Hindawi
Journal of Engineering
Volume 2023, Article ID 8833753, 27 pages
https://doi.org/10.1155/2023/8833753

Research Article

Methodology for Estimating the Cost of Construction Equipment Based on the Analysis of Important Characteristics Using Machine Learning Methods

Nataliya Boyko and Oleksii Lukash

Lviv Polytechnic National University, Lviv 79013, Ukraine

Correspondence should be addressed to Nataliya Boyko; nataliya.i.boyko@lpnu.ua

Received 9 May 2023; Revised 24 July 2023; Accepted 24 August 2023; Published 22 September 2023

Academic Editor: Weili Zeng

This paper considers the current market pace, which requires a corresponding competitive advantage. This study forecasted the cost of heavy machinery depending on geolocation and essential characteristics by the field of activity. This study analyzes specific categories of heavy machinery for important price characteristics. The study classified them by keywords in the text description as essential characteristics. Accordingly, a dataset was formed based on the data obtained. The research objective is to collect and structure data from web resources for the sale of heavy equipment. This paper describes in detail the preliminary data processing. The main stages of preprocessing are presented in detail: detection and processing of missing data, removing anomalous data, coding of categorical data, and scaling. The method of the average value of a specific grouped set was applied to fill in the gaps according to the characteristics and available data. The mode value from the grouped items was used to fill in the gaps. The interquartile range and standard deviation were used to detect anomalies. We used the Kolmogorov–Smirnov, KS_Test, and Lilliefors tests to check the data for normality. In this study, the assessment of abnormal data was applied separately to each set of grouped data with the same parameters. The study built and analyzed models using machine learning methods (linear and polynomial regression, decision trees, random forest, support vector machine, and neural network). Two data encoding methods were used to achieve maximum model accuracy: Label Encoder and One Hot Encoder. The work of each algorithm is considered on the example of the created dataset. In this study, the parameter used for coding was the geolocation of heavy equipment. The study pays additional attention to the specific characteristics of heavy machinery by the sector of the economy. The existing methods and tools for price forecasting, depending on the specific characteristics of the equipment, were analyzed. The practical significance of this work lies in developing an algorithm for predicting the cost of heavy machinery by assessing several parameters.

1. Introduction

The need for accurate forecasting characterizes modern society. For example, governments want to anticipate the trend of many indices, such as unemployment, inflation, industrial production, and expected tax revenues, to shape effective policies. Marketing managers want to anticipate product demand, sales, and changes in consumer preferences to make appropriate decisions about current and future policies and, more generally, to formulate adequate strategic planning. Furthermore, forecasting the prices of heavy machinery is necessary for people engaged in agricultural, construction, freight, or transport activities. After all, technology, its quantity, variability, and technological capabilities are the main subjects of the development of their companies[1, 2].

Since the market for heavy equipment is vast and widely variable, it creates difficulties in finding equipment for the customer's needs. Therefore, there is a need for forecasting using machine learning methods. Its use is not limited to one category of technology and is flexible. For best results, there are needs to adjust essential parameters for the study and adapt to any changes in the data. When forecasting the value of vehicles, they may encounter specific problems, primarily related to unscrupulous sellers. Such problems include the fact that sellers may need to find important information about their heavy equipment, namely, technical and visual conditions, history of use and maintenance, condition of power units, and more. All these parameters can affect the cost estimate of the equipment, although they can try to simulate the missing data; after that, the cost estimate will not be accurate [3, 4].

The task of the study is to collect and structure data from web resources for the sale of heavy equipment, in addition to preprocessing of the above data, construction, and analysis of the results of machine learning methods (linear and polynomial regression, decision trees, random forest, and neural network) [5, 6].

Accurate forecasting is crucial in modern society, and machine learning methods can help. However, the market for heavy equipment is vast and widely variable, making it challenging to find equipment that meets customer needs. Machine learning can help overcome this difficulty [7, 8].

The task of this study is to collect and structure data from web resources for the sale of heavy equipment. The aim is to preprocess the data and use machine learning methods such as linear and polynomial regression, decision trees, random forest, reference vector method, and neural networks to accurately forecast the prices of heavy equipment. However, when forecasting the value of vehicles, specific problems can arise, primarily related to unscrupulous sellers who may miss or try to hide critical information about the equipment. This information includes technical and visual conditions, history of use and maintenance, and condition of power units.

Since the use of heavy machinery in various spheres of life is growing in quantity in the primary and secondary markets, there is an urgent need to analyze and forecast prices for special equipment. The chosen research is relevant, which indicates a high demand for this type of equipment and a realistic forecast for its prices.

The purpose of the study is to forecast the cost of heavy machinery depending on the geolocation and essential characteristics of the field of activity.

In order to achieve this goal, specific tasks need to be performed:

(i) Analyze specific categories of heavy machinery and classify its price characteristics

(ii) Structure the characteristics of the equipment given in the text description by keywords and to create a dataset based on the data obtained

(iii) Conduct preliminary data processing using different methods at each stage

(iv) Analyze the proposed methods and tools for price forecasting depending on the specific characteristics of the equipment

(v) Create and apply a model for predicting the cost of heavy machinery

The work's scientific value lies in comparing and applying machine learning methods to create a model for accurately determining forecast data.

The practical value of the work lies in developing an algorithm for predicting the cost of heavy equipment with the estimation of a large number of parameters.

2. Literature Analysis

The topic of vehicle valuation forecasting is prevalent and has been repeatedly studied by academics and businesses. There are several reasons for this: the desire of businesses to anticipate market trends and understand the formation of prices in the secondary market for both used and new vehicles. In addition to business, ordinary citizens also want to be able to find out the objective value of a vehicle so that they do not overpay or lose money when selling their vehicle. This topic is exciting and constantly under research because the economic background of the world is constantly changing, and with it, the market trends, including the secondary vehicle market, are changing.

When analyzing scientific literature sources, we considered articles and scientific papers that are similar to this work's topic or related to certain aspects of its implementation. These include forecasting the value of passenger vehicles, vehicle analysis, application of machine learning methods for price prediction, in particular, the use of linear, multivariate regression, the use of decision trees, and the use of combinations of such methods.

Predictive models based on machine learning methods can now consider time-dependent parameters (seasonality, trends, and cycles) [9] to maximize the accuracy of forecasts based on the given data. This process is called machine learning prediction. It should be applied in all aspects of business, including sales and demand forecasting, recruitment forecasting, weather forecasting, content consumption forecasting, predictive planning and maintenance, and more.

In papers [10, 11], the research focuses on studying cotton price forecasting using five different ML regression algorithms: linear regression, Bayesian linear regression, decision tree regression, and decision forest regression, metrics for evaluating ML model performance. Although the data studied in this paper are entirely incompatible with those proposed in the thesis, they have a similar structure of dependencies. The data set contains the prices in different states of India, i.e., the dependence of price on geolocation and demand for products in different locations.

The article's authors [12] analyze the car market in India using a machine learning method, supervised learning. The researchers propose to predict the price of vehicles using historical data of the Indian market collected from various car sales platforms and building ensemble machine learning models, namely, random forest and extra trees. The result of the study is a model of the ensemble algorithms random forest and extra trees, an evaluation and comparison of the results obtained, and an analysis of the advantages of using ensemble models compared with other machine learning algorithms.

The article [13] describes the disadvantages of using linear models for price forecasting because the price in the world is formed under the influence of many factors. Accordingly, nonlinear models are better at predicting arbitrary pricing in real life. The researchers also suggest using the S-Curve model as an alternative nonlinear model for estimating the value of used cars. The authors also investigated linear and cubic regression and conducted a comparative analysis with the S-Curve model.

Article [14] suggests that we study the Chinese car market to ensure objective pricing of cars in the same market. The researchers decided to test the pricing concepts by building the following machine learning models: taking into account features specific to particular car brands, features specific to certain types of cars, and a general model that includes all the features in the set. The models were built using linear regression and decision tree methods. The authors provide detailed results of each model and conclude that the decision tree is much more effective when building a model using all available features to predict objective car pricing on a large data set. The decision tree outperforms linear regression using fewer observations in the dataset, but their results are close.

In article [15], the authors conducted a large-scale study of the secondary car sales market, the dependence on car parameters such as mileage, the initial price of the car, the price for which the car was sold, the age of the car, the type of fuel, i.e., as well as the impact of the abovementioned parameters on price formation. After preliminary data preparation, the researchers started training the models. The researchers used many machine learning algorithms as models, such as linear regression, lasso regression, decision tree, Bayesian linear regression, XGBoost, and gradient boosting regression. All methods showed promising results. Researchers pay special attention to gradient boosting regression and decision trees, as these models predict test data most accurately. The coefficients of determination of the models are 0.9355 and 0.9544, respectively, and the average absolute deviation is 0.6378 and 0.6711, which indicates that the models provide reasonably accurate predictions. The other models have an average coefficient of determination of 0.86 and an average absolute deviation of about 1.1.

In [16], researchers describe using a machine learning algorithm called KNN to solve the problem of predicting the price of a car on the secondary market. The essence of their approach is to select K cars with the most similar characteristics and find the arithmetic mean of these k cars' prices. The researchers collected a dataset of over 4 million car sales records in the Indian market for 2018-2019. The authors carried out standard steps to prepare the data, such as converting categorical data and removing units of measurement such as “km” and “hp” from numerical data. The authors also used the K-fold cross-validation algorithm to build models. The arithmetic mean of the results of each model is chosen as the forecasting result. The researchers experimentally chose the value of K and tested several values. The best results, namely, 82% prediction accuracy, were shown by the model with $k = 4$, i.e., when four nearest neighbors are considered.

Work [17] was studied, describing various machine learning methods for price forecasting. A commonly used approach for price forecasting is multiple linear regression analysis. However, there are a large number of factors that affect the price and complicate the task. The paper mentions that the standard regression approach may not be suitable for high-dimensional data, and to overcome this potential problem, a modern method of data analysis that does not depend on the input dimension will be applied, namely, support vector regression. Prediction accuracy will be compared with a statistical regression model. The study [17, 18] also presents a fully automatic approach to the setup and application of SVR. The market analysis is carried out on actual data of cars of the German manufacturer.

In the article [8, 19], the research subject is machine learning methods' analysis and comparative characteristics, such as lasso regression, multiple regression, and regression trees.

This paper [5, 20] is about a platform created using machine learning technology. The study was based on several machine learning algorithms, such as linear regression, KNN, random forest, XG boost, and the decision tree algorithm. Based on these algorithms, statistical models were built that predicted the price of a used car. To do this, use previous consumer data and a given set of features. A comparative characterization of the prediction accuracy of these models was also conducted to determine the optimal one.

Paper [21] describes a study of the purchase and sale of cars on the secondary market using modern data mining technologies. This study's primary goal is to predict a vehicle's value using attributes highly correlated with the price. Linear regression, ensemble random forest, and ensemble bagging regressor methods were chosen as machine learning algorithms. The decision tree was used as a prediction algorithm in the bagging regressor. The initial data were preprocessed; namely, records with empty fields were removed and redundant attributes were removed. During the experiment, the data were split in an 80/20 ratio, where 80 percent is training data and 20 percent is testing data, respectively. Also, 40 different data splits were performed with different random state values for the splits. MSE, MAE, and RMSE were used as evaluation metrics. The best results were obtained for each method: random forest, approximate accuracy was 95%, 0.025 MSE, 0.0008 MSE, and 0.0378 RMSE; bagging regression, approximate accuracy was 85%; linear regression, 85%. The solution was to be integrated into a mobile or web application for public access.

In the article [22], the authors built a linear regression model that predicts the cost of vehicles with high accuracy, with a determination coefficient of 90%. The innovation of the article lies in a nonstandard approach to the selection of features to be taken into account in the model. The researchers use a particular sequence of several methods to identify potentially essential features for forecasting. The recursive feature elimination (RFE) method is the first in this sequence. The next step is to build an OLS regression, after which the variance inflation factor (VIF) is calculated, which indicates the dependence of one independent variable on another, which negatively affects the construction of the regression model. The last step is to discard the features with a high p value. After the abovementioned steps, the data are ready to build a machine-learning model. The data are divided into training and testing data, and then, the machine learning model is trained.

3. Materials and Methods

This section describes the methods used in a study that aimed to predict heavy machinery prices using machine learning models. The study utilized a large dataset of heavy machinery sales records covering various locations. The data were preprocessed using different techniques. Several machine learning models were trained, including linear regression, decision tree regression, and random forest regression, and their hyperparameters were optimized using a grid search. The performance of the models was assessed using various evaluation metrics such as mean absolute error and R-squared, and the best-performing model was selected and assessed using k-fold cross-validation to mitigate overfitting.

This research solves the problems of heavy equipment price forecasting, and this research uses a machine learning approach that involves collecting and structuring data from web resources, preprocessing the data, and using various machine learning algorithms to forecast prices accurately. The study pays additional attention to the specific characteristics of heavy machinery in management [6, 7].

The field of data science has revolutionized how we approach complex problems by allowing us to leverage the vast amounts of data generated by modern systems and technologies. One of the key challenges in data science is the ability to process and analyze large and complex datasets to extract meaningful insights and predictions [11, 23].

This mathematical model addresses this challenge by providing a comprehensive pipeline for data preprocessing and machine learning. This pipeline involves several stages of data processing, including parsing, structuring, tokenization, stemming, and data preprocessing. By following this structured approach to data processing, the mathematical model ensures that the data used in machine learning models are of high quality and relevance, which is critical for achieving accurate predictions.

Furthermore, the model employs advanced techniques such as feature selection and model evaluation to ensure that only the most relevant features are used in machine learning models (Figure 1). This helps reduce the dimensionality of your data and improve the accuracy and interpretability of your models.

