

Truck Fuel Consumption Prediction Using Logistic Regression and Artificial Neural Networks

Sheunesu Brandon Shamuyarira, National University of Science and Technology, Zimbabwe

Trust Tawanda, National University of Science and Technology, Zimbabwe*

 <https://orcid.org/0000-0003-2665-9711>

Elias Munapo, North West University, South Africa

ABSTRACT

Rising international oil costs and the transport industry's recovery from the effects of Covid-19 resulted in the efficient management of fuel by logistics companies becoming a significant concern. One way of managing this is by analyzing the fuel consumption of trucks so as to better utilize the costly resource. Twenty-three driving data variables were gathered from 210 freight trucks and analyzed this data. Relevant variables that impact truck fuel consumption were extracted from the initial 23 variables gathered using stepwise regression, and then a prediction model was built from the identified relevant variables utilizing a binary logistic regression model. In addition, a back propagation neural network was employed in this study to create a second model of truck fuel use, and comparisons between the two models were made. The outcomes showed that the binary logistic regression model and the back-propagated neural network model prediction accuracy were 68.4% and 77.2%, respectively.

KEYWORDS

Artificial Neural Networks, Fleet Management System, Logistic Regression, Truck Fuel Consumption Prediction

1. INTRODUCTION

The invasion of Ukraine by Russia has had a negative ripple effect in the global oil market. Russia has been one of the world's largest oil producing country and now the war combined with economic sanctions on Russia, has had huge repercussions on the global economy as Ukraine and Russia are major players in the food, energy and mining sectors. Since Zimbabwe does not mine petroleum, it is a net importer and as such a price taker of the foregoing global oil prices, thus the efficient management of fuel by logistics companies has become a significant concern. One way of managing this is by analyzing the fuel consumption of trucks so as to better utilize the costly resource. The freight

DOI: 10.4018/IJORIS.329240

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

industry in Southern Africa is largely dominated by road transport and there are many cross-border transporters with both domestic and international origins. Transportation networks have become the major lifeline of modern societies, not only by ensuring individual well-being, but by also fostering economic growth through fast and reliable transportation activities, Muriel-Villegas *et al.* (2019) Since freight trucks carry the majority of commodities by road, they often consume much more fuel than other kinds of vehicles because of their characteristics such as large loads and long distance travel. Zimbabwe does not mine petroleum, thus it is a net importer of fuel and as such a price taker of the ongoing global oil prices. Since Russia invaded Ukraine there has been a turbulence in the global petroleum market which has been causing global spikes in crude oil prices. To add on this, Zimbabwe has had the highest regional fuel prices thus there is need for management to address the increasing cost of operation rising from both high fuel consumption and high fuel cost. Moreover, to reduce emissions thus lowering the carbon footprint, as the sustainable growth of the environment and energy has become a requirement all over the world, particularly in the transportation sector.

2. LITERATURE REVIEW

The fuel consumption of freight trucks is influenced by a number of different factors. Fuel consumption can differ significantly from truck to truck, even when comparing two that are of the same make, model, year, and fuel type. There are a number of factors that may contribute to this. Some of them fall under many categories, including those relating to drivers, weather, traffic, travel, and vehicles, among others. Barbado and Corcho (2021) evaluated the literature on a number of factors that might affect a vehicle's fuel consumption. Considering the effects related to travel factors, the authors identified eco-routing as a crucial element in reducing fuel consumption. By considering the optimum route, one could save fuel not only in terms of distance and travel time, but also in comparison to other viable routes. According to Faria *et al.* (2019), with regards to driver related factors, aggressive driving compared to calmer driving can account for up to 40% of a vehicle's fuel consumption. A mesoscopic fuel consumption estimation model was created by, Chen *et al.* (2017), that took into account factors like the number of lanes and free-flow speed that had previously received little attention. The study's findings demonstrated that these elements had an effect on motor vehicle fuel consumption as well. Freight transport companies now employ telematics services to track their trucks and keep an eye on variables like fuel usage in an effort to minimize the costs connected with fuel consumption. The data gathered by this state-of-the-art digital monitoring and collection technology is more diverse and accurate than manual recording. It is crucial for fleet managers to process and make use of the data generated by the fleet intelligent management system. By carefully examining the data and learning the rule of fuel consumption during truck operation, it is possible to lower the fuel consumption of trucks to a certain extent. Malekian *et al.* (2016) developed a wireless on-board diagnostic system (OBD II) fleet management system. The system was aimed at measuring speed, distance, and fuel consumption of vehicles for tracking and analysis purposes. The findings demonstrated that the system could successfully read a variety of parameters and analyze, transmit, and display readings. In order to anticipate the fuel consumption of diesel buses, Sun *et al.* (2021) computed the relevant fuel consumption model utilizing real-time driving and fuel consumption data collected by on-board sensors. Based on the mobile phone terminals and on-board diagnostic system (OBD) installed in taxis, Yao *et al.* (2020) extracted driving behavior and fuel consumption data and were able to predict vehicle fuel consumption based on mobile phone data. Many scholars have volunteered their time to estimate and calculate the fuel consumption of trucks in an effort to identify the best driving behaviors that can reduce both fuel consumption and greenhouse gas emissions. These can be divided into two categories namely historical models and modern data driven models. Examples of historical models include Vehicle Specific Power (VSP) Model, Comprehensive Modal Emission Model (CMEM), and Emissions from Traffic (EMIT) Model. Zhang *et al.* (2023) by using vehicle acceleration and jerk as the defining parameters, created a unique computational model for the volatile condition (defined by

