Data Analysis
Project - 2

ADVANCED ANALYSIS OF PRODUCTIVITY OF GARMENT WORKERS

Group 11

- Samudika Wanasinghe - 16075
- Meedum Keerthisiri - 16220
- Pasindu Madusanka - 16333

Abstract

This report presents an advanced analysis utilizing machine learning algorithms to predict the productivity of garment industry employees. Built upon insights from a prior descriptive analysis, this study develops predictive models using a comprehensive garment workforce productivity dataset sourced from Kaggle. Key findings from both descriptive and advanced analyses offer valuable insights for management, HR professionals, investors, and researchers in the garment industry. The results contribute to improved decision-making, resource allocation, and a deeper understanding of the factors influencing workforce productivity.

Content

List of Figures

List of Tables

Introduction

The garment industry is a vital part of the global economy, involving the design, production, and distribution of clothing to meet consumer demand. With its fast-paced nature and evolving trends, maintaining efficiency is essential for sustained growth. This industry relies heavily on labor and technology, making workforce productivity a crucial factor in overall performance. Understanding the factors that influence productivity can help businesses improve operations, reduce inefficiencies, and maximize output. This study analyzes garment employee productivity using data-driven methods to identify key patterns and provide insights that can support better decision-making and workforce management strategies.

Description of the question

The objective of this study is to develop a reliable predictive model to estimate garment employee productivity. In the previous descriptive analysis, two distinct clusters were identified based on departments. Given this finding, predictive models will be built separately for each department to improve accuracy.

By identifying key productivity factors, this model will help decision-makers enhance planning, resource allocation, and operational efficiency. The insights gained will support management, HR professionals, researchers, and investors in making data-driven decisions to optimize workforce performance.

Description of the data set

The “Productivity Prediction of Garment Employees” dataset, available on Kaggle, provides detailed information on factors affecting employee productivity in the garment manufacturing sector. The dataset covers the period from January 1, 2015, to March 4, 2015, and includes 1,197 observations with 15 attributes.

Attribute	Description	Type of variable
date	Date in MM-DD-YYYY	Categorical-Ordinal
Quarter	A portion of the month, where each month is divided into four quarters	Categorical-Ordinal
department	Associated department with the instance	Categorical-Nominal
Day	Day of the Week	Categorical-Ordinal
Team	Associated team number with the instance	Categorical-Nominal
Targeted_productivity	Targeted productivity set by the Authority for each team for each day.	Numerical-Continuous
Smv	Standard Minute Value, it is the allocated time for a task	Numerical-Continuous
Wip	Work in progress. Includes the number of unfinished items for products	Numerical-Discrete
Over_time	Represents the amount of overtime by each team in minutes	Numerical-Discrete
Incentive	Represents the amount of financial incentive (in BDT) that enables or motivates a particular course of action	Numerical-Continuous
Idle_time	The amount of time when the production was interrupted due to several reasons	Numerical-Continuous
Idle_men	The number of workers who were idle due to production interruption	Numerical-Discrete
No_of_style_changes	Number of changes in the style of a particular product	Numerical-Discrete
No_of_workers	Number of workers in each team	Numerical-Discrete
actucal_productivity	The actual percentage of productivity that was delivered by the workers. It ranges from 0-1	Numerical-Continuous

Table 1: Attribute Information

Important results of the Descriptive analysis

In a garment factory, two departments shape production: sewing—where creativity meets fabric, and finishing—where precision ensures perfection. Understanding the different workflows, we carried out factor analysis for mixed data and found that two clusters were created based on department classification. With this in mind, we provided insights to achieve our goals based on these departments.

