Garment Production: Building a regression prediction model

Mark Movh
March 2022

ABSTRACT

The aim of this report is to help generate insight through the creation of a regression prediction model. The model created hopes to accurately predict actual productivity of workers. The dataset from which the model is built is a real-life dataset collected from a garment production company, spanning across 3 months in 2015. Through the analysis of the dataset and selection of important features, 3 continuous variables were chosen. The data of the input variables was scaled through normalisation. The dataset was then split on 80% training and 20% testing. The degree of transformation, k was decided through a 5-fold cross validation from a created training model. The mean absolute error (MAE) and correlation of determination were examined. A degree of 2 was chosen, and input into the final model. The final model predicted values with an accuracy of 88%, while have a mean absolute error value of 0.011.

1 INTRODUCTION

The clothing and garment industry are one of the largest, and most profitable on the global scale. Profits may range in the billions as the industry must uphold to the needs and wants of a customer, while additionally keeping track of trending data. High consumption meets high demand, which implies a high supply. [1] Efficiency of the industry must be at its peak in order to satisfy the global demand, meaning that workers have set expectations on performance and delivery. The work of an employee or team of employees may be tracked, information collected analysed, and from such data predictions may be made. Such data allows for decision makers and top branches to see the productivity performance within different factories. Such information can be used to create a prediction model that can analyse and predict actual productivity of workers. In turn this may present information from data that is deemed relevant, informative and may bring insights to pursue change. The premise of this report takes these factors into account, and investigates a dataset on a garment factory in order to model a regression prediction model to determine productivity.

2 DATASET DESCRIPTION

The report follows a real-life dataset, confirmed and validated by industry experts. The dataset is called "Productivity Prediction of Garment Employees Data Set" based in 2015, and was collected from UCI's Machine Learning Repository [2]. The data aims to capture the industrial data, by focusing on a garment factory, and its processes. The dataset notes down performances almost every day, with dataset spanning over the course of roughly 2 and a half months. The data brought forth keeps track of several teams across different departments (in terms of manufacturing process) by comparing the resulting productivity to the target productivity. Each team was assigned a Standard Minute Value (SMV) and number of workers. From this several other pieces of data were noted ranging from number of products still being made, to the amount of overtime being done and financial incentive they are paid. The dataset the report utilises contains 1197 records and 15 attributes to help explain the end productivity result. Such variables are useful as they contain information which may be deemed valuable as an input to the regression model.

Several attributes exist within the data, ranging from categorical (non-numerical) to continuous (numerical). Categorical variables mostly referred to the time period, including date, quarter (5 quarters per month) and the actual day. Descriptive variables of the team included team (number of team), and department (either sewing or finishing). Continuous variables contained expectations and outputs. "targeted_productivity" (productivity target set by management for every day), "smv" (standard minute value – time given for a standard worker to complete a product) fall under expectations. Whereas outputs contained variables such as "wip" (work-in-progress products, that are unfinished), over time (by team in minutes), incentive (financial incentive in BDT), idle time (referring to some interruption causing a stall in production), idle men (represents how many workers were idle as a result of interruption). "No_of_style_change" means number of changes (if made) of style of some product. "no_of_workers" holds the number of workers on that specific team for that day, and lastly the "actual_productivty" which was the productivity output by workers in %. Such number of variables may provide insight towards how the actual productivity can be predicted, however since there are several attributes a data analysis must be performed to narrow down the variables that are most important. Additionally, the report focuses on a data analysis which will aim to answer the following questions:

(1) Does the individual teams' actual productivity exceed their targeted productivity? And which team is the most/worst productive?
(2) Which of the two garment production processes,between sewing and finishing, is the hardest?
(3) Which of the days is mostly the rest day for the workers and how does the resting affect their productivity afterwards?
(4) When does the company pay more incentive to its workers? And does both sewing and finishing departments enjoy the same incentives?

Before any statistical data analysis was performed, the overall data needed to be modified and transformed as there were found to be some missing features and errors. This included fixing a grammatical error where "finishing" and "finishing " departments were found. These were combined. Additionally, the department "sweing" was assumed to mean "sewing", considering this was a garment industry. Furthermore, the column "wip" had several values which were missing. This report concluded that considering the values always fell into the "finishing" department, the work in progress products were finished, and hence there was no value. Taking this into account, all of these missing values were replaced with a value of 0.

3 DATA ANALYSIS

The aim of this report is to build a regression model that would be capable of predicting the actual productivity of workers. In the above sections, the variables which exist in the dataset were mentioned. The model hopes to contain variables from both the categorical and continuous categories. However, variables that are only deemed relevant should be included. The data analysis section focuses on narrowing down the number of variables found in the dataset. As such both continuous and categorical attributes of the dataset are analysed. Firstly, the continuous variables were looked at. The dataset was limited and modified to contain only numerical variables which provided sufficient data. The data was normalised meaning that the data is scaled, and then input to a heatmap so that correlations may be visible between attributes. To be noted, "T_prod" relates to targeted productivity, whereas "A_prod" refers to the actual productivity. "num_workers" represents the number of workers. Heatmap of correlations can be seen in figure 1.

