➔ *Regular Research Paper – NS*

Predicting Used Car Prices with Heuristic Algorithms and Creating a New Dataset

Mehmet BILEN
Mehmet Akif Ersoy University, Golhisar School of Applied Science, Turkey
mbilen@mehmetakif.edu.tr

Abstract

Turkey is one of the countries with a high-volume second-hand car market. Today, used car sales advertisements given over the internet have accelerated this market even more. This situation has caused difficulties in determining the most suitable price for the vehicle to be bought or sold. The problem of determining the price of second-hand vehicles causes both buyers and sellers to have difficulties since it contains many variables. In this study, it is aimed to determine the best prediction model by using heuristic algorithms for the solution of this problem. In this direction, a new dataset including car features and price information was created by compiling the advertisements on the sites that provide car buying and selling advertisement services over the internet, and this dataset was shared for researchers to use in the model development phase. As a result of the estimation processes made with different algorithms and models on the dataset, it was seen that the Fisher+ANN model provided the lowest estimation error with MAE 0.01050, MSE 0.000281 and the highest performance value observed with R2 was 0.8958.

Keywords: *prediction, used car dataset, fisher, ann*

1. INTRODUCTION

Automobiles have an important place in our lives because they have been used to meet the transportation needs of humanity for a long time. With the increase in the human population, the number and variety of vehicles produced are also increasing rapidly. Looking at the second-hand vehicle market in Turkey, it is seen that 1,851,222 second-hand vehicle sales advertisements were created in the first 6 months of 2021 and 43 per cent of these vehicles were sold. Many variables need to be considered in deciding the price of a used vehicle in such a large market. To overcome this difficulty, heuristic algorithms, which have a great advantage in revealing the relationships between variables, are frequently used in the literature.

Studies on second-hand vehicle prices estimation are divided into two in the literature. The first of these is the studies on how and in what direction the price of a vehicle will change depending on time [1-3]. The other field of study is to determine the current price of a vehicle in the same period according to its characteristics. Two methods stand out in the studies conducted in this area. The first is the hedonic method, which directly examines the effect of each feature on the price [4-6]. The other is the intuitive approaches that constitute the method of this study [7-10].

This study is presented under 5 different titles. While the heuristic algorithms for the estimation of vehicle prices in the used car market are given in the important introduction, the stages of compiling the dataset and the methods and approaches used are given in the Material Method section. In the Findings section, the estimation results made with the models created are shared.

In the discussion part, these findings were tried to be clarified. Finally, in the conclusion part, all the findings obtained during the study were evaluated, the disadvantages of the study and the future studies were mentioned.

2. MATERIAL AND METHODS

2.1. Data Set

The samples in the dataset in this study were derived from 4 different sites that provide second-hand sales advertisements over the internet. Samples were selected from four different vehicle brands that represent 50 per cent of the used vehicle market. A car contains many numerical technical features and qualities expressed by linguistic knowledge. This situation both increases the size of the data set and makes it difficult for heuristic algorithms to work with these features. For this reason, the most important features affecting the price of the vehicle were determined by taking the opinions of the experts who are interested in second-hand car buying and selling. 9225 examples, including horsepower, engine capacity, brand, gear type, fuel type, model year and price information of the vehicle, were included in the dataset (Table 1.).

Table 1. Samples from the raw dataset

Index	Price	Horsepower	Engine Capacity	Brand	Gear Type	Engine Type	Model Year
0	167500	110 hp	1461 cc	B1	Semi-Automatic	Gasoline	2016
1	159500	115 hp	1598 cc	B2	Manuel	Gasoline	2018
2	136000	90 hp	1461 cc	B2	Manuel	Diesel	2015
3	179000	90 hp	1461 cc	B3	Automatic	Gasoline + LPG	2018
	...	...	...	...	...	...	...
9224	60500	68 hp	1399 cc	B4	Manuel	Gasoline	2004

After compiling the dataset, each attribute was examined one by one. Price information was determined as the attribute to be estimated by the algorithms, and no missing or incorrect information was encountered in this column. However, there are many different types of data or missing entries in other attributes, especially in the Horsepower and Engine Capacity columns, which should be numerical. Linguistic expressions have been cleared from these headings so that heuristic algorithms can work successfully. But, expressions such as "110-120hp" and "1450-1550 cc" indicating the range have been replaced with the numerical equivalent of the mean. To eliminate missing data, the average imputation method, which is frequently used in the literature, was preferred [11]. The general average of the missing data is written in the place of the missing data (Figure 2.1.).

