INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 274-284

DRIVING TRANSPARENCY: A MACHINE LEARNING APPROACH TO OPTIMIZE USED VEHICLE PRICING WITHVEHICLE WORTHAI

Aryan Singh[1], Vaibhav Singh[2], Vaibhav Singh[3]

[1,2,3]Students: Jagan Institute Of Management Studies Sector-5 Rohini., India.

ABSTRACT

The proposed project aims to develop a web-based platform that facilitates the evaluation of the best possible price for used or second-hand vehicles. The platform will leverage user-provided information such as fuel type, mileage, engine size, vehicle age, location, repairing history, tier status, build quality etc. to generate accurate price estimates. We will use image processing to make our model more accurate. Our project helps the users to get the best price for their used old vehicles in a matter of seconds. This synopsis outlines the key components and steps involved in creating this innovative vehicle valuation system.

Key words- Machine Learning, Artificial Intelligence, Re- gressors

1. INTRODUCTION

According to the report, the pre-owned vehicle market has experienced remarkable growth and transformation over the past few years. In fiscal year 2027, the pre owned to new cars market in India was projected to be approximately 1.9 times larger. Fiscal year 2022 saw the sale of 4.4 million pre-owned cars. The number of pre-owned vehicles increased consistently during the time period. Several factors have contributed to this shift, making the used car market an attractive option for buyers and investors alike. These factors include, Economic factors, rapid depreciation of new vehicle and improved vehi- cle durability. The major concern while buying a second hand or used vehicle is value for money. We aim at solving this problem, by creating an application called Vehicle WorthAI. Here a buyer can enter the details of the vehicle and willknow the estimated value of the vehicle. This way he/shewill be prevented from buying a vehicle at higher price. The major concern while selling a second hand or used vehicle is difference in the expected value and market value. We aim at solving this problem too. Our application will provide the accurate price value for your vehicle. Now he/she can decide whether to sell or not. Since the pre-owned vehicle market is growing rapidly, the demand for reliable used vehicle pricing information is also going to increase. Everyone aims at getting the price he/she deserves, so we will introduce the idea of a web platform that employs advanced algorithms to calculate estimated prices based on various vehicle attributes. This will be a user-friendly application which will be easy to use.

2. PREVIOUS WORK

- Enci Liu's car price prediction model prioritizes accuracy over time efficiency when choosing the hidden layersfor the BP neural network. This approach may lead toa highly accurate model but could require more compu- tational resources and longer training times. Balancing accuracy and training time is a crucial factor in the model's design. [2]
- Prashant Gajera utilized various models for predicting used car prices, but the limited dataset restricted the model's predictive power. Gathering a larger dataset could enhance the model's accuracy and reliability. Ad- ditionally, exploring and incorporating additional relevantfeatures could further improve the predictive capabilities by capturing more nuances and factors influencing used car prices. [4]
- Enis Gegic achieved a 92% accuracy in his model while maintaining lower resource consumption compared to a single, resource-intensive model. He achieved this by em-ploying multiple models, distributing the resource load, and optimizing the overall efficiency of the system. This approach allowed for a balance between high accuracy and resource efficiency. [5]

3. RESEARCH METHODOLOGY

A. Machine Learning

Machine learning, a transformative subset of artificial intel- ligence, has emerged as a cornerstone technology reshaping how computers perceive, analyze, and respond to data. Inessence, machine learning is a dynamic field that allows com- puter systems to learn and improve from experience without being programmed by themselves. A change in paradigm has occurred with the shift away from traditional rule-based programming, where machine learning models can identify patterns and relationships within data to make decisions or predict future outcomes. The main idea centers on the use of iterative learning algorithms that modify their internal parameters to

improve performance and better handle newinformation, both visible and hidden.[20]

The field of machine learning spans a wide range, incorpo- rating diverse methodologies. This includes supervised learn- ing, where models are trained on labeled datasets for predictive or classification tasks; unsupervised learning, which centers on extracting patterns from unlabeled data; and reinforce- ment learning, where agents acquire optimal decision-making strategies through interactions with their environment.[21][22] Techniques such as deep learning, a subfield involving neural networks with multiple layers, have gained prominence for their capacity to handle complex and high-dimensional data, showcasing the versatility of machine learning in tasks rang- ing from image and speech recognition to natural language processing and autonomous systems. The rapid evolution of machine learning techniques and their integration into diverse domains underscore their potential to revolutionize industries, facilitate data-driven decision-making, and pave the way for advancements in artificial intelligence.[23]

The applications of machine learning span across an array of industries, transforming the landscape of how organizations harness and interpret data. In healthcare, machine learning aids in diagnostics, predicting disease outcomes, and personalized treatment plans by analyzing vast datasets of medical records, genetic information, and imaging data. Financial entities em- ploy machine learning for tasks such as detecting fraud, as- sessing risks, and engaging in algorithmic trading. In these ap- plications, models analyze patterns within financial markets to guide strategic decision-making. In transportation, autonomous vehicles rely on machine learning algorithms to navigate complex environments, interpret sensor data, and make split- second decisions to ensure passenger safety. Furthermore, machine learning is pivotal in natural language processing applications, enabling virtual assistants, language translation services, and sentiment analysis in social media.[24]