acceleration and jerk) and classified it into eight varieties. They discovered that the percentage and level of contribution of each form of jerk to fuel usage varied. Additionally, they compared their model to the VSP and VT-Micro models and discovered that their model offered more precise estimates for new routes. An EMIT model was proposed by Cappiello *et al.* (2002) to forecast the fuel consumption and emissions for light-duty vehicles. The performance of this model was comparable to CMEM. Examples of the modern data driven models include the decision tree models, random forest models, artificial neural network models and others. Ma *et al.* (2014), evaluated the impact of driving behavior on the fuel of urban buses during acceleration using the C4.5 decision tree. With fewer training data and a good capacity for generalization, this model had a prediction accuracy of more than 85%. In order to estimate car fuel consumption based on environmental, driver, and vehicle views, Li *et al.* (2022) suggested a multi-view deep neural network (MVDNN). To create several perspectives of the driving scenario, they employed many sorts of data, including weather, road conditions, traffic flow, driver behavior, vehicle speed, acceleration, etc. They then integrated these observations using a fusion layer and sent them into a deep neural network to forecast the fuel use. They discovered that MVDNN performed better than other approaches like linear regression and the random forest model. The best approach for predicting truck fuel consumption relies greatly on a variety of factors.

There is no one strategy that is always the best. What is best depends on the specifics of the problem, the data structure, the characteristics employed, the degree to which those features can be used to distinguish the classes, as well as the classification's goal. Logistic regression and neural networks are two common methods for predicting vehicle fuel consumption based on driving cycle data. An empirical comparison of logistic regression and neural networks in predicting vehicle fuel consumption can reveal their strengths and weaknesses in terms of accuracy, robustness, interpretability, and computational efficiency. However, such a comparison is not very common in the literature. Most studies focus on either logistic regression or neural networks, but not both. For example, Capraz *et al.* (2016), used logistic regression to classify driving cycles into high or low fuel consumption categories based on speed and acceleration features. Topic *et al.* (2022), used neural networks to predict vehicle fuel consumption based on speed, acceleration, and road slope inputs. Both studies reported good performance of their models, but they did not compare them with other methods. Therefore, a possible research gap is to conduct a comprehensive and systematic comparison of logistic regression and neural networks in predicting vehicle fuel consumption using real-time driving cycle data. Such a comparison can provide insights into the advantages and disadvantages of each method, as well as the factors that affect their performance, such as data quality, feature selection, model complexity, parameter tuning, validation method, etc. A comparison can also help management to identify the best method for different applications and scenarios, such as eco-driving, eco-routing, transport planning, etc.

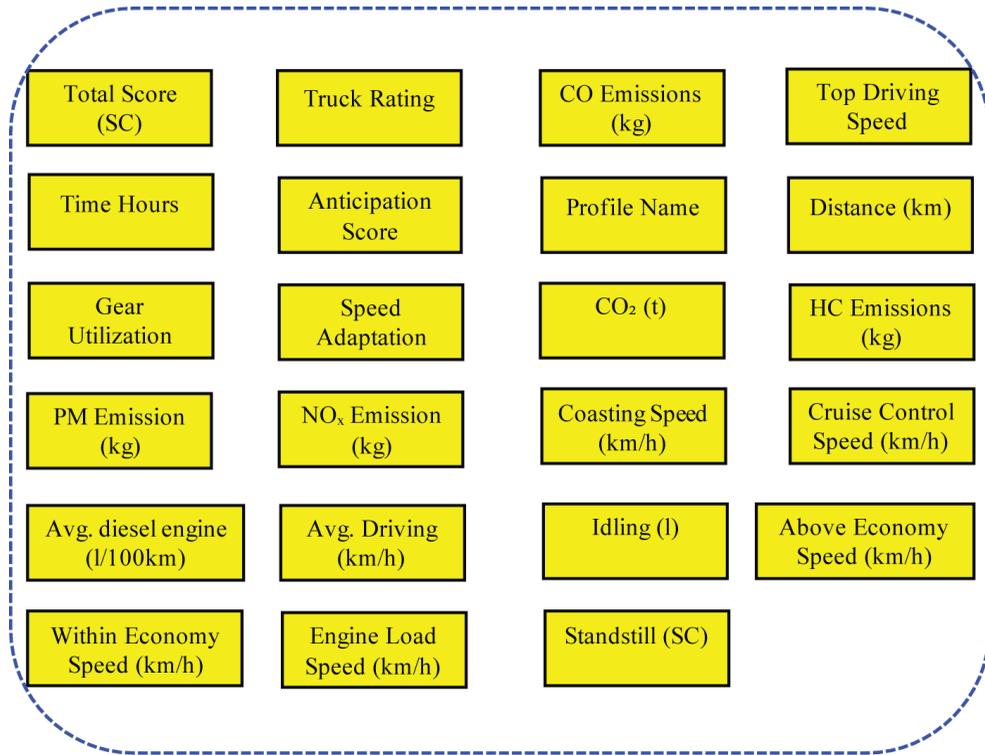
3. DATA AND METHODS

This study examined the fuel consumption characteristics of freight trucks from a medium sized freight company in Zimbabwe. The company's fleet management system was used to collect information for 210 trucks between January 1 and January 31, 2023. This web-based system, interfaces to truck-mounted mobile tracking technology. It carries out positioning monitoring in real time and provides back statistics. The hardware also keeps track of the truck's cargo at each stage of the route, the number of stops it makes, the amount of fuel utilized, etc. Figure 1 shows the set of independent variables collected from the fleet management system.

3.1 Data Summary Statistics

The variables shown in Table 1 were used throughout the analysis and construction of the models with fuel consumption being the dependent variable which is a dichotomous variable and 23 independent variables as indicated in the summary of naturalistic data.

