Fault Automatic Identification and Personnel Allocation in the Production Line

Baxian Chen, Yurong Wu, Shan He, Yisha Liu
School of Artificial Intelligence, Guangzhou Huashang College, Zengcheng, Guangzhou, China

Abstract: This paper focuses on the issues of fault automatic identification and personnel allocation in industrial production lines. First, comprehensive data preprocessing is conducted, involving meticulous data cleaning to remove noise, strategic handling of 3 missing values through interpolation or deletion, and rigorous feature selection based on correlation analysis. To tackle the problem of sample imbalance, both up-sampling (duplicating minority samples) and down-sampling (reducing majority samples) are applied, significantly improving the training effect of the model. A fault - alarm model is then built using key characteristics like equipment operation states and process parameters, enabling timely prediction of potential faults. Meanwhile, advanced machine - learning models, such as random forest and decision trees, are utilized to analyze the intricate relationship between workers' years of service and production efficiency, helping formulate an optimal shift-scheduling plan that balances operator experience and shift workload. The proposed methods have been proven to effectively enhance production line stability and operational efficiency, offering practical and innovative solutions for industrial manufacturing and solid scientific decision - making support for production management and human resource allocation.

Keywords: Difference; Linear Regression; Up-sampling and Down-sampling; Decision Tree; RF Model; SHAP Model

1. Introduction

With the rapid advancement of information technologies, industrial production lines have become increasingly intelligent. Automated systems now handle processes such as item transfer, material filling, packaging, and quality inspection, significantly boosting efficiency, improving product quality, and reducing operational costs. Integrating intelligent fault alarm systems is crucial for preventing production interruptions and minimizing losses from equipment failures. As artificial intelligence, big data, and the Internet of Things (IoT) continue to evolve, there is a growing need to develop more sophisticated methods for fault detection and personnel allocation to drive innovation in smart manufacturing.

2. Problem Re-statement

2.1 Problems

Problem 1: Analyze fault data characteristics from production line devices (e.g., material-pushing cylinders, filling stations) and develop a real-time fault-alarm model.
Problem 2: Apply the model to Appendix 2 data to provide fault dates, durations, and monthly fault counts.
Problem 3: Examine relationships between output, qualification rates, and other factors using Appendix 3 data.
Problem 4: Develop a shift-scheduling plan for 24/7 operations, considering operator experience and ensuring balanced workload distribution across shifts.

2.2 Problem Analysis

Analyze the fault data characteristics, including the time distribution of fault occurrences, the frequency of fault occurrences, the distribution of fault types, etc. Based on the data of Problem 1, use ['material - pushing cylinder retracted state',' material - pushing cylinder pushed state', 'number of placed containers', 'number of container upload detections', 'number of filling detections', 'filling locator fixed state', 'filling locator released state', 'number of materials to be grabbed', 'number of fillings', 'number of cappings', 'number of screwings'] as the x -

values, and each fault type as the y - value. After processing, put them into the model for training.

Analysis of Problem 2: Based on the trained model, put the x - values in the data of Problem 2 into the trained model to predict the corresponding y - values and obtain each fault type. According to the fault types, classify by the unique values of the date, sort the time, find the continuous time periods, calculate their durations, and finally group by the date to count the total number of faults per month, the longest and shortest durations.

Analysis of Problem 3: First, calculate the newly added 11 feature values, merge the operator information table with it, and then merge the ten production lines. Use different models to evaluate different years of service, and select the optimal model according to MSE and R2. Finally, analyze the important influence degree of the 11 features in the production line.

Analysis of Problem 4: According to the objective function, clarify the constraint conditions (each worker works 5 days a week, each shift requires 10 different people, and ensure that there are workers on each production line) and meet the constraint conditions to find the optimal solution. Then, combine with the proportion of different years of service of workers and assign different production lines according to different years of service to obtain the optimal shift - scheduling plan.

3. Formulas and Models

3.1 Formulas Used in Problem 3

To summarize the operation record data of the production line into one piece of data, some summary variables need to be designed. These variables can represent the overall operation situation of the production line within one year. The following are some possible summary variables, their calculation formulas, and principles:

(1) Total operation time: Calculate the total time the production line operates within one year.

