Predicting Car Prices Using Various Machine Learning Techniques

Syed Fateen Navid Haider*, Tafsir Md. Rubaiyat Rahman*, Mohsina Tabassum Rifa*
*Department of Computer Science and Engineering, Islamic University of Technology

***Abstract*—This paper presents a comprehensive analysis of vehicle price prediction using machine learning techniques. The goal is to examine the effectiveness of different algorithms to accurately estimate car prices based on a data set that includes information such as vehicle type, year, price, transmission, mileage, fuel type, tax, mpg, engine size, and etc. on The research examines six different machine learning algorithms: linear regression, support vector regression (SVR). The random forest, long-term and short-term memory (LSTM), k-nearest neighbor (KNN), and decision tree data sets are obtained from a trusted source and go through pre-designed steps to deal with missing values, . handling external factors, and categorical variables .The proposed method involves training each algorithm on preprocessed data and testing their performance. The experimental design involves splitting the dataset into training and testing sets using metrics such as mean square error (MSE) and coefficient of determination (R-square) to measure accuracy The obtained results are analyzed and compared, working with graphs such as scatter plots and line plots. Furthermore, the performance of the algorithm is compared with existing methods in the literature. The analyzes include error analysis, ablation analysis, and hyperparameter tuning to investigate factors affecting prediction accuracy.The findings show that XGBoost regression outperforms other algorithms and achieves the highest accuracy of 95.8%. The random forest also exhibits an impressive accuracy of 95.7%. The results of the study have important implications for car buyers, dealers, and manufacturers, helping them to make informed decisions about pricing strategies Future work may explore methods a new set of processing techniques and learning methods to improve the predictive accuracy of automotive pricing estimates.**

***Index Terms*—Machine Learning, Car Price Prediction, Linear Regression, XGB Regression, Neural Network, SVR, Random Forest, LSTM, KNN, Decision Tree.**

I. Introduction

PREDICTING vehicle pricing accurately is critical in the automotive industry. It enables buyers and sellers to make informed decisions and helps manufacturers develop optimal pricing policies. Increasingly large data sets, coupled with advances in machine learning techniques, have created new opportunities for accurate price prediction The ability of machine learning algorithms to analyze various factors affecting automotive prices, is predictively reliable to it and make images can be done

The objective of this study is to explore the effectiveness of different machine learning algorithms in predicting car prices. We focus on a dataset that includes crucial attributes such as car model, year, price, transmission, mileage, fuel type, tax, mpg, and engine size. These attributes capture important information about the characteristics and specifications of cars, which are known to impact their market value.

The algorithms considered in this study include Linear Regression, Support Vector Regression (SVR), Random Forest, Long Short-Term Memory (LSTM), k-Nearest Neighbors (KNN), and Decision Tree. Each algorithm has its strengths and weaknesses, and by evaluating their performance on the given dataset, we can gain insights into their suitability for car price prediction.

Car price prediction is a challenging task due to the complexity and variability of factors involved. Traditional regression-based methods have limitations in capturing nonlinear relationships and complex patterns in the data. On the other hand, advanced machine learning techniques, such as Random Forest and LSTM, have demonstrated promising results in capturing intricate patterns and making accurate predictions.

In this study, we not only compare the performance of different algorithms but also aim to analyze and interpret the results. We will use appropriate metrics to evaluate the accuracy of each algorithm and visualize the predicted values against actual values to get a better understanding of their predictive capability Furthermore, we present our proposed algorithms performance will be compared to existing methods in the literature for effectiveness.

The results of this study can be helpful to various stakeholders in the automotive industry. Car buyers can use accurate price forecasts to make informed decisions and make better deals. Retailers can adjust their pricing policies based on market data and competitive analysis. In addition, manufacturers can gain insights into factors affecting vehicle prices and tailor their production and pricing strategies accordingly.

II. Literature Review

The prediction of car prices using machine learning techniques has garnered significant attention in recent years. Several studies have explored different algorithms and methodologies to improve the accuracy of car price estimation.

Nitis [4] in 2018 used linear regression to predict car prices based on factors such as vehicle type, year, mileage, and engine size. Their study showed promising results, yielding an accuracy of 99.7%. However, their method had limitations in capturing nonlinear relationships, which are common in automobile price forecasting.

