PCTS A.P. Shah Institute of Technology

Department of Computer Engineering

PROJECT REPORT ON

Car Price Prediction using Machine Learning

Submitted by:

Vivek Behera

B.E. in Computer Engineering

Under the Guidance of

Department of Computer Engineering

Academic Year: 2024–2025

Abstract

This project focuses on building a machine learning-based predictive model to estimate the selling price of used cars. By analyzing various car attributes such as manufacturing year, kilometers driven, fuel type, and transmission type, the model helps users determine accurate prices. Random Forest Regression was used for its robustness and high predictive performance. The results show that the model provides consistent and accurate predictions, which can be beneficial for car dealers, resellers, and buyers.

1. Introduction

The used car market is one of the fastest-growing sectors, and determining the right price is essential for both buyers and sellers. Traditional methods rely heavily on manual estimation, which can be inaccurate. Machine Learning techniques provide a data-driven solution that leverages past sales data to predict the resale value of vehicles accurately.

This project demonstrates the complete process of developing a car price prediction model, from data preprocessing and visualization to training, tuning, and evaluating a Random Forest Regressor.

2. Problem Definition

The problem addressed in this project is inaccurate and inconsistent used car pricing. There is a lack of analytical tools to determine fair prices for used cars. This system uses a trained model to automatically estimate car prices based on significant features extracted from the dataset.

3. Objectives

• To preprocess and clean the car dataset. • To apply feature engineering and exploratory data analysis. • To train and test machine learning models for regression. • To evaluate model performance using accuracy metrics. • To deploy a prediction-ready model.

4. Scope

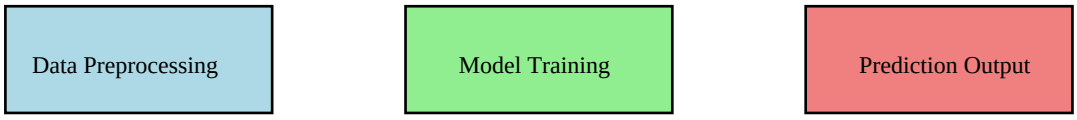
The project is designed to help car sellers, buyers, and dealerships make informed decisions. It can also be extended to predict prices for other assets, such as real estate and electronics, demonstrating the versatility of machine learning in regression problems.

5. Literature Review

Several studies have proven that machine learning models such as Linear Regression, Decision Trees, and Random Forest are highly effective in predicting continuous values. According to research papers published in IEEE and Springer, ensemble models like Random Forest achieve better generalization and handle non-linear relationships efficiently.

6. System Architecture

The system follows a data-driven workflow from data collection to prediction, as shown below:



7. Methodology

The project methodology consists of the following steps: 1. Data Collection from CarDekho dataset. 2. Data Cleaning and preprocessing to remove null or redundant values. 3. Categorical Encoding using one-hot encoding. 4. Data Splitting into training (80%) and testing (20%) sets. 5. Model Training using RandomForestRegressor. 6. Hyperparameter Tuning with RandomizedSearchCV. 7. Model Evaluation using R^2 and RMSE metrics.

8. Implementation

Python libraries such as pandas, numpy, matplotlib, seaborn, and scikit-learn were used for data handling, visualization, and model building. The dataset 'cardekho_dataset.csv' was processed, cleaned, and split into training and testing subsets. The RandomForestRegressor was trained and optimized through RandomizedSearchCV to achieve high accuracy.

9. Results and Discussion

The model achieved strong predictive performance with high R^2 values, indicating accurate predictions. The feature importance graph revealed that 'year', 'km_driven', and 'fuel_type' were key predictors of selling price. This confirms that newer cars with fewer kilometers and specific fuel types tend to have higher resale values.

10. Conclusion and Future Scope

The Car Price Prediction model demonstrates how machine learning can revolutionize the used car market by automating price estimation. Future work includes expanding the dataset, experimenting with deep learning models, and deploying the system as a web or mobile application using Streamlit or Flask.

11. References

1. Breiman, L. (2001). Random Forests. Machine Learning Journal. 2. CarDekho Dataset, www.cardekho.com. 3. Scikit-Learn Documentation, https://scikit-learn.org. 4. Han, Kamber & Pei (2012). Data Mining: Concepts and Techniques. 5. IEEE Papers on Regression and Prediction Models (2018–2024).