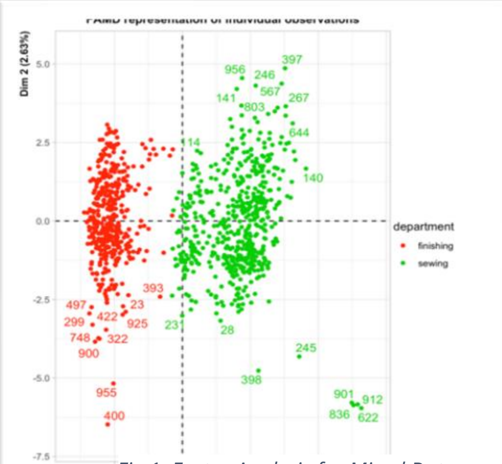


Fig 1: Factor Analysis for Mixed Data

Fig 2: Boxplot of Actual productivity vs Department

Despite higher median productivity, finishing faces inconsistencies as it was seen with its higher variation in boxplot, with 40% of its work deemed unproductive. We addressed this inconsistency under 03 sections. The main discrepancy lies in motivation. Over 80% of sewing workers receive incentives, fostering a reward driven and engaged environment. In contrast, finishing has little to no incentive structure, which might lead to decreased morale and inconsistent performance.

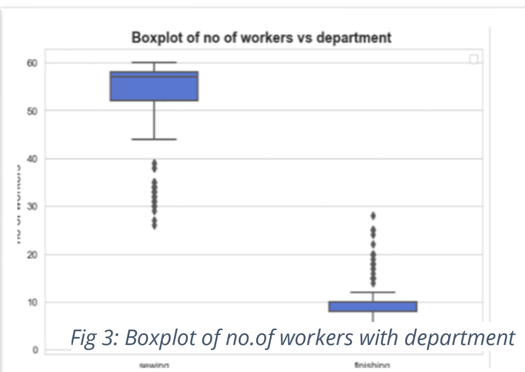


Fig 3: Boxplot of no.of workers with department

Workflow differences further widen the gap. Sewing always had work-in-progress (WIP), ensuring continuous engagement and momentum. On the other hand, finishing often experiences idle periods, disrupting efficiency and mastery over tasks.

The nature of work and workforce size also play crucial roles. Sewing involves creative, evolving tasks that keep workers engaged. Meanwhile, finishing is repetitive and handled by a much smaller team (8 vs. 57), increasing individual pressure and leading to instability in productivity.

To address these issues, under the Data Project 1, we suggested performance-based reward systems, goal charts so that every worker feels worthwhile. Optimizing workflows by introducing WIP and balancing workloads will create a smoother process in the finishing department.

