Car Resale Value Prediction Using Machine Learning Algorithm

A PROJECT REPORT

Submitted by

DHARSHINI L	**(19205007)**
DIVYADHARSHINI S	**(19205009)**
KARTHIKA M	**(19205019)**
KEERTHANA S	**(19205021)**

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in INFORMATION TECHNOLOGY

ERODE SENGUNTHAR ENGINEERING COLLEGE

(An Autonomous Institution)

PERUNDURAI, ERODE – 57.

ANNA UNIVERSITY: CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report **“CAR RESALE VALUE PREDICTION USING MACHINELAERNING ALGORITHMS”** is the bonafide work of **DHARSHINI.L(19205007),DIVYADHARSHINI.S(19205009),KARTHIKA.M (19205019),KEERTHANA.S (19205021)**,, who carried out the project work under my supervision.

SIGNATURE

URESH Dr.M.P.THIRUVENKATASURESH

HEAD OF THE DEPARTMENT

Professor and Head,

Department of Information Technology

Erode Sengunthar Engineering College,

Erode-638 057

SIGNATURE

Dr.M.P.THIRUVENKATASURESH

SUPERVISOR

Professor and Head,

Department of Information Technology

Erode Sengunthar Engineering College,

Erode-638 057

Submitted for the End Semester Viva Voce examination held on_________ for **19IT706– Project Phase 1** during the academic year 2021-2022

INTERNAL EXAMINER

EXTERNAL EXAMINER

Abstract:

The production of cars has been steadily increasing in the past decade, with over 70 million passenger cars being produced in the year 2016. This has given rise to the used car market, which on its own has become a booming industry. The recent advent of online portals has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of a used car in the market. Using Machine Learning Algorithms such as Lasso Regression, Multiple Regression and Regression trees, we will try to develop a statistical model which will be able to predict the price of a used car, based on previous consumer data and a given set of features. We will also be comparing the prediction accuracy of these models to determine the optimal one.

Table of contents:

Literature review:5

Literature review:6

Literature review:7

Literature review:8

Literature review:9

Literature review:10

CHAPTER 1
Introduction

The used car market is an ever-rising industry, which has almost doubled its market value in the last few years. The emergence of online portals such as CarDheko, Quikr, Carwale, Cars24, and many others has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of the used car in the market. Machine Learning algorithms can be used to predict the retail value of a car, based on a certain set of features. Different websites have different algorithms to generate the retail price of the used cars, and hence there isn't a unified algorithm for determining the price. By training statistical models for predicting the prices, one can easily get a rough estimate of the price without actually entering the details into the desired website The main objective of this paper is to use three different prediction models to predict the retail price of a used car and compare their levels of accuracy.

CHAPTER 2

Literature Review:1

Title:Used Cars Price Prediction using Supervised Learning Techniques

Abstract: The production of cars has been steadily increasing in the past decade, with over 70 million passenger cars being produced in the year 2016. This has given rise to the used car market, which on its own has become a booming industry. The recent advent of online portals has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of a used car in the market. Using Machine Learning Algorithms such as Lasso Regression, Multiple Regression and Regression trees, we will try to develop a statistical model which will be able to predict the price of a used car, based on previous consumer data and a given set of features. We will also be comparing the prediction accuracy of these models to determine the optimal one.

Keywords: ANOVA, Lasso Regression, Regression Tree, Tukey's Test

Limitations: parameter estimates of the 67 levels are tabulated here Since Lasso regression heavily relies on the training set to find the best fit levels of attributes, it might miss out on some levels of categorical variables which do not show much association in the trai The ning dataset, due to random sampling.

This might cause our model to be slightly (maybe even statistically insignificant) underfit, since in-group variance might have been overlooked. Hence, an iterative process is needed to determine the mean error rate.

Benefits: The result of the GLM procedure with P-value and R2 values are tabulated along with the type 1 and type 3 error rates. From this model, we can see that the variable Price and the selected variables are highly correlated since the R-Square (coefficient of determination) value is around 0.9927. This implies that these variables account for about 99.27% of the variance in the Price. Moreover, both Type 1 and Type 3 SS tables show us that all the variables are significantly correlated with Price (P values < 0.05), except Cruise control, which is confounded when the other variables are held at their mean. Similar to the GLM Select procedure, this procedure also returns a set of parameter estimates, for numerical variables and every level of the categorical variables.

