Pak. j. life soc. Sci. (2024), 22(2): 19239-19251 E-ISSN: 2221-7630;P-ISSN: 1727-4915

Pakistan Journal of Life and Social Sciences
www.pjlss.edu.pk

https://doi.org/10.57239/PJLSS-2024-22.2.001411

RESEARCH ARTICLE

Car Depreciation Price Prediction Using Multiple Machine Learning Algorithms

Ting Tin Tin[1], Joel Kwoh Lik Xun[2], Gan Bruse[2], Jason Pong Soon Hui[2], Low Weng Chee[2], Chai Kian Hun[2], Ali Aitizaz[3], Umar Farooq Khattak[4], Ayodeji Olalekan Salau[5,6]

[1,5]Faculty of Data Science and Information Technology, INTI International University, Negeri Sembilan, Malaysia
[2]Tunku Abdul Rahman University of Technology and Management, Kuala Lumpur, Malaysia
[3]School of technology, Asia Pacific University, Malaysia
[4]School of IT, Unitar International University, Malaysia
[5]Department of Electrical/Electronics and Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria
[6]Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India

ARTICLE INFO

Received: Sep 18, 2024

Accepted: Nov 3, 2024

Keywords

Used car prices
Prediction of depreciation
Reuse of resource
Pricing decisions
Used car market

***Corresponding Author**

tintin.ting@newinti.edu.my

ABSTRACT

The main objective of this paper is to build a predictive model to estimate car depreciation by comparing different types of machine learning algorithms. The research analyses a dataset of used car prices and related variables, including age, mileage, make and model, and region. The study finds that age and mileage are the most significant factors that affect used car prices, with newer and lower mileage cars commanding higher prices. The make and model of the car, as well as the region where it is being sold, also have a significant impact on prices. Based on these findings, the study develops a predictive model using machine learning algorithms, including linear regression, MLP regression, Ridge, support vector regression, and Random Forest model. The models are trained on the dataset and evaluated using different metrics namely Root Mean Square Error (RMSE), Coefficient of determination (R2), and Mean Square Error (MSE). The results of this study showed that the Random Forest Model had the best accuracy out of the other models. This study provides insight into pricing strategies, inventory management, and decision-making in the related industry, especially in the used car market that encourages vehicle reuse. Furthermore, the model developed in this study can be used to make more accurate predictions of used car prices, helping buyers and sellers make more informed decisions.

1. INTRODUCTION

The used car market in the UK is thriving as it exceeds USD 117 billion in the year 2021. With consumers changing their vehicles every 7.7 years on average, it is expected to increase to $ 226 billion in the year 2027 (Mordor Intelligence, 2024). Given the market share of the and the magnitude of the used-car market, determining a used car price and its depreciation rate is very critical as it brings interest to various stakeholders such as the individuals who are looking to buy and sell cars, as well as the car dealer to determine their buying price and selling price with account of depreciation rate in order to not take a loss. However, the pricing process is often based on gut feeling or comparisons with other advertisements on various websites, which may not always lead to the optimal price for a particular car. To address this problem, a new project aims to determine the key features that influence the price of a used car and develop a machine learning model to predict prices accurately. By making this information freely available, the project aims to promote fair and transparent pricing in the used car market for all parties involved.

This study aims to propose a model that estimates the depreciation price and the rate of used cars in the UK by using multiple machine learning algorithms. Machine learning is a field of Artificial Intelligence that can be used to make predictions and useful inferences for large amounts of data using various algorithms and technologies. This study uses a data set named '100,000 UK Used car dataset" from Kaggle (Aditya, 2020). This study uses 9 car attributes from this dataset to predict the price of the used car and this study added one more attribute, that is, age, and use it with the 9 car attributes together to predict its rate of depreciation. After preparing the dataset which this study will have a detailed elaboration at section 3.0, this study uses this dataset to train five machine learning models and evaluate them using three performance matrices: accuracy, mean squared error (MSE) and root mean squared error (RMSE). The machine learning algorithms that this study used ranged from the simple model such as linear regression, ridge regression and SVR to the more complex model which includes deep learning and decision tree model such as the MLP regressor and the random forest regressor. To enable the use of categorical variables, attributes are converted into dummy variables before training a model. The model with the highest accuracy, MSE and RMSE will be the best model as it is best suited to predict used car prices and its corresponding depreciation rate.

The existing literature on this field of predicting used car prices uses a wide variety of models and datasets from all over the world, such as India, United States, Dubai, and China. Pudaruth (2014) uses a K-Nearest Neighbors (KNN) algorithm and naïve bayes algorithm to predict the prices of used cars in India and achieve a score of 60 to 70 %. AlShared (2021) proposed a used car prediction model using random forest, linear regression, and bagging regression. According to the study, random forests have the highest accuracy of 95%. Random forest regression is also the best model in the Gajera et al. (2021) research. Liu et al. (2022) uses a more sophisticated model to predict used car prices which is Particle Swarm Optimization-Gray Relational Analysis-Backpropagation Neural Network (PSO-GRA-BPNN) and concluded that this model is a much better model to predict the price of a used car.

