AUTOMATED SPARK-BASED DATA FILTERING TOOL FOR TRIP QUALIFICATION

Raja Ali, Avanish Mallya, Vihaan Pradeep, Jazmin Pulido, Muhammad Rizwan, Prisha Singh, Joshua Tsou, Vishal Wagh, Hana Zoaib

INTRODUCTION

The problem:

- Noisy data complicates analysis.
- Duplicate entries cause inaccuracies.
- High null values require extensive cleaning.

The goal:

- An interactive UI
- Eliminate noisy data
- Enhance ML model decision making
- Improve model accuracy & remove false positives

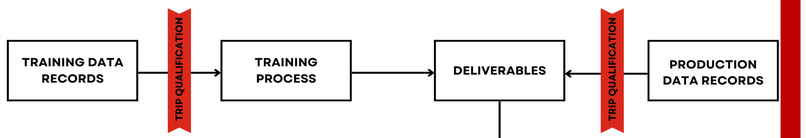


Fig 1. "Trip Qualification" filters noisy records from training and production data, improving model accuracy and decision metrics

METHODS

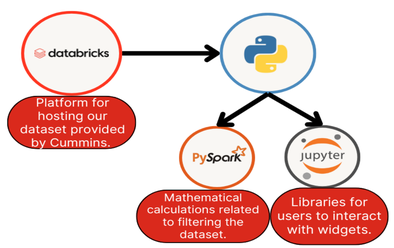


Fig 2. Tech Stack For Data Engineering Pipeline

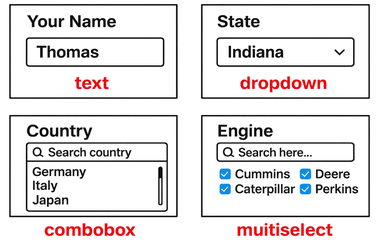


Fig 4. Overview of Databricks Interactive Widgets used in our project

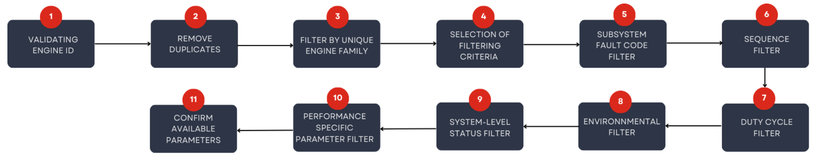


Fig 3. Filtering Order

CONCLUSIONS

Conclusion:

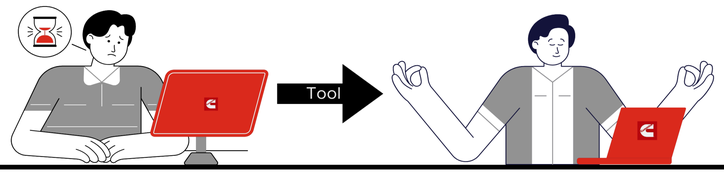
- The filtered data set retains less than 15% of the original data, significantly reducing noisy data while preserving only the most relevant trip information.
- The filters enhance the quality and accuracy of trip data provided to Cummins' machine learning model, enabling more precise predictions and improved overall performance.

Future Work:

- The automated filtering tool will be used by Cummins' non-data scientists and engineers to efficiently preprocess trip data.
- The user Interface will be enhanced to improve usability, making it more intuitive and accessible for non-technical employees at Cummins.

RESULTS

Fig 5. Employees are enthusiastic to have a centralized tool for data filtering tool instead of various people trying different approaches that is 'not vetted'



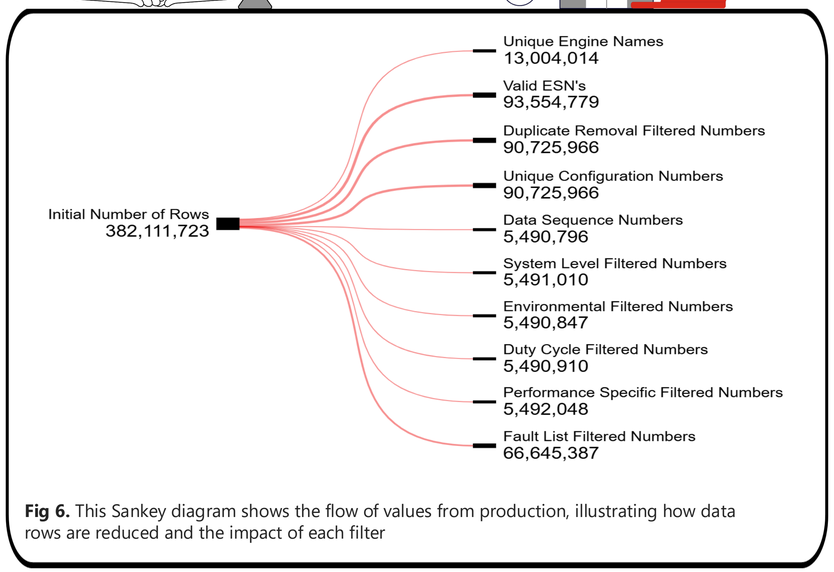


Fig 6. This Sankey diagram shows the flow of values from production, illustrating how data rows are reduced and the impact of each filter

Fig 7. Initially it cost Cummins $20,000 to complete the tasks manually versus $12,000 after our tool. This is calculated as the product of the average salary of a Cummins engineer which is around $100 (Cummins) and the total time it took before and after the tool, i.e., 200 hours versus 120 hours

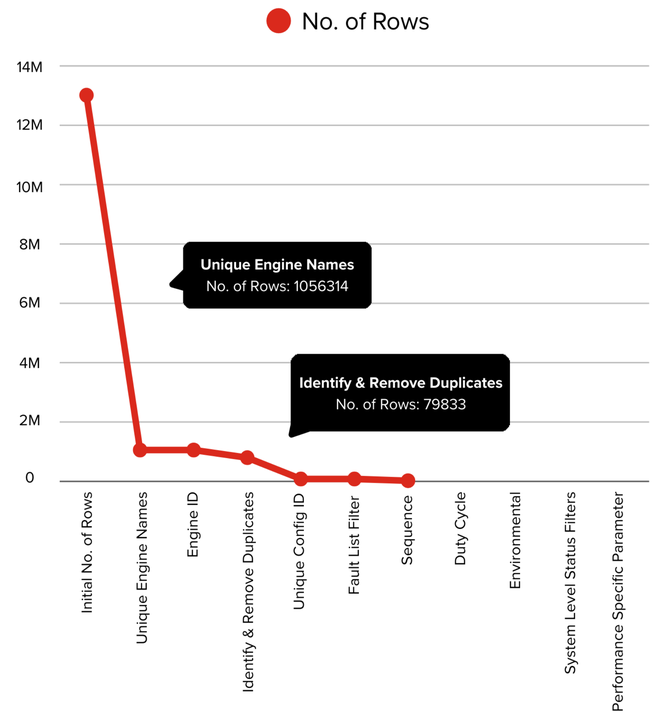


Fig 8. X15 2021 EGR Cooler Diagnostics Model Data Filtering

ACKNOWLEDGMENTS AND REFERENCES

We gratefully acknowledge the SSDS (Self Service Data Science) team, including Yash Chawla, Rich Thielmeyer, Michael Akinola, and Claudia Franchville, for their invaluable support and insights throughout this project.