CAR PRICE PREDICTION

Aditya Arora, Akriti Singh, Aman Goel, Kirti Kushwah

Adiya Arora, Dept. of CSE, Inderprastha Engineering College, Sahibabad, Ghaziabad, India
Akriti Singh, Dept. of CSE, Inderprastha Engineering College, Sahibabad, Ghaziabad, India
Aman Goel, Dept. of CSE, Inderprastha Engineering College, Sahibabad, Ghaziabad, India
Kirti Kushwah, Dept. of CSE, Inderprastha Engineering College, Sahibabad, Ghaziabad, India

---

Abstract - *Over 70 million passenger automobiles were created in 2016. The number of cars produced has been rising over the last tenyears. The secondhand automobile market has emerged as a result of this, and on its own has grown to be a prosperous sector. The emergence of online portals has made it easier for both buyers and sellers to learn more about the patterns and trends that influence a used car's market value. Our goal is to create a statistical model that can forecast the price of a used car by utilizing machine learning algorithms like regression trees, multiple regression, and Lasso regression. This model will be based on past consumer data and a predetermined set of features. We intend to additionally contrasting these models' prediction accuracy in order to identify the best one. In the industry, the manufacturer sets the price of a new car, with the government bearing some additional expenses in the form of taxes. Customers who purchase new cars can therefore be sure that their financial investment is worthwhile. However, sales of used automobiles are rising globally as a result of rising new car prices and consumers' inability to afford them. A used car price prediction system that accurately assesses the car's worthiness based on a range of factors is therefore desperately needed. The current system has a procedure where a vendor chooses a price at random and a buyer as no notion whatthe car is worth in the current market. In actuality, neither the car's current worth nor the asking price are known to the seller. We have created a model that will work incredibly well tosolve this issue. The reason machine learning algorithms are chosen is that their output is continuous rather than categorized. This makes it feasible to forecast an automobile's true cost rather than just its pricing range. Additionally, auser interface has been created that takes input from any user and displays the price of a car based on that input.*

Keywords—Car Price, ML Algorithm, Regression, Prediction, Category.

1.INTRODUCTION

There are so many variables that influence a used car's pricing on the market, it can be difficult to determine whether the quoted priceis accurate. This project's main goal is to create machine learning models that can effectively estimate a used car's price based on its attributes, enabling them to make well- informed decisions. We use and assess learningmethodologies on a dataset comprising the selling prices of various models andmanufactures. We assess how well machine learning algorithms perform. Regression in Linear Form. Various factors will be taken into consideration while determining the car's pricing. Regression The reason algorithms are utilized is that their output is a continuous value rather than a categorized value, which makes it easy to estimate the true cost of an automobile rather than just the range of prices.Additionally, a user interface that gathers input from users and shows car prices based on inputfrom users has been built. The market for old cars is expanding rapidly; in the previous several years, its market value has nearly doubled. The rise of internet portals like CarDheko, Quicker, Carwale, Cars24, and numerous others has made it easier for buyers and sellers to learn more about the patterns and trends that influence a used car's market value. Based on a certain set of features, machine learning algorithms can be used to anticipate an automobile's retail value. Various websites There isn't a single algorithm utilizedto determine the pricing because different companies use different algorithms to create the retail price of used cars. Without actually entering the details into the desired website, one can easily get a reasonable estimate of the price by training statistical models for price prediction. Kaggle generated the data set that was utilized in the prediction models [1]. 9104 automobile records with computed retail prices are included in the data.

The variables that are used are as follows:

3. Cost: The GM vehicles computed retail cost.
4. Mileage: The total kilometers driven by thevehicle driven;
5. Model: The particular models for eachautomobile;
   Fuel: The kind of fuel the car runs on, such as petrol or diesel.
6. Year: The year the actual owner of the car purchased it.

2. LITERATURE REVIEW

2.1 Using Machine Learning Techniques toPredict Used Car Prices

In this work, we examine the usage of supervised machine learning methods to forecast Mauritius used vehicle prices The forecasts are predicated on historical information gathered from daily publications. The predictions have been

made using a using a variety of methods, including decision trees, k-nearest neighbors, naïvebayes, and multiple linear regression analysis. Next, the Predictions are compared and assessed to see which yield the best results it proves to be quite difficult to solve a seemingly simple problem with excellent precision. Each of the four approaches delivered performance that was similar. We plan to make the forecasts using advanced algorithms infuture.

