Paper No: SC-17 **Smart Computing**

Feature selection in automobile price prediction: An integrated approach

Sobana Selvaratnam*
Department of Physical Science
Vavuniya Campus of the University of Jaffna, Sri Lanka
sobanaselvaratnam@gmail.com

B. Yogarajah
Department of Physical Science
Vavuniya Campus of the University of Jaffna, Sri Lanka
yoganbala@yahoo.com

T. Jeyamugan
Department of Physical Science
Vavuniya Campus of the University of Jaffna, Sri Lanka
tjeyamugan@vau.jfn.ac.lk

Nagulan Ratnarajah
Department of Physical Science
Vavuniya Campus of the University of Jaffna, Sri Lanka
rnagulan@univ.jfn.ac.lk

***Abstract* - Machine learning models for predictions enable researchers to make effective decisions based on historical data. Automobile price prediction studies have been a most interesting research area in machine learning nowadays. The independent variables to model the price and the price predictions are equally important for automobile consumers and manufacturers. Automobile consulting companies determine how prices vary in relation to the independent variables and they can then adjust the automobile's design, commercial strategy, and other factors to fulfill specified price targets. Furthermore, the model will assist management in comprehending a company's pricing patterns. The ability of machine learning systems to predict outcomes is entirely dependent on the effective selection of features. In this paper, we determine the influencing features on automobile price using an integrated approach of LASSO and stepwise selection regression algorithms. We use multiple linear regression to build the model using the selected features. From the experimental results using the automobile dataset from the UCI machine learning repository, the influencing features on automobile price are width, engine size, city mpg, stroke, make, aspiration, number of doors, body style, and drive wheels. Training data accuracy for predicting price was found to be 92%, and testing data accuracy was found to be 87%. The proposed approach supports selecting the most important characteristics of predicting the price of automobiles efficiently and effectively. This research will aid in the development of a model that uses the selected attributes to predict the price of automobiles using machine learning technologies.**

Keywords - automobile price prediction, feature selection, LASSO, stepwise selection

I. INTRODUCTION

One of the greatest and most important innovations in human history is the automobile. In 2020, almost 78 million automobiles were produced worldwide [1]. The price of an automobile is determined by a number of distinct features and elements and the accurate car price prediction necessitates specialist expertise. Customers who purchase a new car may assure their investment to be worthy. The automobile consulting companies must comprehend the aspects that influence automobile pricing. The manufacturers always have attention to the elements which are important in estimating the price of automobiles and interest on how well those variables accurately predict an automobile's price. An automobile price prediction system is, therefore, needed to accurately estimate the automobile's price based on a range of factors.

In the field of computer science, machine learning approaches have revolutionized the discipline. Automobile price prediction studies using machine learning approaches [2-6] guide better decisions and take smart actions for high accuracy predictions in real-time. Feature selection is one of the initial steps for the machine learning model assessment to reduce model complexity and increase model performance when it comes to generalization, model fit, and prediction exactness [7]. The problem of feature selection has been extensively researched in the literature [8-9]. Wrapper methods, filter methods, and embedding techniques are the most common feature selection approaches [10]. However, predicting an automobile's pricing and selecting the optimal features are complex tasks since automobiles have many properties but some of the factors only can describe the automobile price.

In supervised machine learning algorithms, when the response variable is a real or continuous value, it is a regression problem. The relationship between one continuous dependent variable and two or more independent variables is explained by multiple linear regression [11], a simple machine learning approach. The goal of this study is to find an appropriate technique for choosing optimal features for the price prediction of automobiles. The technique of selecting the smallest number of effective explanatory variables can more properly characterize a response variable. Stepwise selection [11], a wrapper method, and the LASSO regression methods [12], an embedded method, are the better feature selection methods, which provide a high prediction accuracy, supports to improve the interpretability of the model by removing extraneous variables that aren't related to the response variable, and prevents overfitting. In this study, the LASSO and stepwise selection methods were used in a hybrid way to build an appropriate model for the dataset to select the optimal features. The LASSO method [12] has been used for selecting the optimal features from the numerical variables and removing the multicollinearity of the variables. The stepwise selection method has been used to find the optimal features from the categorical variables. The stepwise selection method is applied again for the selected features from the numerical and the categorical variables to tune the final optimal feature set since the feature set chosen does not contain any multicollinearity. We proved the effectiveness of this integrated approach feature selection method for predicting automobile prices using the multiple linear regression approach with the selected features.

