UTILISATION OF MACHINE LEARNING AND XAI TECHNIQUES TO PREDICT OCCURRENCES OF BOTTLENECKS IN A WAFER FABRICATION PLANT

NUS National University of Singapore

Micron

Department of Industrial Systems Engineering and Management (ISEM) | IE3100M Systems Design Project (Group 10)
Team members: Png Jian Cheng Christopher, Cheng Tze Ning, Iona Dorothy Putri Tanan, Toh Sin Yi
Supervisors: Assistant Professor Cheung Wang Chi (NUS), Dr. Hyeongtae Park (Micron)

Problem Description

Micron's wafer fabrication plant faces a bottleneck when the desired quantity of completed wafer exceeds the maximum capacity of a workstation group (WSG). This causes delay in the production process of wafer products.
This project focuses on using supervised machine learning classification techniques to predict bottlenecks for a particular product DID using simulated datasets over 10-month time period.

Objective

Identify WSGs that are likely to be bottlenecks in a wafer production plant.

Identify key factors that may cause a bottleneck to occur for a WSG

Summary of Methodology

Data Wrangling → Feature Selection → Data Preprocessing → Model Building & Selection → Performance on Hold-out Test Data → XAI Application

Data Wrangling: identification of bottlenecks

For simplification purposes, this project only focuses on one product line: Product_ID_1

Whole dataset → Filtered to only Product_ID_1

Product_ID_1's product line is attached to Route_ID_1, which comprises of different WSGs and their step counts.

? A step count is the number of times the product is being processed in that WSG. If there are 2 step counts, then the product is produced twice in the WSG.

The WSGs and step counts of Product_ID_1 products in Route_ID_1 are filtered based on another dataset.

Each WSG can produce multiple product ID, so the next step is to calculate the proportion of Product_ID_1 in a specific time period. There is also a 3-month gap between production plan and actual production.

$$Ratio = \frac{Amount\ of\ Product\_ID\_1}{Total\ Amount\ of\ all\ products}\ for\ each\ WSG$$

Desired quantity of a product line is then obtained by multiplying *StartedWafers* with the step count and ratio of each WSG.

Concluding code: identifying bottlenecks
If desired quantity > maximum capacity (*CompletedWafers*), it is classified as a bottleneck.
Else, it is not a bottleneck.

Feature Selection

Covariance Matrix

assesses the strength of the relationship between predictors and response variable. Features with High covariances with bottlenecks are selected.

\+

Univariate Feature Selection

evaluates the relationship between a predictor and the response variable (bottleneck). This project uses SelectKBest method that filters only the k highest scoring features according to an ANOVA F-test.

\+

Domain Knowledge

= 5 features are chosen from 34 variables: 'AvgWIPWafers', 'AvgProcTime', 'AvgIntervalTime', 'AvgQueueTime', 'UD%'.

Data Pre-processing

Step 1 The five numerical features identified are standardised.
Step 2 Conduct an 80-20 train-test split on original dataset. Results are reported on the hold-out test data.
Step 3 Conduct a further train-validation split on the train data. Hyperparameter-tuning will be conducted based on results on the hold-out validation data.

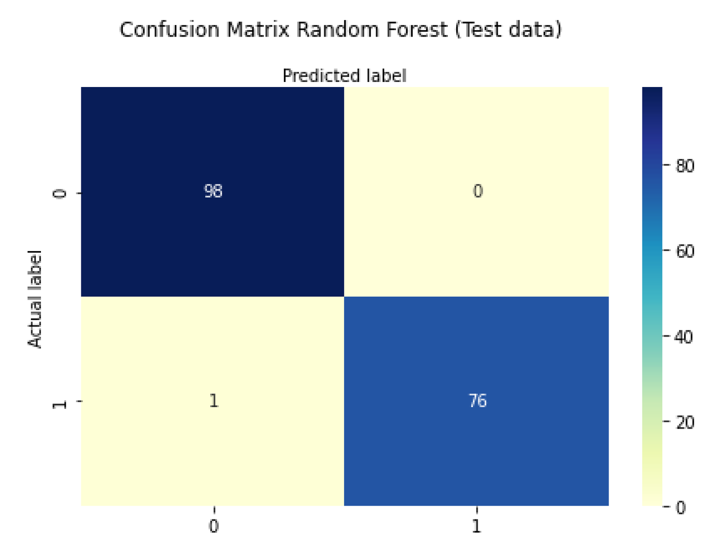
Summary of Model Accuracy

List of models that were built, their test data accuracies, and the best parameters that produced the results (using GridSearchCV)

