**Analysing the predictability of automobile prices using a Data Mining Approach**

**By:**

**Daniel Falcone**

**Truc Le**

**Hikma Abajorga**

**Peter Neal**

**Abstract**

Objective: Automobile cost prediction has particular importance in second-hand pricing decisions, for both vendors and buyers. Furthermore, it has potential applications in odometer fraud detection. Traditionally, assessing the proper cost of a car involves specific domain expertise that most individuals do not poses. With almost one automobile per person in the USA, auto cost prediction is of growing significance. This paper endeavours to enrich the limited research in scientific literature on car price prediction, using a data mining approach. So far, car make and engine size, type and power were identified as the strongest predictors. Unlike previously published papers, this report does not include age among its independent variables, analysing a dataset of 205 vehicles manufactured around the early 80s. Although the size of the dataset is on the shorter end, the report aims to achieve a main objective of providing a new perspective on automobile price prediction.

Methods and Material: The CRISP-DM industry standard framework is followed. The dataset is reduced to 201 cars due to missing samples. The price is divided into 3 categories: cheap, average and expensive. In order to properly evaluate the models, the dataset is split into a train set (150 samples) and a test set (51 samples), representing a 75-25% split. The employed algorithms are k-Nearest Neighbors (kNN), Multi-layer Perceptron, Logistic Regression, and Support Vector Machines (SVM,) achieving 72.549%, 68.627%, 80.392% and 86.275% accuracy respectively. Sensitivity analysis reveals the most important variables for each model. Feature importance aggregation weighted for accuracy between SVM and Logistic Regression reveals that a total of 14, out of the initial 25 variables, are relevant individual price predictors. The best, in descending order, are fuel system, make, curb weight, highway miles per gas (mpg) and wheel drive distribution. An improved SVM model (94.118% accuracy) is attained by removing the non-relevant columns. These findings are compared with previous research in order to inform managerial implications. Further recommendations for improving model performance are made, which include adding more vehicles and variables to the model.

Key Words: automobile price prediction, data mining, curb weight, Support Vector Machines, CRISP-DM, fuel system

**1. Introduction**

*1.1 Motivation*

Accurately predicting automobile prices is especially important when it comes to used vehicles. Correctly predicting car prices traditionally involves expert knowledge, as the price depends on many distinct factors, such as engine specifications, horse power, speed, acceleration, brand, safety, aspect, and many others. Buyers lacking such domain expertise, traditionally had to rely on mechanical engineers to ensure they are paying a fair price. Nowadays however, owing to machine learning technology advancements, the use of algorithms can be employed to determine an approximate correct price before making a purchase. Sellers can also leverage predictive analysis in order to decide how much to ask for their vehicles.

Most research in this field focuses on data whose most sensitive variable is mileage. However, as stated above, there are many important aspects to determining the price of an automobile. Significant insights can be gained through analysis focused on the specifications of older cars; as they were when they were brand-new.

The statistical methods employed by researchers include decision trees, k-Nearest neighbors and linear regression. This research report aims to build upon this research in order to verify the efficiency of linear regression and k-Nearest neighbors, by comparing them to a newer machine learning method, that has been gaining more and more ground in the previous years: the neural network. Specifically, in this report, a multi-layer perceptron (MLP) is used.

Furthermore, automobile price prediction could aid in odometer fraud detection. Also popularly referred to as “busting miles” in the US, in 2002, odometer fraud was estimated to have cost the economy around $1 billion dollar, with approximately 3.5% of vehicles’ having their odometer tampered with at some point during their lifetime [1]. Mechanics may be able to tell when the milage is not accurate, by checking the state of the engine, breaks, wheels and car body. While buyers can do their own investigation, coupled with research on the vehicle’s documents and history, odometer fraud can often be difficult to detect for an inexperienced person.

The price prediction model developed in this report can aid potential scam victims detect odometer fraud, by highlighting abnormally high prices for a given mileage. Previous research on odometer tampering detection focused on applying statistical tests from the fields of election and accounting fraud detection [2]. The main employed method is Benford’s law, which states that, the frequency of first significant digits generated by random data, should follow a rule of decreasing frequency, the number 1 appearing 30% of the time, while 9, only 5%. If they were to follow an even distribution, they should all appear as the leading significant digit about 11.1% of the time [3]. While search for statistical anomalies can represent an efficient method of digital fraud detection, knowledgeable scammers may use counter-measures to avoid being caught. In this context, an additional method is required to strengthen the toolsets at the disposal of fraud detection officers and researchers.

The number of motor vehicles registered in the US has been on the rise, reaching 273,6 million in 2018 [4], indicating the need for more car value estimation research. Automobile cost prediction has the potential to become an effective decision support system, for both automobile-sellers’ pricing decisions, as well as for customer support and advice.

*1.2 Literature Review*

Although fairly limited, research in this domain includes multiple machine learning techniques, used on a wide variety of datasets from countries all across the globe. The literature review section details the most recent and popular studies, in order to set the stage for the new analysis conducted in this report.

Pudaruth, 2014 collected over 400 records, but ended up only using 97 records, due to missing data: 47 Toyotas, 38 Nissans and 12 Hondas. The independent variables used were car-maker, cylinder volume (CC), year, and mileage. Using linear regression, a strong correlation between price and year of manufacture was observed, along with weak positive correlation between CC and worth. Decision trees and Naïve Bayes were the method which achieved the least loss, after normalising data and splitting the price into 6 categories [5].

Another study analysed a much larger dataset of 370,000 cars sold on the e-Bay German subsidiary website, Kleinanzeigen. Data was scraped and published on Kaggle. The dataset contained useful attributes such as fuel type, whether transmission is manual or automatic, power and whether the car had any repaired damages. The method employed is Random Forests (RF), an ensemble machine learning method build on top of decision trees, meant to trade off model interpretability for performance. Although usually used for classification, it can also be tweaked for regression. Data was split into training, testing and cross-validation sets, with a 70:20:10 ratio, obtaining an R2 score of 83.63% on the test data, representing the expected accuracy on new data that the model hasn’t seen before. The most important features were price, kilometer, brand and vehicle type [6].

Researchers from the University of Mauritius were particularly interested in predicting used car prices, due to the increasing prevalence of the second-hand automobile industry in the country. They gathered data from 200 cars and used linear regression, support vector regression (SVR) and artificial neural network (ANN) to predict automotive fees. Their variables were make, engine measured in cubic centimetres, paint type, transmission type and mileage. Cross-validation with ten folds was used and the average residual value was reasonably low for all 4 approaches, however some outliers were noted, suggesting that while machine learning price prediction is viable, it bears some risks, and may have to be used in conjunction with other techniques [7].

In his Master Thesis, Listiani, 2009, employed SVR to predict car prices, in the context of helping leasing companies determine optimal charges. The used dataset comprises 123,386 samples and 180 columns, including details such as number of previous owners, transaction types, types of car use, paint color, cushion and optional equipment. Independent multinomial variables were converted into dummy binary variables, meaning columns were divided into multiple ones containing a 0 or 1 value, to indicate whether the vehicle has the corresponding variable. Moreover, continuous values were normalised to be between 0 and 1, to prevent swamping by larger scale variables. Linear, Polynomial and radial basis function (RBF) kernels were tried, the last one obtaining the lowest root-mean-squared-error, (RMSE), 6.97. [8]

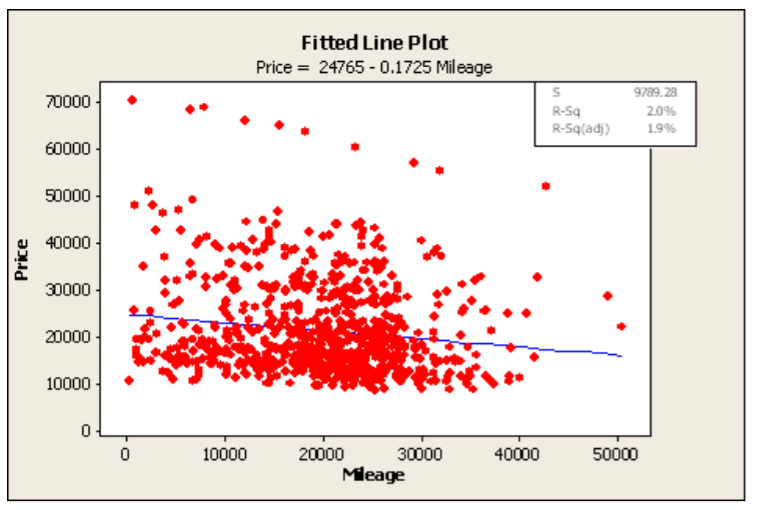
Kuiper, 2008, looked to demonstrate the benefits of multiple regression by using car prediction as an example. In this process, by looking at over 800 hundred General Motors cars, he showed that mileage alone is not a strong predictor of price. Although a clear correlation exists, R2 is just 2% (Figure 1). 

Figure . Scatterplot of retail price and mileage, with a linear regression fitted [9].

By integrating more variables, such as doors, cruise control and leather, while accounting for multicollinearity, a much better model, with R2 = 91.6%, is achieved. [9]

Some researchers have even attempted to build web applications on top of price prediction algorithms, in order to make it easily accessible to the wider public. Dholiya et. al, 2019 used a dataset of 1700-1750 automobiles to train a Multiple Linear Regression model in Python and Java. The model was leveraged to create a web interface that allows potential buyers and sellers to get a quick quote, compare cars and view the features that determined the cost prediction [10].

Muley, 2017, a postgraduate student as Oklahoma State University analysed the same dataset as [6], using instead Decision Trees and regularised Linear Regression, methods that make the model easier to be reversed engineered, facilitating the analysis of feature importance. The strongest correlation with price was that of age (-0.461), followed by horsepower (0.445) [11].

Car price prediction is of tremendous interest in China, where the rapid economic growth in the past decades had determined the national trade volume of used cars to grow by 30% per year. Scientists employed a back propagation (BP) neural network to optimize a model and establish feature importance, which achieved an accuracy higher than other models at the time [12].

