Used Car Price Prediction Using Machine Learning

A PROJECT REPORT

Submitted by

CB.SC.I5DAS21034

NAME: Kavileswarapu Bhaswanth Durga Pavan Kumar

In partial fulfillment of the requirements for the award of the Degree of

INTEGRATED MASTER OF SCIENCE
IN
DATA SCIENCE

Department of Mathematics

AMRITA SCHOOL OF PHYSICAL SCIENCES

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE 641112

April 2025

DEPARTMENT OF MATHEMATICS

AMRITA SCHOOL OF PHYSICAL SCIENCES,

Coimbatore - 641112

BONAFIDE CERTIFICATE

This is to certify that the project report entitled "**Used car price Prediction using machine learing**" submitted by **Kavileswarapu Bhaswanth Durga Pavan Kumar**, **(CB.SC.I5DAS21034)** in partial fulfillment of the requirements for the award of the **Degree of Integrated Master of Science** in **Data Science is** a bonafide record of the work carried out under my guidance and supervision at Amrita School of Physical Sciences, Coimbatore.

Project Coordinator

Project Advisor

Chairperson
Department of Mathematics
Dr. K. Somasundaram

The project was evaluated by us on:

Internal Examiner

Internal Examiner

DEPARTMENT OF MATHEMATICS

AMRITA SCHOOL OF PHYSICAL SCIENCES,

Coimbatore - 641112

DECLARATION

I, **Kavileswarapu Bhaswanth Durga Pavan Kumar, CB.SC.I5DAS21034** hereby declare that this project report entitled "**Used car price Prediction using machine learing**" is the record of the original work done by me at Department of Mathematics, Amrita School of Physical Sciences, and Coimbatore. To the best of my knowledge this work has not formed the basis for the award of any degree / diploma / associateship / fellowship or a similar award to any candidate in any university/institutions.

Date:

Place:

Signature of the Student

COUNTERSIGNED

Dr. Sumathi I.R.
Project Coordinator,
Department of Mathematics,
School of Physical Sciences
Amrita Vishwa Vidyapeetham,
Coimbatore-641112, Tamil Nadu, India.

Acknowledgements

I would take this opportunity to thank everyone who helped me with this project. I would like to thank my Project Coordinator Dr.Sumathi I.R., Assistant Professor, Department of Mathematics, Amrita Vishwa Vidyapeetham, who constantly guided me. I would like to thank him for taking a considerable amount of time from his busy schedule to help with this mini project.

I take this opportunity to thank Dr. Sasanian Ramanathan, Dean Engineering, for providing the opportunity to access these facilities for the well-being of my mini project.

I would like to express my sincere gratitude towards Dr. Somasundaram, Chairperson and all the faculty members of the Department of Mathematics, Amrita Vishwa Vidyapeetham for supporting the vision that I believed in which has made this mini project a reality.

Lastly, I am especially indebted to my parents, family members and friends for their support, motivation, love and affection that have providing me with immense motivation and strength to finish this project successfully.

(Bhaswanth Durga Pavan Kumar K)

Contents

List of Figures

ABSTRACT

The Indian government makes use of taxation and it's also the car manufacturer who keeps setting the base price for new cars so that as soon as someone starts buying a car they feel the value for that investment is worth it. But with the price at which new cars are priced making these out of reach for many people now more and more people from all over the world are choosing to use used cars as an acceptable alternative.

But in order to make this happen, we will employ advanced Machine Learning algorithms to assist in accurately determining a good car price for Indian Car buyers. Random Forest, KNN, Decision trees, and linear regression were used to predict the best price of the car. This will then be converted to an interactive user-friendly web app using Streamlit framework as a major focus.

The basic idea behind this project is to design a useful, smart and efficient solution to help Indian car buyers to make informed decisions while buying a car.

CHAPTER 1

INTRODUCTION

1. Introduction

The second-hand car market has grown steadily in recent years, thereby making accurate pre-sale pricing indispensable for buyers as well as sellers. For this purpose, we devised a predictive ML algorithm for used-car price predictions based on attributes such as brand, model, year, fuel type, transmission, and mileage. The solution features a simple-to-use web app created with Streamlit that enables users to enter car details and receive price predictions instantly.

This project compares supervised machine learning algorithms like Linear Regression, Decision Tree, KNN, and Random Forest, trying to find the most accurate and operational one for deployment. Based on the evaluation metrics applied, Random Forest performed the best with a test set R^2 score of 0.94.

Several data preprocessing steps were done, such as handling missing values, encoding categorical data, and normalizing numeric features. Further, features engineered for the task improved model performance. Lastly, the Streamlit web app was built to assist users in estimating car prices so that individuals and dealerships can make informed decisions in the rapidly changing used car market.

