Performance Analysis of CatBoost Algorithm and XGBoost Algorithm for Prediction of CO_2 Emission Rating

G Sandeep Kumar
Research Scholar
Department of Computer Science and Engineering
Saveetha School of Engineering
Saveetha Institute of Medical and Technical Sciences
Saveetha University
Chennai, Tamil Nadu, India.
sandeepkumarg19@saveetha.com

R Dhanalakshmi
Project Guide
Corresponding Author
Department of Machine Learning
Saveetha School of Engineering
Saveetha Institute of Medical and Technical Sciences
Saveetha University
Chennai, Tamil Nadu, India.
dhanalakshmir.sse@saveetha.com

2023 6th International Conference on Contemporary Computing and Informatics (IC3I) | 979-8-3503-0448-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/IC3I59117.2023.10398160

***Abstract* - The objective of this project is to improve the accuracy of CO2 Emission Rating predictions through the development of a model that incorporates various features by comparing CatBoost algorithm over XGBoost algorithm. Materials and Methods: The clincalc tool was used to determine the total number of iterations N equals 20 (10 iterations for each group) and to predict novel CO_2 Emission Rating with the improved accuracy. The significance value is obtained with the help of an independent sample T-test. Here the G-power analysis was carried out with 80%, alpha rate of 0.05. The novel CO_2 Emission Rating dataset with a size of 7530 was collected from Kaggle. Results: According to the findings, the difference in accuracy between the CatBoost algorithm is 98.01% and the XGBoost algorithm is 94.11% is statistically significant. For accuracy, the significant values are calculated asp=0.000($p<0.05$) which concludes that they are statistically significant. Conclusion: The accuracy value of the CatBoost algorithm is 98.01% whereas the accuracy value of XGBoost algorithm is 94.11%. CatBoost performs better when compared to XGBoost.**

Keywords: Machine Learning, Novel CO_2 Emission Rating, Prediction, CatBoost, XGBoost, Ecosystem.

I. INTRODUCTION

The challenges related to global warming have now spread to all countries. According to the Intergovernmental Panel on Climate Change (IPCC), scientists make up over 95% of the field. positive that rising greenhouse gas Ecosystem concentrations and other anthropogenic activities are what are primarily responsible for the majority of the world's warming , Prediction, Food and drink, desalination, cooling, cryogenic cleaning, welding and cutting, and healthcare are some of the key applications for liquid CO_2. Analysis and Prediction Model of Light-Duty Vehicle Fuel Consumption and Carbon Dioxide Emissions(Hien, Le Huy Hien, and Kor 2022).Additionally, increased oil recovery, chemicals, greenhouse horticulture, and other industrial sectors require purified gaseous CO_2. Increases in acid gasses such as nitrous oxide (N2O), carbon dioxide (CO_2), and methane (CH4) simulation of a HEV's fuel consumption(Lisowski et al. 2022). Perfluorocarbons (PFC) and hydrofluorocarbons (HFC), which are more commonly referred to as greenhouse gasses, have an impact on the equilibrium between the earth and atmosphere. IOP Publishing, 2016, "New prediction of carbon dioxide emissions using the support vector model."(Venkatraman and Alsberg 2017). CO_2 in particular is a significant contributor to global warming.Support vector machines for carbon dioxide emission prediction(Saleh, Dzakiyullah, and Nugroho 2016). Carbon dioxide emissions from burning fossil fuels worldwide each year are about eight billion tonnes.fuels used worldwide in the generation of heat and power as well as for transportation. The remaining products are the combustion of carbon monoxide with water, emissions of carbon dioxide, a greenhouse gas also called carbon dioxide.

