Car Price Predictor

Submitted for

Statistical Machine Learning CSET211

Submitted by:

(E23CSEU2117) RIJUL ARORA

(E23CSEU2123) GURLAL SINGH

Submitted to

DR. SUSMITA DAS

July-Dec 2024

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

INDEX

Table of Contents

Abstract

The rapid growth in the automobile market has led to increased demand for accurate car price predictions. Car price prediction models help buyers and sellers make informed decisions based on the intrinsic and extrinsic features of a vehicle, such as brand, model, age, mileage, and engine specifications. In this project, we developed a predictive model using machine learning algorithms like Linear Regression, Ridge Regression, Lasso Regression, Decision Tree, Random Forest, and K-Nearest Neighbors (KNN). The dataset was preprocessed, and various features were encoded to train and evaluate the models. Experimental results show that ensemble methods like Random Forest outperform traditional regression models in accuracy and error minimization. This report discusses the methodology, results, and potential applications of the developed system. The implementation code is shared on GitHub for reproducibility and further research.

Introduction

Pricing in the automotive industry is a dynamic process influenced by numerous factors such as market trends, vehicle specifications, and economic conditions. Buyers and sellers often rely on subjective evaluations or third-party platforms, which may lead to biased or inaccurate price estimates. Hence, developing a robust machine-learning-based car price prediction model is a practical solution.

This project explores the use of supervised machine learning techniques to predict car prices using structured datasets. By comparing algorithms such as Linear Regression, Ridge Regression, Lasso Regression, Decision Tree, Random Forest, and KNN, the study identifies the most efficient approach for price prediction. The objectives include:

1. Analysing key factors influencing car prices.
2. Building a dataset for training machine learning models.
3. Comparing the performance of different algorithms.
4. Recommending an optimized model for real-world applications.

Related Work

Previous studies in car price prediction have employed statistical models and machine learning techniques. Regression models, such as Linear and Ridge Regression, have been widely used for numerical predictions, providing interpretable results. However, these methods often fail to capture non-linear relationships.

1. **Machine Learning in Automobile Valuation**
   Researchers have shown that ensemble models like Random Forest can effectively handle feature interactions and non-linear dependencies in car price prediction tasks.
2. **Feature Selection for Vehicle Pricing**
   Studies highlight the importance of selecting relevant features such as mileage, age, and brand while excluding redundant data to improve model accuracy.
3. **Use of Decision Trees and KNN**
   Decision Tree algorithms are often preferred for their interpretability, whereas KNN is effective for small datasets but computationally expensive for larger datasets.

These prior works informed our choice of algorithms and preprocessing techniques for this study.

Methodology

4.1 Dataset Description

- **Source:** [e.g., Kaggle, UCI repository]
- **Features:**
  - Brand
  - Model Year
  - Mileage (in km or miles)
  - Engine Capacity (in cc)
  - Fuel Type (e.g., Petrol, Diesel, Electric)
  - Transmission (e.g., Manual, Automatic)
  - Price (target variable)

4.2 Data Preprocessing

1. **Handling Missing Values:** Imputation techniques like mean/mode for numerical/categorical features.
2. **Feature Encoding:** Label encoding for categorical variables like fuel type and transmission.
3. **Feature Scaling:** Min-Max scaling for numerical variables to normalize feature values.

4.3 Model Selection

The project employs the following machine learning algorithms:

1. **Linear Regression:** Establishes a linear relationship between features and price.
2. **Ridge & Lasso Regression:** Adds regularization terms to reduce overfitting and improve generalization.
3. **Decision Tree:** Uses a tree-based structure to split data for predictions.
4. **Random Forest:** An ensemble method combining multiple decision trees to improve accuracy.
5. **K-Nearest Neighbours (KNN):** A distance-based method to predict prices based on similar data points.

4.4 Model Evaluation Metrics

- **Mean Squared Error (MSE)**
- **R-squared**
- **Mean Absolute Error (MAE)**

Selection Hardware/Software Required

Hardware Requirements:

- Processor: Intel i5 or higher
- RAM: 8GB minimum (16GB recommended)
- Storage: 500GB HDD or 256GB SSD
- GPU (Optional): NVIDIA GTX 1050 or higher for faster computation

Software Requirements:

