Comparison of Different Machine Learning Models for Predicting the Price of Car

Abhiram Reddy Konda,
210962003
Department of Computer Science and Engineering
Manipal Institute of Technology,
Manipal Academy of Higher Education
Manipal,India
abhiramrk123@gmail.com

Nehal Chandan Murdeshwar,
210962021
Department of Computer Science and Engineering
Manipal Institute of Technology,
Manipal Academy of Higher Education
Manipal,India
nehalmurdeshwar@gmail.com

***Abstract*— This report presents a study on developing a machine learning-based predictive model for determining the price of used cars accurately, utilizing a dataset from Kaggle. This research focuses on evaluating a comprehensive set of vehicle features, such as age, make, model, mileage, and overall condition, to build a robust model. The primary objective is to identify the most effective machine learning method from techniques like Linear Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, and K-Nearest Neighbors (KNN), for price prediction within the confines of the provided dataset. The study emphasizes the importance of a predictive tool tailored to the static nature of the dataset, rather than adapting to real-time data, and aims to offer a user-friendly interface for individuals and businesses. By comparing different machine learning approaches on the Kaggle dataset, this research seeks to find the model that provides the highest accuracy and generalizability in a controlled data environment. Additionally, the study explores the development of a model that balances interpretability, transparency, scalability, and computational efficiency. The ultimate aim is to equip stakeholders in the used car market with a reliable, data-driven tool for car price valuation based on static historical data, aiding informed decision-making. The findings and advancements from this research are expected to contribute to the academic field of predictive analytics and offer practical insights for the automotive industry, particularly in the context of dataset-based model performance evaluation.**

Keywords*— *Predictive Analytics, Machine Learning, Used Car Valuation, Dataset Evaluation, Model Comparison

I. INTRODUCTION

In the dynamic world of automotive sales, the ability to accurately predict the price of a car remains a cornerstone for consumers, dealers, and manufacturers alike. With the surge in the production and subsequent resale of vehicles, the market for used cars has expanded significantly, creating a pressing need for reliable pricing models. Traditional methods of car valuation, which often rely on manual appraisal and rule-of-thumb assessments, are rapidly giving way to more sophisticated, data-driven approaches. This report delves into the realm of machine learning (ML), exploring how various models such as Linear Regression, K-Nearest Neighbors (KNN), Decision Trees, Support Vector Machines (SVM), and Random Forests can be harnessed to predict car prices with unprecedented accuracy.

The impetus for utilizing ML in car price prediction stems from the intricate interplay of factors that determine a vehicle's value. Unlike new cars, where the price is largely fixed by manufacturers and additional costs are transparently outlined by government taxes, the used car market is characterized by a myriad of influencing variables. Factors such as mileage, year of manufacturing, fuel consumption, transmission type, road tax, fuel type, engine size, and a host of other features each play a role in shaping the final price tag. The complexity of this pricing puzzle makes it an ideal candidate for ML models, which can simultaneously analyze multiple dimensions of data to uncover underlying patterns and relationships.

Linear Regression, one of the most elementary yet powerful ML techniques, serves as a foundational model for price prediction. By establishing a relationship between independent variables (car features) and the dependent variable (car price), Linear Regression can provide a baseline from which the value of a vehicle can be gauged. Its simplicity, however, is both a strength and a limitation; while it excels in capturing linear relationships, it may falter when faced with the non-linear intricacies that often manifest in real-world data.

To address the potential shortcomings of Linear Regression and to capture the nuanced dynamics of car pricing, K-Nearest Neighbors (KNN) offers a more flexible approach. KNN operates on the premise of similarity, predicting the price of a car based on the known prices of 'nearest' vehicles within the feature space. This model's reliance on localized data patterns allows it to adapt to complex, multi-faceted correlations between features and prices, providing a more granular perspective on valuation.

Decision Trees further expand the repertoire of ML models for car price prediction. By segmenting the dataset into subsets based on feature values, Decision Trees create a hierarchical structure of decisions that can lead to more accurate predictions. The intuitive nature of Decision Trees also offers transparency into the decision-making process, allowing users to understand which features are most influential in determining car prices.

For scenarios where the data's dimensionality and complexity require a more advanced approach, Support Vector Machines (SVM) become invaluable. SVMs are adept at managing high-dimensional data, using kernel functions to

facilitate the separation of data points into different price categories. This capability is particularly crucial when dealing with non-linear relationships that are not easily deciphered by other models.

In parallel, Random Forests offer a sophisticated ensemble method, leveraging the collective power of numerous decision trees to produce a consensus prediction. The strength of Random Forests lies in their capacity to mitigate errors from individual trees, resulting in a model that is not only accurate but also generalizable across diverse datasets. As this report unfolds, we will scrutinize each of these models, assessing their individual and collective capabilities in the context of car price prediction. Our investigation is driven by the overarching goal to equip stakeholders in the used car market with a reliable, analytical toolset that transcends traditional valuation methods, paving the way for a more informed, data-centric approach to vehicle sales and purchases.

As this report unfolds, we will scrutinize each of these models, assessing their individual and collective capabilities in the context of car price prediction. Our investigation is driven by the overarching goal to equip stakeholders in the used car market with a reliable, analytical toolset that transcends traditional valuation methods, paving the way for a more informed, data-centric approach to vehicle sales and purchases.

II. LITERATURE REVIEW

The literature on car price prediction, particularly in the used car market, demonstrates a growing reliance on machine learning (ML) techniques to address the complexities involved in accurately assessing vehicle value. The paper titled "Used Car Price Prediction Using Random Forest Algorithm" contributes to this evolving field by developing a model that predicts used car prices with considerable accuracy.

The necessity for such a model is driven by the rising prices of new cars and the financial constraints that compel consumers towards the used car market. This market has witnessed a substantial upsurge globally, necessitating a reliable system to determine a car's worth based on numerous factors such as mileage, year of manufacturing, fuel consumption, transmission, road tax, fuel type, and engine size. The study in question reports the development of a model that can output a relatively accurate price prediction based on these variables.

In this industry, where there is no standardized mechanism for calculating the retail price of used cars, the paper underscores the potential of ML to automate operations, enhance processes, and make judgments based on historical data. The researchers compared various ML algorithms, including Linear Regression, Lasso Regression, Support Vector Machine (SVM), and Random Forest, to identify the most effective approach. The final model, built using the Random Forest algorithm due to its superior R^2 score, demonstrates the power of ensemble learning in providing accurate predictions.

