Data-Driven Evaluation of Truck Driving Performance with Statistical and Machine Learning Methods

Vid Nemec
vidotti.nemec@gmail.com
Jožef Stefan Institute
Ljubljana, Slovenia

Gašper Slapničar
Jožef Stefan Institute
Ljubljana, Slovenia
gasper.slapnicar@ijs.si

Mitja Luštrek
Jožef Stefan Institute
Ljubljana, Slovenia
mitja.lustrek@ijs.si

Figure 1: Truck driving simulator developed by AAER Research d.o.o.

Abstract

This paper investigates which driving features (e.g. speed, acceleration, braking) most strongly affect driving efficiency in a truck simulator environment. The work systematically compares statistical methods (thresholding based on percentiles, IQRs, expert rules) with machine learning methods (clustering using K-means) for driver assessment. In addition to practical machine learning experimentation, the analysis incorporates expert knowledge and insights from recent research. This approach evaluates the agreement and differences between the two approaches and aims to interpret them.

Keywords

Driving simulation, fuel efficiency, percentiles, K-Means, SHAP, statistical thresholds, machine learning, clustering

1 Introduction

Reducing fuel consumption in road transport is a critical goal for sustainability and cost-efficiency [1]. Prior research, such as [2, 3], highlights the impact of driver behaviour - particularly acceleration, braking, and speed profiles on overall fuel efficiency. Yet, how to most effectively quantify and compare drivers remains an open question [4]. This paper addresses which driving features most strongly influence efficiency in a simulated truck driving environment, comparing classical statistical thresholding, based on expert knowledge, with clustering - based machine learning. Applying known methods, we test whether unsupervised ML can identify driver features with stronger influence on fuel consumption than fixed-threshold rules, providing a data-driven baseline for future model-based feedback.

In addition, we compare the empirical outcomes of our ML analytics with insights from recent literature and the practical judgement of a driving expert, to pinpoint where domain knowledge aligns or conflicts with the models. This dual perspective enables a richer interpretation of driver assessment tools and informs the design of future vehicle feedback and incentive systems.

2 Related Work

Recent studies have evaluated driver behaviour for fuel efficiency using both statistical rules and machine-learning approaches. Sullivan et al. present a TORCS-based simulator with a realistic fuel-economy model, enabling safe, repeatable analysis of eco-driving strategies [5]. Maisonneuve characterises driver energy efficiency across driving events and proposes a grading/ranking method based on identified parameters [6]. Zhao et al. develop a simulator-based eco-driving support system with real-time feedback and post-drive reports, demonstrating measurable reductions in fuel use and emissions [7]. Ma et al. provide a scoping review of energy-efficient driving behaviours and applied AI methods [8]. Prototype driver-training systems have been proposed [9], and large-scale, data-driven frameworks to incentivise efficient driving have been developed [3, 10].

Most studies agree that key features include speed, throttle, brake usage, and sometimes gear selection, but differ on methods for quantifying and weighting these features. Machine learning clustering (e.g., K-means) and feature importance analysis (e.g., SHAP) are increasingly used, offering potential improvements in objectivity and interpretability of drivers.

Information Society 2025, Ljubljana, Slovenia

https://doi.org/DOI10.70314/is.2025.skui.4765

3 Methods

3.1 Data Collection and Preprocessing

Driving data were collected from a high-fidelity truck simulator developed by AAER Research d.o.o., which continuously recorded multiple parameters including pedal positions, steering wheel angle, vehicle speed, location, and segment identifiers. To ensure data quality, missing or zeroed pedal values were imputed. The signals were then resampled into 1-second windows, where for each parameter we computed the minimum, maximum, mean, and median values. This aggregation approach was chosen over raw resampling because the signals are irregular, zero-inflated, and not normally distributed, making window-based statistics more representative of driver behavior. In addition, the last observed cumulative distance within each window was retained to preserve distance continuity. Finally, the processed signals were aligned with the boundary of the scenario segment, allowing a consistent basis for later efficiency evaluation.