The implications of machine learning extend beyond tech- nological advancements, raising ethical considerations and societal impacts. The responsible and fair use of machine learning algorithms, avoidance of biases in training data, and ensuring transparency in decision-making processes are crucial aspects to address. As machine learning continues to evolve, the collaboration between technologists, policymakers, and ethicists becomes imperative to shape a future where the benefits of this technology are harnessed for the greater good while mitigating potential risks. The transformative potential of machine learning is undeniable, with its influence perme- ating various facets of our lives, from personalized recom- mendations in online platforms to innovative breakthroughs in scientific research and beyond.[25]

B. Supervised machine learning

Supervised machine learning stands as a cornerstone in the realm of artificial intelligence, distinguished by its capacity to autonomously discern patterns and relationships from labeled training datasets. In this paradigm, the algorithm is bestowed with a wealth of examples, each paired meticulously with its corresponding outcome or label. The fundamental objective ofsupervised learning is to empower the algorithm to extrapolate and comprehend intricate patterns embedded in the training data, enabling it to accurately predict labels for entirely new and unseen instances.

The training phase unfolds as an iterative process wherein the algorithm systematically scrutinizes the features of the labeled dataset, deciphering latent connections between these features and the associated labels. This entails adjusting the internal parameters of the algorithm in a step-by-step manner to progressively reduce the discrepancy between its predictions and the actual labels within the training set. Essentially, the algorithm undergoes a learning process, absorbing the under- lying relationships inherent in the labeled data and becoming adept at mapping input features to the correct output.

Post-training, the supervised learning model emerges as a robust predictor, capable of offering predictions on novel, unseen data by applying the acquired patterns. The model's prowess is gauged through evaluation against a separate test dataset, assessing its ability to generalize and accurately predict outcomes beyond the confines of the training data. Undoubtedly, supervised learning has wide applications in areas like image and speech recognition, natural language processing (like statistics), and predictive analytics. The ar- senal of algorithms at its disposal, including linear regression for regression tasks and a myriad of classification algorithms, renders supervised learning an indispensable tool for harness- ing data to inform predictions and guide decision-making in multifaceted real-world scenarios.

C. Unsupervised machine learning

Unsupervised machine learning stands as a transformative paradigm in artificial intelligence, distinguished by its ability to extract meaningful patterns and structures from datasets devoid of explicit labels or outcomes. In contrast to super- vised learning, where algorithms are trained on labeled data, unsupervised learning operates on unlabeled

datasets, aiming to uncover inherent relationships, groupings, or representationswithin the data without predefined guidance.[30]

Clustering and dimensionality reduction are common tech- niques employed in unsupervised learning. Clustering algo- rithms segregate data points into meaningful groups based on inherent similarities, revealing natural patterns that might not be apparent a priori. Dimensionality reduction methods aim to distill essential features from the dataset, simplify-ing its complexity while retaining crucial information. These unsupervised techniques are invaluable in revealing hidden structures and providing insights into the intrinsic organizationof data.

One prevalent application of unsupervised learning is in clustering, where data points are grouped together based on similarities, allowing for the identification of natural sub- divisions within the dataset. Another key application is in anomaly detection, where deviations from normal patterns are identified, aiding in the identification of irregularities or outliers in complex datasets.[29]

Unsupervised learning's versatility extends to fields such as natural language processing, image and speech recognition,

and recommendation systems. In natural language processing, for instance, unsupervised algorithms can uncover latent se- mantic relationships between words or documents without ex- plicit guidance. In image processing, unsupervised techniques contribute to feature extraction and pattern recognition, en- abling machines to discern complex visual structures without labeled examples.

The power of unsupervised learning lies in its ability to autonomously discover and highlight intrinsic structures within data, offering a valuable complement to supervised methods. By revealing patterns and relationships that may elude manual scrutiny, unsupervised learning plays a pivotal role in uncovering insights, driving innovation, and advancingour understanding of complex datasets in diverse domains.

D. Semi-supervised machine learning

Semi-supervised learning represents a nuanced and prag- matic approach within the broader landscape of machine learning, bridging the realms of labeled and unlabeled data to enhance model performance. In this paradigm, the algorithm is confronted with a dataset that comprises both labeled exam-ples, where the input features are paired with corresponding outcomes, and unlabeled instances lacking explicit annota- tions. The distinctive feature of semi-supervised learning lies in its adeptness at leveraging the limited labeled data availablein conjunction with the larger pool of unlabeled data to construct more robust and generalizable models.[26]

During the training phase, semi-supervised learning algo- rithms endeavor to extrapolate the underlying structure from the labeled data, using it as a scaffold to guide the interpre- tation of the unlabeled instances. This process enables the algorithm to discern latent patterns and relationships within the unlabeled data, effectively augmenting its understanding beyond the confines of explicit supervision. The incorporation of unlabeled data introduces a level of flexibility, allowing the algorithm to generalize more effectively and adapt to the inherent complexity of real-world datasets.[27]

