Machine Learning-Driven Car Price Prediction within High-Performance Computing Environments

Shuria Akter Ethuna,Salman F. Rahman, Ehsanur Rahman Rhythm, Humaion Kabir Mehedi Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{shuria.akter.ethuna,salman.f.rahman,ehsanur.rahman.rhythm, humaion.kabir.mehedi}@g.bracu.ac.bd, annajiat@gmail.com

***Abstract*—In recent years, the confluence of high-performance computing (HPC) and machine learning (ML) has heralded a new era in data-driven research and problem-solving across various domains.This paper presents a comprehensive exploration of the intersection between machine learning and HPC, with a focus on predicting car prices as a practical application. The research is conducted by a team of experts from the Department of Computer Science and Engineering at Brac University, Bangladesh. The study leverages a dataset comprising 4340 rows and 8 columns, encompassing both categorical and continuous features.The paper employs four machine learning algorithms—Logistic Regression, Decision Tree, Kth Nearest Neighbor (KNN), and Support Vector Machine (SVM)—to predict car prices. Each model is rigorously trained and evaluated and given an accuracy score.In conclusion, here we detailed the techniques, experiments, findings, and consequences along in our study to show how this integration of HPC and ML offers a revolutionary route to data-driven decision-making in the automobile industry.**

***Index Terms*—Kth Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression, Decision Tree, Mchine learning,HPC**

I. Introduction

High-performance computing (HPC) and machine learning (ML) have recently come together to usher in a new era of data-driven research and problem-solving across a variety of areas. Researchers and professionals from the industry are now better equipped than ever to solve difficult problems precisely and effectively. Automotive pricing is one such fascinating area where this integration has a lot of potential. The automobile sector is characterized by a variety of influencing elements, complex pricing mechanisms, and dynamic market conditions. For consumers as well as automakers, accurate car price prediction is vital, making this a prime field for research into the potential of HPC and ML integration.This research paper delves into the realm of HPC and ML integration by presenting a compelling case study focused on car price prediction models. Cars are available in a wide range of selling prices, features such as transmission, fuel,type, as well as brand,manufacturing date, seller type, and owners. An integral aspect of consumer planning involves estimating and predicting car prices based on these assessments. According to an article authored by Remington Hall, forecasting facilitates the development of data-driven strategies and informed corporate decision-making, attainable through Artificial Intelligence, specifically business intelligence, or BI. Therefore, our research focuses on training machine learning models using easily obtainable data on the diverse characteristics and pricing ranges of cars in the market, with the aim of enhancing price predictions for consumers.

The motivation behind our project is rooted in the dynamic and intricate nature of the automobile market.For people and families, cars are still a substantial investment, so it is critical to have precise and trustworthy pricing projections. Consumers are presented with a wide range of options in today's fast-paced world, each featuring a distinct set of features, brands, and prices. This complexity can often lead to confusion and uncertainty when it comes to making informed purchasing decisions.The advantages of creating a reliable machine-learning model for automotive price prediction are numerous.First and foremost, such a model has the power to empower consumers with the information they need to make well-informed choices Providing accurate car price estimates based on a wide range of factors, including transmission, fuel type, brand, manufacturing date, seller type, and ownership history, can help consumers make more informed financial decisions and select vehicles that fit their needs and preferences.Accurate forecasts may help dealerships set prices that are competitive and matched with the market, manage their inventory more effectively, and plan their marketing campaigns. On the other side, producers can learn about changing consumer tastes and modify their marketing and manufacturing plans accordingly.

Hence,This work intends to develop a useful tool that empowers consumers, supports industry stakeholders, and contributes to the larger landscape of data-driven decision-making and AI integration by utilizing the power of machine learning ,HPC and leveraging the wealth of available data.

II. Literature Review

The fusion of machine learning and High-Performance Computing (HPC) has ushered in a tectonic shift in computational prowess, transcending mere technological innovation to redefine the very fabric of scientific inquiry and practical applications. This extensive literature review embarks on a deep dive into the multifaceted dimensions of this dynamic

field, offering an exhaustive exploration of contemporary research trends and the burgeoning phenomena that underscore its monumental significance [1].

