Project Report (UCS622)

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology

(Deemed to be University), Patiala – 147004

Abstract

This project focuses on car price prediction by utilizing data scraped from various websites such as OLX, Cars24, CarDekho, and AutoPortal. The data scraping process was implemented using Python and the Selenium library. Over 5000 car listings were collected, including information such as model, year, ,price etc

After gathering the raw data, preprocessing techniques were applied to clean and transform the data into a suitable format for analysis. This involved handling missing values, analysis of outliers, and encoding categorical variables.

Two different regression algorithms, namely K Nearest Neighbors (KNN) Regressor and Random Forest Regressor, were employed to build predictive models for car price estimation. The KNN Regressor utilizes the concept of similarity to predict the price based on the features of similar cars. On the other hand, the Random Forest Regressor constructs an ensemble of decision trees to make predictions.To optimize the performance of the Random Forest model, a grid search technique was employed to tune the hyperparameters. Grid search involves systematically searching through a predefined set of hyperparameter combinations to identify the best configuration that yields the highest accuracy or lowest error.

Through this project, we aim to develop accurate and reliable models for predicting car prices, leveraging the power of web scraping, data preprocessing, and machine learning algorithms. The findings and insights gained from this project can be valuable for both car buyers and sellers, as well as automotive industry professionals.

Introduction

Problem Description:

The costs of new vehicles in the business is fixed by the producer for certain extra expenses brought about by the Government as assessments. Along these lines, clients purchasing another vehicle can be guaranteed of the money/investment they contribute to be commendable. Be that as it may, because of the expanded cost of new vehicles and the ineptitude of clients to purchase new vehicles because of the absence of assets, utilized vehicles deals are on a worldwide increment.

There is a requirement at a pre-owned vehicle cost expectation framework to successfully decide the value of the vehicle utilizing an assortment of highlights. Despite the fact that there are sites that offer this assistance, their expectation technique may not be awesome. Additionally, various models and frameworks might contribute to anticipating power for a pre owned vehicle's genuine market esteem. It is essential to realize their genuine market esteem while both trading.

Problem Challenges:

The various Challenges faced by us during the making of this model are as follows

- Gathering web scraped data from multiple sources and converting it into a dataset .
- Analysis of data to apply appropriate pre-processing techniques for it to become suitable for applying machine learning techniques .
- Choosing an appropriate machine learning model for the problem.

- Working with cuda to reduce model training time for the problem .
- Increasing the model performance through hyper parameter tuning .

Novelty in work :

The project involves scraping data from multiple popular car listing websites, including OLX, Cars24, CarDekho, and AutoPortal. By aggregating data from various sources, a larger and more diverse dataset is obtained, potentially leading to more accurate predictions.The project compares the performance of two regression algorithms, K Nearest Neighbors (KNN) Regressor and Random Forest Regressor, for car price prediction. This analysis provides insights into the strengths and weaknesses of each algorithm in the context of car price estimation.

By leveraging CUDA, the project takes advantage of the parallel processing capabilities of GPUs, thereby accelerating the model training process. This is particularly beneficial when working with large datasets and complex machine learning models, such as Random Forest Regressor.

