Final Year Project Report

**Full Unit – Interim Report**

**Using Machine Learning Algorithms to Predict the Price of Pre-owned Cars**

Rizwan Bagdadi

A report submitted in part fulfilment of the degree of

**BSc (Hons) in Computer Science**

**Supervisor:** Li Zhang



Department of Computer Science

Royal Holloway, University of London

March 06, 2023

**Declaration**

This report has been prepared on the basis of my own work. Where other published and unpublished source materials have been used, these have been acknowledged.

Word Count: 5365

Student Name: Rizwan Bagdadi

Date of Submission: 09/12/2022

Signature: Rizwan Bagdadi

**Contents**

[**Abstract** 3](#_Toc129613863)

[**1.** **Introduction** 4](#_Toc129613864)

[**2.** **Literature Review** 5](#_Toc129613865)

[**3.** **Methodology** 5](#_Toc129613866)

[**3.1.** **Outliers** 5](#_Toc129613867)

[**3.2.** **Data Wrangling** 7](#_Toc129613868)

[**3.3.** **Normalisation** 8](#_Toc129613869)

[**3.4.** **Data Visualisation** 9](#_Toc129613870)

[**4.** **Implementation** 11](#_Toc129613871)

[**4.1.** **Linear Regression** 12](#_Toc129613872)

[**4.2.** **K-Nearest Neighbour** 12](#_Toc129613873)

[**4.3.** **Decision Tree** 15](#_Toc129613874)

[**4.4.** **Random Forest** 20](#_Toc129613875)

[**4.5.** **Support Vector Machine** 21](#_Toc129613876)

[**4.6.** **User Interface** 21](#_Toc129613877)

[**5.** **Analysis** 21](#_Toc129613878)

[**6.** **Engineering tools** 22](#_Toc129613879)

[**7.** **Planning and time scale** 23](#_Toc129613880)

[**8.** **Conclusion** 23](#_Toc129613881)

[**9.** **Bibliography** 25](#_Toc129613882)

[**10.** **Appendix** 27](#_Toc129613883)

[**10.1.** **Video links** 27](#_Toc129613884)

[**10.2.** **My diary** 27](#_Toc129613885)

# **Abstract**

In this paper, I investigate the application of supervised Machine Learning (ML) algorithms to predict the price of pre-owned cars in the UK. The goal of this project is to find the most suitable algorithm that can carry out the most accurate predictions, while providing the best performance and overall efficiency, by comparing their benchmark data. I use two techniques, K-nearest neighbour and decision trees, to make these predictions, using a dataset containing the features and prices of over 100,000 pre-owned cars to train the algorithms.

Throughout this paper, I first discuss why I chose to research this area, including the problems, the client, and the data. I then extensively describe and explain how each algorithm is effectively used, discussing the drawbacks and benefits of each, alongside extracts from my code with explanations on how they work. To give a more comprehensive view of the research, I also include a detailed explanation of the pre-processing techniques used to clean and prepare the data before training. Finally, I provide a full analysis and comparison of their outcomes, including a discussion on potential areas for future research in this field.

1. **Introduction**

Predicting the price of a pre-owned car is an important problem that benefits both the buyer and the seller. Unfortunately, individuals or dealerships often ask for an unreasonable price compared to the actual worth of the vehicle, taking advantage of those who are unsure of the approximate value of the car. When it comes to human-to-human interactions in quoting a fair and accurate price of a pre-owned car, there are several problems that must be highlighted: lack of knowledge, lack of awareness, lack of communication, or lack of resources. Discrimination is another problem that affects both the seller and the buyer. Discrimination based on age, gender, race, or religion can occur. From a seller's perspective, they may be unsure of the worth of the vehicle they are selling, leaving them vulnerable, which commonly occurs with the elderly trying to sell their cars and facing age discrimination. From the buyer's perspective, they could be quoted an unfair price due to the seller's malicious intent or prejudice against certain groups.

When selling or buying a car, it's important to take into account a variety of factors to ensure that you get an accurate and reliable price. While the brand and model of the car are important, there are several other features that can significantly affect the value of the car. For example, the mileage of the car is a key factor that can affect the price, as a car with high mileage may have more wear and tear on its components compared to a car with low mileage. The engine size of the car is also important, as larger engines tend to be more powerful and may be more expensive to maintain. Other important factors to consider include the registration year, transmission type, fuel type, and fuel economy [1]. Considering all of these features manually can be time-consuming and inefficient for calculating an estimate.

A system that utilizes a dataset containing hundreds of thousands of pre-owned car selling prices can benefit both sellers and buyers by predicting the selling price of a used vehicle. To ensure fair, reliable, and accurate results, I have opted to use machine learning to create this system. This approach eliminates all bias factors by removing user control from the system. The system uses supervised algorithms and requires training with a reliable dataset. Machine learning is useful in making accurate predictions, as it can learn from massive amounts of data and continue to learn. Advanced machine learning can even predict stock market trends, which would be extremely helpful in determining the selling prices of cars regardless of market conditions.

The aim of this project is to utilize machine learning (ML) algorithms to identify the most suitable algorithm for making accurate predictions while also providing the best overall efficiency and performance by comparing their benchmark data. To achieve this, I have utilised four different supervised machine learning algorithms, including Linear Regression, K-Nearest Neighbour, Decision Trees and Regression Forest, will be applied. Each algorithm used will undergo regression analysis and cross-validation to determine which provides the best performance. To train the algorithms, a dataset last updated in 2020 will be used, which contains the selling price and features of over 100,000 pre-owned cars [2].

1. **Literature Review**

Write here…

1. **Methodology**

I sourced my data from Kaggle [2], which was uploaded by user Aditya and last updated in 2020. The data was collected in the same year from live listings. To ensure the dataset was appropriate, I verified that all the data was well-organized and included information on price, mileage, road tax, miles per gallon (mpg), and engine size. Additionally, I checked that there were no duplicate listings that could impact my results. Table 1 shows a sample of the data used from the Audi dataset.

