e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

USED CAR PRICE PREDICTION USING MACHINE LEARNING

Prof. Priyanka Gupta[*1], Aditi Ghorwade[*2], Amit Gupta[*3], Digvijay Mudde[*4], Pratiksha Gore[*5]

[*1,2,3,4,5]Artificial Intelligence & Data Science, DR. D. Y. Patil Institute Of Technology, Pune, Maharashtra, India.

DOI : https://www.doi.org/10.56726/IRJMETS57635

ABSTRACT

The surging popularity of used cars, driven by the increasing preference for car ownership, has created a demand for efficient pricing models. This research delves into the potential of machine learning for used car price prediction, aiming to empower key stakeholders in the market. We propose a supervised learning model that analyzes a vast amount of seller data, encompassing pertinent vehicle attributes like mileage, make, model, year, and condition. This comprehensive data analysis helps the model predict used car prices with high accuracy, providing valuable insights into the current automotive market and future price trends. By comparing the linear regression approach with other classification algorithms, we explore the most effective model for price prediction. This comparative analysis ensures the chosen model offers the most accurate and reliable predictions for users. With accurate price predictions, sellers can maximize profits and ensure a faster turnaround time. For buyers, the platform can provide valuable insights to negotiate a fair price and make informed decisions in the dynamic used car market. This research contributes to a more efficient and transparent used car market experience for all participants.

Keywords: Machine Learning, Regression, Feature Engineering, R2 Score.

I. INTRODUCTION

Advances in electric vehicles and automation are significantly reshaping the automotive industry, placing an increasing emphasis on safety and efficiency. However, accurately measuring the value of a wide range of vehicles remains a challenging task. Existing websites do not provide comprehensive cost information, highlighting the limitations and inefficiencies of current methods. This study explores how sophisticated machine learning algorithms can be harnessed to achieve accurate and personalized price predictions in an ever-evolving market. Our models, which encompass both fuel-efficient daily drivers and high-performance sports cars, will take into account a variety of features. By integrating a multitude of data points, we aim to develop a robust system that clearly delineates the unique characteristics of each vehicle. Through the analysis of historical events and market fluctuations, our model can offer a valuable index for both buyers and sellers. With precise price information, buyers can negotiate with greater confidence, and sellers can refine their strategies effectively. Furthermore, the model is capable of detecting inconsistencies or anomalies in existing data, thereby enhancing business transparency. By providing reliable cost estimates to stakeholders, machine learning models can facilitate smoother business operations, reduce transaction costs, and ultimately lead to improved business outcomes.

II. A REVIEW OF THE LITERATURE

In the field of predicting vehicle sales using machine learning, researchers and businesses look at how ML algorithms can help forecast sales for financial stability. When working on such projects, we look at various research papers to understand the latest developments.

Used Car Price Prediction Using Machine Learning

One important thing we consider is the balance between over-fitting and under-fitting in our models. Over-fitting happens when our model learns too much from the training data and doesn't perform well on new data. Underfitting occurs when our model doesn't capture enough information from the data and also performs poorly. We aim for a balance between these two, known as the Bias-Variance trade-off. Choosing the right variables or attributes for our model is crucial. Techniques like Lasso help us pick the most important

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

attributes, reducing errors in our predictions. Similarly, decision trees can suffer from overfitting if they become too complex, so we may need to prune them to improve performance.