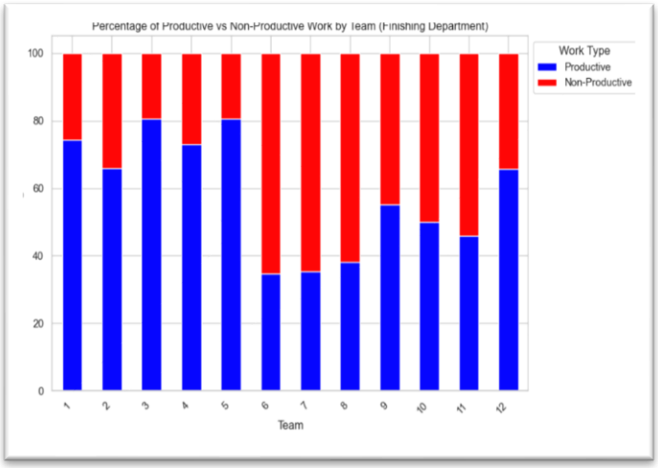


Fig 4: Productivity with teams - Finishing Department

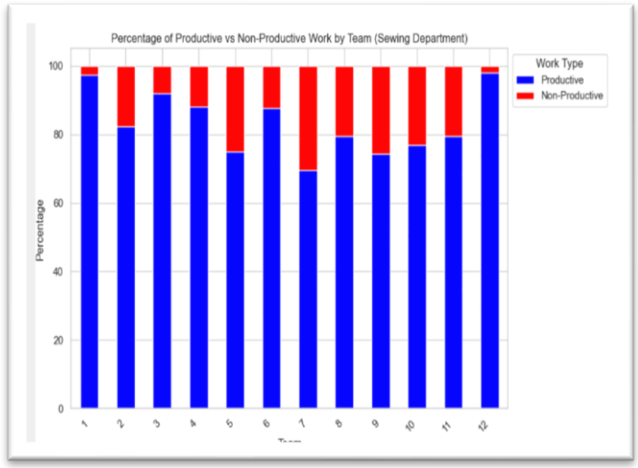


Fig 5: Productivity with teams - Sewing Department

To Address high pressure and instability, we first identified teams struggling with finishing. Teams 1, 3, and 5 performed well, while Teams 6, 7, and 8 had lower productivity. Despite being a top performer, Team 5 had the lowest median targeted productivity. Based on this, we recommend reallocating some of the workload from Teams 6, 7, and 8 to Team 5. Further investigation into underperformance causes and necessary training should be conducted.

Similarly, in the sewing department, Teams 1, 3, and 12 were top performers, while Teams 5, 7, and 9 underperformed. Team 5, in particular, struggled despite having low performance targets. Special attention should be given to Team 5 to improve efficiency through targeted interventions.

<u>Important results of the Advanced analysis</u>

Since multicollinearity was detected among several variables, fitting a standard multiple linear regression model was not suitable. To address this issue and improve model performance, advanced analytical techniques will be applied, including:

- Ridge Regression: Handles multicollinearity, no feature selection.
- Lasso Regression: Performs feature selection, may struggle with multicollinearity.
- Elastic Net Regression: Combines benefits of Ridge and Lasso.

For outliers and non-linearity:

- Regression Trees: Good for non-linearity, sensitive to outliers, may prioritize one feature in multicollinearity.
- Random Forest: combines multiple regression trees to improve accuracy and reduce overfitting
- XGBoost: Handles non-linearity, outliers, and multicollinearity well

<u>Sewing Department</u>

Results obtained for Ridge regression, Lasso regression and Elastic Net Regression

Model	**Best λ**	**Train_MSE**	**Test_MSE**	**Train_R^2**	**Test_R^2**
Ridge	0.01260142	0.004538757	0.00480081	0.8157705	0.76667059
Lasso	0.002778778	0.004694613	0.004851261	0.8094443	0.76421859
Elastic Net	0.004631	0.004708	0.004868	0.808888	0.763419

Table 2: Evaluation metrics for Ridge, Lasso, & Elastic Net – Sewing Department

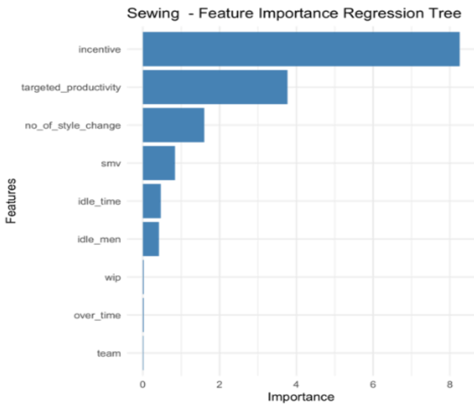
Ridge Regression is the best model for the Sewing Department as it achieves the lowest Test MSE (0.00480081) and the highest Test R^2 (0.7667), indicating better generalization and predictive performance. Its ability to handle multicollinearity ensures a balanced and stable model, making it the most reliable choice compared to Lasso and Elastic Net.

Regression Trees

Regression trees partition the dataset into smaller groups and fit a constant value for each subgroup. After training and tuning the model, the following results were obtained.

Department	**Best CP**	**Train_MSE**	**Test_MSE**	**Train_R^2**	**Test_R^2**
Sewing	0.03152387	0.008111911	0.007279573	0.6707352	0.6461975

Table 3 : Evaluation metrics for Regression Trees - Sewing



The Regression Tree model identifies "Incentive" as the most influential predictor, followed by "Target Productivity" and "No. of Style Changes." Variables like "SMV" and "Idle Time" have lower importance, while "Over Time" and "Team" have minimal impact and could be dropped without much affecting performance. This suggests the model prioritizes key features.

