|  |
| --- |
| Western Governors University |
| Advanced Data Analytics |
| D214: CAPSTONE |

|  |
| --- |
| Allison Casey  3-11-2025 |

**A: RESEARCH QUESTION**

This analysis aimed to determine if an accurate neural network model could be constructed using the used car data set. Since cars depreciate over time, it is common for consumers to buy used cars and dealers to sell used vehicles. As a result, it is beneficial to buyers and sellers of used cars to determine the value of a used car accurately. The model will be considered accurate if the R² value is 0.8 or the model explains 80% of the variance in the test data. This study will utilize a predictive neural network to see if this type of model can be used to predict this value accurately. The null hypothesis is that a model with greater than 80% accuracy cannot be created, and the alternate hypothesis is that it can.

**B: DATA COLLECTION**

The data set, car\_price\_dataset.csv, was used and obtained from Kaggle. This public data set is available for anyone to use and analyze. This data set contains 10000 rows with 10 columns that contain various attributes of used cars, including the price. The primary advantage of using a pre-existing, public data set is the convenience and accessibility to data that may not have been feasible to collect otherwise. As a result, there were no real challenges associated with collecting the data other than generally assessing the quality of the data set. The data was confirmed to have 10000 rows, no missing values, and was relevant to the research question. A disadvantage of this data collection method is that there was no control over the data collection process to understand how the data was sampled or what quality control was involved, which could mean there could be unknown biases or flaws in the data.

**C: DATA EXTRACTION AND PREPARATION**

The data was extracted and explored before being used for the model. Initially, the data was read into a data frame using the Python Pandas library and its functions. The data was imported into a data frame using the read\_csv() function and explored to check for null values, duplicates, and other issues. The functions used to do this were the info(), describe(), duplicated(), isna(), and sum() functions. Since these yielded no issues, the data did not need further treatment. Pandas is advantageous for this sort of initial exploration since it is an efficient and widely used library for data analysis. A disadvantage to using Pandas can be the learning curve to understand more complex uses, but in this case, there wasn't a great deal of complexity with what it was needed for. There wasn't a huge disadvantage in this scenario.

1. A screen shot of a computer

   AI-generated content may be incorrect.
2. A screenshot of a computer screen

   AI-generated content may be incorrect.
3. A screenshot of a computer program

   AI-generated content may be incorrect.
4. A screenshot of a computer program

   AI-generated content may be incorrect.

One hot encoding and scaling techniques were also used to prepare the data for analysis. One hot encoding was performed because much of the data was categorical and needed to be transformed into numerical data that would be useful for the model. One disadvantage of one hot encoding is that it increases the dimensionality of the data, which, in this case, could result in overfitting and greater computational resource consumption.

1. A computer screen shot of a computer program

   AI-generated content may be incorrect.…

Scaling was used because the continuous numeric columns were not all scaled the same, and scaling prevents features from disproportionately influencing a model based on magnitude. The disadvantage of scaling would be that it can make the feature values and model output less interpretable than their original values. The tool used for this was the StandardScaler() method from the sklearn.preprocessing library.

1. A computer screen with text

   AI-generated content may be incorrect.

Finally, the data was split into training and test sets using the train\_test\_split method from the sklearn.model\_selection library. This is advantageous to prevent overfitting and test the model against data it was not trained against. Still, it can be a disadvantage because it limits the testing and training data.

1. A screenshot of a computer program

   AI-generated content may be incorrect.

See D214\_Capstone.ipynb for complete code. See PREPARED\_car\_price\_dataset.csv for prepared data.

**D: ANALYSIS**

A Keras Functional API model was created to analyze the data. This neural network model can handle models with non-linear topology, shared layers, and even multiple inputs or outputs (Keras, 2025). This model is ideal for this scenario because of its flexibility and handling of various inputs and complex relationships that are likely present within this data set. The disadvantage of using this model is that it is more complicated and difficult to interpret than other methods, such as linear regression. The model was created with an input layer, three hidden layers, and an output layer, which will be discussed below.

A screenshot of a computer program

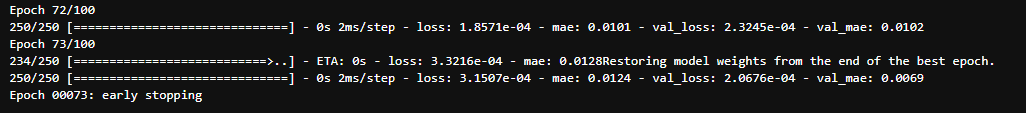
AI-generated content may be incorrect.

