22AIE213 - MACHINE LEARNING

PROJECT REPORT

CAR PRICE PREDICTION

GROUP 8

SUBMISSION DATE:
17-07-2024

Team Members

TEAM MEMBERS	ROLL NUMBER
Amal Shaji	AM.EN.U4AIE22104
Gayathri Krishna G	AM.EN.U4AIE22118
Harigovind C B	AM.EN.U4AIE22119
Satvik Mishra	AM.EN.U4AIE22148

I. Abstract

The goal of this project is to develop a car price prediction algorithm aimed at accurately determining car prices based on various vehicle attributes. By analyzing historical car sales data, the model will consider factors such as make, model, year, mileage, and additional features to estimate the market value of cars. This predictive tool will be valuable for buyers, sellers, and dealers, enabling informed decision-making and fair pricing in the automotive market. The model's performance will be evaluated using Root Mean Squared Error (RMSE), ensuring its effectiveness and reliability. This project leverages datasets from sources like Kaggle and various used car price databases, with the ultimate goal of integrating the model into a scalable web application for widespread user access and utility.

II. Table of contents

III. Introduction

The used car market has witnessed remarkable growth in recent years, becoming a significant segment of the global automotive industry. This expansion is driven by various factors, including economic fluctuations, increasing environmental awareness, and the rapid depreciation of new vehicles. Consumers are increasingly turning to used cars as a cost-effective and environmentally friendly alternative to new car purchases. This trend has made the accurate prediction of used car prices an essential task for both buyers and sellers. Accurate price predictions enable consumers to make informed purchasing decisions, ensuring they receive fair value for their investment. Simultaneously, sellers can set competitive prices that attract buyers while optimizing their returns. This project aims to address this critical need by developing robust machine learning models to predict the prices of used cars using three distinct datasets.

The importance of accurate used car price predictions cannot be overstated. For buyers, understanding the fair market value of a used car helps in negotiating prices and avoiding overpayment. For sellers, particularly dealerships and individual sellers, setting the right price is crucial to remain competitive in a crowded market. Overpricing can deter potential buyers, while underpricing can lead to significant financial losses. Moreover, accurate pricing is beneficial for financial institutions involved in auto financing and insurance companies assessing the value of vehicles for coverage policies. Thus, precise price predictions can streamline transactions, reduce uncertainties, and enhance the overall efficiency of the used car market.

In this project, we focus on regression techniques to predict used car prices, leveraging their ability to model and capture the complex relationships between car attributes and their respective prices. Regression analysis is a supervised machine learning technique used to predict the continuous value of the dependent variable based on one or more independent variables. By applying regression models, we aim to understand how various factors, such as make and model, year of manufacture, mileage, condition, and location, influence the price of a used car. The datasets employed in this project are rich with features that are indicative of a car's value, making them ideal for building predictive models.

IV. Literature review

[1]This paper contains a comparative analysis of Random Forest and LightGBM models for predicting used car prices. The evaluation metrics used in this study included Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The findings revealed that the Random Forest model achieved an MSE of 0.0373, an MAE of 0.125, and an R^2 of 0.936. In comparison, the LightGBM model recorded an MSE of 0.0385, an MAE of 0.117, and an R^2 of 0.933. These results indicate that both models are highly effective in predicting used car prices, with LightGBM having a slightly lower prediction error.The study concluded that the LightGBM model, due to its lower prediction error, could be considered for application in other fields in future research. This comparison demonstrates the robustness and potential of both models in handling regression tasks in the automotive market and underscores the relevance of exploring these algorithms further in various applications.

[2]The study utilized 40 sample sets to compare the Novel XGBRegressor Optimizer with the ExtraTree Regressor algorithm, a variant of the Random Forest algorithm that uses extremely randomized trees to reduce prediction variance. The Novel XGBRegressor Optimizer was found to enhance the system's overall execution by using the best parameters during prediction. The evaluation, conducted using ClinCalc software, considered an alpha value of 0.05, a G-Power of 0.8, and a 95% Confidence Interval (CI). The results showed that the Novel XGBRegressor Optimizer achieved an accuracy of 82.7%, while the ExtraTree Regressor achieved 78.2%. The statistical significance of the results (p=0.000) further underscores the robustness of the Novel XGBRegressor Optimizer.This study's findings demonstrate the potential of optimized machine learning algorithms in achieving higher prediction accuracy, which is particularly relevant to our project. By focusing on model optimization, the study highlights how gradient boosting and parameter tuning can significantly enhance predictive performance.

[3]This paper proposes a more cutting-edge machine learning algorithm based on LightGBM to predict used car prices authoritatively and innovatively using actual transaction records of a used car trading platform. Firstly, the original dataset is cleaned, and the features are filtered by analyzing the importance of each feature. The filtered features are then fed into LightGBM, and the parameters of

LightGBM are tuned using grid search technology to improve prediction accuracy. Extensive experimental results have demonstrated that the proposed used car price forecasting algorithm has better prediction accuracy compared with classical linear regression, SVM, RandomForest, GBDT and other algorithms.

Our project aims to leverage the strengths of LightGBM, along with other models such as XGBoost and ExtraTrees, to develop robust predictive models for used car prices. By comparing the performance of these models across multiple datasets, we intend to identify the most effective approach for accurate price prediction, thereby contributing to the existing body of research in this domain.

V. Methodology

Initially the data was not clean, so to ensure the data was clean we had to preprocess the data. In the data preprocessing step, we performed several critical tasks to ensure our dataset was ready for model training. First, we addressed missing values using K-Nearest Neighbors (KNN) imputation and mode imputation techniques. KNN imputation helped fill in missing values by considering the nearest neighbors of each data point. The algorithm identifies 'k' samples in the training set that are closest to the point with missing data and averages their values to impute the missing entries. This method maintains the data's inherent structure and relationships.