3.1. Data Source. The dataset used in this study was obtained from an extensive database of heavy machinery sales records covering five years from 2017 to 2022. The dataset included information on the type, age, condition, and location of the machinery and other relevant features such as manufacturer, model, horsepower, and weight. The data covered several locations, with 50,000 observations.

The raw data were preprocessed using several techniques to clean and prepare it for analysis. Missing data points were removed using list-wise deletion. Outliers were removed using two popular methods, the Z-score method and the IQR method.

Categorical variables were encoded using one-hot encoding. The data were then standardized to have a mean of 0 and a standard deviation of 1 and split into training and testing sets using a 70/30 ratio.

Several new features were created from the existing data to improve the model's predictive power. These included a variable that calculated the distance between the machinery location and the nearest city center and another variable that represented the level of demand for machinery in the local market. In addition, polynomial features were added to capture nonlinear relationships between the features.

3.2. Model Selection. Several machine learning models were trained on the preprocessed data to predict heavy machinery prices. These included linear regression, decision tree regression, random forest regression, and gradient boosting regression. The hyperparameters for each model were optimized using a grid search, and the model with the best performance on the testing set was selected.

In our research, the choice of machine learning models was driven by a combination of factors, including the problem's nature, the dataset's characteristics, and the established best practices in construction equipment cost estimation. We selected models that have demonstrated effectiveness in similar regression tasks and are widely recognized in the construction industry.

Linear regression, decision trees, random forests, and neural networks were chosen for their interpretability, ability to handle nonlinear relationships, and scalability. These models have been extensively used in various construction-related studies, enabling us to establish a meaningful comparison with existing literature and industry practices. Moreover, they provide insights into the relative importance of input features, which is valuable for cost estimation in the construction equipment domain.

While it is true that other machine learning models, such as SVM, KNN, Light Boost, and CNN, have shown promising results in different domains, including some related to construction, their selection was not arbitrary. Our study aimed at balancing model complexity, interpretability, and performance, considering the specific requirements and constraints of estimating construction equipment costs.

Given the recent success of deep learning-based models, including CNNs, in various domains, we acknowledge their potential applicability to our problem. However, it is essential to note that deep learning models typically require large amounts of data to achieve their full potential and often exhibit increased complexity. In construction equipment cost estimation, where data availability can be limited and interpretability is crucial, the selected models provided a more practical and meaningful approach to achieving accurate and explainable results.

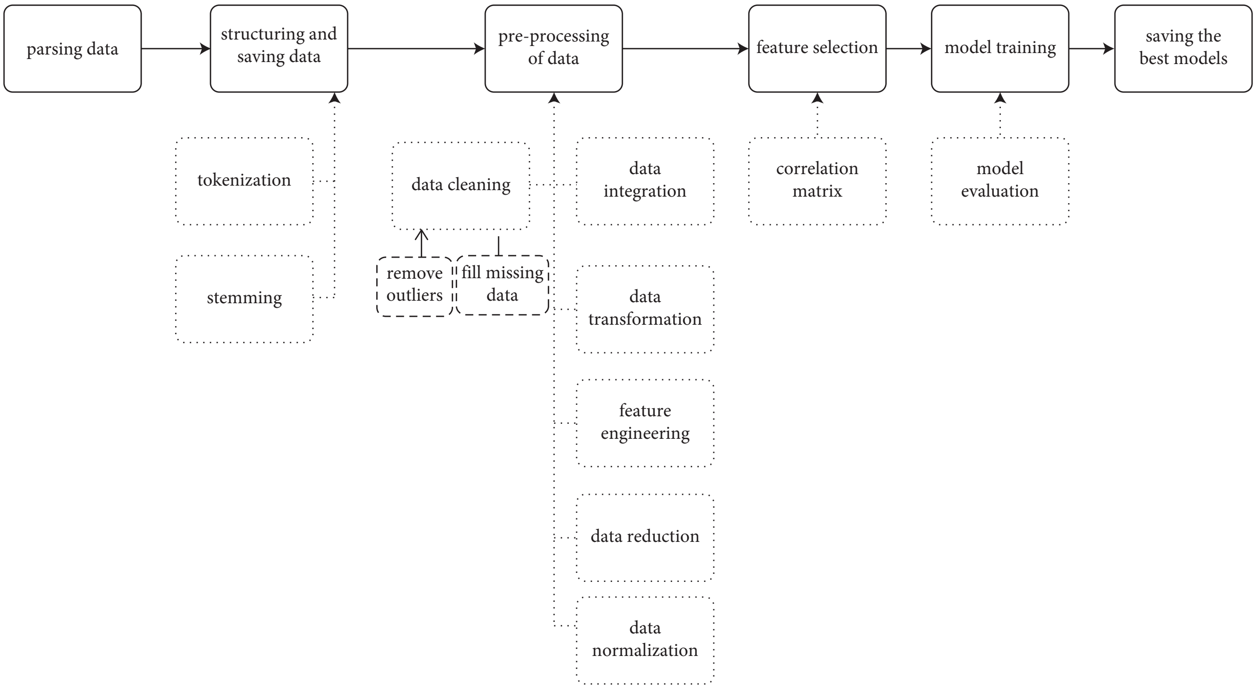


Figure 1: Mathematical model of research.

In summary, the selection of linear regression, decision trees, random forest, and neural networks as the chosen models in our study were driven by their interpretability, ability to handle nonlinear relationships, scalability, and their established usage in similar regression tasks within the construction industry. While we recognize the potential of other models, including deep learning-based approaches, the selection of models was carefully considered to meet the specific requirements and constraints of estimating construction equipment costs using machine learning methods.

3.3. Model Assessment. The performance of the selected model was assessed using several evaluation metrics, including

(i) Mean absolute error (MAE), MAE $= \sum_{i=1}^{n} |y_i - \hat{y}_i|/n$;

(ii) Mean squared error (MSE), MSE $= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/n$;

(iii) R-squared (R^2), $R^2 = 1 - (\sum_{i=1}^{n} (\hat{y}_i - y_i)^2)/(\sum_{i=1}^{n} (y_i - \overline{y}_i)^2)$.

The MAE and MSE were calculated as the average absolute and squared differences between the predicted and actual prices, respectively. The R^2 was calculated as the proportion of variance in the target variable explained by the model. In addition, a residual plot was generated to visually inspect the distribution of the residuals and check for any patterns or outliers.

3.4. Cross-Validation. To further assess the model's performance and mitigate overfitting, k-fold cross-validation was used. The data were divided into k subsets, and the model was trained and tested k times, with each subset used once for testing and the remaining subsets used for training. The mean and standard deviation of the evaluation metrics were calculated across the k iterations.

In summary, this study used a large dataset of heavy machinery sales records and several preprocessing and feature engineering techniques to prepare the data for analysis. Several machine learning models were trained and evaluated using various evaluation metrics, and the best-performing model was selected and assessed using cross-validation.

To further assess the efficacy of the ensembled models and mitigate overfitting, we employed k-fold cross-validation. The data were divided into k subsets, and the ensembled models were trained and tested k times. In each iteration, one subset was used for testing, while the remaining subsets were used for training the ensemble.

To clarify, the ensembled models combine multiple machine learning models, such as bagging or boosting techniques, to improve predictive performance. In our study, we utilized an ensemble approach to leverage the strengths of different models and enhance overall accuracy.

Now, let us address the concern regarding determining the number of folds (k). The choice of k depends on various factors, including the size of the dataset and computational resources. Our study carefully considered these factors and determined an appropriate value for k.

To determine the optimal number of folds, we conducted a preliminary analysis by evaluating the ensembled models' performance using different k values. We started with a conservative value of $k = 5$ and gradually increased it to assess the impact on model performance.

For each value of k, we calculated the evaluation metrics across the k iterations. We analyzed the results and selected the value of k that provided stable performance metrics without excessive computational demands.

After thorough experimentation, we found that a value of $k = 10$ yielded robust performance while ensuring reasonable computational efficiency. This value allowed us to obtain reliable estimates of the models' efficacy without excessively inflating the computational requirements.

In summary, this study utilized a large dataset of heavy machinery sales records and employed several preprocessing and feature engineering techniques to prepare the data for analysis. Using various evaluation metrics, we trained and evaluated several machines learning models, including ensembled models. To assess the efficacy of the ensembled models and mitigate overfitting, we employed k-fold cross-validation with a carefully chosen value of $k = 10$, which provided reliable performance estimates without excessive computational demands.

3.5. Linear Regression. A linear regression model is a simple and widely used approach for modeling the relationship between a dependent variable and one or more independent variables. In simple linear regression, there is one independent variable, while in multiple linear regression, there are multiple independent variables. The model assumes that the relationship between the dependent and independent variables is linear, which means that it can be represented as a straight line.

The linear regression model finds the best-fitting line through the data by minimizing the sum of the squared differences between the observed and predicted values (i.e., the residuals). The model equation can be represented as follows (equation):

$$y_i = b_0 + b_1 * x_{1i} + b_2 * x_{2i} + \ldots + b_n * x_{ni}, \tag{1}$$

where (i) y_i is the dependent variable (the predicted value on data index "i"). (ii) b_0, b_1, b_2, …, b_n are the coefficients (parameters) of the model. (iii) x_{1i}, x_{2i}, …, x_{ni} are the independent variables on data index "i."

Linear regression conducts the learning process through the ordinary least squares (OLS) method and does not have traditional hyperparameters like those used in iterative models.

Learning process in linear regression aims to find the best-fitting line (Figure 2) through the data, minimizing the sum of squared differences between the observed and predicted values (residuals). The learning process in linear regression involves finding the model's optimal coefficients (b-parameters) that minimize the sum of squared residuals (also known as the loss function).

In linear regression, there is no iterative optimization process like gradient descent, so the concept of epochs does not apply. The model's parameters (b-coefficients) are directly calculated using closed-form solutions like the ordinary least squares (OLS) method. Similarly, there is no learning rate (alpha-parameter) to tune in linear regression since the parameters are not iteratively updated. Therefore, the learning rate does not apply to linear regression.

Instead, some considerations related to feature selection and regularization can be crucial.

3.5.1. Feature Selection. In linear regression, feature selection is a vital aspect. We must decide which independent variables (features) to include in the model. Including irrelevant or highly correlated features can lead to overfitting and reduced interpretability. Feature selection techniques such as forward selection, backward elimination, or LASSO (Least Absolute Shrinkage and Selection Operator) can be used to select the most relevant features.

3.5.2. Regularization. Although not a hyperparameter, regularization techniques such as L1 (LASSO) and L2 (ridge) regularization can prevent overfitting and improve the model's generalization. These techniques add penalty terms to the cost function, encouraging the model to keep the coefficient values small.

In summary, the learning process in linear regression involves finding the optimal coefficients using ordinary least squares (OLS). The model does not require an iterative optimization process, and there are no traditional hyperparameters to tune. However, feature selection and regularization techniques play a crucial role in the performance and interpretability of the linear regression model.

3.6. Polynomial Regression. A polynomial regression model is an extension of linear regression where the relationship between the dependent variable and the independent variable is modeled as an nth-degree polynomial ($n > 1$). The degree allows for more complex curves to be fit to the data, going beyond straight lines (Figure 3).

The model equation (2) takes the form (3-th degree):

$$\begin{aligned} y_i = {} & b_0 + b_1 * x_{1i} + b_2 * x_{2i} + \ldots + b_n * x_{ni} \\ & + b_{11} * x_{1i}{}^2 + b_{12} * x_{1i} * x_{2i} + b_{13} * x_{1i} * x_{3i} + \ldots + b_{nn} * x_{ni}{}^2 \\ & + b_{111} * x_{1i}{}^3 + b_{112} * x_{1i}{}^2 * x_{2i} + b_{113} * x_{1i}{}^2 * x_{3i} + \ldots + b_{nnn} * x_{ni}{}^3, \end{aligned} \tag{2}$$

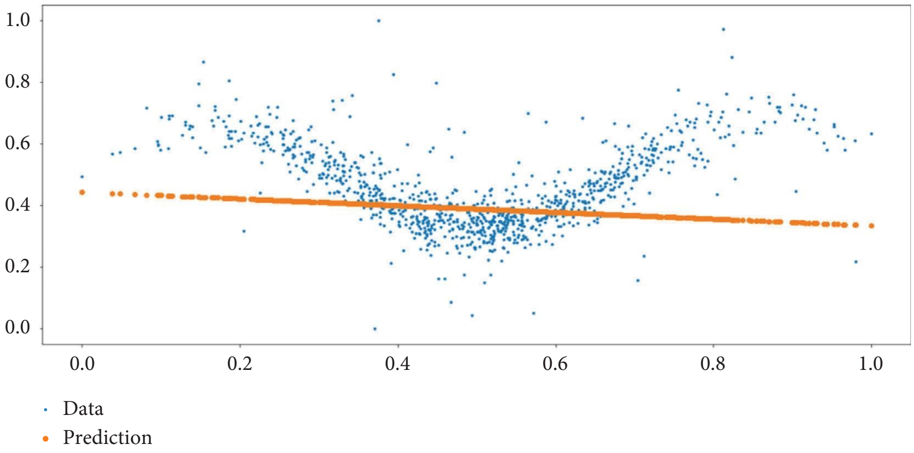


Figure 2: Visualization of the work of linear regression.

