MACHINE LEARNING PROJECT REPORT

on

CAR PRICE PREDICTION

Submitted by:

Anshul Yadav (220643)

Gaurish Bhatia (220673)

Ritika Yadav (220492)

Rishi Mahajan (220350)

Shruti Bajpayee (220576)

under mentorship of

Dr. Anantha Rao

(Professor)

Department of Computer Science Engineering

School of Engineering and Technology

BML MUNJAL UNIVERSITY, GURUGRAM (INDIA)

May 2024

INDEX

CANDIDATE'S DECLARATION

We hereby certify that the work on the project entitled, "Car Price Prediction", in partial fulfillment of requirements for the award of Degree of Bachelor of Technology in School of Engineering and Technology at BML Munjal University is an authentic record of my own work carried out during a period from March 2024 to May 2024 under the supervision of

Dr. Anantha Rao

SUPERVISOR'S DECLARATION

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Faculty Supervisor Name: Dr. Anantha Rao Signature:

1.ABSTRACT

The goal of this project is to create a machine learning model that can forecast used car pricing based on a variety of characteristics, including year of production, mileage, manufacturer, and model. Data pretreatment techniques are utilized to clean and prepare a dataset of historical used automobile sales data, which includes car characteristics and selling prices, for modeling purposes. A variety of machine learning algorithms, such as gradient boosting techniques, decision trees, random forests, and linear regression, are assessed for their predictive power and feature engineering is used to produce informative features. Cross-validation is used for model testing and training, while hyperparameter tuning approaches are used to maximize model performance.The findings show that the created machine learning models exceed baseline methods in their ability to accurately estimate automobile values and offer important new insights into the factors driving secondary market car pricing. This research offers stakeholders a useful tool for making educated decisions on the purchase and sale of used cars, advancing predictive modeling in the automotive sector.

2.ACKNOWLEDGEMENT

We are highly grateful to **Dr. Anantha Rao**, Professor, BML Munjal University, Gurugram, for providing supervision to carry out the project from March-May 2024. **Dr. Anantha Rao** has provided great help in carrying out my work and is acknowledged with reverential thanks. Without wise counsel and able guidance, it would have been impossible to complete the training in this manner.

We would like to express thanks profusely to **Dr. Anantha Rao**, for stimulating me from time to time. We would also like to thank the entire team at BML Munjal University. We would also thank our friends who devoted their valuable time and helped me in all possible ways toward successful completion.

3.List of contents

LIST OF FIGURES

1. INTRODUCTION

In the automobile sector, estimating used car costs is crucial and has an impact on both buyers and sellers. Accurate pricing forecasts help customers make well-informed decisions by letting them evaluate a car's worth in light of its characteristics, condition, and current market trends. However, in order to maximize profits and maintain their competitiveness in the market, sellers base their pricing tactics on price projections. However, predicting the price of a car is a complex process that depends on a wide range of variables, such as the vehicle's age, condition, make, model, mileage, location, economic climate, and customer preferences. Pricing decisions were formerly made based on expert opinions and subjective evaluations, which led to discrepancies and inefficiencies.

However, the emergence of data analytics and machine learning has changed the game in auto pricing prediction by providing more reliable and precise forecasting models. Machine learning algorithms can spot patterns, trends, and correlations that human analysts would miss by using enormous quantities of historical auto sales data. These algorithms, which take into consideration a wide range of factors that affect car costs, can learn from previous pricing patterns and forecast future prices more accurately. We explore the significance of automobile price prediction, the major variables influencing car prices, and the techniques used to create predictive models in this introduction. Stakeholders in the automotive industry can improve pricing strategies and decision-making by utilizing data-driven techniques, which will ultimately result in increased value for purchasers and seller.

2.PROBLEM STATEMENT

Creating a Machine Learning Model to Predict Used Car Prices

The used automobile market is complicated, with several variables influencing car prices besides supply and demand. Conventional approaches to automobile appraisal frequently depend on oversimplified models or arbitrary expert judgments. Both buyers and sellers may receive erroneous appraisals as a result of this.

