Comparing Regression and Neural Network Models in Used Car Price Prediction

1. Introduction

Artificial Intelligence (AI) is gradually permeating every area of life from agriculture to health, law to finance, social media to education, manufacturing to entertainment, etc. Centrally, artificial intelligence is seen as the simulation of human intelligence processes by computer systems. AI use includes the chatbot that assists on a webpage, the Google map that helps direction, to more advanced use in detection of diseases in humans and farms as well as price forecasting. Just like the calculator has not replaced man in doing mathematics, AI does not aim to replace man but rather to assist human in daily activities to be proficient, and thereby accomplishing more.

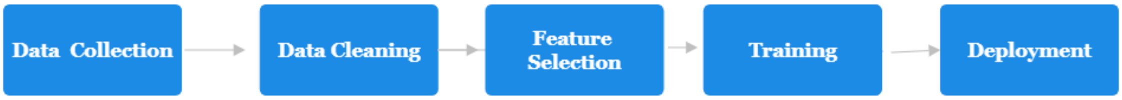
In prediction, computer systems are fed some features of an item or event, and these systems make a forecast about a chosen target (price in this experiment). With increasing work schedule and abundance of goods in the market, it has become particularly useful for both seller and buyer to use a system to guide them in sale and purchase of goods especially used goods like cars. In law, a buyer of used goods is expected to have confirmed through inspection that the goods offered by a seller is fit for use as the law cannot come to rescue the buyer if after purchase, the buyer turns around to complain about any feature of the used goods, including the price, except if the buyer can substantiate misrepresentation or fraud on the part of the seller (see UK case of **Caveat Emptor v. Godbold (1807),** and US case of **Hurley v. Ball (1859)**).

2. Related Work

There exist many materials where scholars and professional have carried out similar tasks in making prediction on used cars and other materials. Kang et al. (2022) in their work used LR, RF, GBT, and FML regression models in making prediction on the price of used cars. At the end, they found that GBT has the highest R2 score of 0.85. However, they considered both R2 score and computing time, and concluded that RF with R2 score of 0.83 and less computing time of 40 secs as against 70 to 355 of the GBT is a better model. Abdullah1 and Salah (2023) in the face of market fluctuation produced a hybrid model CNN-LSTM to provide better predictions to guide investors in the money market. They found that this hybrid model outperforms traditional, statistical, ML, and DL models as the convolutional ability of CNN to dig out useful features hidden in data plus the advantage of LSTM in identifying long-term dependencies in data will bring about more accuracy in price prediction. Similarly, Limsombunchai, Gan and Lee (2004) compared prediction using the hedonistic theory and ANN. Hedonistic theory assume a bundle of features to determine the price of an item. In the instant case, the price of a house is determined by the features of the house like the number of rooms, bathrooms, parking facilities, size of the compound and other facilities present in the house. In the end, they found that the R2 of ANN is higher than the R2 of hedonistic theory. The higher values of R2 for ANN were obtained from tuning and the higher value may not be the case without the tuning which produced a model with higher R2 over the hedonistic model. Zhu(2023) used SVM, XGBoost and neural network to predict the price of used cars. His research found that XGBoost gave a better prediction. It is noted that this work did not mention the neural network used nor did the work provide information on whether any of this model was tuned or otherwise.

Building on the shoulders of my predecessors, I used simple linear regression (SLR), simple polynomial regression (SPR), multiple linear regression (MLR), random forest (RF), support vector machines (SMV), gradient boosting regression (GBR) and artificial neural network (ANN) for prediction of used car price. Hyperparameter tuning (for ANN) and GridSearchCV (for SVM, RF and GBR) were invoked to tune some models for better performance. The models were evaluated to identify the model best suited for prediction. This work aimed at evaluating the supervised learning algorithm that is best suited for prediction and paid attention to the ensemble/boosting algorithm in RF/GBR and then verified how these models compared with ANN.

