Available online at http://www.advancedscientificjournal.com
http://www.krishmapublication.com
IJMASRI, Vol. 2, issue 5, pp. 571-577, May -2022
10.53633/ijmasri.2022.2.5.006

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY ADVANCED SCIENTIFIC RESEARCH AND INNOVATION (IJMASRI)

ISSN: 2582-9130 **IBI IMPACT FACTOR 1.5** **DOI: 10.53633/IJMASRI**

RESEARCH ARTICLE

USED VEHICLE PRICE PREDICTION SYSTEM

Vaibhav Gupta[1], M.L. Sharma[2], K.C. Tripathi[3]
[1, 2, 3] *Department of Information Technology, Maharaja Agrasen Institute of Technology*

Abstract

Cars have become a necessity in this modern world. They are no longer a luxury but rather an asset. Everyone uses cars as the primary vehicle of transportation nowadays but cars are still not very affordable. The problem now becomes that car is needed, but it is not affordable. The economic solution to this would be to buy used cards. But then the question arise, where to buy them from? Stores that sell used cars have become outdated. They have very limited stock of vehicles available to sell and the quality can't be assured. Plus the data can be tampered with by the dealer. To overcome such problems a proper solution is to have an online system and a website where you can buy and sell cars. But there is still one problem that remains. How to price the used cars. It is not as easy to just divide the initial price by a number based on the number of years the vehicle has been used for. There are a lot of factors that determine the current state of the vehicle and these factors are used to predict the prices for the cars too. This system used to predict the car prices is called Car Price Prediction System. The quality of a vehicle deteriorates with age of course but that is not all. Every single vehicle is different even when it is manufactured and sold as a new product and even more so when the same vehicle is used over time. In Car Price Prediction System, we have used the Year of Production of vehicle, the base MRP of the vehicle, the fuel type of the vehicle and a lot of other factors.

Keywords: car price prediction, lasso regression, linear regression, regression analysis, machine learning, correlation analysis

Introduction

To overcome this problem we have come up with a model that will be highly effective. Regression Algorithms are used because they provide us with continuously evaluated value as an output and not a categorized value or a value within a range. So, it will help us in predicting the actual price of a car rather than the price range of a car. We will also be providing a user interface which has also been

developed which takes input from any user and displays the actual Price of a car according to user's inputs.

The objective of this project is to predict an accurate price for a used car depending on the various features. In order to draw a conclusion, we implement several learning methods on our sample data set. The models that will be used for comparison are Lasso Regression, Ridge Regression, Linear Regression (Gelman 2006), Tree Regression (Quinlan 1993) and a few other models. We will train the models using each of the above techniques one by one and then observe the predictions and the accuracy of the trained model to figure out which regression model gives the best results from the Car Price Prediction System (Monburinon et al., 2018). The factors affecting the prediction system will be studied in detail and the relationship curves will be plotted between different factors to determine which ones are useful for our prediction system and which ones are not. Regression analysis is the technique that is used to solve regression based problems in ML. The regression technique is used primarily to calculate the strength of the predictor, forecast of the trend, time series, and in case of cause & effect relation.

Requirements

Hardware Requirements:
Windows 7,8,10,11
Mac OS X 10.11 or higher, 64-bit
Linux: RHEL 6/7, 64-bit
x86 64-bit CPU (Intel / AMD architecture)
4 GB RAM
5 GB free disk space

Software Requirements:
Python 3.6 or above
PyCharm
Anaconda
PIP 2.7
Jupyter Notebook
Chrome
Notepad++
Vs Code/Sublime text editor

Methodology

The dataset is taken from Kaggle which is based on the online car sales. Features in the dataset are name, year, selling price, kms driven, fuel, etc. Few changes were made according to the needs of our models and then the data was divided into two parts, one for training of model and other for testing of model.

The data was then used to train various models namely Multiple Linear Regression, Bayesian Ridge Regression, Decision Tree Regression, Lasso Regression, Ridge Regression. Hence we will be comparing the different regression algorithms and their performance in this experiment.

Objectives

To use the dataset and compare the various factors to determine the best fits for the model. To compare several different regression algorithms.

To present a working software of the proposed car price prediction system.
To make sure that the model has sufficient accuracy.

To organize the work done and make a website or a software that uses the prediction system to make predictions.

Procedure

First step:

To collect a concrete and authentic dataset for the used cars and their features. The quality of the dataset is the cornerstone of the entire model and it's essential to obtain a good dataset first

Second Step:

Data Processing. Data Processing means converting the raw data available into data that is readable by the machine. It has various subparts. Data cleaning: Means removing the unwanted fields from the data and any other undesirable elements from the data. Data reduction: Reduction of data into simpler terms and taking only a part of the data for the first or second phase of the model. Data Transformation:

Conversion of data into numerical forms that are easier for the machine to process and perform operations on.

