International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 531-544 March 2026-Sept 2026 www.csjournals.com

Predicting the Productivity of Garment Employees Using Machine Learning Algorithms

Manisha, Dr. Naveen, Dr. Anupam Bhatia
MCA Student, Asst Professor, Associate Professor
Deptt. of Comp. Sc. and Application, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract:

In textile mills, failure to meet daily targets is routine. This results in production delays and economic losses for the company. The UCI dataset that I used contains information on garment worker productivity in UCI, including 1197 records with 15 features such as SMV, overtime, incentive, WIP, team size, and so on.

Five models of machine learning - linear regression, decision tree, random forest, gradient boosting, and SVM were tested. I cleaned the data by filling missing values for the WIP column with median values, removing outliers using the IQR method, and encoding the categorical features such as department, day, or quarter in the dataset. I used an 80:20 split on the data for modelling and testing.

The results from my analysis indicated that random forest had the lowest error, as indicated by the MAE of 0.0573. Gradient Boosting has the largest variance explained at an R2 value of 0.5535. I also did an analysis on the features that have the most impact on productivity: first was targeted productivity, which accounted for 23.37%, followed by incentive for 13.60%, followed by SMV. Created a Streamlit web application for factory managers to submit worker information and receive real-time predictions on worker productivity.

Keywords: Streamlit, Machine Learning, Random Forest, Gradient Boosting, Feature Importance, Predictive Analytics, and Garment Productivity.

1. Introduction

Garment companies are under continuous pressure with respect to their production goals. Worker productivity affects job schedules and business profitability; most manufacturers do not have the means to predict worker productivity in advance. It's usually when managers realize the shortfall that they notice, but by then, damage has been caused.

Most garment units today use manual monitoring and end-of-day reporting for tracking production.
It's a reactive approach, with no opportunity for early intervention. Consequently, factories incur massive losses over and over again with no means to stop them.

Machine learning can come in handy in this instance. ML models can detect trends and

correlations in historical worker data, which can then predict outcomes in order to improve productivity. This provides the manager with advanced warning, enabling it to deal with problems before they get to critical proportions. [1]

Several factors influence worker output: the size of the team, work in progress, incentives, overtime, and task complexity. The complexity of these interactions with real productivity necessitates the use of a data-driven method that traditional approaches are unable to provide. [2]

1.1 Background of Study

The food processing industry is a competitive industry in which efficiency is a critical aspect of business success. Achieving the work objectives set by management, which can be interpreted as worker productivity, has been shown to change with variations in a number of human and operational factors. Production delays can have a ripple effect, leading to delays in schedules, missed delivery deadlines, and maybe even some financial hits.

The detailed operational data provided by modern manufacturing information systems affords a chance to use machine learning techniques to make predictions about productivity. Predictive models can help identify potential production slowdowns, and if done based on past trends, can warn of future issues, allowing you to take corrective measures, which may include:

- **Adjusting financial incentives to motivate workers**
- **Reallocating workers between teams**
- **Reducing style changes to minimize disruption**
- **Providing additional training or resources**
- **Optimizing overtime allocation**

The study strives to create a precise and interpretable machine learning model to predict the productivity of the garment worker, benchmark several machine learning algorithms, investigate significant determinants of worker productivity, and, in order to use it in industry, create a practical web app.

1.2 Factors Influencing the Productivity of Garment Workers

Building successful prediction models requires an understanding of what motivates productivity. The key factors studied in this research include:

Standard Minute Value (SMV): Standard Minute Value (SMV) is the standard time for producing a garment with one unit when working under normal conditions. The higher the SMV, the more complicated the operation is, which takes more time and effort. The overall productivity rate generally declines as the complexity of tasks increases.

Work in Progress (WIP): Items of work or garments in progress which have yet to be completed. Too much WIP leads to production line delays, a drop in flow efficiency, and all kinds of problems. An optimal WIP level is crucial for productivity improvement.

Financial Incentive: Rewards for exceeding or meeting production targets beyond what is usually expected or due are considered financial incentives. This promotes and motivates employees to work better and conduct themselves more productively. A suitable remuneration system will help to increase motivation and overall productivity in the factory.