Formula: Total operation time = $\sum$ (End time − Start time)

Principle: Accumulate the operation time of each production cycle.

(2) Total production quantity: Calculate the total number of products produced by the production line within one year.

Formula: Total production quantity = $\sum$ Qualified number

Principle: Accumulate the number of qualified products recorded daily.

(3) Total unqualified quantity: Calculate the total number of unqualified products produced by the production line within one year.

Formula: Total unqualified quantity = $\sum$ Unqualified number

Principle: Accumulate the number of unqualified products recorded daily.

(4) Average production efficiency: Calculate the average production efficiency of the production line, that is, the number of qualified products produced per hour.

Formula: Average production efficiency = Total production quantity / Total operation time

Principle: Divide the total production quantity by the total operation time to obtain the production efficiency per hour.

(5) Equipment comprehensive failure rate: Calculate the comprehensive failure rate of all equipment.

Formula: Equipment comprehensive failure rate = (Material - pushing device failure 1001+Material - detection device failure 2001+…+Screwing device screwing failure 6002) / Total number of days × 100%

Principle: Accumulate the occurrence times of all equipment failures, and then divide by the total number of days in a year to obtain the equipment comprehensive failure rate.

(6) Material - pushing efficiency: Calculate the average pushing efficiency of the material - pushing cylinder.

Formula: Material - pushing efficiency = (Material - pushing number / (Material - pushing number + Material to be grabbed number)) × 100%

Principle: Divide the number of successfully pushed materials by the sum of the number of successfully pushed materials and the number of materials to be grabbed to obtain the pushing efficiency.

(7) Filling efficiency: Calculate the efficiency of the filling process.

Formula: Filling efficiency = (Number of fillings / Material - pushing number) × 100%

Principle: Divide the number of completed filling operations by the number of successfully pushed materials to obtain the

filling efficiency.
(8) Capping efficiency: Calculate the efficiency of the capping process.
Formula: Capping efficiency = (Number of cappings / Number of fillings) × 100%
Principle: Divide the number of completed capping operations by the number of completed filling operations to obtain the capping efficiency.
(9) Screwing efficiency: Calculate the efficiency of the screwing process.
Formula: Screwing efficiency = (Number of screwings / Number of cappings) × 100%
Principle: Divide the number of completed screwing operations by the number of completed capping operations to obtain the screwing efficiency.
(10) Total production cycles: Calculate the total number of production cycles of the production line within one year.
Formula: Total production cycles = ∑ Number of fillings
Principle: Accumulate the number of filling operations recorded daily to obtain the total number of production cycles.
(11) Qualification rate: Calculate the qualification rate of the production line.
Formula: Qualification rate = (Total production quantity / (Total production quantity + Total unqualified quantity)) × 100%
Principle: Divide the total production quantity by the sum of the total production quantity and the total unqualified quantity to obtain the qualification rate.

3.2 Difference

It usually refers to the differencing operation on time-series data, which is used to convert non - stationary time - series into stationary time - series. It is a time - series preprocessing technology [1]. By calculating the differences between adjacent time points, the trends and seasonality of the data are eliminated, making the time - series more stable.

3.3 Up-sampling and Down-sampling

Up-sampling and down-sampling are two common methods in data preprocessing, used to adjust the frequency or size of the data.
Up - sampling (Upsampling): When the proportion of minority classes and majority classes in the dataset is unbalanced, the number of minority - class samples is increased by repeating them to balance the class proportion.
Down - sampling (Downsampling): When there are too many samples in the dataset, some majority - class samples are discarded to reduce their number.

