Prediction of monthly vehicle valorization/devaluation in Brazil with a MultiLayer Perceptron Regressor: a case study based on past sales, inflation, and interest rate

André Roberto Ortoncelli[1], **Franciele Beal**[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
Estrada para Boa Esperança, Km 04 – CEP 85.660-000 – Dois Vizinhos – PR– Brazil

{ortoncelli,fbeal}@utfpr.edu.br

Abstract. *This work presents a comparison between the valuation/depreciation prediction results (from one month to another) of vehicles in Brazil considering the combination of four groups of characteristics: i) previous sales; ii) the number of vehicle sales; iii) basic interest rate; and iv) national consumer price index. We create a comparison baseline training a MultiLayer Perceptron Regressor (MLPR) based only on the vehicle's value in the previous month, and then we train the MLPR by combining the previous vehicle value with combinations of the characteristic groups. Experiments were performed from 2013 to 2022 and evaluated in terms of Mean Squared Error (MSR) and Median Absolute Error (MAE). The combination of characteristics that presented the best MSR for the 2018-2022 period (COVID-19 period) was among the worst from 2014 to 2017. It is possibly concluded that data scientists must periodically adjust parameters according to the current economic conditions to obtain the best automatic forecast results of the monthly valorization/depreciation of vehicles in Brazil.*

1. Introduction

Depreciation is an asset's value loss due to its use, natural wear, or obsolescence. Knowing the vehicle accuracy pattern is essential for buyers and other stakeholders, such as financial institutions. In this context, predicting the future price of a car is of interest to consumers and companies that sell or buy vehicles.

The COVID-19 pandemic had an economic impact on several sectors, including the automobile market, with an emphasis on a decrease in production and commercialization in the sale of new cars, causing an increase in demand for used vehicles and consequent an increase in their price. Some surveys indicate that consumers have changed their buying habits during the pandemic, which impacted vehicle sales [Raza and Masmoudi 2020].

In this context, predicting the market value of used vehicles is a non-trivial task, as it depends on related to natural wear and can also be affected by the world's economic aspects. These challenges motivated this work, which presents a study of the impact of different characteristics groups in the prediction of cars prices variation.

We evaluated the impact of different sets of features in predicting the percentage of depreciation/valorization of a vehicle from one month to another with a MultiLayer

Perceptron Regressor (MLPR). To create a comparison baseline, we first train an MLPR using only one characteristic, referring to its value in the immediate previous month. We then combined the prior value of the vehicle with features related to previous car sales, changes in the country's basic interest rate, and a national price index.

We chose these parameter sets because the number of cars sold and economic variations can impact the value of used vehicles. To the best of our knowledge, this is the first work that presents a combination of these groups of characteristics to predict vehicle valuation/devaluation in Brazil.

In experiments, we try to predict the mean car price variation between 2013 and 2022 (10 years). Results were evaluated regarding Mean Squared Error (MSR) and Median Absolute Error (MAE). We can observe that the combination of characteristics that presented the best MSR from 2018 to 2022 (COVID-19 period) was the worst performance between 2014 and 2017. Our results allow us to analyze the impact of the combination of different characteristics in predicting average vehicle sales value over ten years.

The remainder of this paper is organized as follows. In Section 2 as the theoretical aspects necessary for understanding the work. Related works are in Section 3. Details about the experimental methodology are in Section 4. Section 5 has results and a discussion about them. Finally, Section 6 concludes the paper with some final remarks and suggests possible future works.

2. Theoretical Aspects

This section presents theoretical aspects for understanding this work. Subsection 2.1 defines an MLPR. Subsections 2.2 to 2.4 describe the indices/rates explored in the experimental dataset.

