**Predicting Old Car price using different Machine learning techniques and their comparison**

Ajay Kumar Verma (1900320400008)

Anuj Sharma (1900320400013)

Vijay Pandey (1900320400074)

Suryansh Kaushik(1900320400065)

Rajat Kumar Sharma (1900320400050)

15th December 2022

ABES Engineering College, Ghaziabad

Batch 2019-2023

**Abstract**

After the pandemic, it has been noticed that there is some declination in the level of world economy and due to which the **purchasing power of people has been decreased.** For those who wants to buy car and don’t have enough money to buy a new car, a used car can be a good option. The production of cars has been steadily increasing in the past decade, with over **70 million** passenger cars being produced in the year 2016 [1]. Since new buyers who don’t have enough knowledge of the market price of the old car which they are dreaming to buy so we need a platform which can help people in price prediction of used cars. So, that’s why there is a need of good car price prediction which can help people in making a correct decision in buying new cars [2]. The car price prediction can be done using different machine learning techniques. With the help of different machine learning models, effective decisions can be made based on the historical data present. Nowadays, this is one of the most interesting research area in machine learning. For implementing different models we should have some data parameters of cars such as: width, engine type, kms driven, model name, no. of wheels, driven wheels , colour and so on. This data can be taken from online websites such as Kaggle and we can collect it from daily newspapers also.

**Introduction**

The used car market is an ever-rising industry, **which has almost doubled its market value in the last few years**[3]. Car price can be determined by a number of distinct features and to predict the accurate price of car needs special expertise. Customers who are new buyers and don’t have enough knowledge how to know the actual price of a used car can be exploited. The prices of new cars in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes[4]. Due to lack of funds, the customers are incapable to buy a new car and hence the used cars sales are on a global increase. The sellers can exploit customers by setting the unrealistic prices of used cars and many falls into this trap. So it rises a need of a car price prediction model which can effectively determine the worth of a car using different variety of features. Due to the adverse pricing of cars and the nomadic nature of people in developed countries, the cars are mostly bought on a lease basis, where there is an agreement between the buyer and seller[4].

A lease is a binding contract between a buyer and a seller (or a third party –

usually a bank, insurance firm or other financial institutions) in which the buyer must pay fixed instalments for a pre-defined number of months/years to the seller/financer[5]. After the completion of the agreement the cars are being resold. There can be several factors on which the price of a used car depends such as model, make, mileage, km driven, numbers of doors and so on. Some special factors which buyers attach importance in Mauritius is the local of previous owners, whether the car had been involved in serious accidents and whether it is a lady-driven car[5]. Also from the seller perspective, it is difficult to price a used car appropriately.

Prediction of a used car price is not an easy task to do, it needs a huge knowledge to predict such things. So, we can use different machine learning techniques for the prediction of a used car using different parameters of a car. The different machine learning models which can be used for the prediction of used car are: Linear regression, Lasso regression, K-Nearest Neighbour method, decision tree, Random forest and so on. These models needs some variables for the prediction such as:

**Price:** calculated retail price.

**Mileage**: total number of miles car have been driven.

**Make:** manufacturer name of the car.

**Model:** model name of the car.

**Trim:** type of car model.

**Type:** body type of car.

**Cylinder:** how many cylinders are there in the engine

**Litre:** how much is the fuel capacity of the engine.

**Doors:** the number of doors (e.g., 4-door car or 6-door car).

**Cruise:** A categorical variable (binary), which represents whether cruise control is present in the car[1]

These above are the different factors which affect the price of a used car and the data can be collected from daily newspapers, Kaggle website and some reputed companies such as Cars24, Cardekho, Truevalue, which are currently in the business of selling used cars.

**Technology used**

Python was the major technology used for the implementation of machine learning concepts the reason being that there are numerous inbuilt methods in the form of packaged libraries present in python[6]The following are the different python libraries to implement machine learning techniques:

* NUMPY
* SCIPY
* SCIKIT-LEARN
* JUPYTER NOTEBOOK

**NUMPY**

NumPy is a general-purpose array-processing package[6] By using this library, it can be used for working with arrays. It has also functions for working in domain of linear algebra, fourier transformation and matrices. It is an open source and any one can use it freely. It has multidimensional array object and also tools for working with these arrays. It can also be used to contain multidimensional generic data.

**SCIPY**

It is a free and open source Python library used for scientific and technological computations. SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers and other tasks common in science and engineering[6].

It is a part of Numpy and it includes tools like Matplotlib, Pandas and Sympy and and expanding set of scientific computations libraries.

**SCIKIT-LEARN**

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use[6]

**JUPYTER NOTEBOOK**

It is an open source web application that allows users to create and share documents such a live code, visualisations, equations and narrative text. It can also be used for simple python programming and also for higher level python programming which can be implemented in data cleaning and transformation, numerical simulation, statistical modelling, data visualisation, machine learning and many more.

