Productivity Prediction of Garment Employees

Documentation

S. M. S. N Senanayake

2024

1. Abstract

In global manufacturing, the garment industry is a vital player, relying heavily on manual labour, and other complex processes. Meeting the immense global demand for garments hinges on the productivity of factory workers. Decision-makers in this industry need to track, analyse, and predict workforce productivity to maintain competitiveness.

The essence of productivity in this sector extends beyond mere output metrics; it encompasses the efficiency and effectiveness of the entire production process, from raw material procurement to final product delivery. The intricate interplay of manual labor, machinery, logistics, and supply chain management underscores the complexity of this industry's operational landscape.

This project utilizes a dataset from the garment industry to delve into productivity patterns. The dataset covers various aspects, from worker demographics to production processes. The main goal is twofold: to predict productivity levels on a scale of 0 to 1, and to classify productivity into distinct categories. By leveraging machine learning, the aim is to provide insights that empower decision-makers to optimize workforce performance effectively.

Through the lens of machine learning, this project endeavours to illuminate the intricate dynamics shaping productivity performance within the garment industry. By harnessing the power of data-driven insights, we aspire to equip decision-makers with the foresight and agility needed to navigate the complexities of this dynamic and ever-evolving domain.

2. Data Description

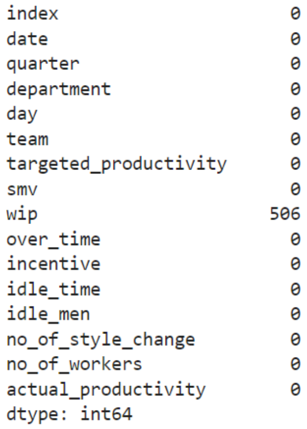
The dataset comprises important attributes related to the garment manufacturing process and the productivity of employees. It has been collected manually and validated by industry experts.

- Subject Area: Business
- Associated Tasks: Classification, Regression
- Feature Type: Integer, Real
- Number of Instances: 1197
- Number of Features: 15
- Missing Values: Yes

01	date	: Date in MM-DD-YYYY
02	day	: Day of the week
03	quarter	: A portion of the month.
04	department	: Associated department with the instance
05	team_no	: Associated team number with the instance
06	no_of_workers	: Number of workers in each team
07	no_of_style_change	: Number of changes in the style of a particular product
08	targeted_productivity	: Targeted productivity set by the authority, each team per day
09	smv	: Standard Minute Value, it is the allocated time for a task
10	wip	: Work in progress. The number of unfinished items for product
11	over_time	: Represents the amount of overtime by each team in minutes
12	incentive	: Represents the amount of financial incentive (in BDT)
13	idle_time	: The amount of time the production was interrupted
14	idle_men	: The number of workers who were idle due to interruption
15	actual_productivity	: The actual % of productivity that was delivered by the workers.

- **data preprocessing**

- Missing Values

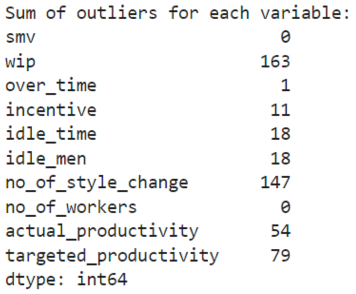


```
index                    0
date                     0
quarter                  0
department               0
day                      0
team                     0
targeted_productivity    0
smv                      0
wip                    506
over_time                0
incentive                0
idle_time                0
idle_men                 0
no_of_style_change       0
no_of_workers            0
actual_productivity      0
dtype: int64
```

There were 506 missing values in ‘wip’ (The number of unfinished items for the product).

Imputed the missing values by the mean of that particular attribute.

- Outliers



```
Sum of outliers for each variable:
smv                        0
wip                      163
over_time                  1
incentive                 11
idle_time                 18
idle_men                  18
no_of_style_change       147
no_of_workers              0
actual_productivity       54
targeted_productivity     79
dtype: int64
```

According to the ‘IQR method’ the outliers are detected as above. No change has been made on the outliers as of the lack of knowledge on the actual domain considered here.

- Data Cleaning

Replacing 'finishing ' with 'finishing' in the department attribute, to correct the inconsistency in the department names to ensure that the data is accurate and consistent before further analysis or modelling.