The total number of papers published on CO_2 Emission Rating is 265, From that 250 papers on IEEE xplore and 15 papers on research gate(Kumar and Muhuri 2019). A New Method for Predicting GDP based on the CO_2 Emission Dataset and Transfer Learning and Utilizing, Light-Duty Carbon dioxide emissions and vehicle fuel use Analysis of Prediction Model(Hien, Le Huy Hien, and Kor 2022). Carbon dioxide emissions from vehicles are a significant contributor to climate change. As such, many countries and regions have implemented regulations aimed at reducing the amount of CO_2 emissions produced by vehicles(Meng and Noman 2022). One important aspect of these regulations is the development of CO_2 emission ratings for vehicles, which provide consumers with information about a vehicle's environmental impact.(Huang, Wu, and Cheng 2021) Support Vector Machine Prediction for Carbon Dioxide Emissions(Saleh, Dzakiyullah, and Nugroho 2016). Estimation of the volume regarding municipal solid trash using a one-dimensional neural network with convolutions, and a model for the attentional mechanism of an ecosystem(Lin et al. 2021). Carbon-di-oxide capture of ionic liquids predictionusing machine learning (Manju, Athira, and Rajendran 2021). CO_2 Emission Ratings for vehicles are typically calculated based on the amount of CO_2 emitted per kilometer driven(Huang, Wu, and Cheng 2021). Vehicles that

979-8-3503-0448-0/22/$31.00 ©2023 IEEE

emit less CO_2 per kilometer are given a higher rating, indicating that they are more environmentally friendly(Venkatraman and Alsberg 2017).

The study states that the disadvantage of XGBoost is that the kernel function must be carefully hand-tuned and it is impossible to obtain a comprehensive model. During this research, the CatBoost algorithm is compared to the XGBoost algorithm to predict the novel CO_2 Emission Rating. CatBoost produces a predictive model within the sort of a collection of faulty prediction models.

II. Materials And Methods

The proposed study is carried out in Chennai at the Saveetha Institute of Medical and Technological Sciences' Machine Learning Lab, which is part of the Saveetha School of Engineering. This division is a part of the computer science and engineering department.. Two algorithms have two groups each. The sample size is ten. The 'WSND' data collection is taken from the Kaggle website(Huang, Wu, and Cheng 2021). The data set shows the acquired values to detect the wireless attack detection.

The prediction of a CO_2 Emission with an improved accuracy rate was based on a sample size of twenty (ten from Group 1 and ten from Group 2), and the calculation was performed using a G-power of 0.8, with a 95% confidence interval(α= 0.05 and β=0.2). The sample size of twenty was ten individuals Ecosystem from Group 1 and ten individuals from Group 2. CatBoost and XGBoost, both with the same amount of data samples (N=10), are used to perform the prediction of novel CO_2 Emission Rating of vehicles to reduce the effect of climate change with CatBoost achieving a better improved accuracy rate. A machine prediction model is suggested for the Yangtze River Economic Zone's information on carbon emissions.

A. CatBoost Algorithm

Yandex created a machine learning system called CatBoost that uses boosted decision trees. Similar to other gradient boosted methods like XGBoost, it operates in a similar manner, except it supports categorical variables out of the box, improves accuracy more without adjusting parameters, and speeds up training with GPU support. CatBoost has proven to be a top performer on numerous Kaggle contests that employ tabular data and is used for a variety of regression and classification tasks. Several instances of CatBoost being used effectively are shown below.

Step 1: Import the necessary files.

Step 2: Bring the dataset into the programming environment.
Step 3: Data should be assigned to Y train, Y test, X test, and X train.

Step 4: Give test and random size as the parameters when using the The training and testing variables should be passed to the split() function after training.

Step 5: Import the Sklearn machine learning classifier. Predict the results of the testing datasets using CatBoost.

Step 6: The determination of required parameters is done so that the model is good to fit.

Step 7: Further analysis is performed and the measurement of improved accuracy is done.

B. XGBoost Algorithm

XGBoost is a effective open-source gradient boosted trees supervised learning algorithm. It combines the forecasts of a number of weaker, simpler models in an effort to accurately estimate a target variable. Regression trees are the weak learners in gradient boosting for regression. Each regression tree transfers a point from the input data to one of its leaves, which contains a continuous score. In order to create a regularized (L1 and L2) objective function that XGBoost minimizes, the convex loss function—using the difference between the predicted and target outputs. The model complexity penalty term are combined (in other words, the regression tree functions). When training progresses iteratively, new trees evolves which predict residuals.