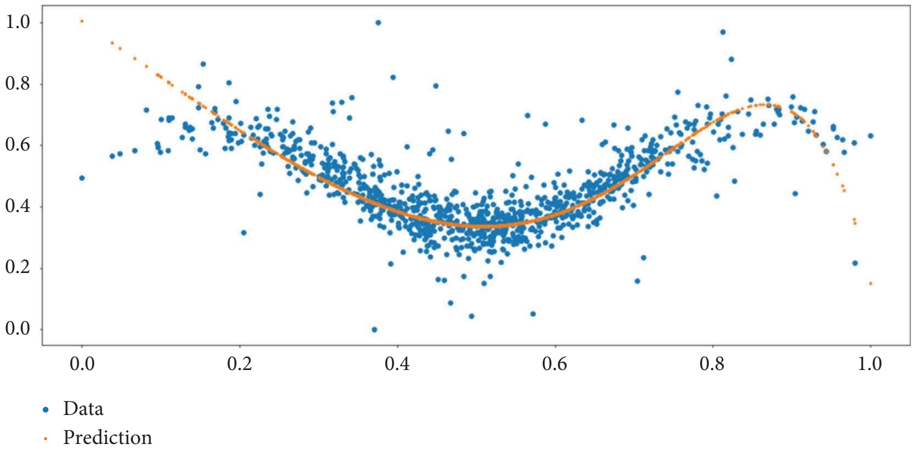


Figure 3: Visualization of the work of polynomial regression.

where (i) y_i is the dependent variable (the predicted value). (ii) $x_{1i}, x_{2i}, \ldots, x_{ni}$ are the independent variables (features). (iii) $b_0, b_1, b_2, \ldots, b_n$ are the coefficients corresponding to the first-degree terms (linear terms) of each independent variable. (iv) $b_{11}, b_{12}, \ldots, b_{nn}$ are the coefficients corresponding to the second-degree terms (quadratic terms) of each independent variable. (v) $b_{111}, b_{112}, \ldots, b_{nnn}$ are the coefficients corresponding to the third-degree terms (cubic terms) of each independent variable.

And so on for higher-degree terms if necessary.

The learning process in polynomial regression is similar to that of linear regression. However, polynomial regression introduces additional complexity due to the higher-degree terms. The learning process in polynomial regression aims to find the optimal coefficients for the polynomial terms that best fit the data. The model equation is a polynomial of higher degrees, and the learning process involves estimating the coefficients that minimize the sum of squared residuals between the observed and predicted values.

The coefficients ($b_0, b_1, \ldots, b_{nnn}$) are estimated from the training data using methods like ordinary least squares (OLS) or other optimization techniques to minimize the sum of squared residuals between the observed and predicted values. The data can be represented in a matrix form to perform the estimation, and the coefficients can be calculated using linear algebra. This process involves solving a system of linear equations, and the closed-form solution for the coefficients can be found.

Hyperparameters in polynomial regression are mainly related to the selection of the degree of the polynomial and the regularization technique used to control model complexity.

3.6.1. Degree of Polynomial. The polynomial degree (often denoted as "*d*") is a crucial hyperparameter in polynomial regression. It determines the complexity of the model and the flexibility to fit the data. Higher degrees can capture more complex relationships but may lead to overfitting, especially with limited data.

3.6.2. Regularization. Regularization techniques can be applied in polynomial regression to prevent overfitting. L1 regularization (LASSO) and L2 regularization (ridge regression) are the most common methods.

L1 Regularization (LASSO): L1 regularization adds a penalty term to the cost function based on the absolute values of the coefficients. It encourages some coefficients to become exactly zero, effectively performing feature selection.

L2 Regularization (Ridge Regression): L2 regularization adds a penalty term based on the squared values of the coefficients. It tends to shrink the coefficients towards zero without necessarily eliminating them.

The choice between L1 and L2 regularization depends on the specific requirements and characteristics of the dataset. Hyperparameter tuning techniques, such as cross-validation, can be used to determine the optimal regularization strength (lambda) for the chosen method.

In summary, polynomial regression conducts the learning process by estimating the coefficients best fitting the higher-degree polynomial equation to the data. The primary hyperparameters in polynomial regression are the degree of the polynomial and the regularization method and strength. Proper hyperparameter selection is essential to create a well-performing polynomial regression model that balances complexity and generalization ability.

3.7. Decision Tree Regressor. Decision trees are a nonlinear model used for both classification and regression tasks. In the context of regression, the decision tree algorithm splits the data into segments based on the independent variables' values to predict the dependent variable's value. The tree structure consists of nodes representing the splits and leaf nodes containing the predicted values.

At each step, the decision tree algorithm selects the best variable and split point to minimize the mean squared error (or other regression metric) of the predicted values within each segment. The prediction for a new data point is made by following the path down the tree until reaching a leaf node and using the average value of the dependent variable within that leaf node as the predicted value (Figures 4 and 5).

The learning process in the decision tree model for regression involves recursively splitting the data based on the selected features to create a tree-like structure, where each leaf node represents a prediction for the target variable. The selection of hyperparameters in decision trees is critical to control the tree's complexity and prevent overfitting.

The learning process in decision tree regression involves the following steps:

> Splitting Criteria: The decision tree starts with the entire dataset at the root node. To create the tree, it iteratively searches for the best feature and split point that reduces the variability in the target variable the most. The commonly used measures for regression are the sum of squared residuals (mean squared error) or mean absolute error.
>
> Recursive Splitting: Once the initial split is made, the process recursively for each subset (child nodes) until a stopping criterion is met. The stopping criteria could be a maximum depth for the tree, a minimum number of samples required to split a node, or a minimum number of samples required to be at a leaf node.
>
> Prediction at Leaf Nodes: At each leaf node, the average value of the target variable (or another suitable value) for the samples within that leaf node is used as the prediction.
>
> Model Representation: The learned decision tree can be represented in a tree-like structure, where each internal node represents a feature and a split point and each leaf node represents the predicted value.

Hyperparameters in decision trees control the model's structure and behavior. Proper selection of hyperparameters is essential to prevent overfitting and achieve better generalization. Common hyperparameters in decision trees for regression include

> Max Depth (max_depth): The maximum depth of the decision tree, i.e., the maximum number of levels between the root node and the deepest leaf node. Limiting the depth helps prevent the tree from becoming too complex and overfitting.
>
> Min Samples Split (min_samples_split): The minimum number of samples required to split an internal node. A node with fewer samples than this value will not be split further. Increasing this parameter can help avoid creating small, less representative splits.
>
> Min Samples Leaf (min_samples_leaf): The minimum number of samples required to be at a leaf node. If a leaf node has fewer samples than this value, it might be merged with its sibling node or its parent node. Larger values can prevent the tree from creating leaves with very few samples.
>
> Max Features (max_features): The maximum number of features to consider when looking for the best split. Limiting the number of features can help prevent the tree from focusing too much on individual features and improve generalization.
>
> Min Impurity Decrease (min_impurity_decrease): The minimum impurity decrease required to split a node. This hyperparameter helps control the tree's growth by considering splits that result in a specific impurity reduction.
>
> Min Impurity Split (min_impurity_split): The minimum impurity threshold for a node to be split. This hyperparameter can prevent further splits if the impurity of a node is below the threshold.

In practice, hyperparameter tuning techniques such as grid search, random search, or Bayesian optimization can explore different hyperparameter combinations and find the optimal settings that result in the best-performing decision tree model for regression. Proper hyperparameter tuning is essential to achieve a balanced and well-generalizing decision tree model.

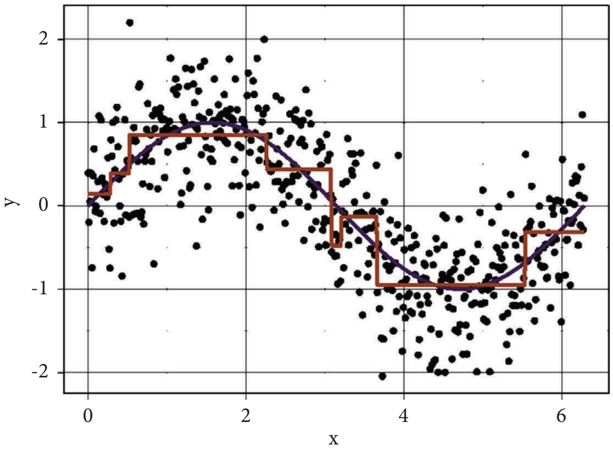


Figure 4: Visualization of the work of the decision tree regressor (point: data value; black line: predicted by polynomial regression; red line: predicted by decision tree regressor).

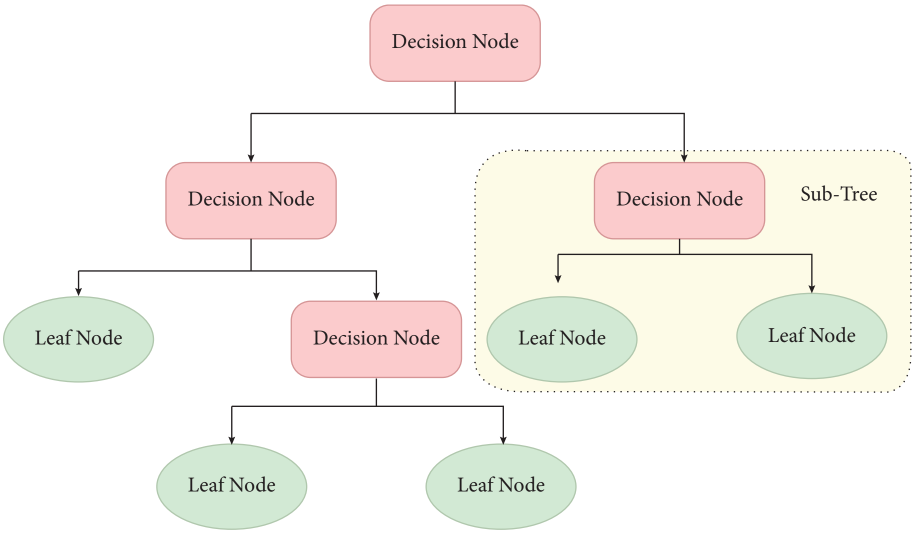


Figure 5: Visualization of the work of decision tree regressor in the schema.

The main hyperparameters are

(i) Maximum depth of the tree: determines the maximum number of levels in the decision tree.

(ii) Minimum number of samples required to split an internal node: determines the minimum number of samples required to split a node.

(iii) Minimum number of samples required to be at a leaf node: determines the minimum number of samples required to be at a leaf node.

(iv) Maximum number of leaf nodes: limits the number of leaf nodes in the decision tree.

(v) Criterion (squared_error, friedman_mse, absolute_error, poisson): determines the function to measure the quality of a split.

3.8. Random Forest Regressor. Random forest is an ensemble learning technique that combines multiple decision trees to improve the accuracy and robustness of the regression model. Each tree in the forest is built using a random subset of the data and a random subset of the features. The final prediction is obtained by averaging (for regression) the predictions of all individual trees.

The randomization and averaging reduce overfitting and make the model more resilient to noise and outliers. Random forest is a powerful regression model suitable for large and complex datasets (Figure 6).

The learning process in random forest regression involves building multiple decision trees on random subsets of the data and features. The random forest model combines the predictions from individual trees to produce the final regression output. The selection of hyperparameters in random forest is crucial to control the ensemble's complexity and ensure good performance.

The learning process in random forest regression involves the following steps:

Bagging (Bootstrap Aggregating): Random forest starts by creating multiple bootstrap samples (random samples with replacement) from the original training

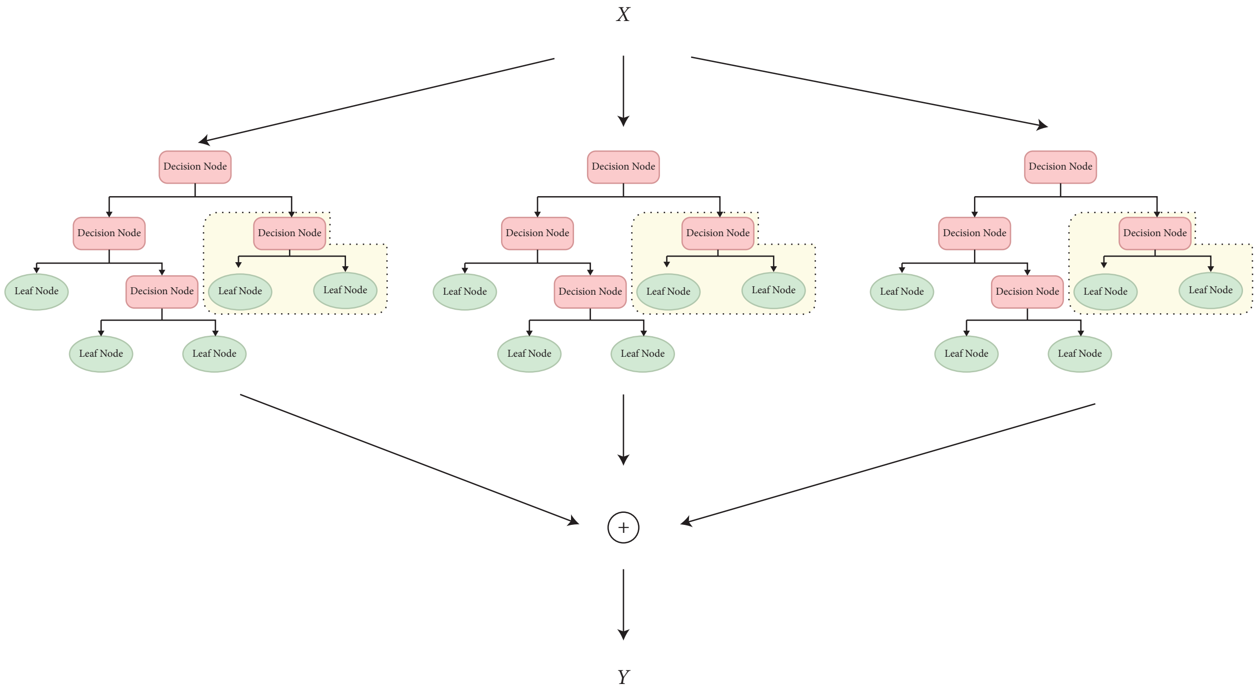


Figure 6: Visualization of the work of random forest regressor in schema.

dataset. Each bootstrap sample will have the same number of data points as the original dataset but with some variations.