CHAPTER 2

Objectives

2. Objectives

This predictive model for pricing of second-hand cars will be made with supervised machine learning techniques.

To test the performance of individual machine learning models against each other in order to come up with the most accurate model for price prediction.

It will develop a real-time prediction application for users based on Streamlit - an interactive and user-friendly web application.

Give understanding through visualization and features importance analysis regarding how the data affect car price estimation.

Develop a dashboard for pricing trend and pattern analysis of used cars users. The model should be made interpretable by means of feature importance and correlations visualizations.

Scalability has to be allowed for the capabilities to incorporate more vehicle attributes and foreign market influences.

Optimize model performance through hyper parameter tuning and feature engineering.

CHAPTER 3

Technologies Used

3. Technologies Used

Programming Language:

Python

Machine Learning Models:

1. Random Forest.
2. Decision Tree
3. K-Nearest Neighbors (KNN)
4. Linear Regression

Libraries Used:

Data Handling & Processing: Pandas, NumPy.

Machine Learning: Scikit-learn.

Visualization: Matplotlib, Seaborn.

Web Application: Streamlit.

CHAPTER 4

Dataset and Data Pre-processing

4. Dataset and Data Pre-processing

4.1 Dataset Collection:

The dataset for this project was collected from Kaggle, a popular site for datasets and data science competitions. It contains 15,411 entries and 14 columns, thus effectively presenting important attributes of used cars. These features play a key role in determining the resale value of the vehicle and act as input variables to train the machine-learning models.

Data Set:

Feature	Description
Car Name	Name of the car (e.g., Maruti Alto, Hyundai i20)
Brand	The car manufacturer (e.g., Maruti, Hyundai, Ford)
Model	The specific model name
Vehicle Age	Age of the car in years
KM Driven	Total distance the car has been driven
Seller Type	Whether the seller is an individual or a dealer
Fuel Type	Type of fuel used (Petrol, Diesel, Electric, etc.)
Transmission Type	Whether the car has a manual or automatic transmission
Mileage	The mileage of the car in km per litter (km/l)
Engine	The engine capacity in cubic centimeters (cc)
Max Power	The maximum power output of the engine (in bhp)
Seats	The number of seats in the car
Selling Price	The actual selling price of the car (target variable)

4.2 Data Cleaning and Preprocessing:

Handling Missing Values: No missing values were present in the dataset.

Encoding Categorical Variables:

One-hot encoding for categorical features (e.g., fuel type, seller type).

Label encoding for ordinal features.

Feature Engineering:

Derived "Car Age" from the manufacturing year.

Feature Scaling:

StandardScaler applied to numerical features for normalization.

Data Splitting:

80% training data, 20% testing data.

CHAPTER 5

Machine Learning Models Used

5. Machine Learning Models Used

Several supervised machine learning models have been identified in this project to determine the most efficient way to predict the sales prices of used cars. The accuracy of prediction, computational speed, and interpretation are the main factors considered when evaluating each of these models. Random forests, decisions trees, k-nikat neighbors (KNN), and linear regression are all part of the models utilized in this work. The selection of these methods was based on their widespread appeal, ease of utilization, and proficiency in managing categories and numerical data.

5.1 Overview of Models:

5.1.1 Random Forest Regressor:

ENSIBAL's Random Forest Learning Medium produces various opinions and compiles them to minimize overfitting and increase prediction accuracy. By gathering a random set of data to train each tree, an average forecasting system will be developed. It was an emotional bone in our tests and scores a very high R2 - 0.94, quite impressively on the test set. Due to the possibility of unlectured connections and ease of data transmission, it was the best choice for deployment.

5.1.2 Decision Tree Regressor:

The arrangement is reminiscent of a tree, consisting of root node, branches, internal nodes, and leaf nodies. This structure. Our classification and regression processes utilize decision trees as well, which offer uncomplicated models. Using conditional control statements, this algorithmic model is non-parametric and has the ability to perform both classifications and regressions through supervised learning. This structure resembles a tree, with alternating levels of branches, roots, and leaves that form overlapping hierarchy. A variety of fields utilize this tool.

