International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 12 | Dec 2022 www.irjet.net p-ISSN: 2395-0072

Prediction of Used Car Prices using Machine Learning Techniques

Eesha Pandit[1], Hitanshu Parekh[2], Pritam Pashte[3], Aakash Natani[4]

[1, 3, 4] *Student, Dept. of IT Engineering, Thakur College of Engineering and Technology, Mumbai, Maharashtra, India.*
[2] *Student, Dept. of Information Technology, St. Francis Institute of Technology, Mumbai, Maharashtra, India.*

Abstract - *The manufacturer sets the price of a new car in the industry, with the government incurring some additional expenditures in the form of taxes. Customers purchasing a new car may thus be sure that their investment will be worthwhile. However, due to rising new car prices and buyers' financial inability to purchase them, used car sales are increasing globally. As a result, a used car price prediction system that efficiently assesses the worthiness of the car utilizing a range of factors is required. The current system comprises a system in which a dealer decides on a price at random and the buyer has no knowledge of the car or its current worth. In reality, the seller has no clue what the car is worth or what price he should charge for it. To address this issue, we have devised a highly effective model. Regression algorithms are employed because they produce a continuous value rather than a classified value as an output. As a result, rather than predicting a car's price range, it will be feasible to estimate its real price. A user interface has also been created that takes input from any user and shows the price of a car based on the inputs.*

Key Words: Used Car Price Prediction, Regression Algorithms, Machine Learning, Linear Regression, Ridge and Lasso Regression, Bayesian Ridge Regression, Decision Tree, Random Forest, XG Boost, Gradient Boosting.

1.INTRODUCTION

Determining if the quoted price of a used car is fair is a difficult process owing to the numerous elements that influence a used vehicle's market pricing. The goal of this research is to create machine learning models that can properly anticipate the price of a used car based on its features so that buyers can make informed choices. We create and analyze numerous learning algorithms using a dataset that includes the selling prices of various brands and models. We will compare and choose the best machine learning algorithms such as Linear Regression, Lasso Regression, Ridge Regression, Bayesian Ridge Regression, Decision Tree Regression, Random Forest Regression, XG Boost Regression, and Gradient Boosting Regression. The price of the car will be determined by a number of factors. Regression algorithms are used because they produce a continuous value rather than a categorized value, allowing us to predict the actual price of a car rather than the price range of a car. A user interface has also been created that takes input from any user and shows the price of a car based on the inputs.

2. LITERATURE REVIEW

[1] Various studies have been conducted in order to predict the price of used cars. Researchers regularly anticipate product prices using past data. Pudaruth predicted car prices in Mauritius, and these cars were not new, but rather used to predict the prices, he employed multiple linear regression, k-nearest neighbours, Naive Bayes, and decision tree techniques. When the prediction results from various strategies were compared, it was discovered that the prices from these methods are quite similar. However, the decision tree technique and the Nave Bayes approach were proven to be incapable of classifying and predicting numeric values. According to Pudaruth's research, the small sample size does not give good prediction accuracy.

[2] Kuiper, S. (2008) demonstrated a multivariate regression model that helps in classifying and predicting values in numeric format. It demonstrates how to apply this multivariate regression model to forecast the price of 2005 General Motors (GM) vehicles. The price prediction of cars does not require any special knowledge. So, the data available online is enough to predict prices. The author of the article did the same car price prediction and introduced variable selection techniques that helped in finding which variables were more relevant for inclusion in the model.

[3] In 2019, Pal et al discovered as a methodology for predicting used cars prices using Random Forest. The paper evaluated usedcar price prediction using Kaggle data set which gave an accuracy of 83.62% for test data and 95% for train-data. The most relevant features used for this prediction were price, kilometer, brand, and vehicle type and identified by filtering out outliers and irrelevant features of the data set. Being a sophisticated model, Random Forest provided good accuracy in comparison to prior work using these data sets.

[4] Gegic, E. et al. (2019) demonstrate the need to create a model to forecast the cost of second hand cars in Bosnia and Herzegovina. They used machine learning techniques such as artificial neural networks, support vector machines, and random forests. However, the aforementioned methods were used in concert. The web scraper, which was created using the PHP programming language, was used to gather the data from the website autopijaca.ba for the forecast. Then, to determine which method best suited the provided data, the respective performances of various algorithms were compared. A Java application contained the final prediction

model. Additionally, the model's accuracy of 87.38% was determined when it was verified using test data. Dholiya et al. demonstrated a machine learning-based method for auto resales in 2019.

