BIG DATA MANAGEMENT

Final report

Group 5

Yan Naing Oo

Linh Cao

Agenda

- Business Problems & Approach
- Data Preprocessing
- Data Exploration
- Model Fitting
- Model Evaluation
- Summary & Limitations

Business problem & approach

Business problem

Our project involves analyzing **garment workers productivity data** using Python, PySpark and machine learning models to get insights into factors that impact actual productivity of workers. By identifying the trends and patterns of data, and the correlation between features (factors impact productivity) and label (productivity), we make a better prediction on the productivity a specific worker may have. The prediction help manufacturers improve their cost efficiency and productivity.

Approach

- **Data clean:** In the dataset, the column **"wip"** contain 42% of missing values. Therefore, we replace these missing data by the mean of the whole column.
- **Data exploration:** We determined the numeric and categorical features and find the correlation between these features and label by using boxplot, histogram, line graph, bar plot, pie chart and correlation heatmap.
- **Our ML approach** involves training 4 different algorithms. We use regression models (Linear Regression, Decision Tree, Random Forest, Gradient Boosting)on train and validation sets. To evaluate models, we applied methods such as R2, RMSE, MSE, and MAE. Finally, we select the best performing model to make predictions on unseen data.

PRE-PROCESSING

Preprocess for variables treatment

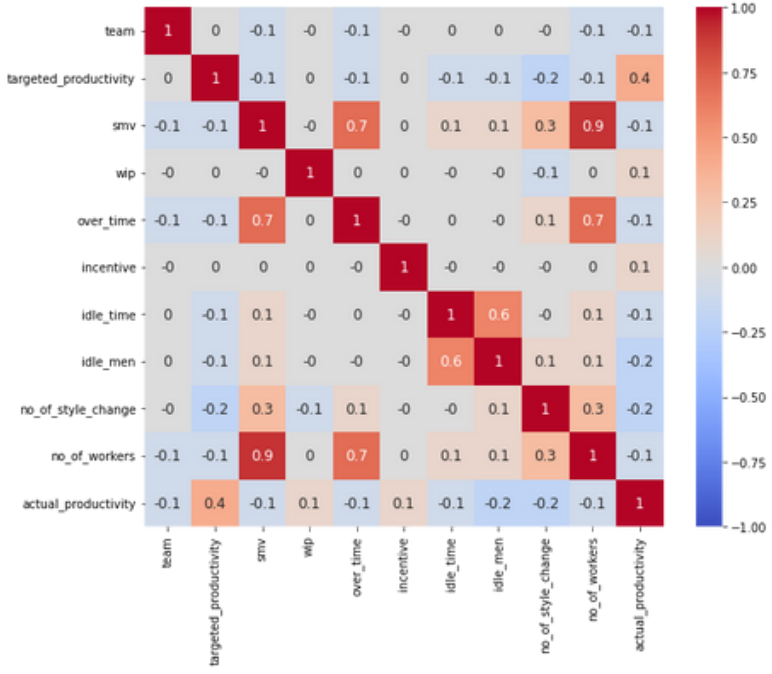
We performed several variable treatment steps, including: check integer type of data, remove column that reduce multicollinearity (no_of_worker) and replace missing value in column wip by mean of the column.