2.1. Multi-Layer Perceptron Regressor (MLPR)

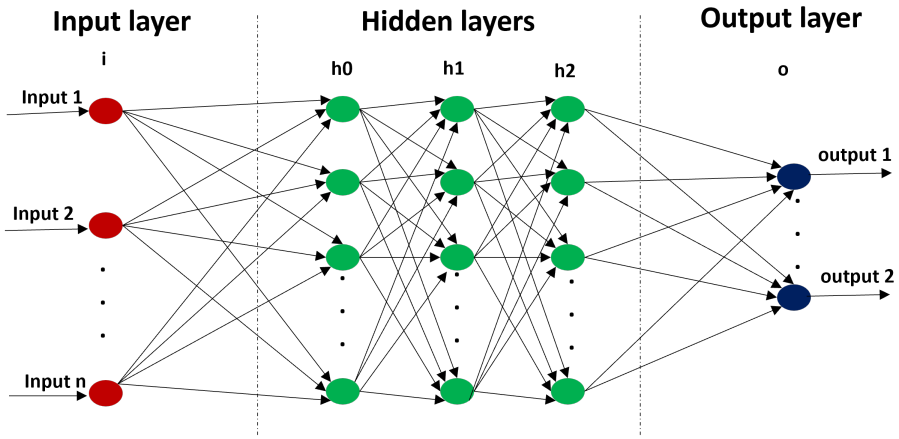
Supervised learning is a subcategory of Machine Learning in which the computer is trained with a dataset with n features ($n > 0$) to identify a given value related to each dataset instance. An Artificial Neural Network (ANN) is a biologically inspired supervised learning algorithm that processes information similarly to human brain neurons. An ANN is a good alternative for problems that are difficult to solve with standard techniques, such as pattern recognition and predictive analysis.

The ANN structure is composed of three layers: i) input layer: composed of n neurons that receive each of the n dataset characteristics; ii) hidden layers: one or more neurons layers between the input and output layers; iii) output layer: returns the final result based on the data transformation performed in the previous layers [Tyagi and Abraham 2022]. Figure 1 shows an ANN architecture.

In an ANN, each neuron returns a value that is the sum of the input values multiplied by a given weight and a bias value. An activation function is applied to the neuron's output to transform it into a non-linear value. The weights and bias values are defined in the training step [Sharma et al. 2017].

A MultiLayer perceptron (MLP) is a fully connected class of ANN. In this work, we use an MLPR – a type of MLP that trains using backpropagation with no activation function in the output layer, in this way, the output value can represent the percentage

Figure 1. Architecture of ANN, adapted from [Tyagi and Abraham 2022]



of car price variation from one month to another. We use the MLPR implemented in the Scikit-learn[1] library, with its default configuration.

In this work, we use an MLPR trained with data from average vehicle sales price (Subsection 2.2), basic interest rate (Subsection 2.3), and national consumer price index (Subsection 2.4).

2.2. Average vehicle sales price

In Brazil, the Economic Research Institute Foundation (FIPE) maintains the FIPE reference table with the average vehicle sales prices in the national market. This table serves as a parameter for negotiations or evaluations.

On the official FIPE page[2], it is to query the average sales value of a vehicle, reporting the reference month, the automaker, the model, and the year of manufacture.

2.3. Basic interest rate

Selic is the basic interest rate of the Brazil economy and influences all interest rates practiced in the country, being an acronym for "Sistema Especial de Liquidação e de Custódia" (in English, Special System of Settlement and Custody), a system managed by the Central Bank.

2.4. National consumer price index

The Extended Consumer Price Index (in Portuguese, Índice de Preços ao Consumidor Amplo – IPCA) is calculated by the Instituto Brasileiro de Geografia e Estatística and is one of the main metrics related to the inflation rate in Brazil.

To calculate the IPCA, the IBGE carries out a monthly survey, in 13 urban areas of the country, of approximately 430 thousand prices in 30 thousand locations. All these prices are compared with the previous month's prices, resulting in a single value that reflects the general variation of consumer prices in the period.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

[2] https://veiculos.fipe.org.br/

3. Related Works

Machine Learning methods have already been explored with different features to automatically predict used car prices with datasets collected in different countries, such as Arabia [Brahimi 2022], India [Varshitha et al. 2022], China [Chen et al. 2017], Malaysia [Khan 2022], and the United States of America (USA) [Pai and Liu 2018].

In [Chen et al. 2017], records of over 100,000 sales of used cars in China validate the empirical results of two algorithms: Linear Regression and Random Forest. Price prediction was carried out in three groups: i) model for a specific car make; ii) model for a particular car series; and iii) universal model. Random Forest had superior results for most experiments.

An approach for car value estimation based on Linear Regression is presented in [Khan 2022]. This approach uses characteristics about the car's maker, model, mileage, manufacture year, engine displacement, engine power, body type, transmission, and combustion type. With a dataset from Malaysian, the authors obtained an accuracy of 79% and identified that the most relevant features were related to car brand, model, and age.