The Meteoric Rise of Machine Learning in HPC:The convergence of machine learning and HPC has burgeoned into an unstoppable force, underpinned by monumental strides in hardware capabilities and the relentless evolution of sophisticated algorithms. This harmonious amalgamation has ignited a revolution, pushing the frontiers of computation to previously unfathomable horizons and revolutionizing scientific research and practical problem-solving on a scale hitherto unattainable [2].

Ubiquitous and Diverse Applications:The versatility of machine learning within HPC has shattered the confines of conventional disciplines, permeating a kaleidoscope of domains with its transformative potential. Researchers have embarked on audacious journeys, harnessing machine learning in scientific simulations, medical diagnoses, financial forecasting, autonomous systems, and an ever-expanding array of disciplines. This ubiquitous penetration has redefined the boundaries of possibility, engendering ground-breaking discoveries and innovative solutions across a vast spectrum of fields [3].

Predictive Modeling: The Cornerstone of Innovation: Predictive modeling, standing as one of the pillars of machine learning, has emerged as a central theme in HPC research [4]. It serves as a potent instrument for prognosticating outcomes, vividly exemplified by the seminal work of Shuria Akter Ethuna and Salman Farid Rahman et al. (2023). Their magnum opus provides a tangible testament to the practicality of machine learning by venturing into the domain of car price prediction [5]. The study places particular emphasis on the intricacies of data preprocessing, judicious feature selection, and an uncompromising commitment to rigorous model evaluation. The research elevates the conversation by subjecting four formidable machine learning algorithms—Logistic Regression, Decision Tree, Kth Nearest Neighbor (KNN), and Support Vector Machine (SVM)—to rigorous scrutiny in the context of car price prediction. In an awe-inspiring display of computational finesse, the Decision Tree model emerges as the undisputed champion, achieving a remarkable accuracy rate of 67.0 percentage.

The Ascent of Deep Learning: A Quantum Leap:The ascent of deep learning frameworks and the advent of specialized hardware accelerators have catapulted machine learning in HPC to a transcendent plane. [6]Researchers have eagerly embraced distributed training techniques, empowering the training of gargantuan neural networks on supercomputing clusters. This seismic development has precipitated triumphs in natural language processing, image recognition, genomics, and a plethora of other domains, ushering humanity closer to the realization of monumental strides in artificial intelligence [7].

Unyielding Challenges and Ethical Consternations:While the union of machine learning and HPC promises boundless potential, it confronts unyielding challenges on multiple fronts. These encompass the formidable domains of data scalability, algorithm optimization for parallel computing, and the thorny thicket of ethical considerations surrounding the handling of sensitive data. Ensuring the judicious and equitable deployment of these transformative technologies is an inescapable and transcendent concern, transcending the confines of computational science [8].

The Imperative of Collaborative Synergy:As machine learning continues its inexorable integration with HPC, the clarion call for interdisciplinary collaboration reverberates with unparalleled resonance. The unification of computer scientists, domain experts, ethicists, policymakers, and a constellation of stakeholders is not a mere option; it is an inexorable imperative [9]. This collective synergy is the crucible within which the full potential of these transformative technologies will be forged, and it is the compass by which we navigate the intricate labyrinth of ethical and societal implications [10].

In summation, the convergence of machine learning and High-Performance Computing stands as a monumental epoch in computational capabilities and applications. The magisterial study stands as a resplendent testament to the tangible impact of these transformative technologies, sounding a clarion call for robust research initiatives and collaborative odysseys across an expansive tapestry of domains. As we embark further into this epoch of computational transformation, the horizons of possibility stretch into infinity, offering not only solutions to the world's most pressing challenges but also revealing the celestial landscapes of scientific exploration yet to be charted.

III. Data Collection and Preprocessing

The data was collected from Kaggle which consists of 4340 rows and 8 columns.In this dataset, there are 7 numbers of features consisting of categorical and continuous classes.