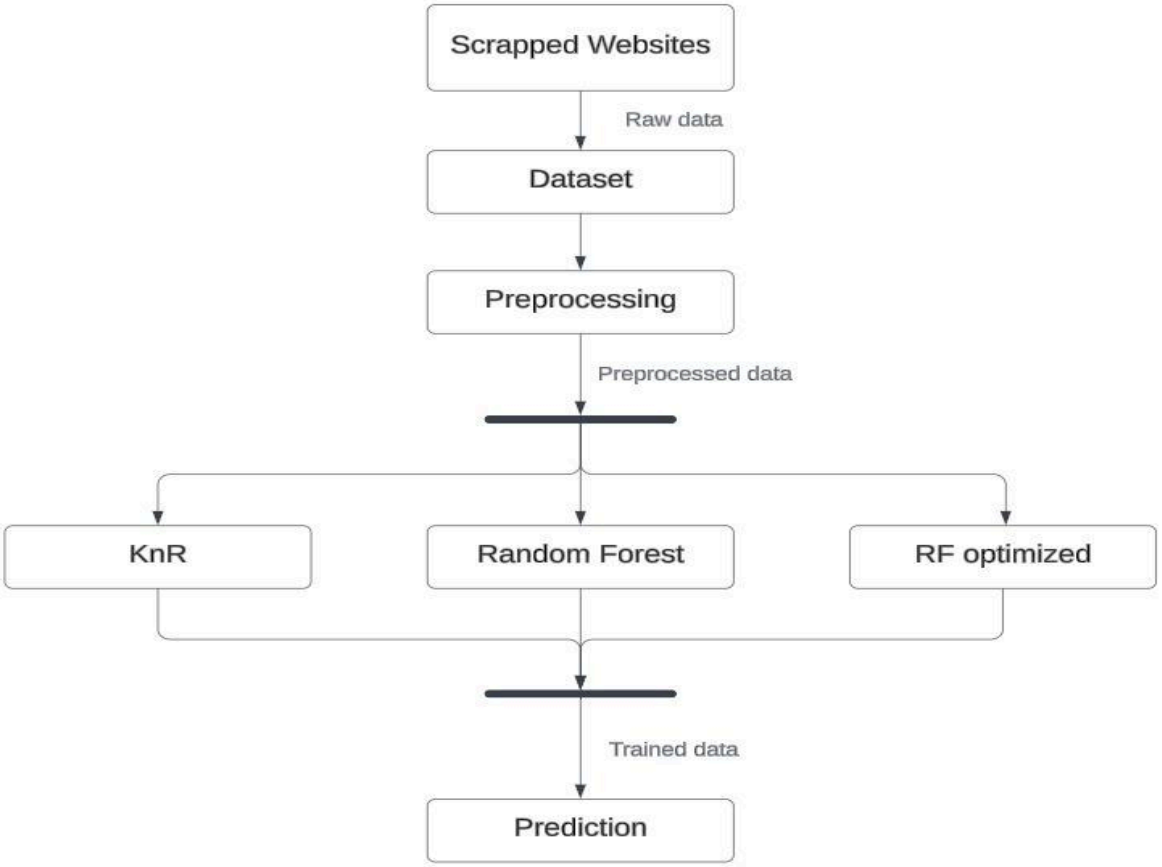
Literature Survey

S.No	Title	Description
1	Das Adhikary, D.R., Sahu, R., Pragyna Panda, S. (2022). Prediction of Used Car Prices Using Machine Learning. In: Dehuri, S., Prasad Mishra, B.S., Mallick, P.K., Cho, SB. (eds) vol 271. Springer, Singapore.	The paper proposes a model to predict the cost of used cars, leveraging machine learning algorithms such as k-nearest neighbor (KNN), random forest regression, decision tree, and light gradient boosting machine (LightGBM).
2	B. Hemendiran and P. N. Renjith, "Predicting the Prices of the Used Cars using Machine Learning for Resale," 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2023, pp. 1-5, doi: 10.1109/SCEECS57921.2023.10063133.	Various models including Random Forest Regressor, Extra Tree Regressor, Bagging Regressor, Decision Tree, and XG Boost are applied. The Random Forest model emerges as the most accurate. This study underscores the significance of selecting the appropriate model for achieving enhanced reliability and precision in forecasting used car prices.
3	N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand, 2018, pp. 115-119, doi: 10.1109/ICBIR.2018.8391177.	Through dataset collection, pre-processing, and exploratory data analysis, various regression techniques including Linear Regression, LASSO Regression, Decision Tree, Random Forest, and Extreme Gradient Boosting are employed with evaluation metrics such as MAE, MSE, and RMSE utilized to assess model accuracy.
4	B. Hemendiran and P. N. Renjith, "Predicting the Prices of the Used Cars using Machine Learning for Resale," 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2023, pp. 1-5, doi: 10.1109/SCEECS57921.2023.10063133.	This paper explores the application of supervised machine learning models to accurately forecast the price of used cars based on their attributes. Multiple models, including Random Forest Regressor, Extra Tree Regressor, Bagging Regressor, Decision Tree, and XG Boost, are employed to generate forecasts. Random Forest emerged as the most accurate predictor.

5	Gongqi, S., Yansong, W., Qiang, Z.: New model for residual value prediction of the used car based on BP neural network and nonlinear curve fit. In: 2011 Third International Conference on Measuring Technology and Mechatronics Automation, vol. 2, pp. 682–685. IEEE (2011)	This paper presents a novel model for predicting the residual value of privately-owned used cars, considering factors such as manufacturer, mileage, and age. The model combines the BP neural network with nonlinear curve fitting to optimize flexibility and accuracy. Initially, distribution curves of residual values over time are analyzed. Subsequently, the BP neural network is established to extract features from these distribution curves under different conditions.
6	Noor, Kanwal, and Sadaqat Jan. "Vehicle price prediction system using machine learning techniques." International Journal of Computer Applications 167, no. 9 (2017): 27-31.	This paper introduces a vehicle price prediction system employing supervised machine learning techniques, specifically multiple linear regression, which achieves a high prediction precision of 98%. Multiple linear regression is utilized, where there are multiple independent variables and one dependent variable (price) whose actual and predicted values are compared to assess precision.
7	N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand, 2018, pp. 115-119, doi: 10.1109/ICBIR.2018.8391177.	Pre-processing involved removing null, redundant, and missing values. In this supervised learning study, three regressors (Random Forest, Linear Regression, and Bagging) were trained, tested, and compared against a benchmark dataset. Random Forest Regressor yielded the highest score of 95%, with 0.025 MSE, 0.0008 MAE, and 0.0378 RMSE. Bagging Regression followed with an 88% score, and Linear Regression with 85%. A train-test split of 80/20 with 40 random states was employed.
8	C. V. Narayana, C. L. Likhitha, S. Bademiya and K. Kusumanjali, "Machine Learning Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business", 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1680-1687, 2021.	Modified Decision Tree (MDT), LightGBM, and XGBoost regressions. Supervised machine learning algorithms were employed to thoroughly analyze datasets through training, testing, modeling, and cross-validation. Four performance metrics (MAE, MSE, MAPE, and R2) were utilized to assess and compare accuracy. With a coefficient of determination of 0.9284, MDT demonstrated the highest prediction accuracy, followed by LightGBM (0.8765) and XGBoost (0.8493).