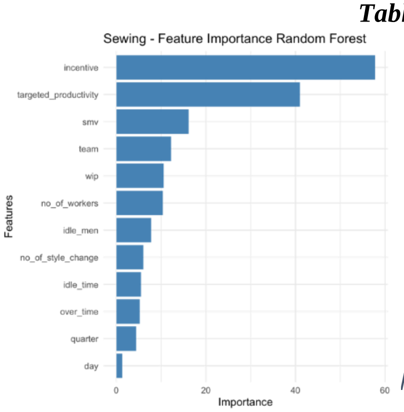
Fig 6: Feature importance plot of Regression Trees - Sewing

Random Forest

Random Forest is an ensemble learning algorithm that combines multiple regression trees to improve accuracy and reduce overfitting. Upon training the model and tuning parameters the following results were obtained.

Department	**Best mtry**	**Train_MSE**	**Test_MSE**	**Train_R^2**	**Test_R^2**
Sewing	6	0.009034374	0.00432204	0.9633292	0.7899398

Table 3 : Evaluation metrics for Random forest - Sewing



The Random Forest model too identifies "Incentive" as the most influential predictor, followed by "Target Productivity" and "smv". Variables like "overtime" and "Idle Time" have lower importance, while "quarter" and "day" have minimal impact and could be dropped without much affecting performance.

Fig 7: Feature importance plot of Random Forest - Sewing

XGBoost

XG Boost is a machine learning algorithm developed by combining decision trees and gradient boosting to enhance prediction accuracy. It excels in handling nonlinear complex relationships and improved model performance on both training and testing data with lower MSE and higher R^2 Values. Additionally, XG Boost provides insights into feature importance helping to identify the most influential variables and also the build in regularization in XG Boost prevents overfitting to ensure better generalization.

Department	Best Parameters	Train MSE	Test MSE	Train R^2	Test R^2
Sewing	colsample_bytree = 1.0 learning_rate = 0.1 max_depth = 5 n_estimators= 100 subsample = 0.8	0.00342	0.00415	0.9914	0.8025

Table 4: Evaluation metrics for XG boost – Sewing Department

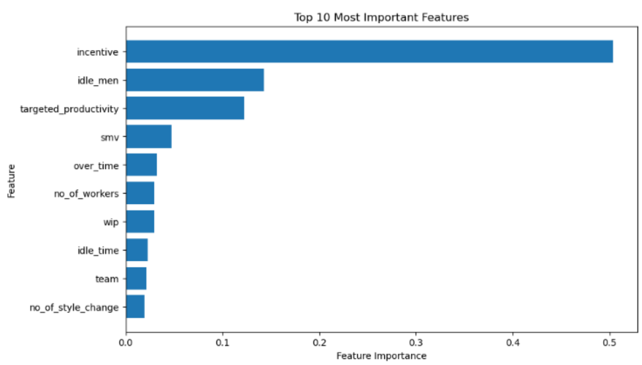


Fig 8 – Feature importance of the XGBoost in sewing Department

The key factors affecting worker productivity are incentives, idle men, targeted productivity, and SMV. Incentives have the highest impact, boosting motivation and output. Idle men highlight the need to minimize downtime, while targeted productivity ensures clear goals for better performance. SMV reflects task efficiency and time management, all crucial for maximizing productivity.

<u>Finishing Department</u>

Results obtained for Ridge regression, Lasso regression and Elastic Net Regression

Model	Best λ	Train_MSE	Test_MSE	Train_R^2	Test_R^2
Ridge	0.02660876	0.030833322	0.03314490	0.2020249	0.05292905
Lasso	0.002539935	0.030655711	0.032641496	0.2066215	0.06731304
Elastic Net	0.005080	0.030707	0.032622	0.205293	0.067866