5.1.3 K-Nearest Neighbors (KNN):

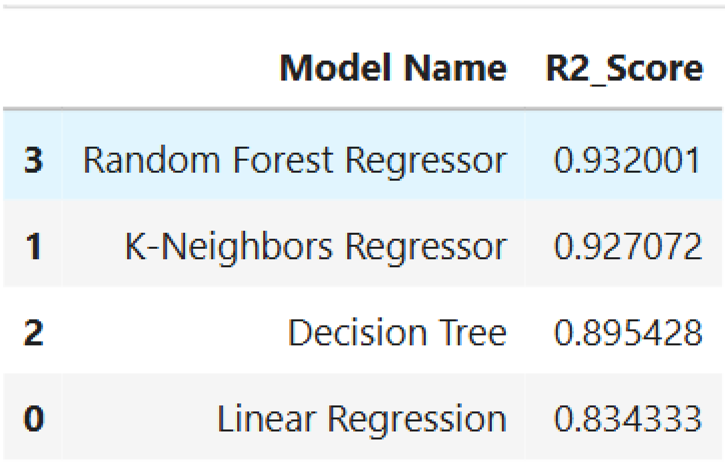
K-Nearest Neighbor(KNN) is a supervised approach in Machine Learning, which is quite unconventionally applied. This is elementary and is applied in classification and regression algorithms. The K-NN relies on the assumption that new data and available data are similar; therefore, the class of the newly created case is allotted to the class most similar to that of the neighboring data. By virtue of simply storing the data, the K-NN algorithm can assign a new input to a class according to whichever category confers more comfort. Once the algorithm takes on the responsibility of classifying new data into the best fit class, it comes into place. While it is used in regression and classification, it is mostly applied in classification.

5.1.4 Linear Regression:

Linear Regression is one of the most fundamental and widely used supervised learning algorithms for regression tasks. It models the relationship between the dependent variable and one or more independent variables using a linear equation. The objective is to fit a straight line (in higher dimensions, a hyperplane) that minimizes the residual sum of squares between the observed and predicted values. Despite its simplicity, Linear Regression is highly interpretable and serves as a strong baseline model. However, it assumes linearity and can perform poorly when data exhibits complex or non-linear relationships.

5.2 Model Performance Comparison:

Random Forest achieved the best performance with an accuracy of 93%.



	Model Name	R2_Score
3	Random Forest Regressor	0.932001
1	K-Neighbors Regressor	0.927072
2	Decision Tree	0.895428
0	Linear Regression	0.834333

Fig 5.2.ML Model R2_score.

5.3 Hyper parameter Tuning:

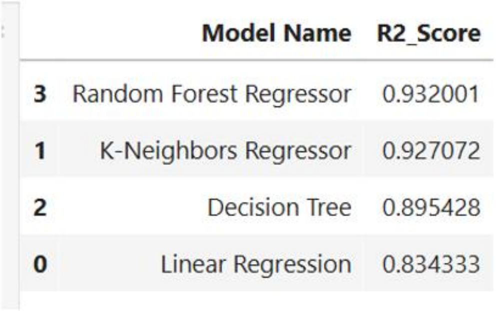
Hyperactive Parameter tuning is a crucial step towards achieving better performance during the machine knowledge model development. Thus exploring through RandomizedSearchCV the varied combinations of hyperactive parameters to correspondingly optimize every swish configuration of the models was performed. The following optimized conditions for the respective model are available:

Random Forest(RF) n_estimators, max_depth, min_samples_split;

Decision Tree(DT) min_samples_split, max_depth;

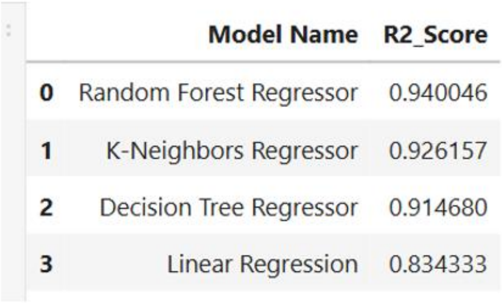
K-Nearest Neighbors(KNN) distance metric, i.e. Euclidean or Manhattan; and number of neighbors, n_neighbors.

Initial exemption for Linear Retrogression was to have tested Ridge and Lasso for the regularization methods to improve on their generality.



	Model Name	R2_Score
3	Random Forest Regressor	0.932001
1	K-Neighbors Regressor	0.927072
2	Decision Tree	0.895428
0	Linear Regression	0.834333

Before Hyper parameter Tuning



	Model Name	R2_Score
0	Random Forest Regressor	0.940046
1	K-Neighbors Regressor	0.926157
2	Decision Tree Regressor	0.914680
3	Linear Regression	0.834333

After Hyper parameter Tuning

Fig 5.3 Hyper Parameter Before and After.