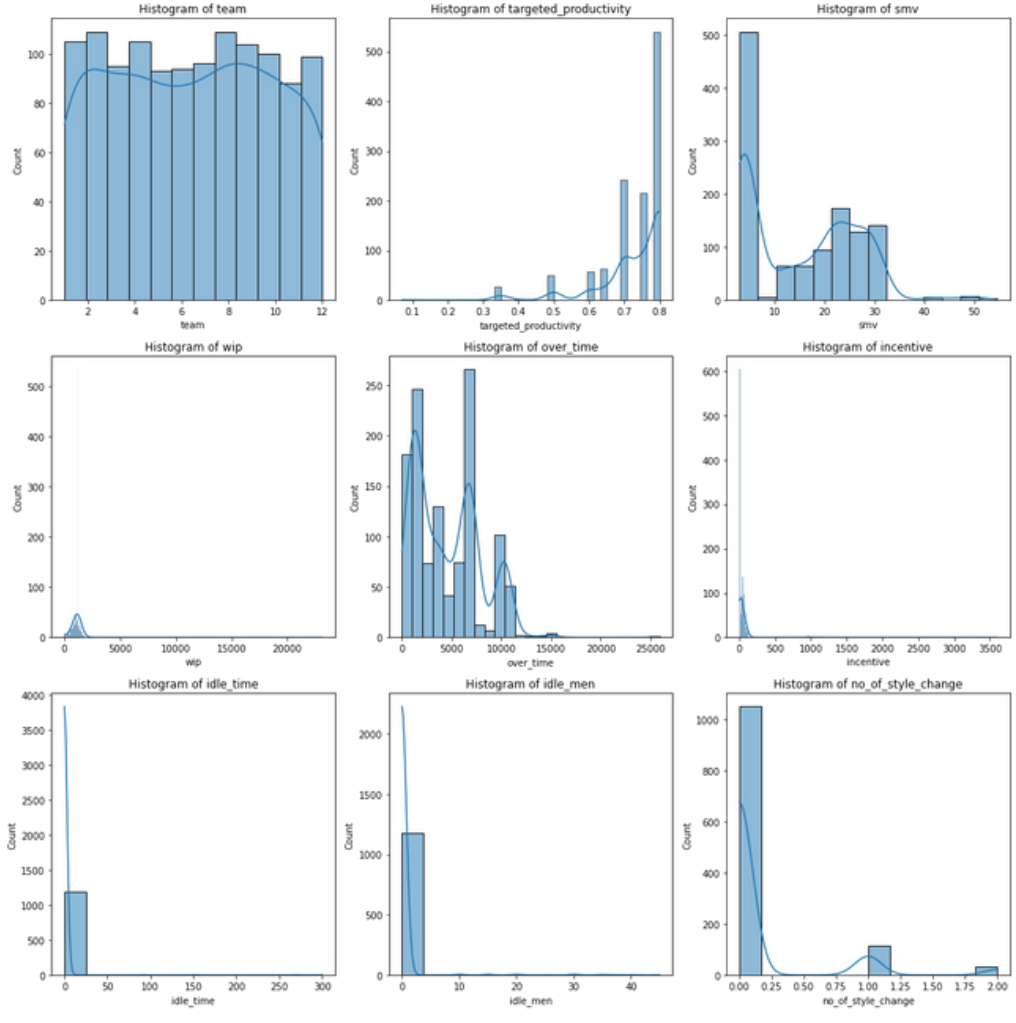
Transform variables

We transformed variables by using **StandardScaler** function from scikit-learn to standardize the numerical variables, and we used the **get_dummies** function to one-hot encode the categorical variables. By using these techniques, we want to ensure all numerical variables are in the same range to avoid bias toward higher-valued variables, and to convert categorical variables into a format that can be used for modeling purposes.

DATA EXPLORATION

Numeric features

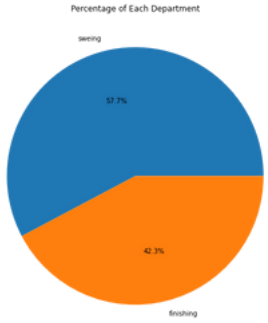


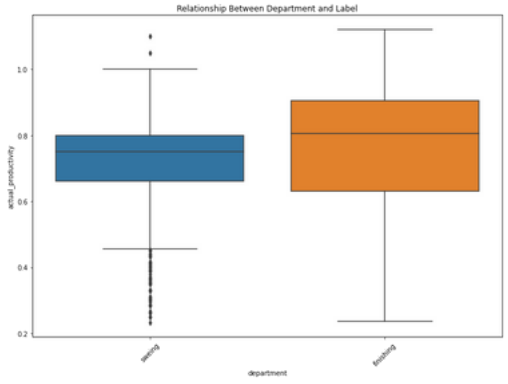


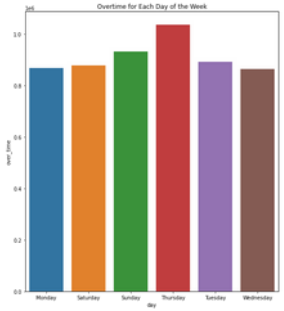
During data analysis, it may be found that there is an area where idle_time and idle_men, no_of_style_change and no_of_worker, no_of_worker and smv, over_time and svm highly correlated with each other. This implies that we should eliminate a feature that has the same impact as other features in training step - no_of_worker.