Methodology: We made sure that all the data was collected in less than one month interval as time itself could have an appreciable impact on the price of cars. In Mauritius, seasonal patterns is not really a problem as this does not really affect the purchase or selling of cars. The following data was collected for each car: make, model, volume of cylinder (funnily this is usually considered same as horsepower in Mauritius), mileage in km, year of manufacture, paint colour, manual/automatic and price. Only cars which had their price listed were recorded. Thus, paint colour and manual/automatic features were removed. The data was then further tweaked to remove records in which either the age (year) or the cylinder volume was not available. Model was also removed as it would have been extremely difficult to get enough records for all the variety of car models that exist. Although data for mileage was sparse, it was kept as it is considered to be a key factor in determining the price of used cars .

Literature Review:2

Title:USED CAR PRICE PREDICTION

Abstract: The price of a new car in the industry is fixed by the manufacturer with some additional costs incurred by theGovernment in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But,due to the increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. Existing System includes a process where a seller decides a price randomly and buyer has no idea about the car and it's value in the present day scenario. In fact, seller also has no idea about the car's existing value or the price he should be selling the car at. To overcome this problem we have developed a model which will be highly effective. Regression Algorithms are used because they provide us with continuous value as an output and not a categorized value. Because of which it will be possible to predict the actual price a car rather than the price range of a car. User Interface has also been developed which acquires input from any user and displays the Price of a car according to user's inputs.

Key Words: Linear Regression, Used car Prediction, Ridge Regression, Lasso Regression, Decision Tree Regressor

Methodology: There are two primary phases in the system: 1. Training phase: The system is trained by using the data in the data set and fits a model (line/curve) based on the algorithm chosen accordingly. 2. Testing phase: the system is provided with the inputs and is tested for its working. The accuracy is checked. And therefore, the data that is used to train the model or test it, has to be appropriate. The system is designed to detect and predict price of used car and hence appropriate algorithms must be used to do the two different tasks. Before the algorithms are selected for further use, different algorithms were compared for its accuracy. The well-suited one for the task was chosen.

Limitations: In future this machine learning model may bind with various website which can provide real time data for price prediction.Also we may add large historical data of car price which can help to improve accuracy of the machine learning model. We can build an android app as user interface for interacting with user. For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.

Benefits: The increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. The proposed system will help to determine the accurate price of used car price prediction. This paper compares 3 different algorithms for machine learning : Linear Regression, Lasso Regression and Ridge Regression

Literature Review : 3

Title:Used Cars Price Prediction using Data Mining TechniquesTechniques

Abstract: Due to the unprecedented number of cars being purchased and sold, used car price prediction is a topic of high interest. Because of the affordability of used cars in developing countries, people tend more purchase used cars. A primary objective of this project is to estimate used car prices by using attributes that are highly correlated with a label (Price). To accomplish this, data mining technology has been employed. Null, redundant, and missing values were removed from the dataset during preprocessing. In this supervised learning study, three regressors (Random Forest Regressor, Linear Regression, and Bagging Regressor) have been trained, tested, and compared against a benchmark dataset. Among all the experiments, the Random Forest Regressor had the highest score at 95%, followed by 0.025 MSE, 0.0008 MAE, and 0.0378 RMSE respectively. In addition to Random Forest Regression, Bagging Regression performed well with an 88% score, followed by Linear Regression having an 85% mark. A train-test split of 80/20 with 40 random states was used in all experiments. The researchers of this project anticipate that in the near future, the most sophisticated algorithm is used for making predictions, and then the model will be integrated into a mobile app or web page for the general public to use.

