Journal of Artificial Intelligence and Data Science (JAIDA)
e-ISSN: 2791-8335
https://dergipark.org.tr/pub/jaida

Vol. 3, No. 2, pp. 107-115, 2023
Research Article
Received: 11/09/2023 Accepted: 19/11/2023

An Analysis for Car Fuel Estimation with Regression Methods

Enes TAŞKIN [1], Vedat MARTTİN [2,*]

Abstract

Fuel consumption and efficiency have emerged as pressing concerns in the context of growing energy sources and increasing environmental awareness. Machine learning, a subset of artificial intelligence, leverages intricate data structures and variable information to make predictions. These algorithms play a pivotal role in modeling and forecasting across diverse industries like healthcare, finance, banking, and energy.

This study offers a comprehensive overview of a typical machine learning project flow, with a particular focus on fuel prediction. The project encompasses key stages such as data collection, data preparation, model development, and evaluation. The methodologies and algorithms employed in this research hold the potential for broader applications in various forecasting projects and industry sectors.

In this investigation, fuel estimation was carried out using a set of features from the Auto MPG Data Set, sourced from the University of California. These features included Mpg (fuel consumption), Number of Cylinders, Engine Volume, Horsepower, Vehicle Weight, Acceleration, Model Year, Vehicle Origin, and Vehicle Name. Various regression algorithms, namely Linear Regression, Ridge Regression, Lasso Regression, and XGBoost, were applied to predict fuel consumption. The study's outcomes were generated by splitting the dataset into training and test data subsets. In the study[1], it was observed that Lasso regression was generally a little more prominent than the others in terms of error metric (RMSE=0.132369, MSE=0.017522, MAE=0.099490) and R^2=0.834900. It was seen that Linear regression was slightly better in terms of training data (RMSE=0.065446, MSE=0.004283, MAE=0.054682, R^2 =0.949617).

Keywords: Artificial intelligence (AI), Machine learning (ML), Modelling, Prediction, Regression

1. Introduction

Fuel consumption and regression models are emerging as an important area of research and application today. Studies in this area emphasize the use of artificial intelligence (AI) and machine learning (ML) techniques. Modeling, forecasting and regression, analyzing fuel consumption data and future "Regression models help us understand and predict the relationship between dependent variables and independent variables. In this context, AI and machine learning offer powerful tools to model complex data structures and improve fuel consumption forecasts. Regression models are widely used in various fields such as marketing, energy, fuel consumption [1].

In the literature, prediction studies have been made with regression models. In the studies, marine vehicles were studied in terms of fuel consumption. In terms of price estimation studies, it has been seen that there are studies such as air ticket, car sales price. This research[2] focuses on the development of predictive models for estimating fuel consumption based on real-world data gathered from a cruise ship during its operation. The selection of input variables for these models was carried out through a combination of statistical analysis and domain knowledge expertise. The study investigated various prediction models, including Multiple Linear Regression (MLR), Decision Tree (DT) approach, Artificial Neural Network (ANN), and ensemble methods. In this research, a comparative analysis of various predictive modeling approaches was conducted across multiple studies. The first study [3] explored the effectiveness of penalized regression techniques, including Ridge, Lasso, and Elastic Net, in predicting flight ticket prices. In the subsequent investigation [4], the performance of Lasso and Linear Regression models was assessed for predicting used car prices. Similarly, another study [5] scrutinized

*Corresponding author

Enes TAŞKIN; Bilecik Seyh Edebali University, Faculty of Engineering, Department of Computer Engineering, Türkiye; e-mail: enestaskin938@gmail.com; 0009-0009-7533-9627

Vedat MARTTİN; Bilecik Seyh Edebali University, Faculty of Engineering, Department of Computer Engineering, Türkiye; e-mail vedat.marttin@bilecik.edu.tr; 0000-0001-5173-2349

[1] "An earlier version of this paper was presented at the ICADA 2023 Conference and was published in its Abstract Book (Title of the conference paper: "Fuel Estimation Study with Regression Algorithms")

the predictive capabilities of Lasso, Multiple Linear Regression (MLR), and Decision Tree methods. Furthermore, a separate study [6] involved a comparison between Linear Regression, Ridge Regression, and Lasso Regression. Meanwhile, in yet another study [7], car price prediction was tackled using Random Forest, K-Nearest Neighbors (KNN), Decision Tree, XGBoost, and Linear Regression models. Lastly, in the study denoted as [8], the performance of Support Vector Machine (SVM) was evaluated for predicting used car prices.

