V 

School of Computing

**Honours Project**

**Report**

***Using an Artificial Intelligence to predict used car sales and provide prediction insight***

**Scott Grant**

This report is submitted as part of the requirements for the degree of

**BSc (Hons) in Computing Science**

at Robert Gordon University, Aberdeen, Scotland

**I confirm that the work contained in this Honours project report has been composed solely by myself and has not been accepted in any previous application for a degree. All sources of information have been specifically acknowledged and all verbatim extracts are distinguished by quotation marks**

Table of Contents

[Table of Figures 4](#_Toc164877554)

[Table of Tables 4](#_Toc164877555)

[Acknowledgements 4](#_Toc164877556)

[Abstract 5](#_Toc164877557)

[1. Introduction 6](#_Toc164877558)

[Project Overview 6](#_Toc164877559)

[Aims and Objectives 6](#_Toc164877560)

[Statement 6](#_Toc164877561)

[Sign posting paragraph 6](#_Toc164877562)

[2. Project Scope 7](#_Toc164877563)

[2.1 Literature Review 7](#_Toc164877564)

[2.1.1 Used Car Valuation Review 7](#_Toc164877565)

[2.1.2 Car prediction papers 8](#_Toc164877566)

[2.1.3 Deep Metric Learning papers 11](#_Toc164877567)

[2.1.4 Conclusion 11](#_Toc164877568)

[2.2 Requirements 12](#_Toc164877569)

[2.2.1 Functional Requirements 12](#_Toc164877570)

[2.2.2 Non-Function Requirements 12](#_Toc164877571)

[3. Design / Methodology 14](#_Toc164877572)

[3.1 Methodology 14](#_Toc164877573)

[3.2 Toolkits/Algorithms/Languages 14](#_Toc164877574)

[Dataset 14](#_Toc164877575)

[Siamese Network 14](#_Toc164877576)

[Triplet Loss 15](#_Toc164877577)

[K-Nearest Neighbour 15](#_Toc164877578)

[Overall Model 15](#_Toc164877579)

[Model Evaluation 15](#_Toc164877580)

[Accuracy Evaluation 15](#_Toc164877581)

[Reproducibility 16](#_Toc164877582)

[4. Implementation 16](#_Toc164877583)

[Dataset Pre-Processing 17](#_Toc164877584)

[Importing 17](#_Toc164877585)

[Removing Outliers 19](#_Toc164877586)

[Preparing for machine learning 19](#_Toc164877587)

[Creating the Triplet Network Deep Metric Learner 20](#_Toc164877588)

[Creating the Triplets 20](#_Toc164877589)

[Creating the Siamese Network with Triplet Loss 21](#_Toc164877590)

[K-Nearest Neighbour Models 22](#_Toc164877591)

[kNN Baseline Model 22](#_Toc164877592)

[kNN using embeddings Model 22](#_Toc164877593)

[Random Forrest 22](#_Toc164877594)

[5. Testing and Results 23](#_Toc164877595)

[Results 23](#_Toc164877596)

[Real World Examples 23](#_Toc164877597)

[6. Evaluation 25](#_Toc164877598)

[Functional Requirements 25](#_Toc164877599)

[Artificial Intelligence Implementation: 25](#_Toc164877600)

[Dataset: 26](#_Toc164877601)

[Output: 26](#_Toc164877602)

[Web Application: 26](#_Toc164877603)

[7. Conclusion and Future work 27](#_Toc164877604)

[Legal, social, ethical and professional issues LSEP 27](#_Toc164877605)

[Commercial relevance – straight forwards 27](#_Toc164877606)

[Future Work 27](#_Toc164877607)

[8. Bibliography 28](#_Toc164877608)

[Appendices 31](#_Toc164877609)

[Project Log 31](#_Toc164877610)

[Source Code, Test Plans and other relevant documents 31](#_Toc164877611)

# Table of Figures

[Figure 1 - Screenshot of WeBuyAnyCar's car valuation offer webpage 8](#_Toc164881903)

[Figure 2 - Distribution of the Fuel Type Column 17](#_Toc164881904)

[Figure 3 - Distribution of the transmission column 17](#_Toc164881905)

[Figure 4 - Distribution of car make 18](#_Toc164881906)

[Figure 5 - Distribution car price count 18](#_Toc164881907)

[Figure 6- Distribution of the price range field 20](#_Toc164881908)

[Figure 7 - Example Triplet 21](#_Toc164881909)

[Figure 8 - Siamese Network Summary 21](#_Toc164881910)

[Figure 9 - MAE against differing values of K 22](#_Toc164881911)

[Figure 10 - Car Makes against average MAE 24](#_Toc164881912)

# Table of Tables

[Table 1 - Sample of the combined dataset 16](#_Toc164877549)

[Table 2- Sample of the processed dataset 16](#_Toc164877550)

[Table 3 - Each model's accuracy on the Kaggle Used Car Dataset 23](#_Toc164877551)

[Table 4 - Each model's accuracy on the real-world example dataset 24](#_Toc164877552)

# Acknowledgements

# Abstract

# Chapter 1. Introduction

The project’s main objective was to create a tool using both Deep Metric Learning (DML) and k-Nearest Neighbour (kNN) to predict used car prices from a dataset. The secondary objective was to determine how well this would perform against comparison algorithms such as a baseline kNN trained on the base dataset, and Random Forrest (RF).

### Project Overview

### Aims and Objectives

### Statement

What degree this was produced for:

Wording on moodle on front page

This work is produced by ‘scott grant’ in partial req for bsc chons degree in comp sci

### Sign posting paragraph

Short para stating which section contains what.

# Chapter 2. Project Scope

## 2.1 Literature Review

### 2.1.1 Used Car Valuation Review

Selling used cars is generally a hard metric to gauge as unlike with brand-new cars there is no Manufacture Suggested Retail Price (MSRP) to give the seller an idea of how much the item is worth. This means that private sellers need to balance many different factors which could affect the value, factors such as condition, age, accident history etc. This means there is no real metric to which a seller can easily price their car to.