Third Step:

Using the test regression algorithms to predict the final output

Fourth Step:

To observe and compare the output values using different algorithms and draw a conclusion for the best suitable algorithm for the Car Price Prediction System.

The Dataset

The dataset used for the research is the used car data from a used car reselling website, Car Dekho.

df.head()

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

Fig: Raw Dataset Head

As observed, the raw dataset has a few useless fields and some vague fields as well. So, we drop the car name field as it has no contribution whatsoever in predicting the price of the car. There are also fields like Fuel Type, Seller Type, Transmission which do not have any mathematical values and have text values instead. But these values can't be eliminated like the car name since these do have an effect on the price of the car. In order to convert these fields into mathematical values for the ease of fitting these into our machine learning model, we use the one hot encoding technique.

final_dataset.head()

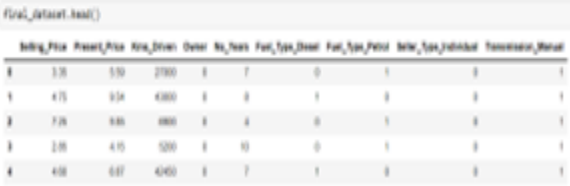

Fig: Dataset after modifications

A correlation is drawn between all the variables in the table (Kanwal Noor and Sadaqat Jan 2017). The correlations help us figure out which of the fields are highly correlated, that is, have a huge impact on the others fields when changed. A highly positive correlation between 2 fields say field A and field B means that change in field A has an impact on field B. And the positive relation here means that an increase in A would also result in increase in B and a decrease in A would give a decrease in B. Correlation between 2 variables doesn't necessarily have to be positive. A highly negative correlation is just as important for a good prediction. Let's say columns A and C and have highly negative correlation with each other. This means that an increase in the value in A would result in decrease in value in C and a decrease in value of A would result in increase in value of C.

In order to find the variables that would be useful for the prediction of car price, we would use the correlation between all the variables and filter out the variables that have a highly positive or negative correlation with the car price.

Fig: Correlation table between the variables

Fig: Correlation heatmap between variables

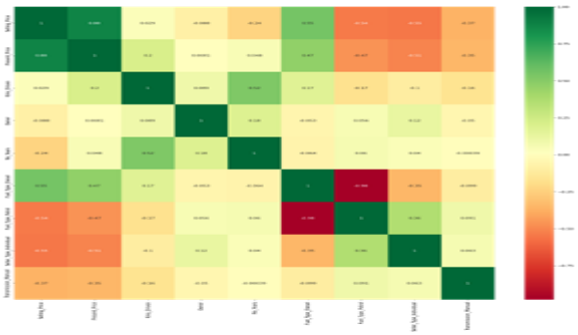

Fig: Correlation heatmap between variables

Since correlation function determines the degree of relation between two variables, it will never have a value exceeding 1 and falling short of -1. A correlation value of 0 means that the variables do not have any impact on each other.

Multiple Linear Regression

In multiple linear regression, multiple independent variables are used to predict the value of the desired variable which is known as dependent variable. The formula used is

$$y=m0 + m1x1 + m2x2 + \ldots + mnxn + c$$

Here, y is the dependent variable, that is selling price in this case and x1,x2, … , xn are the dependent variable. m1,m2, … , mn are the weights calculated by the model.

Result for used car price prediction using multiple linear regression:

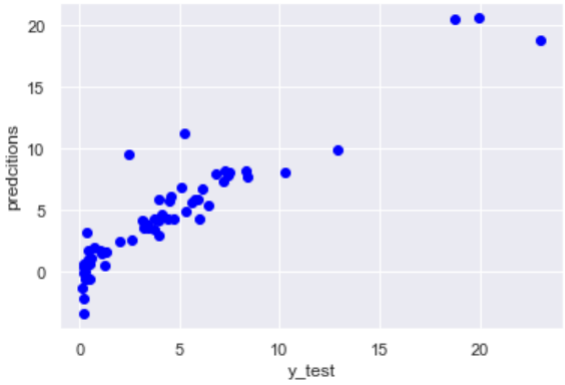

Fig: Plot of predictions using Linear Regression

Lasso Regression

LASSO stands for Least Absolute Shrinkage and Selection Operator. Lasso regression uses regularization and variable selection techniques to enhance the accuracy of the model. It also uses shrinkage which shrinks the data values towards a central point and the L1 regularization adds a penalty of the absolute value of coefficient's magnitude. It results in making sparse models that have fewer coefficients as some coefficients can become zero and thus gets eliminated from the model.