Figure 1. Independent variables collected



3.2 Binary Logistic Regression

According to Maalouf (2011), statisticians and researchers use logistic regression as it is one of the most significant statistical and data mining approaches for the analysis and classification of binary and proportional response data sets. It belongs to the class of generalized linear models (GLMs), which are commonly known as linear regression models for continuous response variables given continuous or categorical predictor variables. This modeling and analysis technique is based on fitting nonlinear relationships and is used to study the interactions and dependencies between independent and dependent variables. Multinomial logistic regression and binary logistic regression are the two models that make up logistic regression. Binary logistic regression was selected because the outcome is a dichotomous variable, or one that can only take the values 0 or 1. Therefore, for the purpose of this research study, 0 represented normal fuel consumption and 1 represented high fuel consumption. This method was also adopted in this research study because it is much easier to set up and train unlike other machine learning models and because it does not assume how classes are distributed in feature space. However, multinomial logistic regression allows the dependent variable to be categorized into multiple groups, and this is typically utilized for cases or issues that are more complicated. It is important to remember that dichotomous variables can also be created from multi-categorical ones. Below is how to derive the probability, P_i , of fuel consumption beginning from the log of odds. The model of predicted probabilities is as follows:

$$\ln \left(\frac{P_i}{1 - P_i} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (1)$$

Table 1. The summary of naturalistic data collected

Variable Name	Definition
Total Score (SC)	Represents each truck's total performance score
Truck Rating	Represents truck performance rating derived over time
CO Emissions (kg)	Represents each truck's carbon dioxide emissions
Top Driving Speed	Represents the average driving speed in open road
Time (Hrs)	Represents the time the truck was on the road
Anticipation Score	Represents the truck's anticipation and braking score
Profile Name	Represents the type of the truck (Long Haul or Construction Rough)
Distance	Represents the total distance traveled by the truck
Gear Utilization Score	Represents the truck's gear utilization score
Speed Adaptation	Represents the truck's speed adaptation score
CO ₂ (t)	Represents the truck's carbon monoxide emission
HC Emissions (kg)	Represents the truck's hydrocarbon emissions
PM Emission (kg)	Represents the truck's particulate matter emissions
NO _x Emission (kg)	Represents the truck's nitrogen emission
Coasting Speed (km/h)	Represents the truck's coasting speed
Cruise Control Speed (km/h)	Represents the truck's cruise control speed
Avg diesel engine (l/100km)	Represents the truck's average diesel engine on (l/100km)
Avg Driving (km/h)	Represents the truck's overall average driving speed
Idling (l)	Represents the truck's idling
Above Economy Speed (km/h)	Represents the truck's above economy fuel consumption speed
Within Economy Speed (km/h)	Represents the truck's within economy fuel consumption speed
Engine Load Speed (km/h)	Represents the truck's engine load speed
Standstill (SC)	Represents the truck's standstill score
Fuel consumption	Fuel consumption per hundred kilometers for each trip

which reduces to:

$$P_i = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n}} \quad (2)$$

where among them $x_1, x_2, x_3, \dots, x_n$ are explanatory variables, while $\beta_1, \beta_2, \beta_3, \dots, \beta_n$ represent the regression coefficients and β_0 being the intercept.

3.3 Neural Network

Artificial neural networks (ANNs), also known as “neural networks” or “neural nets” (NNs), are computer models that replicate how nerve cells in the human brain function. By allowing a computer to learn from observational data, recognize patterns, and make predictions, these biologically inspired programming methods can help make sense of confusing, contradicting, or fuzzy logic. There are three

or more interconnected layers in a neural network. Input neurons make up the first layer. The input neurons send information to the second layer, also known as the hidden layer(s), which then sends information to the third layer, also known as the output layer, as the final output data. Because they can adapt to internal and external data, neural networks are flexible systems. The neural network life cycle comprises two stages: the training phase and the prediction phase. The procedure of determining the weight and bias values takes place throughout the training stage. In the prediction phase, the neural network model processes the input data and generates predictions. A back-propagated feed-forward neural net was employed for the analysis in this study. Data in a feed-forward neural network only travels in one direction, forward, from the input nodes through the hidden nodes to the output node(s). The neural networks can modify their output results by accounting for errors through a technique called back-propagation, Ma *et al.* (2018). By doing this, information is transmitted backward each time a training-phase error is made. Then, each weight is adjusted proportionally to the degree to which it contributed to the inaccuracy. In order to account for the discrepancy between the desired and actual outcomes, the error is utilized to recalibrate the weight of the unit connections of the neural network, Yao *et al.* (2020). The neural network model was adopted in this research study because it has the ability to detect complicated nonlinear correlations between dependent and independent variables and also because it requires less formal statistical training. Figure 2 shows an illustration of a neural net.

4. DATA ANALYSIS

4.1 Binary Logistic Regression Model

The first step before conducting any analysis was to check for sampling bias or data imbalance, because to build a strong model, the proportion of high and normal fuel consumption trucks needed to be statistically balanced, and for this data set, 51.4% of the trucks were high fuel consumers, while the remaining 48.6% were normal consumers. The data was then randomly separated into two sets, the training set and the testing set, using **R** software package instructions with a split ratio of 0.8. The model was created using the data from the training set, and it was validated using the data from the testing set. The observations' composition after the random split is shown in Table 2.