3.2 Rule-based Aggregation of Segment Labels

We aggregated per-segment labels (*PASS*/*WARN*/*FAIL*) into an overall per-driver rating using a linear severity score. A *FAIL* indicates a strong threshold exceedance and is therefore weighted twice a *WARN*, yielding a simple, interpretable metric that tolerates occasional minor deviations.

$$S = 2\,\#\text{FAIL} + \#\text{WARN},$$

$$\text{Rating}(S) = \begin{cases} \text{Good}, & S \le 2, \\ \text{Warning}, & 3 \le S \le 5, \\ \text{Bad}, & S \ge 6\,. \end{cases}$$

This 2:1 weighting reflects relative severity (a *FAIL* is a clearer breach of the threshold than a *WARN*) and preserves stability: small label fluctuations do not flip a driver from *Good* to *Bad*. The middle band (*Warning*) collects borderline cases for review.

Table 1: Per-driver severity summary ($S = 2{\cdot}\#\text{FAIL}+\#\text{WARN}$).

Driver	#WARN	#FAIL	S	Rating
1	4	1	6	Bad
10	5	1	7	Bad
2	7	2	11	Bad
3	4	0	4	Warning
4	4	0	4	Warning
5	6	2	10	Bad
6	3	0	3	Warning
7	3	0	3	Warning
8	4	0	4	Warning

3.3 Machine Learning

3.3.1 K-means clustering. Unsupervised clustering of K-means (k = 3) was applied per segment on standardized aggregated characteristics (acceleration / braking variability, coasting, use of cruise control, speed-related measures). Clusters were assigned semantic labels *Good*/*Warning*/*Bad post hoc* by ordering clusters by their mean fuel rate (`fuel_mean`): lowest → *Good*, middle → *Warning*, highest → *Bad*. We then examined cluster centroids (mean feature profiles) and visualised the result as per-segment heatmaps.

3.3.2 SHAP with LightGBM model. As an orthogonal check of feature relevance, we applied SHAP to a separate LightGBM model predicting fuel rate; this diagnostic analysis is independent of clustering and highlights variables linked to higher consumption (Table 2).

4 Results

4.1 Statistical Thresholding Approach

Based on the analysis of related worke outlined in Section 2, we decided to benchmark driver efficiency based on selected driving features.We investigated two methods covering complementary metrics of acceleration and braking, namely:

- Percentile-based thresholds for gas pedal
- IQR method for brake pedal

Percentiles were chosen for the gas pedal because the signal is highly zero-inflated and not normally distributed, making distribution-aware thresholds more suitable. Braking behavior is irregular and heavy-tailed, where IQR offers a robust way to capture abnormal events. In essence, the IQR rule sets a dispersion-anchored cut-off above Q3-robust to heavy tails-whereas percentile thresholds fix the share of events flagged. Thresholds were determined by examining histograms of pedal deltas (Figure 2), ensuring that cutoffs meaningfully separated typical from extreme behavior. This procedure enabled transparent, segment-level benchmarking of driver performance.

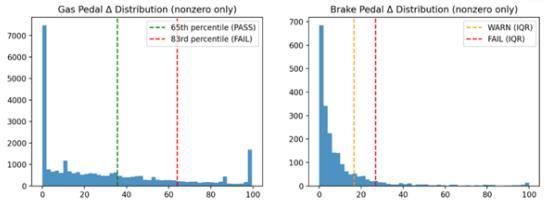


Figure 2: Histograms for both pedals

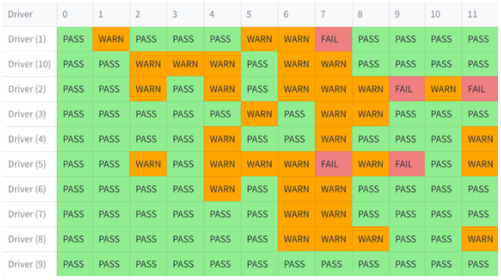
Threshold characterisation:

- Gas Pedal: We applied percentile-based thresholds (65th for WARN, 83rd for FAIL) to the gas pedal delta (change in 0,1 second). This approach better captures outlier acceleration behavior while avoiding over-penalizing normal operation. We removed windows where cruise control was active for more than 30% of the time to reduce automation bias in pedal measurements. It was chosen to balance isolating manual control with keeping enough observations.
- Brake Pedal: We applied an interquartile–range rule computed from the empirical distribution in each segment: with the third quartile $Q3$ and the interquartile range $\text{IQR} = Q3 - Q1$, we set *WARN* at $Q3 + 0.5\,\text{IQR}$ and *FAIL* at $Q3 + 1.5\,\text{IQR}$. It flags both frequent moderate excesses (*WARN*) and rare but severe braking events (*FAIL*) without over-penalising normal behaviour.

Certain segments in the driving scenario required strong braking due to test design (e.g., safety-critical stops). These were labelled as SAFETY and excluded from efficiency scoring, as they reflect controlled conditions rather than natural driving quality.

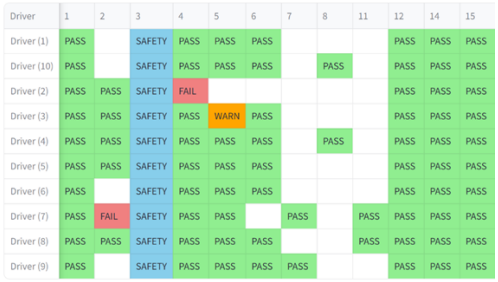
The resulting classifications are summarised as heatmaps (Figures 3 and 4), where rows correspond to drivers and columns to scenario segments. Cells are coloured green (PASS), orange

(WARN), and red (FAIL), providing an intuitive visual overview of performance variability. PASS/WARN/FAIL are segment-level, per-driver labels that state whether the segment was driven efficiently in terms of fuel use: PASS = efficient, WARN = borderline, FAIL = inefficient. These labels refer only to fuel consumption, not safety or travel time. Blank (white) cells indicate cases without an assigned label—either *SAFETY* segments excluded from scoring or driver–segment pairs with too few events to make a reliable decision.



Driver	0	1	2	3	4	5	6	7	8	9	10	11
Driver (1)	PASS	WARN	PASS	PASS	PASS	WARN	WARN	FAIL	PASS	PASS	PASS	PASS
Driver (10)	PASS	PASS	WARN	WARN	WARN	PASS	WARN	WARN	PASS	PASS	PASS	PASS
Driver (2)	PASS	PASS	WARN	PASS	WARN	PASS	WARN	WARN	WARN	FAIL	WARN	FAIL
Driver (3)	PASS	PASS	PASS	PASS	PASS	WARN	PASS	WARN	WARN	PASS	PASS	PASS
Driver (4)	PASS	PASS	PASS	PASS	WARN	PASS	PASS	WARN	PASS	PASS	PASS	WARN
Driver (5)	PASS	PASS	WARN	PASS	WARN	WARN	WARN	FAIL	WARN	FAIL	PASS	WARN
Driver (6)	PASS	PASS	PASS	PASS	WARN	PASS	WARN	WARN	PASS	PASS	PASS	PASS
Driver (7)	PASS	PASS	PASS	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	PASS
Driver (8)	PASS	PASS	PASS	PASS	PASS	PASS	WARN	WARN	WARN	PASS	PASS	WARN
Driver (9)	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS

Figure 3: Heat map of the gas pedal through segments using percentiles method



Driver	1	2	3	4	5	6	7	8	11	12	14	15
Driver (1)	PASS		SAFETY	PASS	PASS	PASS				PASS	PASS	PASS
Driver (10)	PASS		SAFETY	PASS	PASS	PASS		PASS		PASS	PASS	PASS
Driver (2)	PASS	PASS	SAFETY	FAIL						PASS	PASS	PASS
Driver (3)	PASS	PASS	SAFETY	PASS	WARN	PASS				PASS	PASS	PASS
Driver (4)	PASS	PASS	SAFETY	PASS	PASS	PASS		PASS		PASS	PASS	PASS
Driver (5)	PASS	PASS	SAFETY	PASS	PASS	PASS				PASS	PASS	PASS
Driver (6)	PASS		SAFETY	PASS	PASS	PASS				PASS	PASS	PASS
Driver (7)	PASS	FAIL	SAFETY	PASS	PASS		PASS		PASS	PASS	PASS	PASS
Driver (8)	PASS	PASS	SAFETY	PASS	PASS	PASS			PASS	PASS	PASS	PASS
Driver (9)	PASS		SAFETY	PASS	PASS	PASS	PASS			PASS	PASS	PASS

Figure 4: Heat map of the brake pedal through segments using IQR method

4.2 Comparison of Thresholding and Clustering

A focused comparison was carried out on three representative track segments: Segment 1, Segment 8, and Segment 4 using the two complementary methods described in Section 3 (statistical thresholding and K-means clustering). For visualization only, we projected standardised features onto two principal components (PCA) per segment; clustering and label assignment were performed in the original standardised space.

4.2.1 Segment 1 (Steady Acceleration). The percentile method flagged only one driver as exceeding the 'FAIL' threshold, while most achieved the 'PASS' status. The clustering of K-means produced a tightly grouped 'Good' cluster for most drivers, with a single 'Bad' outlier (visible in PCA as an isolated point on the positive PC1 axis). Agreement between methods was high (>85 %), suggesting that, in simpler acceleration scenarios, single-feature metrics and multidimensional clustering agree well.

4.2.2 Segment 4 (Prolonged Uphill Driving). Here the disagreement was most pronounced. The percentile rule classified many drivers as *PASS* because their maximum throttle did not exceed the cut-off. In contrast, K-means frequently assigned them

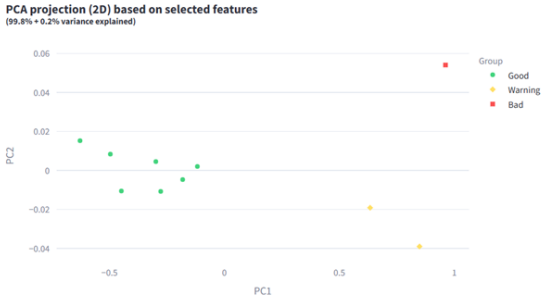


Figure 5: K-means graph for 1st segment

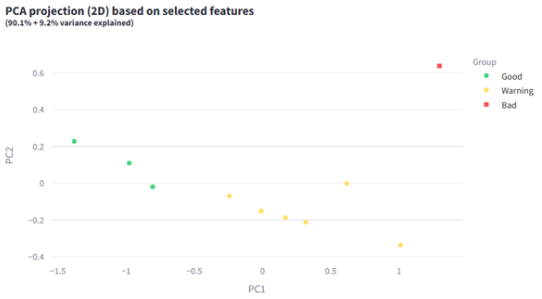


Figure 6: K-means graph for 4th segment

to *Warning* or *Bad*. The 2D PCA projection (Figure 6) shows these drivers displaced from the *Good* centroid, driven by sustained high-load throttle (elevated accelerator mean/variance), low coasting, and reduced cruise-control usage—patterns that the single-peak percentile metric does not penalize. This highlights clustering's sensitivity to cumulative demand and multi-feature context, whereas the percentile approach captures only isolated exceedances.

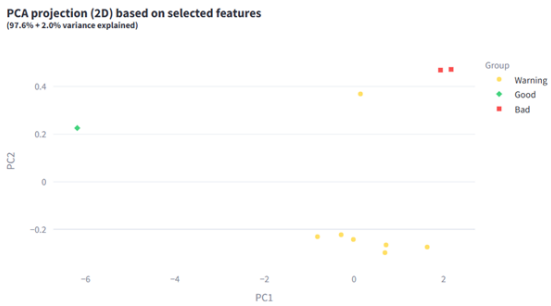


Figure 7: K-means graph for 8th segment

4.2.3 Segment 8 (Complex Curve–Acceleration Mix). This segment showed more divergence. The percentile method marked several drivers as 'WARN' due to short bursts of high throttle, while K-means placed some of these drivers in the 'Good' cluster. PCA visualization revealed that these drivers exhibited smoother braking and higher coasting ratios, which the clustering model positively weighted. This highlights a key difference: the statistical approach penalizes isolated peaks, whereas clustering balances them against compensatory behaviors.

