Journal of the Society of Automotive Engineers Malaysia
Volume 8, Issue 1, pp 41-61, January 2024
e-ISSN 2550-2239

A Comparative Study on Used Car Price Prediction Model

A. A. Ishak[1], **Z. Othman***[2] and S. S. S. Ahmad[2]

[1] Infineon Technologies, Free Trade Zone Batu Berendam Melaka, 75350 Melaka, Malaysia
[2] Universiti Teknikal Malaysia Melaka (UTeM), 76100 Durian Tunggal, Melaka, Malaysia

*Corresponding author: zuraini@utem.edu.my

ORIGINAL ARTICLE ***Open Access***

Article History:

Received
1 Aug 2023

Accepted
25 Dec 2023

Available online
1 Jan 2024

ABSTRACT – *The number of used car prices in the market keeps increasing due to the launch of a new model by the car manufacturer. The sales price is determined by the car's specifications and present state. The objective of this study is to make a comparison of the machine learning algorithm that can be implemented for used car price prediction. Previous studies on used car price prediction commonly perform the comparison of the machine learning model, meanwhile, the study on stock price prediction utilizes hyperparameter tuning. It shows that hyperparameter tuning can increase the performance of the machine learning model. The expected outcome from the study is the best machine learning model will be used for used car price prediction. The machine learning model will be trained by using Azure Machine Learning Studio. Therefore, the study compares four different machine learning models, including linear regression, neural network regression, boosted decision tree regression, and decision forest regression. As a result, boosted decision tree regression is indicated as the most effective model, exhibiting high R-squared values and superior performance compared to the other machine learning algorithms. This study also will perform a comparative study on the prediction model with hyperparameter tuning to get the most feasible and accurate model for the prediction model. There are two different hyperparameter tuning being compared such as the entire grid and random sweep and the study shows that random sweep provides the best R-squared values at 0.874548. The best machine learning model will be deployed and integrated with the web application that is developed by using ASP.NET. This study will be beneficial to the community in providing valuable insight into the factors that influence used car prices. These insights can be utilized by industry professionals and market analysts to make informed decisions and develop effective strategies.*

KEYWORDS: Used car price prediction, machine learning algorithms, hyperparameter tuning, Boosted Decision Tree Regression, web application integration

1. INTRODUCTION

The rate of car production has significantly increased over the past few years. Statista's Research Department estimated that 66.7 million cars will be produced worldwide in 2021 (Statista Inc., n.d.). The overall number of cars sold worldwide each year is a very significant amount. The number of people who wish to acquire a car is growing at the same quick rate as the population. As a result, the used car market is currently seen as a sector that is flourishing. There are some scenarios in which the community will use a new car for around five to seven years before selling it and purchasing another new model because the car manufacturer normally launches a new model every five years.

The newly produced car is unable to reach the customers for various reasons such as high prices, less availability, and financial incapability (Varshitha et al., 2022). On the other hand, citizens also have an option to purchase the used car instead of the new car. The price of the used car usually will be quite low if compared with the new car. Therefore, the best option for customers in many nations is to purchase a used car due to the fair price (Monburinon et al., 2018).

There are clear benefits to acquiring a secondhand car as compared to a new one. Of course, the most significant distinction is that secondhand products are typically less expensive because the vehicle is widely known as one of the assets that is depreciated. New cars will be having higher depreciation value if compared to the used car. A new car loses 9 to 11% of its value the moment you drive it off the showroom. The standard car will have lost 20% of its value in just a year. After then, a car will lose between 15 and 25 percent of its present value annually for the following five years. Depreciation continues after that, but much more slowly. The resell value of a used car depends on the current condition of the car or total mileage of the car. The more mileage cars also will require a lot more maintenance and repair.

Typically, a car's owner will sell the vehicle to a car dealer instead of selling the vehicle by using the online platform. There is a slight difference on the total resell value to individual and a car dealer. Car dealers frequently pay less than market value for used cars. It helps the car can be sells at a higher price in the future. Before a car dealer buys the vehicle from the original owner, there are a few details that may need to be disclosed. The sales price is determined by the car's specifications and present state. The sales price of a used car is being debated around the world because the resell value is subjective.

The implementation of information technology in all applications is a trend that is currently being used around the world to solve any problem that occurs. These modern technologies will bring a new era and changes in daily activities which will benefit all people around the world. Technology also contributed to solving various problems in the world such as a prediction for a used car price. Recent studies show, there is a various method that is widely used for the prediction model such as machine learning algorithm. This prediction will assist in enhancing the used car market's operational effectiveness and competitiveness.

The rapidly expanding discipline of data science includes machine learning as a key element. Algorithms are trained to generate classifications or predictions using statistical techniques, revealing important insights in data mining operations. In the previous paper, a comparison of a few models or algorithms is performed to have a more accurate prediction model of a used car price. According to a study on used car price prediction that was made by Li et al. (2022), gradient boosted decision model can provide a smaller prediction error if compared to the random forest and it can be considered the best model for price prediction.

2. RELATED WORKS

2.1 Recent Studies on Algorithms in Price Prediction

The recent studies on algorithms in price prediction have been widely performed by other researchers. These include used car prediction, stock price prediction, and house price prediction. Each algorithm has its different performance on the prediction model. The machine learning algorithm is being used in the prediction model because it can predict output values from the input data.

Firstly, the prediction of used car prices using artificial neural networks and machine learning by (Varshitha et al., 2022). In this paper, the author is using deep neural networks and a few machine learning models such as linear regression and random forest algorithms. The random forest model provides the best model for this prediction model since it gives minimum possible error. The performance metrics obtained are at R^2 – 0.917 and MAE - 0.746.

Next, prediction of used car price based on a supervised learning algorithm by Wang and Wang (2021) is using few algorithms such as Extra Trees Regression, Random Forest Regressor, and Ridge Regression. In the end, the author found that Extra Trees Regression is the best model for second-hand car price prediction. R^2 for Extra Tress Regression is 0.9807.

Besides, Monburinon et al. (2018) performed the research on prediction of prices for used cars by using regression models. A regression tree is a type of prediction tree that effectively applies the idea of recursive partitioning to handle nonlinear regression problems. The results are compared by using mean absolute error as a criterion. Regression trees that have been gradient-boosted performed best with an MAE of only 0.28. Then came multiple linear regression, which had 0.55 errors, and random forest regression, which had 0.35 mistakes. As a result, the author concluded that gradient-boosted regression trees are suggested for developing price evaluation models.

Furthermore, regression methods are often used in the prediction of home prices. Based on historical information, regression models are utilized to determine a correlation between the independent variables and the dependent variable (home price). Based on the study by Dwivedi et al. (2022), different types of regression models are being used for house price prediction such as Multiple Linear, Ridge, LASSO, Elastic Net, Gradient Boosting, and Ada Boost regression. This comparison study is performed across all six different algorithms that provide different values of coefficient determination to find out the most efficient regression technique for prediction. The value of coefficient determination is being used for the comparison and it shows that Gradient Boosting Regression provides the highest R^2 value at 0.9177022 among other algorithms.

In addition, Maheshwari et al. (2022) performed studies on the prediction of stock prices using the Prophet Model with Hyperparameter Tuning. Prophet is a Python and R-based forecasting tool that is completely accessible to everyone. It is used to forecast time series data using an additive model to suit non-linear trends with yearly, monthly, and daily seasonality as well as holiday impacts. It performs best when combined with multiple seasons of historical data and time series with significant seasonal impacts. This study also performed a comparison of the results from the prophet model and the prophet model with hyperparameter tuning. It uses the same dataset, and the result of the model is measured by using RMSE and MAPE values.

TABLE 1: Comparison of Prophet Model vs. Prophet Model with hyperparameter

Prophet Model		**Prophet Model with Hyperparameter**	
RMSE	404.4906	RMSE	386.2658
MAPE	12.48	MAPE	12.1500

Table 1 shows the result for RMSE and MAPE on the prophet model and prophet model with hyperparameter. It also shows that the prophet model with hyperparameter helps to improve the forecasting model accuracy.

Based on a previous study that is being completed by other researchers on the implementation of machine learning algorithms in price prediction. It can be concluded that various machine learning algorithms can be used for prediction models. Each of the models will produce different performance results and its commonly measured by using the coefficient of determination. Hyperparameter tuning also can be implemented with the machine learning model and it can increase the performance of the model. The study on stock price prediction can be a guideline for this study to consider the implementation of hyperparameter tuning. It also provides a clearer picture for this study and allows it to focus on more than just comparing machine learning models to find the best useful algorithm to predict used car prices.