Step 1: Import the necessary files.

Step 2: The dataset into the development environment

Step 3: Get the data for the X train, X test, Y train, and Y test..

Step 4: Pass the training and testing variables to the function train test split(), and then specify The parameters are test size and random size..

Step 5: Import the Sklearn machine learning classifier. Predict the results of the testing datasets using XGBoost.

Step 6: The determination of required parameters is done so that the model is good to fit.

Step 7: Further analysis is performed and the measurement of accuracy is done.

III. Statistical Analysis

The output is generated using Matlab software. SPSS is utilized for statistical methods of CatBoost and XGBoost approaches. CatBoost and XGB are the independent variables, while output accuracy serves as the dependent variable. For the purpose of determining which algorithm has

the highest output accuracy, two separate group analyses are carried out. The means, standard deviations, and standard errors of means were calculated in SPSS in order to compare the two samples and perform an independent sample t-test. For the Yangtze River Economic Zone's carbon emission information, a machine prediction model is suggested(Huang, Wu, and Cheng 2021).

IV. Results

Table 1 describes the CatBoost algorithm's accuracy has been measured to be around 98.01%. with the XGBoost algorithm is roughly 94.11%. When proposed to run the algorithms with various test sizes, as shown in Comparison of Accuracy Achieved When Evaluating CatBoost and XGB boost Algorithms for Predicting CO_2 Emissions With Different Iterations.

Table 2 represents the analysis of the means, standard errors, and accuracy of the CatBoost and XGBoost algorithms using statistics. use both algorithms' N=10 sample sizes. A statistically significant accuracy difference exists between the algorithms. The XGBoost algorithm is 94.11% accurate, compared to 98.01% for the CatBoost algorithm.

Table 3 shows the comparison significance levels with values greater than 0.05 The significance threshold of Accuracy is >0.05%, and the confidence intervals for the CatBoost and XGBoost algorithms are also 95%. 0.000 is the significance value.

Figure 1 displays the XGBoost algorithm (94.11%) and CatBoost algorithm (98.01%) mean accuracy comparison. When compared to the XGBoost algorithm model, CatBoost seems to have less of a standard deviation The CatBoost algorithm produces more trustworthy results than the XGBoost technique. Detection accuracy on the Y-axis is mean +/- 1 SD.

V. Tables And Figures

TABLE I. Comparison Of Accuracy Achieved When Evaluating Catboost And Xgboost Algorithms For Predicting CO2 Emissions With Different Iterations.

S.No	Sample_size	CatBoost accuracy in percentage	XGBoost accuracy in percentage
1	1	98.42	94.80
2	2	98.00	95.80
3	3	100.00	93.30
4	4	97.80	94.00
5	5	98.40	94.10
6	6	98.30	93.60
7	7	97.40	94.10
8	8	98.02	93.10
9	9	97.20	94.20
10	10	98.10	94.15

TABLE II. Statistical Analysis Of The Catboost Algorithm And Xgboost Algorithms' Means, Standard Deviations, And Accuracy Standard Errors. Between The Algorithms, There Is A Statistically Significant Variation In Accuracy. The Accuracy Of The Catboost Algorithm Is 98.01%, While That Of The Xgboost Algorithm Is 94.11%.

Algorithm Accuracy	N	Mean	Std Deviation	Std. Error Mean
CatBoost	10	98.0160	.40670	.12861
XGBoost	10	94.1150	.76668	.24245

TABLE III. Comparison Of Significance Level With Value P<0.05. Both Catboost And Xgboost Algorithms Have A Confidence Interval Of 95% And The Significance Level Of Accuracy Is < 0.05% Significance Value Is .000.