For categorical features, we used mode imputation, which replaces missing values with the most frequently occurring value in the feature. This approach is simple yet effective in preserving the distribution of categorical variables.

Next, we transformed categorical features into numerical. We applied one-hot encoding to convert categorical variables into binary vectors, ensuring that the models could interpret the categorical information without assuming any ordinal relationship. Additionally, label encoding was used for features with an intrinsic order, converting each unique category into a numerical label.

By performing these preprocessing steps, we ensured that our dataset was clean, complete, and appropriately formatted for model training, leading to more accurate predictions.

We have chosen three models namely LightGBM, XGBoost and ExtraTrees Regressor mostly because of their ability to evaluate large datasets. LightGBM(Light Gradient Boosting Machine) is a highly efficient gradient boosting framework. We are using LightGBM Regressor, a part of the LightGBM library as it is well-suited for handling large datasets with high-dimensional features, making it an excellent choice. It uses tree-based learning where the tree grows leaf-wise (best-first). This leads to deeper trees and potentially better accuracy by capturing more complex relationships. It is particularly chosen for its features like speed and efficiency, regularization technique, hyperparameter tuning etc.

XGBoost (Extreme Gradient Boosting) is highly efficient and scalable, designed for speed with parallel processing capabilities. It includes built-in regularization and is adept at managing sparse datasets. It sequentially builds decision trees to correct prediction errors, updating predictions iteratively until optimal model performance is achieved through features like early stopping and cross-validation. LightGBM grows trees leaf-wise (best-first), whereas XGBoost grows trees level-wise (depth-wise).

ExtraTrees (Extremely Randomized Trees), is an ensemble learning method that builds multiple decision trees using random subsets of features and thresholds. It introduces extra randomness by selecting splits and thresholds randomly rather than based on optimization criteria like entropy, which can lead to reduced variance and improved generalization. It is computationally efficient and effective for handling high-dimensional data or noisy datasets where traditional decision trees or Random Forests may struggle to find optimal splits.

In the final step of our methodology, we focused on hyperparameter tuning to optimize the performance of our machine learning models. Hyperparameters are crucial parameters that govern the learning process and the structure of the models, and tuning them correctly can significantly enhance model performance. We employed grid search for this purpose, which is a systematic approach to hyperparameter optimization. Grid search exhaustively tests a predefined set of hyperparameter values and evaluates the performance of the model for each combination. This allows us to identify the optimal hyperparameter settings that yield the best model performance.

After completing the grid search, we proceeded with model evaluation to assess the effectiveness of the optimized models. We used Root Mean Squared Error (RMSE) as the evaluation metric for this purpose. RMSE is a standard measure of the differences between predicted and observed values, and it effectively captures the model's prediction accuracy. By calculating the RMSE for each of the three models, we were able to compare their performance quantitatively.

This comprehensive approach, combining hyperparameter tuning with grid search and evaluating using RMSE, enabled us to fine-tune our models and select the best-performing one.

VI. Data analysis

In the field of machine learning and data science, the quality and structure of the datasets used are crucial for building accurate and robust models. For our used car price prediction project, we have gathered three distinct datasets, each with unique features and attributes. This analysis delves into the specifics of these datasets, examining their characteristics and evaluating their suitability for the task at hand.

Dataset-1 Analysis:

Dataset-1 comprises 4009 records and 12 attributes. These attributes include a mix of categorical and numerical data, with 7 categorical attributes and 5 numerical attributes. This dataset covers a comprehensive range of features essential for predicting car prices, such as brand reputation, model specifics, and condition indicators.

The presence of both categorical and numerical attributes in Dataset-1 allows for a multifaceted analysis. For instance, Brand and Model, being categorical, can be encoded to facilitate machine learning algorithms, while numerical attributes like Mileage and Engine size provide quantitative measures directly influencing the car's price. The balance between categorical and numerical data suggests that Dataset-1 is well-rounded for exploratory data analysis and feature engineering.

With 4009 records, Dataset-1 offers a moderate-sized sample that can help in identifying general trends and patterns within the used car market. The inclusion of accident history and detailed descriptions of the car's interior and exterior conditions enhances the dataset's depth, allowing for more nuanced predictions. However, the relatively smaller number of records compared to the other datasets may limit its effectiveness in capturing rare occurrences or outlier behaviors.

Dataset-2 Analysis

Dataset-2 includes 6019 records and 13 attributes, with a slightly different composition compared to Dataset-1. It has 5 categorical attributes and 8 numerical attributes. This dataset emphasizes location and engine power, which are crucial factors in car valuation but are not present in Dataset-1.

The increased number of records and attributes in Dataset-2 provides a broader scope for analysis. The higher proportion of numerical attributes facilitates statistical modeling and regression analysis, essential for predicting continuous variables like price. Attributes such as kilometers driven and Engine power are critical indicators of a vehicle's performance and longevity, directly impacting its market value.

Dataset-2’s additional attributes and larger size enhance its reliability and the robustness of any derived models. The inclusion of the Location attribute introduces geographic variability, which is significant given the regional differences in car pricing. This dataset's richer numerical data allows for more sophisticated machine learning techniques, such as gradient boosting or deep learning models, which can handle complex interactions between features.

Dataset-3 Analysis

The most extensive of the three, Dataset-3, contains 16735 records with 18 attributes. It has a balanced mix of 10 categorical attributes and 8 numerical attributes. This dataset provides the most detailed snapshot of the used car market.

The large number of records in Dataset-3 allows for more granular insights and more robust statistical analysis. The diverse range of attributes, including specifics like Drive Type and Fuel Consumption, means this dataset can capture the subtleties of car features and their impact on pricing. The extensive categorical data enables comprehensive segment analysis and clustering, while the rich numerical data supports precise regression models.

