An Intelligent Vehicle Price Estimation Approach Using a Deep Neural Network Model

Thuraya Alnajim 1,*, Nouf Alshahrani 1 and Omar Asiri 2,∗

1 Department of Computer Science, Faculty of Computers & Information Technology, University of Tabuk, Tabuk 71491, Saudi Arabia; 421010117@stu.ut.edu.sa
2 Department of Information Technology, Faculty of Computers & Information Technology, University of Tabuk, Tabuk 71491, Saudi Arabia
* Correspondence: 421009993@stu.ut.edu.sa (T.A.); oasiri@ut.edu.sa (O.A.)

Abstract: In recent years, the market for used-vehicle trade in the Kingdom of Saudi Arabia has grown significantly. This is due to the high cost of new vehicles that are not affordable by most buyers and lifting the ban on women drivers. Recently, several online websites for selling vehicles are available with different functions. However, estimating the vehicle price is based on traditional calculation methods, and this is inaccurate in several selling situations, as there are many factors that may affect the vehicle price, and these factors must be taken into consideration when estimating the vehicle’s price. Therefore, there is high demand to develop an automated vehicle price estimation system through adopting artificial intelligence (AI) technologies. Hence, this paper proposes an efficient vehicle price estimation system through developing an efficient deep neural network (DNN) model. The developed DNN model has been trained using a recent collected dataset for used-vehicle prices in the Kingdom of Saudi Arabia. The developed system has been validated using a recent vehicle price dataset, and the obtained results are compared with seven different machine learning models and showed a promising regression accuracy. In addition, we developed a reliable graphical user interface (GUI) for the purpose of allowing the user to estimate the price of any vehicle using the pre-trained DNN model.

Keywords: vehicle price estimation; estimation approach; machine learning; deep learning

1. Introduction

The Kingdom of Saudi Arabia is not only one of the most important countries in the Middle East but also the fifth largest in Asia. Valued at $4.90 billion in 2021, the Saudi Arabian used-vehicle industry is projected to reach USD $8.69 billion by 2027 [1]. Consumers in Saudi Arabia prefer to buy used vehicles rather than new vehicles, as used vehicles offer them better and affordable prices, aftermarket maintenance support, and easy financing.

Used vehicles refer to vehicles that have previously had one or more retail owners. Usually, used vehicles are sold through different outlets, such as rental vehicle companies, independent vehicle dealers, auctions, leasing offices, and private party sales.

Typically, the price estimation of used vehicles is a complicated task, since there are many factors that may impact the vehicle’s price significantly, and it requires significant effort and background in the field. In addition, there are several factors that affect the final price of a vehicle, which need to be considered prior to the estimate the price of vehicles. Therefore, the process of vehicle price estimation is a complicated task, mainly when the vehicle is located in another city.

According to a recent study [2], more than 557,000 pre-owned automobiles were sold in 2021. This figure has gone up in the Kingdom of Saudi Arabia during the past several years for a variety of reasons. One of them is the increase in exportation and transportation fees, as the prices of new vehicles have increased by 20% [3].
The adoption of artificial intelligence (AI) methods has succeeded in many ranges of applications, including industrial, education, healthcare, military, social, and many other applications [4]. AI technology can be mainly categorized into machine learning (ML), deep learning (DL), natural language processing (NLP), computer vision, and many others. AI technologies allow machines to recognize human languages, make predictions, and learn from examples [5,6].

Recently, there has been a high demand for offering accurate and reliable vehicle price estimation systems. Therefore, the employment of ML and DL technologies in vehicle price estimation systems will significantly enhance vehicle price prediction accuracy. The work presented in this paper involves developing an efficient deep neural network (DNN) approach that can estimate the price of a vehicle, in the Kingdom of Saudi Arabia, through adopting an efficient vehicle price dataset. Therefore, the main contribution of this paper lies on the following aspects:

1. Research, discuss, and analyze the recent developed vehicle price estimation approaches;
2. Investigate the available vehicle price datasets and analyze their reliability;
3. Develop a DNN approach for the estimation of the used-vehicle prices;
4. Validate the developed DNN approach using a real dataset.