Collectively, these studies contribute to a broader understanding of predictive modeling methodologies in various fields. It is aimed to contribute to the literature by using regression models on car fuel consumption of the data set used in this study, which we have not seen used in the literature.

The subsequent sections of this paper are structured as follows: The succeeding chapter delves into the intricacies of data preparation and exploration, elucidating the processes involved in data processing and predictive modeling. The penultimate segment provides a comprehensive account of the findings, while the ultimate section encapsulates the conclusions drawn and offers recommendations for future research.

2. Methodology

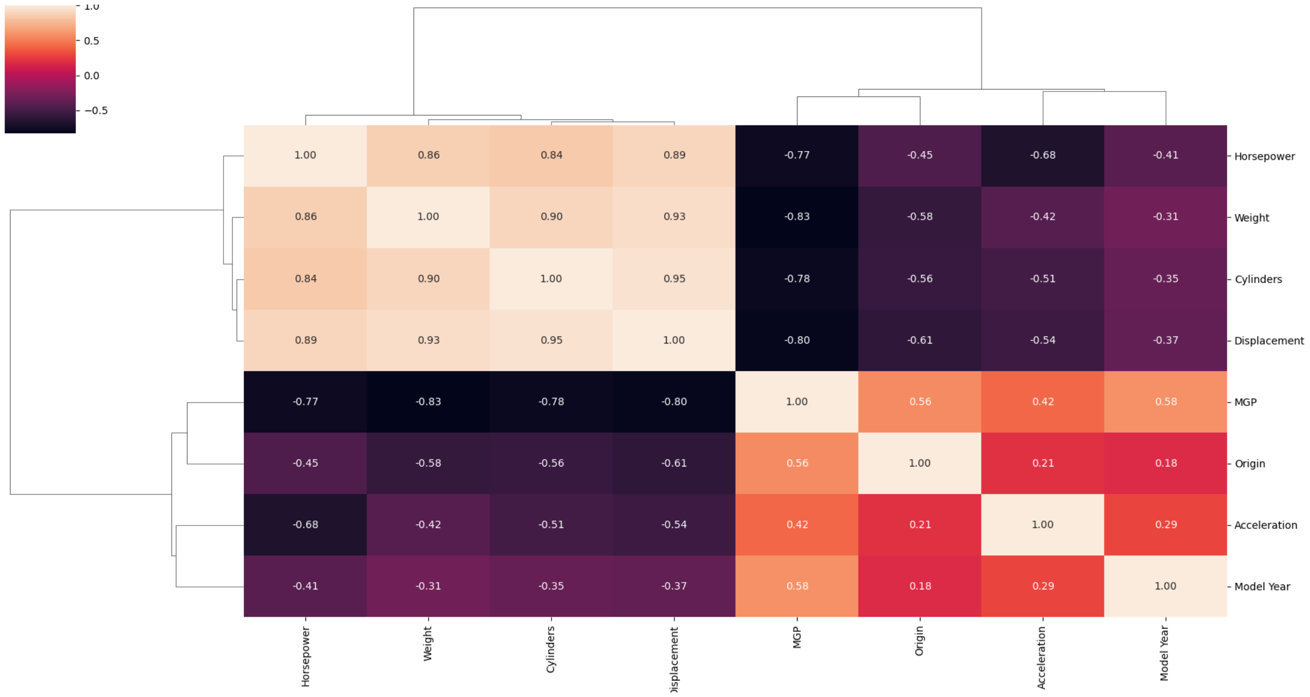
2.1. Data Preparation and Exploration

In this section, Auto MPG Dataset is selected and this data is taken from University of California. This dataset is a slightly modified version of the dataset provided in the StatLib library. Derivatives of the dataset can be found on the kaggle web page [9]. It was used by Ross Quinlan to estimate the "mpg" attribute. For this purpose, 40 of 398 data (10%) were determined as training data, and the remaining 356 data (90%) were reserved as test data. The variables in the data set are multivariate, categorical and real data. Table 1 shows the variables in the data set.

