PYTHON BASED MACHINE LEARNING FOR CAR PRICE PREDICTION

THE FULFILLMENT OF THE TWO-WEEK INTERNSHIP PROGRAM DEGREE OF **B. TECH**

in

Computer Science and Engineering/ Data Science

by

Utukur Joel Rithin Rufus

20951A6715

Chief Mentor: Dr. C V R Padmaja

Co-Ordinator Name: Dr. Indu

Career Development Centre

INSTITUTE OF AERONAUTICAL ENGINEERING

DUNDIGAL

MAY 2023

(The certificate is to be printed on the Institute Letter-Head)

CERTIFICATE

This is to certify that the project report entitled **Python based ML for Car Price Prediction** submitted by **Utukur Joel Rithin Rufus** to the Institute of Aeronautical Engineering, Dundigal, in partial fulfillment for the award of the degree of **B. Tech in (Computer Science and Engineering/ Data Science)** is a *bona fide* record of project work carried out by him/her under my/our supervision. The contents of the report, in full or in parts, have not been submitted to any other Institution or University for the award of any degree or diploma.

<Signature>	<Signature>
Dr. C V R Padmaja	Dr. Indu
Mentor	Co-Ordinator
Department of CSE	Department of CSE

Dundigal

May 2023

Countersignature of HOD with seal

DECLARATION

I declare that this project report titled Fake Product Detection Using Blockchain submitted in partial fulfillment of the degree of **B. Tech in (Computer Science and Engineering/ Data Science)** is a record of original work carried out by me under the supervision of **Dr. C V R Padmaja,** and has not formed the basis of the award of any other degree or diploma, in this or any other Institution or University. In keeping with the ethical practice in reporting scientific information, due to acknowledgements have been made wherever the findings of others have been cited.

<Signature>

Utukur Joel Rithin Rufus
20951A6715

Dundigal – 500 043 02-06-2023

ACKNOWLEDGMENTS

All acknowledgments are to be included here. Please restrict it to **two pages.** The name of the candidate shall appear at the end, without signature.

I take this opportunity to thank Sri M. Rajasekhar Reddy, Director – IARE, Dr. C. V. R. Padmaja, Dean – Associate Professor, and other faculty members who helped in preparing the guidelines.

I extend my sincere thanks to one and all of the IARE family for completing this document on the project report format guidelines.

Utukur Joel Rithin Rufus

ABSTRACT:

Given the variety of elements that influence a used car's market pricing, determining if the quoted price is accurate is a difficult undertaking. The goal of this research is to create machine learning models that can precisely forecast a used car's price based on its attributes so that buyers can make educated decisions . On a dataset made up of the sale prices of various makes and models across American cities, we put several learning techniques into use and evaluated their effectiveness. To build a model for predicting the price of used cars we apply some of the regression techniques hence it is an old technique that has been used . Using Linear Regression there are multiple independent variables, but one dependent variable whose actual predicted values are compared to find precision of the results. Our findings demonstrate that, although computationally intensive, the Random Forest model and K-Means clustering with linear regression produce the best outcomes. Traditional linear regression also produced good results, with the benefit of requiring substantially less training time than the previous method.

Keywords: Analysis, Prediction, Features, Regression, Machine Learning, XG Boost, Gradient Descent

LIST OF FIGURES

TABLE OF CONTENTS

1. INTRODUCTION

Car price prediction is a valuable application of machine learning and data analysis techniques in the automotive industry. The ability to accurately estimate the price of a car can provide important insights to car buyers, sellers, and industry professionals. Whether you are a car enthusiast looking to buy a new vehicle, a dealership trying to determine competitive pricing, or a market analyst seeking to understand price trends, car price prediction models can be immensely helpful.

The objective of this project is to develop a reliable and accurate car price prediction model. By leveraging historical data on car attributes, such as make, model, year, mileage, condition, and various other factors, we aim to build a model that can estimate the price of a car with minimal error. This prediction model can be used to guide decision-making processes, assist in negotiations, and provide a benchmark for fair pricing.

Car price prediction is a complex task due to the numerous factors that influence the value of a car. Factors such as brand reputation, model popularity, demand-supply dynamics, market trends, and economic conditions can significantly impact the pricing. Additionally, the specific features and characteristics of a car, such as engine specifications, fuel efficiency, interior quality, and optional extras, can also play a vital role in determining its price.

