# Car Price Prediction

Yi Chen Wu, Jingkai Wang, Xiaoli Fang, Qixiang Jia
DS3000, Northeastern University

## ABSTRACT

This report presents a machine-learning model for predicting car prices. The dataset used in this study was taken from Kaggle and contains information about various car characteristics, such as mileage, engine size, and horsepower. The primary purpose of this study is to develop an accurate predictive model that car dealers can use to determine the fair market value of a car. To achieve this, we utilize machine learning algorithms such as decision tree regression, random forest regression, and linear regression to develop and evaluate the performance of our predictive models. Random Forest Regressor creates a set of decision trees, each trained on a random subset of features and data, and outputs the average prediction of all decision trees. Linear Regressor assumes a linear relationship between input features and target variables. We use these algorithms to calculate the predicted selling price of a car and evaluate its performance against various metrics. The results show that Random Forest Regressor accurately outperforms other predictive models. The solution's impact is enormous, as it can help car dealers and potential car buyers make informed decisions about car pricing.

## INTRODUCTION

The demand for the automotive industry has grown significantly over the years, making it an important sector of the global economy. *Car pricing* is a complex process that depends on various factors, including the car's make, model, age, and mileage. Accurate pricing is critical for car dealers and potential buyers, ensuring a car is priced fairly and reasonably. Therefore, pricing cars accurately have become critical for car dealers and potential buyers. In order to help with this process, our team uses machine learning algorithms to predict how much a car will sell based on various characteristics. This report describes the approach we took to obtain and prepare the data, our methodology, and the algorithm we used for model selection. We used three algorithms: Decision Tree Regression, Random Forest Regression, and Linear Regression, and we will employ cross-validation techniques to train and evaluate our models. We will also tune the hyperparameters to achieve the best model performance. We aim to achieve high accuracy in predicting car selling prices, with a mean square error (MSE) value of less than 10 in all algorithms we employ. The solution's impact is enormous because it can help car dealers and potential car buyers make informed decisions about car pricing, helping them accurately determine a car's fair market value.

Define the problem:

The problem is to develop an accurate machine-learning model that can predict a car's price based on various car characteristics, such as mileage, engine size, horsepower, etc. Car dealers can use the model to determine a car's fair market value, and potential buyers can use it to make informed decisions about car pricing.

Motivation and goals:

The motivation for this study was to provide car dealers with a tool that could help them accurately determine a car's fair market value. The main goal of this research is to develop an accurate forecasting model that can predict a car's price based on various car characteristics.

## METHODOLOGY

In this project, our goal is to predict the selling price of a car based on various characteristics such as the car's year, current price, kilometers are driven, fuel type, seller type, gearbox type, and owner.

Our first step is to import all necessary libraries such as pandas, numpy, seaborn, and sklearn. We then loaded the car dataset and checked its shape to ensure we loaded the correct number of rows and columns and looked at the shape of the dataset. We also use the describe function to get some statistics about the dataset. We use EDA to check if any missing data and data types are used. Since no missing values exist, we do not need to impute or drop any rows. However, we converted the Fuel_Type, Seller_Type, and Transmission data to strings and removed the "Car_Name" column as it was irrelevant to our analysis. Next, we visualize the relationship between the year of the car and its selling price using a scatterplot. We have observed that the price of new cars is generally higher than that of old cars.

Then, we select the desired features from the dataset and encode the categorical features using one-hot encoding. We also split the data into train and test sets using the train_test_split function. Then recheck the data type. For our model selection, we first used a scatterplot matrix to visualize pairwise relationships between features in the car data. We then use the multiple linear regression model as our baseline model to predict the car's selling price. In this type of regression, we try to predict the target variable based on two or more independent variables. We first create a scatterplot matrix using the Seaborn library to visualize the pairwise relationships between features in the car data. The data is then split into train and test sets using the train_test_split function from the Scikit-learn library. Then use the LinearRegression() function from the Scikit-learn library to create a linear regression model, and then train the model with the training data by calling the fit() function. Finally, call the predict() function, use the trained model to predict the target variable of the test data, and calculate the mean square error (MSE) to evaluate the model's performance.

We also used decision trees and random forest regression to see if we could improve the model's performance. Perform cross-validation on decision tree and random forest regression models.

For decision trees, we first split the data into train and test sets using the train_test_split function. Then, train a decision tree regression model on the training data using the appropriate method. Next, use the cross_val_score function to perform 5-fold cross-validation on the training data and the mean square error for each fold. Finally, print the mean square error for each fold by taking the absolute value of the scores array and the mean square error by taking the mean of the absolute scores. Finally, the average MSE showing all folds is obtained.

We then add cross-validation to DecisionTreeRegressor and an example of using cross-validation to find the best hyperparameters for RandomForestRegressor and the best value for the hyperparameter n_estimators using random search with ten iterations and 3-fold cross-validation. The custom_scoring_function below is defined as mean squared error, returning negative MSE values to maximize the score rather than minimize it. It also demonstrates how to evaluate the performance of the best RandomForestRegressor on the test set and how to visualize the importance of features in the model. Finally, we show how to use the KNeighborsRegressor to predict the selling price of a car and how to plot the mean square error, a KNN regressor with different values of k, to determine the optimal value for the number of neighbors.