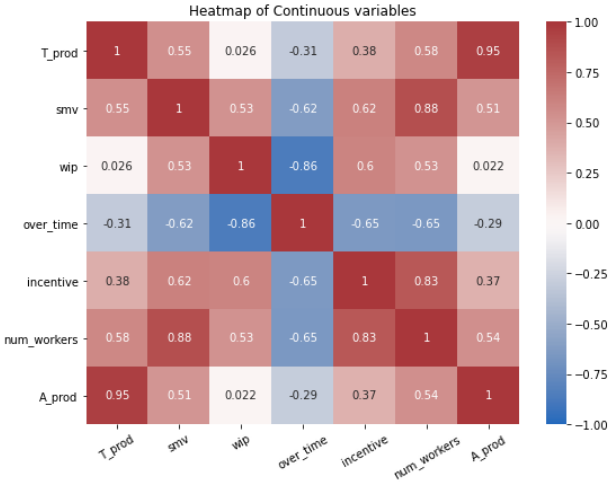

Figure 1: Pearson's correlations of all continuous variables in dataset

Firstly, three variables were omitted from the heatmap, and overall will be excluded from the model. These variables were idle men, idle time and the number of style changes. Due to the limited amount of data, and not enough difference between the data deemed these variables inadequate to be used in the final model. In terms of the heatmap, the continuous variables were normalised, and then calculated for Pearson's correlation coefficient. Generally, what can be seen is that there are several correlations, especially between the work environment features. Smv, wip, incentive and number of workers all impact each other in some way (all correlate above 0.5). Such results of the correlation are explained through the data analysis of the categorical variables. Overtime, interestingly had negative correlations with all other variables. Implying that it decreases as other increases. The more efficient the company is, it seems the less the overtime values are. However, the model mainly focuses on actual productivity, hence this is the most important variable to look at in terms of correlations.

Unlike the other variables, actual productivity did not correlate with all variables. Target productivity seemed to hugely have an impact on actual productivity (Correlation coefficient of 0.95), which may be coherent with the fact that it is used as an expectation and standard of what productivity should be. Target productivity in this sense would be a very important variable to the prediction model. Other correlations weren't as strong, but are still worth mentioning. Smv had 0.51, num_workers 0.54, and incentive lower at 0.37. While these are not necessarily strong correlations, it does not mean they do not impact data. By themselves they would not be good indicators of the actual productivity, however considering the correlations between them (num_workers and incentives's correlation is 0.83, and num_workers and smv is 0.88, and smv against incentive is 0.62) combining these variables and using them in one model may generate higher accuracies because of their strong relationships. Therefore, considering the findings of the heatmap, the following continuous variables were found to be most relevant and had the most sufficient results: targeted productivity, standard minute value, financial incentive and number of workers.

Regarding categorical attributes, several things were looked at to gain the most amount of insight. The first categorical aspect of the dataset that was looked at was worker performance. Specifically, looking into the 12 different teams and how their productivity values can resemble their performance scores. The reason for doing this part of the data analysis is to see whether there are performance differences between teams. Such information can show which teams are deemed most efficient, and should be worth praising in terms of incentive for example. The factory may also take notice of workers which are not carrying out a satisfactory performance, and need to be checked on more often in order to improve their scores. Regardless, the overall idea for checking worker performance is to improve productivity from the data, which may lead to a higher output ratio. A higher output can result in larger profits.

Firstly, before performance can be measured, there must be some criteria that can be followed. There should be enough distinction between teams to determine which is more effective, and which is less effective. In this report there are three things that are inspected between teams. Firstly, how many times to different teams go under targeted productivity, and how many times to they go over. Secondly, the average target productivity and average actual productivity are viewed in order to see a difference. Lastly, the ratio at which the teams go over may help determine efficiency of that team. The ratio being important as a team may have worked more than another, which would be the reason for one team having a higher number of days where they were more efficient than the other. The results of over and under productivity are displayed in figure 2.

The first figure shows the different counts of productivity between teams. What was found was that there were several teams that had a high number of over target days. Team 1 had the highest number, with 90 days being Over Target Productivity (OTP). Followed by team 4 (86), team 3 (84) and team 12 (83) close behind. From this it would be determined that team 1 was the most productive, however when taking into account the Under-Target Productivity (UTP) days the interpretations may differ. Team 3 had the lowest amount of UTP days at 11. Whereas team 1, which had the highest OTP days, had 15. From viewing this, it can be stated

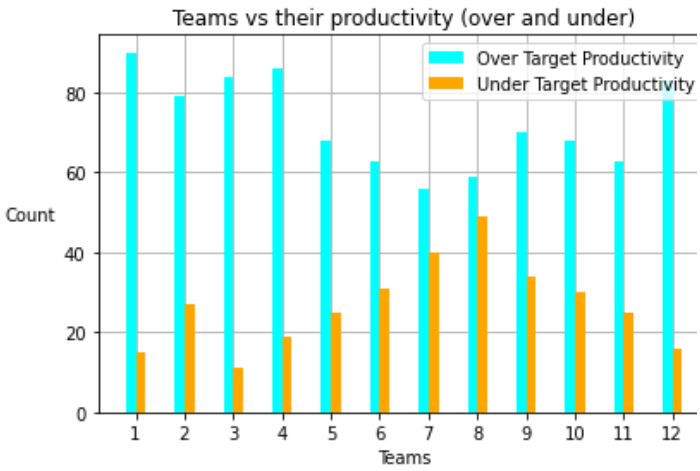

Figure 2: Productivity count of all teams, showing differences between teams

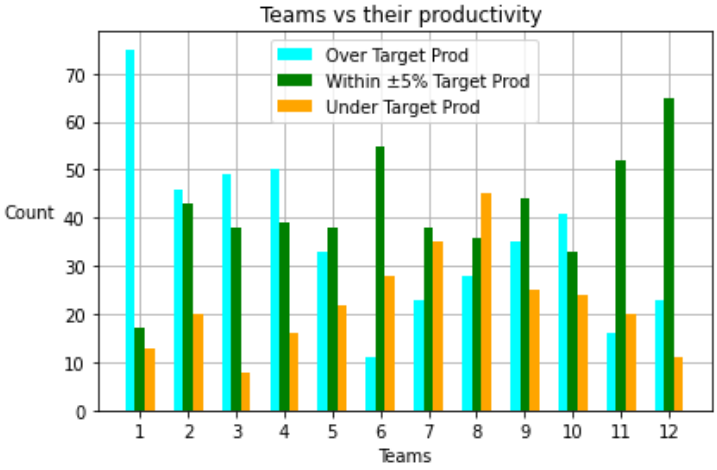

Figure 3: Updated productivity chart, including a range from target productivity

that team 1 and 3 seem to be most efficient workers. Team 7 and 8 on the other hand, from this graph, show to be the least productive. Team 7 had the lowest number of OTP days (56), whereas Team 8 had the highest number of UTP days (49). Taking these findings into account, teams 7 and 8 would be deemed least productive.

