

A MACHINE LEARNING MODEL FOR CAR PRICE PREDICTION

Pavitra Vaishnavi¹, Singireddy Patak Madhav Reddy², Shaik Sammer³, Nakka Rajeev⁴,
M. Santhosha⁵

^{1,2,3,4}Student, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100, India.

⁵Professor, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100, India.

DOI: <https://www.doi.org/10.58257/IJPREMS39481>

ABSTRACT

In recent years, the rising cost of new cars has made it increasingly difficult for many individuals to afford them. As a result, the demand for used cars has significantly grown in India and across the world. Accurately determining the fair price of a used car is essential for both buyers and sellers to ensure transparency and value for money. This paper aims to develop a ML model using Linear Regression, Decision Tree Regressor, Random Forest Regressor to predict the price of used cars in India. The model takes into account various factors such as the model name, year of purchase, fuel type, and number of miles driven to estimate the car's value. By analyzing these features, the model provides a data-driven and reliable price prediction, helping consumers make informed decisions. The implementation of this project demonstrates the effectiveness of data analysis and predictive modeling in solving real-world problems. It offers a practical solution for car dealerships, individual buyers, and sellers to determine a car's market value accurately.

Keywords: Machine Learning, Linear Regression, Decision Tree Regressor, Random Forest Regressor, Predictive Modeling, Car Valuation, Fuel Type, Mileage, Year of Purchase, Model Name, Market Value Estimation, Price Transparency.

1. INTRODUCTION

Determining the price of a used car has traditionally been challenging due to reliance on manual estimation methods such as dealership assessments, newspaper listings, and personal recommendations. These methods lacked accuracy, transparency, and consistency, often leading to undervaluation for sellers and overpricing for buyers. Additionally, valuations varied between dealerships due to differing business strategies, making the process unreliable. With rising new car prices, the used car market has expanded, making accurate price estimation essential. Traditional methods face limitations such as lack of standardization, real-time updates, and human bias. Car valuation depends on various interrelated factors, including brand, model, year, fuel type, mileage, and additional features. Advancements in Machine Learning and Big Data have revolutionized car price prediction, making it more systematic and data-driven. This project aims to develop a Machine Learning model to predict used car prices based on key attributes such as company name, model, year, fuel type, and kilometers driven. The model uses Linear Regression, Decision Tree Regressor, and Random Forest Regressor. Linear Regression identifies relationships between variables, Decision Tree Regressor handles complex patterns, and Random Forest Regressor improves accuracy through ensemble learning.

The dataset includes historical car sales data, processing key factors like mileage, ownership records, and market trends. The model eliminates human bias, ensuring fair market valuations and reducing inefficiencies in price negotiations. Buyers and sellers benefit from a transparent, data-driven approach to car valuation.

According to the National Transport Authority (2014), registered cars in Mauritius increased by 254% from 2003 to 2014, reflecting growing demand for used vehicles. Key resale factors include manufacturing year, make, model, mileage, fuel efficiency, and condition. Additional attributes such as safety features, accident history, and modifications also impact value. This project explores neural networks for price prediction and compares them with traditional models like Linear Regression and Support Vector Regression. The study follows a structured approach, analyzing historical data to train models, compare their accuracy, and evaluate real-world effectiveness. By integrating Machine Learning with extensive market data, this project enhances pricing transparency, efficiency, and accuracy in the used car market. It provides a scalable, unbiased, and automated solution that benefits consumers, dealerships, and online marketplaces, reshaping the automotive resale industry.

2. LITERATURE SURVEY

The National Transport Authority [1] provides statistical data on transportation trends in Mauritius, including vehicle registrations, road usage, and licensing patterns. Their reports serve as a valuable resource for researchers and policymakers analyzing vehicle ownership trends and infrastructure growth. Pudaruth, S. [2] applied machine learning techniques to predict used car prices in the International Journal of Information & Computation Technology. The study utilized regression models, decision trees, and support vector machines (SVM) to estimate prices based on attributes