LITERATURE REVIEW

The initial start of the project proposed by Pudaruth (2014) was to collect data on used cars. Approximately 400 used car data were collected from the team within about a month, all the data were collected in Mauritius. Variables from the data collected from each car consist of specifications of cars such as model, make (manufacturer), volume of cylinder, mileage, car manufacture year, paint color, the type of transmission, and price of the cars. After collecting the data, data cleansing was then performed to remove unwanted variables such as paint color and transmission of the car. Some records and variables were also removed from the data set due to inconsistency and availability. In the end, about 100 records were left which consisted of variables like make, cylinder volume, year, mileage, and price. Numeric encoding, which is the conversion of words into numeric code, was not required as every variable was already in numeric form (Low-Level Academy, 2020). After that, data analysis was then performed on the finalized dataset. Analyzing the correlations between variables using the Pearson correlation coefficient, they found that year and mileage had a weak negative correlation of -0.33. However, the correlation between car price and manufactured year however had a much higher correlation, which is expected since new cars tend to have higher prices (Bobbitt, 2020). They also proceeded to remove some outliers and tested changing the prices of the cars to logarithmic values, which they noticed a small increase of the regression coefficient (from 0.814 to 0.851) or Rs57468 to Rs51084 using mean deviation (Abhigyan, 2020). The next step from them was to fit the data set onto the machine learning algorithms and evaluate the scores of the algorithms by using Mean Absolute Error and accuracy score. The first machine learning algorithm that they tested was the K-Nearest Neighbours (KNN) algorithm, which is a type of supervised learning algorithm (Geeksforgeeks, 2023). The KNN algorithm was split into Toyota and Nissan cars to compare the results of both cars. The best result obtained from Toyota cars was Rs45189 with a value of k being 1 and for Nissan cars, the best result obtained was Rs25741 with k value of 5. With this, KNN worked much better compared to Toyota cars. The next machine learning algorithm model they tested was the Decision tree algorithm with the same car make being tested on. Two different decision tree

algorithms were used, J48 and Random Forest. For each algorithm, 3 types of training were used: 80-20 percentage split, cross-validation with 10 folds, and the whole training set. For J48, the results of the 80-20 percent split, cross-validation, and the whole training set were about 61%, 66% and 83%, respectively. For random forest algorithms, the percentage of success was also about 62%, 60%, and 70%. The last machine learning model that was used was the Naive Bayes model, where the same type of training was used. The results for the 3 types of training were 65%, 60%, and 77%. In conclusion, there were no machine learning algorithmic structures that stood out the most with models like linear regression and KNN algorithm having mean error of Rs 50000 and Rs27000 respectively, while models of J48 and Naive Bayes have accuracy percentages of about 60 to 70 percent. The small amounts of records in the data set had also been one of the struggles they had to face, since they only took one month to gather all the data which they hope to increase for future projects (Chaw et al., 2024; Ting et al., 2024).

AlShared (2021) proposed the prediction and valuation using data mining techniques. This research paper proposed a used car price prediction based on the Dubai dataset. The benchmark dataset was gathered from dubizzle.ae and buyancar.com sells used cars and analyzes them. The data set consists of 21277 rows, 11 attributes that could be factors. After dropping all duplicates like car name and null records, the size of the data set was reduced to 21241. To check the correlation between price and other independent variables, the non-numerical data type variables which are binary attributes will need to be converted to numerical features if needed, otherwise remove the non-numerical attributes (Brownlee, 2017). After converting, show the relationships among factors by heatmap. The heatmap shows that only four factors used that negatively affect the price, which are gear type, fuel type, specs origin, and mileage, are correlated with the target output (CFI Team, 2022). The correlation between price and each factor mentioned above are –0.1, -0.2, -0.2, and -0.5, respectively. Since the price of used cars is the only target to predict, supervised learning is used in this paper. Next, use data mining techniques to build three regressors, which are random forest, linear regression, and bagging regression (decision tree) (Sharma, 2022). The first algorithm used is the Random Forest Regressor, whose precision, mean squared error (MSE), mean absolute error (MAE) and Root Mean Square Error (RMSE) are 95%, 0.025504, 0.000844, and 0.0374 respectively (Acharya, 2021). The second algorithm used is linear regression with an accuracy of 85%, MSE is 0.13, MAE is 0.11 and RMSE is 0.07. The last algorithm is bagging regression with an accuracy of 88%, MSE is 0.14, MAE is 0.19 and RMSE is 0.07. From the above result, the Random Forest Regressor has the lowest errors and the highest accuracy. Thus, it has been chosen as the final model for this prediction.

Other than that, Collard (2022) also performs the price prediction for used cars using multiple machine learning regression models such as Linear Regression, Ridge Regression, Lasso Regression and Random Forest Regression. In this research, both linear regression and Random Forest Regression are used by other researchers as well, so this study can compare the results produced by a couple of researchers. There are also 2 new additions such as the Lasso Regression, and Ridge Regression so this study can compare the scores of these 2 models against the current model that is done by other researchers. Based on the result, he found out that Random Forest Regression performs the best when compared against Linear Regression, Ridge Regression and Lasso Regression across all metrics used. It can represent average depreciation much closer than the other 3 models at a 13.7% predicted annual geometric depreciation for the dataset independent of vehicle age. He also mentioned another factor which is revaluation and stated that it will affect the depreciation of cars. The score of these models is evaluated by root mean squared error (RMSE), R-squared Error and mean absolute percentage error (MAPE). These are the common evaluation metrics to evaluate the trained model, which is also used in research by AlShared (2021) and Milunovich et al. (2023). The data set that he used is obtained from Kaggle which contains 122,144 car listings and 13 attributes from the years 2018 to 2020 from the whole United States. The dataset includes a wide range of vehicles such as vans, pickup-trucks, and cars. After obtaining the dataset, he removes all the attributes that are not related to prices or have too many missing values. This is one of the ways to clean data to enhance the quality of the data. He also performs the base-2 logarithm to the selling price column for the selling price skew to the middle. An 80-20 train test split is done to the dataset prior to the training process. This is in line with the Pareto principle, which suggests that the ratio of