4.2.4 Cross-approach Observations. The alignment was strongest in steady demand scenarios (Segment 1), weaker in mixed behavior contexts (Segment 8), and lowest in sustained load conditions (Segment 4). Statistical thresholding offers high interpretability and segment-level clarity, but may overlook multi-feature inefficiencies. K-means clustering captures complex, composite behavior and can sometimes reclassify drivers that the percentile method labels as efficient. It would be interesting for future work to implement more driver features and analyse in depth which have a different effect.

We additionally investigated the alignment between model-based feature importances and expert knowledge/domain expectations using SHAP.

Table 2: Top-5 features per class

Class	Top 1	Top 2	Top 3	Top 4	Top 5
Bad	AccelerationPedal	Speed	Acceleration	SteeringWheelAngle	BrakePedal
Medium	Speed	AccelerationPedal	Acceleration	SteeringWheelAngle	BrakePedal
Good	AccelerationPedal	Speed	Acceleration	SteeringWheelAngle	BrakePedal
Perfect	AccelerationPedal	Speed	Acceleration	SteeringWheelAngle	BrakePedal

Table 2 presents the five most influential features for each consumption class (*Bad*, *Medium*, *Good*, *Perfect*), ranked by their mean absolute SHAP value. The model consistently identifies *AccelerationPedal* and *BrakePedal* among the top-ranked features across multiple classes, in line with the statistical benchmark results from Section 4.1, where pedal usage was also the dominant indicator of inefficient driving events. This agreement confirms that the machine learning approach captures the same domain-relevant control inputs as the thresholds defined by the expert, while also highlighting secondary but relevant factors such as *Speed*, *Acceleration*, and *SteeringWheelAngle*.

4.3 Pareto Front of Time–Fuel Trade-Offs

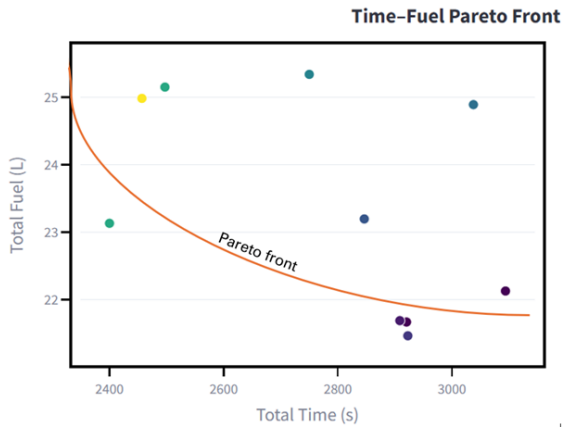


Figure 8: Pareto front

An interesting point of view would be to also consider the temporal information. Fuel consumption may reduce costs, but time is also quite important. Figure 8 plots the total time against the total fuel per driver. A driver is Pareto efficient if no other driver is faster and uses less fuel; these form the lower-left frontier. The points to the upper-right are dominated and can improve at least one objective without worsening the other. We obtain the frontier by non-dominated sorting of per-driver $(Time, Fuel)$ totals and colour points by their K-means group, explicitly linking global efficiency to the segment-level patterns identified earlier.

5 Discussion

This comparative study shows that rule-based thresholding remains highly interpretable and aligns with prior work, while K-means clustering reveals multi-feature patterns that affect efficiency. In practice, percentile rules flag isolated exceedances, whereas clustering captures cumulative demand and co-variation, explaining the discrepancies observed in segments such as Figure 6. Together, the methods are complementary: thresholding offers transparent guardrails; clustering provides a broader, context-aware view.