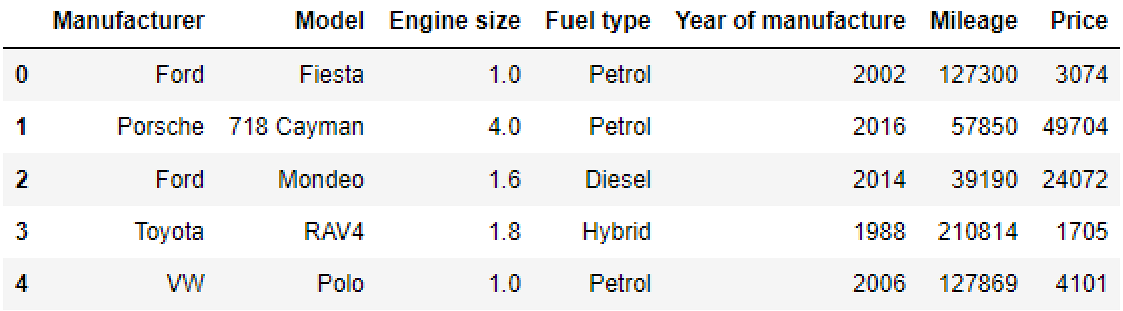
3. Workflow



i. Data Collection: The first stage involved collecting data. In this case data used was sourced from a previous classwork.
ii. Data Cleaning: Following data acquisition, data was checked to ensure it was clean and fit for use.
iii. Feature selection: Features to be used for the experiment were identified. The goal was to see which feature or number of features helped to improve the performance of the ML algorithm.
iv. Training: Data was split and 80% was used for training and 20% for test.
v. Deployment: the model was deployed after test to make prediction of the target variable

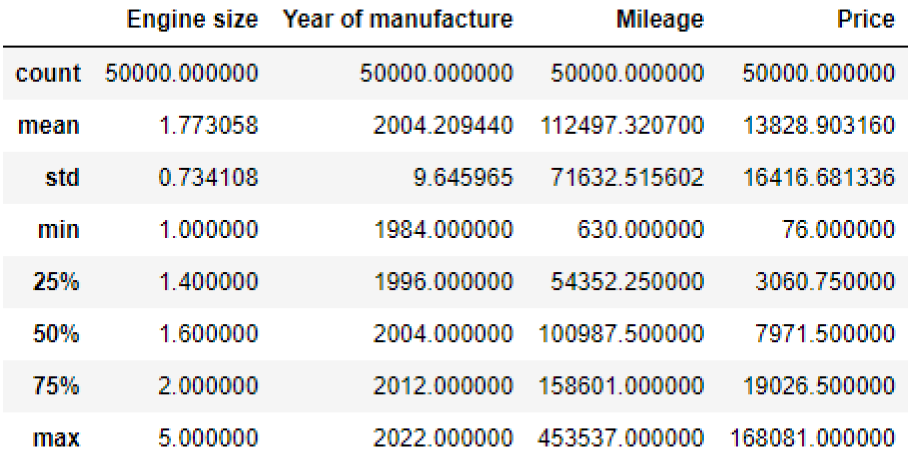
4. Data Exploration

The data contained in the csv file has a total of 50000 used car records. It has a total of 7 columns. A tabular representation of our data is best shown by a pandas dataframe (table) as seen below. This was arrived at using by info() method in pandas that by default shows the first five rows of the dataframe:



	Manufacturer	Model	Engine size	Fuel type	Year of manufacture	Mileage	Price
0	Ford	Fiesta	1.0	Petrol	2002	127300	3074
1	Porsche	718 Cayman	4.0	Petrol	2016	57850	49704
2	Ford	Mondeo	1.6	Diesel	2014	39190	24072
3	Toyota	RAV4	1.8	Hybrid	1988	210814	1705
4	VW	Polo	1.0	Petrol	2006	127869	4101

Using the describe() method, we get a summary of all the numeric columns



	Engine size	Year of manufacture	Mileage	Price
count	50000.000000	50000.000000	50000.000000	50000.000000
mean	1.773058	2004.209440	112497.320700	13828.903160
std	0.734108	9.645965	71632.515602	16416.681336
min	1.000000	1984.000000	630.000000	76.000000
25%	1.400000	1996.000000	54352.250000	3060.750000
50%	1.600000	2004.000000	100987.500000	7971.500000
75%	2.000000	2012.000000	158601.000000	19026.500000
max	5.000000	2022.000000	453537.000000	168081.000000