80-20 is known as the 80-20 rule (Harvey & Sotardi, 2018). Based on the results, the linear regression and ridge regression RMSE value for the testing data set is 5953, while Lasso Regression has a score of 5950 which is slightly better than. Random Forest has a much lower value of 4799 which is the best. The same applies to the R-squared error, which shows that Random Forest Regression had better performance. Using the MAPE metric, Linear Regression, Ridge Regression and Lasso Regression perform like each other. The MAPE value for random forest regression is 37.65% which is better than the other 3 models with scores of 44.45%. He also noticed that Linear Regression, Ridge Regression and Lasso Regression tend to predict prices lower than actual. Lastly, the average depreciation for Linear Regression, Ridge Regression and Lasso Regression, independent of the age of the vehicle is around 9.7% while Random Forest Regression shows 13.2% depreciation for cars that were 2-4 years old and 14.6% for cars that are 13-15 years old. In conclusion, the author says the Random Forest Regression is the best model to evaluate from many perspectives.

Gajera et al. (2021) proposed to predict the used car prices by using machine learning models such as K-Nearest Neighbor Regressor (KNN), Random Forest Regressor, Linear Regression, XGBoost Regressor, and Linear Regressor. In this research, he used a total of five supervised models so that this study can compare the results. This study has a data set that contains 90000 plus data that will be used by the model to train. There are some variables that make the model predictable such as distance travel, type of fuel, model of car, registration year, brand of car, etc. to determine the worth of a car (Gegic et al., 2019). Based on the results, he found out that Random Forest Regressor has the highest predicted accuracy as 93.11% and the Root Mean Squared Error (RMSE) value 3702.34 which is the lowest result out of five models. The random forest regression considers many features in the data set and compares the bagging technique with the gradient boosting method (Meimi, 2021). For the result of RMSE, this study can see that the model with highest RMSE is KNN-Regressor, which is 7771.91, this model is higher than Linear Regression, which is 6864.23. Next, the Random Forest has 3702.34, while the Decision Tree Regressor is better than it, with the RMSE value of the Decision Tree Regressor being 5590.43. Compared to all the five models, the Random Forest Regressor has the lowest RMSE and has the best accuracy, which is 93%. The author says that the accuracy of the results is not strong because this research data are not large for the model. In conclusion, data cleaning is performed to remove outliers and null values from the prepared data, and the models are implemented to predict the price of cars (Anil, 2020). Random Forest Regressor is the best model than others, with a few more variables that may increase the accuracy of the model, example: number of doors, wheelbase, car length, car width, car height, engine size, and more (Akhtar, 2020).

Liu et al. (2022) discovered that there are numerous factors that influence used vehicle prices, which is why they are declining. Conventional trade methods for selling used cars are declining, leading to the expansion of the online used car market. The evaluation of a used vehicle is becoming more crucial as consumers demand a reasonable price given the value of the used car. There is no single primary factor that determines the price of a used car, making the used car market extremely complex (Özçalıcı, 2017). Instead, there are numerous factors that influence price, including mileage, use time, car brand, horsepower, and more. Gray relation analysis (GRA) is a technique that is used to analyze the correlation between various variables in a system, and it can be used to predict the price of used cars (Yao et al., 2019). The BP neural network is employed as a machine learning algorithm in used car price prediction to discover the connection between input variables (like vehicle brand, mileage, age, etc.) and output variables (Used car price). To increase the precision of the forecast, PSO can be used to optimize the weights and biases of the BP neural network. By repeatedly updating the particle's velocity and location and evaluating the results of various solutions using the fitness function - in this case, the BP neural network's prediction error - the PSO algorithm can assist in the search for the best weights and biases. To obtain a high degree of precision for the prediction of used car prices, the Particle Swarm Optimization-Gray Relational Analysis-Backpropagation Neural Network (PSO-GRA-BPNN) has emerged as a new model that compares favorably with multiple linear regression, random forest, and BPNN (Chen et al., 2017). How does this novel paradigm work? The correlation coefficient between a vehicle with an engine and one with mileage, a registration date, and a gearbox is found to be between 0.6 and 0.7. Additionally, the MAPE of this model is almost 4%, which is roughly 30% lower than that of the other three models. According to the accuracy, the data on which it is trained, and how this model can make up for the

shortcomings of the other three models, this study can conclude that the brand-new Particle Swarm Optimization-Gray Relational Analysis-Backpropagation Neural Network (PSO-GRA-BPNN) model is a much better model to predict the price of a used car. When it comes to a good used car price prediction, the market for used cars can become much longer-term and its expansion will be enormous. This is why it is important to keep finding new models of used car price prediction. The precision of a used car price acts as a fair bridge between the buyer and the seller (Xu et al., 2021).

RESEARCH METHODOLOGY

This section explicitly details the processing of the proposed RSSI-based geometric localization (RGL) scheme of the received signal strength indicator. Specific problems with respect to error analysis and approaches to avoid obstacles are examined, and the proposed scheme is refined. The initial phase is to find and gathering the data then analyze the data. After that, this study proceeds to preprocess, split, and build the models. The models are then validated using metrics. The best model is then chosen to predict the future depreciation rate of the cars. Lastly, a discussion about the results and a comparison of the results of other papers was carried out.