Another car price predictor is in [Varshitha et al. 2022], which uses Artificial Neural Network and Random Forest (which had the best results) to predict the car value-based on characteristics related to the sale price, mileage, type of fuel, the buyer (individual or legal entity), type of transmission, and the number of previous owners.

In addition to works that use datasets with characteristics about the vehicle and previous sales values history, there are also works that innovate in the data types and approaches for extracting characteristics, exploring different techniques to collect reference information, like Web Scrapping [Nasiboglu and Akdogan 2020] and Text Mining [Brahimi 2022], and analysis of sentiment scores of tweets [Pai and Liu 2018].

In [Nasiboglu and Akdogan 2020], Web Scrapping (with Beautiful Soup and Selenium libraries) was explored to extract features from online car sales websites. Twelve characteristics were extracted to train Machine Learning algorithms: Linear Regression, Ridge, Lasso, Elastic Net, KNN, Random Forest, XGBoost, and Gradient Boosting Machine. For most of the experiments carried out, the best results were obtained with the Gradient Boosting Machine.

Also, based on the analysis of vehicle advertisement content, [Brahimi 2022] uses Text Mining techniques for improving price prediction. For the experiments, car sales advertisements and lots of equipment were collected, applying four prediction algorithms (Linear Regression, Pace Regression, KNN, and Neural Networks) to estimate prices. Results indicate that the integration of text mining techniques significantly improves prediction.

In [Pai and Liu 2018], opinions posted on social networks (tweets) are an indicator for predicting sales, combined with other factors that influence the purchasing power of vehicles, such as stock market values. Multivariate regression models were trained with three types of data (sentiment scores of tweets, stock market values, and hybrid data) to forecast monthly total vehicle sales in the USA.

There are other reports about Machine Learning applied to predict vehicle prices, but, to the best of our knowledge, this is the first work that presents an analysis of the

combination of characteristics groups related to past sales, inflation, and basic interest rate (better described in the Subsection 4.1) to predict vehicle cars' valuation/devaluation in Brazil.

4. Methodology

For our experiments, we implemented a software in the Python programming language. We use the MLPR of the Scikit-Learn library (with the default parameters). The strategy used to produce the experimental dataset is in Section 4.1. Section 4.2 is the evaluation metrics. The approach used to create the training and test sets is in Section 4.3.

4.1. Experimental dataset

Our experimental dataset was composed of four data groups extracted from three other datasets. We describe each data group below:

- $\mathbf{C_1}$: from the dataset "Tabela Fipe – Histórico de Preços"[3], we extracted the following characteristics, considering that we want to compute the vehicle price variation in the month m:
  - c_{1a}: the car age;
  - c_{1b}: car value in previous month;
  - c_{1c}: the % of depreciation/appreciation of the car in previous month in relation to the month m-3; and
  - c_{1d}: the % of depreciation/appreciation of the car in previous month in relation to the month m-6.
- $\mathbf{C_2}$: from the dataset "Brazil car sales records from 1990 to 2022"[4], we extracted the following characteristics:
  - c_{2a}: the amount of vehicle sales in the previous month;
  - c_{2b}: average amount of vehicle sales in the three previous months; and
  - c_{2c}: average amount of vehicle sales in the previous six months.
- $\mathbf{C_3}$:from the dataset "SELIC & IPCA - Série Mensal Histórica"[5], we extracted the following characteristics:
  - c_{3a}: SELIC rate in the previous month;
  - c_{3b}: SELIC rate three months ago; and
  - c_{3c}: SELIC rate six months ago.
- $\mathbf{C_4}$: also from the dataset "SELIC & IPCA - Série Mensal Histórica", we extracted the following characteristics:
  - c_{4a}: IPCA in the previous month;
  - c_{4b}: IPCA accumulated in the last three months; and
  - c_{4c}: IPCA accumulated in the six three months.

For each average vehicle sale value recorded, we create a script to synchronize the datasets and extract all the above features. With this script, it was possible to create a complete experimental dataset with more than 291 thousand records.

We named the experimental dataset S. For each possible combination between the sets (C_1, C_2, C_3, and C_4), we created a subset of S. As we have four groups of features,

[3] https://www.kaggle.com/datasets/franckepeixoto/tabela-fipe
[4] https://www.kaggle.com/ds/2499941
[5] https://www.kaggle.com/datasets/rmsoler/selic-srie-histrica

it is possible to create 2^4 subsets of S. For example, for the combination between the sets of C_1 and C_2, the subset $S_{c1,c2} \subset S$ is created, with only the characteristics of the groups C_1 and C_2. In this way, it is possible to evaluate the impact of each characteristic group on the prediction results of the monthly depreciation/valorization of cars.