The remainder of this paper is structured as follows: Section 2 discusses the recent developed vehicle price estimation systems. In Section 3, we present the developed system architecture for vehicle price estimation. Section 4 shows the main findings achieved from employing several ML models and the developed deep neural network model. In Section 5, we discuss and analyze the obtained results and compare them with the recent developed regression systems. And finally, Section 6 concludes the work presented in this paper and draws a list of future works.

2. Related Works

ML and DL technologies have recently been implemented in a number of research projects with the objective of predicting the prices of used vehicles. This section classifies the recently developed systems into two categories: ML-based and DL-based approaches.

2.1. ML-Based Vehicle Price Estimation Systems

In general, ML-based regression approaches can provide accurate prediction for observations that have never seen before [7]. There are many vehicle price estimation approaches have been developed recently with the aim of predicting the price of used vehicles. For instance, the work presented in [8] investigates the performance of three different ML methods, including random forest (RF), linear regression (LR), and decision tree (DT), where the authors concluded that the random forest regression model provides the most accurate predictions.

In [9], authors developed a vehicle price estimation system based on the employment of K-nearest neighbors (KNN) and DT methods, revealing that the KNN regression model offers the best mean squared error. In contrast, the authors of [10] developed a vehicle price estimation system by investigating three different AI models, artificial neural network (ANN), support vector machine (SVM), and RF, and achieved a regression accuracy close to 92.38%.

The authors of [11] investigated the performance of five different ML models, including LR, KNN, RF, XB-boost, and DT. According to the obtained results, the RF model achieves the best vehicle price estimation with a minimum root-mean-squared error of $3702. A KNN ML model has been developed in [12] using three different datasets. The achieved root-mean-squared error was around 4.01 and the mean absolute error was 2.01.

The authors of [13,14] involves the employment of LR for vehicle price estimation, whereas the authors of [15] developed a ForeXGBoost regression model that exploits the advantages of the designed data-filling algorithms for the recovery of missing and incomplete data. In [16], three different ML models (LR, lasso regression, and ridge regression) have been investigated for the purpose of vehicle price estimation tasks. The
obtained results offer 83.65% for linear regression, 87.09% for lasso regression, and 84.00% for the ridge regression models.

The work presented in [17] involves the employment of the gradient boosting (GB) model to estimate the vehicle price. The authors revealed that the GB model presented a high R-squared score and a low root-mean-squared error. On the other hand, the authors of [18] investigated the performance of three different ML models: GB, RF, and multiple LR. As a result, the GB model achieves the best mean absolute error (MAE) with 0.28, whereas the RF offers MAE with 0.35, and the multiple LR with MAE of 0.55.

A number of three different ML models (KNN, NB, and DT) have been investigated for the purpose of vehicle price estimation in [19]. In [20], a multiple LR approach has been adopted for the task of vehicle price estimation with a prediction accuracy of 98%. The SVM model has been adopted in [21], with a regression accuracy higher than LR and a neural network. The work presented in [22] involves the employment of SVM in order to overcome the problem of used-vehicle price estimation.

As presented above, several ML-based regression systems for vehicle price estimation tasks have been developed recently. The created methods vary in terms of both the accuracy of their regressions and the datasets that they use.

2.2. Deep Neural Network-Based Vehicle Price Estimation Systems

Deep neural network-based systems may offer accurate regression results; however, these approaches require an intensive background in the area of neural networks, in order to build an efficient network structure [23]. Hence, according to our intensive research investigation, we have only found two research works that have employed the neural network in the developed vehicle price estimation systems.

The work presented in [24] involves the development of an ANN model for the purpose of vehicle price estimation. The developed ANN model consists of two hidden layers with thirty and twenty-five neurons in the first and second hidden layers, respectively. The developed neural network model achieved a high prediction accuracy of 91.38%.

