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

TRAINING/INTERNSHIP/PROJECT REPORT

Submitted in partial fulfillment of the requirements for the award of the degree

Of

-BACHELOR OF TECHNOLOGY-

In

-Mechanical and Automation-
By

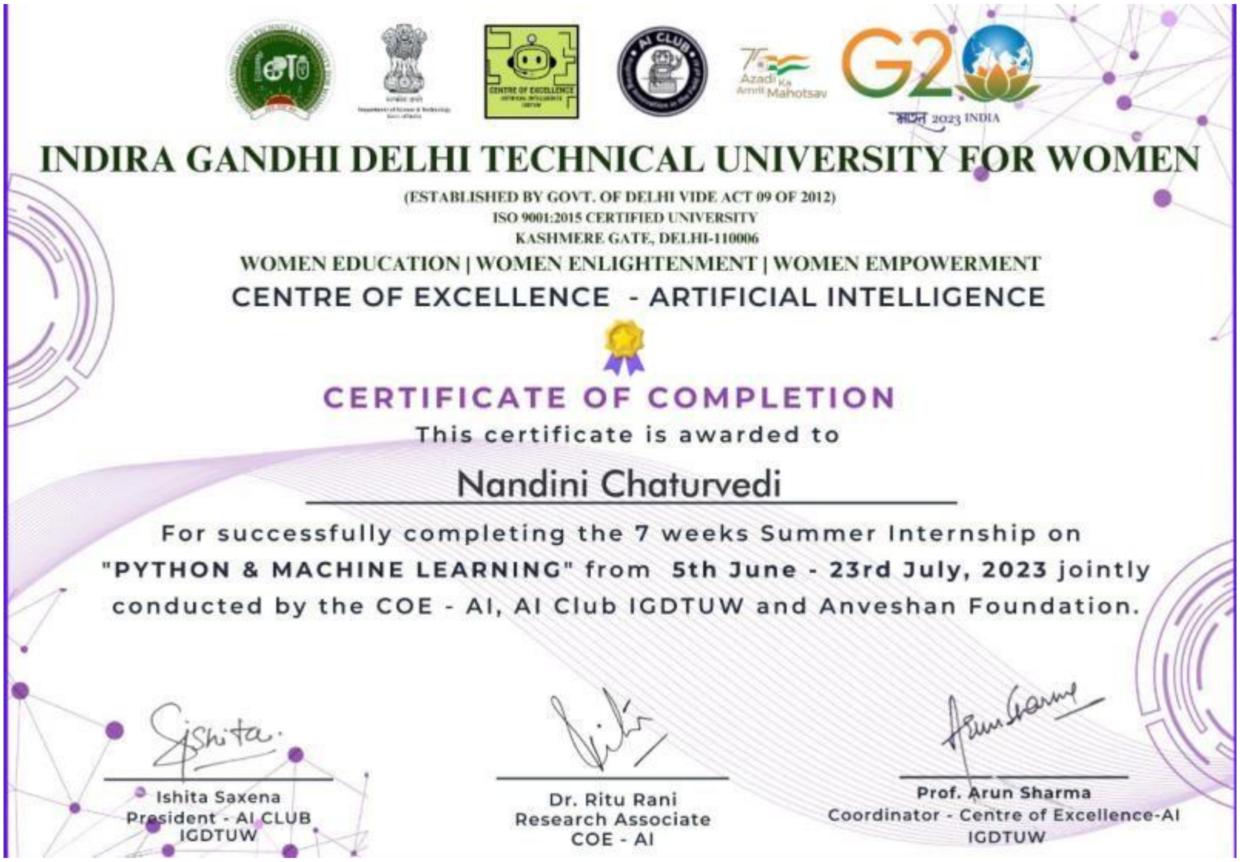
-Nandini Chaturvedi-
-04401042022-

Guided by

-Dr.Ritu Rani-
-Research Associate-
-COE-AI , IGDTUW-

INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN
NEW DELHI – 110006
- JULY 2023 -

CERTIFICATE


AI CLUB
CENTRE OF EXCELLENCE
Azadi Ka Amrit Mahotsav
G20
भारत 2023 INDIA

INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN

(ESTABLISHED BY GOVT. OF DELHI VIDE ACT 09 OF 2012)
ISO 9001:2015 CERTIFIED UNIVERSITY
KASHMERE GATE, DELHI-110006

WOMEN EDUCATION | WOMEN ENLIGHTENMENT | WOMEN EMPOWERMENT

CENTRE OF EXCELLENCE - ARTIFICIAL INTELLIGENCE

CERTIFICATE OF COMPLETION

This certificate is awarded to

Nandini Chaturvedi

For successfully completing the 7 weeks Summer Internship on "PYTHON & MACHINE LEARNING" from **5th June - 23rd July, 2023** jointly conducted by the COE - AI, AI Club IGDTUW and Anveshan Foundation.