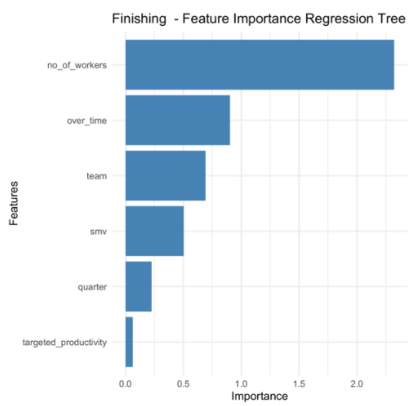
Table 5: Evaluation metrics for Ridge, Lasso, & Elastic Net – Finishing Department

Elastic Net is the best model for the Finishing Department as it achieves the lowest Test MSE (0.032622) and the highest Test R^2 (0.067866), indicating better generalization and predictive performance. By combining the strengths of both Ridge and Lasso regression, Elastic Net effectively balances feature selection and regularization, making it the most reliable choice compared to Ridge and Lasso.

Regression Trees

Department	**Best CP**	**Train_MSE**	**Test_MSE**	**Train_R^2**	**Test_R^2**
Finishing	0.06167450	0.033044044	0.031393102	0.1448108	0.1029843

Table 6 : Evaluation metrics for Regression Trees - Finishing



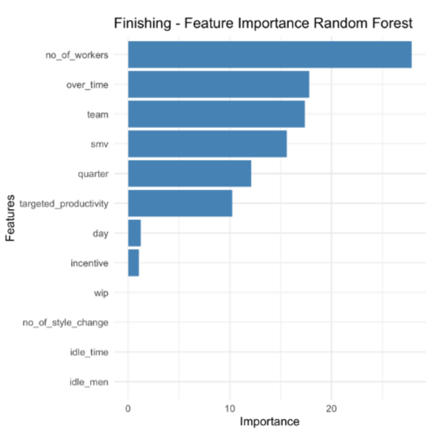
The Regression Tree model identifies “no_of_workers” as the most influential predictor, followed by “over_time” and “team.” Variables like “dmr” and “quarter” show moderate importance, while “targeted_productivity” has minimal. This suggests the model places a strong emphasis on workforce size, overtime usage, and team-related factors in its predictions.

Fig 9: Feature importance plot of Regression tree - Finishing

Random Forest

Department	**Best mtry**	**Train_MSE**	**Test_MSE**	**Train_R^2**	**Test_R^2**
Finishing	3	0.0120107800	0.02770349	0.6891576	0.2084101

Table 7 : Evaluation metrics for Random forest - Finishing



The model shows "no_of_workers" as the most influential predictor, followed by "over_time" and "team." Variables like "wip," "no_of_style_changes," "idle_time," and "idle_men" have minimal impact and could be removed without affecting performance. This suggests workforce-related factors are key drivers, while other variables contribute little.

Fig 10: Feature importance plot of Random Forest - Finishing

XG Boost

Department	Best Parameters	Train_MSE	Test_MSE	Train_ R^2	Test_ R^2
Finishing	colsample_bytree = 1.0 learning_rate = 0.01 max_depth = 5 n_estimators= 200 subsample = 0.8	0.0158	0.0274	0.5904	0.2170

Table 8 : Evaluation metrics for XGBoost - Finishing

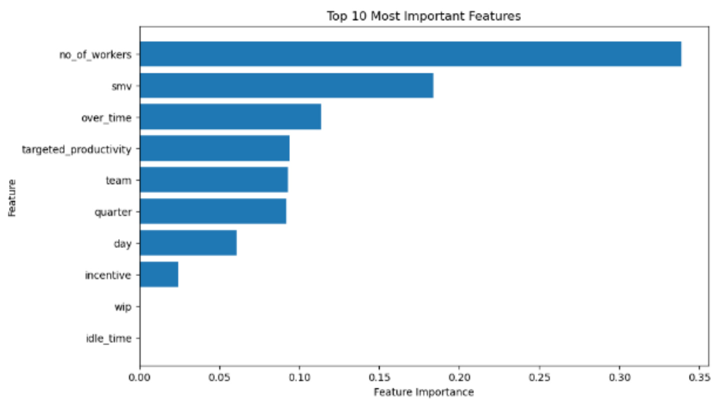


Fig 11 – Feature importance plot of XG Boost for finishing department

In the Finishing department, workforce size (0.338652) is the strongest predictor of productivity. SMV (0.183640) and overtime (0.113771) also play key roles, while targeted productivity (0.093822), team (0.092879), and quarter (0.091999) have moderate influence. Day (0.060884) has a smaller effect, and incentive (0.024352) is minimal. WIP, idle time, idle men, and style changes have no impact.

