International Journal of Innovative Research in Engineering
Volume 4, Issue 2 (March-April 2023), PP: 364-369.
www.theijire.com ***ISSN No: 2582-8746***

Car Resale Value Prediction Using Applied Data Science

Dhaneshkumar K[1], Naveen K[2], Dharmadurai D[3], Gowtham M[4]
[1, 2,3,4]*Computer science and engineering, The Kavery Engineering College, Tamilnadu, India.*

How to cite this paper:
Dhaneshkumar K[1], Naveen K[2], Dharmadurai D[3], Gowtham M[4], "Car Resale Value Prediction Using Applied Data Science", IJIRE-V4I02-364-369.

Abstract: *In a difficult economic environment, it is possible that used and imported pre- owned (reconditioned) cars will sell more. In many developed countries, leasing a car is more common than outright purchase. The buyer will have the option of buying the vehicle for its residual value or expected resale value when the lease term is up. Therefore, from a business perspective, it is in the best interest of sellers and financiers to be able to forecast the salvage value (residual value) of vehicles. We suggested a system that forecasts the resale worth of the car using regression algorithms, making it intelligent, flexible, and effective. It is necessary to build a regression model that takes into consideration the main factors that might affect a vehicle's resale value.*

Key Word: *Cars, Price, Analysis, Prediction, Features, Python, Algorithm, Regression.*

INTRODUCTION

In this endeavor, we created algorithms and techniques to predict the resale value of cars while taking into consideration a variety of car features. In a nutshell, auto resale Value prediction allows the user to estimate the car's resale value based on a variety of factors, including the mileage and fuel type.

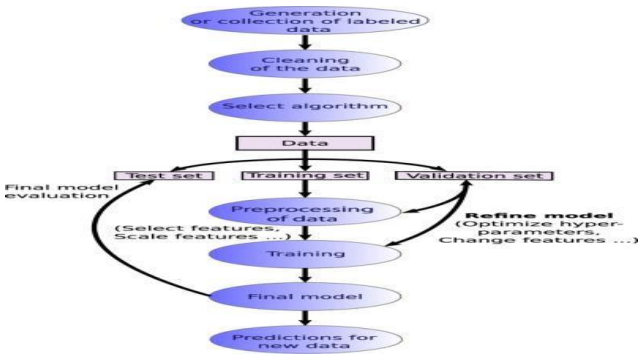
I.PROJECT OVERVIEW

Car Resale Value Prediction:

A car resale value prediction system's objective is to forecast the used car market's accurate valuation, allowing users to sell their vehicles online, with an unbiased assessment and no need for human involvement. Due to the lack of data, the system only takes a small number of features into account when predicting the car's resale worth. The current method, being an online system, does not take into consideration any physical damage to the car's body or engine when calculating its resale value. Data collection and prediction using machine learning-based algorithms make up the novel system we developed.

II.PURPOSE

The main goal of creating a system to estimate car resale value is to practice Python and Data Science. The user-provided characteristics form the basis of the system that predicts the amount of resale value for automobiles. When the user fills out the form with information about the vehicle, the system makes a resale value prediction.



III.OBJECTIVE

The limited amount of information supplied determines how well the system can forecast the value of the car's resale.

The current system does not take into account any physical damage to the car's body or engine when calculating its resale value because it is an online system.

Data collection and prediction using machine learning-based algorithms make up the novel system we developed.

Using web scraping libraries, data from the pages of the cars24 website were gathered. The script runs and uses the URL to get the HTML div's data. The URL must be entered by users. We have gathered information thus far by entering URLs for Swift Dzire automobiles in 5 cities.

IV.TECHNOLOGY USED:

Python is mostly utilised in this project to implement machine learning techniques because it has a lot of built-in tools in the form of packaged libraries and modules.

The libraries used during the project implementation are the following:
Pandas: One of the most popular Python libraries in data science is Pandas. It supports a number of different structures and data analysis tools, all of which are simple to use and offer a high level of performance.