Each row $s_x \in S$ was labeled with the percentage of depreciation/appreciation of the vehicle in the respective month (month m) concerning its average sales value in the previous month ($m - 1$).

4.2. Evaluation metrics

We use two metrics to evaluate the MLPR results with each of the experimental instances:

- **Mean Squared Error (MSR)**: the mean squared difference between all estimated values and the actual value. The MSE measures the quality of an estimator in that it is always a positive value that decreases as the error approaches zero.
- **Median Absolute Error (MAE)**: is the value halfway through the prediction absolute errors ordered. About the Mean Absolute Error, the Median Absolute Error is particularly interesting because it is robust to outliers.

4.3. Training/Test Methodology

We created 116 instances of training and testing sets for each of the 2^4 instances of S created based on combinations of the four feature groups. The number 116 represents the number of months between January 2013 to August 2022 – the period of the last ten years for which it was possible to synchronize the characteristics extracted from the three datasets presented in Subsection 4.1.

For each month m of the set of 116 months considered in the case study, we disregard or allocate each line $s_x \subset S$ in the training or test set based on the following conditions:

- if the month for which the average sales value of the line s_x was recorded is earlier than m, then s_x is added to the training set;
- if the month for which the average sales value of the line s_x was recorded is equal to m, then s_x is added to the test set; and
- if the month for which the average sales value of the line s_x was recorded is later than m, in this iteration, s_x is not added to either the training set or the test set.

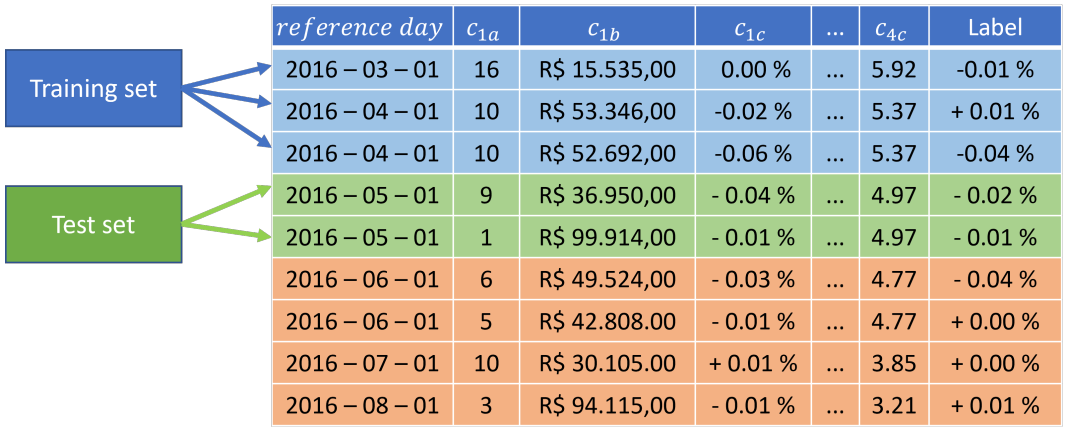
Figure 2 represents one iteration of creating the test and training sets.

For each 2^4 combination of feature sets, we trained and tested the MLPR with each of the 116 pairs of test/training sets used in our experiment, so it was possible to assess the performance of each combination of feature sets, month-to-month, or at intervals of months.

In order to create a comparison baseline, we first train an MLPR using only one characteristic, referring to the car value in the immediate previous month. We then combined the previous value of the vehicle with each one of 2^4 combination of feature sets.

All the features used by MLPR were previously normalized. We just didn't normalize the percentage of depreciation/appreciation that the vehicle had in the current month to the immediately previous month – this was the value that the MLPR estimated, that it was already in a small interval, between -0,1 and 0,1.

Figure 2. Example of an iteration for the creation of test/training sets pair. For example, in this iteration, we consider m = [May 2016] and |S| = 10. Blue lines were added to the training set. Green lines were added to the test set. Red lines were disregarded.



5. Results

Tables 1 and 2 present our MSR and MAE experimental results, respectively. In each Table, the first column indicates the sets of characteristics used, and the last three columns refer to the average of the results obtained in the periods from: i) 2013 to 2022; ii) 2013 to 2017; iii) and 2018 to 2022.

