

**JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY**

**COLLEGE OF PURE AND APPLIED SCIENCES**

**SCHOOL OF COMPUTING AND INFORMATION TECHNOLOGY**

**DEPARTMENT OF COMPUTING**

**BACHELOR OF SCIENCE COMPUTER TECHNOLOGY**

**COMPUTER SYSTEMS PROJECT DOCUMENTATION**

**USED CARS PREDICTOR USING MACHINE LEARNING**

**AUTHOUR: MARTIN MWARANO**

**REG NO: SCT212-0216/2018**

**SUPERVISOR: EUNICE NJERI (PH. D)**

# ABSTRACT

Predicting the value of used cars is a hotly debated subject because of the extraordinary volume of vehicles being bought and sold. People tend to buy used automobiles more frequently in developing nations since they are more affordable. This project's main goal is to calculate used car costs using attributes that have a strong correlation with a price. Data mining technology has been used to achieve this. During pre-processing, missing, redundant, and null values were eliminated from the dataset. The value of an automobile can be evaluated more precisely by using Machine Learning algorithms to better utilize data on all the uncommon qualities of a car.

In this project, I anticipate that in the not-too-distant future, the most complex algorithm is used for producing predictions, and then the model is integrated into a mobile application or website for usage by the general public.

# INTRODUCTION

## Motivation

The retail price of a new car, excluding extra equipment, is the same for all vehicles of the same make, model, and year. The producer determines this cost.

However, the price of used cars is determined by supply and demand. Additionally, there are other characteristics of secondhand cars that affect the price. These factors distinguish cars that may have had a similar retail price, such as their condition, mileage, and repair history. Due to the delicate nature of the manual automobile pricing process, car dealers and manufacturers run the risk of undervaluing good cars while simultaneously taking advantage of buyers who are less knowledgeable about used cars prices,

Numerous tools offer an estimate but do not take the specifics of the car into account when calculating the price. Because car markets are somewhat regional, location also influences price. Therefore, a valuation approach that can gather data from all prior sales of cars with similar features and utilize more of the unique qualities of each car is required.

In summary, there is a high demand for a system that can accurately predict the prices of used cars, as many people prefer to buy pre-owned vehicles due to affordability or economic conditions. However, accurately predicting used car prices can be difficult due to the number of factors that can affect the price, such as the car's age, fuel type, condition, mileage, horsepower, and more. Building an intelligent system to predict used car prices efficiently requires a comprehensive dataset that includes all of these relevant features. The main challenge in creating such a system is collecting this dataset, as it can be difficult to obtain all of the necessary information about a car's features.

Before using the collected data for the used car price prediction model, it was necessary to pre-process and transform it into the appropriate format. The dataset was first analyzed and plotted, and any missing, duplicate, or null values were identified and addressed. Relevant features were selected using correlation matrices and others were discarded. Since this was a supervised learning problem, it was treated as a regression problem and three different regressor models (random forest, linear regression, and bagging regression) were trained and compared. The random forest regressor model performed the best and was therefore chosen as the primary algorithm model for the project.

## Social Impact

Nowadays, almost everyone wants their own automobile, but many people choose to buy used cars due to issues with price or the state of the economy. Because used car prices depend on so many different features and conditions, it takes expertise to anticipate them accurately. Prices for used cars fluctuate on the market, therefore both buyers and sellers require an intelligence system to accurately anticipate the price.

## Technology

The technology to be used in this project is Machine Learning

A machine learning model for a used car price predictor would be trained on a dataset of used car sales, including the various features of the cars (e.g., make, model, age, mileage, etc.) and their corresponding sale prices. The model would learn to identify patterns in the data that relate the car's features to its sale price, in order to make predictions about the sale price of new cars that it has not seen before.

To build such a model, one would first need to collect and clean a large dataset of used car sales. This dataset would need to include a variety of features that might affect the sale price of a used car, such as the make and model of the car, the age of the car, the mileage, the condition of the car, and any additional features or accessories that it has.

Once the dataset is prepared, one would then need to select a machine learning algorithm to build the model. There are many different algorithms that can be used for this task, including linear regression, decision trees, and artificial neural networks. Each of these algorithms has its own strengths and weaknesses, and the choice of which algorithm to use will depend on the specific characteristics of the data and the requirements of the problem.

