Global Perspectives on Used Car Price Prediction: A Machine Learning-based Comparative Study of India, Pakistan, and Germany's Automotive Market

Manjusha Iyer
Mechanical and Automation Engineering
IGDTUW
New Delhi, India
manjusha042btmae22@igdtuw.ac.in

Nandini Chaturvedi
Mechanical and Automation Engineering
IGDTUW
New Delhi, India
nandini044btmae22@igdtuw.ac.in

Samiya Malhotra
Mechanical and Automation Engineering
IGDTUW
New Delhi, India
samiya056btmae22@igdtuw.ac.in

***Abstract*— this research paper presents a comprehensive study on the prediction of used car prices using machine learning techniques in three distinct countries: India, Pakistan, and Germany. The rise of used cars sales is exponentially increasing. Car sellers sometimes take advantage of this scenario by listing unrealistic prices owing to the demand. Therefore, arises a need for a model that can assign a price for a vehicle by evaluating its features taking the prices of other cars into consideration. Using this knowledge base, accurate price prediction models can provide valuable insights to both buyers and sellers in these diverse automotive markets. Leveraging a variety of machine learning algorithms, this study aims to explore the challenges and opportunities associated with predicting used car prices across different economic and cultural contexts. The analysis considers a range of features, including car specifications, mileage, and year of manufacture, to develop models tailored to each country's market conditions.**

Keywords*— *Machine Learning, Supervised Learning model, Price Prediction, Multiple Linear Regression, Decision Tree Regression, KNN Model, Random Forest Regression, Light Gradient Boosted Machine and XGBoost Regression.

I. INTRODUCTION

The global automotive industry is undergoing a transformative shift, increasingly characterized by global mergers and relocation of production centers to emerging developing economies, leading to a growing interest in pre-owned vehicles across the world. As individuals seek cost-effective alternatives and sustainable mobility solutions, used cars have gained prominence as a viable option. Accurate prediction of used car prices plays a pivotal role in facilitating informed decision-making for both buyers and sellers. Machine learning algorithms have emerged as powerful tools for developing predictive models that capture the complex relationships between car attributes and their market values. The advantage of machine learning (ML) is the ability of a computer to learn without explicit instructions using mathematical models and processed data [1]. Artificial intelligence is a subset of machine learning. The data are analyzed using pattern identification algorithms, which are then used to create predictive models. Similar to a human, the conclusions of machine learning enable accurate prediction with as much data and experience as possible. With machine learning, the model can adapt to situations where data are constantly changing or coding a solution is not feasible.

In this context, our research embarks on a journey through the landscape of predictive modelling, employing ML methodologies to estimate the prices of used cars in three distinct countries: India, Pakistan, and Germany. Used car price prediction is a complex challenge influenced by an array of factors, including vehicle age, mileage, horsepower, brand, model, features, and market trends. Leveraging the power of ML, we aim to unravel the intricate relationships among these variables and establish predictive models that can accurately estimate prices. Moreover, this research takes a step further by delving into the variations observed when these models are applied to diverse datasets representing different national markets.

The accurate estimation of used car prices is a critical concern for both consumers and stakeholders in the automotive industry. With the emergence of machine learning techniques, predictive modelling has become an essential tool for estimating these prices, considering the numerous influencing factors involved.

II. LITERATURE REVIEW

In this literature review, we delve into recent studies that apply machine learning algorithms to predict used car prices, spanning many distinct countries. Many studies on understanding pre-owned vehicle market requirements have been carried out for India as well as other nations such as Indonesia [2], Turkey [3], Croatia [4] and Mauritius [5]. The works discussed contribute to a deeper understanding of the challenges and opportunities in this domain.

In the context of used car sales in Indonesia, Puteri and Safitri [2] investigated the applicability of linear regression. Their study serves as a foundation, showcasing the utilization of a simple baseline model to analyze used car sales patterns. While linear regression provides insights into the linear relationships, it may not fully capture the complexities inherent in car pricing. This research serves as a reminder of the need to explore more advanced techniques, as addressed by subsequent studies.

Pudaruth [5] used different machine learning algorithms, which are multiple linear regression analysis, k nearest neighbors, decision trees, and naïve bayes for

predicting car price in Mauritius. This author collected the dataset manually from local newspapers in less period. He considered the following variables to create a model which are brand, model, cubic capacity, mileage in km, manufacturing year, exterior colour, transmission type and price. But, the author found out that Naive Bayes and Decision Tree were unable to handle numeric values. Also, fewer records were used in his model. Accuracy was reached 70%.

Patil et al. [6] conducted a study focusing on the estimation of car prices using a diverse range of machine learning algorithms. The authors highlight the significance of employing various algorithms to capture intricate relationships between car specifications, age, condition, and market trends. By exploring beyond conventional linear regression, their research underscores the potential of more sophisticated models in improving price prediction accuracy. The paper serves as a guide for researchers and practitioners aiming to harness the capabilities of machine learning in the context of used car pricing.

Ganesh (2019) [7] explored the application of supervised learning techniques in predicting used car prices. The study emphasizes the need for feature selection and engineering to enhance prediction accuracy. The research contributes to the understanding that predictive modelling goes beyond algorithm selection; it necessitates careful consideration of data preprocessing steps. Ganesh's work underscores the importance of a holistic approach in constructing effective predictive models.