The selling of used cars has also become increasingly complex in recent years, with prices fluctuating due to factors such as the COVID-19 Pandemic chip shortage (Sillars 2021) or the new Ultra-Low Emission Zone (ULEZ) (Simpson 2023). Historically, there were resources such as the Parkers monthly car valuation book (Adams 2022) or the Kelley Blue Book (KBB) (Kelley Blue Book 2023) in which they publish books which advise readers on what price to expect for used cars. The problem with paper published books is that the books are out of date the moment they are printed, and so cannot keep updated with rapid changing prices. KBB state on their website that the used car valuation they offer is not based on a formula or algorithm instead based on vast knowledge of the industry (Kelley Blue Book 2019). This method of pricing is vague, as with no insight to the actual valuation process it cannot be said that the KBB is an un-bias party, and so prices may be swayed.

In 2021, the UK was recovering from the economic impact of the COVID-19 pandemic lockdown, with the economy growing 7.6% that year (Clark 2023). This recovery helped lead to the number of used car sales exploding like never seen before. The Society of Motor Manufactures and Traders (SMMT) said that “April saw growth of 307.4%, followed by the best May and June since records began.”(SMMT 2021). This explosion in demand lead to prices soaring that year, with the average used car price increasing by 30% over the previous year (Muir 2023). This rise in used car sales and prices could not have been predicted, and so sellers would find gauging used car prices extremely difficult.

New measures being implemented by the UK government have also affected used car sales. The Ultra-Low Emissions Zone in London was rolled out to improve the air quality (Transport for London 2023b) with fines of £12.50 per day to Internal Combustion Engine (ICE) cars that don’t meet the set criteria. Most petrol cars manufactured after 2005 and diesel cars from 2015 are exempt (Transport for London 2023a). BBC News states that ULEZ compliant car prices have increased on AutoTrader by over 40%, they state that the median price beforehand was “£12,989” and afterwards was “£18,295” (Rufo 2023). Factors such as the ULEZ’s or rising fuel prices (Race 2023) could be reasons to why used Battery Electric Vehicles (BEV) sales are up by 81.8% as of Q2 2023 (SMMT 2023).

To make it easier for sellers to value their cars, there are many online free valuation tools available. Some of the most popular of these tools are RAC Cars Online Valuation (RAC Cars 2021), We Buy Any Car (We Buy Any Car 2023), AutoTrader UK (AutoTrader 2023) and Parkers UK (Parkers 2023). These tools work by collecting data about the vehicle such as the car brand, model, model variant, year, and condition, and then the details are processed through an undisclosed algorithm with a price estimation given back to the user.

The main problem with this process is that there is no transparency or justification in their pricing, as these tools are closed source. Many of these online tools offer to buy the car from the user at the end of the valuation process offering a quick process (See Figure 1.).

A screenshot of a computer

Description automatically generated

Figure 1 - Screenshot of WeBuyAnyCar's car valuation offer webpage

This means that these car valuation tools could be affected by alternative motives, such as valuing the car lower on purpose to increase resale profits. My problem would solve this problem by making a free and open-source car price prediction tool for anyone to value cars with. It would allow a fair and transparent estimation process as the goal is to create a highly accurate value estimation tool that also justifies the pricing, so the value is properly understood.

### 2.1.2 Car prediction papers

Due to car price prediction being a complex metric to gauge, there is much previous work done on car prediction models using machine learning. Most of these papers are using regression-based algorithms, such as Linear Regression, Lasso Regression and Random Forrest. Others are based off using the K-Nearest Neighbours algorithm, Neural Networks or Deep Learning.

In (Ashok Kumar and Samruddhi 2020) the K-Nearest Neighbours (KNN) algorithm is applied to a dataset obtained named “The Used Cars dataset” with the goal of analysing used car prices. To compare their results, a Linear Regression algorithm was used as a benchmark. KNN ended up achieving an 85% accuracy, 14% higher than the benchmark set by linear regression.

The conclusion on further work is brief in its suggestions for accuracy improvement however it is stated that “advanced machine learning techniques” would be applied and validating the model with different methods. This however is very vague and would benefit from more specific future goals such as defining what machine learning techniques could be applied to this problem. Techniques such as Deep Metric Learners would be an ideal to use for future as to my knowledge this has not been applied to price prediction before.

The referencing is also poor, with claims such as “Most of the people prefer to buy the used cars because of the affordable price” without any sources. This would infer that the paper is a less reliable source. However, this fact could have been confirmed by a source such as (Armstrong 2022). Here from the Statista Global Consumer Survey about US Consumers when purchasing cars, a low-price rank third overall, only beaten by safety and fuel efficiency.

In (Ganesh and Venkatasubbu 2019) multiple machine learning algorithms are applied to a 2005 dataset made by the Kelly Blue Book that has 805 entries. The goal of the research was to prove the hypothesis of “Multiple and Lasso Regressions are better at

predicting price than the Regression Tree.” by applying these three different supervised learning techniques on predicting used car retail prices and comparing the accuracy of each. The hypothesis was found to be correct, with Mean Error Rates (MER) used to rate the accuracy. Multiple Regression was found to be the best with a MER of 3.48%, while Lasso Regression had a MER of 3.51% and Regression Trees with 3.78%. In the conclusion for future work, it is suggested that algorithms such as Random Forrest should be applied to this problem to improve accuracy, and that to select more data from more recent sources to increase reducibility.

There is no discussion on what data pre-processing was done before the experiment, which could lead to the results of this paper being less accurate and reliable (Maharana, Mondal and Nemade 2022).

In (Varshitha, Jahnavi and Lakshmi 2022) multiple algorithms are applied to a used car dataset from Kaggle. The goal of the paper was to create a model that can predict used car pricing, whilst ensuring that there is no bias to either the owner of the vehicle or buyer. The pre-processing of the dataset was well documented and improves the reproducibility of the experiment. 5 different models were used, Linear Regression, Ridge Regression, Lasso regression, Deep Neural Network and Random Forrest. Deep Neural Networks performed second best, with a Mean Absolute Error (MAE) of 0.766 and a Root Mean Error (RSE) of 0.842. Random Forrest gained the best accuracy rating, with a MAE of 0.746 and RSE of 0.917.

