Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Research for Car Price Prediction Base on Machine Learning

Xianshun Jiang

Rose-Hulman Institute of Technology, Terre Haute, IN, 47803, the United States

Jiangx6@rose-hulman.edu

Abstract. This report details a machine learning project aimed at predicting car prices, which is crucial and meaningful for business owners to forecast the market, avoid financial loss, and for buyers to have a rough overview of each car's price. To accomplish the goal, this project using the Car Price Prediction Challenge dataset from Kaggle. The dataset includes a variety of automobile attributes such as brand, model, year of manufacture, engine specifications, mileage, and other relevant factors. The project's goal was to create an accurate and reliable model to predict car prices based on these characteristics. To accomplish this, the paper utilized multiple machine-learning algorithms and techniques to analyze the dataset. This research's findings highlighted the effectiveness of these machine-learning methods in car price prediction. The random forest regressor emerged as the best-performing model, achieving high accuracy with a low mean squared error and a high R-squared value. Additionally, the feature importance analysis sheds light on the key factors influencing car prices, offering a deeper understanding of market dynamics.

Keywords: Machine Learning, Car Price, Supervised Learning.

1. Introduction

It's important for people who are willing to buy, sell, or foresee the market. When purchasing a car, particularly a used one, there are numerous important features to consider. Safety features such as airbags, seat belts, and braking systems are paramount to ensure passenger protection in the event of an accident. Buyers also often consider the car's fuel efficiency and environmental impact, as these factors affect long-term costs and contribute to sustainability. Additionally, the car's age, classification, size, color, and capacity must align with the buyer's needs and lifestyle, whether for transporting a family or hauling cargo. Other notable features include technology and entertainment systems, interior and exterior design, and performance specifications. These features are also important factors in the prices, varied values may result in a significantly different final price [1].

However, determining which of these features most significantly affects a car's value can be challenging. Understanding the key aspects that influence a car's price is crucial for both buyers and sellers. For buyers, recognizing that a lower-priority feature might significantly impact the car's price can be enlightening. For both parties, avoiding disadvantageous deals is essential. This paper proposed a machine learning model to effectively and accurately predict car prices for both parties for reference.

Given the necessity of affordable cars, this project explores various regression models to predict car prices based on their features. The paper's goal was to accurately predict a car's price using the data available in the Car Price Prediction dataset. This dataset includes numerous features but notably excludes car condition, which is difficult to quantify objectively. The closest available feature to the condition is age.

The paper aimed to identify the best regression model for predicting car prices, focusing on achieving the highest R-squared value. The paper tested a variety of regressors and analyzed their performance and behavior concerning key features, such as mileage, to determine the most effective model.

2. Literature Review

Asghar uses the RFE method, which is Recursive feature elimination, to find the weights of the features [2]. In the statistical test performed for their model, he did recursive feature elimination, ordinary least squares regression, variance inflation factor, and dropping features if the p-value is high. Meanwhile, in the project, the paper uses R2 and the accuracy rate as the criteria instead of the p-value during the feature selection process. Moreover, the paper's research did not build based on this paper because they focused too much on the feature engineering and selection process and did not demonstrate the choice of machine learning algorithm clearly.

Another paper is written by Sameerchand Pudaruth [3]. This paper used five methods: multiple linear regression, K-Nearest Neighbor (KNN), Naive Bayes, decision tree, and random forests, and compared their performance on which algorithm is better by listing their accuracy in tables. The paper's approach was similar since it also compared the algorithm's performance by listing the results and the paper used all the algorithms except naive Bayes. The difference is that the article used cut to cut price into six categories, so they were doing a classifier problem, but the paper was doing a regression. In addition, the data in that article was older and from a different country.

Gao's paper is about a multi-task learning model, which beats all state-of-the-art models [4]. They did not try to improve existing models, but they developed their own instead. Their approach inspired us to think about new models and have a data set that captures the crucial features that affect the car's price. The paper instead compared several models and selected the best one as the final model.

Kalehbasti, et al. used several models to predict the prices of Airbnb rental houses, including linear regression, decision trees, and k-means clustering [5]. They used linear regression with lasso and ridge regression to perform feature selection and then used the selected features on their other prediction models. Their approach inspired us to compare multiple models, fine-tune them, and finally select the best model among these models.