II. Related Work

Supervised learning is used in the vast majority of actual machine learning applications. Supervised machine learning techniques were utilized in the literature to predict the price of automobiles such as linear regression analysis [2,4,6], k-nearest neighbours [4,6], naïve Bayes [6], artificial neural network [5], support vector machine [5], and random-forest [2-5] and decision tree [4,6]. However, most of the research studies are highly interested in used automobile datasets [2-6]. For the feature extraction, different strategies were used by these studies, such as descriptive statistics [2], the correlation between variables [3,4,6], and data pre-processing [5]. There were no specific methods, only the heuristic, and basic statistical methods, used in these studies for selecting the optimal features. These research studies utilized different datasets and filter out the different sets of features such as (price, kilometre, vehicle type, and brand) [3], (number of doors, colour, mechanical and cosmetic reconditioning time, used to new ratio and appraisal to trade ratio) [4], and (brand, model, car condition, fuel, age, kilowatts, transmission, miles, colour, doors, drive, leather seats, navigation, alarm, aluminum rims, AC and more) [5]. Moreover, the main weakness of these studies is the low number of records that have been used [4,6].

Various approaches for solving the feature selection problem have been proposed in the literature. Wrapper methods [13], which use the output of an estimator or model in the selection process, and filter methods, which use heuristics to choose an ideal subset, are the standard strategies of feature selection. Popular regression methods have been used to extract the features for various prediction problems such as LASSO, OLS regression, ridge regression for Diabetes [14], LASSO for Diabetes [15], and LASSO for heart disease [16]. Muthukrishnan et al [14] proved, by decreasing the coefficients to zero, LASSO outperforms the other approaches. Valeria Fonti et al [17] showed the LASSO approach aids in the selection of a model with the most important properties, reduces the overfitting, increases the model interpretability, and has a very good prediction accuracy. New correlation matrices have been introduced in recent years that may have greater expressive capacity when measuring correlations between variables and feature selection [18-19]. However, these new correlation methods focus only on non-linear relationships rather than linear relationships.

Many of the unique algorithms have been constructed using only one form of selection strategies, such as a filter, wrapper, or embedded optimal feature collection procedure. Ensemble methods recently developed strategies [20] to choose influenced variables for machine learning purposes. In an ensemble method, multiple types of feature selection approaches are not taken into account. Furthermore, the use of ensemble feature selection is associated with automobile problems has not been investigated. Recently, optimal feature subsets formed by hybrid approaches combining filters, wrapper, and embedded feature selection approaches in medical datasets [21] and Gene expression data [22], which were performed well for feature selection. Hybrid filter-wrapper cluster-based feature selection method was applied for software defect prediction [23], short-term load forecasting [24], and intrusion detection systems [25]. Best of our knowledge, this study is the first attempt to select the optimal features to efficiently predict the price of a new automobile using a hybrid wrapper-embedded method in a unique approach.

III. Methodology

A. Dataset

The primary dataset was gathered from the automobile dataset from the UCI machine learning repository [26]. Each column in our collection represents a feature of the automobile, and each row represents one automobile. The dataset consists of 26 parameters, as listed in Table I, and the details of 205 automobiles. The outcome of the prediction on the automobile dataset is the price which is a continuous variable and predictors with both numerical and categorical values.

TABLE I. Description of the Attributes and the Datatype of the Automobile Dataset.

Attribute	Attribute Range
Symboling	3, -2, -1, 0, 1, 2, 3
normalized-losses	continuous from 65 to 256
Make	alfa-romero, audi, bmw, chevrolet, dodge, honda,isuzu, jaguar, mazda, mercedes-benz, mercury,mitsubishi, nissan, peugot, plymouth, porsche,renault, saab, subaru, toyota, volkswagen, Volvo
fuel-type	diesel, gas
Aspiration	std, turbo
num-of-doors	four, two
body-style	hardtop, wagon, sedan, hatchback, convertible
drive-wheels	4wd, fwd, rwd
engine-location	front, rear
wheel-base	continuous from 86.6 120.9
Length	continuous from 141.1 to 208.1
Width	continuous from 60.3 to 72.3
Height	continuous from 47.8 to 59.8
curb-weight	continuous from 1488 to 4066
engine-type	dohc, dohcv, l, ohc, ohcf, ohcv, rotor
num-of-cylinders	eight, five, four, six, three, twelve, two
engine-size	continuous from 61 to 326
fuel-system	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi
Bore	continuous from 2.54 to 3.94
Stroke	continuous from 2.07 to 4.17
compression-ratio	continuous from 7 to 23
Horsepower	continuous from 48 to 288
peak-rpm	continuous from 4150 to 6600
city-mpg	continuous from 13 to 49
highway-mpg	continuous from 16 to 54
Price	continuous from 5118 to 45400

B. Mathematical background

Multiple Linear Regression Model: Multiple Linear Regression is a statistical approach that predicts the outcome of a response variable by combining numerous explanatory variables. Multiple Linear Regression models can be described as below:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon i \qquad (1)$$

where dependent variable y_i explanatory variables x_i, regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$, a number of explanatory variables k, and error term εi .