2.2 Machine Learning-Based Car PricePrediction

It necessitates observable effort and professional expertise in the subject, automobile price prediction research has garnered a lot of interest. A large variety of different characteristics are looked at in orderto provide an accurate and trustworthy prediction. We employed the Artificial Neural Network, Support Vector Machine, and Random Forest machine learning techniques to create a model that forecasts used automobile prices in Bosnia and Herzegovina. Nonetheless, the a forementioned methods were used in group projects. A web scraper built in PHP wasused to get the data for the forecast from the website autopijaca.ba online. The optimal algorithm for the given data set was then determined by comparing the respective performances of the several methods. Additionally, Test data were used to evaluate the model, and an accuracy of 87.38% was found.

2.3 Regression Models for Predicting Used Car Prices

For this investigation, we carried out a comparison analysis on Regression using supervised machine learning models is executed. Every model is trained with used automobile market data gathered from German online retailers. With mean absolute error (MSE) = 0.28, gradient enhanced regression trees perform the best as a consequence. Then camemultivariate linear regression (MSE = 0.55), and random forest regression (MSE = 0.35). Use qualified qualitative data and a knowledge- based system to forecast car prices. Summary: The machine learning process of pricing cars is closely linked to the expert system's knowledge acquisition process. Lately, the most common method of gaining knowledge has been the laborious recommendation posting process for purchasing an automobile. or selling on websites for online markets. We can categorize the data into two groups once ithas been found: structured and unstructured, which both need knowledge-based analysis. The methods for meaning extraction, data inference, and rules for qualitative data will all be covered in this study. The current study's primary goal is to investigate various automotive data types in order to develop an automated method for predicting car pricing.

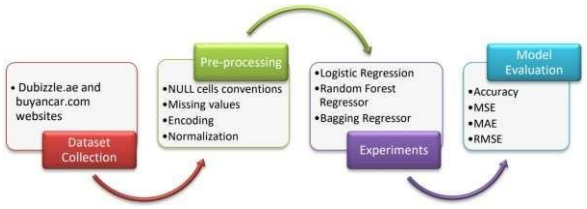
3. OBJECTIVE OF CAR PRICE PREDICTION SYSTEM

Creating a machine learning model that can reliably forecast an automobile's price is the main goal of a project that aims to predict car prices. on a range of characteristics and elements. With the help of this initiative, buyers, sellers, and other industry participants should be able to make more informed judgements about car price

1. Prediction accuracy create a model that can reliably estimate automobile prices based on many features such as brand, model, mileage, year, condition, etc.
2. Assistance in Making Decisions: Help vendors set competitive prices for their listings and buyers in determining fair prices for cars they are interested in buying.
3. Efficiency: Save time by providing a prompt estimate of car costs. in contrast to manual appraisal procedures.
4. Insight Generation: Learn about thevariables that have a big impact on auto costs and comprehend market dynamics to make smarter decisions.
5. Scalability: Create a model that is flexible and scalable for various automotive industry conditions, enabling it to be utilized for a broad variety of automobile listing. The ultimate goal is to use machine learning techniques to develop a trustworthy tool that, by projecting fair values based on numerous variables, improves comprehension and efficiency of the car buying and selling process.

4. METHODOLOGY

This method's primary objective is to provide consumers with an accurate estimate of the amount that must be paid for the specified car.The model might provide the buyer with a list of options for different cars depending on the specifics of the vehicle that the buyer desires. The system helps the customer get enough information to enable him to make a decision. The market for secondhand cars is growing exponentially, and car sellers may benefit from this by underpricing their vehiclesin order to take advantage of the demand. Consequently, there is a need for a system thatcan forecast an automobile's price based on itsspecifications and also account for the expenses of rival models. Our method closes the gaps by giving an estimate of the car's value based on the most advanced price prediction system to buyers and sellers.