However, this graph was deemed insufficient to come to a concrete conclusion. The reason being that actual productivity is seldom exactly the same as target productivity. Only 6 accounts of this were found. Hence, it is usually a miniscule difference between them. For example, on one day Team 11 had a target productivity of 0.8, and their actual productivity was 0.80057. This number is misrepresented on the graph, as the graph implies large differences. Therefore, it was determined that another category be added to counter this issue, and to also bring further distinction between teams. A column would be added, where actual productivity was within 5% of the target. 5% was chosen as a measurement as this would imply a satisfactory difference to deem the team over-productive or under-productive.

Before, it would have been difficult to determine which team was deemed most efficient, and which least efficient. Team 1 and 3 were seen as most productive. Similarly, team 7 and 8 were deemed least productive. However, with this modification, there are clear differences. Team 1 had the highest Over Target Productivity count, barely changing. This further indicates that team 1 was the most productive team. With them being OTP on 75 of the days, 17 days being within +/-5% of target productivity and only 13 days of underperformance. From figure 3 it was found that team 6 was one of the least productive teams with this new adjustment. Being over-productive on only 11 days, while having one of the higher values of underperformance (28 days). However, even with this new information team 6, 7, and 8 all perform poorly and have high numbers of underperforming days (Team 7: 35, Team 8: 45). As mentioned, the means of target and actual performance were also looked at; with ratios of over performance included. To note, in table 1, AP means Actual Productivity, while TP means Target Productivity.

Table 1: Results of teams' average productivities, difference (actual – target) and the over target productivity rate

Team	Average AP	Average TP	Difference	OTP Rate
1	0.821	0.747	0.0743	0.714
2	0.771	0.74	0.031	0.422
3	0.804	0.742	0.0618	0.515
4	0.770	0.718	0.0524	0.476
5	0.698	0.674	0.0243	0.355
6	0.685	0.731	-0.046	0.117
7	0.668	0.714	-0.046	0.24
8	0.674	0.708	-0.0341	0.257
9	0.735	0.758	-0.0237	0.337
10	0.72	0.739	-0.0188	0.418
11	0.682	0.704	-0.022	0.182
12	0.779	0.774	0.00481	0.232

The results of the multibar chart in figure 3 can be compared against table 1. Team 1 was found to be most productive. The data in table 1 supports this premise, as it had the largest positive difference (0.0743) and contained the highest OTP rate (71.4% over-productive days). With such data it can be concluded that team 1 was the most productive team in the factory over the time period of the dataset. Team 6, 7 and 8 were highlighted for the teams that were deemed least productive. In the table presented, all mentioned teams have negative differences. Meaning they were on average less productive than they should have been. Team 8 had somewhat better results with difference being -0.0341 and an OTP rate of 0.257 (25.7%). Hence, the focus is between teams 6 and 7. Each had extremely similar differences, only deferring by a margin. The values were rounded up for the table; team 6 had an average productivity difference of -0.0459977, and team 7 had a difference of -0.0462653. The difference does not imply anything in this case. The least productive team was then determined by combining this with the OTP rate. Team 6 had one of the lowest differences, and the lowest OTP rate

(only 11.7% of days were over target productivity). As such, the data supports the idea that team 6 was the least productive team.

To answer the first question; it was found that every team exceeded their target performance several times. With team 1 containing the highest number of days where they exceeded their target, even with a range included. Furthermore team 1 was found to have the highest positive difference, and an over-productivity rate of 71.4%. Therefore, they were deemed the most productive. Given the criteria, and the process followed, the data collected showed that team 6 was the least productive team. They had one of the highest negative differences in productivity, and an OTP rate of 11.7%. As mentioned, there is reasoning for such analysis. The results demonstrate that certain teams are doing more than others. These teams should be awarded for their effort, whereas teams that underperformed should be notified and communicated with the proper expectations. Certain actions may need to be taken, and plans developed for further improvement. Then progress may be measured to ensure increasing productivity. Ultimately, what can be realised was that the actual productivity does differ between teams, and as such it may be an important categorical variable to include in the prediction model's input.

Different aspects of teams were further looked at. One of these aspects was the count for all teams and how many times they worked in a department. Through continuation of the analysis, it can be further confirmed that different teams perform differently, while also demonstrating how department plays a role. It is important to look into this as the two departments may have separate ranges for target productivity, which could mean that actual productivity is impacted by this. As a lower productivity would mean lower expectations.

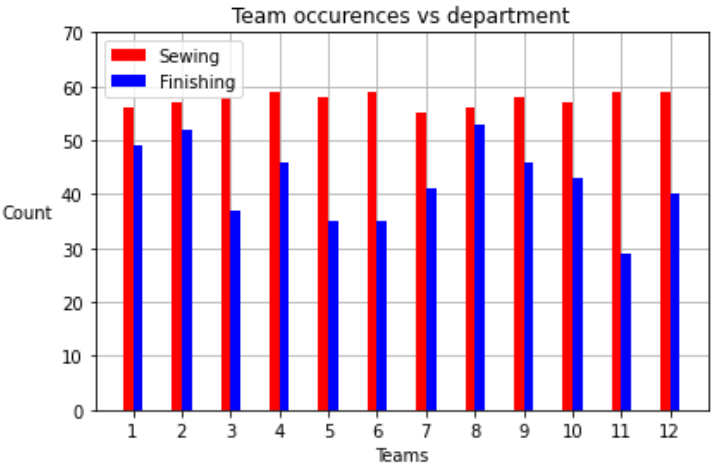

Figure 4: Chart showing how different teams have different count in sewing and finishing

Figure 4 demonstrates that in the sewing department, the numbers were roughly the same. There weren't any high or very low numbers. In the finishing department however, the numbers varied across the teams. Team 11 had the lowest count for finishing jobs, with there being only 29. The highest was team 8, with 53. If the productivity rates differ between departments, then that could explain why some teams witnessed a higher amount of productivity. These numbers could also be unfairly represented. As an imbalance in the actual occurrence of departments could explain the chart. It was found that the sewing department did have a higher count than the finishing department. Sewing was counted to have 691 rows of data. Whereas finishing only had 506 rows of data. Perhaps if the data was balanced and each department had an equal number of rows, the chart would look more balanced.