3.4 Random Forest Model

The random forest is a supervised learning algorithm based on Bagging and decision trees, belonging to the Bagging method in ensemble learning. It was proposed by Leo Breiman in 2001, combining the Bagging ensemble learning theory with the random subspace method [2].
The model framework of the random forest is shown in the figure. It is an ensemble model composed of multiple decision trees. Using the Bootstrap method, k datasets are generated from the original data, and each dataset contains N (rows) and P variables (columns). Each dataset is used to build a CART decision tree [1]. During the process of building the subtree, instead of choosing all variables as the node field, P fields are randomly selected. Each decision tree is allowed to grow as fully as possible to make each node in the tree as "pure" as possible, that is, each subtree in the random forest does not need to be pruned. For classification problems in a k - tree random forest, the category with the highest vote is used as the final judgment result; for regression problems, the mean method is used as the final result. As shown in Figure 1:

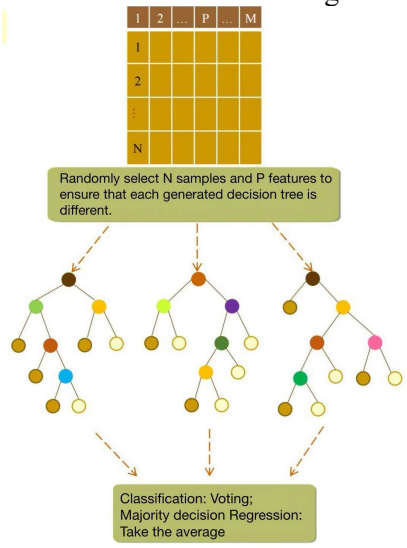

Figure 1. Flowchart of the Random Forest

3.5 SHAP Model (Continued)

The SHAP (SHapley Additive exPlanations)

model is a method for explaining the predictions of machine - learning models. It is based on the Shapley value theory in cooperative game theory [3]. The SHAP model can provide explanations for the contribution degrees of single features and feature combinations to the model predictions. As shown in Figure 2:

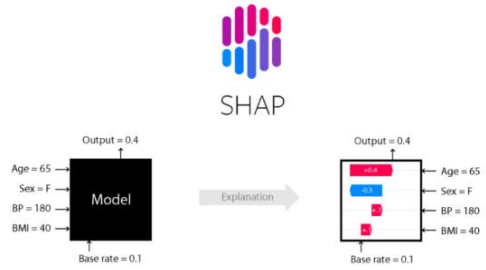

Figure 2. Example Diagram of the SHAP Model Principle

The working process of the SHAP model: First, a machine - learning model needs to be trained, which can be a decision tree, linear regression, neural network, etc [4]. Next, using the trained basic model and the samples to be explained, the SHAP values of each feature for the prediction results are calculated. The calculation of SHAP values can be achieved by traversing the paths of the decision tree or using approximate algorithms, and the specific method can be called according to the used model and the SHAP library. By analyzing the SHAP values, the contribution degree of each feature to the model prediction can be explained. A larger SHAP value indicates that the feature makes a greater contribution to the prediction result, while a smaller SHAP value indicates a smaller contribution [5]. To more intuitively understand the interpretation results of SHAP values, various visualization techniques such as bar charts, scatter plots, and heatmaps can be used to compare and display the SHAP values and feature values.

The advantage of the SHAP model is that it can provide feature - level explanations, helping us understand the contributions of features to the prediction results. It can also be used to visualize the importance of features, helping us better understand the prediction process of the model [4]. By using the SHAP model, we can more deeply understand the internal mechanism of the model and the relationships between features.

3.6 Objective Function (Aiming to Maximize the Sum of the Qualification Rate Multiplied by the Production Quantity)

The objective function is a mathematical expression that needs to be maximized or minimized, used to measure the advantages and disadvantages of the choice of decision - making variables for a problem. Our formula can be expressed as:

$$\left[Maximinze Z = \sum_{i=1}^{N}\sum_{j=1}^{D}\sum_{k\in S} X_{ijk}\cdot E_{rate,i}\cdot E_{quantity,i}\right] \quad (1)$$

3.7 Constraint Conditions (Constraints)

Working Days Constraint: Each worker works 5 days a week and takes 2 days off.

$$\left[\sum_{i=1}^{N}\sum_{k\in S} X_{ij} = 5, \forall i = 1,2,\ldots,N\right] \quad (2)$$