Once the model has been trained on the dataset, it can be used to make predictions about the sale price of new used cars by inputting the features of the car into the model and receiving a predicted sale price as output. The model can then be evaluated on its performance, and if necessary, fine-tuned to improve its accuracy.

## Related Works

Machine learning has been applied in a wide range of research areas, and has been used to solve a variety of problems. Here are a few examples of areas where machine learning has been applied, along with some references:

* Health care: Machine learning has been used to predict patient outcomes, diagnose diseases, and identify patterns in electronic health records that could inform treatment decisions. For example, a study published in the Journal of the American Medical Association used machine learning to predict in-hospital mortality for patients with sepsis (<https://jamanetwork.com/journals/jama/article-abstract/2734013>).
* Finance: Machine learning has been used to predict stock prices, detect fraudulent transactions, and analyze market trends. For example, a study published in the Journal of Financial Economics used machine learning to predict stock prices using news articles (<https://www.sciencedirect.com/science/article/pii/S0304405X14000390>)

## Problem Statement

Develop a machine learning model that can accurately predict the sale price of a used car given its characteristics and features."

To create a solution for this problem, one would need to gather a dataset of used car sales, including information about the various features of the cars (e.g., make, model, age, mileage, etc.) and their corresponding sale prices. The model would then be trained on this dataset, learning to identify patterns in the data that relate the car's features to its sale price. Once trained, the model would be able to make predictions about the sale price of new used cars by inputting the features of the car into the model and receiving a predicted sale price as output.

## Proposed Approach

Project aims at building a model of used car prices to predict car prices using a provided dataset. The model should learn from the data and be able to predict the value of a used car given all the metrics. The metrics include, year of model of the car, mileage, transmission, dealer type and fuel type.

The worth of a car can be predicted with more accuracy by using machine learning to better utilize data on all the uncommon qualities of a car. Customers would clearly profit from this, especially those who must rely on a tool since they are unable to determine the value of the vehicle they are buying or selling on their own. A more specialized tool can offer a more accurate price and make the market more equitable for all players by taking into account the non-standard features of the car.

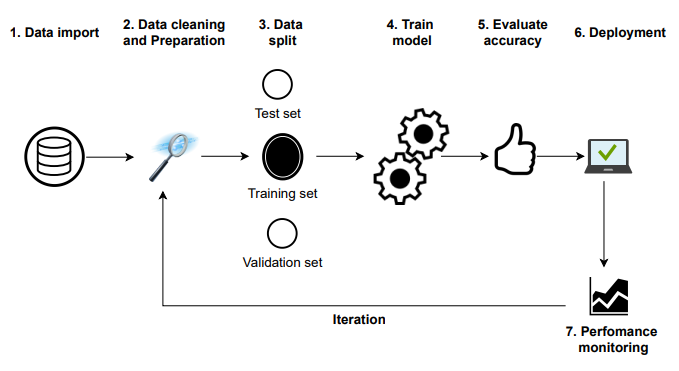


Figure 1: visualization of solution

1. Data import: dataset is downloaded from Kaggle.com as a .csv file
2. Data cleaning and preparation: at this stage, it involves cleaning the data to handle any missing values.
3. Data splitting: this is the third stage which involves separating the data into two sets one for training the model and the other for testing the model.
4. Training the model: which is the fourth stage, the models are trained using the 2 chosen algorithms which are Random Forest and Linear regression
5. Evaluating the accuracy: at this stage the 2 models are compared using the r2\_score and the one with a higher percentage is the most accurate.
6. Deployment: at this stage the more accurate model is used in creating a web application in which the user will be able to input various car features and the system will predict the price by a click of a button
7. Performance monitoring: the last stage in which I will be checking how the system performs given different inputs.