Model	Accuracy	Best parameters
Logistic Regression	64%	LogisticRegression(C=10, max_iter=100000, penalty='l1', solver='liblinear')
Decision Tree	98.3%	DecisionTreeClassifier(ccp_alpha=0.0025, max_depth=8, max_features='sqrt', max_leaf_nodes=210, min_samples_split=3)
Adaptive Boosting	99.4%	AdaBoostClassifier(learning_rate=1.1, n_estimators=110)
Gradient Boosting	100%	GradientBoostingClassifier(max_features='sqrt', min_samples_split=7)
Random Forest	99.4%	RandomForestClassifier(ccp_alpha=0.005, criterion='entropy', max_depth=7, max_leaf_nodes=680, min_samples_split=3, n_estimators=55)
SVM Linear Kernel	74.9%	SVC(C=0.1, kernel='linear', probability=True)
SVM Polynomial Kernel	80%	SVC(C=0.5, coef0=5, gamma='auto', kernel='poly', max_iter=400, probability=True)
SVM Sigmoid Kernel	56%	SVC(C=0.1, coef0=-60, kernel='sigmoid', probability=True)
SVM Radial Kernel	88%	SVC(C=130, gamma='auto', probability=True)

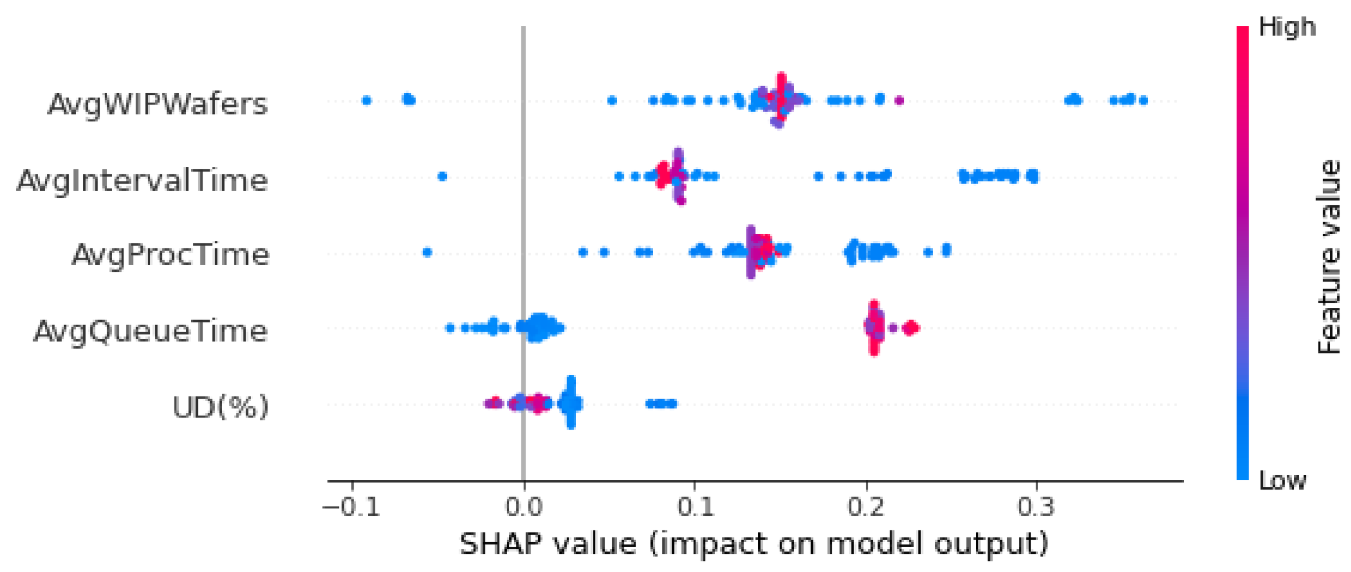
Random Forest metrics

Random forest is chosen since it has the one of the highest test accuracy and is more fitting to the dataset according to domain knowledge.