	Attribute 1	Attribute 2	Attribute n
Sample 1	1	3	600
Sample 2	2	5	
Sample 3		4	662
Sample 4	1		900
Sample 5	2	4	500
Sample 6			
Average	1,5	4	665,5

Figure 1. Average Imputation

All the operations done up to this part allow the data to be evaluated with any research method, independent of the algorithm. For researchers who want to work on the dataset, this version of the dataset has been shared on the UCI Machine Learning site [12].

To analyze the dataset with heuristic algorithms and a training process to predict, a series of operations still need to be performed on the dataset. For this reason, firstly, the titles with linguistic expressions such as "Gear type", "Fuel type" and "Brand" were coded with the dummy coding method and converted into numerical values. The coding made are presented in Table 2., Table 3. and Table 4 respectively.

Table 2. Coding of gear type

	Simple Coding	**Dummy Coding**		
Sample	**Gear Type**	**Manuel**	**Automatic**	**Semi-Automatic**
Manuel	1	1	0	0
Automatic	2	0	1	0
Semi-Automatic	3	0	0	1

Table 3. Coding of the fuel type attribute

	Simple Coding	**Dummy Coding**			
Sample	**Fuel Type**	**Diesel**	**Gasoline**	**Hybrid**	**Gasoline + LPG**
Diesel	1	1	0	0	0
Gasoline	2	0	1	0	0
Hybrid	3	0	0	1	0
Gasoline +LPG	4	0	0	0	1

Table 4. Coding of brand attribute

	Simple Coding	**Dummy Coding**

Sample	Brand	B1	B2	B3	B4
B1	1	1	0	0	0
B2	2	0	1	0	0
B3	3	0	0	1	0
B4	4	0	0	0	1

The attributes of "Horsepower", "Engine Capacity" and "Model Year" have numerical values at different scales. Minimum and maximum scaling have been applied to these three headings using the calculation method given in Eq. 1 to scale all numerical expressions into the same range [13]. With this method, all numerical values are scaled to the range of 0 and 1, and it is aimed that heuristic algorithms will be affected by the disproportionateness created by numerical expressions at different scales and to obtain more accurate results.

Eq. 1

$$N(X_i) = \frac{X_i - min_X}{maks_x - min_x}$$

While X_i given in the equation is the i. sample of the X attribute to be scaled, min_X and $maks_x$ represent the lowest and the highest value in the same attribute. The final state of the dataset, which is formed after the compilation, cleanup, coding, and scaling operations on the data, is given in Table 2.5.

Table 5. Samples from the final dataset

Price	Horsepower	Engine Capacity	Model Year	Manuel	Auto.	Semi-Auto	Gasoline	Diesel	Hybrid	Gasoline+ LPG
167500	0,213	0,239	0,889	0	0	1	0	1	0	0
159500	0,232	0,298	0,933	1	0	0	1	0	0	0
136000	0,137	0,240	0,867	1	0	0	0	1	0	0
179000	0,137	0,240	0,933	0	0	1	0	1	0	0
60500	0,053	0,212	0,622	1	0	0	0	1	0	0

2.2. Heuristic Algorithm

2.2.1 Linear Regression

There is a cause-and-effect relationship in many events that occur in nature. The determination and interpretation of this relationship have been the subject of many types of research. Regression analysis is a statistical analysis method that statistically examines the relationship between two or more variables that have a cause-effect relationship and produces inferences and predictions about that subject with this examination and analysis [14]. With the regression analysis, it is determined whether there is a mathematical and statistical correlation between two or more variables, and if so, it is also important to clarify the rate. Based on this determined ratio, the estimation analysis of the related variable is performed. The accuracy of the estimation results made by regression analysis varies according to the strength of the link between the variables [15]. In linear regression, the effect of the free variable on a single dependent variable can be analyzed, as well as in cases where free variables are more than one [16]. The calculation given in Eq. 2 is used to create the linear regression model.

Eq. 2. $$Y = \propto + \beta X + \varepsilon$$

While Y given in the equation represents the result variable, X refers to the attribute value used for training. ε represents the random error value. The β and $\propto$ values are the coefficients of the model. They are calculated by the methods given in Eq. 3. and Eq. 4., respectively.

Eq. 3. $$\beta = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

Eq. 4. $$\propto = \bar{Y} - \beta\bar{X}$$

2.2.2 Support vector machine

Support Vector Machines (SVM) is one of the algorithms that have a strong place in the literature in solving regression problems [17]. This algorithm, which works with the principle of calculating the most suitable vectors that will divide the samples into two on a plane, can produce predictions by looking at the positions of the samples on this plane, thanks to the vectors it has created. The decision function used during the estimation process of the SVM algorithm is given in Eq. 5.

Eq. 5. $$f(X) = \sum_{i=1}^{n} \alpha_i \, y_i \, \emptyset(X) \, \emptyset(X_i)$$

The $f(X)$ given in the equation represents the decision-making function of the X sample that is desired to be predicted. Where n represents the total number of samples in the dataset, y_i represents the price information of each sample. To optimize the $\alpha_i$coefficient, the Larange function given in Eq. 6. is used. And the constraints are given in Eq. 7.