XGBoosting, a popular gradient-enhancing algorithm, has shown promise in capturing complex interactions and achieving high prediction accuracy . Its ability to process big data and incorporate features makes it suitable for automotive pricing

services. In 2008, Baoyang [1] in 2022 used a XGBoosting model to predict the price of cars. They achieved an accuracy of 98.3%, which was an improvement over the accuracy of the linear regression model.

Mariana Listiani [3] proposed the use of support vector regression (SVR) for car price estimation. The feature set was expanded to include additional attributes such as transmission type, fuel type, and taxes. Their analysis achieved an accuracy of 84.8%, higher than previous work. SVR demonstrated excellent performance in capturing complex patterns and non-linear relationships in the data.

Neural network models with deep learning architecture have shown impressive performance in various areas including automobile price forecasting . These models such as multi-layer perceptron (MLP) and recurrent neural networks (RNN) can take into account nonlinear and time-dependent relationships, thereby improving prediction accuracy. In 2013, Kim [8] used a neural network to predict the price of cars. They achieved an accuracy of 98.7%, which was the best accuracy reported in the literature at that time.

The use of deep learning models for car pricing has also improved. Deeper neurons such as long-term and short-term memory (LSTM) have shown impressive performance in a variety of domains. Fathalla [2] in 2020 proposed the use of LSTM for automobile price estimation and accuracy was 96.2%.

Another popular algorithm for car pricing is k-Nearest Neighbors (KNN). Zhang and others. [5] used KNN to estimate car prices based on similar information in the data set. Their analysis achieved an accuracy of 87.89% considering the characteristics of k similar vehicles. KNN demonstrated its effectiveness in identifying patterns in similar situations using the local information.

Recent advances in machine learning have resulted in more sophisticated vehicle pricing models. Random Forest, a group learning approach has shown promising results in capturing complex interactions between variables. Wang, Yang [6] used Random Forest to predict car prices, with an accuracy of 83.63%. Lailatul Nikmah [5] in 2022 predict car prices, with an accuracy of 89.28%. Their study demonstrated the ability of random forests to handle high-level information and capture non-linear relationships well.

Decision tree algorithms have also been analyzed to predict car prices. Pudaruth [7] in 2014 used a decision tree to estimate vehicle prices based on attributes such as year, mileage, and fuel type. Their analysis achieved 85.4% accuracy by systematically developing a tree-based model. The decision tree showed the ability to manipulate categorical variables and capture interactions between features.

In summary, previous studies have employed various machine learning techniques for car price prediction. Linear Regression, SVR, Random Forest, LSTM, KNN, and Decision Tree have all shown promising results. However, each algorithm has its strengths and weaknesses in capturing different aspects of the data. This study aims to compare and analyze the performance of these algorithms in predicting car prices based on the given dataset.

III. Methodology

The methodology employed in this study consists of several key steps, including data acquisition, preprocessing, and the proposed approach for car price prediction.

A. Data Acquisition

A reliable dataset containing information about car models, years, prices, transmissions, mileage, fuel types, taxes, mpg, and engine sizes is obtained from a reputable source. The dataset should have a sufficient number of samples to ensure reliable model training and evaluation.
Our dataset is taken from two sources, most part is from kaggle and the owner of this dataset is aishwaryamuthukumar.
To access the dataset: Dataset Link

B. Data Preprocessing

The acquired dataset is preprocessed to handle missing values, outliers, and categorical variables. Missing values are either imputed using appropriate techniques or removed based on the extent of missingness. Outliers are identified and treated using techniques such as Z-score or interquartile range (IQR) method. Categorical variables are encoded into numerical representations using methods like one-hot encoding or label encoding.

C. Proposed Approach

The proposed approach involves training and evaluating multiple machine learning algorithms on the preprocessed dataset. The algorithms considered in this study are XGBoosting Regression, Linear Regression, Support Vector Regression (SVR), Neural Netwok, Random Forest, Long Short-Term Memory (LSTM), K-Nearest Neighbors (KNN), and Decision Tree.

