors that impact the headcount reduction in apparel industry, and to develop a novel machine learning-based model for predicting optimum headcount which helps achieving a target sale more accurately. We believe that her research would have valuable insights and contribute to the development of more efficient and effective practices in the apparel manufacturing process as well strategic decision making.

As such, we hereby grant her permission to access and use our company data for the purpose of her research project. We trust that she will treat our data with the utmost confidentiality and professionalism, and that her research findings will be used for academic purposes only.

Please do not hesitate to contact us should you have any further inquiries or concerns. We wish her every success in her research endeavors.

Thank you.

Sincerely,

EMJAY INTERNATIONAL (PVT) LTD


.....
POSHITHA BATUWATTA
Chief Financial Officer

Poshitha Batuwatta,
Chief Financial Officer
Emjay International & Penguin Sportswear (Pvt) Ltd

Penguin Sportswear (Pvt) Ltd

Office: 341/5, M & M Centre, Kotte Road, Rajagiriya, Sri Lanka
Tel: +94 (0) 11 4409600 / 11 2887850 Fax: +94 (0) 11 4409663 Email: info@penguinsl.com Web: <http://www.emjayi.com>

Factories: Pallethalawinna, Katugastota, Kandy, Sri Lanka Tel: +94 (0) 81 4484900 Fax: +94 (0) 81 2499581
Appalabeddahena, Panwila, Kandy, Sri Lanka Tel: +94 (0) 81 4484500 Fax: +94 (0) 81 2472117

REFERENCES

- April, J. et al., 2006. *Enhancing Business Process Management With Simulation Optimization*. s.l., s.n., pp. 642-649.
- Ardiyono, S., 2022. Covid-19 pandemic, firms' responses, and unemployment in the ASEAN-5. *Economic Analysis and Policy*, Volume 76, p. 337–372.
- Ayough, A. & Khorshidvand, B., 2019. Designing a Manufacturing Cell System by Assigning Workforce. *Journal of Industrial Engineering and Management*, 12(1), pp. 13-26.
- Baier, M. et al., 2012. Sales-force performance analytics and optimization. *IBM Journal of Research and Development*, 56(6), pp. 8-1.
- Chicco, D., Warrens, M., Jurman, G. & , 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, Volume 7, p. e623.
- Chien, C.-F., Chen, W.-C. & Hsu, S.-C., 2008. *An indirect workforce (re) allocation model for semiconductor manufacturing*. s.l., IEEE, pp. 2201-2208.
- Chien, C.-F., Chen, W.-C. & Hsu, S.-C., 2010. Requirement estimation for indirect workforce allocation in semiconductor manufacturing. *International Journal of Production Research*, 48(23), p. 6959–6976.
- Drexler, A. & Mundschenk, M., 2008. Long-term staffing based on qualification profiles. *Mathematical Methods of Operations Research*, Volume 68, pp. 21-47.
- Emjay International, 2023. *Emjay International*. [Online]
Available at: <http://emjayi.com/>
[Accessed 13 06 2023].
- Goubko, M. & Mishin, S., 2009. Optimal Hierarchies in Firms: a Theoretical Model. *Contributions to Game Theory and Management*, Volume 2, p. 124–136.
- Gröger, C., Niedermann, F. & Mitschang, B., 2012. *Data Mining-driven Manufacturing Process Optimization*. London, U.K, WCE 2012.
- Hertz, A., Lahrichi, N. & Widmer, M., 2010. A flexible MILP model for multiple-shift workforce planning under annualized hours. *Eur. J. Oper. Res.*, Volume 200, pp. 860-873.
- Jamal, W. & Saif, M. I., 2011. Impact of Human Capital Management on Organizational Performance. *European Journal of Economics, Finance and Administrative Sciences* , 5(34), pp. 13309-13315.
- Jensen, T., 2022. *Feature Importance for Any Model using Permutation*. [Online]
Available at: https://medium.com/@T_Jen/feature-importance-for-any-model-using-permutation-7997b7287aa
[Accessed 24 02 2024].
- Johansson, J., 2022. *FTE vs. headcount: which model is best for you?*. [Online]
Available at: <https://resourceguruapp.com/blog/headcount-vs-fte-whats-the-best-resource-management-model-for-you#:~:text=FTE%20vs.->

.headcount%20calculation,56%20(50%20%2B%206).

