Brandon Stanyer

**IST 707 – Applied Machine Learning - Final Project**

**Using Machine Learning Models to Predict the Price of Used Vehicles**

**Abstract**

The prices of used vehicles are consistently changing, and they vary greatly due to a multitude of factors. The complexity and variability surrounding used vehicle pricing make it a challenge for both buyers and sellers to determine a fair market value. Machine learning algorithms have the potential to transform the way buyers and sellers price vehicles. The used vehicle market has a need for models which can accurately price a vehicle. In this study, machine learning models were created using vehicle listing data, which contained the vehicle specifications and listing price. Analysis was done on decision tree, k-nearest neighbor, and support vector machine models on their ability to predict vehicle price. The resulting models show great potential in their ability to meet this goal, especially with larger datasets and adequate data preprocessing.

**Introduction**

Each year millions of used vehicles are sold in the United States alone. Reputable sources estimate that upwards of 30 million vehicles were sold just in the year 2022. The market for used vehicles presents opportunities for a multitude of businesses to facilitate and aid in these transactions. With such a large amount of business occurring around the used vehicle market, large quantities of data are generated and can be used to aid this process.

Other large markets such as the United States housing market see trends which correspond to economic conditions such as supply and demand of housing options, inflation, and interest rates for loans. Despite these changes, companies like Zillow and Apartments.com consistently find ways to establish trends in these markets and determine what prices are reasonable. The market for automobiles is similarly large and can be similarly analyzed. With such a large market for used vehicles, enough data is readily available to establish trends in pricing of vehicles.

There is consistent opportunity for people and businesses to make use of the data surrounding used vehicle transactions. Many businesses have already established themselves as key players in this market through their use of this data, such as Kelley Blue Book, cars.com, or eBay Motors. One of the key areas of focus by these companies is vehicle pricing. Vehicle dealers and private individuals need to know the approximate market value of the vehicle they are selling. Large inefficiencies occur in this business when used vehicles are not accurately priced. When sellers overestimate the price of their vehicles, resources are spent to maintain and store the vehicle until it sells. On the other hand, when vehicles are underpriced, sellers’ profits are reduced.

**Literature Review**

There have been numerous research efforts to find suitable models which can estimate a market value for used vehicles. Companies like Kelley Blue Book and Cars.com have developed programs which are used commonly in the United States. Both companies have websites where users can input their vehicle information and a program will return a range of prices that are considered fair market value. Since both of these companies have established their worth through these programs, not much is disclosed about their algorithms. On Kelly Blue Book’s website, all that is said about their program is that “Our Values are the results of massive amounts of data, including actual sales transactions and auction prices, which are then analyzed and adjusted to account for seasonality and market trends” (2023). It can be speculated that they use at least one, if not many machine learning algorithms to aid in their analysis.

In the past decade, a few researchers have published their efforts to utilize machine learning algorithms to predict vehicle pricing. Given that used vehicle markets can be drastically different depending on location, most of these analyses are focused on countries other than the United States. These research publications are aimed to fill a gap in understanding that has not yet been filled in their location.

One group of researchers aimed to use machine learning methods to predict the prices of vehicles in India (Dehuri et al. 2022). These researchers aimed to predict prices using vehicle attributes of Name, Location, Year, Kilometers Driven, Fuel Type, Transmission, Owner Type, Mileage, Engine, Power, Seats, and Price. With appropriate data cleansing and exploratory data analysis, they found vehicle data suitable to make reasonably accurate predictions on price. Using a dataset of 6019 vehicles and their attributes, they applied various machine learning algorithms and determined their effectiveness at predicting vehicle price.

The machine learning algorithms they utilized were k-nearest neighbor (KNN), random forest regression, decision tree, and light gradient boosting machine (LightGBM). Since the task of predicting price falls on a continuum, each of the models applied some form of regression analysis. Each of the models were able to predict a vehicle’s price to a varying degree of accuracy. The results of the KNN model were the worst of the methods tested, based on the model’s root mean squared error value (RMSE). Based on the RMSE values, each of the other methods were suitable for this task. The most effective methods were gradient boosting and random forest respectively.