Semi-supervised learning finds application in scenarios where obtaining labeled data is labor-intensive or expensive, as is often the case in fields such as healthcare or natural language processing. By judiciously incorporating both labeled and unlabeled data, semi-supervised learning strives to strike a balance between the benefits of explicit guidance and the broader insights gleaned from the unlabeled data, offering a pragmatic compromise in situations where fully labeled datasets are challenging to obtain.[28] The effectiveness of semi-supervised learning underscores its potential to unlock deeper understanding and more accurate predictions across diverse domains, contributing to the ongoing evolution of machine learning methodologies.

E. Reinforcement learning

Reinforcement learning represents the essence of machine learning, which is the ability to help people learn effec- tive decision-making strategies by interacting with the en- vironment. Essentially, reinforcement learning involves the agent passively navigating the environment, performing tasks, receiving feedback in the form of rewards or punishment, and adjusting its behavior over time for maximum benefit. This iterative process of trial and error allows employees tofind effective strategies to achieve specific goals and enablesgreater learning inherent in open or poorly recorded data. [31]Further study is based on the concept of agents interacting with the environment.[32] The agent makes decisions by choosing actions, and the environment responds with feedback, often in the form of digital gifts. The agent's goal is to learn thepolicy (a map from situation to action) that leads to the best possible outcome over time. The inherently dynamic nature of motivational learning allows it to solve complex problems

where the correct policy may not be known. Key components of reinforcement learning include exploration-exploitation trade-offs, where the agent must balance between trying new actions to discover better strategies (exploration) and choosing actions it believes to be optimal based on current knowledge (exploitation). Reinforcement learning algorithms utilize various approaches,including value functions, policy gradients, and deep reinforcement learning, which integrates neural networks to handle high-dimensional and continuous state spaces.

Applications of reinforcement learning span diverse do-mains, from robotics and game playing to finance and au-tonomous systems. In robotics, reinforcement learning enablesrobots to adapt to changing environments and learn complex tasks. In gaming, reinforcement learning has demonstratedsuperhuman performance in games like Go and Dota 2. In finance, it aids in portfolio optimization and algorithmic trad- ing. The versatility and adaptability of reinforcement learning underscore its potential to revolutionize decision-making in scenarios where adaptive strategies and autonomous learn-ing are paramount.[33]

F. Regressors

On the contrary, regressors are utilized when dealing witha continuous target variable, with the objective of predicting a numerical value. Regression models seek to define the connection between input features and a continuous output. These models leverage historical data to make predictionsabout forthcoming numerical values, refining their parame- ters to minimize the disparity between predicted and actual outputs.[12] Examples of regression tasks include predicting house prices, stock market trends, or temperature. Linear Re- gression, Decision Trees, and Random Forests are commonly used regression algorithms.

G. Classifiers

Classifiers are designed for tasks where the goal is to predict the categorical label or class of a given input. The output of a classifier is discrete and falls into predefined categories.[18] Common applications include spam detection, sentiment analysis, image recognition, and disease diagnosis. In these scenarios, the model is trained on labeled data tolearn patterns and relationships that allow it to categorize

new, unseen data into predefined classes. Popular algorithms for classification tasks include Decision Trees, Support VectorMachines, and Neural Networks.

H. Choosing Between Classifiers and Regressors

Deciding between classifiers and regressors is contingent upon the nature of the problem at hand and the specific typeof output needed. If the objective is to classify inputs into distinct categories, a classifier is appropriate. On the other hand, if the goal is to predict a numerical value, a regressor is more suitable. Some tasks, such as predicting customersatisfaction scores, may offer flexibility, allowing practitionersto choose between classification and regression based on the desired output representation.

4. RANDOM FOREST

Random Forest is an ensemble technique that can perform regression and classification using multiple decision trees and techniques called bootstrapping and pooling (often referred to as bagging). [6] [7] The main idea behind this is to combine multiple decision trees to determine the final outcome, rather than relying on a single decision tree. Random forest contains many trees determined by the learning model. We perform random line sampling and feature sampling from the datasetto create a sample dataset for each sample. This is calledBootstrap.[8]

Random forest is a versatile and powerful combination in machine learning and belongs to the general class of decision tree models. It works by creating multiple decision trees duringtraining outputting the average prediction of individual trees for a regression problem or model. Prediction for distribution problems. This combination has many advantages, including stability, accuracy, and the ability to reduce overfitting.[6] One key strength of Random Forests lies in their capability to handle high-dimensional data and manage collinear featureseffectively.[8] They are less prone to overfitting compared to individual decision trees and are inherently parallelizable, making them computationally efficient. Random Forests excel in both classification and regression tasks, and their predictive performance often rivals more complex models, making them a popular choice across various domains. Interpreting a Random Forest model can be challenging due to its ensemble nature, but feature importance analysis provides insights into which features contribute the most to predictive accuracy. Additionally, Random Forests offer built-in methods for assessing the model's performance on unseen data, such as out-of-bag error estimation during training.[7] In practical applications, Random Forests find use in di- verse areas, including finance for credit scoring, healthcare for disease prediction, and ecology for species classification. The adaptability, robustness, and accuracy of Random Forests make them a valuable tool in the machine learning toolkit, particularly when faced with complex datasets and the needfor reliable predictions.[8]