Samruddhi and Dr. Kumar [8] ventured into the use of K-Nearest Neighbour (KNN) based models to predict used car prices. This work introduces the concept of utilizing the proximity of similar cars for price estimation. Their study emphasizes the importance of selecting an appropriate value for "k" and dealing with noise in the dataset. By implementing KNN, the authors provide a novel perspective, shedding light on the potential of leveraging similarity-based approaches for price prediction. introduces the concept of utilizing the proximity of similar cars for price estimation. Their study emphasizes the importance of selecting an appropriate value for "k" and dealing with noise in the dataset. By implementing KNN, the authors provide a novel perspective, shedding light on the potential of leveraging similarity-based approaches for price prediction.

Vanitha S [9] studied the comparative results of regression and ML models applied on a dataset that contains used car details taken from eBay- Kleinanzeigen, a German e-commerce company and made several noteworthy observations. They showed Boosting algorithms, particularly XGBoost and Gradient Boosting, demonstrated superior performance with less overfitting. Random Forest, while effective, required more computational time and showed some overfitting. K-Nearest Neighbors (KNN) also demanded more time for optimization but exhibited less overfitting compared to Random Forest. Multiple Linear Regression (MLR) was computationally efficient but offered lower accuracy.

The studies reviewed collectively highlight the evolving landscape of used car price prediction using machine learning techniques. These works contribute to the broader understanding of the challenges and potentials of predictive modelling in diverse geographical contexts. From advanced machine learning algorithms to foundational linear regression and innovative approaches like KNN, these studies provide insights that guide future research endeavours in achieving more accurate and robust used car price predictions across different countries.

However, the challenge of comparing multi-nation market behaviour has not been explored enough. The comparative aspect of our study holds substantial significance. India, Pakistan, and Germany offer diverse economic landscapes, cultural as well as consumer preferences, regulatory environments and automotive market dynamics. These regional disparities translate into variations in the factors that impact used car prices. Our study contributes to the burgeoning field of comparative ML by elucidating how differing datasets from these countries necessitate adaptable modelling techniques to yield accurate predictions.

Through a systematic exploration of ML algorithms including regression-based models, decision trees, and ensemble methods, our research not only seeks to uncover predictive accuracy but also delves into the interpretability and transferability of these models across diverse markets. Work by both Naidu Totakura et al [10] and Prof Bharambe et al [11] examine application of multiple machine learning techniques for predicting price.

By analyzing how various ML techniques respond to different national datasets, we offer insights into the complexities that underscore the relationship between ML, regional disparities, and predictive modelling.

By examining these markets through the lens of machine learning, this study aims to uncover global perspectives on used car price prediction, shedding light on both common trends and country-specific variations.

The remainder of this paper is organized as follows:

Section III outlines the methodologies and dataset used for training and evaluating the machine learning models. Section IV presents the results of the exploratory analysis EDA, conducted on the respective datasets of India, Pakistan, and Germany. Section V discusses the implications of the findings and provides insights into the factors that contribute to accurate used car price prediction. Finally, Section VI concludes the paper with a summary of the key takeaways and avenues for future research.

III. METHODOLOGIES AND DATASET

In this section, we delve into the methodologies employed to develop, train, and evaluate the machine learning models for predicting used car prices in the context of India, Pakistan, and Germany. We also provide an overview of the datasets used, including their sources and the preprocessing techniques applied.

A. Methodologies

The prediction of used car prices involves a multi-faceted approach that encompasses various machine learning algorithms, preprocessing techniques, and evaluation procedures. Significant research has been done in predicting the price of used cars. In most of the research papers, ML algorithms used are Linear Regression (LR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Light Gradient Boosted Machine (LGBM) and k-nearest neighbor (KNN). In addition to these models, another variation of Gradient Boosting model called XGBoost (XGB) regression model is considered in this study. All these supervised ML models LR, DTR, RFR, LGBM and XGB are selected for comparison to identify an optimal algorithm for price prediction of used cars. Also,

the range of methodologies selected, is to ensure robust model development and accurate price predictions across different countries' markets. The process of building a robust model is depicted in Fig.1.

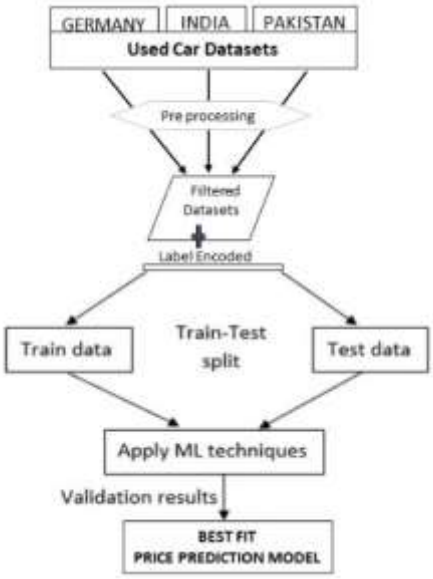

Fig. 1. Building price prediction model

Machine Learning Models:

We explore the performance of several machine learning algorithms known for their effectiveness in regression tasks. The models applied include:

Multiple Linear Regression:

A linear approach to model the relationships between input features and target prices.

Decision Tree:

A tree-based model that recursively splits the data into subsets based on feature conditions.

Random Forest Regression:

An ensemble model combining multiple decision trees to improve prediction accuracy.

K-Nearest Neighbors (KNN):

A non-parametric algorithm that predicts based on the proximity of training instances.

Gradient Boosting:

A boosting algorithm that combines weak learners to create a strong predictive model.

XGBoost Regression:

An optimized gradient boosting implementation known for its speed and accuracy.