Results for used car price prediction using Lasso Regression:

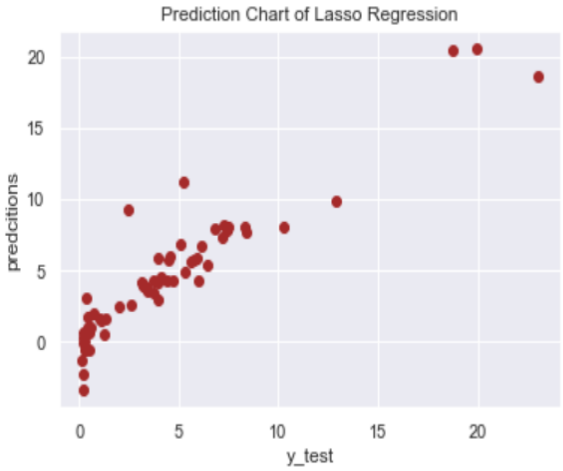

Fig: Plot of predictions using Lasso Regression

Ridge Regression

Ridge regression is the method used for multiple regression methods that is having the problem of multicollinearity. In case of multicollinearity, variances are large that lead to predicting the value which is far from the true desired value. It prevents multicollinearity by shrinking the parameters. It also makes use of L2 regularization technique, which adds L2 penalty of value square of the magnitude of coefficients and thus helps in dealing with multicollinearity which happens when independent values are highly correlated.

Result for used car price prediction using ridge regression:

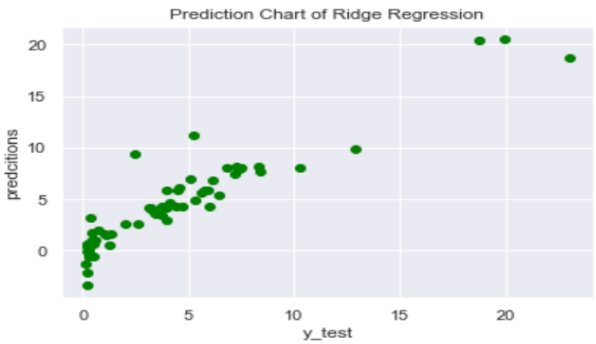

Fig: Plot of predictions using Ridge Regression

Bayesian Ridge Regression

Bayesian Regression makes use of regularization parameters for estimation. Bayesian regression is beneficial to use when we have to deal with insufficient or poorly distributed data. Here, the output is calculated from probability distribution in spite of making prediction as a single value. It makes use of ridge regression and its coefficients under the Gaussian Distribution to estimate the value of desired variable.

Result for used car price prediction using Bayesian Ridge regression:

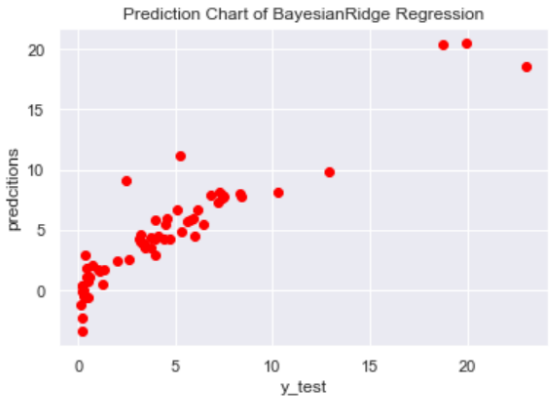

Fig: Plot of predictions using Bayesian Ridge Regression

Decision Tree Regression

In Decision Trees, the regression model is built in the structure like a tree. The dataset is broken into smaller subsets and simultaneously a decision tree is developed incrementally. The Decision tree looks like a flow-chart in which the internal nodes represent the test on an attribute. Branches represent the outcomes of the tests and the leaf nodes represent a class label. It is used to fit a sine curve with additional noisy observations.

Result for used car price prediction using Decision Tree regression:

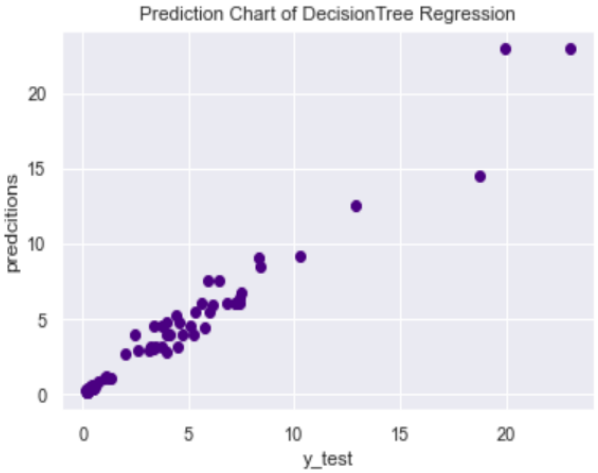

Fig: Plot of predictions using Decision Tree Regression

Random Forest Regression

Random Forest Regression is a type of supervised learning algorithm. It uses the ensemble learning technique for regression. Ensemble Learning technique is a method that combines the result from multiple models to make a more accurate prediction.