LASSO Estimator [12]: The LASSO estimator can be defined by the solution to the l_1 optimization problem,

Minimize $\left(\frac{\|Y-X\beta\|_2^2}{n}\right)$ subject to $\sum_{j=1}^{k}\|\beta\|_1 < t$ (2)

where t is the upper bound for the sum of coefficients.

This optimization problem is equivalent to the parameter estimation that pursues,

$$\hat{\beta}(\lambda) = \underset{\beta}{argmin}\left(\frac{\|Y-X\beta\|_2^2}{n} + \lambda\|\beta\|_1\right) \quad (3)$$

where $\|\boldsymbol{Y}-\boldsymbol{X}\beta\|_2^2 = \sum_{i=0}^{n}(\boldsymbol{Y}_i - (\boldsymbol{X}\beta)_i)^2$,

$\|\beta\|_1 = \sum_{j=1}^{k}|\beta_j|$ and $\lambda \geq 0$ is that the parameter that controls the strength of the penalty, the high value of λ, the greater amount of shrinkage.

Stepwise Selection [11]: Forward and backward selections are combined in a stepwise selection. It starts with no predictors and then adds the most significant predictors one by one (like forwarding selection). Remove any predictors that no longer improve the model fit after each new predictor is included (like backward selection).

C. Integrated approach for feature selection

Predictive model training and deployment pipelines often include data pre-processing techniques, exploratory analysis, and feature engineering. In practice, cleansing data sets before feeding them to a learning algorithm is typical to increase model predictive performance and generalization potential. The pre-processing of the automobile dataset included removing inconsistent and noisy data and managing missing values. The detail of the pre-processing is described in the Feature selection section. A correlation matrix was used to investigate the dependency between variables and detect multicollinearity. A multiple linear regression model was initially built with all the 26 features in the dataset to check the r-squared value and the most significant features for the model.

When we applied the LASSO [12] and stepwise approaches [11] separately to the automobile dataset, they did not perform well. Many numerical factors were substantially correlated with price in the correlation analysis; however, this was not the case for categorical variables. Furthermore, in the automobile dataset, numerical parameters were more strongly influenced by pricing than category factors. The LASSO and stepwise selection methods were therefore used in an integrated way to build an appropriate model for the dataset to select the optimal features and get predictions. The approach is further described with the intermediate results in the Feature Selection section.

The pre-processed data split into a training dataset and testing dataset with ratios of 70 and 30 respectively. The training dataset has been used for model fitting and feature selection and the test data has been used for evaluating the prediction accuracy. An integrated approach using LASSO and stepwise methods was used to select the appropriate feature for predicting the price of automobiles. Data preparation and model building are processed by using the R programming language in Rstudio. We implemented the LASSO method making use of the glmnet package and the plotmo package in R.

D. Price prediction process

The selected optimal features from the integrated approach were used to build a model using multiple linear regression for predicting the price. The training set was evaluated first using the accuracy as r-squared. The test set was evaluated using the same subset of features and computed the accuracy. The model performance indicator for regression issues is based on the coefficient of determination, r-squared, and the percentage of r-squared of the price variation is explained by the variation in the optimum selected independent variables.

IV. Feature Selection

A. Data preprocessing

The data pre-processing step consists of removing the inconsistent and noisy data. The missing data were also removed if any variable has missing values above 50%. The imputation process was performed for other predictors with a small percentage of missing values. In our dataset, we first find out the variables with missing values, and then it regresses on other variables. The missing values of that variable were replaced by predicted values. Moreover, influencing outliers are revealed, and taken action to remove non-influencing outliers.

The price and log (price) histograms are shown in Fig.1. While the price range varies widely with a lengthy tail, log (price) appears to follow a normal distribution. As a result, the outcome of the model development and evaluation procedure will be log (price).