XGboost Algorithm

The XGboost, a scalable tree boosting system, uses a limit gradient lifting algorithm. It mainly uses the training set to predict the change and trend of target variables in the future. The essence of this model is to construct multiple Decision trees. The model predicts each tree separately, and finally combines the prediction results of each tree to achieve the final prediction value. Taking the decision tree as the base learner, multiple weak learners are constructed, and then the model is trained continuously along the descending direction of the gradient. The details of the model are as follows, and the structure is shown in Fig.2. [12]

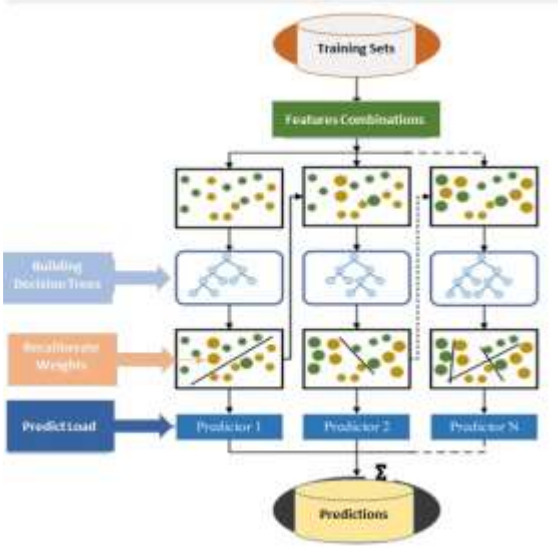

Fig. 2. Schematic representation of XGBoost model

The mechanism of XGboost is keep adding and training new trees to fit residual errors of last iteration. A predicted value is assigned to each instance by adding all corresponding leaves' scores together:

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i),\ f_k = F \tag{1}$$

In (1), k is a tree of the decision tree, f_k is an independent function in the function space, and F is the function space with the following equation:

$$F = \{f(x) = w_{q(x)}\} \tag{2}$$

In (2), q(x) indicates that the sample x is assigned to a leaf node and w is the leaf node weight.

The objective function of the XGboost algorithm contains two parts: the first part is training loss, which describes the difference between the predicted and true load of the model in the test set. The second part is regularization, which not only has the effect of preventing overfitting, but also plays a crucial role in controlling the complexity of the model. The objective function equation is:

$$L(\phi) = \sum_{i} l(\hat{y}_i, y_i) + \sum_{k} \Omega(f_k) \tag{3}$$

$$\Omega(f) = \Upsilon T + \frac{1}{2}\lambda \parallel \omega \parallel^2 \tag{4}$$

In (3), T denotes the number of leaf nodes, ω denotes the leaf node weights, λ denotes the regularization penalty term coefficient, and Y is the penalty factor for the number of leaf nodes.

The minimization of the objective function is achieved through continuous iterations, and the objective function after each iteration is:

$$L^t = \sum_i l(y, \hat{y}_i^{t-1}) + f_t(x_i) + \Omega(f_t) \tag{5}$$

The objective function is expanded as follows:

$$L^t \simeq \sum_{i=1}^{n} \left[l\left(y, \hat{y}_i^{t-1}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \qquad (6)$$

In order to get the minimum value of the objective function and make its derivative equal to 0, the weight of each leaf node is:

$$\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \qquad (7)$$

Substituting into the objective function, the minimum loss of the solution is:

$$\tilde{L}' = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} \qquad (8)$$

Based on the above, the construction steps of XGboost prediction model are as follows:

1. The model starts the initial iteration, and a sub prediction model is constructed in each iteration.
2. Before each iteration, calculate the first derivative g_i and the second derivative h_i of the loss function at each training sample point.
3. Generate a new decision tree, and calculate the corresponding prediction value of each leaf node through (6).
4. After each iteration, the newly generated model is added to the previous model. After several rounds of iteration, the final prediction model can be obtained.

Machado, Karray and Sousa [13] in 2019 concluded in their study that LGBM provides higher accuracy than regression algorithms in prediction problems. Fan, Ma, Wu, Zhang, Yu and Zeng in 2019 concluded GBRM yields high accuracy while using large datasets for predictions [14]. Both these studies motivates this study to expect GBRM algorithm as an optimal algorithm in predicting price of used cars and compare it with performance of XGB algorithm.

The performance of these 6 selected algorithms, obtained after implementing on cleaned dataset for each country individually, are compared at two levels

a) based on the performance metrics (R-squared or Mean Square Error (MSE)), the best-fit model for the given country is discovered and

b) based on cross validation (CV) score among the 3 best-fit models, in case of same accuracy score of the best-fit models.

Dataset Splitting:

The preprocessing of each of the dataset has been performed to handle the missing and null values besides cleansing outliers and data aggregation. This filtered dataset is then divided into training and testing subsets using the "train-test-split" technique. These subsets are trained and tested on the selected ML algorithms for used cars price prediction. The "test-train-split" methodology is employed to partition the dataset into training and testing subsets. This ensures that the models are trained on one subset and evaluated on another, minimizing overfitting and enabling robust performance evaluation.

Encoding Categorical Variables:

Categorical variables are encoded using the label encoding technique to convert categorical data into numerical format suitable for all the machine learning algorithms [10]

B. Datasets File

Accurate prediction models require high-quality datasets that accurately represent the used car markets in the respective countries. The datasets used for this study are collected through web scraping from reputable automotive websites:

→ India Dataset (CarDekho): The India dataset is obtained through web scraping of CarDekho, a prominent automotive portal. The dataset includes various features such as car specifications, mileage, year of manufacture, and additional attributes relevant to the Indian market. The size of the dataset was 15,411 rows and 14 columns.