Keywords: Car Price Prediction, supervised learning, linear regression, bagging regression, classification

Methodology: The project deals with UAE used cars. Using Parse Hub, the benchmark dataset from dubizzle.ae and buyanycar.com was scraped in order to build the effective intelligent model. The project's methodology is as follows: any discrepancies in the units, as well as removing attributes that doesn’t affect the price evaluations if needed to reduce the complexity of the model. Data Understanding and preparation is an essential part of building a model as it gives the insight into the data and what corrections or modifications shall be done before designing and executing the model, preliminary analysis of the data must be done to have deeper understanding into the quality of the data, in terms of outliers and the skewedness of the figures, descriptive Statistics of categorical and numerical variables was done for that to be achieved. As well as the ability to understand the main attributes that affect the results of the price. Afterwards when the data is organized and transformed into a form that could be processed by the data mining technique. Different data mining models were designed to predict prices and values of used cars. In this study three models are proposed to be built using Logistic Regression model technique, Random Forest Regressor and Bagging Regressor. Firstly, the data was portioned into section for training and the other part for testing, portioning percentage can be tested with different ratios to analyse different results. All three models were evaluated on four evaluation matrices known as model score, Mean Square Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). From all, the Random Forest Regressor outperformed .

Limitations: In the past year the world of automobiles has seen a drastic change with the semiconductor shortages after the pandemic, which led to spike in used car prices. Hence, there was fast change in car prices during this study which will affect the actual car pricing prediction future. As the current dataset will undervalue the cars in the market. Therefore, a model that is built on real time data can be best integrated into a mobile app for public use would be the idea solution .

Benefits: Using data mining and machine learning approaches, this project proposed a scalable framework for Dubai based used cars price prediction. Buyanycar.com website was scraped using the Parse Hub scraping tool to collect the benchmark data. An efficient machine learning model is built by training, testing, and evaluating three machine learning regressors named Random Forest Regressor, Linear Regression, and

Bagging Regressor. As a result of pre-processing and transformation, Random Forest Regressor came out on top with 95% accuracy followed by Bagging Regressor with 88%. Each experiment was performed in realtime within the Google Colab environment. In comparison to the system's integrated Jupyter notebook and Anaconda's platform, algorithms took less training time in Google Colab.

Literature Review:4

Title:Used Car Price Prediction using K-Nearest Neighbor Based Model

Abstract: Predicting the price of used cars is one of the significant and interesting areas of analysis. As an increased demand in the second-hand car market, the business for both buyers and sellers has increased. For reliable and accurate prediction it requires expert knowledge about the field because of the price of the cars dependent on many important factors. This paper proposed a supervised machine learning model using KNN (K Nearest Neighbor) regression algorithm to analyze the price of used cars. We trained our model with data of used cars which is collected from the Kaggle website. Through this experiment, the data was examined with different trained and test ratios. As a result, the accuracy of the proposed model is around 85% and is fitted as the optimized model.

Keywords: Regression, Cross-validation, K- K Nearest Neighbor, Prediction, Machine Learning, Used Cars Accuracy, Preprocessing, Fold

Methodology: The Used Cars data set was taken and data processing has done to filter the data and to remove some unnecessary data. The model was trained with the processed data using the KNN algorithm to predict the sales of used cars with higher accuracy.

Limitations: Here, we have validated our proposed model for 5 folds and 10 folds. It is getting accuracy 82%, RMSE rate 4.73 and MAE rate 2.13 for 10 folds with a K value of 4. It is seen that the proposed model getting the best results after cross-validation.

Benefits: Here, we have estimated the accuracy of the model by training with different values of k from 2 to 10 to find a comparative as best performance. Here prediction is made by looking whole training set to find the k most similar values. To find the most similar values for the new data, the distance is measured using the Euclidean Distance metric. Then the average of the measure is taken to find the estimation value. The formula for Euclidean Distance is as shown below where A and B are two points for which the distance should be calculated.