Another group of researchers aimed to use similar methods to predict prices of vehicles in Bosnia and Herzegovina (Gegic et al. 2019). They focused on the machine learning methods of random forest, support vector machines (SVM) and artificial neural networks (ANN). Although the objectives were similar to previously mentioned study, some of the data preprocessing choices differed. Namely, the data was gathered using a web scraping program and brands were removed which had less than 10 vehicles found. Instead of removing the expensive vehicles from the dataset, Gegic and others chose to create three models for different price ranges: cheap, moderate, and expensive.

These researchers found that random forest, SVM, and ANN were not effective at predicting prices in isolation. The models were assessed as an ensemble and the resulting predictions showed that the price prediction improved when multiple models were used together. In the end, this work was able to produce a model with greater than 90% accuracy on the test data. However, it was emphasized that there was insufficient data to classify complex examples, and more data would be needed to improve the model.

There have been other recent research studies that have attempted to answer the same problems. In most cases, research studies have struggled with the complexity and variability found in used car prices. Accordingly, each research group has attempted to remedy this through data preprocessing and choice of algorithm. Each analysis differs slightly in their data cleaning and data preprocessing techniques. Outside of the previously mentioned studies, others such as Dutulescu et al., removed vehicles with model year older than 2000, mileage over 450,00km and horsepower over 600 (2023). In addition to the varied choices in data preprocessing, researchers often opted for more complex algorithms such as ensemble methods or neural networks. These types of models were chosen due to their efficacy with capturing the complex and multilayered relationships between vehicle attribute and price. Based on all of the aforementioned studies, it is clear that an effective model for predicting price needs to meet these criteria.

**About the Dataset**

For this analysis, a used car dataset was utilized from the website Kaggle.com. This dataset is a is a comprehensive collection of automotive information extracted from the popular automotive marketplace website, cars.com. This dataset contains 4,009 rows, each representing a unique vehicle listing with 12 distinct features about each vehicle. The link to access this dataset is found below.

<https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset>

As mentioned previously, this dataset contains information for 4,009 different vehicle listings. These listings were collected recently, with all listings from the year 2023. Accordingly, vehicle valuations should be nearly identical to the time of this analysis. There are 12 different attributes for each vehicle which are outlined below. Some attributes are purely numeric, string or a combination of both.

**Brand & Model:** Brand or company name along with the specific model of each vehicle.

**Model Year:** year is crucial for assessing depreciation and technology advancements.

**Mileage:** mileage is an indicator of wear and tear and potential maintenance requirements.

**Fuel Type:**  gasoline, diesel, electric, or hybrid.

**Engine Type:** engine specifications, shedding light on performance and efficiency.

**Transmission:** transmission type: automatic, manual, or another variant.

**Exterior & Interior Colors:** exterior and interior color options.

**Accident History:** prior history of accidents or damage, crucial for valuation.

**Clean Title:** availability of a clean title, which impacts the vehicle's value and legal status.

**Price:** Access the listed prices for each vehicle, aiding in price comparison and budgeting.

**Purpose of Analysis**

The purpose of this analysis is to create models which can predict the price of a vehicle based on vehicle information and features. Most used car sites employ similar models to provide estimates for pricing on vehicles and determine a fair market range. A model that can predict the price of a vehicle within a reasonable degree of error would be helpful to buyers and sellers of used vehicles. With an effective model, buyers and sellers of used vehicles can know that the vehicle they are buying or selling is priced fairly. A secondary purpose of this analysis is to determine the biggest factors affecting the price of a vehicle. Knowing which factors influence price most will be helpful for tuning future models, as well as for informing buyers and sellers.