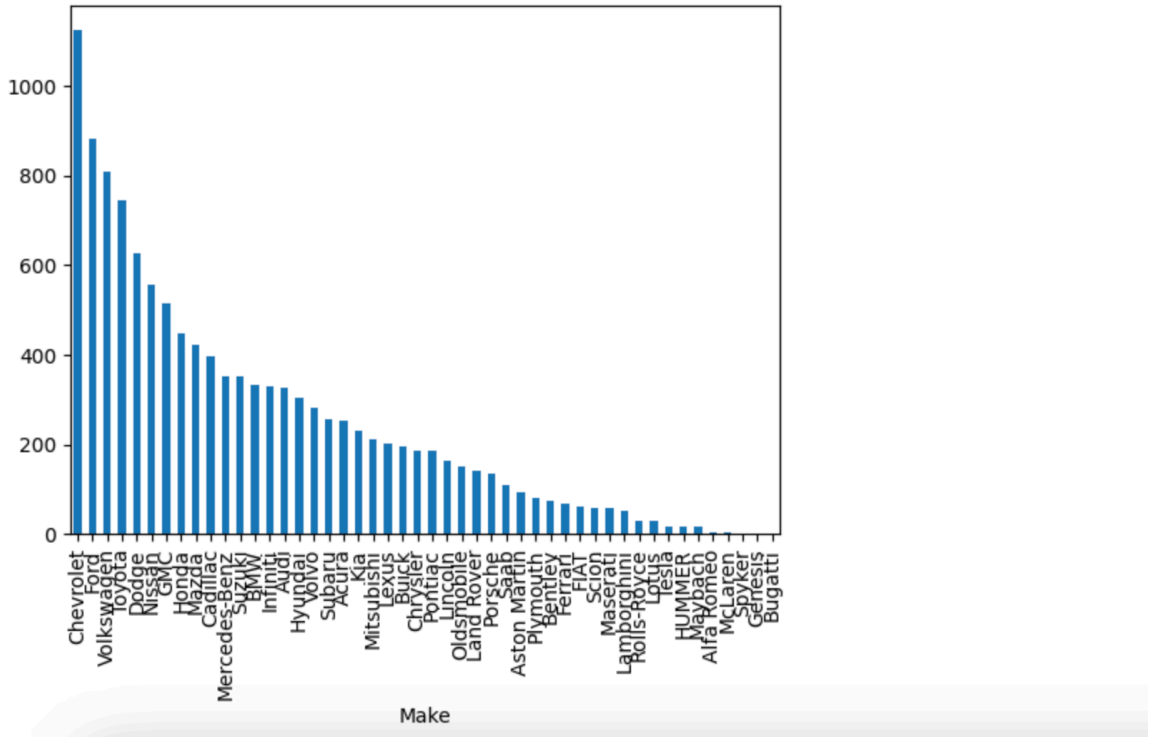
- **Programming Language:** Python (Version 3.x)
- **Libraries:** NumPy, pandas, scikit-learn, matplotlib, seaborn
- **IDE:** Jupyter Notebook / PyCharm / Google Colab
- **Version Control:** Git and GitHub