[5] The goal of the system that Dholiya , M., et al. developed is to give the user a realistic estimation of how much the vehicle might cost them. Based on the specifics of the automobile the user is looking for, the system, which is a web application, may also offer the user a list of options for various car kinds. It assists in providing the buyer or seller with useful information on which to base their decision. This system makes predictions using the multiple linear regression algorithm, and this model was trained using historical data that was obtained over an extended period of time. The raw data was initially gathered using the KDD (Knowledge Discovery in Databases) process. Afterward, it underwent preprocessing and cleaning in order to identify patterns that are valuable and then derive some meaning from those patterns.

[6] Richardson conducted his analysis under the presumption that automakers are more inclined to produce cars that don't lose value quickly. He demonstrated, in particular, that hybrid cars are better equipped to maintain their value than conventional vehicles by utilising multiple regression analysis. This is perhaps because there are increasing concerns about the environment and the climate, as well as because it uses less gasoline. In this study, the significance of additional variables including age, mileage, make, and MPG (miles per gallon) was also taken into account. All of his information was gathered from several websites.

[7] Listiani published another study that is comparable and uses Support Vector Machines (SVM) to forecast lease car pricing . This study demonstrated that when a very large data set is available, SVM is significantly more accurate at price prediction than multiple linear regression. SVM is also superior at handling high dimensional data and steers clear of both under- and over-fitting problems. Finding crucial features for SVM is done using a genetic algorithm. However, the method does not demonstrate why SVM is superior to basic multiple regression in terms of variance and mean standard deviation.

3. TECHNOLOGY USED

Python is mainly used in this project to implement machine learning algorithms since it contains a lot of built-in methods in the form of packaged libraries and modules. During project implementation, Python, Pandas, NumPy, Matplotlib, Seaborn, Scikit-Learn, Plotly, and Pickle libraries were used.

The following technologies were used to build the web application: HTML, CSS, Flask, Jsonify, and Requests.

4. METHODOLOGY

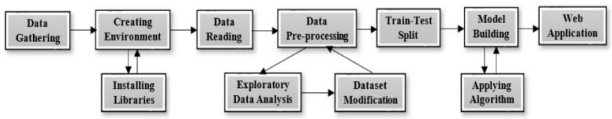

Figure 1: Workflow of Study

4.1 Data Gathering

The source of the data is the web portal Kaggle.com, where vehicle data sets are provided by Cardekho for the sale and purchase of cars. The dataset contained the following features: car name, year, selling price, present or current price, kilometres driven, fuel type: diesel, petrol, or CNG (compressed natural gas), seller type: dealer or individual, transmission: automatic or manual, owner (number of previous owners).

4.2 Create Environment

An environment is created using the Anaconda prompt. This environment would separate our project area from the other default environment (base) or other previously created environments. All the packages, libraries, and modules that we need can be manually installed in the environment created in this way, making it an advantageous step. In such an environment, we can make changes according to our needs.

4.3 Data Reading

The first step is to import and read the csv file for the research. The dataset is extensively examined in terms of null values, shape, columns, numerical and categorical features, dataset columns, unique values of each feature, data information, and so on.

4.4 Data Pre-processing

Some of the data features were renamed for clarity (Present Price = Initial Price, Owner = Previous Owners), and some features that were not important for analysis were removed. In exploratory data analysis, we use statistical graphics and other visualisation techniques to describe the important aspects of data. Top Selling Vehicles, Year vs. Number of Available Vehicles, Selling Price vs. Initial Price, Vehicle Fuel Type, Transmission Type, Seller Type, Age, Selling Price v/s Age, Selling Price v/s Seller Type, Selling Price v/s Transmission, Selling Price v/s Fuel Type, Selling Price v/s Previous Owners, Initial Price vs Selling Price, Selling Price v/s Kilometers Driven, pairplot, heatmaps, and other visualisations are used to gain a better understanding of data. Following EDA, One Hot Encoding approach is used to deal with the dataset's categorical features. After that, the dataset's correlation characteristics are generated and

thoroughly analysed by visualising several plots. Then the features allocation of data is where the dependent feature (Selling Price) and independent features (Initial Price, Kilometers Driven, Previous Owners, Age, and so on) are then allocated for further processing.

4.5 Train-Test Split

Once the dependent and independent features have been assigned, we proceed with the splitting of the dataset into training and testing data. We use 80% of the data to train our model and 20% to test it.