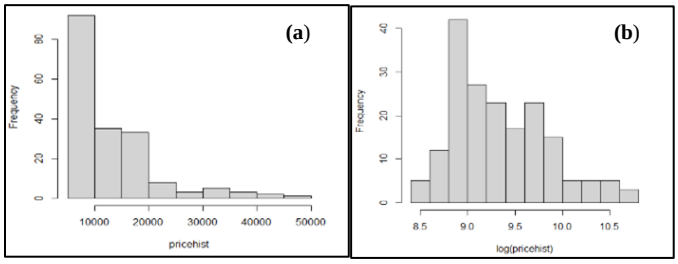

Fig. 1. (a) Price and (b) log(price) histograms

B. Data visualization and exploration

We evaluated the variable visually using matrix linear plots and bar plots. The wheelbase, length, width, curb weight, bore, and horsepower variables have a positive linear relationship with price than height, compression ratio, and peak rpm. City mpg and highway mpg have a negative linear relationship with price.

The correlation matrix of numerical attributes is visualized in Fig. 2. From the correlation matrix we can deduce that the response variable price is highly correlated with horsepower, bore, engine size, curb weight, width, and length. The price is also negatively correlated with highway mpg and city mpg. Some independent variables are highly correlated with each other such as wheelbase, length, width, height, curb weight, engine size, and bore. As a result, the vast majority of numerical variables are multi-correlated covariance variables.

The correlation matrix of the categorical variables was created using Kendall's Tau-b, which is visualized in Fig.3. From the correlation matrix, we can deduce that that price is highly correlated with the number of cylinders and drive

wheels. The price is also negatively correlated with the body style, make, and fuel type.

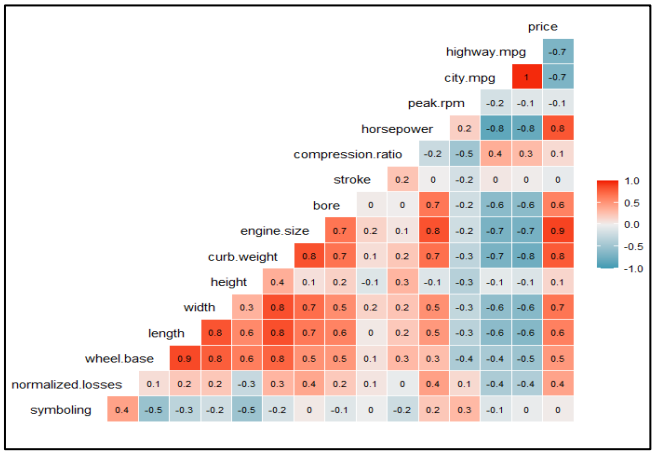

Fig.2. Correlation matrix for numerical variables

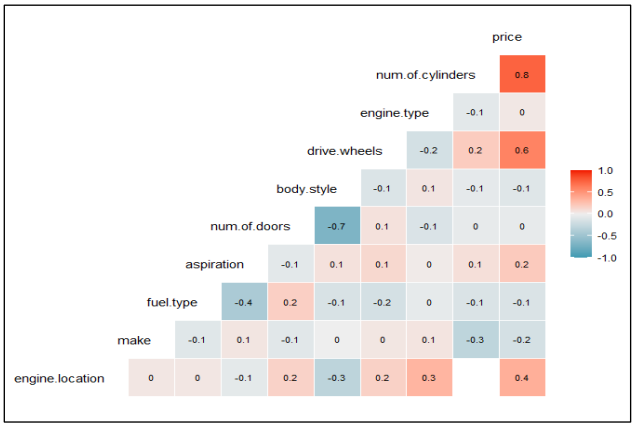

Fig.3. Correlation matrix for categorical variables

C. *Variable analysis using multiple linear regression*

Using all of the 26 variables in the dataset, a multiple linear regression model was created. We find out the most significant variables for the model using analysis variance (ANOVA). Table II presents the results of ANOVA.