- Scaling and Normalizing

As the value ranges and types are unique to each column, with the purpose of taking each attribute to common scale, used Min-Max scaling and Standardization (Z-score normalization).

3. Methodology

Shapiro Wilk statistic – To check the normality assumption of a distribution.

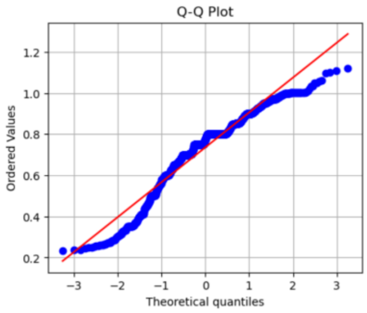
Mann Whitney U test -To check whether there is a difference in the median between two distributions.

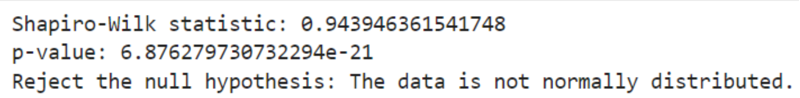
Kruskal Wallis H test - To check whether there is a difference in medians of more than two distributions.

Spearman's rank correlation - Assess the strength and direction of the monotonic relationship between two variables.

4. Exploratory Data Analysis

- Checking the normality assumption for ‘actual_productivity’

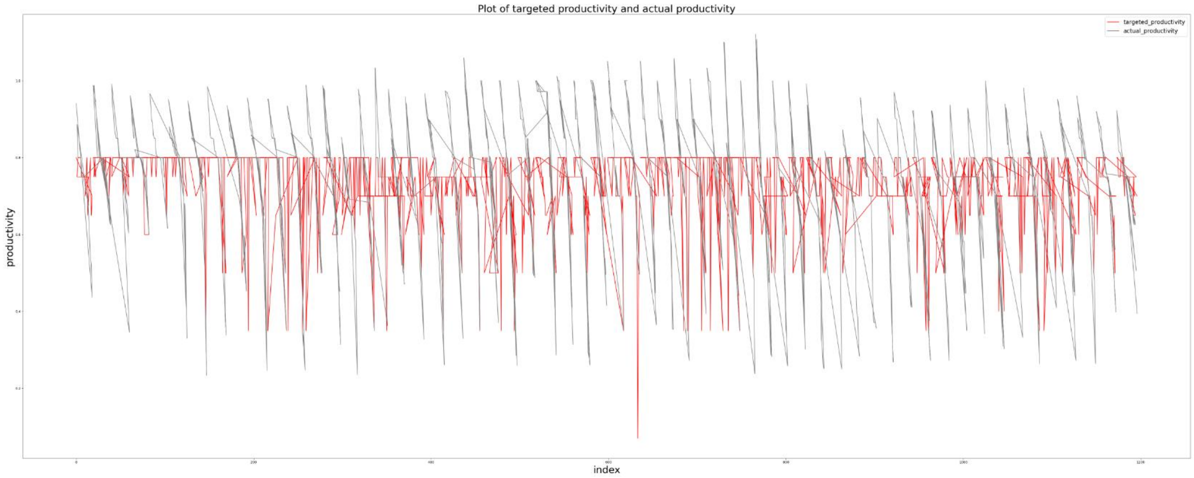




```
Shapiro-Wilk statistic: 0.943946361541748
p-value: 6.876279730732294e-21
Reject the null hypothesis: The data is not normally distributed.
```

The Shapiro-Wilk statistic confirms that the data are not normally distributed. Hence the statistical tests that used in the future will be non-parametric.