This implies that some difference does exist between finishing and sewing. A factor that can impact productivity is difficulty of a job. If a job is more difficult, then usually productivity would be lower, unless given the proper resources. To answer the second question, differences between the two departments were looked at in order to determine which job was the hardest. Two main factors were chosen to measure difficulty. The first was expectations. The higher the expectation of the department, the more difficulty it would be. Secondly, output. Expectations were the standard minute value, and the work in progress products. Output contains variables of number of products, amount of over time, and the number of workers. These 5 variables were compared between departments. If a large enough difference existed, then conclusions may be drawn or further interpreted. For all the factors mentioned above, the averages were taken. Table 2 contains these values.

Table 2: Averages of different continuous variables between departments

Factors (Mean)	Sewing	Finishing
Targeted Productivity	0.724	0.737
Standard Minute Value	23.25	3.88
Work In Progress Products	1190	0
Amount of overtime	6508	1917
Number of workers	52	10

From table 2, what was found was that all of the results except one varied quite largely. The standard minute value of sewing (23.25) was almost six 6 times greater than that of the finishing department (3.88). Indicating that a product, made by a standard sewing worker, requires much longer to complete than that of a standard finishing worker when working at standard performance. A sewing worker must work longer on a product, and can therefore be interpreted as being more difficult. In the work in progress attribute, finishing did not have any values. This analytical report interprets the dataset, that since the department is called "finishing", they are completing the product and hence have no products in progress. Hence since the actual number of products being made/completed is not shown, it would not be a fair comparison to include this value.

On average it was found that sewing had more overtime than finishing (sewing – 6508, finishing – 1917). However, the sewing department also had a greater number of workers, which would mean that the overtime may be greater because of the number of workers. This was further investigated by getting the average overtime per worker. It was found that a sewing worker, works on average 180 minutes of overtime. Whereas a finishing worker works on average 125 minutes of overtime. The difference between

these values implies that a sewing worker works more overtime than that of a finishing worker. All values had some large difference, however targeted productivity had almost no difference.

Interestingly, it appears that the average targeted productivity is roughly the same between the departments. This at least seems so from looking at the data from the table. However, the distributions of these productivities are also important. As an average may be similar, but the spread of data may not be, which may indicate the same results, but a difference may still exist. To confirm whether a difference exists or not a 2-Sample t-test was carried out to test for significant difference. Two hypotheses are brought forth, H_0: No difference exists between the two means; and H_1: There exists som difference between the two means. Mathematically, a significan difference is determined when the null hypothesis (H_0) is rejecte

Table 3: Variable inputs to calculate the z-score betwee sewing and finishing

Department	Mean	Standard Deviation	Sample
Sewing	0.724	0.102	691
Finishing	0.737	0.0915	506

The Z-score in this case is used due to the abundance of dat The significance level is set at 0.05 / 2 = 0.025, with the critical rang being +/-1.96. If the calculated z-score of the t-test goes above o under this limit, some difference between the averages exists. Us ing the inputs from table 3, the z-score was calculated to be 2.3 Since the z-score is larger than the critical range (1.96), the null hypothesis is rejected, as there is not enough sufficient evidence. The calculation in this case implies that there is a significant difference between the average targeted productivities. With this calculation, it can be concluded that enough data has been collected to come to a conclusion on which job is more difficult.

A question that the report focused on was to determine which garment production process was the hardest. A criterion was created, taking into account department expectations and output variables. The averages were calculated between the two department so a difference could be seen. The data collected supports the premise that the sewing production process is deemed more difficult. The standard minute value is higher, implying longer time must be spent creating a product, which results in greater overtime values. A larger number of workers is required to keep up with the production demand, while often also set with high targeted productivity expectations. This information can help manage productivity expectations, while also maintaining compensations as these factors can help improve the overall productivity of the processing line. In essence, through visual difference between the averages across several components, and a 2-sample t-test to confirm significant difference between productivity, it is evident that sewing requires larger demand, time and overall resources.

With this in mind it is important that workers are not overworked. A worker should be capable of working at standard performance, with the target productivity in mind. Hence, proper rest is essential to this. The next categorical variable exploration was into different days, and how they may impact the productivity of the worker. Through such an analysis it is possible to determine how efficient workers will be before rest, and after rest. The reason for this investigation is because it can provide information on when workers are most productive, and how targeted productivity may be adjusted to these values. It was found that workers never worked on Friday. Hence the report determine that this is the workers rest day. However, there was an occurrence of a Saturday being taken off as well. It is not certain whether this data is missing, but it was taken into account as an extended period of rest. The first part of the analysis was to look into average productivity across different days. With keeping specific focus on the days before and after Saturday.