The model was then fit to the data using an early stopping monitor created using the EarlyStopping method from the tensorflow.keras.callbacks library. The benefit to this is that the model will stop training when the performance of the model stops improving, which will prevent overfitting and save computational resources. The disadvantage to this is that the training could potentially be stopped prematurely.

A screenshot of a computer program

AI-generated content may be incorrect.

…

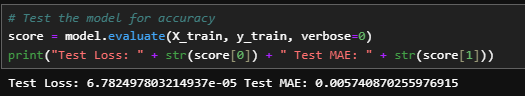


An in depth look at the hyperparameters of the model (The advantage being the reason each was chosen for this model):

* Activation functions: ReLu
  + Advantage: ReLu was used because it is simple and efficient for hidden layers
  + Disadvantage: Neurons in the network can become inactive and only output zero if there are many negative inputs and stopping learning.
* Number of nodes per layer: 128,64,32 determined through trial and error
  + Advantage: Easy and hands-on with the model to see how it is being affected
  + Disadvantage: It is not an easily repeatable process and is time-consuming
* Loss function: Mean Standard Error
  + Advantage: Simplicity and ease to implement
  + Disadvantage: Sensitivity to outliers
* Optimizer: Adam
  + Advantage: Easy and straightforward to implement and is commonly used (Elleh, 2025).
  + Disadvantage: Can be slower due to taking longer to converge than other optimizers
* Stopping criteria: val\_loss
  + Advantage: This will stop the model when the validation loss (MSE) stops improving to prevent overfitting
  + Disadvantage: This is less interpretable than val\_mae
* Evaluation metric: Mean Absolute Error
  + Advantage: Easy to understand and interpret
  + Disadvantage: It is not as helpful for optimization since it ignores the magnitude of errors

Model Output:

Loss explains the distance between the predictions and the actual labels, and the MAE is the average of how much a models predictions differ from the actual values.



The R² explains how much variance the model explains, so 1 would indicate a perfect prediction in this case; the value can be interpreted as 99% of the price variation can be predicted correctly by the model. So, the model is approximately 99% accurate. A black screen with white text

AI-generated content may be incorrect.

Visualizations:

Training Process

A graph of a graph

AI-generated content may be incorrect.

Prices Comparison

A blue line graph with numbers

AI-generated content may be incorrect.

**E: DATA SUMMARY AND IMPLICATIONS**

The results show that the null hypothesis can be rejected, and a model with greater than 80% accuracy can be created. Since regression models don't have accuracy in the same way that a classification model does, the accuracy, in this case, is defined by how well the model explains the variance in the target variable, so the R² value, which was calculated as .9998 or approximately 99% using the test data. One limitation of the analysis would be the data itself since it doesn't have every single make and model of every car, so it is limited to the car models in the data set. The data also limits the analysis because of the amount of data and how it limits the amount of training and testing that can be done overall. However, the data resulted in an accurate model that could be used. The recommended course of action would be for the model to start to be used while also using whatever previous methods were being used for evaluation to continue to evaluate the model and determine if the model is genuinely accurate or overfit to the data set it was trained with. With further assessment like this and further evaluation with more data gathered, the model can start to be used even more reliably. Along these lines, one direction for future study of the data could be to continue to gather more data to test and potentially re-train the model. Another approach for future research of the data could be to create other models, such as a random forest or a linear regression model, to predict the price and compare it to this model.

**Sources**

Elleh, Festus. “Advanced Data Analytics - Task 1.” D213 Task 2 Cohort Webinar. 2025.

“ReLU Activation Function in Deep Learning.” *GeeksforGeeks*, GeeksforGeeks, 29 Jan. 2025, www.geeksforgeeks.org/relu-activation-function-in-deep-learning/.

Team, Keras. “Keras Documentation: Earlystopping.” *Keras*, keras.io/api/callbacks/early\_stopping/. Accessed 11 Mar. 2025.

Team, Keras. “Keras Documentation: The Functional Api.” *Keras*, keras.io/guides/functional\_api/#training-evaluation-and-inference. Accessed 11 Mar. 2025.