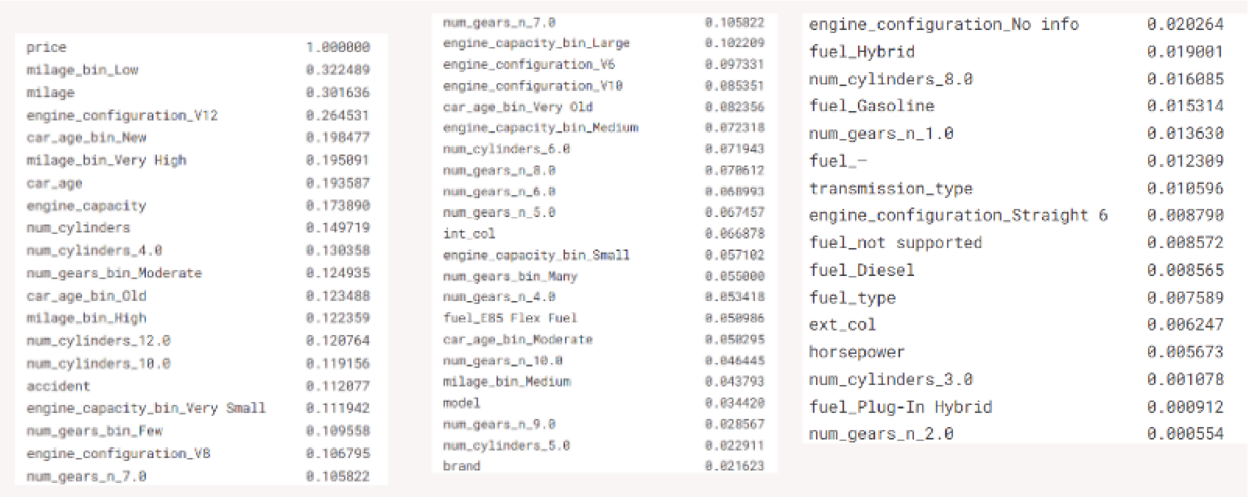
With 16735 records, Dataset-3 offers a significant advantage in training machine learning models due to its size and diversity. The comprehensive attribute set allows for a holistic analysis of the factors influencing car prices. This dataset is particularly suited for advanced machine learning techniques, such as ensemble methods and neural networks, which can leverage the extensive data to improve prediction accuracy. Moreover, the large sample size helps mitigate the risk of overfitting, enhancing the generalizability of the models.

The three datasets provide a robust foundation for building a used car price prediction model. Dataset-1 offers a balanced introduction with essential attributes, Dataset-2 provides additional numerical data and geographic considerations, and Dataset-3’s extensive records and attributes allow for sophisticated modeling and deep insights. Together, these datasets enable a comprehensive analysis and robust model development, essential for accurate and reliable used car price predictions.

VII. Result

Dataset-1

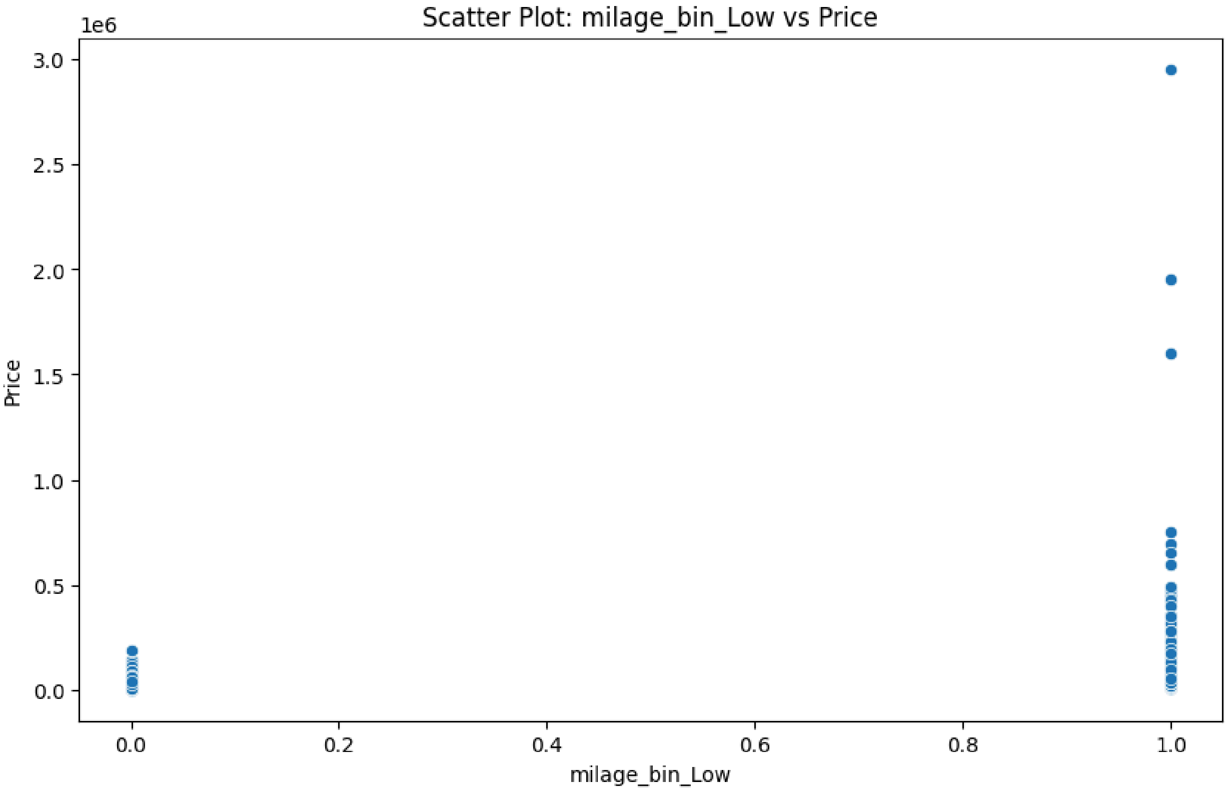
After cross checking with the Correlation matrix and the feature importance graph, we can see that the features mileage and engine capacity dominate.