4.1.1 Stepwise Regression

The backward elimination approach was utilized to identify the variables important to the investigation. Backward elimination is a step-wise regression technique that starts with a full model that includes all predictor variables and gradually removes variables to obtain a smaller model that best fits the data. The variable that considerably improved the AIC (Akaike Information Criterion) value was

Figure 2. Neural net illustration

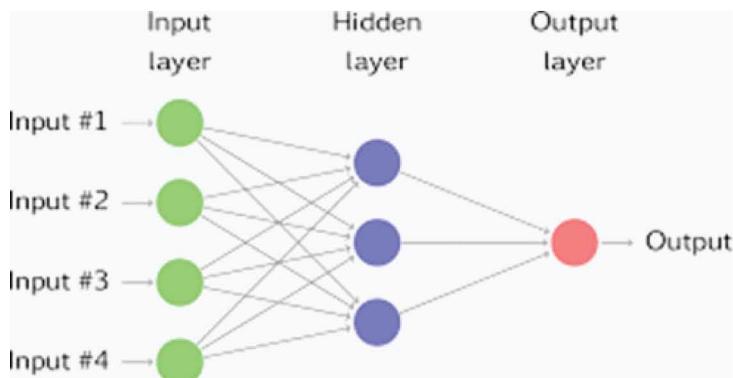


Table 2. Data grouping

	High Fuel Consumption	Normal Fuel Consumption	% of High Fuel Consumption	Total
Training	90	82	52.3%	172
Testing	18	20	47.4%	38
Total	108	102	51.4%	210

then removed before moving on to the next step. The model initially included all 23 explanatory variables. The step-wise method was chosen because it decreased the number of predictors, solved the over fitting issue, and minimized the multi-collinearity issue. The AIC value was chosen as the criterion for variable removal because the analysis was done in **R** software package. Since the objective was to create a model with the lowest AIC value possible, the variable that significantly reduced the AIC value was then eliminated. On the first step 1, the full model including all the theoretical explanatory variables had an AIC value of **257.84**. Tables 3 and 4 show all the 19 steps of the backward elimination procedure.

Truck Rating, Carbon Emission, Top Driving Speed, Time (Hrs), and Idling were found to be the most important variables after a successful elimination operation, since the removal of any of these led to a weaker model with higher AIC values. Out of a total of 23 predictors, only 5 were found to be significant to the model.

4.1.2 Model Estimation

The fitted logistic regression model is summarized in Table 5 below, and the model’s coefficients were estimated using the **R** statistical software. The table has seven columns, and the co-efficient column lists the estimated value of each independent variable’s coefficient. The log chances of the outcome for a one unit increase in the predictor variable are given by the logistic regression coefficients. The z statistic (also known as the Wald z statistic) and the standard errors of the coefficients are also included in the table along with their p -values. The odds ratio is shown in the second from last column and is calculated using $exp(\beta_i)$, where β_i is the estimated regression co-efficient of the variable χ_i .

The intercept is the average log of odds of a truck having a high consumption when all the explanatory factors are taken into account. Without taking into account any of the truck’s features, travel information, or emissions statistics, the odds ratio of 0.0000000327 approximately 0 from Table 5 represents the likelihood that a truck will consume more fuel than budgeted. Any truck has a high likelihood of consuming normal fuel when no other features are known because the coefficient’s value is negative and the odds are close to zero.

- **Truck Rating:** At the 1% level, it was determined that the variable was significant ($p = 0.00203$). The coefficient’s negative value, as shown in Table 5, indicates that there is an inverse correlation between truck rating and likelihood of being a high fuel-consumption truck. Since truck rating has an odd ratio of 0.5405 when all other factors are assumed to be constant, the chance of utilizing more fuel than necessary decreases by a factor of 0.5405 for every unit increase in truck rating. To put it another way, highly rated trucks are less likely to be high fuel-consumption vehicles. This is because a higher truck rating indicates that the vehicle is in better condition, is well-maintained, and is operating efficiently, which lowers the likelihood that it would use more fuel on any journey.
- **Carbon Emissions:** The observed carbon emissions were determined to be statistically significant at the 1% level ($p = 0.00167$). The estimate is positive, which suggests that the likelihood that the truck will be a high fuel-consumption vehicle will grow as carbon emissions rise. While the estimates exponential was calculated, an odd ratio of 1.2135 was found. This means that, while all other factors are held constant, the likelihood of consuming more fuel increases by a factor

Table 3. Steps 1 to 9 of backward elimination

Predictors	Step 1 AIC=257.84	Step 2 AIC=255.84	Step 3 AIC=253.84	Step 4 AIC=251.85	Step 5 AIC=249.87	Step 6 AIC=247.91	Step 7 AIC=246.06	Step 8 AIC=244.29	Step 9 AIC=242.71
Total Score	257.63	255.66	253.69	251.70	249.72	247.73	245.93	244.15	242.32
Truck Rating	256.60	254.60	252.60	250.61	248.64	247.07	245.18	243.32	242.09
Co_Emission	257.41	255.41	253.42	251.42	249.46	248.21	246.31	244.42	243.08
TopDrivingSpeed	257.34	255.34	253.35	251.35	249.37	248.03	246.13	244.22	243.01
Time (Hrs)	256.97	254.97	252.97	250.95	249.00	247.44	245.54	243.67	242.47
Anticipation Score	256.51	254.51	252.51	250.53	248.56	246.66	244.85	243.12	241.47
Profile Name	255.88	253.88	251.88	249.88	247.91				
Distance	256.16	254.31	252.31	250.32	248.34	246.43	244.71	243.02	241.36
Gear Utilisation	255.84								
SpeedAdaptation	256.14	254.21	252.21	250.23	248.23	246.29	244.59	242.92	241.24
CO ₂ (t)	257.38	255.38	253.38	251.38	249.44	247.48	245.61	244.09	242.47
HC_Emission	255.84	253.84							
PM_Emission	256.09	254.09	252.10	250.11	248.13	246.17	244.29		
NO _x _Emission	257.88	255.88	253.88	251.90	249.92	247.96	246.04	244.23	242.27
Coasting	257.08	255.09	254.29	252.29	250.37	248.41	246.50	244.53	242.88
Cruise Control	256.14	254.15	252.55	250.17	248.25	246.29	244.53	242.71	
AvgDieselEngine	256.64	254.64	252.65	250.65	248.66	246.80	244.94	243.10	241.55
AvgDriving	255.84	253.85	251.85						
Idling	258.43	256.51	254.51	252.51	250.92	249.42	247.48	245.78	243.87
Above_Eco_Speed	256.01	254.01	252.01	250.01	248.05	246.06			
Within_Eco_Speed	255.87	253.87	251.87	249.87					
EngineLoadSpeed	256.67	254.71	252.72	250.73	248.77	246.77	244.90	243.04	241.59
StandStill	256.48	255.10	253.10	251.18	249.19	247.22	245.27	243.46	241.78
< none >	257.84	255.84	253.84	251.85	249.87	247.91	246.06	244.29	242.71

