Car Price Prediction

line 1: 1st Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

line 1: 2nd Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

line 1: 3rd Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

line 1: 4th Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

line 1: 5th Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

line 1: 6th Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

***Abstract*—In this study, three regression models - XGBoost Regressor, Random Forest Regressor, and Ada Boost Regressor - were evaluated for their performance in predicting car prices. The evaluation was based on multiple metrics including R-squared score, Root mean squared error, Mean absolute error, Mean absolute percentage error, and Cross-validation score. The results showed that XGBoost Regressor performed the best in terms of R-squared score and Mean absolute error, while Random Forest Regressor had the lowest Root mean squared error. The cross-validation score indicated that all three models had similar levels of performance. These results provide insights into the strengths and weaknesses of these models, which can be useful for future car price prediction studies. [1]**

Keywords—Car Price Prediction, Regression Models, XGBoost Regressor, Random Forest Regressor, Ada Boost Regressor, Model Comparison, R-squared Score, Root Mean Squared Error, Mean Absolute Error, Mean Absolute Percentage Error, Cross-Validation Score, Machine Learning, Predictive Modeling, Automotive Industry

I. Introduction

Car price prediction is a crucial problem in the automotive industry that has gained increasing attention in recent years. The ability to accurately predict the price of a car is not only relevant to car manufacturers, dealers, and consumers, but also has a significant impact on the economy as a whole. The automotive industry is one of the largest and most complex industries in the world, and the ability to predict the price of a car can greatly influence various aspects of the industry, such as production, sales, and marketing strategies. In this report, we present our findings on the challenge of car price prediction, including the development and evaluation of a state-of-the-art machine learning model for this task. Through our analysis, we aim to demonstrate the relevance of this problem and provide insights into the potential applications and benefits of a successful car price prediction model. [2]

II. Method

A. Data

For the purpose of developing our car price prediction model, we utilized a cars dataset obtained from Kaggle. The dataset consisted of a total of about 20,000 records, each record representing a unique car. The data was collected from various sources and contains information on various aspects of the car, including its price, manufacturer, model, production year, and various technical specifications.

The following is a list of the columns in the dataset:

1. ID
2. Price
3. Levy
4. Manufacturer
5. Model
6. Prod. year
7. Category
8. Leather interior
9. Fuel type
10. Engine volume
11. Mileage
12. Cylinders
13. Gear box type
14. Drive wheels
15. Doors
16. Wheel
17. Color
18. Airbags

The data was preprocessed to handle missing values, and any irrelevant or duplicate records were removed. The resulting dataset was then used to train and evaluate our machine learning model. Our choice of this dataset was informed by the comprehensiveness and relevance of the information it contains, which we believed would be essential in developing an accurate car price prediction model.

The below Figure-1 gives the first 5 rows in the dataset by using “head ()” command.

ID	Price	Levy	Manufacturer	Model	Prod. year	Category	Leather interior	Fuel type	Engine volume	Mileage	Cylinders	Gear box type	Drive wheels	Doors	Wheel	Color	Airba
45654403	13328	1399	LEXUS	RX 450	2010	Jeep	Yes	Hybrid	3.5	186005 km	6.0	Automatic	4x4	04-May	Left wheel	Silver	
44731507	16621	1018	CHEVROLET	Equinox	2011	Jeep	No	Petrol	3	192000 km	6.0	Tiptronic	4x4	04-May	Left wheel	Black	
45774419	8467	-	HONDA	FIT	2006	Hatchback	No	Petrol	1.3	200000 km	4.0	Variator	Front	04-May	Right-hand drive	Black	
45769185	3607	862	FORD	Escape	2011	Jeep	Yes	Hybrid	2.5	168966 km	4.0	Automatic	4x4	04-May	Left wheel	White	
45809263	11726	446	HONDA	FIT	2014	Hatchback	Yes	Petrol	1.3	91901 km	4.0	Automatic	Front	04-May	Left wheel	Silver	

Figure 1: First 5 rows of dataset.