Figures 3 and 4 present a line graph of the mean annual results of each combination of characteristics, with the best results (for MSR or MAE) at the intervals presented in Tables 1 and 2. Figures 3 and 4 have the graphs for MSR and MAE, respectively.

Features	2013-2022	2013-2017	2018-2022
previous value - - - -	0,0004641	0,0004371	0,0004932
previous value - C1 - - -	0,0004669	0,0004655	0,0004683
previous value - C1 - C2 - -	0,0005012	0,0004497	0,0005564
previous value - - C2 - -	0,0004582	0,0004540	0,0004628
previous value - - - C3 -	0,0004835	**0,0004127**	0,0005594
previous value - C1 - - C3 -	0,0004667	0,0004639	0,0004698
previous value - - C2 - C3 -	0,0004994	0,0004911	0,0005083
previous value - C1 - C2 - C3 -	0,0004723	0,0004853	0,0004585
previous value - - - - C4	0,0005714	0,0004623	0,0006883
previous value - C1 - - - C4	0,0004520	0,0004585	0,0004451
previous value - - C2 - - C4	0,0005627	0,0005522	0,0005739
previous value - C1 - C2 - - C4	0,0004760	0,0005073	0,0004425
previous value - - - C3 - C4	0,0005218	0,0005026	0,0005424
previous value - C1 - - C3 - C4	**0,0004515**	0,0004642	**0,0004380**
previous value - - C3 - C3 - C4	0,0006097	0,0006048	0,0006149
previous value - C1 - C3 - C3 - C4	0,0004697	0,0004992	0,0004381

Table 1. Experimental results in terms of Mean Squared Error

Features	2013-2022	2013-2017	2018-2022
previous value - - - -	**0,00643**1	**0,005511**	0,007416
previous value - C1 - - -	0,007053	0,006392	0,007761
previous value - C1 - C2 - -	0,007954	0,007601	0,008333
previous value - - C2 - -	0,007501	0,007322	0,007693
previous value - - - C3 -	0,007767	0,006702	0,008908
previous value - C1 - - C3 -	0,007393	0,006921	0,007899
previous value - - C2 - C3 -	0,009133	0,008680	0,009619
previous value - C1 - C2 - C3 -	0,00799	0,008116	0,007855
previous value - - - - C4	0,009282	0,007691	0,010986
previous value - C1 - - - C4	0,007731	0,008091	**0,007345**
previous value - - C2 - - C4	0,01117	0,010900	0,011469
previous value - C1 - C2 - - C4	0,008867	0,009600	0,008080
previous value - - - C3 - C4	0,009910	0,009793	0,010035
previous value - C1 - - C3 - C4	0,008152	0,008491	0,007789
previous value - - C3 - C3 - C4	0,012343	0,011999	0,012711
previous value - C1 - C3 - C3 - C4	0,008680	0,009347	0,007964

Table 2. Mean of experimental results in terms of Median Absolute Error

5.1. Analysis of the results

For MSR in the whole period (2013-2022) and between 2018 and 2022, the best result was obtained with characteristics about the previous vehicle value combined with C_1, C_3, and C_4. Between 2013 and 2017, the best MSR was obtained only with the previous value of the vehicle combined with the C_3. Regarding the last five years (2018-2022), the set of characteristics C_1 proved to be the most important for the MSR metric because, in the eight experimental combinations in which it was used, the best MSR values were obtained.

For MAE, the best results were obtained using only the previous vehicle value for the whole period (2013-2022) and the first five years (2013-2017). Between 2018 and 2022, the best MAE was obtained by combining the previous vehicle value with the C_4.

The combination that presented the third MAE for the third-best last five years (2018-2022) was the same as the best MRS for the same period, so we can conclude that, among the combinations analyzed, this one is more suitable for the current scenario, but it is vital that data scientists that work that this type of forecast remain attentive, because as can be seen in the experiments, the best combination for forecasting vehicle prices varied over the years analyzed.

6. Conclusion

Predicting variations in the average car's selling price is a non-trivial task because different economic factors, such as the impacts of COVID-19 in recent years, can impact it. Our work differed from most of the others existing in the literature, as well as using data on the vehicle's sales history, it also analyzed the impact of data related to inflation and basic interest rate.