Table 4. Steps 10 to 19 of backward elimination

Predictors	Step 10 AIC=241.24	Step 11 AIC=239.38	Step 12 AIC=238.01	Step 13 AIC=236.59	Step 14 AIC=235.35	Step 15 AIC=234.82	Step 16 AIC=234.08	Step 17 AIC=233.7	Step 18 AIC=233.33	Step 19 AIC=231.35
Total Score	240.94	239.04	237.83	236.28	235.04	234.08				
Truck Rating	240.88	238.93	241.67	239.70	238.85	237.99	241.38	241.13	240.28	242.32
Co_Emission	241.98	240.01	244.65	242.84	242.18	240.99	242.90	242.57	241.52	243.88
TopDrivingSpeed	242.03	240.03	241.44	239.47	238.03	236.97	236.97	236.77	236.20	234.54
Time (Hrs)	241.38	239.41	243.21	241.28	240.68	239.46	239.86	239.60	238.83	239.87
Anticipation Score	239.71	238.01								
Profile Name										
Distance	239.38									
Gear Utilisation										
SpeedAdaptation										
CO ₂ (t)	241.02	239.08	237.88	236.16	234.90	234.39	233.70			
HC_Emission										
PM_Emission										
NO _x _Emission	241.02	239.12	237.70	236.35	235.40	234.72	233.87	233.34	231.35	
Coasting	241.40	239.45	238.25	236.51	235.28	234.75	234.05	233.33		
Cruise Control										
AvgDieselEngine	240.12	238.39	236.92	235.35						
AvgDriving										
Idling	242.11	240.74	239.59	238.39	236.45	239.44	239.62	239.50	239.29	237.97
Above_Eco_Speed										
Within_Eco_Speed										
EngineLoadSpeed	240.69	238.39	237.47	236.13	234.82					
StandStill	240.20	238.39	236.59							
< none >	241.24	239.38	238.01	236.59	235.35	234.82	234.08	233.70	233.33	231.35

Table 5. Parameter estimates for the fitted logistic regression model

	Co-efficient	Standard Error	Z-Value	P-Value	Odds Ratio	Sig
Intercept	-17.23456	8.11601	-2.124	0.03371	0.000 000 0327	*
Truck Rating	-0.61525	0.19942	-3.085	0.00203	0.5405	**
C0_Emissions	0.19348	0.06155	3.144	0.00167	1.2135	**
TopDrivingSpeed	0.18401	0.08757	2.101	0.03562	1.2020	*
Time (Hrs)	0.19841	0.07202	2.755	0.00587	1.2195	**
Idling	0.28536	0.10067	2.835	0.00459	1.3302	**

Sig Codes: **** p < 0.001, *** p < 0.01, * p < 0.05
 Null deviance: 238.07 on 171 degrees of freedom
 Residual deviance: 219.35 on 166 degrees of freedom
 AIC: 231.35
 Number of Fisher Scoring iterations: 4

of 1.2135 for a unit increase in the average carbon emission output. In other words, a truck is 1.2 times more likely to use more fuel for every additional unit of carbon it emits.

- **Top Driving Speed:** With a p-value of 0.003562, it was determined that this variable was statistically significant at the 5% level. The estimate's positive value suggests that there is a relationship between the top driving speed and the log of the probability of using more fuel. Therefore, the chance of consuming more fuel is 1.2 times higher with every rise in the average top driving speed. This is due to the fact that a faster truck would burn more fuel, and according to this model, a unit increase in average speed results in a 1.202 increase in fuel consumption for the truck, all other factors remaining constant.
- **Time (Hrs):** With a p-value of 0.00587, this variable was determined to be significant at the 1% level. The estimate is positive, indicating that the likelihood of the truck consuming more fuel will rise as the number of hours it spends on the road increases. In this case, an additional hour spent driving will increase the likelihood of using more fuel by a factor of 1.2195. This is mostly caused by the fact that longer travel times result in longer travel distances, as well as potential delays and traffic, which cause the truck to use more fuel than necessary without traveling an equivalent distance.
- **Idling:** The predicting factor with a p-value of 0.00459 was shown to be statistically significant at 1%. The co-efficient indicates that there is a positive correlation between the likelihood of consuming more and an increase in truck idling. Therefore, for trucks that idle the most, there is a ratio of 1.3302 higher chance of consuming more fuel than usual. In other words, if all other factors remain unchanged, a unit increase in idling increases the risk of using fuel by 1.33 times.

4.1.3 Model Evaluation: Likelihood Ratio Test

In general, three tests can be utilized: the likelihood ratio test, score test, and Wald test. According to Menard (1995), the likelihood ratio test was employed, and the results are displayed in Figure 3 below. Due to the statistical significance of the difference ($p < 0.05$), it can be concluded that the estimated model offers a better fit for the data than the null model since it has more parameters and can predict the output more accurately than the null model.

