Used Indian Car Price Prediction using Machine Learning

Sancheet Kumar Baidya [1]
School of Computer Science and Engineering
(PG Scholars)
Vellore Institute of Technology
(VIT University, Vellore)
Tamil Nadu, India
sancheet.kumar2022@vitstudent.ac.in [1]

Saman Qaiser [2]
School of Computer Science and Engineering
(PG Scholars)
Vellore Institute of Technology
(VIT University, Vellore)
Tamil Nadu, India
saman.qaiser2022@vitstudent.ac.in [2]

Anushree Paul [3]
School of Computer Science and Engineering
(PG Scholars)
Vellore Institute of Technology
(VIT University, Vellore)
Tamil Nadu, India
anushree.paul2022@vitstudent.ac.in [3]

Dr. R Saravanan [4]
School of Computer Science and Engineering
(Professor Higher Academic Grade)
Vellore Institute of Technology
(VIT University, Vellore)
Tamil Nadu, India

rsaravanan@vit.ac.in [4]

***Abstract*—** The manufacturer determines the cost of a new automobile, with taxes paid by the government. However, sales of used automobiles are rising due to new car price increases and consumers' inability to afford them. A system has been created to accurately assess the value of the vehicle using regression algorithms and a user interface that accepts input from any user.

Keywords: Used car price prediction, Regression Algorithms, Machine Learning, Deep Learning, Data Analysis, Data Visualization.

I. INTRODUCTION

Over 42 lakh consumers have purchased used cars in India o ver the course of the last year, reflecting a jump in demand.
The current ratio of new to used vehicles is 1:2.2, which me ans that when 10 new vehicles are sold, 22 used vehicles re main on the market.
Calculating a car's resale value is not a simple process becau se it depends on a number of different factors, including the vehicle's age, make, origin, mileage, horsepower, fuel econo my, interior style, braking system, acceleration, safety rating , size, number of doors, paint colour, weight, and other optio ns.
The pricing is also influenced by unique elements like major accidents and the way the automobile feels and looks.
Unfortunately, the buyer must base his or her decision to pur chase at a certain price just on a small number of these facto rs since information on all of them is not always easily avail able.

II. LITERATURE SURVEY

The first research uses artificial neural networks and machin e learning to estimate used automobile prices.
Deep Neural Networks, Linear Regression, Ridge Regressio n, Lasso Regression, and Random Forest Algorithm are five machine learning algorithms that are employed.
Ridge regression has the lowest accuracy, whereas Random Forest Algorithm has the greatest.
The second article uses Random Forest and LightGBM to an ticipate prices. Gradient boosting regressor had a high R-squared score and low root mean square error, according to a Comprehensive Analysis of Machine Learning Methods fo r Predicting the Resale Price of Used Cars.
The study's findings demonstrated that the gradient boosting regressor excelled every model that was put to the test, havi ng the greatest R2 and the lowest root mean squared error.
Four different machine learning approaches were used to for ecast the cost of used vehicles in Mauritius, with a mean err or of Rs. 27, 000 for Nissan cars and Rs. 45, 000 for Toyota cars.
The main shortcoming of naive bayes accuracy, which varie d between 60 and 70%, was their inability to handle output c lasses having numerical values. The small sample size of records employed in this research is its primary shortcoming. Future studies should include
more advanced techniques including genetic algorithms, fuz zy logic, and artificial neural networks..