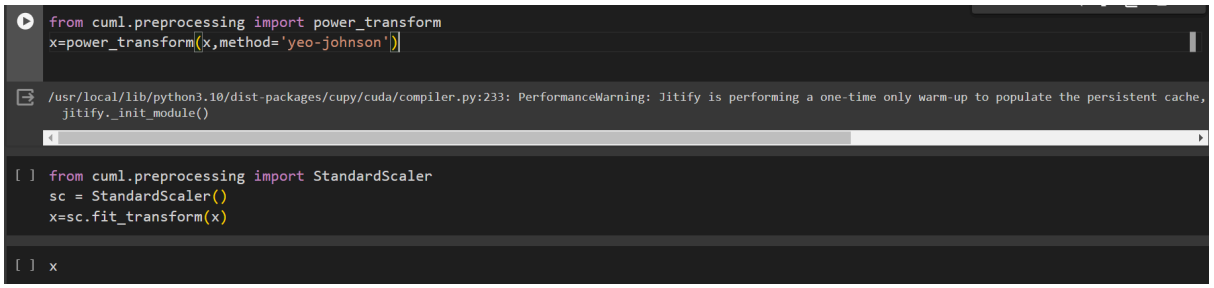
Methodology

Data Flow Diagram



In the preprocessing step of the car price prediction project, power transforms and standard scaling are applied to enhance the data quality and improve the performance of the machine learning models. To address skewness in the dataset, power transforms such as the Box-Cox or Yeo-Johnson methods are utilized. Additionally, standard scaling is employed to standardize the continuous numeric features. By employing power transforms and standard scaling techniques, the preprocessing phase aims to optimize the data

distribution, reduce skewness, and normalize the features, creating a more suitable input for the subsequent machine learning algorithms.



```
from cuml.preprocessing import power_transform
x=power_transform(x,method='yeo-johnson')
```

```
/usr/local/lib/python3.10/dist-packages/cupy/cuda/compiler.py:233: PerformanceWarning: Jitify is performing a one-time only warm-up to populate the persistent cache,
  jitify._init_module()
```

```
[ ] from cuml.preprocessing import StandardScaler
    sc = StandardScaler()
    x=sc.fit_transform(x)

[ ] x
```

The model will be trained using a dataset comprising over 5000 tuples. The value of a car is determined by factors such as the number of kilometres driven, the year of registration, the kind of gasoline used, and the financial strength of the owner. We created regressor methods and compared the two on different car models because this is a regression problem.

KNeighborsRegressor Algorithm:

Calculating the average of the numerical goal of the K nearest neighbours is a straightforward implementation of KNN regression. An inverse distance weighted average of the K closest neighbours is another method. The distance functions used in KNN regression are the same as those used in KNN classification.

```
[ ] from cuml.neighbors import KNeighborsRegressor
    from cuml.metrics import r2_score

[ ] knr = KNeighborsRegressor()
    knr.fit(xtrain,ytrain,convert_dtype=True)
    pred_train_knr=knr.predict(xtrain)
    pred_test_knr=knr.predict(xtest)

    ytest.dtype

    dtype('int64')

[ ] ytest = np.array(ytest, dtype=np.float64)
    pred_test_knr = np.array(pred_test_knr, dtype=np.float64)
```

Random Forest Regressor A regressor with a random forest. A random forest is a meta estimator that employs averaging to increase predicted

accuracy and control over-fitting by fitting a number of classification decision trees on various sub-samples of the dataset.

```
[ ] from cuml.ensemble import RandomForestRegressor

rf=RandomForestRegressor()
rf.fit(xtrain,ytrain)
pred_train_rf=rf.predict(xtrain)
pred_test_rf=rf.predict(xtest)
print('Random Forest Regressor r2_score:',r2_score(ytest,pred_test_rf))
```