Data Understanding and Preprocessing

The dataset that this study chose to use was "100,000 UK Used car dataset" from Kaggle (Aditya, 2020). The dataset contains 100,000 scraped data of different types of used car listings from different car manufacturers, for example, Audi, BMW, Focus, Ford Hyundai, Mercedes and many more. Each car manufacturer has their own used car data set which consists of the car's specifications such as model, year, price, transmission, mileage, fuel type, tax, mpg, and engine size. Therefore, this study dependent variable is price, independent variables are: model, year, transmission, mileage, fuel type, tax, mpg, engine size.

In this paper, the Audi data set is chosen. When looking at the structure of the dataset, most of the independent variables were of string, integer, and floating data type. More specifically, there were 5 columns that were numerical, 3 columns that were categorical, and 1 year column. Duplicate rows and rows that contain empty values were also eliminated during the process of identifying the structure of the dataset. The relationship between the independent variables (transmission, year, mileage, mpg, engine size, tax) and dependent variables (price) was also analyzed using boxplots, and graphs for better visualization.

Machine Learning model used in this research

Linear Regression

The Linear Regression model is one of the most common and simple machine learning algorithms used for prediction. The linear regression model basically predicts the dependent variable using the independent variables by plotting the independent variable (x) on the horizontal axis and the dependent variable (y) on the vertical axis and plotting a line between them. The mathematical formula for the linear regression model is (Pudaruth, 2014):

$$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_i X_i + \hat{\epsilon}_i$$

With a dataset that contains more attributes, this study would use an alternative model, which is a multiple linear regression model. It basically uses the same technique as simple linear regression but with more independent variables. The mathematical formula shown below (Kuiper, 2008):

$$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_i X_i + \hat{\beta}_2 X_i \ldots + \hat{\beta}_n X_n + \hat{\epsilon}_i$$

MLP Regression

Multi-layer Perceptron Regression (MLP) is a supervised learning algorithm (Raj & David, 2020). Figure 3 shows an example of how MLP regression works. The neurons in the input layer (leftmost) represent the independent variables or input features. The next layer is called the hidden layer which might consist of 1 or more hidden layers; Figure 2 shows a model which has 1 hidden layer only. The

last layer (rightmost) is the output layer and is where the values from the last hidden layer are sent to be transformed into output values or predicted values (Kumar, 2022).

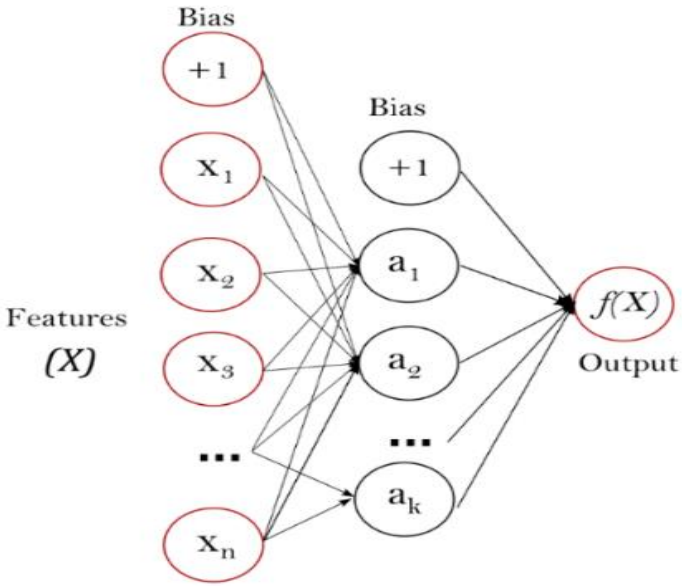

Fig. 2. Multilayer Perceptron Regression

Ridge Regression

Ridge regression is a type of linear regression but is a method used to analyze and reducing the multicollinearity between independent variables (Kumar, 2022). By reducing the standard error of the variables, this would help prevent overfitting and increase the reliability of the prediction estimates. This is known as L2 regularization (Ashok, 2024). This regression method is useful when there are more independent variables that have a highly correlated relationship and when there is a lot of noise in the data set. Therefore, this regression method seems to be suitable for the dataset chosen in this paper. Figure 3 shows the cost function for the ridge regression (Raj, 2020).

$$Min(||Y - X(theta)||^2 + \lambda||theta||^2)$$

Fig. 3. Cost function for ridge regression

SVR

Support vector regression (SVR) is a popular supervised learning algorithm. The basic idea behind SVR is like linear regression which is finding the best line or hyperplane (for higher dimensions) between the independent variable and dependent variable. SVR tries to fit the line or hyperplane within a value which is the distance between the hyperplane and the boundary line and maximize the distance of it (Sharp, 2020). Figure 4 shows an example of a support vector regression model where the black line represents the boundary line, and the red line represents the optimal hyperplane (Pham et al., 2020).