In [25], authors developed an enhanced BP neural network model to find out the best number of hidden neurons in the BP neural network, which enhances the speed of convergence of the network topology, in addition to enhancing the regression accuracy of the developed model.

In conclusion, only two research works have considered the area of neural network to develop a vehicle price estimation system. Although neural network-based systems are complicated in structure and require knowledge in the area of neural networks, neural network-based approaches offer accurate regression results. Therefore, it is significant to develop an efficient deep neural network model for vehicle price estimation.

3. System Design

This section discusses the system design, starting from the adopted dataset, preprocessing task, to the analyses of the obtained results. The main adopted phases are presented in Figure 1, where each phase is discussed in detail below.

![Figure 1. The main phases for developing an efficient vehicle price estimation system.](image-url)
3.1. Cleaning of theSelected Dataset

In this work, we employed a recent collected local vehicle price dataset [26] with a total number of 8248 records including 15 attributes (14 features and a single label). The collected dataset aims to predict the price of vehicle prices through the development of ML and DL approaches. This dataset consists of various information ranging from vehicle brand and type to options and capacity of the engine. Figure 2 presents a screenshot for the selected vehicle price estimation dataset.

![Figure 2. A screenshot for the selected vehicle dataset.](image)

It is worth mentioning that the employed vehicle price dataset includes missing and incomplete values, which negatively affect the vehicle price estimation process. Therefore, a high demand has been paid to employ different intensive preprocessing approaches. The total number of records in the selected dataset is 8248 records with 15 attributes (14 features and one label). However, the selected dataset consists of 2544 null values that exist in different columns (attributes), where the null values affect the performance of the machine learning and DNN models and consequently reduce the classification accuracy.

As previously described, the vehicle price estimation dataset is corrupted with various errors, such as absent, incorrect, or inconsistent values; consequently, the errors that exist in the original dataset can have a significant impact on the accuracy of the vehicle price estimation. Data scientists often engage in many data-cleaning procedures prior to initiating model training [27]. Data cleaning is a significant step for ML and deep learning models; it must be carried out to increase forecast accuracy.

Therefore, this stage involves eliminating columns that are referred to as zero-variance predictors, removing empty rows, and recovering values for the empty records in the dataset. In order to process the missing data values in the dataset, we cover the missing information by adopting several strategies. The cleaning and preprocessing tasks involve several stages: First, the removal of duplicate records is performed, since duplicates will unavoidably skew the dataset and confuse the results. Second, irrelevant data attributes are removed, as the existence of irrelevant data values in the selected dataset will significantly slow down and confuse the analysis process. Third, data types are converted, since the employed dataset consists of different data types, including texts, where the ML and DL models will not be able to be trained using the text data values. And fourth, missing values are handled, as the selected vehicle price dataset consists of a large number of missing and incomplete values. Therefore, it is important to remove observations that have missing or incomplete values.

After considering the above four preprocessing approaches, a new vehicle price dataset is obtained, which contains rich and valuable information with a total number of 8035 records. In addition, two features were removed from the dataset, which include unusable data values (link and sequence features). The total number of 12 features, including make, year, type, origin, color, options, engine size, fuel type, gear type, mileage, region, and negotiable. A single label exists in this dataset, which is the vehicle price. Table 1 presents the general statistics of the original and modified datasets.
Table 1. General statistics for the original dataset.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Original Dataset</th>
<th>Modified Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total records</td>
<td>8248</td>
<td>8035</td>
</tr>
<tr>
<td>Total attributes</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>Total features</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>Total labels</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total values</td>
<td>121,176</td>
<td>104,455</td>
</tr>
<tr>
<td>Total NaN values</td>
<td>2544</td>
<td>0</td>
</tr>
</tbody>
</table>

The selected dataset contains many ‘Toyota’ brand vehicles, which, with a market share of 25%, is the most popular vehicle brand in the Kingdom of Saudi Arabia. Recent studies [28] indicate that Toyota was the most popular vehicle brand in the Kingdom of Saudi Arabia, and this is the primary cause for the large proportion of Toyota vehicles in the selected dataset. As shown in Figure 3, the selected dataset contains more than 2000 sold Toyota vehicles.