Text

Description automatically generated

**Table 1.** *Sample Data Used of Audi Vehicles*

To use the data in my models, I had to first make it suitable. This involved performing data preprocessing, which includes the steps needed to parse the raw data easily by the algorithms and system [3]. Most real-world data tend to have missing values, noise, inconsistency, or unusable formats that cannot be directly used in machine learning algorithms. As shown in Table 1, there are a few problems that preprocessing can take care of. For example, certain features have string values, and the integer variables vary from 1.0 to 35000+ for different features. This process increases the accuracy and efficiency of the models.

## **Outliers**

Outliers are data points that differ significantly from the other values in a dataset. They can affect the accuracy of our results, leading to inaccurate findings. Therefore, it is crucial to identify and handle outliers appropriately. The first step in preprocessing is to find any possible outliers in the dataset that could influence the results and remove them.

To identify outliers, we first calculate the interquartile range (IQR). This is done by finding the difference between the 75th percentile and the 25th percentile of the dataset and multiplying the result by a constant of 1.5. This provides the maximum and minimum values that are not considered outliers. Any values found outside this range are considered outliers. The code sample below shows the steps taken to find and eliminate the outliers.

**for i in ['year']:**

q75,q25 = np.percentile(file.loc[:,i],[75,25])

      IQR = q75-q25

      max = q75+(1.5\*IQR)

    min = q25-(1.5\*IQR)

  file.loc[file[i] < min, i] = np.nan

    file.loc[file[i] > max, i] = np.nan

file = file.dropna(axis = 0)

While identifying outliers is a key step, it is also necessary to understand the reasons behind their occurrence. In some cases, outliers may be valid data points that represent unique characteristics of the dataset. For example, the tax price of an expensive car model may be an outlier, but it is a valid data point that should not be removed.

Chart, box and whisker chart

Description automatically generated

**Figure 1.** *Box Plot demonstrating outliers in Audi dataset.*

Figure 1 displays the outliers identified in the Audi dataset. After analysing the data, I determined that the outliers in mileage, tax, mpg, and engine size are useful for feature analysis. However, the dataset includes only one data point for the years 1997, 1998, and 2002-2014. Since the year greatly impacts the algorithm's accuracy, I removed these outliers. While this reduces the available options for the user to select, it is necessary to ensure the accuracy of the algorithms.

Once the outliers have been identified, there are various ways to handle them. They can be removed from the dataset, treated separately in the analysis, or transformed using mathematical functions. As shown in the code sample above, I replaced all the outliers with Numpy's nan (Not a Number). I then handled these outside the for loop by removing all null values from the dataset.

Chart, box and whisker chart

Description automatically generated

**Figure 2.** *Box Plot demonstrating outliers in Audi dataset.*

After removing the outliers within the year feature, Figure 2 shows the Audi dataset. As a result, we can see that the extreme outliers in mileage have also been removed. By removing these outliers, we can ensure higher accuracy and reduce the chance of extremely inaccurate predictions. In addition, it is important to note that removing outliers may not always be the best solution. In some cases, it may be more appropriate to transform the data or analyse the outliers separately. Therefore, handling outliers requires careful consideration and a thorough understanding of the dataset being used.

## **Data Wrangling**

For the next step of preprocessing, I performed the data wrangling process, which involved a series of complex operations that transformed raw data into a usable and workable format. This was a crucial step in the data analysis process as it made it easier to process and obtain reliable output from the data. The dataset I was working with is split into car brands, each containing thousands of samples. To reduce the sample size and make it more manageable, I initially chose to use only one car brand, Audi. However, as I progressed with my analysis, I realized that limiting my analysis to just one car brand would not give me a complete picture of the data. As a result, I changed my code to work with all the car brands provided in the dataset.

However, within each car brand, the dataset was too large for my laptop to process, with a shape of (10668, 9) for the Audi dataset. To address this issue, I had to find a way to reduce the sample size without compromising the integrity of the data. After much consideration, I decided to use the Pandas head() function to select a sample size of 5000 for each car brand.

Text

Description automatically generated with low confidence

**Table 2.** *Data Types Used*

The first step I took in analysing the data was to examine the shape and type of variables included in the dataset, as shown in Table 2. This involved a careful examination of the data types found in the dataset, which were identified using the dtypes function from the Pandas library. I discovered that most of the variables were integers, with some float and object types. However, I encountered a problem with the variables that contained string values, such as the transmission, fuel type, and car model. These variables were critical in determining the selling price of the cars, but I could not use them in my algorithms because they were not numerical values.

To overcome this problem, I used Scikit-Learn's LabelEncoder() feature, which assigns numerical values to strings between 0 and the number of values – 1. This allowed me to include the variables with string values in my algorithms and obtain more accurate results in predicting the selling price of the cars. Overall, these adjustments helped me to obtain a more comprehensive understanding of the data and make more informed decisions in my analysis.

modelEncoder = LabelEncoder()

modelEncoder.fit( file [ "model" ])

file[ "model" ] = modelEncoder.transform( file[ "model" ])

file = file.head(5000)

The sample code above demonstrates how I used LabelEncoder to modify the variables of my model. For each feature (transmission, fuel type, and car model), I used a separate LabelEncoder to fit each feature, ensuring that there was no influence in transforming the data.

## **Normalisation**

Another thing I had to consider before using the dataset, is the spread of data. Table 1 provides an overview of the dataset and its wide range of values. For instance, the integer variables vary from 1.0 for engine size to over 35,000 for mileage, resulting in a significant difference in their ranges. To address this, I used the process of normalization, which transformed the variables in columns to have the same scale, making them more comparable. Note that not every dataset needs to be normalized for machine learning, only when the characteristics have different ranges [4].

Some machine learning algorithms, especially K-Nearest Neighbours and Support Vector Machines, benefit from normalization. This is because they use distance from data points to determine their similarities [5]. Figure 1 below shows the effects of scaling on a set of data points when comparing the Euclidean distance. We can see that scaling has brought both features closer to each other, making them more comparable than before.

A picture containing text

Description automatically generated

**Figure 1.** *Example of before and after of scaling [g]*

On the other hand, tree-based algorithms are not affected much by the scale of the features, as the decision process is not influenced by other features [5]. Therefore, I only normalised the data for the K-nearest neighbour algorithm. There are many different methods of normalisation, but for my project, I chose to use Min-Max scaling. Min-Max scaling works by setting the minimum of each feature to 0 and the maximum to 1, transforming the range into 0-1 [4].