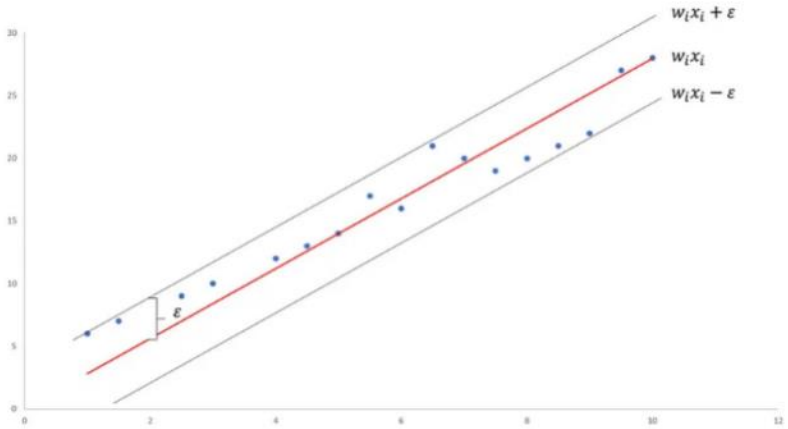

Fig. 4. Support Vector Regression Model

Random Forest

The last model that is used in this paper is the Random Forest model. The Random Forest regression algorithm is a supervised learning algorithm that uses multiple random decision tree models to predict values more accurately compared to the decision tree algorithm alone (Ashok, 2024). The basic idea of the decision tree model is that it builds the regression model in the form of a tree structure with three different kinds of node present in the tree structure. The decision node represents each of the variables in the data set. The leaf node represents the outcome of each decision node. Finally, the root node represents the upper part of the tree structure. This makes the Random Forest model much more accurate in predicting the values of the outcome, as it stacks several decision trees on top of each other.

Validation Method

The validation methods that are used to validate the machine learning models in this article are the MSE, RMSE, and the R2 score. MSE and RMSE are basically metrics that measure the difference between the predicted value and the actual value; therefore, the smaller the difference, the more accurate the predictive model (Dua et al., 2017). RMSE is usually preferred as the better metric as it calculates the square root of the average squared difference between the actual and predicted values, which is usually in the same units as the target variable. The coefficient of determination score or the R2 score is a metric that measures the statistical relationship between the actual and predicted values (Bobbitt, 2021). The value of the R2 score is measured between 0 and 1. For example, a highly correlated would have a value closer to 1 and a lower correlated relationship would have a value closer to 0.

Model Fitting

The prediction of used car price and the depreciation rate was then carried out in four different steps. (1) Split the dataset, (2) selecting best feature, (3) fit the training dataset into models, (4) evaluate the accuracy of models, (5) choose the best model to predict the rate of depreciation. First Step: The first essential step was to divide the data set into a training set and a testing set. The training set would contain 80% of the data set, while the remaining 20% would be the testing set. The selection of data for training testing set was also randomized or shuffled so that the data were not in any order. Second Step: In this step, the best features or independent variables were evaluated. In this paper, there were 37 features after preprocessing. The features are then converted into polynomial features (multiplying the features to increase the number of features), and polynomial simply means higher degree order than 1. This is to help improve the score due to the increase in the number of features used. Third Step: The next step was to fit the training data set into the actual machine learning algorithm models which are Linear regression, MLP regression, Ridge regression, Support vector regression, and Random Forest regression model. Both linear and polynomial features were also tested to see the difference of result between both. Fourth Step: The fourth step was to evaluate and compare the results of the machine learning models with MSE, RMSE, and R2 scores so that further discussion can be made on which models work the best. Fifth Step: After the evaluation of the results

of the models, the final step was to choose the best algorithm to build the model to predict the future depreciation rate of the used cars.

RESULTS

Figure 5 below shows the relationship between the independent variables (transmission, year, mileage, mpg, engine size, tax) and the dependent variables (price). From Figure 5 below, the relationship of transmission and price shows that automatic and semiauto transmission had more used cars listed as more of a higher price compared to manual transmission cars. The year and price relationship shows a positively correlated relationship, which is expected as newer cars are more expensive. On the other hand, the relationship between price and mileage and the relationship between price and mpg show a negative correlation between independent and dependent variable. A lower mileage means that the used car is not driven or used that much, therefore producing a higher listing. For the relationship between price and engine size, it shows that the bigger the engine size, the more expensive the listing is therefore positively correlated. Lastly, for tax and price relationships, it appears that most of the higher listing used cars had taxes valued between 100 to 200.

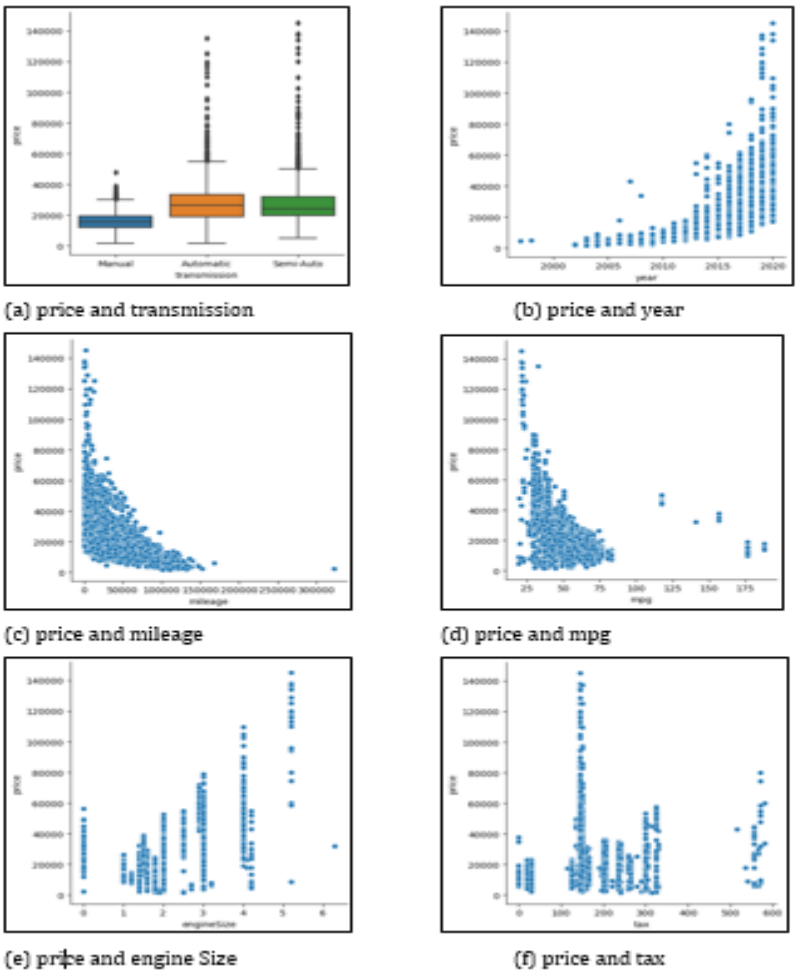

Fig. 5. Relationship between the independent and dependent variable.