CHAPTER 6

Feature Importance Analysis

6. Feature Importance Analysis:

Feature importance analysis identified the key factors affecting price prediction:

- Top Features: Car Age, Mileage, Fuel Type, Transmission Type
- Visualization:

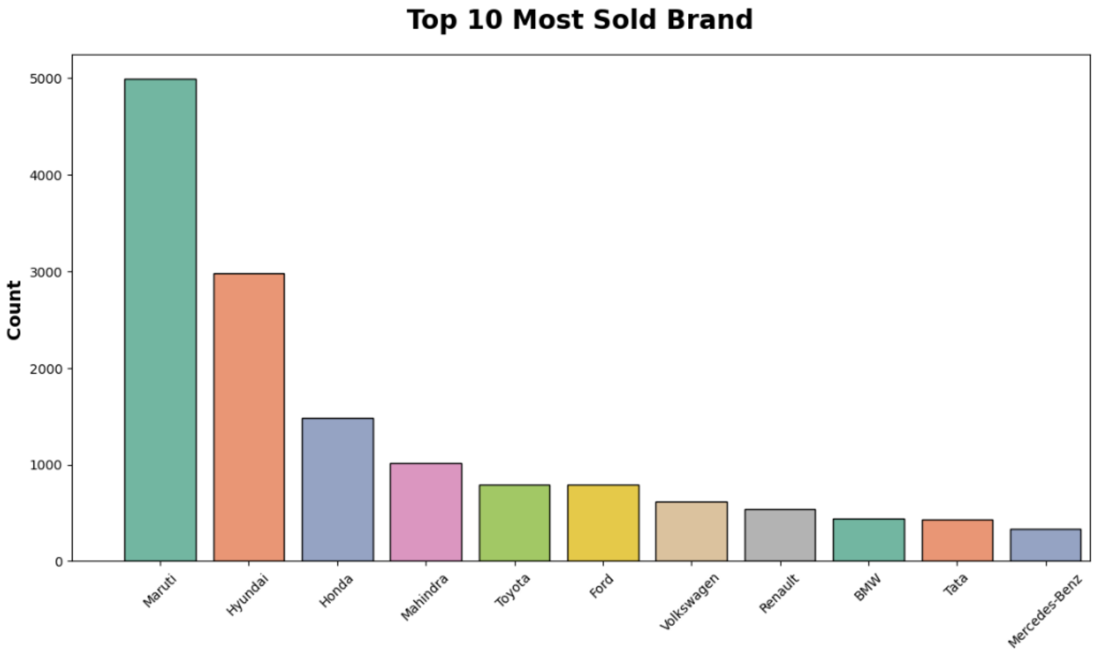


Fig 6.1. Most selling brand

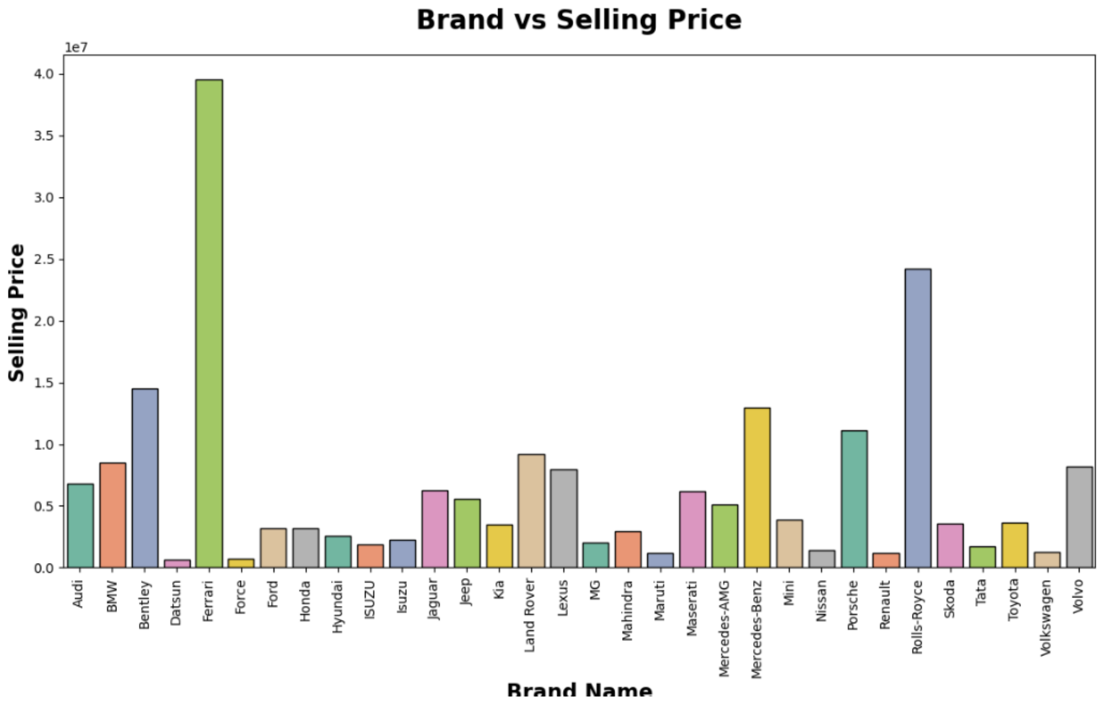


Fig 6.2. Brand and Costliest Car.