Overtime:
Overtime work hours are those that are worked after regular work. A moderate amount of overtime can help factories meet urgent production demands by providing extra facilities. In addition to the upsurge in production, factories can meet the urgent demand for production by providing a moderate amount of overtime. But the limiting factor would be that if there is an excessive amount of overtime, then there would be a decline in employee morale over time, which could bring a reduction in efficiency, and then worker fatigue.

Idle Time and Idle Men - sometimes machines break down or material runs out. During this time workers sit idle and nothing gets produced, which hurts overall output.

1.3 Problem Overview:

Despite the availability of operational data, garment factories face several challenges in productivity management:

1. **Reactive Management:**

   The majority of the garment factories work on a reactive type of management system, which basically identifies productivity problems after it has an effect on production. This puts off timely remedial measures and results in reduced loss prevention capabilities of managers. This, in turn, results in a decrease in operational efficiency and production planning.

2. **No Predictive Tools:**

   Unfortunately, there are not many practical predictive systems to estimate worker productivity prior to the beginning of a production shift at many garment factories. This means that management decisions are made, unfortunately, without the benefit of facts. This restricts planning resources and making a decent production goal.

3. **Unknown Key Drivers:**

   There is often a lack of information for factory managers on the quantitative aspects of productivity affecting their workers. Incentives, working conditions, team composition, and overtime are not systematically analyzed. As a result, it makes it very difficult to develop specific measures to increase productivity.

4. **Financial Losses:**

   When garments actually do not meet the expected rate of production, serious economic losses are experienced in the garment factories. Machines that are not in production are not being used for that production, which leads to overtime costs, inefficiencies, and in some cases, late delivery penalties. These losses are ultimately translated into the profitability and competitiveness of the organization.

This research addresses these challenges by developing a predictive machine learning system and deploying it as an accessible web application for practical use in factory settings.

2. Literature Review

Five models of machine learning - linear regression, decision tree, random forest, gradient boosting, and SVM were tested. I cleaned the data by filling missing values for the WIP column with median values, removing outliers using the IQR method, and encoding the categorical features such as department, day, or quarter in the dataset. I used an 80:20 split on the data for modelling and testing.

3. Research Gap

There is a large body of literature available, and through a complete review of this literature, several gaps were identified. These gaps indicate the need for better knowledge of machine learning methods that can predict the productivity of a garment worker more accurately, reliably, and practically.

3.1 Low Prediction Improvement Scope

Many machine learning models provide moderate prediction errors (higher MAE and MSE values) before good prediction accuracy has been achieved in previous studies. There has been little research focused on building a model with better predictive accuracy by using advanced data preprocessing techniques, feature engineering, and optimization of the model.

3.2 Limited Algorithm Comparison

Previous research studies have tested lesser-known one- to two machine learning algorithms and selected the best after execution without comparing their performance across all. No systematic comparison or evaluation of several algorithms on a common set of data, and using the same performance criteria to check the reliability of the model.

3.3 Lack of Explainability

Most prediction models are “black boxes” and make few, if any, attempts to explain their predictions. If the factory managers don't understand what makes the models work, they may

be reluctant to believe and implement these approaches. Thus, there is an outstanding research need to make models more interpretable.

3.4 Weak Preprocessing Focus

Though it is widely recognized that data preprocessing contributes significantly to the performance of the machine learning algorithm, several previous works have addressed only a handful of data preprocessing methods. Problems like Missing Value, Outliers, encoding, and data normalization are either not addressed enough or insufficiently addressed, which can affect the accuracy and reliability of the model.

3.5 No Real Deployment

Published studies are predominantly based on experimental testing, and work has not been successfully developed or fully disseminated into gearing into points to support practical decisions made by garment factories. There is thus a large divide between academic research and actually implementing the predictive models into industry, reducing the usefulness of these models.

3. Problem Statement

As worker productivity is difficult to predict, it is difficult to meet production targets in garment factories regularly. Supervisors rely on manual observation and daily reporting systems after a problem has occurred.

No tool currently exists that can help managers predict, before the change in shift, if workers will achieve their goals. This can result in multiple losses and late delivery.

My study is an attempt to close this gap by creating models that predict productivity before the shift is over. This will allow time for managers to intervene and prevent any serious damage.

4. Research Objectives

Primary Objective

To develop an accurate and explainable machine learning system for predicting the actual productivity of garment employees using operational and workforce-related features.