III. ARCHITECTURE DIAGRAM

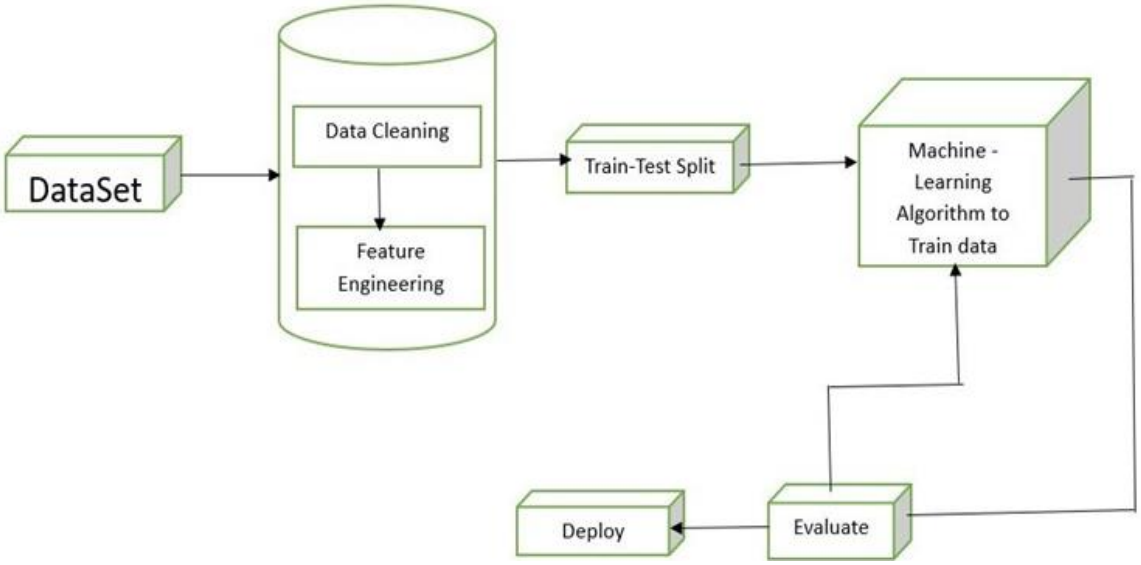


Fig: Machine Learning- Car Price Prediction Framework

IV. METHODOLOGIES

Accurate vehicle sales forecasting relies on several key factors: brand and model identification, vehicle mileage, and usage duration. Analyzing a well-organized dataset with diverse features, such as transmission type, safety features, door count, dimensions, navigation system presence, and vehicle cost, allows for precise sales predictions. Fuel type significantly impacts the cost per mile due to fluctuating fuel prices. This study employs various techniques to enhance prediction accuracy. The proposed car price prediction approach, detailed in Figure 1, involves several steps to create a comprehensive and reliable forecasting model.

4.1 Data Collection

The dataset, sourced from the Quicker website, pertains to used items and encompasses various product attributes: name, company, year, kilometers driven, fuel type (e.g., gasoline, diesel), image, and price. During the textual data processing phase, we combined the textual features (title and description) into a single attribute called "product description." The dataset consists of 892 observations across 6 distinct product types, with item prices ranging from $1 to $20,000.

4.2 Data Preprocessing:

Data preprocessing is a crucial step in preparing a dataset for machine learning models. Here are specific points for converting objects to integers, removing unwanted data, and handling outliers:

- Converting Object to Integer:

Identify columns with object data types that need to be converted to integers. For categorical variables, consider using techniques like one-hot encoding or label encoding to convert them into numerical representations. For ordinal variables, assign numerical values based on their inherent order..

- Removing Unwanted Data:

Identify and remove irrelevant or redundant columns that do not contribute meaningful information to the model. Handle missing data by either imputing values or removing rows/columns with missing data, depending on the extent of missing values and their impact on the analysis.

- Handling Outliers:

We identify outliers using Z-scores or box plots. Depending on their impact and data nature, outliers might be removed, transformed (e.g., log-transformed), or capped to minimize their influence on the model.

- Converting Data Types:

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

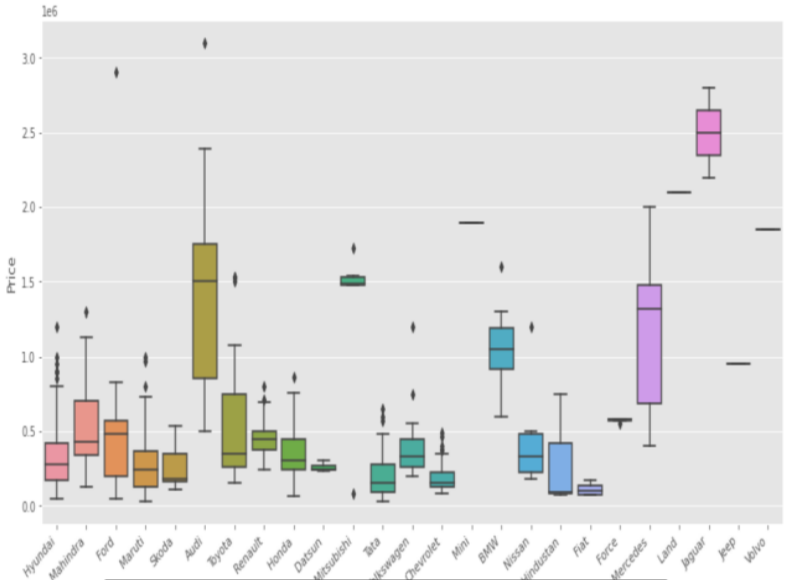
Ensure that numerical variables are in the correct data type (e.g., converting float to integer if appropriate).