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 274-284

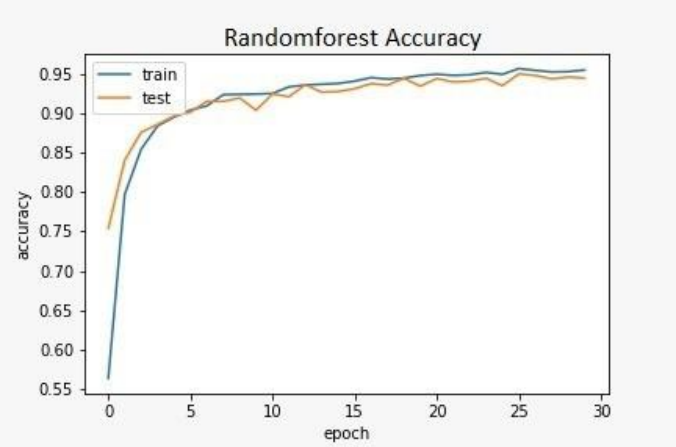

Fig. 1. Accuracy of training vs test plot of RandomForest model.

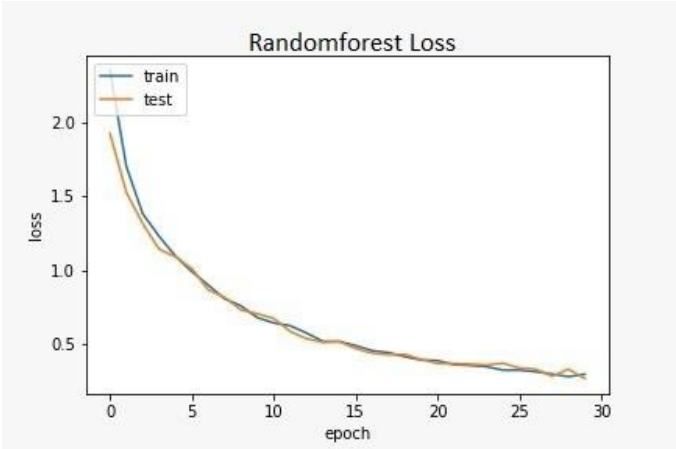

Fig. 2. Loss of training vs test plot of RandomForest model.

5. LINEAR REGRESSION

Linear regression is a statistical method widely used in model estimation and data analysis in machine learning. It belongs to the category of supervised learning, where al- gorithms are trained on datasets to establish relationships between different ideas and regular results. The principle of linear regression is to fit a line across the data points so thatthe squared difference between the predicted and actual valuesis minimized.

This algorithm assumes that the relationship between in- dividual variables (traits) and variable variables (objectives)is linear. This assumption enables the creation of a linear equation, often represented as y = mx + b, where 'y' is the predicted output, 'm' is the slope of the line, 'x' is the input feature, and 'b' is the y-intercept. The algorithm aims to determine the optimal values for 'm' and 'b' that best describethe relationship within the given data.

Linear regression finds application in various domains, from economics and finance to biology and social sciences. For example, in economics, it can be utilized to model the relationship between a company's advertising expenditure and its sales revenue. In healthcare, linear regression might be employed to analyze the correlation between patient age anda specific health outcome.

Despite its simplicity, linear regression remains a powerful tool in the data scientist's toolkit. Its interpretability and ease of implementation make it a popular choice for initial

$$Cost\ Function(MSE) = \frac{1}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})^2$$

Replace $y_{i\,pred}$ with $mx_i + c$

$$Cost\ Function(MSE) = \frac{1}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))^2$$

Fig. 3. General Formula for Linear Regression.

exploratory data analysis and as a benchmark model for more complex algorithms. Overall, linear regression's ability to uncover relationships within data and make predictions based on those relationships underscores its enduring significance in the realm of statistical modeling.

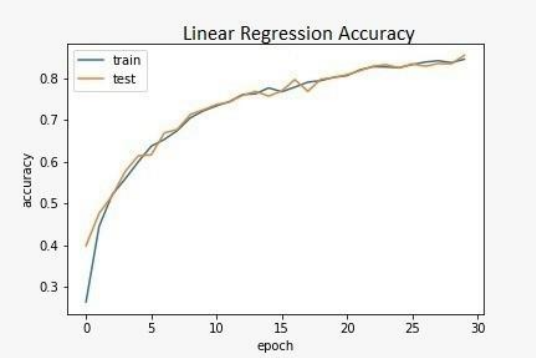

Fig. 4. Loss of training vs test plot of Linear Regression model.