When normalising the data, I had the option of normalising the data before or after splitting the data. However, I chose to normalise the data after splitting the dataset, as normalising it before can cause details of the test set to be leaked. This is because the mean and standard deviation that is used in the normalisation process would be used on the entire dataset, including the test data [6].

## **Data Visualisation**

After cleaning the data, we can visualize it much more clearly and observe the relationships and correlations between each aspect of the cars. Figure A below shows the distribution of each feature, excluding fuel type, transmission, and model.

Looking at the histogram graphs, we can already deduce a number of facts. Most used cars being sold are from 2019 and have a 2-litre engine with low mileage, but with moderately high fuel consumption. User-entered specifications that match these modes of features will have a higher accuracy prediction compared to cars with features like a 2012, 2.2-litre engine, 120,000+ miles, and extremely high fuel consumption. It should never be unexpected for a vehicle with such specifications, and it should be handled appropriately. Fortunately, with machine learning, all dimensions are considered and taken into account when formulating a prediction. This isn’t to say the prediction would be extremely accurate.

Chart, histogram, waterfall chart

Description automatically generated

**Figure A.** *Distribution of Features in Audi Dataset*

Figure 1 depicts the relationships between each feature using Seaborn library's Heatmap feature. By examining the heatmap, we see that price has a positive correlation with engine size and year, indicating that people are willing to pay a premium price for newer cars with larger engines. However, it has a negative correlation with MPG and mileage, suggesting that people avoid cars with high fuel consumption and high mileage. The strongest correlation is between year and mileage, showing that older cars generally have higher mileage.

**Chart, bar chart

Description automatically generated**

**Figure 1.** *Relations Heatmap of Audi Dataset*

When purchasing a used car, mileage is often a key consideration as it can impact the car's reliability. Cars with higher mileage may have endured more stress, which can cause issues with the engine or other components. Figure 2 provides a more detailed view of how mileage impacts the selling price of a used car. As expected, cars with higher mileage tend to be older and sell for less. This supports the idea that buyers are willing to pay more for newer cars with lower mileage. However, it is important to note that other factors such as the car's make and model can also impact its selling price.

Chart, scatter chart

Description automatically generated

**Figure 2.** *Scatter Plot of Relation Between Mileage, Price, and Year of Audi Dataset*

Furthermore, the data reveals that most cars are sold within five years of ownership, with the mode year being 2019. This suggests that people tend to sell their cars after a few years of use, possibly to upgrade to newer models or to obtain a different type of vehicle. This trend may also be influenced by factors such as changes in lifestyle or financial circumstances. In summary, while mileage is an important factor when purchasing a pre-owned car, it is just one of many factors to consider. By taking into account the car's mpg, tax, transmission, fuel type, and other factors, buyers can make a more informed decision when buying a used car.

Figure B shown below further illustrates the correlation between price and year, highlighting the stark difference between the two factors. As we can see, the prices of newer cars tend to be higher than those of older cars. This could be attributed to a number of factors. Firstly, newer cars tend to have less mileage and therefore less wear and tear, which could lead to higher reliability and in turn, higher prices. Additionally, newer cars often come with more advanced features and technology, which could also contribute to their higher prices. It is also important to note that the demand for newer cars may also be a factor in the higher prices, with many consumers willing to pay a premium for the latest models.

Chart, bar chart

Description automatically generated

**Figure B.** *Bar graph showing the relationship of price and year in the Audi Datatset.*

1. **Implementation**

In order to ensure that the outcomes of the project were reliable, I split the dataset used into two subsets: a training subset and a testing subset. The training subset consisted of 75% of the data, while the testing subset was made up of the remaining 25%. By using such a large training set, I was able to increase the reliability of the outcomes.

To create and analyse algorithms, I focused on three models: K-Nearest Neighbours, Decision Trees, and Random Forest. Each of these models has a unique approach and is likely to produce different results. For instance, K-Nearest Neighbours is a type of instance-based learning, while Decision Trees and Random Forest are more rule-based [g][h]. By using three different models, I was able to ensure that my analysis was thorough and that I considered all possible outcomes.

To begin my analysis, I started by testing the K-Nearest Neighbours algorithm. This algorithm works by classifying new data points based on the nearest neighbours in the training set [7]. By testing this algorithm first, I was able to establish a baseline for my analysis and gain an understanding of how the other algorithms compared.

* 1. **Linear Regression**

Linear Regression

Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable. However, in practice we often have more than one predictor. One option is to run three separate simple linear regressions. However, the approach of fitting a separate simple linear regression model for each predictor is not entirely satisfactory. Instead of fitting a separate simple linear regression model for each predictor, a better approach is to extend the simple linear regression model (3.5) so that it can directly accommodate multiple predictors. We can do this by giving each predictor a separate slope coefficient in a single model. Multiple Linear regression takes the form: Y = β0 + β1X1 + β2X2 + ··· + βpXp + ϵ



* 1. **K-Nearest Neighbour**

K-nearest neighbour (KNN) is a widely used and effective non-parametric method of supervised learning that uses proximity to predict the grouping of an individual data point [7]. It is a versatile algorithm that can be used for both classification and regression cases.



**Figure 3.** *Visual of KNN [8]*

KNN works by selecting the closest k data points to the test new point. This method is highly customizable, as the value of k can be adjusted based on the specific problem at hand. To do this, we calculate the Euclidean distance to each training data point and choose the k closest ones. There are other methods for calculating distance, such as Manhattan (for continuous data) and Hamming distance [9]. The Euclidean distance is calculated by taking the square root of the sum of the squared differences between the new point and the training points, as shown in Figure 4. Once the k closest points have been identified, we calculate the average of their labels to determine the prediction. In this case, the label is the price of the item being predicted.

Chart, line chart

Description automatically generated

**Figure 4.** *Euclidian distance [10]*

One advantage of KNN is its simplicity and ease of interpretation. It is fairly easy to understand how it works and how it makes predictions. Additionally, it can handle both categorical and numerical data, as long as the data is transformed into dummy variables, making it a versatile approach. However, KNN can be sensitive to outliers and has extremely poor run time performance with large training data [i].

Below is an extract from my KNN algorithm, that is slightly modified to be more readable.