We further used Grid search which is a widely used hyperparameter optimization technique that systematically explores a predefined grid of hyperparameter combinations to identify the optimal configuration for a machine learning model. In the context of our model, grid search is applied to the Random Forest Regressor model to maximize its predictive accuracy.

```
from cuml.model_selection import GridSearchCV

parameter = { 'bootstrap': [True, False],
 'max_features': ['auto', 'sqrt'],
 'min_samples_leaf': [1, 2, 4],
 'min_samples_split': [2, 5, 10],}

gvc = GridSearchCV(RandomForestRegressor(),parameter,cv=5)
gvc.fit(xtrain,ytrain)
gvc.best_params_
```

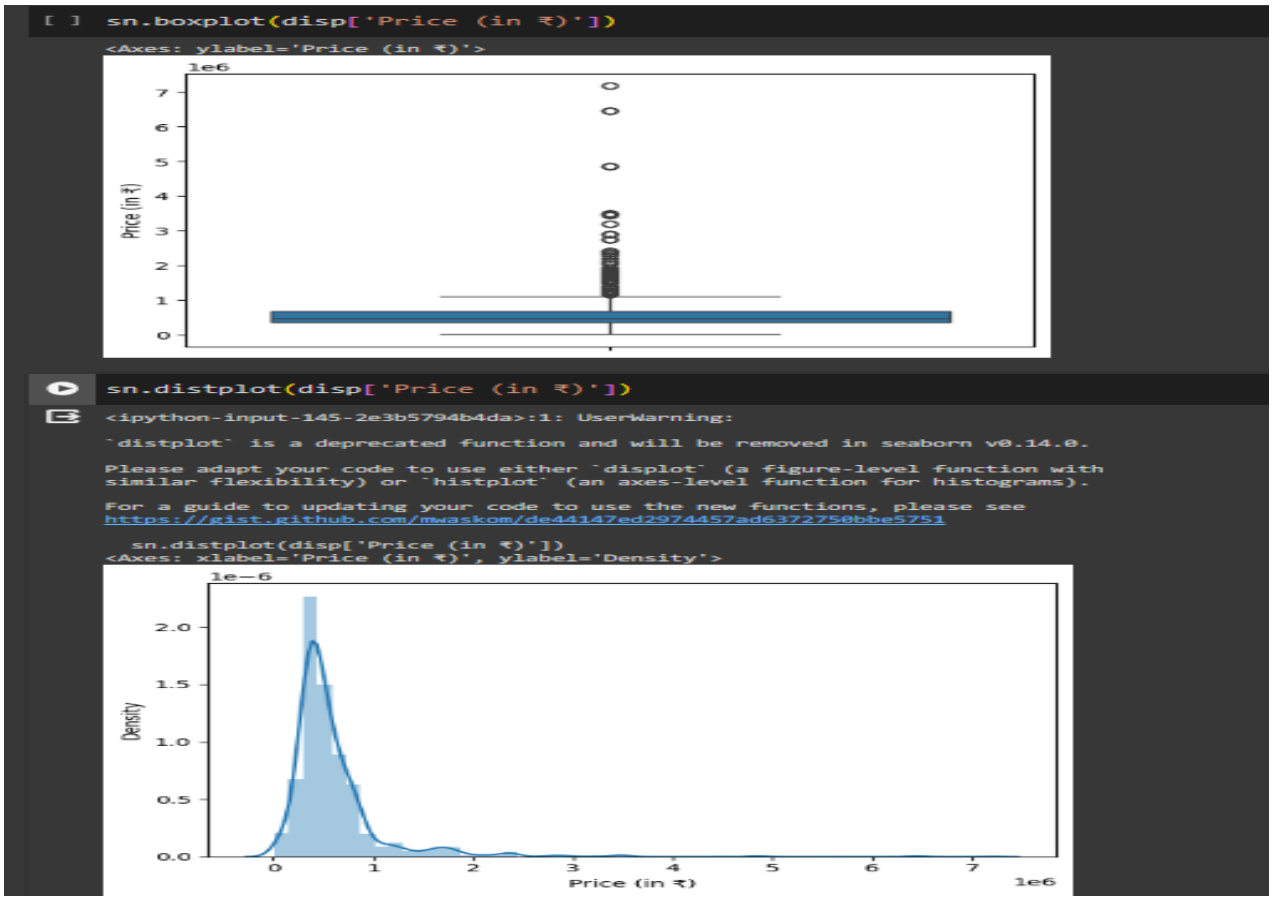
```
/usr/local/lib/python3.10/dist-packages/cuml/internals/api_decorators.py:188: UserWarning: To use pickling first train using float32 data to fit the estimator
  ret = func(*args, **kwargs)
{'bootstrap': True,
 'max_features': 'auto',
 'min_samples_leaf': 1,
 'min_samples_split': 2}
```

