Production Line Performance

Qianying Diao
Xinrong Chen
Shirley Zhang
Jiaru Xu

Agenda

- ❑ Introduction
- ❑ Feature Selection and Engineering
- ❑ Modeling and Evaluation
- ❑ Further Improvement

BOSCH

Data Source

- Kaggle competition: https://www.kaggle.com/c/bosch-production-line-performance/overview
- Predict **internal failures** using thousands of measurements and tests made for each component along the assembly line
- Features represent measurements of parts as moving through production lines
- **Target** is to predict which parts (ID) will fail quality control
- An extremely large number of **anonymized** features with information of production line, station, and feature number
- Features separated into 3 files: **numerical, categorical, and date**

BOSCH

Target and Tools

- Target: predict which parts (ID) will fail quality control
- Tools we use:
  - Loading Data with Amazon S3 bucket
  - Exploratory Analysis: **Spark** SQL, **Spark** Dataframe
  - Feature Selection and Engineering: **Spark** SQL, **Spark** Dataframe, UDF in Pandas, visualization with matplotlib
  - Modeling: **Spark** ML (Logistic Regression, Random Forest, Gradient Boosted Tree), XGBoost

Biggest Challenges

...

- **4,265** features in the dataset
- Extremely **imbalanced** data

```
In [12]: cat.printSchema()
```

```
root
 |-- Id: integer (nullable = true)
 |-- L0_S1_F25: string (nullable = true)
 |-- L0_S1_F27: string (nullable = true)
 |-- L0_S1_F29: string (nullable = true)
 |-- L0_S1_F31: string (nullable = true)
 |-- L0_S2_F33: string (nullable = true)
 |-- L0_S2_F35: string (nullable = true)
 |-- L0_S2_F37: string (nullable = true)
 |-- L0_S2_F39: string (nullable = true)
 |-- L0_S2_F41: string (nullable = true)
 |-- L0_S2_F43: string (nullable = true)
 |-- L0_S2_F45: string (nullable = true)
 |-- L0_S2_F47: string (nullable = true)
 |-- L0_S2_F49: string (nullable = true)
 |-- L0_S2_F51: string (nullable = true)
 |-- L0_S2_F53: string (nullable = true)
 |-- L0_S2_F55: string (nullable = true)
 |-- L0_S2_F57: string (nullable = true)
```