*# K-Nearest Neighbour method*

**def KNN(train, testRow, yTrain, yTest, K):**

distance = list() #Stores distance of each point

for i in range(len(train)-1):

dist = eucDistance(train[i], testRow)

distance.append((train[i], dist, yTrain [i]))

distance = sort(distance) *# Gets sent to sort method – sorts in ascending order*

kNeighbours = list() #list to store K amount of neighbour results

for i in range(K):

kNeighbours.append((testRow, distance[i][1], distance[i][2]))

return kNeighbours

*#Predictes the price of the given car*

**def predict(train, test, yTrain, num\_neighbors):**

predictions = list()

for i in range(len(test)):

neighbour = KNN(train, test[i], yTrain, num\_neighbors)

labels = [] #Stores yTrain for each test variable

for i in range (len(neighbour)):

t = labels.append(neighbour[i][2]) #Appends yTrain

predictions.append(mean(labels))

return predictions

*# Evaluating 100 nearest neighbours*

**def eva(train, test, yTrain, yTest):**

rmseValues = []

for k in range(1, 101):

y\_pred = predict(train, test, yTrain, k)

error = rmse(yTest, y\_pred)

rmseValues.append(error)

print('RMSE value for k =' , k , 'is:', error)

return

In the code above we can see how the evaluation process begin by creating a list of the RMSE values, inside the “eva()” method, then proceeds to loop through 100 values of K, in creating predictions. The prediction method takes all the training values and only the features of the test set, with this it creates a loop that iterates over each row (car) in the test set and finds the K nearest neighbours to that row using the KNN method.

The KNN method works by iterate rating over all the training set rows and finding their Euclidean distance to the test row. This is a costly process as it must work out the distance for 750 points to just one of the 250 rows of test data. Meaning the Euclidean distance must be worked out for 187,500 times for only one value of K. This is a huge factor when considering using KNN, as big datasets will be very time consuming. If I hadn’t reduced the Audi dataset to 1,000 vehicles, the Euclidean distance would have had to be calculated an astonishing 18,750,000 times for one value of K.

Once the KNN method has calculated the Euclidean distance to each training row, it then sorts the list of distances and takes the K nearest points to the test row and returns these values, along with their labels, to the prediction method. Finally, the prediction method takes the labels of the K nearest neighbours and calculates their mean, returning the prediction of that test row or user input.

For the evaluation process, the prediction method stores all given predictions in a list, of which it is returned to the evaluation after the completion of all predictions. With the given predictions, the evaluation method then calculates the RMSE value for each value of K and proceeds to plot these values, enabling a visualisation of the accuracy of my algorithm.

Text, letter

Description automatically generated

**Figure 5.** *Demonstration of user input for my KNN algorithm*

After testing my algorithm against a test set and finding the optimal number of neighbours for my dataset, I created a temporary user interface, shown in figure 5, that allows users to provide details of their own vehicles, currently limited to Audis only, and receive a reasonably accurate prediction. Figure 5 shows an interaction where the details for a 2016 Audi RS6, with 49,050 miles, and an actual worth of £44,985 was inputted into my algorithm using 4 neighbours, outputting a prediction of £44,717. This result being -£268 from the actual worth and with a calculation time of 0.0938 seconds (3 sf) demonstrates promising results from my KNN algorithm. In order to calculate the prediction time, I used the time library to measure the time it took for the prediction method to run and return a prediction.

* 1. **Decision Tree**

A decision tree is a non-parametric form of supervised learning, similar to KNN, that uses labelled input and output datasets to train models. Decision trees are mainly used for classification problems, but can also be used in regression problems to predict outputs from unseen data. Essentially, a decision tree is a tree where “each node shows a feature (attribute), each link (branch) shows a decision (rule) and each leaf shows an outcome (categorical or continues value)” [11]. Decision trees are designed to emulate human-like thinking, resulting in a much more natural thought process, similar to how humans think. Figure 6 illustrates the basic layout of a decision tree, showing sub-trees that fine-tune the predictions from the previous node, starting from the root node, and ending with the leaf nodes.

Diagram

Description automatically generated

**Figure 6.** *Decision Tree Illustration [12]*

For my decision tree algorithm, I used the same dataset as in my KNN algorithm, which had a size of 1000. However, before splitting the data, I decided to perform some preprocessing steps to obtain better results. To start, I converted the features variable (X) using the to\_numpy() function from the Pandas library. This function provides a more efficient way to work with the data, converting a data frame into a NumPy array. This allowed me to handle the data more effectively and avoid unnecessary overhead.

Furthermore, I also reshaped the labels variable (Y) using values.reshape(-1, 1), shaping it into an array with 1 column. This step was necessary to ensure that the data was properly structured and processed by the algorithm, which required a specific format for the input data. Unlike the KNN data, I did not normalize the data, as the decision process isn't influenced by other features, and the results aren't affected by other features. ADD TO THIS – how it doesn’t affect.

When building a decision tree using the training set, a greedy approach is used to make the best split at each step. This means that the algorithm does not look ahead to choose a split that may produce a more desirable tree later on. The desirability of a tree is measured by calculating the information gain, which determines how many features became more accurate than their parent nodes prior to being split.

In classification, we use a method called entropy to measure information gain on a new example's class [13]. This method measures the randomness within the information being processed [14]. In regression, we assume that we will produce an output given the input variables, rather than attempting to predict a class label. We examine whether a split would result in a decrease in the variance of the data. A split is considered useful if the collective weighted variance of its children has a smaller value than the parent's original variance.

The dataset is split recursively using binary splitting until the current node reaches a predetermined minimum number of splits required to split an internal node or a maximum depth allowed. This method follows a top-down, greedy approach to grow the tree with the training set.

Diagram

Description automatically generated with medium confidence

**Figure 7.** *Perspectives of recursive binary splitting [15]*

Figure 7 depicts perspectives of the recursive binary splitting process. The top right panel shows the splitting on a two-dimensional plane, which is then translated into the tree in the bottom left, and then into a 3D perspective of that [15]. This figure provides a vague visualization of the depth of recursive binary splitting that occurs on many planes with the dataset of cars being used.

Timeline

Description automatically generated

**Figure 8.** *Illustration of my decision tree using the training set*

Figure 8 shows a view of my training set, where the maximum depth is 3. We can observe how the algorithm splits at each node, based on the value of the current feature.

