Garment Worker Productivity Prediction

Project Report

(Code Link)

01. Overview

Summary of the project

The primary objective of this project is to use a structured dataset to construct a predictive machine learning model. By following a methodical data science workflow that involves data pretreatment, exploration, model implementation, evaluation, and fine tuning, the main goal is to produce accurate predictions of the target variable. To find the best model for this dataset, a range of machine-learning models were assessed and their performance was methodically compared using a number of measures. Each stage of the project is outlined in this report, including data preparation, model selection, hyperparameter tuning, and a performance evaluation of the finished model.

02. Data Overview

Summary of the dataset and key features

The dataset used for this project is the Garment Worker Productivity Dataset. It includes **1,197 records** and **14 features**.

Key Features	
date	Date of the record.
quarter	The quarter of the year (e.g., Q1, Q2).
department	Department of workers (e.g., sewing, finishing).
team	The number representing the team.
targeted_productivity	The target productivity (between 0 and 1).
smv	Standard Minute Value (time required to complete the task).
wip	Work In Progress (missing values present).
over_time	Overtime in minutes.
incentive	Bonus paid to workers.
idle_time	Time during which no work was done.
idle_men	Number of idle workers.
no_of_style_change	Number of style changes in production.
no_of_workers	Number of workers in the team.
actual_productivity	Target variable representing the productivity achieved (between 0 and 1)

03. Data Cleaning and Preprocessing steps

Summary of the dataset and key features

To guarantee that the dataset is correct, comprehensive, and in a format appropriate for machine learning models, preprocessing is crucial. Every preprocessing procedure, such as handling missing values, identifying and managing outliers, encoding categorical variables, and normalizing numerical data, is covered in detail in this section.

1. **Handle missing values**

   The median of the column was utilized to fill in the missing values in the **'wip'** column. As the median is less susceptible to outliers and promotes consistency, this strategy was selected.

2. **Detect and handle Outliers**

   Outliers were identified in columns **'smv', 'wip', 'over_time'**, **'incentive', 'idle_time', 'idle_men', 'no_of_style_change'** and **'no_of_workers'** using the interquartile range (IQR) method. Values falling outside the 1.5*IQR range were considered outliers and were either removed or capped.

3. **Encoding Categorical Variables**

   Categorical features such as **'day', 'quarter'**, and **'department'** were encoded using label encoding.