like year, mileage, and fuel type, highlighting AI's advantages over traditional valuation methods. Jassibi, J., Alborzi, M., and Ghoreshi, F. [3] developed an AI-based system for car paint thickness control using artificial neural networks (ANNs) and regression techniques. Published in the Journal of Industrial Engineering International, their study enhances quality control in automobile manufacturing. Listiani, M. [4] explored Support Vector Regression (SVR) for car leasing price prediction in an MSc thesis at Hamburg University of Technology. The study demonstrates SVR's effectiveness in handling non-linear relationships in pricing models, optimizing rental costs, and improving financial decision-making. Iseri, A. and Karlik, B. [5] proposed an ANN-based approach for automobile pricing in the Expert Systems with Applications: ScienceDirect Journal of Informatics. Their model analyzes vehicle attributes such as mileage and brand, achieving high accuracy in price estimation compared to traditional statistical methods. Yeo, C.A. [6] investigated the role of neural networks in automobile insurance pricing in the Encyclopedia of Information Science and Technology. The study explores how ANNs model insurance premiums based on driver history, vehicle specifications, and accident records, improving risk assessment and pricing fairness. Doganis, P., Alexandridis, A., Patrinos, P., and Sarimveis, H. [7] introduced an ANN-based sales forecasting model for short shelf-life food products in the Journal of Food Engineering. While focused on the food industry, their findings demonstrate AI's broader applicability to demand forecasting, including automotive sales. Rose, D. [8] examined neural networks for predicting car production rates in a technical thesis at TARDEC. The study highlights AI's role in optimizing manufacturing efficiency, production planning, and adapting to market demand fluctuations. Lexpress.mu (2014) [9] and Le Defi Media Group (2014) [10] are leading news platforms in Mauritius, providing real-time updates on transportation trends, economic affairs, and automotive industry developments. These sources offer essential data for studying vehicle registration trends and pricing models. He, Q. [11] explored neural networks in information retrieval (IR) in a BSc thesis at the University of Illinois. The study examines AI's role in improving search efficiency and extracting pricing insights from automotive datasets. Cheng, B., and Titterington, D.M. [12] provided a statistical review of neural networks in Statistical Science, discussing their mathematical foundations, performance evaluation, and applications in predictive analytics. Anyaeche, C.O. [13] compared linear regression and neural networks for performance forecasting in the African Journal of Engineering Research. The study demonstrates that ANNs outperform regression models in handling complex datasets, reinforcing their effectiveness in used car price prediction. Bharambe, M.M.P., and Dharmadhikari, S.C. [14] studied stock market prediction using ANNs and big data at the Fourth Post Graduate Conference in Pune, India. Their research highlights AI's ability to capture market patterns, reducing investment risks. Ahangar, R.G., Mahmood, and Y, Hassen P.M. [15] conducted a comparative study of ANNs versus linear regression for stock price prediction in the International Journal of Computer Science and Information Security. Their findings illustrate ANN's superiority in handling complex financial data, which has implications for predictive modeling in various industries, including car price estimation.

Table .1. Literature Survey

Study	Key Contribution	Year
National Transport Authority	Provides transportation statistics for Mauritius.	2015
Pudaruth, S.	Used ML (regression, decision trees, SVM) for car price prediction.	2014
Jassibi, J., Alborzi, M., & Ghoreshi, F.	Developed AI for car paint thickness control.	2011
Listiani, M.	Applied SVR for car leasing price prediction.	2009
Iseri, A. & Karlik, B.	Proposed ANN-based car pricing model.	2009
Yeo, C.A.	Used ANNs for automobile insurance pricing.	2009
Doganis, P. et al.	Developed ANN-based sales forecasting model.	2006
Rose, D.	Used ANNs to predict car production rates.	2003
Lexpress.mu	Reports on transportation and automotive trends.	2014
Le Defi Media Group	Covers vehicle registration and pricing trends.	2014
He, Q.	Studied ANNs in information retrieval.	1999
Cheng, B. & Titterington, D.M.	Reviewed statistical foundations of ANNs.	1994
Anyaeche, C.O.	Compared regression and ANNs for forecasting.	2013
Bharambe, M.M.P. & Dharmadhikari, S.C.	Used ANNs for stock market prediction.	2015
Ahangar, R.G. et al.	Compared ANNs vs. regression for stock prices.	2010

3. METHODOLOGY

The methodology for this research focuses on leveraging machine learning techniques, statistical analysis, and data-driven approaches to analyze and predict price of a car. The process involves several key steps, including data collection, data cleaning, Exploratory data analysis- feature selection using recursive feature elimination, model selection, and evaluation.

