## Machine Learning: Predicting Worker's Productivity in the Apparel Manufacturing Industry

## INTRODUCTION

The garment manufacturing industry is characterized by being labor intensive. Even though the apparel industry is a multi-billion-dollar industry, it still relies on human labor to complete most of the tasks needed to produce a garment. As described in the article "Impact of increasing labor costs on apparel supply chains": "labor cost have always been an essential concern for garment companies"(https://finance.yahoo.com/news/impact-increasing-labour-costs-apparel-131538775.html). These costs determine where retailers and apparel brands allocate their productions. Building a model that determines the right allocation of workers is key for business sustainability. We are given data, and we will proceed to create a model that determines worker productivity, helping to determine worker allocation and creates economic value.

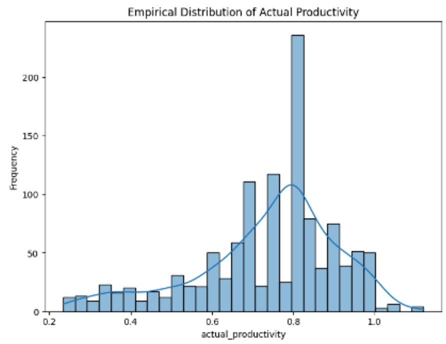
## DATA PREPARATION

After loading the data, we proceed to do some data preprocessing.

Initial assumptions involved data accuracy, identifying a productivity indicator, and binary classification allocation to facilitate the classification model to be applied to Client B. We began by dropping columns such as "wip, date, and idle_men". Satisfied with the remaining variables we identified that "actual_productivity" would be an appropriate and reliable measure of employee performance. For binary classification, we assumed that productivity levels could be categorized as high for those above 0.8 for Client B.

We scan our data for any missing values, from our data we see that the work in progress variable has significant missing values, therefore we proceed to remove this variable from our data set. Additionally, since workday and quarter are already captured in different features, we can delete "date". 'Idle_men' was another feature that we dropped, since it did not contain extra useful information. For better model fit, Categorical values such as 'quarter', 'department', 'day', 'team', 'targeted_productivity', 'no_of_style_change' were transformed utilizing one-hot encoding. Also, we standardized numerical values with "StandardScaler".

It is useful to analyze the data distribution to graph the data. For this graph, we are using a histogram. The histogram shows the 'actual_productivity' data on the x-axis and the frequency (amount workers) on the y-axis. From the histogram we can conclude that the data is normally distributed, as it follows a bell-shaped curve. From the graph, we can also observe that many workers' productivity ranges from approximately 0.7 to 0.85. Our graph is also slightly right skewed, meaning most workers have higher 'actual_productivity'.




In order to verify which variables had a higher correlation, we plotted a correlation matrix. The matrix shows us the correlation of 'actual productivity' and other variables. The features that showed the highest correlations with 'actual_productivity' were:

- 'idle_time', more lost time will impact productivity negatively,
- 'incentive', has a positive correlation with 'actual productivity' this is intuitive as higher incentives will motivate workers to obtain higher productivity,
- 'number of workers', although it is not a strong correlation, it can suggest that the larger a team is the better the team's productivity."

## MODEL SELECTION

Objective: Finding a model that can predict worker productivity best.

### Modeling for Client A

Random Forest model including all features, shows the best model performance. This model has the highest R-squared and the lowest MSE when compared to other models. Having a low MSE means that the predictions made by the model are closer to the values in 'actual productivity', making Random Forest Regression model the most accurate between tested models. The second metric we can focus on is R-squared. Random Forest captures better than the other models the factors that influence productivity, explaining 44.3% of variability in the productivity of workers. A high R-square means our model is capturing a fair amount of the relationship between the features and the label. It is also important to mention that Random Forest avoids overfitting automatically.

| Model for (A) | | | |
|---|---|---|---|
| Model Selected | Features | MSE | R-Squared |
| OLS Regression | All variables but 'actual_productivity' | 0.02586 | 0.13657 |
| OLS Regression | 'smv', 'over_time', 'incentive' | 0.02579 | 0.02888 |
| Random Forest Regression | All variables but 'actual_productivity' | 0.01479 | 0.44305 |
| Lasso Regression | All variables but 'actual_productivity' | 0.03616 | -0.00188 |
| Lasso Regression (Adjusting for Best Alpha) | All variables but 'actual_productivity' | 0.03207 | 0.11143 |
| Neural Networks | All variables but 'actual_productivity' | 0.04560 | -0.26336 |

## Modeling for Client B

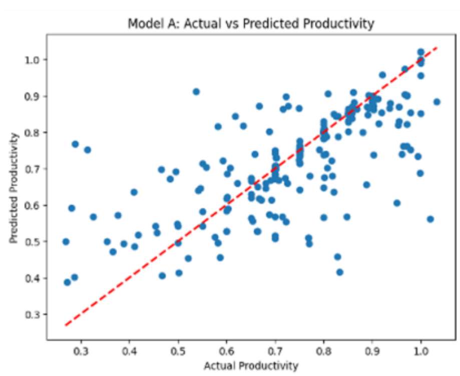
For Client B, our goal is to predict if a worker's annual productivity exceeds 0.80. This is a classification problem. We will proceed to test out classification models to find the best result.