Ishita.

Ishita Saxena
President - AI CLUB
IGDTUW

Dr. Ritu Rani
Research Associate
COE - AI

Arun Sharma

Prof. Arun Sharma
Coordinator - Centre of Excellence-AI
IGDTUW

ACKNOWLEDGEMENT

I would like to express my sincere gratitude and appreciation to all those who have contributed to the successful completion of my internship and the preparation of this report.

First and foremost, I extend my heartfelt thanks to Dr. Ritu Rani, my internship mentor, for providing me with valuable guidance, support, and the opportunity to work on challenging and meaningful projects. Your expertise and mentorship have been instrumental in shaping my learning experience during this internship.

I am also thankful to the entire AI Club IGDTUW team in collaboration with Coding Minutes for creating a conducive and collaborative work environment. The insights and knowledge gained from interacting with professionals in the field have been invaluable in broadening my understanding of machine learning.

I extend my appreciation to my teammates for their support.The teamwork and camaraderie within the organization have made my internship both enjoyable and rewarding.

Lastly, I am grateful to my academic institution- IGDTUW for providing me with the opportunity to undertake this internship, allowing me to apply classroom knowledge to real-world scenarios.

Nandini Chaturvedi

04401042022

DECLARATION

Here, the student should declare that the work presented in the report is original and has been completed entirely by the student, with the help of the mentioned supervisors and references.

I, **Nandini Chaturvedi**, solemnly declare that the project report,**Global Perspectives on Used Car Price Prediction: A Machine Learning-based Comparative Study of India, Pakistan, and Germany's Automotive Market,** is based on my own work carried out during the course of our study under the supervision of **Dr.Ritu Rani,Research associate,COE-AI,IGDTUW**. I assert the statements made and conclusions drawn are an outcome of my research work. I further certify that:

I. The work contained in the report is original and has been done by me under the supervision of my supervisor.

II. The work has not been submitted to any other Institution for any other degree/diploma/certificate in this university or any other University of India or abroad.

III. We have followed the guidelines provided by the university in writing the report.

IV. Whenever we have used materials (text, data, theoretical analysis/equations, codes/program, figures, tables, pictures, text etc.) from other sources, we have given due credit to them in the report and have also given their details.

Nandini Chaturvedi

04401042022

LIST OF ABBREVIATIONS

Abbreviation	Description
KNN	K-nearest neighbour
MLR	Multiple Linear Regression
DT	Decision tree
RFR	Random Forest Regressor
XGB	Extreme Gradient Boosting

LIST OF FIGURES AND TABLES

ABSTRACT

The project 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.

INDEX

CHAPTER 1: INTRODUCTION

1.1 Literature Review

In recent years, the application of machine learning algorithms to predict used car prices has garnered significant attention, with studies extending across various countries, offering valuable insights into the challenges and opportunities within this domain. The following review synthesizes key findings from diverse research efforts, emphasizing their contributions to the broader understanding of used car price prediction.

a) Linear Regression in used car dataset of Indonesia:

Puteri and Safitri [2] laid the groundwork by investigating the applicability of linear regression in analyzing used car sales patterns in Indonesia.

While linear regression provided valuable insights, its limitations in capturing the complexities of car pricing were acknowledged.

b) Machine Learning Algorithms in Mauritius:

Pudaruth's study in Mauritius [5] expanded the scope by utilizing multiple machine learning algorithms, including multiple linear regression, k-nearest neighbors, decision trees, and naive Bayes.

Challenges were identified, including difficulties handling numeric values and limitations in dataset size, resulting in 70% accuracy.

c) Diverse Algorithms in General Context:

Patil et al. [6] extended the exploration by estimating car prices using various machine learning algorithms, emphasizing the significance of capturing intricate relationships between car specifications, age, condition, and market trends.The study highlighted the potential of sophisticated models beyond linear regression.

d) Supervised Learning Techniques:

Ganesh [7] contributed insights by exploring supervised learning techniques, emphasizing the importance of feature selection and engineering for enhanced prediction accuracy.The research underscored the holistic approach required in constructing effective predictive models.

e) K-Nearest Neighbor Models:

Samruddhi and Dr. Kumar [8] introduced K-nearest neighbor (KNN) based models for predicting used car prices.