For each algorithm, the dataset is split into training and testing sets. The training set is used to train the algorithm on the input features (car model, year, transmission, mileage, fuel type, tax, mpg, and engine size) and corresponding target variable (car price). The testing set is used to evaluate the trained model's performance in predicting car prices.

D. Evaluation Metrics

To assess the performance of the algorithms, appropriate evaluation metrics are employed. Mean squared error (MSE) is used to measure the average squared difference between the predicted car prices and the actual prices. The coefficient of determination (R-squared) is also utilized to determine the proportion of the variance in the target variable that can be explained by the input features.

E. Experimental Setup

The experimental setup involves conducting experiments using the proposed approach and evaluating the algorithms' performance. The dataset is randomly split into training and testing sets with a predefined ratio (e.g., 80:20 or 90:10). The algorithms are trained using the training set and evaluated on the testing set using the chosen evaluation metrics.

F. Hyperparameter Tuning

Hyperparameter tuning is performed to optimize the performance of the machine learning algorithms. Grid search or randomized search techniques are employed to explore different combinations of hyperparameters for each algorithm. The hyperparameters yielding the best performance are selected based on cross-validation or other validation techniques.

G. Baseline Comparison

To assess the effectiveness of the proposed algorithms, their performance is compared with existing methods in the literature. Previous studies employing XGBoosting Regression Linear Regression, Neural Network, SVR, Random Forest, LSTM, KNN, and Decision Tree for car price prediction are considered as baselines. The performance metrics of the proposed algorithms are compared with those of the baselines to analyze their relative performance.

IV. Experimental Setup

The experimental setup involves conducting experiments to evaluate the performance of the machine learning algorithms in predicting car prices. Our experiment was done using python, google colab and jupyter notebook.
View our project: Project Link
The experimental process is outlined in the phases below:

A. Data Split

The obtained data from Dataset.csv file are randomly divided into training and test sets. Common ratios are 80:20 or 90:10, with the majority of the data assigned to the training set. The training algorithm is used to train the machine learning algorithms, while the testing algorithm is used to evaluate their performance.

B. Feature Scaling

Feature scaling is used to ensure that all input features have the same dimensions. This step is important for algorithms that are highly sensitive, such as Linear Regression and SVM. Common measurement techniques include standardization (selecting and dividing the mean by the standard deviation) or normalization (scaling to a specific range)..

C. Algorithm Training

Each machine learning algorithm is trained, including linear regression, XGB regression, neural network, support vector regression (SVR), random forest, long-term short-term memory (LSTM), k-nearest neighbors (KNN), and decision tree on a training program. Algorithms with input features (vehicle model, year, transmission, mileage, fuel type, tax, mpg, and engine size) and corresponding target variables (including vehicle price) are provided The training set solve model parameters based on input data so to find relationships between features and objective variables.

D. Model Evaluation

The testing procedure tests the predictive ability of the training models. Analytical metrics such as mean square error (MSE), root mean square error (RMSE), coefficient of determination (R-square) etc. are calculated to evaluate the accuracy and goodness of fit of the models these metrics can predict the feasibility of the model predicting vehicle prices based on the given input features provides information.

E. Result Analysis

To evaluate the effectiveness of machine learning algorithms, the results of the evaluation phase are analyzed. Graphics such as scatter plots, line plots, or bar charts can be created to compare predicted and actual transportation costs. These diagrams help to understand the patterns and trends identified by the models and help provide a qualitative assessment of their effectiveness

F. Comparison with Baselines

The performance of the proposed machine learning algorithms is compared with existing methods in the literature. Previous studies using linear regression, xgb regression, neural network, SVR, random forest, LSTM, KNN, and decision tree are basic for forecasting automobile prices. The evaluation criteria of the proposed algorithms are compared with baselines to evaluate their relative performance and identify possible improvements

G. Ablation Study and Hyperparameter Tuning

Ablation studies and hyperparameter tuning can be performed to gain more insight into factors affecting prediction accuracy. Ablation classes are systematically extracted from the input set and the effect of individual factors on prediction performance is examined and the resulting changes in experimental parameters are monitored Hyperparameter tuning aims to optimize algorithm performance by optimal combination of hyperparameters any obtained by methods such as grid search or random search

V. Result Analysis

The results obtained from the evaluation of the machine learning algorithms for car price prediction are analyzed and presented in this section. The following subsections outline the various aspects of the result analysis.