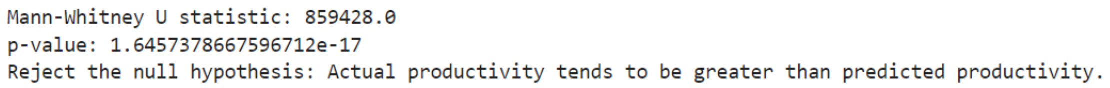
- Targeted productivity and the actual productivity



Hypothesis:

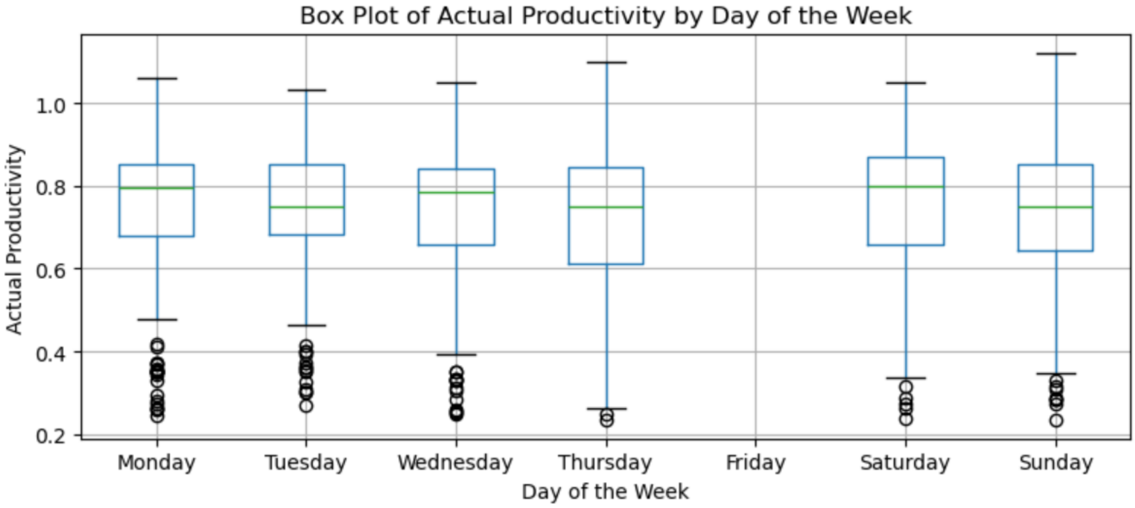
Null Hypothesis (H0): There is no significant difference between the distributions of Actual productivity and Targeted Productivity.

Alternative Hypothesis (H1): Actual productivity tends to be greater than Targeted productivity.



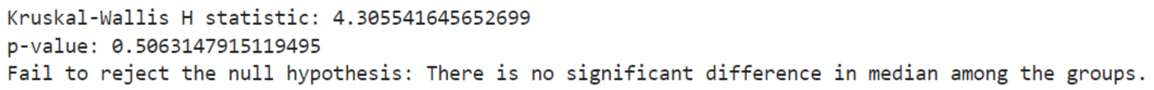
```
Mann-Whitney U statistic: 859428.0
p-value: 1.6457378667596712e-17
Reject the null hypothesis: Actual productivity tends to be greater than predicted productivity.
```

- Actual productivity by day of the week



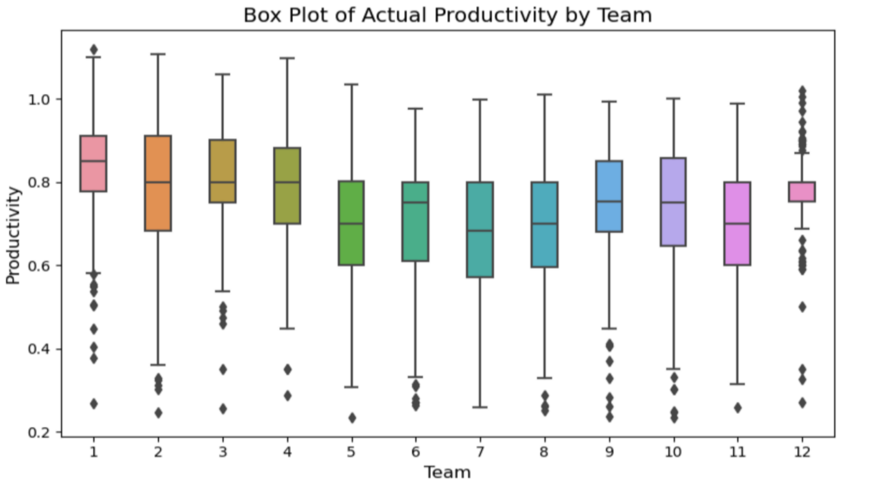
Hypothesis: Null Hypothesis(Ho) – There is no significant difference in the median productivity of 7 days of the week.