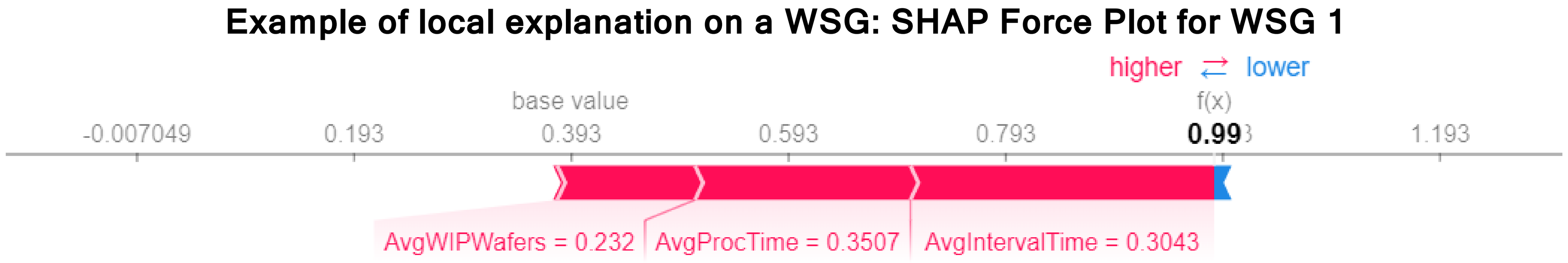
Explainable Artificial Intelligence (XAI) for Random Forest

Global explanation: SHAP Summary Plot for Bottleneck Occurrences



- The summary plot on the right shows that 'AvgWIPWafers' is the most crucial factor used by the random forest model in classifying bottlenecks. SHAP values for "AvgWIPWafers" can also be seen to be mostly positive, thus this meant that there is higher probability of the prediction being a bottleneck regardless of its value.
- Higher 'AvgQueueTime' leads to higher SHAP value, which infers that there is a higher probability of a bottleneck occurring as 'AvgQueueTime' increases

Example of local explanation on a WSG: SHAP Force Plot for WSG 1



- The SHAP force plot for WSG 1 shows that there is a 0.99 chance of a bottleneck occurring.
- Features 'AvgIntervalTime', 'AvgProcTime, 'AvgWIP' are the three most crucial factors that may cause a bottleneck to occur.

Predictions on Test Dataset

After applying the random forest model, WSG_1 is predicted to have the highest occurrences of bottlenecks with 13 instances of bottlenecks.

Model Performance on Hold-Out Data

To validate the robustness of the random forest model, the model is trained and tested on Product_ID_2, which involves a different combination of WSGs.
The results are as follows:

```
Accuracy: 0.816
Precision: 0.831
Recall: 0.855
Confusion matrix:
 [[132  41]
 [ 34 201]]
```

Although the accuracy is lower than the training and testing for Product_ID_1, the model is still performing very well. We can then conclude that the use of classification models such as random forest is able to predict occurrences of bottlenecks in other WSGs and Product IDs, given the availability of training data.

Recommendations and Future Work

This project uses simulation data to train the model. If real-world fabrication level data can be obtained, a more robust classification model that considers the complexities occurring in the real-world can be developed and trained.

Achievements & Outcomes

- ✓ Developed a classification model with high accuracy
- ✓ Provided interpretable explanations for model predictions
- ✓ Identified top factors contributing to a bottleneck occurrence

Benefits

- ✓ Random forest model serves as a robust and effective tool to identify bottlenecks
- ✓ Allow for more efficient planning of wafer production plan
- ✓ Potential increase in wafer manufacturing output

Skills obtained

- ✓ Machine Learning (in Python)
- ✓ Data Analytics
- ✓ Project Management