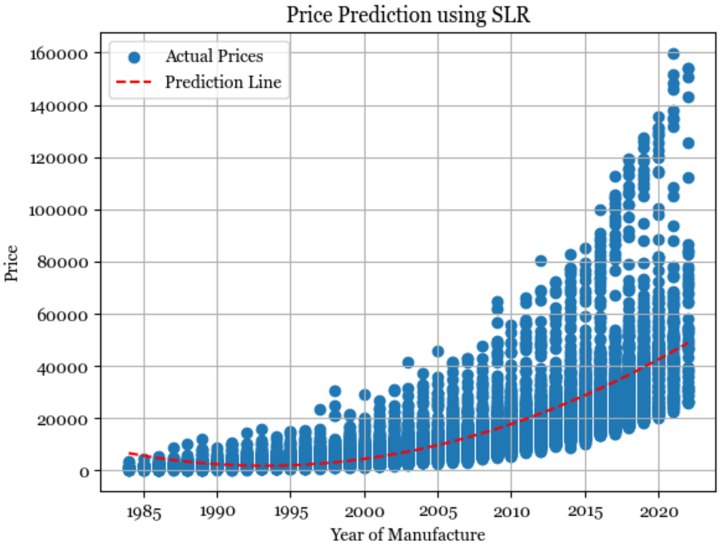
5. Evaluation Metrics

Metrics used to evaluate models include Root Mean Square Error (RMSE), Symmetric Mean Absolute Percentage Error (SMAPE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), and Coefficient of Determination known as R2 score. Chicco, Warrens and Jurman (2021) argue that R2 indicate how well a model responds to new observation and remarked R2 is a better metric when compared to SMAPE, RMSE, MAE, MAPE, and MSE. A model is good if it accurately predicts the output of a test set and to the extent it matches the ground truth (Russel and Norvig, 2021). A good model ought to show the prediction line near majority of the data points. The closer the R2 score is nearer to 1 the better the model prediction. Apart from R2, MSE will also be used here to measure error of each model. MSE is the mean of the squared errors. The squared term penalises large errors; the closer MSE is to 0, the better the model.

6. Experiments

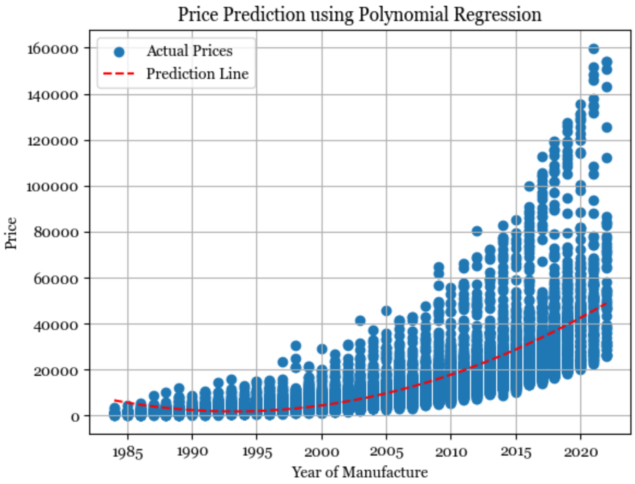
a. **Simple Linear Regression (SLR**)

SLR is a supervised learning model. It shows the relationship between one variable called the independent variable (input variable, features, or predictors) and another variable called the dependent variable (target, output, response, or outcome variable). In this experiment, the 3 numeric predictor columns were each used to predict the target variable (Price). The best model here was the prediction done with the 'Year of manufacture' feature which has an R2 score of 0.51 and MSE of 132678999.95. Visually, this model is represented by the plot below. The low R2 score is validated by the many data points that are far from the prediction line. This model is not close to the ground truth.



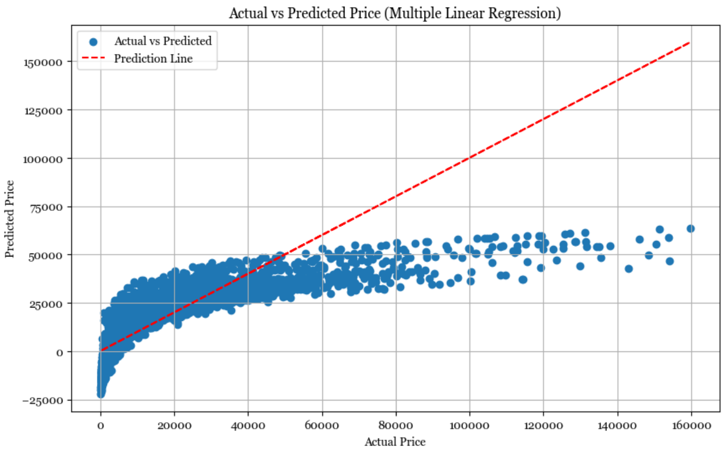
b. Simple Polynomial Regression (SPR)