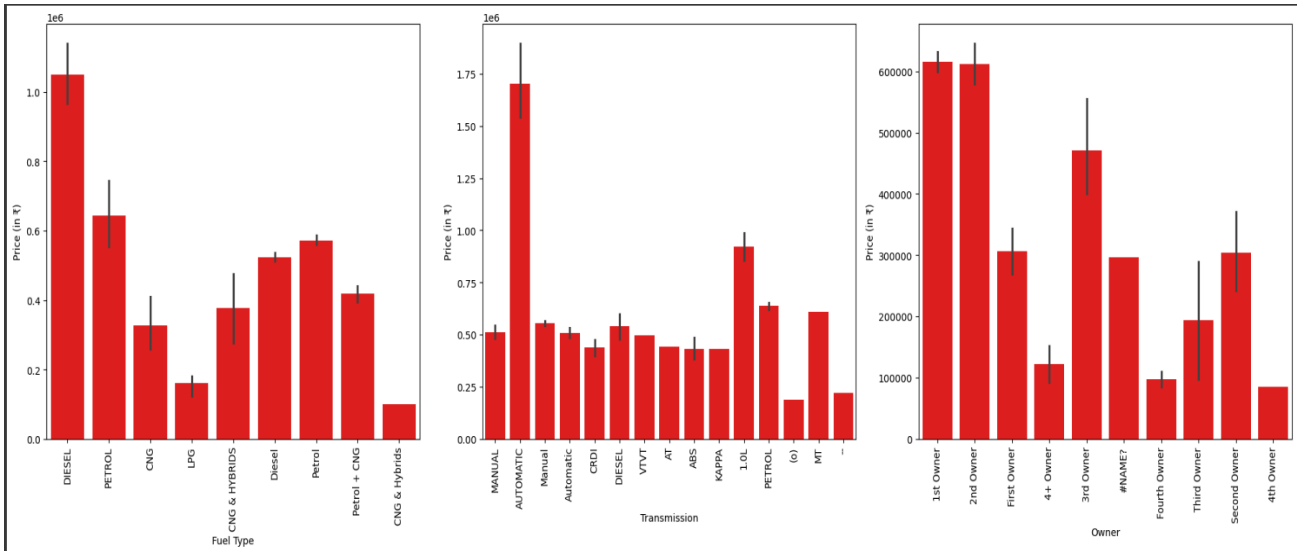
```
[ ] pricecar = RandomForestRegressor(bootstrap=True ,min_samples_leaf=1, max_features='auto', min_samples_split=2 , n_estimators=1000)
pricecar.fit(xtrain,ytrain)
pred=pricecar.predict(xtest)
acc=r2_score(ytest,pred)
print('Random Forest Regressor Tuned r2_score:',acc)
```

Dataset Analysis

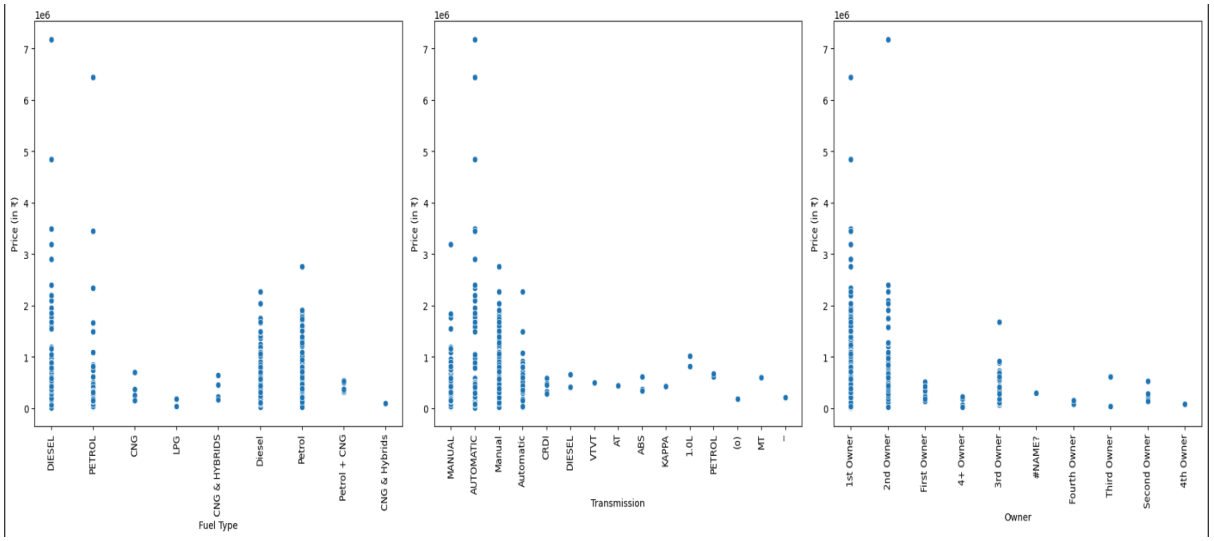
The dataset for the car price prediction project consists of various attributes related to the cars, including the brand, model name, manufacturing year, variant, fuel type, number of owners, location,, transmission type, driven kilometers, and, most importantly, the price. The price attribute is the target variable and is represented as an integer. The remaining attributes are of object data type, indicating categorical or textual information about the cars. These attributes collectively provide valuable insights into the characteristics and specifications of the cars, which will be utilized in the analysis and modeling stages of the project to predict the prices accurately.