Accuracy, MSE, and RMS analysis

In this research paper, this study has tested the accuracy, MSE and RMSE of these models which were Linear Regression, MLP Regression, Ridge, SVR and Random Forest Regression with two types of features, linear and polynomial. From Table 1, this study can see that Random Forest Regression has the highest accuracy and the lowest MSE and RMSE among all models. The lower the MSE and RMSE means, the better the model. For the polynomial feature, the accuracy of Ridge and SVR models is

increasing, but Random Forest was still the highest one which was 96%. Therefore, Random Forest has been chosen for this study final model to predict the car price depreciation rate.

Table 1: The result

Features	Algorithm type	Accuracy (%)	Mean Square Error	Root Mean Square Error
Linear	Linear Regression	88.22%	17793840	4218.274
	MLP Regression	12.46%	132298100	11502.09
	Ridge	88.15%	17905530	4231.493
	SVR	-00.35%	151667200	12315.32
	Random Forest	96.34%	5538473	2353.396
Polynomial	Linear Regression	-1.391358e+18	2.102752e+26	1.450087e+13
	MLP Regression	57.75%	63855090	7990.938
	Ridge	93.96%	9121508	3020.183
	SVR	2.50%	154904200	1244.605
	Random Forest	96.28%	5612518	2369.075

Accuracy, MSE and RMS Analysis

The results predicted by using Random Forest Regression were presented in Figure 6. The comparison between actual and predicted prices. From this graph, this study can see that most of the predicted prices are close to the actual price. Therefore, it means that this model accurately capturing the underlying patterns in the data, since their predicted price and actual are very close.

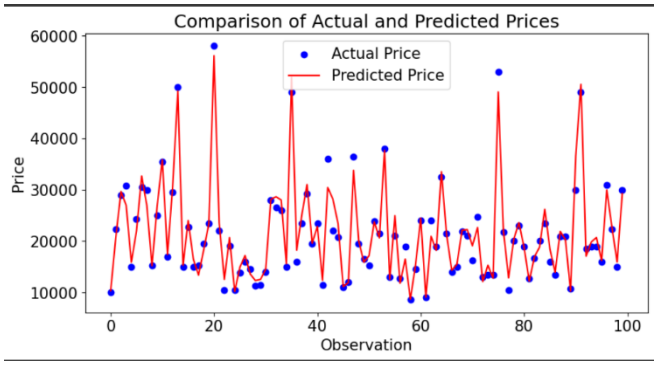

Fig. 6. Comparison of Actual and Predicted Prices.

Predict Car Price Depreciation Rate

Lastly, this study also predicting the depreciation rate based on how long the car has been bought (the data set was collected in 2020; by subtracting that with the car registration date, this study would get the 'age' of the car which this study added in a column called "age"). The way that this study would get the depreciation rate was by fitting the dataset that contains the age of the car so that the machine would know how to predict based on the age assuming the user wants to know the price of one of the cars in the dataset in the future (years). One of the existing cars in the data set is shown in Table 2.

Table 2: Sample record

model	year	price	Transmission	mileage	Fuel-Type	tax	mpg	Engine-Size	age
A1	2017	12500	Manual	15735	Petrol	150	55.4	1.4	3

The price of this used car that was previously listed was 12500 with an age of 3 years (2020-2017). Now after fitting the data set into the model, this study would then try input the same car details into the model but with the age of the car being 10 years old and allow the model to predict the price of it.

In this prediction, the random forest was used as a prediction model. The random forest predicted that the used car will be listed for 10759.317 which is a 13.94% drop. Since the original age of when the used car was listed was 3 years, this study subtract that by the age input by the user which is 10 years, this study would get 7 years. Using the result predicted by random forest model, this study would say that the price will fall at a rate of 1.99% annually (13.94 / 7).

DISCUSSION

The price of a used car is one of the key problems in the car field. Especially, there is more than one factor to influence the price of the used car. In this research paper, random forest regression was applied and developed for use in prediction since it has been proven that it is the best model to predict used car prices because it has the highest accuracy and the lowest MSE and RMSE among the other models. There are two types of characteristics that this study used, linear and polynomial. The performance of the model is presented in Table 3.

Table 3: Model Performance

Features	Algorithm type	Accuracy (%)	Mean Square Error	Root Mean Square Error
Linear	Random Forest	96.34%	5538473	2353.396
Polynomial	Random Forest	96.28%	5612518	2369.075

After exploring the relationship between the variables in this dataset, the car price is a dependent variable. This study also found that there was car model, year, engine size, mileage, mpg (miles per gallon), transmission, and fuel type were mostly influence the used car. The 'age' features were added as it is one of the price of the features that affected the most used car. Since the data set was from 2013 to 2020, this study used 2020 to subtract the car produce year of production of the car age. The correlation coefficient between the variables and the price of the car is presented in Table 4. According to Table 4, mileage, MPG and age exhibited negative correlations with car prices. On the other hand, years, tax and engine size had a positive correlation to car prices.