For future work, the paper mentions how the algorithms used to create the used car price prediction tool, could be adapted for use in selling other used goods such as electronics too. It also mentions that by training the model on parts of a larger dataset rather than one small one, it could increase accuracy furthermore.

The research paper (Enoch Li and Bradford Lin 2021) implements two deep learning algorithms and compares them to linear regression on used car price prediction. The goal of the research is to more accurately be able to predict the car prices over existing literature on the subject. The deep learning algorithms used are two Multilayer Perceptrons (MLP), one is a TurnKey solution from sklearn (MLP1) and the other is a customised version of TensorFLow’s Keras framework (MLP2). The dataset used is 16,000 used car data points scraped from the Kelly’s Blue Book website with a 90/10 train test split. Dropping features such as the car colour. As there is no industry standard for naming colours, there could be “Ice White” and “Artic White”. This however could have been standardised in data pre-processing instead of simplify removing the feature by banding similar colours together. By removing this, the model’s performance would be less accurate as it is widely accepted across the industry that the colour of a car can affect the value(We Buy Any Car 2022) (AA 2018).

In their findings, linear regression performed very poorly with a Mean Squared Log Error (MSLE) of 2.61, while both MLP’s performed much better with MLP1 scoring 0.16 and ML2 with 0.08. While these MLP’s outperformed the linear regression algorithm, they were beaten by (Patch 2018) implementation of Ridge Regression with a MSLE of 0.7 on the same dataset. The papers justification for this was Patch had done different dataset pre-processing, with the removal of outliers and the model was tested on a trimmed dataset. It was then speculated that the MLP1 and 2 would have performed better than Patch’s on the same processed dataset. Rather than speculating, it would have been more accurate to implement Patch’s solution into their own and test the results. Patch’s paper is available on GitHub and with all the code used, meaning it is open source and therefore straightforward to implement.

The paper (Enci Liu et al. 2022) uses 3 neural network-based algorithms Back Propagation Neural Network (BPNN), GRA-BPNN (GRA) and PSO-GRA-BPNN (PSO) to predict used car prices on data obtained from China’s largest online used car website. Extensive data pre-processing is used here, with a final dataset of size 10,620. The dataset includes a field for car location, which is a unique but important field as some locations are more desirable to buy from than others (Motorway 2022). By including this, the results of the paper are more likely to have a higher accuracy by taking this into account. PSO obtained the most accurate score of a MSE of 0.48, compared to Random Forrest’s 0.84 which gives a strong argument in the ability of Neural Networks to perform price prediction task better than more traditional algorithms. However, Random Forrest took just 21 seconds to complete the task with PSO taking over 4 times longer at 94 seconds. In the conclusion it’s mentioned that PSO while very accurate also was worse at price prediction for high market cars, meaning more features may need to be added to increase accuracy.

In (Chuyang Jin 2021) 5 regression-based machine learning models are applied to a UK based user car dataset numbering 100,000 entries scraped from websites. However, the paper focuses on specifically the Mercedes part of the dataset numbering 13,120. This way of training a model on just a specific make could help negate the issue raised by (Enci Liu et al. 2022) of high market cars being harder to predict. Sufficient data pre-processing took place, with excellent data visualisation to show the effect this had.

Agreeing with the results of (Enoch Li and Bradford Lin 2021), (Varshitha, Jahnavi and Lakshmi 2022) and (Ashok Kumar and Samruddhi 2020) Linear regression performed the poorest with a R-Squared value of 0.72. This would back up that used car valuation is a complex factor and is not a linear scaling problem.

Of the 5 algorithms, Random Forrest performed the best with a R-Squared value of 0.90 however in the conclusion it is said that this could be improved upon by using a larger dataset or implementing other algorithms such as Naïve Bayes.

The paper (Varshitha, Jahnavi and Lakshmi 2022) was written with the goal of reducing fraud when buying a used car by the implementation of a highly accurate model with no bias to either the owner or the buyer. The algorithms used are a supervised learning artificial neural network based Keras Regression (using 10 layers), Random Forest, Linear, Lasso and Ridge Regression. The dataset used isn’t named, instead just referenced as from Kaggle. This is an acceptable method of referencing, however failing to detail the name of the specific dataset used causes an issue as there is many used car datasets on Kaggle (Kaggle 2023). The paper is a less reliable source as any reproduction of the work is not possible and so the claims and accuracy claimed cannot be verified.

In the results, Random Forest beat the Neural Network’s accuracy with a MAE of 0.75 and R-Squared of 0.92 compared to the MAE of 0.77 and R-Squared of 0.842 (with MAE generally the closer to 0 the better, and with R-Squared generally the closer to 1 the better). However, there is no explanation to why Random Forest is better, and no future improvements suggested that could improve the Neural Network to better predict used car prices. It is suggested that the models could be linked to a live web scraper, intaking current data about used cars. This would be a very useful idea, as the model would consistently be as up to date with prices as possible and increasing the dataset size consistently which would also help to improve the model’s accuracy.

Above, the suggestions made by the papers for future improvements are usually suggesting certain algorithms or frameworks that could be more accurate than current ones. Or by using a larger dataset which could improve performance by allowing the algorithm more references for the price prediction. Some papers mention impartiality and not ensuring no bias however, none of these papers focus on making the reasoning of the used car price prediction open. By adding some insight to the decision process made by the algorithm, the buyers and sellers could understand the features that are impacting the price the most.

To add insights to algorithms such as Random Forest would be near impossible, it works by combining multiple decision trees together (IBM 2023). This means that understanding what features are affecting the price prediction extremely difficult.

The paper (Martin, Wijekoon and Wiratunga 2020) while not related to used car price prediction, proves that DML’s can be very accurate especially when in comparison to algorithms such as KNN. (Ashok Kumar and Samruddhi 2020) found by using KNN in used car prediction an accuracy of 85% was obtained. If DML’s could be used here to further improve this accuracy, which is important as this would help make the tool a more reliable source for price predictions.