We have plotted histograms and distribution plots in univariate analysis, which interpreted that all the columns are equally important but the columns like brand, variant, location, date and total driven kilometers have a wide range of data spread hence we will not perform its bivariate analysis.



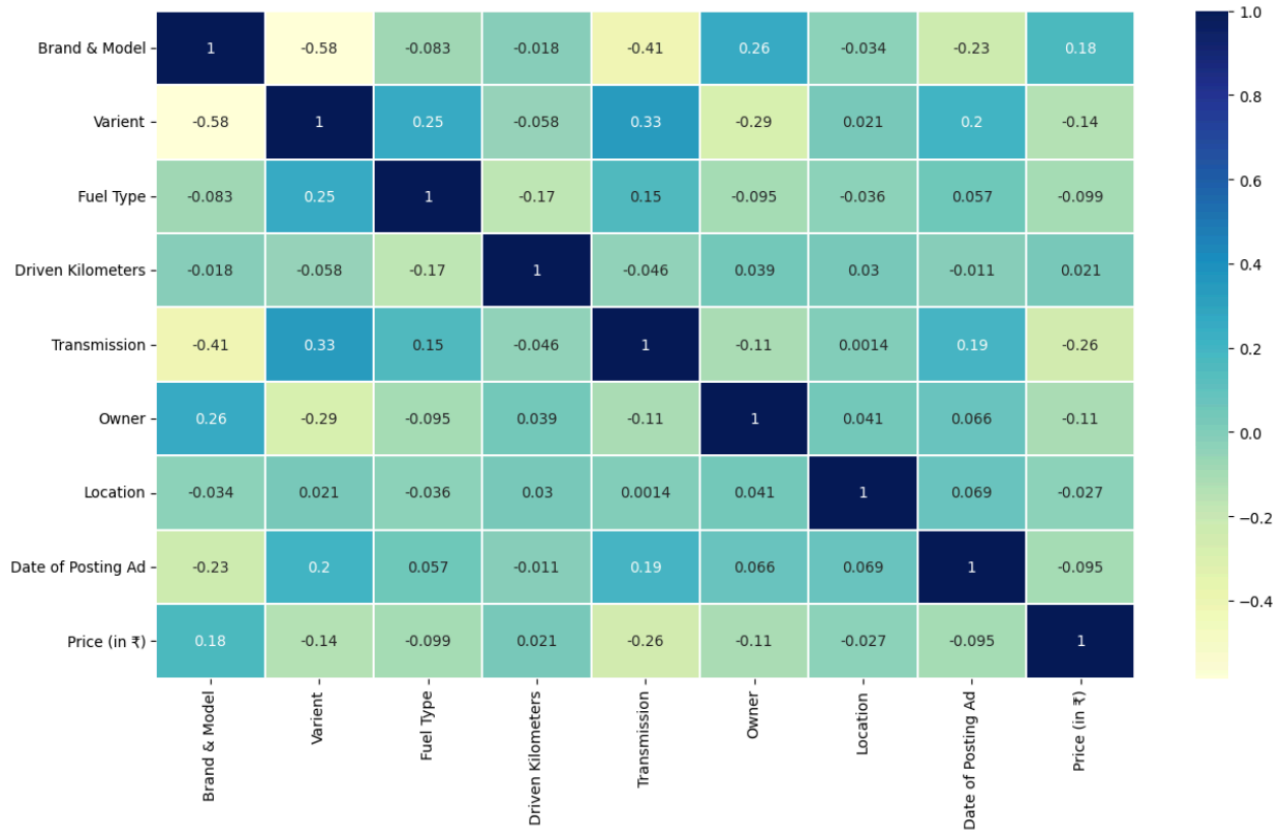


From bivariate analysis we conclude that, Since Brands, Variants, Driven Kilometers & Location have a wide range of values in them, we will not perform bivariate analysis for them as they will not give us any specific details. Now by plotting graph of Fuel type, Transmission and Owner against Price, we conclude that Car that uses Diesel, have automatic Transmission and Has only 1 owner is more likely to have a high price