Specific Objectives

- To collect, explore, and preprocess the UCI Garment Worker Productivity dataset
- I wanted to properly handle missing values, outliers and encoding issues in the data.
- I built and trained five models.
- Linear regression, Decision tree, Random forest, Gradient Boosting and SVM.

- I compared all models using different error metrics like MAE, MSE, RMSE, MAPE and R2.
- To identify the most influential features affecting worker productivity using feature importance analysis
- To analyze the impact of incentive, SMV, and targeted productivity on actual worker output
- To develop and deploy a Streamlit web application for real-time productivity prediction
- To address identified research gaps including model comparison, preprocessing quality, and practical deployment.

5. Research Methodology

5.1 Dataset Description

We adopt the UCI Garment Worker Productivity dataset that is made available by A. Al Imran (2020. It's a comprehensive set comprising 1,197 records from a garment production unit based on 15 features including operational and workforce aspects. The features of the dataset are described in Table 1.

Table 1: Dataset Features and Descriptions

Feature	Rewritten Description	Data Type
Quarter	Production period divided into five quarterly segments (Q1–Q5).	Categorical
department	Production section responsible for the assigned task, such as Sewing or Finishing.	Categorical
Day	Weekday on which the production activity was performed.	Categorical
Team	Identification number assigned to each production team.	Numerical
targeted_productivity	Expected productivity level established by factory management before production begins.	Numerical
Smv	Standard time, expressed in minutes, required to complete one unit of work.	Numerical
Wip	Number of garment items that are currently under production but not yet completed.	Numerical
over_time	Total overtime worked by the production team, measured in minutes.	Numerical
incentive	Additional monetary reward provided to workers for achieving production targets.	Numerical
idle_time	Total time during which production activities remained inactive.	Numerical
idle_men	Count of workers who remained idle during the production process.	Numerical
no_of_style_change	Frequency of style or product design changes during production.	Numerical
no_of_workers	Total number of workers assigned to a production team.	Numerical

actual_productivity	**Target variable** representing the productivity level actually achieved by the production team, expressed on a scale from 0 to 1.	Numerical

5.2 Data Preprocessing

One important step that is closely related to the performance of the model is data preprocessing. The following pre-processing steps were used:

Fixing Department Column: The Department column had padding, which caused there to be three separate values when there should only have been two. To obtain two clean categories, sewing and finishing, the str.strip() function was used to remove spaces.

Handling Missing Values: The WIP column had 506 missing values, which is around 42% of the data. For the gaps, I used the median value of 1022.0, not the mean, as there were some very high values such as 23,122 for WIP, which would have skewed the mean.

Outlier Removal: I removed outliers from the target variable using the IQR method. I removed data points from the dataset as Outliers using the IQR method. Essentially, data points that fell below a row distance of Q1 - 1.5IQR or above a row distance of Q3 +1.5IQR were omitted. The total number of rows dropped was 54, and the number of records in the dataset came out to 1,143.

Feature Removal: The Date column was eliminated since the date was already represented by the Quarter and Day columns.

5.3 Feature Engineering and Encoding

It was found that machine learning models need to have a numerical representation for input variables, so categorical variables were transformed into numerical representation by Label Encoding. Unique category labels were retained for each of the variables (Quarter, Department, and Day), which were coded as integers. This shift allowed the algorithms to sense this categorical info in an efficient way during model training.

5.4 Train–Test Split and Normalization

A dataset was split into 80% training and 20% testing data to be used for monitoring the performance of the model on data that had not been seen. The states were randomly selected to achieve 42 states, which gave 914 training and 229 testing records for reproducibility. All numerical features were standardized to ensure their equal contribution to learning and to reduce the need to learn efficiently, using StandardScaler before model training.

5.5 Machine Learning Models

1. Linear Regression

A simple model for forecasting the yield was chosen in order to compare it with the other models and interpret it simply Linear Regression. It assumes a linear relationship between the input variables and the output variable to estimate worker productivity. It works well for patterns of linear relationships, but is less accurate when the relationship between the variables is nonlinear.

2. Decision Tree Regressor

The Decision Tree Regressor predictsively splits the data set repeatedly on the basis of the different features until the smallest datasets remain. Because of this hierarchical organization of the model, it can be easily understood and used to implement complex decision rules. But if the tree is too deep, it might overfit on the training data and get poor performance on test data.