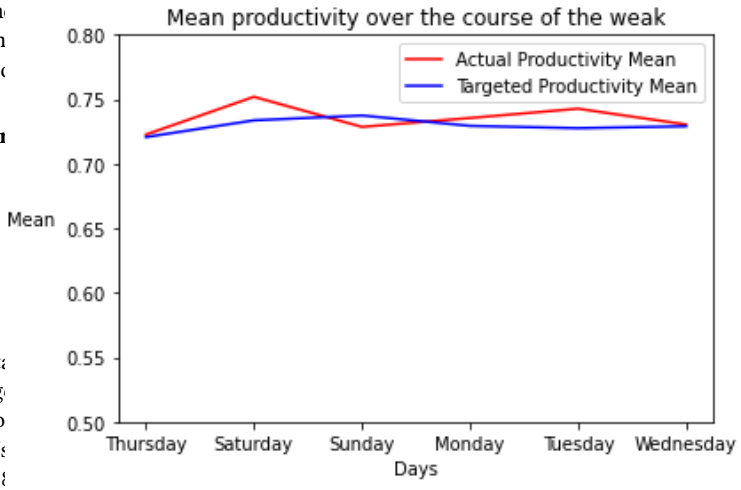

Figure 5: Actual Productivity compared against Targeted Productivity, on average across the week

All the data was grouped into 6 days, excluding the rest day. A line chart can be seen on figure 5 which shows the productivity over time across the week. To be noted, the y-axis was limited at 0.5 such that the difference between target and actual productivity may be better visible. In reality, the difference between these values is miniscule. While a rise is seen from Thursday to Saturday, the day after the resting period, the difference between these two is only 0.018, or 1.8%. Such data would show some slight increase in productivity, and then a decrease from Saturday to Sunday (-0.88%). Drawing any conclusions would in turn be redundant, and further exploration would be necessary. As target productivities and actual productivities vary across different days, a more focused and precise analysis could bring out further information. In term, the average productivities were looked at throughout the entire course of the dataset, spanning across 59 days.

This way, any surges can be monitored and better interpreted, than the limited perspective of figure 5. Figure 6 shows that target productivity is often very close to the actual productivity. With some exceptions occurring with larger differences in the negatives and positives. When looking at the days after the resting day, these being Saturdays most often (one Sunday), the difference between the target and actual productivity was 0.0052 or 0.52%, indicating some improvement from resting. However, the following day's average (2 days after resting), the average difference was -0.009 or -0.9% showing that productivity generally did not seem to improve. 3 days

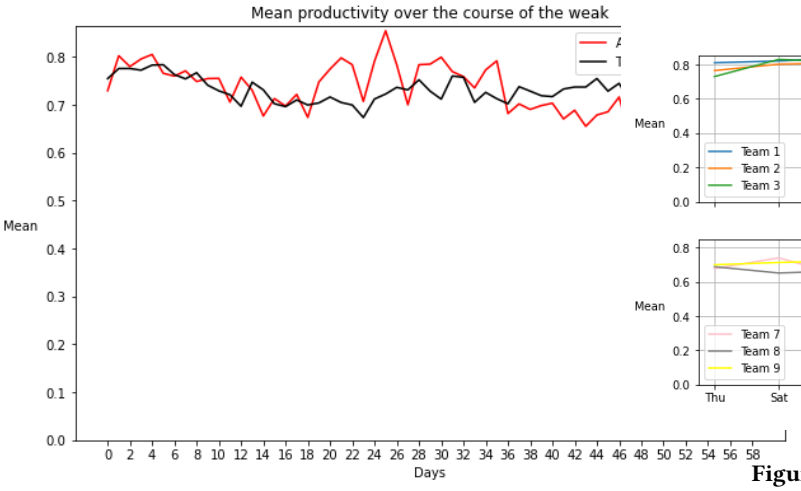

Figure 6: Comparison between productivities shown over all days of the dataset

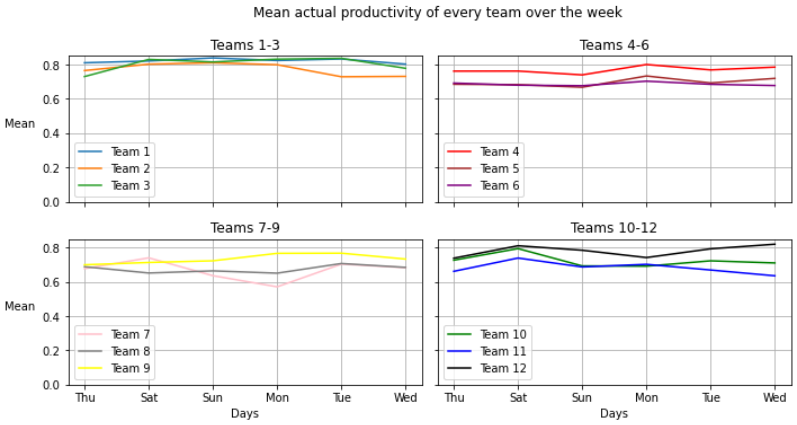

Figure 7: Average actual productivities of all teams across the week

after the difference was 0.89%, showing some more productivity. Contrary, the data did not show any patterns that may indicate that resting impacts worker performance.

One interpretation of such results is that the average difference between the productivities could potentially be a result of management setting the target productivities inadequately. For example, the target productivities may be set too high on the days right after Friday, which implies that expectations may need to be lowered to correspond better to the output productivity of workers. The values between days 19 and 25 best present an ideal week in this case. As the week is goes on, targeted productivity should be decreasing and then after the resting day, the target should increase, if the rest shows that actual productivity does increase. However, even with the data collected from the graph there is still no certainty that productivity increases as a result of the resting day. On the other hand, it could imply that other variables have more impact on productivity. The data analysis earlier demonstrated that performance differed between teams. As such, target productivities varied, as did actual productivities. It was determined that taking this into account and checking productivity of all teams across the week might provide some insight and help finalise a conclusion on what impact rest has on productivity.

The teams were split up into 4 different sections in order to better view the results and avoid clashing lines and misinterpretations. Targeted productivity in this case was omitted as only the productivity after the resting day matter. What was found from the graph was that most teams did witness an increase on Saturday, or having a slight rise. Only three teams were found to go below the average productivity on Thursday (Teams: 5, 6, 7). Team 3 witnessed the greatest difference of 0.1 or 10%. The average increase, between positive teams, was found to be roughly 5%. From Saturday to Sunday, mostly all teams witnessed productivity decreases. Therefore, if a conclusion was to be interpreted from the data in figure 7, it would support the idea that generally productivity increased after the resting day, but only for Saturday. Beyond Saturday, the productivity decreases and stabilises for most teams.