4.6 Model Building

Following the Train-Test split, data modeling is complete, and the process of building the model begins. The model is defined, along with a few parameters, for future implementation. After the model is built, various algorithms are used to create the final results. After building the model, the following algorithms are used for predictive analysis.

Linear Regression: It is a linear approach in statistics for modeling the relationships between a scalar response and dependent and independent variables. In linear regression, relationships are modelled using functions such as linear predictor, and unknown model parameters are estimated from data.

Lasso Regression: It is a sort of linear regression in which the data values are shrunk towards a data point in the center, or, in simpler terms, the mean of the data. The Lasso procedure supports simple and sparse models with fewer parameters. When a model has a high amount of multicollinearity, this regression provides the best fit for that model. This approach can also be used if some aspects of model selection, such as variable selection or parameter elimination, need to be automated. The abbreviation 'LASSO' stands for Least Absolute Shrinkage and Selection Operator.

Ridge Regression: It is a regression approach used for tuning a model and analyzing multicollinear data. This function implements L2 regularization. The multicollinearity of the data results in unbiased least-squares, a huge variance, and hence the predicted values are considerably far from the actual values.

Bayesian Ridge Regression: This regression is used to estimate any probabilistic model of any regression issue using linear regression formulation with the use of probability distributors, providing a natural process that survives data insufficiency or poor data distribution.

Random Forest Regression: Random Forest is a Supervised Learning Algorithm that employs the ensemble learning approach for classification and regression. Random forests are made up of trees that run parallel to each other and have no interaction while they develop. Random Forest is a meta-estimator that aggregates the outcomes of several predictions. It also aggregates numerous decision trees with certain modifications.

XGBoost Regression: XGBoost is a very powerful technique for creating supervised regression models. XGBoost is an ensemble learning strategy that includes training individual models and then merging them (base learners) to get a single prediction.

Gradient Boosting Regression: This is a machine learning approach used to construct a prediction model for regression and classification problems. The prediction model generates an ensemble of weak prediction models, which are often decision trees. This method outperforms the random forest method in most cases.

5. IMPLEMENTATION

Adding a new feature Age, which determines the number of years the vehicle has been used, is stored in the final dataset, and the year attribute is dropped.

5.1 Exploratory Data Analysis

In this stage, we summarize the major characteristics of data using statistical graphics and other visualization tools. Various graphs and charts are plotted to gain a better understanding of the dataset and the relationships between its features.

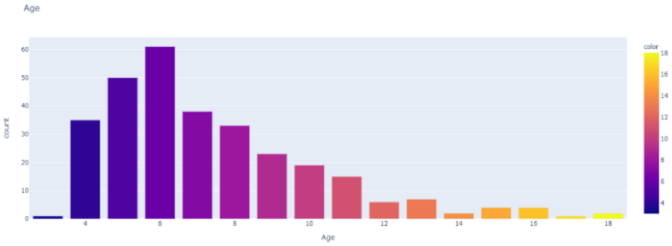

Figure 2: Count w.r.t Age

Vehicle count in relation to vehicle age: The following bar graph depicts the number of vehicles of a certain age.

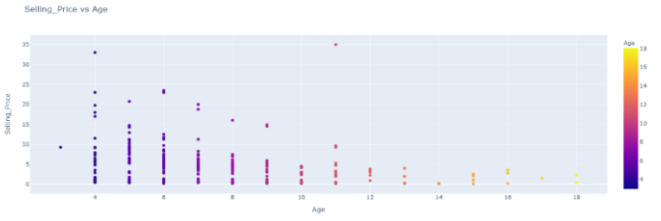

Figure 3: Selling price v/s Age

Comparison of each vehicle's selling price vs. age: The chart below depicts the selling price and age of a certain car. And it is easy to conclude that the selling price is high for a car of a young age.

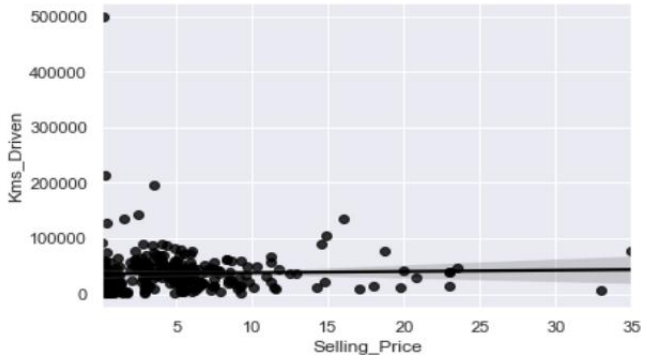

Figure 4: Initial price v/s Selling price

Comparison of Initial Price and Selling Price: The graph below demonstrates the direct proportionality between Initial Price and Selling Price, which suggests that a higher initial price will result in a higher selling price.

