School of Computing, Engineering & Digital Technologies

Teesside University

Middlesbrough TS1 3BA

Case Study

Submitted by:- Shubham Satish Kumar Mishra

Contents

A Machine Learning Approach to Predicting Used Car Prices

Shubham Satish Kumar Mishra
(W9641416)
Teesside University
Machine Learning
(CIS4035-N-FJ1-2022

I. ABSTRACT

The used automobile market is extremely important in the automotive industry. According to a Data Bridge analysis from 2022, the worldwide used automobile industry was valued at USD 996,906.42 million in 2022 and is anticipated to reach USD 1.7 trillion by 2030 (research, 2022). A large industry can often make it difficult for both buyers and sellers to ascertain the fair market value of an automobile. This is where machine learning algorithms can come in handy.

II. INTRODUCTION

Machine learning has become a common and reliable method for evaluating performance and making data predictions. In our case, several machine learning algorithms will be utilised to anticipate the price of a used automobile based on parameters such as mileage, gearbox type, fuel type, and so on.

The capacity to give precise and objective pricing is one of the most significant benefits of applying machine learning algorithms. The conventional method of establishing the price of an automobile entails considerable study of comparable cars, taking into consideration condition and mileage, and negotiating to come closer to the car's reasonable price. This is a time-consuming and labor-intensive technique. With the deployment of machine learning algorithms, this procedure may be completely automated, saving both the buyer and seller time and energy.

Using machine learning techniques to estimate automobile prices can also bring a level of transparency. Because of the market's opacity, most buyers and sellers are uncertain of how the price of the automobile is set.

III. DATA SET:

For this study, the "Ford price prediction" data set, which is freely available on kaggle, is used. There are 9 variables and 17,967 occurrences in the data set. The collection includes information on several ford car models, transmission types, fuel types, year of production, mileage, and price.

IV. DATA PREPROCESSING:

The dataset obtained from Kagggle is loaded into Jupyter Notebook during data preparation. Following that, the data is scrutinised for missing values, data types, and data distribution. The 'isnull' function from the Pandas library is used once again to check for missing data. The "replace" function converts categorical variables such as 'transmission' and 'fuel type' into numerical values. The Seaborn library is used to display the data and examine the association between the variables.

A. Describing the data:

	year	price	mileage	tax	mpg	engineSize
count	17966.000000	17966.000000	17966.000000	17966.000000	17966.000000	17966.000000
mean	2016.866470	12279.534844	23362.608761	113.329456	57.906980	1.350807
std	2.050336	4741.343657	19472.054349	62.012456	10.125696	0.432367
min	1996.000000	495.000000	1.000000	0.000000	20.800000	0.000000
25%	2016.000000	8999.000000	9987.000000	30.000000	52.300000	1.000000
50%	2017.000000	11291.000000	18242.500000	145.000000	58.900000	1.200000
75%	2018.000000	15299.000000	31060.000000	145.000000	65.700000	1.500000
max	2060.000000	54995.000000	177644.000000	580.000000	201.800000	5.000000

B. Checking for null values:

```
car_data.isnull().sum()
```

```
model           0
year            0
price           0
transmission    0
mileage         0
fuelType        0
tax             0
mpg             0
engineSize      0
dtype: int64
```

