Machine Learning for Used Car Price Prediction

Peter Yao, Xuan Bai, Tianyuan He, ZiQian Zhao

1. Introduction

The used car market is a significant sector of the US economy, with 40.9 million cars sold and a value of $196 billion in 2021 [2][3]. A wide set of features determine the price of a used car, including age, mileage, and model [7]. Several methods have been used in the past to predict the value of used cars based on these features, including machine learning tools like XGBoost, random forest, and lasso regression [3][4]. In this study, we evaluated the effectiveness of several such machine-learning models in predicting the value of used cars. Three regression models were selected and trained on a dataset of used car transactions on eBay. The accuracy and speed of the models were then tested, and one of them was selected to be used as the basis of a fourth ensembling model that we developed. The accuracy, speed, and feature importance of the four models were then analyzed.

2. Dataset and Cleaning

2.1 Dataset

Our dataset was sourced from Kaggle [1] and consists of approximately 160,000 sales records of US and Canadian used cars on eBay over a 20-month period from 2019 to 2020. The dataset contains 13 columns that list ID (a unique value assigned to each transaction), body type (sedan, SUV, etc), number of cylinders, drive type (RWD, 4WD, etc), price sold, year (the year the car was manufactured), zipcode, mileage, make, model, year sold, trim, and engine.

2.2 Data Cleaning

Not all features in our dataset were useful, and many entries contained null values or errors. Several features deemed largely irrelevant or which had too many errors and null values were dropped from the dataset. We further excluded rows with Canadian zip codes and removed the last 2 digits of each zip code, since many zip codes lacked the last 2 digits. In this way, we were able to obtain meaningful information about the location of each car sale. We considered dropping the number-of-cylinders feature, as many gas-powered cars were listed as having 0 cylinders in what were likely errors. However, as this would cause all-electric car data to be lost, cars with 0 cylinders were dropped from the dataset unless they were Teslas. Note that the number of non-Tesla electric vehicles in the dataset was very small.

Rows with outlier values (these often contained typos) were dropped, as well as any rows with null values. In the end, about 60,000 of the original ∼160,000 entries remained.

