Pre-Owned Car Price Prediction Using Machine Learning Techniques

Andy Zhu
Computer Science, Rensselaer Polytechnic Institute, Troy, New York, U.S.A.

Keywords: Pre-Owned Car Price Prediction, Machine Learning, Linear Regression, Decision Trees, Random Forest.

Abstract: As the car industry continues to evolve, car price prediction models have been a highly focused topic for research. In the current pre-owned car marketplace, buyers are impelled to decide whether a vehicle is within a reasonable price range. From visiting car dealers to websites, the task of finding the true worth of a car is laborious, especially for those who have limited knowledge about cars. The rise of online car listing platforms calls for buyers and sellers to be informed about the values of vehicles in the market. A tool that could determine the honest value of a vehicle would be crucial to the automotive marketplace. This paper proposes several machine learning algorithms that aim to predict the prices of used vehicles based on the features of a car. By using linear regression, decision trees, and neural network, this study explores models that best capture the price of used cars from pre-existing car sales data. The models' performances are assessed to determine the most suitable model for future price prediction applications.

1 INTRODUCTION

In today's growing automotive market, pre-owned vehicles make up the majority of auto sales in the United States. In 2021, pre-owned auto sales consisted of nearly 78% of all vehicle sales in the U.S. (Carlier and Mathilde 2023. Through the advent of online car listing platforms, people looking to purchase second-hand cars can conveniently search car listings around their areas on their personal computers. While these platforms facilitate a search experience to help consumers find cars within affordability, there is no unified algorithm shared by these platforms for determining the price of a car. Competition exist between selling platforms, and each platform has their own measurement of how much a car is worth.

The demand for transparency in the U.S.' ever-rising pre-owned car industry is an issue worth emphasizing. Consumers deserve to know the worthiness of a car without the convolution of online car selling platforms advertisements when, at the end of day, the goal of the platform is to make the consumer make a purchase. In this paper, different machine learning techniques are applied to discover trends and patterns on pre-existing car sales data, exploring the models which predict the true cost of a vehicle. Several machine learning models are implemented to best determine the retail price of a used car based on features such as make, model, year, and mileage.

2 RELATED WORKS

Price prediction models for used vehicles have been at the core of other studies that utilize various techniques of machine learning.

Venkatasubbu and Ganesh used Lasso regression, multiple regression, and regression tree to develop their models. Samruddhi proposed supervised machine learning through K Nearest Neighbor algorithm to determine the used car prices. Cross-validation was ultimately used to evaluate the model. Gegic et al made prediction models through an artificial neural network, support vector machine, and random forest in ensemble fashion. Heavy classification was done to categorize prices into broader classes (cheap, moderate, expensive) for separate model training thereby minimize price fluctuations. Sharma and Sharma utilized linear regression with one-hot encoding and attained a relatively high accuracy of 0.86. Pandit et al compared the methods of Lasso Regression, Ridge Regression, Bayesian Ridge Regression, and the output can be visualized through a web application.

Zhu, A.
Pre-Owned Car Price Prediction Using Machine Learning Techniques.
DOI: 10.5220/0012810100003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 356-360
ISBN: 978-989-758-705-4

3 METHODS AND MATERIALS

This study's data comes from a dataset which contains completed auto sales in the United States in recent years (Sharma and Sharma 2020). Figure 1 shows the overall process of approach and Table 1 lists the details of the cars from the data.

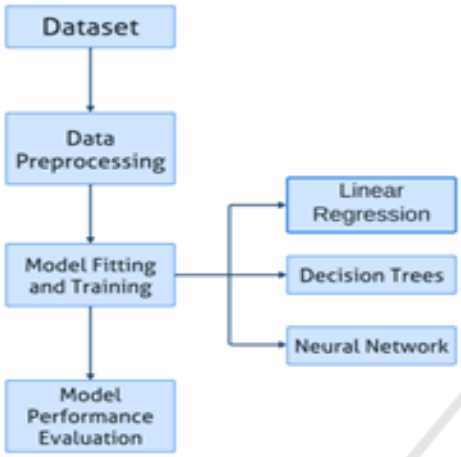

Figure 1: Process Diagram (Picture credit: Original).

A car has multiple features from the dataset.

