International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi :https://doi.org/10.32628/IJSRST229149

Car Acceptability Prediction System Using Machine Learning

Shaikh Md Azhar Md Hamid
M.E. Data Science, Department of Computer Engineering, Zeal College of Engineering and Research, Narhe, Savitribai Phule Pune University, Pune, Maharashtra, India

Article Info
Volume9, Issue 2
Page Number: 545-551

Publication Issue
March-April-2022

Article History
Accepted :03April2022
Published :20April2022

ABSTRACT
A car price prediction has been a high-interest research area, as it requires noticeable effort and knowledge of the field expert. Considerable number of distinct attributes are examined for the reliable and accurate prediction. To build a model for predicting the price of used cars in Bosnia and Herzegovina, we applied three machine learning techniques (Artificial Neural Network, Support Vector Machine and Random Forest). However, the mentioned techniques were applied to work as an ensemble. The data used for the prediction was collected from the web portal autopijaca.ba using web scraper that was written in PHP programming language.
Respective performances of different algorithms were then compared to find one that best suits the available data set. The final prediction model was integrated into Java application. Furthermore, the model was evaluated using test data and the accuracy of 87.38% was obtained.

Keywords – car price prediction, support vector machines, classification, machine learning

I. INTRODUCTION

Car price prediction is somehow interesting and popular problem. As per information that was gotten from the Agency for Statistics of BiH, 921.456 vehicles were registered in 2014 from which 84% of them are cars for personal usage [1]. This number is increased by 2.7% since 2013 and it is likely that this trend will continue, and the number of cars will increase in future. This adds additional significance to the problem of the car price prediction.

Accurate car price prediction involves expert knowledge, because price usually depends on many distinctive features and factors. Typically, most significant ones are brand and model, age, horsepower and mileage. The fuel type used in the car as well as fuel consumption per mile highly affect price of a car due to a frequent changes in the price of a fuel. Different features like exterior colour, door number, type of transmission, dimensions, safety, air condition, interior, whether it has navigation or not will also influence the car price. In this paper, we applied different methods and techniques in order to achieve higher precision of the used car price prediction.

This paper is organized in the following manner:
Section II contains related work in the field of price prediction of used cars. In section III, the research methodology of our study is explain. Section IV

elaborates various machine learning algorithms and examine their respective performances to predict the price of the used cars. Finally, in section V, a conclusion of our work are given, together with the future works plan.

II. RELATED WORK

Predicting price of a used cars has been studied extensively in various researches. Listian discussed, in her paper written for Master thesis [2], that regression model that was built using Support Vector Machines (SVM) can predict the price of a car that has been leased with better precision than multivariate regression or some simple multiple regression. This is on the grounds that Support Vector Machine (SVM) is better in dealing with datasets with more dimensions and it is less prone to overfitting and underfitting. The weakness of this research is that a change of simple regression with more advanced SVM regression was not shown in basic indicators like mean, variance or standard deviation.

Another approach was given by Richardson in his thesis work [3]. His theory was that car producers produce more durable cars.

Richardson applied multiple regression analysis and demonstrated that hybrid cars retain their value for longer time than traditional cars. This has roots in environmental concerns about the climate and it gives higher fuel efficiency.

Wu et al. [4] conducted car price prediction study, by using neuro-fuzzy knowledge-based system. They took into consideration the following attributes: brand, year of production and type of engine. Their prediction model produced similar results as the simple regression model. Moreover, they made an expert system named ODAV (Optimal Distribution of Auction Vehicles) as there is a high demand for selling the cars at the end of the leasing year by car dealers. This system gives insights into the best prices for vehicles, as well as the location where the best price can be gained. Regression model based on k-nearest neighbour machine learning algorithm was used to predict the price of a car. This system has a tendency to be exceptionally successful since more than two million vehicles were exchanged through it [5].

Gonggie [6] proposed a model that is built using ANN (Artificial Neural Networks) for the price prediction of a used car. He considered several attributes: miles passed, estimated car life and brand. The proposed model was built so it could deal with nonlinear relations in data which was not the case with previous models that were utilizing the simple linear regression techniques. The non-linear model was able to predict prices of cars with better precision than other linear models.