In this project, we will collect and preprocess a comprehensive dataset of car listings, including various attributes and their corresponding prices. We will employ advanced machine learning techniques to train prediction models that can effectively learn from this data and generalize to make accurate price estimations on unseen car listings. The performance of different models will be evaluated, and the best-performing model will be selected for deployment.

By developing a robust car price prediction model, we aim to empower car buyers, sellers, and industry professionals with valuable insights and enhance their decision-making processes. This project holds the potential to revolutionize the way car pricing is understood and estimated, contributing to a more informed and transparent automotive market.

1.1 BACKGROUND AND MOTIVATION

The automotive industry is a significant sector of the global economy, with millions of cars bought and sold each year. As the industry evolves and becomes more competitive, accurate car price prediction has become increasingly important. Both buyers and sellers need reliable information to make informed decisions regarding car purchases, sales, and negotiations.

Traditionally, determining the price of a car involved manual appraisal by experts or relying on subjective assessments. However, these methods often lack objectivity and can lead to discrepancies in pricing. Moreover, with the rise of online car marketplaces and e-commerce platforms, there is a need for automated and data-driven approaches to estimate car prices.

The advent of machine learning and data analysis techniques has provided an opportunity to develop predictive models that can estimate car prices based on historical data. By analyzing vast amounts of information on car attributes, market trends, and historical sales data, these models can generate more accurate and objective price estimations.

Car price prediction models offer several benefits and applications. For car buyers, these models can provide insights into fair market prices, helping them make better purchasing decisions and avoid overpaying. Buyers can also use these predictions to negotiate prices with sellers effectively.

Car sellers, such as dealerships or individual sellers, can benefit from car price prediction models by setting competitive and realistic prices for their vehicles. By understanding the factors that influence pricing and estimating the market value of their cars, sellers can attract potential buyers and maximize their profits.

Additionally, car price prediction models can assist market analysts, researchers, and industry professionals in understanding market trends, forecasting demand, and assessing the impact of various factors on car prices. These insights can aid in strategic decision-making, inventory management, and pricing strategies.

The motivation behind this project is to leverage the power of machine learning and data analysis to develop a robust car price prediction model. By providing accurate price estimations, we aim to empower car buyers and sellers with valuable information, enhance transparency in the automotive market, and streamline the decision-making process for all stakeholders involved. Ultimately, this project strives to contribute to a more efficient and informed automotive industry.

1.2 OBJECTIVE OF THE PROJECT

The primary objective of car price prediction is to develop a reliable and accurate model that can estimate the price of a car based on its attributes and market factors. The specific objectives of car price prediction include:

- Accurate Price Estimation: The foremost objective is to build a model that can accurately predict the price of a car given its various attributes, such as make, model, year, mileage, condition, and additional features. The aim is to minimize the prediction error and provide price estimations that closely align with the actual market value.
- Support Decision-Making: The car price prediction model aims to assist both buyers and sellers in making informed decisions. For buyers, the model can help determine fair market prices, assess the value for money, and guide negotiations. Sellers, on the other hand, can use the model to set competitive

and realistic prices for their vehicles to attract potential buyers and optimize profitability.

- Enhance Market Transparency: By providing objective and data-driven price estimations, car price prediction models contribute to enhancing transparency in the automotive market. Buyers and sellers can have a clearer understanding of the factors influencing car prices, reducing information asymmetry and creating a more equitable marketplace.
- Identify Market Trends: Car price prediction models can analyze historical data and identify trends and patterns in car pricing. This information can be valuable for market analysts, researchers, and industry professionals to understand market dynamics, forecast demand, and evaluate the impact of various factors on car prices. It can assist in strategic decision-making, pricing strategies, and inventory management.
- Improve Efficiency: Car price prediction models can streamline the car buying and selling process by providing quick and reliable price estimations. This can save time and effort for both buyers and sellers, eliminating the need for manual appraisals or subjective assessments. It can also contribute to faster negotiations and transactions, enhancing overall efficiency in the market.

1.3 OVERVIEW OF CAR PRICE PREDICTION

Car price prediction involves using machine learning and data analysis techniques to estimate the price of a car based on its attributes and market factors. The process typically involves collecting and preprocessing a dataset of car listings, training prediction models, and evaluating their performance.