→ Pakistan Dataset (PakWheels): The Pakistan dataset is sourced from PakWheels, a well-known automotive platform. It includes data on used cars available in Pakistan, encompassing features specific to the Pakistani context. The dataset contained 76,690 records of 10 variables.

→ Germany Dataset (AutoScout24): The Germany dataset is collected by scraping AutoScout24, a reputable website for buying and selling cars in Germany. It comprises features relevant to the German automotive market. The shape of the dataset is (46405, 9).

C. Preprocessing and Data Analysis Tools

The datasets undergo rigorous preprocessing and exploratory data analysis using a variety of Python libraries:

- ✓ Pandas is utilized for data manipulation and analysis.
- ✓ NumPy is employed for numerical computations and array operations.
- ✓ Matplotlib.pyplot and Seaborn are used for data visualization, allowing us to gain insights into feature distributions and relationships.
- ✓ scipy.stats is applied for statistical analysis and hypothesis testing.

These tools collectively enable us to clean, pre-process, and analyze the data effectively, preparing it for training and evaluating the machine learning models. As mentioned earlier, the preprocessing steps are handling the missing and null values either by imputation (wherever applicable) or by dropping the concerned record. For cleansing outliers the z-score measure for the concerned numerical feature was obtained. Records were removed if the calculated z-score does not lie between (-3, 3). Box-plot approach also helped in identifying outliers. To ensure uniformity in unique values for variables common to all 3 datasets, data aggregation was carried out. The final step in feature-engineering was identifying the underlying correlation among the numerical variables through heat map or

generating correlation table. Subsequently the filtered datasets are applied on selected ML algorithms and then the performance of the selected algorithms are compared based on performance metrics.

In summary, Section III outlines the methodologies employed in this research, encompassing the machine learning algorithms, dataset acquisition through web scraping, preprocessing techniques, and the Python libraries utilized for data analysis. These methodologies set the foundation for the subsequent sections where we present the results and implications of our study on predicting used car prices across India, Pakistan, and Germany.

IV. Exploratory Data Analysis

In this section, we present the results of the exploratory data analysis (EDA) conducted on the individual datasets of India, Pakistan, and Germany. EDA serves as a crucial preliminary step to understand the characteristics of the data, identify patterns, outliers, and relationships among features. Through EDA, we gain insights that guide the subsequent steps of preprocessing, feature selection, and model development.

A. *India Dataset EDA*

The EDA process for the India dataset, sourced from CarDekho, revealed several key observations:

Feature Distribution: We analyzed the distribution of features such as "Mileage," "Year," and "Engine Capacity." The majority of cars in the dataset exhibited a skewed distribution of mileage, with the majority falling within a certain range.

Price Distribution: The distribution of used car prices exhibited a right-skewed pattern, with a majority of cars priced in the lower range. We identified potential outliers with unusually high prices that required further investigation.

Correlation Analysis: Correlation analysis revealed interesting relationships between features. For instance, we observed a negative correlation between "Mileage" and "Price," indicating that higher mileage tends to correlate with lower prices.

B. *Pakistan Dataset EDA*

The EDA process for the Pakistan dataset, obtained from PakWheels, yielded valuable insights:

Feature Insights: We examined the distribution of features such as "Transmission," "Fuel Type," and "Engine Capacity." There was a variety of fuel types represented in the dataset. The distribution of car years (Fig 4) displayed a diverse range, indicating the presence of outliers across different manufacturing years. The oldest used car recorded was 83 years old which was dropped as part of feature engineering.

Price Variation: The distribution of used car prices indicated variations across different car categories. Luxury cars tended to have higher prices compared to compact and economy models.

Feature Relationships: We explored relationships between categorical features and their impact on price variation. For instance, we observed that cars with automatic transmissions tended to have higher average prices than those with manual transmissions.

C. *Germany Dataset EDA*

The EDA process for the Germany dataset, scraped from AutoScout24, provided insights specific to the German automotive market:

Feature Overview: We studied features such as "Body Type," "Fuel Type," and "Year." The distribution of body types showed a diverse range, encompassing sedans, SUVs, and hatchbacks. The distribution of car years displayed least range as compared to other 2 datasets, indicating the presence of various car models across different manufacturing years.

Price Distribution: The distribution of used car prices exhibited variations based on body types. Luxury vehicle categories displayed a wider range of prices compared to more common body types.

Correlation Analysis: The correlation matrix obtained is shown below in Fig 3.

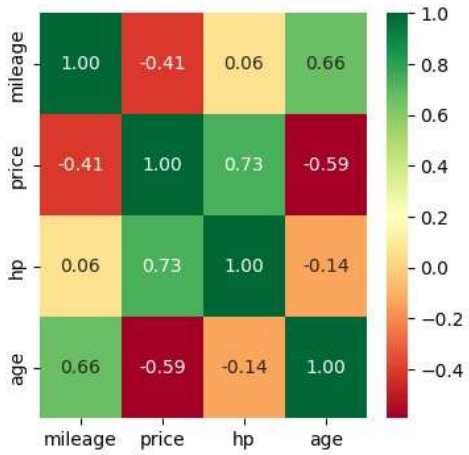

Fig. 3. Heat map for Germany dataset

Geographical Variation: We analyzed the distribution of cars across different regions of Germany. This analysis revealed potential regional price variations, suggesting the influence of geographical factors on used car prices.

Given are the visualization highlights of the datasets under study. The Fig. 4 shows distribution of number of cars available in the dataset categorized by age.