Alternative Hypothesis(Ho) – There is a significant difference in the median productivity of 7 days of the week.



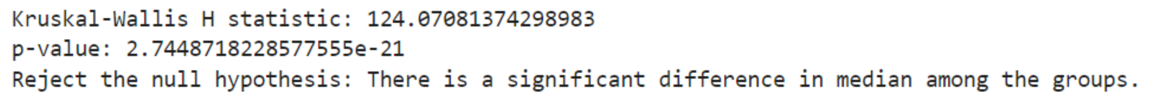
```
Kruskal-Wallis H statistic: 4.305541645652699
p-value: 0.5063147915119495
Fail to reject the null hypothesis: There is no significant difference in median among the groups.
```

- Actual productivity by team



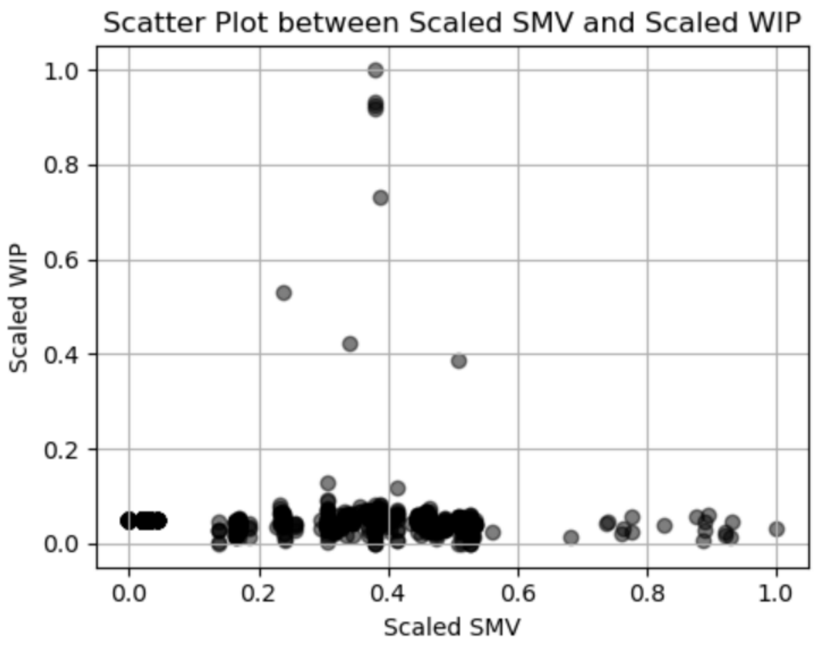
Hypothesis: Null Hypothesis(Ho) – There is no significant difference in the median productivity of 12 teams.

Alternative Hypothesis(Ho) – There is a significant difference in the median productivity of 12 teams.

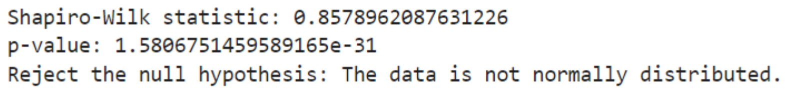


```
Kruskal-Wallis H statistic: 124.07081374298983
p-value: 2.7448718228577555e-21
Reject the null hypothesis: There is a significant difference in median among the groups.
```

- ➢ Allocated time for a task and number of unfinished items

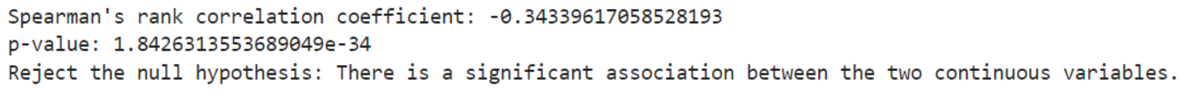


Normality test for ‘smv’,



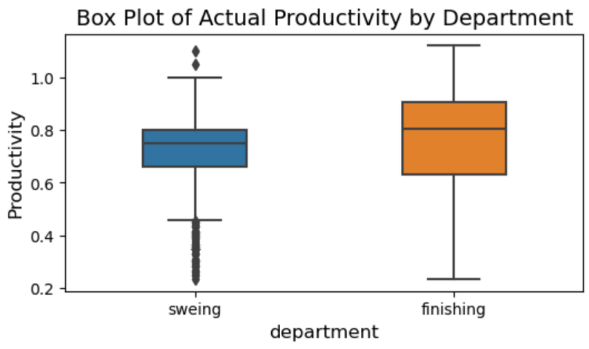
```
Shapiro-Wilk statistic: 0.8578962087631226
p-value: 1.5806751459589165e-31
Reject the null hypothesis: The data is not normally distributed.
```