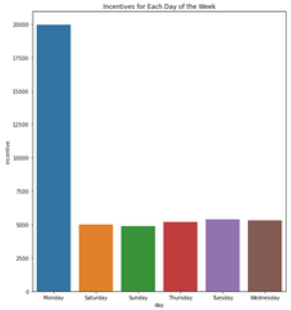
DATA EXPLORATION

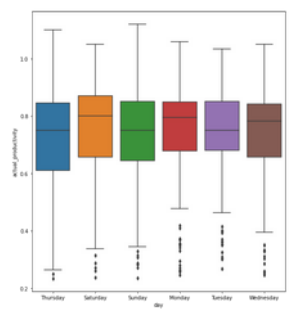
Categorical features

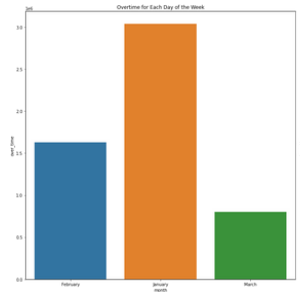


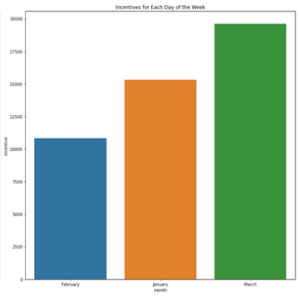


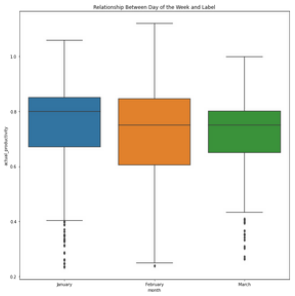








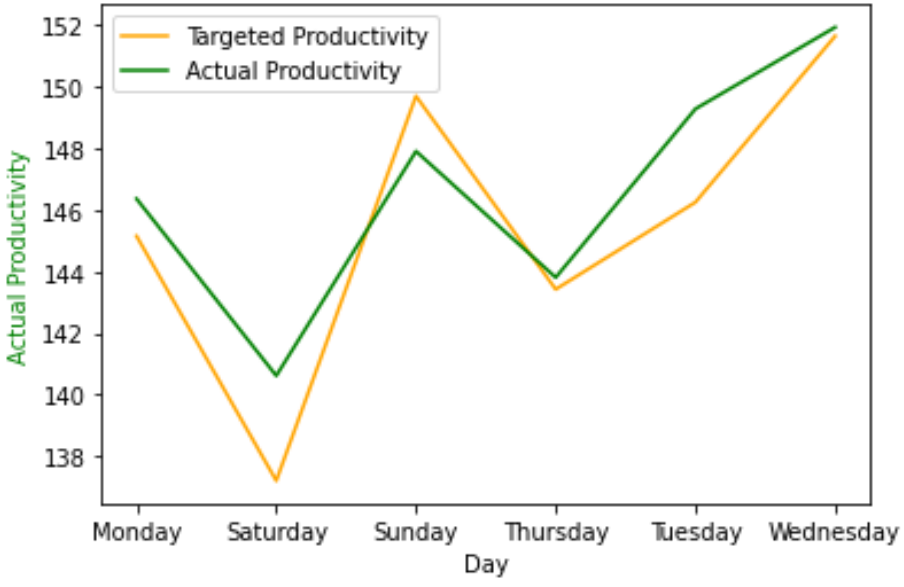




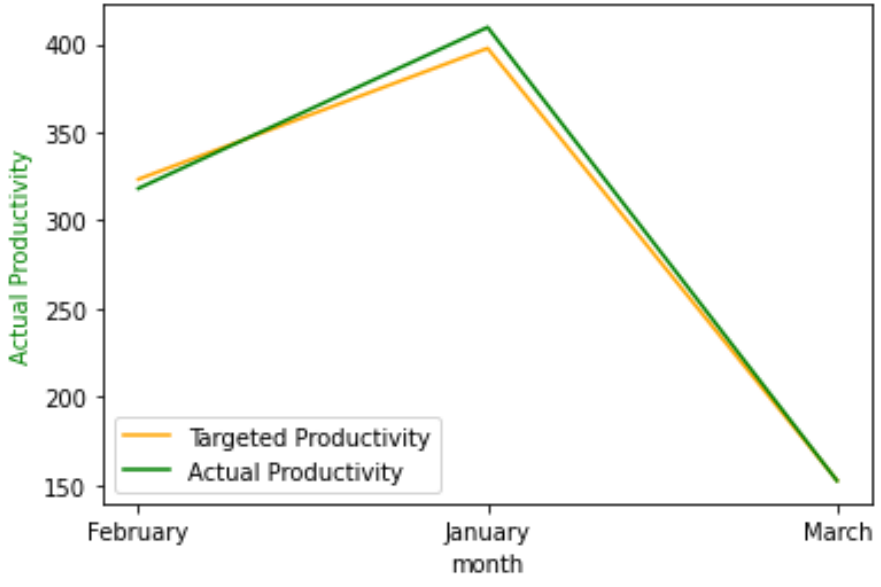
From the visualization:

