Anomaly Detection in Agricultural Applications

Data Infrastructure: Daniel Choi, Sarayu Byreddy

Data Science: Thanh Bui Minh, Tristan Do, Aarav Mohanty, Ashika Gogineni, Piyush Deshpande

Data Visualization: Winston Lin, Tyler Grabowski, Rick Sun

PURDUE UNIVERSITY | The Data Mine

Introduction & Motivation

What is CTB, Inc.?

CTB, Inc. is a global company that develops innovative technologies to help agricultural producers and processors operate more efficiently. With a growing population and decreasing farmland, their mission is to create solutions that ensure a safe, affordable, and abundant global food supply.

This project focuses on enhancing **Cynergy™**, CTB's cloud-based platform, by improving **predictive maintenance**. Using **data modeling and anomaly detection**, the system will analyze equipment data to identify potential failures early, reducing downtime, increasing efficiency, and preventing costly disruptions.

Figure 1: Poultry processing equipment

Motivation

High-frequency, independent farmhouse & processing plant sensors generate abundant data that can signal when something is deviating from the norm. An anomaly detection system can notify users of insights leading to better resource management, higher yields, and improved sustainability.

Objectives

Data Infrastructure: Preliminary development of data pipelines and data storage

Data Science: Develop anomaly detection system for poultry houses

Data Visualization: Design and deploy an interactive dashboard and email notification system

Research Methodology

Data Infrastructure: Merged, cleaned, and structured data using Python, ensuring quality, consistency, and accessibility for analysis

Data Science: Identify key features, filter anomalies, research models, develop task-specific anomaly detectors, and validate using Chore Time/MEYN data with precision, recall, and F1-score analysis

Data Visualization: Create and assess visualizations to highlight key data relationships, refine insights, and transition from PowerBI to Python (Shiny, Dash, PySpark) for improved analysis

References & Acknowledgements

Acknowledgements

Mentors – Rob Janek, Bob Christenson

Data Mine Staff – Cai Chen, Ashley Arroyo

Teaching Assistant – Tristan Bridewester

References

Sprang, R. van. (2017). Condition Based Maintenance at Marel Poultry? https://pure.tue.nl/ws/portalfiles/portal/230378638/Master_Thesis_Lard_van_Zuylen.pdf

Zhu, T., Ran, Y., Zhou, X., & Wen, Y. (2024, March 22). *A survey of Predictive Maintenance: Systems, purposes and approaches*. arXiv.org. https://arxiv.org/abs/1912.07383

Website LinkedIn

Findings & Solutions

Data Infrastructure

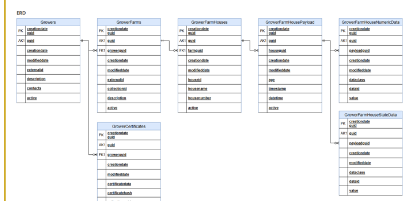


Figure 2: Here is an Entity Relationship Diagram (ERD) of the initial data we received. Every major component of the grower entity is broken down and linked with unique identifiers (UIDs). In this ERD, the payload numeric data contains one dataid and its respective value. Later, we reshaped the payload numeric data to contain all dataid values that correspond to the same payloadguid to reduce the number of rows and repeated UIDs, dates, and dataclasses.

Data Science

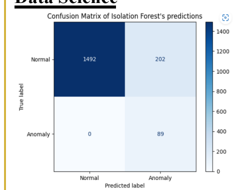


Figure 3a: The confusion matrix above shows that 100% of actual anomalies have been captured/predicted by the model. Also, only 15% of normal data is misclassified as anomalies.

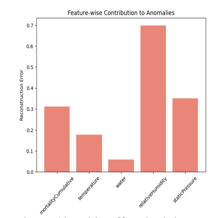


Figure 3b: This offers insights into the contributions of relative humidity to anomalies, having the highest error. Water contributes the least, implying it is less impactful for anomaly detection

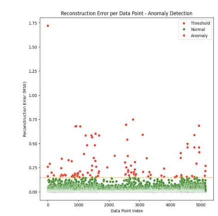


Figure 3c: The vast majority cluster below the threshold, while the anomalies scatter above the line. Therefore, the errors stem from failing to reconstruct from anomalous inputs, making reconstruction error a good metric for detection

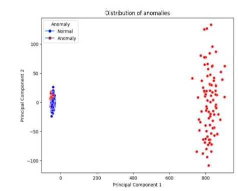
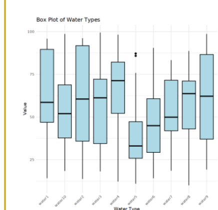


Figure 3d: The scatterplot of the distribution of testing data using PCA. This shows how the Isolation Forest decides to separate normal data from anomalies on a 2D spatial field.

Data Visualization



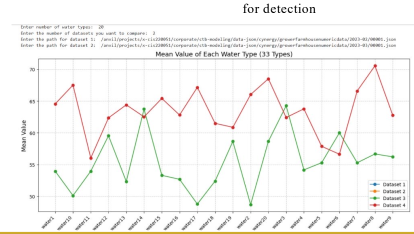


Figure 4a: The figure on the left contains a series of box plots demonstrating the distributions of several water lines within the poultry houses. This information is valuable to understand the average value and variance of the water levels.

Figure 4b: The figure on the right similarly depicts water level values. This graph shows the variability in water levels between different payloads.

Conclusion

Achieved

- Partitioned data in memory efficient manner
- Successfully captured all anomalies, achieving an 89% overall accuracy on predictions
- Built interactive dashboard to easily interpret conditions causing anomalies

Future Plans

- Develop system to automatically notify maintenance about anomalies detected
- Further develop dashboard to display data with the most significant correlation with anomalies
- Test model on live data

The Data Mine Corporate Partners Symposium 2025