Brandon Stanyer

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

**Using Classification Models to Predict the Price of Used Vehicles**

**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 like the housing market. 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 used applied some form of regression analysis. Each of the models used were able to predict a vehicles 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.

Other pieces of recent research 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.