**Data Preprocessing & Exploratory Data Analysis**

For this analysis, data preprocessing and model creation was done solely in R. After initial viewings of the data, a few areas were clearly in need of preprocessing before use in a machine learning model. Originally, all variables were given the wrong attribute type. In addition to this, many of the attributes were overly long strings or contained unnecessary information. The attributes of price and mileage originally had *$* and *mi* respectively for each entry. These strings were removed from each of the two columns so the data could be treated numerically.

When the data was initially read into R, all attributes were treated as strings. The attributes of *brand, model, engine, fuel type, transmission, exterior color, interior color, accident, clean title* were converted to the nominal attribute of factor. The *model year* attribute was converted to the ordinal variable of ordered factor. Lastly, the mileage and price attributes were converted to numeric attributes.

In addition to categorizing the attributes differently, a few of the attributes needed to be modified for better use in the models. The labels for *accident* and *clean title* were shortened for ease of interpretation. The attribute of *fuel type* also contained some errors, with NAs being represented by an empty string, a dash (-), or with the phrase “not supported”. Each of these three categories were grouped under the one label of NA. Aside from removing rows to better tune the model, no additional data preprocessing steps were taken. It should be noted that the *engine* attribute had very lengthy strings which included other information, such as horsepower and displacement. This led to over 1,000 different options for engine. The scope and scale of this analysis did not allow for the investigation or processing of this attribute. As such, it was not used in any of the models.

After these steps, a few key computations and visualizations highlighted areas of concern for the dataset. The first concern was the presence of 1898 different vehicle models within the dataset. With a dataset of 4,009 vehicles, this meant that many most vehicles had very little or no other vehicles of the same model to reference. Figure 1 shows just how few models there are of the same type. Most vehicle models are only represented 1-3 times in the dataset. It can be speculated that since vehicle model likely plays a significant role in pricing, the machine learning models could struggle to classify vehicles with so few references. With that mind, it would not be feasible to remove vehicles based on the number of models found in the dataset, since this is such an important attribute and too many vehicles would be lost.

A blue and black graph

Description automatically generated

*Figure 1. Count of each vehicle model in the dataset*

Following concerns about models not being represented enough in the dataset, attention was drawn toward brand representation. Figure 2 shows the count of each brand represented within the dataset. Brands were clearly represented much more within the dataset as compared to vehicle model. Many of the brands have 50 or more vehicles within the dataset. However, a significant amount had very few vehicles, such as Ferrari, FIAT, and Karma. Initial machine learning models attempted to use all brands of vehicles. Later analysis found that removing brands with less than 10 vehicles made significant improvements to model performance.

A graph with many different colored bars

Description automatically generated with medium confidence

Figure 2. Count of vehicle brands in the dataset

Based on other researchers’ findings, it seemed likely that the different fuel types may not be represented enough in the dataset. With some analysis, this speculation was found to be true. Figure 3 shows the count of vehicles with each of the different fuel types. The only fuel type with significant presence in the dataset was gasoline. As a result, vehicles with the other fuel types were removed for this analysis.

A graph of a number of fuel prices

Description automatically generated

Figure 3. Count of vehicle fuel types

The last attribute of the dataset which was chosen to investigate more thoroughly was *model year*. Since other research groups had chosen to restrict their models to vehicles within a certain year range, this prompted investigation. Figure 4 shows the distribution of vehicles by model year. The leftward skew shows a clear bias in the dataset toward newer vehicles. Instead of attempting to normalize this variable, a clear cutoff point was chosen for model year. This analysis focused only on vehicles with model year 2000 or later.

A graph of a number of years

Description automatically generated

Figure 4. Count of vehicles by model year

One final data preprocessing step was to remove vehicles over $75,000. It is commonly known that the luxury cars exceeding this price are much more scarce and variable in their pricing. Since this analysis is targeted toward commonly transacted vehicles, removal of these vehicles was of clear benefit. With each of these data preprocessing steps, the dataset was made more ready for use in the models. Some of the attributes which were originally unusable were now made usable. In addition, by narrowing the scope of the machine learning models to vehicles within certain brands, fuels, and model years, significant improvements in performance were achieved. As a final preprocessing step, the data was split into training and testing datasets, with 80% set aside for training and 20% for testing.