Figure 3. The total number of sold vehicles in the selected dataset.

3.2. Feature Engineering

The goal of feature engineering is to turn information into insight [29] information. In a broad sense, a feature can be defined as a numerical representation of a certain characteristic observed in real-world data or occurrences [30]. Feature engineering is divided into two broad categories:

- **Feature transformation**: This involves converting the original feature to the functions of the original feature. The cleaned dataset contains nonnumerical data values that are not suitable for training and testing ML and DL approaches. Therefore, it is important to transform the nonnumerical data values into numerical data values.

- **Feature selection**: The weights of features are not equal. Therefore, the most important features need to be selected to perform efficient regression.
3.2.1. Feature Transformation

The selected dataset is engineered in order to convert the data in the selected features into numerical values where the employed ML model can accept such data in the training and testing tasks. Hence, the following features are transformed: origin, color, options, region, fuel type, gear type, negotiable, make and type.

3.2.2. Feature Selection

A new dataset file is obtained from the previous task with numerical data values. Then the feature significance process takes place; this task focuses on the investigation of the feature significance in the new dataset. Feature significance refers to the techniques that estimate the score for all the input features using several ML models. The score usually represents the significance (importance) of each input feature, where a higher score means that the selected feature will have a larger effect on the machine learning model that is being employed to predict a certain variable. Finding the feature importance score is extremely useful for many reasons, including:

1. Model enhancement: Incorporating the most significant features in the training process will increase the classification/regression accuracy and precision. In addition, this will minimize the training and testing times.
2. Data understanding: The feature importance allows the researchers to understand the relationship between the input features and the output (target) variable.
3. Model interpretability: Calculating the scores of features can be used to determine the feature attributes that offer the best predictive power to the employed machine learning model.

Therefore, we employed several ML models to obtain the significance of the input features in the selected dataset, as follows:

1. Logistic Regression (LogR): Using the LogR, we can obtain the coefficient that includes the coefficients obtained for each single input variable. The obtained coefficients may offer the source for a crude feature significant score. Figure 4 presents the feature importance scores for each single input feature using the LogR model.

![Figure 4. Feature importance using the LogR model.](image)

2. Decision Tree (DT): This includes the classification and regression trees that offer feature significance scores based on the reduction in the criterion employed to select split points, like Gini or Entropy. We employed the decision tree regressor in order to obtain the most significant features in the selected dataset. Figure 5 presents the feature scores for the input features using the DT regression model.

![Figure 5. Decision Tree feature scores.](image)
Figure 5. Feature importance using the decision tree regression model.

3. Random Forest (RF): The RF algorithm can be used to extract the importance of the input features, through the adoption of the RF regression method. Figure 6 depicts the importance scores of the input features for the selected dataset using the RF.

Figure 6. Feature importance using the RF regression model.

4. XGBoost: This is a feature importance function that offers an efficient implementation of the stochastic gradient boosting algorithm. We employed the XGBoost regressor to obtain the feature importance. Figure 7 presents the feature importance score for each input feature using the XGBoost model.

Figure 7. Feature importance using the XGBoost regression model.
5. Permutation: This is a technique that is used for estimating the feature significance scores that are independent of the employed model. Permutation feature selection can be adopted through employing the permutation importance function that takes a dataset as an input, fit model, and a scoring function. Figure 8 presents the scores of feature importance using the KNN model.

Figure 8. Feature importance using the permutation regression model.

As noticed above, the feature importance scores slightly vary with each of the machine learning models. However, if we consider the decision tree, random forest, and XGBoost models, we notice that these three models agreed that input feature 2 (mileage) is the most significant feature with the highest score value. The decision tree achieves a 0.25573 importance score for the mileage feature, whereas the random forest model achieves a 0.25579 importance score for the mileage feature, and finally, the XGBoost model achieves a 0.28589 importance score for the mileage feature.