Figure 3. Annual Mean Squared Error from January 2013 to August 2022

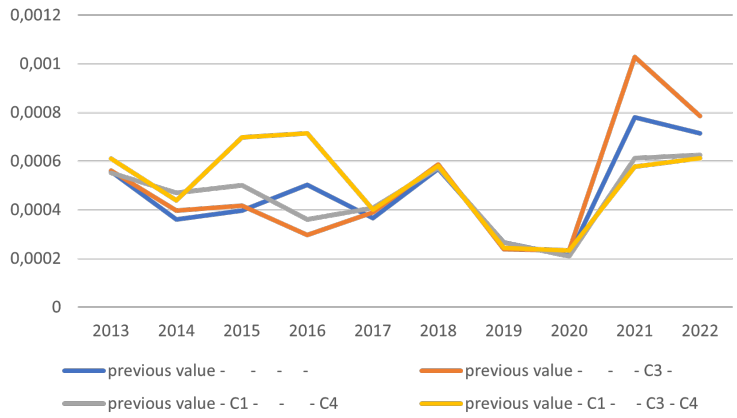
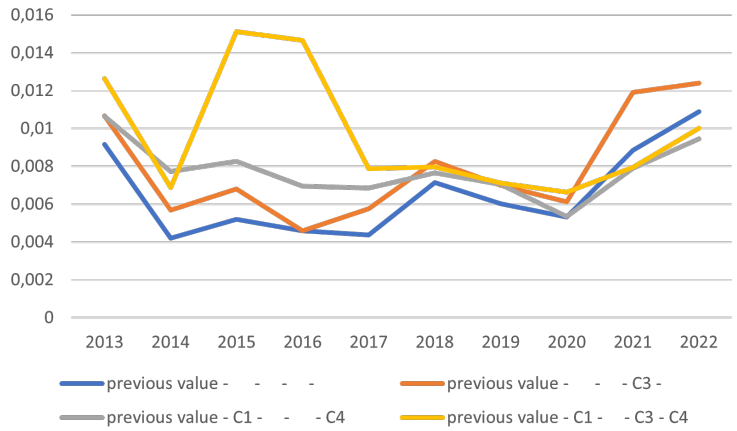


Figure 4. Annual mean of mensal Median Absolute Error from January 2013 to August 2022



In our experiments, the combination of characteristics that presented the best MSR for predicting car values importance between the years 2018 to 2022 presented one of the worst MSR between 2014 and 2017, which highlights the need for ongoing studies from data scientists to the updating of models used for this type of forecast.

We work with the average sale price of vehicles in Brazil, but the prices practiced vary depending on the region, vehicle model, color, accessories, or any other factor that may influence the conditions of supply and demand for a specific vehicle. In this context, to forecast the sales value of a specific vehicle, it is also essential to consider all these characteristics. A suggestion for future work is to combine different sets of related features with the other features explored by us to predict variations in the value of a specific car.

Other data sets can be explored, such as the content of posts on social networks and stock exchange quotes. Besides using MLPR, evaluating the results of other Machine Learning algorithms is also relevant as the set of vehicle values is a time series (as well as other characteristics used). Studies related to adjustments in the hyperparameters of the algorithms are also relevant.

References

Brahimi, B. (2022). Arabic text mining for used cars and equipments price prediction. *Computación y Sistemas*, 26(2).

Chen, C., Hao, L., and Xu, C. (2017). Comparative analysis of used car price evaluation models. In *AIP Conference Proceedings*, volume 1839, page 020165. AIP Publishing LLC.

Khan, Z. (2022). Used car price evaluation using three different variants of linear regression. *International Journal of Computational and Innovative Sciences*, 1(1).

Nasiboglu, R. and Akdogan, A. (2020). Estimation of the second hand car prices from data extracted via web scraping techniques. *Journal of Modern Technology and Engineering*, 5(2):157–166.

Pai, P.-F. and Liu, C.-H. (2018). Predicting vehicle sales by sentiment analysis of twitter data and stock market values. *IEEE Access*, 6:57655–57662.

Raza, S. and Masmoudi, M. (2020). Consumer vehicle purchase decision-making during covid-19. In *International Conference on Decision Aid Sciences and Application*, pages 692–696. IEEE.

Sharma, S., Sharma, S., and Athaiya, A. (2017). Activation functions in neural networks. *Towards Data Sci*, 6(12):310–316.

Tyagi, A. K. and Abraham, A. (2022). *Recurrent Neural Networks: Concepts and Applications*. CRC Press.

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