The Random Forest algorithm, aptly named for its method of constructing multiple decision trees on different data subsets and averaging them for improved predictive accuracy, is highlighted for its efficacy. It operates on the principle of bagging, where the predictions from each tree are combined, based on the majority votes, to produce the final output. This method not only improves prediction accuracy but also combats overfitting, a common issue in predictive modeling.

The study also emphasizes the need for rigorous data collection and preprocessing, critical steps that significantly affect the performance of ML algorithms. By employing multiple subsets of data, the Random Forest algorithm demonstrates a robust approach to handling complex datasets. It is an exemplary instance of a bagging strategy, known for its capability to deal with large amounts of data and its resilience against the overfitting of models.

The research provides a comprehensive view of ML's role in predicting used car prices, exploring the advantages and methodologies of various algorithms. It lays the groundwork for future advancements in the field, suggesting that the integration of more sophisticated ML techniques and a wider array of predictive factors could further refine the accuracy and reliability of such predictive models.

Adding to this, the complexity of the used car market is further illuminated by the multitude of factors that influence a vehicle's price. Advanced machine learning techniques like the Random Forest algorithm have proven to be highly effective in managing these factors due to their ability to process large datasets and the collaborative nature of multiple decision trees. This collaborative approach is key to the algorithm's success, as it ensures that the model benefits from the collective insights of diverse trees, each considering a different subset of features and data points, thus reducing the likelihood of overfitting and enhancing the model's generalizability.

Lasso Regression, also explored in the study, extends the capabilities of traditional linear regression models by introducing a penalty term that constrains the size of the coefficients. This regularization technique helps in producing a model that focuses on the most significant features, thereby simplifying the model and reducing the risk of overfitting. Such an approach is particularly beneficial when dealing with datasets with collinear features, as it selectively includes only the most predictive ones, thus improving the model's predictive accuracy.

Support Vector Machine (SVM) offers yet another approach, known for its excellence in finding the optimal hyperplane for classification and regression tasks. Its kernel functions allow for the transformation of the input space into higher dimensions where linear separation is possible, making it highly adaptable to varied data types. The SVM's ability to deliver robust predictions, even in high-dimensional spaces, makes it a valuable asset in car price prediction.

Decision Trees, often used in conjunction with Random Forest, offer interpretability that is unmatched by more complex models. Their straightforward, logical structure allows for an easy understanding of the decision-making process, shedding light on the most influential variables in determining car prices. Despite being surpassed by ensemble methods in terms of predictive performance, Decision Trees remain a critical tool for gaining insights into the factors that drive car valuation.

The convergence of findings across these studies validates the superiority of ensemble methods in creating models with greater predictive accuracy. These methods, by amalgamating the strengths of individual algorithms, have excelled in

classifying vehicles into distinct price categories, effectively addressing the challenges presented by intricate datasets.

The paper also highlights the importance of user-friendly interfaces that facilitate access to predictive analytics for users without technical expertise. An intuitive interface that simplifies the input process and delivers accurate price predictions can significantly increase transparency and trust in the used car market.

Lastly, the suggestion to integrate unsupervised learning techniques for identifying latent patterns and time series analysis to consider temporal trends points towards a more comprehensive approach to car price prediction. These techniques could lead to dynamic pricing models that are sensitive to temporal fluctuations in car valuation, marking a significant advancement in the field.

In sum, the synthesis of literature indicates a strong consensus around the effectiveness of ML models, particularly ensemble methods, for predicting used car prices. The evolution of the automotive industry, coupled with these advanced data-driven approaches, will undoubtedly continue to shape the future of car sales analytics, providing a sophisticated toolkit for navigating the intricate landscape of vehicle valuation.

III. Research Gaps and Objectives

A. Research Gaps

Since this topic is well documented and researched. As of now, there are so research gaps regarding the machine learning models that we have decided to implement. There well could be some research gaps present in more complex models or models pertaining to deep learning but we will not be covering those in our project hence we will not be documenting those research gaps.

B. Objectives

- *Model Comparison Focus:* To concentrate on a comparative analysis of various machine learning models, including Linear Regression, K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Decision Trees, and Random Forest, to determine which is most effective in predicting car prices.
- *Feature-Model Interaction Analysis:* To explore how different models handle the interplay of diverse features such as age, make, model, mileage, and condition in predicting car prices, and to assess their ability to capture complex interactions.
- *Data Quality and Representation:* To emphasize the collection and preprocessing of a comprehensive dataset that accurately represents various market segments and car types, serving as a robust foundation for model training and comparison.
- *Comparative Performance Evaluation:* To rigorously evaluate and compare the performance of the selected machine learning models in the context of car price prediction, identifying strengths and weaknesses of each approach.
- *Focus on Model Interpretability:* To prioritize the interpretability and transparency of each model, offering clear insights into how different features and their interactions influence predicted car prices.

IV. Methodology

In the proposed methodology of this research, the initial step involves acquiring a dataset from Kaggle. This dataset will serve as the foundation for the subsequent analytical processes. Following the data acquisition, a comprehensive exploratory data analysis (EDA) will be conducted. The purpose of this EDA is to uncover underlying trends and delineate the critical relationships among various features within the dataset. This phase is crucial for gaining insights into the dataset's structure and the interdependencies of its variables.

Subsequently, the dataset will undergo meticulous preprocessing and cleaning. This stage is essential to refine the data, ensuring that it is in an optimal format for analysis by the machine learning algorithms. Such preprocessing includes the handling of missing values, normalization of data, and encoding of categorical variables, among other tasks. This process is fundamental to enhancing the quality and reliability of the data, thereby facilitating more accurate model predictions.

Upon the completion of these preparatory steps, the dataset will be subjected to a series of machine learning algorithms. These algorithms include, but are not limited to, Linear Regression, K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Decision Trees, and Random Forest. Each of these models will be applied to the dataset, and their performance in predicting car prices will be rigorously evaluated.

The results obtained from these machine learning models will then be thoroughly analyzed in the subsequent section of the study. This analysis aims to compare and contrast the effectiveness of each algorithm in the context of car price prediction, identifying the most proficient model based on various performance metrics.

The architecture of the proposed methodology is systematically illustrated in Figure 1.