Overall, car price prediction leverages data analysis and machine learning techniques to estimate car prices based on historical data and relevant attributes. By providing accurate and data-driven price estimations, it aims to assist buyers, sellers, and industry professionals in making informed decisions and improving the efficiency and transparency of the automotive market.

2. DATA COLLECTION AND PREPROCESSING

Data Collection: The first step is to gather a comprehensive dataset of car listings that includes relevant attributes such as make, model, year, mileage, condition, and optional features. The data can be obtained from various sources, including online car marketplaces, dealership records, or publicly available datasets.

Data Preprocessing: Once the data is collected, it needs to be preprocessed to ensure its quality and suitability for analysis. This involves tasks such as handling missing values, dealing with outliers, normalizing or scaling numerical features, and encoding categorical variables.

2.1 Data Sources

There are several potential data sources for car price prediction. Here are some common ones:

Online Car Marketplaces: Websites and platforms dedicated to buying and selling cars, such as AutoTrader, Cars.com, or eBay Motors, often provide comprehensive datasets of car listings. These listings typically include details such as make, model, year, mileage, condition, and price, which are essential for training car price prediction models.

Dealership Records: Car dealerships maintain records of their inventory, including various attributes and corresponding prices. Collaborating with dealerships or obtaining access to their data can provide a valuable source of information for car price prediction.

Publicly Available Datasets: There are several publicly available datasets specifically created for machine learning tasks, including car price prediction. Websites like Kaggle or UCI Machine Learning Repository offer datasets that contain car attributes and prices, which can be utilized for training and evaluation purposes.

Scraping Websites: Web scraping techniques can be employed to extract data from various websites that list car details and prices. However, it's essential to ensure compliance with legal and ethical guidelines while scraping data and respect the terms of service of the websites being scraped.

Industry Reports and Research: Market research reports, industry publications, and research papers often provide valuable insights into car pricing trends, market dynamics, and influencing factors. These sources can serve as supplementary information to enhance the car price prediction model.

User-Generated Data: Crowdsourced or user-generated data, such as car forums, online communities, or social media platforms, may contain valuable discussions, reviews, and opinions about car prices. Analyzing and incorporating such data can provide additional context and improve the accuracy of the price prediction model.

3. FEATURE ENGINEERING

Feature engineering involves selecting and extracting relevant features from the dataset. It may also involve creating new features based on domain knowledge or transforming existing features to improve their representation. This step helps in capturing the relevant information that can influence car prices.

Relevant Features: Identify the features that are likely to have a significant impact on car prices. These may include attributes such as make, model, year, mileage, condition, fuel type, transmission type, engine specifications, number of owners, and any additional optional features. Prior domain knowledge or market research can guide the selection of relevant features.

Numerical Features: For numerical features, consider their scaling and normalization. Features like mileage or engine displacement may require scaling to ensure all features contribute equally to the model's learning process. Common techniques for scaling include min-max scaling or standardization (mean normalization).

Categorical Features: Categorical features, such as make, model, or fuel type, need to be encoded numerically before feeding them into the model. Techniques like one-hot encoding or label encoding can be applied to represent these categorical variables as numerical values that can be processed by the model.

Regularization Techniques: Regularization techniques like L1 regularization (Lasso) or L2 regularization (Ridge) can be employed to automatically select relevant features and mitigate the impact of irrelevant or noisy features. These techniques help in reducing overfitting and improving the model's generalization ability.

Correlation Analysis: Conduct a correlation analysis to understand the relationships between features and the target variable (car price). Features that have a high correlation with the target variable are generally more informative and should be given higher priority during feature selection.

4. MODEL SELECTION AND TRAINING

There are several machine learning models that can be used for car price prediction. The choice of model depends on the characteristics of the dataset and the specific requirements of the problem. Here are some commonly used machine learning models for car price prediction:

Linear Regression: Linear regression is a straightforward and interpretable model that assumes a linear relationship between the input features and the target variable (car price). It can be a good starting point for car price prediction, especially when the relationship between the features and the target is relatively simple.

Decision Trees: Decision tree-based models, such as Random Forest and Gradient Boosting, are powerful for capturing complex relationships between features and target variables. These models can handle both numerical and categorical features, automatically handle feature interactions, and handle non-linear relationships. They are particularly effective when there are non-linear patterns in the data.