<u>Issues Encountered and Proposed Solutions</u>

During our advanced analysis, we encountered few challenges that required careful problem-solving. First, data quality issues, such as inconsistent entries and missing values, posed obstacles. We addressed these by implementing rigorous data cleaning and imputation techniques to ensure accuracy. Additionally, in applying regularization techniques, unexpected equality in results prompted us to use cross-validation for optimizing parameters, enhancing model performance. Our troubleshooting approach at every stage of the analysis reflects our commitment to delivering reliable, high-quality insights despite analytical hurdles.

Discussion and Conclusion

Models for the Sewing Department

Model	Train_MSE	Test_MSE	Train_R^2	Test_R^2
Ridge	0.004538757	0.00480081	0.8157705	0.76667059
Lasso	0.004694613	0.004851261	0.8094443	0.76421859
Elastic Net	0.004708	0.004868	0.808888	0.763419
Regression Tree	0.008111911	0.007279573	0.6707352	0.6461975
Random Forest	0.009034374	0.00432204	0.9633292	0.7899398
XG Boost	0.00342	0.00415	0.9914	0.8025

Table 9 : Summary of Models for the sewing department

For the sewing department, XGBoost demonstrates the best performance with the lowest test MSE (0.00415) and the highest test R^2 (0.8025), making it the most effective model for predicting productivity. Random Forest follows closely with a test R^2 of 0.7899, but its higher train MSE (0.00903) suggests potential overfitting. Among the linear models, Ridge Regression performs the best with a test R^2 of 0.7667, while Lasso and Elastic Net show similar but slightly weaker results. Regression Tree has the poorest performance, with the highest test MSE (0.00728) and the lowest test R^2 (0.6462), indicating weaker predictive power for this department.

Models for the Finishing Department

Model	Train_MSE	Test_MSE	Train_R^2	Test_R^2
Ridge	0.030833322	0.03314490	0.2020249	0.05292905
Lasso	0.030655711	0.03264149	0.2066215	0.06731304
Elastic Net	0.030707	0.032622	0.205293	0.067866
Regression Tree	0.033044044	0.0313931	0.1448108	0.1029843
Random Forest	0.0120107800	0.02770349	0.6891576	0.2084101
XG Boost	0.0158	0.0274	0.5904	0.2170

Table 10 : Summary of Models for the Finishing department

For the finishing department, XGBoost and Random Forest outperform the other models, with XGBoost achieving the lowest test MSE (0.0274) and the highest test R^2 (0.2170). Random Forest follows closely with a test MSE of 0.0277 and a test R^2 of 0.2084, making these two models the best choices for prediction. The linear models, including Ridge, Lasso, and Elastic Net, show weaker performance, with test R^2 values around 0.05 to 0.07, indicating limited predictive power. Regression Tree performs the worst, with the lowest test R^2 (0.1029), suggesting that it struggles to capture the underlying patterns in the data

References

- https://pmc.ncbi.nlm.nih.gov/articles/PMC7425836/
- https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/
- https://www.ibm.com/think/topics/random-forest
- https://www.geeksforgeeks.org/xgboost/

Appendix

- Link for the dataset: Productivity Prediction of Garment Employees
- Link for the Google Collab Notebook:
    1. https://colab.research.google.com/drive/1beMDzJlbvHdqXeD0Pu-5Apra4N4dTNXj?usp=sharing
    2. https://colab.research.google.com/drive/1G5hg98anU61f7lRNn0fa_BgUa8PR_Jsh?usp=sharing
    3. https://colab.research.google.com/drive/1jakgKsV0pHfIaOwBeIBTC-aRwoEXpVjn?usp=sharing