Eq. 6. $$\max_a L_D = \max_a \left(\sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j \, K(X_i, X_j)\right)$$

Eq. 7. $$0 \le a_i \le C, \qquad \sum_{i=1}^{n} a_i y_i = 0$$

The C given in the equation represents the upper bound coefficient from the SVM parameters.

2.2.3 Artificial neural networks

Most of the heuristic algorithms are inspired by nature and living things. Artificial Neural Networks (ANN) is one of these algorithms. It is based on the idea of simply imitating the communication between the nerve cells of the human brain and the learning events that occur through communication [18-19]. When simulated the ability of the human brain to reveal correlations between attributes and the outputs led many problems such as classification and prediction to be solved by computers as well. As mentioned before, ANN's consist of artificial neurons organized in a certain order. The part where the nerves that receive the information are

processed first is called the input layer. Afterwards, a hidden layer with the number of nerves in direct proportion to the complexity of the information should be created. The neurons in the output layer are used to generate the output information. The neurons produce the information in two steps. First, an aggregate function is used (Eq. 8), in which the information from the neurons in the previous layers is summed, then the resulting sum must be passed through an activation function. The sigmoid function used as the activation function in this study was calculated as given in Eq. 9.

Eq. 8.

$$Net_i = \sum_{j=1}^{n} w_{ij} Ç_j$$

While Net_i calculates the sum, n gives the total number of all the information coming to the i. neuron. w_{ij} refers to the weights between the nerves in the previous layers and the i. neuron, and $Ç_j$ refers to the output signals produced by these neurons.

Eq. 9.

$$Ç_i = \frac{1}{1 + e^{-(Net_i + \beta_i)}}$$

β_i is the bias/correction value of the i. nerve. If the signals produced after the sum function and activation function are in the hidden layers, they are transmitted to the next layers and the same operations are performed again in the next layers. However, if these signals are generated at the output layer, they are accepted as the prediction value or classification result of the network in a traditional ANN algorithm. To increase the success of this prediction or classification result, the same operations are repeated, but beforehand, it is necessary to know the error and score the network has and to optimize the result. The error of each neuron at the output of the network is calculated by the method given in Eq. 10. By using this error, changing the values of weights between the neurons, the backpropagation of the error used for learning is carried out with the calculation method presented in Equation 11.

Eq. 10.

$$\delta_i = Ç_i(1 - Ç_i)(Y_i - Ç_i)$$

The δ_i given in the equation shows the error of the neuron, Y_i shows the real value that is expected to produce, and $Ç_i$ shows the predicted value produced by the ANN.

Eq. 11.

$$\Delta w_{ij}(t) = \gamma \delta_i Ç_j + \alpha \Delta w_{ij}(t-1)$$

On account of training to take place in the ANN algorithm, the weights between the neurons must be updated. Because the memory of the network is kept on these weights. In this context, $\Delta w_{ij}(t)$ shows the change in weight between two nerves, $\Delta w_{ij}(t-1)$ shows the value of the same weight in the previous repetition. γ is the learning coefficient of the network and α is the momentum coefficient.

2.3. Approach

In this section, the approach followed to perform the training of the heuristic algorithms and the prediction of the price using the examples in the dataset created will be explained. Although expert opinion is taken when deciding on the compiled features in the dataset, the actual effect of these features on the price information may vary. For this reason, filtering, selection or moving to

a different dimension (feature extraction) of the existing features in the dataset before starting the training of the heuristic algorithms increases the success rate and reduces the processing cost and makes them work faster. Accordingly, in this study, it has been tried to determine the combination that gives the best results on the data set with different preprocessing steps and different heuristic algorithms.

Fisher Correlation Score, which is a statistical approach, was preferred for filtering due to its simple and fast operation [20]. The Fisher Score of each feature was calculated by the method given in Eq. 12. It was decided by looking at this score whether the features should be used in the training of heuristic algorithms.

Eq. 12.

$$fisher(x_i) = \frac{\sum_{j=1}^{s}(x_{ij} - u_i)^2}{(\sigma_i)^2}$$

While x_i represents the attribute i, s represents the total number of samples, u represents the mean, and σ represents the standard deviation.

For the feature selection process Forward Feature Selection (FFA) algorithm is preferred. This method simply evaluates each attribute on its own and selects the one with the highest success. It then measures the performance again by adding a second attribute. By repeating this forward-looking process, it is aimed to select the repeat feature combination that exhibits the best performance.