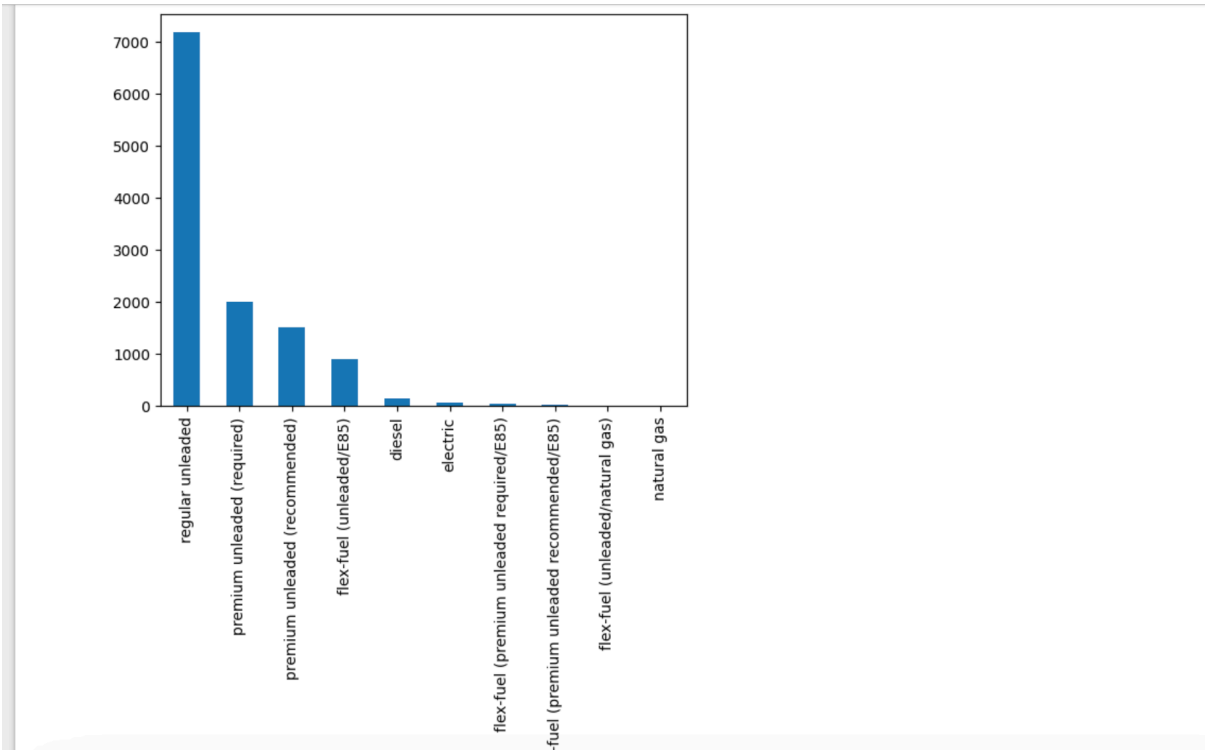
Experimental Results

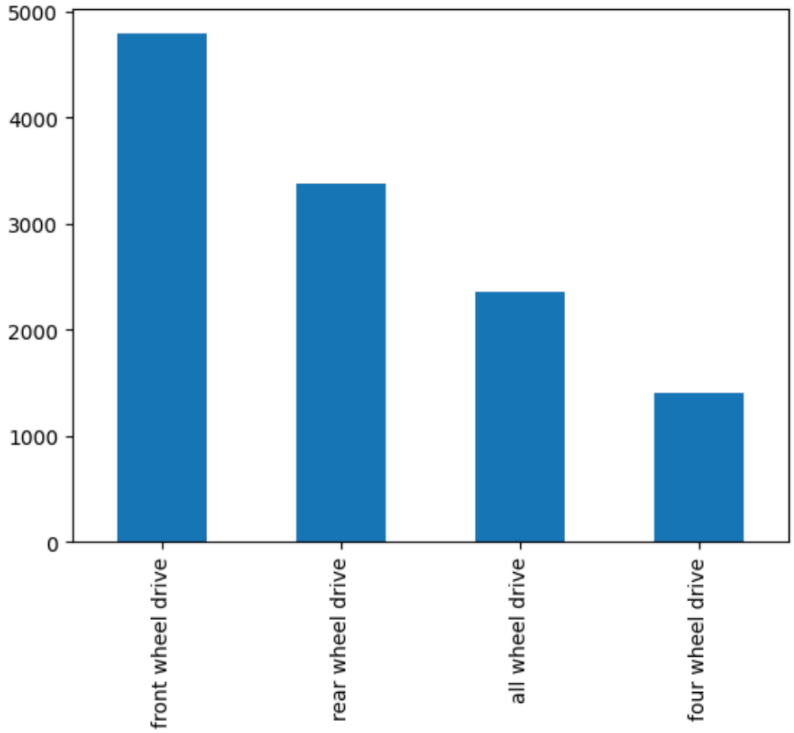
6.1 Model Performance

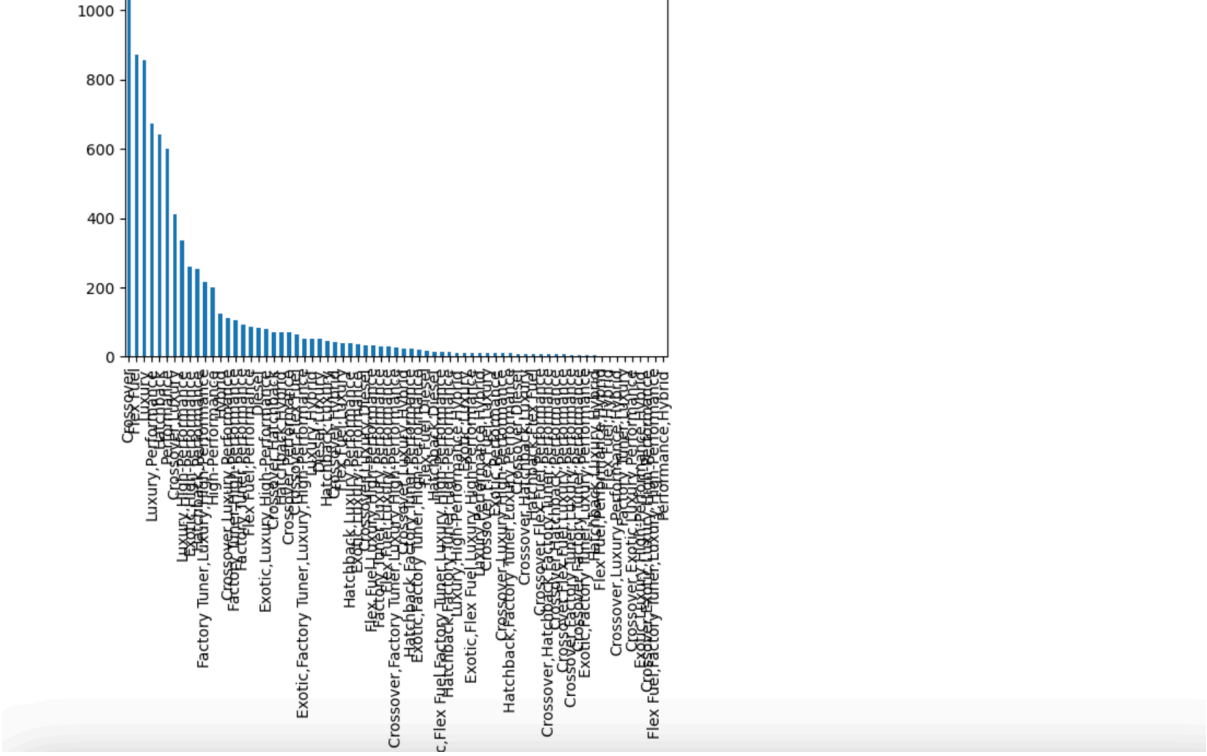
Algorithm	R-Squared
Linear Regression	0.78
Ridge Regression	0.67
Lasso Regression	0.79
Decision Tree	0.68
Random Forest	**0.992**
KNN	0.78