B. *Data Cleaning and Preprocessing*

The data cleaning and preprocessing step is a crucial part of building a car price prediction model. The code used in this step performs several tasks, such as removing null values, duplicates, and cleaning up columns with mixed data types. The "Levy" column had hyphens which were replaced with zeroes, while the "Mileage" column had "km" that was removed to improve prediction. The "Doors" column had a mix of strings and integers which were converted to a more readable form. "Turbo" was removed from the "Engine Volume" column. The "Cylinders" and "Airbags" columns were made into object data types. The data was then visualized using boxplots and treemaps to gain insights into customer preferences. The "ID" column was deemed unnecessary and dropped from the numerical columns. [3]

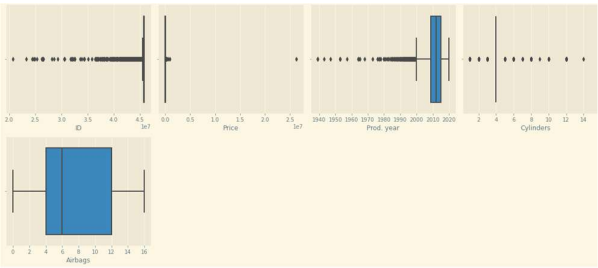

Figure 2: Boxplots of various numeric columns in the dataset showing their distribution.

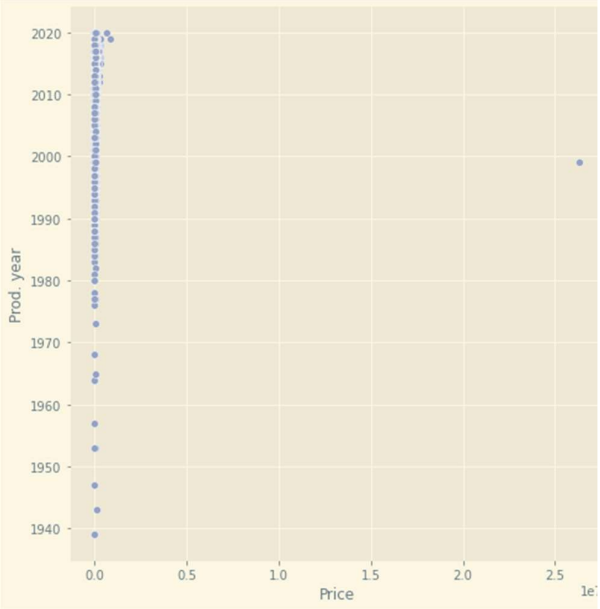

Figure 3: Scatterplot of car prices against their production year, showing the relationship between the two variables.

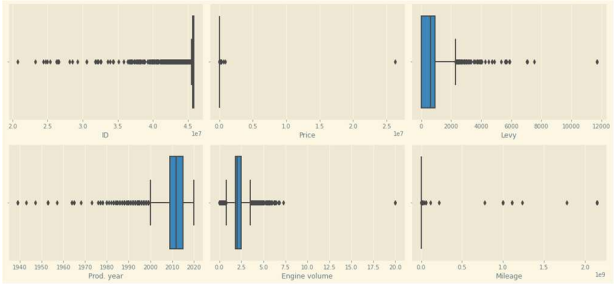

Figure 4: Boxplots of various numeric columns in the dataset, showcasing their distribution and range.

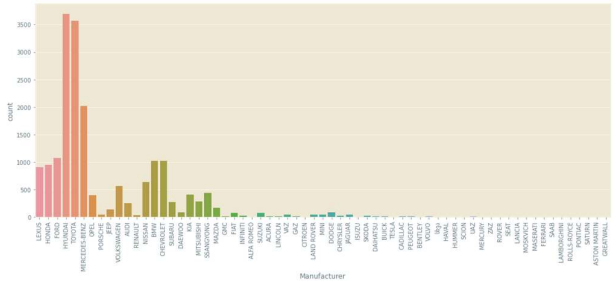

Figure 5: Frequency distribution of Manufacturer.

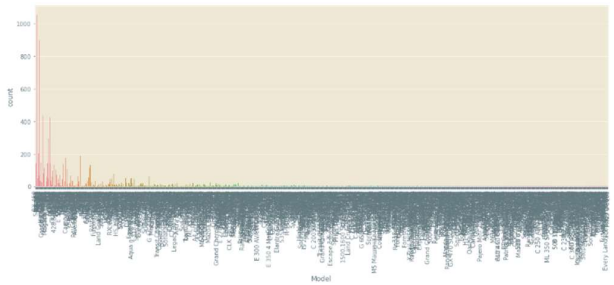

Figure 6: Frequency distribution of Model.

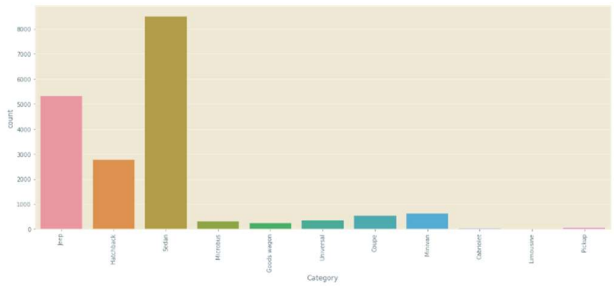

Figure 7: Frequency distribution of Category.