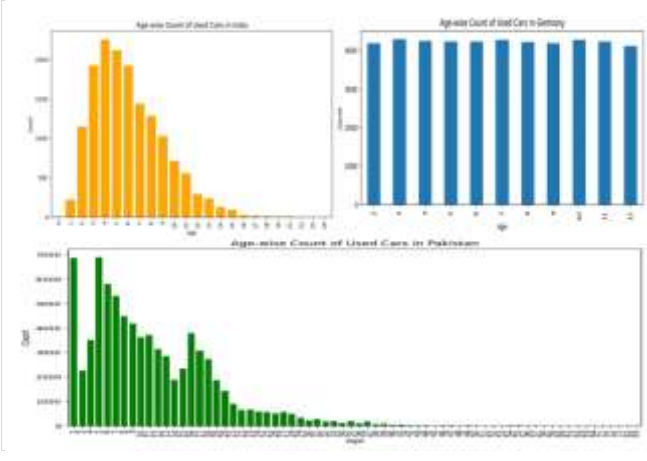

Fig. 4. Distribution of used cars age feature

It is interesting to look closer at correlation between ages of the car with the quoted selling price. A consistent trend emerges from the provided graphs in Fig.5 across all three datasets: the highest prices are consistently associated with cars aged 2-3 years. However, an interesting anomaly surfaces in both Indian and Pakistani datasets. In the Indian context, the mean price of 22+ old car is very close to mean price of cars aged between 10 and 12. In the Pakistani data the price of a 16-year-old car exceeds that of a 15-year-old car. These anomalies underscores a crucial point: car pricing is not solely contingent upon age but is influenced by a multitude of other factors. Variables such as the car's brand, overall condition, and other relevant factors significantly contribute to its market value. Also presence of outliers in either Price or Age column or both columns may also be the reason for the anomaly.

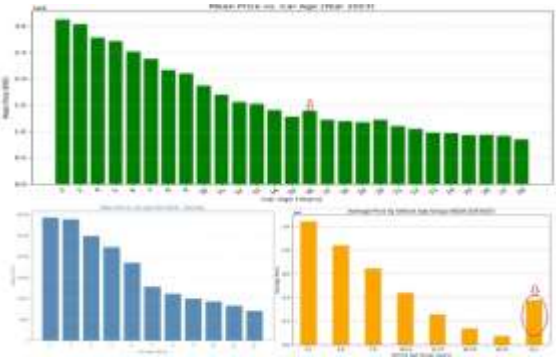

Fig. 5. Age wise price variation across the three nations

The boxplot representation of the relationship between the transmission type and Price is displayed as Fig 6. The graphical data unmistakably illustrates that, across all three countries, automatic cars command a higher price than their manual counterparts. However, it's worth noting a unique pattern in the case of Germany, where manual cars are priced higher than semi-automatic ones. This phenomenon may be attributed to the strong preference for manual transmissions among the populace in Germany, leading to increased demand and, subsequently, higher prices for manual cars compared to semi-automatic options.

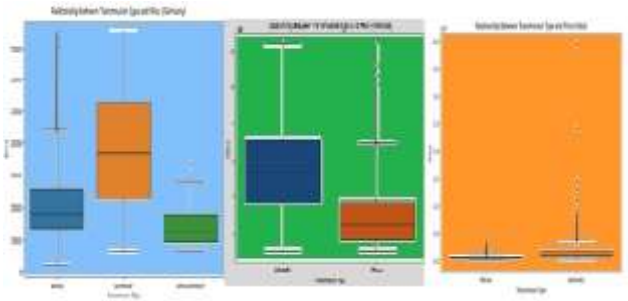

Fig. 6. Relationship between the transmission type and price of the cars

The availability of extensive information on car mileage along with the kind of fuel used by these used cars helped in finding any notable performance differences between petrol and diesel cars in terms of mileage and price among the 3 different datasets. Fig. 7 depicts the pattern in the form of scatterplot and boxplot. These visualization tools allows to comprehend the quality of data values and convert the datasets into clean, dependable data frame. Relationship between the Transmission type and Price of the cars

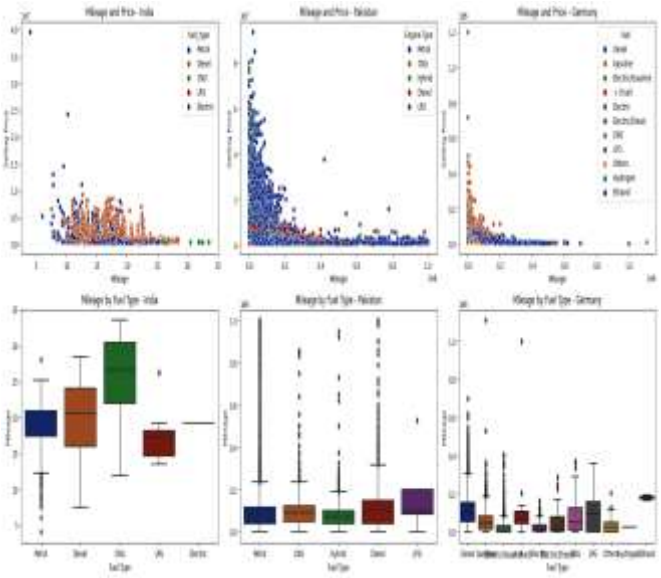

Fig. 7. Country wise visualization of mileage vs price and mileage vs fuel type

It is an obvious assumption that in all 3 countries the majority of used cars are powered by petrol/gasoline and diesel. The second-hand electric cars are for resale in Germany and India but absent in Pakistan. The distribution of available used cars based on fuel type is shown in Fig.8.