SPR belongs to the supervised learning family. Like SLR, SPR shows the relationship between a predictor and another variable called the target variable. Polynomial regression is used when there is no linear relationship between the variables. The relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x. Like in SLR above, the 3 numeric predictor columns were each used to predict the target variable (Price). The ‘Year of Manufacture’, again, was a better predictor as in SLR with R2 score of 0.60941 and lesser MSE 105,993,894.2019.



c. Multiple Linear Regression (MLR)

MLR is another supervised learning algorithm like SLR used for regression. While MLR shows the relationship between variables like SLR, MLR takes 2 or more input variables or predictors. In this case, all the 3 numeric predictor columns were jointly used to predict the target variable (Price). Evaluation shows an MSE of 89158615.76017143 and an R2 score of 0.671456306417368. The plot below confirms the low R2 score as most the data points are far from the prediction line.



d. Support Vector Machines (SVM)

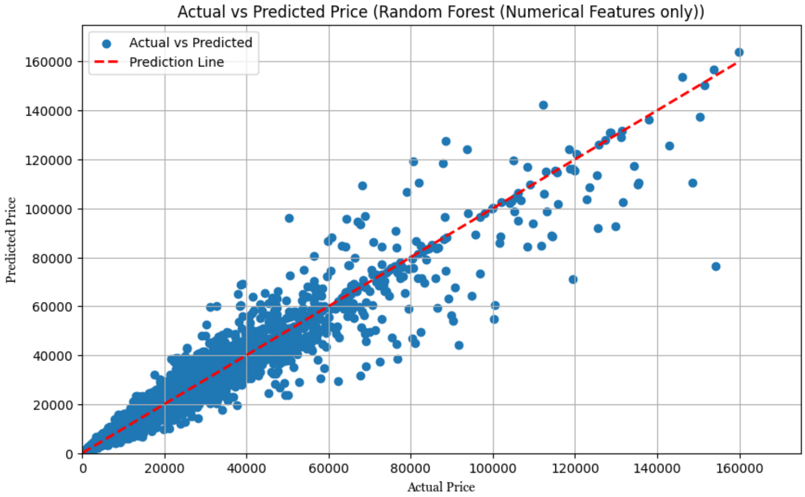
SVM is a supervised learning algorithm originally designed for predicting numerical and categorical target but presently used for classification (Theobald, 2017). The SVM algorithm create a decision boundary called the hyperplane. As this hyperplane is established, new instances are then classified into matching group. It handles high-dimensional data and is very efficient in cases where there is a clear boundary between groups of data like in spam and non-spam.

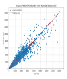
In this experiment, SVM behaved differently from others with regards to inclusion of more variables. While with numerical variables alone it had an R2 score of 0.20784375447380532 and MSE of 214971572.11184293, it recorded an R2 score of 0.09096523576446158 and MSE of 246689505.35411796 with addition categorical variables to the model. This departs from all other algorithms used in this work which usually record an improved R2 score with addition of categorical features to the numerical features in building a model. Upon tuning of the model with all features using Gridsearch CV we recorded some improvement as R2 came to 0.5811702329834011 and MSE of 113660018.42602134.

e. Random Forest (RF/Num Variables Only)

Still continuing with supervised learning model, RF was used in this experiment. RF operates on the concept of ensemble technique. A forest we know is made up of many trees. True to this picture of a forest, RF works by training many decision trees on a dataset and then arrives at a result by voting -i.e., it computes the average of the prediction of each of these trees as its prediction. RF is used for both classification and regression. It takes care of overfitting and handles high dimensional data very well.