Verify that the data types align with the requirements of the machine learning algorithms being used.

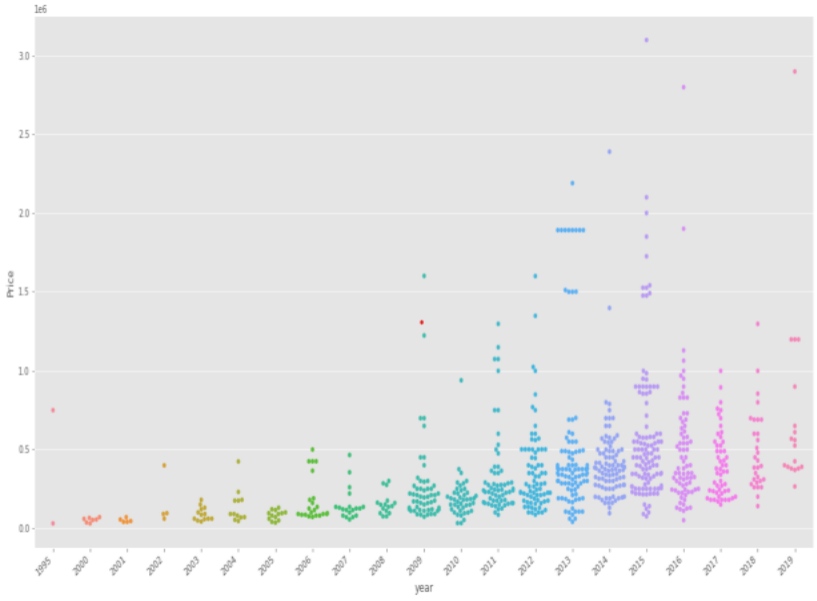
4.3 Data Analysis:

Car price prediction relies on data analysis to extract valuable insights for accurate model development. Exploratory Data Analysis (EDA) helps understand the distribution and relationships between variables like car make, model, year, mileage, and fuel type. Statistical techniques and visualizations, such as scatter plots and correlation matrices, reveal key factors influencing price. Data preprocessing ensures a clean dataset for training, including handling missing values, converting categorical variables, and addressing outliers. Machine learning algorithms, ranging from regression models to ensembles, then predict car prices based on the analyzed features. Regular model validation and testing with unseen data guarantee reliability for real-world price predictions.

Analysing relation of Company with Price:

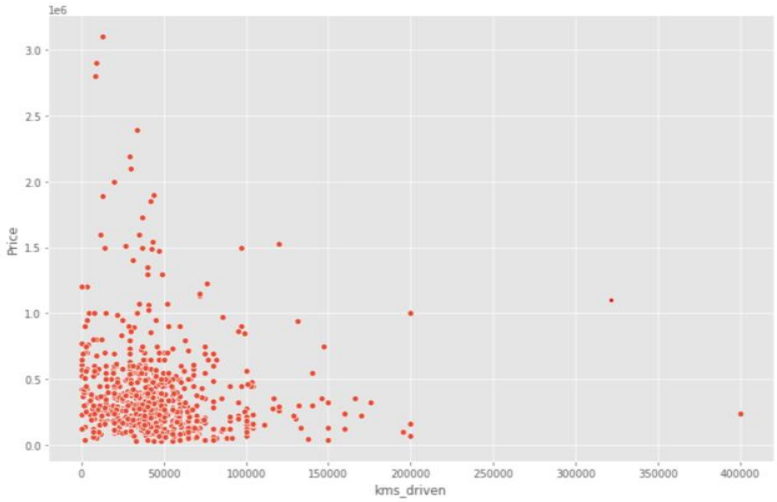


Analysing relationship of Year with Price:

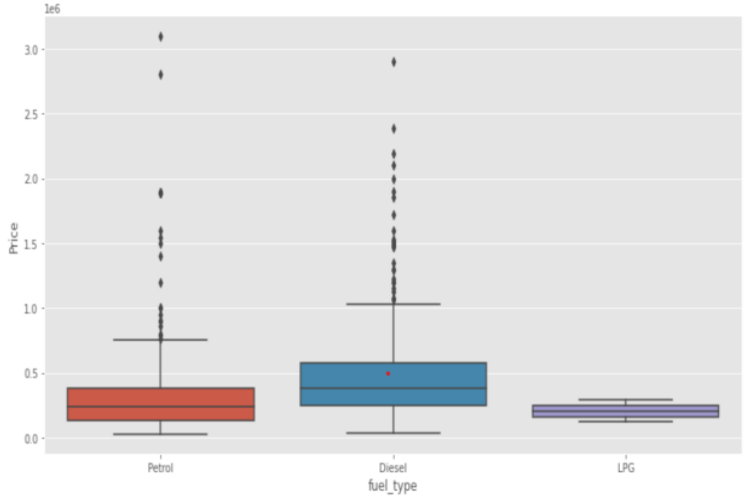


Analysing relationship of Kms_driven with Price:

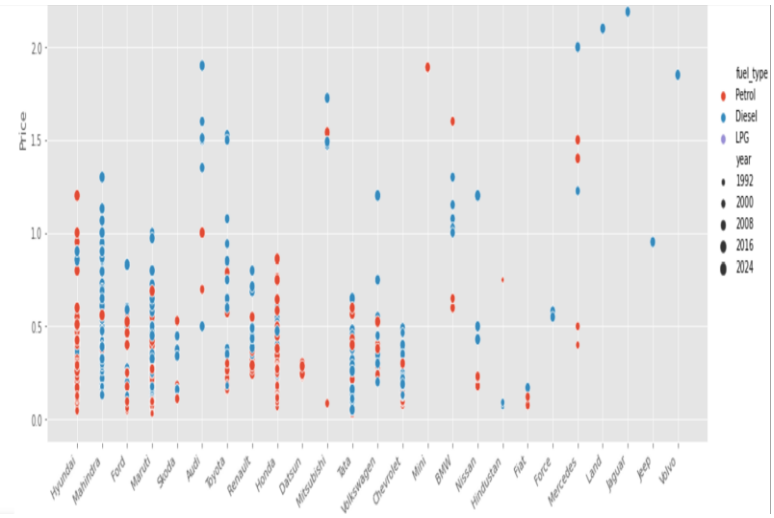
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com



Analysing relationship of Fuel type with Price:



Analysing relationship of Price with Fuel type, Year and Company mixed:



4.4 Feature Engineering:

Feature engineering involves generating new features or modifying existing ones to enhance a model's pattern recognition capabilities. For example, you can derive an 'age' feature by subtracting a car's manufacturing year from the current year. This 'age' feature may more accurately reflect how a vehicle's age affects its price4.9 Model Selection:

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

1. Linear Regression:

Linear regression plays a crucial role in car price prediction. It establishes a linear relationship between features (make, model, year, mileage) and the target variable (car price). This approach allows analysts to quantify the impact of each feature on price. During training, the algorithm finds the optimal coefficients for a linear equation that minimizes the difference between predicted and actual prices. Once trained, the model predicts car prices based on new data. Linear regression offers a straightforward and interpretable method, revealing how individual features influence a car's overall price. Regular validation and testing ensure accuracy and reliability across diverse datasets.

2. Random Forest:

Random Forest is a widely used machine learning algorithm for car price prediction. It works by creating numerous decision trees during training, with each tree using different subsets of features like make, model, year, and mileage to predict car prices. The final prediction is made by averaging the predictions from all the trees. This ensemble method improves predictive accuracy and reduces overfitting. Random Forest is effective at capturing complex, nonlinear relationships in the data, making it a reliable choice for car price prediction. Regular validation and testing ensure the model's accuracy and applicability to different datasets, which enhances its use in automotive pricing analytics.