6 Conclusions

The results suggest that integrating both statistical and machine learning perspectives offers a more robust and nuanced driver assessment for fuel efficiency. While classical thresholding offers transparency, machine learning enables the discovery of complex patterns. Future work should further validate these findings to develop hybrid driver feedback systems. We only used SHAP diagnostically; a more systematic SHAP analysis would be interesting across models, segments, and time, to stabilize attributions and translate them into actionable feedback.

Acknowledgements

We thank the AAER Research d.o.o. team, led by CEO Matej Vengust, for access to simulator data and expert support. We also acknowledge support from the EDIH DIGI–SI project.

References

[1] Oscar Delgado, Felipe Rodríguez, and Rachel Muncrief. 2017. Fuel Efficiency Technology in European Heavy-Duty Vehicles: Baseline and Potential for the 2020–2030 Timeframe. White Paper. The International Council on Clean Transportation (ICCT), (July 2017). https://theicct.org/publication/fuel-efficiency-technology-in-european-heavy-duty-vehicles-baseline-and-potential-for-the-2020-2030-timeframe/.

[2] Hung Nguyen, George Tsaramirsis, Ilir Mborja, Dhimitraq Dervishi, Eriona Hoxha, Stavros Shiaeles, Anastasios Kavoukis, and Stamatios Vologiannidis. 2023. A data-driven framework for incentivising fuel efficient driving behaviour in heavy-duty vehicles. *J. Clean. Prod.*, 420, 139942. DOI: 10.1016/j.jclepro.2023.139942.

[3] Shuyan Chen, Hongru Liu, Yongfeng Ma, Fengxiang Qiao, Qianqian Pang, Ziyu Zhang, and Zhuopeng Xie. 2024. High fuel consumption driving behavior identification and causal analysis based on lightgbm and shap. *Res. Sq.* Preprint. DOI: 10.21203/rs.3.rs-4010652/v1.

[4] Alexander Meschtscherjakov, David Wilfinger, Thomas Scherndl, and Manfred Tscheligi. 2009. Acceptance of future persuasive in-car interfaces towards a more economic driving behaviour. In *AutomotiveUI 2009.* (Sept. 2009), 81–88. DOI: 10.1145/1620509.1620526.

[5] Charles Sullivan and Mark Franklin. 2010. An extended driving simulator used to motivate analysis of automobile fuel economy. In *Session 1: Tools, techniques, and best practies of engineering education for the digital generation.* (May 2010). DOI: 10.18260/1-2-1153-53783.

[6] Mathieu Maisonneuve. 2013. *Characterization of drivers' energetic efficiency: Identification and evaluation of driving parameters related to energy efficiency.* Master's thesis. Chalmers University of Technology. https://hdl.handle.net/20.500.12380/185531.

[7] Xiaohua Zhao, Yiping Wu, Jian Rong, and Yunlong Zhang. 2015. Development of a driving simulator based eco-driving support system. *Transportation Research Part C: Emerging Technologies*, 58, 631–641. Technologies to support green driving. DOI: https://doi.org/10.1016/j.trc.2015.03.030.

[8] Zhipeng Ma, Bo Nørregaard Jørgensen, and Zheng Ma. 2024. A scoping review of energy-efficient driving behaviors and applied state-of-the-art ai methods. *Energies*, 17, 2. DOI: 10.3390/en17020500.

[9] A McGordon, J E W Poxon, C Cheng, R P Jones, and P A Jennings. 2011. Development of a driver model to study the effects of real-world driver behaviour on the fuel consumption. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 225, 11, 1518–1530. DOI: 10.1177/0954407011409116.

[10] Thomas J. Daun, Daniel G. Braun, Christopher Frank, Stephan Haug, and Markus Lienkamp. 2013. Evaluation of driving behavior and the efficacy of a predictive eco-driving assistance system for heavy commercial vehicles in a driving simulator experiment. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, 2379–2386. DOI: 10.1109/ITSC.2013.6728583.