TABLE II: ANOVA FOR THE AUTOMOBILE DATASET

Response Variable: log(price)				
Predictors	**Df**	**F value**	**Pr(>F)**	
symboling	1	41.8091	1.793e-09	***
normalized.losses	1	1102.8527	< 2.2e-16	***
make	15	140.0633	< 2.2e-16	***
fuel.type	1	0.6761	0.41243	-
aspiration	1	148.3762	< 2.2e-16	***
num.of.doors	1	52.6042	3.102e-11	***
body.style	4	14.2010	1.149e-09	***
drive.wheels	2	67.7883	< 2.2e-16	***
engine.location	1	4.5218	0.03533	*
wheel.base	1	169.4448	< 2.2e-16	***
length	1	90.6747	< 2.2e-16	***
width	1	44.9671	5.324e-10	***
height	1	2.2505	0.13596	-
curb.weight	1	114.3238	< 2.2e-16	***
engine.type	2	0.5032	0.60576	-
num.of.cylinders	3	1.6966	0.17086	-
engine.size	1	2.5527	0.11250	-
fuel.system	3	3.0498	0.03094	*
bore	1	1.6475	0.20155	-
stroke	1	2.0016	0.15948	-
compression.ratio	1	4.7318	0.03139	*
horsepower	1	0.7001	0.40427	-
peak.rpm	1	0.0857	0.77019	-
city.mpg	1	2.7325	0.10070	-
highway.mpg	1	3.8492	0.05187	.
Residuals	132	-	-	-

Based on the p-values of Table II, symboling, normalized losses, make, aspiration, num.of.doors, body style, drive wheels, engine.location, wheel.base, length, width, curb.weight, fuel.system, and compression.ratio are the most significant variables and other variables are not significant in the model. Thus, we can concern these significant variables for the final model.

D. *LASSO implementation*

We create a model to predict the price for the automobile dataset and to find out which explanatory variables to include in the final model using the LASSO regression method (In glmnet, alpha = 1 for the LASSO regression and alpha = 0 for the Ridge regularization). Glmnet generates a series of various models based on the tuning parameter λ. To determine the influencing features, we first utilized the function on all of the numerical explanatory factors in the automobile dataset. The analyses' findings are depicted in Fig. 4 and Fig.5. We can see when each variable entered the model and how much it changed the response variable using these charts. From Fig. 4, lasso included only 10 predictors out of 15 predictors which removed the following predictors such as normalized losses, length, curb weight, horsepower, peak rpm.

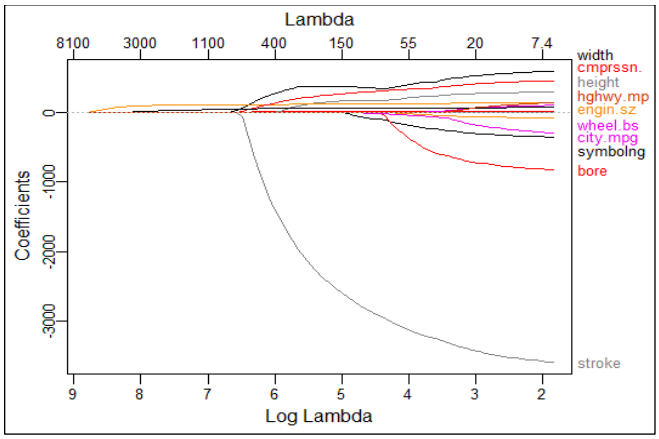

Fig.4. Glamnet graph for the numerical variables

A Correlation matrix was computed for the removed predictors by LASSO. Curb weight and horsepower are highly correlated (0.7326893) with price but they are highly correlated with each other. LASSO handles, therefore, the multicollinearity problem efficiently.

Fig.5 shows the top nine influencing predictors of automobile price. width, compression ratio, highway mpg, engine size have positively affected the model, and city mpg, symboling, bore, and stroke has negatively affected the model. To determine the value of λ, use k-fold cross-validation to find the λ value that generates the lowest test mean squared error (MSE).

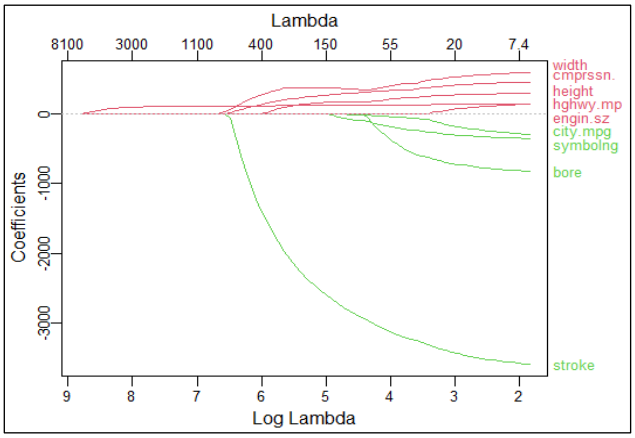

Fig.5. Glamnet graph for influencing explanatory variables

The LASSO approach extracts different values for λ to determine the best acceptable value for, such as λ-min (first vertical line in Fig.6), which offers the minimum mean cross-validated error, and λ-1se (second vertical line in Fig.6), which produces a model with error within one standard error of the minimum. At this point, we can select the value for λ that is most appropriate for the problem.