### 2.1.3 Deep Metric Learning papers

In the paper (Martin, Wijekoon and Wiratunga 2020) three different Deep Metric Learner (DML) algorithms are used with the goal of finding which is the highest performing in similarity-based return. The algorithms are used across three Human Activity Recognition datasets all of which are based on different activities called SelfBACK, Multi-Model Exercise Data (MEx) and a Physical Activity Monitoring dataset (PAMAP2). KNN was used as a benchmark model in which to compare the three DML algorithms against. Across all datasets, KNN was significantly outperformed by the DML’s across all datasets. The highest difference being in the MEx dataset, KNN scored 68.56% accuracy whilst highest performing DML was Matching Network (MN) using MLP with 94.19% and using Convolutional neural network (CNN) with 95.49%.

The paper (Wang et al. Oct 2017) aims to use Angular Loss (AL) to improve DML’s ability. The DML’s are used across three image-based datasets: a bird species dataset (CUB-200-2011), Stanford Car dataset and an online product-based dataset. There were 7 algorithms used, but the main comparison ones were the AL and N-Pair & Angular Loss (NP&AL). NP&AL was used to show that Angular Loss can also be applied to other frameworks such as N-Pair.

In the discussion about the results, it’s said that AL across all three datasets had increased performance over the baseline, however by applying it to NP&Al the best performance can be achieved across all tested algorithms. For future work it is stated that ideas such as implementing a clustering-based framework could be an improvement to performance.

The paper is well written with mathematical formulas, tables, and figures to help explain the narrative to the reader.

### 2.1.4 Conclusion

According to my research, currently there are no Deep Metric Learner papers on price prediction, and certainly not any on used car price prediction so I feel this subject area is a gap in research. By using DMLs it has potential to outperform the most accurate algorithms in this research area, while also having the ability to share insights about the decision-making process not available by algorithms such as Random Forrest. To the best of my knowledge no car price prediction papers look at insights for the users about the algorithm, therefore this is also a gap in the research that my project aims to resolve.

## 2.2 Requirements

The requirements are for a regressor based problem, in which the requirement prioritisation technique MoSCoW (Agile n.d.) is used to sort requirements by must have, should have, could have and won’t have.

### 2.2.1 Functional Requirements

* Artificial Intelligence Implementation
  + Must be using a Deep Metric Learner based algorithm for price prediction
  + Must use a dataset of used car selling data
  + Must process car input data given and give a price prediction
  + Must be an open-source tool with version control
  + Should manage an accuracy above 80% to ensure tool useability
  + Should be compared against a different machine learning algorithm
  + Could obtain a higher accuracy than the comparison algorithm as this would increase tool viability
  + Could use Random Forrest for comparison as this has shown to be one of the highest performers from the literature review
  + Won’t have any buyer or seller bias implemented as this tool is made to be completely impartial
  + Won’t perform a price prediction for anything other than a private car
* The dataset:
  + Must be populated with used car sales
  + Must pre-process the data to maximise the potential accuracy of the tool
  + Should combine multiple datasets. This could help to remove any bias one source may have.
* The output
  + Must contain a price prediction from the given dataset
  + Must contain at least one performance measure
  + Should use Mean Average Error
  + Could use additional performance error methodologies
  + Could use figures to display the performance differences
* Web Application
  + Could be implemented into a web application
  + Could use the Driver Vehicle and Licencing Agency (DVLA) dataset api to return a price based on a registration number input

### 2.2.2 Non-Function Requirements

* Security
  + No sensitive information will be stored or transmitted
  + Car information such as registration plates are public knowledge
  + The datasets obtained are from public data repositories such as Kaggle
* Capacity
  + System Requirements
    - If made into a web application the project should support modern browsers such as Chromium Based Browsers, Firefox
    - Safari support could be available
    - Mobile phone and tablet browsers will not be supported
    - Must be able to run on a 2020 MacBook Air M1 as that will be the development machine
    - Could also run on a windows-based Intel i5 Desktop as this would be the backup development machine
* Maintainability and Manageability
  + The project will be open sourced throughout the development of the project
  + This means that maintainability will be easier as bug reports, or improvements can be suggested by peers
* Scalability
  + It is only expected for this project there will be no more than one person using the application at a single time as this is a proof of concept
* Useability
  + The web application should be simple and clear which would help to increase accessibility by reducing clutter
* Accuracy
  + The application should be as accurate as possible as this would increase the reliability of the tool

# Chapter 3. Design

## 3.1 Methodology

How I went about the work and development methods

Link to best practice methodologies either waterfall or agile

Think about which one I actually did

Experiment design was, what did you try to find out (goal) and how did you decide what machine algorithm you were gonna implement and what comparison would be

How am I evaulatiing these, and talk through what it means

The project was developed using an Agile methodology (Atlassian 2024), in which I

## 3.2 Data and Toolkits Is this an appropriate title?

### Dataset

The dataset used in this project is a collection of cleaned used car advert data collected through data scraping of UK used car markets (Aditya 2020). The collection is split into 13 CSV files, 9 car manufacturers and the other 4 on specific car models however these were not used as they are duplicates. Each CSV contains 9 features, 7 categorical: ‘model’, ‘year’, ‘transmission’, ‘mileage’, ‘fuelType’, ‘tax’, and ‘engineSize’ and 2 numerical: ‘price’ and ‘mpg’. Together all the sets have 99,187 cars.

### Siamese Network

A neural network is a type of supervised machine learning that works by combining multiple layers of nodes each with an associated weights to classify and cluster data (IBM 2024a). A Siamese neural network works by comparing two or more subnetworks which generate feature vectors for each input (Essam and Valdarrama 2021). The Siamese network used in this project uses Keras’s Siamese network documentation example (Essam and Valdarrama 2021), which has been adapted for use with this dataset instead. I’m unsure how to reference this, ask Kyle.