Using a non-parametric technique to assess the correlation



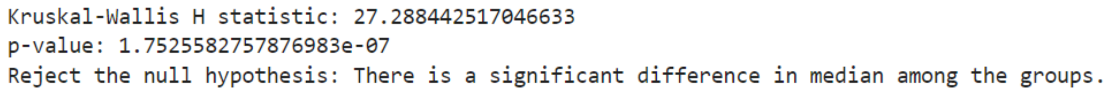
```
Spearman's rank correlation coefficient: -0.34339617058528193
p-value: 1.8426313553689049e-34
Reject the null hypothesis: There is a significant association between the two continuous variables.
```

- ➢ Actual productivity by department



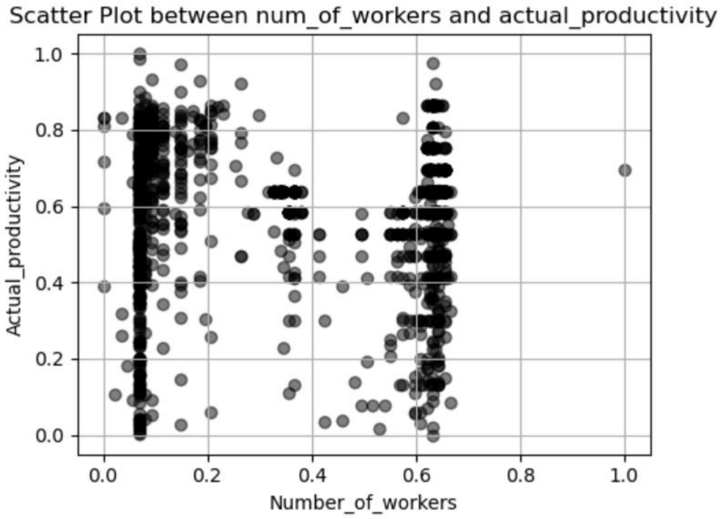
Hypothesis: Null Hypothesis(Ho) – There is no significant difference in the median productivity between two departments.

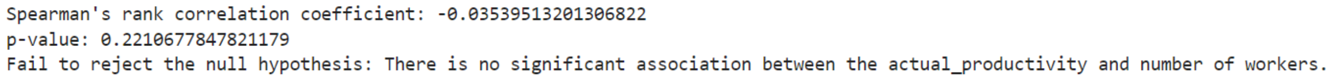
Alternative Hypothesis(Ho) – There is a significant difference in the median productivity between the two departments.



```
Kruskal-Wallis H statistic: 27.288442517046633
p-value: 1.7525582757876983e-07
Reject the null hypothesis: There is a significant difference in median among the groups.
```

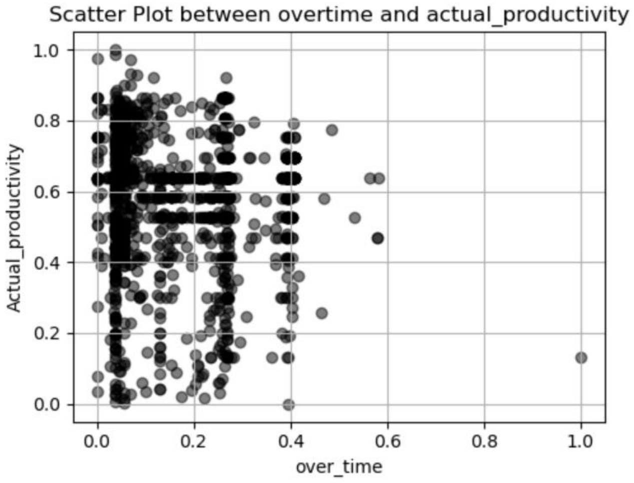
- ➢ Actual productivity and number of workers