Noor & Jan, 2017 utilised a dataset with features which included price, cubic capacity, exterior color, date when the ad was posted, number of ad views, power steering, mileage, rims type, type of transmission, engine type, city, registered city, model, version, make and model year. They found that only mode, year and engine type were relevant independent variables and achieved a model with 98% accuracy by using multiple linear regression [13].

Researchers from Bosnia and Herzegovina recently (2019) published a paper in which they analysed 797 samples and classified the price into 15 categories, in order to use classification algorithms to predict it. They compared support vector machine classifier (SVM), RF and ANN. Instead of choosing the best method, an ensemble of the three was created in order to achieve an overall of 87.38% on test data [14].

Previous studies in this field thus convey the importance of creating an accurate model for predicting car price, for its many case uses such as helping people looking to purchase a car compare the expected price between different vehicles and gain an idea of what to expect for a fair price. Multilinear regression, SVM, ANN, RF, SVR and decision all exhibit good potential. Furthermore, an ensemble from multiple methods may significantly improve the accuracy of the final prediction. The most important features for predicting a car’s price appear to be age, engine and model. Properties such as type of transmission and paint color have been shown to be mostly irrelevant. This study aims to fill some of the need for further research indicated in previous papers.

**2. Methodology**

*2.1 CRISP-DM*

In this report, CRISP-DM which stands for Cross-Industry Standard Process for Data Mining, is the research method employed, because of its wide application and acceptance in the industry. The approach was conceived in the late 1996 by DaimlerChrysler, SPSS and NCR. It consists of sets of tasks categorized in four levels of abstractions, from general to specific: phase, generic task, specialized task and process instance [15]. The life cycle of a data mining consists of 6 steps (Figure 2), which are followed consecutively and allow for cyclicity, meaning that steps can be reverted back to, whenever the business problem calls for it. Nonetheless, each step must be thoroughly completed before moving on the next one, in order to ensure a high-quality progression flows smoothly. CRISP-DM provides researchers with a solid framework to organize business data mining processes for maximum efficiency. Furthermore, it provides a clear structure that people from different companies and even different domains can all easily follow and understand. The following section explains how the CRISP-DM 6 steps and how the approach was employed in this analysis.

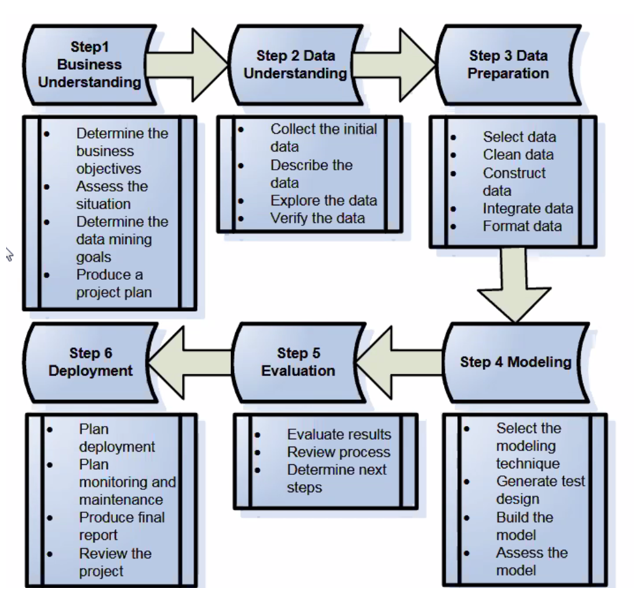


Figure . Graphical illustration of the CRISP-DM steps

Step 1. The first step is clearly understanding the business context in which the data mining is conducted. It consists of determining the business objectives, which in the case of this report, represents the desire to assist automobile buyers and sellers, as well as to assist in odometer fraud detection. Assessing the situation comprises the literature review section. The data mining goals are determined to be predicting automobile price. The project plan is produced, consisting of determining the report’s structure and how it will help achieve the specific business objectives.

Step 2. Next, data must be gathered and clearly understood. Initial data collection is conducted from the public data.world website [16]. Data will be thoroughly described in a following section, 2.2: each variable is explained, including their potential connection with the dependent variable. Then, once data migration is done, it is verified for accuracy and consistency.

Step 3. The third step embodies the data preparation process. It consists of selecting the data, cleaning it, which involves removing corrupt and inaccurate records from the set, constructing, integrating data, and finally formatting it, consisting of scaling continuous variables to have values within the same range, here 0 to 1, as well as turning all data types into a format that the model can understand. Moreover, the continuous dependent variable, price, is split into 3 classes (Figure 3), in order to improve accuracy and allow the use of classification models. The terminology is explained in a future section.

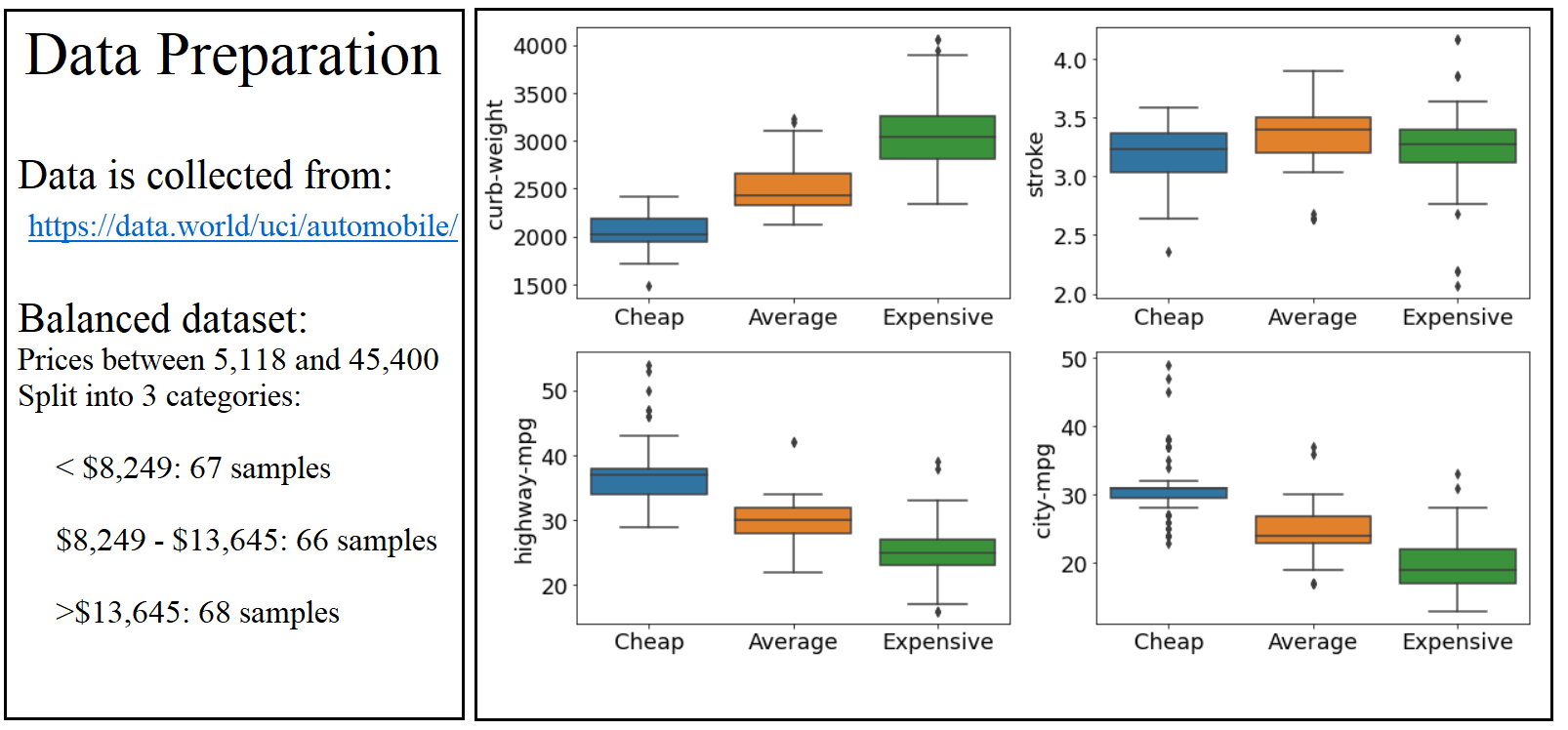


Figure . Data Preparation. Showing the 3 price categories against 4 features: curb weight, stroke, highway mpg, city mpg

Step 4. Modelling comes next and it includes first of all selecting the modelling technique(s). In this report, k-nearest neighbors, multilayer perceptron, logistic regression and support vector machines are chosen. The test design opted for is splitting data into training and test sets, so the model can be validated on samples it hasn’t seen before. The models are then built on the training dataset and assessed on the testing set.

Step 5. In this step the results from the three different models are evaluated on a multitude of scoring techniques: accuracy (the ratio of correct to total predictions), precision/specificity, a measure of how many predictions are made correctly (ratio of true negatives to all negatives), recall/sensitivity, measuring how many samples are predicted correctly (ratio of true positives to all positives) and f1 score, which combines the former two to provide an alternative to accuracy when classes are unevenly distributed. To illustrate the performances, a confusion matrix is drawn for each model. Sensitivity analysis is performed with the purpose of identifying the most important variables in predicting car price. The next steps are then determined.

Step 6. Deployment is the last step, which includes plan monitoring and maintenance, producing this final report and reviewing the project. In this paper, it mainly consist of describing the managerial implications, specifically how it can be used by clients wishing to know the worth of an automobile, either by itself, or as part of an ensemble of models.

*2.2 Data*

The dataset used in this report contains 205 cars and 26 attributes. As 4 rows lacked a value for the target variable (price), they were dropped, leaving n=201. It was originally created by Jeffrey C. Schlimmer in 1987 and was collected from data.world [16]. This section describes the variables of this dataset and how they may relate to car price.

The first attribute is symbolling, a measure from -3 to 3, representative of the insurance risk associated with its price. A higher score is suggestive of more risk. While it may not directly relate to the price, it could provide users with a measure of uncertainty.

A normalized factor of relative average loss payment per insured vehicle is included, however many values are missing, and for the purposes of this report, the column is dropped.

Make was shown to be a crucial variable in previous studies, which is naturally to be expected, as the price of a BMW is going to be higher than a Volkswagen with similar features. Besides corresponding to varying quality of materials, the brand name itself has a significant importance in determining the correct price for the vehicle.

