



Predicting Second-hand Car Prices

Parmis Mehdiyar (mehdiya2): parmis.mehdiyar@mail.utoronto.ca

Praniti Agarwal (agarw307): praniti.agarwal@mail.utoronto.ca

Pierre Sarrailh (sarrailh): pierre.sarrailh@mail.utoronto.ca

Motivation



- Research conducted by the Bureau of Economics, Bureau of Consumer Protection and the Federal Trade Commission talks about high opaqueness and complexity in the decision-making of a trade of a second-hand car, and the market in general.
- This happens due to multiple reasons: deals are highly customized (cars can have additional features), asymmetric information (dealers have more information regarding the car), sellers & buyers negotiate the price (advertised price is not the transaction price), etc.
- Consequently, buyers benefit from researching a price of a car, in order to lower the asymmetry of information which helps in negotiations and for sellers, as online selling grows, it is beneficial for them to advertise the right price.
- Hence, this project was motivated to solve this problem and build a comprehensive objective system that produces the right price based on different attributes and qualities after considering various machine learning methods.
- The vast amount of data found online along with the various number of features, makes it a machine learning problem.

Research Questions

How can a used car's various features be used to predict the car's resale price?

This question explores the fitting of machine learning models and tackles a prediction problem.

What are the most important determinants of a second-hand vehicle's price?

This question explores which of the input variables are the strongest predictors and looks towards examining a causal relationship.

Answer to the RQs

We built models that help predict used-car prices based on various input features (car model, mileage, year, state, quarterly GDP of state, number of past owners of the car, number of accidents, whether used for personal or business purposes). Consequently, users will have to just input their car and its features to attain an approximated price.

Based on existing research, we hypothesize that car mileage and age will be the most important features, since cars depreciate the fastest in the first few years in value.

Contribution

Sellers can determine when is the best time to sell their car for maximum benefit and which are the best models which can be resold when determining a new purchase. Further, they can obtain an approximate price for their car, based on which they can formulate their selling strategy.

Buyers not only get a reference price which can help them in the negotiation process but also learn the most important features to look out for in potential purchases.



Literature review

Price Determinants on Used Car Auction in Taiwan - (Meng, Shiang-Min, et al, 2019)

Examines the determinants of used car prices in Taiwan's auction market, and finds that the age and mileage of a used car are the most important determinants of its price, with newer and lower mileage cars commanding higher prices.

Like our paper, they assume car prices is a deterministic function of various factors, such as age, brand, model, and condition.

However, they use an OLS regression model which likely does not capture complex relationships between these features and car price.

Machine Learning for Used Cars Price Prediction: Moroccan Use Case (Benabbou et al, 2022)

This paper gathers used car prices from various online platforms in Morocco to find the XGBoost model to be most accurate with the highest accuracy of 90%.

Like our paper, they use different multiple regression methods.

However, their sample is highly limited to about 500 observations in Morocco and they do not account for classification techniques and macro-economic variables we have tried to account for.

Comparative Analysis of Used Car Price Evaluation Models (Chen et al, 2017)

This paper compares different Machine Learning Methods used for Car Price Evaluation to determine which is the best based on complexity of question and database.

Like our paper, they tried to test different methods to see which is the best method by changing the complexity of the dataset.

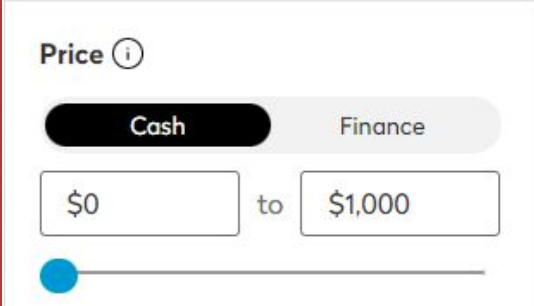
However, they only focused on the linear regression and forest methods, and did not account for KNN which can help in categorical variables important in cars.