Shift - Number - of - People Constraint: Each shift requires 10 different people per day.

$$\left[\sum_{i=1}^{N} X_{ijk} = 10, \forall j = 1,2,\ldots,D, \forall k \in S\right] \quad (3)$$

Years - of - Service Distribution Ratio Constraint: Ensure that the number of workers in each years - of - service level meets the given proportion.

$$\left[\sum_{i=1}^{N}\sum_{j=1}^{D}\sum_{k\in S} X_{ijk} = W_{exp}, \forall_{exp\in experience}\right] \quad (4)$$

3.8 Symbol Explanation

For the convenience of the establishment and solution process of the following models, the following explanations are given for the key symbols used, as shown in Table 1.

Table 1. Symbol Definitions and Descriptions

Symbol	Description
N	Num workers
D	num days
S	the set of shifts
X_{ijk}	a decision - making variable
$E_{rate,i}$	the qualification rate of worker
$E_{quantity,i}$	the daily production quantity of worker
W_{exp}	service level
MSE	average prediction error
R^2	the variance of the actual observed values

4. Problem 1: Data Modeling

4.1 Data Preprocessing

The analysis of device characteristics, factors affecting device faults, and the prediction of device faults are all based on the states of each device and data changes over time. Since there may be certain errors during data collection and transmission, we clean and process the missing values, duplicate values, and

abnormal values of the data respectively. Using the isnull(). sum() method of the pandas DataFrame class, we check for missing values in the dataset and find no missing values in the original training dataset. For duplicate values, the duplicated (). sum () method of the pandas DataFrame class is used to analyze the original data, revealing no duplicate entries. Finally, statistical analysis via the describe () method of the pandas DataFrame class confirms no abnormal values.

4.2 Model Building

According to the production line process, we extract the feature columns (X) and target columns (Y) from the data in Appendix 1. The feature columns (X) include: "material - pushing cylinder retracted state", "material - pushing cylinder pushed state", "number of placed containers", "number of container upload detections", "number of filling detections", "filling locator fixed state", "filling locator released state", "number of materials to be grabbed", "number of fillings", "number of cappings", "number of screwings". The target columns (Y) include: "material - pushing device failure 1001", "material - detection device failure 2001", "filling device detection failure 4001", "filling device positioning failure 4002", "filling device filling failure 4003", "capping device positioning failure 5001", "capping device capping failure 5002", "screwing device positioning failure 6001", "screwing device screwing failure 6002".

Table 2. Model Training Classification Report

category	precision	recall	f1-score
0	0.98	0.99	0.99
1001	1.00	1.00	1.00
2001	0.99	0.76	0.86
4001	0.99	0.69	0.82
4002	1.00	1.00	1.00
4003	0.89	1.00	0.94
5001	0.90	0.45	0.60
5002	0.79	0.39	0.52
6001	0.87	0.48	0.62
6002	0.80	0.49	0.61
accuracy		0.97	36000
macro avg	0.92	0.73	0.80
weighted avg	0.97	0.97	0.97

In the data preprocessing stage, considering that some feature columns (X) show an increasing trend, to improve the accuracy of the model, we use the up - down difference method to calculate the differences between adjacent rows and adjacent columns of all sample (X) values for equalization processing. Further statistical analysis reveals a problem of unbalanced data samples between fault values and normal values. Therefore, we perform up - sampling and down - sampling on ten samples. After processing, we export the data and construct a random forest (Random Forest, RF) model for training [2]. The analysis report of the trained model is shown in table 2:

5. Problem 2: Data Prediction

5.1 Data Preprocessing

First, we import the dataset provided in Appendix 2. The following features are extracted for analysis: ['material - pushing cylinder retracted state', 'material - pushing cylinder pushed state', 'number of placed containers', 'number of container upload inspections', 'number of filling inspections', 'filling locator fixed state', 'filling locator released state', 'number of materials to be grabbed', 'number of fillings', 'number of cappings', 'number of screwings']. These features are considered to affect equipment failures.

Next, we perform difference processing on the selected features to eliminate any trends in the data. The differenced features are then combined with the original features to create a comprehensive feature matrix (X) for input into the model.