NumPy: An open-source Python tool called NumPy performs extremely rapid mathematical operations on matrices and arrays. "Numeric Python" or "Numerical Python" is what NumPy stands for. The Python Machine Learning Ecosystem is made up of NumPy and other machine learning modules like Scikit-learn, Pandas, Matplotlib, etc.

Matplotlib: Plotting the essential elements of data visualization, such as bars, pies, lines, scatter plots, etc., is the main use of Matplotlib. For Python data visualization, it is a graphics package that integrates seamlessly with NumPy and Pandas libraries. Thepyplot module closely mimics the plotting functions of MATLAB.

Seaborn: A module called Seaborn offers many visualisation patterns. It has simple syntax and easy-to-understand themes by default. Seaborn's area of expertise is statistical visualization, which is used to describe the data distribution in addition to summarising data with the use of various visuals. By extending the Matplotlib module in Python, Seaborn makes it possible to create excellent graphics quickly and easily.

Scikit-learn: Through a uniform Python interface, the Scikit-learn module offers a number of learning techniques that are either supervised or unsupervised. Since SciPy is the foundation upon which Scikit-learn is constructed, SciPy or Scientific Python must first be installed before using the scikit-learn package. The robustness and support required for use in production systems is the library's vision.

Pickle: By implementing binary protocols, the pickle module is used to serialise and de-serialize a Python object structure. The conversion of the Python object hierarchy into a byte stream is known as "pickling," and "unpickling" is the opposite of the aforementioned procedure. Serialization, marshalling, and flattening are other terms for pickling. These technologies were used to implement the web application.

HTML: Hyper Text Markup Language's abbreviation For developing and producing texts that would be viewable on any web browser, it is a common markup language. It can also be supported by tools like JavaScript, a programming language, and Cascading Style Sheets.

CSS: It stands for Cascading Style Sheets, a type of style sheet language used to specify how a document written in a markup language like HTML should be presented.

Flask: It is a Python-based microweb framework that is categorised as a microframework because it does not require any specialised libraries or tools. Flask lacks the database abstraction layer, form validation, and other similar components that rely on third-party libraries for functionality.

Jsonify: It depends on the flask. using Flask's json module. Jsonify performs the serialisation of data to the JavaScript Object

Notation (JSON) format and wraps it in a response object with the json/application mimetype. The flask module enables direct import of Jsonify. Requests: Using Python, this module enables users to send HTTP requests. A Response Object is produced in response, containing information like as content and encoding status.

V.PROJECT FLOW

Find below the project flow to be followed while developing the project:

- Download the dataset.
- Preprocess or clean the data.
- Analyze the pre-processed data.
- Train the machine with preprocessed data using an appropriate machine learning algorithm.
- Save the model and its dependencies.
- Build a Web application using a flask that integrates with the model built.

VI.TEST CASES

- **Missing values**

Four feature inputs are needed by the trained ML model to predict the output. If that doesn't work, the model returns an incorrect input error. The user must fill out every field because every one of the html form's necessary fields has been marked as such using CSS. Output: The user must fill out every field; otherwise, the form will display a warning that states, "This field needsto be filled." Therefore, there can be no model prediction mistakes.

- **Invalid Input**

For all 4 features, the trained ML model simply needs numerical input. Thus, the model may generate an error if the user inputs symbols like commas. Preprocessing script is deployed in the backend to eliminate all undesirable characters, such as commas and whitespace, in order to ensure that the model receives the input it needs.

Output: The python preprocessing script will ensure that the model receives the correct input and can make accurate predictions.

- **Unseen year of purchase**

With data from vehicles acquired between 2011 and 2020, the model was trained. Since that data is fairly recent and unfamiliar to the model, it may become confused if the user enters information about a car they bought after that, in 2021. Output: Because the model was trained using the boosting approach, it provides results that are fairly accurate, with an RMSE of only about 65,000 INR.