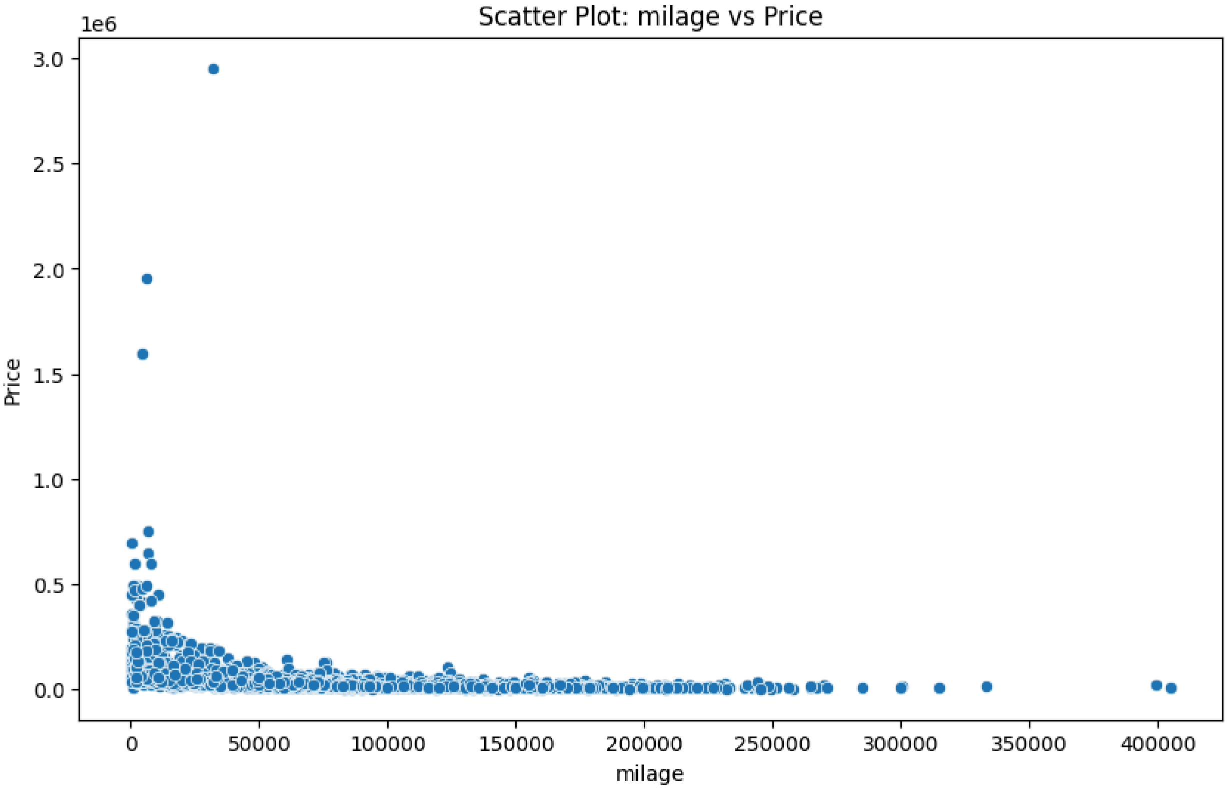


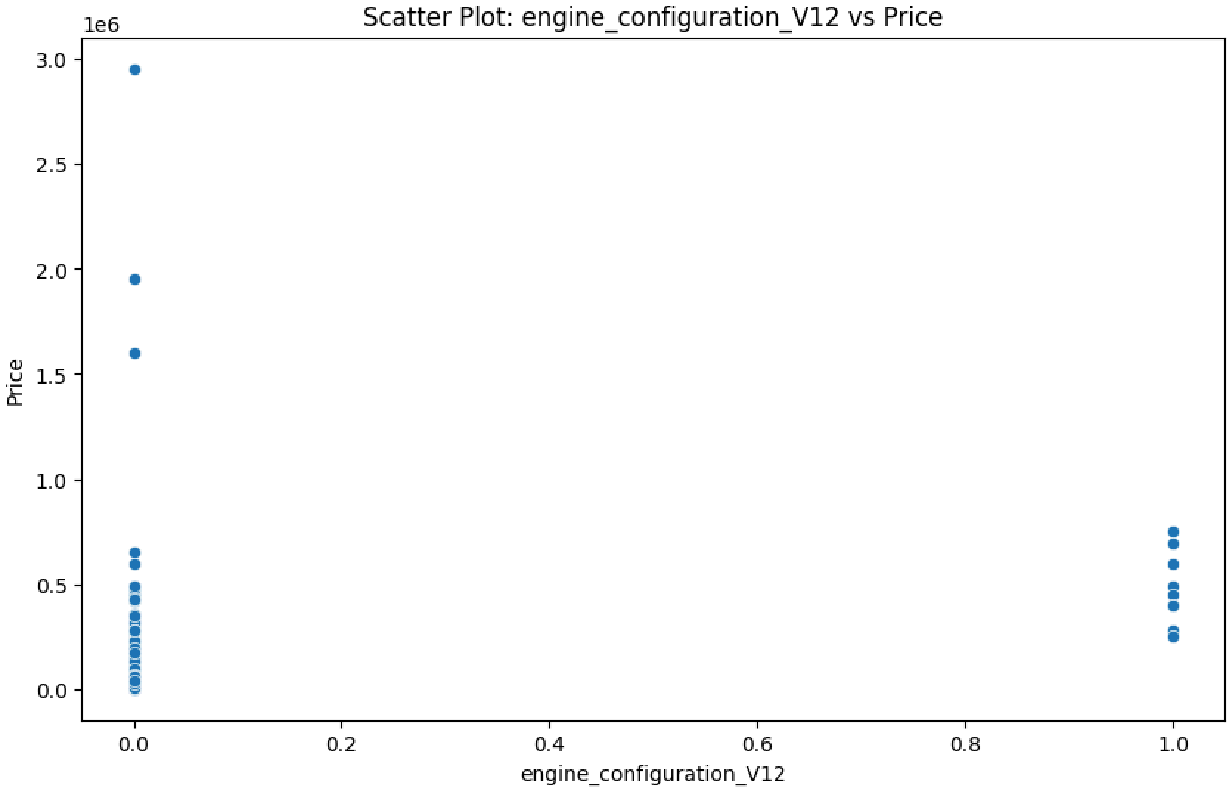
```
price                                    1.000000
milage_bin_Low                           0.322489
milage                                   0.301636
engine_configuration_V12                 0.264531
car_age_bin_New                          0.198477
milage_bin_Very High                     0.195091
car_age                                  0.193587
engine_capacity                          0.173890
num_cylinders                            0.149719
num_cylinders_4.0                        0.130358
num_gears_bin_Moderate                   0.124935
car_age_bin_Old                          0.123488
milage_bin_High                          0.122359
num_cylinders_12.0                       0.120764
num_cylinders_10.0                       0.119156
accident                                 0.112077
engine_capacity_bin_Very Small           0.111942
num_gears_bin_Few                        0.109558
engine_configuration_V8                  0.106795
num_gears_n_7.0                          0.105822
```