4.1 LINEAR REGRESSION

The method of employing independent factorsto predict a dependent factor is calledregression. Usually, the method is used for estimating and computing the correlations between the independent and dependent variables. The regression models determine relationship between independent and dependent variables. Regression analysis that involves only one independent variable and a linear connection between the independent (x) and dependent (y) variables is known as linear regression. The red line in the accompanying graph is referred to as the best fit straight line. Plotting a line that mostaccurately predicts the data points given the data points is our goal. The line can be represented by the linear equation given below y is equal to

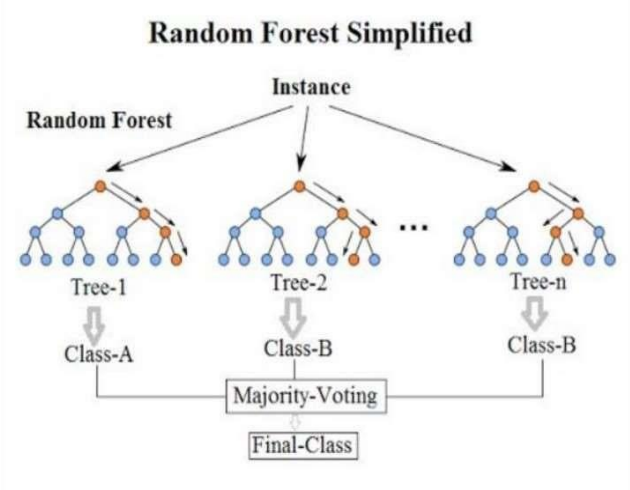
a0 + a1 * x.

4.2 COST FUNCTION

The greatest viable values for a0 and a1 are found using the cost function, and these valuescan be utilized to create the most feasible fit line for the dots that are plotted against the data. Since our goal is to find the optimal values for a0 and a1, we use this to formulate a minimization problem in which our goal is to reduce the difference between the expected (anticipated) and actual (truth) values. We employ the previously stated function to minimize. The difference between the expected and ground truth values is used to calculate the error difference. The error difference is squared, the datapoints are added together,and the sum is divided by the total number of data points. You are now presented with the average squared error for each of your data points. Consequently, the Another term for thiscost function is the Mean Squared Error (MSE) function. The values of a0 and a1 will now be adjusted using the MSE function until the MSE-value achieves the minimum.

a. RANDOM FOREST

A popular supervised machine learning algorithm for classification and regression issues is called random forest. It develops decision trees on several samples and uses the average in the case of regression and the majority vote for classification. When interpretability is not a big problem and we have a vast dataset, Random Forest is a good fit.Decision trees are far simpler to comprehend and analyses. A random forest is more challenging to read since it mixes several decision trees whether there is a linear or exponential.



4.3. Lasso Regression

Lasso Regression On the training data set, we first use Lasso regression to identify the subsetof attributes that result in the best/least sum of squared mistake when making a price prediction. Ten-fold cross-validation is used to "lasso" the ideal subset of attributes. L1 regularization is used.

LAR Selection Summary

Step	Effect Entered	Number Effects In	CV PRESS
0	Intercept	1	5.47454E10
1	Cylinder_8	2	2.94477E10
2	Make_Cadil	3	2.54198E10
3	Type_Conve	4	1.70491E10
4	Make_SAAB	5	1.0723E10
5	Liter	6	5716511688
6	Model_XLR-V8	7	4455568638
7	Cruise_0	8	4462900633
8	Mileage	9	3141496232
9	Make_Chevr	10	3102016376
10	Model_Corvette	11	2836230790
11	Type_Wagon	12	2434477976
12	Model_STS-V8	13	2241697550
13	Model_Park Ave	14	2022249850
14	Model_9_5	15	2016211182
15	Trim_SS Sedan 4D	16	1870278120
16	Model_STS-V6	17	1767874656
17	Model_Grand Pr	18	1706384400
18	Model_CST-V	19	1594252419
19	Trim_Arc Sedan 4	20	1537014571
20	Trim_Arc Conv 2D	21	1432488055
21	Trim_GT Coupe 2D	22	1357217957
22	Trim_Special Ed	23	1341923945
23	Model_9-2X AWD	24	1207730522
24	Model_Deville	25	1192390216
25	Model_Malibu	26	1140423212
26	Model_Lacrosse	27	1079213317
27	Model_Vibe	28	1057127900
28	Trim_SS Coupe 2D	29	991121705
29	Trim_SVM Hatchba	30	968461697
30	Trim_Sedan 4D	31	906490734
31	Model_Cavalier	32	900413251
32	Model_AVEO	33	895244688
33	Trim_CXS Sedan 4	34	886133715
34	Model_Sunfire	35	866898872
35	Trim_Custom Seda	36	849386379
36	Trim_SVM Sedan 4	37	842938940
37	Model_Grand Am	38	834462359
38	Trim_LS Coupe 2D	39	820399420
39	Trim_LT Coupe 2D	40	809493685
40	Trim_GXP Sedan 4	41	785550309
41	Model_Century	42	780838005
42	Model_L Series	43	757522044
43	Model_G6	44	734553436
44	Trim_GTP Sedan 4	45	709162714
45	Trim_Limited Sed	46	691701586
46	Trim_AWD Sportwa	47	687992677
47	Trim_CXL Sedan 4	48	680838392
48	Trim_DTS Sedan 4	49	674240643
49	Leather_0	50	664002260
50	Trim_Arc Wagon 4	51	662795933