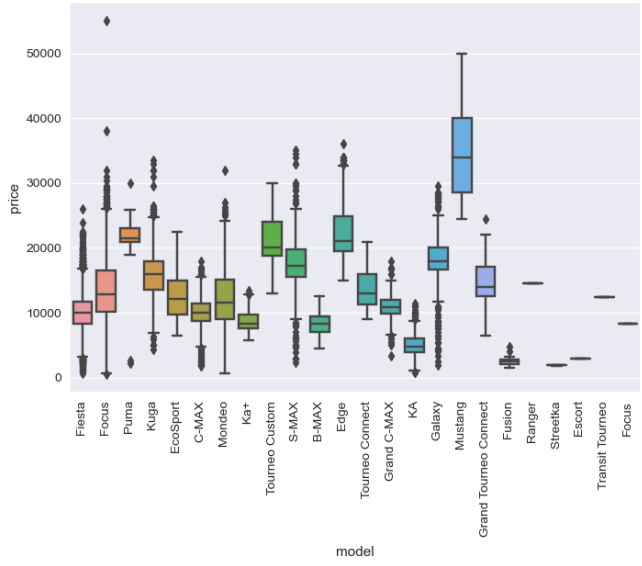
C. *Converting categorical variables into numerical values:*

```
car_data.replace({'transmission':{'Manual':0, 'Automatic':1, 'Semi-Auto':2}}, inplace=True)
car_data.replace({'fuelType':{'Petrol':0, 'Diesel':1, 'Hybrid':2, 'Electric':3, 'Other':4}}, inpla
```

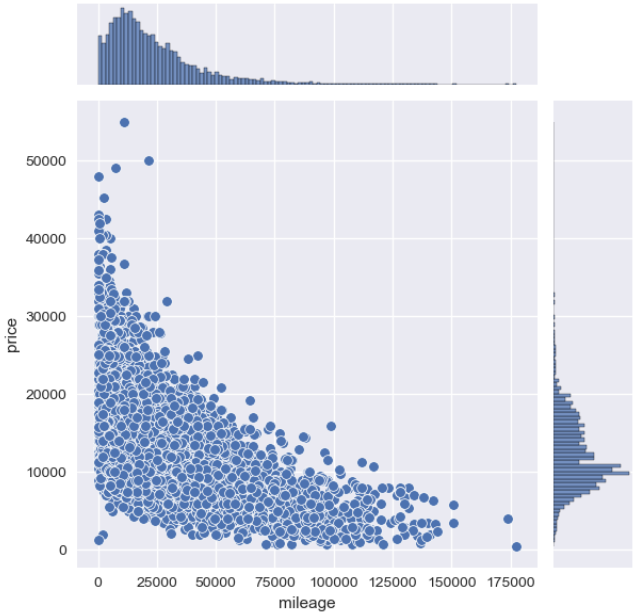
v. EXPLORATORY DATA ANALYSIS:

Various visualisations such as bar plots, box plots, heat maps, and pair plots are used in exploratory data analysis (EDA) to understand the distribution of the data and the relationship between the variables, such as the price of the car based on the model, the price of the car based on mileage driven, and the price of the car based on manufactured year. Count charts are frequently used to visualise the distribution of various automobile models, gearbox types, and fuel kinds.