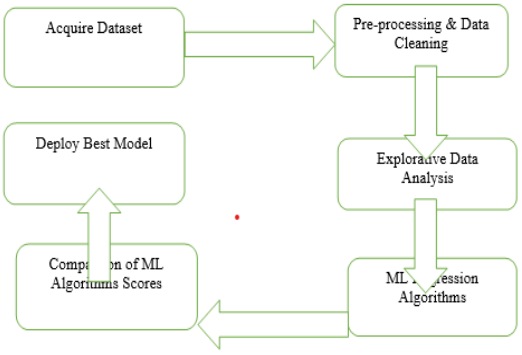

Fig 1. Architecture of Proposed Methodology

A. Sample Dataset

The dataset that we obtained from Kaggle[11] has the dimensions 11914 x 16 as shown below in Figure 2 and Figure 3 displays all the 16 features present in the dataset and the first 4 rows of data.

```
data = pd.read_csv("data.csv")
dimensions = data.shape

print(f'Dimensions of dataset is: {dimensions}')
```

```
Dimensions of dataset is: (11914, 16)
```

Fig 2. Dimensions of Dataset

Make	Model	Year	Engine Fu	Engine HP	Engine Cy	Transmiss	Driven_W	Number o	Market Ca	Vehicle Si	Vehicle St	highway M	city mpg	Popularity	MSRP
BMW	1 Series M	2011	premium	335	6	MANUAL	rear wheel	2	Factory Tu	Compact	Coupe	26	19	3916	46135
BMW	1 Series	2011	premium	300	6	MANUAL	rear wheel	2	Luxury,Pe	Compact	Convertib	28	19	3916	40650
BMW	1 Series	2011	premium	300	6	MANUAL	rear wheel	2	Luxury,Hig	Compact	Coupe	28	20	3916	36350
BMW	1 Series	2011	premium	230	6	MANUAL	rear wheel	2	Luxury,Pe	Compact	Coupe	28	18	3916	29450

Fig 3. Features present in the dataset.

B. Exploratory Data Analysis

In the Exploratory Data Analysis (EDA) section of the research, various visualizations were created to explore the dataset comprehensively. The EDA focuses on understanding the distribution and trends within the data, which is crucial for the subsequent modeling phase.

The first visualization is a count plot displaying the number of cars from each company. This plot provides a clear representation of the prevalence of different car manufacturers within the dataset, highlighting which companies have more vehicles represented. This information is pivotal in understanding market dominance and diversity in car manufacturing. The plot, with 'Car Company' on the y-axis and 'Number of Cars' on the x-axis, offers an intuitive understanding of the data's composition regarding car manufacturers.

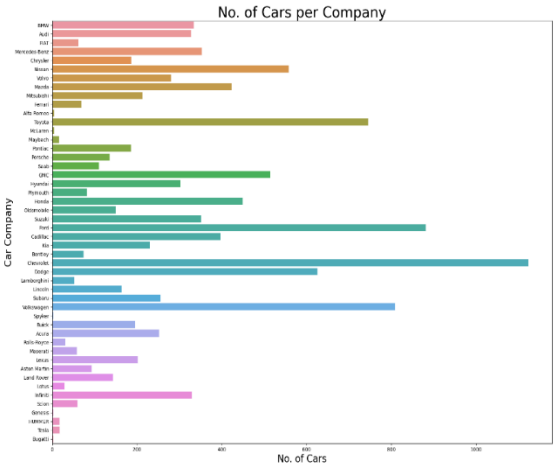

Fig 4. Count plot for cars per company

Next, the dataset was analyzed for the distribution of cars across different years. A count plot depicting the number of cars from various years reveals trends in the car industry over time, such as spikes in production or popularity of certain models in specific years. This plot, with 'Year' on the y-axis and 'Number of Cars' on the x-axis, aids in understanding the temporal dynamics of the car market.

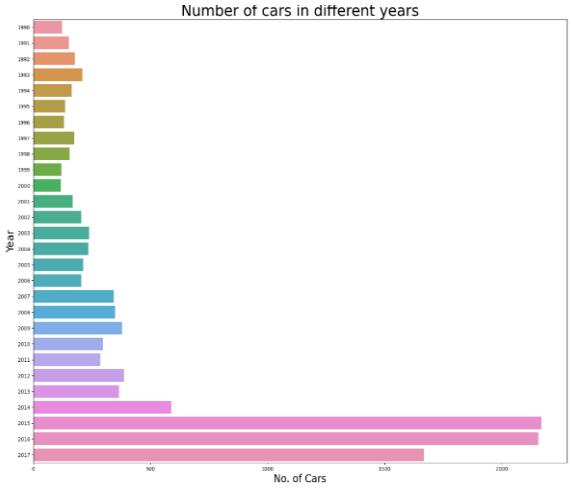

Fig 5. Count plot for no. of cars in different years

Further analysis was conducted on the transmission types of the cars in the dataset. The count plot for transmission types illustrates the distribution of cars based on their transmission, providing insights into the popularity and market availability of different transmission systems. This information is valuable for understanding consumer preferences and technological trends in the automotive industry.

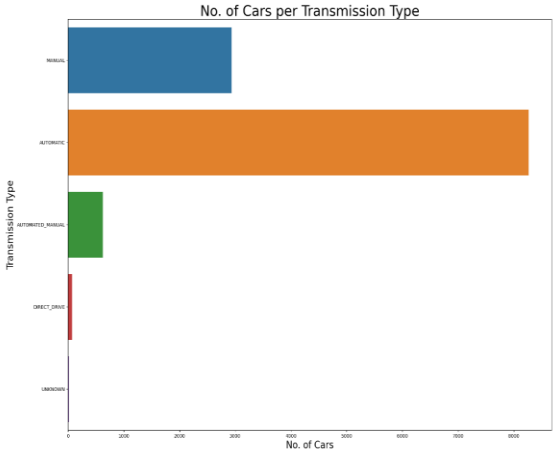

Fig 6. Count plot for no. of cars per transmission type

Lastly, a count plot was created to show the number of cars per vehicle size. This plot categorizes cars based on their size, offering an overview of the variety of vehicle sizes available in the market. This visualization helps in understanding market trends and consumer preferences regarding vehicle size.