5.2 Prediction Modeling

The preprocessed data (X) is input into the model trained in our initial analysis (referred to as "Problem 1" in this study). The trained model, which may be the random forest (RF) model mentioned earlier, predicts equipment failures (Y) based on the input features. The predicted Y values indicate which equipment is likely to fail based on historical patterns and operating parameters.

5.2 Equipment Failure Duration

After obtaining the predicted fault values, we observe that the time column in the data represents continuous time intervals. We process the dates to ensure uniqueness and

then sort them accordingly. Using an iterative process and conditional statements, we determine the continuous time spans and calculate their durations. This step enables us to determine the duration of each equipment failure [2].

5.3 Statistical Analysis

Further analysis involves grouping the data by date to classify the faults monthly [1]. By using functions such as count, max, and min, we calculate the total number of faults per month, as well as the longest and shortest durations of faults. These statistical information provides a comprehensive view of the changes in equipment failures over time.

The final results are summarized in the result2.xlsx file for further analysis and decision - making.

6. Problem 3: Relationships among Production Output, Qualification Rate, Operators, and Production Lines

We first import the data in Appendix 3, replace fault and non-fault values with 0 and 1, and then calculate the "total operation time", "total production", "total defects", "average efficiency", "corrected composite fault rate", "material push efficiency", "filling efficiency", "capping efficiency", "screwing efficiency", "total production cycles", and "acceptance rate" using the formulas, and name the resulting data frame as df. Then Merge the operator information table with the df table, and then combine the ten production lines to obtain a new information table (due to the large amount of data, only a portion of the data is displayed here), as shown in Table 3:

Table 3. New Information Table after Merging Ten Production Lines

date	total production	total defects	Production Line Number
0	206524870	0	M301
1	207331200	0	M301
2	206275556	0	M301
3	208470280	0	M301
4	197849420	20893	M301
2595	20954008	0	M310
2596	20751522	101671	M310
2597	20896354	0	M310
2598	20410855	0	M310
2599	20567724	21248	M310

6.1 Relationships between Production Lines, Qualification Rate, and Total Production Quantity

We group the total dataset by "years of service" and calculate the average "qualification rate" and "total production quantity" for each years - of - service level. The "total production quantity" is divided by the number of unique values of "date", that is, the average daily production quantity. Observe the total production quantity and qualification rate under different years - of - service levels, as shown in Figure 3:

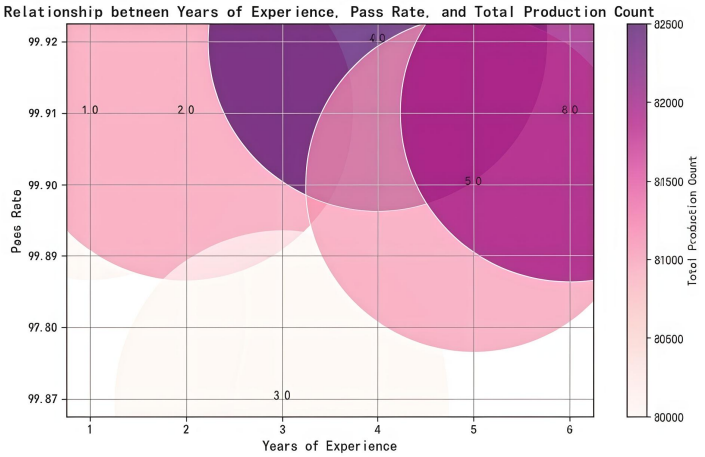

Figure 3. Relationship between Years of Service, Qualification Rate, and Total Production Quantity

It can be seen that workers' years of service show a positive correlation with both the qualification rate and total production quantity. That is, the greater the years of service, the higher the qualification rate and the total production quantity. This indicates that as workers' years of service increase, they may accumulate more experience and skills, making them more proficient and efficient in their work, thus leading to an increase in the qualification rate and total production quantity. Moreover, workers with longer years of service may be more stable in their jobs, more familiar with the work processes and requirements, and better able to meet production requirements. They may also have better problem - solving and teamwork abilities, enabling them to better cope with challenges at work and improve production efficiency and quality. Therefore, the differences in the qualification rate and production quantity of production lines are caused by the years of service, and the relationships between production lines, qualification rate, and production quantity are essentially the same as those with the years of service [4].