Table 1. *The variables in data set*

Variables
Mpg(fuel consumption)
Cylinders(number of cylinders)
Displacement(inches of engine)
Horsepower
Weight(vehicle weight)
Acceleration
Model year
Origin(vehicle origin)
Car name(vehicle name)

There are missing values in this data set. These missing values were removed from the data set and processed. It has been observed that the chance of success decreases in models with missing data. Figure 1 shows the Correlation matrix of the variables in the data set.



	Horsepower	Weight	Cylinders	Displacement	MGP	Origin	Acceleration	Model Year
Horsepower	1.00	0.86	0.84	0.89	-0.77	-0.45	-0.68	-0.41
Weight	0.86	1.00	0.90	0.93	-0.83	-0.58	-0.42	-0.31
Cylinders	0.84	0.90	1.00	0.95	-0.78	-0.56	-0.51	-0.35
Displacement	0.89	0.93	0.95	1.00	-0.80	-0.61	-0.54	-0.37
MGP	-0.77	-0.83	-0.78	-0.80	1.00	0.56	0.42	0.58
Origin	-0.45	-0.58	-0.56	-0.61	0.56	1.00	0.21	0.18
Acceleration	-0.68	-0.42	-0.51	-0.54	0.42	0.21	1.00	0.29
Model Year	-0.41	-0.31	-0.35	-0.37	0.58	0.18	0.29	1.00

Figure 1. *Correlation matrix for MGP data set*

It is drawn as a 0.75 filter in the figure and when you look at this table, there is a correlation of -0.83 between MGP and cylinder. Likewise, there appears to be a -0.77 correlation between MGP and horsepower, a -0.83 correlation between MGP and vehicle weight, and a -0.80 correlation between MGP and engine inches. The features appearing in this table seem to be the features that have the most relationship with MPS.

2.2 Data Processing and Predictions

At this stage, real artificial intelligence and machine learning algorithms are used to process data and give predictions. The data set is divided into training/sample set and testing set. Training of algorithms is done with sample data set and the test set is used to verify the results. Evaluation of the results is done using . performance measurements and predictions are obtained. In the selection of these algorithms, regression methods that are expected to be main, effective and have high success rates were preferred, taking into account the size of the data set. Below, information is given about Linear, Lasso, Ridge equations and XGBoost, which is a decision tree structure.

Linear regression is a machine learning approach employed to establish a mathematical connection between dependent and independent variables. This method encompasses two primary categories: simple linear response variable based on a single predictor, while multiple linear regression is particularly useful when predicting a response variable utilizing multiple predictors [10]. The process of linear estimation is formally defined as follows in Eq.(1):

$$PRegLinear = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{1}$$

where the p th predictor is outlined by X_p, and also an association between a variable and the response is measured by β_p.

Rigde regression serves as a valuable tool for determining the coefficients in multiple-regression models, particularly when the independent variables are strongly correlated with one another. This technique often results in smaller variances and mean square estimators compared to those obtained through least square estimations [11]. The ridge estimate is formally defined as follows in Eq.(2):

$$PRegRidge = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j{}^2 \tag{2}$$

In this context, the parameter λ (where $\lambda \geq 0$) plays a crucial role as a tuning parameter, and the term $\lambda\sum_{j=1}^{p}\beta_j{}^2$ is known as a shrinkage penalty. This penalty serves the purpose of mitigating overfitting and addressing issues related to multidimensionality. Ridge regression constructs a model that involves all parameters without excluding any variables and simultaneously nudges the coefficients closer to zero. Consequently, it becomes essential to carefully determine an appropriate value for alpha (the penalty term) during the model-building process [12]. It's worth noting that utilizing ridge regression doesn't offer any advantages when λ equals zero.