Literature Review: 5

Title:Car Price Prediction Using Machine Learning

Abstract: The demand for used cars has increased significantly in the past decade and it is prognosticated that with Covid-19 outbreak this requirement will augment considerably. Hence to enhance the reliability, with the expansion of the used car market, a model that can forecast the current market price of a used automobile on the basis of a variety of criteria. This analysis can be used to study the trends in the industry, offer better insight into the market, and aid the community in its smooth workflow. The aim of this research paper is to predict the car price as per the data set (previous consumer data like engine capacity, distance traveled, year of manufacture, etc.). The result of these algorithms will be analyzed and based on the efficiency and accuracy of these algorithms, the best one of them can be used for the said purpose.

Keywords:

Machine Learning, Linear Regression, Lasso Regression, Correlation.

Methodology: The main goal of this method is to give users an accurate estimate of how much has to be paid for the given vehicle. The model may give the customer a record of possibilities for various automobiles based on the details of the automobile the customer wants. The system assists in providing the customer with sufficient data to help him to reach a conclusion. The used automobile market is expanding at an exponential rate, and vehicle vendors may profit from this by offering incorrect prices to capitalise on the demand. As a result, a system that can predict the price of a car based on its parameters while also taking into consideration the costs of competing vehicles is necessary. Our system fills in the gaps by providing buyers and sellers with an estimate of the car's value based on the best algorithm available for price prediction.

Limitations: In this research, we used linear regression and lasso regression to develop a price model for used automobiles in a comparative research. Data was gathered from Kaggle for each algorithm. The main goal of this study is to discover the best predictive model for estimating the price of a used.

Benefits: With the rise in auto ownership, the used automobile market is ripe for growth. The healthy development of the used car market requires an accurate used car pricing evaluation. Since the developed system can be realtime and user friendly in terms of its handling, it is an overall unique proposal idea that is simple to implement and gives overall customer satisfaction, proving to be a profitable business idea.

Literature Review: 6

Title:Vehicle Price Prediction using SVM Techniques

Abstract: The prediction of price for a vehicle has been more popular in research area, and it needs predominant effort and information about the experts of this particular field. The number of different attributes is measured and also it has been considerable to predict the result in more reliable and accurate. To find the price of used vehicles a well defined model has been developed with the help of three machine learning techniques such as Artificial Neural Network, Support Vector Machine and Random Forest. These techniques were used not on the individual items but for the whole group of data items. This data group has been taken from some web portal and that same has been used for the prediction. The data must be collected using web scraper that was written in PHP programming language. Distinct machine learning algorithms of varying performances had been compared to get the best result of the given data set. The final prediction model was integrated into Java application.

Keywords: Artificial neural network, Support vector machine, Random forest

Methodology: Based on the varying features and factors, and also with the help of experts knowledge the vehicle price prediction has been done accurately. The most necessity ingredient for prediction is brand and model, period usage of vehicle, mileage of vehicle. The fuel type used in the vehicle as well as fuel consumption per mile highly affect price of a vehicle due to a frequent changes in the price of a fuel. Different features like exterior color, door number, type of transmission, dimensions, safety, air condition, interior, whether it has navigation or not will also influence the vehicle price. In this paper, we applied different methods and techniques in order to achieve higher precision of the used vehicle price prediction.

Limitations: When the application is executed it starts running. The used browser is internet explorer and the server will start its process. The project will run on the local area network so the server machine will serve as the administrator while the other connected systems can act as the clients.

Benefits: In this paper, the insufficient set of complex data is the drawback here. We will get only 50 percent result on applying the single machine algorithm. Therefore, we proposed multiple groups of machine learning algorithm to gain more accuracy and it achieved 93 percent of efficiency. This comparison of single and multiple groups of machine learning algorithm is significant. And also it overcome the drawback of single machine algorithm whch is given in proposed system. Although, this system has achieved valuable performance in vehicle price prediction, our aim for the future work is to test this system to work successfully with various data sets.