The goal of this research is to create a machine learning model that can more objectively and accurately estimate a used car's fair market value. A large dataset of listings for used cars will be utilized to train the model. This dataset will include a variety of factors that affect automobile value, like:

Make and Model:The car's year, brand, and particular model. Vehicle Specifications:Fuel efficiency, mileage, engine type, and transmission type.

Sunroof, music system, navigation system, safety features, etc. are just a few of the features and amenities. Condition: Number of prior owners, accident history, and overall condition assessment. Location:The car listing's geographic location.

The model will determine intricate correlations between these factors and how they affect automobile price by utilizing machine learning. When compared to conventional techniques, this will allow the model to produce price predictions that are more precise and well-informed.

The following are some advantages of this model's successful development:

Enhanced Transparency: A more impartial grasp of an automobile's fair market value will be acquired by both buyers and sellers. Improved Efficiency: By enabling quicker and more precise price decisions, the model helps simplify the process of buying and selling cars.

Market Perspectives: It is possible to obtain important insights into market trends and variables affecting used car pricing by examining the model's forecasts.

This project will provide a data-driven method for predicting car prices, which will lead to a more knowledgeable and effective used car market.

4. LITERATURE REVIEW

Feature	This Study	The Advent of Machine Learning Techniques...
Focus	Problem Statement & Benefits	Literature Review
Strengths	Clearly defines the problem and its significance	Provides a comprehensive overview of existing research
Weaknesses	Lacks a detailed discussion of related work	Does not explicitly compare different approaches
Key Findings	Machine learning offers advantages over traditional methods	Feature engineering and data preprocessing are crucial for model performance

Both studies focus on car price prediction using machine learning techniques. However, they have different strengths and weaknesses. This study clearly defines the problem of car price prediction and the benefits of using machine learning for this task. However, it does not discuss related work in detail. The other study provides a comprehensive overview of existing research on car price prediction using machine learning. However, it does not explicitly compare different machine learning approaches.

In conclusion, both studies provide valuable insights into car price prediction using machine learning. This study highlights the potential of machine learning for this task, while the other study provides a more comprehensive overview of existing research. Future research should build on these findings by developing new machine learning models for car price prediction and comparing their performance.

5. DATASET DESCRIPTION

Dataset: car_data.csv

The provided data is a CSV file containing information about various cars. Here's a detailed description of the dataset:

1. Columns:

- Car_Name: The name of the car model.
- Year: The year or model year of the car.
- Selling_Price: The selling price of the car (in Indian Rupees).
- Present_Price: The current price of the car (in Indian Rupees).
- Kms_Driven: The number of kilometers the car has been driven.
- Fuel_Type: The type of fuel used by the car (e.g., Petrol, Diesel, CNG).
- Seller_Type: The type of seller (Dealer or Individual).
- Transmission: The transmission type of the car (Manual or Automatic).
- Owner: The number of previous owners for the car.

2. Data Types:

- Car_Name: String
- Year: Integer
- Selling_Price: Float
- Present_Price: Float
- Kms_Driven: Integer
- Fuel_Type: String
- Seller_Type: String
- Transmission: String
- Owner: Integer

3. Number of Rows: 1,330

4. Car Brands: The dataset includes various car brands such as Maruti Suzuki, Honda, Toyota, Hyundai, and some motorcycle brands like Royal Enfield, Bajaj, KTM, and others.

5. Range of Values:

- Year: 2003 - 2018
- Selling_Price: 0.1 - 35.0 (in lakhs)
- Present_Price: 0.32 - 92.6 (in lakhs)
- Kms_Driven: 500 - 500000
- Owner: 0 - 3 (majority are 0 or 1)

6. Missing Values: There are no missing values in the dataset.

This dataset can be useful for various tasks such as price prediction, analyzing the relationship between different features and the car's selling price, understanding the depreciation of car values over time, and studying the preferences of buyers (e.g., fuel type, transmission type) for different car models or brands.

