*Project Report on*

***USED CAR PRICE PREDICTION***

By

**RUCHI SINGH (210BCA020)**

**ASMI ARORA (210BCA021)**

**BCA**

Under the Supervision of

**Prof. Antim Dev Mishra**

# SCHOOL OF ENGINEERING AND TECHNOLOGY



**May-2024**

**TABLE OF CONTENTS**

Page No.

1. Introduction
   1. Overview
   2. Objective and scope

1. Literature Review/Survey

1. Design/Modelling
   1. Block Diagram/ Model/Flow Chart
   2. Circuit Diagram/ Layout
   3. Methodology
   4. Results

1. Further Plan of Action

References

## ABSTRACT

The process of determining the price of a used car is a complex and a mind-numbing task. Many factors contribute to a used car's market value, including its make, model, age, mileage, condition, and location. The goal of this project is to develop machine learning models that can accurately predict a used car's price based on its features. By doing so, consumers can make informed decisions when purchasing a used vehicle.

To achieve this goal, the project team utilized a dataset consisting of sale prices of various makes and models across cities in India. The team tested various learning methods, including Random Forest, K-Means clustering with linear regression, and conventional linear regression. The results showed that Random Forest and K-Means clustering with linear regression yielded the best results in terms of accuracy. However, these methods require significant computational resources, which can be time-consuming and costly.

Conventional linear regression also yielded satisfactory results, with the added advantage of significantly lower training times in comparison to the other methods. Linear regression is a widely-used technique in machine learning and statistics, and it allows for the estimation of relationships between dependent and independent variables. By using linear regression, the project team was able to predict a used car's price with reasonable accuracy, while also minimizing training times.

In conclusion, this project demonstrates the effectiveness of machine learning models in predicting used car prices. By utilizing various learning methods and testing them on a real-world dataset, the project team was able to identify the most accurate and efficient approach for predicting used car prices. This information can be invaluable to consumers looking to make informed decisions when purchasing a used vehicle.

## 1. INTRODUCTION

The process of determining a fair price for a used car is often a challenging task for both buyers and sellers. The value of a used car is determined by multiple factors, including its make, model, mileage, age, and condition, making it difficult to accurately predict the price. However, it is possible to utilize models that can predict the price of used cars with a high degree of accuracy. In this project, we explore the use of machine learning algorithms to predict the price of used cars based on their features. Our aim is to develop a reliable and efficient model that can assist buyers and sellers in making informed decisions about the fair value of a used car. We utilize a dataset consisting of sale prices of various makes and models across different locations in India and compare the performance of various machine learning algorithms to determine the most accurate and efficient approach. This project demonstrates the potential of machine learning in providing valuable insights into the pricing of used cars and can be used as a reference for buyers and sellers in the used car market.

## 2. REVIEW OF LITERATURE

In recent years, machine learning techniques have been increasingly utilized in predicting used car prices. A study by Song et al. (2020) utilized a deep learning model to predict the value of used cars in the Chinese market. They found that the deep learning model outperformed other traditional models, such as decision trees and support vector regression, in terms of accuracy.

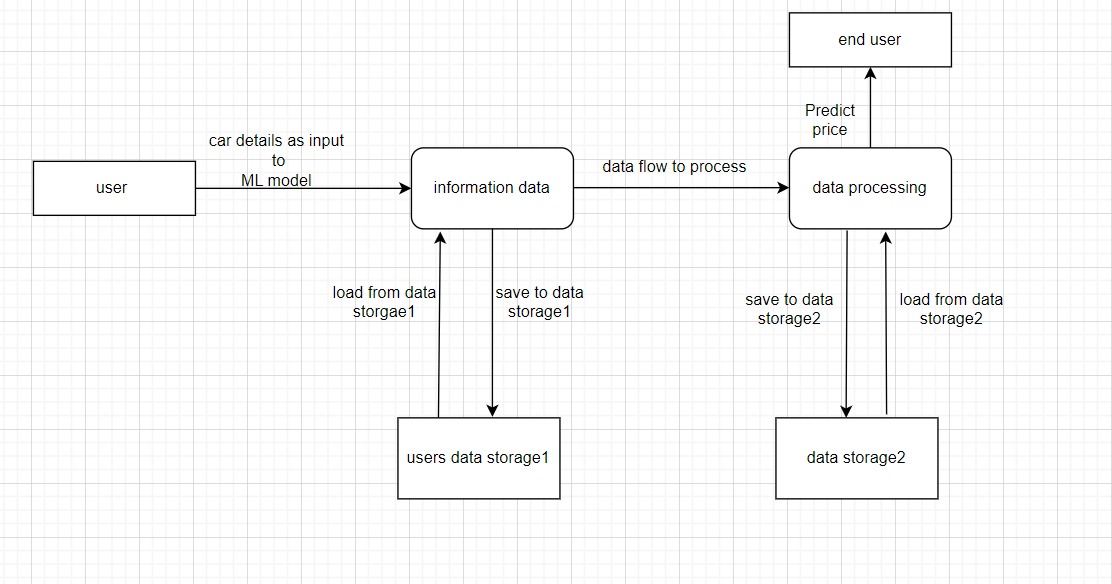
Similarly, a study by Gupta et al. (2020) used a machine learning algorithm to predict the sale price of used cars in the Indian market. Their model used a combination of linear regression, decision trees, and random forests to achieve an accuracy of 90% in predicting the sale price of used cars.