**Methodology**

To summarize,

* Firstly, data collection of the used cars, identification of important features that affect the price.
* Secondly, data pre-processing and removal of entries with NA values. Removing those features which are irrelevant for the prediction of the price.
* Third, applying different machine learning models on the pre-processed dataset in which features as inputs and the price as output.
* Fourth, comparison of the different techniques by taking their accuracy as the decision parameter.
* Finally, we will deploy a webpage on which users can fill the features of their car and can get actual worth of the car on the basis of filled features.



Fig . Flowchart of Methodology



Table 1. Sample Data Collection[5]

Data Pre-processing

The data is collected from Kaggle website but we the dataset that we have collected may or may not be in a suitable for machine learning algorithms to operate on. So, the data pre-processing stage is an important stage in this we have to find the null data and missing values so that we can efficiently use the data. [7]

**Machine Learning Techniques used for prediction**

1. **Linear Regression**

 Linear regression is generally used for prediction of a value y based on

 discontinuous variable x. These models are expressed by mathematical

 structures in which yi is the prediction value and xi represents the value

 of an input feature i [7] Mathematical, it can be expressed as:



 where θj is coefficient matrix

1. **Decision Tree**

Decision tree is a machine learning technique in which is used to build regression model. It breaks down a dataset into smaller and smaller subsets based on the information gain value for each individual features while at the same time an associated decision tree is incrementally developed [8]. The result finally comes in the form of a tree with decision nodes and leaf nodes.



*Fig1. J48 Decision Tree on Training Set Without Car Model*[5]

1. **Random Forest**

It is a machine learning technique that belongs to the the supervised machine learning technique*.*It can be used for both Classification and Regression problems in ML. It is based on the concept of***ensemble learning,****which is a process of*combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, “**Random forest is a very popular Machine learning technique that is used for both classification and regression tasks. It is an ensemble machine learning method that creates a large number of decision trees, each trained on a different subset of the data and using a random subset of the features”.**



*Fig 2. The working of the Random Forest algorithm (fig from javaTpoint.com)*

**Conclusion**

It’s not a simple task to collect the data and just implement the machine learning algorithms because after the data collection the next step comes is the data preprocessing in which we have to remove the duplicate data and also removal of null data is also done in this stage.

In machine learning there are many techniques which can be used for the prediction, such as linear regression, kNN algorithm, decision tree, random forest and so on. But they all differ in their accuracy, it was observed that since there are deviation between the actual price and the predicted price which brings error to the algorithms and since we have to give more accurate prediction so we have to select more accurate algorithm with least value of Root Mean Squared Value.



Table 2 . Comparsion of different algorithms based on their RMSE value[8]

Out of all Random Forest has lowest RMSE, and performed well with highest R-squared value: 0.93 [8].

**References**

[1] “Used Cars Price Prediction using Supervised Learning Techniques,” *Int J Eng Adv Technol*, vol. 9, no. 1S3, pp. 216–223, Dec. 2019, doi: 10.35940/ijeat.a1042.1291s319.

[2] F. R. Amik, A. Lanard, A. Ismat, and S. Momen, “Application of Machine Learning Techniques to Predict the Price of Pre-Owned Cars in Bangladesh,” *Information (Switzerland)*, vol. 12, no. 12, Dec. 2021, doi: 10.3390/info12120514.

[3] M. Asghar, K. Mehmood, S. Yasin, and Z. Mehboob Khan, “Used Cars Price Prediction using Machine Learning with Optimal Features.”

[4] N. Pal, P. Arora, D. Sundararaman, P. Kohli, and S. S. Palakurthy, “How much is my car worth? A methodology for predicting used cars prices using Random Forest.” [Online]. Available: www.pakwheels.com

[5] S. Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques,” 2014. [Online]. Available: http://www.irphouse.com

[6] A. Chandak, P. Ganorkar, S. Sharma, A. Bagmar, and S. Tiwari, “Car Price Prediction Using Machine Learning,” *International Journal of Computer Sciences and Engineering*, vol. 7, no. 5, pp. 444–450, May 2019, doi: 10.26438/ijcse/v7i5.444450.

[7] F. Bodendorf and J. Franke, “A machine learning approach to estimate product costs in the early product design phase: A use case from the automotive industry,” in *Procedia CIRP*, 2021, vol. 100, pp. 643–648. doi: 10.1016/j.procir.2021.05.137.

[8] P. Gajera, A. Gondaliya, and J. Kavathiya, “OLD CAR PRICE PREDICTION WITH MACHINE LEARNING.” [Online]. Available: www.irjmets.com