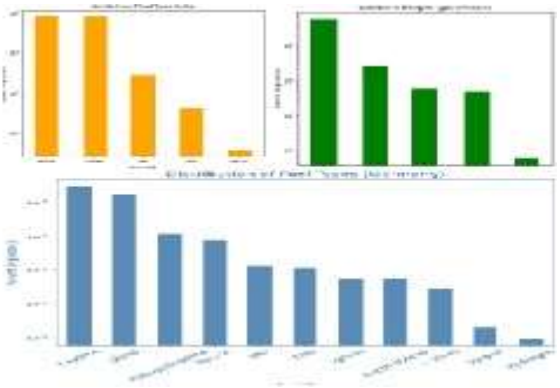

Fig. 8. Country wise distribution of all fuel types

The availability of so many car models in all the three datasets explained why the top 5 most popular models (Table 1) have low presence in the datasets. In the German and Pakistani market, there were only near about 2% of all models and 5% in Indian market.

TABLE I. TOP 5 CAR MODELS OF INDIA, PAKISTAN AND GERMANY

model	brand	count	model	brand	count	model	brand	count
i20	Hyundai	906	Corolla GLi 1.3 VVTi	Toyota	2954	Golf	Volkswagen	1452
Swift Dzire	Maruti	890	Civic Oriel 1.8 i-VTEC	Honda	2177	Corsa	Opel	1451
Swift	Maruti	781	City 1.3 i-VTEC	Honda	1648	Fiesta	Ford	1289
Alto	Maruti	778	Corolla XLi VVTi	Toyota	1597	Astra	Opel	1151
City	Honda	757	Vitz F 1.0	Toyota	1493	Focus	Ford	985

The leading car brands are Maruti, Toyota and Volkswagen in India, Pakistan and Germany respectively.

In summary, the exploratory data analysis conducted on the datasets of India, Pakistan, and Germany provided valuable insights into the distribution of features, price trends, and relationships among variables. These insights guide our subsequent steps of data preprocessing, feature engineering, and the selection of appropriate machine learning algorithms. The findings from EDA contribute to the development of accurate used car price prediction models tailored to the unique characteristics of each country's automotive market.

V. FINDINGS AND ITS IMPLICATION

The finding of factors that contribute to accurate used car price prediction for each country may provide detailed insights for each country.

A. Features

Common Features: The common features across all three datasets (Price, age of the car, fuel, mileage, make, model, transmission) indicate that these factors are universally important in determining used car prices. This suggests that the age, condition, brand and specifications of a car play a significant role in its pricing across different markets.

Country-Specific Features:

Germany: "hp" and "offerType" are specific features for Germany. Higher horsepower (hp) is associated with higher prices, which is common as more powerful cars tend to be priced higher. The "offerType" may reflect the influence of the type of listing on the price, indicating that certain listing types might fetch higher prices.

India: "hp," "seller_type," "engine capacity," "km_driven," and "seats" are specific features for India. Higher horsepower, engine capacity, and features like seating capacity might contribute to higher prices. "Seller_type" could indicate that cars sold by dealers or owners might be priced differently. "km_driven" and "age" are expected to negatively impact prices.

Pakistan: "engine capacity" and "city" are specific features for Pakistan. Engine capacity might have a significant influence on prices, and the "city" might reflect regional pricing variations. The model's equation suggests that these features have a relatively lower impact on prices in Pakistan compared to other countries.

B. Model Building and Comparison:

Building a well-suited price prediction mode where in all the 3 countries dataset had combination of both numerical and categorical features of a car, involved 3 data-processing steps-

Use ‘Label encoder’ method to transform categorical columns into numerical columns.

Variable splitting of the dataset – Separate target column(y) from feature array(X).

Data splitting – use test-train splitting technique to create training set and test set with test-size= 0.2. The standard test-size = 0.2 allows the feature matrix to be split in the ratio of 80:20 for training set : testing set using random selection.

The next step consisted in applying these 5 models besides multiple linear regression model. Comparison of all the 6 prediction scheme was based on testing score metrics such as R^2 scores of each of the model since testing score quantifies how well the model's predictions match the actual values.

All the required techniques preprocessing, model_selection, linear_model, tree, neighbors, ensemble are imported from python library sklearn. For applying XGB, xgboost library was used. The metrics module from sklearn provided the statistical measures under the module model_selection. From the ensemble module, functions such as RandomForestRegressor, GradientBoostingRegressor are imported.

C. Result of MLR and XGB Model:

Multiple Linear Regression (MLR):

The purpose of linear regression analysis is to measure the intensity of the relationship between two or more variables and contain predictions / estimates of the value of Y and the value of X. The general form of multiple linear regression equations includes two or more variables [15].

In this study the country-wise Multiple Linear Regression Equations obtained are -

y (Price_German)= 12618.87 -0.04 * mileage + 14.38 * make + 1.45 * model + 1438.90 * fuel -2732.38 * transmission -892.40 * offerType + 116.69 * hp -1007.08 * age

y (Price_Indian) = 450979.42 + 935.60 * make – 665.38 * model -52568.31 * age -0.65 * kmdriven -39743.29 * fuel -162258.45 * transmission -5260.29 * mileage + 116.45 * engine_capacity + 7953.57 * hp

y (Price_Pakistan) = -2310.73 + 0.00 * make + 0.00 * model -6.52 * transmission -0.32 * Fuel + 0.02 * engine Capacity(CC) -0.00 * mileage(km) -0.00 * City + 1.15 * age

The coefficients in the regression equations indicate the impact of each feature on the predicted car price. For example, in the Germany equation, higher make, model, fuel, and horsepower contribute to higher prices, while higher mileage, age, and certain transmission types lead to lower prices.