Literature Review: 7

Title:USED CAR PRICE PREDICTION AND LIFE SPAN

Abstract: The main objective of this project is to predict the Prices of Used Cars, compare the prices and also estimate the life span of a particular car, keeping in mind various statistics of that car. It is said that a new car loses its value by 10% the moment the car is taken out from the showroom. We can easily say that the main predictor of prices in this scenario is the number is kilometers the car has been driven

Keywords: multiple linear regression analysis, naive bayes, k-nearest neighbors and decision trees

Methodology: There are two primary phases in the system: 1. Training phase: Based on the algorithm chose the system is trained by using the data in the data set and fits a model (line/curve) accordingly. 2. Testing phase: the inputs are provided to the system and is its working is tested. The accuracy has checked. Therefore, the data must be appropriate which is used to train the model or test it. The designed system is to detect and predict price of used car. In order to do this appropriate algorithm must be used to do the two different tasks. Different algorithms were compared for its accuracy Before the algorithms are selected for further use. The well-suited one for the task was chosen.

Limitations: The process starts by collecting the dataset. The next step after this is to do Data Preprocessing which includes Data cleaning, Data reduction, Data Transformation. Then, we will predict the price using various machine learning algorithms. The algorithms involve Linear Regression, Ridge Regression and Lasso Regression. The best model is selected which predicts the most accurate price. After selecting the best model, the predicted price will be display to the user according to user's inputs. User can give input through website to for used car price prediction to machine learning model.

Benefits: This Project In machine learning model that will be connected with may dataset and with various website which can provide real time data for price prediction Will Stored in their site or GitHub. Also, we may add big amount of data of car price which can help an improve accuracy of the machine learning model. We also trying to develop an android app as user interface for interacting and user-friendly with user. For better performance of the model, we also plan a to use neural network.

Literature survey:8

Title:Predicting the Price of Used Cars using Machine Learning Techniques

Abstract: In this paper, we investigate the application of supervised machine learning techniques to predict the price of used cars in Mauritius. The predictions are based on historical data collected from daily newspapers. Different techniques like multiple linear regression analysis, k-nearest neighbours, naïve bayes and decision trees have been used to make the predictions. The predictions are then evaluated and compared in order to find those which provide the best performances. A seemingly easy problem turned out to be indeed very difficult to resolve with high accuracy. All the four methods provided comparable performance. In the future, we intend to use more sophisticated algorithms to make the predictions.

Keywords: component; formatting; style; styling; insert (key words)

Methgodology: Data was collected from <<petites announces>> found in daily newspapers such as L'Express [8] and Le Defi [9]. We made sure that all the data was collected in less than one month interval as time itself could have an appreciable impact on the price of cars. In Mauritius, seasonal patterns is not really a problem as this does not really affect the purchase or selling of cars.

The following data was collected for each car: make, model, volume of cylinder (funnily this is usually considered same as horsepower in Mauritius), mileage in km, year of manufacture, paint colour, manual/automatic and price. Only cars which had their price listed were recorded Because many of the columns were sparse they were removed. Thus, paint colour and manual/automatic features were removed. The data was then further tweaked to remove records in which either the age (year) or the cylinder volume was not available. Model was also removed as it would have been extremely difficult to get enough records for all the variety of car models that exist. Although data for mileage was sparse, it was kept as it is considered to be a key factor in determining the price of used cars. Predicting the resale value of a car is not a simple task. It is trite knowledge that the value of used cars depends on a number of factors. The most important ones are usually the age of the

car.

Limitations: Its make (and model), the origin of the car (the original country of the manufacturer), its mileage (the number of kilometers it has run) and its horsepower. Due to rising fuel prices, fuel economy is also of prime importance.

Benefits: In this paper, four different machine learning techniques have been used to forecast the price of used cars in Mauritius. The mean error with linear regression was about Rs51, 000 while for kNN it was about Rs27, 000 for Nissan cars and about Rs45, 000 for Toyota cars. J48 and NaiveBayes accuracy dangled between 60-70% for different combinations of parameters. The main weakness of decision trees and naïve bayes is their inability to handle output classes with numeric values. Hence, the price attribute had to be classified into classes which contained a range of prices but this evidently introduced further grounds for inaccuracies. The main limitation of this study is the low number of records that have been used. As future work, we intend to collect more data and to use more advanced techniques like artificial neural networks, fuzzy logic and genetic algorithms to predict car prices.