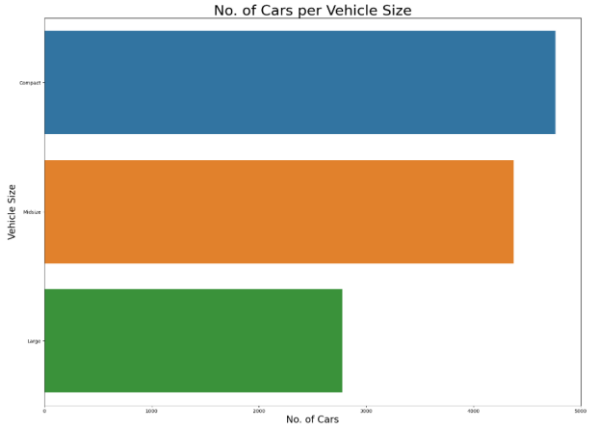

Fig 7. Count plot of no. of cars per vehicle size

These visualizations form a critical part of the EDA, as they offer a foundational understanding of the dataset's characteristics. This understanding is essential for the effective application of machine learning models in the later stages of the study.

C. *Preprocessing Data*

A series of comprehensive steps were meticulously undertaken to prepare the dataset for effective machine learning modeling. This phase of preprocessing is crucial in ensuring the quality and integrity of the data, which directly influences the accuracy and reliability of the predictive models.

a) Missing Data Analysis and Handling: The first step in the preprocessing involved a thorough analysis of missing data, utilizing the Missingno library to visualize missingness patterns. This process was vital to understand the extent and nature of missing data across different columns, facilitating informed decisions on handling such gaps. Strategies for missing data included imputation for columns with manageable levels of missing values and the removal of columns where missingness was too extensive to salvage.

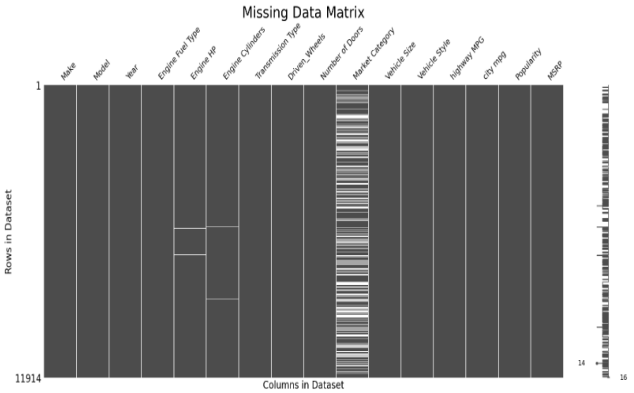

Fig 8.Visualisation of Missing Data per Feature

*b) Outlier Identification and Removal:*Outliers can severely affect model predictions. The dataset was scrutinized for outliers, particularly in crucial variables such as 'Highway MPG' and 'City MPG'. Advanced scatterplot visualizations were employed for this purpose along with boxplot visualisations.

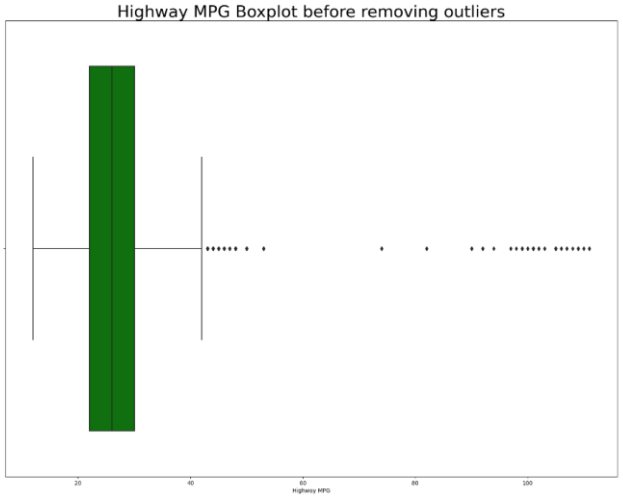

Fig 9. Boxplot of 'Highway MPG' showing outliers

The removal process involved using percentile-based thresholds to identify and eliminate data points that were significantly deviating from the norm. This refined the dataset, ensuring a more representative and balanced data set for model training.

```
The 90.0th percentile value is 35.00
The 91.0th percentile value is 36.00
The 92.0th percentile value is 36.00
The 93.0th percentile value is 37.00
The 94.0th percentile value is 37.00
The 95.0th percentile value is 38.00
The 96.0th percentile value is 39.00
The 97.0th percentile value is 40.00
The 98.0th percentile value is 42.00
The 99.0th percentile value is 46.00


The 99.0th percentile value is 46.00
The 99.1th percentile value is 46.00
The 99.2th percentile value is 48.00
The 99.3th percentile value is 48.00
The 99.4th percentile value is 50.00
The 99.5th percentile value is 85.52
The 99.6th percentile value is 97.35
The 99.7th percentile value is 101.00
The 99.8th percentile value is 103.35
The 99.9th percentile value is 107.09
```

```
#Boxplot after removing outliers

data = data[data['highway MPG'] < 60]
```

Fig 10. Detecting percentile-based thresholds and applying them to the dataset

c) Null Value Management: Null values were prevalent in several features. A targeted approach was adopted to address these null values, with methods tailored to the nature of each feature. Median values were used for numerical features like 'Number of Doors', while categorical features such as 'Engine Fuel Type' were filled with the most frequent values. This nuanced approach to null value treatment helped preserve data integrity and ensured that the dataset remained robust and comprehensive.

```
Make                  0
Model                 0
Year                  0
Engine Fuel Type      3
Engine HP            21
Engine Cylinders     20
Transmission Type     0
Driven_Wheels         0
Number of Doors       1
Market Category    3737
Vehicle Size          0
Vehicle Style         0
highway MPG           0
city mpg              0
Popularity            0
MSRP                  0
dtype: int64
```

Fig 11. Display no. of null values per feature

d) Feature Dropping: Features with a high percentage of missing values, like 'Market Category', were carefully evaluated and subsequently dropped from the dataset. This decision was made to enhance the overall quality of the dataset, as retaining such features could potentially introduce bias or inaccuracies in the model's predictions.

e) Data Shuffling and Splitting: Shuffling the dataset was a critical step, ensuring that the model training and testing phases were not biased by any pre-existing order in the data. The data was then split into an 80-20 ratio for training and testing sets, respectively. This split ensured that a significant portion of the data was used for model training, while still reserving an adequate subset for unbiased evaluation of the model's performance.

f) Categorical Data Encoding: The presence of categorical data necessitated the use of encoding techniques to convert these into a machine learning-friendly numerical format. Target encoding was implemented for features with numerous categories, ensuring that these features were accurately represented without inflating the feature space.

	Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	Number of Doors	Vehicle Size
1354	10812.757938	30176.543012	36784.190660	regular unleaded	140.0	4.0	AUTOMATIC	front wheel drive	4.0	Midsize
896	28423.023983	25245.937696	2558.613101	regular unleaded	185.0	4.0	MANUAL	front wheel drive	2.0	Compact
2635	28230.392090	5061.892819	2558.613101	regular unleaded	200.0	6.0	MANUAL	four wheel drive	2.0	Large
11165	196884.138144	97860.899828	46953.929157	premium unleaded (required)	430.0	8.0	MANUAL	rear wheel drive	2.0	Compact
2554	26660.798742	22687.787683	46953.929157	regular unleaded	143.0	4.0	AUTOMATIC	front wheel drive	4.0	Compact

Fig 12. Target encoding of Make,Model and Year

As shown in Fig.12, Make, Model and Year were converted to numerical values through target encoding as they had a lot of different categories.

One Hot Encoding was applied to other categorical variables, effectively transforming these into a format conducive for use in various machine learning algorithms.

Engine Fuel Type_1	...	Vehicle Style_7	Vehicle Style_8	Vehicle Style_9	Vehicle Style_10	Vehicle Style_11	Vehicle Style_12	Vehicle Style_13	Vehicle Style_14	Vehicle Style_15	Vehicle Style_16
1	...	0	0	0	0	0	0	0	0	0	0
1	...	0	0	0	0	0	0	0	0	0	0
1	...	0	0	0	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0	0	0	0
1	...	0	0	0	0	0	0	0	0	0	0

Fig 13. One hot encoding of Vehicle Style, Engine Fuel Type

g) Standardisation and Normalisation: To ensure that all features contribute equally to the model's predictions, standardization and normalization were applied. The MinMaxScaler was used to scale the features in the training and test sets, maintaining the range between 0 and 1. This scaling is crucial for models sensitive to the scale of data, such as KNN or SVM, ensuring that no single feature disproportionately influences the model's predictions.

D. Training Models

a) Linear Regression: The Linear Regression model, known for its simplicity and interpretability, was the first to be trained. The model was fit using the standardized and normalized training data (X_train_new, y_train). Predictions were then made on the test set (X_test_new). A comparison of the predicted values against the actual MSRP values was visualized using a regression plot. This plot provided a clear visual assessment of the model's predictive accuracy. Key error metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-Squared were calculated to quantify the model's performance.

b) K-Nearest Neighbors Regressor: The KNN model, which operates based on proximity to nearest data points, was then trained. The algorithm's reliance on feature scaling made the prior standardization process critical for its performance. Similar to the Linear Regression model, the KNN model was trained, predictions were made, and performance metrics were evaluated to understand its efficacy in predicting car prices.

c) Support Vector Regressor(SVM): SVR, known for its effectiveness in handling non-linear data, was applied next. The model was fit to the training data, and its performance was evaluated on the test set. SVR's complexity and robustness against overfitting made it a vital model to compare with the simpler Linear Regression and KNN models.

d) Decision Tree Regressor: The Decision Tree Regressor, which builds a tree-like model of decisions, was also employed. This model is particularly useful for its interpretability, as it makes decisions based on feature values. After training, the model's predictions were compared against the actual values, and its performance was assessed using the standard error metrics.

e) Random Forest Regressor: Finally, the Random Forest Regressor, an ensemble learning method known for its high accuracy and ability to run efficiently on large databases, was trained. By averaging the predictions of multiple decision trees, Random Forest tends to offer improved accuracy and robustness over a single Decision Tree.

V. Discussion and Analysis of Results

A. Linear Regression Analysis

The Linear Regression model showed varying degrees of accuracy in its predictions. While some predictions were close to the actual MSRP values as shown in Fig 14, others, like the notably lower prediction for the first car, indicated a

divergence from the expected prices. This variability suggested potential underfitting or model simplicity issues.

	Predicted Output	MSRP
0	11476.951720	24660
1	14522.146184	2000
2	70454.251781	49770
3	29525.276266	20875
4	307776.371569	284976

Fig 14. Predicted vs Actual Output table for Linear Regression

The regression plot for this model was particularly revealing, as shown in Fig 15, showing a spread of data points around the line of best fit. It highlighted the model's general trend in predicting car prices and areas where improvements could be made, especially in capturing the nuances of higher-priced or feature-rich vehicles.

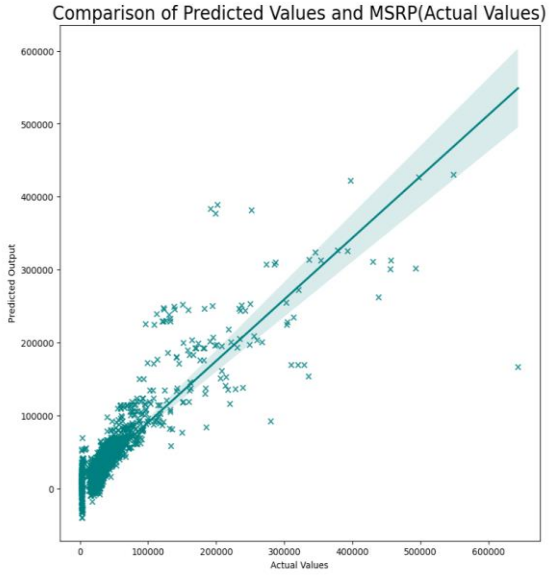

Fig 15. Regplot of Linear Regression Model

B. K-Nearest Neighbours Regressor Analysis

The KNN model's performance was a blend of precision and deviation. It excelled in predicting prices for cars at the lower and higher ends of the price spectrum but was less accurate for mid-range vehicles as shown in Fig 16.

	Predicted Output	MSRP
0	24416.0	24660
1	2039.5	2000
2	51845.0	49770
3	18685.0	20875
4	277078.0	284976

Fig 16. Predicted vs Actual Output for KNN Model

This pattern was visually depicted in the regplot, which illustrated the proximity-based prediction mechanism of KNN. The plot, as shown in Fig 17,showed areas where the model accurately captured the market trends and those where it struggled, potentially due to the influence of neighboring data points or the choice of 'k' value in the algorithm.