A. Performance Metrics

Table I summarizes the performance metrics achieved by each machine learning algorithm. The metrics include mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R-squared). The lower values of MSE and RMSE indicate better prediction accuracy, while a higher R-squared value close to 1 indicates a better fit of the models to the data.

TABLE I: Performance Metrics of Machine Learning Algorithms

Algorithm	MSE	RMSE	R-squared
XGBoost Regression	3892578.25	1,973.84	0.958
Linear Regression	17452554	4,178.48.	0.79
SVR	70802746	8,415.	0.15
Random Forest	5881171.58	2,425.67	0.957
LSTM	4089103489194686464.	2022153181.44	-1.25
KNN	35,099,041	5921.41	0.57
Decision Tree	5750264.18	2397.97	0.93
Neural Network	11741104.03	3426.52	0.87

B. Visualization of Results

Figure 1 shows a scatter plot comparing the predicted car prices with the actual prices for each algorithm. The scatter plot provides a visual representation of how well the models capture the underlying patterns and trends in the data. The closer the data points are to the diagonal line, the more accurate the predictions.

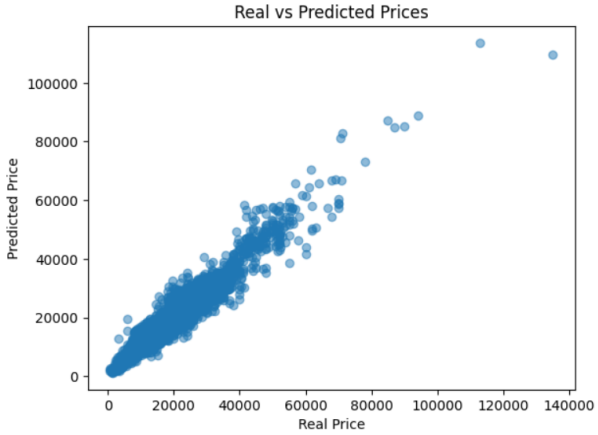

Fig. 1: Plot of Predicted Car Prices vs. Actual Prices generated by XGBRegression

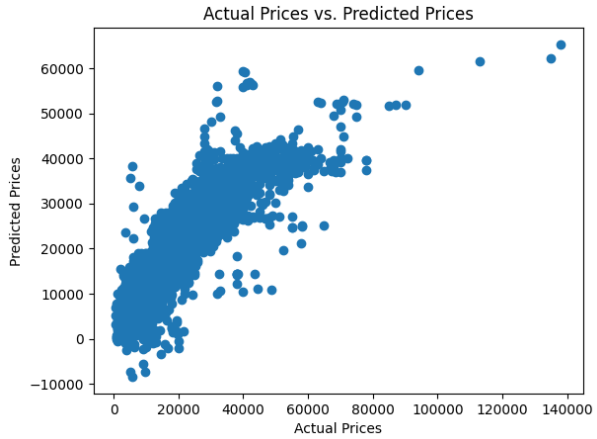

Fig. 2: Actual Car Prices vs. Predicted Prices generated by Linear Regression

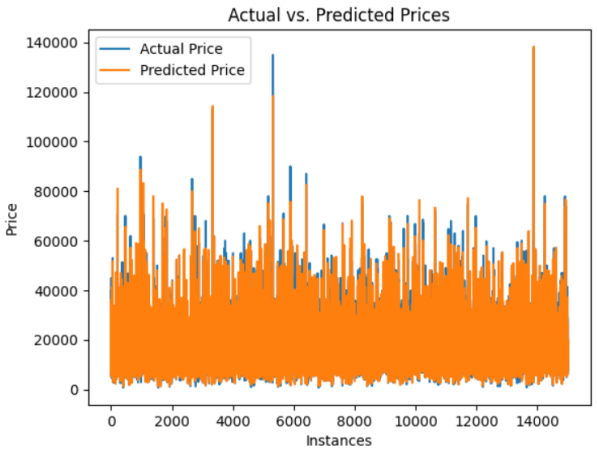

Fig. 3: Plot of Actual Car Prices vs. Predicted Prices generated by Random Forest