### Triplet Loss

The Siamese network implements triplet loss to learn the embeddings necessary to cluster the dataset. The loss function uses a Triplet which contains three samples the anchor, positive and negative to compute the embeddings. The anchor is the selected datapoint, a positive is a randomly selected similar datapoint and negative is a random dissimilar datapoint. The triplet loss function is defined in Equation 1 below. Ask Kyle how to reference this formula correctly here, and if it needs to be in the current format then do I need a table of formula too?

Equation 1 - Triplet Loss Formula

To create the Triplets a custom triplet generator method was created, where it will return 3 nested arrays, inside the triplet array. The triplet is created by iterating through the dataset selecting an anchor. The positive is randomly selected from where the ‘make’ and ‘price\_range’ are the same as anchor, while the negative is randomly selected where neither of those conditions apply.

### K-Nearest Neighbour

The KNN algorithm selected was from the Sklearn library and is called ‘KNeighborsClassifier’. KNN works by using a distance function to calculate a particular datapoint’s nearest k number of neighbouring datapoints. The neighbours are then analysed and the class label with the largest number of neighbours is then predicted to the datapoints label. If the k value is 1, then the label of the nearest neighbour is assigned to the datapoint (IBM 2024b; Srivastava 2018).

### Overall Model

The Siamese Network with Triplet Loss is trained on the generated triplets and saves embeddings. The dataset is then embedded, and the KNN model is then trained on this. This model is then what is used for price predictions. Siamese networks work well with classifying similar datapoints, as they base the predictions based upon similarity to other datapoints while KNN labels the data based on the closest neighbour.

These two points are why this these technologies were chosen for this project works well, as generally a car with similar features to another will be priced similarly.

### Model Evaluation

To provide prediction comparison, a baseline KNN and RF models were chosen. KNN was chosen as this will provide a comparison to how well the embedded dataset compares with the original. RF was chosen as in my literature review it was found to be the highest performing algorithm across multiple papers in used car price prediction. The RF model used was selected from the Sklearn ‘ensemble’ library and is called ‘RandomForrestClassifier’. These two models provide a baseline and a strong contender to compare the embedded KNN model against.

### Accuracy Evaluation

To evaluate the accuracy of the models, Mean Absolute Error (MAE) and Real Mean Squared Error (RSME) were used. These methods are used over a percentage of correct predictions as is it is more useful to examine how close each model is to the correct figure. The used car prices are set artificially by sellers, not by a MSRP and so identical cars can have differing prices.

### Reproducibility

This project was developed in a Juypter Notebook file on the online cloud service ‘Google Colab’ using the T4 GPU Runtime. To use the libraries required for this project the following commands must be run at the start of each runtime to upgrade the environment.

Links to the Google Collab file and the GitHub Page are available in Appendix A.

# Chapter 4. Implementation

## 4.1 Dataset Exploration and Pre-Processing

### Dataset Import

When the collection of datasets is imported, the 9 CSV files are combined into a single pandas dataframe, with a new column made for the car manufacturer. The dataframe is then explored, by checking distributions of columns such as fuelType, transmission and make.

Table 1 - Sample of the combined dataset

A screenshot of a graph

Description automatically generated

### Feature Distribution

Many features of the dataframe, were analyzed and plotted to display any imbalances, find outlying data points, and look for mislabeled data.

A graph of fuel type

Description automatically generated

Figure 2 - Distribution of the Fuel Type Column

In Figure 2 there is a clear imbalance in the distribution of ‘fuelType’, however both the two largest ‘Petrol ‘and ‘Diesel’ are very balanced in comparison to one another. There is very little representation for ‘Other’ and ‘Electric’, numbering less than 300.

A graph with blue bars

Description automatically generated

Figure 3 - Distribution of the transmission column

Figure 3 shows that is a balanced feature with many data points for each type of transmission other than the outlying ‘Other’ type.

A graph of blue bars with white text

Description automatically generated

Figure 4 - Distribution of car make

Figure 4 shows the unbalanced distribution of the make of car, with the largest brand having almost triple the representation of the smallest. Even with this difference, there is still a enough data points for a good representation for each make.

A graph of a car price

Description automatically generated

Figure 5 - Scatterplot of Price against count

Figure 5 shows the price distribution throughout the dataset, with the highest numbers of occurrences under £20,000. The average prices are a mean of £16,805 and a median of £14,495 which shows that the price distribution is positively skewed.

### Removing Outliers

Cars without enough representation are seen as outliers as there is such a large disparity between them and the other types in that feature that it would hamper results. Cars that met the following conditions were dropped from the dataset:

* A ‘fuelType’ of ‘Electric’ or ‘Other’,
* An ‘engineSize’ that occurs fewer than 50 times.
* Transmission value of ‘Other’
* Priced either under £2,000 or over £60,000
* Over 100,000 miles
* Over 20 years old
* Models with less than 10 datapoints

In total 2184 cars were removed from the dataset.

### Conversion into a Categorical Problem

To change the task of price predication from a regression-based to categorical, a ‘price\_range’ was created to replace the ‘price’ field. Every price was rounded to the nearest £500, as this would reduce the number of price points to 120 possible labels, while trying to retain as close to the original price point as possible.

A graph with blue and black lines

Description automatically generated

Figure 6- Distribution of the price range field

After conversion into the ‘price\_range’ field, figure 6 shows the new distribution. By removing the outlying price points, the overall distribution is now less positively skewed, and closer to a normal distribution. The averages remain similar with a mean average of £16,629 and a median of £14,500.

## Preparation for machine learning

The conversion of string features into numerical representations needed to be done for the machine learning models. This is done as machine learning models are designed to handle numerical data. Mapping Dictionaries were used to do this by relating each string value to a number. This method was applied to convert the ‘transmission’, ‘make’, ‘fuelType’ and ‘model’ fields.

The ‘year’ field while containing numerical data, is represented categorically. This was converted to an age column instead to better display the relationship of price to year. This then will more clearly display the car’s age as a range of 20 different categorical values. A mislabelled car ended up with a negative age, and so this was dropped from the dataset. Figure 2 shows a sample of the dataset after this processing.