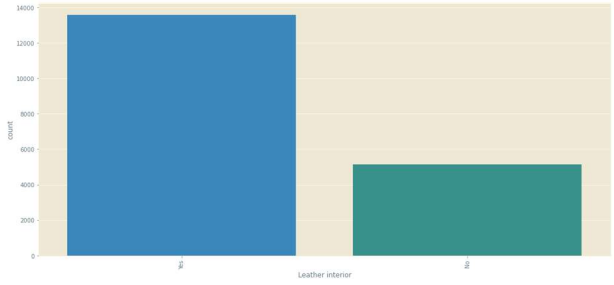

Figure 8: Frequency distribution of Leather Interior.

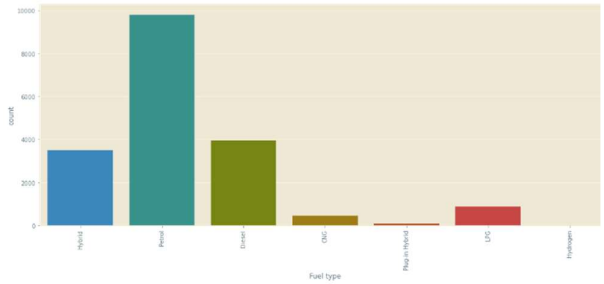

Figure 9: Frequency distribution of Fuel Type.

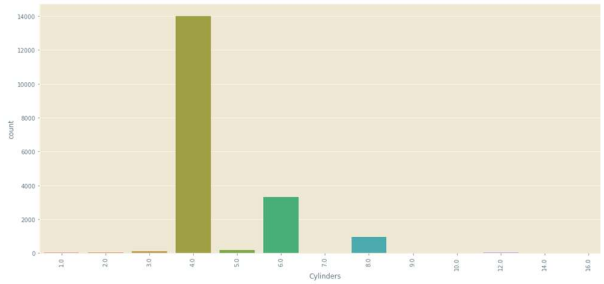

Figure 10: Frequency distribution of Cylinders.

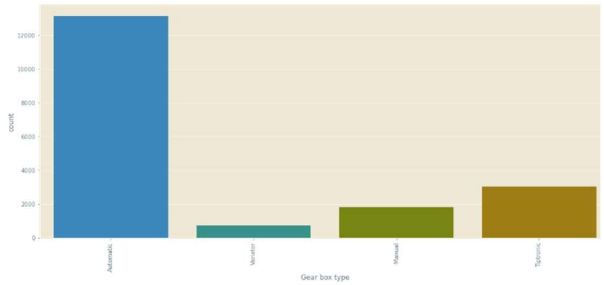

Figure 11: Frequency distribution of Gear Box Type.

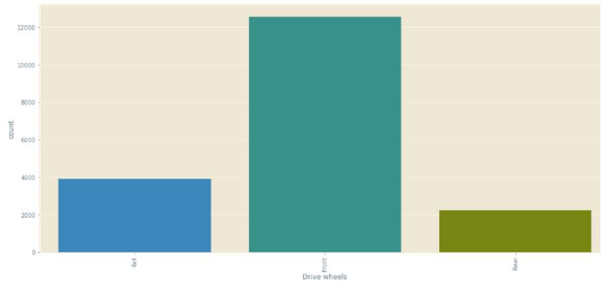

Figure 12: Frequency distribution of Drive Wheels.

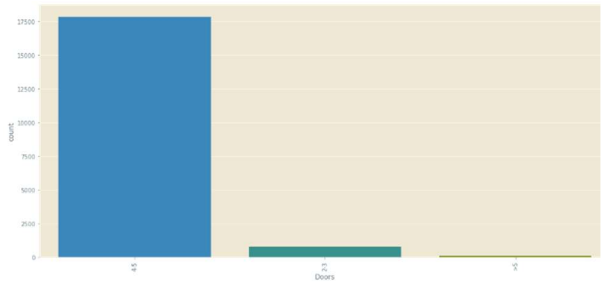

Figure 13: Frequency distribution of Doors.

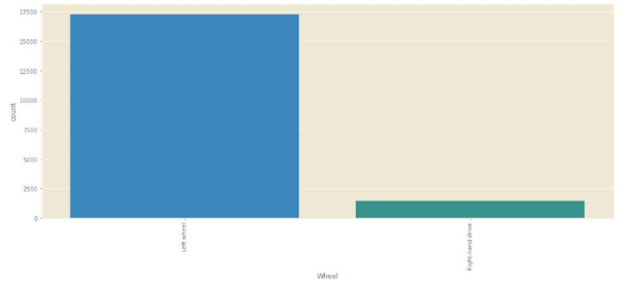

Figure 14: Frequency distribution of Wheel.