A Random Forest works by creating multiple decision trees during training and returning the mean of the predictions of all trees.

Result for used car price prediction using Random Forest regression:

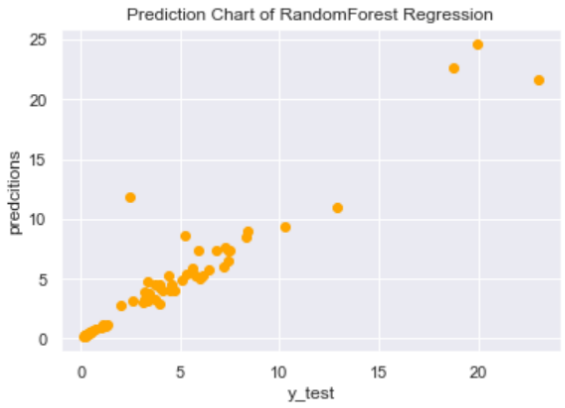

Fig: Plot of predictions using Random Forest Regression

After training various models, the remaining data was used to test the accuracy of the models and checking if ensembling methods can provide us with better results and the following data was recorded after testing all the models with the test data:

Errors Models	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	1.0998575552990952	2.982384861859748	1.726958268708236
Lasso Regression	1.0934873952604163	2.907197959149361	1.7050507204037542
Ridge Regression Model	1.108094193398559	2.963295353286871	1.7214224796042576
Bayesian Ridge Regression	1.075017607433412	2.8302932475517473	1.6823475406561355
Random Forest Regression	0.7446229508196723	2.5644490327868863	1.6013897192085649
Decision Tree Regression	0.6027868852459017	0.9108311475409837	0.9543747416717312

Future Scope

In future, having better and concise data would result in the improvement of this proposed model. Adding more to the training set would improve the accuracy of the model by a lot. Due to the lack of data, the fuel types were limited to Petrol, Diesel, CNG in the current model. But in the future, having more data on electric vehicles would be beneficial for the car price prediction model. Improved ensembling techniques can be applied to further work on improving the accuracy. Currently, vehicles running on exhaustible resources such as Petrol, Diesel, CNG would be discarded and the sales of electric vehicles would also increase. Which would create a huge wave in the used car buying and selling industry making the used car price prediction system a must. In this research paper and in reference to the mentioned research papers, conclusions were drawn to make the car price prediction system accurate. A comparison was done between all the variables in the dataset and the variables that could be used in the prediction of car price were determined. The regression algorithms were compared. The proposed system will help in determining the prices of the used cars. All things which were said to affect the price of a used car were affecting the prices in our model too. We tried ensembling techniques but still the Decision trees Regression gives the best results.

We can see in the model vs errors table that the decision tree is giving us the least error among all the algorithms that we have tried for the model training. Along with that, it can be seen from the graph of all models that the decision tree is the most consistent method to predict the selling prices of used cars.

the vehicles such as Trucks, Autorickshaws, E-Rickshaws do not have enough data for us to incorporate them within the dataset. Hopefully, with time, the dataset for the same would improve to help with the further development of the system.Conclusion

The modern scenario has resulted in a need for the selling and buying of second hand vehicles. Because of this, the system is valuable and with society moving away from fossil fuels, it is inevitable that the old

References

1. Listiani, M., (2009). “Support Vector Regression Analysis for Price Prediction in a Car Leasing Application.” Thesis (msc). Hamburg University of Technology.
2. Richardson, M., (2009). “Determinants of Used Car Resale Value.” Thesis (bsc). The Colorado College.
3. Gelman, A. And Hill, J., (2006). “Data Analysis Using Regression and Multilevel Hierarchical Models.” Cambridge University Press, New York, USA.
4. Quinlan, J. R., (1993). C4.5: “Programs for Machine Learning.” Morgan Kaufmann.
5. Monburinon, N., Chertchom, P. Kaewkiriya, K, Rungpheung,S. Buya S and Boonpou, (2018). "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), 2018, pp. 115-119, doi: 10.1109/ICBIR.2018.8391177.
6. Kanwal Noor and Sadaqat Jan. (2017) Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications 167(9):27-31.

******