3.1. Data Collection

It involves gathering relevant details about used cars from various sources such as online marketplaces, dealerships, and user-generated datasets. The dataset typically includes attributes like Car Name, Location, Year, Kilometers Driven, Fuel Type, Transmission, Owner Type, Mileage, Engine Capacity, Power, Seats, and Selling Price. High-quality, diverse, and well-structured data is crucial for building an accurate machine-learning model.

3.2. Data Cleaning

Data cleaning is crucial in car price prediction to handle inconsistencies, missing values, duplicates, and outliers. Since data comes from multiple sources like **L'Express** and **Le Defi**, discrepancies are common. Missing values in numerical features (e.g., mileage, engine capacity) are filled using mean, median, or mode imputation, while categorical features (e.g., paint type, transmission) use mode imputation. Duplicate listings are identified and removed based on attributes like make, model, year, mileage, and price. Outliers, such as unrealistic car prices, are detected using boxplots, Z-score analysis, and IQR methods. Categorical features are converted into numerical values using one-hot or label encoding. Finally, numerical features are standardized using Min-Max Scaling or Standardization for balanced model training. These steps ensure a structured and consistent dataset for analysis.

3.3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) helps uncover patterns and relationships in the dataset using statistical methods and visualizations. Histograms and boxplots reveal the distribution of car prices, mileage, and engine capacity, with log-scaling applied for skewed data. A correlation heatmap identifies relationships, such as the negative correlation between mileage and price, aiding in feature selection. Scatter plots and line charts visualize pricing trends based on factors like year, mileage, fuel type, and engine size. Outliers are detected using boxplots to prevent misleading model training. **Feature selection using Recursive Feature Elimination (RFE)** improves model accuracy and efficiency by identifying the most relevant attributes. A machine learning model first assigns importance scores to features, and less significant ones are removed iteratively. The process continues until the most predictive attributes, such as mileage, engine capacity, year, and transmission type, remain while eliminating redundant factors like paint color. This ensures a streamlined and effective predictive model.

3.4. Data Modeling and Evaluation

After feature selection, machine learning models are trained using an 80-20 train-test split to evaluate performance. Data modeling follows thorough cleaning, EDA, and RFE to identify key factors influencing car prices. The correlation matrix helps interpret relationships between variables.

Correlation strength ranges from -1 to 1. A value near +1 indicates a strong positive relationship, meaning as one variable increases, so does the other. A value near -1 shows a strong negative correlation, where one variable decreases as the other increases. A value near 0 suggests little or no linear relationship. These insights refine model accuracy and feature selection.

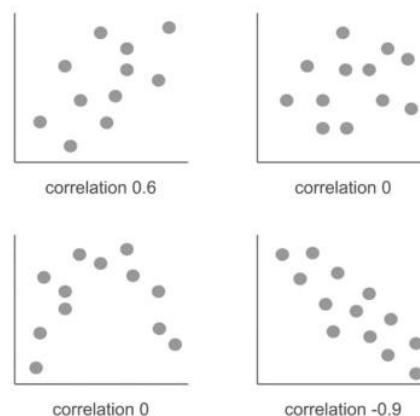


Fig. 1 Visualization of correlation

The correlation coefficient is calculated using the formula below:

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

R2 is a measure of the goodness of fit of a model. In regression, the R2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An R2 of 1 indicates that the regression predictions perfectly fit the data. Three regression models were implemented: Linear Regression, Decision Tree Regressor, and Random Forest Regressor.

3.4.1 Linear Regression

Linear Regression models the relationship between car price and features like mileage, engine size, and horsepower by fitting a straight line or hyperplane. It assumes a linear relationship and optimizes coefficients using the Ordinary Least Squares (OLS) method to minimize prediction errors. The formula for the regression equation is:

$$y = a + bx \quad (2)$$

$$b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} \quad (3)$$

$$a = \frac{\sum y - b(\sum x)}{n}$$

3.4.2 Decision Tree Regressor

Decision Tree Regressor predicts car prices by recursively splitting data based on the best feature and split point. It captures non-linear relationships well, achieving a test accuracy of 0.8840, outperforming Linear Regression. However, it may overfit, making it less reliable for generalization. Ensemble methods like Random Forest help mitigate this issue.