Fuel type, either diesel or gas should have a small but clear impact. Diesel fuel is more expensive and of higher quality than gas. Diesel cars tend to be more fuel efficient in localities, but consume more on the runway. Gas cars are usually sold for cheaper than diesel.

Aspiration can be either turbo charger or natural. In standard aspirated engines, the air intake depends entirely on atmospheric pressure. On the other hand, turbo’s are “essentially forced induction devices powered by small turbines that increase an internal combustion engine’s efficiency and power output by forcing more air into the combustion chamber” [17], which leads to a boost. Since turbo is a precision made component and a more sophisticated technology, it costs significantly more.

The number of doors was previously shown to not be relevant in predicting car price. It can be either two, also called a coupe, or four, a sedan. While coupes are on average more expensive than sedans, when it comes to the same model, the former are usually slightly cheaper. Given the vast number of variables in this dataset, it may be expected that the number of doors should be slightly positively correlated with price, however given the relatively small number of samples (201), the model may not pick up on this subtility.

The body style was proven to be a good price predictor. The types are hardtop, wagon, sedan, hatchback and convertible. Convertibles are likely to be the most expensive body style.

Drive wheel represents the way the vehicle transmits force to the wheels, which determines the tractive force that tires exert on the road. Four-wheel drive (4WD) constitutes in the force being transmitted to all four wheels simultaneously. It is common in rally racing and off-road vehicles due to the better traction on slippery surfaces such as muddy terrain. Two-wheel drive is further categorized into front-wheel drive (FWD), which drives the force to the front wheels, allowing for easy steering control, and is what most modern automobiles operate on. As the name suggests, in rear-wheel drive (RWD), the force is driven to the rear wheels. It was particularly popular in the 19th century, before the fuel crises of the 1970s, when engines were usually mounted in the back to the car [18]. While it’s unclear whether front or rear-wheel automobiles are more expensive, four-wheeled drive cars are known to cost more than their two-wheeled counterparts.

In this dataset, the engine is located either at the front or at the rear. Rear-engines provide RWD and although they provide a lot of power and traction to the back wheels, making them accelerate quicker, they can be relatively difficult to control and prone to oversteer, unless proper suspension and chassis tuning is installed. Front-engines are by far the most popular. Most come with FWD, which makes the vehicle more stable and helps maintain a relatively balanced weight distribution when accelerating. Nevertheless, since FWD loses traction when accelerating, due to the car’s weight shifting to the rear wheels, it is sub-optimal for racing cars. Thus, RWD is sometimes used in front-engine vehicles to achieve a more even weight distribution and therefore better balance. Generally, rear-engines are common in sports cars and are more expensive [19].

The wheelbase (Figure 4) is the distance between the front and rear wheels, between 86.6 and 120.9 inches in this dataset. It is crucial in maintaining the weight distribution of the vehicle. Furthermore, a longer wheelbase is indicative of more interior space. It can be argued that the wheels being further apart “gives the space to create an elegant, flowing shape” [20]. Although longer wheelbase is correlated to car length, the relationship is not entirely linear, as the wheelbase to length ratio varies. Generally, cars with a larger ratio have better control and are more spacious. although exceptions such as front engine sports car exist [20]. There is concern of collinearity between this variable and other size related variables. Therefore, feature engineering that combines some columns such as wheelbase, weight, length, etc., has the potential to reduce multicollinearity and improve the performance of the model.

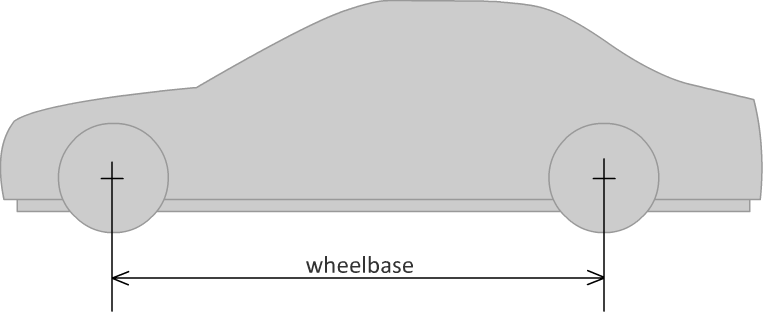


Figure . Drawing showing the wheelbase. From Wikimedia [21].

Length is a continuous variable of the distance from the rear to the front. It varies here from 141.1 inches to 208.1, therefore the data contains cars of varying dimensions, from minis, roughly 150 inches, to some cars in the executive and even luxury range, around 200 inches and higher. A strong positive with price is obviously expected.

Car width values range between 60.3 to 72.3 inches. This variable is highly correlated to length. Collinearity can lead to skewed results. Therefore, for achieving the best performance, this variable is dropped from the dataset.

Height varies between 47.8 and 59.8. It’s unclear whether this has enough relevance in predicting car price, especially in smaller datasets, where the model is probably unable to pick up the more nuanced aspects of an automobile. Expensive sports cars are shorter, however at other side of the height spectrum, some large off-road vehicles should be more costly too. Perhaps a convex function of price against height is to be expected, although mini-sized vehicles may decrease the average price of shorter cars.

Curb-weight, between 1488 and 4066, is the total weight of the vehicle as it is when not being used, meaning not including any passengers or cargo. It does however include all necessary operating consumables such as oil, brake fluid, coolant, antifreeze, etc. A convex relationship with price may be noted in this case too, as very heavy cars represent expensive, large off-road vehicles, while light ones are mostly race cars. However, as many light vehicles may be just minis, a more linear relationship could instead occur.

Engine types are DOHC, DOHCV, L, OHC, OHCF, OHCV and rotor (Figure 5). As the engine is perhaps the most important part of a car, its type is expected to be at least somewhat related to its price. Nonetheless, it is unclear which engine type is best, with even engineers not agreeing on any one in particular. Any variation of Overhead Cam (OHC), such as Double Overhead Cam (DOHC), or a variation of Over Head Valve (OHV), could be better depending on specific design and applications.

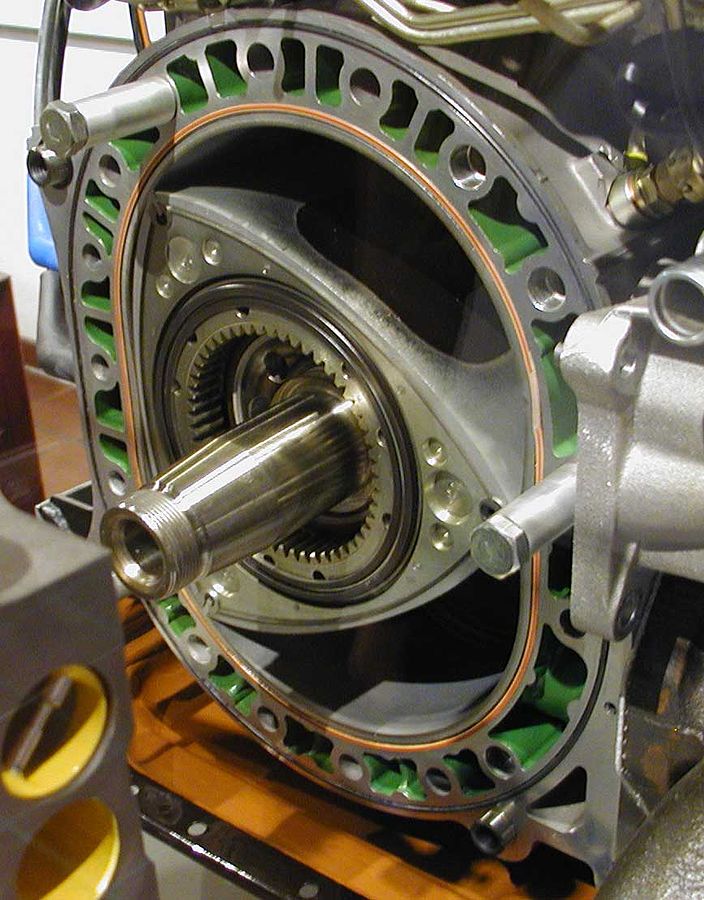


Figure . Mazda Wankel engine. A family of Wankel rotary engines. From Wikipedia [22]

The cylinders count can be 2, 3, 4, 5, 6, 8, or 12. The cylinder is the main working component in an engine. Each cylinder “has a piston inside which pumps oil into the crankshaft” [23]. A higher number of cylinders means more power, and therefore should mean a higher price.

Engine size is positively correlated to the space for air and fuel in it. Although a larger engine therefore usually means more power, that is not always the case due to the engine efficiency factors. In this dataset, the size is between 61 and 326.

The fuel system is responsible for pumping fuel to the engine. Newer and more sophisticated systems, such as sequential fuel injection (SPFI) may indicate a more expensive car than older ones, such as 1-barrel fuel injection (1bbl).

The bore represents the inner diameter of the engine cylinders. A wider bore usually means more power as it allows more fuel and air to be brought into the cylinder. However, the stroke must also be taken into account within an engine system to determine its power. The stroke (Figure 6) is the distance that the piston travels within the cylinder. A longer stroke determines more fuel consumption efficiency as it reduces surface area during combustion [24]. Four rows were missing a value for bore and stroke. Instead of dropping the entire rows and risk losing valuable data, the missing variables were assigned their respective mean values.

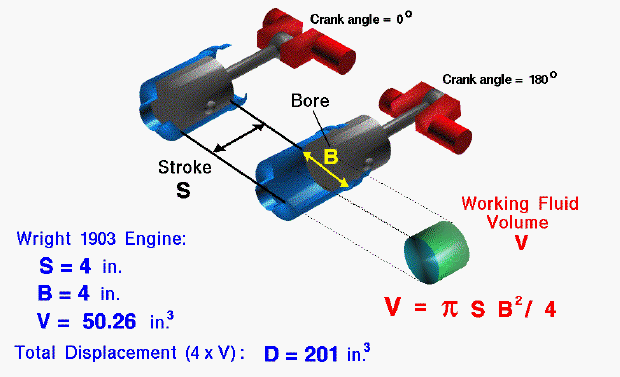


Figure .Technical drawing of an internal combustion engine. Shows bore and stroke. From NASA [25]

Compression ratio (CR) is the ratio between the cylinder volume and its head space “when the piston is at the bottom of its stroke to the volume of the head space when the piston is at the top of its travel“ [26]. CR affects the thermodynamic cycle in the engine and the geometry of the combustion chamber. Therefore, a higher compression ratio means better fuel economy [27]. Cars with better fuel economy may have a more expensive retail price, as the buyer would be saving money on oil.