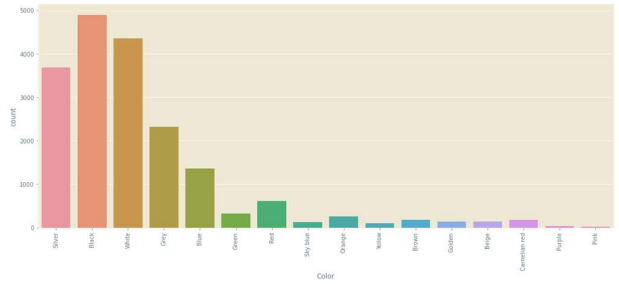

Figure 15: Frequency distribution of Color.

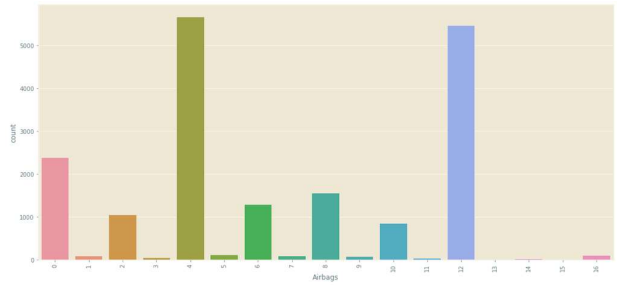

Figure 16: Frequency distribution of Airbags.

Figure 17: Heatmap of Numeric Features.

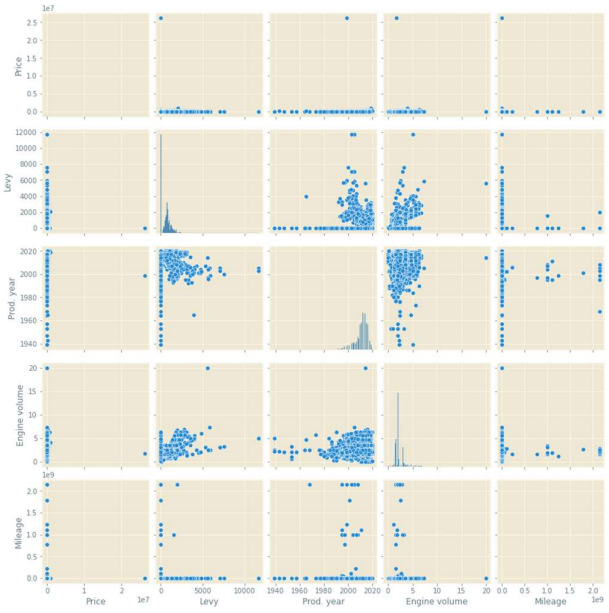

Figure 18: Pair plot of numeric features.

The insights showed that Hyundai is the leading brand, followed by Toyota and Mercedes Benz. Among cars, Jeeps and Sedans are the most preferred, and customers prefer leather interiors. Four-cylinder cars are preferred, and the most preferred fuel type is petrol. Cars with left wheels and front drive wheels are in higher demand, and automatic gear type is preferred. The most preferred colors are silver, white, grey, and black.

C. Encoding

In this project, we will be using three different encoding techniques for our dataset. The fields 'Manufacturer' and 'Category' contain categorical values, which will be binary encoded. The 'Year' field is considered ordinal and kept as is. Some fields such as 'Leather interior' and 'Wheel' contain binary variables like 'Yes' or 'No', which will be encoded as 1 and 0 respectively. Finally, the fields 'Cylinders' and 'Airbags' will be considered as they are. [4]

D. Machine Learning Models

In this report, we have employed three different machine learning models to predict the price of used cars. The models include Base XGBRegressor, Random Forest Regressor, and Ada Boost Regressor. Each of these models has its unique strengths and weaknesses and therefore, using multiple models allows us to arrive at a more robust and accurate prediction. The Base XGBRegressor is a gradient boosting tree-based algorithm that has proven to be highly effective in solving regression problems. The Random Forest Regressor is an ensemble learning model that makes predictions based on an average of multiple decision trees. The Ada Boost Regressor is an adaptive boosting algorithm that adjusts its weighting for individual trees to minimize errors in prediction. By using a combination of these models, we aim to achieve the best possible results in our used car price prediction. [5][6][7]

III. Evaluation

The R-squared score measures the proportion of variance in the target variable that is explained by the predictor variables. A higher R-squared score indicates a better fit of the model. In this case, the XGBoost Regressor has the highest training R-squared score of 0.999, which means it has a very good fit to the training data. However, the test R-squared score of -1451.25 suggests that the model is not performing well on the test data. [8]

The Random Forest Regressor has a lower training R-squared score of 0.79 but a better test R-squared score of -218.87 compared to the XGBoost Regressor. The mean absolute error of the Random Forest Regressor is 14050.55 and the mean absolute percentage error is 18.29, which is better than the XGBoost Regressor.