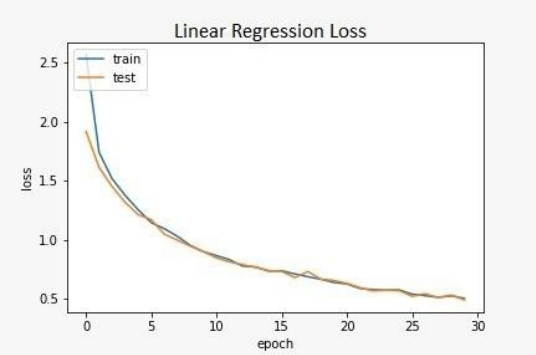

Fig. 5. Loss of training vs test plot of Linear Regression model.

6. LOGISTIC REGRESSION

Logistic regression, a widely used statistical method, is a fundamental algorithm in machine learning specifically designed for binary classification problems. Unlike linear re- gression, which predicts continuous output, logistic regression predicts the probability that an input belongs to one of two classes, typically denoted as 0 or 1. This makes it particu- larly suitable for scenarios such as spam detection, medical diagnosis, and credit risk assessment.[11]

The logistic regression algorithm employs the logistic func-tion, also known as the sigmoid function, to map the linear combination of input features to a value between 0 and 1. The logistic function's characteristic S-shaped curve ensures that the predicted probabilities are bounded, making them interpretable as likelihoods. The logistic regression equationis expressed as:

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Fig. 6. General Formula for Logistic Regression

where 'P(Y=1)' represents the probability of belonging to class 1, 'e' is the base of the natural logarithm, 'm' is the slope, 'x' is the input feature, and 'b' is the intercept.

Logistic regression is versatile and extends beyond binary classification through techniques like one-vs-all or one-vs-one for multi-class classification problems. Its simplicity, inter- pretability, and efficiency make it a popular choice, especiallywhen the relationship between features and the binary outcomeis expected to be linear.

In practical applications, logistic regression is applied in diverse fields. For instance, in healthcare, it might be employed to predict the likelihood of a patient developing a specific med- ical condition based on various health indicators. In marketing,logistic regression can assist in predicting whether a customer is likely to purchase a product or churn from a service.

In conclusion, logistic regression is a powerful algorithm for binary classification tasks, providing a probabilistic interpreta- tion that enhances decision-making. Its ease of implementation and ability to handle complex relationships make it an essentialtool in the toolkit of data scientists and analysts.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 274-284

7. SVM

- Classification with SVM:

  In the case of a classification, SVM works by finding a general plane that best separates different data points in a given domain into clusters. The ideal hyperplane is the plane that produces the edge, which is the distance between the hyperplane and the nearest datum of each class. [10] The data points closest to the hyperplane and the resulting edges are called support vectors.

  SVM can be run distributed and non-distributed using kernel functions. The kernel function transforms input data into a high-dimensional space, allowing SVM to detect biased decisions. [13] Kernel functions include linear, polynomial, and radial basis function (RBF) kernels.

- Regression with SVM:

  Support Vector Machines (SVM) can also be utilized for regression tasks, aiming to forecast a continuous output. In regression, the SVM aims to fit a hyperplane that captures the general trend of the data, minimizing deviations from the actual output values. Similar to classification, SVM regression can employ different kernel functions to model non-linearrelationships in the data.[14]

- Advantages of SVM:

  A distinctive benefit of SVM lies in its efficiency in high- dimensional spaces, rendering it well-suited for tasks involvinga substantial number of features. Additionally, SVM is less prone to overfitting, as it aims to maximize the margin between classes. The ability to handle both linear and non- linear relationships gives SVM a competitive edge in diverse applications.

- Applications of SVM:

  SVM finds application in various fields, including image recognition, handwriting recognition, bioinformatics, and fi- nance. For example, in image classification, SVM can be employed to distinguish between different objects or scenes based on their features.

In conclusion, Support Vector Machines offer a robust solution to both classification and regression problems. Their ability to handle complex relationships and perform well in high-dimensional spaces makes SVM a valuable tool in the machine learning landscape. As technology advances, SVM continues to be a cornerstone algorithm in addressing a wide array of real-world challenges.

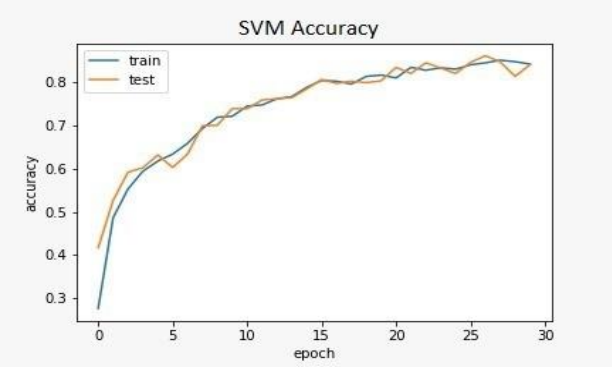

Fig. 7. Accuracy of training vs test plot of SVM model.