For the three ML models (decision tree, random forest, and XGBoost), the feature significance for the decision tree is as follows: engine size, mileage, negotiable, make, type, origin, options, region, fuel type, gear type, year, and color.

On the other hand, the feature importance in the RF regression model are as follows: negotiable, mileage, engine size, make, type, and options. Table 2 presents the feature importance for the three machine learning models (decision tree, random forest, and XGBoost).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Decision Tree</th>
<th>Random Forest</th>
<th>XGBoost</th>
<th>Logistic R</th>
<th>Permutation</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engine size</td>
<td>0.14530</td>
<td>0.17124</td>
<td>0.08164</td>
<td>0.00520</td>
<td>0.00051</td>
<td>0.080778</td>
</tr>
<tr>
<td>Mileage</td>
<td>0.06247</td>
<td>0.06865</td>
<td>0.02000</td>
<td>0.00098</td>
<td>0.09214</td>
<td>0.048848</td>
</tr>
<tr>
<td>Negotiability</td>
<td>0.25520</td>
<td>0.25715</td>
<td>0.36203</td>
<td>0.00221</td>
<td>0.00129</td>
<td>0.175576</td>
</tr>
<tr>
<td>Make</td>
<td>0.16162</td>
<td>0.13404</td>
<td>0.13990</td>
<td>0.00399</td>
<td>0.00854</td>
<td>0.089618</td>
</tr>
<tr>
<td>Type</td>
<td>0.08313</td>
<td>0.07558</td>
<td>0.02570</td>
<td>0.00174</td>
<td>0.00625</td>
<td>0.03848</td>
</tr>
<tr>
<td>Origin</td>
<td>0.00536</td>
<td>0.00859</td>
<td>0.01831</td>
<td>0.00163</td>
<td>0.00098</td>
<td>0.006974</td>
</tr>
<tr>
<td>Options</td>
<td>0.08320</td>
<td>0.08068</td>
<td>0.16848</td>
<td>0.00081</td>
<td>0.00055</td>
<td>0.066744</td>
</tr>
<tr>
<td>Region</td>
<td>0.01010</td>
<td>0.01285</td>
<td>0.00595</td>
<td>0.00021</td>
<td>0.00091</td>
<td>0.006004</td>
</tr>
<tr>
<td>Fuel type</td>
<td>0.00168</td>
<td>0.00186</td>
<td>0.01123</td>
<td>0.00790</td>
<td>0.00005</td>
<td>0.004544</td>
</tr>
<tr>
<td>Gear type</td>
<td>0.00086</td>
<td>0.00115</td>
<td>0.02684</td>
<td>0.00081</td>
<td>0.00006</td>
<td>0.005944</td>
</tr>
<tr>
<td>Year</td>
<td>0.16997</td>
<td>0.17171</td>
<td>0.12549</td>
<td>0.16851</td>
<td>0.12801</td>
<td>0.152738</td>
</tr>
<tr>
<td>Color</td>
<td>0.02112</td>
<td>0.01650</td>
<td>0.01442</td>
<td>0.00584</td>
<td>0.00548</td>
<td>0.012672</td>
</tr>
</tbody>
</table>
According to several recent studies [31,32] random forest, decision tree, and XGBoost regression models are the best models in terms of estimating the feature significance for regression problem scenarios. On the other hand, based on the feature importance scores obtained from employing several machine learning models, we built a new dataset that includes the most significant features in order to enhance the accuracy for the training and testing of the machine learning models.

3.3. ML Regression Models

The previous section presents the feature importance scores using five different approaches. However, this section discusses the ML regression models that have been customized to predict the vehicles’ prices.

The final obtained dataset consists of nine total attributes (eight features and a single label). This dataset consists of the most significant features for enhancing the regression accuracy. The employed ML models are as follows:

1. **Linear Regression (LR):** LR is a fundamental and widely used category of predictive analyses. LR examines the relationship between a dependent variable and one or more independent variables.