3. Process

3.1. Data Source

The dataset the paper uses, titled "Used Car Information," is available on Kaggle and contains 19,237 records and 18 features [6]. These features include categorical data such as manufacturers, models, and colors, as well as numerical data like mileage, cylinders, and production year. To facilitate predictive modeling, this research applied techniques such as one-hot encoding for the categorical data and standardization for the numerical features, except for the production year. By framing the problem as a regression task, it focused on predicting car prices using the available features. Regression analysis allows this research to create models that estimate the continuous numerical value of a car's price by considering the relationships between the features and the target price.

3.2. Preprocessing

The data preprocessing phase involved several key steps to ensure quality and relevance for the car price prediction task. The modifications made to the dataset included:

1. **Removal of Unique and Non-useful Features**: Features such as IDs and levies, which were unique identifiers or had limited relevance for price prediction, were removed.
2. **Filtering Based on Car Year**: To focus on cars with a regular valuation curve and avoid complications from classic cars, only cars manufactured from the year 2000 or newer were considered.
3. **Removal of Redundant Features**: The "Manufacturer" feature was removed to avoid redundancy with the "Model" feature, as the model alone captures the necessary manufacturer information.

4. **Handling Outliers and Impossible Data**: Outliers, such as cars priced below $1,000 (indicative of junk data) or above $150,000 (considered unaffordable outliers), were excluded. Additionally, a car using "Hydrogen" as fuel was removed.
5. **Conversion of Text to Floats**: Textual data with units like "km" associated with numeric values were converted to floats by removing the "km" designation, ensuring uniformity for analysis and modeling.
6. **Removal of Models with Insufficient Data**: Models with fewer than 20 instances were removed to enhance the significance of the results and prevent inconsistencies between the train and test sets.

After implementing these preprocessing steps, the final dataset comprised 12,401 records and 163 features, resulting in a more focused and reliable subset ready for subsequent analysis and modeling in the car price prediction challenge.

3.3. Classifiers/Regressors and Tuning

1. Simple Bias Regressor:

To establish a baseline for comparison, this research employed a simple bias regressor to evaluate the performance of more sophisticated models against a basic approach. This research selected the bias to be the mean price of the dataset to minimize the mean squared error (MSE) and determine whether more complex models could outperform this straightforward baseline.

2. Linear Regression:

In addition to the simple bias regressor, this research employed linear regression as it is a fundamental and widely used algorithm for regression tasks. The linear regressor assumes a linear relationship between the input features and the target variable, which aligns with the objective of predicting car prices based on various features. By utilizing a linear regressor, it aimed to capture any linear patterns or trends that may exist within the dataset, allowing us to make informed predictions about car prices.

3. Hyperparameter Tuning:

To enhance the performance of the linear regressor, this research incorporated L1 and L2 regularization techniques known as lasso and ridge regression. These regularization methods help prevent overfitting by introducing a penalty term that constrains the magnitude of the coefficients. To determine the optimal hyperparameter alpha (α) for both lasso and ridge regression, this research employed a grid-search approach combined with cross-validation. The grid search involved evaluating a range of alpha values while using cross-validation to estimate the performance of each combination. This process allowed us to identify the alpha values that yielded the best performance in terms of minimizing the mean squared error (MSE) and achieving optimal regularization for the linear regressor.

4. Lasso Regression:

Lasso Regression is a linear regression technique that incorporates regularization by adding a penalty term to the loss function. It aims to both fit the data and perform feature selection by driving some of the coefficients to exactly zero. This property makes Lasso Regression useful for feature selection in high-dimensional datasets.

The hyperparameter alpha in Lasso Regression represents the regularization strength or the level of penalty applied to the model's coefficients. It controls the trade-off between fitting the training data well and keeping the coefficients small. A higher alpha value increases the amount of regularization, driving more coefficients to zero and promoting sparsity in the model.

5. Ridge Regression:

Ridge regression is another linear regression variant that employs L2 regularization. It adds a penalty term to the cost function that is proportional to the square of the magnitude of the coefficient, thereby encouraging smaller and more evenly distributed coefficient values.