Building Decision Trees: A decision tree is built for each bootstrap sample, typically using a subset of the available features. At each node split, only a random subset of features is considered. This introduces diversity among the trees.

Voting (Aggregating Predictions): When predicting new data points, each decision tree in the random forest makes an individual prediction. Regression predictions from all trees are averaged to produce the final output. This ensemble approach helps reduce variance and improve generalization.

Hyperparameters in random forests control the ensemble's behavior and individual tree's characteristics. Properly selecting hyperparameters is vital to avoid overfitting and improve the model's performance. Common hyperparameters in random forest regression include

Number of Trees (n_estimators): The number of decision trees to be included in the random forest ensemble. More trees generally lead to better performance, but there is a diminishing return with additional trees. However, more trees also increase computation time, so it is essential to find a balance.

Max Depth (max_depth): The maximum depth of each decision tree in the random forest. Limiting the depth can prevent overfitting and promote a more interpretable model. If not set, the trees can grow until they contain very few samples in the leaves.

Min Samples Split (min_samples_split) and Min Samples Leaf (min_samples_leaf): Like decision trees, these hyperparameters control the minimum number of samples required to split an internal node and the minimum number required to be at a leaf node.

Max Features (Max_Features): The maximum number of features to consider when looking for the best split. Randomly selecting a subset of features at each split helps introduce diversity among the trees and improves the model's robustness.

Bootstrap Samples (bootstrap): This option specifies whether to use initial samples to generate the training data for each tree or not. If set to False, the entire original dataset is used to build each tree, which may result in trees with high similarity and decreased diversity.

Random State: The random seed is used to initialize the random number generator. Setting a random seed ensures the reproducibility of the results.

Proper hyperparameter tuning is crucial to achieving a well-generalizing and accurate random forest regression model.

Some of the hyperparameters of a random forest regressor include

(i) n_estimators: The number of decision trees in the forest.

(ii) max_features: The maximum number of features considered for splitting a node.

(iii) max_depth: The maximum depth of the decision trees.

(iv) min_samples_split: The minimum number of samples required to split an internal node.

(v) min_samples_leaf: The minimum number of samples required to be at a leaf node.

(vi) bootstrap: A Boolean parameter indicating whether bootstrap samples are used when building decision trees.

3.9. Neural Network. A neural network for regression is a supervised machine learning algorithm that uses an artificial neural network to predict a continuous output variable. The structure of a multilayer perceptron was taken as the basis for building the network. A neural network consists of an input layer, a hidden layer, and an output layer. There can be several hidden layers. Each layer consists of the nth number of neurons and has an activation function associated with the neurons. The activation function is responsible for creating nonlinearity in the relationship. In our case, the output layer should contain a linear activation function (Figure 7).

The construction of a neural network consists of two stages: forward propagation and backward propagation.

(i) Forward propagation moves input data through a neural network to produce an output prediction. During this process, the input data are multiplied by the weights of the neurons in the hidden layers, and an activation function is applied to the result. The output from one layer becomes the input to the next layer until the output layer is reached.

(ii) Backpropagation is the process of calculating the gradient of the loss function with respect to the weights of the neural network. It is used to update the weights during training. During backpropagation, the derivative of the loss function is calculated with respect to the output of the network. This derivative is then propagated backward through the network, layer by layer, using the chain rule of calculus to calculate the gradient of the loss function with respect to the weights in each layer. The weights are then updated in the opposite direction of the gradient to minimize the loss function.

In short, the input layer is fed with our data; in the hidden layers, there is a combination of various features to train the model, and at the output, we get the prediction results.

A neural network for regression is a supervised machine learning algorithm that uses an artificial neural network to predict a continuous output variable. The structure of a multilayer perceptron was taken as the basis for building the network.

A neural network model for regression consists of an input layer, one or more hidden layers, and an output layer. In the context of regression, the output layer typically contains a single neuron that directly gives the predicted value.

Each neuron in the hidden layers applies a weighted sum of its inputs, adds a bias term, and then applies an activation function to produce an output. The weights and biases are learned during training using optimization techniques like gradient descent.

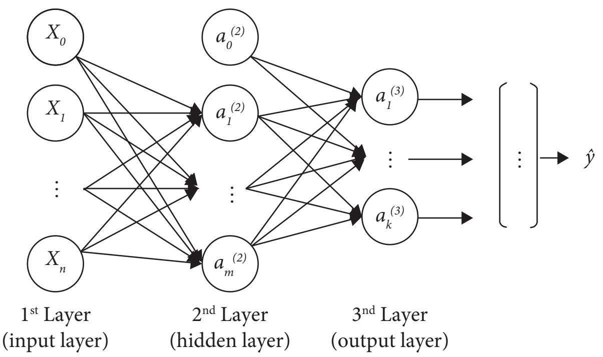


Figure 7: Visualization of the work of neural network in schema.

The activation function is responsible for creating nonlinearity in the relationship. In our case, the output layer should contain a linear activation function (Figure 7). The choice of activation function in the output layer depends on the specific requirements of the regression task. For unbounded continuous predictions, a linear activation function can be used. If the output needs to be bounded within a specific range, other activation functions like sigmoid or tanh might be more suitable.

Neural network (perceptron) model regression conducts the learning process through forward and backward passes, also known as forward propagation and backpropagation. The selection of hyperparameters in a neural network is critical to control its architecture, learning rate, and regularization, ensuring adequate training and preventing overfitting.

The learning process in a neural network (perceptron) regression involves the following steps:

Model Architecture: Define the architecture of the neural network, including the number of layers, the number of neurons in each layer, and the activation functions used in each neuron. For regression, the output layer typically consists of a single neuron with a linear activation function (or no activation function).

Forward Propagation: During forward propagation, input data are fed through the neural network, layer by layer, to generate predictions. The activations of neurons in each layer are calculated using weighted sums of the previous layer's outputs and appropriate activation functions.

Loss Function: A loss function is defined to measure the difference between the predicted output and the actual target values. Commonly used loss functions for regression tasks are mean squared error (MSE) or mean absolute error (MAE).

Backpropagation: Backpropagation updates the neural network's weights and biases to minimize the loss function. It calculates the gradients of the loss for the network's parameters and uses gradient descent (or its variants) to update the weights and biases in the direction that minimizes the loss.

Update Weights and Biases: The updated weights and biases are obtained by multiplying the gradients with the learning rate (step size) and subtracting them from the current weights and biases.

Repeat: The forward propagation and backpropagation steps are repeated for multiple epochs or until the model converges to a satisfactory solution.

Hyperparameters in a neural network govern the model's architecture and learning process. The selection of hyperparameters is essential for successful training and model performance. Common hyperparameters in neural network regression include

A Number of Layers and Neurons: Each layer's number of layers and neurons determines the network's architecture and complexity. Too few neurons may not capture complex patterns, while too many may lead to overfitting. Hyperparameter tuning can help find an optimal balance.

Activation Functions: The choice of activation functions in hidden layers can impact the network's ability to model complex relationships. Common choices include ReLU (Rectified Linear Unit) and its variants. A linear activation function is typically used for the regression output layer.

Learning Rate (Alpha): The learning rate determines the step size for updating the weights during gradient descent. A high learning rate may result in overshooting the optimal weights, while a low learning rate can cause slow convergence.

Batch Size: The batch size determines the number of samples used in each forward and backward pass. It can affect training speed and memory usage. Common choices include batch gradient descent, mini-batch gradient descent, and stochastic gradient descent (SGD).

Number of Epochs: The number of epochs defines the number of times the entire training dataset is used during training. Too few epochs may result in underfitting, while too many may lead to overfitting.

Some of the hyperparameters of a neural network for regression include

(i) A number of hidden layers: The number of layers in the neural network between the input and output layers.

(ii) Number of neurons per hidden layer: The number of nodes in each hidden layer.

(iii) Activation function: The function introduces nonlinearity in the neural network.

(iv) Learning rate: The step size taken during gradient descent optimization to update the weights in the neural network.

(v) Batch size: The number of samples used to update the weights during each iteration of the optimization algorithm.

(vi) Number of epochs: The number of times the entire training dataset is passed through the neural network during training.

4. Methodology

4.1. Input Data. The input data in the central part of the implemented system should be considered perfectly formed and structured data sets with all the essential characteristics for further system operation. Since finding a large amount of similarly structured data is impossible, it was decided to develop additional parts of the system that will be responsible for data retrieval, processing, and structure.

The data processing module will receive all unstructured data from the previous module (Figure 8). It was also decided to create a convenient user interface for working with the system.

Many of these data columns duplicate information; for example, the columns "Manufacturer" and "Original manufacturer" and "Price" and "Original price" contain the same information, so combining each pair into one was decided. Under the statement "had the same information," it meant that they could have a difference in writing names or currency in different languages depending on the country where the ad was located. All columns that contain "Original" in their name fall under such criteria.

The dataset is quite extensive, and several files containing about 50,000 observations were collected based on the web ads selling such equipment. The data contain technical information about the vehicle, such as manufacturer, model, weight in tons, an extra column with unique details for a model of vehicle, rated operating capacity in kilograms, types of tires, and width, transmission type, transport length, and width on meters, bucket width in meters, and capacity on cubes meter, driver protection is a cabin type, load capacity, and maximum lifting height in meters. It also includes such essential attributes as the year of vehicle manufacture, time of operation, cost in euros and US dollars, country of sale, and others.

4.2. Output Data. The search part of the system will return significant amounts of partially structured and unstructured data from various web portals.

Then, the module responsible for processing and structuring the data will return a single data structure containing all the essential vehicle price prediction characteristics.

The next and central part of the system is to build and train a machine learning model for predicting the cost of heavy equipment. The output data are the model itself and the model evaluation results.

The final part of the system is the interface. The output data will be the predicted market value of the equipment the model provides on the vehicle's obtained characteristics.

The methodology for processing the data can be divided into several stages.

4.3. Parsing Data. Data from web resources will be collected for heavy equipment categories, such as excavators, tractors, bulldozers, cranes, loaders, and rammers. Each category has specific important characteristics for forecasting value and standard features such as model, brand, year of manufacture, power, and number of hours worked. A separate module will be used to process the data and structure their essential characteristics according to the category of heavy machinery. Unstructured data will be tokenized based on key features of the vehicles, and the frequency of repetition of keywords in the description of each line will be analyzed. The most common and important words will be extracted to structure the data into the desired set of "key-value" pairs for further operation.

Two approaches are used to obtain data from web resources:

(i) The first and more straightforward approach is to use off-the-shelfsolutions-web scraping tools designed to extract and collect any open information from websites intended for human viewing [11]. This software searches for information under the user's control or automatically. Data that meet the configured parameters are selected and stored in the desired form and according to a specific structure. Such tools allow you to retrieve information from a website using a user-friendly interface without writing a single line of code. Such solutions are expensive and need the flexibility that solutions developed specifically for a particular website have.

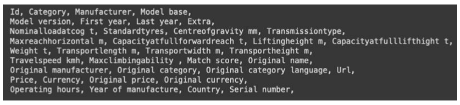
(ii) Another method of information extraction is to use algorithms to search and navigate through the desired websites by mimicking human behavior and copying the information needed using keywords. This approach is suitable for finding information on any website, as it returns unstructured data, regardless of the data structure of a particular web portal.

4.4. Data Preprocessing. After receiving data from web resources, they must be brought to the same form and key-value structure. A separate module was used to process the data and structure their essential characteristics according to the heavy equipment category.

First, it is necessary to tokenize the unstructured data on the critical features of vehicles and analyze the frequency of keywords in the description of each line. The most common and important words should be taken into account. Next, a splitting operation should extract and structure these values into the desired key-value dataset for further processing.

The next step is to transform the data into the appropriate form using various preprocessing methods to implement machine learning algorithms further.

Preprocessing in machine learning is an essential step that helps improve data quality by retrieving only relevant information. Data preprocessing [3] in machine learning refers to preparing (cleaning and organizing) raw data to make it suitable for building and training various machine learning models.



```
Id, Category, Manufacturer, Model base,
Model version, First year, Last year, Extra,
Nominalloadatcog t, Standardtyres, Centreofgravity mm, Transmissiontype,
Maxreachhorizontal m, Capacityatfullforwardreach t, Liftingheight m, Capacityatfulllifthight t,
Weight t, Transportlength m, Transportwidth m, Transportheight m,
Travelspeed kmh, Maxclimbingability , Match score, Original name,
Original manufacturer, Original category, Original category language, Url,
Price, Currency, Original price, Original currency,
Operating hours, Year of manufacture, Country, Serial number,
```

Figure 8: Columns of data before the structure and preprocessing for the loader's heavy machine type.

Data preprocessing in machine learning consists of several stages:

(i) Finding and removing duplicate data

(ii) Removal of anomalous data (outliers)

A critical stage of data preprocessing is the removal of so-called outliers. Outliers are specific objects that are very different from the general population. They have characteristics that differ from most other objects in the data set (Figure 9).

Outliers can be of two types:

(i) A univariate outlier is a data point that consists of an extreme value of a single variable.

(ii) A multivariate outlier is a combination of unusual values for at least two variables.

The presence of outliers in a dataset can negatively impact the quality of machine learning model training. Therefore, the main stage of data preprocessing is outlier removal.

For the detection of emissions, it is used:

(i) IQR: Interquartile range or interquartile range method.

Not all data are normally distributed; in this case, the standard deviation method cannot be used. An excellent way to summarize the distribution of Gaussian data is the interquartile range.