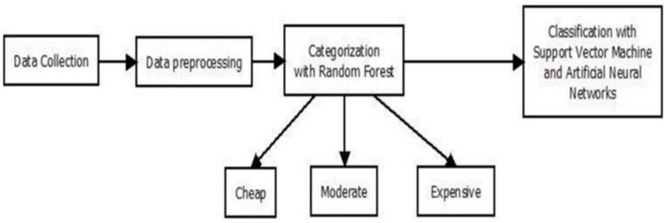
Furthermore, Pudaruth [7] applied various machine learning algorithms, namely: k-nearest neighbours, multiple linear regression analysis, decision trees and naïve bayes for car price prediction in Mauritius. The dataset used to create a prediction model was collected manually from local newspapers in period less than one month, as time can have a noticeable impact on price of the car. He studied the following attributes: brand, model, cubic capacity, mileage in kilometres, production year, exterior colour, transmission type and price. However, the author found out that Naive Bayes and Decision Tree were unable to predict and classify numeric values. Additionally, limited number of dataset instances could not give high classification performances, i.e. accuracies less than 70%.

Noor and Jan [8] build a model for car price prediction by using multiple linear regression. The dataset was created during the two- months period and included the following features: price, cubic capacity, exterior colour, date when the ad was posted, number of ad views, power steering, mileage in kilometre, rims type, type of transmission, engine type, city, registered city, model, version, make and model year. After applying feature selection, the authors considered only engine type, price, model year and model as input features. With the given setup authors were able to achieve prediction accuracy of 98%.

In the related work shown above, authors proposed prediction model based on the single machine learning algorithm. However, it is noticeable that single machine learning algorithm approach did not give remarkable prediction results and could be enhanced by assembling various machine learning methods in an ensemble.

III. MATERIALS AND METHODS

Approach for car price prediction proposed in this paper is composed of several steps, shown in Fig. 1.


Data Collection
Data preprocessing
Categorization with Random Forest
Classification with Support Vector Machine and Artificial Neural Networks
Cheap
Moderate
Expensive

Data is collected from a local web portal for selling and buying cars autopijaca.ba [9], during winter season, as time interval itself has high impact on the price of the cars in Bosnia and Herzegovina. The following attributes were captured for each car: brand, model, car condition, fuel, year of manufacturing, power in kilowatts, transmission type, millage, colour, city, state, number of doors, four wheel drive (yes/no), damaged (yes/no), navigation (yes/no), leather seats (yes/no), alarm (yes/no), aluminium rims (yes/no), digital air condition (yes/no), parking sensors (yes/no), xenon lights (yes/no), remote unlock (yes/no), electric rear mirrors (yes/no), seat heat (yes/no), panorama roof (yes/no), cruise control (yes/no), abs (yes/no), esp (yes/no), asr (yes/no) and price expressed in BAM (BosnianMark).

Since manual data collection is time consuming task, especially when there are numerous records to process, a "web scraper" as a part of this research is created to get this job done automatically and reduce the time for data gathering. Web scraping is well known technique to extract information from websites and save data into local file or database. Manual data extraction is time consuming and therefore web scrapers are used to do this job in a fraction of time. Web scrapers are programmed for specific websites and can mimic regular users from website's point of view.

After raw data has been collected and stored to local database, data pre-processing step was applied. Many of the attributes were sparse and they do not contain useful information for prediction. Hence, it is decided to remove them from the dataset. The attributes "state", "city", and "damaged" were completely removed.

The collected raw data set contains 1105 samples. Since data is collected using web scraper, there are many samples that have only few attributes. In order to clean these samples, PHP script that is reading scraped data from database, perform cleaning and saves the cleaned samples into CSV file. The CSV file is later used to load data into WEKA, software for building machine learning models [10].

After clean-up process, the data set has been reduced to 797 samples. In particular, all brands that have less than 10 samples and where the price is higher than 60 000 BAM were removed due to the