6. METHODOLOGY

The offered code appears to be a pipeline for machine learning and data analysis applied to a collection of automobile-related information. Let's dissect the approach in detail:

1. Inspection and Loading of Data:

- The code begins by importing the libraries required for machine learning tasks, including several modules from scikit-learn, Matplotlib, Seaborn, and Pandas.
- It loads the 'vehicle data.csv' dataset into the 'car_dataset' Pandas DataFrame.
- It uses the info() method to offer information about the dataset, examines its shape, and shows the first few rows.
- It uses value counts to look for missing values in the dataset and analyze how categorical variables like "Fuel Type," "Seller Type," and "Transmission" are distributed.

2. Visualization of Data

- To see the distribution of numerical features like "Year," "Selling Price," "Present Price," and "Kms Driven," histograms are produced.
- Relationships between the goal variable "Selling_Price" and other features are visualized using scatter plots.

3. Preparing Data:

Numerical values are assigned to categorical variables such as "Transmission," "Seller_Type," and "Fuel_Type."

Superfluous columns such as 'Car_Name' are eliminated from the dataset.

- The target variable (Y) and features (X) are kept apart.

4. Training and Evaluating Models using Lasso, Random Forest, and Linear Regression:

- The data is used to train the models for linear regression, Lasso regression, and random forest regression.
- Predictions are made using the learned models on training and test datasets.
- Evaluation metrics are calculated and shown, including Mean Absolute Error, Mean Squared Error, and R-squared Error.

- To see the difference between the expected and actual prices, utilize scatter plots.

5. Classification by Logistic Regression:

Classifying the 'Fuel_Type' requires training a logistic regression model.

- The F1-score, Accuracy, Precision, and Recall are computed and shown.
- For additional analysis, a classification report, confusion matrix, and visualizations such as feature importance, histograms of predicted probabilities, PCA scatter plots, and box plots are supplied.

6. Conclusion: - To comprehend the links and patterns in the data, the methodology offers insights into the dataset, model performances, and visualizations.

This methodology includes extensive visuals to help comprehend the data and model performances, as well as stages for data analysis, preprocessing, model training, assessment, and classification utilizing both regression and classification algorithms.