- The number of rows in cat: 1,183,747
- The number of columns in cat: 2,141

BOSCH

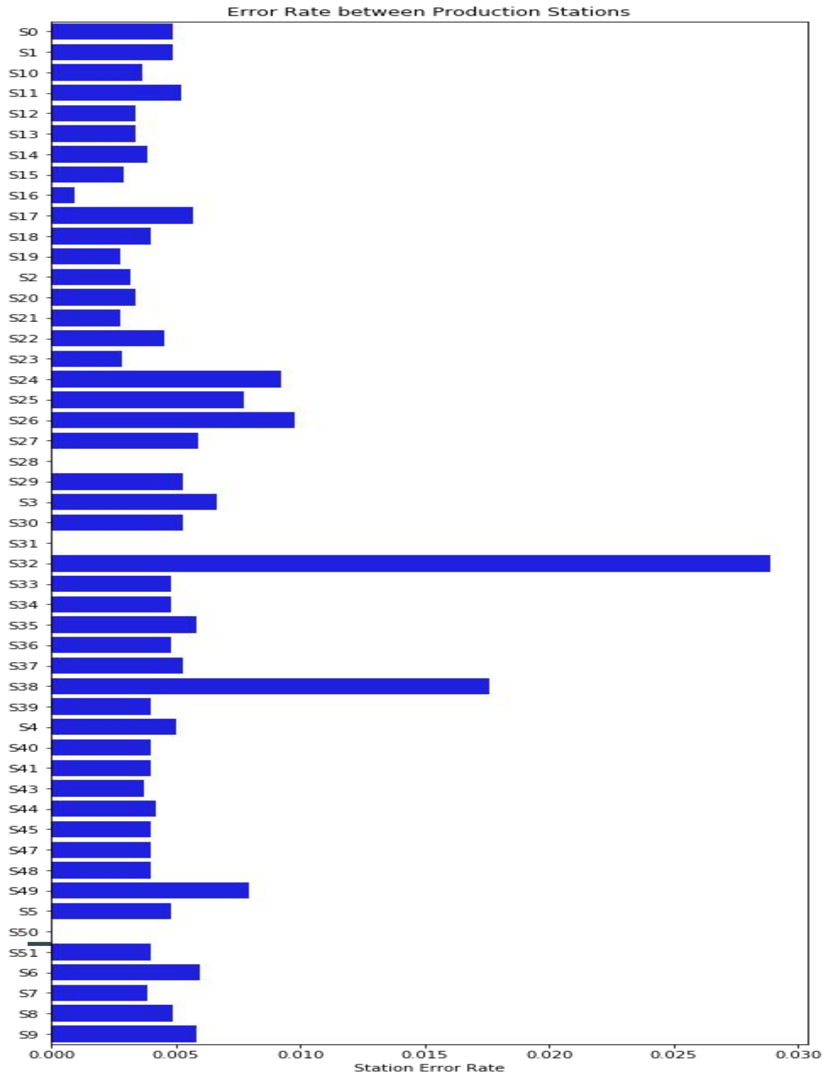
0.58%

v.s

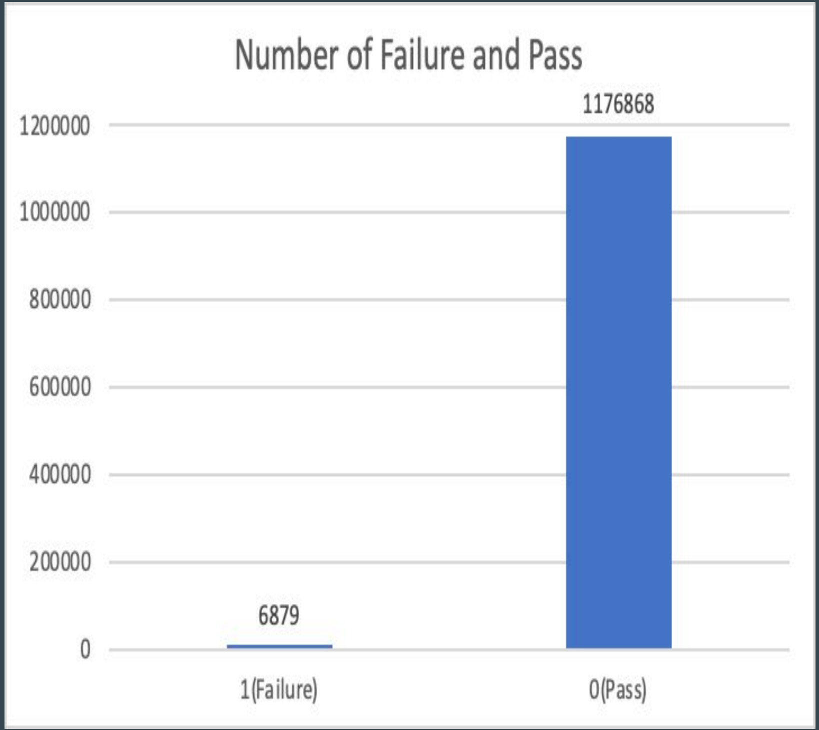
Success Rate=99.42%

Failure rate at different stations

Top 1: S32
Top 2: S38
Top 3: S26



Imbalanced Response

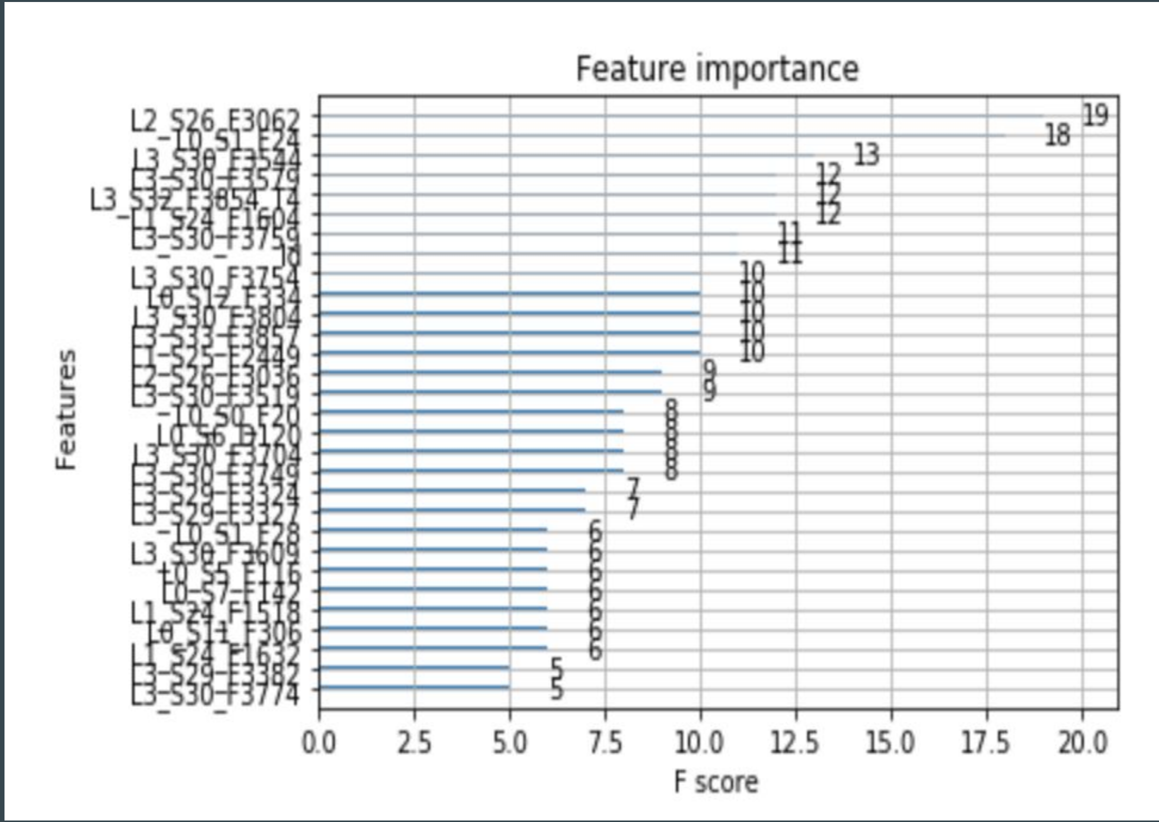


- Data Preparation:
  Oversampling (SMOTE)

- Model Evaluation:
  Matthews correlation coefficient (MCC)

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Feature Selection --- XGBoost



Feature Engineering

01 Numeric Features

- Failure rate differs based on stations
- Generate Features for each station

02 Categorical Features

- Type of features has most NaN
  (Only two have missing values less than 900k.)
- Pick columns with largest variance

03 Date Features

- A lot of duplicate columns for each station
  (Prob. for station to have the same values: 0.98)
- Observations are unique for station Id pair

BOSCH

Modeling & Evaluation

Model	MCC Score
Random Forest	0
Logistic Regression	0.019
Gradient Boosted Tree	0.097
XGBoost	0.15

Further Improvement

- Alternative methods to deal with imbalanced, sparse data
- Hyperparameters tuning

Thank you

Q & A