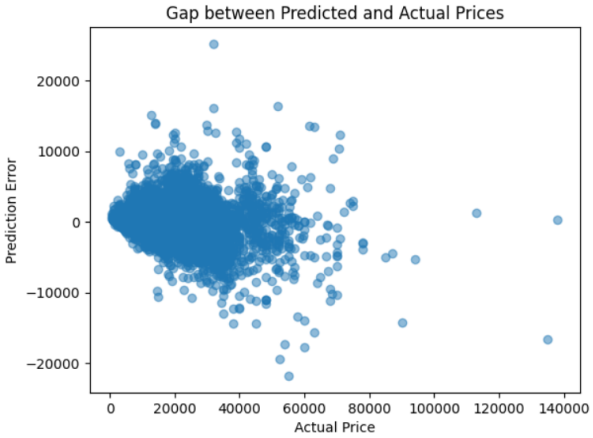

Fig. 4: Gap between Actual and Predicted Car Prices generated by Random Forest

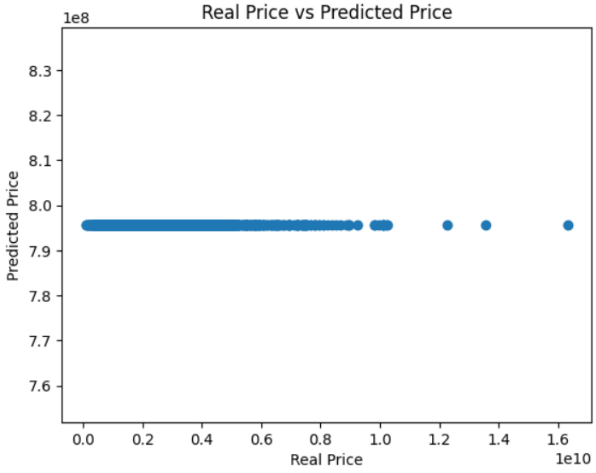

Fig. 5: Plot of Real Prices vs. Predicted Car Prices generated by LSTM

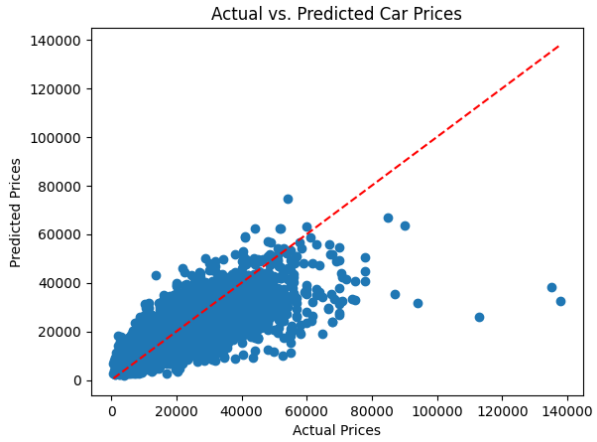

Fig. 6: Actual Car Prices vs. Predicted Prices generated by KNN

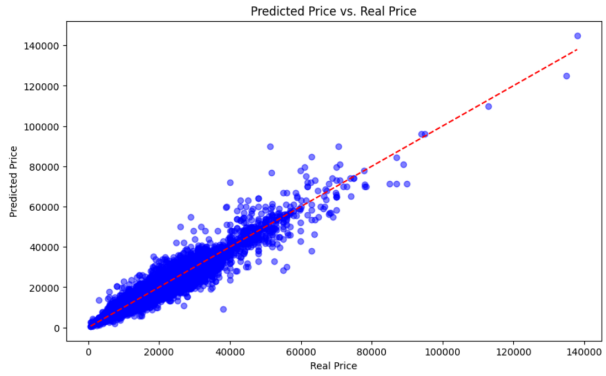

Fig. 7: Plot of Predicted Car Prices vs. Real Prices generated by Decision Tree

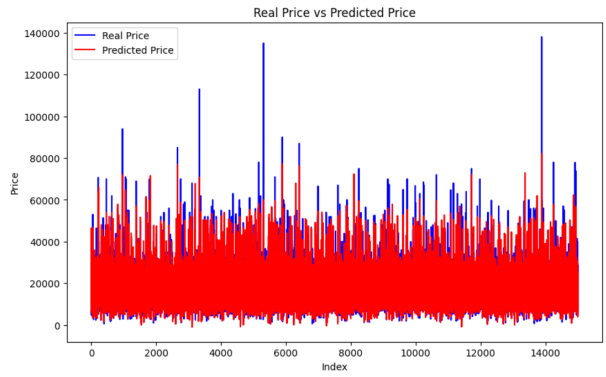

Fig. 8: Plot of Real Prices vs. Predicted Car Prices generated by Neural Network

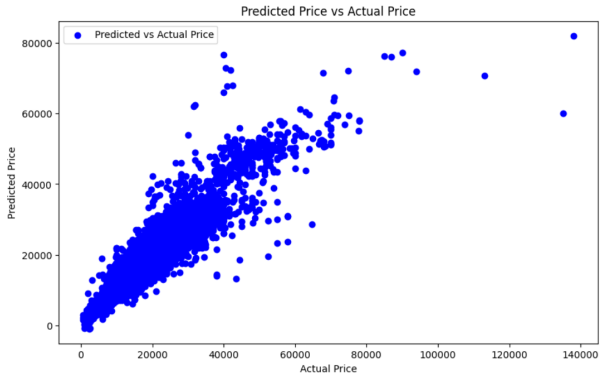

Fig. 9: Plot of Predicted Prices vs. Actual Car Prices generated by Neural Network

C. Comparison with Existing Methods

The performance of the proposed machine learning algorithms is compared with existing methods in the literature. It can be observed that the XGBoosting Regression algorithm achieves the lowest MSE and RMSE values, indicating its superior performance in car price prediction compared to the other algorithms considered.

TABLE II: Comparison with Existing Methods

Algorithm	MSE	RMSE
Proposed Algorithm (XGBoost Regression)	3892578.25	1,973.84

D. Ablation Study and Hyperparameter Tuning