Similar to Lasso regression, this research performed hyperparameter tuning for Ridge regression by searching for the optimal alpha value. Through cross-validation and grid search, this research selected the alpha value that maximized the model's performance.

6. Random Forest Regressor:

Random Forest Regression is an ensemble learning method that combines multiple decision tree models to make predictions. It works by creating a forest of randomized decision trees, where each tree is trained on a random subset of the training data and a random subset of the input features. The final prediction is obtained by taking the mean of all predictions in the ensemble.

This research again employed a grid-search approach with cross-validation to optimize the hyper-parameter "n_estimators" which refers to the number of decision trees to be included in the ensemble. Increasing the number of estimators leads to a more accurate and stable model at the cost of additional training time.

7. KNN Regressor:

The k-nearest algorithm is an unsupervised algorithm that can be used for both classification and regression tasks. It finds the k nearest points and predicts based on their categories or values. Here it uses a KNN regressor.

In KNN, all the numerical variables need to be on the same scale to make sure they are in the same weight when calculating the distance between points. Different kinds of distance, like the city block distance or the Euclidean distance, can be used. KNN is an efficient algorithm and can produce satisfactory results in a short amount of time.

8. Decision Tree Regressor:

Decision Tree is a hierarchical tree-like model that is used in predicting outcomes. This can be used for decision analysis but is also popular in machine learning. Here it uses a decision tree regressor.

The most important aspects of the decision tree are the decision rules, i.e. which attributes to divide the sub-categories. This process will continue and make a tree with many nodes. For all the leaf nodes, the decision tree will predict the majority class or value.

9. Hyperparameter Tuning:

To optimize the Random Forest regressor, it performed hyperparameter tuning using a grid search. The key hyperparameter it focused on was the number of estimators (trees) in the Random Forest. It created a grid of values ranging from 10 to 200 with a step size of 10. The goal was to find the optimal number of estimators that maximized the model's predictive performance.

During hyperparameter tuning, it utilized k-fold cross-validation with 5 folds. This technique allowed us to train and evaluate the Random Forest models on different subsets of the training data. By assessing the models' performance across multiple folds, it obtained more reliable estimates of their performance and selected the configuration with the highest mean cross-validated R-squared score.

10. Gradient Boosted Trees:

In the car price prediction project, it also employed the Gradient Boosted Trees algorithm. Gradient Boosted Trees is an ensemble learning method that combines multiple decision trees sequentially to make predictions. Each subsequent tree is trained to correct the errors made by the previous trees, gradually improving the model's performance.

The algorithm works by initially fitting a simple decision tree to the training data. Subsequently, additional trees are trained to minimize the residuals (the differences between the predicted and actual values) of the previous trees. The final prediction is obtained by summing the predictions of all the trees.

4. Experimental Setup and Results

4.1. Feature Importance

It identified the key factors contributing to a car's value by evaluating the significance of each feature in this research's models. Feature importance scores were calculated for the Random Forest Regressor and Gradient-Boosted Trees Regressor, quantifying the relative contribution of each feature to the models' predictive abilities.

The analysis revealed several significant insights:

1. **Production Year**: This emerged as the most influential feature with an importance score of 0.267. Newer cars generally have higher prices, underscoring the significant impact of a car's age on its value.
2. **Mileage**: With an importance score of 0.099, mileage was the second most important feature. As expected, higher mileage tends to lower a car's value, reflecting the common understanding that more usage leads to depreciation.
3. **Engine Volume**: This feature had an importance score of 0.085. Cars with larger engine volumes, which are typically more powerful and expensive, were associated with higher prices.
4. **Airbags**: Airbags scored 0.082 in importance. More airbags, contributing to greater passenger safety, positively influenced car prices. However, this finding was not consistent throughout the project.

Additionally, other features such as brand, model, fuel type, and car category showed varying degrees of importance in predicting car prices, though their impact was relatively lower compared to the key features mentioned above.