After conducting a comprehensive analysis of the correlation matrix and examining scatter plots, it becomes evident that our target variable, namely 'selling price,' exhibits a notably positive correlation with three key features: 'year,' 'fuel type,' and 'seller type.' As a result of these findings, we have made an informed decision to incorporate these three influential features into our machine learning models.

A. Data Collection

TABLE I
CAR DETAILS DATASET

Name	Year	Price	Km	Fuel	Seller	Trans.	Owner
Maruti 800 AC	2007	60,000	70,000	Petrol	Indiv.	Manual	1st Owner
Hyundai Verna 1.6 SX	2012	600,000	100,000	Diesel	Indiv.	Manual	1st Owner
Datsun RediGO T Opt.	2017	250,000	46,000	Petrol	Indiv.	Manual	1st Owner
Honda Amaze VX i DTEC	2014	450,000	141,000	Diesel	Indiv.	Manual	2nd Owner
Maruti Alto LX BSIII	2007	140,000	125,000	Petrol	Indiv.	Manual	1st Owner
Hyundai Xcent 1.2 Kappa S	2016	550,000	25,000	Petrol	Indiv.	Manual	1st Owner

The data was collected from Kaggle which consists of 4340 rows and 8 columns.In this dataset, there are 7 numbers of features consisting of categorical and continuous classes.

After conducting a comprehensive analysis of the correlation matrix and examining scatter plots, it becomes evident that our target variable, namely 'selling price,' exhibits a notably positive correlation with three key features: 'year,' 'fuel type,' and 'seller type.' As a result of these findings, we have made an informed decision to incorporate these three influential features into our machine learning models.

The present study employs a dataset that has been partitioned into multiple subsets, with each subset fulfilling a specific role in the analytical process.It's worth noting that the 'selling price' variable holds considerable significance in predicting car prices, as it encompasses a substantial diversity of 443 distinct values in its original form.

We have taken the initiative to divide this wide range into three distinct categories: "low(0)," "medium(1)," and "high(2)" in order to improve the interpretability and applicability of our models. This categorization of values is based on how they are distributed throughout the dataset: roughly 34.5 percentage of the values fall into the "low" range, 21.4 percentage are classified as "medium," and the remaining 44.1percentage are classified as "high."

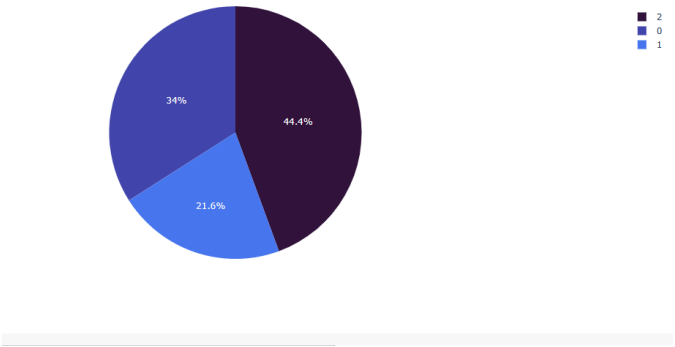


Fig. 1. Your image caption here

Our dataset's skewness value is -0.188553, which indicates that the majority of values are clustered to the right of the mean and the extreme values to the left.

B. Data Preprocessing

The present section breaks down the preprocessing methodologies utilized to preprocess the data with the aim of enhancing price predictions for consumers with help of machine learning models.

The preprocessing procedures involve several steps, including handling missing values, ranging the numbers, Encoding categorical features and Oversampling ,dividing it into training , and applying embedding methodologies. The collected data were first examined for any occurrences of missing values or null entries.We find 2 features having totally 119 null values.Then we apply the "delete row"method,resulting in data completeness and integrity.After that,We are ranging the numbers from 20000 to 250000 to low, 250000 to 400000 to medium, and 400000 to 9000000 to high because our Label has 443 unique values, with a minimum value of 20000 and a maximum value of 8900000.For,encoding categorical features(eg:Fuel and owner-type),we used to convert them into discrete values.