Table 2- Sample of the processed dataset

A screenshot of a graph

Description automatically generated

The final preparation for machine learning was to split the ‘price\_range’ from dataset and then split the sets again into a training and testing sets. This was done using a test size of 20% and a random state of 42 to keep the dataset split the same for each test run.

## Creating the Triplet Network Deep Metric Learner

Based of Keras’s Siamese Network with Triplet Loss

### Creating the Triplets

* A triplet consists of three components
  + Anchor is the selected data point
  + Positive is a similar datapoint
  + Negative is a random dissimilar datapoint
* The Logic chosen to do this is:
  + The positive sample is a randomly chosen car of the same make and price\_range as the anchor
  + The negative is randomly selected from a different make and price range
  + Example below

A white background with black text

Description automatically generated

Figure 7 - Example Triplet

* The triplets are then generated based on X and Y train

### Creating the Siamese Network with Triplet Loss

* Distance Function
  + Calculates Euclidean distances between the anchor embedding and the positive embedding, as well as the anchor embedding and the negative embedding.
* Layers
  + As shown below

A screenshot of a computer

Description automatically generated

Figure 8 - Siamese Network Summary

* Model
  + Computes loss based on the triplets
  + Uses Adam optimizer with a value of 0.0001
  + 50 epochs used, loss flatlines after this.
* Add in PCA graph’s with a description

## K-Nearest Neighbour Models

### kNN Baseline Model

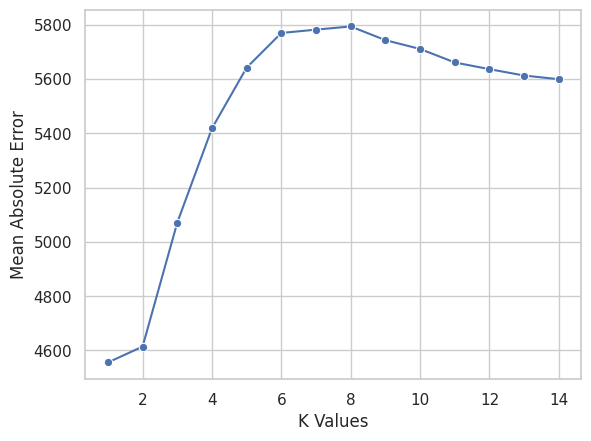


Figure 9 - MAE against differing values of K

* Sklearn kNN used
* Trained on X and Y train
* K values of 1 – 15 tested for highest performance
* K value of 1 best for both MAE and RMSE
* Accuracy of 7% however this is useless (explain why MAE and RSME are better metrics instead)

### kNN using embeddings Model

* Same as baseline kNN however X train and X test are embedded using the DML
* Ends up being generally 2 times as accurate

### Random Forrest

* Shown as best model in studies for used car price prediction
* Sklearn Random Forrest Classifier used

# 5. Testing and Results

* How do you know that what you did
  + Works
  + Is it fit for purpose

Marking Grid: Comprehensive evaluation methods, appropriate to the project's goals and type, are employed strategically. These may range from software quality assurance methods like focus groups, user testing or black box testing to empirical experiments for theory validation. Results are meticulously analysed and used to inform suggestions for future work. The conclusion offers deep analysis of outcomes against initial objectives and a reflection on lessons learned.

To test the differing models, MAE and RSME were taken as evaluation metrics. This is because as it is a used market I am evaluating on, prices even for the exact same car could differ, and so rather than checking how often it is right, a better evaluation metric would be to check how close it is to being correct. By using MAE and RSME, an acceptable margin can be made. i.e. this car is £22,000 (+- £1500) or state a range to the user (£21,500 – £23,500).

### Results

Table 3 - Each model's accuracy on the Kaggle Used Car Dataset

|  |  |  |
| --- | --- | --- |
| Model | MAE | RSME |
| Baseline kNN | 4572 | 6714 |
| Embedded kNN | 2543 | 4240 |
| Random Forrest | 1282 | 2035 |

### Real World Examples

* Important to note the car market will have changed since the dataset was created in 2020, and so it is likely all models will be inaccurate here
* 10 cars were manually taken off auto trader by me and implemented into a dataset to test each model on.
* Dataset has high categorical columns which could be why RF is better
  + Multiple dec trees algorithm
  + Very good at cat data as it is a branching path
  + That allows it to learn various types of
  + Not so good on numerical data
* My DML
  + May not of handled one of the cat the right way
  + Embedding may not have learned the relationships between the categories and the outcomes
  + Highlight the triplet didn’t outperform the rf but always outperms the base knn. It is optimised representation however it needs to improve how it handles cat variables

Table 4 - Each model's accuracy on the real-world example dataset

|  |  |  |
| --- | --- | --- |
| Model | MAE | RSME |
| Baseline kNN | 6000 | 6629 |
| Embedded kNN | 4667 | 5593 |
| Random Forrest | 3000 | 3431 |

A graph with different colored bars

Description automatically generated

Figure 10 - Car Makes against average MAE

Worse for prestige car dealers overall (a contributing factor to this is size of the makes and that cars above £60k were dropped)

Good for the average consumer

This is reflected in the positive skewed distribution – much more data in the lower end market

Relate this back to figure 4 also, with some brands being more represented than others

If there is time, could also graph MAE against grouped mileage (every 5 or 10k or so)

# 6. Evaluation

## Functional Requirements

Overall, 14/20 (70%) of the functional requirement objectives were met.