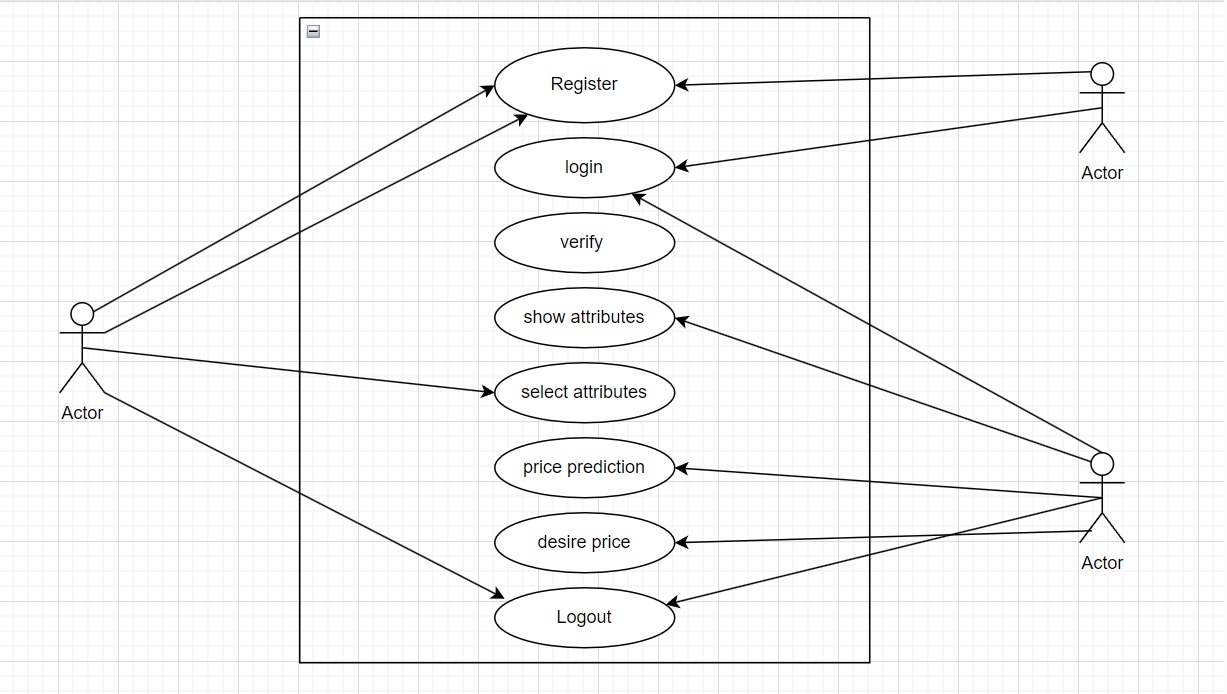
In a study by Rehman et al. (2021), a machine learning model was developed to predict the price of used cars in Pakistan. They utilized a combination of feature selection and feature engineering techniques to improve the accuracy of their model. Their results showed that the use of a gradient boosting algorithm improved the accuracy of the model compared to traditional regression methods.