Principal Component Analysis (PCA) was used for feature extraction. PCA can be defined as a new feature creation process by looking at the change in the values of the features. This process, which is done by preserving the correlations and change information between the features carried by each other, causes the heuristic algorithm to be used especially in large data sets to work faster. Although the dataset in this study is not large in terms of data size, PCA was preferred to reveal the relationship between the features. Eq. 13 was used to treat the features as a vector and to find the mean, and Eq. 14. to reveal the relationships between the features with the covariance [21].

Eq. 13.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Eq. 14.

$$C_X = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$$

In the equations, X_i represents the vector of each feature, while n gives the total number of features. After these calculations, the Eigen component vectors providing the equation given in Eq. 14. is used to determine the features that best represent the data set.

Eq. 15.

$$C_X V_m = \lambda_m V_m$$

V_m represents the Eigen component vector and λ_m represents the eigenvalues.

2.4. Metrics

Three different criteria were determined to evaluate the performance of the algorithms and preprocessing steps while performing the estimation operations of the samples in the dataset with heuristic algorithms. The first two of them, Mean Absolute Error (MAE), Mean Square Error (MSE) were used to measure the error resulting from the estimation process, while the last criterion, R2 (Consistency Coefficient) was used to calculate the success score of the model. The absolute error of an estimation operation is most basically calculated by taking the difference of the estimated value from the true value into the absolute value. This method is applied to all estimation results and the value obtained when the average is taken gives the MAE of the model (Eq. 16).

Eq. 16.

$$MAE(y,\check{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \check{y}_i)$$

The y given in the equation represents the actual prices,$\check{y}$ represents the price predictions made by the model, and n represents the total number of samples/estimates. The MSE value of the model is calculated by squaring the absolute error, which is different from the MAE (Eq. 17.).

Eq. 17.

$$MSE(y,\check{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \check{y}_i)^2$$

Unlike the other two criteria, the R2 value gives the model a success score instead of calculating error. It is interpreted that the model, which varies between -1 and 1, is more successful as it gets closer to 1, and more unsuccessful as it gets closer to -1 (Eq. 18).

Eq. 18

$$R^2(y,\check{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \check{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The $\bar{y}$ the symbol represents the average of all real price values in the equation.

3. FINDINGS

In this section, firstly, the distribution of the features contained in the dataset was examined, then the results were shared by applying the preprocessing methods mentioned in the approach step. Finally, the new data subsets created as a result of the preprocessing steps were separately estimated with Linear Regression, SVM and ANN. The error and performance values obtained were shared.

The distribution of the continuous features that the dataset has and its relationship with the price is presented in Figure 2. It is seen that the numerical increase in all 3 attributes is directly proportional to the price information.

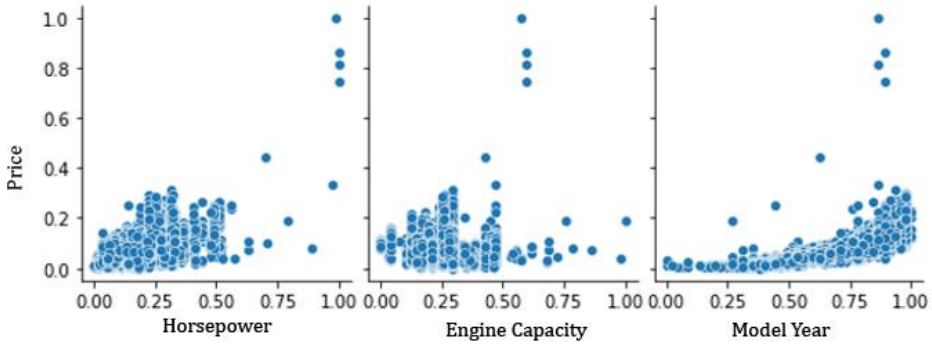

Figure 2. Correlation between horsepower, Engine Capacity, Model Year and Price

The frequencies of discontinuous data in the data set are presented in Figure 3. It is seen that cars with manual gear have the highest frequency, while vehicles with hybrid fuel types have the least frequency.