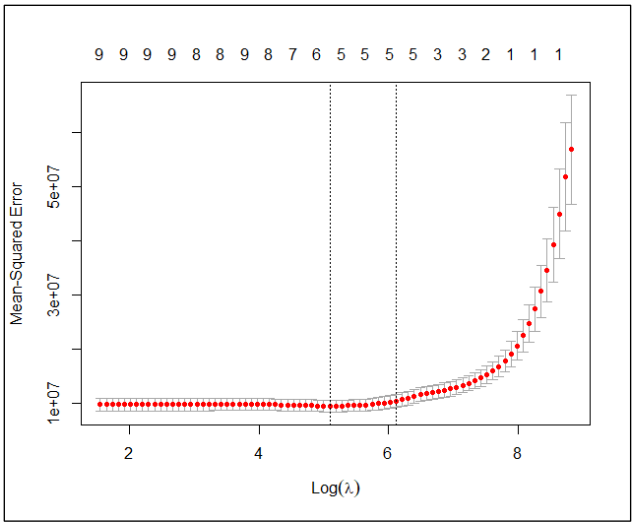

Fig. 6. Cross-validation

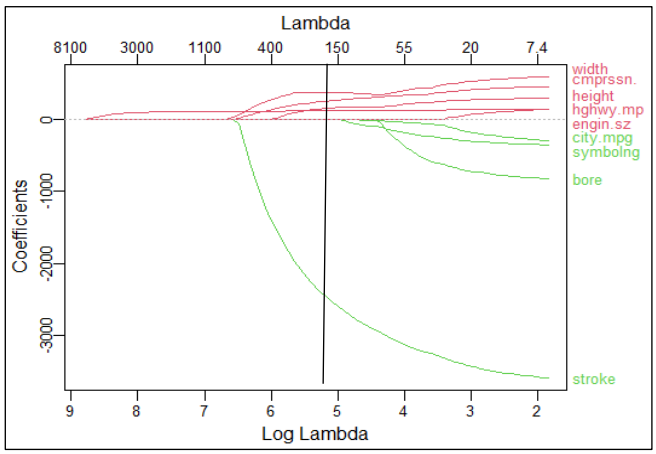

Fig. 7. Most important features

Because the aforementioned Fig. 6 plot exhibits an exponential trend, λ-min is not obvious in our analysis. So, we compute those two λ values (λ-min value =162.7058 and λ-1se value= 543.3253)

TABLE III: 10 X 1 SPARSE MATRIX OF CLASS

Predictors	Coefficient
(Intercept)	1525.35581
symboling	-
width	94.73858
height	-
engine.size	147.21568
bore	-
stroke	-2760.41510
compression.ratio	268.16365
city.mpg	-277.45107
highway.mpg	-

From Table III, no coefficient is shown for the predictors symboling, height, bore, and highway mpg because as a result of the lasso regression, the coefficient was reduced to zero. This means it was deleted entirely from the model because it did not influence it. By combining the plots in Fig. 7 and Table III, we can conclude. The most significant numerical variables for the price prediction from the automobile dataset are Width, compression ratio, engine size, city mpg, and stroke, which have been selected according to the λ-min value.

E. Stepwise selection implementation

The stepwise selection method was utilized for the categorical variables in the automobile dataset to find out the influencing features. The results of the stepwise selection regression method are shown in Table IV.

TABLE IV: STEPWISE SELECTION METHOD'S OUTCOME OF THE CATEGORICAL VARIABLES

log(price) ~ make + aspiration + num.of.doors + body.style + drive.wheels + num.of.cylinders + fuel.system

	Df	Sum of Sq	RSS	AIC
+ engine.location	1	0.00903	2.3681	-436.75
+ engine.type	3	0.01134	2.3658	-432.87
- body.style	4	0.29426	2.6714	-431.56
- aspiration	1	0.28457	2.6617	-426.02
- num.of.doors	1	0.51683	2.8940	-415.48
- fuel.system	4	0.72231	3.0994	-412.84
- num.of.cylinders	3	0.71775	3.0949	-411.02
- drive.wheels	2	0.78058	3.1577	-406.49
- make	15	2.88326	5.2604	-368.19

The above results (Table IV) present the final step of the stepwise selection for categorical predictors from the automobile dataset. From this method, seven influencing predictors are filtered such as make, aspiration, number of doors, body style, drive wheels, number of cylinders, fuel system. After that, we have applied the stepwise selection method to selected numerical and categorical predictors selected from lasso and stepwise selection.