Below is an extract of the Node class from my Decision Tree algorithm.

*# Node class to initialise instances of each node*

**class Node():**

**def \_\_init\_\_(self, feature = None, limit = None, leftSide = None, rightSide = None, gain = None,**

**leaf = None):**

self.feature = feature

self.limit = limit

self.leftSide = leftSide

self.rightSide = rightSide

self.gain = gain

self.leaf = leaf

This extract of code above is used to initialise the instances of each node and store all their values. The feature variable stores the current features of the node. The limit stores the local threshold of the node, where the threshold is calculated as the largest value from the training set that is below the limit. The “leftSide” and “rightSide” variables store the subtrees on either side of the node, given from the split method. The gain variable stores the information gained with on each node. Finally, the leaf variable stores the mean of the labels/prices of all the leaf nodes.

Below is an extract of the DTRegressor class from my Decision Tree algorithm along with technical descriptions of how they work

**class DTRegressor ():**

**def \_\_init\_\_ (self, minSamples, maxDepth):**

self.root = None

self.minSamples = minSamples

self.maxDepth = maxDepth

**def fit (self, X, Y):**

trainingSet = np.concatenate((X, Y), axis=1)

self.root = self.treeBuild(trainingSet)

The class begins by creating an instance of the class using “*myTree = DTRegressor(3, 3)*”,

where the *init* method takes the 3’s to define the minimum samples and the max depth of the

tree. This is then followed by a call to the *fit* method, with the training sets features and

labels, which concatenates the training set into a whole data set, using the NumPy library,

and then calls the *treeBuild* method with the training set.

**def treeBuild (self, trainingSet, currentDepth = 0):**

X = trainingSet [:,:-1] *# everything but the last value*

Y = trainingSet [:,-1] *# only the last value*

*# Iterates until this condition is met*

if X.shape[0] >= self.minSamples and currentDepth <= self.maxDepth:

bestSplit = self.bestSplit(trainingSet, X)

if bestSplit["gain"] > 0:

leftTree = self.treeBuild(bestSplit["leftSide"], currentDepth + 1)

rightTree = self.treeBuild(bestSplit["rightSide"], currentDepth + 1)

node = Node(bestSplit["feature"], bestSplit["limit"], leftTree, rightTree,

bestSplit["gain"])

return node

*# Calculates mean of leaf nodes*

leafValue = np.mean(Y)

val = Node (leaf = leafValue)

return val

The *treeBuild* method begin by splitting the dataset back into features and labels. The method then moves into a recursive binary loop that only stops once these conditions are met: when the shape of the tree/sub-tree reaches the minimal sample and the depth of the sub-tree reaches the max depth. As soon as the code enters this loop it calls the bestSplit method which finds the best splits to build the tree. After finding the best split of the current node, it then finds the best split of the left and right children, and then saves and returns the node along with its children/sub-tree values. This recursive binary loop keeps occurring until the conditions are met. Once met, the mean value for each leaf node is calculated. In classification, the mode is used instead of the mean, just like with KNN.

**def bestSplit (self, trainingSet, X):**

bestSplitt = {}

biggestGain = -1

for feature in range(X.shape[1]):

featureValues = trainingSet [:, feature]

thresholds = np.unique(featureValues)

for j in thresholds:

leftSide, rightSide = self.splitTree(trainingSet, feature, j) #splits node into 2 sub-trees

if (len(leftSide) > 0 and len(rightSide) > 0 ):

parent = trainingSet [:, -1]

leftNode = leftSide[:, -1]

rightNode = rightSide[:, -1]

currentGain = self.infoGain(parent, leftNode, rightNode)

if currentGain > biggestGain:

bestSplitt["feature"] = feature

bestSplitt["limit"] = j

bestSplitt["leftSide"] = leftSide

bestSplitt["rightSide"] = rightSide

bestSplitt["gain"] = currentGain

biggestGain = currentGain

return bestSplit

**def splitTree (self, trainingSet, feature, limit):**

left Branch = []

rightBranch = []

for i in trainingSet:

if i[feature] <= limit:

leftBranch.append(i)

else:

rightBranch.append(i)

rightBranch = np.array(rightBranch)

leftBranch = np.array(leftBranch)

return leftBranch, rightBranch

The best split method starts by creating a dictionary to store the best splits and a variable that stores the highest gain. The code then goes into a for loop that iterates over every feature storing and iterates over each unique features, finding the split of the left and right branch using the *splitTree* method. After splitting, I appended the current node labels to a parent list and the left/right nodes labels to a left/right node list. With these lists, I measured the info gain on them and then compared it to the biggestGain variable. If it was bigger than the biggest stored gain, that would then become the biggest gain variable and the best split.

**def infoGain (self, parent, leftNode, rightNode):**

leftWeight = len(leftNode) / len(parent)

rightWeight = len(rightNode) / len(parent)

information\_gain = np.var(parent) - (leftWeight \* np.var(leftNode) + rightWeight \*

np.var(rightNode))

return information\_gain

The information gain method works by taking the weight of both the left and right nodes computing the variance reduction.

**def prediction Loop(self, test Row, root):**

if root.leaf != None: #not empty

return root.leaf

featureVal = testRow[root.feature]

if featureVal <= root.limit:

return self.predictionLoop(testRow, root.leftSide)

else:

return self.predictionLoop(testRow, root.rightSide)

**def predict(self, xTest):**

predictions = [ ]

for row in xTest:

predictions.append(self.predictionLoop(row, self.root))

return predictions

The prediction method takes the testing variable and iterates over each row in the set given and calls the *predictionLoop* method for each row. The *predictionLoop* method also uses a recursive binary loop that repeats until it reaches a leaf. It searches through both the left and right branches of the already trained tree, by checking whether or not it feature value is less than or greater than the limit/threshold.

* 1. **Random Forest**

Random forest (RF) regression is an ensemble machine learning algorithm that is constructed using multiple decision trees together to produce one final prediction, hence the reason it’s called a forest. The algorithm works by building a set number of decision trees, of which each tree predicts its output based on a subset of input variable. Each decision tree in the random forest is built using a different sample of rows, which is selected through the process of bagging. Bagging, also known as bootstrap aggregating, is a process that involves taking several random samples of the original dataset, with replacement. This minimises the variance of the model and prevents overfitting [J]. The algorithm then selects the best split at each node, based on a random sample of features selected for splitting. This technique is known as feature bagging and further minimises overfitting and helps the model’s generalisation to new data [J].