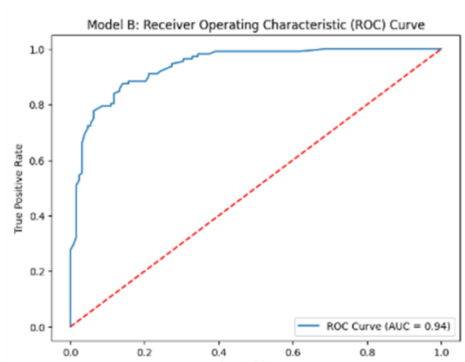
The Random Forest Classification model shows superior metrics, achieving an accuracy of 0.8625, this means that the model is correctly classifying workers based on predicted productivity. This can be explained by the fact that random forest models can capture complex patterns in datasets. Another metric we can review is the ROC-AUC, this metric indicates if the model can effectively distinguish between classes, for this case productivity bigger than 0.8, the model has an AUC-ROC of 0.9388, meaning it does an effective job at distinguishing between workers productivity classes.  These two metrics indicate the model's good performance, at predicting non-productive workers.

## REVENUE CALCULATION

Once we define our models and make predictions using test data, we can calculate the revenue per worker obtained from Client A and Client B.

- Client A Revenue per worker: $87.83
- Client B Revenue per worker: $113.54



**PLACEMENT DECISION**

As we previously discussed, the garment industry is highly labor intensive. Therefore, correctly predicting labor productivity is key to being profitable. Considering that each worker’s cost is $70.00 we must decide, aided by our models, where to place each worker.

The following decision rules have the objective of maximizing profitability, placing workers only when the expected profit exceeds the placement cost.

**CLIENT A**

Using our Random Forest Regression model, we can decide if we place workers for client A based on their predicted productivity.

Revenue for a worker is obtained from calculating the predicted productivity * $120 minus the placement cost of the worker ($70.00). A worker will be placed if the result, the expected profit is larger than 0.

**CLIENT B**

Our objective was to predict with our Random Forest Classifier model if a worker's productivity exceeds 0.80.

The decision if a worker is placed for client B, is based on the predicted probability of that working achieving a higher efficiency than 0.80. The revenue is the multiplication of the predicted probability of exceeding 0.8 productivity times $250.00 minus the placement cost. A worker will be placed for client B, if the profit exceeds the cost of placement for the worker.

**RESULTS**

- For Client Type A, the number of workers placed equals 198, with a total profit of $4,669.10. The Average profit per worker placement is $23.58.
- For Client Type B, the number of workers placed is 137, with a total profit of $16,410.00. The Average profit per placement: $119.78.
- The prediction models demonstrate financial value, by placing workers where they are more likely to generate profits based on predicted productivity and our decision rules. The model allows for efficient resource allocation.

**RANDOM PLACEMENT**

To prove that the use of our model and our decision rules creates financial value. We will randomly generate results to compare to the profits obtained from running our model.

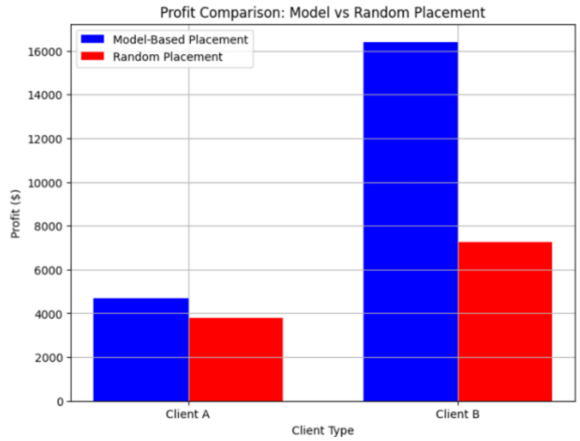
Random Placement Results:

- Client A - Total profit: $4,027.82

- Client B - Total profit: $6,690.00

Model Value:

- Client A - Model adds $641.28 in value
- Client B - Model adds $9,720.00 in value




**CONCLUSIONS**

The following predictive models were utilized in productivity and profit analysis of our Garment Manufacturing client:

- ***Random Forest Regressor*** to forecast worker's productivity for *clients of type (A)*
- ***Random Forest Classifier*** to assess the likelihood that this productivity exceeds 0.80 for *clients of type (B)*

After the creation and deployment of our models, we proceeded to place workers based on our decision rules with the objective of maximizing profitability. To prove our models efficiency we lastly analyzed our model’s financial value by comparing its results to a random placement of workers. We can conclude our models do provide financial value, creating more profit than random worker selection.