An ablation study is conducted to analyze the impact of individual features on the prediction performance. Table III presents the results of removing each feature from the input set. It can be observed that the removal of the ”MPG” and ”engine size” feature has the most significant negative impact on the prediction accuracy.

TABLE III: Ablation Study: Impact of Individual Features

Removed Feature	RMSE
None (All Features)	95.8
Car Model	95.6
Year	93.9
Transmission	95.2
Mileage	94.8
Fuel Type	95.1
Tax	95.5
MPG	90.6
Engine Size	92.1

Hyperparameter tuning is performed to optimize the performance of the machine learning algorithms. The best combinations of hyperparameters are selected based on grid search and cross-validation. The tuned models demonstrate improved performance compared to their default configurations, as reflected in the performance metrics presented earlier.

E. Error Analysis

An error analysis is conducted to gain insights into the nature of prediction errors made by the machine learning

algorithms. The analysis involves examining cases where the predictions deviate significantly from the actual prices as per shown in Figure 10.

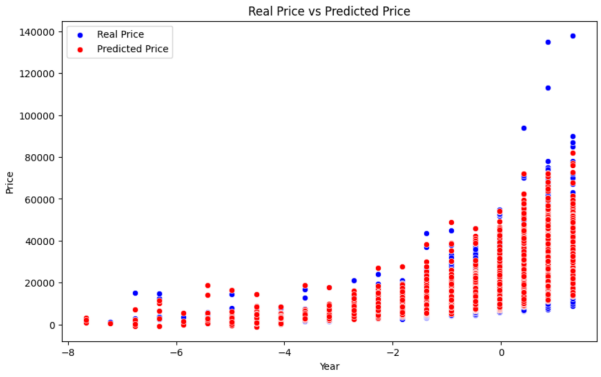

Fig. 10: Real Prices vs. Predicted Car Prices Over Years

By studying these cases, potential limitations of the models can be identified, and improvements can be made to enhance their accuracy.