The interquartile range is calculated as the difference between the 75th and 25th percentile data. Statistics-based emissions detection methods assume that emissions occur in regions with low probability stochastic patterns, and therefore, data occur in regions with high probability. Thus, the interquartile range method can identify emissions by defining limits for sample values, the so-called k IQR parameter, that is below the 25th percentile or above the 75th percentile. The general value of the k factor is 1.5. Any value that falls outside the range of $-1.5 \times \text{IQR}$ to $1.5 \times \text{IQR}$ is considered abnormal (Figure 10).

(ii) Standard deviation or the method of standard deviation

Standard deviation is a valuable measure of distribution for normal distributions. Because in normal distributions, data are distributed symmetrically without skewness. Most values are centered around the central region, decreasing as they move away

Figure 9: Detection of outlier values.

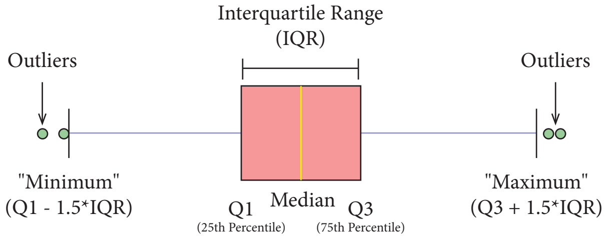


Figure 10: Interquartile range method.

from the center. The standard deviation indicates how many positions, on average, the data are located from the center of the distribution.

The normal distribution contains two important parameters, the mean and standard deviation (mean and standard deviation, respectively). In a normal distribution, these two parameters can be used to evaluate atypical data in a sample. The standard deviation and mean can indicate the values' position in a normal distribution.

The rule of thumb, or the 68-95-99.7 rule, states that 68% of the data are within one standard deviation, two standard deviations-95%, and three standard deviations-99.7% (Figure 11).

Data outside the three standard deviations are also part of the population, but these are atypical or rare cases. Therefore, as a rule of thumb, all data outside three standard deviations are considered outliers and should be removed from the population. This value may vary depending on the size of the data set: if the data set is too large, the data may fall outside four standard deviations and vice versa; if the data set is small, the data may fall outside two standard deviations.

The standard deviation can be represented as a bell curve, with a flatter or more open shape representing a significant standard deviation and a steep, high bell curve representing a slight standard deviation (Figure 12).

To eliminate outliers, can use the following methods:

(i) *Anomaly removal*: remove all outlier values from the dataset.

(ii) *Value transformation*: take the natural logarithm as the new data, as it reduces the deviation values that were caused by extreme values

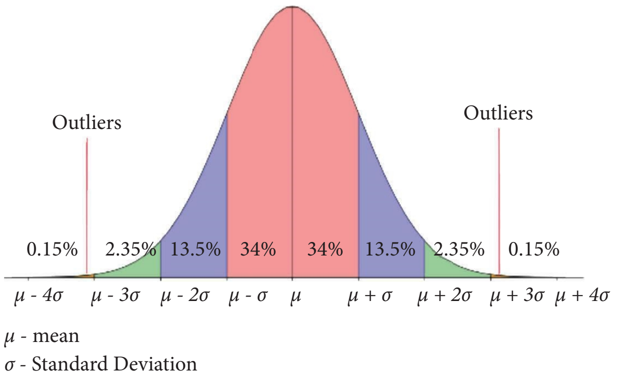


Figure 11: Standard deviation method.

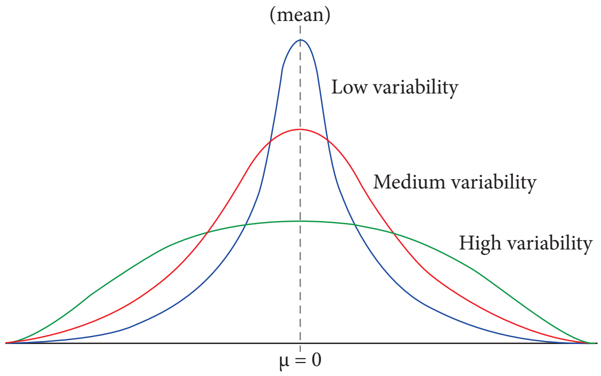


Figure 12: Variations in the standard deviation of data.

(iii) *Data splitting*: if many grouped emissions exist, they should be considered separately. The approach is to treat the groups differently and build the model individually for each. In the end, we need to combine the results.

4.4.1. Detection and Processing of Missing Values of Certain Characteristics. The next and no less important stage of data preprocessing is detecting and processing missing values of specific characteristics.

In real-world tasks, data may contain gaps. This may be because the user still needs to fill in all the fields in the ad or account or because not all parameters have been digitized throughout the system's life cycle. So, the question is how do you handle these gaps.

(i) The simplest options are to exclude objects with incomplete information (i.e., delete rows from the feature matrix with gaps in the columns) or to exclude features with incomplete information (i.e., delete columns with gaps); this method is called deleting a particular row. The advantage of this approach is that it is straightforward. The disadvantages are obvious: if many records have gaps, it can remove essential patterns hidden in the data. As a result, we need an accurate approximation. If you need to drop columns containing key features, you risk losing information about the data's dependencies.

(ii) The second option for handling gaps is to replace them with interpolation. This can be the average or median of the columns. If the property is a function of time and a single object, you can only interpolate gaps to adjacent time values. Advantages: no data loss, works well with small data sets, and is easy to implement. Disadvantages: works only with numeric data and does not consider covariance between attributes. If the data are categorical, missing values can be replaced with the most frequent value. However, this method can increase the variance of the data set. At the same time, data loss can be avoided.

(iii) The third option includes open economic, social, or other parameters that can be found in other sources. This way, the dataset is supplemented with data.

(iv) The fourth option is to encode the space using numerical or categorical values.

(v) The fifth approach involves an expert in the relevant field who can tell you how best to interpolate the data. For example, for some functions, you can generate pseudorandom values influenced by properties that consider other known features. This also includes creating synthetic data. Long-term and careful data analysis can reveal the behavior of hidden parameters and identify ways to model them with a certain probability. However, this approach is dangerous because it needs to be more accurate to put other patterns into the data and then find them using machine learning methods. In other words, instead of solving the original problem, you need to solve a contrived problem.

(vi) Sometimes, a variation of the abovementioned approaches is used in practice. It all depends on the specifics of the task. Some records or columns are removed because interpolating or simulating their values is difficult. As practice shows, these are features with many values that need to be added (for example, where there are more gaps than actual values). Sometimes features are characterized by specific values and parameters, so it is only possible to simulate them correctly by studying the scope of the application. But if the application of the methods does not help achieve the expected result, then there is probably not enough data.

4.4.2. Categorical Data Coding. Since machine learning models involve complex mathematical calculations, the values in a dataset must be in numerical form to be calculated. However, not all features of objects are described by a numerical value. If we are talking about the size of an object or the cost of a product, these features will undoubtedly be numerical. If we are talking about the color, type of product (category), or textual description of an object, then such features are usually not digitized.

Therefore, it is essential to convert all text values to numeric values. There are several ways to convert categorical data to numeric. Each approach has its advantages and impact on the set of functions.

Let us look at two pretty popular approaches to converting categorical data into numerical data:

(i) Label Encoder

This simple approach involves converting each value in the column to a number. In label encoding, we must replace the categorical value with a numeric value from 0 to the number of classes minus 1 (Table 1). If the value of a categorical variable contains four different classes, then 0, 1, 2, and 3 should be used.

However, this coding algorithm has a rather significant disadvantage. The problem with using numeric data is that it introduces a relationship/comparison between categories. For example, we have four options for vehicle colors: red, blue, green, and yellow. Moreover, we need to find out which is better and more popular, which is worse, and how it affects the cost of the product.

There is no connection between the colors, but looking at the number, one of the colors has a higher priority than the other. Thus, the machine learning model can assume some correlation between these variables.

If the column needs to be ordered/prioritized, for example, to create a security level (high, medium, and low), then this coding method is entirely appropriate.

(ii) One Hot Encoding

One Hot Encoding is another popular method for processing categorical variables. It creates additional functions based on the number of unique values in the categorical object. Each unique value in the category will be added as a function.

In this approach, a new column (sometimes called a dummy variable) with a binary encoding (0 or 1) must be created for each category of the object to determine whether a particular row belongs to that category (Table 2).

While this approach eliminates hierarchy/order issues, it has the downside of adding additional columns to the data set. This can significantly increase the number of columns if the category column has many unique values.

4.4.3. Splitting Data. Each dataset is divided into training and test datasets before being fed into the machine learning model.

The training dataset selects the appropriate machine learning model parameters that best fit the test dataset. The model selects the appropriate parameters using optimizers. Most often, this is gradient descent.

The test data set provides an unbiased evaluation of the machine learning model after training. The evaluation will be unbiased because the model did not use these data for training.

Table 1: Example of using the label encoder.

Color		
Red	➡	0
Blue	➡	1
Green	➡	2
Yellow	➡	3

Table 2: Example of using the one hot encoding.

Color		Red	Blue	Green	Yellow
Red	➡	1	0	0	0
Blue	➡	0	1	0	0
Green	➡	0	0	1	0
Yellow	➡	0	0	0	1

Splitting the data into so-called validation datasets is a good tactic. As a result, an unbiased evaluation is provided during the training phase of the model, which will be used for regularization by stopping early to avoid overtraining and deterioration of the original model.

4.4.4. Scaling a Data Set. The values of the raw data vary greatly, and this can lead to biased model training or, ultimately, to increased computational complexity. Raw data have a different scale and distribution for each attribute. Therefore, it is essential to normalize them. Feature scaling is a technique that brings data values into a shorter range.

For example, one characteristic may have measurement ranges from 0.0001 to 0.2, while another may range from −100 to 100. For example, a customer's age may be between 16 and 40, but most customers are between 18 and 25, so the mathematical expectation is shifted to the center of the distribution. This characteristic difference can cause significant errors in many models (e.g., for regression, neural networks). Therefore, it is necessary to integrate all functions into one form.

There is some confusion about the term's "standardization" and "normalization." Standardization and normalization are often seen as different things and sometimes as part of normalization. Therefore, it is essential to understand these methods' general nature and purpose.

Data standardization is the process of making the vector of each attribute look like a vector so that its mathematical expectation becomes zero and its variance becomes one.

Let us look at how we can scale data to standardize it:

(i) Scaling using minimum and maximum values (min_max_scaled) (equation).

$$X_i\text{scaled} = \frac{X_i - \min(X)}{\max(X) - \min(X)}. \tag{3}$$

This approach scales each variable so that it falls within a certain range of the training dataset, for example, from 0 to 1.

(ii) Scaling variables by maximum absolute value (max_absolute_scale) (equation).

$$X_i\text{scaled} = \frac{X_i}{\max(|X|)}. \tag{4}$$

This function scales each variable so that the maximum absolute value of each attribute in the training set is 1. Since the algorithm does not move or center the data, it does not violate sparsity.

(iii) Scaling the variable to the standard deviation (standard_deviation_scale) (equation).

$$X_i\text{scaled} = \frac{X_i - \text{mean}(X)}{\text{std}(X)}. \tag{5}$$

Data normalization is the process of scaling the vector of each feature. All vectors must have the same scale; normalization is required. There are several types of norms and normalizations:

(iv) Max-norm (equation)

$$X_i\text{norm} = \frac{X_i}{\max(X)}. \tag{6}$$

To ensure that all values are within the range, we need to find the maximum possible value and divide all other values by it. Therefore, the maximum value is one, and all other values fall within the range from 0 to 1, provided that there are no negative values.

(v) L1-norm (equation)

$$X_i\text{norm} = \frac{X_i}{\sum_{i=1}^{m} X_i}. \tag{7}$$

Each value must be divided by the sum of the values of the given distribution.

(vi) L2-norm (equation)

$$X_i\text{norm} = \frac{X_i}{\sqrt{\sum_{i=1}^{m} X_i^2}}. \tag{8}$$

This approach is also called the Euclidean norm.

Attribute scaling is the final stage of data preprocessing in machine learning. This method brings the independent variables of a dataset into a specific standard range of values. In other words, attribute scaling limits the range of variables so that the model can compare them with all others, regardless of the scale of the value.

Less popular methods can also be used:

(i) Data integration, combining data from multiple sources into a single dataset for analysis. This can involve merging datasets based on standard variables, such as location or machinery type, and resolving discrepancies or inconsistencies between data sources. For example, data on heavy machinery's age, condition, and location can be combined with data on local market conditions, such as demand, competition, and economic indicators. These datasets can be merged based on standard variables such as machinery type, location, and time period.

(ii) Feature engineering, identifying and creating new variables relevant to the analysis. For example, a new variable can be created that calculates the average machinery price in a given location or a variable that represents the distance between a machinery's location and the nearest industrial park. Another example could be to create a variable for the demand for machinery in a particular location based on the number of inquiries or sales for machinery of a similar type.

These are just a few examples of how data preprocessing techniques can be applied to raw data related to forecasting heavy machinery prices based on geolocation and market features. The actual values used in these techniques will depend on the specific dataset being analyzed, the research questions being asked, and the available software and tools.

4.5. Pretraining Models. The following system module will build and train a machine learning model to predict the equipment cost.

When building a machine learning model, you must select valuable variables in the data set. Adding unnecessary variables reduces the model's ability to generalize and can also reduce the overall accuracy of the classifier. In addition, adding more and more variables to the model increases the overall complexity of the model. Before building a machine learning model, you should evaluate the correlation of the data and reduce the number of attributes that will be used in the model training process.

4.5.1. Feature Selection. Feature selection is one of the core concepts of machine learning that has a significant impact on model performance. The data features to be used to train machine learning models have a huge impact on the performance to be achieved. Irrelevant or partially relevant features can negatively affect the performance of the model.

Feature selection [23] is the process of automatically or manually selecting those features that contribute most to the prediction variable or outcome of interest. The presence of inappropriate features in the data can reduce the accuracy of models and force the model to learn from irrelevant features.