2.2 Coefficient of Determination

A statistical measure known as the coefficient of determination, or R^2, measures the percentage of variance in the dependent variable that can be explained by the independent variables in a prediction model. The coefficient of Determination can be used to evaluate the model performances. In other words, it evaluates the model's degree of predictability of fit. Higher numbers suggest a better fit, and R^2 values range from 0 to 1. According to Katagiri and Fujii (2022), the smaller the error between the measured and estimated values, the higher the R^2 value. The following is what various R^2 refer to:

- R^2 = 0 indicates that the model does not take into consideration any of the variations in the dependent variable.
- R^2 = 1 refers to the model that completely accommodates the dependent variable's variability.

The formula to calculate the coefficient of determination is shown in Figure 1 below:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Figure 1: Coefficient of Determination Formula

Details of each variable in the formula for the coefficient of determination are shown in Table 2.

Table 1: Coefficient of Determination variable datasets

Variable	Details
n	Total number of observations
Σx	Total of the First Variable Value
Σy	Total of the Second Variable Value
Σxy	Sum of the Product of First & Second Value
Σx2	Sum of the Squares of the First Value
Σy2	Sum of the Squares of the Second Value

It's crucial to remember that R^2 does not give a complete view of the model's performance. Consider the following important factors when analyzing R^2:

1) Model Complexity
    - A greater R^2 does not always reflect a better model. Even if the extra predictors aren't useful or significant, it's possible to artificially increase a model's R^2.
2) Overfitting
    - If the model has been overfitting to the training set of data, the R^2 may be misleading. High R^2 values on training data may not relate well to new data.

A measure of how well the independent variables reflect the variation in the dependent variable is given by the coefficient of determination. To fully comprehend the predictive performance of the model, it must be comprehended when compared with other model evaluation metrics and factors, such as model complexity, overfitting, sample size, and residual analysis.

3. PROPOSED METHOD

The focus of the research process for the study is to explain each phase taken to carry out the research. It consists of a few phases such as data collection, data preparation, data modeling, and data evaluation. The comparison of each data model will be performed to get the best model with the highest accuracy which aligns with the objective of this study. The objective of this study is to have a prediction model with high accuracy on the used car price prediction.

The research design flowchart for this study is shown in Figure 2. The first process will be data collection which will be used for the data modeling in this study. Based on the data that has already been collected and gathered, pre-processing on the dataset is required for this study. Next, data transformation is needed based on the dataset. Once the dataset is ready, the dataset will be an input for the data modeling for the study.

Various machine learning models are available in the data mining process. This study will use linear regression, neural network regression, boosted decision tree regression, and decision forest regression as an algorithm for the prediction model. Every model will produce its result and a comparison of the four models will be performed. This activity is known as data evaluation and the only model that provides the highest performance will be chosen for this study. The best performance model will be implemented with hyperparameter tuning such as Random Sweep and Entire Grid. Both hyperparameter tuning performances will be compared and the high R^2 values will be considered as the best model for prediction models.

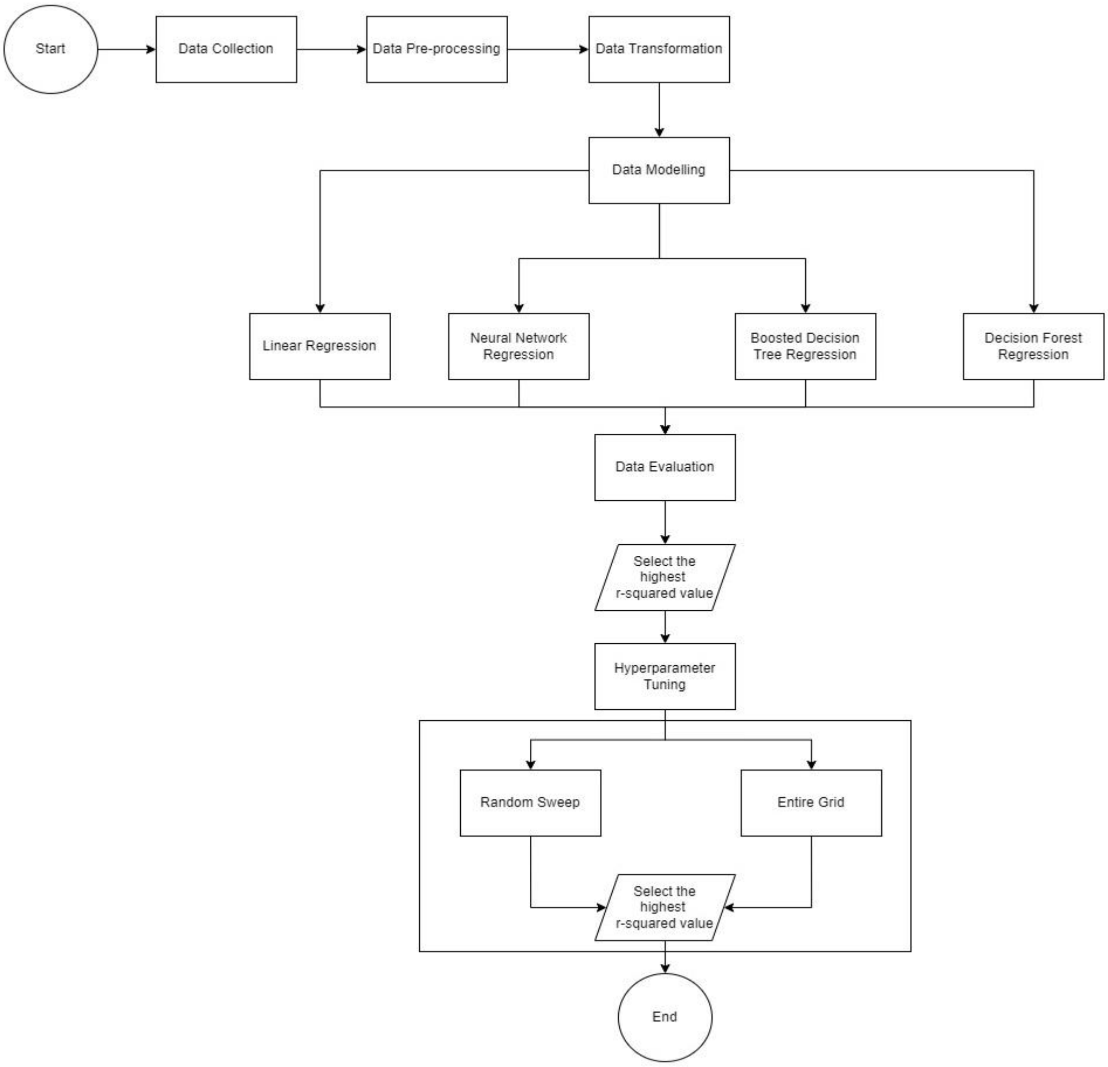

Start
Data Collection
Data Pre-processing
Data Transformation
Data Modelling
Linear Regression
Neural Network Regression
Boosted Decision Tree Regression
Decision Forest Regression
Data Evaluation
Select the highest r-squared value
Hyperparameter Tuning
Random Sweep
Entire Grid
Select the highest r-squared value
End

FIGURE 2: Research design flowchart

3.1 Data Collection

The dataset for this study is being collected from an online platform for the community of data scientists and machine learning practitioners. The dataset was gathered in 2019 and is being shared on the website. The dataset consists of 7,253 rows and 12 attributes. This dataset consists of various brands of used cars that exist around the world. The dataset for used cars was collected by Avi (n.d.) from 1998 until 2019.

3.2 Data Preprocessing

For data experts or business users, data preprocessing is frequently known as a time-consuming task, but it is necessary to store data in context to generate insights and remove bias brought on by inaccurate information. There are a few processes that can be performed in data pre-processing such as removing outliers and feature selection by using Exploratory Data Analysis (EDA). The dataset also might have outliers that need to be removed from the raw data. In a random sample taken from a population, an outlier is an observation that is abnormally distant from other values. The outliers can be detected if the data is visualized by using a boxplot. A solid line is drawn across the box to indicate the median and a box is drawn between the upper and lower quartiles to construct a box plot. Figure 3 below shows the example of the outliers by using the boxplot. The figure below also shows that the data values for 1,500 outliers for the dataset.