Lasso regression simultaneously conducts variable selection and regularization to enhance the accuracy and interpretability of the statistical model being developed. Its primary objective is to identify coefficients that minimize the sum of squared errors while applying penalties to these coefficients [12]. The lasso regression function is outlined as follows in Eq.(3):

$$PRegLasso = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{3}$$

Unlike ridge regression, lasso regression differs in that it forces the coefficients of irrelevant variables to assume a value of zero. In more precise terms, the component $\beta_j{}^2$ in Eq.(2) is replaced with $|\beta_j|$ in Eq.(3) in the context of lasso regression.

XGBoost, stands for Extreme Gradient Boosting and stands for gradient boosting and decision tree It is a machine learning technique based on algorithms. Extreme gradient boosting (XGBoost) is an ensemble learning technique that creates a comprehensive model by combining a collection of individual models, typically decision trees. This method employs gradient-based optimization, akin to how neural networks utilize gradient descent for weight optimization [13]. It calculates second-order gradients of the loss function to minimize errors and incorporates advanced regularization techniques, including L_1 and L_2 regularization [14]. These regularization methods help reduce overfitting, ultimately enhancing the model's ability to generalize and perform well.

2.3 Performance Evaluation Methods

Several error metrics such as RMSE (Root Mean Square Error), MSE (Mean Squared Error), MAE (Mean Absolute Error), and R^2 (Coefficient of Determination) are employed to evaluate the predictive performance of the developed models. These reference error metrics provide insights into the extent of model training and the nature of errors. The metrics used for performance assessment are detailed as follows in Eqs.(4)-(7):

RMSE represents the standard deviation of prediction errors (residuals), indicating how tightly the data points cluster around the best-fitting line [15, 16]. RMSE estimate is defined by Eq.(4)

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(P_i - T_i)^2} \tag{4}$$

MSE quantifies the average of the squared discrepancies between predicted values and actual observations [17]. It is alternatively referred to as mean squared deviation (MSD). MSE estimate is defined by Eq.(5).

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(T_i - P_i)^2 \tag{5}$$

MAE assesses the absolute magnitude of discrepancies between pairs of data points, offering a direct comparison between predicted values and actual observations within the current context [15, 17]. MAE estimate is defined by Eq.(6).

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|P_i - T_i| \tag{6}$$

R-squared (R^2) evaluates the fraction of the variability observed in the dependent variable that can be accounted for or predicted by the independent variable(s) [17]. This metric is commonly referred to as the Coefficient of Determination. R^2 estimate is defined by Eq.(7).

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(Pi-Ti)^2}{\sum_{i=1}^{m}(Ti-\mathrm{T}^*)^2} \tag{7}$$

where 'P_i' represents the predicted values, 'T_i' represents the actual or tested values, and 'm' signifies the total number of data points. The mean value of all the tested values is calculated as the sum of the tested values divided by '*m*'.

2.4 K-Fold Cross-Validation

K-fold Cross Validation is not an error metric, but it is useful in training models and measuring performance. It helps to determine whether the high performance of the model is purely by chance [18]. This technique involves randomly dividing the dataset into 'k' approximately equally sized subsets or multiples. During each iteration, *'k-1'* of these subsets are used to train the model while the remaining subset is reserved for testing. The process is repeated 'k' times and the average error value in these experiments serves as an indicator of the validity of the model. Typically 'k' is chosen as 3 or 5, but can also be set to larger values such as 10 or 15.

2.5 Experimental Setup and Hyperparameter Tuning

In the study, we used the pandas library to draw the data set with the Python language, the seaborn and pairplot libraries to easily visualize the data in the form of plots, the scipy library to eliminate the skewed values in our data, and finally to standardize our data, perform cross-validation testing and regression. The sklearn library was used to perform the operations. The computer features and program information used are shown in the Table 2.

Table 2. *System Configuration*

Hardware and Sofware	Characteristics
Memory(RAM)	8 GB DDR4
Processor	Amd Ryzen5 2500U
Graphics	Readeon RX560
Operating System	MS Windows 10
Integrated Development Environment(IDE)	Anaconda Navigator, Spyder
Programming Languge	Python
Library	Pandas, Seaborn, Pairplot,Numpy, Matplotlib, Scipy, Sklearn.