Price: Price of the car
Levy: Tax or levy imposed on the car
Manufacturer: Brand or maker of the car
Model: Specific version or variant of car
Production Year: Year which car was produced
Category: Body style of car (Jeep refers to SUV)
Leather Interior: Leather seats and interior upholstery
Fuel Type: Gasoline/Hybrid/Diesel
Engine Volume: Capacity of car's engine
Mileage: Total number of miles driven

Cylinders: Number of combustion chambers
Gear Box Type: Transmission system type
Wheels: Drivetrain Configuration (4x4, FWD, RWD)
Color: Airbags: Number of airbags

When assessing a car's value, all its attributes come into play. Variables such as price, year, mileage, etc. are numerical features which can be readily applied in the model building. The remaining features fall into categorical and require conversions. One-hot encoding is used to transform the categorical data into a format that is machine-readable. Table 2 shows the structure of the data after one-hot encoding is applied. Table 3 contains the training and testing set sizes.

Table 1: Data overview of preprocessed data.

Price	Levy	Manu.	Model	Prod. Year	Category	Leather Interior	Fuel Type	Engine Volume	Mileage (km)	Cylinders	Gear Box Type	Wheels	Color	Airbags
13328	1399	LEXUS	RX 450	2010	Jeep	Yes	Hybrid	3.5	186005	6.0	Auto	4x4	Silver	12
16621	1018	CHEVY	Exqui.	2011	Jeep	No	Petrol	3	192000	6.0	Tipt.	4x4	Black	8
3607	862	FORD	Escape	2011	Hatchback	Yes	Hybrid	2.5	168966	4.0	Auto	4x4	White	8
11726	446	HONDA	FIT	2014	Hatchback	Yes	Hybrid	1.3	91901	4.0	Auto	Front	Silver	4
39493	891	HYUNDAI	Sante Fe	2016	Jeep	Yes	Diesel	2	160931	4.0	Auto	Front	White	4

Table 2: Sample data after preprocessing.

	Price	Year	Engine Volume	Mileage (km)	Fuel Type	Category	Gear Box Type	Wheels	Color
0	13328	2010	3.5	186005	1	1	1	0	1
1	16621	2011	3	192000	0	1	1	0	0
2	3607	2011	2.5	168966	1	0	0	1	0
3	11726	2014	1.3	91901	0	1	0	1	1
4	39493	2016	2	160931	1	0	1	0	1

Table 3: Partitioning of data into training sets.

	Ratio of dataset	Number of entries
Training Set	0.8	15389
Test Set	0.2	3847

The following are the techniques used in this study:

Linear Regression: Linear regression is a supervised learning technique that predicts a continuous variable based on a number of predictor variables. It offers simplicity in implementation and computation. Linear Regression can be applied to the entire dataset without additional parameter tuning. However, the simplicity means it does not capture more complex relationships between data as effective as other models.

Random Forest: Random Forest is an ensemble learning algorithm that combines the decisions from multiple decision trees to improve predictive performance. Each tree is built on a random subset of data, ensuring diversity to reduce variance of the model. The model captures both linear and complex nonlinear relationships, make it a great fit for predicting car sales prices based on various input features.

XGBoost: Extreme gradient boosting (XGBoost) is a powerful algorithm that excels in speed and predictive accuracy. Unlike Random Forest which builds trees in parallel, XGBoost builds trees sequentially, where new trees correct errors of previous ones. This model handles bad and missing data automatically, reducing the need for data imputation. Overall, gradient boosting is a robust method well-suited for predictive tasks.

Neural Network: Neural networks function like human brains. They consist of interconnected nodes (neurons) that process data in layers. They employ multiple hidden layers between input and output layers, allowing the model to learn precise patters from data. A neural network is also more acquainted to unstructured data, leveraging heavy computational power for its performance. While neural networks can capture complex relationships, they also require more time to tune and refine. This study's implementation is inspired by Tamoghno's neural network models (Tamoghno 2020).

4 MODEL IMPLEMENTATION

The models are based on supervised learning and make predictions about car prices using information from the training sets and subsequently are applied to the test dataset for evaluation. Figure 2 shows the results from the linear regression model.

5 LINEAR REGRESSION

The linear regression model is based on Animesh's implementation and ideas are derived from Ali's work.