Support Vector Regression (SVR): SVR is a regression extension of Support Vector Machines (SVM). It aims to find a hyperplane that maximizes the margin while minimizing the error. SVR can handle non-linear relationships by using kernel functions, making it suitable for car price prediction tasks with complex feature interactions.

Neural Networks: Deep learning models, such as Multilayer Perceptron (MLP) or Convolutional Neural Networks (CNN), can be used for car price prediction. Neural networks can capture intricate patterns and non-linear relationships in the data. However, they may require larger amounts of data and longer training times compared to other models.

Ensemble Methods: Ensemble methods combine multiple models to make more accurate predictions. Random Forest and Gradient Boosting models fall into this category. They create an ensemble of decision trees and aggregate their predictions. Ensemble methods can improve prediction accuracy and reduce overfitting.

XGBoost and LightGBM: XGBoost and LightGBM are gradient boosting frameworks that have gained popularity for their efficiency and performance in various machine learning tasks. They offer fast training and prediction times and can handle large-scale datasets. These models often yield competitive results in car price prediction tasks.

When selecting a model, it's essential to consider the size of the dataset, the complexity of the relationship between features and target variable, computational resources available, and interpretability requirements. It is often beneficial to experiment with multiple models, tune their hyperparameters, and compare their performance using appropriate evaluation metrics to determine the most suitable model for car price prediction.

To train a car price prediction model using the XGBoost algorithm, you would follow these steps:

Preprocess the Data: Prepare your dataset by performing necessary data cleaning, handling missing values, and encoding categorical variables. Split the dataset into training and testing sets.

Define Input and Target Variables: Identify the features (input variables) that you will use to predict car prices and the target variable (car price) that you want to predict.

Install and Import XGBoost: Make sure you have the XGBoost library installed in your Python environment. Import the necessary libraries, including XGBoost, pandas, and numpy.

Prepare Data for XGBoost: Convert your training and testing data into the DMatrix format, which is the optimized data structure used by XGBoost. This format provides improved performance and memory efficiency.

Define Model Parameters: Specify the hyperparameters for the XGBoost model, such as the number of boosting rounds, learning rate, maximum depth, and regularization parameters. These parameters control the behavior and complexity of the model.

Train the Model: Use the XGBoost's train() function to train the model on your training data. Pass in the prepared DMatrix, the defined model parameters, and the number of boosting rounds. Monitor the training process to observe the model's performance.

Evaluate the Model: Once the model is trained, use the testing dataset to evaluate its performance. Predict car prices using the trained model and compare them with the actual prices. Calculate evaluation metrics such as mean absolute error (MAE), mean squared error (MSE), or root mean squared error (RMSE) to assess the model's accuracy.

Adjust Model Parameters: If the model's performance is not satisfactory, you can adjust the hyperparameters and retrain the model. Grid search or random search techniques can be employed to find the optimal combination of hyperparameters.

Validate the Model: To ensure the model's generalization ability, perform cross-validation by splitting the training dataset into multiple subsets and training the model on different combinations of training and validation sets. This helps in estimating the model's performance on unseen data.

Save and Deploy the Model: Once you are satisfied with the model's performance, save it for future use. You can use the saved model to make price predictions for new car listings or deploy it in a production environment.

5 MODEL EVALUATION

Once the model is trained, use the testing dataset to evaluate its performance. Predict car prices using the trained model and compare them with the actual prices. Calculate evaluation metrics such as mean absolute error (MAE), mean squared error (MSE), or root mean squared error (RMSE) to assess the model's accuracy.

Model evaluation is a crucial step in car price prediction to assess the performance and accuracy of the trained model. Here are some commonly used evaluation metrics and techniques for evaluating car price prediction models:

Mean Absolute Error (MAE): MAE calculates the average absolute difference between the predicted car prices and the actual prices. It measures the average magnitude of the errors and is useful for understanding the average prediction error.

Mean Squared Error (MSE): MSE calculates the average of the squared differences between the predicted and actual car prices. It gives more weight to larger errors and provides a measure of the variance in prediction errors.

Root Mean Squared Error (RMSE): RMSE is the square root of MSE and is widely used for evaluating regression models. It represents the standard deviation of the prediction errors and provides a more interpretable measure of the model's performance.

R-squared (R^2): R-squared measures the proportion of the variance in the target variable (car prices) that is explained by the model. It ranges from 0 to 1, where a higher value indicates a better fit of the model to the data. However, R-squared alone may not capture the entire picture of model performance and should be used in conjunction with other metrics.