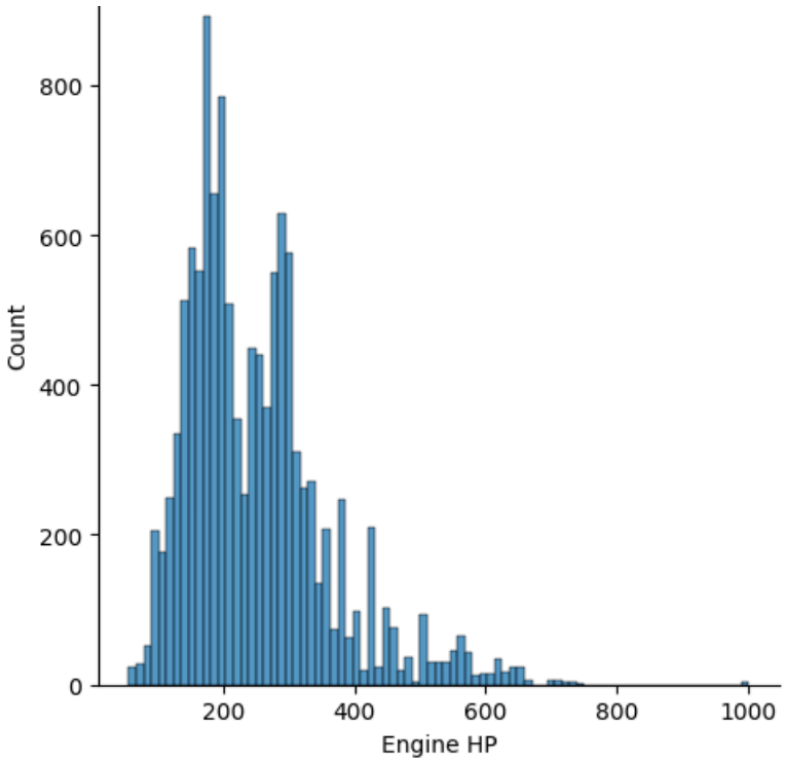
6.2 Visualizations:




7000
6000
5000
4000
3000
2000
1000
0
regular unleaded
premium unleaded (required)
premium unleaded (recommended)
flex-fuel (unleaded/E85)
diesel
electric
flex-fuel (premium unleaded required/E85)
-fuel (premium unleaded recommended/E85)
flex-fuel (unleaded/natural gas)
natural gas


5000
4000
3000
2000
1000
0
front wheel drive
rear wheel drive
all wheel drive
four wheel drive





Conclusions

The study demonstrated that Random Forest consistently outperformed other models in terms of accuracy and error

minimization, making it a suitable choice for car price prediction tasks. Simpler models like Linear Regression and Ridge Regression also performed reasonably well but struggled with non-linear relationships. The project highlights the significance of ensemble methods in improving prediction robustness.

Future Scope

1. **Feature Expansion:** Include additional features like market trends, insurance costs, and resale value.

2. **Deep Learning Models:** Explore advanced techniques like neural networks for better generalization.
3. **Real-Time Data Integration:** Incorporate live data from car dealerships or APIs for real-time predictions.
4. **Deployment:** Deploy the model using web frameworks like Flask or Django for user accessibility.

Github Link-:

https://github.com/gurlalsingh1313/ml-project.git