National College of Ireland

# Pre-Owned Bike Price Prediction Using Machine Learning

MSc Research Project
Data Analytics


Keerthana Sathyanrayanan
Student ID: x21195234

School of Computing
National College of Ireland

Supervisor: Dr. Anh Duong Trinh

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Keerthana Sathyanrayanan |
| **Student ID:** | x21195234 |
| **Programme:** | Data Analytics |
| **Year:** | 2023 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Dr. Anh Duong Trinh |
| **Submission Due Date:** | 14/08/2023 |
| **Project Title:** | Pre-Owned Bike Price Prediction Using Machine Learning |
| **Word Count:** | 8580 |
| **Page Count:** | 28 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | Keerthana Sathyanarayanan |
| **Date:** | 17th September 2023 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | ✓ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | ✓ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | ✓ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Pre-Owned Bike Price Prediction Using Machine Learning

Keerthana Sathyanrayanan
x21195234

**Abstract**

Middle class is the economic and social backbone of India. They form a substantial portion of India's population. Due to the cost-effectiveness and affordability, pre-owned bikes market has attracted the middle class. Not only for the reason of affordability, pre-owned bikes are becoming more popular in India for a number of reasons.The focus is also placed on the requirement for pre-owned bikes for both personal and business transportation or commercial purposes. Due to this surge, it is crucial to ensure fair pricing of pre-owned bikes that benefits both buyers and sellers. Many studies concentrate on determining the values of secondhand cars and automobiles, but there is not much research on predicting the price of pre-owned bikes. Thus, this project explores the use of machine learning models for predicting the prices of used bikes and comprehensive comparative analysis is performed. The study evaluates five classic machine learning models, including Linear Regression, Elastic Regression, Support Vector Regressor, Random Forest, and XG Boost, to identify the best models to provide sellers and buyers with fair pricing. The findings of the study include two top performing models, Random Forest and XG Boost.

**Keywords:** Pre-Owned Bikes, Machine Learning Models, Price prediction, Comparative Analysis.

# 1 Introduction

## 1.1 Background and Motivation

Middle class is the significant segment of India. Middle class sector drives economic growth and serves as a base of stability and development. Beyond the confines of economic contributions, the middle class plays a pivotal role in driving societal advancement by virtue of their access to crucial services like education, healthcare, and more. This access acts as a catalyst for transformative changes, leading to better health outcomes, increased literacy rates, and amplified social mobility. Bikes have developed as a symbol of adaptability and usefulness in the realm of transportation, serving both personal

and business objectives with the same flair. Within the realm of delivery services and courier enterprises, bikes are the key to efficient operations. Renowned platforms like Rapido, Uber, Ola, Swiggy, Dunzo and Zomato, heavily reliant on prompt deliveries, emphasize how essential bikes are to their business. Herein, the adoption of pre-owned motorcycles holds the potential for substantial cost savings. This transition can lead to diminished capital outlay on vehicles and commensurate reductions in maintenance and insurance expenditures. Particularly advantageous for small-scale businesses, this approach empowers them to channel their resources into development, frees them from the burden of high operating costs. This trend reflects the pivotal role that second-hand motorbikes can play in fostering entrepreneurship and business expansion. The need for fair and balanced pricing of secondhand bikes is at the heart of this discussion. Such an approach demonstrates the commitment to providing these bikes at prices that the expanding middle class can afford. Facilitating middle-class access to essential transportation options becomes crucial for creating diversity and a strong and healthy society as a result of their growing socioeconomic impact and footprint in India. Maintaining a system that is supportive of their development has the potential to direct India toward a trajectory of all-encompassing development and shared prosperity as the middle class continues to increase, both in terms of numbers and influence.

PewResearchCenter 2021 says [1], Middle class is characterized according to the income range which is between 700 rupees and 1400 rupees that is, ten and twenty dollars per day while the rich is categorized as earning more than twenty dollars per day which is above 1400 rupees. India's middle class is remarkable with a projected number of 300 million people that is approximately 22% of the population. People earning more than $20 per day are defined as high income persons and their population is estimated to be 40 million which is 3% of population. The publication [2]highlights the fact that the number of Indians in the middle class has increased. As per Ernst & Young, the group of middle-income people, which was around 50 million in 2015 which is 5% of India's population will grow rapidly to 200 million in 2020. By 2030, this number will touch 475 million. There will be more middle-class residents in the entire nation than there are in China by that point.

In India, the middle class has had the most rapid rise between 1995 and 2021, both in terms of percentage and overall numbers. During this time, it grew by around 6.3 percent annually. Currently, over 31% of the population is middle class and it is projected to cross 38% by 2031 and 60% by 2047. By the time India hits 100 million people, more than a billion of its citizens would be middle class, according to the Economic Times.

[1]https://www.pewresearch.org/global/2021/03/18/the-pandemic-stalls-growth-in-the-global-middle-class-pushes-poverty-up-sharply/

[2]https://www.asianstudies.org/publications/eaa/archives/the-middle-class-in-india-from-1947-to-the-present-and-beyond/

[3]

All of the information presented and the prevalent areas of attention highlighted lead to the formulation of my research topic.

## 1.2 Research Question

How effective are the machine learning models in accurately predicting the resale value of the used bikes?

## 1.3 Proposed Approach

Using the five classic machine learning models, we aim to perform a comparison research and identify the best models for estimating used bike prices. The models employed include Linear Regression, Elastic Regression, Random Forest, Support Vector Regressor and XG Boost.

## 1.4 Report Structure

This part will provide an overview of the overall structure of the research study, which will be detailed in the following sections.

- Section 2: RELATED WORK- This section presents a comprehensive review of the work on used vehicle prediction using several methodologies. This section discusses innovative techniques to answering my research question.
- Section 3: METHODOLOGY AND IMPLEMENTATION- This section delves into the methodology and implementation. It covers all the process of the experiment conducted.
- Section 4: TECHNICAL SPECIFICATION: This section presents the hardware and software specifications required to conduct this experiment.
- Section 5: IMPLEMENTATION: This section presents information about number of trials performed in the experiment and their results are reported.
- Section 6: EVALUATION- This section discusses the evaluation step. It showcases the numerical fluctuations prior to and after the hyperparameter tuning.
- Section 7: RESULTS AND DISCUSSION- The results section displays the interpretation and discussion of the findings.
- Section 8: CONCLUSION AND FUTURE WORK- This section goes into the conclusion, limitations of current research and future enhancements.