The study emphasized the importance of selecting an appropriate value for "k" and dealing with noise in the dataset, providing a novel perspective on similarity-based approaches.

f) Comparative Results in Germany:

Vanitha S [9] conducted a study comparing regression and machine learning models applied to a German dataset from eBay-Kleinanzeigen.

Boosting algorithms showed superior performance, with XGBoost and Gradient Boosting exhibiting less overfitting. Random Forest demanded more computational time but was effective, and KNN exhibited less overfitting compared to Random Forest.

Despite the extensive individual studies, the comparative aspect of market behavior across diverse nations like India, Pakistan, and Germany remains largely unexplored.

This study aims to fill this gap, considering the diverse economic landscapes, cultural preferences, regulatory environments, and automotive market dynamics in these countries.

1.2 Problem Statement

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

Addressing the increasing complexity and exploitation in the used car markets of India, Pakistan, and Germany is very important. The surge in demand has led to sellers charging inflated prices, creating a disparity between listed and actual vehicle values. This decreases transparency and efficiency in these markets, impacting both buyers and sellers. A solution is crucial in the form of a robust model that accurately assigns prices to vehicles, considering economic variations and cultural preferences. The predictive model aims to enhance pricing strategies and offer insights for improved decision-making in these diverse automotive markets.

CHAPTER 2: METHODOLOGY

- **Preprocessing and dataset splitting:**

Preprocessing involves handling missing and null values by imputation or deletion and cleansing outliers.Filtered datasets are split into training and testing subsets using the "train-test-split" technique.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.

This ensures models are trained on one subset and evaluated on another to prevent overfitting and ensure robust performance.

- **Encoding Categorical Variables:**

Categorical variables are encoded using label encoding to convert them into numerical format suitable for machine learning algorithms.

- **Datasets File:**

High-quality datasets are crucial for accurate prediction models.

Datasets are collected from Kaggle.

➔ **India Dataset (Kaggle):**

Obtained through Kaggle.

Includes features such as car specifications, mileage, year of manufacture, and other attributes.

Dataset size: 15,411 rows and 14 columns.

➔ **Pakistan Dataset (Kaggle):**

Sourced from Kaggle.

Contains features specific to the Pakistani context.

Dataset size: 76,690 records of 9 variables.

➔ **Germany Dataset (Kaggle):**

Collected from Kaggle.

Comprises features relevant to the German automotive market.

Dataset shape: (46,405 rows, 9 columns).

- **Data Analysis Tools:**

Rigorous preprocessing and exploratory data analysis are performed using various Python libraries.

➢ Pandas for data manipulation and analysis.

➢ NumPy for numerical computations and array operations.

➢ Matplotlib.pyplot and Seaborn for data visualization.

➢ scipy.stats for statistical analysis and hypothesis testing.

CHAPTER 3: FINDINGS AND THEIR SIGNIFICANCE

3.1 Features

The common features in all the three datasets are:

- Price, age of the car, fuel, mileage, make, model, transmission: Universally important in determining used car prices.
- Suggests that age, condition, brand, and specifications significantly influence pricing across different markets.

Country-Specific Features:

- **Germany:**

"hp" (horsepower) :

Higher horsepower associated with higher prices.

- **India:**

"hp," "seller_type," "engine capacity," "km_driven," and "seats":

Higher horsepower, engine capacity, and seating capacity contribute to higher prices.

"Seller_type" suggests pricing differences based on dealers or owners.

"km_driven" and "age" expected to negatively impact prices.

- **Pakistan:**

"engine capacity" and "city":

Engine capacity has a significant influence on prices.

"City" reflects regional pricing variations.

3.2 Models used and Results

- **Multiple Linear Regression (MLR):**

MLR is an extension of simple linear regression where multiple independent variables are used to predict the dependent variable.

- **Decision Tree (DT):**

A decision tree is a tree-like model where each node represents a decision based on a particular feature, leading to a branch for each possible outcome.

- **K-Nearest Neighbors (KNN):**

In KNN, a data point is classified or predicted by the majority class or average of its k-nearest neighbors in the feature space. It is simple and effective but can be computationally expensive for large datasets.