The results (Table V) present, width, engine size, city mpg, stroke, make, aspiration, number of doors, body style, drive wheels, number of cylinders, fuel system are filtered out by stepwise selection method from the dataset. Predictors selected by the stepwise method are analysed by ANOVA. From the ANOVA (Table VI), the number of

cylinders and fuel systems are not significant. So, we removed them from the model.

TABLE V: STEPWISE SELECTION METHOD'S OUTCOME OF THE SELECTED NUMERICAL AND CATEGORICAL VARIABLES

log(price) ~ width + engine.size + city.mpg + stroke + make + aspiration + num.of.doors + body.style + drive.wheels + num.of.cylinders + fuel.system

	Df	Sum of Sq	RSS	AIC
- num.of.cylinders	3	0.07271	1.4078	-502.28
+ compression.ratio	1	0.00588	1.3292	-501.52
- city.mpg	1	0.05763	1.3927	-499.63
- fuel.system	4	0.13305	1.4681	-498.99
- width	1	0.06693	1.4020	-498.80
- stroke	1	0.10576	1.4408	-495.35
- num.of.doors	1	0.12324	1.4583	-493.83
- drive.wheels	2	0.15806	1.4931	-492.86
- aspiration	1	0.23138	1.5665	-484.82
- body.style	4	0.36160	1.6967	-480.76
- engine.size	1	0.34923	1.6843	-475.68
- make	15	1.44497	2.7801	-440.54

V. PRICE PREDICTION

The integrated approach's best attributes (width, engine size, city mpg, stroke, make, aspiration, number of doors, body style, and drive wheels) were used to create a model that employed multiple linear regression to forecast the price. We obtained 92% accuracy for the price prediction using the training set. We evaluated the final model using the testing dataset and we obtained 87% testing accuracy. The high r-squared values show that the selected independent variables with the hybrid feature selection method truly determine the price of automobiles. To reduce the overfitting problem and improve interpretation capabilities, the number of features in the chosen algorithms was maintained as minimal as possible.

The experimental results suggest that a hybrid approach integrating LASSO (embedded method) and Stepwise (wrapper method) regression techniques provides a high level of prediction accuracy and a reasonable rate of feature reduction. For the proper comparison with other approaches in the literature, no study in the literature uses the UCI machine learning [26] repository data to predict automobile prices. Some researchers used this automobile dataset for various purposes such as data-guided approach to generate multi-dimensional schema for targeted knowledge discovery [27], mapping nominal values to numbers for effective visualization [28], and attribute identification and predictive customization [29].

VI. CONCLUSIONS

The study presented a hybrid approach to select the optimal features to build an efficient model for the price prediction of automobiles. First, the dataset is analysed and pre-processed for the model building and then split the dataset into train and test datasets. Next, the feature selection was conducted using the training dataset using lasso and stepwise selection regression methods in an integrated way. The most relevant features for the prediction of automobiles are width, engine size, city mpg, stroke, make, aspiration, number of doors, body style, and drive wheels. These optimal features were evaluated in the multiple linear regression model with training dataset accuracy of 92% and testing dataset accuracy of 87% respectively. The findings show that combining embedded and wrapper feature selection to build a hybrid form of feature selection yields better outcomes.