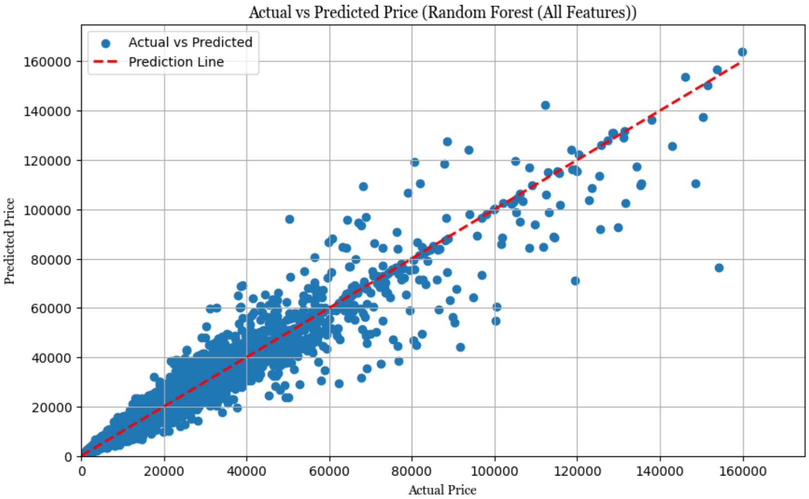
In this inquiry with only numerical variables, MSE was 20159131.522860516 and of R2 score was: 0.9257149129843565. Upon tuning, we got an R2 score of 0.9411 and MSE of 15,994,568.8387. We are beginning to see majority of the data points near the prediction line confirming the higher value of the R2 score. See below plot.





f. Random Forest (RF/Numerical + Categorical)

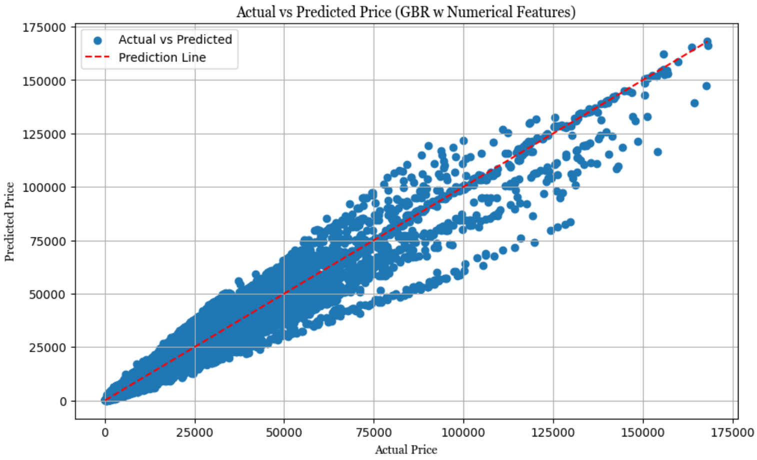
This trial processed to check the effect of adding categorical features in making prediction with RF. It was discovered that the inclusion of categorical variables produced a model which had MSE score of 14529057.213402066 and R2 score of 0.9464613702168213. This is shown graphically in the plot below:



g. Gradient Boosting Regression (GBR/Numerical Only)

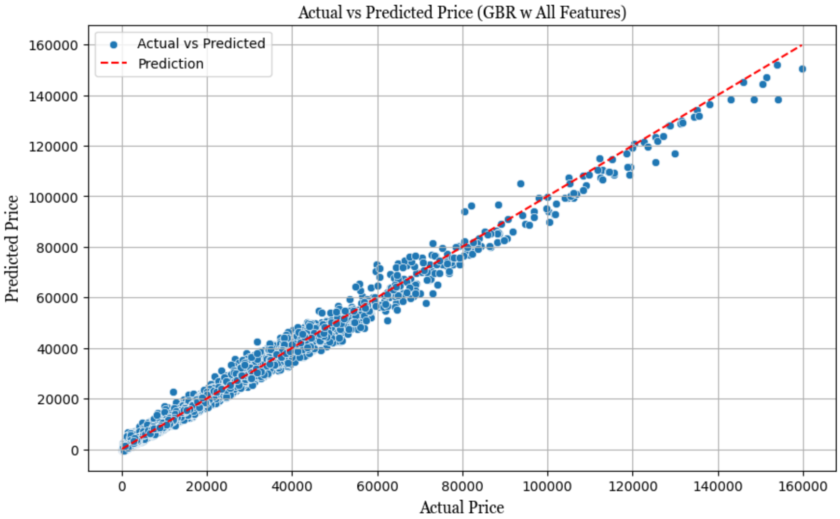
This experiment moved on to consider prediction with GBR. GBR is a supervised learning algorithm that runs on boosting algorithm. Like RF in operation, it also combines the output of several trees as its prediction. Author, Theobald (2017), describes boosting algorithm by liking it to a teacher offering extra classes to students whose performance were poor in an examination with the intention of improving class performance. GBR starts by fitting a model to data, and then successively adds more model to the data, and each time improving on the errors of the predecessor model. In the end, it combines the predictions of these trees to get its pred iction. GBR is ideal for building models with high degree of accuracy. It can be used in many regression problems, and it handles missing values well. Kang et al. (2022) remarked that it can consume significant computation time while working with big data.