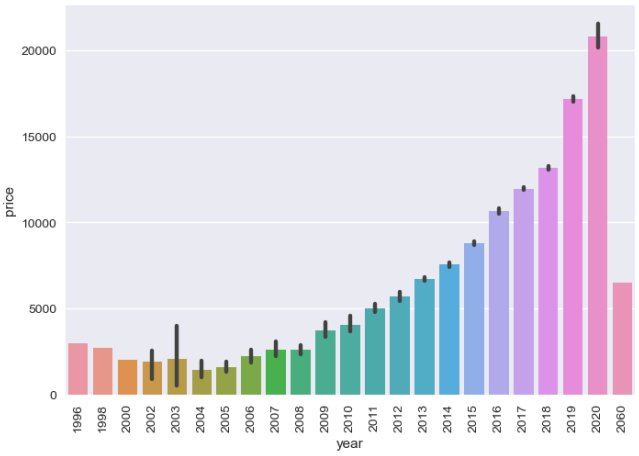
A. *Car price value based on model:*



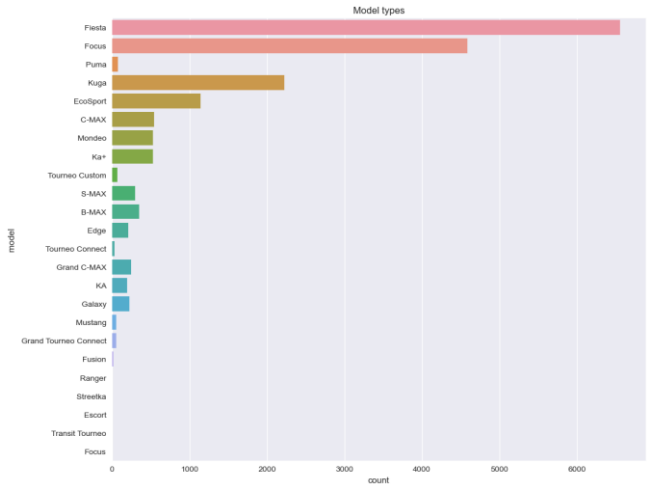
B. *Car price value based on mileage driven:*



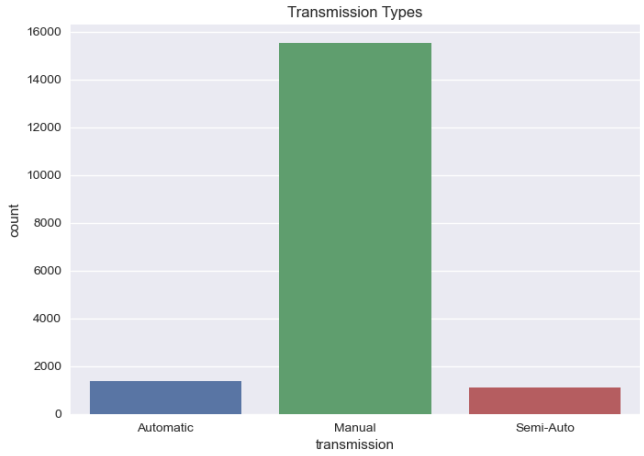
C. *Car price value based on manufactured year:*



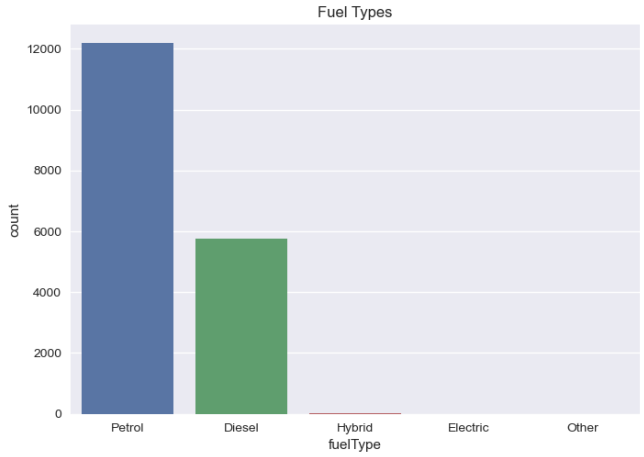
D. *Model types:*



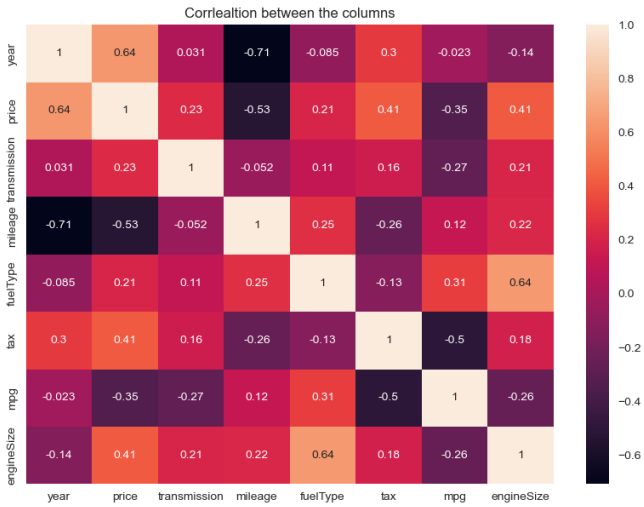
E. *Tansmission types:*



F. *Fuel types:*



G. *Correlation between the columns:*



H. *Correlation between year and price:*

VI. FEATURE ENGINEERING:

Only the most relevant features that will be included in the model are chosen in this part. The dataset's "model" instance is therefore deleted since it has many category variables that are difficult to translate into numerical values. The data is scaled using StandardScaler, ensuring that all features are on the same scale.