		Levene's Test for Equality of Variances		t=test for the equality of Means						
		F	sig	t	df	sig.(2-tailed)	Mean Difference	Std.Error Difference	95% Confidence interval of the Difference	
									lower	upper
Accuracy	Equal variances assumed	1.130	.302	14.214	18	.000	3.90100	.27445	3.32441	4.7759
	Equal variances not assumed			14.214	13.693	.000	3.90100	.27445	3.32441	4.7759

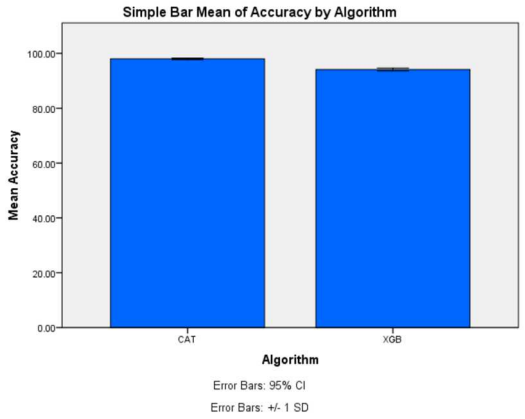

Figure 1.XGBoost algorithm (94.11%) and CatBoost algorithm (98.01%) mean accuracy comparison. When compared to the XGBoost algorithm model, CatBoost seems to have a lower standard deviation. Results from the CatBoost algorithm are more reliable than those from the XGBoost method. Y-axis: Mean Accuracy with +/-1 SD.

VI. Discussion

The research found that the proposed CatBoost algorithm performed better at forecasting novel CO_2 Emission Rating than the XGBoost model. Results from numerous repetitions of the experiment indicate that CatBoost has a greater accuracy of 98.01% than XGBoost, which has an improved accuracy of 94.11%.The Independent Sample T-Test yields a significance value for the CatBoost algorithm of 0.894.. The significance value of the Independent Sample T-Test is 0.000. Secondly, the goal was to create a network architecture with three parts, including the sections on the past, present, and outside, based on the correlation between vehicle CO_2 data on emissions and road conditions. The objective is to monitor CO_2 emissions derived from the use of coal and electricity in the manufacturing Ecosystem. The electrical and energy statistics were gathered in the setup to train and test the model. 60% of the statistics were from training, and 40% were from testing.

The similar study in the Positivity of novel CO_2 Emission RatingAlgorithms are using the machine learning technique for getting better predictions.To calculate the performance metrics More than two machine learning algorithms are compared(Nayyar et al. 2017). Opposing this research on the Negativity of novel CO_2 Emission Rating, This process is very complicated to find the improved accuracy of both the Algorithms(Hien, Le Huy Hien, and Kor 2022). Since it is essential to assess the environmental impact of intelligent transportation systems, the creation of a vehicle emission model with high enhanced accuracy has been a persistent problem in transportation research(Van Aerde and Baker,). Based on the findings the CatBoost algorithm appears to have a somewhat greater mean error rate than the XGBoost algorithm. Various methods may be used in the future to approve the mean error rate. Prediction using support vector machines(Saleh, Dzakiyullah, and Nugroho 2016). By providing consumers with information about a vehicle's CO_2 emissions, they can make more informed choices and select vehicles that are more environmentally friendly. Additionally, CO_2 emission ratings can incentivize automakers to develop more fuel-efficient and low-emission vehicles, helping to drive progress towards a more sustainable future(Wysocki, Deka, and Elizondo 2019).

The limitations of the proposed model of CO_2 Emission are computational expense, information type jumble and dataset size. There are several shortcomings in the study, Ecosystem despite the fact that the results of the suggested method are superior in statistical and experimental analyses. The mean error rate in the CatBoost algorithm seems to be a little bit higher than in the XGBoost. The mean error rate may be approved in the future utilizing a variety of techniques. Global warming-related problems are now present in all nations. Scientists are over 95% certain of growing greenhouse gas concentrations in ecosystems, according to climate change, according to the Intergovernmental Panel on Climate Change (IPCC).

VII. Conclusion

The study showed that the CatBoost Algorithm achieved a higher accuracy rate of 98.01%, outperforming the XGBoost Algorithm's accuracy of 94.11% for the dataset obtained from kaggle. Independent sample t-test was done in the SPSS tool. The independent sample t-test had given a significance p value of 0.000 ($P<0.05$), which is statistically significant. Thus, the results indicate that the CatBoost Algorithm is a superior option compared to the XGBoost algorithm and can lead to improved outcomes.