3. Decision Tree Regression:

Decision tree regression estimates used car prices by building a tree structure that represents decision-making processes. Each node signifies a feature (such as age, mileage, or brand), and branches denote the decision criteria based on these features. This tree splits the data into smaller, more manageable subsets, with the leaves (end nodes) providing the predicted prices. The model effectively captures non-linear relationships and feature interactions, which helps in managing complex patterns in car price data. Through recursive data partitioning, it identifies significant factors influencing prices, enabling precise and interpretable predictions.

4. R2 Score:

When assessing the effectiveness of automobile price prediction models, the R2 score, also known as the coefficient of determination, is frequently employed. It gauges how much of the variation in car pricing the model can account for. A model is considered effective when it catches changes in vehicle pricing when its R2 score is greater, ranging from 0 to 1. A score of one indicates an ideal fit. An essential indicator of the model's accuracy and dependability in automobile price prediction is its high R2 score, which indicates that the chosen variables largely explain the variance in vehicle prices. To guarantee that the model appropriately reflects and anticipates the intricacies of automobile pricing, regular testing using R2 scores is conducted.

Regression Analysis Validation Results:

Model	R2 Score
Linear Regression	0.89
Random Forest	0.24
Decision Tree Regression	0.70

Regression analysis relies on specific assumptions to ensure the accuracy and validity of its results:

1) Linearity:

For accurate interpretation, the relationship between the independent and dependent variables needs to be linear. This means changes in the independent variable result in proportional changes in the dependent variable.

2) Homoscedasticity:

Consistent error variance is crucial for reliable analysis in linear regression. This translates to a constant spread of data points around the fitted line.

3) Independence:

Errors in linear regression should be independent of each other. This ensures the error associated with one observation doesn't influence the errors of other observations.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

4) Normality:

Error terms should follow a normal distribution. Requires the distribution of errors to exhibit a bell-shaped curve.

5) Grid Search for Hyperparameter Tuning:

Hyperparameter tuning involves systematically searching through combinations of hyperparameters to find the optimal set for each model, thereby improving predictive performance.

Grid search involves defining a grid of hyperparameter values to explore. For each combination, the model is trained and evaluated. The combination yielding the best performance metrics is selected as the optimal set of hyperparameters.

V. GAPS IDENTIFIED AND LIMITATIONS

After reviewing the literature, we have identified the following gaps and limitations. By identifying and addressing these gaps and limitations, we aim to encourage research that aligns with the following:

1. **Data Quality and Availability**: Accessing comprehensive and reliable data on used car sales, especially in developing countries, can be challenging. This challenge could potentially lead to biases or inaccuracies in the predictive model.
2. **Feature Selection and Engineering**: Determining which attributes truly influence used car prices and how to represent them effectively in the model can be challenging. This could potentially result in suboptimal feature selection and engineering.
3. **Model Selection and Evaluation**: The paper does not provide insights into the selection process for choosing specific algorithms. This gap could potentially lead to suboptimal model choices for the prediction task.

Limitations:

1. **Bias and Fairness**: Training data may have its own bias, which will affect the model's prediction. This may lead to unfair consequences for some citizens or interests.
2. **Generalization and Robustness**: It can be difficult to evaluate a model's ability to capture complex patterns in training data versus its ability to perform well on completely new data. This can limit the model's effectiveness in real-world situations.
3. **Interpretability and Explainability**: Some complex models (like decision trees) can resemble black boxes; decision-making processes are not clear. Lack of transparency can undermine user trust and make it difficult to ensure regulatory compliance.

Addressing these gaps and limitations requires a comprehensive approach that involves careful data collection and preprocessing, rigorous model selection and evaluation, attention to fairness and bias mitigation, and efforts towards interpretability and transparency. Additionally, ongoing monitoring and refinement of the predictive platform based on real-world feedback and new data are essential for its effectiveness and reliability.

VI. CONCLUSION

This research is motivated by the under-exploration of price prediction for second-hand goods, particularly used cars. Sellers often rely heavily on the brand when setting prices, creating a gap in data-driven approaches. While existing studies predict used car prices, this paper introduces a novel approach. It leverages exploratory data analysis and features derived from both current and historical data to forecast future trends in the used-car market. By employing supervised machine learning techniques and rigorous validation methods, the model ensures statistically robust predictions.