4.1.4 Goodness of Fit Test: Hosmer-Lemeshow Test

This test employs the Pearson chi-square method to determine whether the observed proportion of events is similar to the predicted probabilities of occurrence in subgroups of the dataset. With this strategy, the assumption that the model adequately describes the data is tested. Small chi-square values with high p-values lead to the rejection of the null hypothesis, which leads to the conclusion

Figure 3. Likelihood ratio test results

```

Likelihood ratio test

Model 1: FuelConsumption ~ TruckRating + CarbonEmmision_Grams + TopDrivingSpeed +
  Time_Hrs + Idling
Model 2: FuelConsumption ~ 1
#Df LogLik Df Chisq Pr(>Chisq)
1 6 -109.67
2 1 -119.03 -5 18.725 0.002163 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    
```

that there is insufficient evidence to demonstrate the model’s poor fit, hence the fitted model offers a good fit. On the other hand, large values with p-values below 0.05 suggest a poor fit.

Since the null hypothesis was rejected by the chi-square value of 0.94876 from Figure 4 above and the corresponding p-value more than 0.05 (p = 0.8136), we deduced that the developed model offers a strong fit to the data.

4.2 Design of the Neural Network Model

The statistical program **R** was used to train the network in this section. The neural network was fitted using the same variables that were used to estimate the logistic regression model for the sake of comparison. As a result, it follows that the classification problem’s input layer, which is made up of the independent variables, would likewise contain five nodes (input neurons), which would be the same number of variables as those employed in the modified logistic regression model.

It is important to choose the number of hidden layers as well as the hidden layer(s) neuronal population. Even though they are not directly connected to the outside world, the hidden layers have a significant impact on the outcome, thus they must be carefully examined. In order to create a better network architecture, trial and error was done with neurons ranging from 1 to 5. The best architecture for the model was chosen based on the mean square error (MSE) and misclassification rate. The below formula was used to determine the misclassification rate:

$$Rate = \left(1 - \frac{Correctly\ Predicted}{All\ Cases} \right) \tag{3}$$

The mean square error formula used was:

$$MSE = \frac{1}{N} \sum_{j=1}^N (e_j)^2 = \frac{1}{N} \sum_{j=1}^N (t_j - Y_j)^2 \tag{4}$$

Figure 4. Hosmer-Lemeshow test results

```

Hosmer and Lemeshow goodness of fit (GOF) test

data: train$FuelConsumption, fitted(reduced)
X-squared = 0.94876, df = 3, p-value = 0.8136
    
```

where t_j and Y_j are the target value and the network output for the testing case respectively and N was the number of testing cases.

The trial and error criteria were used to select the network that offered the best model fit. Table 6 shows that a single hidden layer neuron with three hidden neurons had the lowest rate of misclassification and the second-best mean square error, making it the ideal design for fitting the neural network based on the data used in this study.

4.2.1 Model Summary

The variables and model details are summarized in Table 7 above. Five independent variables and three hidden neurons were combined into one hidden layer. In order to generate an output with a range of 0 to 1, the sigmoid activation function, which gives the model non-linearity, was used. Cross entropy was also employed as the error function because it is generally preferred to sum of squared errors in classification issues.

The variables and model details are summarized in Table 7 above. Five independent variables and three hidden neurons were combined into one hidden layer. In order to generate an output with a range of 0 to 1, the sigmoid activation function, which gives the model non-linearity, was used. Cross entropy was also employed as the error function because it is generally preferred to sum of squared errors in classification issues.

An overview of the neural network model, including run time, error, and the total number of training steps, is given in Table 8.

The model was fitted in 8702 steps, as shown in Table 8 above, and the cross entropy error was determined to be 16,597. A mean square error of 0.2668 was discovered in the identical situation,

Table 6. Trial and error results

Number of Hidden Neurons	Misclassification Rate	MSE
1	0.421	0.262
2	0.5	0.341
3	0.395	0.267
4	0.421	0.359
5	0.526	0.341

Table 7. Network information

Input Layer	Covariates	Truck Rating, CO Emissions, Top Driving Speed, Time (Hrs), Idling
	Number of Neurons	5
	Rescaling Methods for Covariates	Min-Max Normalization
Hidden Layer	Number of Hidden Layers	1
	Number of Hidden Neurons	3
	Activation Function	Sigmoid
Output Layer	Dependent Variable	Fuel Consumption
	Number of Neurons	1
	Activation Function	Sigmoid
	Error Function	Cross Entropy

indicating relatively minor discrepancies between the anticipated output and the actual result based on the training data set. The software package’s automation and the training method (resilient back-propagation) made fitting the model simple and straightforward. Following model construction, predictions based on training data produced a misclassification rate of 39.5%.

The neural network built with **R** is shown in Figure 5, along with the weights given to each synapses that link any two neurons.

4.3 Empirical Comparisons

The end objective of the study was to empirically compare the two fuel consumption forecasting models. The two models were compared using the receiver operator characteristic (ROC) curve and its related area under the curve (AUC). Figure 6 displays the ROC curves for the two models created using binary logistic regression and an artificial neural network. An area under the curve of 0.8 denotes an excellent prediction by the model, 0.9 denotes an extraordinary prediction, and 0.7 denotes a subpar prediction. Since the ROC curve for the neural network had a larger area under the curve (AUC) value of 0.772 (approximately 0.8) than the logistic regression model’s AUC value of 0.684 (approximately 0.7), it is clear from Figure 6 that the neural network outperformed the latter. As a result, the neural network model accurately forecasted the data while the logistic regression model did so just fair.