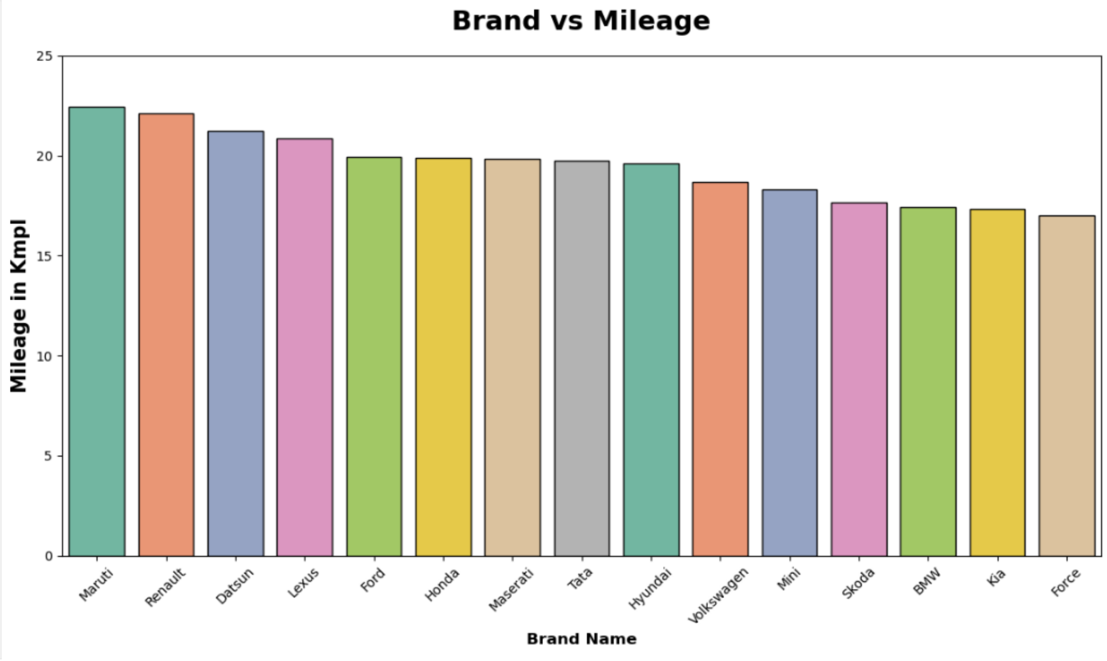


Fig 6.3. Most Mileage Brand and Car Name.