The hyperparameters used to optimize the regression types used in the study are shown in the Table 3. The selected algorithm limits were determined by comparing the values that gave high success rates, which were used in many projects before, and the success rates obtained with various Monte Carlo methods used in the study.

Table 3. *Hyperparameters of Machine Learning Models*

Models	Hyperparameters	Optimal Values
Linear	Learning_rate	0.01
	Random State	42
	Max. Iteration	1000 epoch
Ridge	Learning_rate	0.001
	Random State	42
	Max. Iteration	1000 epoch
Lasso	Learning_rate	0.0001
	Random State	42
	Max. Iteration	1000 epoch
XGBRegressor	Criterion	Friedman
	Splitter	Random
	Min_sample_split	100

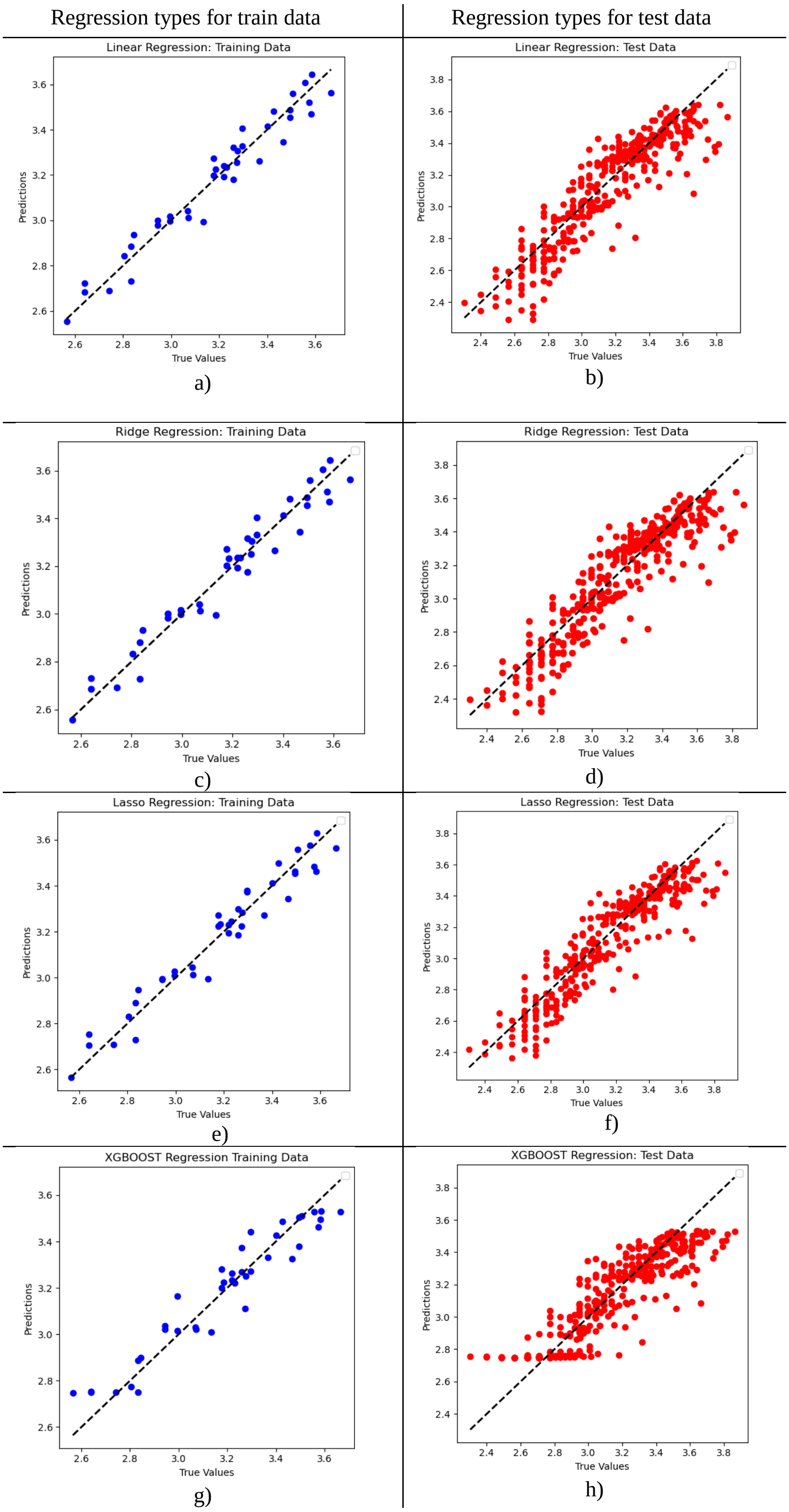
3. Results

The results of the regression models in the study according to performance metrics are shown in Table 4. It has been observed that the results of the regression models are close to each other according to the regression type. In regression performance analysis, error metrics such as RMSE, MSE, MAE are expected to be close to zero (0), while the R^2 value is desired to be close to one (1). As can be seen from Table 4, in the general case, Lasso regression is good in terms of RMSE, MSE, MAE, R^2 (0.132369-0.017522-0.099490-0.834900) values, respectively. It is seen that linear regression is slightly better in terms of training data, with RMSE, MSE, MAE, R^2 (0.065446-0.004283-0.054682-0.949617) values, respectively.

Table 4. *Performance evaluation measures for Regressions.*

Regression Types/Metrics	General				Train			
	RMSE	**MSE**	**MAE**	**R^2**	**RMSE**	**MSE**	**MAE**	**R^2**
Linear	0.143639	0.020632	0.106551	0.805590	0.065446	0.004283	0.054682	0.949617
Ridge	0.140447	0.019725	0.104693	0.814135	0.065554	0.004297	0.054940	0.949450
Lasso	0.132369	0.017522	0.099490	0.834900	0.068098	0.004637	0.057570	0.945451
XGBRegressor	0.021799	0.021799	0.111439	0.794593	0.087112	0.007589	0.070412	0.910735

The curves for regression models training and testing data is shown in Figure 2. The graph shows the following models based on the letters; a) Linear Regression applied to the training data, b) Linear Regression applied to the testing data, c) Ridge Regression applied to the training data, d) Ridge Regression applied to the testing data, e) Lasso Regression applied to the training data, f) Lasso Regression applied to the testing data, g) XGBoost Regression applied to the training data, h) XGBoost Regression applied to the testing data.