3.4.3 Random Forest Regressor

Random Forest Regressor improves Decision Trees by averaging predictions from multiple trees, reducing variance and overfitting. It achieved the highest test accuracy of 0.9066 in car price prediction, handling complex interactions and missing data effectively. Though robust and stable, it requires more computation than a single Decision Tree.

4. RESULT ANALYSIS AND DISCUSSION

The dataset "CarPrice_Assignment.csv" is used for car price prediction and is loaded into a pandas DataFrame. It contains various attributes describing cars, including "car_ID" (unique identifier), "symboling" (risk rating), and "CarName" (model name). Key specifications include "fueltype" (gas/diesel), "aspiration" (standard/turbo), and "doornumber" (two/four doors). Structural details such as "carbody" (sedan, hatchback, etc.), "drivewheel" (rwd, fwd, 4wd), and "engine location" (front/rear) are also present. Numerical attributes like "wheelbase," "carlength," "carwidth," and "carheight" contribute to price estimation. For predictive modeling, steps like data cleaning, EDA, and feature engineering are crucial. Identifying correlations and handling missing values will help in training a machine learning model for car price prediction.

```
In [2]: data = pd.read_csv("C:\MAJOR PROJECT\Car Prices Prediction\Dataset\CarPrice_Assignment.csv")

In [3]: data.head()

Out[3]:
```

	car_ID	symboling	CarName	fueltype	aspiration	doornumber	carbody	drivewheel	engine location	wheelbase	carlength	carwidth	carheight
0	1	3	alfa-romero giulia	gas	std	two	convertible	rwd	front	88.8	168.8	64.1	48.8
1	2	3	alfa-romero stelvio	gas	std	two	convertible	rwd	front	88.8	168.8	64.1	48.8
2	3	1	alfa-romero Quadrifoglio	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4
3	4	2	audi 100 ls	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3
4	5	2	audi 100ls	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3

Fig. 2. Data Loading

The Seaborn bar plot visualizes car brand frequencies, with Toyota being the most common, followed by Nissan, Mazda, Mitsubishi, and Honda. Less common brands include Mercury, Renault, and Alfa-Romeo. The x-axis represents brand counts, while the y-axis lists brand names. This analysis helps understand brand distribution and its impact on car prices. The distribution plot of car prices, created with Seaborn's distplot(), shows most prices below \$20,000, peaking around \$10,000. The data is right-skewed, indicating fewer high-priced cars. Normalization techniques like log transformation may be needed for better model performance. High-priced outliers might require robust regression models or filtering. Further steps involve feature engineering and training machine learning models for accurate price prediction.

Visualise different car names

```
df=pd.DataFrame(data['CarName'].value_counts().reset_index().rename(columns={'index':'car_name','CarName':'count'}))
plot = sns.barplot(y='car_name',x='count',data=df)
plt=plt.set(plot.get_xticklabels(), rotation=90)
```