R^2 or the determination coefficient is how much the ability of all independent variables to explain the variance of the dependent variable y (price) [13]. Tool to measure the level of compatibility / perfection of the regression model or to state the proportion of total diversity of the variable y values which can be explained by the values of variable X (feature columns) through that mathematical relationship. And the R^2 measure for these 3 Multiple LRs are 0.83, 0.75 and 0.71 for Germany, India and Pakistan respectively. These measures indicated need for finding a more suitable predictive model.

The test-score (R^2) calculated for each model is taken as model accuracy score.

TABLE II. MODEL ACCURACY SCORE

Nation	Model					
	LR	*DT*	*KNN*	*RFR*	*GBRM*	*XGB*
India	0.75	0.89	0.92	0.93	0.92	0.94
Pakistan	0.71	0.94	0.94	0.95	0.92	0.96
Germany	0.83	0.89	0.92	0.92	0.91	0.94

The finding that XGBoost models outperforms all other models applied in all three countries underscores the non-

linear relationships between features and used car prices. XGBoost's ability to capture complex interactions makes it well-suited for this prediction task.

In order to compare and understand further the level of perfection of the predicted XGBoost regression model, among these 3 different country specific data, another evaluation measure the cross-validation score was obtained. The cross-validation score assesses how well a model generalizes to new data.

In the stratified cross-validation approach, each of the 3 dataset is split into ‘k=5’ distinct subsets or "folds." This decision is made to comprehensively evaluate our model's performance across different portions of the data. Each fold represents approximately 20% of the dataset, ensuring an equitable distribution of data for training and testing purposes, while preserving the original class distribution. The choice of ‘k=5’ folds strikes a balance between robust evaluation and computational efficiency.

The cross-validation is executed as follows:

The XGB model is trained on four of these folds (the training set). The trained model is then used to make predictions on the fifth fold (the testing set). This process is repeated five times, with each fold serving as the testing set exactly once, while the remaining folds constitute the training data for each iteration. The outcome of the cross-validation procedure is a set of five R-squared scores, each corresponding to a distinct fold and the mean R^2 score referred as cross-validation (CV) score is taken as performance metric.

The calculation outcome shows average Cross-Validation-R^2 (CV) Score was highest for Pakistan (0.96) followed by Germany (0.94) and India (0.93). The high CV scores in all regions suggest that the model generalizes well to new, unseen data, which is a positive sign for its application in real-world scenarios.

Top 3 Feature Importance:

Importance of features reveal importance scores for each attribute used in a predictive model. The provided score indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance.

The top three important features from the 3 XGBoost models (Fig. 5) give us insight into the most influential factors affecting used car prices in each country.

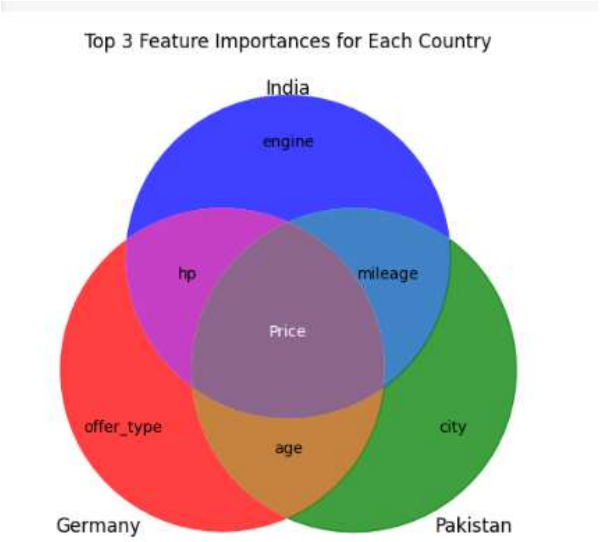

Fig. 9. Top 3 feature importance for each country

In Germany, "age," "hp," and "offer_type" have the highest importance, emphasizing the significance of car age and engine power.

In Pakistan, "age," "city," and "mileage" are most important. Location and car age seem to play crucial roles in pricing.

In India, "hp," "engine," and "mileage" hold the highest importance, indicating the influence of engine power and efficiency.

A plausible takeaway from this could be that Indian buyers prioritize performance and low mileage and they can get a relatively newer car with good performance at a reasonable price, even if the used car is a bit older. On the other hand, both Pakistanis and Germans may opt for ‘aged’ used cars with plausible reasoning such as Pakistanis prefer older used cars for it is relatively low priced and Germans often prefer newer vehicles with the latest safety and technology features. This aligns with the country's automotive culture, including its famed luxury and sports car manufacturers. The common feature preference (attribute ‘hp’) between Indians and Germans indicate both appreciate high-performance cars with powerful engines. Further this Indian consumer behavior may be attributed the contribution of many different factors like sales incentives, introduction of new models as well as variants coupled with easy availability of low cost finance with comfortable repayment options.

VI. CONCLUSION

In response to the expanding car resale global market this paper, proposes a peek into the challenge of comparing three-nation namely, India, Pakistan, and Germany second hand car market behavior. Six forecasting models were applied and based on feature filtering and error corrections, and model accuracy metrics obtained the best fit model for each country. The main conclusions are as follows.