Table 8. ANN model summary

Error	16.597
Steps	8702
Training time	0.05 seconds
Mean Square Error (MSE)	0.2668
% of incorrect predictions	39.5%

Figure 5. Neural network model for fuel consumption

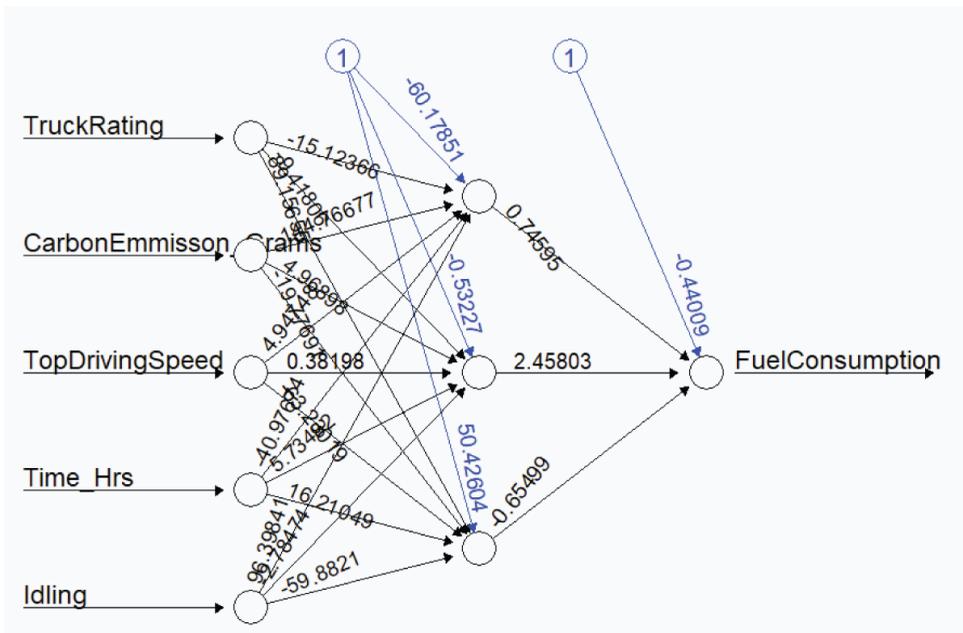
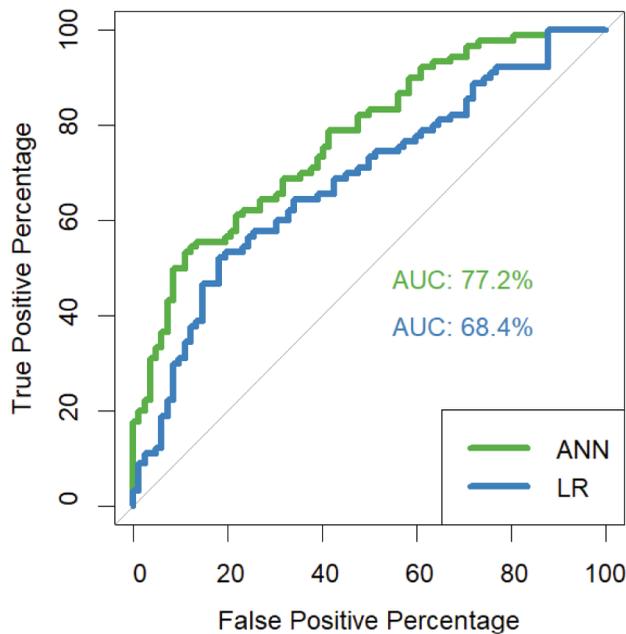


Figure 6. ROC curves and AUC values



5. CONCLUSION

The purpose of this research study was to address the issue of increased cost of operation due to high fuel consumption and high fuel cost by developing fuel consumption prediction models. By using the driving data acquired by the fleet management system, the study showed how to analyze the fuel usage of freight trucks. In this study, machine learning and conventional statistical methods were merged. A similar study on predicting the fuel consumption of heavy duty diesel trucks was conducted by Gong et al. (2021) By employing logistic regression in a manner akin to that used in this work, they were able to isolate the important variables influencing vehicle fuel usage and create models for predicting it. In this study, it was found that truck rating, carbon emissions, top driving speed, time, and idling were the most important variables influencing truck fuel use. As a result, the findings of this study imply that neural networks offer a more exact and accurate method of predicting truck fuel use than logistic regression. This research raised knowledge about how to manage pollution from trucks and lower exhaust gas emissions, which could also help to achieve the goal of decreasing cost and enhancing efficiency of truck driving. Neural networks are a viable option for businesses interested in forecasting truck fuel use as identified in this research study. Below are some of the recommendations to management on why they should adopt the use of neural networks in forecasting fuel use:

- Neural networks can be used to predict fuel consumption because of their high degree of accuracy. As a result, they can be used to improve routing, scheduling and driver behavior, all of which can lead to reduced fuel costs.
- Telematics systems together with the information obtained from the neural network model, can be used to monitor driver behavior, such as speed, acceleration and braking. This information can be used to identify drivers who drive inefficiently and then provide them with training on how to improve their driving habits.

- A fuel management program can be implemented that could help organizations track their fuel usage, identify areas where they are wasting fuel as companies and implement changes to reduce their fuel costs.
- Organizations might look into the prospect of using alternative fuels. If this is viable, alternative fuels such as natural gas, electricity or propane can cost less than traditional fuels such as diesel. Switching to alternative fuels can help to reduce fuel costs.
- Drivers can have a big impact on fuel consumption, thus educating them about fuel efficiency is important. By educating them about fuel efficient driving practices, the organization can help them to reduce their fuel usage.

It is possible that other characteristics are also utilized to forecast truck fuel usage in addition to those included in this study. As a result of data mining's cutting-edge nature and methodological intricacy, there is still considerable space for improvement in this study, as several elements including the impact of the weather and traffic on fuel use were overlooked. Additionally, there is room for improvement through the application of various extra machine learning models. Future studies will expand the type and volume of data collected under the presumption that the data are reasonable and accurate. Trying different data mining techniques will increase the model's predictive accuracy. Future changes to the sample size, the prediction model, and the relevant factors extracted will allow for the weighting of various indicators.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The authors certify that this research does not involve animal or human participants.