[Accessed 24 02 2024].

Kawas, B., Squillante, M. S., Subramanian, D. & Varshney, K. R., 2013. *Prescriptive Analytics for Allocating Sales Teams to Opportunities*. s.l., IEEE.

Li, M., 2022. Research on the Impact of the Epidemic on Marketing. *Advances in Economics, Business and Management Research*, Volume 648, pp. 65-70.

Lu, H. & Sturt, B., 2022. *On the Sparsity of Optimal Linear Decision Rules in Robust Inventory Management*, s.l.: arXiv preprint arXiv:2203.10661.

Mundschenk, M. & and Drexl, A., 2007. Work force planning in the printing industry. *International Journal of Production Research*, 45(20), pp. 4849-4872.

Olya, M. H., Badri, H., Teimoori, S. & Yang, K., 2022. An integrated deep learning and stochastic optimization approach for resource management in team-based healthcare systems. *Expert Systems with Applications*, Volume 187.

Pac, M. F., Alp, O. & Tan, T., 2009. Integrated workforce capacity and inventory management under labour supply uncertainty. *International Journal of Production Research*, 47(15), pp. 4281-4304.

Peterson, S. J. et al., 2011. Psychological Capital and Employee Performance: A Latent Growth Modeling Approach. *Personnel psychology*, 64(2), pp. 427-450.

Prabhu, T. N., 2020. *Normalization vs Standardization, which one is better*. [Online] Available at: <https://towardsdatascience.com/normalization-vs-standardization-which-one-is-better-f29e043a57eb#:~:text=If%20your%20dataset%20has%20extremely,values%20into%20a%20small%20range.> [Accessed 24 02 2024].

Rodrigo, D. S. & Ratnayake, G. S., 2021. *Employee Turnover Prediction System: With Special Reference to Apparel Industry in Sri Lanka*. Maharashtra, India, s.n., pp. 1-9.

scikit-learn.org, n.d. *Choosing the right estimator*. [Online] Available at: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html [Accessed 24 02 2024].

scikit-learn.org, n.d. *Gradient Boosting Regression*. [Online] Available at: https://scikit-learn.org/stable/auto_examples/ensemble/plot_gradient_boosting_regression.html#sphx-glr-auto-examples-ensemble-plot-gradient-boosting-regression-py [Accessed 24 02 2024].

Shaikh, R., 2018. *Cross Validation Explained: Evaluating estimator performance: Improve your ML model using cross validation*. [Online] Available at: <https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85> [Accessed 24 02 2024].

Sjödin, D., Parida, V., Leksell, M. & Petrovic, A., 2018. Smart Factory Implementation and Process. *Research-Technology Management*, 61(5), pp. 22-31.

Streamlit, 2024. *Streamlit Documentation*. [Online]
Available at: <https://docs.streamlit.io/>
[Accessed 24 02 2024].

Trevor, C. & Nyberg, A., 2008. Keeping your headcount when all about you are losing theirs: Downsizing, Voluntary Turnover Rates and the Moderating Role of HR Practices. *Academy of Management Journal*, 58(2), p. 259–276.

Wright, P., Gardner, T. & Moynihan, L., 2003. The impact of HR practices on the performance of business units. *Human Resource Management Journal*, 13(3), pp. 21-36.

Wright, P., Gardner, T., Moynihan, L. & Allen, M., 2005. *The Relationship Between HR Practices and Firm Performance: Examining Causal Order*, Ithaca, NY 14853-3901: Cornell University, School of Industrial and Labour Relations.

Zhao, Y. et al., 2022. Manpower forecasting models in the construction industry: a systematic review. *Engineering, Construction and Architectural Management*, 29(8), pp. 3137-3156.