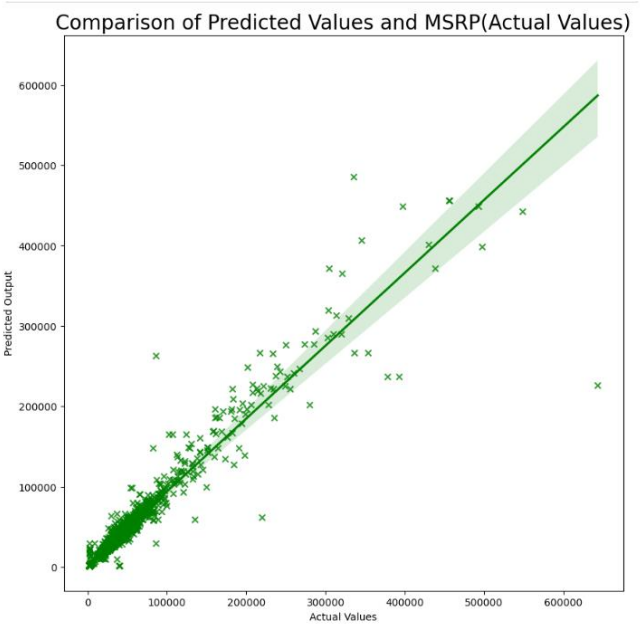

Fig 17. Regplot of K-Nearest Neighbours Model

C. Support Vector Regressor(SVM) Analysis

SVR predictions displayed a tendency towards consistent overestimation, particularly evident in high-value cars. This overestimation might reflect the model's sensitivity to outliers or the need for parameter tuning.

	Predicted Output	MSRP
0	29749.232558	24660
1	29560.908265	2000
2	30533.052210	49770
3	29157.826878	20875
4	30391.080446	284976

Fig 18. Predicted vs Actual Outputs for SVR Model

The regplot for SVR provided a comprehensive view of this trend across various price ranges, underscoring the need for model refinement to better capture the complex pricing structures of different car segments.

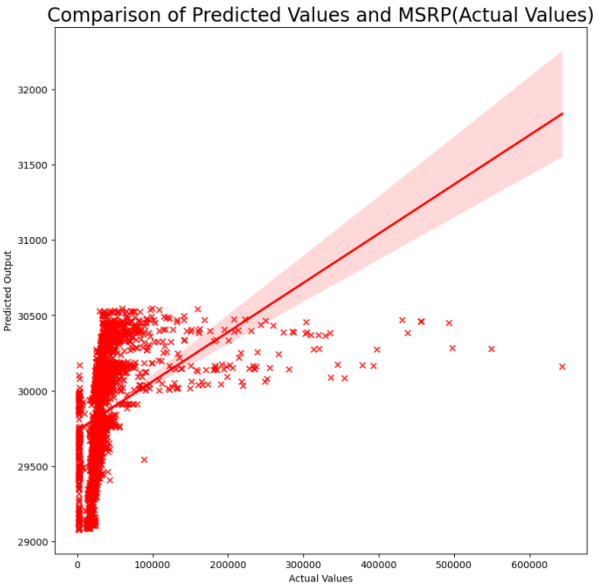

Fig 19. Regplot for SVR model

D. Decision Tree Regressor Analysis

The Decision Tree Regressor showed a high degree of accuracy in certain instances, particularly for lower-priced vehicles, but also significant deviations, especially at the upper end of the price spectrum. This variability in prediction accuracy could stem from the model's depth or the complexity of the decision rules it formed.

	Predicted Output	MSRP
0	24462.0	24660
1	2000.0	2000
2	50800.0	49770
3	16495.0	20875
4	275861.0	284976

Fig 20. Predicted vs Actual Output for Decision Tree Regressor Model

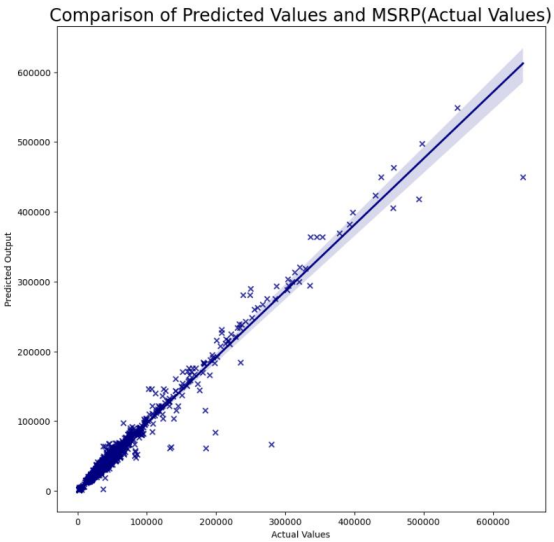

Fig 21. Regplot for Decision Tree Regressor Model

The corresponding regplot offered insights into the model's decision-making patterns, indicating areas of overfitting or oversimplification, especially in handling cars with a wide array of features or unique attributes.

E. Random Forest Regressor Analysis

The Random Forest Regressor exhibited a commendable level of accuracy, with most predictions closely aligning with the actual MSRP values. This model's ensemble approach effectively mitigated individual decision trees' biases and errors, resulting in a more balanced and accurate prediction across the board.

	Predicted Output	MSRP
0	24650.764603	24660
1	2000.000000	2000
2	51843.705820	49770
3	17179.268889	20875
4	276523.928667	284976

Fig 22. Predicted vs Actual Outputs for Random Forest Regressor Model

The regplot for this model was a testament to its robustness, displaying a tight clustering of predictions around the actual values, and highlighting its superiority in handling diverse data with varying features and price points.

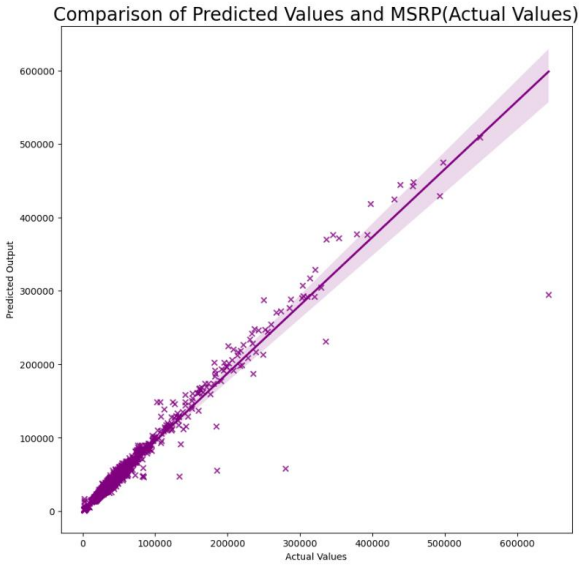

Fig 23. Regplot for Random Forest Regressor Model

F. Accuracy Analysis

a) Mean Absolute Error(MAE): The MAE values were visualized using a barplot as shown in Fig 24, providing a clear comparison of the average absolute errors across models. Linear Regression exhibited the highest MAE, suggesting a significant average deviation in its predictions. In contrast, the Decision Tree and Random Forest Regressors showed the lowest MAE values, indicating their predictions

were closer to the actual values on average. The KNN Regressor also performed well, with a relatively low MAE. This metric was crucial for understanding the average magnitude of errors in the models' predictions.