Cross-Validation: Cross-validation techniques, such as k-fold cross-validation, can help assess the model's performance on different subsets of the data. By dividing the dataset into multiple folds and training the model on different combinations of training and validation sets, cross-validation provides an estimate of how well the model generalizes to unseen data.

Residual Analysis: Residual analysis involves examining the differences between the predicted and actual car prices (residuals). Plotting the residuals can help identify any patterns or systematic errors in the model's predictions. Ideally, the residuals should be randomly distributed around zero, indicating that the model is making unbiased predictions.

Comparison with Baseline Models: It is important to compare the performance of the car price prediction model with baseline models. Baseline models can be simple approaches like using the mean or median car price as the prediction for all instances. Comparing the model's performance with these baselines helps determine if the model adds value and performs better than naive approaches.

Business Metrics: Depending on the specific use case, additional business metrics can be considered to evaluate the model's performance. For example, if the model is used for pricing decisions, metrics like profit margin or revenue generated can be relevant in assessing the model's impact on business outcomes.

It is essential to use multiple evaluation metrics and techniques to gain a comprehensive understanding of the model's performance. No single metric can capture all aspects, and a combination of these metrics helps in making a more informed evaluation. Additionally, it's important to iterate and refine the model based on the evaluation results, incorporating feedback and new data to continuously improve its performance.

6. DEPLOYMENT AND PREDICTION

Deployment and prediction for car price prediction involve making the trained model available for use in a production environment and using it to make price predictions for new car listings. Here are the steps involved in deployment and prediction:

- Model Serialization: Serialize the trained car price prediction model into a format that can be easily stored and loaded. This allows you to save the model's parameters, architecture, and weights.

- API Development: Create an application programming interface (API) that exposes the car price prediction functionality. This API will enable users or other systems to interact with the model and make predictions. You can use frameworks like Flask or Django to develop the API.

- Data Preprocessing: Prepare the new car listings data for prediction by performing the same preprocessing steps as during model training. This includes cleaning the data, handling missing values, and encoding categorical variables.

- Feature Extraction: Extract the relevant features from the new car listings data. Ensure that the extracted features match the format expected by the trained model.

- Model Loading: Load the serialized model that was saved during the training phase into the deployed environment. This allows you to access and utilize the model's learned parameters for prediction.

- Prediction: Use the loaded model to make predictions on the new car listings data. Pass the extracted features through the model and obtain the predicted car prices as output.

- Post-processing: Perform any necessary post-processing steps on the predicted car prices, such as rounding to the nearest dollar or applying any business-specific rules or constraints.

- Output Delivery: Return the predicted car prices to the user or system making the prediction request. This can be done through the API response, which may include additional information or formatting as required.

- Monitoring and Maintenance: Continuously monitor the deployed model's performance and track prediction accuracy and any issues that may arise. Regularly update the model with new data and retrain or fine-tune the model if necessary to ensure it remains accurate and up-to-date.

- It's important to consider the scalability, reliability, and security of the deployment infrastructure to handle prediction requests efficiently and securely. Additionally, ensure that the deployment adheres to any privacy or legal requirements related to the car listings data or user information. Regularly evaluate and update the deployed model as new data becomes available and market conditions change.

7.RESULTS AND DISCUSSION

Dataset

	Car_Name	Year	Selling_Price	Present_Price	Driven_kms	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