III. DATASET

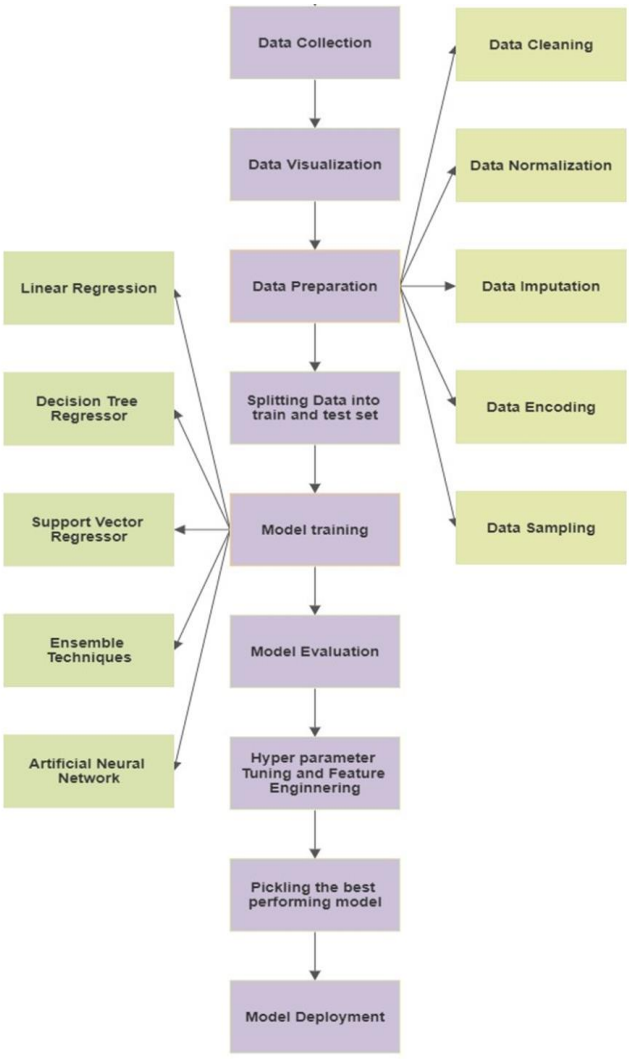
We are going to utilise a real dataset for the project. Information on around 21000 used automobiles was c ollected from www.cardekho.com and included in this dataset..Link of dataset: https://tinyurl.com/3jjvh9k2

The above link has 3 datasets. Out of them we are going to use the cardekho_imputated.csv dataset. It has 16 columns which are

- **sno:** Serial number of rows
- **car_name**: Full name of the car (brand + model)
- **brand:** Manufacturer of the car

- **model:** Model name of the car
- **min_cost_price**: Price of the base model
- **max_cost_price**: Price of the top model
- **vehicle_age:** Number of years for which the vehicle was owned
- **km_driven:** Total kilo meters driven by the car
- **seller_type:** 'Individual' or 'Dealer' or 'Trustmark Dealer'
- **fuel_type:** Petrol or Diesel
- **transmission_type:** Manual or Automatic
- **mileage:** Number of kilometres travelled per litre of fuel
- **engine:** Capacity of the engine in cc (cubic centimetres)
- **max_power:** Torque of the engine in Newton metres.
- **seats:** Total passenger carrying capacity
- **selling_price:** Listed price of the car

IV. METHODOLOGY



The goal of this research is to create machine learning models that can reliably forecast a used car's price based on its attributes.
It assesses and chooses the best machine learning techniques, such as Lasso Regression, Ridge Regression, and Linear Regression.
Additionally, a user interface has been created to show a car's price based on user inputs.
There are two primary phases in the system: training and testing. Different algorithms were compared for accuracy and the best one was chosen.

V. DATA PREPROCESSING

The procedure required to analyse the data from the Pandas library are the most crucial information in this work.
Utilising the head () function to get a quick overview of the data is one of these procedure, dropping the unnamed column, checking for null values and data types, dropping the car_name column, forming a new column called "avg_cost_price", performing descriptive statistics for categorical and numerical columns separately, and forming two separate data frames one for categorical and one for numerical columns.

VI. DETECTION AND REMOVAL OF OUTLIERS

It is important to remove outliers in a dataset as they can cause overfitting and erroneous predictions. In the given dataset, there were many outliers which were removed with the help of exploratory data analysis. These outliers included selling price more than 2 crore, average cost price more than 8 crores, km driven more than 10 lakhs, number of seats in a car less than 2, max power greater than 500, and engine cc more than 6000.

VII. ANALYSIS OF CATEGORICAL FEATURES

There are two types of categorical variables: nominal (having no specific order) and ordinal (having some o rder).Ordinal data consists of a collection of orders or scales, while nominal data just contains the name vari able without any numerical values.
A list of patients could include information like a pers on's blood sugar level, which can be categorised as hi gh, low, or medium.