- The company has more sewing team and less finishing team, however, the productivity of finishing team is higher than that of sewing team. This maybe the sewing teams take more time to finish their tasks than finishing team. The different between work demand can affect the productivity of each team. However, we can consider these factors when training the model and make prediction.
- Monday, despite having the highest incentives, is one of the days with the lowest productivity. The company may use incentive to motivate their employees in the first day of the week to boost the productivity of the whole week. Furthermore, Saturday offers the lowest incentives and Thursday stands out the day with the highest overtime having the highest productivity. We can make the hypothesis that: overtime may positively affect productivity while incentive is used for other purpose such as increase motivation of employees or for workers with higher skills.
- The dataset showcases the data of January, February and March. According to the plot, January has the highest overtime and second-highest incentive but the productivity is in the second place. This can imply that the higher overtime but lower incentive negatively impact the motivation of workers, leading to the decrease in the productivity. In March, the incentive is highest among three months, however, the overtime is lower. This implies that the overtime is important in increasing the productivity. While having the second-highest overtime and incentive, February balances well these two factors, leading to the highest productivity.

Compare targeted and actual productivity



In overall, the actual productivity in day is higher than the targeted productivity except Sunday.



In overall, the actual productivity in month is similar as targeted productivity except January.

MODEL FITTING & EVALUATION

By using machine learning models to train our system, we can predict the productivity of garment workers more accurately. This can be a useful tool for manufacturers to optimize their production strategy and reduce their operation cost. The machine learning algorithms can help and learn from the patterns in historical data and predict the productivity of workers on a range of factors such as Over Time, Incentive, Standard Minute Value, Day, Month, etc. By providing manufacturers with accurate predictions, they can make informed decisions about productivity, human resources and operation. Additionally, machine learning models can reduce the potential human error and consitent the reliable strategy.

Fitting models and parameters

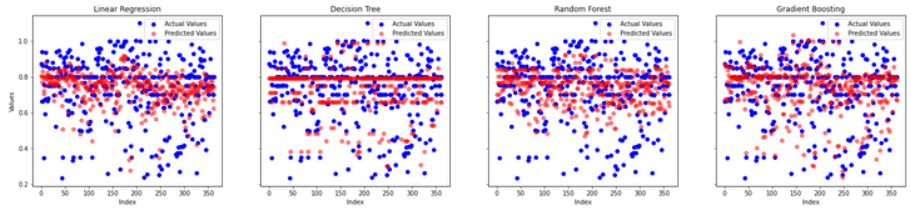
Fitting models and parameters Fitting multiple models with varying parameters is an important step in any machine learning project. In this case, we trained and tested 4 different models including: Linear Regression, Random Forest, Gradient Boosting and Decision Tree. Before model fitting the data need to be adjusted. Therefore, normalized all numerical values are on the same scale which can help improve the performance of the machine learning model. Categorical data is also converted into a format that the machine learning model can understand, using a process called "one-hot encoding." After all of the processed data, including the normalized numerical values and dummy variables, are combined into a single set of features to be used in the model.

Lastly, the Spark ML pipeline is set up to streamline the execution of these preprocessing steps, efficiently and easier to manage by using machine learning models.