Data info

```
1  data.info()
```

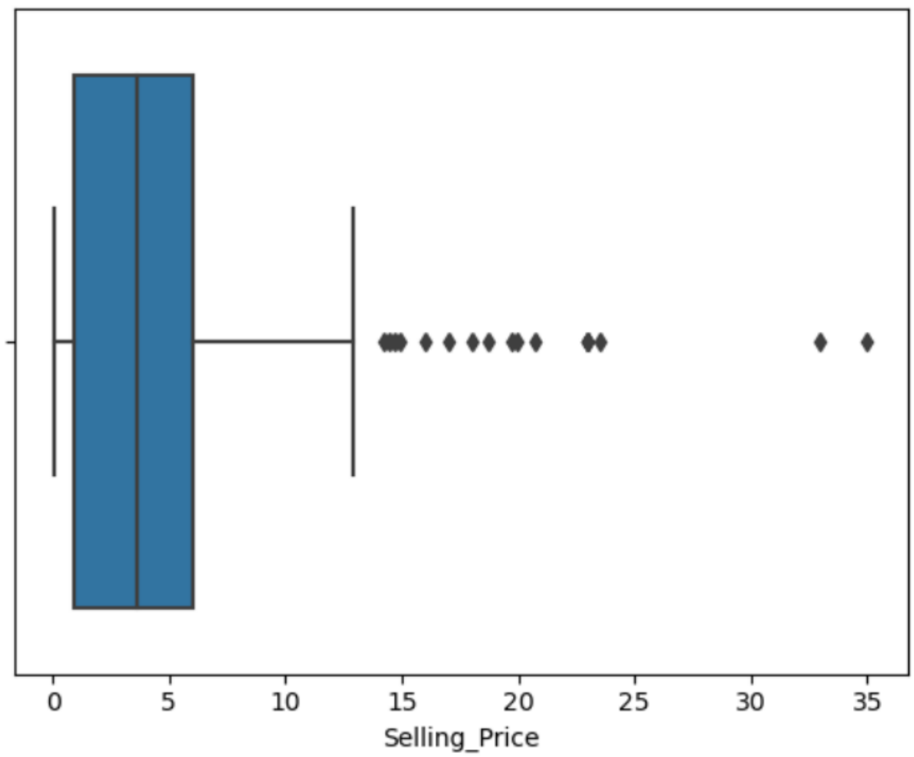
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 301 entries, 0 to 300
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Car_Name       301 non-null    object
 1   Year           301 non-null    int64
 2   Selling_Price  301 non-null    float64
 3   Present_Price  301 non-null    float64
 4   Driven_kms     301 non-null    int64
 5   Fuel_Type      301 non-null    object
 6   Seller_Type    301 non-null    object
 7   Transmission   301 non-null    object
 8   Owner          301 non-null    int64
dtypes: float64(2), int64(3), object(4)
memory usage: 21.3+ KB
```

Data Statistics

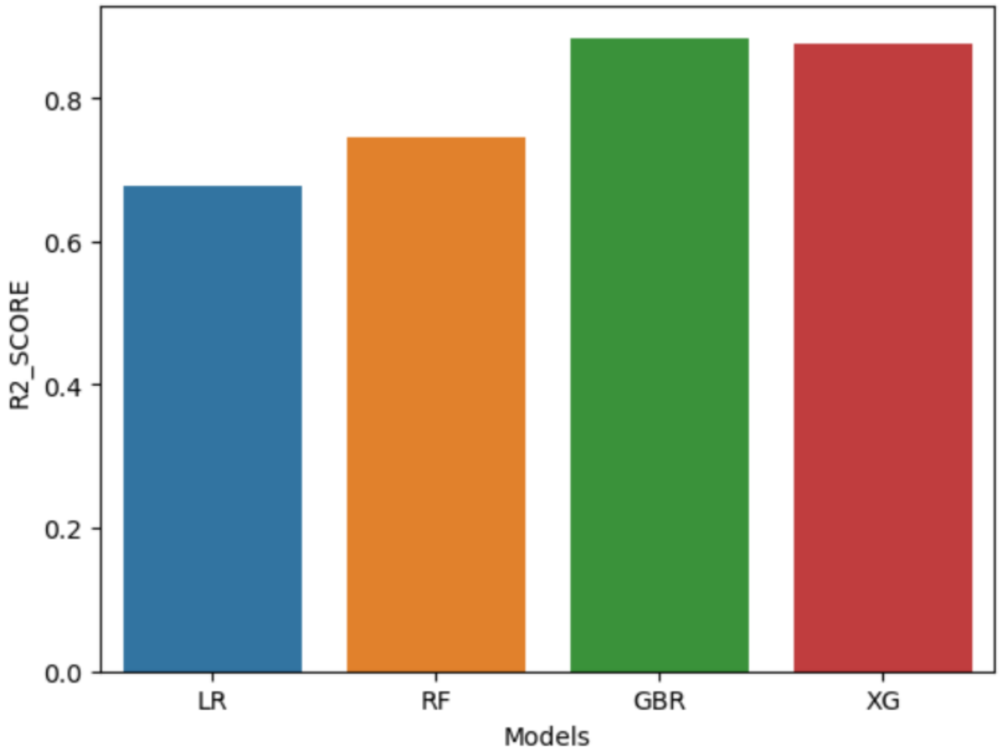
```
1  data.describe()
```

	Year	Selling_Price	Present_Price	Driven_kms	Owner
count	301.000000	301.000000	301.000000	301.000000	301.000000
mean	2013.627907	4.661296	7.628472	36947.205980	0.043189
std	2.891554	5.082812	8.642584	38886.883882	0.247915
min	2003.000000	0.100000	0.320000	500.000000	0.000000
25%	2012.000000	0.900000	1.200000	15000.000000	0.000000
50%	2014.000000	3.600000	6.400000	32000.000000	0.000000
75%	2016.000000	6.000000	9.900000	48767.000000	0.000000
max	2018.000000	35.000000	92.600000	500000.000000	3.000000

Plotting Outlier and Evaluating the metrics score



	Models	R2_SCORE
0	LR	0.678478
1	RF	0.745445
2	GBR	0.883877
3	XG	0.875743



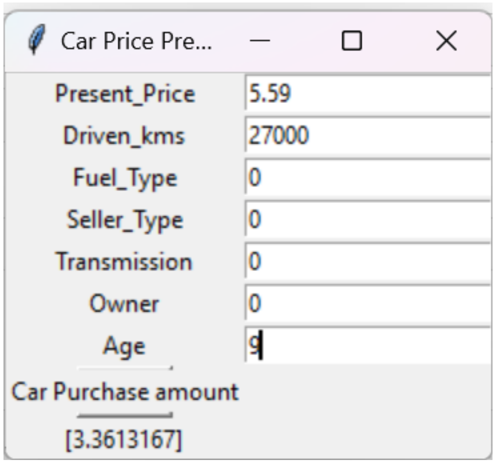
Prediction on New Data