A. Implications and Insights:

Engine Specifications: Engine power (hp), engine capacity, and mileage play key roles in predicting car prices across all three countries. More powerful and efficient engines tend to increase a car's value.

Age and Condition: Car age consistently impacts prices negatively. Older cars are generally priced lower due to wear and tear.

Location Matters: In Pakistan, the city where a car is sold influences its price. Regional preferences and economic factors may contribute.

Listing Type: In Germany, the listing type (offer_Type) affects prices. This could be due to differences between private and dealer sales.

Brand Value: Brand (make) and model significantly influence prices. Well-established brands and popular models tend to command higher prices.

Complex Relationships: XGBoost models outperform linear regression due to their ability to capture complex relationships between features and prices.

Overall, a combination of technical specifications, condition, brand value, and market dynamics determine used car prices. The differences observed in each country's model and feature importance highlight the need for localized models that consider region-specific factors when predicting car prices accurately.

B. Future Scope:

Areas of enhancement that can be considered for future research based on the findings and insights from the current study:

It would be desirable to study the impact of features such as car history, maintenance records, and accident history on the accuracy of predictions. More relevant features such as, features related to advanced safety technologies, infotainment systems, and interior/exterior aesthetics may also provide valuable insights. The dynamic nature of the used car market in a global setup can be better represented by conducting a temporal analysis to understand demand and sales of all kinds of automobiles and how car prices change over time. Factors like inflation, economic conditions, and market trends should be analyzed and be part of any price predicting model. Further exploring the integration of external data sources such as economic indicators, gas prices, and consumer sentiment indexes besides those already mentioned factors could lead to more robust and universally acceptable price model. Another area of interest could be to consider segmenting the used car market into different categories (e.g., economy, luxury, SUVs) and developing separate models for each segment. This approach can capture finer nuances in pricing based on market segments. These areas of enhancement offer exciting opportunities to build upon the current research and further improve the accuracy and applicability of used car price prediction models.

REFERENCES

[1]. Aurelien Geron. Hands-On Machine Learning with Scikit-Learn, Keras, TensorFlow, 2nd Edition, O'Reilly Publication.

[2]. Puteri, C. K., & Safitri, L. N. (2020). Analysis of linear regression on used car sales in Indonesia. Journal of Physics: Conference Series, 1469, 012143.

[3]. M. C. Satioglu, Y. Ar and B. Tugrul, "Automobile Price Prediction in Turkey Marketplace with Linear Regression," 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 2021, pp. 329-333, doi: 10.1109/ISMSIT52890.2021.9604688.

[4]. Bukvić, L.; Pašagić Škrinjar, J.; Fratrović, T.; Abramović, B (2022). Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning. Sustainability 2022, 14, 17034.

[5]. Sameerchand Pudaruth. Predicting the Price of Used Cars using Machine Learning Techniques. International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753-764.

[6]. Patil, R., Bade, R., Pawar, S., Aitwad, R. (2023). Estimation of Car Price Prediction Using Various Machine Learning Algorithms. International Journal of Scientific Research in Engineering and Management (IJSREM), 07(05), Page 1. DOI: 10.55041/IJSREM21574.

[7]. Ganesh, M. (2019). Used Cars Price Prediction using Supervised Learning Techniques. International Journal of Engineering and Advanced Technology, Volume (Issue), Page Range. DOI: 10.35940/ijeat.A1042.1291S319.

[8]. Samruddhi, K., & Dr. R. Ashok Kumar. (2020). Used Car Price Prediction using K-Nearest Neighbor Based Model. International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE), 4(2), 629-632. DOI: 10.29027/IJIRASE.v4.i2.2020.629-632.

[9]. Krishnan, Jayashree & Selvaraj, Vanitha. (2022). Predicting resale car prices using machine learning regression models with ensemble techniques. 2nd International Conference on Mathematical Techniques and Applications. AIP Conference Proceedings. 2516. 240001. 10.1063/5.0108560. https://doi.org/10.1063/5.0108560.

[10]. Sri Sai Ganesh Satyadeva Naidu Totakura, Harika Kosuru (2021). Comparison of Supervised Learning Models for predicting prices of Used Cars. Thesis submission, Faculty of Engineering, Blekinge Institute of Technology, Sweden.

[11]. Prof. Pallavi Bharambe, Bhargav Bagul, Shreyas Dandekar, Prerna Ingle (2022). Used Car Price Prediction using Different Machine Learning Algorithms. International Journal for Research in Applied Science & Engineering Technology (IJRASET), ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538; Volume 10 Issue IV Apr 2022.

[12]. XIAOTONG YAO , XIAOLI FU, AND CHAOFEI ZONG (2022).Short-Term Load Forecasting Method Based on Feature Preference Strategy and LightGBM-Xgboost. IEEE Access, vol. 10, pp. 75257-75268, 2022, doi: 10.1109/ACCESS.2022.3192011.

[13]. Marcos Roberto Machado, Salma Karray, and Ivaldo Tributino de Sousa: LightGBM: an effective decision tree gradient boosting method to predict customer loyalty in the finance industry. In 2019 14th International Conference on Computer Science Education (ICCSE), pages 1111–1116. ISSN: 2473-9464.

[14]. Junliang Fan, Xin Ma, Lifeng Wu, Fucang Zhang, Xiang Yu, and Wenzhi Zeng: Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. 225:105758.

[15]. Riduwan (2008). Dasar-dasar Statistika. (Bandung: Alfabeta) .