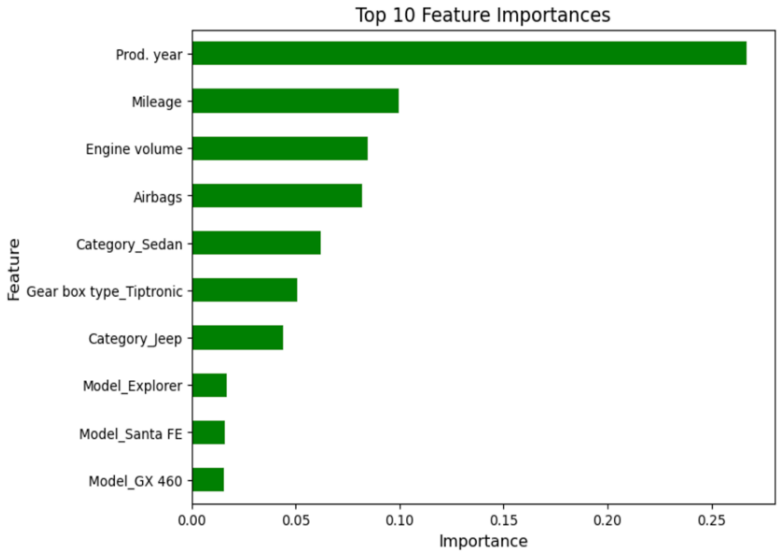

Fig. 1 Bar chart of feature importance (Photo/Picture credit: Original).

Fig. 1 visually depicts the feature importance scores derived from the models, offering a clear and intuitive understanding of each feature's relative importance in determining car prices. Fig. 1 highlights the top influential features, which are production years, mileage, engine volume, airbags, etc., and their corresponding scores, emphasizing their significance in the prediction process.

Tables 1 and 2 list the five most significant and five least significant features. Many of the most important features are commonly inquired about by potential car buyers, representing fundamental aspects of a car. In contrast, the least important features are more abstract, such as color and some less common car models.

Table 1. Most important features

Feature	Importance
Production Year	0.2670
Mileage	0.0994
Engine Volume	0.0849
Airbags	0.0818
Category: Sedan	0.0622

Table 2. Least important features

Feature	Importance
Model: Colt	0.000006
Color: Pink	0.000005
Color: Purple	0.000005
Model: Airtrek	0.000004
Model: Pajero IO	0.000002

Understanding the relative importance of these features provides valuable insights for car buyers, sellers, and industry professionals. It allows stakeholders to identify the most influential factors affecting car prices and make informed decisions regarding pricing strategies, marketing campaigns, and customer preferences.

By incorporating feature importance analysis into the car price prediction models, it was able to develop more accurate and interpretable models. The analysis helped us identify the features that drive price dynamics and provided a solid foundation for understanding the relationships between features and car prices.

4.2. KNN Regressor

This research used grid search to find an optimal number of neighbors K ranging from 1 to 20. Fig. 2 for training and testing R2 with K is attached below. The optimal K this research found was K=3 with an error rate=0.367. Using that, this research applied cross-validation and got a train R2 of 0.819 and a test R2 of 0.598. This is not a high result, but for a simple algorithm, this is satisfactory and higher than the baseline. When k is greater than 3, the model is overfitting, and both training and testing accuracy drops.

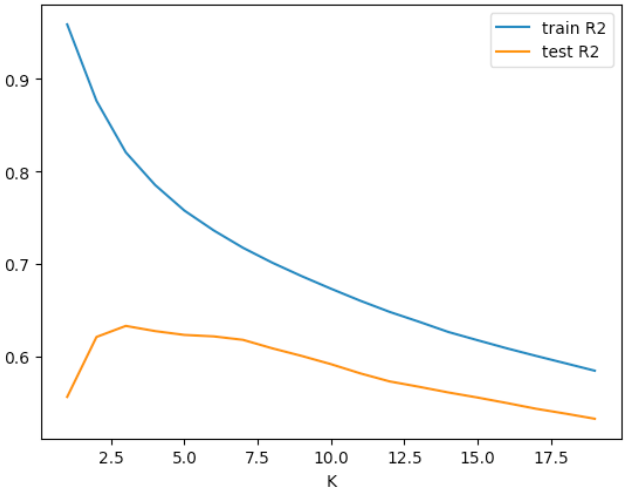

Fig. 2 KNN Regressor (Photo/Picture credit: Original).