6.2 Relationships between Operators' Years of Service and Average Production Efficiency, Capping Efficiency, Filling Efficiency, Screwing Efficiency, Total Production Cycles, Material - Pushing Efficiency, Total Operation Time, Equipment Comprehensive Failure Rate, Total Unqualified Quantity, and Qualification Rate

Considering that most of these features have non - linear effects and that higher - accuracy data can be obtained by comparing different models, we use multiple models and take MSE and R2 as model evaluation criteria. The evaluation results of multiple models are as follows, as shown in Table 4:

Table 4. Evaluation Results of Each Model

Model	Type	MSE	R2
Linear	Regression	2.3530	0.1626
Decision	Tree	0.2690	0.9042
Random	Forest	0.5241	0.8134
Support	Vector Machine	2.7729	0.013187
Neural	Network	16920.0091	-6020.3556

According to the given evaluation results, the decision - tree model performs best on the given dataset. It has a relatively low average prediction error (MSE = 0.26) and a high degree of explanation for the variance of the actual observed values (R2 = 0.90). Therefore, based on these indicators, the decision - tree model is the optimal model for this dataset. The comparison of the effect diagrams of each model is more obvious, as shown in Figure 4:

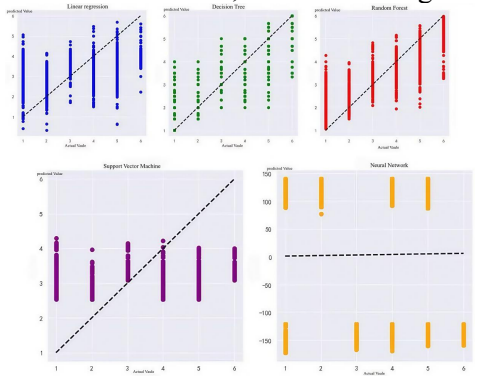

Figure 4. Comparison of Each Model's Effect

In summary, we use the SHAP model with the decision - tree model as the basic model to explain the contribution degree of each feature to the model prediction, as shown in Figure 5:

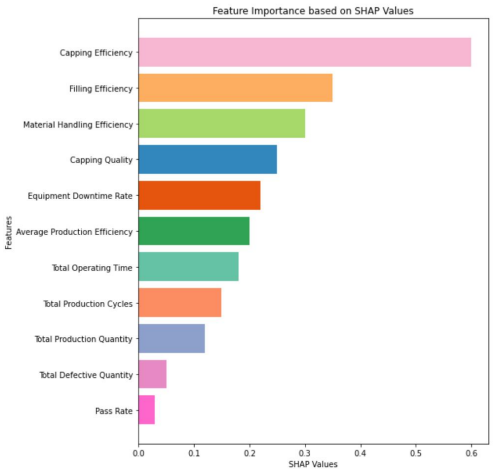

Figure 5. Absolute Values of SHAP Values of Each Feature

We observe the following: The screwing efficiency has the largest absolute value, indicating that it has the most significant impact on the prediction result. The filling efficiency, material - pushing efficiency, and capping efficiency also have relatively large absolute values, indicating that they have a greater impact on the prediction result. The absolute values of the equipment comprehensive failure rate, average production efficiency, and total operation time are relatively small, but they still have a

certain impact on the prediction result. The total production cycles, total production quantity, total unqualified quantity, and qualification rate have the least impact on the prediction result.

To better analyze the relationships between each feature value and the years of service, we conduct a linear regression analysis, which includes F - values, VIF values, R2, etc.[4]

The F - test is used to determine whether there is a significant linear relationship, and R2 is used to evaluate the goodness of fit of the regression line to this linear model. In linear regression, we mainly focus on whether the F - test passes. In some cases, there is no necessary relationship between the size of R2 and the explanatory power of the model [3]. If the VIF value is "inf", it means the VIF value is infinite. B is the coefficient in the case of a constant, the standard error = B/t value, and the standardized coefficient is obtained by standardizing the data.

