ECO481: Final Paper

Predicting Second Hand Car Prices

Group 5

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

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

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

Introduction

Motivation

The global used car market is valued at $1.57T in 2021 and is expected to have an annual growth rate of 6.1% from 2022-30 (Milenkovic, 2022). Buying second-hand cars is very different from buying a new car from a car-dealership. And with online marketplaces bridging the gap between buyers and sellers, many countries see more used-car sales than new in recent years due to affordability. Research conducted by the Bureau of Economics, Bureau of Consumer Protection, & the Federal Trade Commission says that the second-hand car market is much more opaque and complex due to multiple reasons. Cars have multiple additional features, quality of the car (mileage, transmission type, engine, etc.) which leads to asymmetry of information with the dealers and sellers having more information than the buyers. Furthermore, sellers & buyers negotiate the price and the advertised price is thus not the same as the transaction price.

Consequently, for buyers, it is beneficial to bridge the gap created by this asymmetry of information by doing research into the existing second-hand car market, and the model and features they would like before going into the negotiation process. Simultaneously, for sellers, as online marketplaces grow and transparency increases, it would be beneficial to get the most activity on their products by advertising the right price. Hence, this project aims to build a comprehensive objective system by taking into account different features of cars to predict the approximate right price for any car based on a set of feature inputs. It can help reduce the chances of buyers buying overpriced cars, and for sellers to help determine their tactic into the market. Further, the study will help us even explore the most important features in the used car market that can provide vital information to sellers.

Research Question

The problem described above is a prediction problem and consequently, can be tackled by treating it as a machine learning problem. Consequently the question we pose to answer is: *"How can various car features (mileage, model, company, etc.) help to predict a used car's price?"*. This question will be answered by the best fitted & most accurate model on testing, and requires a large amount of data to properly fit numerous features. Subsequently, with the study, we hope to answer: "*What are the strongest predictors in the price of a second-hand car?*" This question explores more towards answering the casual relationship between the outcome and the inputs.

Answer to the Research Question

Upon the completion of this paper, we hope to build a model that helps to 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 approximate price. Currently, according to our understanding & past literature we believe that the random forest model will be the most accurate due to its ability to handle numerous features effectively, however, decision trees have the easiest interpretability. Further, based on existing research, we think that car mileage and age will be the most important features, since cars depreciate the fastest in the first few years in value.

Contribution

The model and the features help both sellers and buyers of the cars as explained above. Sellers can determine when is the best time for them to sell their car to get the most benefit and which are the best models which can be resold when determining a new purchase. Further, with increasing transparency, they can obtain an approximate price for their car, based on which they can formulate their selling strategy. Similarly, due to the opaqueness of the car market, 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

The "*Price Determinants on Used Car Auction in Taiwan*" by Deng et. al, published in 2018, builds a polynomial log-linear multiple regression model to determine the best price determinants of used cars in Taiwan. Like our paper, it assumes that used car price is a deterministic function of various features, and concludes that age (lower) and low mileage are the most important features. However, this study does not build a model to predict prices. Further, the study is restricted to only cars built in Taiwan and costing under 1M New Taiwan Dollars (CAD 44K).

Another similar study, "*Machine Learning for Used Car Price Prediction: Moroccan Case Study"* by Benabbou, Sael, & Herchy, published in 2022, builds a model using the XGBoost machine learning method for Morocco specifically. Like our paper, it compares different regression methods but does not account for classification methods along with macro-economic factors. Further, our study is focused on a completely different geographical area.

Lastly, the study "*Comparative analysis of used car price evaluation models*" by Chen, Hao, & Xu, in 2017, compares two methods in depth - linear regression and random forest using data scraped on websites in China to conclude that random forest is better to handle complex models with multiple variables when modeling something like the a universal car model, but with specific models for a certain model or car make, it has no obvious advantage. Our paper hopes to build on that by using KNN method to judge a universal model for all car types.

Data & Data Collection Methodology

Web scraping and data collection

To collect our data, we scraped the website truecar.com, an American website where consumers can resell used cars or new cars to anyone around the country and have them delivered. Users fill out information to present about their car, and for our data we scraped the car's price, mileage, model, brand, past history (like crashes & number of previous owners), location, year it was made, and model ID. To do this we iterated through all the different price combinations of the cars and slowly scraped cars at each price point. Due to technological constraints, as it took 28 hours of continuous running time to get this amount of data, we were not able to scrape more than 133,695 data points. Consequently, we were unable to obtain data points of cars costing less than or equal to $11,000 dollars. To select the different attributes to scrape from the website, we used a combination of class names and complete XPATHs. Further, we also downloaded the GDP of every American state from Statistica to use in our regression to incorporate a macro-economic variable in our model.