Ultimately while the discussed line chart does demonstrate some change in productivity, the previous collected data in figure 5 and 6 show almost no increase from resting. Therefore, to answer the third question, it was found that Fridays are mostly the rest days of the workers. A data analysis of three different perspectives was brought forth to attempt to see how productivity changes after the resting day. The analysis exhibited that there was often no change when compared to average across week, and when across all 59 days. Contradictory, when including teams, some increase was witnessed. It may be concluded that while some increase is visible, the premise of certain change after the rest day remains vague. This is not to say that a change does not exist. The analysis in this report witnessed almost no change. Further investigation would need to be conducted, with other variables checked to determine how resting may impact productivity.

The last question focuses on incentive, and how a difference may be witnessed between the two departments. Furthermore, it explores when the company will pay more incentive, meaning give financial payments to encourage workers. Knowing when to pay incentive is important as it can let the worker feel properly accommodated for their efforts. The analysis aims at finding certain factors in variables, whether continuous or categorical to see some pattern of when incentive is higher or lower. With such information a company may better adjust their incentive payments, or search for outliers that may have been too high. Adjustment and improvement of the system can better increase productivity as workers can be better motivated. Firstly, the two different departments were checked. Results of the exploration is visible in table 4.

Table 4: Incentive information between finishing and sewing

Department	Number of Incentives	Sum of Incentives
Sewing	583	30739
Finishing	10	15000

The data was found to be fairly uneven. Out of the 691 rows that were under the sewing department, 583 of these got paid incentives. Whereas the finishing department only got paid incentives 10 times. On the other hand, the incentives for the finishing department were much larger, ranging from 960 to 3600. The sums also differ between departments, with finishing having little under half of what sewing got. Such data reveals that finishing and sewing do not enjoy the same incentives, with sewing getting better benefits. Workers under sewing collect incentives more often, and they should collect more. In the above sections, there was an analysis that determined which department was hardest. The conclusion was reached that the sewing department faced a more difficult task. Larger amounts of overtime, lengthier tasks (bigger smv values) and overall higher number of workers. As the job is more taxing and difficult, the sewing department should enjoy more incentives. The results in the table are justifiable by prior acquired information. However, it is still important to understand when the company pays more incentive.

Productivity was the first variable to look at. Mainly when the target productivity is exceeded. It was found that 535 incentives were paid to the workers when the actual productivity was higher than the targeted productivity. 53 incentives were paid even when the actual productivity was lower. Showing that productivity has some effect on when incentives are paid, but it doesn't necessarily state when the payments were larger. In the heatmap from figure 1, correlations were checked, and there did not seem to be any very strong relationships between any of the continuous variables. Except with overtime having a correlation of -0.65. Why this may have occurred is explained further below. With no very strong correlations, the categorical variables were checked. Firstly, the individual days were checked. It was found that Monday had the largest average incentive payment, at 100, whereas the lowest was Saturday with 24. All other days ranged between 25 and 27. Furthermore, some difference was also found between teams. Team 9 on average got an incentive value of 60. Team 7, being the lowest at 16. The values were more balanced in this case, but still showing a great difference. Lastly quarter variable also showed that quarter 2 had the largest incentive payment, with 70 whereas quarter 3 was the lowest at 24. Such large difference may seem to show when the company will pay more incentive, but the explanation could be the same reasoning as the negative correlation with overtime.

The finishing department had 10 incentive payments, as mentioned these were extremely large. When investigating into further detail, it was found that these were not spread out, and all occurred within the same day, and quarter. Hence explaining why Monday and quarter 2 had such large averages. As there were only 10 payments, only 10 teams enjoyed these benefits. Team 9 had the largest average because they are the team that got paid 3600. Whereas Team 7 did not get anything. When these values occurred, the overtime values were set to 0. This could be an indication of why the relationship was -0.65. It was found that when these seeming outliers were removed, new information was generated. There was no longer a difference between days. All averages ranged between 51, and 55, with Monday being the lowest at 51. Regarding teams, Team 1 was paid higher than the rest. With an average of 73, with the lowest being Team 7, followed closely behind by Team 6. These values may be better explained, as the analysis from before showed Team 1 being the most productive, while Team 6 and 7 were one of the least productive and efficient teams. The data collected from removing the present outliers, shows that to some extent productivity may represent incentive payments. Saturday's average was higher than the rest, and on average Saturday was more productive than usually. However, there isn't enough evidence to fully support such a conclusion, therefore it is difficult to ultimately answer when the company paid more incentives.

Additionally, the quarter variable was found to have an increased incentive payment in quarter 5. However, when checking actual productivity, this was not reflected. Quarter 5 was the only quarter in which the productivity was under the target. Taking all of the above-mentioned regarding incentive, the conclusion depends on the interpretation of data. A company paid more depending on day, and quarter. At least that is what the data found stated. Yet when removing the large outliers, it was only found that payments differed between teams and quarters. However, neither of these points had sufficient data to come to a solid conclusion. It may be interpreted in different ways, and hence more a extended data analysis would be required. Regardless of when the company pays more, the data displayed a difference between departments, indicating that that the sewing departments enjoyed incentives more often, and overall larger sums. Taking these results into account, and with the fact that one production process was deemed harder than the other, the department variable was seen as an important inclusion to the input of the prediction model.

In essence of the data analysis section, various columns of the dataset were gone over and analysed. These ranging from continuous to categorical, in order to determine which had the most relevance and would be most suitable to include in the regression prediction model. These were mainly determined on how they impacted the actual productivity, as that would be the predicted variable. The chosen variables at the end of the analysis were determined: (Continuous) target productivity, standard minute value, financial incentive, and number of workers; (Categorical) department, quarter, and team. While these variables were chosen to be most relevant, in terms of interpretation of the data analysis, they can be further narrowed down for a more accurate model.