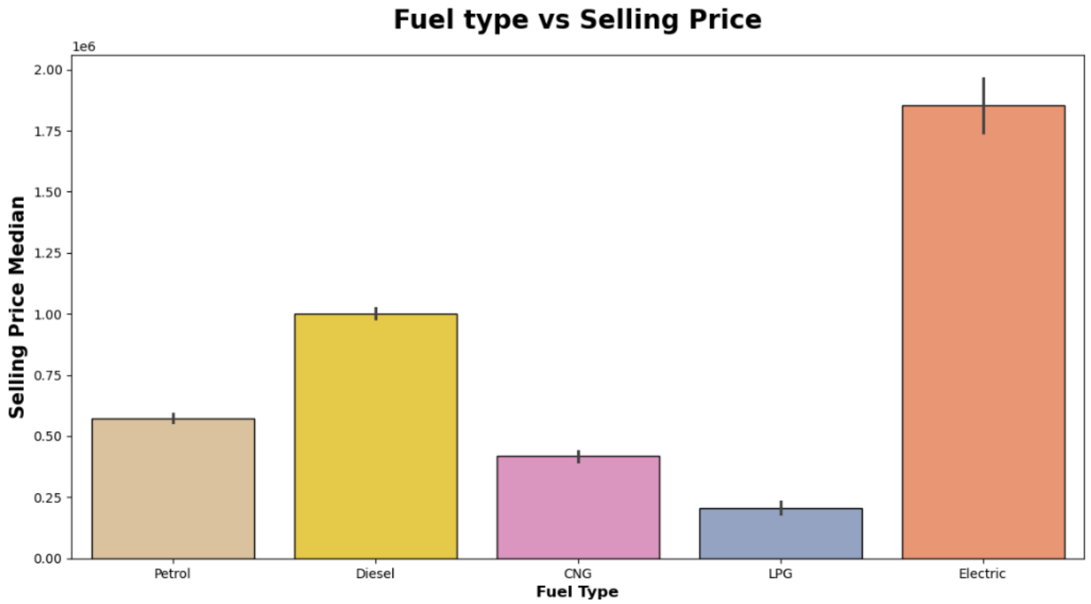


Fig 6.4. Fuel Type Selling Price.

Car Age:

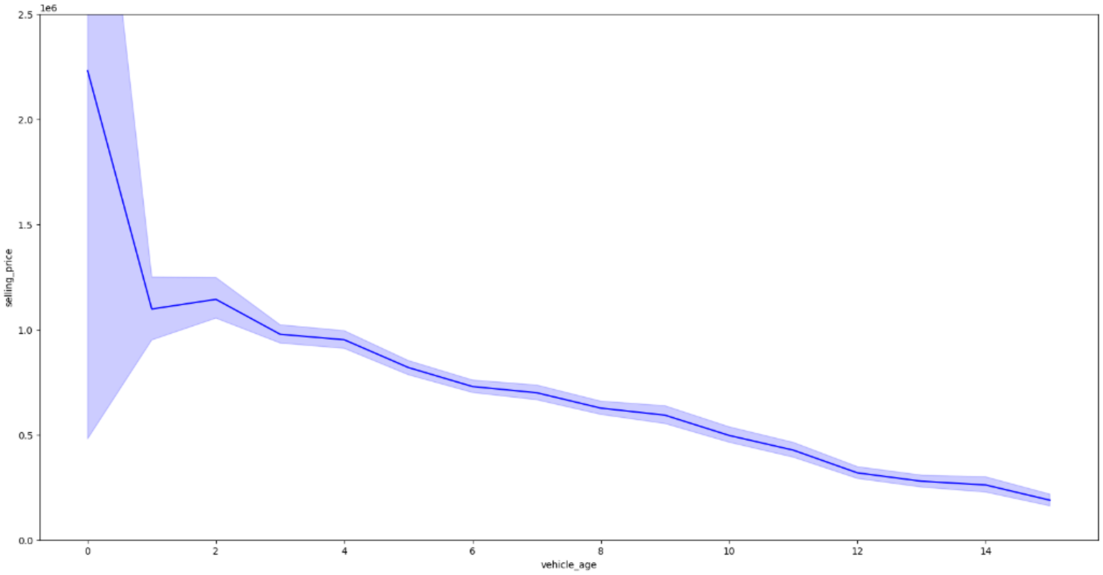


Fig 6.5. Vehicle age vs Selling Price.

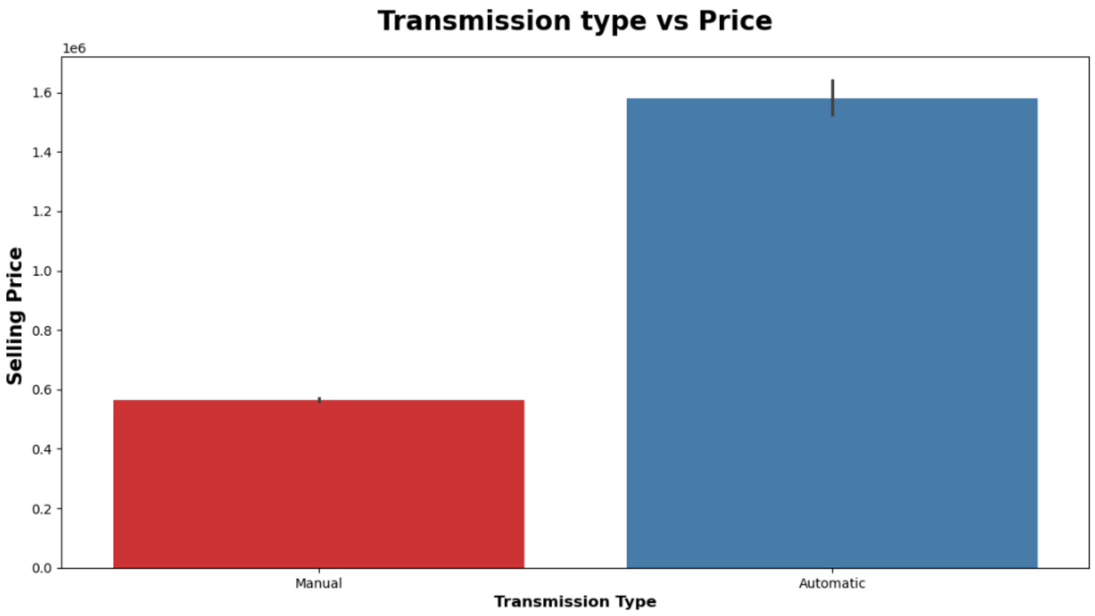


Fig 6.6. Transmission Type.