**Models & Results**

The objective of this research was to find a model well-suited to predict vehicle price based on a vehicle’s attributes. This study focused on three different machine learning models. These include decision trees, k-nearest neighbor (KNN), and support vector machines (SVM). It should be noted that each of these models were set up as regression models. Although data could be discretized and preprocessed for classification models, the objective of predicting price falls on a continuum. As such, regression models were best to predict the continuous variable of price.

**Decision Trees**

The first machine learning models used on this dataset were decision trees. This analysis used the rpart() function in R with the anova method. Decision trees were chosen initially for their simplicity and ease of interpretation. Early decision tree models showcased key parameters and attributes for use in later models. With careful experimentation, the attributes of *brand, model, mileage, model year, accident,* and *clean title* all led to improvements in model performance. The attributes of *fuel type, engine, transmission, exterior color,* and *interior color* decreased model performance when used as predictors.

The final decision tree model used the attributes listed above and the parameters of minimum splits = 2, maximum depth = 20, and a complexity parameter = 0. The model showed clear improvements over initial decision trees. The primary indicator used for evaluating model performance was root mean squared error (RMSE). For this application, RMSE can be interpreted as how far the predicted price was from the actual price on average. Although modifications to the model led to more than a 5,000 decrease in RMSE, the final model still had an RMSE value of 10,286. This shows that the model is making relatively accurate decisions for much of the data, but there are likely many outliers or difficult to classify vehicles.

To assess the decision tree model further, a visualization of how far each prediction was from the actual value was created. Figure 5 shows a scatter plot of the predicted prices to the actual price. There is a clear correlation between the predicted prices and the actual prices, which indicates some successful prediction. However, since each point represents a vehicle in the dataset, there are clearly many vehicles which have predicted prices vastly different than their actual price. This is seen predominantly with more expensive vehicles.

A graph showing a number of black dots

Description automatically generated

Figure 5. Scatter plot - Decision tree prediction vs actual price

**K-Nearest Neighbor**

The second type of machine learning model used was k-nearest neighbor (KNN). This was done using the knn.reg() function. This machine learning algorithm is well-suited for this analysis because it involves finding similar data points to classify new data. This is similar to how experts price vehicles based on previously sold vehicles of similar attributes. For the KNN models, the same vehicle attributes were used as in the decision tree models. Experimentation was done to see which value of k led to the best results.

No value of k generated a model with an RMSE comparable to the decision tree model. Analysis revealed two similar models with different biases. The first of these models used a k value of 3. Since the dataset has so few of each vehicle model, it seemed fitting to choose a small value for number of nearest neighbors to classify new data. This model resulted in an RMSE of 18,948. This model did much worse than the decision tree model. The weak correlation in Figure 6 shows how the majority of vehicles had predicted prices which were drastically different than their actual price. It should be noted that this model did appear to work best at predicting prices of lower priced vehicles.

A graph showing a number of black dots

Description automatically generated

Figure 6. Scatter plot – KNN model (k=3)

The second KNN model used a k value of 200. This improved the model and lowered the RMSE down to 15,968. Although this is a significant improvement, this is still much less effective than the decision tree model. The scatter plot in Figure 7 shows a clearer correlation between the predicted value and the actual price. Given the KNN algorithm’s efficacy toward classifying complex, multi-dimensional data, it was surprising to see such poor results.

A graph showing a number of black dots

Description automatically generated

Figure 7. Scatter Plot – KNN model (k=200)

**Support Vector Machines**

The last machine learning model used was support vector machines. SVM models tend to outperform other machine learning methods, but their lack of information about classification makes them difficult to interpret. This also makes SVM models difficult to use in business settings because the reasoning behind data classification is very complex. With this difficulty in interpreting results, tuning models becomes more difficult as well. Using the information gathered from previous models, some SVM models were constructed.