3. Random Forest Regressor

Random Forest is an ensemble method for regression that includes a collection of decision trees to make a single prediction. To prevent overfitting and make the predictions more stable, a forest made of 100 decision trees was created in this study. In addition, the algorithm can determine feature importance, aiding in the identification of variables that have the biggest impact on worker productivity.

4. Gradient Boosting Regressor

Gradient Boosting is a family of decision tree methods that builds a series of trees where a new tree is trained to offset the prediction error of the already built trees. The algorithm does not build all of the trees individually; rather, it builds the trees one step at a time, using iterative learning. In order to get a better prediction of productivity, a total of 100 boosting trees were employed in this study.

5. Support Vector Regression (SVR)

Support Vector Regression (SVR) was employed to model the complex relationships present in the garment productivity dataset. The model used the Radial Basis Function (RBF) kernel, which is effective for capturing nonlinear patterns in the data. By identifying an optimal regression function with minimum prediction error, SVR provides reliable estimates even when the relationship between variables is highly complex.

5.6 Evaluation Metrics

Five widely used regression indicators were used to measure the performance of the machine learning models. The metrics are used from various angles of view to compare the accuracy of prediction and offer a comprehensive evaluation of model performance.

1. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is used to determine the average size of the prediction errors made by a regression model. The absolute difference between the observed and predicted productivity values is calculated, and all of the errors are given equal weighting, regardless of whether they are negative or positive. The lower the MAE, the more the model is predicting the actual value, as the prediction is

2. Mean Squared Error (MSE)

The Mean squared error (MSE) is used to measure how well the model predicts the data by computing the average of the squared differences between actual and predicted values. A larger prediction error will have a larger influence on the final prediction than a smaller error in prediction because it is squared before it is averaged. Therefore, the lower the MSE value of a model, the closer the approximation is to the correct estimation of the model.

3. Root Mean Squared Error (RMSE)

Root Mean Squared Error (MSE) is calculated by taking the square root of the Mean Squared Error (RMSE). This transformation is done to represent the prediction error in the same unit as the target variable so that it is easier to understand. The smaller the RMSE, the closer the productivity values are predicted to the observations.

4. Mean Absolute Percentage Error (MAPE)

The Mean Absolute Percentage Error (MAPE) is a metric used to measure the accuracy of predictions, based on their average percentage error. The MAPE is a relative error measure that can be compared with other datasets and scales, unlike an absolute error measure. Lower MAPE will lead to more accurate and reliable forecasts.

5. Coefficient of Determination (R^2 Score)

The Coefficient of Determination, commonly represented as R^2, assesses how well a regression model explains the variation in the dependent variable. Its value ranges between 0 and 1, where higher values indicate that a greater proportion of the variability in worker productivity is explained by the model. Consequently, a higher R^2 score reflects better predictive performance and a stronger fit between observed and predicted values.

6. Results and Discussion

6.1 Model Performance Comparison

All five models were trained on the same preprocessed dataset and evaluated on the held-out test set. Table 2 presents the complete performance comparison.

Table 2: Model Performance Comparison

Model	MAE	MSE	RMSE	MAPE(%)	R2
Linear Regression	0.0863	0.0148	0.1217	13.26	0.2291
Decision Tree	0.0716	0.0159	0.1262	10.36	0.1712
Random Forest	**0.0573**	**0.0091**	**0.0956**	**8.60**	**0.5250**
Gradient Boosting	**0.0603**	**0.0086**	**0.0927**	**9.20**	**0.5535**
SVM	0.0785	0.0133	0.1153	11.96	0.3089

The results demonstrate that the ensemble learning methods were more successful than the simpler models. Out of them, random forest had the lowest average error of 0.0573, indicating that it predicted with an average difference of 0.57 on the range of 0- 1 from the actual values of productivity. For practical industrial applications, this is an adequate error of about 5.7%pe. From the results, it can be concluded that ensemble models performed better than simple ones. The lowest MAE was 0.0573, which indicates the accuracy of the predictions by random forest, and it was only 5.7% away from the real value. This amount of error would be suitable for a real factory operation.

Gradient boosting recorded the highest R2 with 0.5535, which means it could account for 55% of the variation in productivity. The scores are also mixed because human productivity is affected by other factors such as mood, health, personal issues, and those are not provided here.