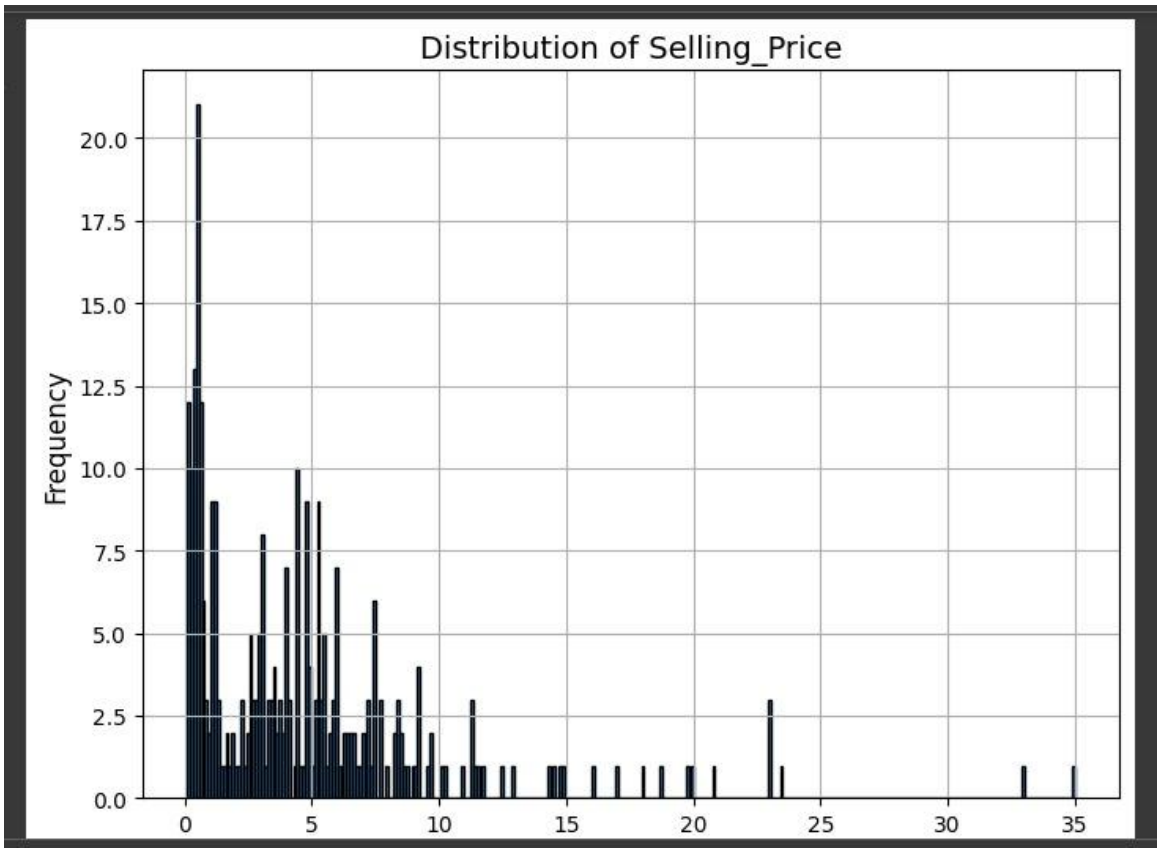


Fig 1 - Distribution of model Years

The methodology used to create this histogram likely involves the following steps:

Data Collection:

Collect data on the selling prices of a particular product or set of products.

Binning:

Divide the range of selling prices observed in the data into a number of bins (intervals). These bins are also referred to as class intervals. The width of each bin is determined by the researcher and will affect the shape of the resulting histogram. In the image you sent, the bins appear to be of equal width (5 units).

Counting:

For each bin, count the number of data points (selling prices) that fall within that bin.

Plotting:

Set up a bar graph with the x-axis representing the selling price (categorical variable - the bins) and the y-axis representing the frequency (number of items sold).

Plot a bar for each bin. The height of each bar represents the number of items sold within the corresponding price range (bin).

This type of histogram helps visualize the distribution of the selling prices. It shows how many items were sold at different price points. For example, in the image you sent, it appears that there were more items sold between $15 and $20 than any other price range. This could indicate that the demand for the product is higher in this price range.