Horsepower (hp) is a measure of the rate at which work is done. One horsepower equals the power required to lift a mass of 33,000 pounds one foot in one minute. It is equivalent to 746 watts [28]. Thus, horsepower is clearly and strongly related to the engines power. Most of the previous studies identified horsepower as one of the strongest price predictors.

Peak revolutions per minute (rpm) is a measure of how fast the engine’s crankshaft can spin and is indicative of the maximum power of an engine. Although it is correlated with horsepower, a larger car will generally have lower rpm than a smaller one with the same horsepower. A positive correlation with price is expected, however less strong than hp. Two rows had no hp and peak-rpm data and were therefore assigned the mean values of the respective variable.

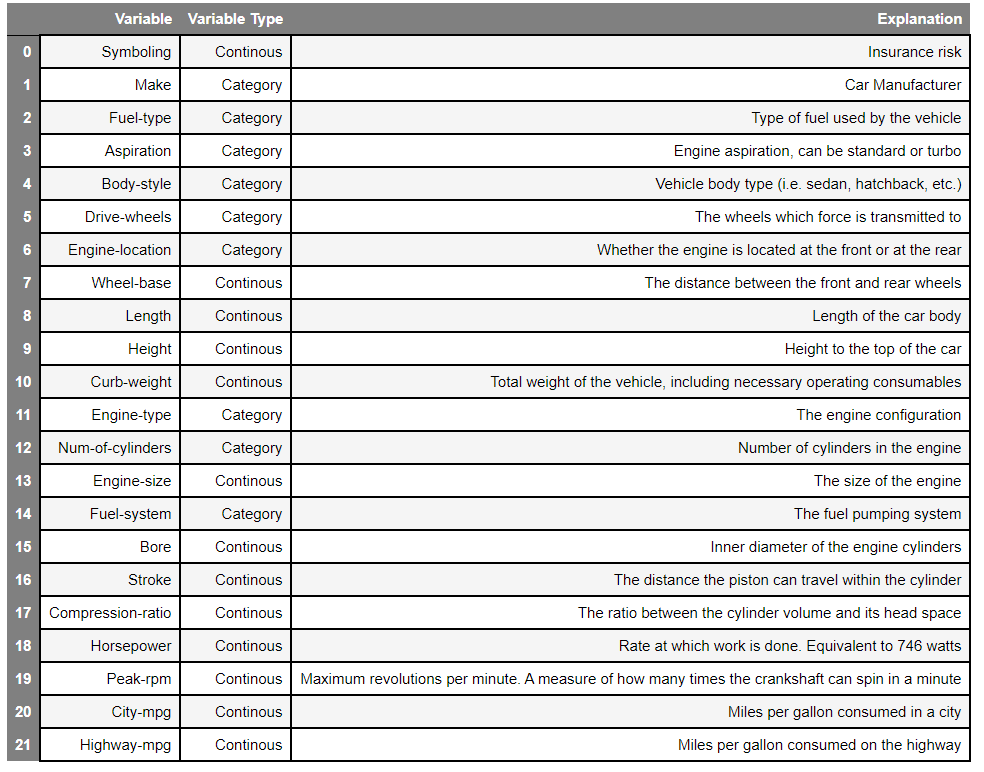
City miles per gallon (mpg) is a measure of how many miles a car travels, on average, using one gallon of fuel, in the city. As larger and more powerful cars generally consume more fuel, it should have a positive relationship with price.

Highway miles per gallon is a measure similar to city mpg, but for highway travel. These two variables are obviously strongly related. Nevertheless, as previously discussed, certain specification can determine various ratios between the fuel consumption measurements.

Finally, the dependent variable, price, ranges between 5,118 and 45,400 dollars. In order to employ classification models and improve accuracy, the price is categorised into 3 roughly equally sized groups: less than $8,249, between $8,249 and $13,645, and $13,645+. This split was performed using pandas qcut method. The categories have 67 samples, 66, and 68 respectively. Relative to one another, these categories will be called “cheap”, being the least expensive among the cars in this dataset, averagely-priced and the ones who are priced highest in this dataset are termed “expensive”.

Not all of the aforementioned variables were considered in the final models (Table 1). As the table suggests, normalized losses, number of doors, and width were dropped, as they are not expected to affect the price. Sensitivity analysis will confirm which of the variables are the strongest automobile cost predictor.

Table . The 22 independent variables used to predict car price. Shows variable name, type and explanation.



Data was processed and modelled using the Python programming language. The libraries utilised for data cleaning and analysis are: NumPy, which provides fast numerical computing tools, Pandas, a flexible data manipulation tool built on top of Numpy, and sklearn, short for scikit-learn, an open source data analysis tool which packages algorithms such as multi-layer perceptron neural networks, for easy and intuitive use. For plots, seaborn is chosen for its versatility and its compatibility with pandas and sklearn.

*2.3 Data transformation*

Before the data can be modelled (Figure 7), it has to be transformed in a format that a computer can understand. Using the SciKit Learn library with Python, continuous and categorical data was transformed accordingly, as described in the following subsections.

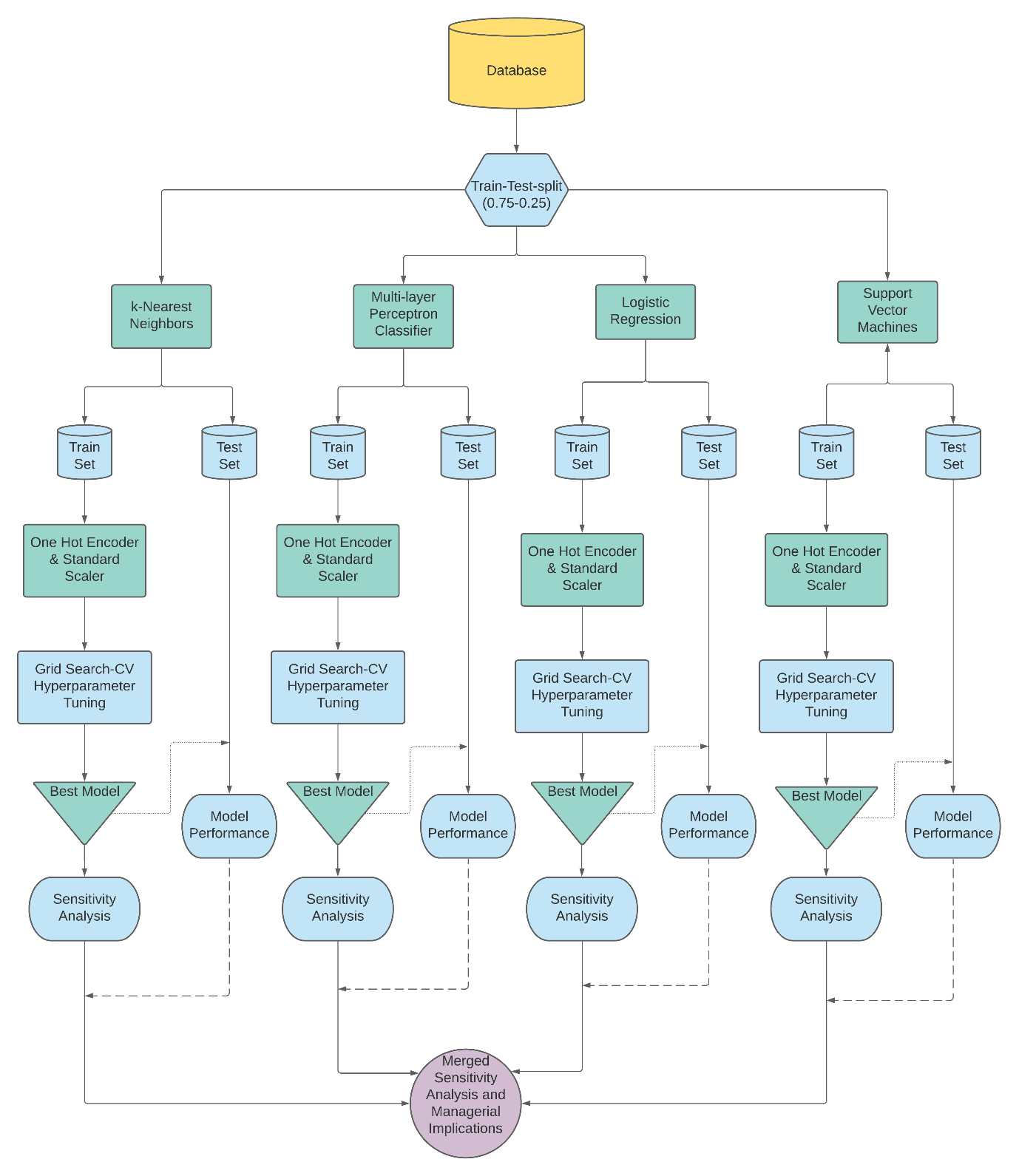


Figure 7. Graph showing the data mining process. Created with Lucid Chart [29].

*2.3.1 One-hot encoding*

Categorical data is in text format, which cannot be processed by Python. One hot encoding transforms a categorical variable into a series of columns, each representative of an individual value, containing binary values (0 or 1) which represent whether the car is of that class or not. For example, the drive wheels column is split into three: 4WD, FWD and RWD. For a car that has FWD, the 4WD and RWD columns are equal to 0, and the FWD column has a value of 1.

*2.3.2 MinMax Scaling*

Although Python recognises continuous variable, they must be scaled before being passed into the model. Scaling ensures values on a larger scale, such as peak rpm (4150-6600), do not unevenly affect the model, compared to variables with values on a smaller scale, such as bore (2.54 to 3.94). MinMax scaling projects all values to a range between 0 and 1, according to the formula:

(1)

where “z” is the scaled value, “x” is the original value, “x.min” is the minimum, and “x.max” is the maximum. MinMax scaling has the downside of compressing many values into a narrow range, therefore potentially increasing variance. However, the dataset has some outliers, which other scaling methods are very sensitive to.