skew class problem. The whole dataset creation process is shown in the Fig. 2

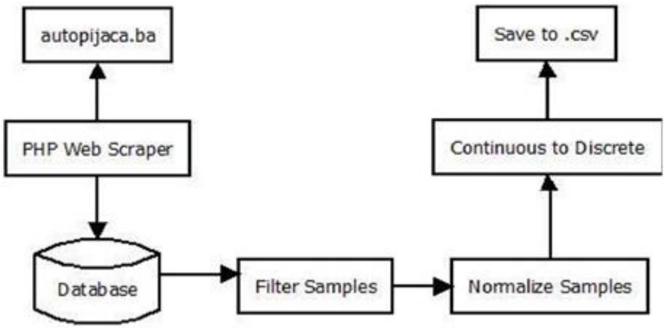
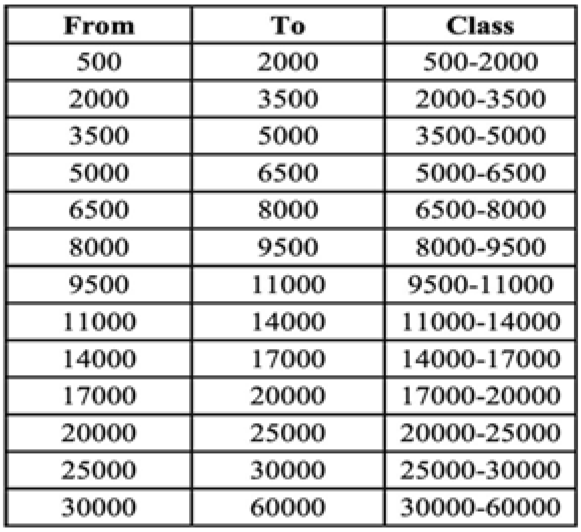


Figure 2. Data gathering and transformation workflow diagram

The colour of the cars was normalized into fixed set of 15 different colours. Continuous attributes such as "millage", "year of manufacturing", "power in kilowatts" and "price" are converted into categorical values using predefined cluster intervals. The millage is converted into five distinct categories, the year of manufacturing has been converted into seven categories and the power in kilowatts is converted into eleven categories. The price attribute has been categorized into 15 distinct categories based on price range. These categories are shown in Table 2 and similar principle was applied to other attributes. This data transformation process converted regression prediction machine learning problem into classification problem.

Table 2. Price classification based on price ranges



From	To	Class
500	2000	500-2000
2000	3500	2000-3500
3500	5000	3500-5000
5000	6500	5000-6500
6500	8000	6500-8000
8000	9500	8000-9500
9500	11000	9500-11000
11000	14000	11000-14000
14000	17000	14000-17000
17000	20000	17000-20000
20000	25000	20000-25000
25000	30000	25000-30000
30000	60000	30000-60000

IV. MODEL IMPLEMENTATION AND EVALUATION

Single machine learning classifier approach that has been used in all previous researches was also tested in this research. The whole data set collected in this research has been split into training (90%) and testing (10%) subsets and Artificial Neural Network, Support Vector Machine and Random Forest classifiers models were built.

Random forest (RF) also known as random decision forest belongs to the category of ensemble methods. RF can be used for classification and regression problems. The algorithm was developed by Ho as an improvement for overfitting of the decision tree algorithms [11]. Artificial Neural Networks is the machine learning model that tries to solve problems in the same way as the human brain does. Instead of neurons, the ANN is using artificial neurons also known as perceptron. In the human brain, neurons are connected with axons while in ANN the weighted matrices are used for connections between artificial neurons.

Information travels through neurons using connections between them, from one neuron information travels to all the neurons connected to it. Adjusting the weights between neurons system can be trained from input examples [12]. Support Vector Machine can be used for solving classification and regression problems. For input data set, the SVM can make a binary decision and decide in which among the two categories the input sample belongs. The SVM algorithm is trained to label input data into two categories that are divided by the widest area possible between categories [12]. In cases when input data is not labelled, SVM algorithm cannot be applied. For unlabelled data, it is necessary to apply unsupervised learning method and SVM has its implementation called Support Vector Clustering (SVC) [13][14].

Table 3. Single classifier approach accuracy results

Classifier	Accuracy	Error
RF	41.18%	8.04%
ANN	42.35%	7.05%
SVM	48.23%	10.55%

Results shown in Table 3. confirm that single machine learning classifier approach is not reliable for prediction of car prices. Therefore, in this paper ensemble method for car prices prediction was proposed. To apply ensemble of machine learning classifiers a new attribute “price rank” with values: cheap, moderate and expensive has been added to the data set. This attribute divides cars into three price categories: cheap (price < 12 000 BAM), moderate (12 000 BAM <= price < 24 000 BAM) and expensive (24 000 BAM <= price).