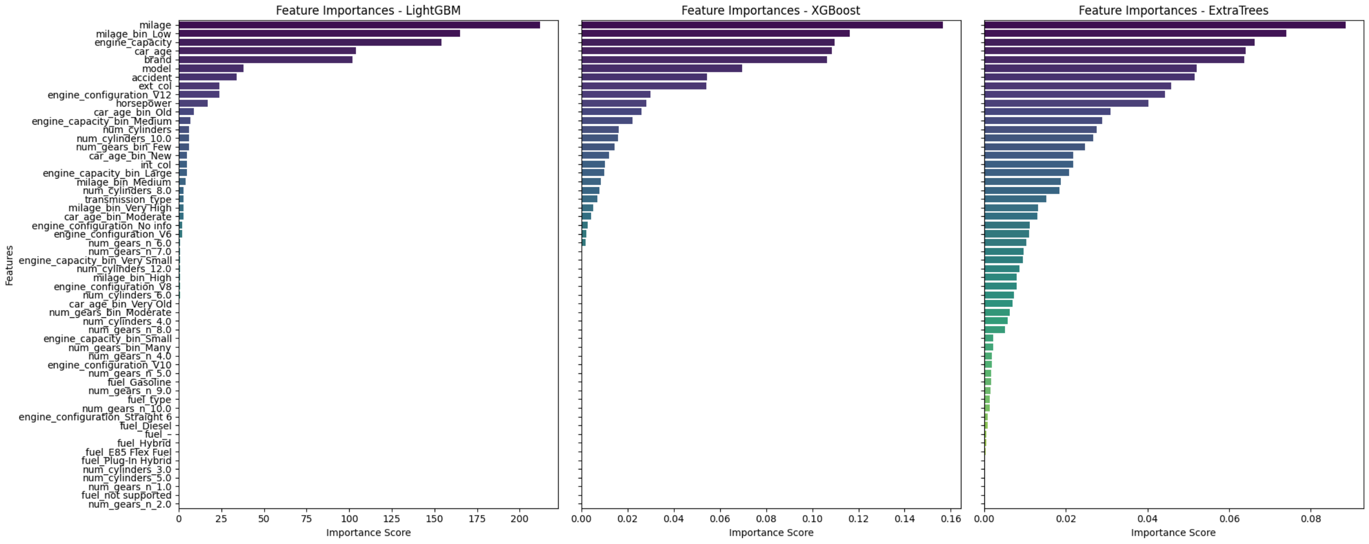
```
num_gears_n_7.0                          0.105822
engine_capacity_bin_Large                0.102209
engine_configuration_V6                  0.097331
engine_configuration_V10                 0.085351
car_age_bin_Very Old                     0.082356
engine_capacity_bin_Medium               0.072318
num_cylinders_6.0                        0.071943
num_gears_n_8.0                          0.070612
num_gears_n_6.0                          0.068993
num_gears_n_5.0                          0.067457
int_col                                  0.066878
engine_capacity_bin_Small                0.057102
num_gears_bin_Many                       0.055000
num_gears_n_4.0                          0.053418
fuel_E85 Flex Fuel                       0.050986
car_age_bin_Moderate                     0.050295
num_gears_n_10.0                         0.046445
milage_bin_Medium                        0.043793
model                                    0.034420
num_gears_n_9.0                          0.028567
num_cylinders_5.0                        0.022911
brand                                    0.021623
```

```
engine_configuration_No info             0.020264
fuel_Hybrid                              0.019001
num_cylinders_8.0                        0.016085
fuel_Gasoline                            0.015314
num_gears_n_1.0                          0.013630
fuel_–                                   0.012309
transmission_type                        0.010596
engine_configuration_Straight 6          0.008790
fuel_not supported                       0.008572
fuel_Diesel                              0.008565
fuel_type                                0.007589
ext_col                                  0.006247
horsepower                               0.005673
num_cylinders_3.0                        0.001078
fuel_Plug-In Hybrid                      0.000912
num_gears_n_2.0                          0.000554
```

