Aurubis Buffalo
Waiting Time Analysis for Coil Production
Executive Summary

Team 17:
Li Zeng, Aijing Song, Fanglei Su, Xinyu Huang, Yuqi Hou

Our project objective is to build a model to predict the waiting time between each operation in order to better predict the time need for orders and to reduce broken promise dates to customers.

I. Data Cleaning and Handling

1.Dropping Unrelated Attributes

First, to make the dataset clearer and easier to work with, we drop the attributes that is obviously unrelated with the operation waiting time such as minor location, instruction.

2.Handling Missing Values

We count the total number of missing values and percentage of missing values of each attribute. For those attributes' percentage missing greater than 50%, they are all categorical variables and we cannot use imputation. So, we decide to drop these attributes. For other attributes with missing value, we will use listwise deletion if attributes are necessary after feature engineering because dropping small portion of data will not substantially loose statistical power.

3.Transform Data Type

Since some of the categorical variables in the dataset are represented by number such as INSNXU (next unit), we need to carefully separate categorical variables and numerical variables in order to build a precise predictive model.

4.Calculating Waiting Time

To calculate waiting time, we should only focus on the rows whose status are IN which means the materials began loading to the machine. And the waiting time for each coil at each machine is equals to start timestamp of this IN status operation minus finish timestamp of the last IN status operation.

II. Exploratory Data Analysis

1.Order Perspective

First, we analyze the data from the order perspective. According the fourth quartile value, we divide wait time to three categories, long, short and medium. 50% of orders' wait time is medium and the other two categories is almost same. When the schedule deliver date is earlier than actual date, we recognize this kind of order as delay. The wait time of on scheduled orders is almost half of wait time of delay orders. So, analysis and prediction on wait time can help to solve order delay problem.

2.Machine Perspective

Over half of machine's wait time is above average which is 58.5 hours. The Wait time for machine 78 is the longest and Wait time for machine 22 and 44 is the shortest. So, for machine 78, we advise to buy more this kind of machine. For machine 22 and 44 which may not be used very often, so we can sell some of them.

3.Process Analysis

In this part, we also use Sankey Plot make a process analysis focusing on alloy type, machine and wait time. From the Sankey plot, we can see that most of alloy types of coils are A-1, A-2, A-K and A-7. A-1 coils are more likely to go to the machine 46 (long wait). A-K coils are more likely to go to machine 36 (long wait).

III. Feature Engineering

1.Explore Dataset

It can be seen that among 48 features, more than 50% are categorical variables. If we directly use those categorical variables as input, machine will not understand those values, thus we need to give them a numerical value.

2.Target Encoding and New Feature Creation

Here we will adopt 'target encoding' to solve those problems and create some new features. For 'INALLY', we use the first digit as mentioned before. We also count how many steps hidden between two 'IN' status because we guess more steps may extend the waiting time for that order to be processed into next machine. In addition, the yield loss is the percentage of lost weight for that steel part.

3.Three Classical Methods Comparison for Feature Engineering

After preparing all the variables, we will use three classical methods to go on feature selection. We compare the pros and cons of those methods from the overfitting/ underfitting and whether ML-based perspective. Finally, we found that INSNXU AND INSSCL, which represent the next unit and scale respectively, appear two times. We aggregated those selected variables and create a new feature list to prepare for our later model building.

IV. Model Selection

1.Three Metrics for Model Evaluation

In our project, we chose three metrics (mean absolute error, R squared score, and root mean squared error) to evaluate our regression model performance.

2. Hyperparameter Tuning for Improving Model

Next step is using default hyperparameters as baseline model and then tuning these hyperparameters. Grid search is commonly used as an approach to hyper-parameter tuning that methodically build and evaluate a model for each combination of algorithm parameters specified in a grid. To increase accuracy, we focus on the changing degree of decreasing root mean square errors.

3. Predictive models and performance

Model Name	MAE	R2_Score	RMSE
Linear Regression	3772.11	0.98869	14963.68
KNN	1830.62	0.6235	167685.23
Bayesian Model	3846.62	0.9889	14859.94
LightGBM	7860.36	-29268.17	181696.44
Random Forest	**2106.16**	**0.9997**	**2110.09**
Neural Network	1185.73	-4.99	140738.01

V. Conclusion

Based on the results we found, we notice wait time for machine 78 is the longest and machine 22 and 24 is the shortest. And the wait time of on-schedule orders is almost half of wait time of delay orders which means wait time prediction can reduce delayed order.

In model building part, after comparing different models, we found the Random Forest model's performance is best.

Overall, we recommended that purchasing more machine 78 if still under budget control and sell some machine 22 and 44 if possible. Finally, adjusting the delivery date according to the predictive time (using random forest model) is also highly recommended for the purpose of reducing other delay.