[3] `https://economictimes.indiatimes.com/defaultinterstitial.cms`

# 2 Related Work

## 2.1 Used Automobiles Price Prediction

With the help of many machine learning models, Satapathy et al. (2022)created an automated automobile price forecast system. The project's dataset was obtained from Kaggle and contains information about secondhand vehicles in India. The models employed in this study include KNN, Random Forest, The Gradient Boosting, and XGBoost. With an accuracy of about 92%, the XGBoost model was determined to be the most accurate and consistent model for second-hand automobile price prediction. The authors used the dataset to build the model, which included the following 12 factors (Location, Kilometers Driven,Name, Year, Fuel Type, Engine Transmission, Engine, Owner Type, Price, Power, Mileage), of which only 6 (Year, Kilometers_Driven, Fuel_Type, Transmission, Mileage, Seat,) were used to implement the model after EDA. They also divided the dataset into 85 percent for training and 15 percent for testing. They made a variety of diagrams to better comprehend the relationship between independent and dependent properties. They are as follows: 1) Heatmap Plotting - to visualize the correlation between numerical columns like Mileage, Seat, etc., features and the Price column. 2) Barplot - to understand how categorical features like Fuel Type, Transmission, etc. affect the Price 3) Counterplots - percentage of each categorical feature such as Fuel Type, Transmission, etc., present in the dataset. 4) Scatterplots - how the numerical columns like Age, Mileage, etc. affect the Price column.The evaluation metrics used are R-Squared and RMSE. The accuracies of the models are as follows, KNN - 85%, Random Forest - 90.17%, The Gradient Boosting - 91%, XGBoost - 92.66%. Though the literature review was not well organised, the methodology of this study proved to be beneficial. The research gap of this study is, it does not consider the history of previous owners which is addressed in my research.

The study presented by Chavare et al. (2023) performed a research that used three machine learning approaches to construct an algorithm to forecast automobile sales prices: Multiple Linear Regression (MLR), Random Forest, and Support Vector Machine (SVM). The information for the study was gathered from cars24.com and included automobile details such as year of manufacturing, mileage, horsepower, and associated selling prices. The study discovered that MLR has an R-Squared value of 0.75. Random Forest, on the other hand, produced a less precise R-Squared value of 0.62. SVM, on the other hand, gave near-accurate predictions and was chosen as the best model to forecast automobile sales prices using the available information. According to the publication, the SVM algorithm is the best one for predicting the selling price of used automobiles. The evaluation measures give insight into model correctness and provide as a baseline for comparison with other machine learning approaches, assisting in the selection of the optimum model. In terms of limitations, the study did not address the influence of outliers

or anomalous data points on prediction models. This research gap is covered by my research.

In the work by Monburinon et al. (2018) primary objective was to estimate used automobile prices using regression models based on supervised machine learning. The authors conducted a comparison study to assess the performance of multiple linear regression, random forest regression, and gradient boosted regression trees. The dataset was gathered from eBay-Kleinanzeigen, a German e-commerce site, and included 304,133 observations and 11 characteristics. The models' performance was compared using the same testing data and the mean absolute error (MAE). The comparison analysis found that gradient boosted regression trees gave the greatest prediction performance, with an MAE of 0.28, followed by random forest regression with an MAE of 0.35, and multiple linear regression with an MAE of 0.55. The authors discovered that regression models based on supervised machine learning predict continuous variables with high accuracy. As a result, the regression model may be applied in real-world circumstances. The comparison of three prominent regression models was a notable strength of the study; this offered background for researchers to evaluate which regression model is appropriate to utilize for a certain application. The restricted dataset of the study, however, is a disadvantage, as the findings were confined to automobiles sold on a single German e-commerce website. Furthermore, it may not be immediately relevant to other nations' automobile marketplaces. The research gap identified is it did not analyze the impact of outliers. This void is filled by my research.

The purpose of this research was to create a used automobile price forecast system utilizing the Random Forest Regressor algorithm. Kinadi et al. (2022) gathered information from two separate Indonesian secondary automotive marketplaces, focusing on cars with attributes from 2015 to 2018 and mileage under 1000 kilometers. Their motivation for performing this study stemmed from the increased usage of private vehicles as a safer means of transportation during the Covid-19 outbreak, as well as the market's different pricing for secondhand cars. The authors' data preprocessing step comprised data selection and cleaning, in which they filtered the dataset to only contain Surabaya data and removed unrelated columns. They then utilized the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) methodologies to analyze their outcomes after using a Random Forest Regressor algorithm to estimate used automobile pricing. Their results demonstrated that the Random Forest Regressor method can handle vast volumes of data with high dimensions, with accuracy values ranging from 0.4520 to 0.5663 for their validation and testing datasets. The study's strength is its use of an efficient algorithm to estimate used vehicle prices, as well as its ability to give useful information for both buyers and sellers. The study's drawback is that it is limited to Surabaya and automobiles with specified model years and mileage, limiting its generalizability to other regions and car types. The potential research gap is its reliance on a single machine learning

algorithm which is taken care of in my study by performing comparative analysis.

The goal of this research was to create a machine learning model that could predict the selling price of used automobiles based on important criteria. The writers Narayana et al. (2022) stressed the importance of such a model in the rapidly expanding Indian second-hand automobile industry. To enhance performance, they applied two machine learning algorithms, Random Forest Regression and Extra Trees Regression, and tweaked their hyperparameters using RandomizedSearchCV. The authors found from their investigation that the suggested model outperformed previous techniques, giving higher performance with fewer error. In terms of research strengths, the study employs a real-world dataset relevant to the Indian automobile sector. However, the study used a small dataset, which limits the generalizability of the findings. The research gap found in this study is the authors did not go into depth about their feature engineering methodologies.

Chen et al. (2017) examined the predictive power of linear regression and random forest models on used automobile pricing. The data for this study were gathered from the Shanghai used automobile market and a used car auction website, totaling over 100,000 used car dealer purchase records between October 2015 and October 2016. The newness, economy value, and function of the automobile were used to choose explanatory factors. Model 1 was built for a certain automobile brand, Model 2 for a specific car series, and a universal model for all used cars. The study's findings suggest that random forest models outperform linear regression in most circumstances, particularly for complicated models with a large number of variables and samples. Because Model 1 (a specific automobile model) only included 5 explanatory variables, the effect of both approaches was easily altered by sample size, however random forest was more stable. Model 2 (a specific automobile series) included 13 explanatory factors and gave improved results as sample size rose, with random forest outperforming when enough data were available. With 19 explanatory variables, the universal model displayed the largest advantage of random forest, explaining 95.6% of the variations in the response variable with a Root Mean Square Error of 16,648. In all models, linear regression produced excellent results only when there were few explanatory variables and a sufficient sample size. The study's features include its big database and comparative examination of algorithms, which should aid academics and modelers in comparing different models for various circumstances. The study, however, has certain limitations due to the limited number of samples in each model, making it relevant for just particular automobile brands or models.

The research done by Wang et al. (2021) used machine learning algorithms to anticipate used automobile prices with less human interaction. The authors pre-processed a data set and used supervised learning to compare the performance of several techniques. They discovered that Extra Trees Regressor and Random Forest Regressor worked rather well and used the hyperparameter function to improve the method. The authors determined that the Extra Trees Regressor method had the highest performance, with an R2 of

0.9807, and that the final algorithm model was confirmed using additional data. Using this method, used car pricing might be created automatically, making the used car market's process quicker and more competitive. The authors stated that future work might increase the algorithm's performance by tweaking the super parameters. Strengths of this study included the use of various algorithms in supervised learning and the optimization of the algorithm using the hyperparameter function. Additionally, the study's contributions towards automating the used car market's workflow were noteworthy. Limitations included the subjectivity of dataset selection and the lack of some significant features in the data set. My research has filled this gap by appropriate selection of features.