|  |  |  |
| --- | --- | --- |
| Artificial Intelligence Implementation: | | |
| Requirement | Was this met: | Comments: |
| Must be using a Deep Metric Learner based algorithm for price prediction | Yes | Implemented Keras’s Siamese Network with Triplet Loss. |
| Must use a dataset of used car selling data | Yes | Dataset is obtained from Kaggle and contains just under 100,000 records of UK Used Car Adverts data scraped. |
| Must process car input data given and give a price prediction | Yes | The real-world data section of the results shows this. |
| Must be an open-sourced tool with version control | Yes | Code has been stored on GitHub and will be made public upon the completion of my honours project. |
| Should manage an accuracy above 80% to ensure tool useability | No | Refer to **Section 5. Testing and Results,** this explains why using an accuracy percentage is not an optimal solution for this problem. |
| Should be compared against a different machine learning algorithm | Yes | The embedded kNN is compared against a baseline kNN model as well as a RF model. |
| Could obtain a higher accuracy than the comparison algorithm as this would increase tool viability | No | The embedded kNN outperforms the baseline kNN however RF is still the highest performing. |
| Could use Random Forrest for comparison as this has shown to be one of the highest performers from the literature review | Yes | RF was implemented as a comparison algorithm. |
| Won’t have any buyer or seller bias implemented as this tool is made to be completely impartial | No | No bias was intentionally implemented however by using a dataset of used car adverts this may negatively skew the price towards the seller. If records of actual car sale prices could be obtained, then this project could be made more impartial to both parties. |
| Won’t perform a price prediction for anything other than a private car | Yes | The dataset used only contains used car advert data. |
| Dataset: | | |
| Must populated with used car sales | Yes | - |
| Must pre-process the data to maximise the potential accuracy of the tool | Yes | Pre-processing was done and is detailed in the implementation. |
| Should combine multiple datasets. This could help to remove any bias that a single source may have. | No | There were no datasets combined in this project however, the project was applied to the Kaggle dataset, as well as the real-world example dataset. This is planned for future work. |
| Output: | | |
| Must contain a price predication from the given dataset | Yes | An example price prediction can be seen in Figure (add in AUDI PREDICTED PRICES). |
| Must contain at least one performance measure | Yes | The performance measures used are MAE and RSME. |
| Should use MAE | Yes | - |
| Could use additional performance error methodologies | Yes | RSME was used. |
| Could use figures to display the performance differences | Yes | Figure 10 displays how MAE changes with make/ |
| Web Application: | | |
| Could be implemented into a web application | No | The implementation of the tool into a website was out of the project’s scope, and so is under future work. |
| Could use the Driver Vehicle and Licencing Agency (DVLA) dataset API to return a price based on registration number input | No | There is no student licensing for the API, and so implementation would require financial backing. This is also under future work. |

Gibbs reflective cycle

Methodology to formalise reflection

Reflection – eveolving your practice based on your experience

This gives a recipe for doing this

6 steps

Description

Looking back I developed this to ...

Onenthing I looked at but didn’t use ..

In the end I am disappointed not sued.

Looking at it I have good reason to believe..

Explain

In conc I feel they would have been valuable to the project

# 7. Conclusion and Future work

Reflect!

Legal, social, ethical and professional issues

Commercial relevance Quality of your engineered solution

Your Planning and Achievement

### Legal, social, ethical and professional issues LSEP

Think about is there any of these issues of what I’ve developed

Ethically/Social could contribute to good practice

Professionally you have a responsibility that people are clear on capabilities (provide mae and whatnot (plus or minus)

Not legal ramifications

Can be bulletpointed list

### Commercial relevance – straight forwards

conclude the reflection

### Future Work

Test triplet generator with and without selecting make as this could impact results

No bias was intentionally implemented however by using a dataset of used car adverts this may negatively skew the price towards the seller. If records of actual car sale prices could be obtained, then this project could be made more impartial to both parties.

Should combine multiple datasets. This could help to remove any bias that a single source may have.

Changing fields such as price range to every 1k or every 250 instead

Being more brutal with removing outliers

The implementation of the tool into a website was out with the project’s scope, and so is under future work.

There is no student licensing for the API, and so implementation would require financial backing. This is under future work

# 8. Bibliography

References

AA, 2018*. Car colour - does it matter when buying a car? | AA.* [online]. Available from: <https://www.theaa.com/car-buying/car-colour> [Accessed Oct 21, 2023].

ADAMS, K., 2022*. Parkers - About us.* [online]. Available from: <https://www.parkers.co.uk/about-us/> [Accessed Oct 14, 2023].

ADITYA, 2020*. 100,000 UK Used Car Data set.* [online]. Available from: <https://www.kaggle.com/datasets/adityadesai13/used-car-dataset-ford-and-mercedes/data?select=cclass.csv> [Accessed 14 Oct 2023].

AGILE, n.d. *Chapter 10: MoSCoW Prioritisation.* [online]. Available from: <https://www.agilebusiness.org/dsdm-project-framework/moscow-prioririsation.html#:~:text=MoSCoW%20(Must%20Have%2C%20Should%20Have,useful%20in%20many%20other%20areas.>

ARMSTRONG, M., 2022*. Infographic: Most Important Factors When Buying a Car.* [online]. Available from: <https://www.statista.com/chart/13075/most-important-factors-when-buying-a-car> [Accessed Oct 24, 2023].

ASHOK KUMAR, R. and SAMRUDDHI, K., 2020. Used Car Price Prediction using K-Nearest Neighbor Based Model. *IJIRASE,*

ATLASSIAN, 2024*. What is Agile?* [online]. Available from: <https://www.atlassian.com/agile> [Accessed Apr 24, 2024].

AUTOTRADER, 2023*. Free Car Valuation Auto Trader UK.* [online]. AutoTrader UK. Available from: <https://www.autotrader.co.uk/cars/valuation> [Accessed Oct 12, 2023].

CHUYANG JIN, 2021. Price Prediction of Used Cars Using Machine Learning. *IEEE,*

CLARK, D., 2023*. Gross domestic product of the United Kingdom from 1948 to 2022.* [online]. Available from: <https://www.statista.com/statistics/281744/gdp-of-the-united-kingdom/> [Accessed Oct 14, 2023].