*2.4 Train-test-split*

In order to test the performance of the models, data has to be randomly split between train and test sets. If the entire model was used for fitting the model, then it would do very well on the data points that it has seen, corresponding to low variance. However, it may incorporate random noise and lead to overfitting, which means it would not do well on new data. By keeping the train and test sets separate, an assessment of how accurate the model truly is can be made. The model is trained on the training set. Then, the model is used to predict price using the independent variables of the test set. The output is compared with the actual price, in order to assess the performance of the model. Out of the 201 automobiles, 75% (150) went into the train set, and 25% (51) into the test set. The split ratio is chosen because the training set is large enough for the model to pick up on the factors that determine the price, while the test set is sufficient to effectively asses the performance of the model.

*2.5 Data Mining Techniques Employed*

Four models are employed in this research: k-nearest neighbors (kNN), multilayer perceptron, logistic regression and support vector machines (SVM). Comparing the results of the four has the potential to inform intelligent managerial implications regarding the importance of the analysed variables. The models and their hyper-parameters are explained below.

Hyperparameters are the details of the models. For example, for kNN these can be the number of neighbours and the kernel function. Hyperparameter tuning is performed using Grid Search Cross-Validation (CV). Cross Validation is a method of splitting data into a user-provided number of folds, in this report, three. The folds are shuffled, and each iteration one of the folds is the test set, while the others are the train set. Many models are tested during this process. For instance, for SVM, a combination of 75 hyperparameter candidates are tested, each 3 times, totalling 225 fits.

Grid search implies that all the inputted hyperparameter combinations are tested. As this dataset is small, the computational power to directly run a grid search CV is easily affordable. If the dataset were larger, different methods of CV, such as random search, which randomly tests a given number of hyperparameters from the given lists, may have to be employed first.

*2.5.1 k-Nearest Neighbors (kNN)*

The kNN model for classification is a predictive analysis method in which a new data point is classified to be equal to the majority category between its k nearest neighbors, where k can be any whole number larger than 1 [30]. Hyperparameter tuning on the automobile dataset revealed that 15 neighbors gives the best result (Figure 8). Unlike other machine learning algorithms, kNN does not need time to “learn”, since it uses existing already existing distances. It is easy to implement and because its classification method is straightforward, the reasons why predictions are made is clear. On the other hand, on larger datasets, calculating the distance of new points can be costly. The same issues arises in datasets with a large number of dimensions. Another weakness of kNN is noise sensitivity: missing values and outliers can significantly reduce performance [31].

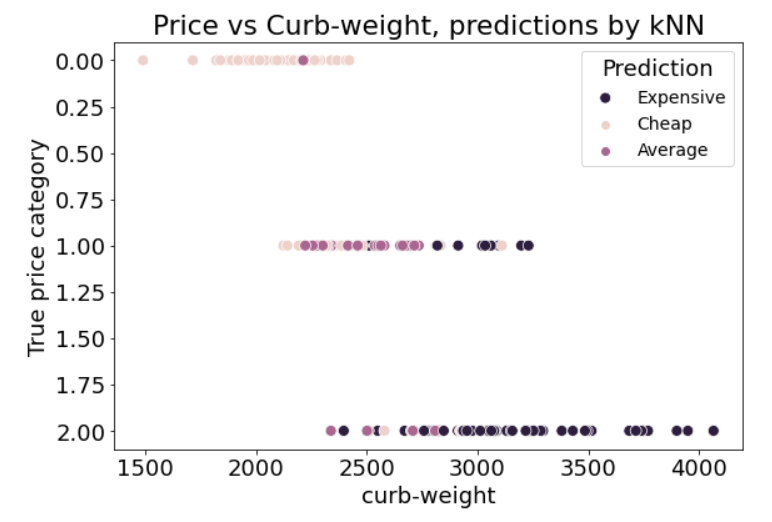


Figure 8. Showing the price category predictions given by the kNN algorithm, when k=15

Besides the number of neighbors, the most optimal leaf size and p value were searched for. In order to increase the time of finding neighbors, sklearn’s kNN provides 2 alternatives to a brute search: K-d tree and Ball-tree. They both use tree data structures to store data points in efficient ways which allow for improved time performance. In this model, the algorithm is allowed to automatically choose the most optimum neighbors finding mechanism. The leaf size hyperparameter controls how many data points the sub-trees in these algorithms can have. The most optimal number found is 20.

In order to calculate the distance between points, the algorithm uses “minkowsi” distance. This can be either Euclidean distance, when p=2, or Manhattan distance, when p=1. Grid search revealed that p=1 gives the best accuracy [32]. The Manhattan distance is measured along the axes at right angles, such that in a plane containing 2 points with coordinates (x1, y1) and (x2, y2), it is given by:

*2.5.2 Multilayer Perceptron Classifier*

A perceptron/neuron (Figure 9) is an algorithm for binary classification. In a multi-layer perceptron model, also known as an artificial neural network (ANN), multiple perceptrons act as a simplified version of the human brain neurons.

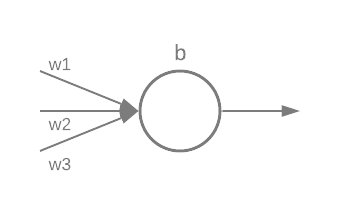


Figure 9. A perceptron taking input from three neurons in the previous layer. Shows the weights and the bias.

A perceptron has a threshold which has to be reached in order for it to activate, which can alternatively be written as a bias. The simplest perceptron activation function is a step function

(3)

,where Σi is the sum of all the products of weight (wi) and previous neurons (xi), plus this perceptron’s bias (b). The activation function used in this model is the rectified linear unit:

,chosen for its simplicity and computational speed.

The output of neurons feeds forward into neurons in the next layer, until the output layer is reached. An ANN learns by tweaking its weights and biases. Each time a series of predictions are made, they are compared with the true values, and a log-loss score is calculated, representing how wrong the predictions are. The purpose of the ANN is to minimise the loss/cost, which is achieved using stochastic gradient descent. In order to find the minimum of the cost function, one approach could be computing derivatives. However, given the many thousands of variables that the multi-layer perceptron has to tweak, this approach would be extremely inefficient. Therefore, an alternative method is required. Gradient descent works by repeatedly computing the gradient of the cost function, and moving in the opposite direction, “falling down” the slope of the valley [33].

The multi-layer perceptron classifier employed (Figure 10) has 69 inputs, corresponding to the number of columns after column transformation, an output layer consisting on one neuron, the price category, and two hidden layers. The hidden layers allow the network to learn specific aspects of the dataset which help predict the price. Their sizes are 256 and 50 neurons. The L2 penalty is set to 0.0001. Its purpose is to prevent over-fitting.

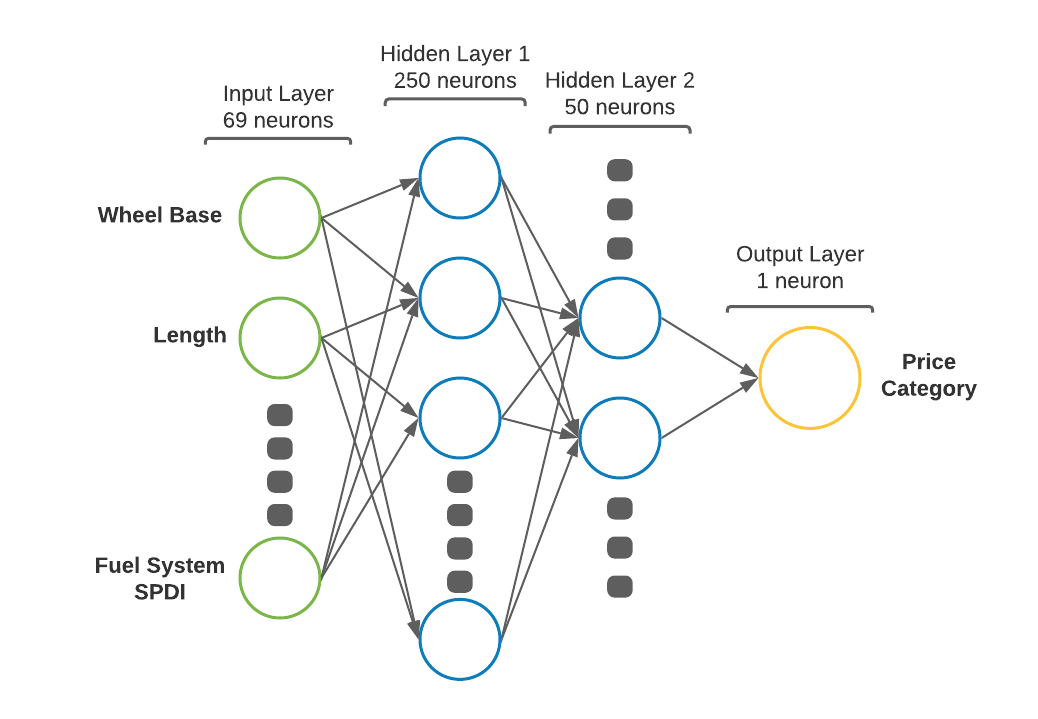


Figure 10. Diagram of the Multilayer Perceptron Classifier. Shows the connections between perceptrons, the 4 layers and their neurons count.

*2.5.3 Logistic Regression*

Logistic regression is one of the simplest classification algorithms and has many applications in the scientific world, although it is most popular in domains that require straightforward models which can be easily reverse engineered, such as medicine. In its basic form, it uses a logistic function to model a binary variable. Thus, in order to predict 3 or more categories, an extension of the logistic regression must be used. During hyperparameter tuning, multinomial regression performed the best on the dataset at hand. The algorithm works by performing a maximum likelihood estimation. Although it does not assume normality, linearity or homoscedasticity (constant noise), it does assume independence among the dependent variable and non-perfect separation [34]. The algorithm used in the optimisation problem is Limited memory Broyden-Fletcher-GoldfarbShanno (L-BFGS). It works by approximating the second derivative matrix using gradient evaluation and its main benefit is minimising memory usage by only storing a limited number of updates [35]. L2 penalty, also called ridge regularisation is added. It adds a coefficient to the loss function in order to reduce the model’s complexity and prevent overfitting.