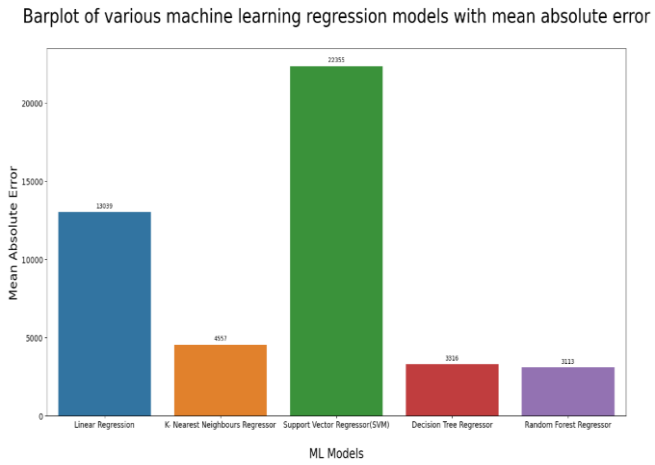

Fig 24. Barplot of MAE's of the different models.

b) Mean Squared Error(MSE): The MSE values, representing the average of the squares of the errors, were similarly displayed in a barplot.

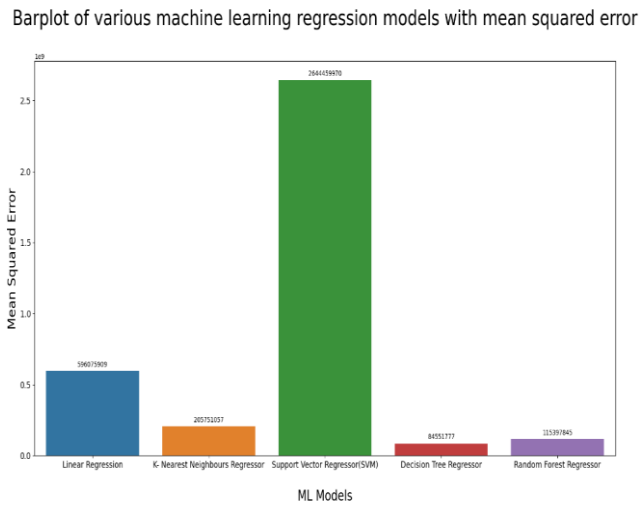

Fig 25. Barplot of MSE's of the different models.

Here, the SVR model displayed a significantly higher MSE, indicating larger squared deviations in its predictions, possibly due to its tendency to overfit. The Decision Tree and Random Forest models showed much lower MSE values, reflecting their higher accuracy and consistency in predictions.

c) Root Mean Squared Error(RMSE): RMSE, providing an understanding of the error magnitude on the same scale as the data, was also analyzed through a barplot. The SVR's high RMSE value reaffirmed its comparatively lower accuracy. The Decision Tree and Random Forest models maintained their lead with the lowest RMSE values, underscoring their precision in predicting car prices.

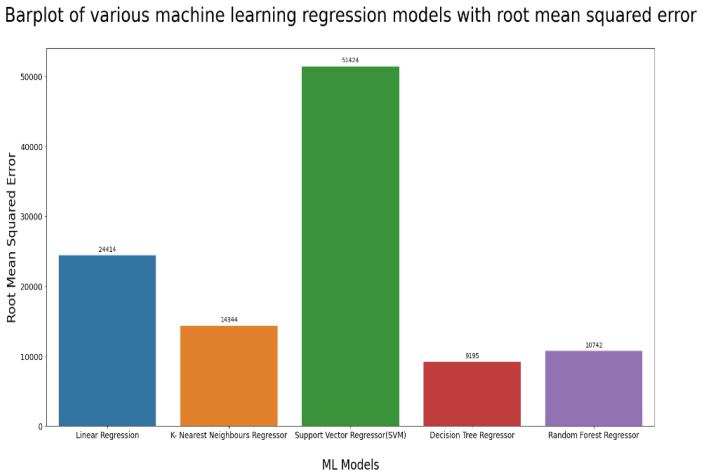

Fig 26. Barplot of RMSE's of different models.

d) R2 Score: The R2 Score, a measure of how well the variations in the dependent variable are explained by the model, was the final metric analyzed. The barplot for R2 Scores revealed that the Decision Tree Regressor had the highest score, closely followed by the Random Forest Regressor, indicating their superior computational power. The SVR model's negative R2 Score highlighted its poor performance in this dataset.

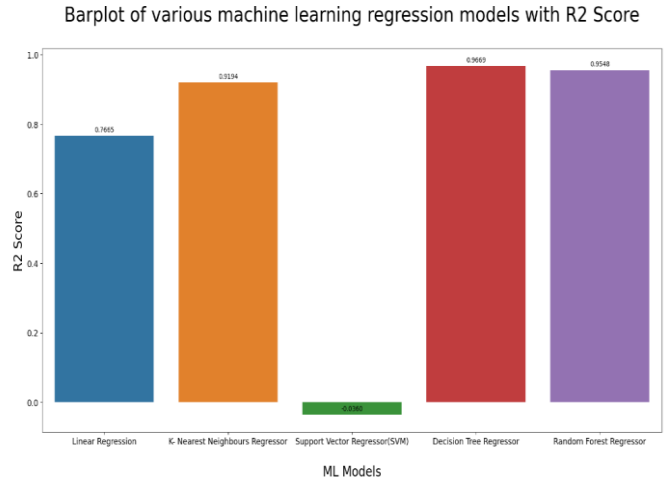

Fig 27. Barplot of R2 Score's of the different models.

e) Final Table: Fig 28. shows the final perfomance metrics calculated on the different models. We can look at this and choose decision tree to be the best possible model for this dataset.