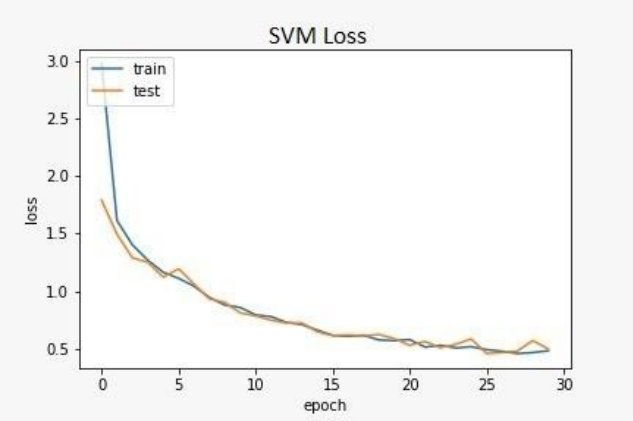

Fig. 8. Loss of training vs test plot of SVM model.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 274-284

8. K-NEAREST NEIGHBORS

K-Nearest Neighbors (KNN) is a straightforward and intu- itive machine learning algorithm applicable to both classifi- cation and regression tasks. KNN works on the principle of proximity by making predictions based on the majority of the class or the average of the k closest points at a given location. In classification, the algorithm assigns points to most of the class's closest neighbors, the choice of "k" being the key that affects the sensitivity of the algorithm to noise.[16] For regression tasks, KNN predicts the target variable by averaging the values of its k-nearest neighbors, making it effective for datasets with local patterns or no discernible global trend. The effectiveness of KNN heavily relies on the chosen distance metrics, commonly using Euclidean distance or Man- hattan distance to quantify the similarity between data points in the feature space. This guides the algorithm in identify- ing neighbors that are closer in the selected metric. KNN finds diverse applications in recommendation systems, pattern recognition, and anomaly detection. In collaborative filtering- based recommendation systems, KNN is employed to suggest items or content based on the preferences of users with similar tastes.

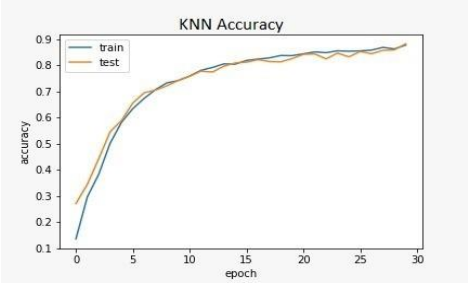

Fig. 9. Accuracy of training vs test plot of KNN model.

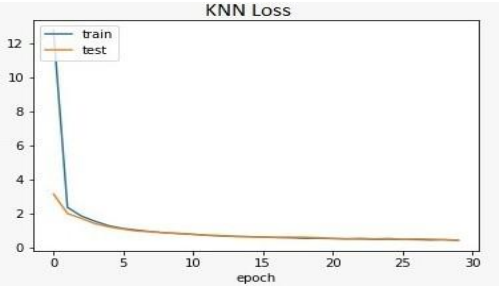

Fig. 10. Loss of training vs test plot of KNN model.

While KNN is known for its simplicity and ease of imple- mentation, it may not be the optimal choice for large datasets or those with high dimensionality due to computational de- mands. Additionally, the algorithm is sensitive to the scale of features, often requiring preprocessing steps such as normal- ization to ensure robust performance. In conclusion, K-Nearest Neighbors serves as an accessible and effective algorithm for solving both classification and regression problems, offering

a valuable tool in the machine learning toolkit, especially for smaller datasets exhibiting discernible local patterns.

a) Data Collection: A comprehensive dataset for training and evaluating AI models was acquired from Kaggle, tapping into the platform's diverse collection of second-hand car data. The dataset encompasses information from online car mar- ketplaces, dealerships, and historical sales records, ensuring a varied representation of brands, models, mileage ranges, and other relevant features.

b) Data Preprocessing: The collected data will undergo preprocessing to handle missing values, outliers, and ensure consistency across features. Categorical variables will be en- coded, numerical values normalized, and features scaled to create a standardized dataset suitable for model training.

c) Feature Selection: Features significantly influencing second-hand car prices, such as mileage, brand reputation, and kilometers driven, will be identified through exploratory data analysis and statistical methods. Prioritization of relevant features will be crucial for model accuracy.

d) Model Implementation: Linear Regression, SVM, KNN, and Random Forest models will be implemented us- ing Python's scikit-learn library. The models will be trained on a subset of the dataset and validated to ensure optimal performance.

e) Model Evaluation: The performance of each model will be gauged through various metrics, such as Mean Squared Error (MSE), R-squared, and accuracy. Through comparative analysis, we aim to identify the algorithm that delivers the most precise and dependable predictions for second-hand car prices.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 274-284