Ensemble method combines three machine learning algorithms that were applied in the first experiment as single classifiers: RF, SVM, and ANN.

Random Forest algorithm was applied on the whole dataset, to test how accurately the classifier can categorize samples into cheap, moderate and expensive car classes. RF is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over- fitting [15].

The following features were used to build model: brand, model, car condition, fuel, age, kilowatts, transmission, miles, color, doors, drive, leather seats, navigation, alarm, aluminium rims, digital AC, manual AC, parking sensors, xenon, remote unlock, seat heat, panorama roof, cruise control, abs, asr, espand price.

Before model training step, numeric attribute price was converted into nominal classes shown in Table 4.

Table 4. Nominal categories of car price attribute

From	To	Class
0	12000	Cheap
12000	24000	Moderate
24000	…..	Expensive

Then, RF classifier is applied, and results are obtained (Table 5.).

Table 5. Classification results with RF classifier

Type of evaluation	% Of correctly classified
Cross validation with 10 folds	85.82
90% percentage split	88.75

Both classifiers, SVM and ANN are further applied to each price category dataset: cheap, moderate and expensive cars datasets.

Applying classification on cheap dataset using SVM and ANN algorithms

Cheap dataset was divided into 2 nominal classes, shown in Table 6.

Table 6. Nominal classes in Cheap dataset

From	To	Class
0	6000	0-6000
6000	12000	6000-12000

In total, 230 samples of Cheap dataset were input to SVM and ANN algorithms.

After running SVM and ANN on given dataset, following results were obtained:

Table 7. Accuracy results for SVM and ANN on Cheap dataset

Type of evaluation	SVM	ANN
Cross Validation with 10 folds	86.96	83.91
90% percentage split	86.96	73.91

Applying Classification on Moderate dataset using SVM and ANN algorithms

The model is further trained on the Moderate dataset. For this purpose, attribute price is ranked into 2 classes, shown in Table 8.

Table 8. Nominal classes in Moderate dataset

From	To	Class
12000	15000	12000-18000
18000	21000	18000-24000

After applying Multilayer Perceptron algorithm on dataset, we got the following results.

Table 9. Accuracy results for SVM and ANN on Moderate dataset

Type of evaluation	SVM	ANN
Cross Validation with 10 folds	78.65	76.41
90% percentage split	83.33	86.11

Applying Classification on Expensive dataset using SVM algorithm

As for the previous datasets, the model is trained on the Expensive dataset. For this purpose, the attribute price is grouped into 2 classes.

Table 10. Nominal classes for Expensive dataset

From	To	Class
24000	28000	24000-32000
32000	36000	32000- …..

SVM and ANN algorithms are further applied to Expensive dataset and results are obtained

Table 11. Accuracy results for SVM and ANN on Expensive dataset

Type of evaluation	SVM	ANN
Cross Validation with 10 folds	79.72	75
90% percentage split	90.48	85.71

After models are built, they have been assembled into the final prediction system, shown in Fig. 3. For the case of 90% dataset split, SVM achieved the highest accuracy in Cheap and Expensive subsets, while ANN performed better in Moderate subset

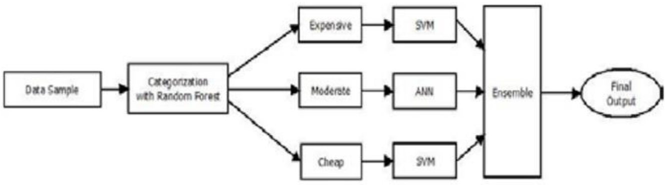
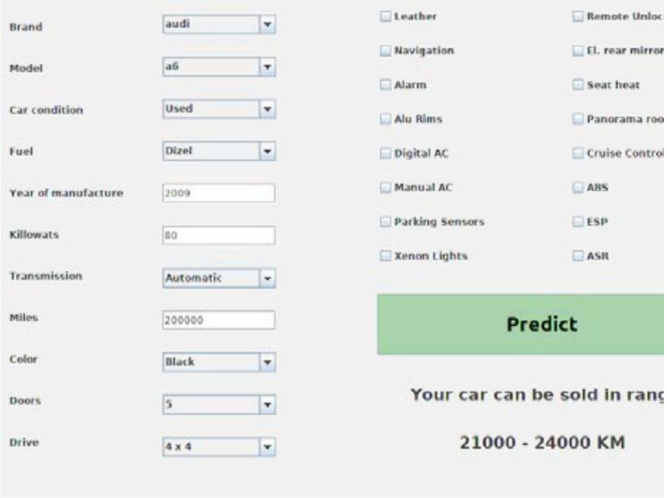


Figure 3. Prediction model for 90% split case

The final prediction system has been incorporated into the Java swing GUI application for the car price prediction. The simple application GUI, shown in Fig. 4. enables potential car buyers to estimate the price of the desired car.