Table 4: Correlation coefficient of variables

Independent variables (numerical)	Dependent variable (price)
Year	0.59
Mileage	-0.54
Tax	0.36
MPG	-0.6
Engine size	0.59
Age	-0.59

However, the data set that this study used was not the most appropriate for the models to predict the future predictions in this study. AlShared (2021) had mentioned that in future work, there should be a real-time data set to increase the accuracy of the prediction. Thus, not only the historical data set was enough for the used car prediction, as it keeps updating the car price in different periods. In a nutshell, the combination of historical and real-time datasets will be a better recommendation for future predictions. Liu et al. (2022) had proven that set of algorithms as a new model. In their study, he combined PSO (Particle Swarm Optimization), GRA-BPNN (Grey Relational Analysis-Backpropagation Neural Network) as a new model called PSO-GRA-BPNN. This new model had higher prediction accuracy compared to BPNN, GRA-BPNN, and has a lower error range. Although it has a better performance, it has lost time and it is mainly analyzed from two as-pects which were hidden layer of BP (Backpropagation) neural network and iteration number and population size. If

the accuracy of the model is prior, the longer it took to process. With the increase of iteration and population size, the accuracy of model increases but at the point that value keeps continuously increasing, the accuracy might look almost stable.

Based on the studies that have mentioned above, a combination of historical and real-time dataset and ensemble of algorithms were recommended to be used in the prediction of the price of used cars. Lastly, most of the present studies that this study has mentioned in this study were conducted using the Random Forest Regression, a combination of single decision trees, and found that Random Forest had higher accuracy, lower MSE and RMSE values.

CONCLUSION

In this paper, five different models of machine learning algorithms were used to predict the used cars from the dataset "100,000 UK Used car dataset" that was obtained from Kaggle. The 5 models that were selected were Linear regression, MLP regression, Ridge, Support vector regression, and Random Forest regression. The brand of the car, the model, the year, the size of the engine and mileage were the independent variables of the price of the dependent variable. Pre-processing of the data set such as removal of duplicate records, removal of outliers, changing the features into numerical values, and converting data into linear and polynomial. The accuracy and performance of the models were also evaluated using metrics. Out of these 5 models, The Random Forest Regression Model outperformed the other 4 models, with accuracy being 96.34% for R2 score, Rs5538473 for MSE and Rs2353.396 for RMSE. The data set was also in polynomial form, which had an increase in accuracy compared to linear form. After the model evaluation, the random forest was then used to predict the depreciation rate of one of the cars used cars in the dataset. The result obtained from the model from predicting the used car was that the depreciation rate will depreciate by 13.94% after 10 years.

In conclusion, the Random Forest model that was built was overall very good at predicting the price of used cars compared to different models. However, one of the limitations of the Random Forest model is that it cannot predict the range of a whole new independent variable that is outside the range of the training data provided. This means that the result might not be accurate for that record. To improve the model, this study can combine it with other models, such as Gradient Boosting Regressor, to create an ensemble model. This can increase the accuracy of the model and reduce the risk of overfitting. For future work, this study wants to test out other models such as Gradient Boosting Regressor, Neural Network, K-means, KNN algorithms and more to compare the performance and find the most suitable for prediction. Finally, this study also hopes to develop a more accurate prediction model so that it can be used in the real world (Ahmad et al., 2024).

REFERENCES

Abhigyan. (2020). Detecting and Removing Outliers. https://medium.com/analytics-vidhya/detecting-and-removing-outliers-7b408b279c9

Acharya, S. (2021). What are RMSE and MAE? https://towardsdatascience.com/what-are-rmse-and-mae-e405ce230383

Aditya. (2020). 100,000 UK Used Car Data set. https://Www.Kaggle.Com/Datasets/Adityadesai13/Used-Car-Dataset-Ford-and-Mercedes

Ahmad, M., Ali, A., Tin, T. T., & Ali, A. (2024). Harnessing the Power of Geospatial Data: The Convergence of SDI, Big Data, and Cloud Computing. *Emerging Trends in Cloud Computing Analytics, Scalability, and Service Models* (pp. 312-330). IGI Global. https://doi.org/10.4018/979-8-3693-0900-1.ch016

Akhtar, T. (2020). Predicting Car Price using Machine Learning. https://Towardsdatascience.Com/Predicting-Car-Price-Using-Machine-Learning-8d2df3898f16

Alshared, A. (2021) Used Cars Price Prediction and Valuation using Data Mining Used Cars Price Prediction and Valuation using Data Mining Techniques. Rochester Institute of Technology. Thesis. https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12220&context=theses

Anil, P. A. (2020). Used Car Price Prediction using Machine Learning. https://towardsdatascience.com/used-car-price-prediction-using-machine-learning-e3be02d977b2

Ashok, P. (2024). What is Ridge Regression? Great Learning Team. https://www.mygreatlearning.com/blog/what-is-ridge-regression/#:~:text=Ridge%20regression%20is%20a%20model,away%20from%20the%20actual%20values

Bobbitt, Z. (2020). What is Considered to Be a "Strong" Correlation? Statology. https://www.statology.org/What-Is-a-Strong-Correlation/#:~:Text=Strong%20positive%20correlation%3A%20When%20the