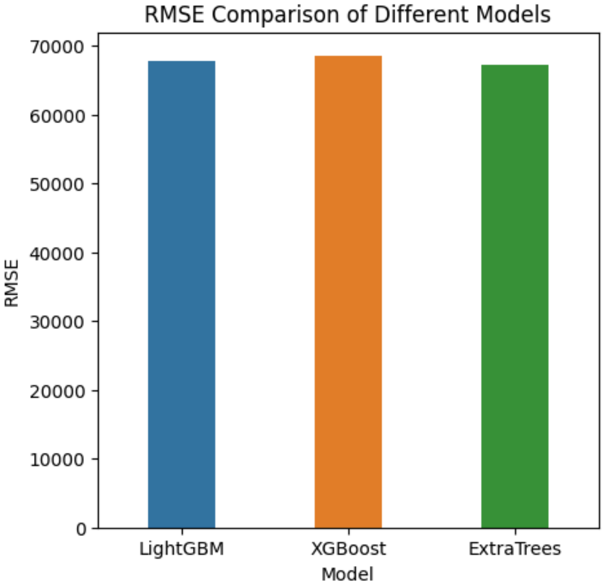

Scatter Plot: milage_bin_Low vs Price
1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Price
0.0
0.2
0.4
0.6
0.8
1.0
milage_bin_Low


Scatter Plot: milage vs Price
1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Price
0
50000
100000
150000
200000
250000
300000
350000
400000
milage