Longani et al. (2021) aimed to propose a methodology that would allow fair prices for used automobiles in Mumbai to be calculated. The authors gathered information on 2,454 automobiles from web sources and used machine learning techniques to forecast fair values. They built two models that predicted used vehicle prices using random forest and eXtreme Gradient Boost approaches. The models' performance was then assessed and compared to guarantee that the most accurate model was chosen. The authors began by reviewing previous research on machine learning algorithms for forecasting car pricing. Previous research has revealed that multiple linear regression, decision trees, Naive Bayes, and ANN are unreliable. Ensemble approaches that integrate several models, such as bagging and boosting, are effective for improving accuracy. One of the strengths of this study is the use of ensemble machine learning techniques to improve performance. The authors found that the eXtreme Gradient Boost technique outperformed the Random Forest Algorithm, displaying a Root Mean Squared Error of 0.53 compared to 3.44 respectively. The paper also considers several factors that influence the price of pre-owned cars, including year of purchase, run of the car (in terms of kilometers), showroom price, and many more. The dataset used was also extensive, with over 2,400 records, making it thorough. The analysis is limited to pre-owned car prices in the Mumbai region. Therefore, generalizing the results may not be possible for a broader audience. The outcomes of this study may not be relevant or applicable to other geographical areas or global markets. This research gap is addressed in my study.

In the research documented by Li et al. (2022) the central goal was to present an accurate model that could estimate the price of used automobiles based on a variety of parameters. For its investigation, the study employed two machine learning algorithms: random forest and LightGBM. The most significant variables were chosen using the random forest approach. LightGBM was used to produce predictions based on the factors that were chosen. The research found that machine learning algorithms may be a very useful tool for precisely calculating the price of used cars. It emphasizes the power of machine learning algorithms in processing large volumes of data, identifying patterns, and making accurate predictions. Machine learning algorithms, unlike traditional approaches that rely on entire data to generate predictions, can manage missing data. The study

also demonstrates that machine learning algorithms can outperform traditional methodologies, particularly in terms of accuracy. The study found one restriction as the necessity for enormous volumes of data that are difficult to obtain. Incomplete or faulty data can impair machine learning model performance, resulting in incorrect predictions. The research also acknowledges the possibility of biased machine learning algorithms producing inaccurate predictions. Bias can be caused by the nature of the data used to train the algorithm, such as mistakes and inconsistencies, or by systematic biases in the dataset. My study addresses a specific research gap identified in this study which is selection of right features.

The work by Han et al. (2022) introduces a novel approach named the weighted mixed regression (WMR) model for the prediction of used automobile prices Feature engineering is accomplished by data preparation and screening procedures such as outlier removal, missing value filling, and correlation analysis. Using the mathorcup Big Data Challenge Model Evaluation Standard, the model is compared to various regression models such as XGBoost, LightGBM, Random Forest and GBDT. The WMR model outperformed the other models, scoring 0.79 on the assessment scale. The suggested WMR model enhances model resilience by merging two regression models with different features (Random Forest and XGBoost). The WMR type also performs well in terms of second-hand automobile worth. The paper provides a unique technique for predicting second-hand automobile prices that utilizes a weighted mixed regression model. Feature engineering is accomplished using a variety of approaches, including data pretreatment and screening procedures, which increase model accuracy. The paper compares the proposed model to existing regression models using a complete assessment standard and shows that the WMR model beats the others. The study does not address the ethical concerns of employing machine learning algorithms to predict used automobile prices. Furthermore, it is unknown if the suggested approach would be useful in projecting used automobile values in other locations or markets. The features chosen for this model may be unique to the Chinese market. It is vital to evaluate the model's performance on data from other nations and determine whether it can be utilized in a larger context. Therefore, potential research gap is the generalizability of the work.

Hankar et al. (2022) estimated the resale value of used automobiles in Morocco based on factors such as mileage, fuel type, fiscal power, brand, model, and production year. To estimate used automobile costs, the researchers examined multiple supervised machine learning methods. Gradient Boosting Regressor (GBR) had a good R-squared score and a low root mean squared error in all evaluated models. According to the study, GBR outperformed other evaluated models and approximated an R2 value of 0.80. However, this is a poor R2 value for determining if the model is a good fit. The study discovered that the GBR model has a poor value for reducing RMSE errors. The study's key strength is that it includes a machine learning model for forecasting used automobile prices. Customers

and sellers may use this model to assess the price worth of an automobile before making a purchase or sale decision. To avoid information loss, the study also includes a thorough data gathering method and preparation activities. However, the study has significant limitations, including the fact that data was collected from only one ecommerce website, Avito, and that it solely focused on Moroccan used car costs. Deep learning techniques, which might improve the model's performance, were not used in this work. The research gap identified is lack of sufficient variables to the feature set.

The study aims to develop a high accuracy model that can predict the price of a used cars with no bias to either the buyer or the merchandiser. According to Varshitha et al. (2022), the used vehicle industry is expanding internationally, but it is still in its infancy in India, and the study intends to establish a model that may assist in evaluating the price of used automobiles with low inaccuracy. The information contained, crucial used cars variables such as gasoline type, seller type, transmission, and the number of prior owners. Outliers were removed from the raw data, fake and superfluous data points were removed, and all string data types were converted to numerical data types. They then create a supervised learning-based ANN model and compare its performance to that of other algorithms such as Lasso, Ridge, Linear regression, and Random Forest. Based on the results of the trials, the authors discovered that the Random Forest model has the lowest Mean Absolute Error (MAE) value of 0.746 and the highest R-squared error value of 0.917, making it the best method.

Wang et al. (2022) conducted a study with aim to assess the worth of secondhand cars. The fundamental purpose of this research is to identify the important elements influencing used automobile transaction costs, compare and filter the critical aspects using several machine learning models, and lastly build a prediction model based on the extracted features. Six models, namely AdaBoost, GBDT, XGBoost, LightGBM, CatBoost, and Decision tree, were used for this aim.The research focuses on the suggested LightGBM-based model and its advantages in terms of decreasing time complexity and unnecessary processing. The authors show that the LightGBM model outperforms other models when it comes to developing used car sales tactics. Furthermore, according to the results of the studies, the LightGBM model outperformed the other models, attaining an assessment criterion result of 0.85257. The paper highlights the Linear regression model that was used for analysis in this study to perform a linear relationship between the dependent variable (price) and the independent variables (features) for five different types of used car sales cycles: "Best-selling," "Easy to Sell," "Sellable," "Difficult to Sell," and "Stagnant Sell." It aided in identifying important influencing elements utilized to forecast used automobile prices. Despite the fact that a thorough model comparison was performed, the presentation of model comparisons and outcomes may be improved. However, a potential research gap could be the limited scope of the study, which is focused on the Chinese used car market.