Linear regression gave a poor R2 of only 0.2291, which shows that productivity data is not linear. Although the decision tree had a lower MAE than the linear regression, its R2 was even lower at 0.1712. This is because the decision tree was overfitting the training data.

6.2 Feature Importance Analysis

Feature importance analysis using Random Forest identified the key drivers of garment worker productivity. Table 3 presents the importance scores for all features.

Table 3: Feature Importance (Random Forest)

Rank	Feature	Importance Score
1	targeted_productivity	0.2337

Rank	Feature	Importance Score
2	Smv	0.1539
3	incentive	0.1360
4	no_of_workers	0.0976
5	over_time	0.0927
6	Team	0.0870
7	Day	0.0845
8	Wip	0.0412
9	quarter	0.0380
10	department	0.0210
11	idle_time	0.0098
12	idle_men	0.0076
13	no_of_style_change	0.0052

The results show three main determinants for the productivity of the cotton. Targeted productivity (0.2337) is the most influential factor, showing that management-prescribed targets back-feed into worker productivity. Third place, at 0.1539, was SMV, whose results indicated that task complexity has a significant impact on workers' productivity, while incentives are the third, with 0.1360, which clearly means money does motivate workers to perform better.

From the results obtained, the factory managers should concentrate on three aspects: (1) setting realistic goals, (2) starting motivation by giving good incentives to workers while controlling the complexity of tasks using SMV, and (3) keeping task complexity low while controlling SMV by example. These three factors can really have an impact on productivity.

6.3 Web Application

It was developed with Streamlit, a web application, to allow for predictions of productivity in real time. The app loads the saved Random Forest using pickle. The manager only has to put in all the necessary information, such as the team size, incentive, overtime, and SMV. The app shows the anticipated productivity score and determines classification into one of four worker performance categories. A productivity score of 0.80 is rated as "excellent", a score between 0.70 and 0.80 as "good", a score between 0.60 and 0.70 as "average", and a score below 0.60 as "low".

The deployment helps integrate research into real applications, allowing factory personnel, who are not tech-savvy, to utilize the predictive model, which in turn leads to proactive production management.

7. Conclusion

The majority of this research was aimed at comparing the best-performing model among the five machine learning models for the prediction of productivity of the garment workers in the UCI dataset.

Predict productivity and identify factors most likely to impact worker productivity.

The key findings are summarized as follows:

- Random Forest achieved the best predictive performance with the lowest MAE of 0.0573
- Gradient Boosting obtained the highest R2 score of 0.5535, explaining 55% of productivity variance
- Targeted productivity (23.37%), SMV (15.39%), and incentive (13.60%) are the three most important productivity drivers
- A practical Streamlit web application was developed for real-time prediction and industrial deployment
- Five identified research gaps were addressed through comprehensive model comparison, quality preprocessing, and real deployment

R2 values are not very high but that is okay because human productivity is affected by many personal things we cannot measure or put in a dataset. The models are adequate, even with moderate scores, and they can support the daily decisions made by managers. Further work may involve the use of features at the worker level, further development of the neural network, and SHAP analysis.

8. Future Scope

- Exploration of deep learning approaches such as LSTM networks for capturing temporal productivity patterns
- Integration of additional features such as worker experience, attendance records, and health data
- Application of SHAP explainability analysis for more granular insights into individual predictions
- Development of real-time data collection systems integrated with the web application
- Extension of the study to other manufacturing sectors beyond garment production
- Implementation of hybrid models combining Random Forest and Gradient Boosting for improved accuracy

9. References

[1] Imanuel Balla, Sri Rahayu, Jajang Jaya Purnama, "Garment Employee Productivity Prediction Using Random Forest," Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, 2021.

[2] Al Imran et al., "Deep Neural Network Approach for Predicting the Productivity of Garment Employees," 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), IEEE, 2019.

[3] Hasibul Hasan Sabuj et al., "Interpretable Garment Workers' Productivity Prediction in Bangladesh Using Machine Learning Algorithms and Explainable AI," 25th International Conference on Computer and Information Technology (ICCIT), IEEE, 2022.

[4] A. Al Imran, "Productivity Prediction of Garment Employees Dataset," UCI Machine Learning Repository, 2020.