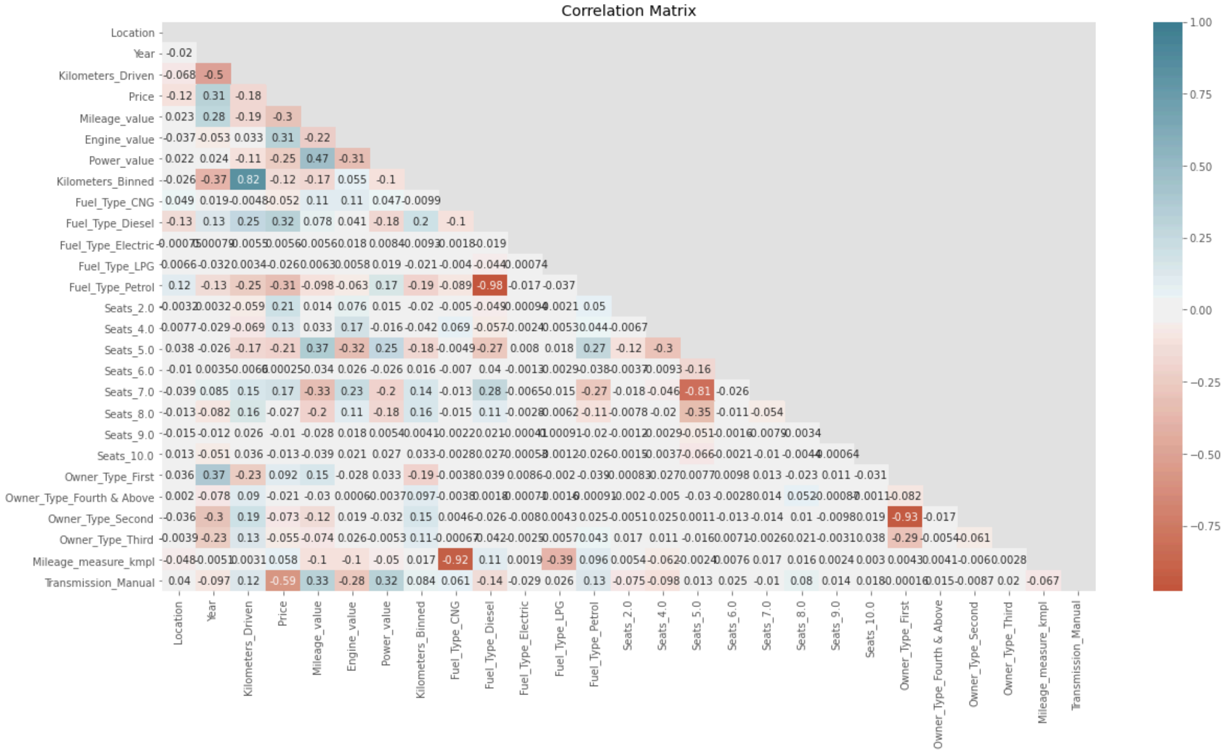


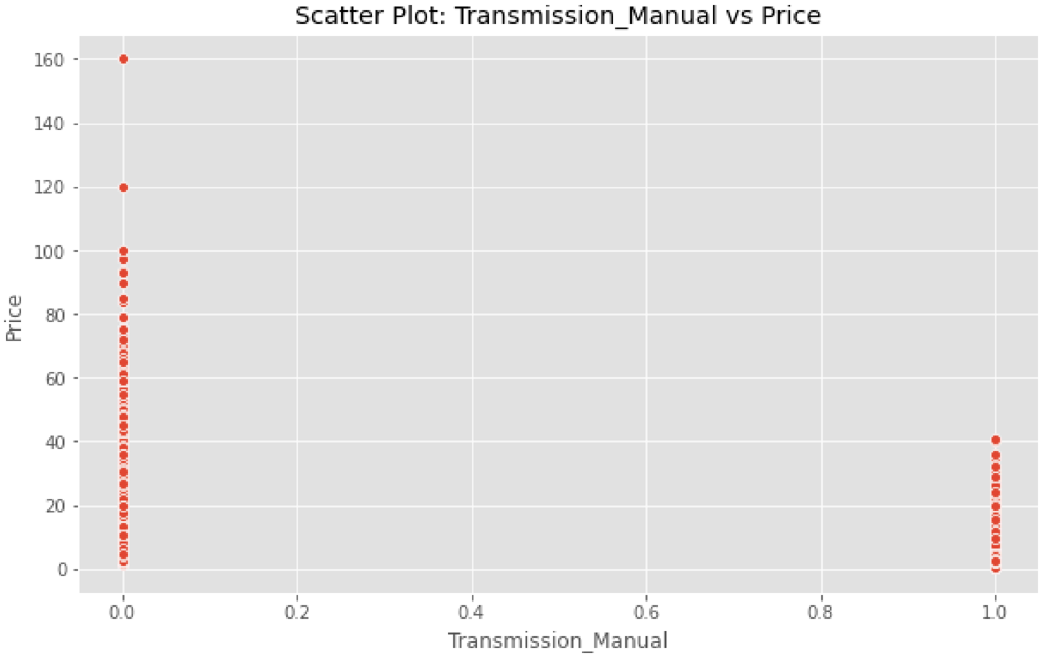
Then we have used Root Mean Square Error as an evaluation metric to compare the results of all three models. In Dataset 1, ExtraTrees Regressor has performed best.

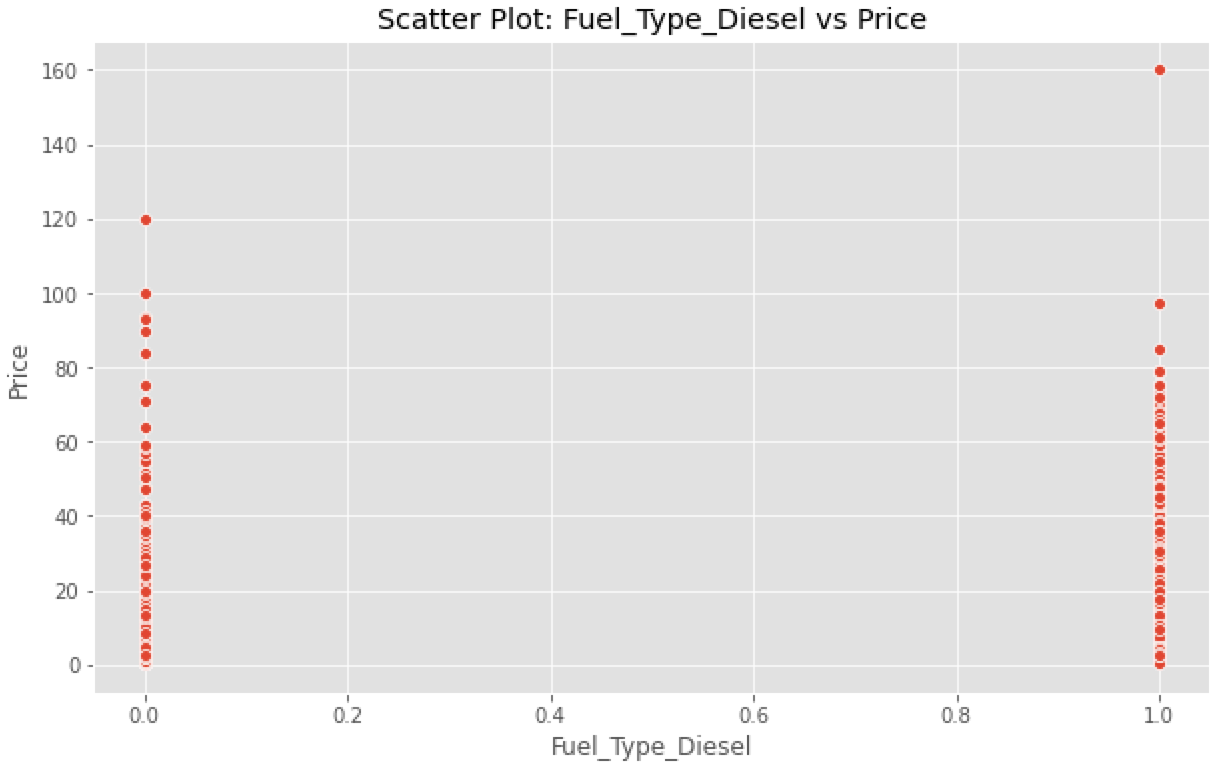

RMSE Comparison of Different Models
RMSE
70000
60000
50000
40000
30000
20000
10000
0
LightGBM
XGBoost
ExtraTrees
Model