VIII. PLOTTING CORRELATION

We first need to locate the correlational matrix, which provides information on the correlation between different characteristics, in order to do bivariate

analysis between numerical columns. It may be seen using a heatmap.

IX. INFERENCE FROM EXPLORATORY DATA ANALYSIS

Exploratory data analysis was done on the dataset to gain valuable insights. It included imputation of missing values, changes of data types, removal of unnecessary columns, and univariate and multivariate analysis. Outliers were removed and a baseline model was formed to analyze performance on different Machine Learning algorithms. Hyper parameter tuning was discussed to improve model performance.

X. MODELLING AND ERROR ANALYSIS

The processes needed to develop a baseline model for a machine learning project are the sections of this article that are most crucial to understanding.

These procedures include categorical variables being binary encoded, partitioning the data into train and test sections, building a baseline model, scaling features, evaluating the model, tweaking hyperparameters, and visualising the residual distribution.

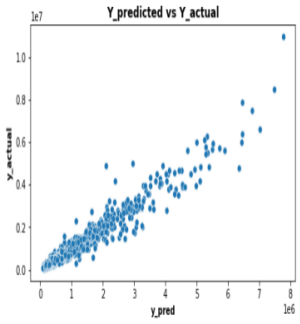
The performance of machine learning algorithms is calculated using the train-test split approach when they are used to make predictions on data that was not used to train the model.

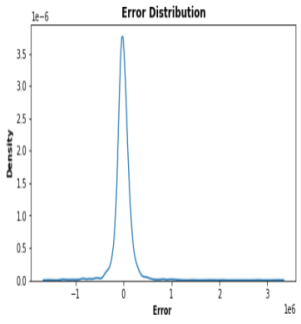
A baseline model is a straightforward model that serves as a reference for machine learning projects and offers a wealth of data that may guide the project's subsequent stages.

One of the most important phases in the preprocessing of data prior to building a machine learning model is feature scaling since it may determine whether a model is strong or weak.

XI. COMPARISON OF THE PERFORMANCE OF VARIOUS MODELS

After comparing the various models, we can conclude that the Radom Forest Regressor model gave the best performance in terms of cross validation score, R2 score and Median Absolute Error.





Some other machine learning models like Bagging Regressor, Extra Forest Regressor, Gradient Boosting Regressor also gave very good performance comparable to the Random Forest Regressor model.

	Regressor	CV_Score	Median_Abs_Error	R2_Score
0	BaggingRegressor	0.94	68800.00	0.93
1	RandomForestRegressor	0.95	67720.83	0.93
2	ExtraTreesRegressor	0.95	112965.00	0.70
3	AdaBoostRegressor	0.77	318242.86	0.71
4	GradientBoostingRegressor	0.94	86684.65	0.91
5	GradientBoostingRegressor	0.94	86684.65	0.91
6	StackingRegressor	0.91	106705.00	0.84
7	VotingRegressor	0.84	349087.80	0.69
8	Artificial Neural Network	NaN	124566.62	0.75

ACKNOWLEDGMENT

I would like to express my heartfelt gratitude to all those especially *Dr. Saravanan R* and those who have supported and guided us throughout this journey. Your unwavering encouragement and belief in my abilities have been instrumental in my accomplishments. I am truly grateful for your invaluable contributions and the lessons learned along the way. Thank you!

REFERENCES

[1] Pudaruth, S. (2022) Making use of machine learning techniques to predict the cost of used cars.

[2] Kumar, S., Kaur, D., & Parvez, A. (2020). Prediction of prices car price prediction with machine learning.

[3] Beni-Hssane, A. H. (2022). Used Car Price Prediction using Machine Learning (p. 282). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ISIVC54825.2022.9800719

[4] Lakshmi ,V. J.. (2022). Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning (p. 282). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICCCI54379.2022.9740817

[5]Virpariya, S., V. R. (2022). Institute of Electrical and Electronics Engineers Inc. (p. 282). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/CISES54857.2022.9844350

[6] Nagasindhu, A., N. C (2022) Using Machine Learning, Predict Second Sale Car Price.