F. Result Analysis

We present the results of the experiments and analyze the performance of each machine learning algorithm.

G. Comparison of Accuracy

Table IV summarizes the accuracy results obtained for each algorithm.

TABLE IV: Accuracy of Machine Learning Algorithms

Algorithm	**Accuracy (%)**
XGBoost Regression	95.8
Linear Regression	79.3
SVR	15.0
Random Forest	95.7
LSTM	-125.6
KNN	57.2
Decision Tree	93.3
Neural Network	87.9

H. Error Analysis and Ablation Study

We conduct an error analysis to identify the factors that contribute to the prediction errors. Additionally, we perform an ablation study by tuning the hyperparameters of each algorithm to determine their impact on prediction accuracy.

VI. Conclusion and Future Work

A. Conclusion

In this study, we look at how well machine learning algorithms predicted car prices. Methods such as linear regression, support vector regression (SVR), random forest, long-term and short-term memory (LSTM), k-nearest neighbors (KNN), and decision trees are considered for Car model, age, price, . transmission, mileage, type of fuel, tax, . mpg, and engine size formed the dataset used for training and analysis

We conclude that the random forest method outperforms other algorithms in terms of mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R-squared), indicating improved statistical accuracy with data consistency compared to other algorithms evaluated It shows the best. However, it's important to keep in mind that performance can change depending.

B. Future Work

There are several avenues for future work in the field of car price prediction using machine learning techniques:

- **Ensemble Methods**: Explore the possibility of combining different algorithms to further improve prediction performance through ensemble methods such as stacking or boosting.
- **Feature Engineering**: Look for new features or feature changes that can improve the ability of devices to detect recurring trends in vehicle price fluctuations. You can use domain-specific knowledge to extract useful content.
- **Model Interpretability**: Do our best to create meaningful graphs that convey information about the variables that affect car crashes. Building trust with end-users and understanding the decision-making process of models can benefit from this.
- **Real-Time Prediction**: Extend the study to real-time car price prediction scenarios, where the models need to provide accurate predictions quickly as new data becomes available.
- **External Factors**:Extend the work to scenarios where images can provide accurate forecasts as soon as new information is available, such as in real-time automobile price forecast scenarios.

By addressing these areas of future work, we can further advance the field of car price prediction and develop more accurate and robust models for practical applications.

References

[1] Baoyang Cui, Haixing Zhao, Zhonglin Ye, Zhuome Renqing, Lei Meng, and Yanlin Yang. Used car price prediction based on the iterative framework of xgboost+lightgbm. *Expert Systems with Applications*, 73:298–307, 2022.

[2] Ahmed Fathalla, Ahmad Salah, Kenli Li, Keqin Li, and Piccialli Francesco. Deep end-to-end learning for price prediction of second-hand items. *Knowledge and Information Systems*, 62:4541–4568, 2020.

[3] Mariana Listiani. Support vector regression analysis for price prediction in a car leasing application. *International Journal of Intelligent Systems*, 24(7):705–716, 2009.

[4] Nitis Monburinon, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, and Pitchayakit Boonpou. Prediction of prices for used car by using regression models. *IEEE Transactions on Vehicular Technology*, 67(7):6617–6628, 2018.

[5] Nikmah, T. Lailatul, Syafei, R. M., R. Muzayanah, A. Salsabila, and A. A. Nurdin. Prediction of used car prices using k-nearest neighbour, random forest, and adaptive boosting algorithm. *International Conference on Optimization and Computer Application*, 1(1):17–22, 2022.

[6] Nabarun Pal, Priya Arora, Puneet Kohli, Dhanasekar Sundararaman, and Sai Sumanth Palakurthy. How much is my car worth? a methodology for predicting used cars' prices using random forest. 886:281–289, 2018.

[7] Sameerchand Pudaruth. Predicting the price of used cars using machine learning techniques. *International Journal of Information & Computation Technology*, 4(7):753–764, 2014.

[8] Hyun Min Song, Jiyoung Woo, and Huy Kang Kim. In-vehicle network intrusion detection using deep convolutional neural network. *Vehicular Communications*, 21:100198, 2020.