Overall, the utilization of bagging and feature bagging in random forests results in accurate predictions and reduces the risk of overfitting by reducing the impact of noise and outliers in the dataset. The final prediction is made by calculating the average of the predictions from all the trees [a]. This is an ensemble modelling technique that aggregates models to improve test accuracy, while reducing the costs associated with storing, training, and obtaining inferences from multiple models [b].

Diagram

Description automatically generated

**Figure 9.** Illustration of bagging

Random Forest Regression is a versatile algorithm that can handle a wide range of data types, including categorical and continuous variable. It also works well with noisy or incomplete data, making it a popular choice for real-world data [c]. Random Forest Regression has several other advantages, such as its robustness to outliers, where output outliers only affect the estimate of the leaf node it’s in and not the value of other leaf nodes; this is where the RF algorithm tends to stand out in comparison to other algorithms [d]. RF’s are also capable of handling non-linear relationships between variables, unlike other machine learning algorithms that are designed specifically for linear relationships, making it a versatile and powerful tool for data analysis [e].

In addition, the RF algorithm is also well suited for large volumes of data, since it can analyse datasets with vast amounts data points efficiently, making it a popular choice for large data applications. Moreover, it is easily parallelizable, allowing it to benefit from multi-core processors and distributed computing power to further maximise its performance on large datasets.

* 1. **Support Vector Machine**

Write here…

* 1. **User Interface**

In order to make my program suitable for a wide audience range, I designed a simple and intuitive user interface. To achieve this, I used PyQt, a toolkit that enables developers to create graphical user interfaces for desktop and mobile applications. It consists of Python bindings for Qt, a set of C++ libraries and development tools that offer platform-independent abstractions. PyQt provides a wide range of libraries, tools, and widgets that can be used to build complex and modern user interfaces.

1. **Analysis**

By working out the RMSE values from each algorithm, I was able to get an accurate view on the accuracy of predictions produced by both algorithms. For my KNN algorithm I worked out the RMSE value of each K value, from 0 to 100, to find the optimal value of K that enhances the accuracy of the predictions. In doing so, I was able to see that 4 nearest neighbours were the best value of K for my data set, giving a value of £3448.40 (2 dp). RMSE is the root mean square error, which shows how far the predictions falls from the actual values, by calculating the average deviation between the actual price and predicting price, using the Euclidean distance between them [16]. The lower the RMSE value, the better the model fits the dataset. In figure 5, we see that the best values of K for this data set are between 3 and 6, with the best being 4 neighbours. After k = 6, there is a huge linear increase in the value of RMSE, showing that the larger the value of K, the less accurate the results.

Chart, line chart

Description automatically generated

**Figure 10.** *RMSE value for values of K*

To analyse my decision tree algorithm, I created an evaluation method that takes the predictions of X\_Test and the original labels/prices to work out their RMSE. Unlike with KNN, I did not manage to fine tune my decision tree algorithm, setting my minimum sample number and maximum depth at 3. With KNN I had created an evaluation method that tested every value of K, between 1-100, to find the optimal number of neighbours. Unfortunately, I did not manage to find the optimal setting for my decision tree, thus not setting a fair comparison between the two algorithms. This is one of many improvements I need to make to my decision tree algorithm.

Looking at figure 10, we can see that the KNN model did not reach the same RMSE as the decision tree model until it used 60+ neighbours, demonstrating a superior ability to make accurate predictions. The less accurate results of the decision tree algorithm could be due to overfitting. Another issue to address with decision trees is that they are not robust and require training each time a change is made to the dataset.

Due to the KNN algorithm having to calculate many metric distances, it takes an extremely long time to process all the calculations for a large dataset, but a positive aspect of the KNN algorithm is that it’s instance-based learning, meaning it does not have to learn anything during the training period, instead it learns from the training set while making predictions. This also means that new data can be added to dataset whenever, without the problem of having to train the algorithm again.

1. **Engineering tools**

The main reason for me using Python as my chosen language, is due to its access of great libraries that are great for machine learning use; specifically SciKit-Learn and Pandas. SciKit-Learn was extremely useful for handling basic ML aspects, such as normalising and splitting my dataset into training and test sets, It was also useful to compare my algorithms too, as they contain functions that calculate regression predictions for many ML algorithms. Another useful library was Pandas, which was very useful for handling my dataset and presenting it in a clear viewable format. The code for Python is also very readable and understandable, which was needed when handling complex codes such as machine learning algorithms.

My main software to program on was Jupyter Notebook, a web-based computing platform that clearly lays out your code. It is extremely useful in illustrating my analysis at each step and helped me view all of my code in one place, and also allows for other people to clearly see and understand all of my code together. It also contained all the libraries I needed, without me having to install each and everyone of them on my personal computer.

Graphical user interface, application

Description automatically generated

1. **Planning and time scale**

As far as following my timeline, proposed in my project plan, I hadn’t followed it very well, as I had not considered external factors that could affect my progress of work. In terms of completing research, I was able to do that in the proposed time but was unable to implement my code in the set target time. Which then affected my other proposed tasks, especially the two weeks to fine tune and tweak my algorithms to be as accurate as possible. I had initially struggled to grasp the concept of how decision trees worked, with the concept of splitting with multiple features, which is why I was unable to improve my decision tree algorithm, giving it an unsatisfactory RMSE value of £5486.19 (2 dp).

1. **Conclusion**

Predicting the price of used cars can be a challenging task to complete, due to many affecting factors. My aim throughout this project was to find the most suitable ML algorithm capable of carrying out the most accurate and reliable predictions. As far as those aims go, I was successful in producing two machine algorithms: k-nearest neighbour, and decision trees. Both algorithms produced promising results with my KNN algorithm producing an RMSE value of £3447.68(2 dp), and £5486.19 (2 dp) for my Decision Tree.