51	Trim_DHS Sedan 4	52	618387221
52	Trim_GT Sportwag	53	612366627
53	Make_Satur	54	609667982
54	Trim_LS Sport Co	55	606909142
55	Model_Classic	56	604819404
56	Trim_SLE Sedan 4	57	601079687
57	Sound_0	58	596479662
58	Trim_GT Sedan 4D	59	597018312
59	Trim_Linear Conv	60	599172036
60	Trim_LT Sedan 4D	61	597976402
61	Trim_Coupe 2D	62	595494703
62	Trim_Conv 2D	63	587687756
63	Trim_LT MAXX Hba	64	586066103
64	Model_9_5 HO	65	585734528
65	Trim_MAXX Hback	66	585894705
66	Trim_LS Sedan 4D	67	586199168
67	Model_Monte Ca	68	582745854*
68	Trim_Quad Coupe	69	583208092
69	Trim_LT Hatchbac	70	583490938
70	Trim_LS Hatchbac	71	585911868
71	Trim_Aero Wagon	72	586112390
72	Trim_LS Sport Se	73	586892601
73	Trim_Aero Sedan	74	587208896
* Optimal Value of Criterion			

5. NOVELTY OF CAR PRICE PREDICTION

The automobile price prediction project is interesting because it creatively uses data analytics and machine learning to anticipateautomotive pricing with precision. This project utilizes advanced algorithms to provide real-time estimations for a variety of car variables, including brand, model, miles, year, and condition. The algorithms are trained on large datasets to accommodate the dynamic nature of the automotive industry. Adaptability is increased by its ability to tailor models to different automotive types and local market preferences. Accuracy is also improved and new insights into changing market trends are fostered by ongoing refining through the integration of current data. This project is noteworthy for its capacity to provide insightful and predictive data that is based on data, so enabling giving stakeholders access to a user-centric tool that improves decision- making in the fiercely competitive and dynamic automotive sector.

6. ADVANTAGES AND DISADVANTAGES OF CAR PRICE PREDICTION SYSTEM

A. Machine learning is used in the automobile price prediction project. Algorithms to deliver a thorough and effective solution that benefits the automobile market's suppliers as well as purchasers. This research provides precise car pricing estimates by examining a number of variables, including brand, model, mileage, year, and condition. This helps decision- makers make well-informed choices by reducing the time and effort required formanual valuation procedures. By taking into account a variety of factors that affect prices, it helps sellers set competitive prices, helps buyers negotiate fair bargains, and improves market understanding through insights into relevant dynamics. This project, which is scalable and constantly improvable, makes cars more accessible and guarantees ongoing improvement, which simplifies the car-buying and selling process for all parties involved.

