CNH Industrial: Data Engineering and Systems Connectivity

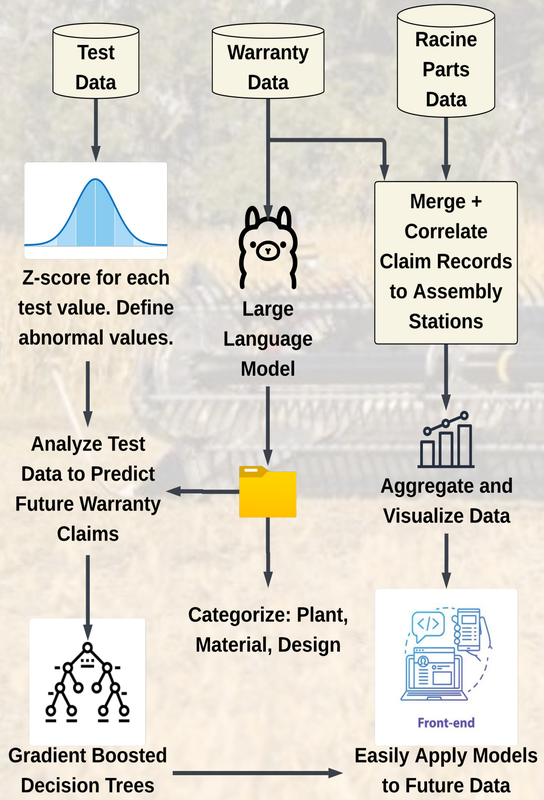
Nhu Nguyen (TA), Jing E Gan, Alan Kang, Kevin Huffman, Ying Zhou, Daren Lim, Lu-Hsia Lo, Yizheng Jiang

The Data Mine

Introduction/Objective

- **CNH Industrial** is a global leader in agricultural and construction equipment.
- Dealer **warranty comments** hold valuable failure insights but are difficult to analyze due to **unstructured text.**
- This project develops an **algorithm** to establish **predictive causality** between test data and warranty failures.
- A user-friendly **front-end software interface** will be delivered so engineers can easily leverage this tool for decision-making.

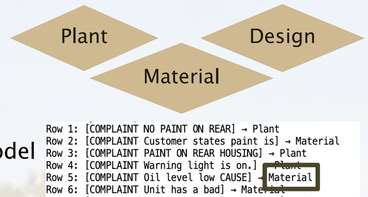
Project Flow



1 Data Processing & Feature Preparation

a. Dealer Comments Interpretation:

- Parse Dealer Comments Using Large Language Model (gemma 3)
- Converting **Unstructured Text** into **Structured Categories**
- 4M + 1D Classification

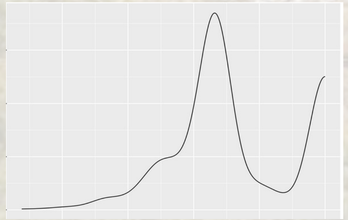



```
Row 1: [COMPLAINT NO PAINT ON REAR] → Plant
Row 2: [COMPLAINT Customer states paint is] → Material
Row 3: [COMPLAINT PAINT ON REAR HOUSING] → Plant
Row 4: [COMPLAINT Warning light is on.] → Plant
Row 5: [COMPLAINT Oil level low CAUSE] → Material
Row 6: [COMPLAINT Unit has a bad] → Material
Row 7: [COMPLAINT Air filter housing broken] → Plant
Row 8: [COMPLAINT Air intake tube bolts] → Plant
Row 9: [COMPLAINT COMPLAINT THAT THE AIR] → Plant
Row 10: [COMPLAINT Transmission fluid leak CAUSE] → Plant
Row 11: [COMPLAINT Tractor will not start.] → Plant
Row 12: [COMPLAINT Fuel filter housing loose] → Plant
Row 13: [COMPLAINT tank would take long] → Plant
Row 14: [COMPLAINT Cover is loose. CAUSE] → Plant
Row 15: [COMPLAINT SEGMENT 4 REROUTE FUEL] → Plant
Row 16: [COMPLAINT THERE IS A FUEL] → Material
Row 17: [COMPLAINT Fuel leak under unit.] → Plant
Row 18: [COMPLAINT Diesel fluid leaking CAUSE] → Plant
Row 19: [COMPLAINT Exhaust leaking from muffler;] → Plant
Row 20: [COMPLAINT Machine has white smoke] → Plant
```

b. Z-score Calculation:

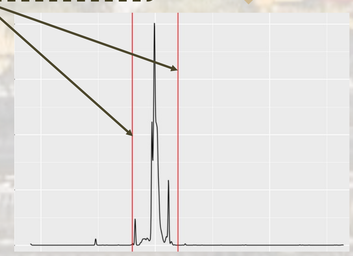
Distribution of C1 Clutch test pass rates