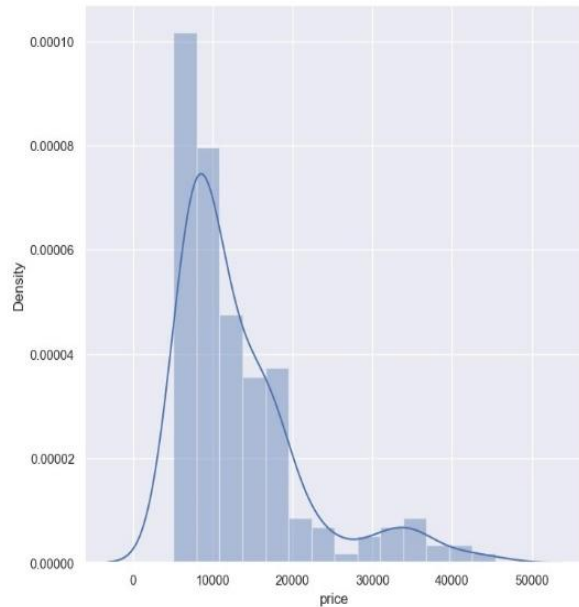
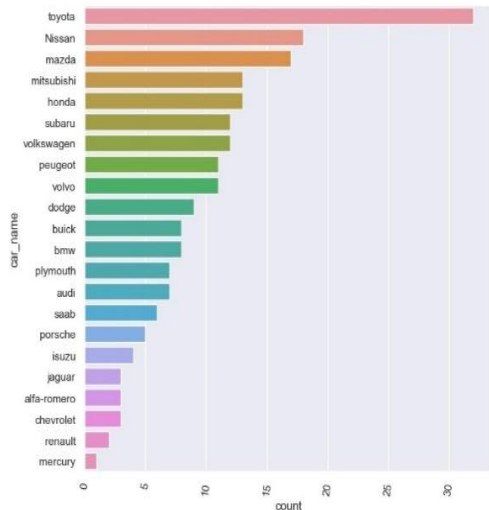
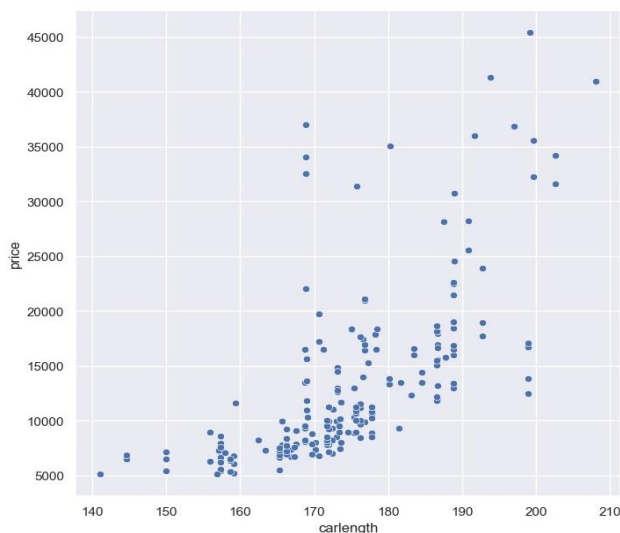


Fig. 3. List of Cars and Details of Car Price

The scatter plots demonstrate a positive correlation between car dimensions and price. Smaller cars (below 170 units in length and 64–66 units in width) are generally priced between \$5,000 and \$15,000. As length exceeds 190 units and width surpasses 70 units, car prices often rise above \$30,000, indicating they belong to premium or luxury categories. This suggests that longer and wider cars are associated with higher-end models, offering more space, comfort, and advanced features. These insights emphasize the importance of car dimensions in price prediction models. Incorporating length and width into machine learning models, along with other attributes like engine size, horsepower, and brand reputation, can improve prediction accuracy.

Carlength vs Car price

```
sns.scatterplot(x="carlength", y="price", data=data,color='b')
0]: <AxesSubplot:xlabel='carlength', ylabel='price'>
```



Car width Vs Price

```
sns.scatterplot(x="carwidth", y="price", data=data,color='b')
1]: <AxesSubplot:xlabel='carwidth', ylabel='price'>
```

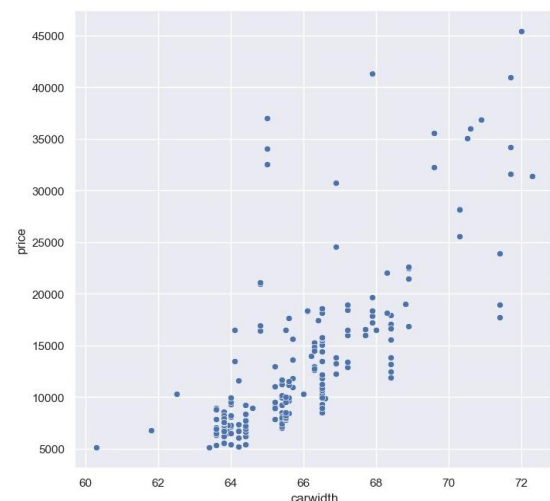


Fig. 4. Scatter plot for car length vs car price and car width vs car price

The correlation matrix highlights key factors influencing car prices. Engine size (0.83) and horsepower (0.81) are the strongest predictors, followed by curb weight (0.76), car width (0.75), and length (0.68), indicating that larger, heavier, and more powerful cars tend to be more expensive. Moderate correlations exist for bore ratio (0.55) and wheelbase (0.57).