With numerical features, the trial recorded an R2 score of 0.9453864886338167 and MSE of 14718452.503718982. Upon tuning using GridSearch CV, marginal improvement was seen with R2 score of 0.9496392419300299 and MSE of 13572326.83198446



h. Gradient Boosting Regression (GBR with all Features)

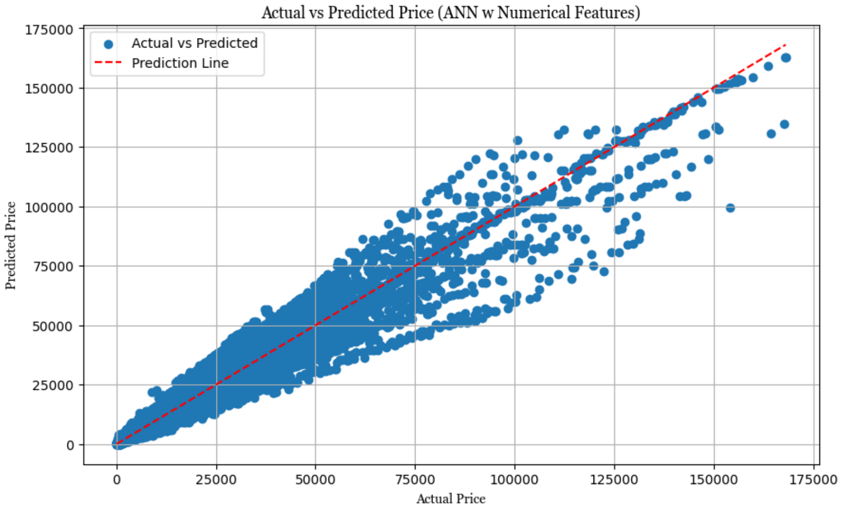
Adding categorical features to test produced a significant improvement and the resultant model had an R2 score of 0.9899257725078963 and MSE of 2733895.660131188



f. Artificial Neural Network (CNN/Numerical Features Only)

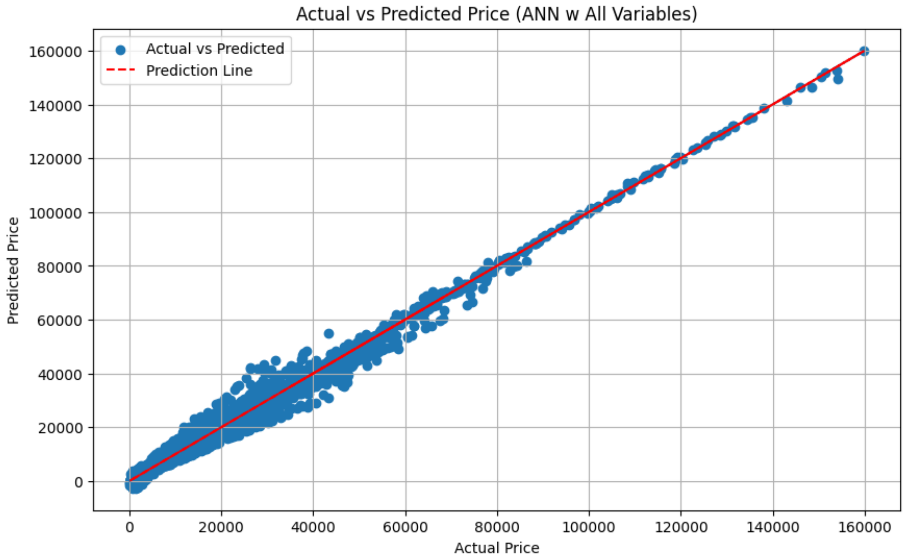
Lastly, this trial built an ANN model to make prediction and to compare with preceding models. ANN is an unsupervised learning approach. It is a computational model that is patterned to mirror the operations of neurons in animals. It can be described as the mathematical model of the human brain, and it arrives at a decision following the way a human brain processes information. ANN is suitable for large data sets and when there is no linear relationship between data, and it discovers hidden connection between data through hidden layers on artificial neurons. It consumes more computing time than other models except when GridSearchCV was used in tuning. It is used in both regression and classification.