	Models	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R2 Score
0	Linear Regression	13039	596075909	24414	0.766485
1	K- Nearest Neighbours Regressor	4557	205751057	14344	0.919396
2	Support Vector Regressor(SVM)	22355	2644459970	51424	-0.035979
3	Decision Tree Regressor	3316	84551777	9195	0.966876
4	Random Forest Regressor	3113	115397845	10742	0.954792

Figure 28. Final Performance Metrics Table

VI. Conclusion and Future Work

The study revealed significant differences in the performance of these models when applied to the task of predicting car prices. The Decision Tree and Random Forest Regressors emerged as the most accurate models, as evidenced by their low Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), along with high R2 Scores. These models proved particularly effective in capturing the complexities and nuances of the dataset, offering predictions that closely mirrored actual car prices. The Linear Regression model, while useful for its simplicity and interpretability, demonstrated limitations in prediction accuracy, as did the SVR model, which was particularly prone to overfitting as indicated by its negative R2 Score.

This project contributes valuable insights into the field of machine learning applied to car price prediction. It underscores the importance of choosing the right model based on the specific characteristics of the dataset and the predictive task at hand. The findings also highlight the potential of ensemble methods, like the Random Forest Regressor, in dealing with complex, real-world datasets where multiple factors influence the outcome variable.

Furthermore, this study lays the groundwork for future research in this area, suggesting areas for further exploration such as feature engineering, hyperparameter tuning, and the use of more advanced machine learning techniques. The methodology and findings of this research have significant implications for stakeholders in the automotive industry, including manufacturers, dealers, and buyers, by providing a more nuanced understanding of the factors influencing car prices and enhancing the accuracy of price predictions.

In terms of future work, this study opens several avenues for further research and development in the domain of predictive modeling for car prices. One potential area of exploration is the integration of more advanced machine learning and deep learning techniques, such as neural networks or ensemble methods that combine multiple algorithms for enhanced prediction accuracy. These techniques could potentially capture more complex patterns and interactions within the data that are not fully harnessed by traditional models.

Another promising direction is the incorporation of real-time data and more dynamic features into the models. This could involve using data that reflects current economic trends, consumer preferences, and evolving automotive technology, thereby making the models more responsive to the rapidly changing automotive market. Additionally, further research could focus on improving the interpretability of complex models, ensuring that the results are not only accurate but also understandable and actionable for industry practitioners.

Expanding the dataset to include more diverse geographical regions and a broader range of vehicle types, including electric and hybrid cars, could also provide more comprehensive insights and enhance the generalizability of the models. Finally, the development of user-friendly applications and tools based on these predictive models could be explored, making advanced analytics accessible to a wider range of users in the automotive sector.

References

[1] V.C. Sanap, M. Munawwar Rangila, Sufiyaan Rahi, Samiksha Badgujar, Yashodhan Gupta, "Car Price Prediction using Linear Regression Technique of Machine Learning," in International Journal of Innovative Research in Science Engineering and Technology, vol. 11, no. 4, April 2022

[2] D. A. Gaikwad, P. S. Suwarnakar, Y. R. Mahajan, A. U. Petkar, S. G. Theurkar, "Used Car Price Prediction Using Random Forest Algorithm," in International Journal for Multidisciplinary Research (IJFMR), vol. 5, no. 3, May-June 2023.

[3] A. D. Sharma and V. Sharma, "USED CAR PRICE PREDICTION USING LINEAR REGRESSION MODEL," in International Research Journal of Modernization in Engineering Technology and Science, vol. 2, no. 11, November 2020.

[4] P. Hasan Putra, Azanuddin, B. Purba, Y. A. Dalimunthe, "Random Forest and Decision Tree Algorithms for Car Price Prediction," in Jurnal Matematika Dan Ilmu Pengetahuan Alam LLDikti Wilayah 1 (JUMPA), vol. 3, no. 2, 2023, pp. 81-89, LLDIKTI WILAYAH 1.

[5] A. Amalia, M. Radhi, S. H. Sinurat, D. R. H. Sitompul, and E. Indra, "Prediksi Harga Mobil Menggunakan Algoritma Regressi Dengan Hyper-Parameter Tuning," Jurnal Sistem Informasi Dan Ilmu Komputer Prima (JUSIKOM PRIMA), vol. 4, no. 2, pp. 28–32, 2022

[6] A. Chandak, P. Ganorkar, S. Sharma, A. Bagmar, and S. Tiwari, "Car Price Prediction Using Machine Learning," International Journal of Computer Sciences and Engineering, vol. 7, no. 5, pp. 444–450, 2019, https://doi.org/10.26438/ijcse/v7i5.444450.

[7] K. K. Dutta, S. A. Sunny, A. Victor, A. G. Nathu, M. Ayman Habib, and D. Parashar, "Kannada alphabets recognition using decision tree and random forest models," Proceedings of the 3rd International Conference on Intelligent Sustainable Systems ICISS 2020, pp. 534–541, 2020.

[8] B. Saireddy, A. Vamshikrishna, G. Abhilash, and D. Vinith Srinivas, "Car price prediction using machine learning," 2022.

[9] K. Samruddhi and R. A. Kumar, "Used car price prediction using K-Nearest Neighbor based model," International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE), vol. 4, no. 3, pp. 686-689, DOI: 10.29027/IJIRASE.v4.i3.2020.686-689, September 2020

[10] S. E. Viswapriya, D. S. Sandeep Sharma, and G. Sathya Kiran, "Vehicle price prediction using SVM techniques," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 9, no. 8, June 2020

[11] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, "Car price prediction using machine learning techniques," TEM Journal, vol. 8, no. 1, pp. 113-118, DOI: 10.18421/TEM81-16, February 2019

[12] Cooper Union, "Car Features and MSRP Dataset," Kaggle, 2023. Available: https://www.kaggle.com/datasets/CooperUnion/cardataset. [Accessed:].

[13] A. Pandey, V. Rastogi, and S. Singh, "Car's selling price prediction using random forest machine learning algorithm," 2023

[14] Analytics Vidhya, "Car Price Prediction: Machine Learning vs Deep Learning," Analytics Vidhya, July 2021. Available: https://www.analyticsvidhya.com/blog/2021/07/car-price-prediction-machine-learning-vs-deep-learning/. [Accessed: 9-11-2023].

[15] The Clever Programmer, "Car Price Prediction with Machine Learning," The Clever Programmer, August 2021. Available: https://thecleverprogrammer.com/2021/08/04/car-price-prediction-with-machine-learning/. [12-11-2023].

[16] ODSC - Open Data Science, "Predicting Car Prices Using Machine Learning and Data Science," Medium, 2021. Available: https://medium.com/odscjournal/predicting-car-prices-using-machine-learning-and-data-science-52ed44abab1b. [Accessed: 12-11-2023].