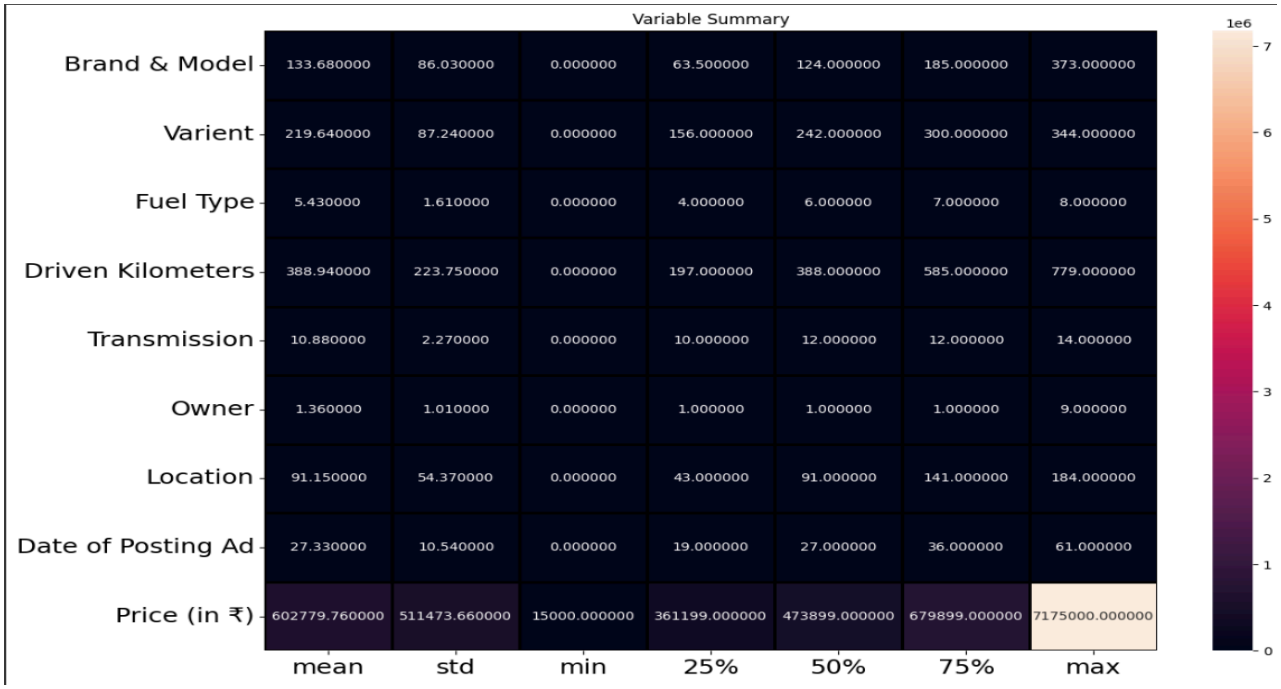


Just like bar graph we can see that Price range is likely to be high for cars using diesel as fuel or having automatic transmission or is owned by one owner .

The multivariate analysis done by plotting heatmap says that there is no multicollinearity in the dataset.



Statistical analysis of the Dataset shows the following result


Variable Summary
Brand & Model
Varient
Fuel Type
Driven Kilometers
Transmission
Owner
Location
Date of Posting Ad
Price (in ₹)
133.680000 86.030000 0.000000 63.500000 124.000000 185.000000 373.000000
219.640000 87.240000 0.000000 156.000000 242.000000 300.000000 344.000000
5.430000 1.610000 0.000000 4.000000 6.000000 7.000000 8.000000
388.940000 223.750000 0.000000 197.000000 388.000000 585.000000 779.000000
10.880000 2.270000 0.000000 10.000000 12.000000 12.000000 14.000000
1.360000 1.010000 0.000000 1.000000 1.000000 1.000000 9.000000
91.150000 54.370000 0.000000 43.000000 91.000000 141.000000 184.000000
27.330000 10.540000 0.000000 19.000000 27.000000 36.000000 61.000000
602779.760000 511473.660000 15000.000000 361199.000000 473899.000000 679899.000000 7175000.000000
mean
std
min
25%
50%
75%
max
1e6
7
6
5
4
3
2
1
0

Results

CUML

Sno	Model/Algorithm	Score(r2)	Execution Time(s)
1	Knr	0.5211	0.04
2	Random Forest	0.8861	1.7581
3	Random Forest Tuned	0.8875	3.5496
4	Grid-Search	-	142.51
5	SGD	0.0524	0.054