Literature Review: 9

Title:Used Cars Price Prediction using Supervised Learning Techniques

Abstract: The production of cars has been steadily increasing in the past decade, with over 70 million passenger cars being produced in the year 2016. This has given rise to the used car market, which on its own has become a booming industry. The recent advent of online portals has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of a used car in the market. Using Machine Learning Algorithms such as Lasso Regression, Multiple Regression and Regression trees, we will try to develop a statistical model which will be able to predict the price of a used car, based on previous consumer data and a given set of features. We will also be comparing the prediction accuracy of these models to determine the optimal one.

Keywords:ANOVA, Lasso Regression, Regression Tree, Tukey's Tes

Methodology: Using One-way Analysis Of Variance (ANOVA) we need to verify whether the error rates of these models differ significantly from each other. The process was run 35 times, and the error rates for lasso regression, multiple regression, and regression tree were noted (Table 8) along with the respective seeds of splitting for reproducibility.

Limitations: The data set primarily comprises of categorical attributes along with two quantitative attributes.

Benefits:The prediction error rate of all the models was well under the accepted 5% of error. But, on further analysis, the mean error of the regression tree model was found to be more than the mean error rate of the multiple regression and lasso regression models. Even though for some seeds the regression tree has better accuracy, its error rates are higher for the rest. This has been confirmed by performing an ANOVA.

Literature Review:10

Title: Predicting Used Car Prices

Absract: Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models across cities in the United States. Our results show that Random Forest model and K-Means clustering with linear regression yield the best results, but are compute heavy. Conventional linear regression also yielded satisfactory results, with the advantage of a significantly lower training time in comparison to the aforementioned methods .

Key words:k-means,linear regression,Random forest algorithm,X-G boost.

Methodology :We utilized several classic and state-of-the-art methods, including ensemble learning techniques, with a 90% - 10% split for the training and test data. To reduce the time required for training, we used 500 thousand examples from our dataset. Linear Regression, Random Forest and Gradient

Boost were our baseline methods. For most of the model implementations, the open-source Scikit-Learn package was used.

Limitations: Compared to Linear Regression, most Decision-Tree based methods did not perform comparably well. This can be attributed to the apparent linearity of the dataset. We believe that It can also be attributed to the difficulty in tuning the hyperparameters for most gradient boost methods. The exception to this is the Random Forest method which marginally outperforms Linear Regression. However Random Forests tend to overfit the dataset due to the tendency of growing longer trees. This was worked upon by restricting the depth of trees to different values and it was observed that beyond limiting depth toresulted in negligible improvement in prediction performance but progressively increased overfitting. As expected lightGBM performed marginally better than XGBoost but had a significantly faster training time. Building up from the relatively good performance of Linear Regression, the KMeans + Linear Regression Ensemble Learning Method (with K = 3) produced the best score on test data without high variance asR2.

Benefits: Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car appropriately[2-3]. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices.

Existing system:

The prices of new cars in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But due to the increased price of new cars and the incapability of customers to buy new cars due to the lack of funds, used cars sales are on a global increase. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a used car's actual market value. It is important to know their actual market value while both buying and selling.

Proposed System:

By performing different models, it was aimed to get different perspectives and eventually compared their performance. With the help of the data visualizations and exploratory data analysis, the dataset was uncovered and features were explored deeply. The relation between features were examined. At the last stage, predictive models were applied to predict price of cars in an order: random forest, linear regression, ridge regression, KNN.

Conclusion:

The prediction error rate of all the models was well under the accepted 5% of error. But, on further analysis, the mean error of the regression tree model was found to be more than the mean error rate of the multiple regression and lasso regression models. Even though for some seeds the regression tree has better accuracy, its error rates are higher for the rest. Also, the post-hoc test revealed that the error rates in multiple regression models and lasso regression models aren't significantly different from each other. To get even more accurate models, we can also choose more advanced machine learning algorithms such as random forests, an ensemble learning algorithm which creates multiple decision/regression trees, which brings down overfitting massively or Boosting, which tries to bias the overall model by weighing in the favor of good performers. More data from newer websites and different countries can also be scraped and this data can be used to retrain these models to check for reproducibility.

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 August 10, 2018].