## Main Objective

Below are the objectives of using machine learning in this project:

1. Building a model that can accurately predict the sale price of a used car given its characteristics and features: The ultimate goal of the project is to develop a model that can make reliable predictions about the sale price of used cars based on their features.
2. Improving the efficiency of the model: The model should be efficient in terms of both the amount of data and computing resources it requires, as well as the time it takes to make a prediction.
3. Ensuring the model is easy to use and deploy: The model should be easy to use and integrate into an existing system or workflow, without requiring extensive technical expertise.
4. Ensuring the model is scalable: The model should be able to handle a large volume of data and make predictions quickly, even as the volume of data increases over time.
5. Maintaining the model's accuracy over time: The model should be able to maintain its accuracy over time, even as the underlying data and patterns in the market change.

# STATE OF THE ART

Used automobile prices have been predicted using a variety of studies and related works utilizing various methodology and approaches, with accuracy ranging from 50% to 90%.

The results of the study showed that the accuracy of the used car price prediction model ranged from 60-70%. The author recommended using more advanced models and algorithms in order to improve the evaluation, as the decision tree and naïve Bayes models had a weakness in that they required discretizing and classifying the prices, which led to inaccuracies. Additionally, the author suggested using a larger dataset to train the models, as the current dataset was not sufficient.

## Review of Related Works

A study conducted by researchers at the International Burch University in Sarajevo (Gegic, Isakovic, Keco, Masetic, & Kevric, 2019) used three different machine learning techniques (Support Vector Machine, Random Forest, and Artificial Neural Network) to predict used car prices. The dataset for the study included 797 used car listings scraped from a Bosnian website, and the results showed that using a single machine learning algorithm had an accuracy of less than 50%. However, when the algorithms were combined and the prices were pre-calculated using the Random Forest model, the accuracy increased to 87.38%.

In (Pudaruth, 2014), the researcher made the suggestion to forecast used car prices in Mauritius. He used the following to get his results: decision trees, K-nearest neighbors, multiple regression, and naive bayes algorithms, which he based on historical data gathered from the newspaper. Accuracy levels reached varied from 60 to 70%.

In a study by Nabarun Pal (2018), the Random Forest supervised learning method was used to predict used car prices using a dataset from Kaggle. The researchers performed exploratory data analysis to identify the impact of each feature on the price and trained 500 Decision Trees with Random Forests. Although this method is typically used for classification, the researchers transformed the problem into a regression problem in order to predict the car prices. The results of the experiment showed that the training accuracy was 95.82% and the testing accuracy was 83.63%. The model was able to accurately predict the car prices by selecting the most correlated features.

Based on the review of previous research, it can be concluded that predicting used car prices is a current area of focus for many researchers. The highest reported accuracy so far is 83.63% on the Kaggle dataset using the random forest technique.

For this project, two models will be trained and tested, which are Linear regression and Random Forest Regressor. I will compare them and evaluate their performance.

# SYSTEM ANALYSIS AND DESIGN

## SYSTEM ANALYSIS

### Feasibility Study

1. Data availability: the dataset that will be used in the following project is from Kaggle.com. It has about 4346 records in which 8 columns contain information about the car. It’s a very large dataset. in the dataset it includes data on the make, model, its year of make, mileage, body type, registration and Engine type.
2. Algorithm Complexity: The machine learning algorithms to be used in this project are Linear Regression and Random Forest Regressor. For these algorithms there are many resources on how to train and use them to get desired results.
3. Resource Requirements: There is enough computing power and time to train and deploy the machine learning model for this project
4. Economic Feasibility: There will be no cost acquiring and preparing the data, training and deploying the model.

### Functional Requirements

Here are the Functional Requirements to be considered:

1. Data input: The system should be able to accept data about used cars, such as make, model, age, mileage, and condition, as input.
2. Machine learning model: The system should be able to use a machine learning model to analyze the data it has collected and make predictions about the prices of used cars.
3. User interface: The system should have a user interface that allows users to input data about a used car and receive a prediction of its price.
4. Accuracy: The system should be able to provide accurate predictions of used car prices, with a low margin of error.
5. Scalability: The system should be able to handle a large volume of data and remain efficient as it processes more and more used car listings.

### Non- functional Requirements

1. Security: The system should ensure the confidentiality and integrity of user data.
2. Usability: The system should be easy to use and navigate for users.
3. Scalability: The system should be able to handle a large volume of data and users without performance issues.
4. Reliability: The system should be reliable and consistently provide accurate predictions.
5. Maintenance: The system should have a robust maintenance plan in place to ensure smooth operation and timely updates.
6. Interoperability: The system should be able to integrate with other systems or platforms as needed.