4.3. Decision Tree Regressor

This research used grid search to find an optimal depth ranging from 1 to 20. Fig. 3 for training and testing R2 with the depth is attached below. The optimal depth found was tree depth=18 with an error rate=0.309. With this depth and cross-validation, the training R2 was 0.94, and the testing R2 was 0.701. This is slightly better than KNN, but still lower than the random forest regressor. Both training and testing accuracy increases as the depth increases, and until death is 18, it reaches the maximum accuracy.

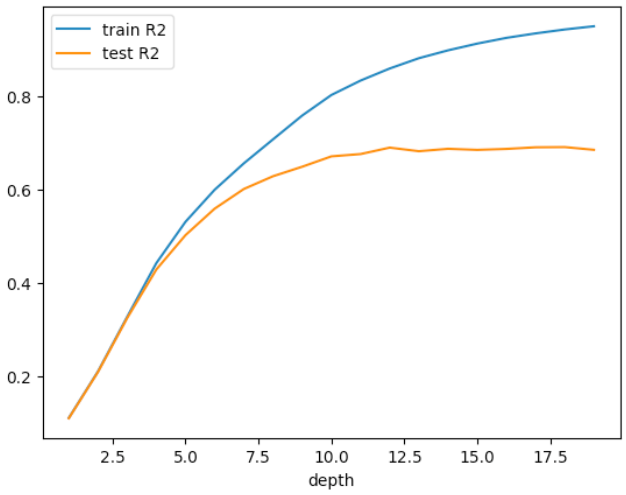

Fig. 3 Decision tree regressor (Photo/Picture credit: Original).

4.4. Decision Tree Classifier

The Decision Tree classifier reflects the Price classification based on different features. Although the Decision Tree technique is not always useful when the target data is a numerical value rather than a categorical, the classification result for this dataset is clear about the importance of each feature for price decision and was mainly used as a reference.

Since Car Price is a numerical value as above, the strategy is to divide values into different groups. For this project, it used 4 groups (0-25%, 25-50%, 50-75%, 75-100%) based on which quantile the price in the dataset stays in. Then it trained the decision tree based on these prices and saw which parts of attributes were essential to decide the car price. And it found that the production year is the most important feature in classifying cars. The production year is used for both the 1st and 3rd levels of the decision tree. This feature's importance is the same as the importance shown in the Importances

of attributes trained in the Random Forest Regressor. Then, for medium levels, Wheels Front, Fuel Type, Cylinder, and Manufacturer decide where the prices fall. The lowest levels are used for later depth mostly all the way, mixing with Airbag, Mileage, and Engine Volume. These were the same as the importance shown in Importance Trained by Random Forest as well. These lower-level attributes had low influences on Car Price Prediction. The resulting tree showed the importance of the Production Year but did not reflect the importance of Mileage or Engine Volume clearly. A possible explanation for this result is, that the Production Year's importance factor for Random Tree is 0.26, almost 3 times of Mileage Engine Volume, Airbags, etc. That made other features not important even for Decision Trees in medium levels, which made the final Decision Trees use Mileage, Engine Volume, and Airbags in lower levels.

5. Discussion

5.1. Coefficient of Mileage Discussion

During the analysis of the combined dataset, this research observed an intriguing anomaly regarding the coefficient of mileage in the models. Common intuition suggests that as mileage increases, the price of a car tends to decrease. However, this research's findings revealed that the coefficient of mileage was near zero, contradicting this conventional understanding.

This research's analysis revealed a range of mileage coefficients among different car models. This research grouped data based on models and applied Linear, Lasso, and Ridge regressors correspondingly to each model. This research found that some models have negative coefficients of mileage while some are positive. This research noticed that some luxury car brands, such as those shown in Table 3 (two Mercedes-Benz and one Lexus), had the largest positive mileage coefficients. This seems extremely counter-intuitive. In the meanwhile, in Table 4, this research sees manufacturers such as Ford and Toyota having the most negative mileage coefficients, which is more intuitive. One possible explanation for this could be that Fords and Toyotas, having less prestige to their name, lose value more easily than luxury cars, which always have some amount of value. The positive coefficient for mileage, however, is difficult to explain, and may simply be chalked up to having not enough data or information, perhaps potentially including condition, about the sold cars.

