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

November 17, 2022

**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 Where other published and unpublished source materials have been used, these have been acknowledged.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:

Student Name: Rizwan Bagdadi

Date of Submission:

Signature: Rizwan Bagdadi

**Contents**

[**Abstract** 3](#_Toc119718768)

[**Project Specification** 4](#_Toc119718769)

[**1.** **Introduction** 5](#_Toc119718770)

[**2.** **Methodology** 6](#_Toc119718771)

[**3.** **Implementation** 6](#_Toc119718772)

# 

# **Abstract**

In this paper, I am investigating 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 use ML algorithms to find the most suitable algorithm able to carry out the most accurate predictions, as well as providing the best performance and overall efficiency, by comparing their benchmark data. Two techniques, K-nearest neighbour and decision trees, have been used 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 have first discussed why it is I have chosen to research this area, discussing the problems, the client and data. I have then extensively described and explained how each algorithm is effectively used, discussing the drawbacks and benefits of each.

# **Project Specification**

Enter here………

1. **Introduction**

Predicting the price of a pre-owned car is an important problem that not only benefits the buyer, but also the seller. Individuals or dealerships often state their buying price at very unreasonable prices - in comparison 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 two main problems that must be highlighted; the first being human error. Figure 1 shows a list designed by Gordon Dupont called the “dirty dozen” which highlights 12 of the most common human errors from which several are applicable when it comes to selling pre-owned cars, such as stress, lack of knowledge, lack of awareness, lack of communication and lack of resources. The second problem being discrimination; this includes race, gender, age, and religion. From the seller's perspective, they may not know the worth of the vehicle they are selling, leaving them vulnerable, this commonly occurs with elderlies trying to sell their cars, therefore the seller is discriminating against their age. From the buyer's perspective, they could be quoted an unfair price due to the seller having prejudice against certain groups.

Text

Description automatically generated

There are several distinct features that must be considered and examined to determine a prediction that is both accurate and reliable. Some of these factors consist of mileage, engine size, condition, registration year, transmission type, fuel type and fuel economy [2]. The number of features that are needed to be taken into consideration will result in a time-consuming process and therefore will be inefficient for the estimation to be calculated manually.

Having a system that uses a dataset of hundreds of thousands of pre-owned cars selling prices, to predict the selling price of a used vehicle, would benefit both sellers and buyers. However, this system should not be user controlled, due to unfair and/or unreliable results. Hence the reason I believe Machine Learning (ML) should be implemented; to eliminate any bias and provide a fair, accurate selling price.  However, the system will use supervised algorithms, and will need to be trained by a reliable data set. ML is useful to use when it comes to making accurate predictions, as it has the ability to learn from massive amounts of data and can continue to learn. Advanced ML can also predict stock market trends - which would be extremely useful when it comes to stating the selling prices of cars no matter what the state of the market is.

The goal of this project is to use ML algorithms to find the most suitable algorithm able to carry out the most accurate predictions, as well as providing the best performance and overall efficiency, by comparing their benchmark data. With this project I will be applying at least three different supervised machine learning algorithms; consisting of logistic regression, k-nearest neighbour, and decision trees. To analyse these algorithms, I will perform a regression analysis and cross-validation of each algorithm used, in order to find which provides the best performance. To train the algorithms, I am going to use a dataset, last updated in 2020, that has the selling price and features of over 100,000 pre-owned cars [3].

1. **Methodology**

Data was collected from Kaggle [ ] by user @Aditya, and was last updated in 2020. When looking for an appropriate dataset, I made sure that all the data was well organised, containing information of price, mileage, road tax, miles per gallon (mpg) and engine size, and ensured there were no duplicate listings that would affect my results. I also wanted the data to contain transmission, fuel type and the car model, as these are huge factors when considering the selling price and I would like to implement these in the prediction phase in the future. I will implement these key features by replacing them with integer e.g. petrol = 0, diesel = 1. The dataset is split into car brands, each containing thousands of samples. As the data was too large for my laptop to process, I had reduced the sample size to 200, but I am looking to solve this issue in the future, using an virtual machine. A sample of the data used is shown below in Table 1.

**Table 1.** *Sample Data Used*

\*INPUT SS OF PART OF DATA SET\*

1. **Implementation**

The data I used has been split into training (90%) and testing (10%) subsets