In Summary:

- Data is sourced from an online platform specializing in used cars, and relevant features influencing pricing are identified.
- After removing non-available values and irrelevant features, the supervised machine learning techniques are applied to the initial dataset, and the validation is compared against the price prediction outcomes from a second dataset, focusing on critical features.

- Linear regression emerges as the most accurate prediction model, particularly when essential features like price and model information are available.

The study encompasses diverse vehicle types, considering usage conditions and prices. Various techniques for numeric data preprocessing and text analysis are employed to handle structured and unstructured data, respectively. The competitive advantage of predicting trends in the second-hand market through data mining and analysis lies in optimizing vehicle prices, averting misclassification and associated risks, and enhancing consumer awareness for informed purchasing decisions. Future research endeavors will involve additional datasets collected over the next two quarters, utilizing data mining, machine learning techniques, and diverse model validation methods for optimization. This necessitates the extraction of additional variables related to vehicle condition and drive type. As the interest in used vehicles grows, influencing perceptions of vehicle valuation and price prediction, machine learning methods play a significant role across various applications.

VII. REFERENCES

[1] Prashant Gajera, Akshay Gondaliya, Jenish Old Car Price Prediction with Machine Learning. 2021.

[2] Jasmina Pasagic, Borna Abramovic, Lucija Bukvic, Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning. arxiv 2021.

[3] Ahmad Salah, Ahmed Fathalla, Piccialli Francesco, Deep end-to-end learning for price prediction of second-hand items. 2021.

[4] Enis Gegic, Jasmin Kevric, Dino Keco. Car Price Prediction Using ML Techniques, 2021.

[5] Mukkesh Ganesh, Pattabiraman Venkatasubbu. Used Car Price Prediction Using Supervised learning Techniques 2021.

[6] Fadi Al-Turjman, Sinem Alturjman, Chadi Altrjman. Vehicle Price Classification and Prediction Using Machine Learning, 2022.

[7] Lucija Bukvi, Tomislav Fratrovi´c, Borna Abramovic, Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning, 2022

[8] Used Car Price Prediction using Different Machine Learning Algorithms Pallavi Bharambe, Shreyas Dandekar, Prerna Ingle, Used Car Price Prediction using Different Machine Learning Algorithms, 2021

[9] Snehit Shaprapawad, Premkumar Borugadda, Nirmala Koshiga, Car Price Prediction:An Application of Machine Learning, 2023

[10] Alamaniotis M, Bargiotas D, Bourbakis NG, Tsoukalas LH Genetic optimal regression of relevance vector machines for electricity pricing signal forecasting in smart grids. IEEE Trans 2015.

[11] Chen C, Li K, Teo SG, Chen G, Zou X, Yang X, Vijay RC, Feng J, Zeng Z Exploiting spatiotemporal correlations with multiple 3d convolutional neural networks for citywide vehicle flow prediction. IEEE 2018.

[12] Chen J, Li K, Tang Z, Bilal K, Li K A parallel patient treatment time prediction algorithm and its applications in hospital queuing-recommendation in a big data environment, IEEE 2016.

[13] Chitsaz H, Zamani-Dehkordi P, Zareipour H, Parikh PP Electricity price forecasting for operational scheduling of behind-the-meter storage systems, IEEE 2018.

[14] He K, Zhang X, Ren S, Sun J Deep residual learning for image recognition IEEE 2016

[15] Kalaiselvi N, Aravind K, Balaguru S, Vijayaragul V Retail price analytics using backpropagation neural network and sentimental analysis, IEEE 2017

[16] Samruddhi, K.; Kumar, R.A. Used Car Price Prediction using K-Nearest Neighbor Based Model, arxiv 2020

[17] AlShared, A. Used Cars Price Prediction and Valuation using Data Mining Techniques, arxiv 2021.

[18] Siva, R.; M, A. Linear Regression Algorithm Based Price Prediction of Car and Accuracy Comparison with Support Vector Machine Algorithm, arxiv 2022.

[19] Pudaruth, S. Predicting the Price of Used Cars using Machine Learning Techniques, arxiv 2014.