Regression types for train data

Regression types for test data

Linear Regression: Training Data

a)

Linear Regression: Test Data

b)

Ridge Regression: Training Data

c)

Ridge Regression: Test Data

d)

Lasso Regression: Training Data

e)

Lasso Regression: Test Data

f)

XGBOOST Regression Training Data

g)

XGBOOST Regression: Test Data

h)

Figure 2. *The curves for regression models training and testing data*

In Figure 2, it is seen that the Linear, Ridge and Lassso curves of the regression models are concentrated in the regression curve region better than XGBoost in terms of training and test data.

4. Conclusions and Discussion

Predictions are made with the help of machine learning algorithms, using the complex structure of the data and the information accumulated in the variables. Today, machine learning algorithms play an important role in fields such as health, finance, banking and energy. Although electric vehicles have become widespread, fossil fuel and related fuel consumption estimates are still made due to reasons such as the widespread use of the service network of charging stations. In this study, the performance of a data set containing the data of cars that human beings need in life was examined using Linear Regression, Ridge Regression, Lasso Regression and XGBOOST algorithms. Auto MPG Data Set collected by the University of California was used as the data set. Application results were obtained by taking part of the data set as training data and the remaining part as test data. According to the research, academic fuel consumption studies have been conducted, but no studies similar to this study have been found. There are studies using different algorithms and variable parameters. In summary, in this study, the results obtained by applying various regression methods and algorithms to the vehicle data set are compared. In the study, it was observed that Lasso regression was generally a little more prominent than the others in terms of error metric and R^2 (RMSE=0.132369, MSE=0.017522, MAE=0.099490, R^2=0.834900). It was seen that Linear regression was slightly better in terms of training data (RMSE=0.065446, MSE=0.004283, MAE=0.054682, R^2 =0.949617). By applying different techniques and regression models to this data set, performance can be increased in future applications. In addition, this algorithm can be tested on different data sets and this study can be turned into an application that everyone can access and use via a website or a smart phone application.