### Technology/ Tools/ Software/ Hardware

* **Python Language**: it’s a popular language for machine learning due to its simplicity, ease of use and has extensive libraries and frameworks for use
* **Jupyter Notebook**: this is the code editor I will use
* The machine learning models to be used are:
  + **Linear regression**: is a statistical model that is commonly used in machine learning to predict a continuous dependent variable based on one or more independent variables. In a used car price predictor, linear regression could be used to predict the price of a used car based on factors such as the make and model, age of the car, and mileage.
  + **Random Forest Regressor**: this model can be used in a used car price predictor to make predictions about the price of a used car based on various features of the car. This model works by building a large number of decision trees and using the average of their predictions as the final prediction. One benefit of using a random forest regressor model is its ability to handle a large number of features and make accurate predictions even when there are many variables at play. It is also resistant to overfitting, meaning it is less likely to make predictions that are too specific to the training data and not generalizable to new data.
* The python Libraries:
  + **Numpy** is a Python library that may be used to work with multi-dimensional arrays, linear algebra, the Fourier transform, and matrices.
  + **Pandas** is a data manipulation and analysis package written in Python.
  + **Matplotlib** is a Python package that allows you to create static, animated, and interactive visualizations.
  + **Seaborn** is a matplotlib-based python data visualization package. It has a high-level interface for creating visually appealing and instructive statistics visuals.
  + **Sklearn** is a Python toolkit that allows you to create machine learning and statistical models including clustering, classification, and regression.
* Visual Studio Code: writing the web app code.
* Hardware: used a hp Laptop for all operations.

## SYSTEM DESIGN

### Importing Libraries

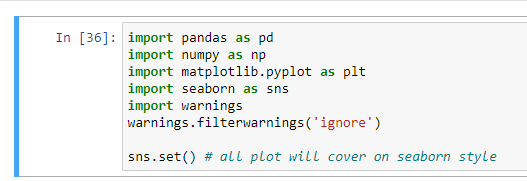


Figure 2: importing libraries

### Reading the Data

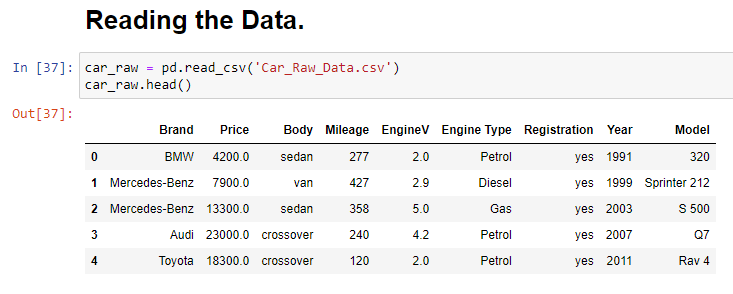


Figure 3: reading the data. shows some of the records

Above shows part of the data set. Contains the car brand e.g., BMW Price is 4200$, its body type is sedan, mileage is 227, Engine Volume is 2.0, Engine Type is Petrol, its has been registered, the Year is 1991 and the Model is 320.

### Data Preprocessing

Data preprocessing entails converting raw data into a comprehensible format that a machine learning model can understand.