B. Predicting automobile prices has a number of intrinsic limits despite its benefits. Reliance on the comprehensiveness and quality of the dataset poses a serious problem, since faulty orInaccurate forecasts resulting from biassed data may affect user confidence and decision- making. Complex machine learning models may be difficult to use in contexts with limitedresources both in terms of compute demands and deployment. Additionally, the consistency and forecast accuracy of the model may be compromised by the volatility of outside factors that affect car pricing, such as changes in the economy, market trends, or unanticipated events. Additionally, certain complex models' opacity could lead to a lack oftransparency, which could lower user confidence and comprehension of how predictions are made. The maintenance of data quality, ongoing adaptation, and initiatives to improve model interpretability are essential for minimizing these restrictions and guaranteeing the project's applicability and relevance in the ever-changing automotive industry

7. CASE STUDY OF CAR PRICEPREDICTION

A used auto e-commerce company wanted to improve its pricing approach, so it put in place a machine learning-based car system for predicting prices.Utilizing a broad dataset that included vehicle attributes including make, model, mileage, year, and condition, they used ensemble learning techniques to create a prediction mode.

The technology used extensive data analysis tooffer sellers with predicted listing pricing. Sales grew and customer satisfaction rose as a resultof this initiative's streamlining of the sales process. Because of the model's accuracy in determining fair prices, the platform's overall confidence and openness increased, giving buyers and sellers alike a more dependable and effective marketplace experience.

8. CONCLUSION

Sales of used automobiles are rising globally due to rising new car prices and consumers' inability to afford them. A used car price prediction system that accurately assesses the car's worthiness based on a range of factors is therefore desperately needed. The suggested method will contribute to the precise estimation of used automobile prices. Three distinct machine learning algorithms— decision trees, random forests, and voting classifiers—are compared in this work.

9. FUTURE OF CAR PRICE PREDICTION

This machine learning model might eventually link to different websites that offer real-time data for price prediction. Additionally, we might include a significant amount of historical car price data to help the machine learning model become more accurate. As a user interface, we can create an Android app to communicate with users. We intend to carefully consider the architecture of deep learning networks, employ adaptive learning rates, and train on data clusters in order to improve performance. Instead of the entire datasets

10. TECHNICAL ARCHITECTURE OF CAR PRICE PREDICTION

What makes up an automobile price predictionsystem's technical architecture is Dataset creation comes after gathering information from multiple sources. This data is used to feedmachine learning models, which use algorithms learned on processed data, such as Random Forest, Gradient Boosting, or Neural Networks. The trained model is made available via an API or web service that is housed on a server or cloud platform and can be accessed via an intuitive front-end interface that allows users to enter information about their cars andget price estimates. Ongoing improvements are facilitated via user input systems, model performance review, and continuous monitoring. Ensuring scalability, user privacy, and data security, the system combines multiple components to produce an accurate, efficient, and user-friendly car price forecast tool.

11. REFERENCES

The "CAR PRICE PREDICTION" project's successful development depends on a strong base of research, current knowledge, and pertinent sources. The following important sources of information and references will help to guide and assist our project:

[1] “Predicting the Price of Used Cars using Machine Learning Techniques" by SameerchandPudaruth; (IJICT 2014)

[2] "Car Price Prediction Using Machine Learning," by Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, and Jasmin Kavcic; (TEM Journal)

[3] Ning Sun, Hongxi Bai, Yuxia Geng, and Huizhu Shi, "BP Neural Network Theory-Based Price Evaluation Model in Used Car System," Hai University Changzhou, China

[4] “Prediction of Prices for Used Car by using Regression Models," by Nitis Monburinon, PrajakChertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, and Pitchayakit Boonpou (ICBIR 2018)

[5] “Prediction car prices using qualify qualitative data and knowledge-based system” (Hanoi National University) by Doan Van Thai, Luong Ngoc Son, Pham Vu Tien, Nguyen Nhat Anh, and Nguyen Thi Ngoc Anh

[6] Journal of Statistics Education, 16:3 Shonda Kuiper (2008), Introduction of Multiple Regression: How much is your car worth?

[7] Bias vs Variance Decomposition for Regression and Classification, Guerts P. (2009). In: Rokach L., Maimon O. (eds) Information extraction and discovery Guidebook. Boston, MA 8 Springer. Regression Shrinkage and Selection through the Lasso, Robert T. In: Royal Statistical Society of Great Britain Journal, Series B (Methodological)