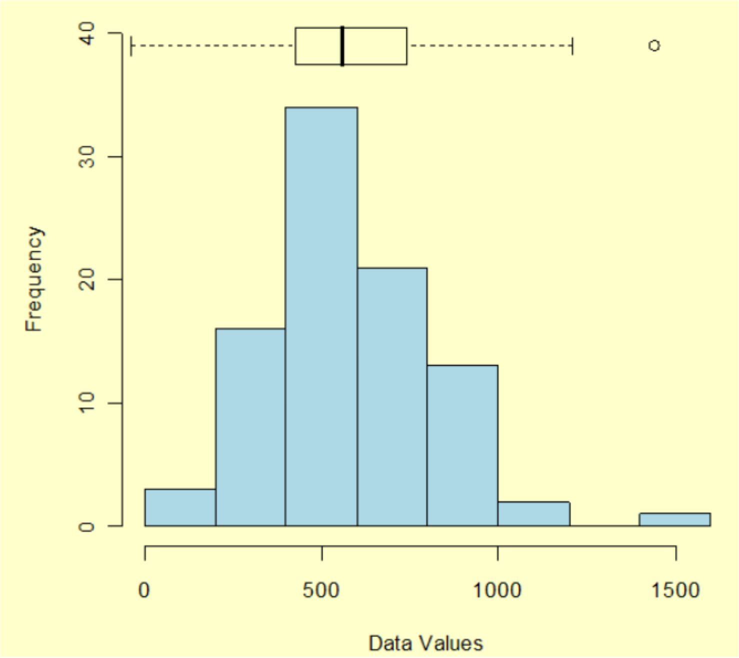


FIGURE 3: Example of Outliers by using boxplot

On the other hand. in statistics, EDA enables data analysts to discover the key aspects of the data generally through visual methods, which might also lead to the creation of hypotheses and the execution of new experiments (Samvelyan et al., 2020). EDA is one of the important activities that is needed in data science. Data visualization and insight extraction are components of EDA.

3.2.1 Data Cleaning

For this study, data cleaning may involve removing missing values, transforming the data, and discarding extraneous data. Every dataset will consist of values that usually refer to the text, dates, number, or Boolean data. There is some scenario in which the missing value is being stored in the dataset. Missing data is when the values of the relevant variables are not measured or recorded for all sample subjects (Psychogyios et al., 2023). Data can be missing due to several reasons, especially on a huge dataset. There are many methods to identify the missing values in the dataset such as by using pandas in Google Collaboratory. As a result, the null values in the dataset are shown as NaN in Table 3.

TABLE 3: Example of missing values in the dataset

Price	Brand	Engine(cc)	Power (bph)
295	Honda	NaN	NaN
211	Maruti	NaN	NaN
360	Ford	1498.0	99.0
175	Maruti	NaN	NaN
2,650	Land	NaN	NaN
320	Honda	NaN	NaN
580	Maruti	NaN	NaN

Missing values can affect the outcomes of analysis, and some algorithms are unable to deal with them, thus it is crucial to discover and manage them appropriately. After determining the missing number, the study will primarily concentrate on two different ways to solve the issue. Based on a study by Miles and Hunt (2015), this scenario needs to attempt to impute (or fill in) the values of the missing data instead of removing data values or cases with missing data. Besides, the common approach to handle missing values is by performing imputation on the dataset. This imputation consists of the mean, median, and mode of the dataset. The common approach to imputation is by using mean imputation because it can be used if the data is symmetrically distributed. The mode and median can be used if the data is skewed but the median is more relevant to use if the data consists of a high number of outliers.

3.3 Data Transformation

Before performing data mining, data transformation is a crucial data pre-processing technique that must be applied to the data to produce patterns that are simpler to comprehend. Data transformation transforms the data into clean, usable data by altering its format, structure, or values. Data transformation may be used in data integration, migration, warehousing, and data wrangling. Businesses can make better data-driven decisions, especially in data transformation, which also improves the efficiency of business and analytical operations.

3.4 Data Modelling

Data modeling is a collection of procedures used to mix and examine various data sets to find connections or trends. By using historical data, data modeling aims to guide future actions. This phase will be using the dataset, and it usually will be split into two datasets which are known as training data and testing data. The data used to train an algorithm or machine learning model to predict the outcome that your model was designed to predict is known as training data. Test data are used to evaluate the effectiveness of the algorithm you are using to train the algorithm, such as its accuracy or efficiency. In this study, 70% of the dataset will be training data and 30% of the dataset will be test data.

3.4.1 Prediction Model Using Machine Learning Algorithm

Four different machine learning algorithms have already been selected and will be used in this study such as linear regression, neural network regression, boosted decision tree regression, and decision forest regression. The algorithm is being selected based on previous studies that were performed by other researchers. A data mining method known as regression is used to forecast the numerical values of a given data collection. Regression may be used, for instance, to forecast the price of a good or service or other variables. Additionally, it is used in many different industries for trend research, financial forecasting, and business and marketing behavior.

3.4.2 Hyperparameter in Machine Learning Algorithm

Hyperparameters refer to the defined parameters before a machine learning algorithm is trained and are not derived from the data. It is one of the optimization techniques that can mitigate the effect of over-fitting and under-fitting the problem (Ahamed et al., 2022). It influences how the algorithm behaves and can significantly affect how well the model works. The number of hidden layers and the number of nodes in each layer, for instance, can be manually modified in neural networks. Hyperparameters play a major role in the performance of the model.

Finding the combination of hyperparameters that produces the greatest performance is referred to as hyperparameter tuning or hyperparameter optimization. Usually, the process requires physical labor and is computationally expensive. There are a few techniques for hyperparameters such as random sweep and entire grid. Random sweep commonly referred to as random search, is a machine-learning technique for hyperparameter optimization. It serves as an alternative to grid search, which thoroughly analyses every potential set of hyperparameters. Meanwhile, the entire grid is known as a grid search, and it is one of the methods for hyperparameter optimization that explores every possible combination. Its significant disadvantage is that it operates very slowly. Checking every possible configuration of the point would take a lot of time, and also will have the possibility that the point is not available.

3.5 Data Evaluation

One of the most important steps in any data mining process is data evaluation. It fulfills two functions, including predicting how well the final model will perform in the future and acting as a key component of numerous learning techniques that help identify the model that most accurately represents the training data. In data science, it is not acceptable to evaluate model performance using the training data because this can quickly lead to overly optimistic and overfitted models. This study will be using the R^2 value which is known as the coefficient of determination to compare four machine learning algorithms to get the best model that can be used for used car price prediction. The model with a high R^2 value will be considered the best model for the prediction model.

4. RESULTS AND DISCUSSION

The findings and analysis of this research on forecasting used car prices are presented in this section. The objective of this study is to develop a realistic and reliable model that can compute the price of a used car based on a variety of factors and characteristics.

The result and analysis obtained from Microsoft Azure Machine Learning Studio and Tableau will be thoroughly assessed by the researcher after analyzing this section. The four primary findings from the study are graphically displayed in Figure 3. Furthermore, the importance of observing the research goal will be emphasized and highlighted in this chapter.

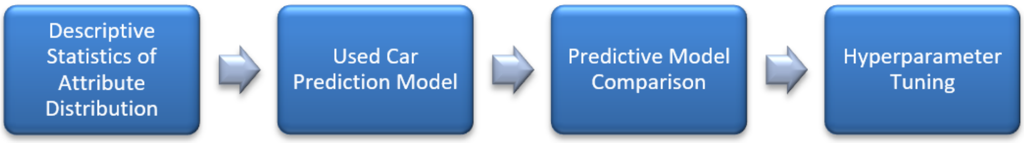

Descriptive Statistics of Attribute Distribution
Used Car Prediction Model
Predictive Model Comparison
Hyperparameter Tuning

FIGURE 4: Result and analysis process flow

Utilizing machine learning is the method to develop a prediction model. There are four primary sections within this section:

1) Feature Transformation
   - In machine learning, feature transformation is the process of changing or modifying a dataset's input features to enhance the functionality and efficiency of a machine learning model.
2) Feature selection
   - To increase the effectiveness and interpretability of a predictive model, feature selection is used to limit the number of input variables. It is not necessary to include all feature that exists in the original dataset because not all features are useful for developing the machine learning model.
3) Dataset splitting
   - Developing the machine learning model requires the researcher to split the dataset into two different datasets which consist of training and testing datasets. A training dataset will be used to train the algorithm. Meanwhile, the testing dataset will be used to measure the accuracy or effectiveness of the model.
4) Performance evaluation
   - To compare the coefficient of determination values, the researcher can compare the performance measure outcomes in this section. This comparison makes it possible to identify the technique that produces the best outcomes. The best method is also enhanced by adding a new parameter known as a hyperparameter to the predictive model.