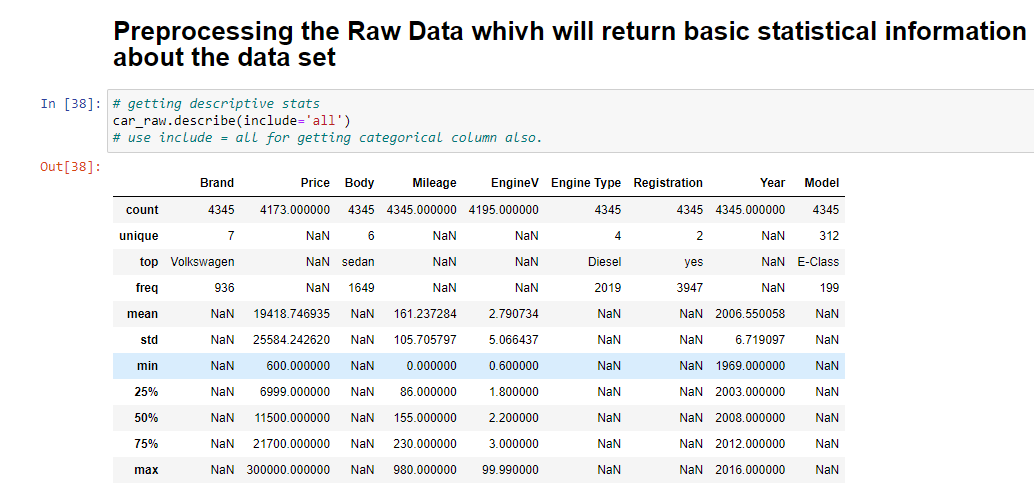


Figure 4: Data Preprocessing

As the Model column is contain 312 unique model of cars, and after converting it on dummy variable the dimension of our data will be very high, it will increase 312 more columns into our data.

That is why for now I will drop the Model column from the dataset.

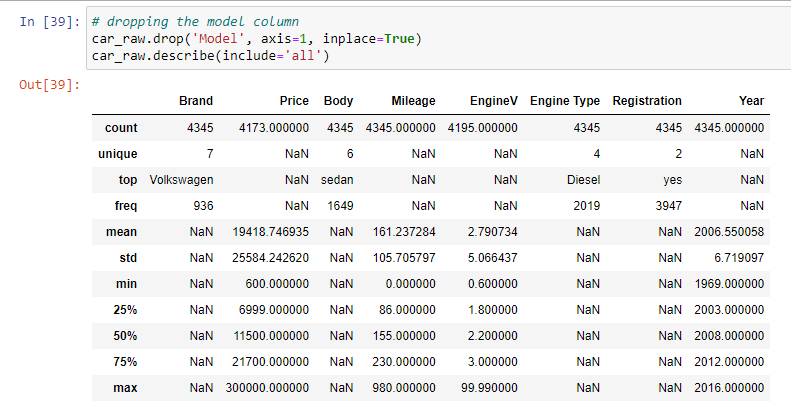
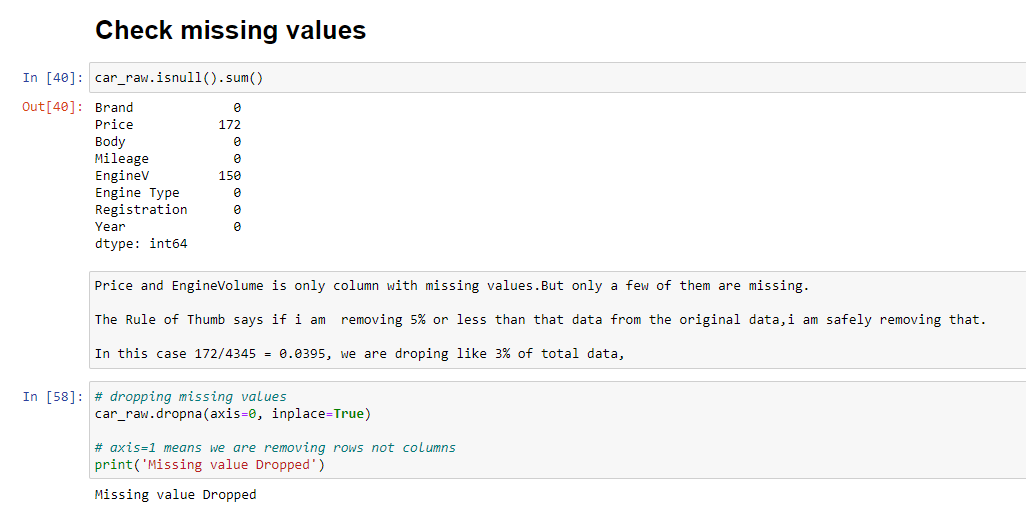


Figure 5: dropping the model column

### Checking for missing values and dropping them



# IMPLEMENTATION

# TESTING

# CONCLUSION

# REFERENCES