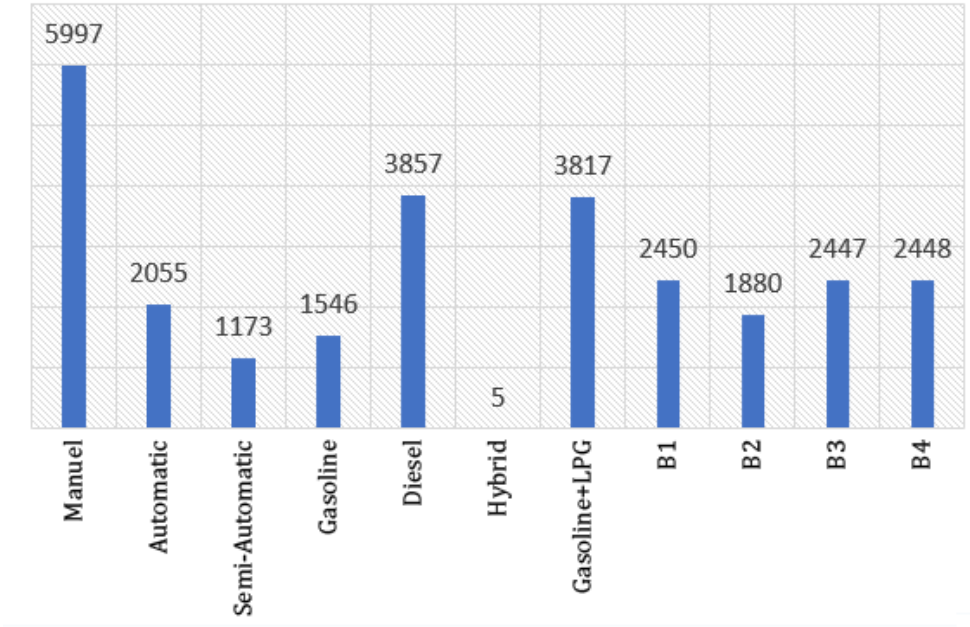

Figure 3. Frequencies

The relationship between the price and the frequency of the attributes in the dataset provides us with a lot of information about the dataset. However, preprocessing steps are needed to understand how important the effect of the change in which attribute combination is on the price. The effect of each attribute on the price was examined with Fisher and the findings are shared in Figure 4.

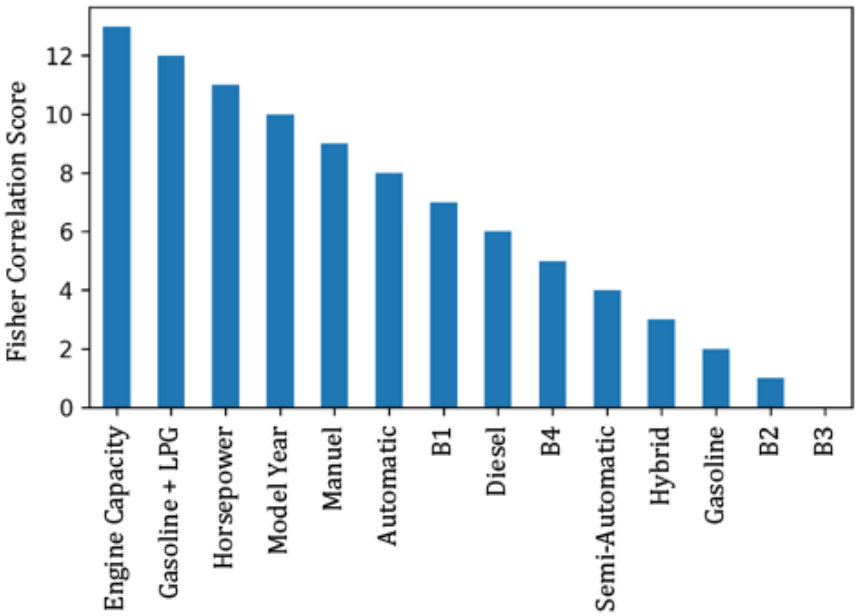

Figure 4. Fisher correlation scores of attributes

When the given graph is examined, Engine Capacity has the highest score according to Fisher Correlation Score, and B3 attribute has the lowest score. Based on this graph, the 5 attributes with the highest score were selected and other attributes were filtered. In this way, it is aimed that heuristic algorithms will produce better results by removing the possible noise created by the parameters that have a low effect on the price.

The results obtained in the feature scoring processes performed with the FFS algorithm, which is another preprocessing method, are given in Table 6. X marks in the table indicate that the relevant attribute has been selected in the relevant step. The score given by the algorithm as a result of this selection is presented at the bottom of the table.

Table 6. Attribute scores obtained FFS

Iteration / Attribute Count	1	2	3	4	5	**6**	7	8	9	10	11	12	13
Horsepower		X	X	X	X	**X**	X	X	X	X	X	X	X
Engine Capacity			X	X	X	**X**	X	X	X	X	X	X	X
Model Year	X	X	X	X	X	**X**	X	X	X	X	X	X	X
Manuel											X	X	X
Automatic												X	X
Semi-automatic					X	**X**	X	X	X	X	X	X	X
Gasoline									X	X	X	X	X
Diesel				X	X	**X**	X	X	X	X	X	X	X
Hybrid							X	X	X	X	X	X	X
Gasoline + LPG								X	X	X	X	X	X
B1										X	X	X	X
B2													X
B3						**X**	X	X	X	X	X	X	X
B4													

Score	0.5651	0.8045	0.817	0.8207	0.8241	**0.8255**	0.8253	0.8250	0.8211	0.8189	0.8047	0.8068	0.8044

There are a total of 14 attributes in the dataset, excluding price information. FFS algorithm has a function to examine which of these features or combinations of these features can yield more successful results. When the table is examined, it is seen that the most successful feature combination is the 6 features selected in Step 6 and the success score of these features is 0.8255. This success rate has never been exceeded with fewer or more features, or with different combinations in the same number of features.