The purpose of the study outlined in Shaprapawad et al. (2023) is to develop the most effective machine learning model that can help vendors, buyers, and automotive manufacturers make fairly accurate car pricing estimates based on user data. The study compares the performance of various machine learning algorithms, including linear regression, Lasso regression, Ridge regression, Elastic net regression, Random Forest, Decision tree, and Support Vector Machine, using metrics such as R-squared score, Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error. It consists of 90% training data and 10% validation data. Based on the evaluation's criteria, the support vector machine model appears to be the best option, with an R-squared score of 95.27%, a Mean Absolute Error of 0.142, a Mean Squared Error of 0.047, and a Root Mean Squared Error of 0.217. The methodology is clearly explained in the publication. However, the model may be enhanced further by incorporating more variables that may influence used car pricing. This research gap is filled in my work by including important variables.

The work conducted bySaini and Kaur (2023) sought to anticipate the resale value of autos correctly to avoid any losses while dealing with a used car. The study developed five distinct models utilizing several machine learning algorithms, including Linear Regression, Decision Tree, Random Forest, KNN, and XG Boost, and evaluated their accuracies to determine the best-performing model. The extreme gradient boost and random forest methods both performed much better than the other techniques. The study reveals that the random forest algorithm performed the best, with an accuracy score of 90.67%. One possible shortcoming of this work is that the assessment metric employed for this investigation is not specified. However, the text is well-organized, and the comparison of accuracy and performance is provided clearly.

In the work presented by Narayana et al. (2021) recommended the use of predictive analytical models to reliably project future automobile costs based on factors such as model, transmission type, and number of years in use, among others. The study examined numerous regression models, including linear regression, random forest, and k-nearest neighbors (KNN) regression, according to the publication. The authors discovered that random forest regression was the most accurate model, with an accuracy rate of 85%. To make the data compatible with the regression models, they performed feature engineering activities such as deleting outliers and assessing feature significance. The research compares and evaluates the efficacy of several regression models in forecasting the price of secondhand cars. This is one of the document's potential strengths. One disadvantage of the paper is that the dataset is restricted and may not include all of the essential aspects that might impact the retail selling price of used automobiles. Some important factors, such as the car's condition, accident history, location, and supply and demand, are not taken into account in the dataset and may have an impact on the car's price. The authors acknowledged this shortcoming and proposed that future efforts include more factors influencing used automobile selling prices.

Hemendiran and Renjith (2023)examined the application of machine learning algorithms to predict the price of a secondhand car. In this work, supervised machine learning models were used to anticipate used automobile prices in India using historical data gathered from daily news stories, magazines, and several standard websites. Five prediction models were used: the Random Forest Regressor, the Extra Tree Regressor, the Bagging Regressor, the Decision Tree, and the XG Boost technique. After analyzing the forecasts of the models used, the most accurate predictions were chosenThe Random Forest model generated the best results. The approach is described in depth, which will aid in the replication of similar studies in the future. The study provided vital insight into the application of machine learning models in forecasting used automobile values based on numerous parameters, particularly those relevant in the Indian market. The study, however, is confined to data obtained in India. Thus the research gap identified is that it may not be generalizable to other countries.

## 2.2 Consistent models from the papers reviewed

Two consistent models can be observed in this section. They are Random Forest and XG Boost. They are backed up by relevant publications and presented with the positive and negative aspects and evaluation metrics employed in each of these papers. In Table 1 the papers in which XG Boost performed exceptionally well are reported. In Table 2 the publications with Random Forest as the best model are showcased.

Table 1: Consistent Model 1- XG Boost

| **Author** | **Title** | **Evaluation Metric** | **Merits** | **Demerits** |
|---|---|---|---|---|
| Santosh Kumar. Satapathy, Rutvikraj Vala, and Shiv Virpariya | An Automated Car Price Prediction System Using Effective Machine Learning Techniques | 1. $R^2$ score<br>2. RMSE | Methodology well explained | Literature Review not well organized |
| Iqbal Singh Saini and Navneet Kaur | Comparison of Various Regression Techniques and Predicting the Resale Price of Cars | Metric used not mentioned | Methodology well structured | Evaluation metric used for prediction of accuracy not mentioned |

| **Author** | **Title** | **Evaluation Metric** | **Merits** | **Demerits** |
|---|---|---|---|---|
| Chetna Longani, Sai Prasad Potharaju, and Sandhya Deore | Price Prediction for Pre-Owned Cars Using Ensemble Machine Learning Techniques | 1. MAE<br>2. MSE<br>3. RMSE | Methodology adequately explained | Dataset is confined to specific city |

Table 2: Consistent Model 2- Random Forest

| **Author** | **Objective** | **Evaluation Metric** | **Merits** | **Demerits** |
|---|---|---|---|---|
| Janke Varshitha, K Jahnavi, and Dr. C. Lakshmi | To provide a high precision model that can estimate the price of a used car | 1. $R^2$ Score<br>2. MAE | Methodology clear and concise | Only few references mentioned in Literature Review |
| Iqbal Singh Saini and Navneet Kaur | To predict the resale value of cars | Metric used not mentioned | Methodology well structured | Evaluation metric used for prediction of accuracy not mentioned |
| Chejarla Venkat Narayana, Chinta Lakshmi Likhitha, Syed Bademiya, and Karre Kusumanjali | To provide a fair price mechanism to predict the selling price of used cars | 1. MAE<br>2. MSE<br>3. RMSE | Extensive review of related works has been performed | Dataset does not include some important features |
| Chuancan Chen, Lulu Hao, and Cong Xu | To find an optimal model to accurately predict the price of used cars | RMSE | Diverse and comprehensive dataset used | Future work requires more information |

| Author | Objective | Evaluation Metric | Merits | Demerits |
|---|---|---|---|---|
| Hemendiran B and Renjith P N | To predict the prices of used cars for resale | 1. MAE<br>2. MSE<br>3. RMSE<br>4. $R^2$ Score | Valuable information provided on methodology | Brief overview of results presented but not detailed explanation given |
| Chejarla Venkata Narayana, Nukathoti Ooha Gnana Madhuri, Atmakuri NagaSindhu, Mulupuri Aksha, and Chalavadi Naveen | To develop a prediction model that can estimate the selling price of used cars | 1. MAE<br>2. MSE<br>3. RMSE | Dataset illustration and feature selection well elaborated | Lack of detailed discussion of results |
| Annisaa Fauziyah Kinadi, Rachmadita Andreswari, Edi Sutoyo, Ramdhan Nugraha, and Anton Abdul Basah Kamil | To develop a used car price prediction system using the Random Forest | 1. RMSE<br>2. MAE | Detailed description of methodology | Less information on dataset used |

# 3 Methodology

## 3.1 Experiment Design

Figure3.1 shows the process flow of the experiment conducted.