The SVM models used the same attributes as prior models. Experimentation with different kernel types revealed that the linear kernel was most effective. Since vehicle price decreases with mileage and age in a linear fashion, this kernel type performs best. The resulting SVM model achieved a RMSE of 7,316, making it the most effective model found. Although this is still too large to be considered highly effective, it is a clear step in that direction. As seen in Figure 7, there is a strong correlation between the predicted price and the actual price. There are also far less outliers, particularly for vehicles on the less expensive side.

A graph showing a number of black dots

Description automatically generated

Figure 8. Scatter Plot – SVM model with linear kernel

**Conclusion**

The task of accurately predicting the price of a vehicle is an incredibly complex task. It involves careful consideration of highly dimensional data, with each dimension containing a multitude of different options and variation. For these reasons, experts have long since been striving to atomate this process. Highly effective models have the potential to take into account more information and make more accurate predictions than human experts can. Effective price predicting models have the potential to transform the way vehicle transactions occur by reducing inefficiencies within these businesses.

This analysis found that regression techniques with decision trees, k-nearest neighbor, and support vector machines algorithms have the potential to meet this need. SVM models show the most promise in their ability to generate highly effective price predictions for used vehicles. However, their difficulty with interpretation and tuning present challenges for use in business applications. Decision tree and KNN models are better in this respect, but the accuracy of the models created with this dataset are still too low to be trustworthy. Overall, these models are a great step toward a price prediction model that would be regularly implemented in the used vehicle marketplace.

Although these models would not be considered highly effective on the dataset used here. They have potential to generate far better results given more data. This dataset is quite small for the objective of price prediction. Many of the outliers which troubled the machine learning models were vehicles which were rarely found in the data. It can be inferred that the reason that most of the models had lower RMSE values was due largely to the lack of comparable models. As mentioned previously, in this dataset of only 4,009, there were 1898 different vehicle models present. The models could be drastically improved just by increasing the amount of data.

Another consideration for future research is combing models together. These models could be used in partnership by averaging their predictions or by using ensemble models. Each of these three models offers some benefit toward the goal of price prediction. Due to the complexity of used car pricing, neural networks and artificial intelligence also have the potential to generate accurate price predictions.

The task of accurately pricing used vehicles is a problem of great scope and complexity. It involves thorough consideration of multitude of factors. As such, machine learning algorithms have great potential to solve this problem. This research is an initial exploration into what is possible with use of machine learning techniques. With the use of larger datasets and different machine learning models, it is very likely that highly effective models could be made to predict vehicle prices. Soon sellers and buyers of used vehicles could have an effective way of determining a fair market value for used vehicles using these techniques.

**References**

Bukvić, L., Pašagić Škrinjar, J., Fratrović, T., & Abramović, B. (2022). Price prediction and classification of used-vehicles using supervised machine learning. *Sustainability (Basel, Switzerland*), 14(24), 17034. https://doi.org/10.3390/su142417034

Dehuri, S., Prasad Mishra, B. S., Mallick, P. K., & Cho, S. (2022). Prediction of Used car prices using machine learning. *Biologically inspired techniques in many criteria decision making* (pp. 131-140). Springer. https://doi.org/10.1007/978-981-16-8739-6\_11

Dutulescu, A., Catruna, A., Ruseti, S., Iorga, D., Ghita, V., Neagu, L., & Dascalu, M. (2023). Car price quotes driven by data-comprehensive predictions grounded in deep learning techniques. *Electronics (Basel),* 12(14), 3083. https://doi.org/10.3390/electronics12143083

Gegic, E., Isakovic, B., Kečo, D., Mašetić, Z., & Kevrić, J. (2019). Car price prediction using machine learning techniques. *TEM Journal, 8(1),* 113-118. https://doi.org/10.18421/TEM81-16

Kelley Blue Book. “Instant Used Car Value & Trade-in Value | Kelley Blue Book.” *Kbb.Com*, www.kbb.com/whats-my-car-worth/. Accessed 11 Dec. 2023.