OMNIe Solutions – Task 01 Project Evaluation Report

Name: Dipin Raj
University: Chandigarh University
Ph. No: 6235876977
E-mail: dipinr505@gmail.com
Wayand, India- 673593

Assignment: Evaluation 01
Job Role: AI/ML Intern
Submitted To: Mr. Chhatra Pratap & Mr. Anand Shukla
Company Name: Omnie Solutions
Noida, India

I. TASK 1: PROBLEM STATEMENT

This dataset consists of the operational statistics of the number of daily operations of several production teams in a garment manufacturing unit. The records are associated with a team and a particular day and they record the values of work time, idle time, team size, production goals and productivity realized.

The goal is to predict team efficiency (or productivity ratio) using production-related and workforce-related features.

Problem Statement:

- Predicting Team Efficiency in a Manufacturing Plant
- You are working as a data analyst for a manufacturing company.
- The management wants to understand the factors that influence the daily operational efficiency of production teams.

II. TASK 1: DELIVERABLES

To create a regression-based model that can be used to strikingly forecast the effectiveness of the team with the help of the features connected to the production and those connected to the workforce, and to comprehend which features have the most significant effect on the results of the efficiency.

The Task is to:

- Preprocess the data
- Carry out exploratory data analysis (EDA) to establish significant drivers of productivity.
- Feature Engineering.
- Evaluate the Model.
- Describe your results and model interpretability knowledge.

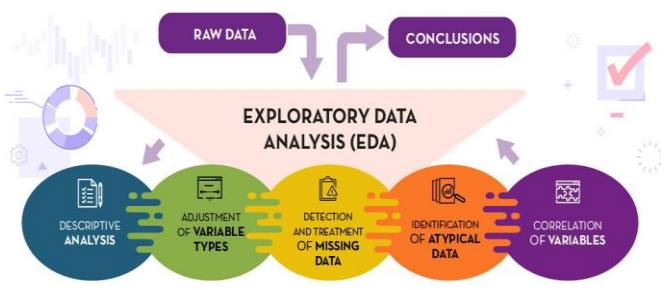

Fig 1: EDA (Exploratory Data Analysis)

II. DATASET ANALYSIS

The data is in the form of operation records of a garment manufacturing floor on a daily basis. Each row is a team-day and carries with its production / workforce measures and metadatas including:

- PlannedEfficiency, standardMinuteValue, workInProgress, overtimeMinutes, performanceBonus, idleMinutes, idleWorkers, workerCount, efficiencyScore (target).
- Categorical / contextual types: team, production department (Stitching / Finishing and Quality), day etc.
- Metadata (recordDate, fiscalQuarter).

III. EXPLORATORY DATA ANALYSIS (EDA)

There were a number of insights that were made during the exploratory data analysis (EDA) that informed the preprocessing and modeling.

The missing values were initially seen to be concentrated mostly under Finishing & Quality department and Stitching department data was mostly complete. Also, we had cases of duplication of column names due to the presence of trailing whitespaces (e.g., the column name “Finishing and Quality” and “Finishing and Quality ”) that were made standard in the process of data cleaning.

Additional examination showed that there were different behavioral patterns in the departments. Stitching and Finishing and Quality, which are production and post-production departments respectively, represented very different distributions on most variables. This difference indicated that a combination of the two under one model would blur out any significant patterns so it was decided to analyze Stitching separately to be able to discover patterns.