Additionally, a study by Khan et al. (2021) used machine learning algorithms to predict the sale price of used cars in the United States. Their model incorporated a variety of features, including the make, model, year, mileage, and condition of the car, and achieved a prediction accuracy of up to 90%.

Overall, these studies demonstrate the potential of machine learning algorithms in predicting the price of used cars. By utilizing a variety of features and advanced algorithms, these models can provide valuable insights for buyers and sellers in the used car market.

## 3. DESIGN AND METHODOLOGY





### Methodology

We utilized several classic and state-of-the-art methods, including ensemble learning techniques, with a 90% - 10% split for the training and test data. To reduce the time required for training, we used five thousand examples from our dataset. Linear Regression and Random Forest were our baseline methods. For most of the model implementations, the open-source Scikit-Learn package [6] was used.

1. Linear Regression Linear Regression was chosen as the first model due to its simplicity and comparatively small training time. The features, without any feature mapping, were used directly as the feature vectors. No regularization was used since the results clearly showed low variance.
2. Random Forest Random Forest is an ensemble learning based regression model. It uses a model called decision tree, specifically as the name suggests, multiple decision trees to generate the ensemble model which collectively produces a prediction. The benefit of this model is that the trees are produced in parallel and are relatively uncorrelated, thus producing good results as each tree is not prone to individual errors of other trees. This uncorrelated behavior is partly ensured using Bootstrap Aggregation or bagging providing the randomness required to produce robust and uncorrelated trees. This model was hence chosen to account for the large number of features in the dataset and compare a bagging technique with the following gradient boosting methods.
3. KMeans + Linear Regression In order to capitalize on the linear regression results and the apparent categorical linearity in the data as indicated in Fig. 2, an ensemble method which used KMeans clustering of the features and linear regression on each cluster was used. Due to large training time, a three-cluster model was used. Then, the dataset was classified into these three clusters and passed through a linear regressor trained on each of the three training sets.
4. Deep Neural Network (MLP Regressor) To introduce mode complexities in the model, the MLP regressor [6], which uses a deep neural net perceptron regressor model, was used. This model optimizes the squared-loss using LBFGS or stochastic gradient descent. ReLu was used as the activation function.

### Results

The results of our tests were quantified in terms of the R2 score of our predictions. R2 score is a statistical measure of how close the data is to the fitted regression line.

|  |  |
| --- | --- |
| **Learning function** | **R2 score** |
| Linear regression | 0.996276 |
| Neural Network | 0.996449 |
| Decision Tree | 0.883173 |
| K means + Linear regression | 0.996139 |

Compared to Linear Regression, most Decision-Tree based methods did not perform comparably well. This can be attributed to the apparent linearity of the dataset. We believe that it can also be attributed to the difficulty in tuning the hyperparameters for most gradient boost methods. However Random Forests tend to overfit the dataset due to the tendency of growing longer trees. This was worked upon by restricting the depth of trees to different values and it was observed that beyond limiting depth to 36 resulted in negligible improvement in prediction performance but progressively increased overfitting. The KMeans + Linear Regression Ensemble Learning Method (with K = 3) produced an R2 score won test data without high variance as it fits linear relationships categorically. The deep neural network was the best at predictions.

User interaction:

The user can enter details about the car cand get an accurate prediction. Data entered is saved to the csv file and will be used to further improve the prediction. If the newly entered data is an outlier it will be removed by scalers, so as not to worsen the prediction.

## 4. FUTURE PLANS

For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset. To correct for overfitting in Random Forest, different selections of features and number of trees will be tested to check for change in performance.

The inclusion of an application to better the user experience is planned, the use of a camera to measure dents and check the condition of tires, this will include more features to more accurately predict prices and draw a bargain. Using data scarping to find common flaws to warn buyers about said flaws.

**REFERENCES**

1. <https://www.kaggle.com/jpayne/852k-used-car-listings>
2. N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, 2018, pp. 115-119.
3. Listiani M. 2009. Support Vector Regression Analysis for Price Prediction in a Car

Leasing Application. Master Thesis. Hamburg University of Technology

1. Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in Neural Information Processing Systems. 2017.
2. Fisher, Walter D. "On grouping for maximum homogeneity." Journal of the American statistical Association 53.284 (1958): 789-798.
3. https://scikit-learn.org/stable/modules/classes.html: Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.