SKlearn

Sno	Model/Algo	Score(R2)	Execution Time(s)
1	Knr	0.5272	0.527
2	Decision Tree	0.7960	0.669
3	Random Forest	0.879	2.97
4	Random-Forest Tuned	0.8969	20.4027
5	Grid-Search	-	175.23
6	SGD	0.0544	0.079

Conclusion

Car price prediction has picked researchers' interest since it takes a significant amount of work and expertise on the part of the field expert. For a dependable and accurate forecast, a large number of unique attributes are analysed Using well-known algorithms from Python libraries, we were able to successfully construct machine learning algorithmic paradigms. On our dataset, we first do pre-processing and data cleaning. We trimmed the tuples that contained null values, which accounted for less than 1% of the total. The findings revealed a positive relationship between price and kilometres travelled, as well as year of registration and kilometres travelled . The respective performances of different algorithms were then compared to discover the one that best suited the existing data set .

Future Scope

As a part of future work, we aim at the variable choices over the algorithms that were used in the project. More advanced algorithms such as xg boost can be applied to further improve the performance of the model .

References

- Das Adhikary, D.R., Sahu, R., Pragyna Panda, S. (2022). Prediction of Used Car Prices Using Machine Learning. In: Dehuri, S., Prasad Mishra, B.S., Mallick, P.K., Cho, SB. (eds) vol 271. Springer, Singapore. https://doi.org/10.1007/978-981-16-8739-6_11

- Zhu, Yian. (2023). Prediction of the price of used cars based on machine learning algorithms. Applied and Computational Engineering. 6. 785-791. 10.54254/2755-2721/6/20230917.

- B. Hemendiran and P. N. Renjith, "Predicting the Prices of the Used Cars using Machine Learning for Resale," 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2023, pp. 1-5, doi: 10.1109/SCEECS57921.2023.10063133.

- Gongqi, S., Yansong, W., Qiang, Z.: New model for residual value prediction of the used car based on BP neural network and nonlinear curve fit. In: 2011 Third International Conference on Measuring Technology and Mechatronics Automation, vol. 2, pp. 682–685. IEEE (2011)

- Noor, Kanwal, and Sadaqat Jan. "Vehicle price prediction system using machine learning techniques." International Journal of Computer Applications 167, no. 9 (2017): 27-31.

- N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand, 2018, pp. 115-119, doi: 10.1109/ICBIR.2018.8391177.

- Pudaruth, Sameerchand. "Predicting the price of used cars using machine learning techniques." Int. J. Inf. Comput. Technol 4.7 (2014): 753-764.

- Noor, Kanwal, and Sadaqat Jan. "Vehicle price prediction system using machine learning techniques." International Journal of Computer Applications 167, no. 9 (2017): 27-31.

- Al-Jarrah, Omar & Yoo, Paul & Muhaidat, Sami & Karagiannidis, George & Taha, Kamal. (2015). Efficient Machine Learning for Big Data: A Review. Big Data Research. 2. 87–93. 10.1016/j.bdr.2015.04.001.

- C. V. Narayana, C. L. Likhitha, S. Bademiya and K. Kusumanjali, "Machine Learning Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business", 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1680-1687, 2021.

- S. K. Satapathy, R. Vala and S. Virpariya, "An Automated Car Price Prediction System Using Effective Machine Learning Techniques", 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), pp. 402-408, 2022.

- N. O. G. Madhuri, C. V. Narayana, A. NagaSindhu, M. Aksha and C. Naveen, "Second Sale Car Price Prediction using Machine Learning Algorithm", 2022 7th International Conference on Communication and Electronics Systems (ICCES), pp. 1171-1177, 2022.

- Pandey, Abhishek, Vanshika Rastogi, and Sanika Singh. "Car's Selling Price Prediction using Random Forest Machine Learning Algorithm (2020)." In 5th International Conference on Next Generation Computing Technologies (NGCT-2019). https://dx.doi.org/10.2139/ssrn, vol. 3702236. 2019.

- Kiran, S. "Prediction of resale value of the car using linear regression algorithm." Int. J. Innov. Sci. Res. Technol 6.7 (2020): 382-386.

- J. D. Wu, C. C. Hsu and H. C. Chen, "An expert system of price forecasting for used cars using adaptive neurofuzzy inference", Expert Systems with Applications, vol. 36, no. 4, pp. 7809-7817, 2009.

- Suma, V., and Shavige Malleshwara Hills. "Data mining based prediction of demand in Indian market for refurbished electronics." Journal of Soft Computing Paradigm (JSCP) 2, no. 02 (2020): 101-110.

- TY JOUR, K. AU-Madhuvanthi, AU-Kailasanathan, AU-N C Nallakaruppan, AU-Somayaji Senthilkumar and PY Siva, "T1 - Car Sales Prediction Using Machine Learning Algorithmns", JO - International Journal of Innovative Technology and Exploring Engineering, vol. 8, 03 2019.