4.1 Descriptive Statistics of Attribute Distribution

There are twelve dependent variables, and one independent variable is used to demonstrate the flowchart results from this study. The dependent and independent variables are listed in Table 4. Both independent and dependent variables are used in the creation or development of the machine learning algorithm.

The measurements or properties used as input to a machine-learning model are referred to as independent variables. The measurements or properties used as input to a machine learning model are referred to as independent variables, also known as features or input variables. These variables are selected based on their possible impact on the target variable as well as their ability to offer insightful data for forecasting or classifying outcomes. Meanwhile, the variable that the machine learning model aims to foresee or classify based on the input data is known as the dependent variable. The machine learning model has utilized the relationship between independent and dependent variables to ensure the accuracy of the model.

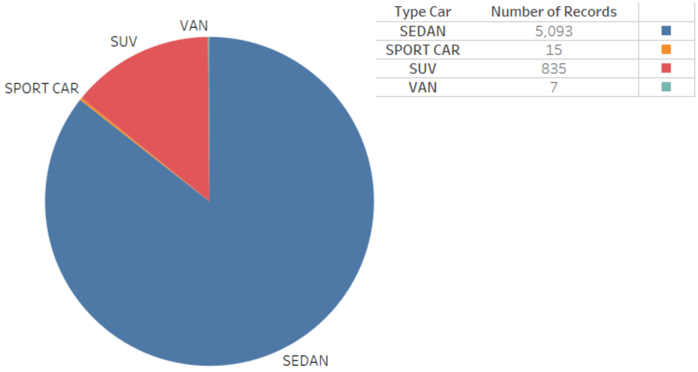
TABLE 4: Dependent and Independent Variables

Types of Variables	Variable Name
Independent Variable	Fuel_Type, Transmission, Owner_Type, Engine(cc), YearRange, Seats ,TypeCar ,Region, Power(bhp), Brand, KilometerDrivenRange, Mileage(km/l)
Dependent Variable	Price

Descriptive statistics are being performed in this study by using a visualization application called Tableau. Tableau is one of the tools that can be used for data visualization and business intelligence. It can provide the tools and functionality that can help end users analyze in the most interactive and intuitive approach. These tools can connect with multiple data sources such as databases, Excel spreadsheets, or any online data source from cloud services.

4.1.1 TypeCar

Nowadays, there are many various car types, and each of these is made to fulfill a certain function and satisfy a variety of preferences and demands. The dataset consists of 5,950 samples and most of it is sedan car type. The most popular kind of vehicle is a sedan, which has a separate trunk and typically has four doors. The dataset consists of sedans, sports cars, Sport Utility Vehicles (SUVs), and vans as shown in Figure 5.



Type Car	Number of Records	
SEDAN	5,093	🟦
SPORT CAR	15	🟧
SUV	835	🟥
VAN	7	🟩

FIGURE 5: Distribution of car type

4.1.2 KilometerDrivenRange

KilometerDrivenRange is one of the independent variables in the dataset. The variable can provide insightful information on the relationship between price and kilometer driven. Figure 6 shows the average price will decrease if the car's kilometer range is higher. The range for kilometers driven also is split into 4 different categories such as "0-50,000 km", "50,000-100,000 km", "100,000-150,000 km" and "150,000-200,000 km".

4.1.3 YearRange

The YearRange distribution for the dataset is shown in Figure 7. It shows that the average price of a used car will be higher in the latest car year. In the last 5 years, cars commonly had much higher prices if compared to old cars. YearRange variable also is classified into five different categories such as "1995-1999", "2000-2004", "2005-2009", "2010-2014" and "2015-2019".

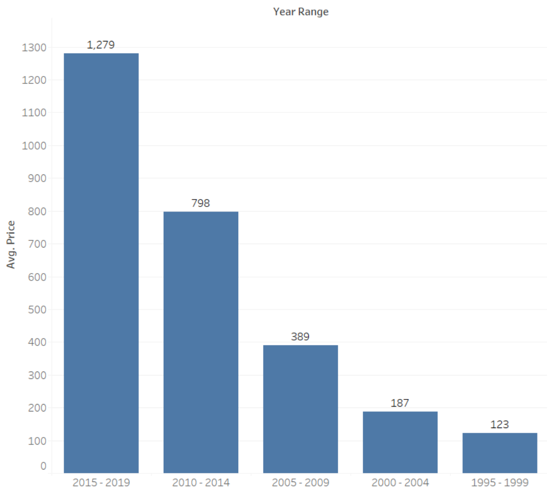


FIGURE 6: Distribution of KilometerDrivenRange

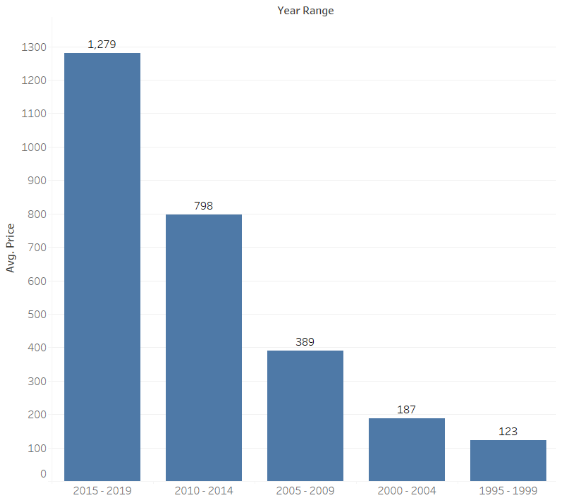


FIGURE 7: Distribution of YearRange

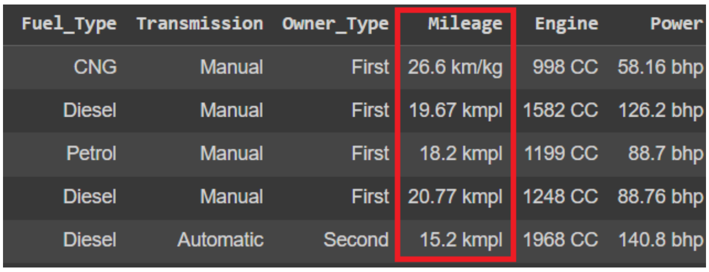
4.2 Used Car Price Prediction Model

This study is focused on the prediction model which is used to predict the price of a used car. This section will explain the outcome of the research on used car price prediction model activity which consists of feature selection, feature transformation, training and testing dataset, and prediction model that is being used.

4.2.1 Feature Transformation

In machine learning, the process of changing or transforming a dataset's original features into a new representation is referred to as feature transformation. The purpose of feature transformation is to improve the features of the data, addressing problems like nonlinearity, abnormal distributions, or scaling inconsistencies, to enhance the performance of machine learning models.

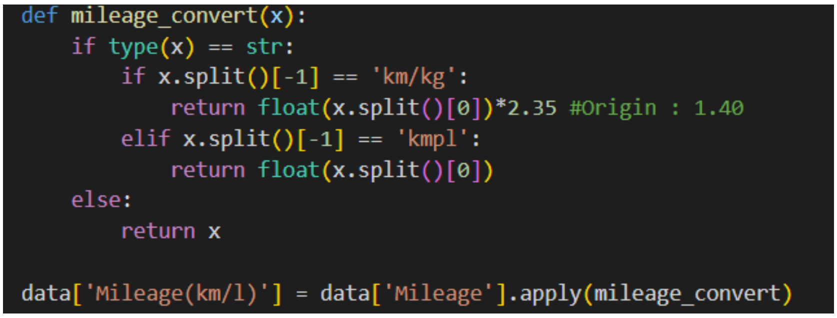
There is a possibility that the data in the dataset is not consistent and requires the transformation of the dataset. This dataset has one feature called Mileage and upon checking, the researcher found that the format value for that attribute is not consistent as shown in Figure 8.



Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power
CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp
Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp
Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp
Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp
Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp

FIGURE 8: Inconsistent feature values

This inconsistency of data has required the researcher to change the value and ensure that the format for all rows in that column is aligned. The values will be changed from km/kg to km/l by using Google Collaboratory tools. The Python code is being executed for this data transformation as shown in Figure 9 below.



```
def mileage_convert(x):
    if type(x) == str:
        if x.split()[-1] == 'km/kg':
            return float(x.split()[0])*2.35 #Origin : 1.40
        elif x.split()[-1] == 'kmpl':
            return float(x.split()[0])
    else:
        return x

data['Mileage(km/l)'] = data['Mileage'].apply(mileage_convert)
```

FIGURE 9: Mileage transformation code

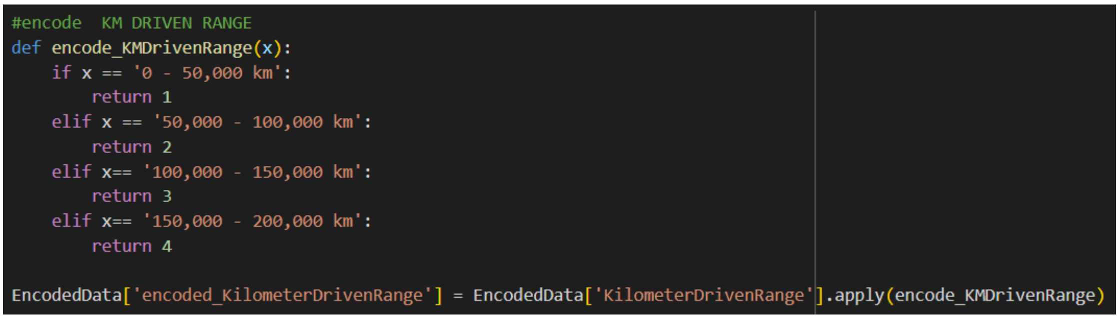
4.2.2 Feature Selection

A crucial phase in machine learning is feature selection, which involves determining which features or variables are from the original or raw dataset. According to Balakumar and Mohan (2019), feature selection is the method of selecting the features from a set of high-dimensional features. The intention to perform feature selection is to increase the model performance, reduce overfitting, and improve the understanding of the dataset.

Although the original dataset has many features, it is not necessary to utilize all the features as input in the machine learning model. This feature selection will require the assessment or evaluation activity to identify the importance or relevance of the attributes. Various techniques are commonly used for this feature selection phase. This study will use chi-square feature selection techniques which can be used to evaluate the relationship between categorical features and categorical target features. This technique is also able to evaluate how closely the features are being observed and expect the data distribution to be matched (Nirmala et al., 2022).

Since the chi-square technique can be used for categorical features only, the dataset also is being transformed to an encoded format to ensure that the dataset can perform the feature selection process. There are a few features that transformed such as Fuel_Type, Owner_Type, Region, Transmission, YearRange, KilometerDrivenRange, TypeCar, and other non-categorical features that exist in the dataset. The code for the data encoded is shown in the figure below.

The categorical data that exist in the dataset will be measured to identify whether it is a significant feature or a non-significance feature. This measurement will be defined by using P-Values as shown in the below table.



```
#encode  KM DRIVEN RANGE
def encode_KMDrivenRange(x):
    if x == '0 - 50,000 km':
        return 1
    elif x == '50,000 - 100,000 km':
        return 2
    elif x== '100,000 - 150,000 km':
        return 3
    elif x== '150,000 - 200,000 km':
        return 4

EncodedData['encoded_KilometerDrivenRange'] = EncodedData['KilometerDrivenRange'].apply(encode_KMDrivenRange)
```

FIGURE 10: Encoded feature code

TABLE 4: Chi-Square feature selection

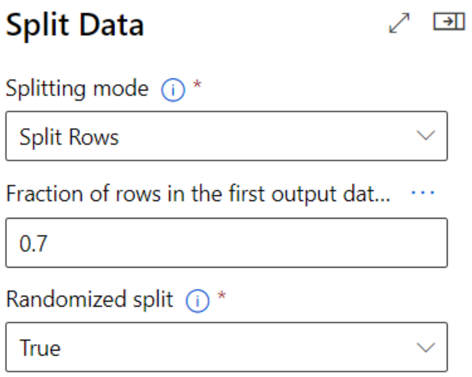
Features	P-Values
KilometerDrivenRange	0.0002
TypeCar	0.0013
YearRange	0.0005
Seats	0.345
Engine(cc)	0.0056
Power(bhp)	0.09
FuelType	0.125
OwnerType	0.242
Region	0.181
Transmission	0.0041
Brand	0.11
Mileage(km/l)	0.15

Based on the above result from Chi-Square feature selection techniques, it shows that five features of 12 features have P-values less than 0.05. If the p-values are less than 0.05, it shows that the features are significant for the prediction model. However, if the P-value is higher than the significance level, it is not considered statistically significant, and there is little to no evidence to support the null hypothesis.

4.2.3 Training and Testing Data

The dataset is split into two different datasets which consist of training data and testing data. The training dataset has a 70% proportion meanwhile the remaining 30% will be the testing dataset. The dataset has a total of 5950 rows which means 4165 rows will be the training dataset and 1785 will be the testing dataset.

This splitting technique is essential in machine learning algorithms since it can prevent or minimize the overfitting issue. When a model performs well on training data but poorly on testing data, it may be overfitting and not generalize well. The splitting dataset is being done by using Azure Machine Learning Studio as shown in Figure 11.



Split Data

Splitting mode *

Split Rows

Fraction of rows in the first output dat... ···

0.7

Randomized split *

True

FIGURE 11: Split dataset

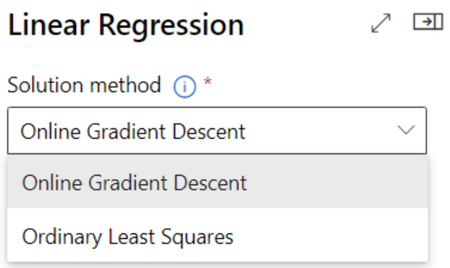
Based on the above figure, it shows that the dataset will have a randomized split. In machine learning, randomized splitting, commonly referred to as random train-test split, is a method for randomly dividing a dataset into training and testing subsets. This method ensures that the division is not biased towards any feature in the data by randomly assigning the data instances to the training and testing sets.

4.3 Predictive Machine Learning Model using Azure Machine Learning Studio

This study will be using Azure Machine Learning Studio (AMLS) as a tool for developing a machine learning model. The application will have different types of machine learning algorithms such as regression and classification. This study will be using a regression algorithm to predict the used car price and will be further explained in the below section.

4.3.1 Linear Regression

A supervised machine learning approach called linear regression is utilized in Azure Machine Learning Studio to predict a numeric target variable based on one or more input features. A linear relationship between the input features and the target variable that linear regression proposes. Two different solution methods exists in AMLS such as ordinary least square and online gradient descent.



Linear Regression

Solution method *

Online Gradient Descent

FIGURE 12: Solution method in linear regression

In linear regression, the estimation of a linear equation's coefficients is carried out using either the Ordinary Least Squares (OLS) method or online gradient descent. Online gradient descent is an incremental processing approach that updates the coefficients iteratively, one data instance or small batch at a time. Online gradient descent is computationally effective, open to new data, and ideal for handling big datasets or streaming data, although it can be more sensitive to noisy data and requires careful modifying of the learning rate and regularization parameters.

Meanwhile, OLS is a batch processing method that computes the coefficients using the full dataset. OLS does not naturally handle new data progressively, requiring the storage of the complete dataset, and is computationally expensive for large datasets. According to Kaba et al. (2023), the comparison to the existing mathematical models in the literature, OLS is seen to be the algorithm that is more accurate for parameter estimation.

These two different methods also provide a different amount of R^2 values for the study as shown in the below table.

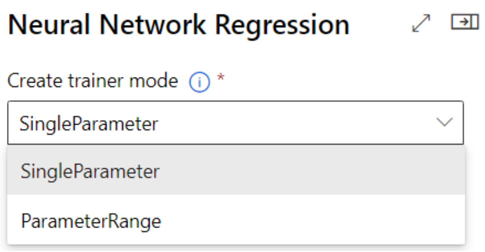
Table 5: R^2 value for linear regression

Solution Method	**R^2 Values**
Ordinary Least Squared	0.650658
Online Gradient Descent	0.636311

It shows that using the ordinary least squared method in the linear regression algorithm will provide a better model if compared to the online gradient descent.