CONSENT FOR PUBLICATION

N/A

AVAILABILITY OF DATA

Data supporting can be provided upon request.

COMPETING INTERESTS

N/A

FUNDING

Publisher has waived the Open Access publishing fee.

AUTHORS' CONTRIBUTIONS

All authors contributed equally.

ACKNOWLEDGMENT

We would like to take this opportunity to acknowledge the time and effort devoted by the editors and reviewers to improving the quality of this manuscript.

REFERENCES

- Barbado, A., & Corcho, Ó. (2021). *Vehicle fuel optimization under real-world driving conditions: An explainable artificial intelligence approach*. arXiv preprint arXiv:2107.06031. <https://doi.org/10.48550/arXiv.2107.06031>
- Cappiello, A., Chabini, I., Nam, E. K., Lue, A., & Abou Zeid, M. (2002, September). A statistical model of vehicle emissions and fuel consumption. In *Proceedings. The IEEE 5th international conference on intelligent transportation systems* (pp. 801-809). IEEE. doi:10.1109/ITSC.2002.1041322
- Çapraz, A. G., Özel, P., Şevkli, M., & Beyca, Ö. F. (2016). Fuel consumption models applied to automobiles using real-time data: A comparison of statistical models. *Procedia Computer Science*, 83, 774–781. doi:10.1016/j.procs.2016.04.166
- Chen, Y., Zhu, L., Gonder, J., Young, S., & Walkowicz, K. (2017). Data-driven fuel consumption estimation: A multivariate adaptive regression spline approach. *Transportation Research Part C, Emerging Technologies*, 83, 134–145. doi:10.1016/j.trc.2017.08.003
- Faria, M. V., Duarte, G. O., Varella, R. A., Farias, T. L., & Baptista, P. C. (2019). How do road grade, road type and driving aggressiveness impact vehicle fuel consumption? Assessing potential fuel savings in Lisbon, Portugal. *Transportation Research Part D, Transport and Environment*, 72, 148–161. doi:10.1016/j.trd.2019.04.016
- Gong, J., Shang, J., Li, L., Zhang, C., He, J., & Ma, J. (2021). A comparative study on fuel consumption prediction methods of heavy-duty diesel trucks considering 21 influencing factors. *Energies*, 14(23), 8106. doi:10.3390/en14238106
- Li, Y., Zeng, I. Y., Niu, Z., Shi, J., Wang, Z., & Guan, Z. (2022). Predicting vehicle fuel consumption based on multi-view deep neural network. *Neurocomputing*, 502, 140–147. doi:10.1016/j.neucom.2022.06.047
- Ma, H., Xie, H., Chen, S., Yan, Y., & Huang, D. (2014). *Effects of driver acceleration behavior on fuel consumption of city buses (No. 2014-01-0389)*. SAE Technical paper. 10.4271/2014-01-0389
- Ma, J., Li, D., Lu, Y., & Chen, J. (2018, December). Intelligent Diagnosis System for Vehicle Network Based on BP Neural Network. In *IOP Conference Series: Materials Science and Engineering* (Vol. 452, No. 4, p. 042004). IOP Publishing. doi:10.1088/1757-899X/452/4/042004
- Maalouf, M. (2011). Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies*, 3(3), 281–299. doi:10.1504/IJDATS.2011.041335
- Malekian, R., Moloisane, N. R., Nair, L., Maharaj, B. T., & Chude-Onkonkwo, U. A. (2016). Design and implementation of a wireless OBD II fleet management system. *IEEE Sensors Journal*, 17(4), 1154–1164. doi:10.1109/JSEN.2016.2631542
- Menard, S. (1995). *Applied logistic regression analysis (Sage university paper series on quantitative application in the social sciences, series no. 106)* (2nd ed.). Sage.
- Muriel-Villegas, J. E., & Correa-Espinal, A. A. (2019). A cross-border, long haul freight transportation problem with transshipments. *International Journal of Logistics Systems and Management*, 32(3/4), 437–464. doi:10.1504/IJLSM.2019.098327
- Sun, R., Chen, Y., Dubey, A., & Pugliese, P. (2021). Hybrid electric buses fuel consumption prediction based on real-world driving data. *Transportation Research Part D, Transport and Environment*, 91, 102637. doi:10.1016/j.trd.2020.102637
- Topić, J., Škugor, B., & Deur, J. (2022). Neural network-based prediction of vehicle fuel consumption based on driving cycle data. *Sustainability (Basel)*, 14(2), 744. doi:10.3390/su14020744
- Yao, Y., Zhao, X., Liu, C., Rong, J., Zhang, Y., Dong, Z., & Su, Y. (2020). Vehicle fuel consumption prediction method based on driving behavior data collected from smartphones. *Journal of Advanced Transportation*, 2020, 1–11. doi:10.1155/2020/9263605
- Yao, Y., Zhao, X., Zhang, Y., Chen, C., & Rong, J. (2020). Modeling of individual vehicle safety and fuel consumption under comprehensive external conditions. *Transportation Research Part D, Transport and Environment*, 79, 102224. doi:10.1016/j.trd.2020.102224

Zhang, L., Peng, K., Zhao, X., & Khattak, A. J. (2023). New fuel consumption model considering vehicular speed, acceleration, and jerk. *Journal of Intelligent Transport Systems*, 27(2), 174–186. doi:10.1080/15472450.2021.2000406

Sheunesu Shamuyarira is an aspiring data analyst with an undergraduate degree in the Bachelor of Science Honours Degree in Operations Research and Statistics. An innate problem solver, and a hardworker.

Trust Tawanda is faculty member at the National University of Science and Technology, Zimbabwe. Has published papers and book chapters, and is currently a PhD student.

Elias Munapo is Professor of OR at North West University, South Africa. Has published in international journals, author of books and edited books and contributed book chapters.