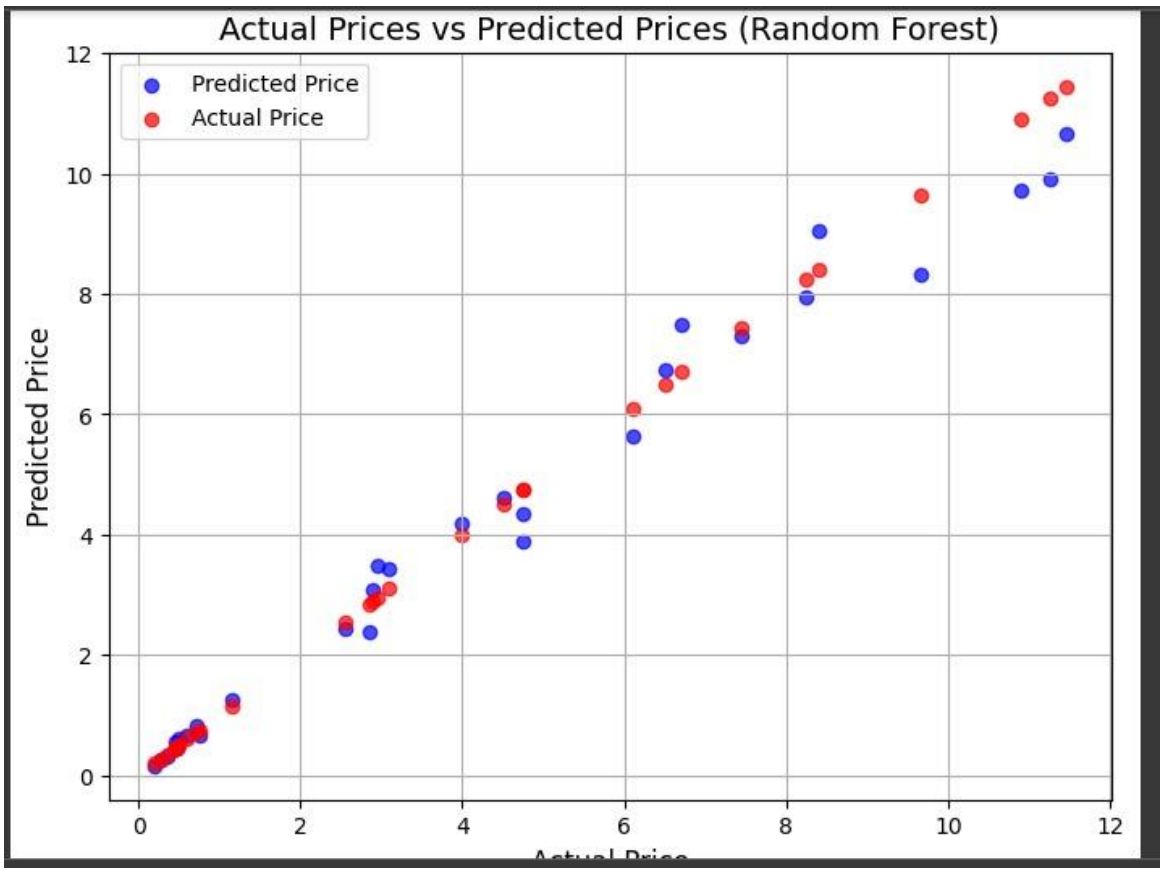


Figure 2-Actual Price vs Predicted Prices (Random Forest)