Declaration of Interest

The authors declare that there is no conflict of interest.

Acknowledgements

An earlier version of this paper was presented at the ICADA 2023 Conference and was published in its Abstract Book (Title of the conference paper: "Fuel Estimation Study with Regression Algorithms")

Author Contributions

Enes Taşkın: Investigation, Modelling, Software. Vedat Marttin: Supervision, Conceptualization, Methodology, Visualization, Investigation, Modelling, Sofware, Writing - Original Draft, Writing - Review &Editing.

References

[1] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions." SN computer science 2.3 ,2021: 160.

[2] Agand, Pedram, et al. "Fuel consumption prediction for a passenger ferry using machine learning and in-service data: A comparative study." Ocean Engineering 284 (2023): 115271.

[3] S. Buyrukoğlu, and Y.Yılmaz,"An Approach for Airfare Prices Analysis with Penalized Regression Methods." Veri Bilimi 4.2 pp.57-61, 2021.

[4] M. Asghar, K.Mehmood, S.Yasin and Z. M.Khan,"Used Cars Price Prediction using Machine Learning with Optimal Features". Pakistan Journal of Engineering and Technology, 4(2), pp:113-119, 2021.

[5] P. Venkatasubbu and M.Ganesh, "Used Cars Price Prediction using Supervised Learning Techniques". Int . J. Eng. Adv. Technol. (IJEAT), 9(1S3),2019.

[6] P. Rane, D.Pandya, D.Kotak," Used car price prediction "International Research Journal of Engineering and Technology (IRJET),2021.

[7] P. Gajera, A. Gondaliya, and J.Kavathiya, "Old Car Price Predict ion with Machine Learning". Int. Res. J. Mod. Eng. Technol. Sci, 3, pp:284-290,2021.

[8] S.Snehit, P.Borugadda, and N.Koshika. "Car Price Prediction: An Application of Machine Learning." 2023 International Conference on Inventive Computation Technologies (ICICT). IEEE, 2023.

[9] Kaggle [Online] Available: https://www.kaggle.com/datasets/uciml/autompg-dataset [Accessed Sept. 11, 2023].

[10] O. G., Uzut, S.Buyrukoglu, "Prediction of real estate prices with data mining algorithms". Euroasia Journal of Mathematics, Engineering, Natural and Medical Sciences.pp:77-84,2020, https://doi.org/10.38065/euroasiaorg.81

[11] T. Hastie, R.Tibshirani, J. H.Friedman and J. H Friedman. The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer,2009.

[12] G. James,D.Witten,T. Hastie, and R.Tibshirani, An introduction to statistical learning, Vol. 112, pp: 18,. New York: springer, 2013.

[13] J. Han, J.Pei, M.Kamber. Data mining: concepts and techniques. Amsterdam: Elsevier; 2011.

[14] F.Pedregosa, G.Varoquaux, A. Gramfort,V.Michel, B.Thirion. O.Grisel,M. Blondel, P.Prettenhofer, R.Weiss,V.Dubourg et al.”Scikit-learn: machine learning in python”. J Mach Learn Res.12:2825–30, 2011.

[15] J. S.Chou, and , A. D.Pham. “Enhanced artificial intelligence for ensemble approach to predicting high performance concrete compressive strength.”, Construction and Building Materials,49, pp:554-563,2013.

[16] P. N.Reddy and J. A. Naqash. “Strength prediction of high early strength concrete by artificial intelligence” Int J Eng Adv Technol, 8(3), pp:330-334,2019

[17] D. C.Feng, Z. T.Liu, X. D., Wang, Y .Chen, J. Q. Chang, D. F.Wei and Z. M. Jiang, “Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach”. Construction and Building Materials, 230, 117000, 2020.

[18] T. Fushiki,” Estimation of prediction error by using K-fold cross-validation”. Statistics and Computing, 21(2), pp:137-146, 2011.