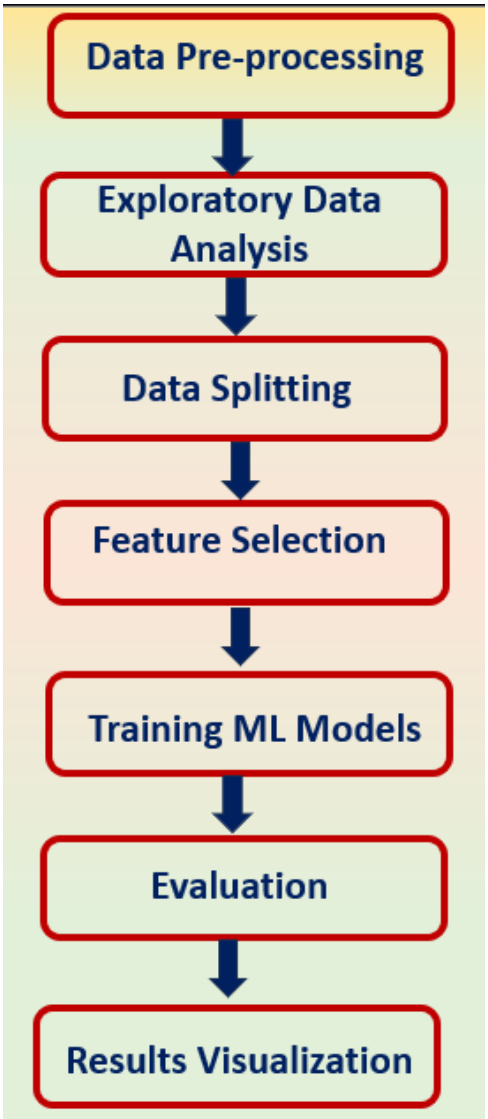


Figure 1: Work Flow Design

## 3.2 Data Acquisition:

- The dataset used in this project is a bike price prediction dataset.
- It includes various features related to bikes, such as model name, model year, kms driven, owner location, mileage power, price.
- It comprises over 5063 records providing data from all across India.
- The source of the dataset is Kaggle. [4]

## 3.3 Data Loading:

- The essential R packages are loaded to facilitate further process.

[4] https://www.kaggle.com/datasets/vinayjain449/bike-prediction-with-linear-regression

- The dataset is loaded into the data frame.
- Inital glance of the data is done to understand the structure and datatype of the data.
- Data quality is assessed by checking for the preseance of missing values.

## 3.4 Data Preprocessing

### 3.4.1 Categorical to Numerical Conversion:

The 'owner' feature initially contains categorical labels such as 'first owner', 'second owner', 'third owner', and 'fourth owner'. To make it suitable for modeling, these labels are transformed into numerical labels using str replace and as.numeric.

The mapping is as follows: 'first owner' - 1 'second owner' - 2 'third owner' - 3 'fourth owner or more' - 4. This transformation converts the 'owner' feature into a numerical representation that can be used in regression models.

### 3.4.2 Outlier Removal:

Outliers are identified and removed for the 'model year' and 'price' features using the Interquartile Range (IQR) method. Outliers can significantly affect model performance and may lead to biased predictions. To handle outliers, the code computes the lower and upper bounds for 'model year' and 'price' features using the Interquartile Range (IQR) method. Before removal of outliers- 5062. After removal of outliers- 4382.

- For the 'model year': Bikes older than 15 years are prone to increased wear and tear and are of least interest to the customers. Hence, they were removed. The first quartile (Q1) and third quartile (Q3) are calculated using quantile. The interquartile range (IQR) is calculated as Q3 - Q1. The lower fence is set as Q1 - 1.5 * IQR, and the upper fence is set as Q3 + 1.5 * IQR. Any 'model year' values below the lower fence or above the upper fence are considered outliers.
- For the 'price': In the price variable, there were outliers with the value as 0 which is invalid. This could impact the prediction by the model significantly. Hence, they were removed. The minimum value of price was 0 and after removal of outliers, minimum value is changed to 19300. The same process is repeated for the 'price' feature to identify outliers. The identified outliers are stored in separate data frames outliers and outliers1 for 'model year' and 'price', respectively. Then, these outliers are removed from the original data frame df. Removing outliers helps improve the quality of the data and enhances the performance of the models by reducing the influence of extreme values.

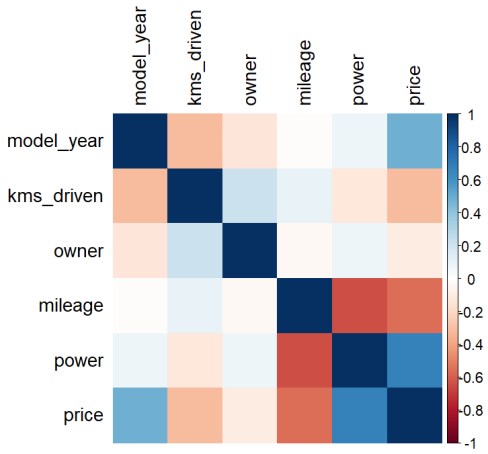


Figure 2: Correlation Plot

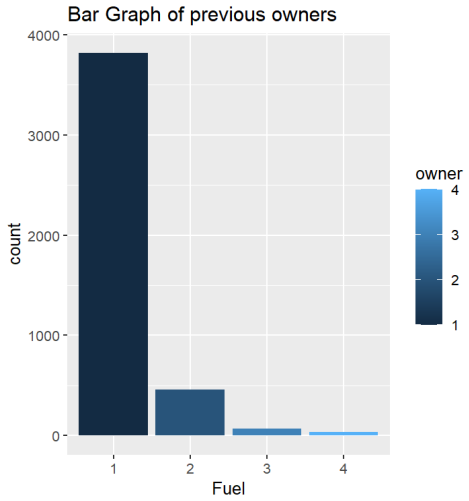


Figure 3: Bar Plot of previous owners

## 3.5 Exploratory Data Analysis (EDA):

### 3.5.1 Correlation Matrix:

- 2 The correlation values are displayed for each pair of features ('model year', 'kms driven', 'owner', 'mileage', 'power', and 'price').
- The features'model year' have a high positive correlation with 'power' and 'price'.

### 3.5.2 Bar Plot:

3A bar plot is generated to visualize the distribution of previous owners ('owner' feature).

## 3.6 Data Splitting

Dataset is split into training and testing subsets into two different ratios namely, 80:20 and 70:30.

## 3.7 Feature Selection

Based on the domain knowledge acquired, the following features are potential influencers in determining the bike prices. The 'location' and 'model_name' does not have significance to my research as name of a bike and location does not influence the price of the bike.

The following features are selected for the model training for the following reasons:

Model year: Old bikes have low market value as they would have already have driven higher mileage. Also, they will lack latest features as in recent release bikes.

Kilometers driven: The distance travelled by a bike is an important feature as it provides information about the usage of the bike. Higher the mileage, more wear and tear of the bike therefore, increased maintenance cost. Hence, the value of the bike decreases.

Owner: The resale value of the bike is directly impacted by the number of prior owners. Bikes with less number of owners are considered to be in better condition hence, good market value.

Mileage: Bikes that provide more mileage are more fuel efficient thus it has the potential to impact the resale value of the bike.

Power: The engine power of a bike can determine its performance. People looking for powerful rides consider the engine power. The value of the bike increases with power of the engine.