ENCI LIU et al., 2022. Research on the Prediction Model of the Used Car Price in View  
of the PSO-GRA-BP Neural Network. *Sustainability,* 14

ENOCH LI and BRADFORD LIN, 2021. Predicting used car prices with deep learning. *Standford University,*

ESSAM, H. and VALDARRAMA, S.L., 2021*. Keras documentation: Image similarity estimation using a Siamese Network with a triplet loss.* [online]. Available from: <https://keras.io/examples/vision/siamese_network/> [Accessed Apr 24, 2024].

GANESH, M. and VENKATASUBBU, P., 2019. Used Cars Price Prediction using Supervised Learning Techniques. *International Journal of Engineering and Advanced Technology,* 9, pp. 216-223.

IBM, 2023*. What is Random Forest? | IBM.* [online]. Available from: <https://www.ibm.com/topics/random-forest> [Accessed Oct 22, 2023].

IBM, 2024a*. What is a Neural Network?* [online]. Available from: <https://www.ibm.com/topics/neural-networks> [Accessed Apr 24, 2024].

IBM, 2024b*. What is the k-nearest neighbors algorithm? | IBM.* [online]. Available from: <https://www.ibm.com/topics/knn> [Accessed Apr 24, 2024].

KAGGLE, 2023*. Kaggle Search: Used Car.* [online]. Available from: <https://www.kaggle.com/search?q=used+car>

KELLEY BLUE BOOK, 2019*. What are Kelley Blue Book Values?* [online]. Available from: <https://www.kbb.com/car-advice/what-are-kelley-blue-book-values/> [Accessed Oct 21, 2023].

KELLEY BLUE BOOK, 2023*. Kelley Blue Book | New and Used Car Price Values, Expert Car Reviews.* [online]. Available from: <https://www.kbb.com/> [Accessed Oct 21, 2023].

MAHARANA, K., MONDAL, S. and NEMADE, B., 2022. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings,* 3(1), pp. 91-99.

MARTIN, K., WIJEKOON, A. and WIRATUNGA, N., 2020. Human activity recognition with deep metric learners. 2567

MOTORWAY, 2022*. What can impact the valuation and price of my vehicle?* [online]. Available from: <https://help.motorway.co.uk/hc/en-gb/articles/4406711501714-What-can-impact-the-valuation-and-price-of-my-vehicle-> [Accessed Oct 21, 2023].

MUIR, J., 2023*. Our analysis of used car prices.* [online]. Available from: <https://www.autotraderinsight-blog.co.uk/auto-trader-insight-blog/used-car-prices-analysis> [Accessed Oct 14, 2023].

PARKERS, 2023*. Free Car Valuation Parkers.* [online]. Available from: <https://www.parkers.co.uk/car-valuation/> [Accessed Oct 14, 2023].

PATCH, T., 2018*. Kelley Blue Book Project.* [online]. Available from: <https://tai-pach.github.io/kbb/> [Accessed Oct 21, 2023].

RAC CARS, 2021*. Free Car Valuations, How Much is My Car Worth? .* [online]. Walsall, England: RAC Cars. Available from: <https://www.raccars.co.uk/value-my-car> [Accessed Oct 12, 2023].

RACE, M., 2023*. Petrol prices rise for fourth month in a row - RAC.* [online]. Available from: <https://www.bbc.co.uk/news/business-66997316> [Accessed 14 Oct 2023].

RUFO, Y., 2023*. Ulez expansion: Contested claims examined.* [online]. Available from: <https://www.bbc.co.uk/news/uk-england-london-64798395>

SILLARS, J., 2021*. COVID-19: New model shortages drive near-record demand for used cars.* [online]. Available from: <https://news.sky.com/story/covid-19-new-model-shortages-drive-near-record-demand-for-used-cars-12377702> [Accessed Oct 14, 2023].

SIMPSON, J., 2023*. Second-hand car prices up 45pc thanks to Ulez expansion.* [online]. Available from: <https://www.telegraph.co.uk/news/2023/06/28/second-hand-car-prices-up-45pc-thanks-to-ulez-bubble/> [Accessed Oct 14, 2023].

SMMT, 2021*. Used car sales: Q2 2021.* [online]. Available from: <https://www.smmt.co.uk/2021/08/used-car-sales-q2-2021/> [Accessed Oct 14, 2023].

SMMT, 2023*. Used car sales: Q2 2023.* [online]. Available from: <https://www.smmt.co.uk/2023/08/used-car-sales-q2-2023/> [Accessed Oct 14, 2023].

SRIVASTAVA, T., *A Complete Guide to K-Nearest Neighbors (Updated 2024).*

TRANSPORT FOR LONDON, 2023a*. ULEZ Information - Cars.* [online]. Available from: <https://www.tfl.gov.uk/modes/driving/ultra-low-emission-zone/cars> [Accessed Oct 14, 2023].

TRANSPORT FOR LONDON, 2023b*. Ultra Low Emission Zone.* [online]. Available from: <https://www.tfl.gov.uk/modes/driving/ultra-low-emission-zone> [Accessed Oct 14, 2023].

VARSHITHA, J., JAHNAVI, K. and LAKSHMI, C., 2022. Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning. *IEEE,*

WANG, J. et al. , Oct 2017. Deep metric learning with angular loss. Oct 2017. ICCV, pp. 2593-2601.

WE BUY ANY CAR, 2022*. Which car colours have the best resale value? | Webuyanycar.* [online]. Available from: <https://www.webuyanycar.com/guides/selling-and-buying-cars/car-colour-value/> [Accessed Oct 21, 2023].

WE BUY ANY CAR, 2023*. We Buy Any Car Free Car Valuation.* [online]. We Buy Any Car. Available from: <https://www.webuyanycar.com/car-valuation/> [Accessed Oct 12, 2023].

# Appendices

## Appendix A – Project Links

Google Colab:

https://drive.google.com/file/d/1O-BxKoA9cwrcvEQ7CAcYGLvFJ0kmPKR9/view?usp=sharing

### UPDATE THIS LINK WITH UPDATED VERSION AFTER COLAB IS CLEANED UP!!!

GitHub: https://github.com/ScottGrant528/honours

## Appendix B: Project Log

### Source Code, Test Plans and other relevant documents