*2.5.4 Support Vector Machines (SVM)*

Support vector machines work by plotting a hyperplane to split data into different categories. Given any N-dimensional data, the hyperplanes will be N-1 in size. SVM plots the hyperplane that has the maximum margin, meaning maximum distance between data from both classes. Points are classified according to which side of the hyper-plane they fall on [36]. SVM can give good scores for complex data and is not prone to overfitting. However, training times can be long and for multi-dimensional data it’s not always easy to see how the model chooses predictions.

To use SVM for multiple classification, the SVM is kernelized in order to increase the number of dimensions of the input data space. For the automobile dataset, 3 kernel functions were tried:

Linear Function:

Polynomial Function:

Radial Basis Function (RBF):

, where the input space X consists of xi and xj, and k(xi, xj) is the kernel function, and γ is a free parameter for the RBF [37].

The kernel function takes two datapoints and calculates a distance score, which is used to map the data points into a higher dimension. The RBF is found to be the best kernel function for the automobile dataset. The regularization parameter C is 10. The regularization strength is inversely proportional to C. Regularization is given by l2, a penalty regarding how many data points are falsely classified. A degree parameter was passed, which controls the degree of polynomial kernel functions. However, since the best function is RBF, this hyperparameter is not used by the model.

**3.Results & Discussion**

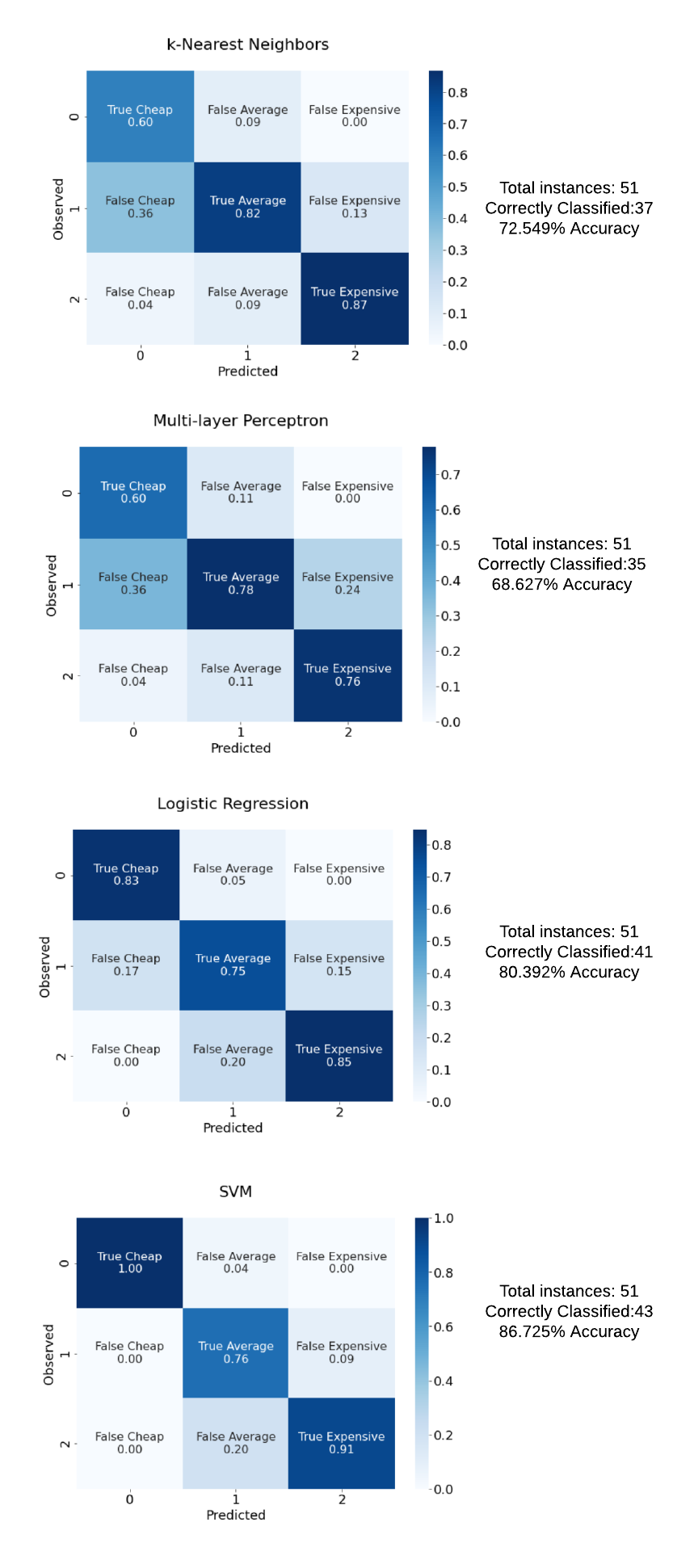
This section comprises the first parts of the CRISP-DM fifth step (Figure 2): evaluating the results and reviewing the process. The results of all the three models (Table 2) are compared, then sensitivity analysis is performed, in order to see which variables are the best price predictors. The classes of the dependent variable are relatively evenly split, the testing set has 16 cheap vehicles, 20 averagely priced, and 15 expensive. Therefore, the performance metric opted for is accuracy, which is the ratio of correct to total predictions.

*3.1 Algorithms performance*

*3.1.1 k-Nearest Neighbors*

The kNN model classifies 37 out of 51 cars correctly, achieving 72.55% accuracy. It has poor precision on cheap automobiles, 60%. It largely overestimates their count, 25 instead of 16. It correctly predicts 15 of them, thus achieving 93.75% recall on the low-price category. However, out of the 25 predictions, 9 are wrongfully made as average cost, and one as expensive. Only 11 cars are predicted to have medium price. Out of these, only 9 are estimated correctly, therefore the sensitivity being only 45% (9/20). The other 15 are classified as expensive. Precision is 87%, as 13 cars are correctly classified as expensive, and 2 incorrectly as averagely priced. As there are 15 expensive automobiles in the dataset, the recall is also 87%. The model therefore does well on identifying cheap and expensive cars, however it overestimates the number of cheap auto-vehicles, and achieves poor recall on averagely priced cars.

Table . Data Mining Results Visualisation: Confusion matrices. Shows the predictions for all four models, and precision/specificity for each price class.



*3.1.2 Multi-layer perceptron*

Although the most complex model, the ANN performed the worst (68.627%), only 35 out of 51 instances being identified correctly. The model overestimates the number of cheap vehicles: 25, instead of 16. Only 15 of the cheap class predictions are accurate (60%). Although it corresponds to a high recall, 93.75%, 9 cheap predictions incorrectly classify average price, and 1, expensive. Only 9 cars are predicted to be averagely priced, 7 of which are correct predictions (78%). Therefore, only 7/20 average price instances are identified correctly (35%). The remaining 17 cars are classified as expensive, of which 13 correctly, and 4 wrongly as average instances. Thus, 13 of 15 expensive cars are recognized (87%). The model is therefore performing well on the two extremes, but is vastly underestimating the amount of averagely priced vehicles.

*3.1.3 Logistic Regression*

The logistic regression algorithm correctly predicts 41 out of the 51 automobile price categories. From 18 cheap prediction, 15 are correct, and 3 are incorrectly classified as average (83%). The recall on cheap vehicles is therefore is 15/16 (93.75%). Exactly 20 average predictions are made, however 15 are right, meaning the recall on the average class and the accuracy of average predictions are both 75%. One vehicle was incorrectly categorised as cheap, and 4 as expensive. The remaining 13 instances are predicted expensive, 11 of which are correct (85% accuracy). The remaining 2 (15%) are classified as average. Thus, the precision on expensive vehicles is 13/15 (85%). Similar with the ANN model, logistic regression performs better for cheap and expensive cars. However, it achieves a decent accuracy for the average price category too: 75%, compared to just 35%

*3.1.4 Support Vector Machines*

The SVM model achieved the best accuracy, correctly predicting 43 instances (86.275%). It achieves 100% precision on the cheap category (15/15), however one cheap automobile is missed (93.75% sensitivity). It overestimates the number of averagely priced cars, making 25 predictions for this class, of which 19 are correct (76% precision). Since there are 20 average cost vehicles, the recall is 95%. Only 11 predictions are made for the expensive category, 10 of them being correct. While this coincides with 91% precision, the recall is only 10/15 (66.6%). Unlike the previous models, which underperform on the average class in terms of sensitivity, the SVM achieves the best recall on the class (95%), however underperforms on expensive vehicles.

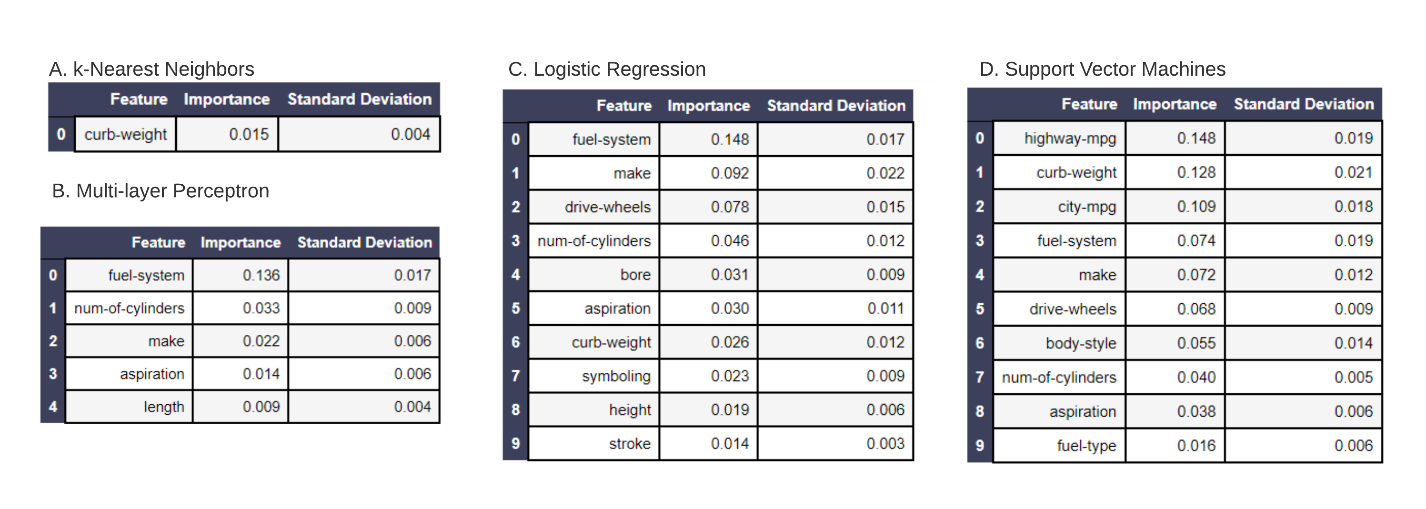
*3.2 Sensitivity Analysis*

Sensitivity analysis, or feature importance, shows which individual variables have the strongest effect on the models. Even though there are similarities between the 3 algorithms, each one has picked on some different features.