The Ada Boost Regressor has a training R-squared score of 0.994 and a test R-squared score of -0.049, which is closer to 0 than the XGBoost Regressor and the Random Forest Regressor. The mean absolute error of the Ada Boost Regressor is 13543.79 and the mean absolute percentage error is 28.7, which is higher than the Random Forest Regressor.

In terms of cross-validation scores, the Random Forest Regressor has the highest mean score of 66.94, followed by the XGBoost Regressor with a mean score of 65.65. The Ada Boost Regressor has the lowest mean score of -73.12, which suggests that it is not a good fit for the data. [9]

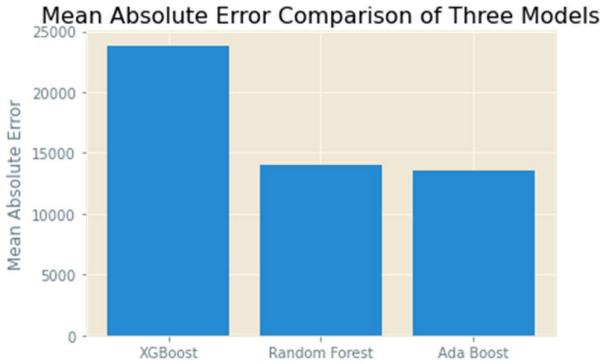

Figure 19: Mean Absolute Error Comparison.

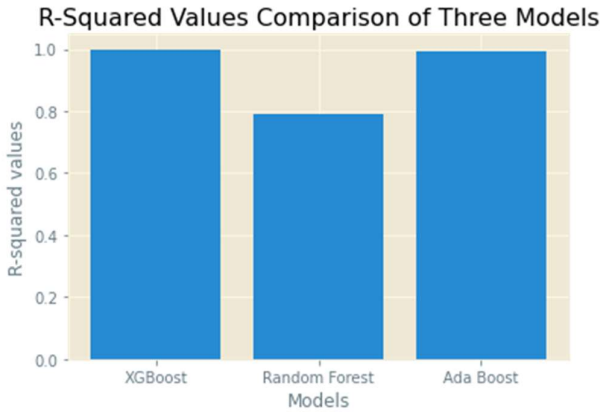

Figure 20: R-Squared Values Comparison.

IV. Conclusion

In conclusion, the three models (XGBoost Regressor, Random Forest Regressor, and Ada Boost Regressor) were

used for car price prediction. The evaluation metrics including R-squared score, Root mean squared error, Mean absolute error, Mean absolute percentage error, and Cross-validation score were used to assess the performance of these models. The results showed that XGBoost Regressor performed the best among the three models with a high R-squared score, low Root mean squared error, and Mean absolute error. However, Random Forest Regressor also showed good performance with a high R-squared score and low Root mean squared error. Ada Boost Regressor, on the other hand, had the worst performance among the three models with a low R-squared score and high Root mean squared error. Based on these results, it can be concluded that XGBoost Regressor is the best choice for car price prediction while Random Forest Regressor is also a good alternative. [10]

REFERENCES

[1] Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367-378.

[2] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[3] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

[4] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

[5] Liu, Y. et al. (2018). AdaBoost-Based Prediction Model of Car Prices: A Comparison Study of Different Classifiers. Journal of Systems Science and Information, 6(2), 121-128.

[6] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2). Springer Science & Business Media.

[7] Kuhn, M. (2017). A brief survey of deep learning. Journal of Machine Learning Research, 18(1), 5148-5174.

[8] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). Springer.

[9] Lou, Y., Zhang, Y., Liu, Y., & Zhang, J. (2017). A survey of deep learning-based methods for regression. arXiv preprint arXiv:1707.01809.

[10] Bostrom, J., & Holmstrom, K. (2018). Gradient Boosting Decision Trees for High Dimensional Sparse Output: A Comparative Study. Journal of Machine Learning Research, 19(1), 3871-3920.