The Eigen components found with the PCA algorithm, which is the pre-processing method and the representation ratios of the dataset/variance of these eigen components are presented in Figure 5. When the graph is examined, it is seen that 8 components represent 99.4% of the data set and this ratio increases slightly in the next components, and it is very low in this representation ratio in the number of fewer components. Therefore, these 8 components were used in the estimation process.

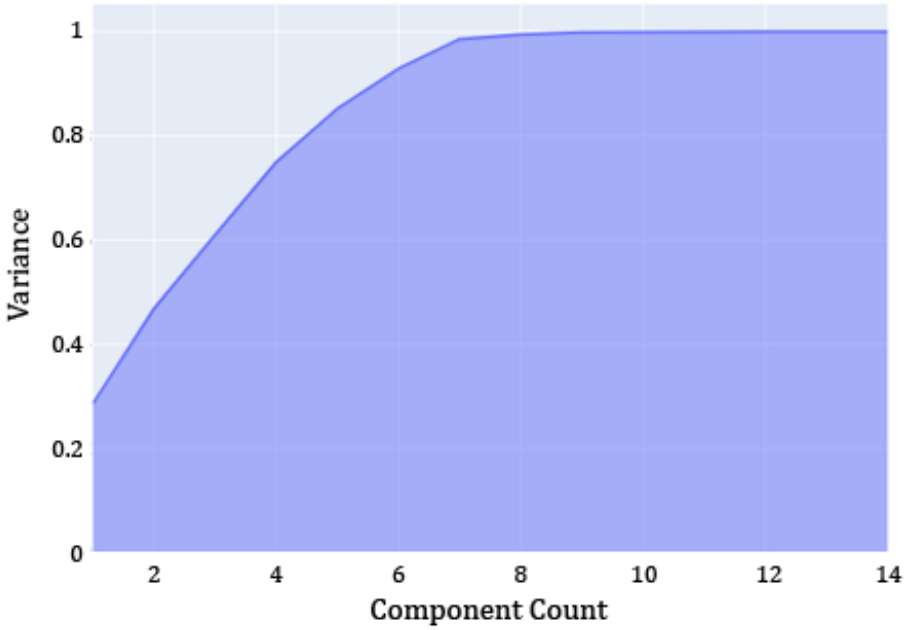

Figure 5. PCA components and Its Variances

After the preprocessing steps, data-subsets with different features were predicted separately by Linear Regression, SVM and ANN algorithms. Before the prediction operations, 80% of the data subsets were used for training the relevant models and 20% for the testing operations. Obtained error and performance values are given in Table 7. for each pre-process and algorithm.

Table 7. Error and Performance values of preprocessing steps and algorithm

Preprocess	Attributes	Algorithm	MAE	MSE	R^2
FCS		Linear Regression	0.01550	0.000660	0.7549

	Engine Capacity, Gasoline + LPG, Horsepower, Model Year, Automatic	SVM	0.01723	0.000474	0.8242
		ANN	**0.01050**	**0.000281**	**0.8958**
FFS	Horsepower, Engine Capacity, Model Year, Semi-Automatic, Diesel, B3	Linear Regression	0.01620	0.000697	0.7413
		SVM	0.01749	0.000539	0.8000
		ANN	0.01171	0.000384	0.8576
PCA	8 Component, Variance: 0.9944	Linear Regression	0.01650	0.000776	0.7122
		SVM	0.01950	0.000748	0.7224
		ANN	0.01189	0.000507	0.8118

When the error and performance table is examined, it is seen that all of the models trained with Fisher-selected features are better than the scores obtained with other features. The best model was Fisher+ANN, as it had both the lowest errors and the highest R2 score. The distance of the predictions made with this model to the actual prices is presented in Figure 6. In addition, the individual absolute error of each prediction is shown on a boxplot in Figure 7.