Requirements

1. Hardware Requirements:-

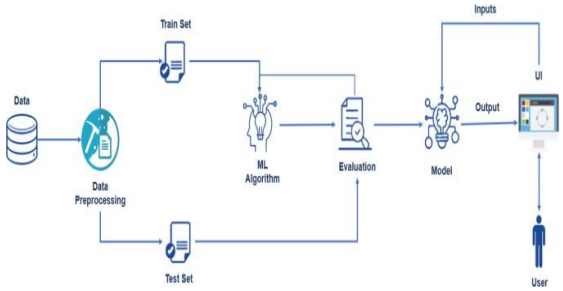
- 2GB RAM (minimum)
- 100GB HDD (minimum)
- Intel 1.66 GHz Processor Pentium 4 (minimum)
- Internet Connectivity

2. Software Requirements:-

- WINDOWS 10 or higher
- Python 3.6.0 or higher
- Visual Studio Code
- Flask Framework
- HTML – CSS
- Chrome , Micro Soft Edge.

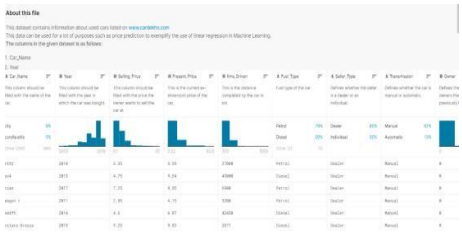
VII.PROJECT DESIGN

Data Flow Diagram



Methodology:

Data Gathering: The data's original source is the Kaggle.com web service, which makes available the Cardekho automobile dataset for car sales and purchases. The dataset provided the features listed below: Name and year of the vehicle, the selling price, the current price, and the number of miles driven fuel types include gasoline, diesel, and compressed natural gas, and seller types include dealers and private individuals. Transmission: Manual or Automatic, Owner (No. of previous owners).



Creating Environment: Anaconda prompt is used to construct an environment. The other default (base) environments and any other previously established environments would be separated from our project space by this environment. This step is advantageous because we may manually install all the packages, libraries, and modules we need in the environment we generatedin this way. In such a setting, we can make the changes necessary to meet our needs.

Data Reading: The first step is to import and read the csv file for the study. The null values, shape, columns, numerical and categorical features, dataset columns, unique values of each feature, data information, etc. are all carefully read for the dataset.

Data Pre-processing: For easier understanding, some of the features in the data were renamed (Present Price = Initial Price, Owner = Previous Owners), and some others that weren't necessary for study were removed. Data for Exploration In order to summarise the key features of the data, we analyse it using statistical graphics and other visualisation techniques. Top Selling Vehicles, Year vs. Number of Available Vehicles, Selling Price vs. Initial Price, Vehicle Fuel Type, Transmission Type, Seller Type, Age, Selling Price vs. Age, Selling Price vs. Seller Type, Transmission, Fuel Type, Selling Price vs. Previous Owners, InitialPrice vs. Selling Price, Selling Price vs. Kilometers are just a few examples of the graphs and charts that are available. To gain a deeper insight, plots like as driven, pairplot, heatmaps, etc.

Train-Test Split: We continue by dividing the dataset into training and testing data when the allocation of dependent and independent features is complete. We utilise 80% of the data to train our model and 20% of the data to test it.

```
# Splitting into test and train data
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=1)
print("x train: ",x_train.shape)
print("x test: ",x_test.shape)
print("y train: ",y_train.shape)
print("y test: ",y_test.shape)
```

```
x train:  (240, 8)
x test:  (61, 8)
y train:  (240,)
y test:  (61,)
```

Model Building: Following the Train-Test split, data modelling is carried out, which is where the model-building process starts. For further implementation, the model is defined along with a few parameters. Once the model is prepared, several algorithms are then used to get the findings that were produced by them. After model creation, the following algorithms are used for the predictiveanalysis.