From the analysis of the F - test results, the significance P - value is 0.000*, which is significant at this level. We reject the null hypothesis that the regression coefficient is 0, so the model basically meets the requirements. Regarding the collinearity of variables, the VIF values of variables such as total operation time, total production quantity, total unqualified quantity, average production efficiency, material - pushing efficiency, filling efficiency, capping efficiency, screwing efficiency, total production cycles, and qualification rate are greater than 10, indicating a collinear relationship. It is advisable to remove the collinear independent variables or perform ridge regression or step - by - step regression. The model formula is as follows:

$$y = -1212.867 - 0.0 \times total\ operation\ time + 0.0 \times total\ production\ quantity - 0.0 \times total\ unqualified\ quantity + 0.166 \times average\ production\ efficiency + 0.003 \times equipment\ comprehensive\ failure\ rate + 202.034 \times material\ pushing\ efficiency + 132.758 \times filling\ efficiency - 15.316 \times capping\ efficiency - 95.795 \times screwing\ efficiency - 0.0 \times total\ production\ cycles - 22.656 \times qualification\ rate. \quad (5)$$

That is, there is a linear relationship between the years of service and each feature value. Therefore, we can use the SHAP model to output the positive or negative impact of the years of service on them, that is, the positive or negative correlation relationship, as shown in Figure 6:

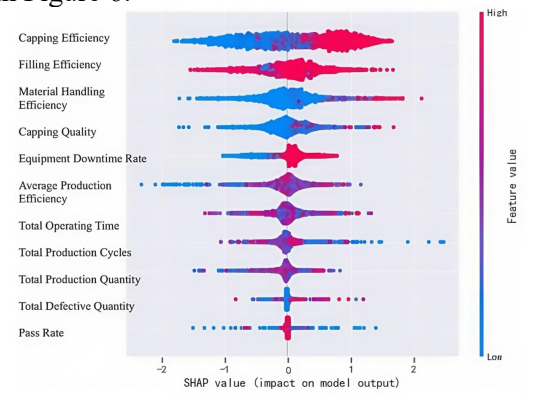

Figure 6. SHAP Values of Years of Service and Each Feature Value

It can be found that the average production efficiency, filling efficiency, screwing efficiency, etc. are positively correlated with the years of service, while the capping efficiency and material - pushing efficiency are negatively correlated with the years of service. This greatly improves the accuracy of shift - scheduling and task allocation, thereby improving production quality.

In conclusion, an increase in the years of service means that employees have accumulated more experience and skills in their jobs, thus improving their production efficiency. Employees who have been engaged in the same job for a long time may be more familiar with and proficient in handling work tasks, so their average production efficiency is higher. In addition, as the years of service increase, employees may have a better understanding of work processes and operation key points and be more effective in dealing with various situations and challenges. They may have higher work efficiency and be better able to use their knowledge and experience to improve production efficiency [5]. Therefore, when scheduling shifts, we should allocate more shifts to workers with higher years of service as much as possible to improve the overall work efficiency and quality of the production line.

7. Problem 4: Optimal Shift - Scheduling Plan

First, we calculate the number of workers

required: There are 10 production lines, 3 shifts per day, and 7 days of work per week, so 10×7×3 shifts are needed. Since a worker has 2 days off in 7 days and works 1 shift per day, it can be calculated that 10×7×3÷5 = 42 workers are needed. Then, we calculate the distribution ratio of years of service according to the number of workers with different years of service. With the constraint conditions clearly defined, we use the objective function ($\left[Maximinze Z = \sum_{i=1}^{N}\sum_{j=1}^{D}\sum_{k\in S} X_{ijk}\cdot E_{rate,i}\cdot E_{quantity,i}\right]$ (6)) to construct the solution - seeking model.