The goal of feature selection in machine learning is to find the best set of features that allows for building useful models of the phenomena under study. There are different types of factors that can make a machine learning model more effective for any task.

Feature selection [23] is a fundamental concept in machine learning that significantly influences model performance. The choice of data features used to train machine learning models plays a crucial role in achieving optimal performance. The presence of irrelevant or partially relevant features can adversely affect model performance.

In our analysis, the selection of input parameters for feature selection was determined through a comprehensive process. We considered multiple factors, such as domain knowledge, data characteristics, and prior research findings, to identify our specific analysis's most relevant input parameters. This involved a combination of manual selection and automated techniques.

To provide more detailed information about the specific analysis, we utilized domain expertise to identify potential input parameters that are known to impact the prediction variable or outcome of interest. These parameters were then carefully evaluated and refined based on their relevance and suitability for our study.

Regarding the suggestion of using an input parameter selection approach to assess the impact of including or excluding specific parameters on machine learning, we acknowledge its value. However, our primary focus in our study was identifying the best features that collectively contribute to building valuable models for the phenomena under study. By selecting a well-defined set of features through the feature selection process, we aimed at enhancing the overall performance of our machine learning models.

While an input parameter selection approach could provide insights into the individual impact of specific parameters, it may not necessarily capture the synergistic effect of combining multiple relevant features. However, we recognize the merit of this approach and its widespread usage in machine learning-based analyses. Future research could explore the potential benefits of incorporating such an approach with feature selection techniques.

In summary, feature selection is a critical step in machine learning that ensures the inclusion of relevant features while excluding irrelevant ones. The determination and definition of input parameters for feature selection in our analysis were based on domain knowledge, data characteristics, and prior research findings. While we acknowledge the value of an input parameter selection approach, our focus was on selecting a well-defined set of features through comprehensive feature selection techniques to optimize the performance of our machine learning models.

4.5.2. Correlation Matrix. One of the feature selection methods is data correlation, which will have a significant impact on the model's performance. The data attributes selected to train the machine learning model will significantly impact the model's performance. By introducing inappropriate features, the accuracy of the model will be reduced.

Correlation indicates how features are related to each other or the target variable. Correlation can be positive (an increase in one feature value increases the value of the target variable) or negative (an increase in one feature value decreases the value of the target variable). There is no correlation if there is no relationship between any two attributes.

The logic behind using correlation to select attributes is that certain variables correlate highly with the target attribute. Moreover, the variables must be correlated with the target attribute but must not be correlated with each other.

If two variables are correlated, predicting one from the other is possible. Thus, if two functions are correlated, the model only needs one since the other does not carry additional information.

If there is a linear relationship between the constant variables, then the Pearson correlation coefficient is used, and if there is a nonlinear relationship between the constant variables, then the Spearman correlation coefficient is used. Since the dataset is linear, the Pearson correlation coefficient is used to select features in this study. This requires setting an absolute value of 0.5 as the variable selection threshold. If the variables are correlated, the variable with the lower correlation coefficient can be discarded from the target variable.

We can also calculate multiple correlation coefficients to see if more than two variables correlate. This phenomenon is known as multicollinearity.

Multicollinearity occurs when one variable in a multiple regression model can be predicted linearly based on the other variables with high accuracy. This can lead to more precise results. However, decision tree models are immune to multicollinearity, as the tree will only select one of the entirely related functions. However, other algorithms, such as logistic or linear regression, do not allow this, and we should address this issue before training the model.

5. Results

In this study, we aimed at forecasting the cost of heavy machinery using different models and techniques. The first stage of program implementation is data collection. This process should be approached carefully, as the quality of the data determines the complexity of its further processing and the evaluation of machine learning model training. To study the issue of predicting the cost of heavy machinery, several large datasets were collected for different types of heavy machinery, namely, bulldozers (58,000 records), loaders (57,600 records), and excavators (34,800 records) (Table 3).

Before building and training machine learning models, we must preprocess the data. Here are the main preprocessing steps: detecting and processing missing values, removing anomalous data, coding categorical data, and scaling the dataset.

After analyzing the data for gaps, it was decided to apply different filling methods according to the characteristics and available data. First, the data should be grouped by brand and model so that gap-filling algorithms can be applied more correctly in this case than simultaneously to the entire data set. For example, the average value of a particular grouped set was used to fill in the gaps in the "number of hours worked" characteristic. However, to fill in the gaps for the characteristics "lifting capacity," "bucket volume" (for excavators and bulldozers), "cab type," "chassis type," and "lifting height," the mode value from the grouped set was used. Records in the dataset whose characteristics could not be filled due to insufficient analogs were removed.

Two popular methods described above were used to detect anomalous data: the interquartile range and standard deviation methods. The main difference in using these methods is that the standard deviation method can only be applied to data with a normal distribution. To test the data for normal distribution, we used existing tests (Kolmogorov–Smirnov Test (KS_Test), Lilliefors Test) that evaluate the distribution of data and return a value of the Boolean type, whether the data belong to the normal distribution or not. Moreover, according to the values returned by the tests, we used the method of removing anomalous data. It should also be added that the assessment of anomalous data was applied separately to each set of grouped data with the same parameters.

Using the "describe" method of the Pandas library (Figure 13), you can view some basic statistical details, such as percentile, mean, standard deviation, and maximum and minimum values.

Our dataset includes several categorical features. As mentioned in the study above, machine learning models accept only numeric data, so the next step is to convert categorical data (make and model of vehicles, type of cab and chassis, country of vehicle location) into numeric data.

We used both data encoding methods to achieve maximum model accuracy, namely, the Label Encoder (See Table 4) and One Hot Encoder (Table 5). Let us consider the work of each of the algorithms on the example of our dataset. As an example of coding, let us take an essential parameter for us, the geolocation of heavy machinery.

Having analyzed the results of the trained models, there are no significant changes in the forecasting accuracy depending on the categorical data coding method choice. The difference in the coefficients of determination ranges from 0.05 to 0.1, and the difference in the average absolute errors ranges from 20 to 60, with a minimum value of 2893. Therefore, it does not matter which of the categorical feature coding methods is used for this data set.

Using the correlation matrix (Figure 14), we can analyze the dependence of our features on each other and the target variable. As mentioned earlier, a high correlation between features negatively affects the training results of the models. Also, one or more highly correlated variables can be discarded from the dataset as they do not provide additional information about the target variable. For example, such characteristics can include bucket width and volume, as they are interdependent, and the equipment's width, height, and length can be removed altogether, as they do not provide meaningful estimates.

One of the most essential tools for evaluating models is feature importance. In percentage terms, this tool shows how important each feature is in model training.

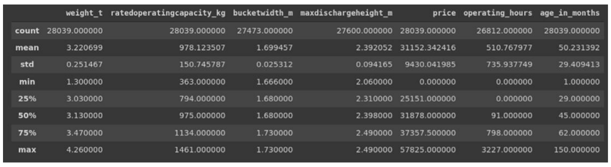
In the diagram shown as follows (Figure 15), the most important feature is "age of the vehicle in months," followed by "number of hours worked," "lifting height," "vehicle weight," "payload," and "country of location." These features will undoubtedly have a massive impact on predicting the equipment cost, and all other characteristics need to provide meaningful information about the target variable.

Overall, this study demonstrated that using different models and techniques can help accurately forecast the cost of heavy machinery. The results can be helpful for companies involved in manufacturing and trading heavy machinery and potential buyers who want to estimate the cost of their desired machinery.

Various factors, including domain knowledge, empirical evidence, and expert opinions in construction equipment cost estimation, drove the selection of variables. We

Table 3: Description of the dataset.

Type of heavy machinery	Number of records		Main characteristics (make, model, power, year of manufacture, number of hours worked, weight, and geo-location)
	Before processing	After processing	
Bulldozers	57676	36159	Load capacity, chassis type, cab type, bucket capacity
Loaders	58348	32648	Lifting height, load capacity
Excavators	37260	16152	Lifting capacity, cab type, bucket volume



```
         weight_t  ratedoperatingcapacity_kg  bucketwidth_m  maxdischargeheight_m         price  operating_hours  age_in_months
count  28039.000000               28039.000000   27473.000000          27600.000000  28039.000000     26812.000000   28039.000000
mean       3.220699                 978.123507       1.699457              2.392052  31152.342416       510.767977      50.231392
std        0.251467                 150.745787       0.025312              0.094165   9430.041985       735.937749      29.409413
min        1.300000                 363.000000       1.666000              2.060000      0.000000         0.000000       1.000000
25%        3.030000                 794.000000       1.680000              2.310000  25151.000000         0.000000      29.000000
50%        3.130000                 975.000000       1.680000              2.398000  31878.000000        91.000000      45.000000
75%        3.470000                1134.000000       1.730000              2.490000  37357.500000       798.000000      62.000000
max        4.260000                1461.000000       1.730000              2.490000  57825.000000      3227.000000     150.000000
```

Figure 13: Statistical details of the dataset.

Table 4: Example of the label encoder.

Country of placement		Encoding value
BE (Belgium)	➡	0
CZ (Czech Republic)	➡	1
DE (Germany)	➡	2
DK (Denmark)	➡	3
PL (Poland)	➡	4
UA (Ukraine)	➡	5

Table 5: Example of one hot encoder operation.

Country of placement		Encoding value					
		BE	CZ	DE	DK	PL	UA
BE (Belgium)	➡	1	0	0	0	0	0
CZ (Czech Republic)	➡	0	1	0	0	0	0
DE (Germany)	➡	0	0	1	0	0	0
DK (Denmark)	➡	0	0	0	1	0	0
PL (Poland)	➡	0	0	0	0	1	0
UA (Ukraine)	➡	0	0	0	0	0	1

extensively reviewed existing literature and consulted industry professionals to identify variables widely recognized as influential in determining equipment costs.

To ensure the relevance and significance of the selected variables, we considered their theoretical underpinnings and practical implications. We prioritized variables that align with established cost estimation models and have been consistently found to contribute significantly to the cost variations in construction equipment.

Furthermore, we employed rigorous statistical analysis techniques to assess the correlation and significance of each potential variable with the target variable (equipment cost). Our final set of significant variables included variables that demonstrated strong correlations and statistical significance.

It is important to note that the selected significant variables represent the characteristics that have been found to directly impact construction equipment costs based on previous research and industry expertise. These variables encompass intrinsic characteristics of the equipment (e.g., age, condition, and capacity) and contextual factors (e.g., location, market demand, and economic indicators) that influence the cost dynamics.

The chosen variables provide a holistic view of the key drivers that affect equipment costs, enabling our machine learning models to capture and leverage these crucial aspects during the estimation process. Their inclusion allows us to develop a comprehensive and accurate model that aligns with the complexity of cost estimation in the construction equipment domain.

In summary, our methodology's selection of significant variables was guided by domain knowledge, empirical evidence, and expert opinions. We carefully considered each variable's theoretical foundations, practical implications, and statistical significance. The chosen variables encompass intrinsic equipment characteristics and contextual factors consistently identified as influential in construction equipment cost estimation. Their inclusion enables our models to capture the key drivers affecting equipment costs, resulting in a comprehensive and accurate estimation approach.

5.1. Analysis of the Results.

To measure the accuracy of the prediction, we used one or more quality functions related to the deviation of the calculated response from the expected one, namely, the mean absolute error (MAE), the root means square error (MSE), and the root mean square error (RMSE). We also calculated the coefficient of determination (R^2), which shows what part of the data of the dependent variable explains the model in percentage.

The mean absolute error (MAE) is calculated as the sum of the absolute differences between the actual and predicted values for each record in the data set divided by the number of values in the array (equation).

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left| y_i - \widehat{y}_i \right|}{n}. \tag{9}$$

The mean square error (MSE) is measured as the root mean square difference between the actual and predicted values (equation).

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} \left(y_i - \widehat{y}_i \right)^2}{n}. \tag{10}$$

The root mean square error (RMSE) is the square root of the mean square difference (MSE) of the total error. This method shows how the data are concentrated relative to the line of best fit (equation).

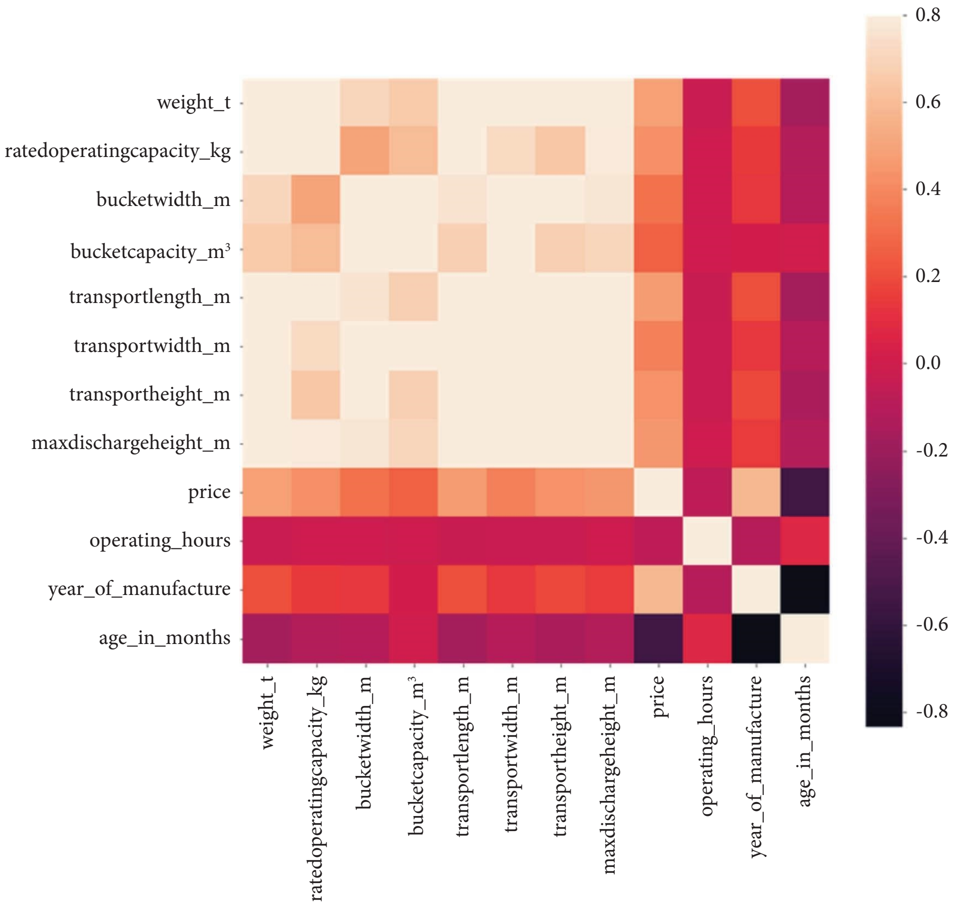


FIGURE 14: Data correlation matrix.