9. SOME COMMON MISTAKES

In a project as complex and multifaceted as the development of Vehicle WorthAI, several common mistakes may occur. It's crucial to identify and address these issues proactively to ensure the project's success. Here are some common mistakes to be aware of:

- Insufficient Data Quality: Issue: Using incomplete or inaccurate datasets for training the machine learning model can lead to poor predictions.
- Mitigation: Ensure thorough data cleaning and validation. Use diverse and comprehensive datasets to enhance the model's accuracy.
- Overfitting or Underfitting the Model: Issue: Overfitting occurs when a model performs well on the training data but poorly on new data. Underfitting happens when a model is too simple to capture the underlying patterns in the data.
- Mitigation: Implement techniques such as cross- validation and regularization to find the right balance. Continuously monitor and adjust the model as needed.
- Lack of Feature Engineering:
- Issue: Inadequate identification and selection of relevant features can lead to inaccurate price predictions.
- Mitigation: Conduct thorough feature engineering to se- lect the most informative attributes. Consider the impor- tance of each feature in relation to the target variable.
- Inadequate User Interface Design:
- Issue: A poorly designed user interface can hinder user interaction and impact the overall user experience.
- Mitigation: Involve UX/UI designers in the development process. Conduct usability testing to gather feedback and make iterative improvements to the interface.
- Poor Image Processing:
- Issue: If image processing is a key component, inaccura- cies in image recognition can affect the overall accuracy of the model.
- Mitigation: Implement robust image processing tech- niques, and validate results against ground truth data. Consider using pre-trained models for image recognition.

10. RESULTS

The results table serves as a comprehensive overview of the performance metrics for each employed model in predicting second-hand car prices based on the Kaggle dataset. The table includes crucial information such as the model name, max- imum accuracy achieved during training and testing phases, and the minimum loss recorded during the model's learning process.

a) Model Name: The first column of the results table displays the names of the implemented models, including Lin- ear Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forest.

b) Maximum Accuracy: The second column provides insights into the maximum accuracy achieved by each model during the training and testing phases. Accuracy reflects the ratio of correctly predicted outcomes and serves as a crucial metric for evaluating the overall performance of the model. A higher accuracy value indicates enhanced predictive capabili- ties of the model.

c) Minimum Loss: In the third column, you can find information about the minimum loss registered by each model throughout the learning process. Loss functions quantify the disparity between predicted and actual values, offering a gauge of the model's performance.

A diminished loss signifies a more accurate alignment of the model with the data.

TABLE.1- ANALYSIS OF DIFFERENT MODELS

Sno.	Model Name	Maximum Accuracy	Minimum Loss
1	Linear Regression	89.54%	1.2%
2	Randomforest Regressor	93.02%	0.7%
3	SVM	90.06%	0.8%
4	KNN	87.88%	1.4%

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 274-284

11. CONCLUSION

In summary, our research culminated in the development ofVehicle WorthAI, a web-based platform for improving used ve-hicle valuation. After exploring various machine learning mod-

els, Random Forest emerged as the top performer, achieving an impressive 93% accuracy. The robust methodology, including data collection, feature engineering, and image processing,contributed to the success. While acknowledging challenges faced, the project sets the stage for further enhancements and positions Vehicle WorthAI as a promising solution inreshaping the pre-owned vehicle market, offering transparencyand accuracy in pricing.

12. REFERENCES

[1] Mahfouz, Mohammed A., Sara M. Mosaad, and Mohamed A. Belal. ”Forecasting Vehicle Prices using Machine Learning Techniques based on Federated Learning Strategy.” International Journal of Computer Applications 975 (2023): 8887.

[2] Liu, Enci, et al. ”Research on the prediction model of the used car price in view of the pso-gra-bp neural network.” Sustainability 14.15 (2022): 8993.

[3] Bukvić, Lucija, et al. ”Price Prediction and Classification of Used- Vehicles Using Supervised Machine Learning.” Sustainability 14.24(2022): 17034.

[4] Gajera, Prashant, Akshay Gondaliya, and Jenish Kavathiya. ”Old car price prediction with machine learning.” Int. Res. J. Mod. Eng. Technol.Sci 3 (2021): 284-290.

[5] Gegic, Enis, et al. ”Car price prediction using machine learning tech- niques.” TEM Journal 8.1 (2019): 113.

[6] Rigatti, S. J. (2017). Random forest. Journal of Insurance Medicine, 47(1), 31-39.

[7] Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25, 197-227.

[8] Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L. (2019). Developing prediction models for clinical use using logistic regression: an overview. Journal of thoracic disease, 11(Suppl 4), S574.

[9] Tondi, K. M., Hatmi, W., & Nurmedika, N. (2023). ANALYSIS OF FACTORS AFFECTING FARMING HOUSEHOLD FOOD SECU- RITY POST-NATURAL DISASTER IN LAMBARA, CENTRAL SU-LAWESI. AGROLAND The Agricultural Sciences Journal (e-Journal).