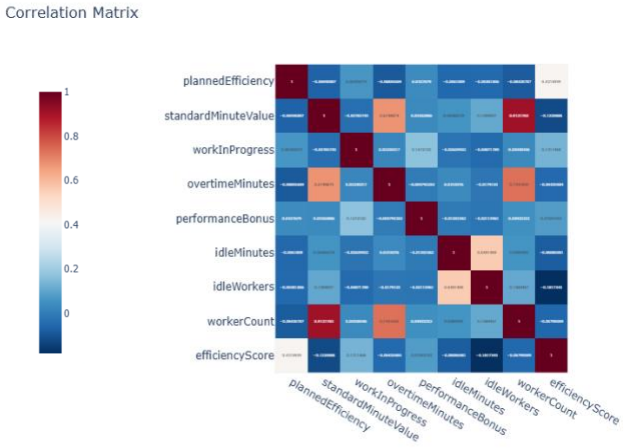

Fig 2: Correlation heatmap before split

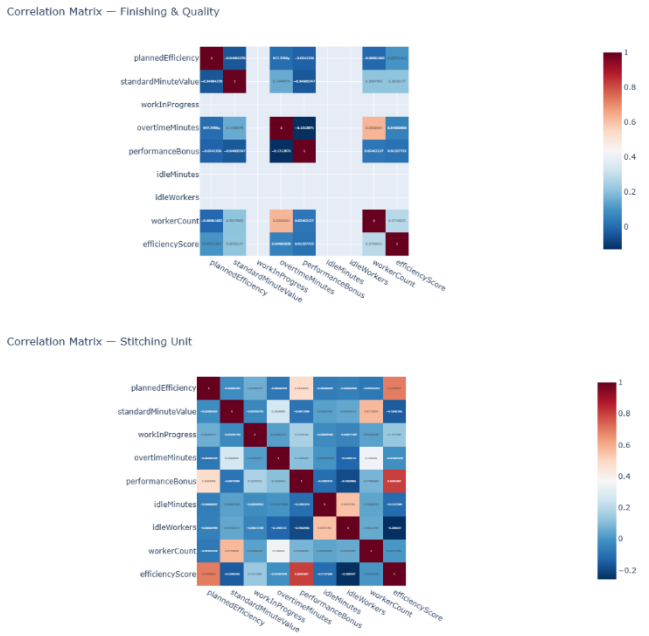

Fig 3: Correlation heatmap after split

There were also some outliers in the data. The summary statistics and visualization, box and violin plots were used to point out extreme values in the 'overtimeMinutes' and 'workInProgress'. These exceptions were probably caused by the uncommon operational accidents or errors by handwriting. Outliers may introduce bias in a regression model hence value ranges were subsequently established to overcome the influence of outliers.

The correlation analysis revealed that there were overall weak pairwise correlations within the entire data set and none of the variables exhibiting strong and linear relationships (greater than 0.5) with the target 'efficiencyScore'. Nevertheless, when the data were filtered to bring only the Stitching department and the outliers eliminated, there were more significant relationships, the most significant of them being with 'plannedEfficiency' and 'performanceBonus', which emerged as the sources of efficiency.

Lastly, it was observed that the size of the dataset was not that big particularly after segmentation into departments. This weakness affected the behavior of some models: decision trees would underfitting because of lack of complexity on data patterns, and extremely deep trees and ensemble techniques would overfitting. Linear regression models, however, were able to offer consistent and understandable baselines to this data.

IV. METHODOLOGY

The approach used in the given project was a systematic process that included preprocessing of data, exploration, and predictive modeling.
During the preprocessing of the data, the data was initially checked on inconsistencies, missing values and duplicate data. Informative indexes like 'recordDate', 'fiscalQuarter', 'dayOfWeek' and 'productionDept' were also found to be non-informative during prediction and thus were eliminated. Whitespaces in the names of categories in the columns such as Finishing and Quality were normalized in order to bring about uniformity.

Then, the descriptive statistics and box and violin plots were used to evaluate the data in terms of outliers. Pragmatic numeric ranges were established based on the distributions observed on some of the key features as 'plannedEfficiency', 'overtimeMinutes', and 'workerCount'. Any values that went beyond these limits were eliminated.

Fig 4: Scatter plots revealing outliers

The Finishing and Quality and Stitching departments had completely different patterns of data, so they were considered individually to eliminate the possibility of blurring irrelevant behaviors together. A subunit of Stitching Unit, which exhibited more regular relationships with the target variable efficiencyScore, was used for trainning the modeles.

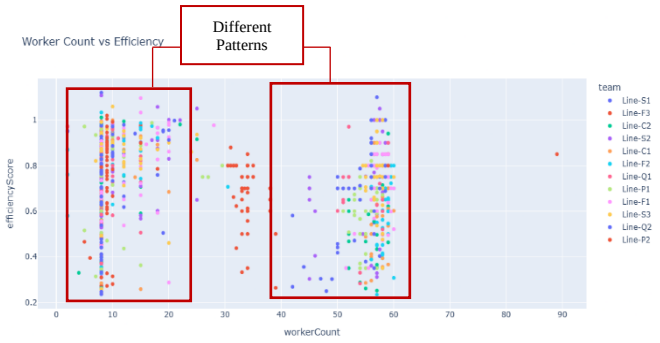

Fig 5: Scatter plots revealing different patterns for both departments

The categorical variables such as team and styleChangeCount were one-hot coded, and the fields that were given out as a boolean converted into 0/1 format. All the numerical characteristics were then scaled by StandardScaler to make the contribution of all the variables similar.

After preprocessing, Linear Regression, Decision Tree Regressor, and Random Forest Regressor models were used to forecast the efficiency of the team and determine the importance of features.

V. RESULTS

The modeling stage was used to assess the predictability of the efficiency score of production teams using various algorithms taking into account the processed data.