2.3 Data Visualization

After cleaning the data, we explored its structure with various plots and charts

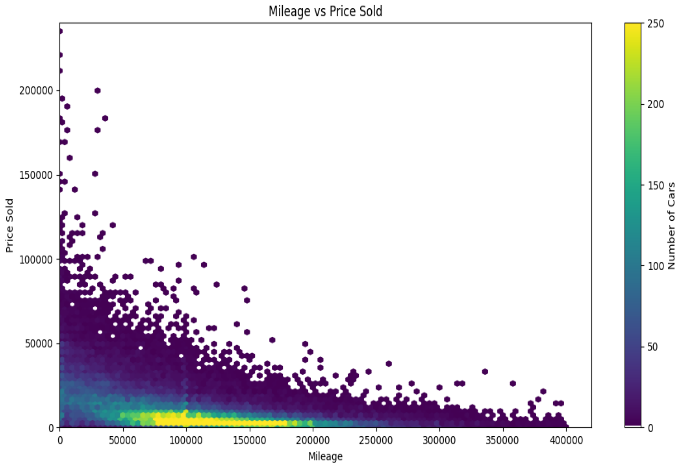


Figure 1: Price over mileage, colored by frequency

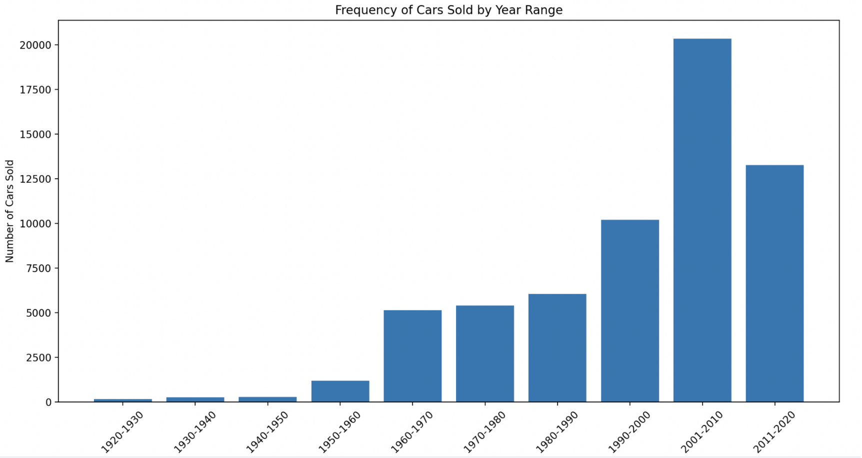


Figure 2: Frequency of cars in different age brackets

3. Models and Results

After cleaning the data, we encoded categorical features (make, model, drive type, and body type) as integer values. We did not use one-hot encoding due to the large number of categories in some of the features. We then split the dataset into a train and validation dataset, and trained three models from scikit-learn on the training dataset: HistGradientBoostingRegressor (HGBReg), RandomForestRegressor (RandForest), and LinearRegression (LinReg). We tested the three models on the validation dataset, recording the mean squared error (MSE), mean absolute error (MAE), time to fit each model, and time to run each model on the validation set. We repeated this process for three trials on each of the three models, thus training a total of nine models (3 of each type). In the last trial, we also evaluated the permutation importance of each feature for the HGBReg and RandForest models, as well as the coefficients for each feature for the LinReg model. We then selected HGBReg to be used in a fourth ensembled model, which was trained and tested in a 3-trial process in the same way the other models were evaluated.

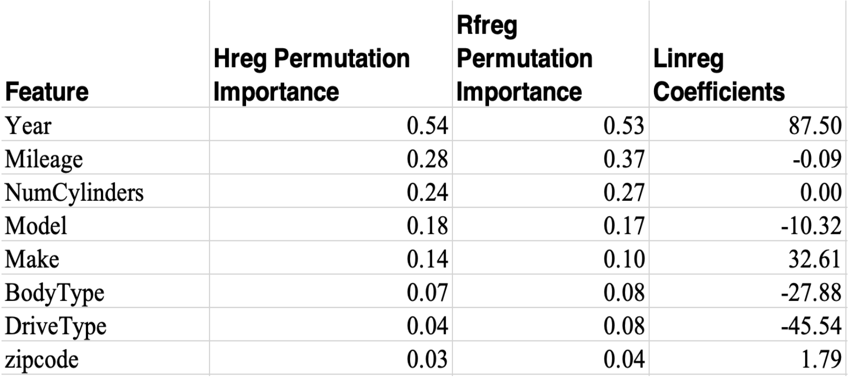
3.1 Model Initialization and Comments

As the training dataset did not change with each trial, we omitted the values for MSE and MAE in the LinearRegression model (as linear regression models train identically on the same data). To prevent RandForest from training identically, we varied its random_state input with each trial. A random state of 0 was used for trial 1, 42 was used for trial 2, and 101 was used for trial 3. We initialized HistGradientBoostingRegressor with a maximum number of trees of 1000, and we used squared error for the loss.

3.2 The Ensembled Model

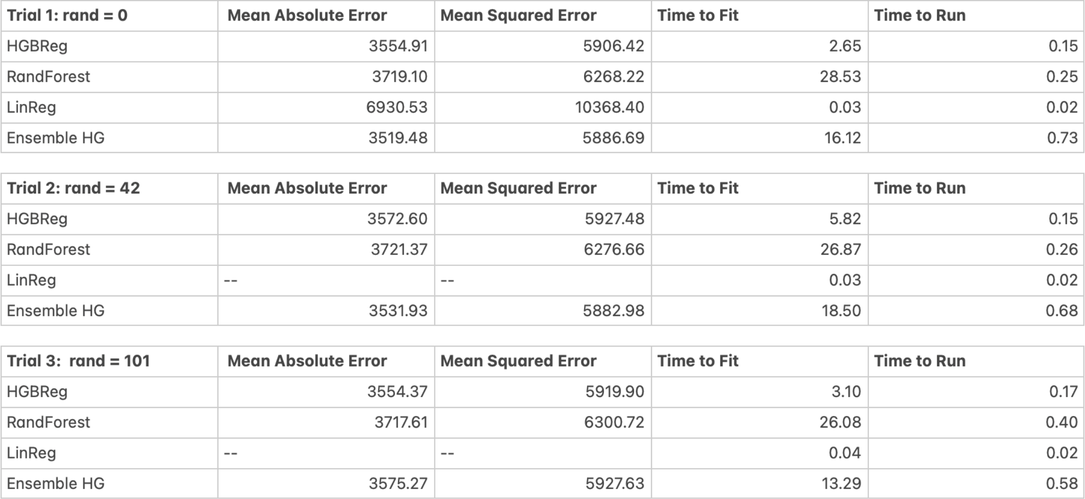
The ensembled model (Ensemble HG) consisted of five HGBReg models, which was selected because it had the best performance of the three initial models. The training dataset was split into 5 folds, and each regressor was trained on a different subset containing 4 of the 5 folds. The prediction value for input to the model was then calculated by averaging the predictions of the five HistGradientBoostingRegressors.