## 3.8 Training ML models

### 3.8.1 Linear Regression

Linear regression is a popular model that involves fitting a linear equation to observed data to represent the connection between a dependent variable (target) and one or more independent variables (predictors). In this scenario, we employ linear regression to forecast bike prices based on model year, kilometers driven, owner, mileage, and power. We initiate the linear regression by defining the following formula:

$$price \sim model\_year + kms\_driven + owner + mileage + power \tag{1}$$

The formula depicts that we attempt to predict price with help of other variables.

Following the initialization, the linear regression model is fit to the training data. The defined formula along with training dataset is given as inputs to the lm() function. It then

calculates the coefficients for each independent variable that minimize the gap between the predicted and actual prices in the training set. The model learns the relationship between dependent and independent variables using the training data. After training step, summary of the model is generated. The summary of the model helps us to analyze the links between features and prices and evaluate the efficacy of the model .

The lm function, a key utility in the R programming language, is essential in predictive modeling using linear regression. Linear regression is the process of producing predictions based on data by identifying correlations between known variables. The lm function can operate without the conventional hyperparameters as used in complex machine learning algorithms. It forecasts the outcomes based on inputs while following to principle of linear regression. It necessitates a precise formulation of the formula that establishes the relationship between the dependent variable (the one to be predicted) and the independent variables (factors influencing the prediction).In essence, while the lm function lacks the traditional hyperparameters seen in more complex algorithms, it excels in flexibility and simplicity. Thus, it serves as baseline model for comparison and providing a simple benchmark against more complex models.

The model built is evaluated using the MAPE and RMSE metrics. The lowest value of MAPE recorded is 33.44 and least RMSE value identified is 28553.16.

### 3.8.2 Elastic Regression

Elastic net regression is a linear regression approach that combines the characteristics of Ridge and Lasso regression. It seeks to handle difficulties such as multicollinearity (the correlation of predictor variables) and feature selection (the selection of essential predictors). The model minimizes a combination of the L1 (Lasso) and L2 (Ridge) regularization terms, which aids in dealing with scenarios with a large number of correlated variables.

Hyperparameter tuning is carried out with caret package. In the first place, data is split into train and test to ensure evaluation of the model in unseen data.The trainControl() method is used to configure the cross-validation parameters. The model is trained and examined ten times using 10-fold cross-validation, which divides the data into ten subsets (folds). This 10-fold cross-validation is performed 5 times to boost the process's reliability (determined by the repeats parameter). Random search is performed. Features are centenred and scaled before modeling. The tuneLength parameter defines that ten distinct hyperparameter combinations will be explored. The method is set to "glmnet," suggesting that elastic net regression is used. As final step, MAPE and RMSE are calculated. 33.76 is the least MAPE score achieved and 28555.25 is the best RMSE value.

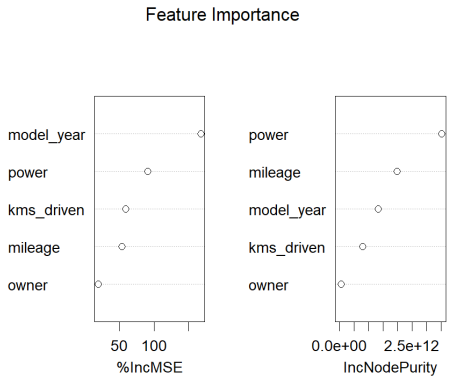


Figure 4: Feature Importance

### 3.8.3 Support Vector Regressor

A hyperparameter grid is created, which includes combinations of the hyperparameters 'epsilon,' 'cost,' and 'kernel'. We then train the model with these parameters and make predictions on test set. Following this, we compute the RMSE and MAPE scores. If the computed RMSE is less than the prior best RMSE, the current hyperparameters, RMSE, and MAPE are recorded as the best values. The best recorded MAPE value is 21.85965 and that of RMSE is 23668.89.

### 3.8.4 Random Forest

Random Forest model is built using the randomForest() function. Hyperparameter tuning is performed using several parameters including, number of decision trees (set to 500), mtry (set to 3), nodesize (set to 5), maxnodes (NULL), importance (TRUE) and replace (TRUE). These factors impact the model's behaviour. Following this is the feature importance visualization. 4This stage entails visualizing the significance of various Random Forest model features. The varImpPlot() method provides a plot that shows how much each attribute adds to the predictive power of the model. The primary goal is to determine which features have the most effect on the model's predictions. Finally, MAPE and RMSE are used to evaluate the model built. The least MAPE value achieved is 15.28 and least RMSE score obtained is 18053.13.

### 3.8.5 XG Boost

The learning rate (eta, set to 0.1), Maximum tree depth (max_depth set to 3) and Number of boosting iterations (nrounds set to 100) are the key hyperparameters employed. Iterative process is used to obtain best combination of hyperparameters. The 'params' list was updated after each cycle to reflect the new hyperparameter values.The combination of hyperparameters that produced the optimum performance was found after many

iterations of hyperparameter tuning. The optimized hyperparameters are used to train final XG Boost model. RMSE and MAPE are used to evaluate the model performance. The best MAPE score attained is 16.81 and the best RMSE score attained is 18287.10.

# 4 Technical Specification

## 4.1 Hardware and Software Specifications:

- CPU: Processor used for the research is 11th Gen Intel(R) Core(TM) i5-11320H @ 3.20GHz 2.50 GHz.
- RAM: RAM used for the study is 16GB.
- Storage: 477 GB
- GPU: NVIDIA GeForce MX450
- Operating System: 64-bit operating system, x64-based processor. Windows 11 is used.
- Environment: R Studio.
- The programming language used is R programming.
- The package dependencies are tidyverse, corrplot, ggplot2, lubridate, gridExtra, caTools, GGally, randomForest, caret, ISLR, xgboost.

# 5 Implementation

- The models built were made to go through several trials to ensure correct numbers.
- Each model was tested with two different splits.
- The split ratios are 80:20 and 70:30.
- In each split, three trials were performed to obtain RMSE and MAPE scores.
- Since the dataset is small, changing the split ratios did not impact the MAPE and RMSE values greatly.

This confirmed that the values obtained were correct and that the model developed was stable.

In Table 3 it shows the trails performed in each split in Linear Regression.

In Table 4 it displays the trials run in each split of Random Forest.

Table 3: Data Split and No. of trials- Linear Regression

| Split | Seed | Metrics | |
|---|---|---|---|
| | | MAPE | RMSE |
| 80:20 | 123 | 33.44 | 30159.58 |
| | 321 | 34.83 | 28553.16 |
| | 1712 | 33.64 | 30392.1 |
| 70:30 | 123 | 33.44 | 30159.58 |
| | 321 | 34.83 | 28553.16 |
| | 1712 | 33.64 | 30392.1 |

Table 4: Data Split and No. of trials- Random Forest

| Split | Seed | Metrics | |
|---|---|---|---|
| | | MAPE | RMSE |
| 80:20 | 123 | 33.44 | 30159.58 |
| | 321 | 15.28 | 18053.13 |
| | 1712 | 17.01 | 19405.75 |
| 70:30 | 123 | 16.92 | 18910.54 |
| | 321 | 15.28 | 18053.13 |
| | 1712 | 17.01 | 19405.75 |

In Table 5 it displays the trials performed in each split of Support Vector Regressor.
In Table 6 shows the trails run on each split in Elastic Regression.
In Table 7it displays the trials performed in each split of XG Boost.