The only feature used by kNN is curb weight (Table 3A). Lighter vehicles in this dataset must predominantly be minis and generally smaller, less expensive cars, while heavier ones are mostly executive automobiles. The large importance of curb-weight is surprising, as most previous studies identified make or engine size as better predictors. Because kNN achieves very poor specificity on averagely priced automobiles (45%), it is not utilised in the aggregated sensitivity analysis.

The most important variable in the multi-layer perceptron classifier is the fuel system (Table 3B). As discussed in the variables analysis, more sophisticated fuel systems may be indicative of a newer, more technological advanced car. Number of cylinders, make, aspiration and length are the other features considered by the algorithm. The variables picked by the ANN do not allow it to identify averagely priced cars significantly better than a random guess (35% vs 33.3%). Therefore, the algorithm’s feature importance will not be taken into account for the merged sensitivity analysis.

Table . Feature importance of the four models. Showing only the features with importance >0. Tables show the features, their importance score, and the standard deviation on the score.

****

Unlike the previous 2 models, which use less than 5 features, the logistic regression model was able to utilise 10 variables. The most important one is the fuel system, the same as for the ANN. The sensitivity analysis (Table 3C) reveals a distribution slightly close to was expected following the literature review and variables description. The make, drive wheel distribution, number of cylinders and bore all play an important role. Nonetheless, features that were expected to be of significant importance, such as horse power or engine size are absent.

The SVM identified miles per gallon on the highway as the most important feature (Table 3D). It is also the only model to utilise this variable. It is followed by curb weight and city miles per gas, the latter of which is also not used in the other models. Among the other features, body style and fuel type are too not identified as relevant by the other models. Unlike the logistic regression, which picked up on engine specifics, the SVM is more focused on fuel characteristics.

A mixed sensitivity analysis is conducted between the SVM and Logistic Regression models (Figure 11), with the purpose of improving general understanding of what the best automobile price indicators may be. The aggregated model is achieved by normalising the importance by the corresponding’s algorithm accuracy. The results of the integrated sensitivity analysis model are generally close to what was observed by previous studies. However, the importance of fuel system and curb weight is somewhat unexpected.

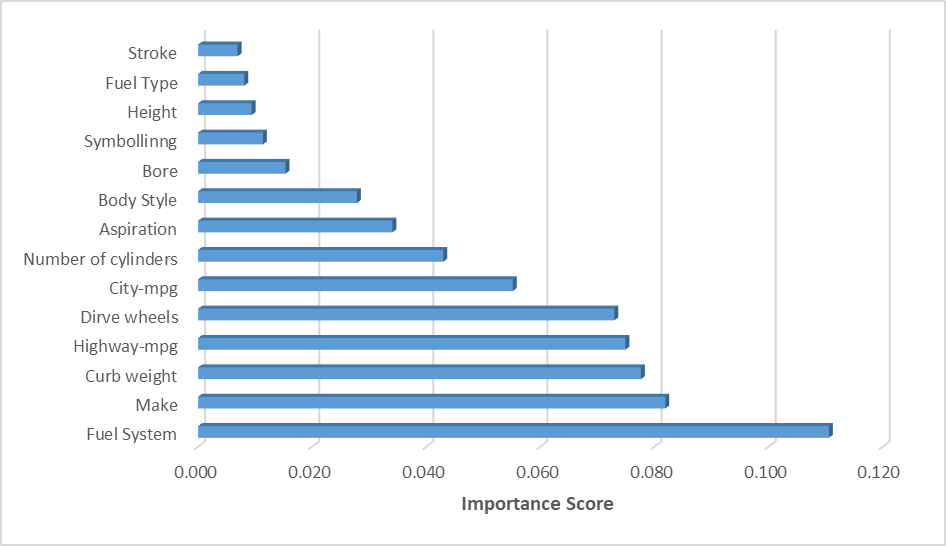
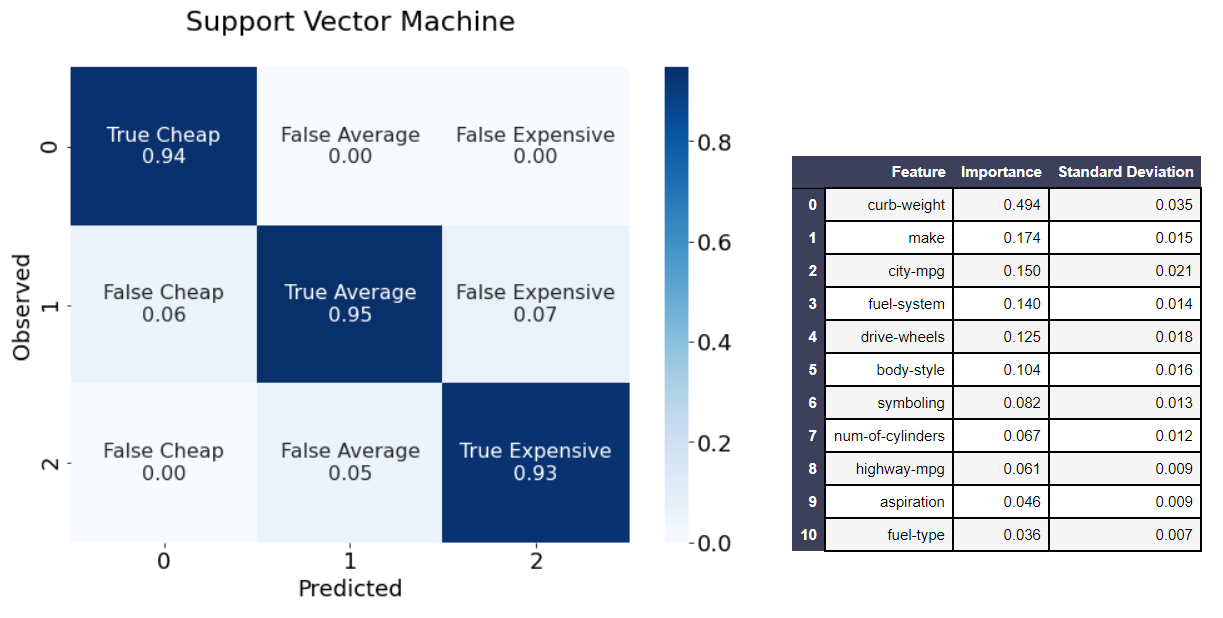


Figure 11. Aggregated Sensitivity Analysis of SVM and Logistic Regression.

In order to attempt improving accuracy, a new model is created by only keeping the 14 variables relevant among either SVM or Logistic Regression. A SVM model that correctly predicts 48 out of 51 is achieved (94.12%). Unlike in the original SVM, the best kernel was polynomial of 2nd degree, with a weaker penalty, C=10,000. The model achieves 100% sensitivity on the cheap category, and 94% precision, as it incorrectly predicts one cheap car to be averagely priced (Table 4). The precision on medium cost cars is 95%, as the model makes 19 predictions, out of which only is incorrectly categorised as expensive. The recall is 90%, as 18/20 averagely priced are accurately identified. A total 15 cars are predicted to be expensive, out of which 14 correctly. Specificity and sensitivity on this class are therefore both 93%. Curb weight is by far the most relevant feature, with an importance score (0.494) more times greater than the runner up, make (0.174). This contradicts previous literature, which usually puts make as the more important variable. The likely reason why this model achieved better performance is that, having to choose from less features, it was able to quickly the large importance of curb-weight, and choose a simpler kernel function, the 2nd degree polynomial.

Table . Shows the performance (left) and sensitivity analysis (right) of a SVM classifier fitted using only the 12 important variables.



Furthermore, the models were ran one more time without their most important features, in order to see the effect on performance (Table 5).

Table . The performance of the models when ran without one of their top 3 features.



Removing curb weight, the only feature used in the original kNN model, leads to a 5.88% drop in accuracy. For the ANN and Logistic Regression, whenever a variable is removed, the model’s accuracy either stays the same or decreased. The largest delta occurs when fuel system is dropped from ANN, 11.76%. However, dropping the same variable from Logistic Regression had no effect on performance. In the case of the SVM, a significant improvement was achieved by dropping its most important feature, highway mpg, from 86.28% to 94.12%. Intestingly, when city mpg was removed, the model’s accuracy decreased by 5.89%. In both cases when curb weight was removed, a drop in performance was recorded, further cementing the importance of the feature.

**4.Managerial Implications**

The results of this report suggest that car price prediction is feasible, the best accuracy results being 86.3% on 3 categories. Given the small number of vehicle data available (201), the performance of the model is encouraging, the expectation being that by obtaining more vehicle data, a much better model can be achieved.

Management is advised to focus on incorporating the 14 important variables within automobile pricing processes and frameworks. The relevant identified features are, as shown in Figure 10, in descending order: fuel system, make, curb weight, highway mpg, drive wheels, city mpg, number of cylinders, aspiration, body style, bore, symbolling, height, fuel type and stroke. These features are only somewhat consistent with previous research, specifically that, cylinder count and make are crucial cost predictors. Nonetheless, some engine force indicators such as horse power and compression ratio were not included. The most probable explanation is that other engine power and quality indicators such as fuel system, bore and stroke were preferred. It is recommended that these variables be included in future models, as they appear to provide superior accuracy compared to simply using only one metric for engine force, such as horse power. Moreover, fuel consumption is recognised as being an important cost predictor.