Unfortunately, I was unable to fine tune my decision tree algorithm and find its optimal minimum sample value and maximum depth value, due to limited time. I was also unable to fully understand certain aspects of Decision tree. My decision tree algorithm was heavily inspired by an algorithm written by @Suji04 from GitHub.com, and due to my time limiting factor, I did was unable to change and optimise certain aspects that I thought weren’t fully efficient [17]. One of my many goals for my final report, is to use this Decision Tree algorithm I produced, and adapt on it. One way I plan on doing this is by performing a grid search to find the optimal parameter values for my algorithm.

Although I was unable to fine tune my decision tree algorithm, it was able to complete the prediction, from figure 6, in a significantly faster time of 0.000997 seconds (3 sf) and a predicted a price of £41,233.30, compared to the 0.0938 (3 sf) seconds from the KNN algorithm with a much more accurate price of £44,717, using 4 neighbours. The cars actual worth was £44,985. Originally, my KNN algorithm did not include any normalisation of data, and I had not planned to use the model, transmission, or fuel as factors for either algorithm. Without normalising my data for my KNN algorithm, the RMSE value increased to £9023.19, giving very unreliable results.

Another limiting factor of mine was not being able to use the dataset to its full extent. My goals for next time is to be able to use all car brand datasets, allowing for multiple car brands to be used, and to increase the limit of vehicles used from 1000 to the full set. Using the full dataset will allow for me to produce more accurate results, but will all result in a longer processing time, especially for the KNN algorithm, but this will allow for a deeper analysis to be made. Also, I had limited my training set to 75%, where a higher training set would also lead to more accurate results.

Overall, my algorithms both produced very promising results, along with decent processing time. However, there are still many improvements I would like to make. First, I would like to fine tune my decision tree algorithm and better understand all aspects of the algorithm. I also want to research into the Random Forest algorithm, an algorithm developed by Ho, proposed to deal with overfitting for large datasets. This algorithm combines several decision trees and reduce overfitting without massively increasing error due to bias. Another proposed concept that I will try out, is pruning, a concept that removes parts of the tree that prevent it from growing to its full potential.

Other than focusing on my decision tree, I want to improve my KNN algorithm by adapting my code to run an evaluation process, with a given set of values of K, which automatically finds the optimal value of K, for the data set being used, and allows the users input to use that value of K. This may induce a time-consuming process for the, as it would have to calculate millions of Euclidean distances, which has proven to be extremely time consuming during this project. Another idea, I would like to adapt on is using multiple machine learning algorithms combined to produce a singular output. I saw this idea proper in a paper written by Gegic et al [18].

To conclude this paper, I believe that a successful comparison was made, showing that that my K-Nearest Neighbour algorithm was able to produce more accurate results, but much slower in comparison to the decision tree, due to the mass number of Euclidean distances to be worked out. There are still improvements that are required to be made to both algorithms, as a difference of £3400 is still a significant when selling a used car, whether it be £3400 more or less, as it could affect either seller or buyer. Also, a deeper analysis of both algorithms is required, using a cross-validation and regression analysis.

**KNN vs RF:** Scale Tolerance Features often have different scales, and some algorithms, such as k-Nearest Neighbors, only do well if features are first transformed to have the same range. While the best implementations of k-NN do that automatically, naïve implementations don’t, and one can get very misleading results (where the variables with the largest ranges dominate the distance, and thus the decisions). One never has to worry about scaling, though, with Random Forests. As they are based on decision trees, they only depend on rank; that is, only a data point’s relative value within a feature, not its magnitude, matters during training. <https://www.elderresearch.com/blog/jump-start-your-modeling-with-random-forests/> [d]

1. **Bibliography**

[1] Brian Martucci writes about credit cards (2021) *12 factors that affect your car's resale value*, *Money Crashers*. Available at: https://www.moneycrashers.com/factors-affect-used-cars-resale-value/ (Accessed: October 2, 2022).

[2] Kaggle.com. 2022. *100,000 UK Used Car Data set*. [online] Available at: *<https://www.kaggle.com/datasets/adityadesai13/used-car-dataset-ford-and-mercedes>* [Accessed 28 September 2022].

[3] Baheti, P. (2022) *Data preprocessing in machine learning [steps & techniques]*, *Data Preprocessing in Machine Learning [Steps & Techniques]*. Available at: https://www.v7labs.com/blog/data-preprocessing-guide#h1 (Accessed: October 2, 2022).

[4] *What is normalization in Machine Learning* (2021) *Deepchecks*. Available at: https://deepchecks.com/glossary/normalization-in-machine-learning/ (Accessed: October 4, 2022).

[5] Bhandari, A. (2022) *Feature scaling: Standardization vs normalization*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/ (Accessed: October 20, 2022).

[6] Tingle, M. (2021) *Preventing data leakage in your machine learning model*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/preventing-data-leakage-in-your-machine-learning-model-9ae54b3cd1fb (Accessed: October 16, 2022).

[7] *What is the K-nearest neighbors algorithm?* (no date) *IBM*. Available at: https://www.ibm.com/uk-en/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20algorithm%2C%20also%20known%20as%20KNN%20or,of%20an%20individual%20data%20point. (Accessed: October 17, 2022).

[8] Christophe , J. (no date) *K-Nearest Neighbors (KNN) in Python*, *JC Chouinard*. Available at: https://www.jcchouinard.com/k-nearest-neighbors/ (Accessed: November 23 2022).

[9] Singh, A. (2022) *K-Nearest Neighbors Algorithm: KNN Regression Python*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/ (Accessed: November 7, 2022).

[10] *Euclidean distance formula - derivation, examples* (no date) *Cuemath*. Available at: https://www.cuemath.com/euclidean-distance-formula/ (Accessed: October 19, 2022).

[11] Patel, H.H. and Prajapati, P. (2018) “Study and analysis of decision tree based classification algorithms,” *International Journal of Computer Sciences and Engineering*, 6(10), pp. 74–78. Available at: https://doi.org/10.26438/ijcse/v6i10.7478.

[12] Jauregui, A.F. (2022) *How to code decision tree in python from scratch*, *Ander Fernández*. Available at: https://anderfernandez.com/en/blog/code-decision-tree-python-from-scratch/ (Accessed: November 28, 2022).

[13] Jordan, J. (2022) *Decision trees.*, *Jeremy Jordan*. Jeremy Jordan. Available at: https://www.jeremyjordan.me/decision-trees/ (Accessed: December 2, 2022).