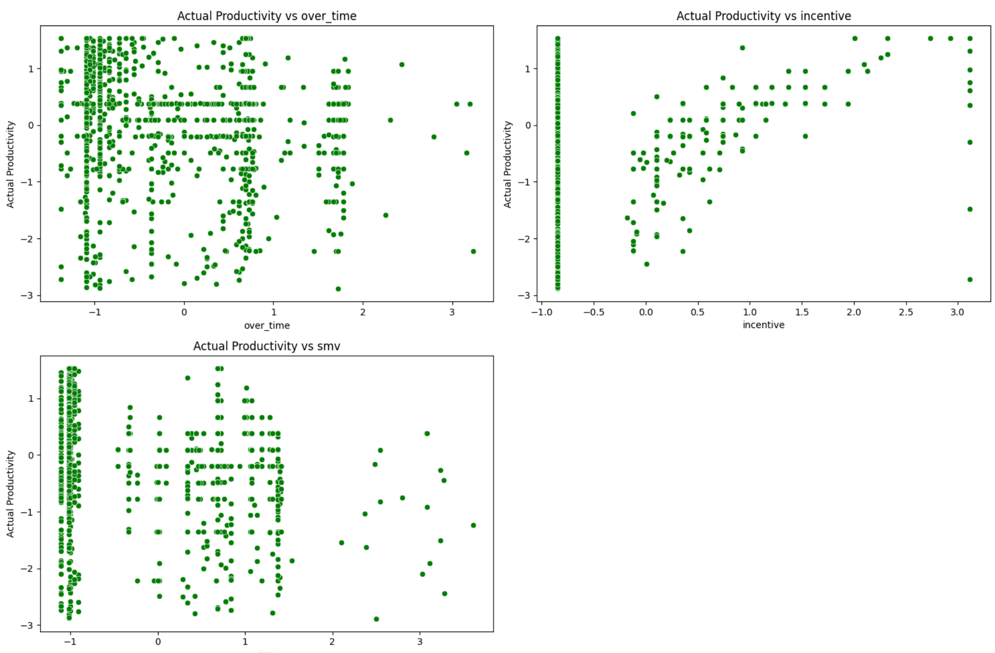
4. **Standardized numerical features**

   The numerical columns **'smv', 'wip', 'over_time', 'incentive', 'idle_time', 'no_of_workers', 'no_of_style_change', and 'actual_productivity'** were subjected to standard scaling, which resulted in a mean of zero and a standard deviation of one. This improves the performance of models like Support Vector Machines (SVM) and Gradient Boosting that are sensitive to feature scales.

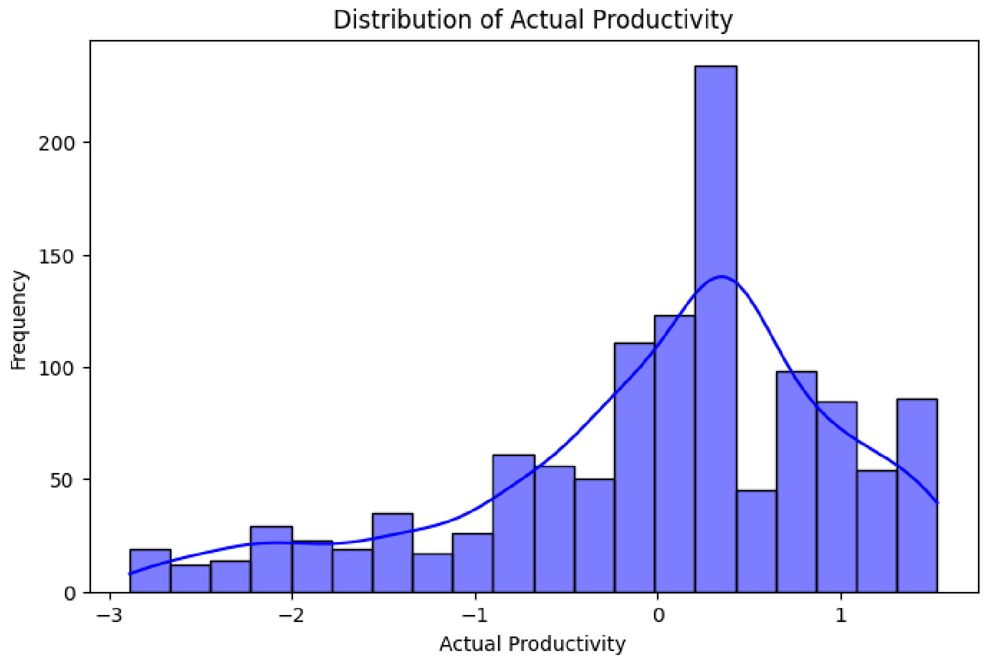
04. Data Visualizations

To understand the relationships between features and the target variable **'actual_productivity'**, several visualizations were created, including scatter plots, distribution plots, and a correlation heatmap. Visualizations provide insights into patterns, trends, and correlations in the dataset.