Table 3. Cars with greatest mileage coefficient

Car Model	Linear Coef.	Ridge Coef.	Lasso Coef.	Mean Coef.
Mercedes E 300	4.289	8.167	8.419	6.965
Mercedes GLE 350	3.110	5.013	4.642	4.255
Lexus RX-400	7.910	-0.710	-0.760	2.147

Table 4. Cars with the smallest mileage coefficient

Car Model	Linear Coef.	Ridge Coef.	Lasso Coef.	Mean Coef.
Ford Mustang	-1.247	-1.867	-2.274	-1.796
Toyota CHR	-1.186	-2.012	-2.157	-1.785
Ford Explorer	-1.032	-1.618	-1.678	-1.443

5.2. R2 Discussion

This research checked the R-squared scores for all Ridge, Lasso, Random Forest, Decision Tree Regressor, and KNN. Table 5 shows the results of these models. The R-squared scores are bad for both Ridge and Lass. R-squared scores are both around 0.44. So, Ridge and Lasso are not suitable for prediction. For the rest models, both the Decision Tree Regressor and Random Forest have better R-squared values. It is 0.701 for the Decision Tree Regressor and 0.799 for the Random Forest. And KNN has 0.602, which is not as good as Random Forest even if it can explain some variances. With this information, it seems that the tree-related models are more suitable for predicting car prices for this dataset.

Table 5. Model comparison

Model	**R2**
Random Forest	0.799
Decision Tree (Regressor)	0.694
KNN	0.602
Ridge	0.442
Lasso	0.441

6. Conclusion

This research found that the random forest regression model performed best. This model could explain approximately 80% of the variation in the car prices, as compared to the linear regression which only accounted for 50% of the data's variance.

Considering the limitations imposed by the dataset and the issues arising from unusual records, this research recognizes the importance of investing additional effort into data inspection and validation. By ensuring the integrity and quality of the dataset, this research could have established a more robust foundation for the modeling endeavors, potentially yielding more accurate and reliable predictions.

In future projects, this research intends to allocate sufficient time for comprehensive data exploration, verification, and cleansing, ensuring that the dataset is reliable, valid, and free from anomalies. This preliminary step is crucial for generating trustworthy results and optimizing the performance of the models. Hyperparameter tuning for all models rather than one or two models is also important, this research only fine-tuned the number of estimators of random forest. And a better way of feature selection could also be developed to improve the results.

Initially, this research did not use cross-validation for some of the models. As a result, it relied solely on the performance metrics from a single train-test split, leading to significant overfitting and poor validation scores. To address this, it later applied cross-validation techniques to all models, reducing the risk of overfitting.

It also encountered the challenge of sharing fitted models without the computational cost of retraining them. By using the pickle Python module, this research were able to save the trained models in a serialized format. This enabled us to share and reuse the trained models easily, saving valuable time in the project.

Additionally, while generating test data for the presentation, this research faced an issue with one-hot encoding. This research's prediction model expected 163 features, but the test data lacked many of

these features due to one-hot encoding. To resolve this, this research one-hot encoded the test data using the original dataset, ensuring consistency with the original model.

References

[1] The Average Car in The U.S. Is 11.5 Years Old. 2015-08-03. Retrieved on 2024-05-26. Retrieved from: https://www.wbur.org/hereandnow/2015/08/03/average-car-age

[2] Asghar M, Mehmood K, Yasin S, et al. Used cars price prediction using machine learning with optimal features. Pakistan Journal of Engineering and Technology, 2021, 4(2): 113-119.

[3] Pudaruth, S. Predicting the Price of Used Cars using Machine Learning Techniques. International Journal of Information & Computation Technology, 2014, 4(7): 753-764.

[4] Gao G, Bao Z, Cao J, et al. Location-centered house price prediction: A multi-task learning approach. ACM Transactions on Intelligent Systems and Technology (TIST), 2022, 13(2), 1-25.

[5] Car Price Prediction Challenge. Retrieved on 2024-05-17. Retrieved from: https://www.kaggle.com/datasets/deepcontractor/car-price-prediction-challenge

[6] Rezazadeh Kalehbasti P, Nikolenko L, Rezaei H. Airbnb price prediction using machine learning and sentiment analysis. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction, 2021, 173-184.