3.3 Results



Feature	Hreg Permutation Importance	Rfreg Permutation Importance	Linreg Coefficients
Year	0.54	0.53	87.50
Mileage	0.28	0.37	-0.09
NumCylinders	0.24	0.27	0.00
Model	0.18	0.17	-10.32
Make	0.14	0.10	32.61
BodyType	0.07	0.08	-27.88
DriveType	0.04	0.08	-45.54
zipcode	0.03	0.04	1.79

Figure 3: Feature importance for HGBReg (denoted Hreg) and RandForest (denoted Rfreg), and the coefficients for LinReg (denoted Linreg)



Trial 1: rand = 0	Mean Absolute Error	Mean Squared Error	Time to Fit	Time to Run
HGBReg	3554.91	5906.42	2.65	0.15
RandForest	3719.10	6268.22	28.53	0.25
LinReg	6930.53	10368.40	0.03	0.02
Ensemble HG	3519.48	5886.69	16.12	0.73

Trial 2: rand = 42	Mean Absolute Error	Mean Squared Error	Time to Fit	Time to Run
HGBReg	3572.60	5927.48	5.82	0.15
RandForest	3721.37	6276.66	26.87	0.26
LinReg	--	--	0.03	0.02
Ensemble HG	3531.93	5882.98	18.50	0.68

Trial 3: rand = 101	Mean Absolute Error	Mean Squared Error	Time to Fit	Time to Run
HGBReg	3554.37	5919.90	3.10	0.17
RandForest	3717.61	6300.72	26.08	0.40
LinReg	--	--	0.04	0.02
Ensemble HG	3575.27	5927.63	13.29	0.58

Figure 4: Results for MAE (dollars), MSE (root dollars squared), time to fit (seconds), and time to run (seconds). Note that LinReg MSE and MAE results are not given for trials 2 and 3 as they would be identical to trial 1

4. Discussion

HGBReg had the best performance out of the first three models tested. Notably, RandForest required a much larger training time than HGBReg or LinReg. The ensembled HGBreg model (Ensemble HG) had marginally better performance than HGBReg, though training time was longer. Runtime for all of the models in all of the trials was less than 1 second. For both decision tree models (HGBReg and RandForest), the feature with the highest permutation importance was the car's year. LinReg also showed a high correlation with car year.

Our results validated prior work which found that gradient-boosted tree models like XGBoost were the most effective at predicting used car prices [5]. Our dataset was however limited, as it was restricted to eBay sales in a 20-month span. In the future, other methods such as neural networks could be explored, as well as car sales data over a longer period.

5. References

1. Ts. (2020, November 16). *US used car sales data.* Kaggle. https://www.kaggle.com/datasets/tsaustin/us-used-car-sales-data

2. Moore, C. J. (2022, January 18). *U.S. used-vehicle sales record set in 2021.* Automotive News. https://www.autonews.com/retail/used-car-sales-set-us-record-2021-cox-automotive-say

3. *US used car market size & share analysis - industry research report - growth trends.* US Used Car Market Size & Share Analysis - Industry Research Report - Growth Trends. (n.d.). https://www.mordorintelligence.com/industry-reports/united-states-used-car-market

4. Used cars price prediction and valuation using data mining techniques. (2021). https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12220&context=theses

5. Jain, S. (2021, May 17). *Used car price prediction using supervised machine learning.* Medium. https://shubh17121996.medium.com/used-car-price-prediction-using-supervised-machine-learning-ea9dace76686

6. Hagerty, M. (2022, January 27). *Factors that can affect used car trade-in value.* Capital One Auto Navigator. https://www.capitalone.com/cars/learn/managing-your-money-wisely/factors-that-can-affect-used-car-tradein-value/1224

7. Gong, J., Peng, L., & Li, J. (2018a). A study on the factors affecting the value of used cars in Panzhihua region. *Proceedings of the 2nd International Forum on Management, Education and Information Technology Application (IFMEITA 2017).* https://doi.org/10.2991/ifmeita-17.2018.17