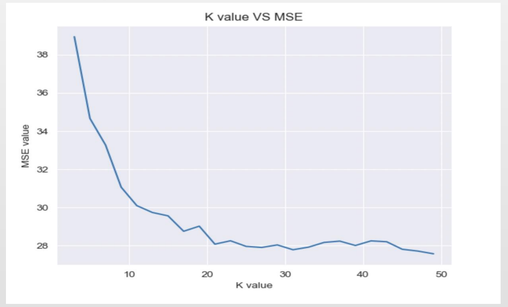
## RESULTS AND EVALUATIONS

The analysis shows that the random forest regression performed best, with an R^2 score of 0.74, an MAE of 1.01, and an MSE of 7.03. The best parameters for this model are max_depth=12, max_features='log2', min_samples_leaf=1, min_samples_split=3, and n_estimators=498. Random forest regression shows that the model can predict car prices within a reasonable margin of error. An R^2 value of 0.74 indicates that the regressor explains approximately 74% of the variance in the data.

The decision tree and multiple linear regression models had high mean square errors, suggesting that they are not as accurate as the random forest model.

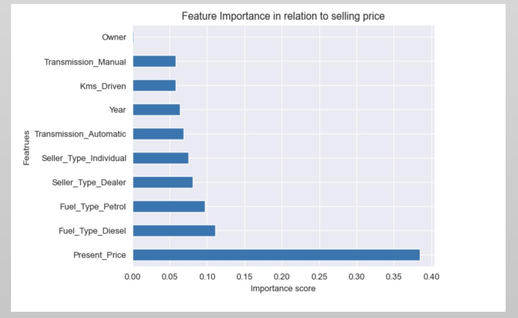
The average mean square error of the decision tree regression was 2.81, while the mean square error of the multiple linear regression model was 7.45, relatively higher than the random forest regression.


For KNN regression, the mean squared error decreases as neighbors increase. The best K value is around 20, which shows that the model can capture some potential patterns in the data and make reasonable choices to accurately predict car prices based on the input features.




The random forest regression is a promising approach for car price prediction with reasonable accuracy compared to the other two, as it has the lowest MAE and relatively low MSE. However, it is essential to note that, yes, the accuracy of any model is highly dependent on the quality and quantity of data used to train it. Therefore, it is essential to continually evaluate and refine the model as more data becomes available.

We also use a random forest regressor to identify important features that predict a car's selling price, and create a visualization that shows that the most important feature affecting selling price is the current price of the car.




## RELATED WORK

The Kaggle project titled "Car Price Prediction - Machine Learning Project" was created by user Swatikhedekar. The project aims to build a machine learning model to predict the selling price of a car based on various factors such as the car's make, model, year, fuel type, and other specifications. The dataset used in this project contains information on scrapped used cars in the Indian market. It is related to our project because both aim to predict the selling price of a car using machine learning algorithms. This project provides a similar dataset of car characteristics, including year, model, mileage, engine size, and selling price.

Additionally, the project utilizes machine learning algorithms, including linear regression, decision tree regression, random forest regression, and XGBoost, to predict car prices. Our project also plans to use similar algorithms and cross-validation techniques to ensure our models stay within the training data. This project is a good reference value and inspiration for our project.
https://www.kaggle.com/code/swatikhedekar/car-price-prediction-ml-project/comments

## IMPACT

The potential impact of our solution is enormous, as accurately predicting car prices can benefit both buyers and sellers in the automotive industry.

For Buyers:

Buyers can benefit from our solution by obtaining more accurate information about the car's actual market value, which can help them make smarter car buying decisions and negotiate better prices. Our solutions help buyers avoid overpaying for their cars, saving them money in the long run.

For Sellers:

Sellers can also benefit from our solution by using it to price their cars, which can help them sell their cars faster and at a more competitive price. Accurate pricing can also help dealers manage inventory more efficiently and avoid holding vehicles for long periods, reducing overall costs.

Other potential beneficiaries of our solutions include insurance companies and lenders. Insurance companies can use our solutions to help determine the value of a vehicle and set premiums accordingly. Lenders can use our solution to assess the value of vehicles when considering loan applications, which can help them reduce risk and make more informed lending decisions.

Overall, our solution has the potential to benefit buyers and sellers in the auto industry as well as other stakeholders, such as insurance companies and lenders, by providing more accurate car price forecasts.

## CONCLUSION

In summary, our goal is to develop a machine-learning model to predict the selling price of used cars and evaluate their performance using mean squared error and cross-validation. According to the results and analysis, the aims and objectives of this project were achieved to some extent; we used four models: Multiple Linear Regression, Decision Tree Regression, Random Forest Regression, and K Nearest Neighbors Regression. Of these, the Random Forest Regressor model performed the best, with an R^2 score of 0.74 and a mean absolute error (MAE) of 1.01.

In general, the purpose and goal of this project have been achieved because we were able to use machine learning techniques to successfully predict the selling price of used cars. Through visual chart analysis, we found that the feature that most affect the selling price of used cars is the car's selling price. However, there is room for improvement. First, we can explore other regression models in future work to see if they can outperform the currently used models. Also, we can try to get more data to train the model, which might lead to more accurate predictions. Secondly, many other factors may affect the price of a used car, such as the vehicle's appearance, maintenance records, accident records, market supply, and demand. We can add these factors that may affect the price of the car to the improvement of this project. However, quantifying or obtaining data from these factors may require more work. While the random forest model provides a good starting point, further improvements can be made to make the model more accurate and helpful in predicting used car selling prices.

## REFERENCE

Swati Khedekar. (2021). Car Price Prediction - ML Project. Retrieved April 14, 2023, from https://www.kaggle.com/code/swatikhedekar/car-price-prediction-ml-project/notebook.

Mahmoud Chami(2023)Car Price. https://www.kaggle.com/code/slaivre/car-price-prediction