4 FEATURE SELECTION

At this stage, the data was almost ready to be input to the prediction model. However, while the variables mentioned above were interpreted to be relevant, the feature selection further focuses on selecting features that are deemed most important through the use of a random forest. The reason for this is that one variable can have a larger impact on the actual productivity than other features, hence the model should only include them if they are most important. Firstly, the data was transformed. The categorical variables were split into numerical values by incorporating them into a binary format. For the prediction model, the data would be normalised. The choice behind this was because it wasn't certain whether the data followed a Gaussian distribution. After the transformations, and normalisation of all data, the importance levels of all input features were calculated. Results of these calculations are visible in figure 8.

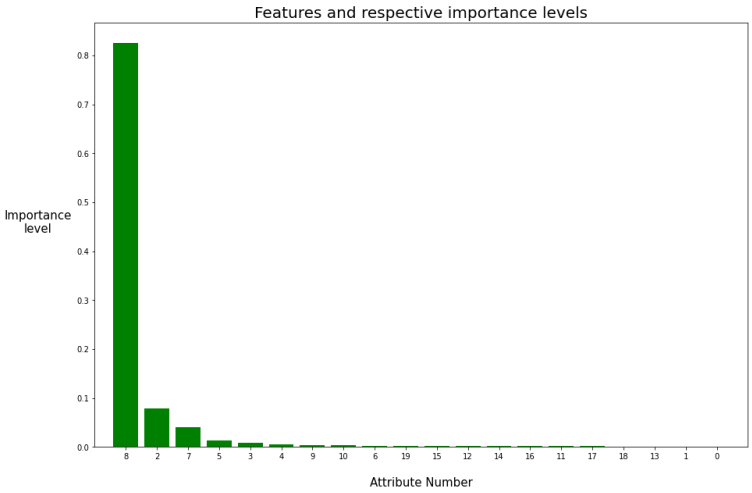

Figure 8: Importance level chart of all input variables from transformed and normalised data

As can be seen in the figure, there are a total of 20 variables/features, for which the importance level was calculated. From the chart it was clear that 1 feature had the largest importance level, with other hardly having an impact. Targeted productivity was the highest scoring feature, with its importance level value at 0.826. Considering its extremely high correlation on the heatmap, the result is logical. This premise can be further confirmed as the 2nd and 3rd variables are smv and no_workers, with each of these having a correlation. The scores of these two features, along with the rest, may be visible in the table presented below. Such a process is carried out because as was shown, certain variables are deemed more important. In the case of this dataset, the continuous variables help explain the actual productivity far better than the categorical. Aside from incentive, all other continuous features ranked high. Whereas department for example, was shown to be least important. Perhaps this could be the result of such immense differences between finishing and sewing, and would also explain the low score of incentive. Considering the results of the graph, it was deemed that only the first 3 values would be used as final inputs to the model.

To conclude the feature selection section, three variables were chosen. Targeted productivity, smv and no_of_workers. The rest of the features were omitted from being used in the final prediction model, as there was an insufficient amount of data to support keeping them. The 3 mentioned features had the largest importance in terms of the data, despite targeted productivity being vastly higher, it was deemed important to keep the other two variables as they had some potential value. Combined with the correlation of the heatmap, it was theorised that retaining these would further improve the accuracy of the regression model. Keeping this in mind, continuous variables were overall found to have the greatest influence on the data, despite the data analysis showing some relevance of categorical variables. These were found to be less important. The conclusions deems that the continuous variables mentioned should be focused on, in order to best interpret results. However, the report does not take into account any outer factors that may have impacted this data. The feature selection section did not delve into or consider such a component. The thought process and critical reasoning is only based on the statistics and data found from analysing the given dataset.

Table 5: Input features and their corresponding importance level rankings and values

I-Ranking	Feature Number	Variable Name	I-Level
1	8	Targeted Productivity	0.826
2	2	Number of workers	0.0789
3	7	Standard minute value	0.0411
4	5	Quarter 4	0.0134
5	3	Quarter 2	0.00811
6	4	Quarter 3	0.00499
7	9	Team 10	0.00333
8	10	Team 11	0.0031
9	6	Quarter 5	0.00279
10	19	Team 9	0.0025
11	15	Team 5	0.00236
12	12	Team 2	0.00232
13	14	Team 4	0.00231
14	16	Team 6	0.00217
15	11	Team 12	0.00193
16	17	Team 7	0.00184
17	18	Team 8	0.0014
18	13	Team 3	0.00118
19	1	Incentive	0.000349
20	0	Department	0.000226

5 MULTILINEAR REGRESSION MODEL DISCUSSION

Finally, the regression prediction model can be built in order to predict actual productivity of a worker. Since there are several input variables, this is classified as a multilinear model. Although there are mentions of multilinear models, the report does not focus on linearity only. Polynomial features are also explored. It can be stated that the model focuses on polynomial regression in this case. The accuracy of a regression model is based on the best fit line calculating distances between all the points. The line in this case may be transformed into a polynomial if the degree that it was transformed to goes above 1. Before a final model can be created it must be known what degree this should be, as this will provide most sufficient results.

A training model, and a final model need to be created. For this purpose, the dataset was split up into two sets; training and testing. Training consisted of 80% of the data, while testing contained only 20% of the data. The training model will apply the training set. In order to test the accuracy of the final mode, the testing set is utilised. The training model will help determine which is the most suitable degree of k. This k being to what degree the best fit line is transformed into an nth degree polynomial. To determine different scores based on different degrees of k, a cross validation method is applied. This way, the training data is further split up into larger numbers of training and test sets. This is done to measure the

outputs as accurately as possible. The training model would utilise the cross validation on a 5-fold split, and check two factors. Firstly, the accuracy of the model (r2), and secondly the mean absolute error (MAE).