Conversely, fuel efficiency metrics like city MPG (-0.68) and highway MPG (-0.70) show negative correlations, meaning fuel-efficient cars are generally cheaper. Peak RPM (-0.08) has minimal impact. The heatmap visually confirms these relationships, emphasizing that engine size, horsepower, and car dimensions are the most crucial factors in price prediction.

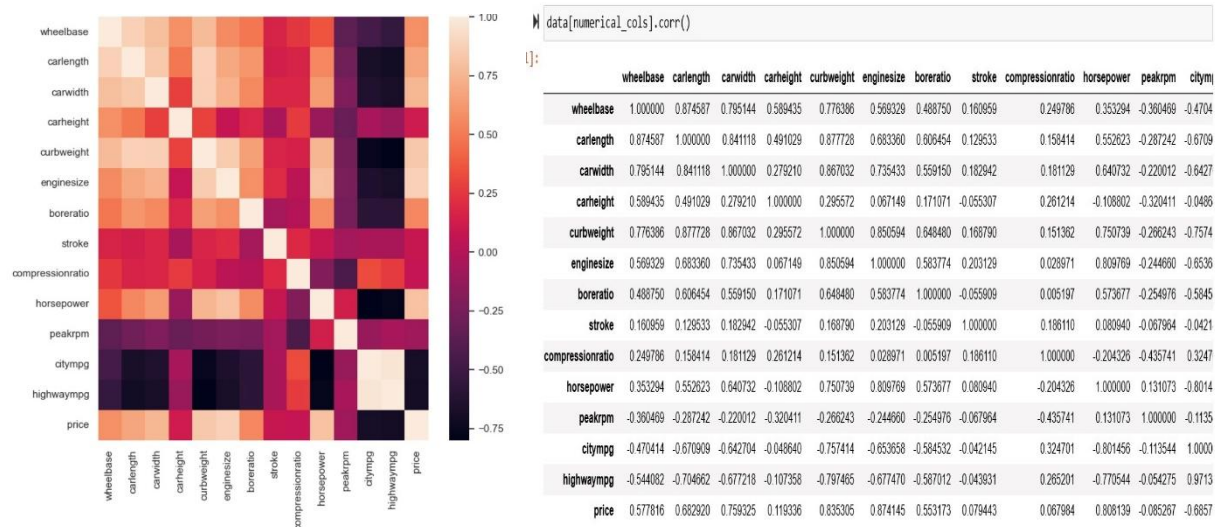


Fig. 5. Correlation Matrix

Car price prediction involves training and evaluating machine learning models. Linear Regression achieved a test accuracy of 0.7375 but struggled with non-linear relationships. The Decision Tree Regressor improved accuracy to 0.8839 by capturing complex patterns but risked overfitting. To enhance performance, Random Forest Regressor was implemented with 15 estimators, reducing overfitting by averaging multiple trees. It achieved the highest R^2 score of 0.9065, making it the most effective model for predicting car prices. In comparison, Linear Regression scored 0.7375, while Decision Tree Regressor achieved 0.8839. The Random Forest model provides better accuracy due to its ability to minimize overfitting by averaging multiple decision trees, thus capturing more complex patterns in the data.

Table 2. Comparative Summary of Models

Algorithm	Accuracy (%)	Key Characteristics
Linear Regression	73.7	Simple, interpretable, assumes linear relationships, struggles with complex patterns.
Decision Tree Regressor	88.3	Captures non-linear relationships, handles high-dimensional data, prone to overfitting.
Random Forest Regressor	90.6	Combines multiple trees for better accuracy, reduces overfitting, computationally intensive.

5. CONCLUSION AND DISCUSSION

The Accurately predicting second-hand car prices is essential in the growing automobile market of Mauritius, where demand rises by 5% annually. Despite multiple car-selling platforms, none offer a data-driven price estimation system. This study developed a machine learning model to bridge this gap, using a dataset of 200 second-hand cars. Key features such as make, year, paint type, transmission, engine capacity, and mileage were analyzed. Four machine learning algorithms were tested, with ten-fold cross-validation ensuring accuracy. Results showed low residual values, confirming machine learning's feasibility for price prediction.