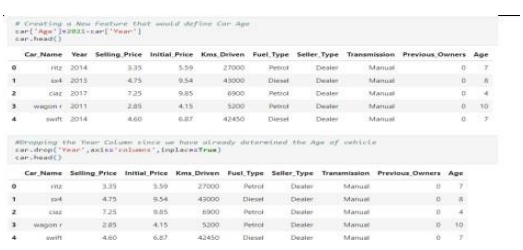
Linear Regression It is a linear strategy for simulating the interactions between a scalar response and dependent and independent variables in the field of statistics. In linear regression, model parameters that are unknown are estimated from the data and functions like the linear predictor are used to model relationships.

Lasso Regression: It is a specific kind of linear regression that employs shrinkage, which means that the values of the data are shrunk towards the center, or more simply, the data's mean. Simple and sparse models with fewer parameters are supported by the Lasso technique. This regression is best suited for any model that exhibits a significant degree of multicollinearity. This model can also be used if it's necessary to automate the selection of variables or the deletion of parameters during the model selection process.Least Absolute Shrinkage and Selection Operator is referred to as "LASSO."

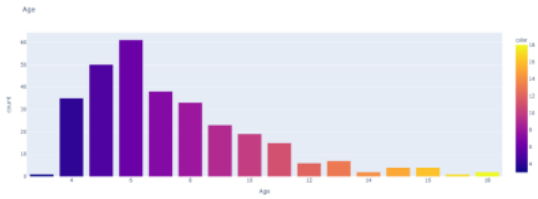
Random Forest Regression: Random-forest is a supervised learning algorithm since it uses the ensemble learning approach for classification and regression. The trees in random forests are parallel to one another and do not interact as they grow. A meta- estimator called random forest compiles the outcomes of numerous predictions. Additionally, it aggregates different decision trees with the aid of various adjustments.

Decision Tree: This approach is used to create tree-structured regression and classification models. A dataset is divided into smaller subsets, and at the same time, an incremental decision tree is also generated using it. Decision nodes or leaf nodes are the tree's final results. The decision tree construction algorithm uses a top-down greedy search to go through all of the tree's potentialbranches without going backwards.

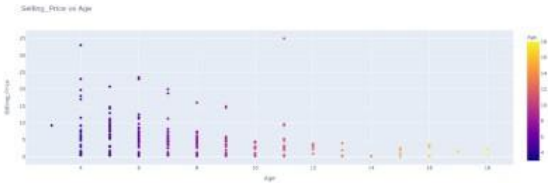
VIII.IMPLEMENTATION



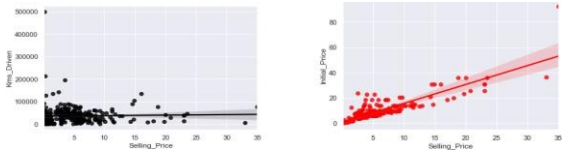
Exploratory Data Analysis: Data for Exploration In order to summarise the key features of the data, we analyse it using statistical graphics and other visualisation techniques. To better comprehend the dataset and the relationships between its properties, various graphs and charts are presented. Vehicle Age and the Number of Vehicles: The following bar graph shows the number of vehicles for a specific age.



Comparing the selling price of each vehicle with its age: The selling price and age of a specific automobile are shown in the accompanying graph. And it's simple to get the conclusion that a vehicle's selling price is high given its young age.



The following graph shows the direct proportionality between the original price and the selling price, indicating that a higher initialprice would also result in a higher selling price.



Comparison of Kilometers Traveled and Selling Price: The graph above demonstrates that a car with a high mileage will sell for less money than one with a lower mileage.

Hot Encoding: The categorical variables in the dataset are dealt with using the one hot coding technique. With a binary column for each category or parameter, it yields either a sparse matrix or a dense array depending on the parameters. Our dataset had three categorical variables: Fuel Type, Seller Type, and Transmission. A binary representation of these variables is created after one hot encoding, meaning that for a car having a fuel type of diesel, the value of Fuel_Type_Diesel would be a binary 1 and the value of Fuel_Type_Petrol would be 0. The other categorical variables are used in the same way.