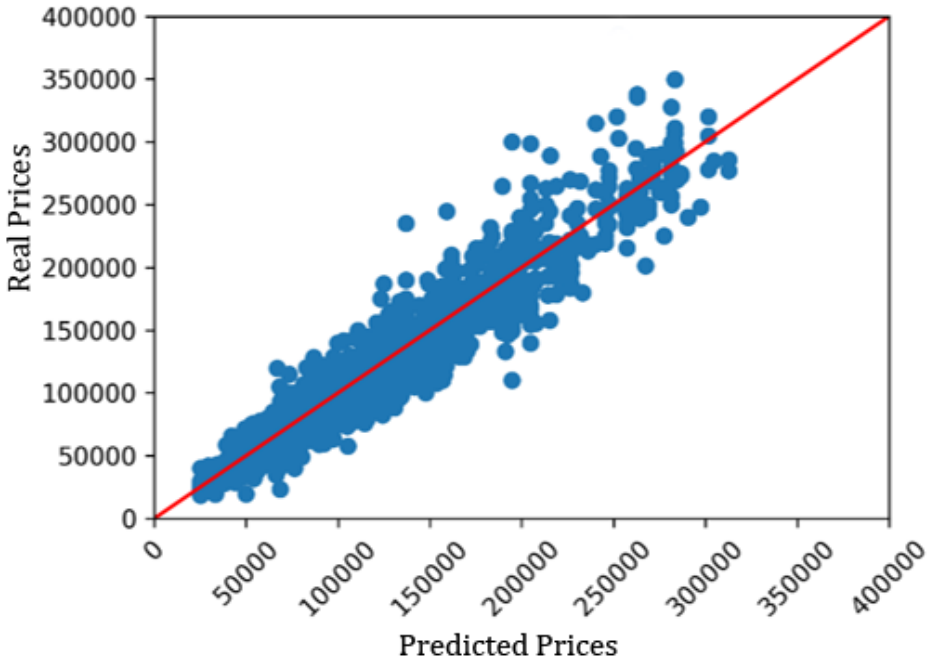

Figure 6. Scatter plot of Predicted and Real Prices

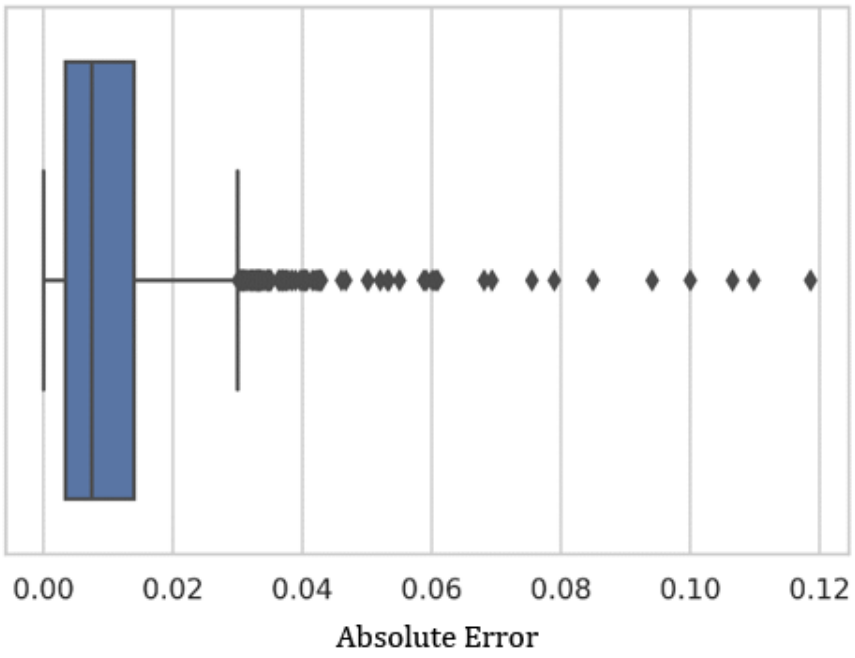

Figure 7. Absolute error values of each prediction

4. DISCUSSION

When the results obtained in the research findings are examined, it is understood that the best features for the created dataset are determined by the Fisher algorithm, and the best prediction is performed by the ANN algorithm using the data subset created from these features. However, ANN successfully passed the SVM and Linear Regression models in all other preprocessing methods. These results show us that the prediction algorithm used should be supported by a good feature selection/filtering method to achieve the best performance. The same algorithms can obtain different performance values in different data sets. For this reason, the most appropriate preprocessing step and estimation algorithm should be determined according to the structure of the dataset.

5. CONCLUSION

In this study, a new predictable dataset was created that can be used in training heuristic algorithms. The most important headings that affect second-hand car prices are included in this dataset, which is formed by the compilation of used vehicle sales advertisements on the Internet, in line with expert opinions. Among these headings, Engine Capacity, Gasoline + LPG fuel type, Horsepower, Model Year, Manual gear type are determined as the most important factors affecting the price of a vehicle.

As a result of the prediction processes using different preprocessing steps and different prediction algorithms, the Fisher+ANN model achieved the best performance with MAE 0.01050, MSE 0.000281 error and R2 0.8958 performance value. In all cases where different preprocessing methods are used, ANN surpasses other algorithms.