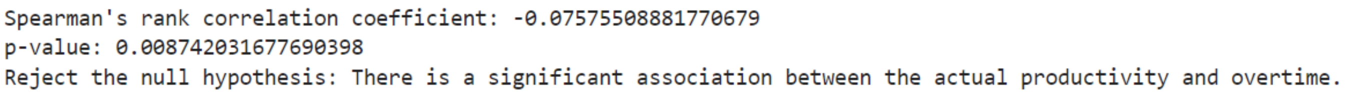




```
Spearman's rank correlation coefficient: -0.03539513201306822
p-value: 0.2210677847821179
Fail to reject the null hypothesis: There is no significant association between the actual_productivity and number of workers.
```

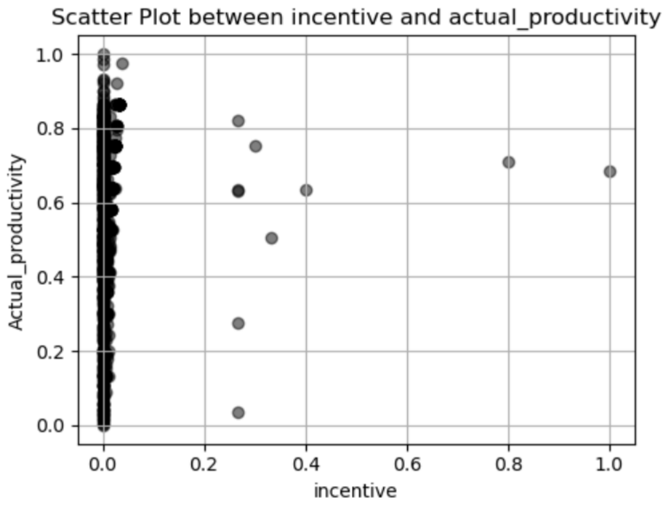
- ➢ Actual productivity with overtime

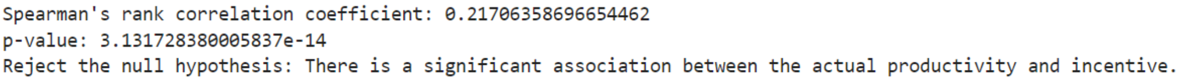




```
Spearman's rank correlation coefficient: -0.07575508881770679
p-value: 0.008742031677690398
Reject the null hypothesis: There is a significant association between the actual productivity and overtime.
```

- Actual productivity with incentive





```
Spearman's rank correlation coefficient: 0.21706358696654462
p-value: 3.131728380005837e-14
Reject the null hypothesis: There is a significant association between the actual productivity and incentive.
```

5. Model Selection and Training

1. Encode categorical variables.
2. Split the data into features (X) and target variable (y)
3. Split the data into training and testing sets

6. Model Evaluation

i) Choose the Linear Regression model

Model Evaluation - Mean Squared Error: 0.0215

ii) Choose DecisionTree Regressor

Model Evaluation - Mean Squared Error: 0.0181

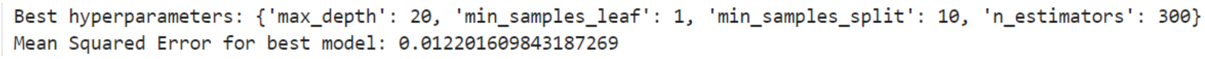
iii) Choose RandomForest Regressor

Model Evaluation - Mean Squared Error: 0.0122

iv) Choose SupportVectorMachine Regressor

Model Evaluation - Mean Squared Error: 0.015

- Hyperparameter Tuning - GridSearchCV



```
Best hyperparameters: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 300}
Mean Squared Error for best model: 0.012201609843187269
```

7. Results and Insights

- Actual productivity tend to be greater than targeted productivity.
- There is no difference in the median productivity of 7 days of the week.
- There is a difference in the median productivity of 12 teams.
- There is a difference in the median productivity between the two departments, ‘sweing’ and ‘finishing’.
- There is an association between svm and wip.
- There is an association between productivity and overtime.
- There is an association between productivity and incentive.
- Best fitted machine learning algorithm is RandomForest Regressor.

8. Code Repository

https://github.com/SenanayakeSMSN/ML-1