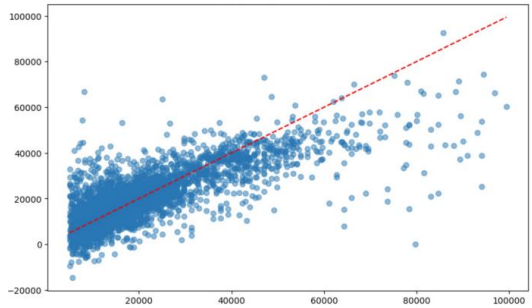

Figure 2: Predicted(y) vs Actual Value(x) of selling price (Linear Regression) (Picture credit: Original).

6 RANDOM FOREST REGRESSOR

Hyperparameter tuning is done using GridSearchCV to optimize the performance of the model. The hyperparameters used are in Table 4, and Figure 3 shows the results of the Random Forest Regression.

Table 4: Optimal hyperparameters for random forest.

n_estimators (number of trees in the forest)	30
max_depth (maximum depth of the trees)	2
min_samples_split (minimum number of samples required to split an internal node)	2
min_samples_leaf (minimum number of samples required at a leaf node)	100

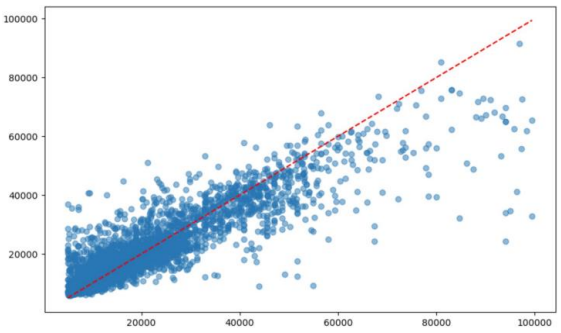

Figure 3: Predicted(y) vs Actual Value(x) of selling price (Random Forest) (Picture credit: Original).

7 XGBOOST

Gradient boosting, inspired by Mello's "XGBoost: Theory and practice", is applied to the dataset. Table V outlines the hyperparameters used for tuning the XGBoost model, and the results are shown in Figure 4.

Table 5: Parameters for XGBoost.

learning_rate (step size for each iteration in gradient boosting)	0.01, 0.05, 0.1
max_depth (maximum depth of individual decision tree)	4, 6, 8
colsample_bytree (fraction of features sampled for tree growth)	0.6, 0.7, 0.8
alpha (L1 regularization term for controlling complexity)	0, 0.5, 1
n_estimators (number of decision trees to include in the ensemble)	500
objective (function specifying the task)	reg:squarederror

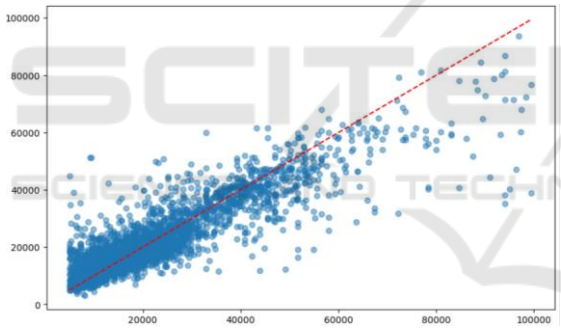

Figure 4: Predicted(y) vs Actual Value(x) of selling price (XGBoost) (Picture credit: Original).

8 NEURAL NETWORK

Table 6 shows the performance of the neural network model across various structures having different number of layers.

Table 6: Neural network evaluation on layer.

Summary	Layers		
	2	3	4
R^2-Value	0.66	0.64	0.65
MAPE	31.54%	27.80%	29.98%

Hidden layers are intermediate processing stages for the neural network to recognize and learn patterns, features, and relationships from the data. The results reveal that the neural network with three hidden layers has the lowest MAPE score, indicating the highest level of accuracy in predictions. The network with four hidden layers has a slightly larger MAPE score but performs well in terms of accuracy. The network with two layers has the largest MAPE score, indicating the lowest accuracy.

9 MODEL PERFORMANCE EVALUATION

Table 7: Overall Model Evaluation.