The test carried out with numerical features produced a model with R2 score of 00.928 504 and MSE of 21772496.34144821. On tuning hyperparameters, we recorded an improved R2 o f 0.9464613702168213 and MSE of 14529057.213402066



f. Artificial Neural Network (ANN/ All Features)

Adding the categorical features to this experiment we arrived at a very improved mode which R2 score was 0.990059303751474 with MSE of 2697658.590083296. The proximity of the predictor line t o the data point confirms the high R2 score.



7. Evaluation of the Models

The table below presents the models and the 2 metrics (R2 and MSE) that we have used to measure the model performances. SVM performed worst in the class of models where all features were used to make a prediction, GBR is the best when only numeric features were used for the test. Overall, ANN with all the features is the best model with R2 score of 0.9908.

S/N	Algorithm	MSE	R2
1	SLR	132,678,999.9479	0.5111
2	SPR	105,993,894.2019	0.6094
3	MLR	89,158,615.7601	0.6715
4	SVM (Numeric Features)	214,971,572.1118	0.2078
5	SVM (All Features)	113,660,018.4260	0.5812
6	RF (Numeric Features)	15,994,568.8387	0.9411
7	RF (All Features)	14,529,057.2134	0.9465
8	GBR (Numeric Features)	13,572,326.8320	0.9496
9	GBR (All Features)	2,733,895.6601	0.9899
10	ANN (Numeric Features)	14,529,057.2134	0.9464
11	ANN (All Features)	2,500,667.5809	0.9908

8. Conclusion

This experiment has proved that increase in relevant features in modelling leads to improved prediction. Except for SVM, the addition of more features to models resulted in an improved model, and this is close to ground truth where buyers usually consider mileage, model, manufacturer, number of doors, colour, year of manufacture, etc in considering a purchase and sale of a used car. SVM performance proved that it is not an ideal candidate for regression, and it is mostly used today for classification. GBR outperformed RF on all fronts. GBR also performed better than ANN when only numerical features where used. While ANN produced the best model when all features were used in building a model, GBR makes a better model when data size is not large and when computing time is a constraint.

In the future, it will be fitting to extend predictive analysis to disease management in a developing country. The goal will be to provide more information to officials in these areas engaged with disease control. Future studies plan to use a hybrid model in exploring prediction especially where regression may not be an ideal candidate like in time series analysis.

References

Russell, S. J. & Norvig, P, (2021) Artificial intelligence: a modern approach, Global:4th edition. Pearson, Harlow.

Theobald, O. (2017) Machine Learning for Absolute Beginners, 2nd edition. Great Britain: Amazon.

Abdullah, W & Salah, A. (2023) A Novel Hybrid Deep Learning Model for Price Prediction. *International Journal of Electrical and Computer Engineering (IJECE)* Vol. 13, No. 3. Available online: DOI: 10.11591/ijece.v13i3.pp3420-3431 [Accessed 14/02/2024].

Chicco, D., Warrens, M. J., & Jurman, G. (2021) The coefficient of determination R-squared is more in formative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Compu ter Science. Available online:* doi: 10.7717/peerj-cs.623 [Accessed 10/01/2024].

Kang, J.I., Parekh, H., Ramdas, P., Lee, S. & Woo, J. (2022) Comparing Regression Models Predicting the Price of Used Cars in Big Data, *IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pp. 01-04. Available online: doi: 10.1109/ICCE-Asia57006.2022.9954633 [Accessed 14/02/2024].

Limsombunchai, V., Gan C. & Lee M. (2004) House Price Prediction: Hedonic Price Model vs. Artificial Neural Network *American Journal of Applied Sciences* 1(3). Available online: DOI:10.3844/ajassp.2004.193.201 [Accessed 14/02/2024]

Zhu, Y. (2023). Prediction of the price of used cars based on machine learning algorithms. *Applied and Computational Engineerin*g. Available online: DOI:10.54254/2755-2721/6/20230917 [Accessed 2/01/2024].