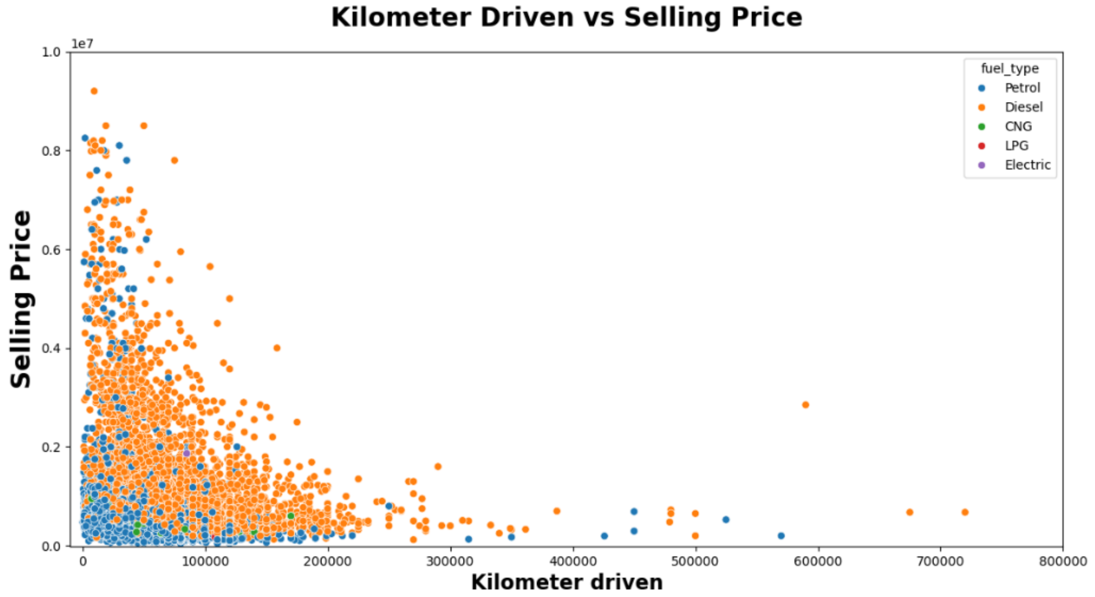
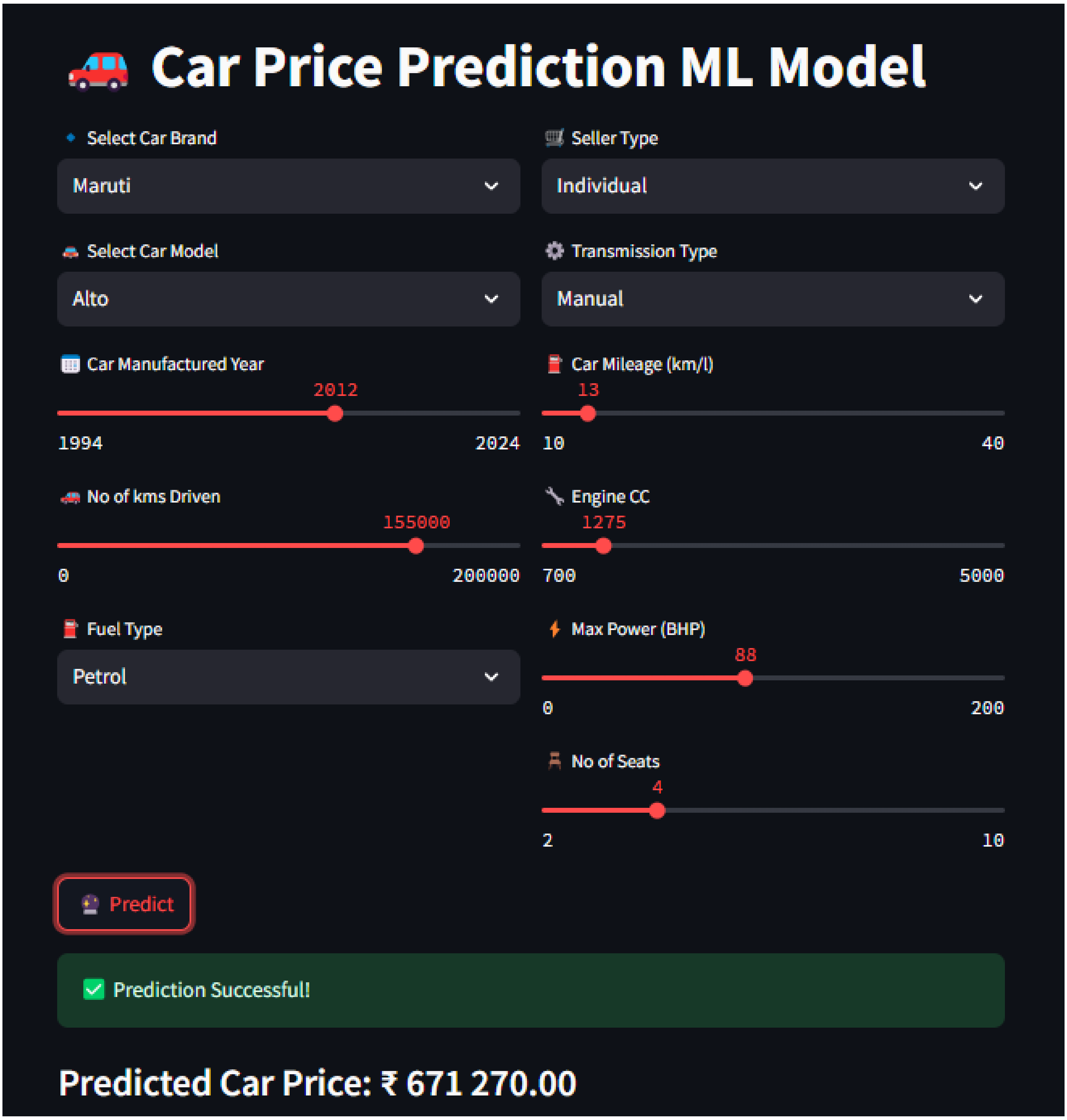