4.3.2 Neural Network Regression

Regression problems can be solved with neural networks by using neural network regression, which is a term used by Azure Machine Learning Studio. A continuous numerical value or a set of continuous numerical values is used to be predicted in a supervised learning problem called regression. Neural networks are effective machine learning models that were influenced by the human brain. It can be trained to recognize complex patterns and correlations in data and to predict the future using those patterns. There are two methods to build and train neural network regression models in Azure Machine Learning. It can be done by using AMLS and Azure Machine Learning Python SDK. This study will be using AMLS as a tool for model training. Neural network regression in AMLS provides two different trainer modes such as single parameter and parameter range as shown in the below figure.



Neural Network Regression

Create trainer mode ⓘ *

SingleParameter

FIGURE 13: Trainer mode in neural network regression

The phrase "trainer mode" refers to a configuration that controls how the model is trained throughout the training phase. There is only one specific value for learning rate in SingleParameter mode. This means that the model is being trained by using a fixed learning rate throughout the training process. On the other hand, ParameterRange allows the user to define a range of values such as minimum and maximum values. Therefore, the model also will be trained by using multiple values within the specified range of values. The result for R^2 values for both trainer modes in neural network regression is shown in Table 6.

TABLE 6: R^2 value for neural network regression

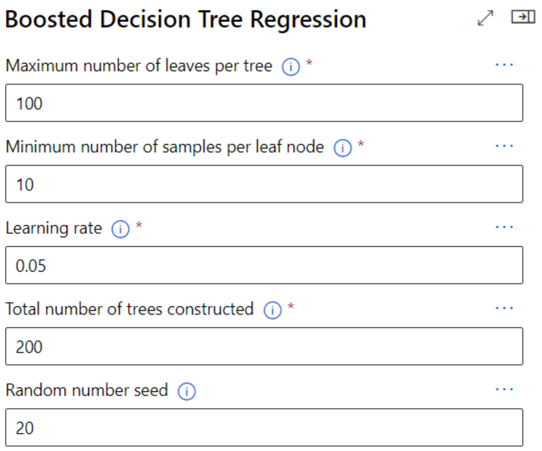
Solution Method	**R^2 Values**
SingleParameter	0.782423
ParameterRange	0.694011

Based on the above outcome, it can be concluded that SingleParameter will provide a high R^2 at 0.782423. The learning rate for SingleParameter is set at 0.01 meanwhile ParameterRange is set in a range from 0.1 to 0.4.

4.3.3 Boosted Decision Tree Regression

Boosted Decision Tree Regression is the other regression technique that is available in Azure Machine Learning Studio. This technique used the concept of decision trees and boosting to create an effective prediction model. AMLS is utilizing the FastRank algorithm to train this technique. It also can handle complex relationships between variables and target variables of the dataset. It works effectively when the data contains non-linear interactions and feature interactions.

Figure 14 shows the configuration that is being used in AMLS for this model. The maximum number of leaves per tree is set at 100 and the minimum number of samples per leaf node is at 10. The learning rate for the model is configured at 0.05 and the total number of trees constructed is at 200. Lastly, the random number of seeds is set at 20.



Boosted Decision Tree Regression

Maximum number of leaves per tree *
100

Minimum number of samples per leaf node *
10

Learning rate *
0.05

Total number of trees constructed *
200

Random number seed
20

FIGURE 14: Boosted decision tree regression configuration

Table 7 below shows the results for R^2 values from AMLS. The R^2 values for boosted decision tree regression are at 0.822809.

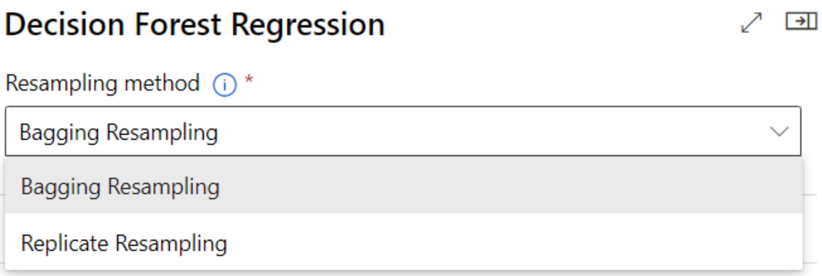
Table 7: R^2 value for boosted decision tree regression

Machine Learning Algorithm	$\boldsymbol{R^2}$ **Values**
Boosted Decision Tree Regression	0.822809

4.3.4 Decision Forest Regression

The Decision Forest technique in Azure Machine Learning Studio refers to the Random Forest algorithm, a common ensemble learning technique for classification and regression tasks. A group of decision trees known as Random Forest collaborate to develop predictions. The Decision Forest algorithm is well-known for being capable of managing complex data relationships, dealing with missing values, and providing insights into the significance of features. It is a flexible algorithm that can be used for a variety of machine learning purposes and has found extensive use in several industries, including e-commerce, healthcare, and finance.

There are two different resampling methods available in AMLS for decision forest regression such as bagging resampling and replicate resampling. Resampling methods are techniques used for sampling variations of the training data to increase the precision and generalizability of the regression model.



Decision Forest Regression

Resampling method *

Bagging Resampling

FIGURE 15: Resampling method in decision forest algorithm

Bagging resampling is also known as bootstrap aggregating. It is a method of resampling used to provide different versions of the training data for the ensemble of decision trees. Bagging is frequently used in random forest techniques, such as decision forest regression, to increase the model's accuracy and robustness. As a result of prediction, each tree in a regression decision forest produces a Gaussian distribution. The goal of the aggregation is to identify a Gaussian whose initial two moments correlate to the moments of the Gaussian distribution mixture obtained by mixing all distributions that individual trees returned. Furthermore, replicate resampling refers to a technique that uses the same input data on each tree. Each tree node's split criteria is chosen at random, and a variety of trees will be produced.

The result for R^2 values for the resampling method in decision forest regression is shown in Table 8. The R^2 values for the bagging resampling method are much higher if compared to the replicate resampling method. It shows that bagging resampling is the best method that can be used for decision forest regression.

TABLE 8: R^2 value for decision forest regression

Resampling Method	R^2 Values
Bagging Resampling	0.811526
Replicate Resampling	0.794878

4.4 Predictive Model Comparison Results

This study will make a comparison of each of the machine learning models that are being used for used car price prediction. As a result, the best and most effective machine learning model will be used to execute the used car price prediction model. Table 9 will show the results of four prediction models such as linear regression with ordinary least square technique, neural network regression with single parameter mode, boosted decision tree regression, and decision forest regression with bagging resampling method. R^2 values for each of the models are being measured to compare and define which model is the most suitable prediction model. The R^2 value that is nearest to 1 is considered to be the best model for the used car prediction model.

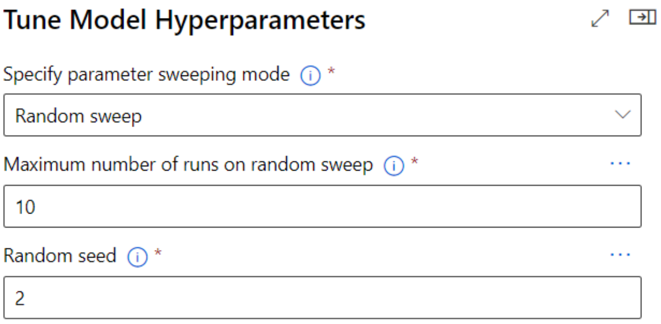
TABLE 9: Used car price prediction model comparison

Machine Learning Model	R^2 Values
Linear Regression	0.650658
Neural Network Regression	0.782423
Boosted Decision Tree Regression	0.822809
Decision Forest Regression	0.811526

According to Table 9 above, Boosted Decision Tree Regression is the best for predicting the used car price with a R^2 value is of 0.822809 which is greater than other machine learning models.