Model evaluation

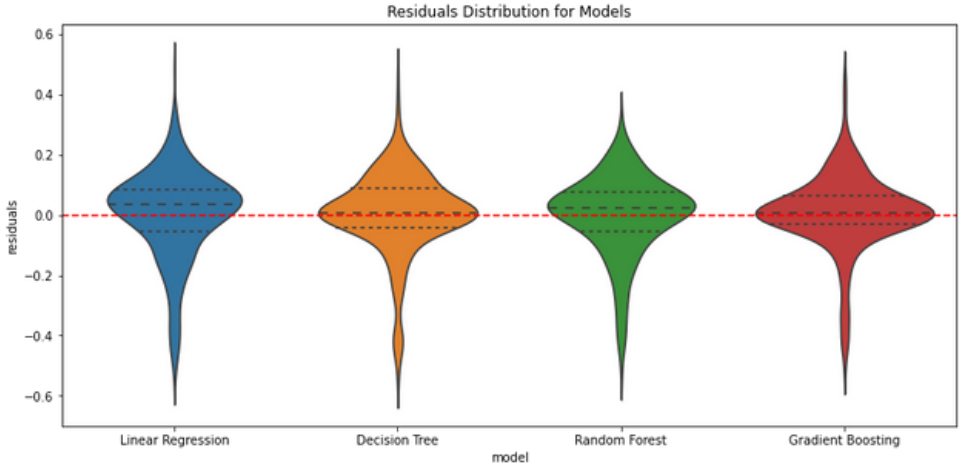
To evaluate the models, we use for matrices including R2, RMSE, MSE, and MAE.

	R2	RMSE	MSE	MAE
Linear Regression	0.272783	0.149500	0.022350	0.110430
Decision Tree	0.386787	0.137282	0.018846	0.092053
Random Forest	0.405376	0.135186	0.018275	0.098132
Gradient Boosting	0.465881	0.128123	0.016416	0.083571



The lowest RMSE, MSE, and MAE values, indicating the best overall performance models. R2 is response variable that exaplained by independent variables. In other words, it represents how well the model fits the data. In this case, Gradient Boosting is the best-performing model among evaluated models.

MODEL RESIDUAL



Residual distribution is approximately symmetrical around zero, which indicates that the model predictions are likely to be above or below the actual productivity values. Quartile lines inside the violin indicate the range of the residuals, and the red dashed line represents the zero line, which indicates that model predicted the actual productivity value perfectly.

Overall models, the mean of the residuals is very close to zero and standard deviation is small, which indicates that the models are performing well in terms of predicting (actual productivity) values.

Linear Regression Residuals:

mean	stddev	min	max
0.002585610570184...	0.14968399428954846	-0.5362930255814364	0.48081755705434953

Decision Tree Residuals:

mean	stddev	min	max
0.00592371121523949	0.1373438634167539	-0.5546710165435356	0.46763087425

Random Forest Residuals:

mean	stddev	min	max
-0.00212051035138...	0.1353555984437276	-0.5289439833948724	0.32524741242896416

Gradient Boosting Residuals:

mean	stddev	min	max
0.007766866705912152	0.12806420840801716	-0.5151483632409208	0.46469057412399856

SUMMARY & LIMITATIONS

- Overtime have more direct impact on productivity, while incentives serve broader purposes such as motivate employees to attracting more skilled workers and encourage workers work overtime. To have a higher productivity, manufacturers should balance overtime and incentive .

- Each department has requirement and demand of work, therefore, the productivity depends on these factors. The manufacturers can use this insights to allocate work, standard minute value, idle time, idle man to leverage the productivity.

- Gradient Boosting model emerged as the best performer with the lowest RMSE, MSE, and MAE values, and the highest R2 value. The manufacturers can use this model in predicting the productivity of their workers.

LIMITATIONS

1 Dataset [1]

The dataset contains import aspects of the garment production process and worker productivity, collected manually and verified by industry experts. However, It only shows the data within 3 months of a year, therefore, may not need to collect more data for building more accurate model.

2 The Model

The machine learning models used may not be fully optimal for predicting garment workers productivity, as there may be other models or parameters that could be more effective but were not explored in the analysis. Since data is collected manually and verified by industry experts, despite its value, potential limitations include human error, limited scope, subjectivity, temporal constraints, and generalizability concerns.

3 The Analysis & Insights

The analysis may not take into account external factors that could affect productivity. The insights and recommendations provided may not be applicable to all garment manufacturer, as each manufacturer may have their own unique circumstances and strategies.

[1] Productivity Prediction of Garment Employees. (2020). UCI Machine Learning Repository

THANK YOU!