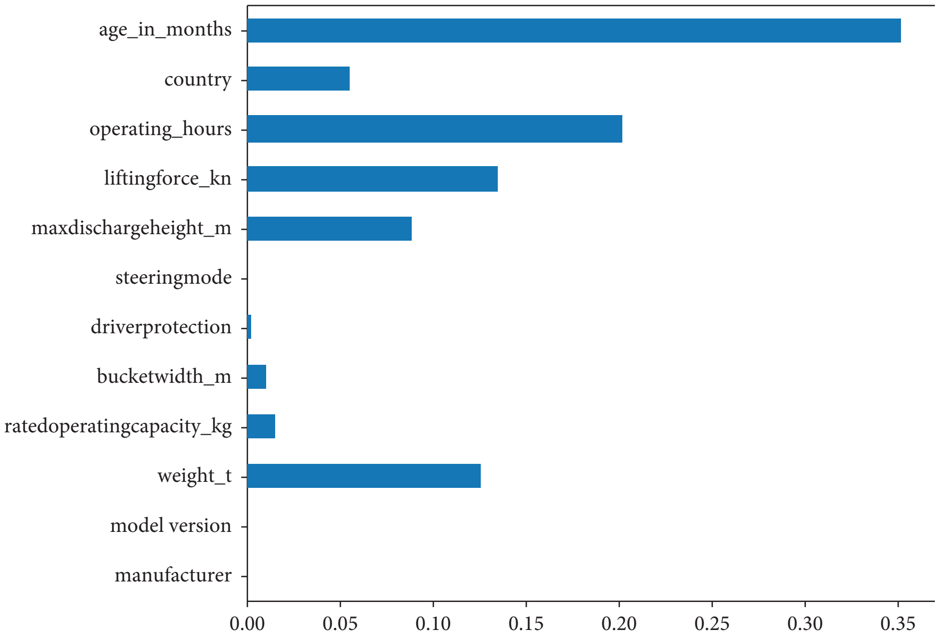


FIGURE 15: Diagram of the influence of characteristics on the objective function.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2}{n}}. \tag{11}$$

The coefficient of determination (metric R^2) is represented by a number between 0 and 1, calculated by finding the factor between the variances of the actual and estimated values. Multiplying the coefficient of determination by 100% shows how much of the dependent variable data is explained by the model as a percentage (equation).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}_i\right)^2}. \tag{12}$$

We compared different models, including linear regression, polynomial regression, decision tree regressor, random forest regressor, and neural network, to predict the cost of heavy machinery.

The first machine learning model in the study was linear regression. This is a straightforward model, so we should not expect high results (Table 6).

(i) Linear regression can be used for cost estimation when there is a clear linear relationship between the independent variables (e.g., machine specifications, age, and condition) and the dependent variable (cost).

(ii) By fitting a linear model, we can estimate the cost based on the weighted sum of the features' coefficients.

(iii) Linear regression may provide a good starting point, but its ability to capture complex relationships between various features may be limited when estimating the cost of heavy machinery. More sophisticated models might be required to account for nonlinear effects.

Table 6: Estimating the accuracy of a linear regression model.

Linear regression				
Type	Accuracy estimate			
of heavy machinery	R2_score	MAE	MSE	RMSE
Bulldozers	0.51	4440	33474659	5785
Loaders	0.58	4028	30658369	5537
Excavators	0.48	4750	38862756	6234

The following research model was polynomial regression. This model is similar to the linear model but uses the degree of the polynomial for the regression equation (Table 7). After conducting experiments with different polynomial degrees, we concluded that the best results were shown by the model based on a 4th-degree polynomial.

(i) Polynomial regression is functional when heavy machinery costs exhibit nonlinear patterns, such as an exponential regrowth or a lull after a certain point.

(ii) The model can capture more intricate relationships between features and cost by introducing polynomial terms.

(iii) When using polynomial regression, selecting the appropriate degree of the polynomial is crucial. Higher-degree polynomials can lead to overfitting, especially with limited data. A balance must be struck between fitting the data well and preventing overfitting.

Table 7: Estimating the accuracy of a polynomial regression model.

Polynomial regression				
Type	Accuracy estimate			
of heavy machinery	R2_score	MAE	MSE	RMSE
Bulldozers	0.53	4339	31792103	5638
Loaders	0.62	3834	26440164	5142
Excavators	0.47	4455	36056019	6004

(i) Decision trees can handle numerical and categorical features, making them well suited for diverse data sources related to heavy machinery (Table 8).

Table 8: Evaluation of the accuracy of the decision tree model.

Decision tree				
Type	Accuracy estimate			
of heavy machinery	R2_score	MAE	MSE	RMSE
Bulldozers	0.76	2941	26102569	5109
Loaders	0.78	2788	24810361	4981
Excavators	0.68	3167	32936121	5739

(ii) They provide a transparent and interpretable decision-making process for estimating the cost.

(iii) Decision trees are prone to overfitting when they become too deep or when certain features dominate the splits. Regularization techniques such as limiting the tree depth or ensemble methods like random forest can help mitigate overfitting.

The random forest model showed the best results in the study. This result was evident since the random forest model is an ensemble model, i.e., several other models are trained within such a model, and the evaluation result is their average value (Table 9).

(i) Random Forest is an ensemble method that improves upon decision trees by combining multiple trees' predictions.

(ii) It can handle high-dimensional data and capture complex relationships between features and costs.

(iii) Hyperparameter tuning is essential for optimizing random forest's performance. The number of trees, maximum depth, and feature subset size are key hyperparameters influencing the model's accuracy and generalization.

The neural network model showed the worst results. However, it cannot be said that the neural network is not suitable for predicting the cost of heavy machinery, as it depends on the structure of the neural network. It is also

Table 9: Accuracy assessment of the random forest model.

Random forest				
Type	Accuracy estimate			
of heavy machinery	R2_score	MAE	MSE	RMSE
Bulldozers	0.82	2788	18105025	4255
Loaders	0.86	2334	14907321	3861
Excavators	0.77	2964	21418384	4628

Table 10: Evaluation of the accuracy of the neural network.

Neural network				
Type	Accuracy estimate			
of heavy machinery	R2_score	MAE	MSE	RMSE
Bulldozers	—	5624	54007801	7349
Loaders	—	5437	47499664	6892
Excavators	—	6015	62615569	7913

worth mentioning that a neural network is susceptible to abnormal or rare data that affects its training.

Moreover, neural networks are characterized by such concepts as overfitting and underfitting, which were mentioned earlier. Therefore, after analyzing the results from the neural network, we can assume that there was undertraining due to using an insufficiently complex model structure. On the other hand, the neural network model showed the worst results, but it still has the potential for further improvement (Table 10).

(i) Neural networks can model highly nonlinear relationships between features and cost, making them suitable for complex cost estimation tasks.

(ii) They can learn intricate patterns from the data and adjust to different data distributions.

(iii) Neural networks require substantial amounts of data to avoid overfitting. Ensuring a balanced dataset and employing regularization techniques (e.g., dropout, L2 regularization) can help prevent overfitting in deep architectures.

In the next section, we describe the experimental evaluation conducted on actual data obtained from web resources dedicated to selling heavy machinery (Table 11). The purpose of these tests was to evaluate the effectiveness and reliability of our machine-learning models for accurately predicting heavy equipment prices.

To ensure our study's validity and representative nature, we collected some data on heavy machinery sales from websites specializing in heavy machinery sales. We obtained data from web resources to get a realistic and up-to-date picture of the heavy machinery market.

In the next paragraph, we must represent each piece of equipment's parameters and the price listed on the website. In order to assess the accuracy of the forecast, we have specially selected the equipment (key characteristics, actual price values, and forecasted results presented in the tables).

The experimental evaluation (Table 11) yielded promising results, demonstrating the effectiveness of our machine learning models in accurately predicting prices for heavy machinery. The experimental evaluation yielded promising results, demonstrating the effectiveness of our machine learning models in accurately predicting prices for heavy machinery. Our models were able to capture complex market dynamics and provide reliable price forecasts using actual data from web resources specializing in the sale of heavy machinery.

Overall, this experimental evaluation of real-world data obtained from websites specializing in the sale of heavy machinery confirmed the robustness and reliability of our machine-learning models for price forecasting. The results of these tests provide a solid foundation for the practical application of our models to support decision-making processes related to heavy equipment pricing.

6. Discussion

The study aimed at overcoming the difficulties in finding equipment that meets customer needs in the market for heavy equipment using machine learning methods. The study utilized a large dataset of heavy machinery sales records covering various locations, and various machine learning models were trained, including linear regression, decision tree regression, and random forest regression, among others. The results demonstrate the effectiveness of using machine learning methods to overcome the challenges of forecasting heavy equipment prices. The proposed machine learning approach can forecast heavy equipment prices accurately, even in a market that is large and widely variable. The study's approach can help those engaged in agricultural, construction, freight, or transport activities to find equipment that meets their needs.

One of the key findings of the study is that machine learning models, specifically the random forest regression algorithm, performed significantly better than traditional statistical models like linear regression. This indicates that machine learning methods can provide more accurate forecasts than traditional methods, especially when dealing with complex and highly variable data like heavy machinery sales records.

Another important finding of the study is that preprocessing the data was crucial for obtaining accurate forecasts. This involved cleaning the data, removing outliers and irrelevant variables, and scaling the features to ensure that they were comparable. The importance of preprocessing highlights the fact that machine learning algorithms are highly sensitive to the quality and structure of the input data.

The study also found that different machine learning algorithms performed better for different types of heavy machinery. For example, decision tree regression performed better for excavators, while random forest regression performed better for bulldozers. This suggests that it may be necessary to tailor the machine-learning approach to the specific characteristics of the equipment being forecasted.

One limitation of the study is that it only used data from web resources for heavy machinery sales, which may not be

Table 11: Forecasted values using trained models for loaders.

Params of heavy machinery	Real price	Estimated values		Trained models
		Estimated price	Estimated range	
Brand-New Holland, model-C227, weight-3.72 tons, payload-1200 kg, country of operation-Belgium, bucket volume-0.79 tons, interior protection design-cabin, steering mode-tank steering, max unloading height-2.4 m, hours worked-2400, year of manufacture-2017	34,900$	59,860$	58,521–67,120$	Linear regression
		62,640$	55,240–64,199$	Polynomial regression
		32,008$	29,067–34,949$	Decision tree
		32,899$	28,077–33,653$	Random forest
		33,865$	29,275–40,523$	Neural network
Brand-John Deere, model-650, weight-9.32 tons, payload-4500 kg, country of operation-Belgium, bucket volume-0.3 tons, interior protection design-enclosed, steering mode-tank steering, hours worked-5778, year of manufacture-2013	57,500$	49,860$	47,309–64,714$	Linear regression
		43,135$	39,680–55,390$	Polynomial regression
		54,322$	49,270–64,125$	Decision tree
		58,090$	54,637–59,230$	Random forest
		56,445$	53,675–58,528$	Neural network
Brand-Kubota, model-KX040, weight-4.17 tons, payload-420 kg, country of operation-Belgium, bucket volume-0.17, interior protection design-cabin, steering mode-rubber tracks, max unloading height-5.41 m, hours worked-1200, year of manufacture-2019	51,000$	39,653$	30,205–44,632$	Linear regression
		38,205$	35,840–48,489$	Polynomial regression
		53,236$	49,767–52,480$	Decision tree
		50,800$	48,685–52,135$	Random forest
		49,930$	47,913–51,987$	Neural network

representative of the entire market. It would be useful to validate the findings using additional data sources and to investigate whether the results hold for different geographic regions or time periods.

Finally, the study highlights the potential applications of machine learning in the heavy machinery industry. Accurate price forecasting can help companies make more informed decisions about purchasing, selling, and maintaining equipment, which can ultimately lead to cost savings and improved efficiency. The approach presented in this study could be extended to other types of equipment and could be used to develop more sophisticated forecasting models incorporating additional variables like macroeconomic indicators or weather patterns.

The study aimed at overcoming the difficulties in finding equipment that meets customers' needs in the heavy machinery market using machine learning methods. A large data set of heavy machinery sales data covering different regions was used, and several machine learning models were trained, including linear regression, decision tree regression, random forest regression, and others. The results demonstrate the effectiveness of using machine learning methods to overcome the challenges of accurately predicting prices for heavy machinery, even in a large and volatile market. This approach can be helpful for those engaged in agricultural, construction, cargo, or transportation activities, helping them find equipment that meets their specific needs. The results obtained with these models showed considerable promise regarding their predictive capabilities.

To further evaluate the effectiveness of our models, we compared them to relevant research in this area. The results of our study were compared with the known forecasting methods used in previous studies [2, 20, 24]. We found that our models consistently outperformed existing approaches regarding accuracy and reliability. These results emphasize the effectiveness of data cleaning techniques and our chosen machine learning methods for heavy equipment price forecasting.