4.5 Proposed Hyperparameter Tuning

This study will implement two different hyperparameter tuning that are available in AMLS such as Entire Grid and Random Sweep. This hyperparameter will be implemented in Boosted Decision Tree Regression since it performs better compared to others. The goal for the implementation of hyperparameter tuning is to identify the best combination of hyperparameter values that can maximize the performance of the model. Hyperparameter tuning will be executed in two separate jobs to generate the R^2 values and determine the most suitable hyperparameter tuning for implementation with Boosted Decision Tree Regression. AMLS required the user to specify the maximum number of runs on a random sweep and the number of random seeds for the Random Sweep hyperparameter. Meanwhile, AMLS already pre-defined the two values for the Entire Grid hyperparameter.



Tune Model Hyperparameters

Specify parameter sweeping mode *

Random sweep

Maximum number of runs on random sweep *

10

Random seed *

2

FIGURE 16: Parameter in random sweep hyperparameter

The parameter for this study for the Random Sweep hyperparameter is shown in Figure 16. It shows that the maximum number of runs on random sweep is set at 10 and the total random seed is at 2. The outcome of the implementation of both hyperparameter tuning with Boosted Decision Tree Regression is shown in Table 10.

TABLE 10: Comparison on R^2 values for hyperparameter tuning

Machine Learning Model (Hyperparameter)	R^2 Values
Boosted Decision Tree Regression	0.822809
Boosted Decision Tree Regression (Random Sweep)	0.874548
Boosted Decision Tree Regression (Entire Grid)	0.868136

Based on the R^2 values that are shown in the table above, it shows that Random Sweep hyperparameter tuning generated higher R^2 values if compared to Entire Grid and the algorithm itself. These may meet the research objective which shows that the implementation of hyperparameter tuning can generate the best prediction model.

4.6 Deployment Machine Learning Model in Azure Machine Learning Studio

Azure Machine Learning Studio has a feature to deploy the trained model. The trained model can be deployed and integrated into a few applications such as Microsoft Power BI, Microsoft Excel, and web applications. This study will use web applications as a platform to integrate with the trained machine learning model. The web application will be developed by using Microsoft Visual Studio and ASP.NET will be used as a programming language.

The trained model needs to be converted to machine learning inference in AMLS as shown in Figure 17. Two processes are available in machine learning inference such as Batch Inference and Real-Time Inference. Batch Inference is an asynchronous process that bases its predictions on a batch of observations. The predictions are stored as files or in a database for end users or business applications. Meanwhile, Real-Time Inference frees the model to make predictions at any time and trigger an immediate response. This pattern can be used to analyze streaming and interactive application data. Real-time Inference is more suitable for this study since the web application is based on user input.

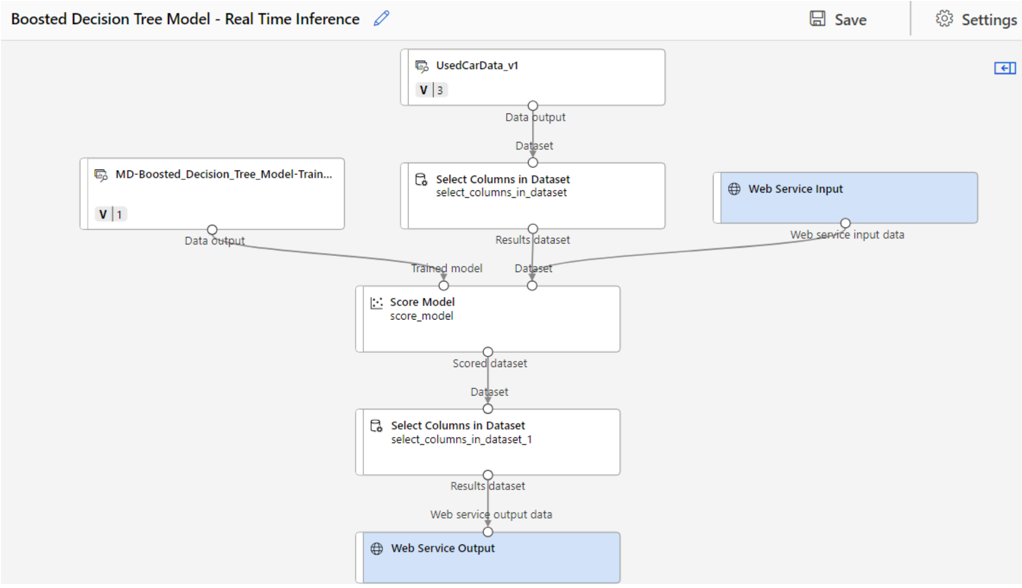
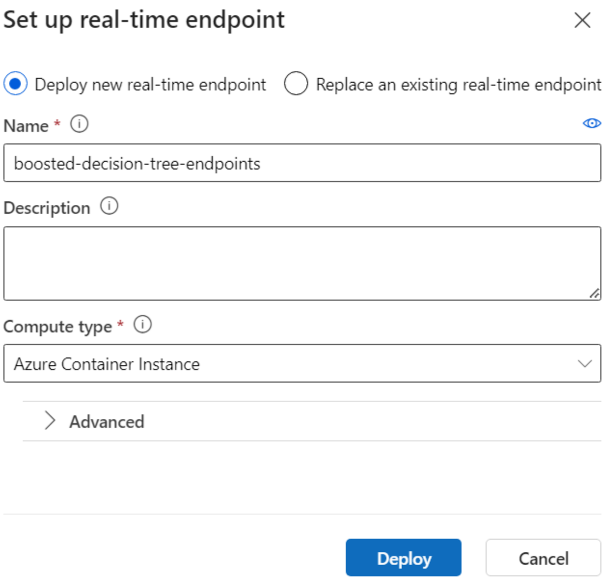


FIGURE 17: Converted trained model to machine learning inference

Next, the real-time inference model needs to be deployed to an endpoint since it will provide us with the Application Programming Interface (API) that can be used in web applications. Azure Container Instance will be used as a compute type for this model as shown in Figure 18.


Set up real-time endpoint

● Deploy new real-time endpoint ○ Replace an existing real-time endpoint

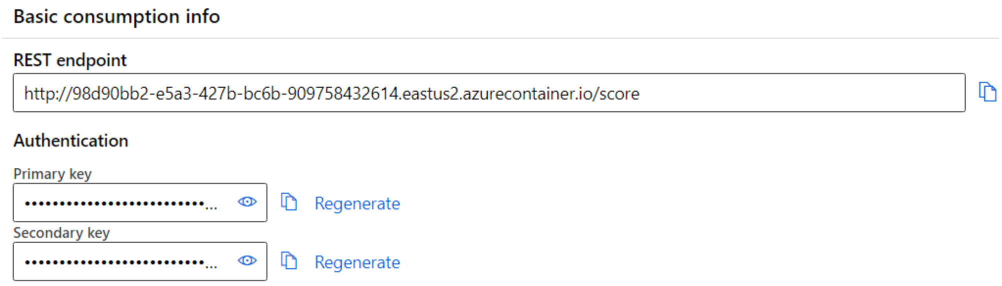
Name *: boosted-decision-tree-endpoints

Description:

Compute type *: Azure Container Instance

FIGURE 18: Deploy endpoint in Azure Machine Learning Studio

The deployed endpoint will be available in AMLS, and it can be accessed by using REST API. Representational State Transfer Application Programming Interface, or REST API, is a design concept and architectural approach for developing networked applications. It offers a common method for creating web services that can be accessed online. The client and server's communication is stateless, which means that each request comes with all the information the server needs to understand it and process it without reference to any previous communications. A significant volume of requests per second can simultaneously access the Web REST API URIs (Varthis et al., 2021).



Basic consumption info

REST endpoint

http://98d90bb2-e5a3-427b-bc6b-909758432614.eastus2.azurecontainer.io/score

Authentication

Primary key

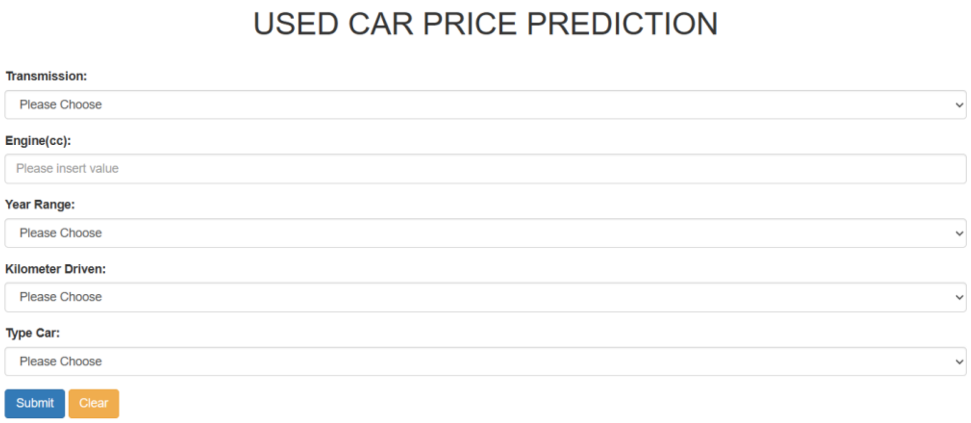
•••••••••••••••••••••••••...