As a result, a new data set suitable for prediction processes was created and shared with heuristic algorithms that researchers can work on. It was seen that the data set could be predicted successfully. But, changes in car prices in short periods under volatile market conditions will cause

these data to become outdated. For this reason, for the models trained with the shared dataset to give better results, the price information should be updated at certain time intervals.

In future studies, it is planned to turn the model trained for used car price prediction into a service and make it available to people who buy and sell these. In addition, it is aimed to measure how the trained models are affected by the changes in the used car market and to develop a method to tolerate this change.

REFERENCES

[1] Lessman, S., & Vob, S. (2017). Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy. *International Journal of Forecasting, 33*(4), 864-877.

[2] Due, J., Xie, L., & Schroeder, S. (2009). PIN optimal distribution of auction vehicle system: Applying price forecasting, elasticity estimation and genetic algorithms to used vehicle distribution. *Marketing Science, 28*, 637-644.

[3] Lessmann, S., Kistiani, M., & Vob, S. (2010). Decision support in car leasing: a forecasting model for residual value estimation. *Proceedings of the international conference on information systems*, (s. 17). Saint Louis.

[4] Murray, J., & Sarantis, N. (1999). Price-quality relations and hedonic price indexes for cars in the United Kingdom. *International Journal of the Economics of Business, 6*(21), 5-27.

[5] Dastan, H. (2016). Türkiye'de İkinci El Otomobil Fiyatlarını Etkileyen Faktörlerin Hedonik Fiyat Modeli ile Belirlenmesi. *Gazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 18*(1), 303-327.

[6] Cumhur, E., & Şentürk, İ. (2009). A Hedonic Analysis of Used Car Prices in Turkey. *International Journal of Economic Perspectives, 3*(2), 141-149.

[7] Çelik, Ö., & Osmanoğlu, U. Ö. (2019). Prediction of The Prices of Second-Hand Cars. *Avrupa Bilim ve Teknoloji Dergisi, 16*, 77-83.

[8] Özçalıcı, M. (2017). Predicting Second-Hand Car Sales Price Using Decision Trees and Genetic Algorithms. *The Journal of Operations Research, Statistics, Econometrics and Management Information Systems, 5*(1), 103-114.

[9] Noor, K., & Jan, S. (2017). Vehicle price prediction system using machine learning techniques. *International Journal of Computer Applications, 167*(9), 27-31.

[10] Pal, N., Arora, P., Kohli, P., Sundararaman, D., & Palakurthy, S. S. (2018). How Much Is My Car Worth? A Methodology for Predicting Used Cars' Prices Using Random Forest. *In Future of Information and Communication Conference*, (s. 413-422). Cham.

[11] Demissie, S., LaValley, M. P., & Horton, N. J. (2003). Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. *Stat Med, 22*, 545-547.

[12] UCI Irvine, Machine Learning Repository, https://archive-beta.ics.uci.edu/ml/datasets

[13] Jain, S., Shukla, S., & Wadhvani, R. (106). Dynamic selection of normalization techniques using data complexity measures. *Expert Systems with Applications, 106*, 252-262.

[14] Çevik, H. (2004). Türkiyenin kısa dönem elektrik yük tahmini. *İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi*, 5-31.

[15] Balcı, H., Esener, İ., & Kurban, M. (2012). Regresyon analizi kullanarak kısa dönem yük tahmini. *Eleco Elektrik-Elektronik ve Bilgisayar Mühendisliği Sempozyumu*, (s. 796-797). Bursa.

[16] Tekin, H. (2019). Yeni bir metot olan geri beslemeli lineer regresyon ile akıllı şebekeye bağlı meskenlerde kısa dönem yük tahmini. *Batman Üniversitesi Fen Bilimleri Enstitüsü*, 53.

[17] Cortes, C., & Vapnik, V. N. (1995). Support-Vector Networks. *Machine Learning, 20*(3), 273-297.

[18] Blanton, H. (1997). An Introduction to Neural Networks for Technicians, Engineers and Other non PhDs. *Proceedings of 1997 Artificial Neural Networks in Engineering Conference.* St. Louis.

[19] McCulloch, W. S., & Pitts, A. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematics and Biophysics, 5*, 115-133.

[20] Xiong, M., Fang, X., & Zhao, J. (2001). Biomarker identification by feature wrappers. *Genome Res, 11*(11), 1878-1887.

[21] Adiwijaya, Wisesty, U. N., Lisnawati, E., Aditsania, A., & Kusumo, D. S. (2018). Dimensionality Reduction using Principal Component Analysis for Cancer Detection Based on Microarray Data Classification. *Journal of Computer Science, 14*(11), 1521-1530.