REFERENCES

[1] Agencija za statistiku BiH. (n.d.), retrieved from: http://www.bhas.ba . [accessed January, 2021.]
[2] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, 2018, "Prediction of prices for used car by using regression models", 5th International Conference on Business and Industrial Research (ICBIR), 2018, pp. 115-119.
[3] N. Pal, P. Arora, S.Sumanth Palakurthy, D.Sundararaman, P.Kohli, 2017, How much is my car worth? "A methodology for predicting used cars prices using Random Forest", CoRR, abs/1711.06970
[4] P. Gajera, A. Gondaliya, J.Kavathiya, 2021,"Old Car Price Prediction With Machine Learning", International Research Journal of Modernization in Engineering Technology and Science,Volume:03, Issue:03, pp.284-290.
[5] E. Gegic, B. Isakovic, D.Keco, Z.Masetic, J.Kevric, 2019, "Car Price Prediction using Machine Learning Techniques", TEM Journal. Volume 8, Issue 1, pp. 113-118.
[6] S. Pudaruth, 2014, "Predicting the Price of Used Cars using Machine Learning Techniques, International Journal of Information & Computation Technology, Volume 4, Number 7 , pp. 753-764.
[7] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, 2007, "Data preprocessing for supervised leaning", International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol. 1, pp. 4104–4109.
[8] A. L. Blum and P. Langley, 1997, "Selection of relevant features and examples in machine learning", Artificial Intelligence, vol. 97, no. 1, pp. 245 – 271.
[9] H. Motoda and H. Liu, 2002, "Feature selection, extraction and construction", Communication of IICM (Institute of Information and Computing Machinery, Taiwan), vol. 5, pp. 67–72.
[10] Guyon, I., Elisseeff, A, 2003, "An introduction to variable and feature selection". Journal of machine learning research, pp.1157-1182.
[11] M.A. Efroymson, 1960, "Multiple regression analysis - Mathematical Methods for Digital Computers", Ralston A. and Wilf,H. S., (eds.), Wiley, New York.
[12] R. Tibshirani, 1996, "Regression shrinkage and selection via the lasso". J. R. Stat. Soc. Ser. B (Methodological), 58, pp. 267–288.
[13] A. Y. Ng, 1998, "On feature selection: Learning with exponentially many irrelevant features as training examples", in Proceedings of the Fifteenth International Conference on Machine Learning, ser. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 404–412.
[14] R. Muthukrishnan and R. Rohini, 2016, "LASSO: A feature selection technique in predictive modeling for machine learning", IEEE International Conference on Advances in Computer Applications (ICACA), pp. 18-20.
[15] P. M. Kumarage, B. Yogarajah and N. Ratnarajah, 2019, "Efficient Feature Selection for Prediction of Diabetic Using LASSO", 19th International Conference on Advances in ICT for Emerging Regions (ICTer), 2019, pp. 1-7.
[16] P. Ghosh et al., 2021, "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques", in IEEE Access, vol. 9, pp. 19304-19326.
[17] V Fonti, E Belitser, 2017, "Feature Selection using LASSO", VU Amsterdam Research Paper in Business Analytics, Volume 30, pp. 1-25.
[18] D. N. Reshef, Y. A. Reshef, M. Mitzenmacher, and P. C. Sabeti, 2013, Equitability analysis of the maximal information coefficient, with comparisons, CoRR, abs/1301.6314.
[19] A. Luedtke and L. Tran, 2013, "The generalized mean information coefficient, arXiv: Machine Learning", [Online]. Available: https://arxiv.org/abs/1308.5712
[20] D.Guan, W. Yuan, Y. Lee, K. Najeebullah, and M.K. Rasel, 2014. "A review of ensemble learning based feature selection". IETE Technical Review, 31(3), 190–198.
[21] C.W Chen, Y.H Tsai, F.R Chang, W.C Lin, 2020, "Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results". Expert Systems; e12553.
[22] S. Shilan Hameed, O. O.Petinrin, A.Osman Hashi and Faisal Saeed, 2018, "Filter-Wrapper Combination and Embedded Feature

Selection for Gene Expression Data", Int. J. Advance Soft Compu. Appl, Vol. 10, No. 1.

[23] F. Wang, J. Ai and Z. Zou, "A Cluster-Based Hybrid Feature Selection Method for Defect Prediction", 2019, IEEE 19th International Conference on Software Quality, Reliability and Security (QRS), pp. 1-9.

[24] Z. Hu, Y. Bao, T.Xiong, R.Chiong, 2015, "Hybrid filter–wrapper feature selection for short-term load forecasting", Engineering Applications of Artificial Intelligence, Volume 40, pp. 17-27.

[25] M.Kamarudin, C. Maple and T. Watson, 2019, "Hybrid feature selection technique for intrusion detection system", Int. J. High Performance Computing and Networking, Vol. 13, No. 2, pp.232 – 240.

[26] Automobile dataset from UCI machine learning repository, ,https://archive.ics.uci.edu/ml/datasets/automobile.

[27] R.L. Pears, M. Usman, A. Fong, 2012, "Data Guided Approach to Generate Multidimensional Schema for Targeted Knowledge Discovery", 10th Australasian Data Mining Conference (AusDM 2012), Sydney, Australia.

[28] G. E. Rosario, E. A. Rundensteiner, D. C. Brown and M. O. Ward, "Mapping nominal values to numbers for effective visualization", IEEE Symposium on Information Visualization 2003 (IEEE Cat. No.03TH8714), 2003, pp. 113-120.

[29] A. A. F. Saldivar, C. Goh, Y. Li, H. Yu and Y. Chen, "Attribute identification and predictive customisation using fuzzy clustering and genetic search for Industry 4.0 environments," 2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA), 2016, pp. 79-86.