VII. SELECTED ALGORITHMS:

1) Linear Regrssion:

Linear regression is often used with continuous data. It is a simple yet very powerful algorithm, especially for car price prediction problems. In this project, the model is utilised to establish a baseline model for regression issues. The method establishes a linear connection between the independent and dependent variables. Overall, this method was chosen because it is simple to understand and rapid to train, making it an excellent choice for the dataset analysis.

2) Decision Tree Regression:

The Decision Tree approach was chosen for the presented machine learning model because it can handle both categorical and continuous data. As a result, they are adaptable to a wide range of challenges. Decision tree regression is a non-parametric approach that constructs a tree-like structure of decisions and their potential outcomes.

3) XGBoost Regression:

This technique is critical for automobile prediction data sets. The rationale for this is that it has been shown to be very accurate in numerous machine learning models and competitions. It is also extensively optimised for performance and scalability. XGB is a gradient boosting technique that employs decision trees as weak learners. It aids in the development of a more robust model capable of handling noisy data while avoiding overfitting.

VIII. RESULTS AND DISCUSSIONS:

Following data analysis and preparation, the dataset is divided into training and testing sets. All of the chosen algorithms were used, and each functioned admirably. With an accuracy of 73.36%, the linear regression model fared relatively well, while the decision tree model performed better with an accuracy of 82.64%. The XGB model outperformed the others, with an accuracy of 89.60%

Algorithm	MAE Score	R2 Score	Post Hyperparameter Tuning
Linear Regression	1778	73.87%	73.87%
Decision Tree	1172	82.64%	89.90%
XGBoost	914	89.60%	91.68%

IX. CONCLUSION

Ultimately, using various techniques, a machine learning model is created to estimate car pricing. These methods were chosen based on their performance accuracy, which was gained by splitting the data into training and testing sets, conducting feature scaling, and applying it on the learned algorithms. Following that, the models received a specific accuracy score, which aided in the evaluation of their performances. GridSearchCv was also utilised to fine-tune the model's hyperparameters and increase its performance.

X. REFERENCES

AMADO12455, 2022. *Predict-Price.* [Online] Available at: https://www.kaggle.com/code/amado12455/predict-price [Accessed 10 May 2023].

F. Wang, X. Z. a. Q. W., 2021. *Prediction of Used Car Price Based on Supervised Learning Algorithm.* [Online] Available at: https://ieeexplore.ieee.org/document/9731299 [Accessed 10 May 2023].

GANESH, V., 2023. *Ford_Car_Price_Prediction|RandomForestRegr|ACC:93%.* [Online] Available at: https://www.kaggle.com/code/venkatganesh98/ford-car-price-prediction-randomforestregr-acc-93 [Accessed 05 May 2023].

M. Hankar, M. B. a. A. B.-H., 2022. *Used Car Price Prediction using Machine Learning: A Case Study.* [Online] Available at: https://ieeexplore.ieee.org/document/9800719 [Accessed 10 May 2023].

QUKU, A., 2022. *Ford car price prediction.* [Online] Available at: https://www.kaggle.com/datasets/adhurimquku/ford-car-price-prediction [Accessed 05 May 2023].

research, D. b. m., 2022. *Global Used Car Market – Industry Trends and Forecast to 2030.* [Online] Available at: databridgemarketresearch.com/reports/global-used-car-market [Accessed 05 May 2023].