While scraping data from the website, several challenges arose. Firstly, the website timed out before loading all the cars on a page, causing the elements being scraped to load in after the pictures had been rendered. To overcome this, the page loadout time in selenium was set to 30 seconds using the line driver.set_page_load_timeout(30). Another challenge was the website only allowing access to the first 10,000 car entries. The work around for this was to incrementally increase the price of the cars and scrape all the cars within that range. To avoid missing any cars, a small interval in price ($1000) was picked.

Moreover, the website would label a user as a bot and stop responding to requests if pages were flipped through too quickly. This was overcome by making the instance of selenium wait 5 seconds every time it opened a new page. Furthermore, the element labels changed on every page of the website, which made it difficult to scrape. For example, the id for the price would not be a consistent like “price” but rather some randomly generated integer between 0 and 1000 for every car display card. To overcome this, all the <li> tags in the <ul> table were scrapped and then iterated through, the XPATH to that <li> was obtained, and the different values were scraped based on class labels. Additionally, the class label for mileage would contain multiple elements, and the value of each element would change depending on whether the car had the attribute “Upfront Price Available”. An if statement was used to catch whether or not that attribute was applied. Lastly, there was no way of knowing when the last page of a given price range was reached, and selenium would crash if the program tried to go to the next page when there wasn’t any. To overcome this, a try-catch statement was used, where the program would try to go to the next page and if selenium returned an error, the loop going through the pages was broken, and the price was increased.

Cleaning

For cleaning our data, we performed several steps. Firstly, we split the "condition" column into three columns and extracted the numeric portion of the accidents and num_owners columns while converting the personal_use column into a binary column. Secondly, we removed non-numeric characters from the "mileage" column using regular expressions and converted it to a numeric column. Thirdly, we created binary columns for each unique value of the "model" column using one-hot encoding. Finally, we extracted the state code from the "location" column, created binary columns for each unique state code, and dropped the original location and state columns. Next, we dropped observations that had “discounted” their price as we did not have the real value. Finally, we merged this data frame with the data frame containing GDP information on state level, thereby adding the state's 2022 Q3 GDP as a new column to the original dataset. These steps allowed us to clean and transform the dataset, making it more consistent and ready for further quantitative analysis.

Descriptive statistics

After cleaning, we had a total of 114,456 observations. The mean year of cars in the dataset is 2009, with the oldest car being from 1997 and the newest from 2022, showing a variety of car models. The average price of cars in the dataset is $7,395, with the least expensive being $1,500 and the most expensive being $11,000. The average mileage of cars is 144,724 miles, with the lowest mileage being 208 and the highest being 472,030 miles. The average state real GDP (from Q3, 2022) of the states in the dataset is about $975 billion, with a range of $31.4 billion to $2,893.95 billion. The average number of accidents for each car in the dataset is less than 1, with the fewest being 0 and the most being 7. The average number of owners is 2.78, with the lowest being 0 and the highest being 14. The majority of cars (78%) in the dataset are for personal use.

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)	$975.60	$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

Table 1: Descriptive statistics

Model Optimization

For optimizing our models, we first separated the "price" column from the dataframe and assigned it to a new variable. Then, we split the remaining data into a training set and a testing set using a 75-25 train-test split and shuffled the data. We optimized three models: decision tree regressor, random forest regressor, and K-nearest neighbors regressor. For each model, we used GridSearchCV to search over a parameter grid and find the best hyperparameters with 5-fold cross-validation. For the decision tree regressor, the parameter grid included "max_depth," "min_samples_split," and "min_samples_leaf." For the random forest regressor, the parameter grid included "n_estimators," "max_depth," "min_samples_split," and "min_samples_leaf." For the K-nearest neighbors regressor, the parameter grid was limited to "n_neighbors". We optimized each model by using the negative mean squared error as the scoring metric. After fitting the GridSearchCV objects to the training data, we printed the best hyperparameters and the corresponding mean squared error for each model.

The grid search gave us the optimal hyperparameter values. For the Decision Tree model, the optimal maximum depth of the tree is limited to 6 to avoid overfitting, and the minimum number of samples required to be at a leaf node is set to 2. The minimum number of samples required to split an internal node is set to 8. For the Random Forest model, the maximum depth of each tree in the forest is limited to 10 to avoid overfitting, and the minimum number of samples required to be at a leaf node is set to 2. The minimum number of samples required to split an internal node is set to 8, and the number of trees in the forest is set to 100. For the K-Nearest Neighbors model, the number of neighbors used to make predictions is set to 5.