# 6 Evaluation

The evaluation metrics used for this study are RMSE and MAPE.The following paper influenced the use of RMSE as an evaluation metric. Chai and Draxler (2014) reported that RMSE is better than MAE. It is mentioned that RMSE impacts variance by giving bigger absolute value errors more weight than smaller absolute value errors, thereby weighting unfavorable situations more strongly. Also it is documented that When the error distribution is predicted to be Gaussian, the RMSE is a better way to characterize model performance than the MAE. The RMSE meets the triangle inequality criteria for a distance measure.Moreover, RMSE avoids using absolute values, which are extremely undesirable in many mathematical calculations, for example computing gradient or sensitivity. Clearly, from this publication, RMSE proves to be one of the promising evaluation metric.

The Mean Absolute Percentage Error (MAPE) formula is given by:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - F_i}{A_i} \right| \times 100\%$$

Table 5: Data Split and No. of trials- Support Vector Regressor

| Split | Seed | Metrics | |
|---|---|---|---|
| | | MAPE | RMSE |
| 80:20 | 123 | 24.32 | 24148.12 |
| | 321 | 21.85 | 23705.36 |
| | 1712 | 22.78 | 23668.89 |
| 70:30 | 123 | 24.32 | 24148.12 |
| | 321 | 21.85 | 23705.36 |
| | 1712 | 22.78 | 23668.89 |

Table 6: Data Split and No. of trials- Elastic Regression

| Split | Data | Metrics | |
|---|---|---|---|
| | | MAPE | RMSE |
| 80:20 | 123 | 35.21 | 30157.29 |
| | 321 | 34.38 | 28555.25 |
| | 1712 | 33.76 | 30379.65 |
| 70:30 | 123 | 35.21 | 30157.29 |
| | 321 | 34.38 | 28555.25 |
| | 1712 | 33.76 | 30379.65 |

where:

$A_i$ : Actual value of the data point $i$,

$F_i$ : Forecasted value of the data point $i$,

$n$ : Total number of data points.

8MAPE was chosen as the second evaluation measure due to its own advantages such as ease of calculation and interpretability.

The lower the MAPE number, the more accurate the model is.[5]

MAPE and RMSE penalize necessary prediction errors more significantly, which is typically essential in price prediction tasks. MAPE and RMSE strike a balance between bias (systematic errors) and variability (random errors).

9The performance difference between the Random Forest and XG Boost models is minimal. The Random Forest model has a lower MAPE of 17.01 and a Root Mean Square Error (RMSE) of 19405.75, whereas the XG Boost model has a MAPE of 17.4 and an RMSE of 18998.57.

[5] `https://stephenallwright.com/good-mape-score/`

Table 7: Data Split and No. of trials- XG Boost

| Split | Seed | Metrics | |
|---|---|---|---|
| | | MAPE | RMSE |
| 80:20 | 123 | 17.83 | 18834.24 |
| | 321 | 16.81 | 18287.10 |
| | 1712 | 17.4 | 18998.57 |
| 70:30 | 123 | 17.83 | 18834.24 |
| | 321 | 16.81 | 18287.10 |
| | 1712 | 17.4 | 18998.57 |

Table 8: MAPE before and after hyperparameter tuning

| **Model** | **Before** | **After** |
|---|---|---|
| Random Forest | 34 | 17.01 |
| ELastic Regression | 34 | 34.37 |
| Support Vector Regressor | 22.80 | 22.78 |
| XG Boost | 17.37 | 17.4 |

# 7 Results and Discussion

Line charts are used to illustrate the performance of various machine learning models before and after hyperparameter tuning. To analyze model performance, two different evaluation metrics are used: Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE).

5The major goal is to use the Mean Absolute Percentage Error (MAPE) to assess the performance of hyperparameter tweaking on various machine learning models. Model IDs, MAPE values prior to tuning, and equivalent values after tuning are all included in the dataset. The chart positions the models on x-axis and MAPE values on y-axis.Thus, this chart provides concise yet comprehensive analysis of hyperparamter tuning.

6 The RMSE measure is employed similar to MAPE hyperparameter tuning. The data is well structured encompassing model names, RMSE values before tuning and corresponding values post tuning. Model names are on the x-axis while RMSE values are on the y-axis and disctinctive colours are applied to differentiate RMSE values before and after tuning.

7A special plot is developed to allow for a thorough comparison of MAPE values for all five machine learning models. The dataset is well-organized, with model designations and associated MAPE values. A line chart is crafted using data reshaping technique. The x-axis contains model identities, the y-axis has MAPE results, and the color variation represents the specific evaluation metric. When MAPE is evaluated, Random Forest is the best performing model since it has the lowest MAPE value. However, there is a minor variation in the results between the Random Forest and XG Boost models, 17.01 and 17.04, respectively.

Table 9: RMSE before and after hyperparameter tuning

| Model | Before | After |
|---|---|---|
| Random Forest | 30392.1 | 19405.75 |
| ELastic Regression | 28555.2 | 28552.68 |
| Support Vector Regressor | 23831 | 23668 |
| XG Boost | 19413 | 18998.57 |

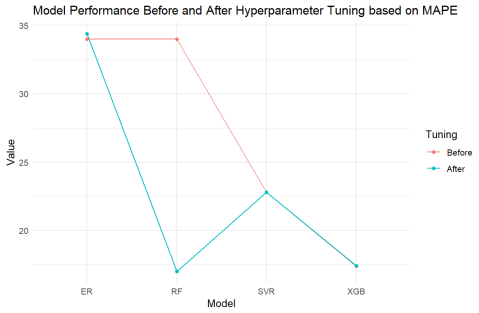


Figure 5: Hyperparameter Tuning: MAPE

8 Similar to the MAPE Comparison line chart, RMSE line graph is plotted mirroring the same procedure as MAPE adopting RMSE as the evaluation metric.The dataset carries model names and corresponding RMSE values. The RMSE value of the XG Boost model is the lowest amongst all the models. This indicates that the XG Boost exhibited better predictive performance than others. Next to XG Boost, is the Random Forest model. Random Forest has RMSE value of 19405.75 which is slightly higher than the XG Boost with 18998.57.

Thus, the intent to visually analyze the performance of machine learning models before and after hyperparameter tweaking using two separate evaluation metrics (MAPE and RMSE) is achieved.The charts reveal the success of the tuning procedure as well as the relative strengths of the models.