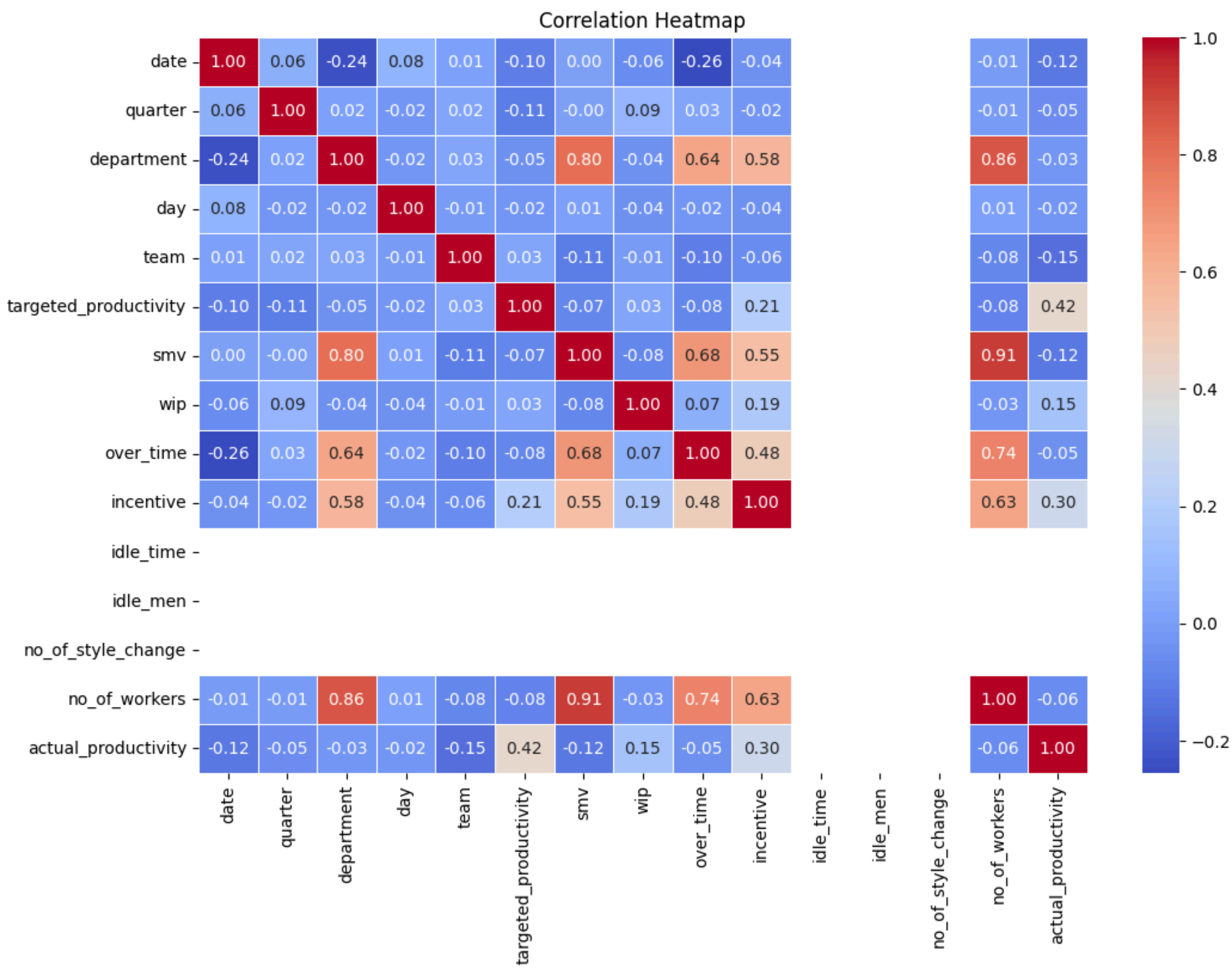
Relationship between features and target variable 'actual_productivity'



Distribution of the target variable ‘actual_productivity’



Correlation heatmap of the features and target variable



05. Modeling

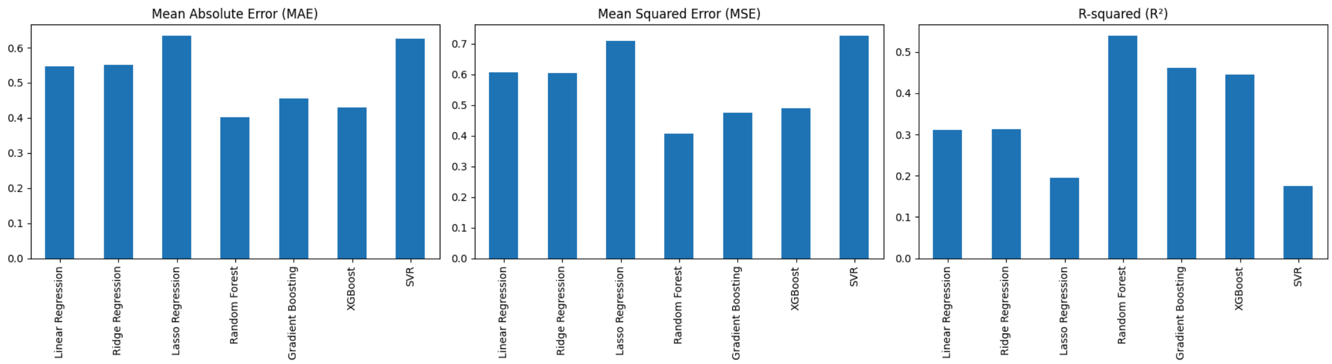
To predict productivity, a range of machine learning models were used, each assessed using a number of performance indicators. Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, Support Vector Regressor (SVR), and XGBoost were among the models that were analyzed. The following performance measures are used to evaluate the efficacy of the model:

Mean Absolute Error (MAE): This metric quantifies the average magnitude of errors in predictions without considering their direction.

Mean Squared Error (MSE): This metric emphasizes larger errors by penalizing them more significantly.

R^2 Score: This metric indicates the proportion of variance in the target variable that is explained by the model.

Models performance



	MAE	MSE	R2
Linear Regression	0.547451	0.605642	0.311706
Ridge Regression	0.550629	0.603926	0.313656
Lasso Regression	0.633687	0.708101	0.195264
Random Forest	0.402419	0.405475	0.539189
Gradient Boosting	0.454254	0.473930	0.461393
XGBoost	0.430052	0.488643	0.444671
SVT	0.624851	0.725237	0.175789

The Random Forest model was selected as the preferred option due to its superior performance across multiple evaluation metrics compared to other models tested. Among all candidates, Random Forest achieved the lowest Mean Squared Error (MSE) and Mean Absolute Error (MAE), alongside the highest R^2 score, indicating its ability to explain a greater proportion of the variance in productivity.

This model's ensemble approach, which combines predictions from multiple decision trees, enables it to effectively capture complex, non-linear relationships in the data while minimizing overfitting. While Gradient Boosting also delivered strong results, Random Forest outperformed it with a higher R^2 score and lower error values, suggesting better generalization to unseen data and a more accurate representation of true productivity.

This combination of exceptional accuracy, robustness, and interpretability establishes Random Forest as the ideal model for predicting productivity in this context.

Fine-tuning Random Forest

To enhance the performance of the Random Forest model for productivity prediction, we conducted hyperparameter tuning using GridSearchCV. This approach aimed to identify the optimal combination of hyperparameters to maximize the R^2 score, reflecting the proportion of variance explained by the model. The following parameters were adjusted:

n_estimators: Representing the number of trees in the forest, this parameter was tested within a range of **50 to 200** to balance computational efficiency, underfitting, and overfitting risks.

max_depth: Dictating the maximum depth of each decision tree, values ranged from **10 to 30**, with the inclusion of None to allow unlimited depth. This governs model complexity and helps address overfitting.

min_samples_split: The minimum number of samples required to split an internal node was tested with values of **2, 5, and 10**. Lower values increase model sensitivity, while higher values reduce overfitting.

min_samples_leaf: Representing the minimum number of samples required to be at a leaf node, this parameter was tested with values of **1, 2, and 4**, helping control tree growth and improve generalization.

Utilizing 5-fold cross-validation on the training dataset with R^2 as the scoring metric, GridSearchCV systematically explored all parameter combinations. The process was parallelized to leverage all available processing cores, ensuring efficiency. The optimal parameter configuration identified was as follows:

n_estimators: 200

max_depth: 20

min_samples_split: 5

min_samples_leaf: 2

The performance metrics associated with this finely tuned model are as follows:

- **Mean Absolute Error (MAE): 0.402419**, indicating a low average prediction error.
- **Mean Squared Error (MSE): 0.405475**, reflecting minimal variance in prediction errors, which is advantageous for accuracy.
- **R^2 Score: 0.539189**, indicating that the model explains approximately 54% of the variance in productivity, representing a significant improvement over prior iterations.

These results demonstrate that the optimized Random Forest model not only succeeded in reducing errors, as evidenced by the MAE and MSE, but also achieved a higher R^2 score compared to other models evaluated. Consequently, this model may be confidently employed to predict productivity with enhanced accuracy and reliability, establishing it as an effective solution for our application.