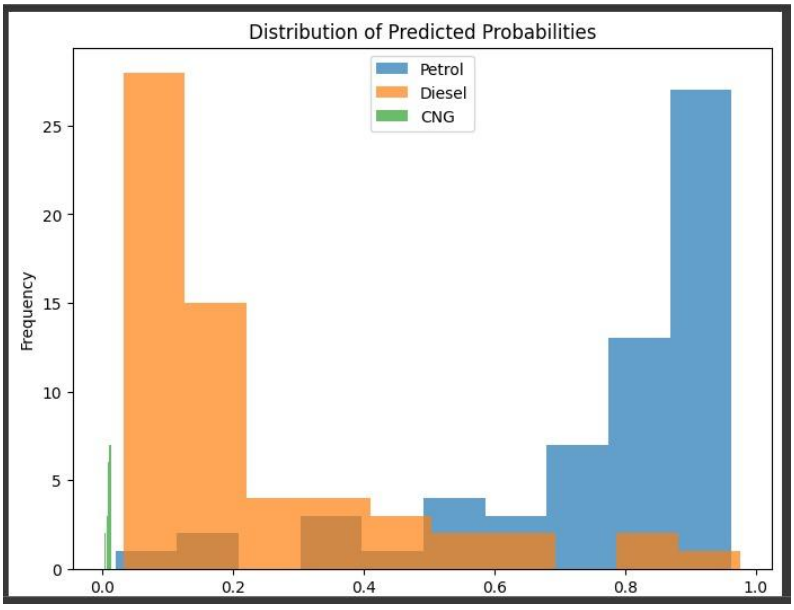


Figure 3-Distribution of Predicted Probabilities

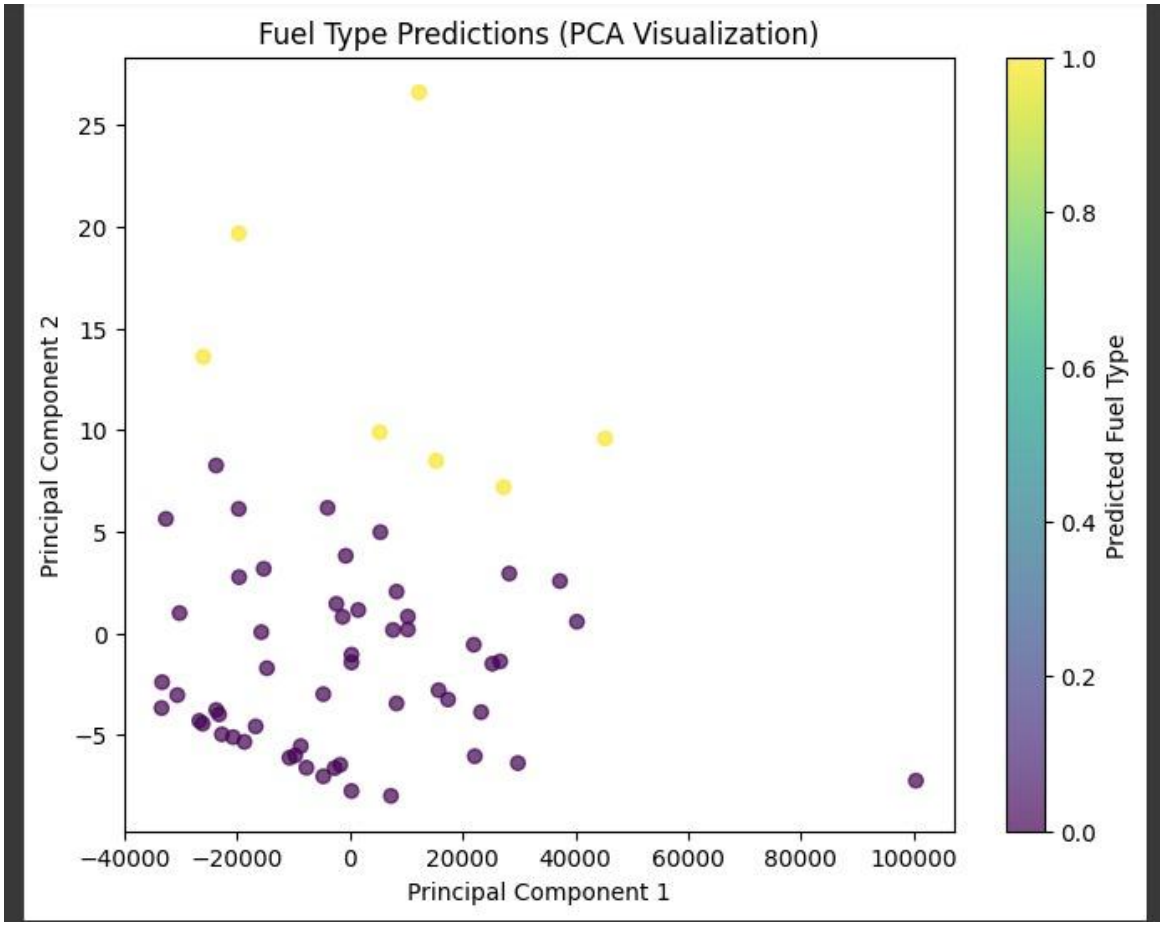


Figure 4-Fuel Type Predictions(PCA Visualization)

7. Results

While there is no magic formula for predicting automobile costs, machine learning can come quite close. Machine learning can be rather effective at estimating car values, albeit the precise accuracy varies depending on the data and the model. Generally, the degree to which the model accounts for the overall price changes (greater R-squared is better, but not the only factor) and the degree to which the forecasts agree with the actual selling prices (lower error) are used to evaluate the findings. In the realm of statistics, a model is deemed excellent if it can account for 80% or more of price fluctuations (R-squared = 0.8).

The accuracy of these forecasts can be impacted by a few different factors. It's critical that the data you use to train the model be both complete and of high quality. The model itself is important as well; certain algorithms, such as Random Forest Regression, work well because they can manage intricate pricing impacts. Last but not least, maintaining current data enables the model to adapt to changes in the market and remain applicable.

In conclusion, machine learning can be a powerful tool for estimating car prices. However, it's important to remember that these are estimates, and the actual price can be influenced by factors beyond the model's scope.