[10] Huang, M. W., Chen, C. W., Lin, W. C., Ke, S. W., & Tsai, C. F. (2017). SVM and SVM ensembles in breast cancer prediction. PloS one, 12(1), e0161501.

[11] Ahmad, S. A., Ahmed, H. U., Rafiq, S. K., Mahmood, K. O. F., Rostam, K. J., & Jafer, F. S. (2023). A Comprehensive Exploration on the Effect of Waste Glass Powder as a Partial Replacement of Cement in Mortar:A Review, Analysis, and Modeling Investigation. Arabian Journal for Science and Engineering, 1-28.

[12] Ziebart, A., Dremel, J., Hetjens, S., Nieuwkamp, D., Linn, F. H., Etminan, N., & Rinkel, G. J. Case Fatality and Functional Outcome after Spontaneous Subarachnoid Haemorrhage: A Systematic Review and Meta-Analysis of Time Trends and Regional Variations in Population- Based Studies. Available at SSRN 4638329.

[13] Lin, Y., Guo, H., & Hu, J. (2013, August). An SVM-based approach for stock market trend prediction. In The 2013 international joint conferenceon neural networks (IJCNN) (pp. 1-7). IEEE.

[14] Otchere, D. A., Ganat, T. O. A., Gholami, R., & Ridha, S. (2021). Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. Journal of Petroleum Science and Engineering, 200, 108182.

[15] Danacı, Ç ., Avcı, D., & Arslan Tuncer, S. Xgboost-Powered Pes Planus Diagnosis: Unearthing the Most Relevant X-Ray Image Features. Avail-able at SSRN 4633291.

[16] Avudaiammal, R., Rajangam, V., Durai Raji, V., & Senthil Kumar, S. (2023). Color Models Aware Dynamic Feature Extraction for Forest Fire Detection Using Machine Learning Classifiers. Automatic Control and Computer Sciences, 57(6), 627-637.

[17] Kushwah, J. S., Kumar, A., Patel, S., Soni, R., Gawande, A., Gupta, S. (2022). Comparative study of regressor and classifier with decision tree using modern tools. Materials Today: Proceedings, 56, 3571-3576.

[18] Liu, X., Li, S., Kan, M., Zhang, J., Wu, S., Liu, W., ... Chen, X. (2015). Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In Proceedings of the IEEE International Conference onComputer Vision Workshops (pp. 16-24).

[19] Van Engelen, J. E., Hoos, H. H. (2020). A survey on semi-supervised learning. Machine learning, 109(2), 373-440.

[20] Mitchell, T. M. (1997). Machine learning. Balcázar, J. L., Bonchi, F., Gionis, A., Sebag, M. (Eds.). (2010). Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010. Proceedings, Part III (Vol. 6323). Springer.

[21] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010: 19th International Con- ference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers (pp. 177-186). Physica-Verlag HD.

[22] Shoeb, A. H., Guttag, J. V. (2010). Application of machine learning to epileptic seizure detection. In Proceedings of the 27th international conference on machine learning (ICML-10) (pp. 975-982).

[23] Firdausi, I., Erwin, A., Nugroho, A. S. (2010, December). Analysis of machine learning techniques used in behavior-based malware detection. In 2010 second international conference on advances in computing, control, and telecommunication technologies (pp. 201-203). IEEE.

[24] Ahmed, N. K., Atiya, A. F., Gayar, N. E., El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. Econometric reviews, 29(5-6), 594-621.

[25] Li, F. (2010). The information content of forward-looking statements in corporate filings—A na¨ıve Bayesian machine learning approach. Journalof Accounting Research, 48(5), 1049-1102.

[26] Van Engelen, J. E., Hoos, H. H. (2020). A survey on semi-supervised learning. Machine learning, 109(2), 373-440.

[27] Reddy, Y. C. A. P., Viswanath, P., Reddy, B. E. (2018). Semi-supervised learning: A brief review. Int. J. Eng. Technol, 7(1.8), 81.

[28] Huo, H., Rong, Z., Kononova, O., Sun, W., Botari, T., He, T., ... Ceder, G. (2019). Semi-supervised machine-learning classification of materials synthesis procedures. Npj Computational Materials, 5(1), 62.

[29] Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. L. A., Elkhatib, Y., ... Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges. IEEE access, 7, 65579-65615.

[30] Hahne, F., Huber, W., Gentleman, R., Falcon, S., Gentleman, R., Carey, V. J. (2008). Unsupervised machine learning. Bioconductor case studies, 137-157.

[31] Qiang, W., Zhongli, Z. (2011, August). Reinforcement learning model, algorithms and its application. In 2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC) (pp. 1143-1146). IEEE.

[32] Mehta, D. (2020). State-of-the-art reinforcement learning algorithms. International Journal of Engineering Research and Technology, 8, 717- 722.

[33] Jordan, S., Chandak, Y., Cohen, D., Zhang, M., Thomas, P. (2020, November). Evaluating the performance of reinforcement learning al- gorithms. In International Conference on Machine Learning (pp. 4962- 4973). PMLR.