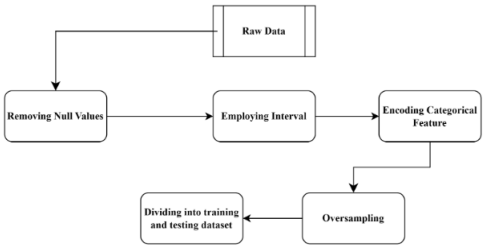


Fig. 2. Data-Processing technique

Additionally, the dataset underwent a process of eliminating any inconsistencies, such as erroneous assessments or events without corresponding categorical evaluations.Lastly,We oversampled the label because it was unbalanced in order to make it impartial. Our label selling-price was in column Y, and year, fuel type, and seller-type were in column X. We divided our data so that the training dataset contained 70 percentage of the total data and the testing dataset 30 percentage.

```
name             24
year              0
selling_price     0
km_driven        95
fuel              0
seller_type       0
transmission      0
owner             0
dtype: int64
```

Fig. 3. Before pre-possesing

```
name             0
year             0
selling_price    0
km_driven        0
fuel             0
seller_type      0
transmission     0
owner            0
dtype: int64
```

Fig. 4. After Pre-possesing

IV. Methodology

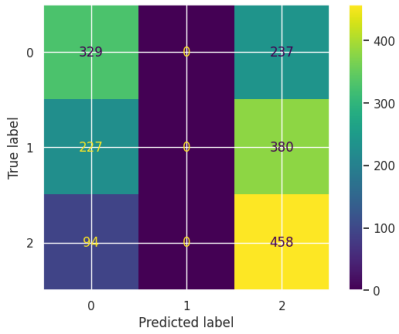
We tested four machine learning algorithms Logistic Regression,Decision tree ,Kth Nearest Neighbor ,SVM Model to identify .....These models were selected to capture various data patterns and improve overall prediction accuracy.

The dataset utilized in this research consisted of different features of a car categorized as in a wide range of selling prices, features such as transmission, fuel,type, as well as brand, manufacturing date, seller type, and owners. Preprocessing of the dataset involved deleting all null values and categorical characters.They were converted into discrete values.With a percentage of 70 percentage and 30 percentage of the total data we divided the dataset into training and test data.The results obtained from applying various machine learning algorithms such as Logistic Regression, Decision Tree, Kth Nearest Neighbor, SVM Model, to predict car prices.Moreover,we used hidden values to evaluate our modeling. As a result, our model should anticipate a medium selling price for the Name:Maruti Celerio Green VXI year:2017,

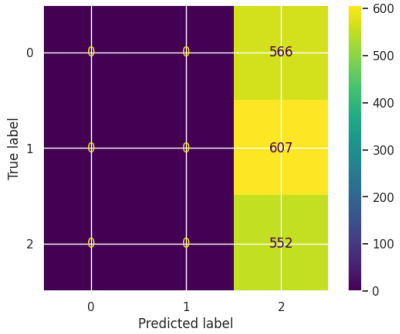
km-driven:78000, fuel:CNG, seller-type:Individual, transmission:Manual, Owner:First vehicle. Consequently, we utilized the KNN technique, and our model was properly predicted.

The Article titled "COMPARATIVE STUDY OF MACHINE LEARNING KNN, SVM, AND DECISION TREE ALGORITHM" discusses the use of machine learning algorithms included our application algorithm KNN,logistic regression,SVM,Decision tree. KNN's model representation is simple, as it stores the entire training dataset without requiring any explicit model learning. Predictions are made by comparing a new instance with the K most similar instances in the training data, using various distance measures like Euclidean, Hamming, or Manhattan.On the other hand, Logistic Regression is ideal for simple binary classification tasks, easy to interpret, computationally efficient, and can be regularized.Support Vector Machine (SVM) is versatile for both binary and multiclass classification, effective in high-dimensional spaces, and robust to outliers.Lastly,Decision Tree is highly interpretable, handles non-linear relationships, reveals feature importance, and forms the basis for powerful ensemble methods.For our work,we split the dataset into training and testing sets using the train-test-split method.Four models are trained on the training data, and predictions are made on the testing data.Model performance is evaluated using appropriate metrics. For Linear Regression, the coefficient of determination (R-squared) is calculated. Decision Tree and KNN models are optimized by varying hyperparameters (e.g., tree depth and number of neighbors) to achieve the best performance.Finally,The model with the highest predictive accuracy on the testing data is selected as the final model for car price prediction. A brief description of the models used is given below:

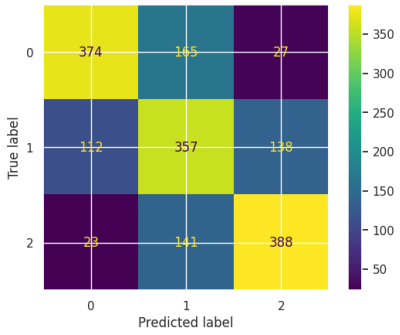
K-Nearest Neighbor (KNN) KNN calculates the Euclidean distance (diagonal distance) between the query point and the k-number of the nearest neighboring points. then decides on the class label based on the label with the highest frequency. This classifier, which uses supervised learning, examines closeness to predict the classification or grouping of a particular data piece. Our software imports data from Scikit-Learn module KNeighborsClassifier from sklearn.neighbors, and model train with 3 being the default value of k. This produces a poor accuracy score of 66.0 percentage before training on the dataset and 60.0 percentage after training on the dataset.



Support Vector Machine (SVM) The SVM algorithm's objective is to establish the optimal line or decision boundary that can divide n-dimensional space into classes so that we may quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary. SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors are the phrase for these extreme circumstances, and as a result, Support Vector Machine is the name of the algorithm. To train the model, we import SVC from sklearn.linear-model. When this is done, the accuracy on the dataset is 32 percentage, which is much less accurate than the KNN model.



Logistic Regression: A statistical model known as logistic regression, sometimes known as a supervised learning classifier, forecasts the likelihood that a binary event will occur on a given input dataset of independent variables. As a result, the dependent variable's output is a probability that ranges from 0 to 1 (inclusive). To train the model, we import LogisticRegression from sklearn.linear-model. When this is done, the accuracy on the dataset is 46 percentage, which is much less accurate than the KNN model.



Decision Tree A decision tree, a supervised learning method, may be used to tackle classification and regression problems. It is a classifier with a tree-like structure, where each leaf node represents the classification result, each internal node a feature, and each branch decision point. To train the model using the decision tree learning approach, we import DecisionTreeClassifier from sklearn.tree. This yields a score of 67.0 percentage for accuracy, which is greater than that of the Naive Bayes model but lower than that of the Logistic Regression model.

V. Results and Analysis

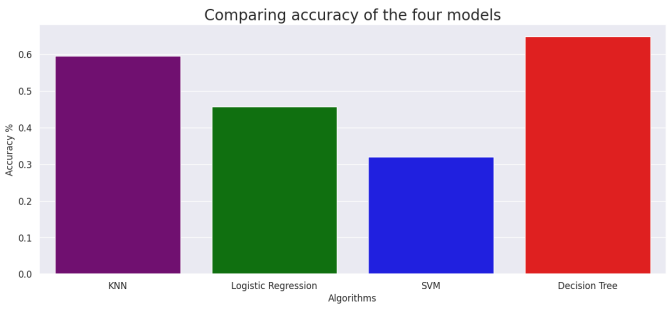
These various models: KNN,SVM,Logistic regression and Decision tree were used for checking the accuracy rate of our analysis.

The Decision Tree model clearly has the highest accuracy score (67.0 percentage) before training on the dataset, while the SVM model has the lowest accuracy score (32.0 percentage), according to demonstrations of the four models. Last but not least, the KNN model yields a score of 60.0 percentage, placing it in the middle of the Decision Tree and Logistic Regression models. In light of the available dataset, the Decision Tree model can forecast automobile prices the most accurately. Below, a bar chart that provides a clearer understanding of the outcomes serves as the visual depiction of the outcomes.

