Capstone Project

{Fab transfer low yield debug} Final Report

1. **Define the Problem Statement:**

This project is to debug a fab transfer yield delta issue. Fab B shows ~0.7% lower yield (higher Bin 11 failure ~0.7%) than the mother Fab A.

Exploration of the data shows some missing values in PCM (process control monitor) data and a lot of information columns which aren't PCM data. Data preprocessing is conducted to clean the data.

PCA with optimized number of components (WAT parameters) are used for feature analysis to reduce the feature dimensions. Ridge and Lasso regression models are used with Pipeline, GridSearchCV to optimize the hyper-parameters of the models. Top 10 WAT parameters which contributed to the BIN 11 fallouts are derived from the optimized Ridge and Lasso model.

However, further debug shows that the yield data isn't the best way to root cause the problem. ATE test program was changed to collect new POR circuit trip point voltage. Preprocessing of the new data shows missing data and non-numerical data, which aren't PCM data needed. Dropped missing data rows and non-numerical columns.

Used Random Forest model, XGBoost model and Ridge model to compare the top 5 features (PCM parameters) affecting the target value (POR trip point).

Used PCA with Ridge to reduce feature dimension, GridSearchCV and Permutation to optimize hyperparameters and optimize top features.

Select the most reasonable model and top features to assist the low yield debug.

1. **Model Outcomes or Predictions:**

With the old data, only linear model Lasso and Ridge were used. Ridge is better in this case. The top 10 features are very similar in these two models. 9/10 top features overlap in these two models.

|  |  |  |  |
| --- | --- | --- | --- |
| **Lasso Feature** | **Lasso Coef** | **Ridge Features** | **Ridge Coef** |
| Rs\_HRI | 0.022375 | Rs\_PODRPO2 | 0.012936 |
| Rs\_PODRPO2 | 0.020197 | Iof\_PTM9A | -0.010889 |
| Cmim40X50 | 0.019557 | Cmim40X50 | 0.010483 |
| Iof\_PTM9A | -0.019451 | Rc\_TV1 | -0.010032 |
| Rs\_PPORPO2 | 0.01897 | Rs\_NODRPO2 | 0.009208 |
| Iof\_P45L | -0.017601 | Rs\_HRI | 0.008754 |
| Rs\_NODRPO2 | 0.016797 | Iof\_P45 | -0.008469 |
| Rs\_NPORPO2 | 0.01665 | Rs\_NW\_STI | 0.008355 |
| Rs\_HRIserp | 0.01383 | Rs\_NPORPO2 | 0.008239 |
| Rs\_NW\_STI | 0.013069 | Rs\_PPORPO2 | 0.007339 |

With the new data, Random Forest, XGBoost and Ridge Model comparison shows that Random Forest and XGBoost have good feature overlaps (3/5), but no overlap with Ridge model.

From the top 5 feature comparisons among the 3 models, the two ensemble models show 3/5 features are the same. The top 2 features are the same Rs\_PDD\_STI and CONT1\_TM1. The Ridge model shows very different features. Since the PCM parameter and POR trip point measurement are non-linear, I think we will use Ensemble model as reference for the problem debug

|  |  |  |  |
| --- | --- | --- | --- |
| Random Forest | XGBoost | Ridge | Importance |
| Rs\_PDD\_STI | CONTI\_TM1 | BV\_N+PW | Top-1 |
| CONTI\_TM1 | Rs\_PDD\_STI | CONTI\_PO | Top-2 |
| Idl\_N45N | CONTI\_M3 | Cmim40X50 | Top-3 |
| CONTI\_M3 | Rs\_NPORPO2 | Rc\_V2 | Top-4 |
| Rc\_TV1 | Ids\_N4L | Ids\_PTM9A | Top-5 |

In all three models, the actual vs predictions aligned very well on the POR trip point voltage (target values)

**Random Forest Model**

Optimize the number of the tree with n\_estimator [50.100,150]. GridsearchCV shows 150 is the best choice.

The mean squared errors are very low:

Train Mean Squared Error: 2.7402834867335024e-05

Test Mean Squared Error: 0.00020651086359358185

Actual vs predictions with new data

A graph of a graph showing the value of a number of points

Description automatically generated with medium confidence

Top 10 features

A blue and black rectangles

Description automatically generated

One of the tree visualizations

A diagram of a diagram

Description automatically generated

**XGboost Model**

The mean squared values are very low:

Train MSE with XGBoost: 2.0023503796164772e-07

Test MSE with XGBoost: 0.00024875414952773043

A graph of blue dots

Description automatically generated

Top 10 features

A graph of a bar graph

Description automatically generated with medium confidence

**Ridge model**

Train MSE with Gridsearch: 0.00017942668525017176

Test MSE with Gridsearch: 0.0002614496578446856

best alpha value 0.0

Number of components to reach 95% variance: 1

Number of PCA components selected: 70

A graph with blue dots

Description automatically generated

Top Features

A blue rectangular object with black text

Description automatically generated with medium confidence

Actual vs prediction on train and test data

A comparison of blue and white dots

Description automatically generated with medium confidence

1. **Data Acquisition:**

Data are from my company’s Fab transfer PCM (process control monitor) data and the ATE testing data on the wafers. Old data is with yield data from ATE. New data is with improved ATE test data on POR circuit trip point voltage measurements.

1. **Data Preprocessing/Preparation:**

Check if there are any Null values in the data by using isnull().

PCM data are numerical data. So I checked the data information by using info(). Drop all unnecessary object columns, and Null values.

Use train\_test\_split to split the data to train and test data.

Since the data are all numerical, no encoding needed.

1. **Modeling:**

Used linear models: Lasso and Ridge models for the old yield and PCM data.

Used both Ensemble and Linear models for the new parametric data and PCM data: Ridge, Random Forest and XGboost

1. **Model Evaluation:**

Model evaluation using five fold cross validation with GridSearchCV. Predictions vs actual target values are compared for MSE.

XGboost:

Train MSE with XGBoost: 2.0023503796164772e-07

Test MSE with XGBoost: 0.00024875414952773043

Random Forest;

Train Mean Squared Error: 2.7402834867335024e-05

Test Mean Squared Error: 0.00020651086359358185

Ridge:

Train MSE with Gridsearch: 0.00017942668525017176

Test MSE with Gridsearch: 0.0002614496578446856

The model MSEs are all good. But the MSEs aren’t the only determined factor for the final model to be used, but just a reference.

Based on the domain knowledge, ensemble models are better for PCM vs ATE test parameter regression analysis to find the top influential PCMs. The reasons are that PCM and the Trip point measurement aren’t linear relationships. The PCM involves complicated process doping steps etc..

In addition, the PCM data parameters have collinearity. Using linear models could have overfitting issues.

From the ensemble models: Random Forest and XGboost, the top influential PCMs are very similar. Decided to use Random Forest, which has slight better MSE than XGboost.

From feature engineering perspective, the top 2 features are the same between XGboost and Random Forest regressors. Will focus on the top 2 features for low yield debug.