- **Random Forest Regression (RFR):**

Random Forest is an ensemble of decision trees. It builds multiple decision trees and merges them together to get a more accurate and stable prediction.

- **Gradient Boosting Regressor Model (GBRM):**

Gradient Boosting builds a model in a stage-wise fashion, combining weak learners (usually decision trees) to create a strong predictive model. It fits the new model to the residuals of the existing model, iteratively reducing the errors and improving overall prediction accuracy.

- **XGBoost Regression Model (XGB):**

XGBoost (Extreme Gradient Boosting) is an optimized and efficient implementation of gradient boosting. XGBoost can handle missing data and is robust to outliers.

- **Cross-Validation Score:**

Assesses how well XGBoost model generalizes to new data.

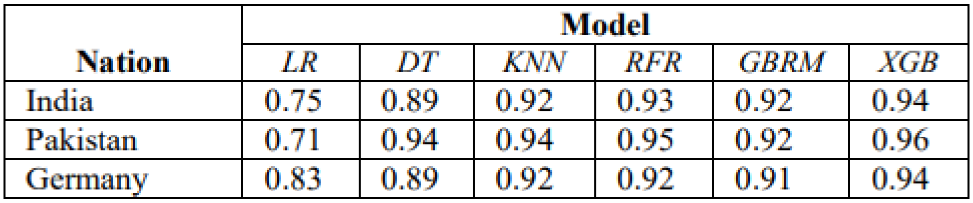
Average Cross-Validation □^{2} Score:

Pakistan: 0.96

Germany: 0.94

India: 0.93

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

Top 3 Features:

- Germany:

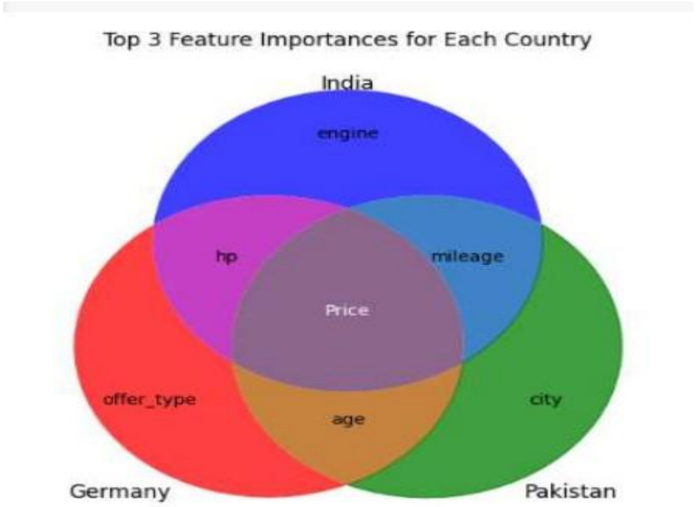
"Age" and "hp" have the highest importance.

- Pakistan:

"Age," "city," and "mileage" are most important.

- India:

"hp," "engine," and "mileage" hold the highest importance.



CHAPTER 4: CONCLUSION

This study provides valuable insights into the preferences of car buyers in India, Pakistan, and Germany in the used car market. Key observations indicate distinct priorities among these nations' consumers.

- **Indian Buyers:**

1) Prioritize performance and low mileage.
2) Favor relatively newer cars with good performance at a reasonable price.
3) Willing to consider older cars if they offer good performance.

- **Pakistani Buyers:**

1) Prefer older used cars due to lower pricing.
2) The age of the car negatively impacts its price consistently.
3) Geographic location influences pricing, indicating regional preferences and economic factors.

- **German Buyers:**

1) Prefer relatively newer vehicles with the latest safety and technology features.
2) Place importance on brand value and popular models, aligning with Germany's automotive culture.

- Both Indian and German buyers prefer high performance cars with powerful engines(indicated by ‘hp’)

Conclusion:

- The study proposes a comparative analysis of used car market behavior in India, Pakistan, and Germany.
- Six models were applied, with XGBoost models outperforming linear regression, emphasizing their ability to capture complex relationships.
- Engine specifications, age, condition, location, listing type, brand value, and market dynamics collectively determine used car prices.

CHAPTER 5: 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 and interior/exterior aesthetics may also provide valuable insights. Factors like inflation, economic conditions, and market trends should be analyzed and be part of any price predicting model. Further exploring 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. 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.

BIBLIOGRAPHY

[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)