Figure 5: Kilometers Driven v/s Selling Price

Comparison of Kilometers Driven vs. Selling Price: The graph above shows that a vehicle with a high number of kilometers driven has a lower selling price than one with a low number of kilometers driven.

5.2 One Hot Encoding

The one hot coding approach is used to deal with the categorical variables in the dataset. It generates a sparse matrix or a dense array based on the parameters while creating a binary column for each category or parameter. Fuel Type, Seller Type, and Transmission were the three categorical variables in our dataset. Following one hot encoding, these variables are given a binary representation, so that for a car with a Fuel Type of Diesel, the value of Fuel_Type_Diesel is a binary 1 and the value of Fuel_Type_Petrol is a binary 0. The same procedure is applied for the remaining category variables.

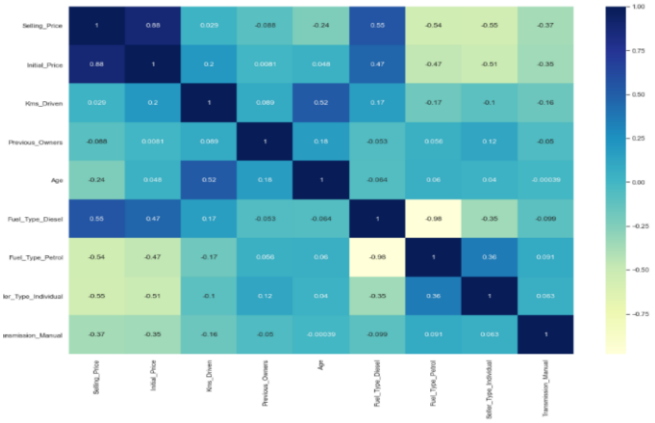

Figure 6: Correlation Heatmap

Heatmap of Correlation Features for the Final Dataset: A dataset's correlation features define how close two variables are to having a linear relationship with each other. Features with a high correlation are more linearly dependent and have the same effect on the dependent variable. If two variables have a high correlation, we can always eliminate one of them. The heatmap of correlation is shown below, with darker colors representing high correlation and lighter colors representing low correlation.

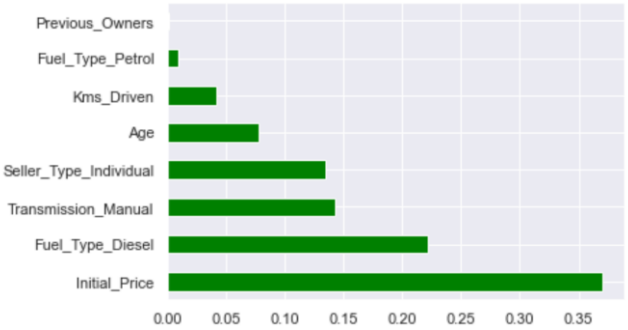

Figure 7: Feature Importance

Feature Importance of dataset: The feature importance technique provides a score to features in a feature set based on their usefulness in predicting the target variable. Initial Price is the most relevant feature in the provided dataset, while Previous Owners is the least important.

5.3 Model Building

After the train-test split of the dataset, modeling is complete, and the process of building the model begins. For final implementation, the model is created with a few parameters, such as the algorithm, x train, y train, x test, and y test. After the completion of the model, various algorithms are used to generate the final results.

5.4 Developing a Web Application:

A web application is then made using HTML, CSS, and JavaScript in the frontend and using the Flask Framework of Python in the backend. This web application allows any user to enter parameters and calculate the estimated selling price of a used car. To view the results, the user must enter values for variables such as year, initial price (in lakhs), kilometers driven, and previous owners, as well as select options for parameters such as fuel type, transmission type, and seller type.

Figure 8: Web Application

6. RESULTS

After applying regression algorithms to the model, the r_2 scores and other assessment metrics such as mean absolute error, mean squared error, and root mean squared error were obtained for comparison of the performance of each method.