The importance of curb weight in price prediction is the most unexpected and consequential observation. The variable performs particularly well in identifying cheap vehicles, as seen in the kNN model, which only uses curb weight and has 93.75% recall. By integrating other variables, SVM is able to achieve 100% recall while only overestimating the number of low-priced vehicles by 1 (6%).

The high relevance of curb weight implies that the variable should be included in future datasets meant to analyse automobile value, as well as in general non-computational assessments.

By including age in the prediction process, accurate price predictions for used cars can be achieved. Management can leverage the results of this report in order to improve upon existing or upcoming second-hand automobile pricing decisions. Furthermore, the best model identified has the potential of aiding in odometer fraud detection, by identifying anomalies in predictions that include age vs those that do not, as well as by employing more sophisticated techniques which are beyond the scope of this report.

**5.Conclusions**

This report analysed automobile price prediction using k Nearest Neighbors, Multi-layer Perceptron, Logistic Regression and Support Vector Machines. The price is split into 3 roughly equal categories and the model performances is measured in accuracy: 72.549%, 68.627%, 80.392% and 86.725% respectively. Because the ANN failed to accurately predict averagely priced cars, not even performing significantly better than the 1/3 random guess change (35%), its feature importance was not included in the aggregate analysis. The same approach was taken for kNN, which had only 45% recall on averagely priced automobiles. After considering the performance of the three models and the sensitivity analysis, a new SVM is created, using a polynomial 2nd degree order kernel and a C regularization parameter of 10,000. Only the 14 variables identified as important were considered. An improvement is achieved as 48 out of the 51 test instances are correctly predicted, accuracy therefore being 94.12%, a 7.395% improvement.

Although previous research suggests that engine power and make are usually the best predictors, the sensitivity analysis conducted in this report identified fuel system as being the most important feature. However, after reducing the dataset to only the 14 relevant variables, the SVM model chooses curb weight as being by far the best price predictor, with an importance score of 0.494, compared to the second best variable, make, which has 0.174.

Performance improvements may be achieved by integrating ensemble methods, which would combine multiple machine learning algorithms. Management is advised to identify any relevant additional variables that might further improve the accuracy of the models, such as more details on engine parts: valve, ignition, piston, etc. Details on a car’s interior, such as material, chairs, steering wheel should be considered. Air conditioning, heating system, radio, and nowadays software, are important variables that could be added for newer datasets. Moreover, safety details such as airbags, ABS, ESP, should be included for more modern vehicles.

References

[1] Morgan, C. (2002, April). Preliminary Report: The Incidence Rate of Odometer Fraud. Retrieved October 28, 2020, from <https://one.nhtsa.gov/cars/rules/regrev/evaluate/809441.html>

[2] Montag, J. (2017). Identifying odometer fraud in used car market data. Transport Policy, 60, 10-23.

[3] Weisstein, Eric W (n.d.). "Benford's Law". MathWorld, A Wolfram web resource. Retrieved 30 October, 2020 from https://mathworld.wolfram.com/BenfordsLaw.html

[4] Statista Research Department, (2020, April). Retrieved 31 October, 2020, from <https://www.statista.com/statistics/183505/number-of-vehicles-in-the-united-states-since-1990>

[5] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764.

[6] Pal, N., Arora, P., Kohli, P., Sundararaman, D., & Palakurthy, S. S. (2018, April). How Much Is My Car Worth? A Methodology for Predicting Used Cars’ Prices Using Random Forest. In Future of Information and Communication Conference (pp. 413-422). Springer, Cham.

[7] Peerun, S., Chummun, N. H., & Pudaruth, S. (2015, June). Predicting the Price of Second-hand Cars using Artificial Neural Networks. In The Second International Conference on Data Mining, Internet Computing, and Big Data (BigData2015) (p. 17).

[8] Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application. Unpublished. <https://www.ifis.uni-luebeck.de/~moeller/publist-sts-pw-andm/source/papers/2009/list09.pdf>.

[9] Kuiper, S. (2008). Introduction to Multiple Regression: How Much Is Your Car Worth?. Journal of Statistics Education, 16(3).

[10] Dholiya, M., Tanna, S., Balakrishnan, A., Dubey, R., & Singh, R. (2019). Automobile Resale System Using Machine Learning.

[11] Muley, A. Jaideep (2017) Prediction of Used Cars’ Prices by Using SAS EM

[12] Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New model for residual value prediction of the used car based on BP neural network and nonlinear curve fit. In 2011 Third International Conference on Measuring Technology and Mechatronics Automation (Vol. 2, pp. 682-685). IEEE.

[13] Noor, K., & Jan, S. (2017). Vehicle price prediction system using machine learning techniques. International Journal of Computer Applications, 167(9), 27-31.

[14] Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car price prediction using machine learning techniques. TEM Journal, 8(1), 113.

[15] Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R (2000). CRISPDM 1.0. SPSS Inc. Retrieved 3 November, 2020 from: <http://www.the-modeling-agency.com/crisp-dm.pdf>

[16] Schlimmer, C. Jeffrey (1987). Ward’s Automotive Yearbook. Retrieved 4 November, 2020 from <https://data.world/uci/automobile/>

[17] Eagle Ridge GM (2020, February). Turbocharged Vs Naturally Aspirated Engines. Retrieved from November 5, 2020 from <https://www.eagleridgegm.com/turbocharged-vs-naturally-aspirated-engines/>

[18] Stewart, Ben (2004, September). Comparison test: front-wheel drive vs rear-wheel drive. Retrieved November 6, 2020 from <https://www.popularmechanics.com/cars/reviews/a54/1266931/>

[19] Leith Cars Blog (n.d.). Front vs Mid vs Rear Engines – Which is Best? Retrieved November 6, 2020 from <https://www.leithcars.com/blogs/1421/lifestyle/front-vs-mid-vs-rear-engines>

[20] King, Graham (2020, July). What Is a Car Wheelbase? Retrieved November 8, 2020 from [www.carwow.co.uk/guides/glossary/what-is-a-car-wheelbase-0282](http://www.carwow.co.uk/guides/glossary/what-is-a-car-wheelbase-0282)

[21] MrBlue & Ashvindx (2004, August). Wheelbase. Wikimedia. Retrieved November 10, 2020 from <https://commons.wikimedia.org/wiki/File:Wheelbase.png>

[22] Softies (2003, July). Wankel. German Wikipedia. Retrieved November 10, 2020 from https://en.wikipedia.org/wiki/File:Wankel-1.jpg

[23] Your Mechanic (2015, December). What Do Different Cylinder Numbers Mean in Regards to Engine Performance or Reliability? Retrieved November 9, 2020 from <https://www.autoblog.com/2015/12/02/what-do-different-cylinder-numbers-mean-in-regards-to-engine-per>

[24] Edelstein, Stephen (2020, February). Bore vs. stroke: which makes more power? Retrieved November 9, 2020 from <https://www.motorauthority.com/news/1127285_bore-vs-stroke-which-makes-more-power>

[25] Hall, Nancy (2015, May). Bore and Stroke. Glenn Research Center, NASA. Retrieved November 10, 2020 from https://www.grc.nasa.gov/www/k-12/airplane/stroke.html

[26] Dell, R. M., Moseley, P. T., & Rand, D. A. (2014). Towards sustainable road transport. Academic Press.

[27] Winterbone, D., & Turan, A. (2015). Advanced thermodynamics for engineers. Butterworth-Heinemann.

[28] Britannica (n.d.). Horsepower. Retrieved November 9, 2020 from [www.britannica.com/science/horsepower](http://www.britannica.com/science/horsepower).

[29] Lucid Chart. Available at <https://www.lucidchart.com/>.

[30] Srivastava, T. (2020). K Nearest Neighbor: KNN Algorithm: KNN in Python & R. *Analytics Vidhya*. Retrieved December 4, 2020 from https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

[31] Kumar, N. (2019). *Advantages and Disadvantages of KNN Algorithm in Machine Learning*. Retrieved December 4, 2020 from http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-knn.html

[32] Linux Uncle (2015). KNeighborsClassifier explained. *ashokharnal.wordpress.com*. Retrieved December 4, 2020 from https://ashokharnal.wordpress.com/tag/kneighborsclassifier-explained/

[33] Nielsen, M. A. (2015). Neural networks and deep learning (Vol. 2018). San Francisco, CA: Determination press.

[34] Starkweather, J., & Moske, A. K. (2011). Multinomial logistic regression. Consulted page at September 10th: http://www.unt.edu/rss/class/Jon/Benchmarks/MLR\_JDS\_Aug2011. pdf, 29, 2825-2830.

[35] Saputro, D. R. S., & Widyaningsih, P. (2017, August). Limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method for the parameter estimation on geographically weighted ordinal logistic regression model (GWOLR). In AIP Conference Proceedings (Vol. 1868, No. 1, p. 040009). AIP Publishing LLC.

[36] Gandhi, R. (2018, July 05). Support Vector Machine - Introduction to Machine Learning Algorithms. Retrieved December 05, 2020, from <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>

[37] Marius, H. (2020, September 09). Multiclass Classification with Support Vector Machines (SVM), Kernel Trick & Kernel Functions. Retrieved December 05, 2020, from https://towardsdatascience.com/multiclass-classification-with-support-vector-machines-svm-kernel-trick-kernel-functions-f9d5377d6f02

Hello, my name is Peter Neal and I am part of the team that presented Analysing the predictability of automobile prices using a Data Mining Approach with Daniel Falcone,Truc Le, and Hikma Abajorga .   
I am a graduate student at University of Massachusetts Lowell who is majoring in Data Analytics. My undergraduate is in Management Information Systems with a minor in marketing. I used to work at Kronos as a Engineering tech specialist, but now I attend my Masters full time. My side job is cryptocurrency and photography. I do that on the weekends. My aspiration is to go into the cryptocurrency space with deep mining techniques to apply in exchanges.