# 8 Conclusion and Future Work

## 8.1 Conclusion:

A comprehensive comparative study is performed to predict the price of pre-owned bikes in India. Through experimentation with various algorithms, the results show that the Random Forest and XG Boost models performed the best with the lowest MAPE and RMSE errors. The decision to use machine learning models rather than deep learning approaches for the investigation can be justified for the following reasons. Deep learning algorithms frequently require a big quantity of data to adequately train. Traditional machine learning models may perform better with a restricted dataset due to their capacity

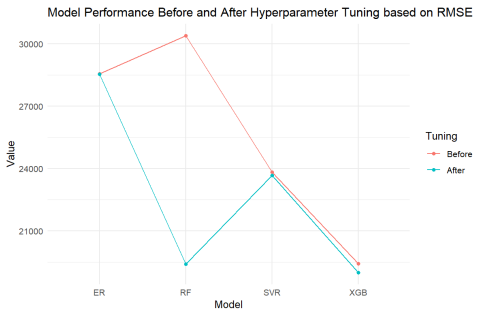


Figure 6: Hyperparameter Tuning: RMSE

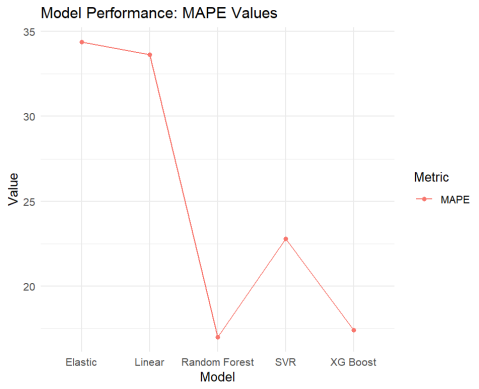


Figure 7: Model Performance Comparison based on MAPE

to generalize from smaller data volumes. Traditional models can obtain competitive results with simpler architectures for relatively basic tasks when deep learning's complexity is not required. Deep learning is often better suited to complicated tasks such as image and speech recognition. Deep learning algorithms are prone to overfitting, especially when data is scarce. Classical models, such as Random Forest and SVR, can provide better generalization and reduce the chance of overfitting, especially when the dataset is limited. A lot of research employed traditional models than complex ones and are considered as first choice for simpler problems. Also, their effectiveness is well documented as in literature review. Thus, while comparing to deep learning models, traditional models like Random Forest, Support Vector Regressor and Linear Regression can work well with limited data. They provide better predictive performance, require less computational resources amd also provide good interpretability.

## 8.2 Limitations of this research:

- The dataset used for this study is small. Since the dataset was small, there were no significant fluctuations in the MAPE and RMSE values when the split ratios were changed. To improve the model's performance, a dataset with a considerably larger

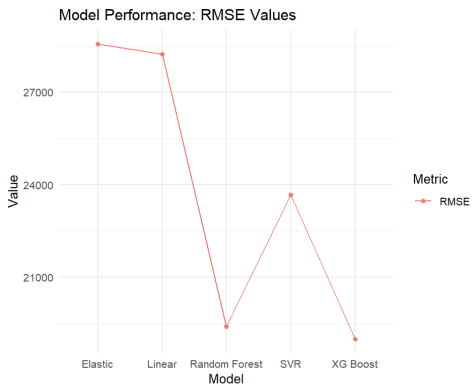


Figure 8: Model Performance comparison based on RMSE

size may be employed.

- Dataset used is specific to India.
- The features such as engine condition and maintenance of the bike have significant influence on the price of the used bike but were not considered in this study due to dataset limitation.

## 8.3 Future Work:

- The scope of features could be expanded. By adding in new features, the model performance could significantly increase.
- Deploying ensemble models can potentially reduce model inefficiency. By combining predictions from multiple models, this can be handled leading to better predictions.
- Development of user-friendly mobile application with this research can benefit people.
- With adequate data, this idea might be used in other locations outside India. Predictions can be improved by integrating geography-specific information.
- In real world scenarios, deep learning can manage enormous datasets.

# 9 Acknowledgment

I would like to express my deepest gratitude to Dr. Anh Duong Trinh for the consistent assistance he has provided throughout each and every step of this research effort. His guidance and advice have allowed me to gather extensive information, which has resulted in an abundance of ideas and perspectives for my research topic.

# References


Chai, T. and Draxler, R. (2014). Root mean square error (rmse) or mean absolute error (mae)?– arguments against avoiding rmse in the literature, *Geoscientific Model Development* **7**: 1247–1250.

Chavare, R., Joshi, R., Wagh, O., Vaishale, A. and Ingale, A. (2023). Car sales price prediction using mlr, random forest and support vector machine, *2023 International Conference for Advancement in Technology (ICONAT)*, pp. 1–4.

Chen, C., Hao, L. and Xu, C. (2017). Comparative analysis of used car price evaluation models, *AIP Conference Proceedings* **1839**(1): 020165.
**URL:** *https://doi.org/10.1063/1.4982530*

Han, S., Qu, J., Song, J. and Liu, Z. (2022). Second-hand car price prediction based on a mixed-weighted regression model, *2022 7th International Conference on Big Data Analytics (ICBDA)*, pp. 90–95.

Hankar, M., Birjali, M. and Beni-Hssane, A. (2022). Used car price prediction using machine learning: A case study, *2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC)*, pp. 1–4.

Hemendiran, B. and Renjith, P. N. (2023). Predicting the prices of the used cars using machine learning for resale, *2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pp. 1–5.

Kinadi, A. F., Andreswari, R., Sutoyo, E., Nugraha, R. and Kamil, A. A. B. (2022). Used car price prediction in surabaya using random forest regressor algorithms, *2022 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS)*, pp. 1–4.

Li, Y., Li, Y. and Liu, Y. (2022). Research on used car price prediction based on random forest and lightgbm, *2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA)*, pp. 539–543.

Longani, C., Potharaju, S. P. and Deore, S. (2021). *Price Prediction for Pre-Owned Cars Using Ensemble Machine Learning Techniques*.

Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S. and Boonpou, P. (2018). Prediction of prices for used car by using regression models, *2018 5th International Conference on Business and Industrial Research (ICBIR)*, pp. 115–119.

Narayana, C. V., Likhitha, C. L., Bademiya, S. and Kusumanjali, K. (2021). Machine learning techniques to predict the price of used cars: Predictive analytics in retail

business, *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1680–1687.

Narayana, C. V., Madhuri, N. O. G., NagaSindhu, A., Aksha, M. and Naveen, C. (2022). Second sale car price prediction using machine learning algorithm, *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, pp. 1171–1177.

Saini, I. S. and Kaur, N. (2023). Comparison of various regression techniques and predicting the resale price of cars, *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 857–861.

Satapathy, S. K., Vala, R. and Virpariya, S. (2022). An automated car price prediction system using effective machine learning techniques, *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, pp. 402–408.

Shaprapawad, S., Borugadda, P. and Koshika, N. (2023). Car price prediction:an application of machine learning, *2023 International Conference on Inventive Computation Technologies (ICICT)*, pp. 242–248.

Varshitha, J., Jahnavi, K. and Lakshmi, C. (2022). Prediction of used car prices using artificial neural networks and machine learning, *2022 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–4.

Wang, A., Yu, Q., Li, X., Lu, Z., Yu, X. and Wang, Z. (2022). Research on used car valuation problem based on machine learning, *2022 International Conference on Computer Network, Electronic and Automation (ICCNEA)*, pp. 101–106.

Wang, F., Zhang, X. and Wang, Q. (2021). Prediction of used car price based on supervised learning algorithm, *2021 International Conference on Networking, Communications and Information Technology (NetCIT)*, pp. 143–147.