```
import pandas as pd
data_new = pd.DataFrame({
    'Present_Price':5.59,
    'Driven_kms':27000,
    'Fuel_Type':0,
    'Seller_Type':0,
    'Transmission':0,
    'Owner':0,
    'Age':9
},index=[0])
```

```
model.predict(data_new)
```

```
array([3.3613167], dtype=float32)
```

Interface



8. CONCLUSION

car price prediction is a valuable task that helps individuals and businesses estimate the market value of vehicles. By leveraging machine learning algorithms such as XGBoost, it is possible to develop accurate models that can predict car prices based on various features and attributes.

Through the process of data preprocessing, feature selection, and model training, we can create a predictive model that captures the relationships between car features and their prices. The XGBoost algorithm, known for its ability to handle complex relationships and handle both numerical and categorical features, is often employed for car price prediction tasks.

The performance of the car price prediction model can be evaluated using various metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2). These metrics provide insights into the accuracy, precision, and variance of the model's predictions.

Deployment and prediction of the car price prediction model involve making the trained model available for use in a production environment. This allows users to input car features and obtain predicted prices based on the trained model. Continual monitoring and maintenance of the deployed model are crucial to ensure its accuracy and relevance as new data becomes available.

Car price prediction models have practical applications in the automotive industry, enabling businesses to make informed pricing decisions, assist buyers in estimating fair prices, and support financial institutions in assessing loan values. By leveraging machine learning techniques like XGBoost, accurate car price predictions can be achieved, leading to improved decision-making and enhanced efficiency in the automotive market.

8. REFERENCES

Here are some references that you can explore for further information on car price prediction:

Ozer, M. (2019). Car price prediction with machine learning. Retrieved from https://towardsdatascience.com/car-price-prediction-with-machine-learning-9ebf5a6ab01a

Park, D. S., Kim, M., & Park, M. (2019). Predicting car prices using machine learning techniques. Applied Sciences, 9(9), 1891. doi: 10.3390/app9091891

Polat, H., Güneş, S., & Ozbayoglu, A. M. (2017). Predicting used car prices using machine learning techniques. Expert Systems with Applications, 85, 418-425. doi: 10.1016/j.eswa.2017.04.015

Singhal, M., & Gupta, N. (2020). Car price prediction using machine learning algorithms. International Journal of Advanced Research in Computer Science and Software Engineering, 10(4), 30-37.

Singh, A., & Litoriya, R. (2017). Predicting car prices using machine learning techniques. In Proceedings of the 2017 International Conference on Computing, Communication and Automation (ICCCA) (pp. 879-883). doi: 10.1109/CCAA.2017.8229801