Model	R^2-Value	MAPE Score
Linear Regression	0.576	38.95%
Random Forest	0.765	25.34%
XGBoost	0.789	24.48%
Neural Network	0.65	28.98%

Table 7 lists the models' performances. Linear regression provided a baseline for this study's prediction assessment. The R^2 suggested the model provided a moderate level of predictive capability and in explaining the variance of the dataset. However, the MAPE score remained relatively high, indicating a significant deviation from the actual car prices in the training set. The Random Forest model showed a considerable improvement over linear regression and captured larger portions of variance. The decreased MAPE score implies better accuracy in predictions. The gradient boosting algorithm further improved the model's predictions, achieving the highest R^2 among the models implemented. The enhanced accuracy can be attributed to XGBoost's ability to capture complex non-linear relationships in the dataset. The neural network model did not perform as well as Random Forest or XGBoost, which suggests the model needs further optimization.

While linear regression provided a reasonable initial insight, it did not capture the intricate relationships influencing car prices. XGBoost balances accuracy and computational efficiency. From the training results, XGBoost appears as the preferred choice in car price prediction tasks.

The models evaluated in the study represented a progression from basic linear regression models to advanced learning techniques such as a neural network. The decision for the best model depends on the trade-offs between a model's accuracy, complexity, and efficiency.

10 CONCLUSION

In today's thriving automotive market where pre-owned car sales account for the majority of vehicles sold every year in the U.S, it is imperative that consumers are fed transparent, data-driven information about car prices. Price control and profit maximization are often at the best interests of car sellers. They list prices of cars according to their preferences, often overlooking the financial situations of buyers. While this practice has been a tradition among car retailers, it is important to ensure that buyers are well-informed about the true costs of a vehicle. Machine learning algorithms are able to achieve this task and predict the costs of used cars based on historical data. This allows buyers to make informed decisions on purchasing, and from a moral standpoint, promotes fairness in the car market industry. A transparent and universal car price evaluation algorithm would foster healthy competition among car retailers and listing platforms. This study tackles this issue by testing machine learning algorithms that predict the true cost of used vehicles. This analysis begins with exploring the linear regression model and delves into more sophisticated techniques involving random forest decision trees. Advanced techniques such as gradient boosting and neural network are implemented to learn intricate patterns more precisely from feature variables in the dataset. This study identifies the best machine learning models to use for evaluating used car prices. The objective of this research is to provide a method for price prediction that ensures fairness between consumers and sellers.

This study delves into machine learning models for predicting used car prices. The models proposed in this study can be adopted in similar price prediction studies such as determining the price of real estate properties, vintage collectibles, or electronic gadgets in the pre-owned market. These models can undergo heavy tuning on a larger dataset to provide more accurate predictions. Clustering algorithms like K-Means can also be considered, involving attribute-based clusters. This opens the option for price range analysis that assesses groups of cars across different price segments. For future applications, this study will explore larger datasets containing a broader range of sales data across any industry.

REFERENCES

Carlier, Mathilde, "U.S. New and Used Car Sales", (Statista 29 Aug. 2023)

P. Venkatasubbu and M. Ganesh, "Used Cars Price Prediction using Supervised Learning Techniques", IJEAT, 2019, pp. 216-223.

K. Samruddhi and Dr. R. Ashok Kumar, "Used Car Price Prediction using K-nearest Neighbor Based Model", IJIRASE, 2020, pp. 686-689.

Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric; "Car Price Prediction using Machine Learning Techniques", TEM, 2019, pp. 113-118.

Ashutosh Datt Sharma, Vibhor Sharma, "Used Car Price Prediction Using Linear Regression Model", IRJET, 2020, pp. 946-953.

Eesha Pandit, Hitanshu Parekh, Pritam Pashte, Aakash Natani, "Prediction of Used Car Prices using Machine Learning Techniques", IRJET, 2022, pp. 355-360.

Bhattacharya Tamoghno, "An introduction to neural networks with implementation from scratch using Python", Towardsdatascience July 1, 2020).

Agarwal Animesh, "Linear regression using python", (Towardsdatascience November 14, 2018).

Arthur Mello, "XGBoost: Theory and practice", (Medium August 17, 2020).

Amir Ali, "Linear regression with practical implementation", (Medium November 24, 2019)