The Linear Regression models, one of which was written manually and another one used the Scikit-learn library, were almost equally performing, which confirms the accuracy of the manual implementation. All of them gave a value of about 0.808 for the R^2 value, and the Mean Squared Error (MSE) was close to 0.00388, which is quite high and thus indicates the good linear correlation between the selected features and the efficiency score.

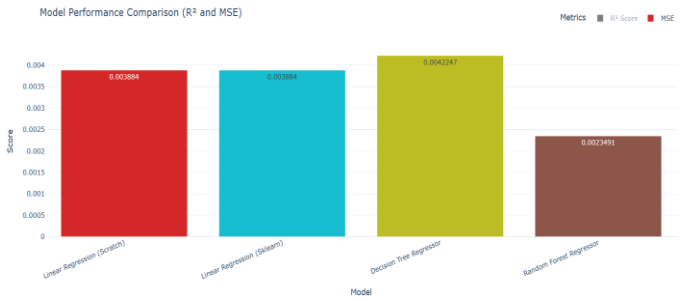

Fig 6: Model Comparison by MSE

The Decision Tree Regressor with maximum depth of 4 had R^2 of 0.74 and MSE of 0.00422. Although moderately accurate, this model showed evidence of underfitting, probably because of limited data and shallow depth made this model poor in learning complicated patterns. This was supported by the fact that performance did not increase significantly with increasing depth.

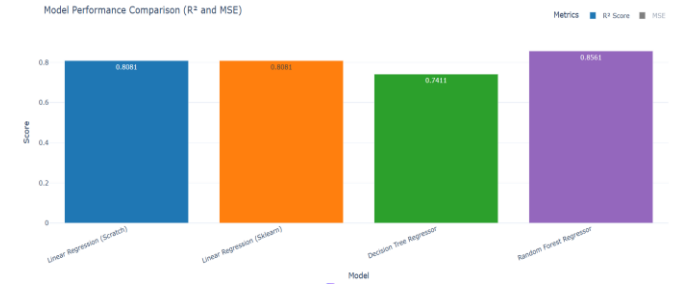

Fig 7: Model Comparison by R^2

Random Forest Regressor showed the best results of all the models with R^2 of 0.856 and MSE of 0.00234. It had slight overfitting initially, and the depth parameter was tuned to stabilize its results.

Table 1: Model Comparison

Model	R^2 Score	MSE
Linear Regression (Scratch)	0.8081	0.0038840
Linear Regression (Sklearn)	0.8081	0.0038840
Decision Tree Regressor	0.7411	0.0042247
Random Forest Regressor	0.8561	0.0023491

VI. CONCLUSION

This project was able to illustrate an end-to-end workflow of predicting operations efficiency in a garment manufacturing unit. In a rigorous EDA, we were able to detect structural discrepancies, department-specific behavior variations, and important outliers that would potentially corrupt the model learning. Filtering, cleaning, and feature engineering helped to address these problems to provide a more reliable modeling base.

The Finishing and Quality departments when broken down showed no significant linear relationships or noticeable nonlinear trends with efficiency. The results did not show any meaningful relationships even after the use of Decision Tree and Random Forest models as evidenced with the help of scatter plot visualization. While introducing synthetic drift variables could have forced correlation, it would compromise the integrity of the analysis and fall outside the logical and ethical scope of the study. Therefore, the modeling procedure specialized to the Stitching Unit, and patterns as well as dependencies were all data-driven and interpretable.

The experiments showed that the strongest predictors of team productivity are planned efficiency and performance bonus, and the use of managerial planning and incentive structure has a direct impact on the performance. Comparison of models showed that although Linear regression offered us interpretability and validation of our scratch-built model, Random Forest Regression was better at prediction as it was able to represent the intricate relationships in the data. Nevertheless, because of the small size of the dataset, tree-based models were sometimes prone to either underfitting or overfitting, which implies that more varied data within departments are necessary to make sound generalization.

To conclude, this report highlights that quality of data and data preprocessing that is domain-specific are as fundamental as the choice of algorithms. The framework can be extended with more valuable data to create real-time production monitoring and workforce optimization predictive dashboards.

REFERENCE

[1] Colab Notebook Link: Predicting Operational Efficiency of Manufacturing Teams

[2] GitHub Repo Link: Predicting Operational Efficiency of Manufacturing Teams

[3] Cousineau, D. and Chartier, S., 2010. Outliers detection and treatment: a review. *International journal of psychological research*, *3*(1), pp.58-67.

[4] Iduseri, A., 2022. Winsorization with Graphical Diagnostic for Obtaining a Statistically Optimal Classification. *Advances in Principal Component Analysis*, p.199.