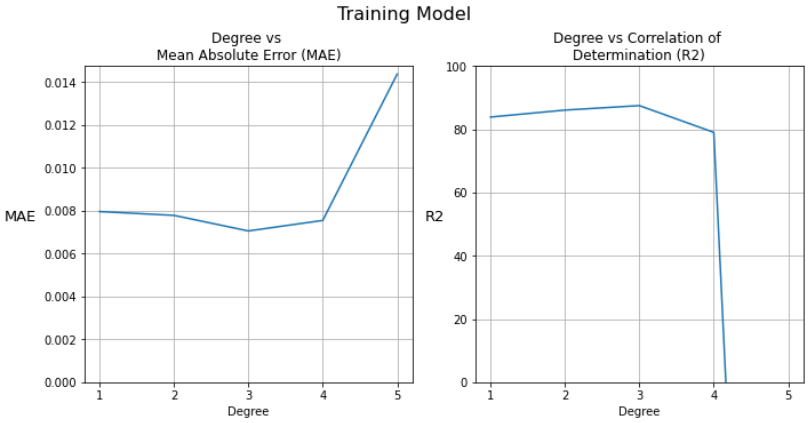

Figure 9: Mean Absolute Error and Correlation of Determination of the multilinear training model

Generally, the Mean Absolute Error and correlation of determination roughly stayed the same as the model was still simple. When the degree of k was increased, meaning a more complex model had to be created, the MAE and R2 values both increased and decreased tremendously, causing decreasing accuracy. From these results, it can be concluded that a more complex model in this case is not favourable. A simpler model with better generalisation capabilities was deemed best. Considering the graph, it was deemed sufficient to select either the degree of k=1 or k=2. The results in these values were stable before the changes. This report selects k=2 as it offers slight decrease in MAE and a slight increase in R2. The final model utilised both the 80% training set and the 20% test set.

Table 6: Final Model results compared against training model

	Training Set		Testing Set	
Degree of k	MAE	R2	MAE	R2
2	0.00779	86.11	0.011	87.99

Three variables were input, with the actual productivity prediction accuracy being measured. From table 6, it is possible to see results of the training and the final model. There are two important things to highlight. Firstly, the MAE is higher on the final model (0.011) than on the training model (0.00779). This difference could be because the training model had a larger dataset to better pinpoint its accuracy. However, it was not more accurate than the final model. The final model had an accuracy of roughly 88%, whereas the training model was slightly below it at approximately 86%. It was a minor improvement from the training model. Such a result may indicate cross validation error. Ultimately, what may be concluded is that the polynomial regression model presented above demonstrates that it is capable of predicting the actual productivity

Table 7: 10 results of the final mode. Predicted values compared against actual values

Index	Actual Value	Predicted Value
679	0.102035	0.076057
893	0.012685	0.019484
865	0.095592	0.076859
736	0.008654	0.008251
420	0.048998	0.046167
884	0.046689	0.046379
286	0.010027	0.010633
970	0.009579	0.009320
567	0.063040	0.066266
294	0.008848	0.009517

correctly 88% of results. The three input variables allow for an accurate regression analysis, which the garment production company may utilise to predict the productivity of workers. A sample of the results may be found below.

6 LIMITATIONS

The discussion of the model brought out several thoughts regarding conclusions and interpretation of the results. While decision making regarding variable relevancy, and feature selection were made, they may exist certain limitations that should not go unnoticed. The incentive column for example was found varying in values between departments. There were many other large differences between "finishing" and "sewing" which could have potentially negatively impacted the model. To mention a few; "smv" is low for finishing, but higher for sewing. The finishing department had no "wip" whereas sewing did. Lastly the number of workers were lower for finishing than for sewing. Such differences are brought up because the actual productivity between these departments was mostly similar, and it would be difficult for predictions to be made based off input value.

Despite the large accuracy, there could have been other mistakes made in the actual method of choosing and selecting features. The data analysis section aimed at narrowing down the useful variables, however some methodical mistakes could have been made. The feature selection process followed a more mathematical approach, collecting the importance of every input variable. The data analysis section referred mostly to interpretation of one's own exploration. Despite efforts of attempting to be as detailed as possible, certain factors or information may still have been missed. Missing something crucial in the variables would lead to different results in the final model. Several complications were discovered while creating the prediction model. Consistency in the results wasn't transparent, and would be shifting a lot. This was especially visible through different scaling methods bringing different r2 scores. Taking everything mentioned above in this section, the results of the prediction model may be open to speculation, and the credibility of its accuracy left to be questioned.

7 CONCLUSION

To conclude this statistical report, the process of creating a prediction model consisted of finding relevant variables and extracting

insightful information. Inputs were collected via feature selection to ultimately be used to predict the actual productivity of workers. A data analysis section was written to describe the investigation into the dataset. The results of the sections regarding various aspects of the dataset were evaluated and interpreted, in order to extract the most relevant variables. To further justify the input of the model, a feature selection process was presented which calculated the importance level of each relevant variable. From the feature selection three continuous variables remained. While targeted productivity was the most important variable; standard minute value and number of workers were found to have some effect on the output. To build the multilinear regression model, the dataset was split into training and testing sets. A training model was developed through the training set, and the use of 5-fold cross validation. Through this the most ideal degree of k to transform the multilinear model to polynomial was found. The final model was used on the testing set with this degree. It was found that the final model's predictions were highly accurate, with having some slight improvement over the training model. Certain limitations were brought to light, and mentioned. While such limitations may be used as quality control, the report concludes that the presented prediction model is deemed accurate in predicting the actual productivity of workers. Through its utilisation the garment company could better adjust its targeted productivity, among other features in order to gain valuable knowledge on how to achieve the best actual productivity.

REFERENCES

[1] Abdullah Al Imran. 2020. Productivity Prediction of Garment Employees Data Set. https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees

[2] Danijela Paunovic. 2012. Foreword. In *Strategic Management in the Garment Industry*, Gordana Colovic (Ed.). Woodhead Publishing India, vii–viii. https://doi.org/10.1016/B978-0-85709-582-4.50009-9