However, challenges remain. Price valuation depends on dynamic factors like market trends, economic fluctuations, and subjective vehicle conditions, including accident history and modifications—data not included in the initial dataset. Future studies should integrate such variables to improve accuracy. Additionally, price trends shift with supply, demand, and seasonal variations, necessitating external data sources like historical prices, inflation, and interest rates for enhanced predictions.

To refine the model, expanding the dataset and leveraging advanced techniques like deep learning and ensemble learning can improve performance. Developing a user-friendly web or mobile app for instant price estimation would add practical value. Real-time updates, trend analysis, and comparison tools could further enhance user experience. In conclusion, machine learning-based car price prediction is feasible and beneficial. Future improvements in data integration, model optimization, and analytics can create a more precise and impactful pricing system, revolutionizing the second-hand car market in Mauritius and beyond.

6. REFERENCES

- [1] NATIONAL TRANSPORT AUTHORITY. 2015. Available at: <http://nta.govmu.org/English/Statistics/Pages/Archives.aspx>. [Accessed 24 April 2015].
- [2] Bharambe, M. M. P., and Dharma dhikari, S. C. (2015) "Stock Market Analysis Based on Artificial Neural Network with Big data". Fourth Post Graduate Conference, 24-25th March 2015, Pune, India.
- [3] Pudaruth, S. (2014) "Predicting the Price of Used Cars using Machine Learning Techniques". International Journal of Information & Computation Technology, Vol. 4, No. 7, pp.753- 764.
- [4] Jassibi, J., Alborzi, M. and Ghoreshi, F. (2011) "Car Paint Thickness Control using Artificial Neural Network and Regression Method". Journal of Industrial Engineering International, Vol. 7, No. 14, pp. 1-6, November 2010
- [5] Ahangar, R. G., Mahmood and Y., Hassen P.M. (2010) "The Comparison of Methods, Artificial Neural Network with Linear Regression using Specific Variables for Prediction Stock Prices in Tehran Stock Exchange". International Journal of Computer Science and Information Security, Vol.7, No. 2, pp. 38-46.
- [6] Listiani, M. (2009) "Support Vector Regression Analysis for Price Prediction in a Car Leasing Application". Thesis (MSc). Hamburg University of Technology.
- [7] Iseri, A. and Karlik, B. (2009) "An Artificial Neural Network Approach on Automobile Pricing". Expert Systems with Application: ScienceDirect Journal of Informatics, Vol. 36, pp. 155-2160, March 2009.
- [8] Yeo, C. A. (2009) "Neural Networks for Automobile Insurance Pricing". Encyclopedia of Information Science and Technology, 2nd Edition, pp. 2794-2800, Australia.
- [9] Doganis, P., Alexandridis, A., Patrinos, P. and Sarimveis, H. (2006) "Time Series Sales Forecasting for Short Shelf-life Food Products Based on Artificial Neural Networks and Evolutionary Computing". Journal of Food Engineering, Vol. 75, pp. 196–204.
- [10] Rose, D. (2003) "Predicting Car Production using a Neural Network Technical Paper- Vetronics (Inhouse)". Thesis, U.S. Army Tank Automotive Research, Development and Engineering Center (TARDEC).
- [11] LEXPRESS.MU ONLINE. 2014. [Online] Available at: <http://www.lexpress.mu/> [Accessed 23 September 2014].
- [12] LE DEFI MEDIA GROUP. 2014. [Online] Available at: <http://www.defimedia.info/> [Accessed 23 September 2014].
- [13] He, Q. (1999) "Neural Network and its Application in IR". Thesis (BSc). University of Illinois.
- [14] Cheng, B. and Titterington, D. M. (1994). "Neural Networks: A Review from a Statistical Perspective". Statistical Science, Vol.9, pp. 2-54.
- [15] Anyaeche, C. O. (2013). "Predicting Performance Measures using Linear Regression and Neural Network: A Comparison". African Journal of Engineering Research, Vol. 1, No. 3, pp. 84-89.