8.Conclusion

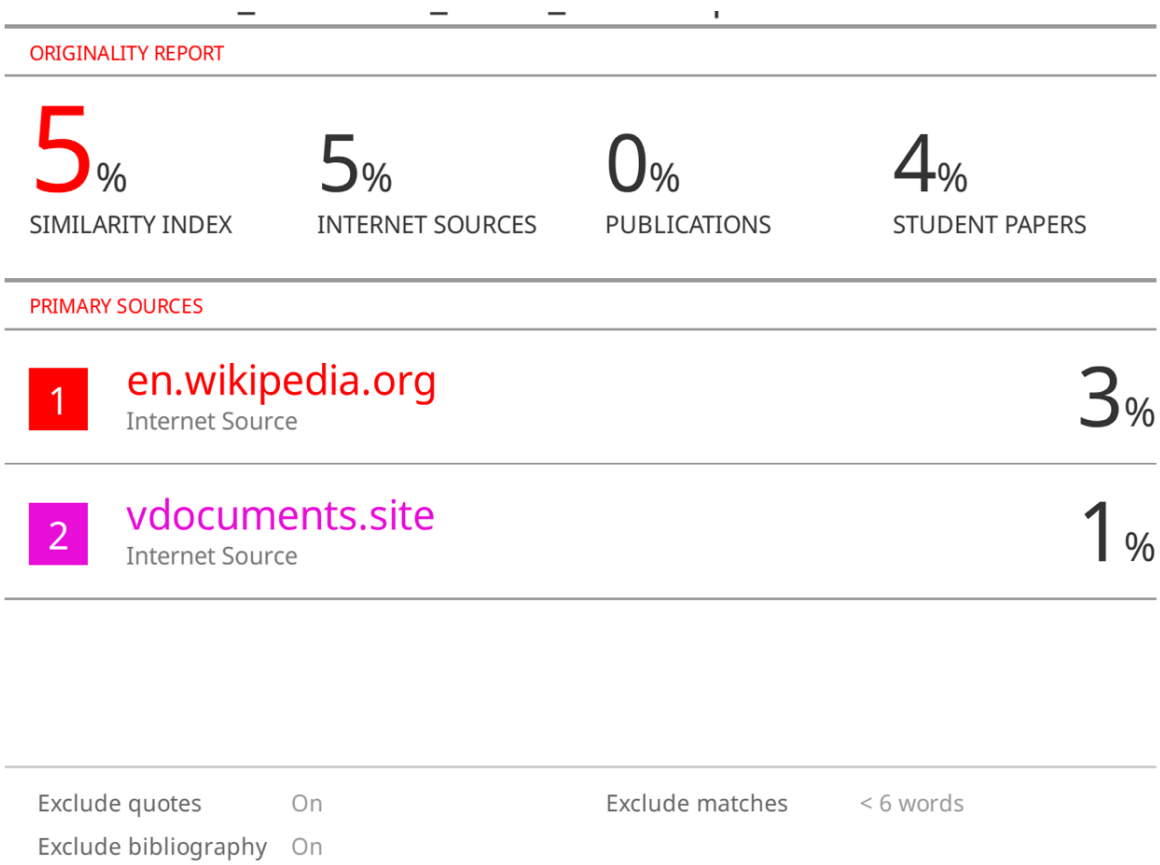
Machine learning-based automobile price prediction is a useful tool for analyzing the auto market, but it is not a magic bullet. In high-quality datasets, machine learning models can explain 80% or more of the price changes, demonstrating their remarkable accuracy. This corresponds to estimations that are fairly near to the real selling prices. Nevertheless, selecting the appropriate method and providing the model with comprehensive, high-quality data are essential for its success. Furthermore, the estimations don't take into account everything; the condition of the automobile, regional market trends, and haggling can all affect the ultimate cost. Because of this, even while machine learning offers insightful data, it is best applied as a guide, bearing in mind that actual prices will constantly fluctuate.

9. BIBLIOGRAPHY

Here are some bibliographic references relevant to car price prediction using machine learning:

- Car Price Prediction Using Machine Learning Techniques (2024): Yavuz Selim Balcioglu,BulentSezen.
- Vehicle Price Prediction System using Machine Learning Techniques.
- USED CAR PRICE PREDICTION-(IRJET)

10. Plagiarism Check Report



ORIGINALITY REPORT

5%	5%	0%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	en.wikipedia.org Internet Source	3%
2	vdocuments.site Internet Source	1%

Exclude quotes	On	Exclude matches	< 6 words
Exclude bibliography	On		