Secondary key

•••••••••••••••••••••••••...

FIGURE 19*:* REST API in Azure Machine Learning Studio

The web application will use the designated key that is shown in Figure 19 to communicate with the deployed endpoint, taking input and showing the results in the user interface. The web application will make use of the Bootstrap framework, which provides design templates for typography, forms, buttons, tables, navigation, modals, and several other features utilizing HTML and CSS, to develop the user interface. Microsoft Visual Studio is being used in this study to develop the web application. The web application will require the user to input and fill in the form to get the predicted price for a used car as shown in Figure 19.



USED CAR PRICE PREDICTION

Transmission: Please Choose

Engine(cc): Please insert value

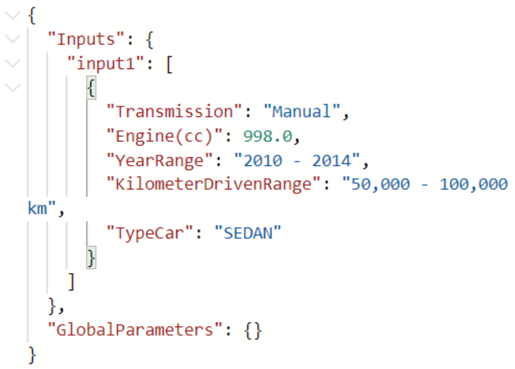
Year Range: Please Choose

Kilometer Driven: Please Choose

Type Car: Please Choose

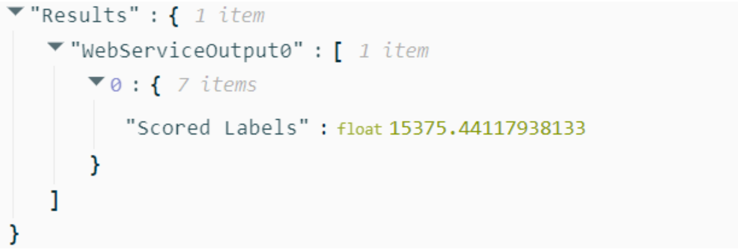
FIGURE 20*:* Web application form

The user's input, functioning as a client, should be in JSON format, as depicted in Figure 20. On the other hand, the server's response or output is illustrated in Figure 21. JSON, which stands for JavaScript Object Notation, is a lightweight data format commonly utilized for data storage and transmission. It is frequently employed when sending data from a server to a web page.



```
{
  "Inputs": {
    "input1": [
      {
        "Transmission": "Manual",
        "Engine(cc)": 998.0,
        "YearRange": "2010 - 2014",
        "KilometerDrivenRange": "50,000 - 100,000
km",
        "TypeCar": "SEDAN"
      }
    ]
  },
  "GlobalParameters": {}
}
```

FIGURE 20: JSON input



```
▼"Results" : { 1 item
    ▼"WebServiceOutput0" : [ 1 item
        ▼0 : { 7 items
            "Scored Labels" : float 15375.44117938133
        }
    ]
}
```

FIGURE 21: JSON output

5. CONCLUSION

In conclusion, this study compared various machine learning regression techniques, including linear regression, neural network regression, boosted decision tree regression, and decision forest regression. The objective was to develop an effective model for predicting used car prices and integrate it with a web application.

The model's performance was evaluated using R^2 values, which measure how well the predicted prices align with the actual prices. Through this evaluation, the study identified boosted decision tree regression as the most effective model, exhibiting high R^2 values and superior performance compared to the other regression techniques.

Furthermore, by implementing hyperparameter tuning techniques such as random sweep and entire grid, the model's performance was further enhanced. The hyperparameter tuning process resulted in improved R^2 values for boosted decision tree regression. Specifically, the R^2 value increased from 0.822809 to 0.874548 with the random sweep approach, while the entire grid approach yielded an R^2 value of 0.868136. These findings demonstrate that adjusting the hyperparameters of boosted decision tree regression, as recommended by the researchers, led to the best performance for the model.

Overall, this study establishes a successful machine learning model for predicting used car prices, with boosted decision tree regression being the most effective algorithm. Implementing hyperparameter tuning further enhances the model's performance, providing valuable insights for future applications in the field.

ACKNOWLEDGEMENT

The authors express their deepest gratitude to ASEAN NCAP Collaborative Holistic Research (ANCHOR) IV (No. ANTARABANGSA-ANCHOR/2020/FTMK-CACT/A00032) for supporting this research.

REFERENCES

Ahamed, J., Mir, R. N. & Chishti, M. A. (2022). Industry 4.0 oriented predictive analytics of cardiovascular diseases using machine learning, hyperparameter tuning and ensemble techniques. Industrial Robot, 49(3), 544-554.

Avi, K. (n.d.). Used Cars Price Prediction.

Balakumar, J. & Mohan, S. V. (2019). Artificial bee colony algorithm for feature selection and improved support vector machine for text classification. Information Discovery and Delivery, 47(3), 154-170.

Dwivedi, R., Gupta, R. & Pal, P. K. (2022). House price prediction using regression techniques. Proc. of the 4th International Conference on Advances in Computing, Communication Control and Networking, ICAC3N, 165-170.

Kaba, A., Yurdusevimli Metin, E. & Turan, O. (2023). Thrust modelling of a target drone engine with nonlinear least – squares estimation based on series expansions. Aircraft Engineering and Aerospace Technology, 95(1), 38-52.

Katagiri, K. & Fujii, T. (2022). Partitioned Path Loss Models based on coefficient of determination. International Conference on Information Networking, 198-203.

Li, Y., Li, Y. & Liu, Y. (2022). Research on used car price prediction based on random forest and LightGBM. In IEEE 2nd International Conference on Data Science and Computer Application, 539-543.

Maheshwari, A., Malhotra, A., Tuteja, S., Ranka, M. & Basha, M. S. A. (2022). Prediction of stock prices using Prophet Model with Hyperparameters tuning. In IEEE North Karnataka Subsection Flagship International Conference, NKCon 2022, 1-5.

Miles, J. N. V. & Hunt, P. (2015). A practical introduction to methods for analyzing longitudinal data in the presence of missing data using a marijuana price survey. Journal of Criminal Psychology 5(2), 137-148.

Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S. & Boonpou, P. (2018). Prediction of prices for used car by using regression models. Proceedings of 2018 5th International Conference on Business and Industrial Research (ICBI), 115-119.

Nirmala, G., Thirumurugan, G. & Jayasree, M. (2022). An analysis comparing traditional and digital marketing (advertising) using Chi-Square test and linear regression model. 1st International Conference on Computational Science and Technology, 99-104.

Psychogyios, K., Ilias, L., Ntanos, C. & Askounis, D. (2023). Missing value imputation methods for electronic health records. IEEE Access, 21562-21574.

Samvelyan, A., Shaptala, R. & Kyselov, G. (2020). Exploratory data analysis of Kyiv city petitions. IEEE 2nd International Conference on System Analysis and Intelligent Computing, 1-4.

Statista Inc. (n.d.). Number of cars sold worldwide from 2010 to 2023, with a 2024 forecast.

Varshitha, J., Jahnavi, K. & Lakshmi, C. (2022). Prediction of used car prices using artificial neural networks and machine learning. International Conference on Computer Communication and Informatics, 1-4.

Varthis, E., Poulos, M., Giarenis, I. & Papavlasopoulos, S. (2021). A novel framework for delivering static search capabilities to large textual corpora directly on the web domain: An implementation for Migne’s Patrologia Graeca. International Journal of Web Information Systems, 17(3), 153-186.

Wang, F. & Wang, Q. 2021. Prediction of used car price based on supervised learning algorithm. International Conference on Networking, Communications and Information Technology, 143-147.