The proposed prediction model has been evaluated on the test subset and model achieved overall accuracy of 87.38%. This proves that combination of multiple machine learning classifiers strengthens the classification performance overall.



Brand	audi
Model	a6
Car condition	Used
Fuel	Dizel
Year of manufacture	2009
Killowats	80
Transmission	Automatic
Miles	200000
Color	Black
Doors	5
Drive	4 x 4

□ Leather
□ Navigation
□ Alarm
□ Alu Rims
□ Digital AC
□ Manual AC
□ Parking Sensors
□ Xenon Lights
□ Remote Unloc
□ El. rear mirror
□ Seat heat
□ Panorama roo
□ Cruise Control
□ ABS
□ ESP
□ ASR

Your car can be sold in rang

21000 - 24000 KM

Figure 4. Graphical user interface of the Java applicat for car price prediction

V. CONCLUSION

Car price prediction can be a challenging task due to the high number of attributes that should be considered for the accurate prediction. The major step in the prediction process is collection and pre-processing of the data. In this research, PHP scripts were built to normalize, standardize and clean data to avoid unnecessary noise for machine learning algorithms.

Data cleaning is one of the processes that increases prediction performance, yet insufficient for the cases of complex data sets as the one in this research. Applying single machine algorithm on the data set accuracy was less than 50%. Therefore, the ensemble of multiple machine learning algorithms has been proposed and this combination of ML methods gains accuracy of 92.38%. This is significant improvement compared to single machine learning method approach. However, the drawback of the proposed system is that it consumes much more computational resources than single machine learning algorithm.

Although, this system has achieved astonishing performance in car price prediction problem our aim for the future research is to test this system to work successfully with various data sets. We will extend our test data with eBay [16] and OLX [17] used cars data sets and validate the proposed approach.

VI. REFERENCES

[1]. Agencija za statistiku BiH. (n.d.), retrieved from: http://www.bhas.ba . [accessed July 18, 2018.]

[2]. Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application (Doctoral dissertation, Master thesis, TU Hamburg-Harburg).

[3]. Richardson, M. S. (2009). Determinants of used car resale value. Retrieved from: https://digitalcc.coloradocollege.edu/islandora/object/coccc%3A1346 [accessed: August 1, 2018.]

[4]. Wu, J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications, 36(4), 7809-7817.

[5]. Du, J., Xie, L., & Schroeder, S. (2009). Practice Prize Paper—PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution. Marketing Science, 28(4), 637-644.

[6]. Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on (Vol. 2, pp. 682-685). IEEE.

[7]. Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764.

[8]. Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 167(9), 27-31.

[9]. Auto pijaca BiH. (n.d.), Retrieved from: https://www.autopijaca.ba. [accessed August10, 2018].

[10]. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (n.d.), Retrieved from: https://www.cs.waikato.ac.nz/ml/weka/. [August 04, 2018].

[11]. Ho, T. K. (1995, August). Random decision forests. In Document analysis and recognition, 1995., proceedings of the third international conference on (Vol. 1, pp. 278-282). IEEE.

[12]. Russell, S. (2015). Artificial Intelligence: A Modern Approach (3rd edition). PE.

[13]. Ben-Hur, A., Horn, D., Siegelmann, H. T., & Vapnik, V. (2001). Support vector clustering. Journal of machine learning research, 2(Dec), 125-137.

[14]. Aizerman, M. A. (1964). Theoretical foundations of the potential function method in pattern recognition learning. Automation and remote control, 25, 821-837.

[15]. 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.19.2 documentation. (n.d.). Retrieved from: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html [accessed: August 30, 2018].

[16]. Used cars database. (n.d.) Retrieved from: https://www.kaggle.com/orgesleka/used-cars-database. [accessed: June 04, 2018].

[17]. OLX. (n.d.), Retrieved from: https://olx.ba.[accessed August 05,2018].