Data and Web Scraping

Car data scraping methodology

- We used selenium to scrape the website truecar.com
- We scraped the car price, location, mileage, model, number of previous owners, car history, car brand,
- Methodology:
 - Open the website truecar.com using selenium
 - Set the search price parameters to 0-1000 \$
 - Scrape the first page of cars then recursively iterate through all the pages for that price
 - Wait 5 seconds every time we enter a new page to get around the anti-scraping algorithms on the website
 - Increase the price search parameter to 1001-2000
 - Recursively repeat the steps until we reach maximum price of \$100'000
- With this methodology we were only able to scrape 133k observations out of a potential 650'000 in the span of 28 hours
- As well as truecar.com we also downloaded the GDP of every state (in billions) for Q3 of 2022 for analysis in our regression

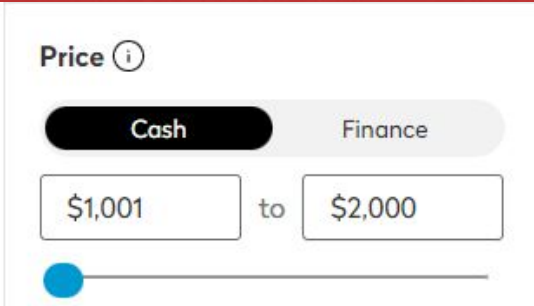


Price ⓘ

Cash Finance

\$0 to \$1,000

Price range slider with a blue dot at the start.



Price ⓘ

Cash Finance

\$1,001 to \$2,000

Price range slider with a blue dot at the start.

Descriptive statistics

114,456 Observations

Variable	Mean	Minimum (least)	Maximum (most)
Year	2009	1997	2022
Price	\$7,395	\$1,500	\$11,000
Mileage	144,724	208	472,030
State GDP in Q3 2022 (in billion)	\$975.603339	\$31.4	\$2,893.95
Number of Accidents	0.474152	0	7
Number of Owners	2.784618	0	14
Personal Use	0.783797	0	1



Methods Used

1. **Linear Regression:** simple and widely used model for predicting continuous values. It works by finding the linear relationship between the input features (such as the car's age, mileage etc) and the output variable (the price of the car).
2. **KNN:** able to handle a mix of categorical and numerical features, which is useful when modeling features such as the make and model of a car coupled with milage, etc.
3. **Decision Tree:** Highly interpretable and easy to identify the most important features that influence the price of a used vehicle.
4. **Random Forest:** As an ensemble learning algorithm that combines multiple trees to, RF may make more accurate predictions.

Hyperparameters optimized through grid search:

Algorithm	Hyperparameters
Decision Tree	max_depth=6, min_samples_leaf=2, min_samples_split=8
Random Forest	max_depth=10, min_samples_leaf=2, min_samples_split=8, n_estimators=100
KNN	n_neighbors=5

Results

Our analysis aimed to identify the most suitable machine learning model for predicting used car prices and the most important features for these predictions:

Model	MSE	Top 5 Features
5-NN	1663419.73	Not applicable
Random Forest	2346769.98	year, mileage, num_owners, model_Ford Focus, loc_CA
Decision Tree	2828896.23	year, mileage, model_Ford Focus, num_owners, loc_CA
Linear Regression	2.637E+14	model_Chevrolet C/K 3500, model_Dodge Ram 3500 Chassis Cab, gdp, year, personal_use

- KNN → 5 most similar observations. Very close to the real-world pricing techniques based on most similar competitors.
- California → highest GDP, perhaps systematically more expensive or certain regulations
- Similar features in DT and RF
- Poor performance of linear regression

Conclusion

- Overall, this research found 5-NN to be an effective tool for predicting price of used cars in the given range (1500-11000) using a very large database.
- Random forest and decision trees also confirmed our hypothesis about year of make, milage, number of past owners and potentially location of car being important features.
- While tested out of sample with separate test and train samples, the model may not perform very well out of the price range, as the interaction between features in different price ranges may be very different.
- Future research could address this issue with access to a larger database that covers all price range that we missed in the scraping
- Future research could also consider more unique features of each car, for example by reading and analyzing descriptions of the car using NLP tools.