2. **K-Nearest Neighbors (KNN):** KNN is an uncomplicated and straightforward machine learning approach that may be utilized to address classification and regression challenges. A KNN algorithm works by assuming that similar things exist in close proximity.

3. **Decision Tree (DT):** A DT is based on the structure of a tree, since it breaks down the required dataset into smaller subsets while an associated decision tree is incrementally developed.

4. **Random Forest (RF):** RF is an ensemble learning technique that may be used for both classification and regression tasks. Random forest (RF) operates by generating several decision trees during the training phase. The resulting class is determined by either selecting the mode of the classes for classification or calculating the average prediction for regression based on the individual trees.

5. **Logistic Regression (LogR):** LogR predicts a dependent data value through analyzing the relationship between one or more independent variables.

6. **Ridge Regression (RR):** The ridge regression model is used to analyze data values that suffer from multicollinearity, where the ridge model performs L2 regularization.

7. **Lasso Regression (LassR):** LassR is considered as a type of linear regression that employs shrinkage. Shrinkage is a method used to shrink data values toward a central point, such as the mean.

The employed ML models have been tailored to fit the selected dataset and the desired outcomes. This section presents the adjusted parameters for each single ML model. Table 3 presents the customized hyperparameter tuning for the LR model, where the number of splits was set to eight, and the number of repeats was set to three. On the other hand, the customized hyperparameters for the KNN model is displayed in Table 4, where the number of neighbors is equal to six, whereas the Euclidian was set as metric distance function.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>n-splits</td>
<td>8</td>
</tr>
<tr>
<td>n-repeats</td>
<td>3</td>
</tr>
<tr>
<td>random-state</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 4. The hyperparameter tuning for the KNN model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>n-neighbours</td>
<td>6</td>
</tr>
<tr>
<td>metric</td>
<td>Euclidian</td>
</tr>
<tr>
<td>sample-weight</td>
<td>None</td>
</tr>
</tbody>
</table>

The customized hyperparameter tuning for the DT model is presented in Table 5, where the max-depth value was set to 6, and the max-leaf-node value was set to 5. Table 6 shows the customized hyper parameters turning for the RF model, with a max-depth of 5 and an n-estimators value of 25, and the minimum sample-split was set to 18.

Table 5. The hyperparameter tuning for the DT model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>max-depth</td>
<td>6</td>
</tr>
<tr>
<td>max-leaf-node</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 6. The hyperparameter tuning for the RF model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>max-depth</td>
<td>5</td>
</tr>
<tr>
<td>n-estimators</td>
<td>25</td>
</tr>
<tr>
<td>min-sample-split</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 7 presents the hyper parameters for the LogR model, where the solver parameter was set to saga which relatively performs well compared to the other methods; in addition, it is an efficient choice for multinomial logistic regression. On the other hand, the penalty value was set to 10. Penalty intends to minimize the model generalization error and its aim is to disincentivize and regulate overfitting. Finally, the C parameter refers to regularization strength which must be a positive float. The C value works with the penalty parameter to regulate overfitting.

Table 7. The hyperparameter tuning for the logistic regression model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solver</td>
<td>lbfgs</td>
</tr>
<tr>
<td>Penalty</td>
<td>11</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 8 presents the hyper parameters for the ridge regression model. A tuning parameter (λ) refers to the penalty parameter that controls the strength of the penalty term in ridge regression and lasso regression models.

Table 8. The hyperparameter tuning for the ridge regression model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>λ</td>
<td>5</td>
</tr>
<tr>
<td>min-sample-split</td>
<td>18</td>
</tr>
</tbody>
</table>

3.4. Deep Neural Network Regression Model

The previous section discusses the employment of tailored ML models for developing an efficient vehicle’s price estimation system. In general, deep learning algorithms produce more efficient regression results in several scenarios. However, deep learning algorithms
are more complex than ML models. In addition, DNN approaches are best suited for handling high-complexity decision-making recommendations, image classification, voice recognition, and recommendation systems [33]. Therefore, this section discusses the design and development of an efficient DNN vehicle price estimation model. Figure 9 shows the proposed DNN model which consists of five layers (one input, one output, and three hidden layers).