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

Table 2: Hyperparameters selected

Results

Our analysis shows that KNN with 5 nearest neighbors had the best performance with a mean squared error (MSE) of 1,663,419.73. Random Forest had an MSE of 2,346,769.98, Decision Tree had an MSE of 2,828,896.23, and Linear Regression had an extremely high MSE of 2.64 x 10^14. The top features varied between models, with year and mileage being consistently important across all tree-based models, while the Linear Regression model showed car make being Chevrolet C/K 3500 as the most important feature. In the context of predicting used car prices, it appears that features related to the vehicle's age, mileage, ownership history, and location are key in determining its value. It's also worth noting that while the KNN model outperformed the other models, it did not have any features with notable importance, suggesting that its success is due to the combined effect of multiple features rather than any single feature.

Finally, the fact that the location feature "loc_CA" consistently had high importance across multiple models suggests that the location of a vehicle being California may be an important factor in predicting its price. It is possible that the price range for used cars in California is systematically different, perhaps due to regulations or demand for vehicles in California compared to other states, differences in the availability of certain types of vehicles in California, or differences in the costs associated with owning and operating a vehicle in California compared to other states. Further analysis would be necessary to determine the specific reasons behind this trend. Overall, our results suggest that tree-based models may be more suitable for predicting used car prices than linear models, and that KNN is a promising option for this type of prediction.

Model	MSE	Top 5 Features
KNN	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

Table 3: Summary of results

Conclusion

In conclusion, our research found 5NN to be the most effective tool for prediction prices for used cars, according to our large database that contained the price ranges from 1500-11,000$. This is not in line with our hypothesis and previous research, since, previously no research considered had an extensive database & considered multiple different methods like ours. However, in line with our hypothesis, random forest & decision trees also determined the year of make & mileage as the most important features with number of past owners & potential location coming close after.

However a glaring limitation is that due to the lack of information, after testing out of sample data, the model may not perform well for car models out of the price range. This could due to reasons like differing importance of features for more important cars, like look, color, brand, etc.

and consequently, further research could extend into including a larger range of cars and into the whole price range. This could easily be done with our algorithm but with better running power of computers. Lastly, we noticed that we did not include descriptions written by the seller into our analysis, and one of the further extensions could be to use Natural Language Processing tools to expand into reading and analyzing descriptions to determine the best words that help sell cars, which will help us build a more rounded model.

References

Milenkovic, D. (2022, May 16). *18 Undeniable Used Car Sales Statistics*. Carsurance. https://carsurance.net/insights/used-car-sales-statistics/

Used Car Market Size & Share Report, 2022-2030. (n.d.). Used Car Market Size & Share Report, 2022-2030. https://www.grandviewresearch.com/industry-analysis/used-car-market

Jin, C. (2021, November 22). Price Prediction of Used Cars Using Machine Learning. *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)*. https://doi.org/10.1109/icesit53460.2021.9696839

Bureau of Economics , Bureau of Consumer Protection, & Federal Trade Commission. (2020, July). *The Auto Buyer Study: Lessons from In-Depth Consumer Interviews and Related Research*. https://www.ftc.gov/system/files/documents/reports/auto-buyer-study-lessons-depth-consumer-interviews-related-research/bcpreportsautobuyerstudy.pdf

Meng, S.-M. ., Liu, L.-J. ., Kuritsyn, M. ., & Pechnikov, V. . (2018). Price Determinants on Used Car Auction in Taiwan. *International Journal of Asian Social Science*, *9*(1), 48–58. https://doi.org/10.18488/journal.1.2019.91.48.58

Benabbou, F., Sael, N., & Herchy, I. (n.d.). Machine Learning for Used Cars Price Prediction: Moroccan Use Case. In Proceedings of the 5th International Conference on Big Data and Internet of Things (pp. 332–346). Springer International Publishing. https://doi.org/10.1007/978-3-031-07969-6_25

Chen, C., Hao, L., & Xu, C. (2017). Comparative analysis of used car price evaluation models. *AIP Conference Proceedings*, *1839*(1). https://doi.org/10.1063/1.4982530

TrueCar. (n.d.). *TrueCar | New & Used Cars for Sale | Car Pricing & Reviews*. https://www.truecar.com/

Statista. (2023, March 29). *U.S. Real Gross Domestic Product 2022, by state*. https://www.statista.com/statistics/248053/us-real-gross-domestic-product-gdp-by-state/