Bobbitt, Z. (2021). MSE vs. RMSE: Which Metric Should You Use? Statology. https://www.statology.org/Mse-vs-Rmse/

Brownlee, J. (2017). Why One-Hot Encode Data in Machine Learning? Https://Machinelearningmastery.Com/Why-One-Hot-Encode-Data-in-Machine-Learning/

CFI Team. (2022). Negative Correlation. https://corporatefinanceinstitute.com/resources/data-science/negative-correlation/

Chaw, J. K., Chaw, S. H., Quah, C. H., Sahrani, S., Ang, M. C., Zhao, Y., & Ting, T. T. (2024). A predictive analytics model using machine learning algorithms to estimate the risk of shock development among dengue patients. Healthcare Analytics, 5, 100290. https://doi.org/10.1016/j.health.2023.100290

Chen, C., Hao, L., & Xu, C. (2017). Comparative analysis of used car price evaluation models. In *AIP Conference Proceedings* (Vol. 1839, No. 1). AIP Publishing. https://doi.org/10.1063/1.4982530

Collard, M. (2022). Price prediction for used cars: a comparison of machine learning regression models. (Dissertation). https://urn.kb.se/resolve?urn=urn:nbn:se:miun:diva-45286

Dua, R., Ghotra, M. S., & Pentreath, N. (2017). *Machine Learning with Spark*. Packt Publishing Ltd.

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. In Frontiers in Energy Research (Vol. 9). Frontiers Media S.A. https://doi.org/10.3389/fenrg.2021.652801

Gajera, P., Gondaliya, A., & Kavathiya, J. (2021). Old car price prediction with machine learning. *International Research Journal of Modernization in Engineering Technology and Science. Sci*, *3*, 284-290.

Geeksforgeeks. (2023). k-nearest neighbor algorithm in Python. Https://Www.Geeksforgeeks.Org/k-Nearest-Neighbor-Algorithm-in-Python/

Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car price prediction using machine learning techniques. *TEM Journal*, 8(1), 113–118. https://doi.org/10.18421/TEM81-16

Harvey, H. B., & Sotardi, S. T. (2018). The Pareto Principle. Journal of the American College of Radiology, 15(6), 931

Kuiper, S. (2008). Introduction to Multiple Regression: How Much Is Your Car Worth? https://www.researchgate.net/publication/265656449_Introduction_to_Multiple_Regression_How_Much_Is_Your_Car_Worth

Kumar, A. (2022). Ridge Regression Concepts & Python example. Https://Vitalflux.Com/Ridge-Regression-Concepts-Python-Example/

Liu, E., Li, J., Zheng, A., Liu, H., & Jiang, T. (2022). Research on the prediction model of the used car price in view of the pso-gra-bp neural network. *Sustainability*, *14*(15), 8993. https://doi.org/10.3390/su14158993

Low-Level Academy. (2020). Number Encoding. https://lowlvl.org/prerequisites/binary-and-hexadecimal-numbers

Meimi Li. (2021). Predicting Vehicle Price With Random Forest Regressor. https://medium.com/analytics-vidhya/predicting-vehicle-price-with-random-forest-regressor-d1c272668be5

Milunovich, G., Wu, L., & Zhao, Y. (2023). Forecasting the Prices of Used Cars: A Comparative Analysis of Supervised Learning Algorithms. *SSRN 4334750*. https://dx.doi.org/10.2139/ssrn.4334750

Mordor Intelligence, *Malaysia used car market size & share analysis – growth trends & forecasts (2024-2029)*. (2024). https://www.mordorintelligence.com/industry-reports/malaysia-used-car-market.

Özçalıcı, M. (2017). Predicting second-hand car sales price using decision trees and genetic algorithms. *Alphanumeric Journal*, *5*(1), 103-114. https://doi.org/10.17093/alphanumeric.323836

Pham, B. T., Le, L. M., Le, T. T., Bui, K. T. T., Le, V. M., Ly, H. B., & Prakash, I. (2020). Development of advanced artificial intelligence models for daily rainfall prediction. *Atmospheric Research*, *237*, 104845. https://doi.org/10.1016/j.atmosres.2020.104845

Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. *Int. J. Inf. Comput. Technol*, *4*(7), 753-764.

Raj, A. (2020). Unlocking the True Power of Support Vector Regression. Using Support Vector Machine for Regression Problems. https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0#:~:Text=Support%20Vector%20Regression%20is%20a,Find%20the%20best%20fit%20line.

Raj, P., & David, P. E. (2020). *The digital twin paradigm for smarter systems and environments: The industry use cases*. Academic Press.

Sharma, R. (2022). Classification in Data Mining Explained: Types, Classifiers & Applications. https://www.upgrad.com/blog/classification-in-data-mining/

Sharp, T. (2020). An Introduction to Support Vector Regression (SVR). https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2

Ting, T. T., Wei, C. J., Min, O. T., Mooi, L. S., Tiung, L. K., Aitizaz, A., ... & Salau, A. O. (2024). Visualization of Personality and Phobia Type Clustering with GMM and Spectral. International Journal of Advanced Computer Science & Applications, 15(9). https://doi.org/10.14569/ijacsa.2024.0150988

Xu, M., Tang, W., & Zhou, C. (2021). Operation strategy under additional service and refurbishing effort in online second-hand market. *Journal of cleaner production*, *290*, 125608. https://doi.org/10.1016/j.jclepro.2020.125608

Yao, Z., Zhang, T., Wang, J., & Zhu, L. (2019). A feature selection approach based on grey relational analysis for within-project software defect prediction. *Journal of Grey System*, *31*(3).