References

[1] Ben-Chaim, Michael, Efraim Shmerling, and Alon Kuperman. 2013. "Analytic Modeling of Vehicle Fuel Consumption." *Energies*. https://doi.org/10.3390/en6010117.Canada, Atlas of, and Atlas of Canada. 2010. "Light-Duty Vehicle Fuel Efficiency Improvement - Model Year 1990 to 2010 (Human Activities Leading to Emissions)." https://doi.org/10.4095/301034.

[2] Hien, Ngo Le Huy, Ngo Le Huy Hien, and Ah-Lian Kor. 2022. "Analysis and Prediction Model of Fuel Consumption and Carbon Dioxide Emissions of Light-Duty Vehicles." *Applied Sciences*. https://doi.org/10.3390/app12020803.

[3] Huang, Huafang, Xiaomao Wu, and Xianfu Cheng. 2021. "The Prediction of Carbon Emission Information in Yangtze River Economic Zone by Deep Learning." *Land*. https://doi.org/10.3390/land10121380.

[4] Kumar, Sandeep, and Pranab K. Muhuri. 2019. "A Novel GDP Prediction Technique Based on Transfer Learning Using CO2 Emission Dataset." *Applied Energy*. https://doi.org/10.1016/j.apenergy.2019.113476.

[5] Lin, Kunsen, Youcai Zhao, Lu Tian, Chunlong Zhao, Meilan Zhang, and Tao Zhou. 2021. "Estimation of Municipal Solid Waste Amount Based on One-Dimension Convolutional Neural Network and Long Short-Term Memory with Attention Mechanism Model: A Case Study of Shanghai." *The Science of the Total Environment* 791 (October): 148088.

[6] Lisowski, Maciej, Wawrzyniec Gołębiewski, Konrad Prajwowski, Krzysztof Danilecki, and Mirosław Radwan. 2022. "Modeling the Fuel Consumption by a HEV Vehicle – a Case Study." *Combustion Engines*. https://doi.org/10.19206/ce-157112.

[7] Manju, B. R., V. Athira, and Athul Rajendran. 2021. "Efficient Multi-Level Lung Cancer Prediction Model Using Support Vector Machine Classifier." *IOP Conference Series: Materials Science and Engineering*. https://doi.org/10.1088/1757-899x/1012/1/012034.

[8] Meng, Yang, and Hossain Noman. 2022. "Predicting CO2 Emission Footprint Using AI through Machine Learning." *Atmosphere*. https://doi.org/10.3390/atmos13111871.

[9] Saleh, Chairul, Nur Rachman Dzakiyullah, and Jonathan Bayu Nugroho. 2016. "Carbon Dioxide Emission Prediction Using Support Vector Machine." *IOP Conference Series: Materials Science and Engineering*. https://doi.org/10.1088/1757-899x/114/1/012148.

[10] Van Aerde, M., and M. Baker. "Modeling Fuel Consumption and Vehicle Emissions for the TravTek System." *Proceedings of VNIS '93 - Vehicle Navigation and Information Systems Conference*. https://doi.org/10.1109/vnis.1993.585599.

[11] Venkatraman, Vishwesh, and Bjørn Kåre Alsberg. 2017. "Predicting CO2 Capture of Ionic Liquids Using Machine Learning." *Journal of CO2 Utilization*. https://doi.org /10.1016/ j.jcou.2017.06.012.

[12] Wei, Siwei, Ting Wang, and Yanbin Li. 2017. "Influencing Factors and Prediction of Carbon Dioxide Emissions Using Factor Analysis and Optimized Least Squares Support Vector Machine." *Environmental Engineering Research*. https://doi.org/10.4491/eer.2016.125.

[13] Wysocki, Oskar, Lipika Deka, and David Elizondo. 2019. "Heavy Duty Vehicle Fuel Consumption Modeling Using Artificial Neural Networks." *2019 25th International Conference on Automation and Computing (ICAC)*. https://doi.org/10.23919/iconac.2019.8895072.