Table 1. Evaluation Metrics of Algorithms

Algorithm	R_2 Scores	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)
Random Forest Regression	0.8576	0.7583	2.6763	1.6359
Linear Regression	0.8625	1.0998	2.9823	1.7269
Ridge Regression	0.8634	1.1080	2.9632	1.7214
Lasso Regression	0.8659	1.0934	2.9071	1.7050
Bayesian Ridge Regression	0.8695	1.0750	2.8302	1.6823
XG Boost Regression	0.8958	0.6822	2.2584	1.5027
Gradient Boosting Regression	0.9355	0.6378	1.4111	1.1878
Decision Tree Regression	0.9544	0.6711	1.3139	1.1462

The Decision Tree Algorithm has the best r_2 score of 0.9544 when all regression methods' r_2 scores are compared, which simply implies that the Decision Tree Algorithm has delivered the most accurate predictions when compared to the other algorithms.

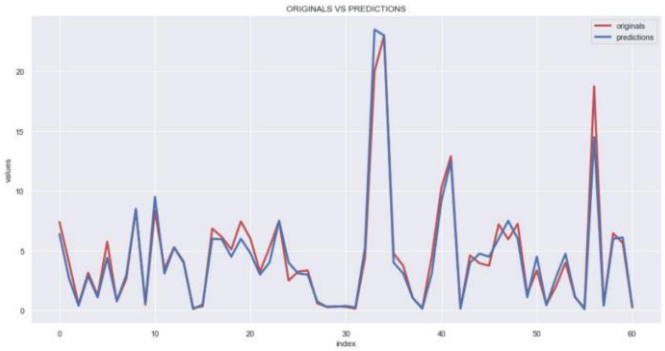

Figure 9: Original v/s Prediction Decision Tree Regression

In the graph above, where the red line represents the original values of the dataset and the blue line shows the values predicted using Decision Tree Regression, we can see that both lines are pretty close to each other, indicating that the predictions are highly accurate.

7. CONCLUSION

Predicting used car prices is a difficult task due to the large number of features and parameters that must be examined in order to get reliable findings. The first and most important phase is data collection and preprocessing. The model was then defined and built in order to implement algorithms and generate results. After executing various regression algorithms on the model, it was concluded that the Decision Tree Algorithm was the top performer, with the greatest r2 score of 0.95, implying that it provided the most accurate predictions, as shown by the Original v/s Prediction line graph. Aside from having the highest r2 score, the Decision Tree also had the lowest Mean Square Error (MSE) and Root Mean Square Error (RMSE) scores, indicating that the errors

in predictions were the lowest of all and that the results obtained were very accurate.

8. FUTURE SCOPE

The developed machine learning model can be exported as a "Python class" and deployed as an open source, ready-to-use price predictor model, which can then be easily integrated with third-party websites. The model can be greatly optimised by using neural networks by designing deep learning network topologies, employing adaptive learning rates, and training on data clusters rather than the entire dataset.

9. REFERENCES

[1] Pudaruth, S. (2014) 'Predicting the Price of Used Cars using Machine Learning Techniques', International Journal of Information & Computation Technology, 4(7), pp. 753–764. Available at: http://www.irphouse.com.

[2] Kuiper, S. (2008) 'Introduction to Multiple Regression: How Much Is Your Car Worth?', Journal of Statistics Education, 16(3). doi: 10.1080/10691898.2008.11889579.

[3] Pal, N. et al. (2019) 'How Much is my car worth? A methodology for predicting used cars' prices using random forest', Advances in Intelligent Systems and Computing, 886, pp. 413–422. doi: 10.1007/978-3- 030-03402-3_28.

[4] Gegic, E. et al. (2019) 'Car price prediction using machine learning techniques', TEM Journal, 8(1), pp. 113–118. doi: 10.18421/TEM81-16.

[5] Dholiya, M. et al. (2019) 'Automobile Resale System Using Machine Learning', International Research Journal of Engineering and Technology(IRJET), 6(4), pp. 3122–3125.

[6] Richardson, M. (2009) Determinants of Used Car Resale Value. The Colorado College

[7] Listiani, M. (2009) Support Vector Regression Analysis for Price Prediction in a Car Leasing Application, Technology. Hamburg University of Technology.

[8] https://www.jigsawacademy.com/popular-regression-algorithms-ml/

[9] https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article

[10] https://www.javatpoint.com/machine-learning-life-cycle

[11] https://www.simplilearn.com/tutorials/machine-learning-tutorial/machine-learning-steps