TABLE II
Model Comparison Analysis

Model Name	Accuracy (%)
Logistic Regression	46.0
Decision Tree	67.0
Kth Nearest Neighbor	60.0
SVM Model	32.0

Model selection/Comparison analysis: The Decision Tree model clearly has the highest accuracy score (67.0 percentage) before training on the dataset, while the SVM model has the lowest accuracy score (32.0 percentage), according to demonstrations of the four models. Last but not least, the KNN model yields a score of 60.0 percentage, placing it in the middle of the Decision Tree and Logistic Regression models. In light of the available dataset, the Decision Tree model can forecast automobile prices the most accurately. Below, a bar chart that provides a clearer understanding of the outcomes serves as the visual depiction of the outcomes.



These findings highlight the potential of machine learning techniques in predicting car prices. As per demonstrations of the four models, it is evident that the Decision Tree model produces the highest accuracy score (67.0 percentage) before training on the dataset, whereas the SVM model produces the lowest accuracy score (32.0 percentage). And lastly, the KNN model generates a score of 60.0 percentage, which sits between the Decision Tree and the Logistic Regression models. In conclusion, the Decision Tree model can best predict the price of car prices relative to the given dataset.However, further research and validation using larger datasets and additional evaluation metrics are necessary to ensure the reliability and generalizability of these models in real-world economical settings. Nevertheless, the results obtained from this study provide a valuable foundation for future research and practical applications of machine learning in predicting car pricing. The fusion of HPC and ML is not just an evolution; it is a revolution that will reshape industries, drive innovation, and redefine what is possible in the era of data-driven decision-making.

References

[1] Y. Etsion and D. Tsafrir, "A short survey of commercial cluster batch schedulers," *School of Computer Science and Engineering, Vol. 44221, The Hebrew University of Jerusalem*, pp. 2005–2013, 2005.

[2] T. Khan, W. Tian, G. Zhou, S. Ilager, M. Gong, and R. Buyya, "Machine learning (ml)–centric resource management in cloud computing: A review and future directions," *Journal of Network and Computer Applications*, 2022, article 103405.

[3] T. Kurth, S. Treichler, J. Romero, M. Mudigonda, N. Luehr, E. Phillips, A. Mahesh, M. Matheson, J. Deslippe, M. Fatica, P. Prabhat, and M. Houston, "Exascale deep learning for climate analytics," in *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis*, Nov 2018, pp. 649–660.

[4] P. Gepner, "Machine learning and high-performance computing hybrid systems, a new way of performance acceleration in engineering and scientific applications," in *2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2021, pp. 27–36.

[5] M. Happ, M. Herlich, C. Maier, J. Du, and P. Dorfinger, "Graph-neural-network-based delay estimation for communication networks with heterogeneous scheduling policies," *ITU Journal of Future Evolution Technology*, vol. 2, no. 4, 2021.

[6] J. Behrmann, P. Vicol, K.-C. Wang, R. Grosse, and J.-H. Jacobsen, "Understanding and mitigating exploding inverses in invertible neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2021, pp. 1792–1800.

[7] D. Zhang, D. Dai, Y. He, F. Bao, and B. Xie, "Rlscheduler: An automated hpc batch job scheduler using reinforcement learning," in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020, pp. 1–15.

[8] S. Salman, C. Streiffer, H. Chen, T. Benson, and A. Kadav, "Deepconf: Automating data center network topologies management with machine learning," in *Proceedings of the 2018 Workshop on Network Meets AI & ML*, 2018, pp. 8–14.

[9] A. Roy, J. Pachuau, and A. Saha, "An overview of queuing delay and various delay based algorithms in networks," *Computing*, pp. 2361–2399, 2021.

[10] K. Haghshenas, A. Pahlevan, M. Zapater, S. Mohammadi, and D. Atienza, "Magnetic: Multi-agent machine learning-based approach for energy efficient dynamic consolidation in data centers," *IEEE Transactions on Services Computing*, 2019.