Figure 9. The developed deep neural network model.

The structure of the developed DNN model is composed of the following parameters and layers:
- Input layer: A single input layer is presented which consists of eight different inputs that are directly fed from the vehicle price dataset.
- Hidden layers: Three hidden layers are included in the developed DNN network model, in different sizes. The first, second, and third layers have 256, 128, and 64 nodes, respectively.
- Output layer: A single output layer includes a single node that predicts the vehicle price.

For the hidden layers, we adopted the ‘ReLU’ as an activation function, whereas the ‘linear’ activation function was employed in the output layer. Rectified linear unit (ReLU) is an activation function that was employed in the hidden layers. ReLU is a piecewise linear function.

\[
f(x) = \max(0, x)
\]  

(1)

where \(x\) is the input value. If the ReLU input function receives a value and if the received value is a negative value, then it returns a 0; however, if \(x\) is a positive value, then the ReLU function returns that value.

3.5. Developed Graphical User Interface (GUI)

For the purpose of efficient validation of the developed DNN model, we developed a graphical user interface (GUI) using the Python development environment and tkinter package. Tkinter is way to create a GUI in Python. Therefore, a GUI has been implemented where the trained DNN model is embedded in order to allow the user to predict the price of any vehicle with certain specifications.

For instance, the user needs to choose the vehicle’s make, type, options, negotiable, mileage, engine size, year, and color. Then, these data will be fed to the developed DNN model to predict the price of the vehicle with the required specifications. Figure 10 presents the developed GUI that contains of eight inputs and a single predicted value.
4. Experimental Results

This section discusses the results obtained from adopting several ML models and the designed DL model in order to estimate the regression accuracy for each single model. We implemented different tailored machine learning models with the selected dataset. The implemented machine learning models have been tailored to predict the price of used vehicles using the adopted dataset. The implemented machine learning models have been assessed using the following metrics:

- **R-Squared Error (RSE):** This represents the fraction of variance of the real value of the response variable that is captured by the regression model.
- **Mean Absolute Error (MAE):** MAE refers to the magnitude in the difference between the estimated observation and the real value of the observation.
- **Root-Mean-Squared Error (RMSE):** For simplicity, RMSE refers to the square root of the MSE.

For each ML model, we calculated the above five metrics. The R-squared error enables us to evaluate the performance of the deep learning or ML model. The R-squared error represents the value of how much the independent variables are able to find the value for the target variable. Figure 11 presents the R-squared error for each ML model. As presented, the LogR achieves the best R-squared value among the employed ML models; however, the designed DNN model achieves better results than the LogR model.

On the other hand, the MAE values have been analyzed for each ML model. MAE is used to estimate the effectiveness of the developed regression model. MAE estimates the average size of mistakes in a collection of predictions for a given dataset. It is concerned with estimating the average absolute difference between the predicted and actual values. Figure 12 shows the MAE for the seven ML models. As noticed, the developed random for-

![Developed GUI](image-url)

**Figure 10.** The developed GUI using the developed DNN model.

![R-squared Error](image-url)

**Figure 11.** R-squared error for the 8 models.
est model almost obtains the best MAE results among the employed ML models; however, the developed DNN model offers better results compared to the RF model.

The RMSE metric is also assessed for all the seven implemented ML models and the developed DNN model. RMSE measures the average difference between the predicted values by a certain MN/DL model and the actual values. In general, RMSE offers an estimation of how well the model will be able to predict the goal value, which refers to the accuracy. Among the implemented ML models, the RF model offers the best RMSE result. However, the DNN achieves better RMSE regression accuracy than the RF model, as presented in Figure 13.

5. Discussion

The main aim of this work is to focus on the area of used-vehicle price estimation, through developing an efficient DNN model that is able to predict the price of a used vehicle based on several factors. The developed prediction system will assist vehicle buyers in estimating the real price for a certain vehicle based on an efficient regression approach, and this will save the buyers’ time, money, and effort.