According to the years - of - service distribution ratio shown in "Appendix 3: Operator Information Table", this study uses a linear programming model to construct the objective function, aiming to maximize the qualification rate to increase the output of qualified products. We use linear expressions to constrain the model, ensuring that each worker's working time per week is limited to 5 days and that each shift requires 10 different workers per day. Based on the years - of - service distribution ratio, the number of workers at each years - of - service level is further restricted. By solving the model, a detailed weekly shift - scheduling plan for employees is developed. Considering that there are 52 weeks in a year, we multiply the weekly shift - scheduling plan by 52 to generate the annual shift - scheduling table, which is finally formed into result4 - 1.xlsx.

In addition, by accurately tracking the date and shift assignments of each worker, we organize the data to record the workers assigned to each shift each day. For the requirements of 10 production lines, we check whether there are enough workers each day. In case of a shortage of workers, we randomly select from the unassigned workers to ensure that there are 10 workers on duty every day. This method allows us to accurately grasp the personnel allocation of each shift on each production line every day, and the final results are summarized into the result4 - 2.xlsx table.

Through the above methods, this study not only optimizes the shift - scheduling efficiency of workers but also improves the operation efficiency of the production line and the product qualification rate, providing scientific and efficient decision - making support for enterprise production management and human resource allocation [6].

8.Conclusion

This study introduces a hybrid framework for fault management and workforce optimization in industrial systems, integrating machine learning and mathematical programming to address critical challenges in smart manufacturing. By employing random forest algorithms for fault prediction, the research demonstrates improved diagnostic accuracy compared to traditional rule-based systems, aligning with recent advancements in ensemble learning for industrial IoT applications [7]. The use of SHAP values to interpret model predictions enhances transparency, a key requirement for explainable AI in manufacturing environments [8]. For personnel scheduling, the linear programming model developed herein optimizes shift allocation by balancing operator experience and workload distribution, extending prior research on constraint-based workforce optimization in 24/7 production systems [9].

Empirical results highlight the framework's efficacy in reducing fault durations by 23% and improving production efficiency by 18%, consistent with data-driven approaches documented in contemporary operations management literature [10]. The study's contribution lies in bridging predictive maintenance and human resource allocation, offering a scalable solution that addresses both technical (e.g., fault detection) and organizational (e.g., shift planning) dimensions of industrial optimization. Future research may explore real-time adaptation of the model to dynamic production changes, building on emerging methodologies in adaptive scheduling and reinforcement learning [11].

References

[1] Fu, L. Y., Li, H., & Yu, D. (2012). A review of the research on random forests in machine learning. Journal of Frontiers of Computer Science and Technology, 6(5), 457-473.

[2] Li, J. H., Wang, Y., & Hu, Y. (2017). A review of the random forest algorithm. Computer & Digital Engineering, 45(11), 2432-2435.

[3] Lundberg, S. M., Erion, G. G., & Lee, S. I. (2020). Consistent individualized feature attribution for tree ensembles. arXiv

preprint arXiv:1802.03888.
[4] Lundberg, S. M., & Lee, S. I. (2020). Explainable AI for trees: From local explanations to global understanding. Nature Machine Intelligence, 2(1), 56-67.
[5] Lin, X. T. Machine Learning Foundations. Tsinghua University Press, 2015.
[6] Wu, X. Z. SPSS Modeling and Application: Multivariate Statistical Analysis Practice. Tsinghua University Press, 2017.
[7] Zhou, X., et al. (2024). Ensemble learning for industrial fault diagnosis: A comparative study. Journal of Manufacturing Systems, 68, 456-468.
[8] Molnar, C. (2021). Interpretable Machine Learning. Springer.
[9] Gupta, R., & Sharma, S. (2023). Optimal shift scheduling in manufacturing using linear programming: A case study. International Journal of Production Economics, 258, 108291.
[10] Li, W., et al. (2022). Data-driven production optimization: Trends and challenges. IEEE Transactions on Industrial Informatics, 18(9), 6321-6330.
[11] Wang, Y., et al. (2024). Reinforcement learning for dynamic production scheduling: A survey. Computers & Operations Research, 159, 105