- Analyze C1 Clutch test data
- Identify which values are "anomalous"



Red Line = Z-score Threshold

- Convert raw data → to Z-scores
- Added two red lines



Distribution of C1 Clutch Z-scores

2 Root Cause & Predictive Modeling

a. Correlation & Root Cause:

Input Test Data

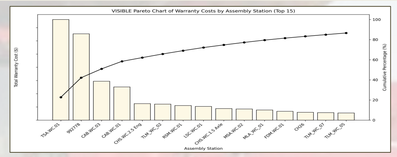
Data Cleaning and Standardization (Racine Parts Data (BOM), Warranty Claims Data)
→ Standardize field names (Part_Number, etc.)
→ Remove whitespace / Standardize to uppercase
→ Remove duplicates (to prevent data bloat)

Part Number Matching (Left Join)
→ Warranty Claims + Assembly Station
→ Retain all claim records
→ Flag unmatched data (unmatched claims)

Map Warranty to Assembly Station
Verify that the record counts are consistent.
Verify that the total cost remains unchanged.

Data Aggregation (Frequency + Cost)
Filter Valid Factory Failure Data
→ Group by Assembly Station
→ Find: Failure Frequency, Total Warranty Cost
→ Compute Percentage, Cumulative Percentage

Pareto Analysis



Output: Critical Issue Workstation

b. Predictive Causality:

Build Label (Y) → Faulty = 1 & Non-faulty = 0

Feature Engineering → Construct 28-Dimensional Test Features

Anomaly Detection (Isolation Forest) → Assess the Relationship Between Anomalies and Faults

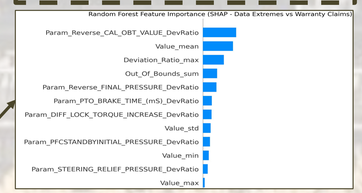
```
--- ML Proof of Concept Results ---
Total Transmissions Analyzed: 1951
Total True Failures (Target=1): 1

1. Unsupervised Anomaly Detection (Isolation Forest)
Top 200 anomalies contained 0 true failures (baseline random expectation: 0.1).

2. Random Forest Feature Importance
Top 10 Mathematically Strongest predictors of Failure:
                                Feature  Importance
Param_Reverse_CAL_OBT_VALUE_DevRatio     0.074341
                           Value_mean    0.067191
                      Deviation_Ratio_max  0.046847
                          Out_Of_Bounds_sum  0.031974
Param_Reverse_FINAL_PRESSURE_DevRatio     0.030333
Param_PTO_BRAKE_TIME_(mS)_DevRatio        0.020211
Param_DIFF_LOCK_TORQUE_INCREASE_DevRatio  0.019431
                            Value_std     0.017354
Param_PFCSTANDBYINITIAL_PRESSURE_DevRatio  0.016002
                            Value_min     0.013269
```

Supervised Learning (Random Forest) → Predict Failure Probability

SHAP Interpretation (Feature Importance) → Identify Key Test Variables



3 Model Improvement & Results

Initial Model Multi-class Prediction (Many Warranty Classes)

The model's performance is very poor. Accuracy = 22%

Problem Analysis Severe Data Imbalance
→ Only 2-3 samples per class
→ Model unable to learn

Improvement Strategy 1: Category Merging (Grouping)

Merge the categories into 5 major groups:

- No Transmission Issue
- Mechanical
- Electrical
- Hydraulic
- Other

≥ 15 samples per class

Model Retraining (Cross-Validation)

Performance = 0.505
(Improved, but still insufficient)

Improvement Strategy 2: Binary Classification

Redefine Task:
Warranty Fail = 1
No Warranty Fail = 0

Model Training (Binary Classifier)

Model Evaluation (ROC-AUC)

Performance = 0.768

FINAL MODEL

Results & Conclusion

- While we cannot yet reliably predict the exact failing component, we found success in using **transmission test data** to predict a **general group of failures.**
- We achieved the **highest reliability** in predicting whether any warranty failure will happen at all **(binary classification).**
- This provides **valuable information to the CNH team**, enabling them to proactively investigate these **tractors of interest** to find the exact issue.

Future Plans

- Focus on developing **predictive models** to better identify **potential failures** before they occur.
- Implement **machine learning algorithms** to improve **defect prediction accuracy** and integrate **real-time data processing.**
- Enhance the **user interface** to allow stakeholders to **interact with and interpret** the model outputs.

Acknowledgements

We would like to thank CNH Industrial for providing the data and project support.

Special thanks to Mike Murray and our corporate mentors for their guidance.

We also appreciate the support from the Data Mine program and our TA, Nhu Nguyen.