In general, ML and DL approaches require a dataset to train a certain model and then initialize the prediction process. However, the existing datasets are either do not suit the issue in the Kingdom of Saudi Arabia, are old, or are corrupted. Therefore, we adopted a dataset that has been collected recently from the Saudi market; however, this dataset contains missing, incomplete, and unnecessary attributes.

The chosen dataset has been processed through several stages in order to process the missing and incomplete data records and remove the unmercenary attributes. The outcome is a clean, reliable, and efficient vehicle price dataset that can be employed for ML and DL approaches for the purpose of vehicle price estimations.
Several ML-based vehicle price estimation systems [8–11] have been developed recently with the purpose of offering an efficient price estimation result. However, the developed systems are either inefficient in terms of prediction accuracy, focused on a certain dataset, or do not provide reliable validation results.

On the other hand, DNN-based approaches offer better estimation results. Although the DNN requires longer time for training the model, it however achieves much better prediction accuracy. Moreover, DNN methods tend to solve the vehicle price estimation problem end to end, and DNN methods offer reliable regression accuracy for complex tasks that necessitate machines to make sense of unstructured data. Hence, according to our research results, two deep neural-based approaches [24,25] have been developed recently for the purpose of vehicle price estimations, where the obtained results are incomplete and were not validated with efficient datasets.

The developed vehicle price estimation system offers several advantages over the existing approaches: First, the developed system achieves an efficient regression accuracy (the minimum prediction cost). Second, an efficient set of features has been selected to improve the regression results. Third, the regression results obtained from the developed DNN-based model has been compared to several ML models, where the DNN-based regression model achieves better regression accuracy. Fourth, a usable GUI system has been designed to offer a reliable way to assess the efficiency of the developed vehicle price estimation system. However, the main drawback of this system is the limited size of the available vehicle dataset.

Therefore, this paper investigated the performance of several ML models for the purpose of used-vehicle price estimation tasks, and the obtained results were reasonable. However, for vehicle price estimation task, a high demand is needed to obtain accurate regression results. Therefore, we developed an efficient DNN model that offers reliable and accurate price estimation for used vehicles.

6. Conclusions

A high demand for developing an efficient vehicle price estimation approach for the purpose of achieving reasonable prediction accuracy has appeared recently. This paper discusses the problem of used-vehicle price estimation for the market of Saudi Arabia as a case study. A recent collected vehicle price dataset has been adopted, where several preprocessing processes have been achieved to offer an efficient and accurate vehicle price dataset. On the other hand, several ML models have been tailored for the purpose of vehicle price estimation with reasonable estimation accuracy. However, the obtained results from the tailored ML models were incomplete; therefore, we developed an efficient DNN vehicle price estimation model that outperforms the estimation accuracy of the seven implemented ML models. The obtained results are promising, and there is a huge demand to obtain an up-to-date dataset to allow for better estimation results. The developed vehicle price estimation system can be used by end users in order to facilitate the estimation process of used vehicles. For future research work, we aim to employ additional datasets in order to accurately validate the efficiency of the developed vehicle price estimation system. Moreover, a mobile application will be developed to facilitate the process of vehicle price estimation.

Author Contributions: Conceptualization, T.A., N.A. and O.A.; methodology, T.A. and N.A.; software, T.A.; validation, T.A.; formal analysis, T.A. and N.A.; investigation, T.A. and N.A.; resources, T.A., N.A. and O.A.; data curation, T.A. and N.A.; writing—original draft preparation, T.A. and N.A.; writing—review and editing, T.A., N.A. and O.A.; visualization, T.A.; supervision, O.A.; project administration, T.A. and O.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: This paper contains the original findings of this study. For more information, please contact the corresponding author.
Acknowledgments: The authors would like to express their sincere gratitude to Saleh Albelwi, a faculty member in the Department of Computer Science at the Faculty of Computers and Information Technology, University of Tabuk, for his invaluable support throughout the research process. Albelwi’s expertise and insightful feedback were instrumental in the successful completion of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.