Model Building: Modeling is carried out where the process of developing the model begins, following the Train-Test separation of the dataset. For the final implementation, the model is produced together with a few arguments, including the algorithm, x train, y train, x test, and y test. Several algorithms are then applied to the fully developed model to provide the results.

Creating a Web Application: HTML and CSS are then used to build a web application. This makes it possible for any user to enter parameters and produce the anticipated selling price of a used car. The user can choose values for the characteristics like FuelType, Transmission Type, and Seller Type and input the desired values for parameters like Year, Initial Price (in Lakhs), Miles Driven, and Previous Owners. After entering the information, the user need only click the Selling Price button to see the final selling price of the used car for which the information was entered.

IX.EXPECTED OUTCOME

- Good at learning complex and non-linear relationships.
- Highly explainable and easy to interpret.
- Robust to outliers.
- No feature scaling is required.
- User Friendly.
- The Model is 80-90% efficiency.

X.CONCLUSION

A used car's price forecast is a challenging task because there are so many features and parameters to consider in order to get accurate findings. The first and most crucial step is data gathering, followed by preparation. Then a model for specifying algorithms and creating output was created.

When multiple regression algorithms were applied to the model, it was found that the Decision Tree Algorithm performed best. This was demonstrated by the algorithm's highest r2 score of 0.95, which simply indicated that the original vs. prediction line graph revealed that its predictions were the most accurate. Decision Tree not only got the greatest r2 score, but also the lowest Mean Squared Error and Root Mean Squared Values, showing that prediction errors are quite uncommon.

References

1. *Sameerchand Pudaruth, Computer Science and Engineering Department, University of Mauritius, Reduit, MAURITIUS. Predicting the Price of Used Cars using Machine Learning Techniques. International Journal of Information & Computation Technology, 2014.*
2. *Saamiyah Peerun, Nushrah Henna Chummun and Sameerchand Pudaruth, University of Mauritius, Reduit, Mauritius. Predicting the Price of Second-hand Cars using Artificial Neural Networks. Proceedings of the Second International Conference on Data Mining, Internet Computing, and Big Data, Reduit, Mauritius 2015.*
3. *Nabarun Pal(Department of Metallurgical and Materials Engineering, Indian Institute of Technology Roorkee, Roorkee, India), Priya Arora(Department of Computer Science, Texas A & M University Texas, United States), Sai Sumanth Palakurthy(Department of Computer Science and Engineering, IIT (ISM) Dhanbad, Dhanbad, India), Dhanasekar Sundararaman (Department of Information Technology, SSN College of Engineering, Chennai, India), Puneet Kohli (Department of Computer Science, Texas A & M University, Texas, United States). How much is my car worth? A methodology for predicting used cars prices using Random Forest. Future of Information and Communications Conference (FICC) 2018.*
4. *Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric, International Burch University, Sarajevo, Bosnia and Herzegovina. Car Price Prediction using Machine Learning Techniques. TEM Journal, February 2019.*
5. *Ashish Chandak , Prajwal Ganorkar , Shyam Sharma , Ayushi Bagmar, Soumya Tiwari, Information Technology, Shri Ramdeobaba College of Engineering, Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur. Car Price Prediction Using Machine Learning. India International Journal of Computer Sciences and Engineering, May 2019.*
6. *Pattabiraman Venkatasubbu, Mukkesh Ganesh. Used Cars Price Prediction using Supervised Learning Techniques. International Journal of Engineering and Advanced Technology (IJEAT), December 2019.*
7. *Laveena D'Costa, Ashoka Wilson D'Souza, Abhijith K, Deepthi Maria Varghese. Predicting True Value of Used Car using Multiple Linear Regression Model. International Journal of Recent Technology and Engineering (IJRTE). January 2020.S.E.Viswapriya, Durbaka Sai Sandeep Sharma, Gandavarapu Sathya Kiran. Vehicle Price Prediction using SVM Techniques. International Journal of Innovative Technology and Exploring Engineering (IJITEE), June 2020.*