Fig 6.7. Kilometer Driven vs Selling Price.

CHAPTER 7

WEB PAGE

7. Web Page:



🚗 Car Price Prediction ML Model

Select Car Brand: Maruti

Seller Type: Individual

Select Car Model: Alto

Transmission Type: Manual

Car Manufactured Year: 2012

Car Mileage (km/l): 13

No of kms Driven: 155000

Engine CC: 1275

Fuel Type: Petrol

Max Power (BHP): 88

No of Seats: 4

✅ Prediction Successful!

Predicted Car Price: ₹ 671 270.00

Fig 7. Web Page

Conclusion

Conclusion:

The **Used Car Price Prediction** project successfully implemented a **machine learning-based system** integrated into a **Streamlit web application** for used car price estimation. The study of **four supervised learning models**—Random Forest, Decision Tree, K-Nearest Neighbors, and Linear Regression—revealed that the **Random Forest Regressor** performed the best, achieving an **R^2 score of 94%**.

The web application was designed with a **user-friendly interface**, allowing users to input various car details such as **brand, year of manufacture, mileage, fuel type, and engine capacity** to instantly obtain a price estimate. The interactive nature of the application ensures ease of use, making it accessible even to non-technical users.

Key Features:

1. High Prediction Accuracy: The Random Forest model demonstrated superior performance with a high **R^2 score of 94%**.

2. **Interactive Web Application:** Developed using **Streamlit**, enabling easy access to price predictions.

3. Feature Importance Analysis: Identified critical factors influencing car prices, such as **Car Age, Mileage, Engine Capacity, and Fuel Type**.

4. Optimized Model Performance: Implemented **hyper parameter tuning** for better accuracy.

References

References:

Data set: https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset.

1. Abishek, R. (2022). Car Price Prediction Using Machine Learning Techniques. International Research Journal of Modernization in Engineering Technology and Science, 4(2), 54-50.

2. Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S., & Boonpou, P. (2018, May). Prediction of prices for used car by using regression models. In 2018 5th International Conference on Business and Industrial Research (ICBIR) (pp. 115-119). IEEE.

3. Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car Price Prediction using Machine Learning Techniques.

4. Chandak, A., Ganorkar, P., Sharma, S., Bagmar, A., & Tiwari, S. (2019). Car price prediction using machine learning. International Journal of Computer Sciences and Engineering, 7(5), 444-450.

References Webpage:

Orange Book Value (OBV). Used Car Pricing Platform. Available at:
https://orangebookvalue.com/used-cars