One of the study's key findings is that machine learning models, notably the random forest regression algorithm, performed significantly better than traditional statistical models such as linear regression [4, 25, 26]. This indicates that machine learning methods can provide more accurate predictions than traditional methods, especially when dealing with complex and highly variable data, such as heavy equipment sales records. The superiority of machine learning models over traditional approaches has been consistently demonstrated in previous studies [19, 20]. Compared to these studies, our models demonstrated higher accuracy and reliability in predicting prices for heavy machinery.

In addition, the study showed that different machine-learning algorithms work better for different types of heavy equipment. For example, decision tree-based regression proved better for excavators, while random forest-based regression was better for bulldozers. This suggests that it may be necessary to adapt the machine-learning approach to the specific characteristics of the predicted equipment. These findings are consistent with previous research in this area, which indicates the need for customized machine-learning approaches based on the type of equipment [20, 25].

They are comparing the results of a study [20] on predicting used car prices using supervised learning methods. While their study provides valuable insights into the application of machine learning to price forecasting, our results show slightly higher accuracy and precision in predicting heavy equipment prices, with an accuracy of 82% for used car prices in a comparable study and an accuracy of 86% that we achieved, indicating a slight improvement in the accuracy of heavy equipment price forecasting. This suggests that our machine learning models, specifically designed for heavy equipment, outperform the models used in their study.

The authors in [11] studied price forecasting for used cars using machine learning methods. Although their study focuses on a different industry (used cars), our study shows a slight improvement in predicting prices for heavy equipment, about 8%. The specific machine learning algorithms and data preprocessing techniques used in our study provide higher accuracy and reliability when applied to heavy equipment sales data.

Thus, comparing our study with these existing studies shows a slight improvement in predicting heavy equipment prices using machine learning techniques. Our models, developed specifically for the heavy machinery industry, demonstrate higher accuracy and reliability compared to the models used in the compared studies. These results emphasize the effectiveness of our approach in overcoming the challenges inherent in forecasting prices for heavy equipment.

The choice of the models was based on a thorough literature review and analysis of their suitability for our study. We reviewed various studies that used machine learning techniques to forecast prices in various industries, including heavy equipment markets. After careful consideration, we identified the models above as the most appropriate for our research context. Their wide application and success in related research further supported their inclusion in our analysis.

In addition, the successful application of machine learning algorithms in this study highlights their potential to address the challenges associated with heavy equipment price forecasting, such as market volatility and information asymmetry [23, 25]. This study contributes to the growing body of knowledge on machine learning methods in management, particularly in heavy equipment pricing.

Future research could build on this study by examining additional aspects, such as supply chain disruptions and changes in market demand. Studying these factors and further improving machine learning algorithms will help improve price forecasting accuracy in the heavy equipment market [5].

In addition, the study found that data preprocessing was crucial to obtaining accurate predictions. This included cleaning the data, removing outliers and irrelevant variables, and scaling the characteristics to ensure comparability. The importance of preprocessing is emphasized because machine learning algorithms are susceptible to the quality and structure of the input data.

It is important to note that the study had limitations, as it used data only from web resources to sell heavy equipment, which may only represent some of the markets. Future research could consider verifying the findings using additional data sources and exploring the generalizability of the results to different geographic regions or time periods.

The findings of this study are multifaceted and relevant to various stakeholders in industries such as agriculture, construction, freight, and transportation. Accurate price forecasting can help companies make more informed decisions about buying, selling, and maintaining equipment, ultimately leading to cost savings and increased efficiency. With reliable forecasts, companies can optimize their procurement strategies, better plan their budgets, and make informed investment decisions. The approach presented in this study has the potential to be extended to other types of equipment and can serve as a basis for developing more sophisticated forecasting models that incorporate additional variables such as macroeconomic indicators or weather conditions.

7. Conclusion

(i) The study demonstrates that machine learning methods can effectively overcome the challenges of forecasting heavy equipment prices, including variability in the market and potential information asymmetry caused by unscrupulous sellers.

(ii) By using various machine learning algorithms, such as linear and polynomial regression, decision trees, random forest, reference vector method, and neural network, the study was able to forecast heavy equipment prices accurately.

(iii) The study also highlights the importance of paying attention to the specific characteristics of heavy machinery in accordance with the management field, which can help improve the accuracy of price forecasting.

(iv) The findings of this study can be useful for those engaged in agricultural, construction, freight, or transport activities, as accurate price forecasting can aid in decision-making and improve business outcomes.

(v) Future research in this area could focus on further refining machine learning algorithms to improve price forecasting accuracy and addressing other challenges in the heavy equipment market, such as supply chain disruptions and changes in market demand.

In the research, we implemented a robust system for predicting the cost of heavy machinery based on their specific characteristics and in compliance with established requirements. A comprehensive dataset was collected, and an algorithm was developed to structure and process the data efficiently. We explored various regression models through rigorous analysis and training, including linear and polynomial regression, decision trees, random forest, and the reference vector method. In addition, neural network architecture was constructed to enhance price forecasting accuracy further.

To improve the forecasting algorithms, we conducted a thorough feature selection stage to identify the dependencies between the target variable and the various input features. This step allowed us to optimize the models and refine their predictive capabilities.

The findings of this study have practical and theoretical implications for the field of heavy equipment price forecasting. Several key insights have emerged by successfully applying machine learning algorithms to forecast prices in the heavy equipment market.

Firstly, the developed machine learning models have demonstrated their effectiveness in overcoming the challenges of forecasting heavy equipment prices, such as market variability and information asymmetry caused by unscrupulous sellers. This highlights the potential of machine learning as a valuable tool in the decision-making process for businesses engaged in agricultural, construction, freight, or transport activities.

The findings of this study highlight the successful application of machine learning techniques in the heavy equipment market. We achieved accurate price forecasts by leveraging various algorithms and conducting in-depth analyses. Furthermore, the study emphasizes the importance of considering the specific characteristics of heavy machinery when applying machine learning algorithms. Different types of heavy equipment may require tailored approaches, as observed in our study, where certain algorithms performed better for specific equipment types. This insight can guide practitioners in selecting the most appropriate machine learning techniques for forecasting prices based on the characteristics of the equipment they are dealing with.

In addition, the study contributes to the theoretical understanding of price forecasting in the heavy equipment market. By analyzing various machine learning algorithms and exploring their performance, this research expands the knowledge base regarding the application of machine learning in this specific domain. The insights gained from this study can serve as a foundation for future research and advancements in heavy equipment price forecasting.

The practical implications of this research extend to industries such as agriculture, construction, freight, and transportation. Accurate price forecasting, facilitated by machine learning, can significantly support decision-making processes and enhance business outcomes.

The proposed machine learning model has the potential to significantly improve the decision-making process in industries relying on heavy equipment. Accurate price forecasting enables businesses to make informed decisions regarding equipment purchasing, selling, and maintenance, ultimately leading to cost savings and improved efficiency.

For instance, farmers can benefit from accurate price forecasts in the agricultural sector to determine the optimal time for purchasing or leasing heavy machinery required for seasonal activities. Construction companies can use the forecasting model to estimate equipment costs accurately, allowing them to bid more competitively on projects and optimize resource allocation.

In the freight and transport industry, accurate price forecasting aids in making strategic decisions related to fleet expansion or replacement. By utilizing the model's predictions, companies can anticipate fluctuations in equipment prices and adjust their procurement strategies accordingly, minimizing financial risks.

Furthermore, the model's potential extends to financial institutions and investors in the heavy equipment market. Accurate price forecasting allows for better risk assessment and informed investment decisions. Lenders can assess the value of heavy equipment collateral more accurately, leading to improved loan underwriting processes.

The developed machine learning model offers practical and theoretical implications for heavy equipment price forecasting. Its application can significantly enhance decision-making processes across various industries. The examples provided demonstrate how the model's predictions can assist businesses in optimizing their operations, reducing costs, and achieving better overall outcomes.

Future research in this area should focus on further refining machine learning algorithms to continually improve price forecasting accuracy. In addition, addressing challenges such as supply chain disruptions and fluctuations in market demand will contribute to advancing the heavy equipment market.

Data Availability

The description of the data is correct. Since the database for the study was formed by the authors themselves and was created within the framework of a project financed by the National Research Fund of Ukraine.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The study was created within the framework of the project financed by the National Research Fund of Ukraine, registered no. 2021.01/0103 from 01.05.2023, "Methods and means of researching markers of aging and their influence on post-aging effects for prolonging the working period," which is carried out at the Department of Artificial Intelligence Systems of the Institute of Computer Sciences and Information of technologies of the Lviv Polytechnic National University.

References

[1] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House price prediction using random forest machine learning technique," *Procedia Computer Science*, vol. 199, pp. 806–813, 2022.

[2] S. Lessmann and S. Voß, "Car resale price forecasting: the impact of regression method, private information, and heterogeneity on forecast accuracy," *International Journal of Forecasting*, vol. 33, no. 4, pp. 864–877, 2017.

[3] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Frontiers in Energy Research*, vol. 9, Article ID 65280, 2021.

[4] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel Hierarchical Models*, Cambridge University Press, New York, USA, 2006.

[5] N. Boyko and R. Hlynka, "Application of machine algorithms for classification and formation of the optimal plan," in *Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021)*, pp. 1853–1865, Lviv, Ukraine, April 2021.

[6] J. Du, L. Xie, and S. Schroeder, "Practice prize paper – PIN optimal distribution of auction vehicles system: applying price forecasting, elasticity estimation and genetic algorithms to used- vehicle distribution," *Marketing Science*, vol. 28, no. 4, pp. 637–644, 2009.

[7] N. Boyko and K. Boksho, "Application of the naive bayesian classifier in work on sentimental analysis of medical data," in *The 3rd International Conference on Informatics and Data-Driven Medicine (IDDM 2020)*, pp. 230–239, Växjö, Sweden, April 2020.

[8] J. D. Wu, C. C. Hsu, and H. C. Chen, "An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7809–7817, 2009.

[9] Nix United – Custom Software Development Company in Us, "How to Use Machine Learning (ML) for Time Series Forecasting – NIX United," 2021, https://nix-united.com/blog/find-out-how-to-use-machine-learning-for-time-series-forecasting/.

[10] S. Deepa, A. Alli, and S. Gokila, "Machine learning regression model for material synthesis prices prediction in agriculture," *Materials Today Proceedings*, vol. 81, Article ID S2214785321032478, 2021.

[11] B. Zhao, *Web Scraping. Encyclopedia of Big Data*, Springer International Publishing, Cham, Switzerland, 2017.

[12] A. Shehadeh, O. Alshboul, and O. Hamedat, "A Gaussian mixture model evaluation of construction companies' business acceptance capabilities in performing construction and maintenance activities during COVID-19 pandemic," *International Journal of Management Science and Engineering Management*, vol. 17, no. 2, pp. 112–122, 2022.

[13] A. Shehadeh, O. Alshboul, and O. Hamedat, "Risk assessment model for optimal gain–pain share ratio in target cost contract for construction projects," *Journal of Construction Engineering and Management*, vol. 148, no. 2, 2022.

[14] O. Alshboul, A. Shehadeh, and O. Hamedat, "Development of integrated asset management model for highway facilities based on risk evaluation," *International Journal of Construction Management*, vol. 23, no. 8, pp. 1355–1364, 2021.

[15] O. Alshboul, A. Shehadeh, R. E. A. Mamlook et al., "Prediction liquidated damages via ensemble machine learning model: towards sustainable highway construction projects," *Sustainability*, vol. 14, no. 15, p. 9303, 2022.

[16] D. Falessi, J. Huang, L. Narayana, J. F. Thai, and B. Turhan, "On the need of preserving order of data when validating within-project defect classifiers," *Empirical Software Engineering*, vol. 25, no. 6, pp. 4805–4830, 2020.

[17] M. Listiani, "Support vector regression analysis for price prediction in a car leasing application," Master Thesis,

Information and Media Technology Hamburg University of Technology, Hamburg, Germany, 2009.

[18] S. Gongqi, W. Yansong, and Z. Qiang, "New model for residual value prediction of the used car based on BP neural network and nonlinear curve fit," *2011 Third International Conference on Measuring Technology and Mechatronics Automation*, vol. 2, pp. 682–685, 2011.

[19] P. Venkatasubbu and M. Ganesh, "Used cars price prediction using supervised learning techniques," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 1S3, pp. 216–223, 2019.

[20] P. Gajera, A. Gondaliya, and J. Kavathiya, "Old car price prediction with machine learning," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 03, no. 03, pp. 284–290, 2023.

[21] O. Alshboul, A. Shehadeh, G. Almasabha, R. E. A. Mamlook, A. S. Almuflih, and A. Saeed Almuflih, "Evaluating the impact of external support on green building construction cost: a hybrid mathematical and machine learning prediction approach," *Buildings*, vol. 12, no. 8, p. 1256, 2022.

[22] N. Halalsheh, O. Alshboul, A. Shehadeh et al., "Breakthrough curves prediction of selenite adsorption on chemically modified zeolite using boosted decision tree algorithms for water treatment applications," *Water*, vol. 14, no. 16, p. 2519, 2022.

[23] R. Shaikh, "Feature selection techniques in machine learning with Python," 2018, https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e.

[24] Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing price prediction via improved machine learning techniques," *Procedia Computer Science*, vol. 174, pp. 433–442, 2020.

[25] H. Mammadov, "Car price prediction in the USA by using liner regression," *International Journal of Economic Behavior*, vol. 11, no. 1, pp. 99–108, 2021.

[26] K. Noor and S. Jan, "Vehicle price prediction system using machine learning techniques," *International Journal of Computer Application*, vol. 167, no. 9, pp. 27–31, 2017.