[14] *Decision trees in python - step-by-step implementation* (2022) *AskPython*. Available at: https://www.askpython.com/python/examples/decision-trees (Accessed: December 3, 2022).

[15] James, G. *et al.* (2021) “An Introduction to Statistical Learning with Applications in R,” *Springer Texts in Statistics*, 2, pp. 303–332. Available at: https://doi.org/10.1007/978-1-0716-1418-1\_1.

[16] Zach (2021) *RMSE vs. R-squared: Which metric should you use?*, *Statology*. Available at: https://www.statology.org/rmse-vs-r-squared/ (Accessed: December 3, 2022).

[17] Suji04 (2021) *decision tree regression.ipynb*, *GitHub*. Available at: https://github.com/Suji04/ML\_from\_Scratch/blob/master/decision%20tree%20regression.ipynb (Accessed: November 24, 2022).

[18] Gegic, E. *et al.* (no date) “Car Price Prediction using Machine Learning Techniques,” *TEM Journal*, 8(1), pp. 113–118. Available at: https://doi.org/10.18421/tem.

[a]<https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why#:~:text=Each%20tree%20is%20created%20from,to%20produce%20a%20single%20result>.

[b] <https://www.mdpi.com/2227-9717/9/11/2015>

[c] <https://www.numpyninja.com/post/all-about-random-forests-and-handling-missing-values-in-them#:~:text=Random%20Forests%20and%20Missing%20Data%20%3A&text=This%20forces%20researchers%20who%20want,lost%20and%20inferential%20power%20compromised>.

[d]<https://www.elderresearch.com/blog/jump-start-your-modeling-with-random-forests/>

[e] <https://www.sciencedirect.com/science/article/pii/S0892687512001987>

[f] <https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/>

[g] <https://www.ibm.com/docs/en/ias?topic=procedures-k-nearest-neighbors-knn>

[h] <https://www.ibm.com/docs/en/spss-modeler/18.0.0?topic=nuggets-decision-tree-model-rules>

[i]<https://www.researchgate.net/profile/Mohammad-Bolandraftar/publication/304826093_Application_of_K-nearest_neighbor_KNN_approach_for_predicting_economic_events_theoretical_background/links/5a296efba6fdccfbbf816edf/Application-of-K-nearest-neighbor-KNN-approach-for-predicting-economic-events-theoretical-background.pdf>

[J] <https://link.springer.com/article/10.1007/s10021-005-0054-1> ARTICLE

1. **Appendix**
   1. **Video links**

Decision Tree algorithm: <https://youtu.be/kfBXdJPYvIw>

K-Nearest Neighbour algorithm: <https://youtu.be/3yFoKoO74Ek>

* 1. **My diary**

|  |  |
| --- | --- |
| Date | Tasks |
| 27/10/2022 | - Begun to research into the different types of Machine Learning Algorithms  - looked in K-Nearest Neighbour, how it works and how to implement it |
| 28/10/2022 | - Installed scikit-learn  - Added dataset to repository |
| 01/11/2022 | - Started coding to Jupyter Notebook - so that all data can be visualised/compiled in one place  - Researched the seaborn library to potentially visualise data on informative statistical graphic  - Also Researched Pandas library - to potentially analyse datasets  - split the data set using scikit-Learn  - Created a method to work out the Euclidean distance between two points  - Created a method to work out the KNN |
| 03/11/2022 | - Trying to convert .csv file into a .txt file - because reading the data set file by row has become difficult  - Created a sort method to sort out the distances in ascending order  - Used numpy's genfromtxt instead, but dataset is too large and takes too long to process |
| 04/11/2022 | - Cleaned the data: getting rid of model, transmission and fuel type.  - Broke down dataset into first 200 entries, because dataset is too large to process  - Created method for 1NN to test  - Numpy's genfromtxt doesn't like the engineSize and tax variables - so I have left them out for the time being |
| 05/11/2022 | - Created a prediction method to predict the price of the cars  - Results from prediction method are returning in an unwanted format  - Using Pandas library to read from library - for efficiency and better readings from the dataset  - Created a seaborn heatmap to visualise dataset  - Looked into Seaborn's distplots graph |
| 17/11/2022 | - Began writing the evaluation method for KNN - is currently suited for classification, not regression  - Began work writing interim report. |
| 18/11/2022 | - Began writing the methodology in the interim report |
| 21/11/2022 | - Begun writing description of KNN algorithm to Interim report |
| 23/11/2022 | - Begun coding the decision tree algorithm: imported & split data, method using scikit-learn, entropy method  - Begun writing description of decision algorithm to Interim report  - Attempted to visualise the decision tree |
| 25/11/2022 | - Normalised data for KNN  - Added preprocessing and normalization to Interim report  - Removed unnecessary .txt files |
| 26/11/2022 | - Fixed issued with reading normalised data  - Added code examples to interim report |
| 29/11/2022 | - Completed the evaluation method for KNN, using RMSE |
| 30/11/2022 | - Changed how I reduced data set to increase accuracy by 0.3 at peak  - Used Pandas feature "get\_dummies()" to generate dummy variables for model, transmission and fuel type  - Created a user input to get prediction of their own car |
| 01/12/2022 | - Removed "get\_dummies()" and replaced with "LabelEncoder()"  - Completed user prediction input - fully working |
| 02/12/2022 | - Fixed an issue with prediction, where predictions were completely wrong - fixed by normalising user input |
| 05/12/2022 | - Worked on decision tree algorithm - understanding what each method does  - Calculated how many times Euclidean distance is calculated and how long a single prediction takes  - Described my KNN algorithm code in the interim report |
| 06/12/2022 | - Created Decision tree class  - Initialised values  - Created fit method to put features and label together  - Created start of build tree method  - Created method to find the best split |
| 07/12/2022 | - Created a method to find split tree into branches - extension of best split method  - Using template entropy method to calculate entropy of a node - temporary to test later methods  - Created method to calculate information gain  - Completed best split and tree build methods  - Created method to predict the price of a car - currently not working fully |
| 08/12/2022 | - Fixed predict and best split methods  - Created method to evaluate and calculate the RMSE of the decision tree  - Completed Interim report |