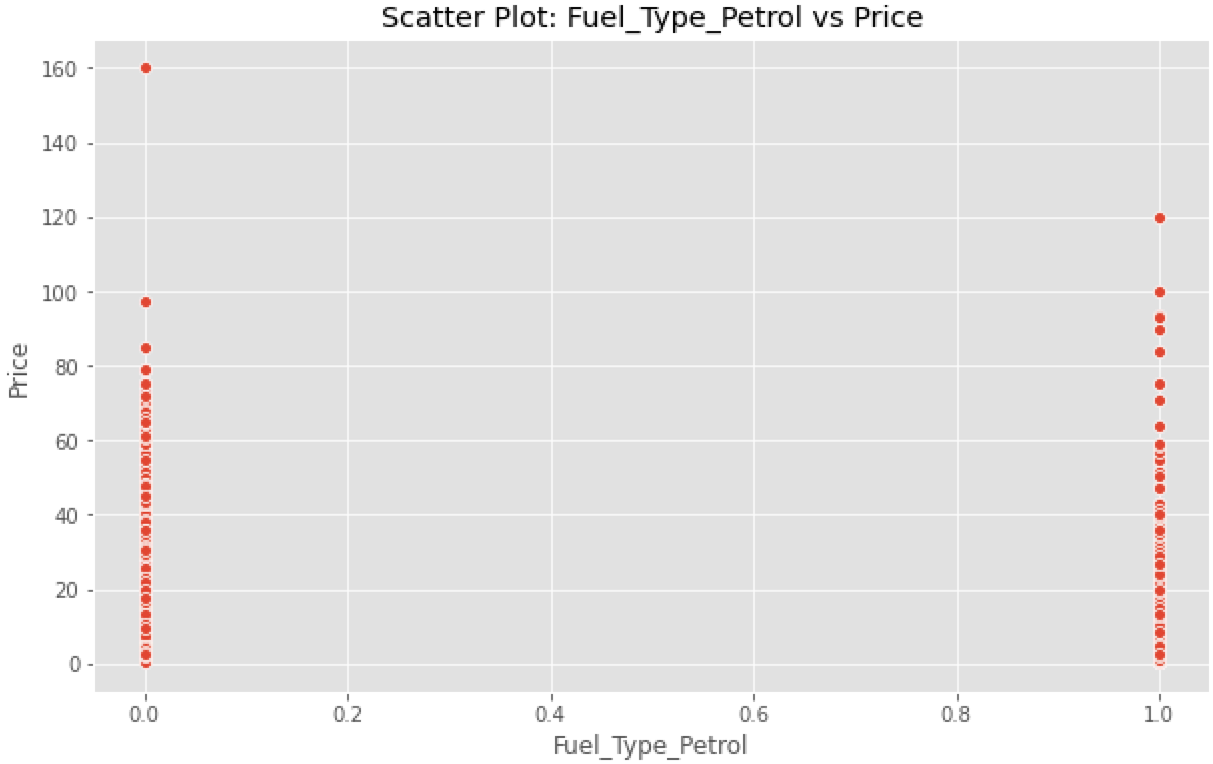
Dataset-2

After cross checking with the Correlation matrix and the feature importance graph, we can see that the features transmission manual, year and power dominate.

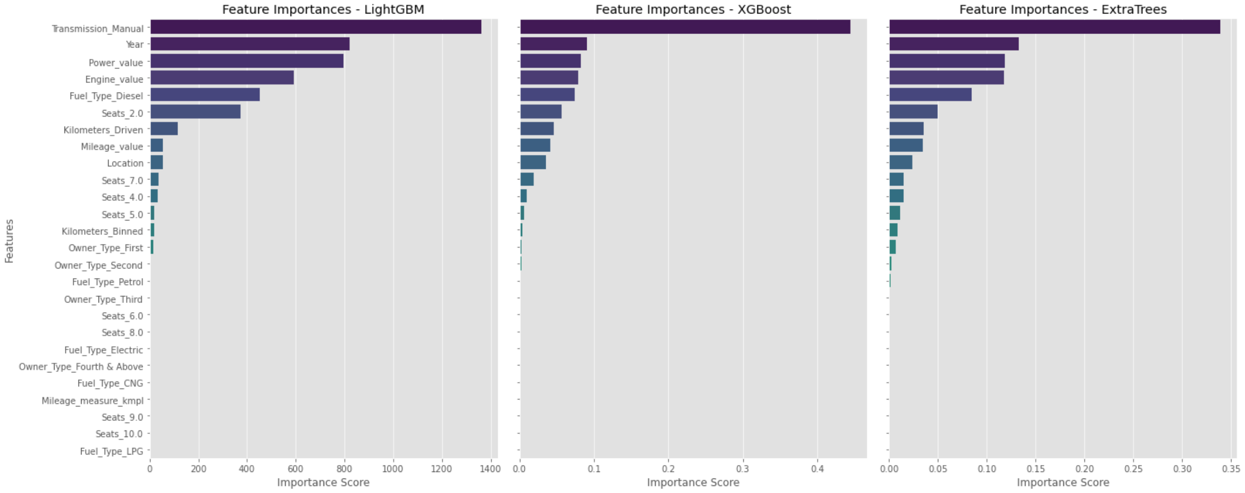




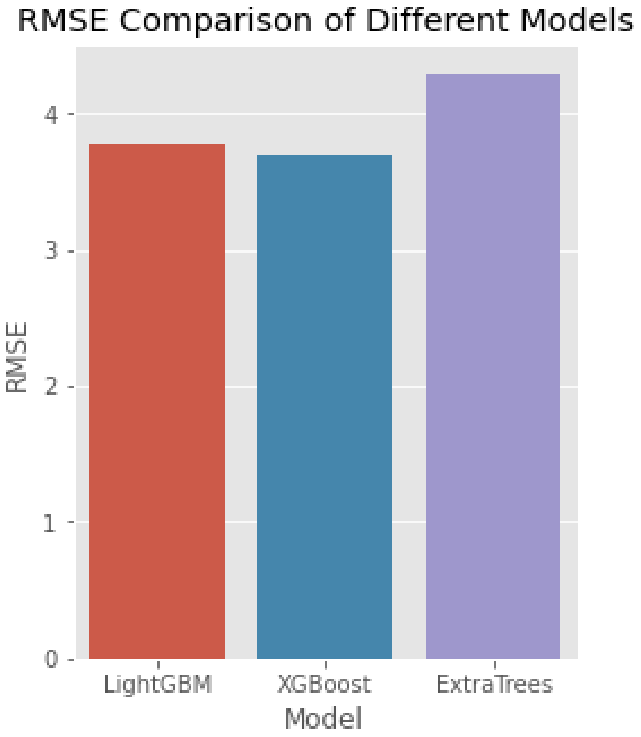




Feature importance:

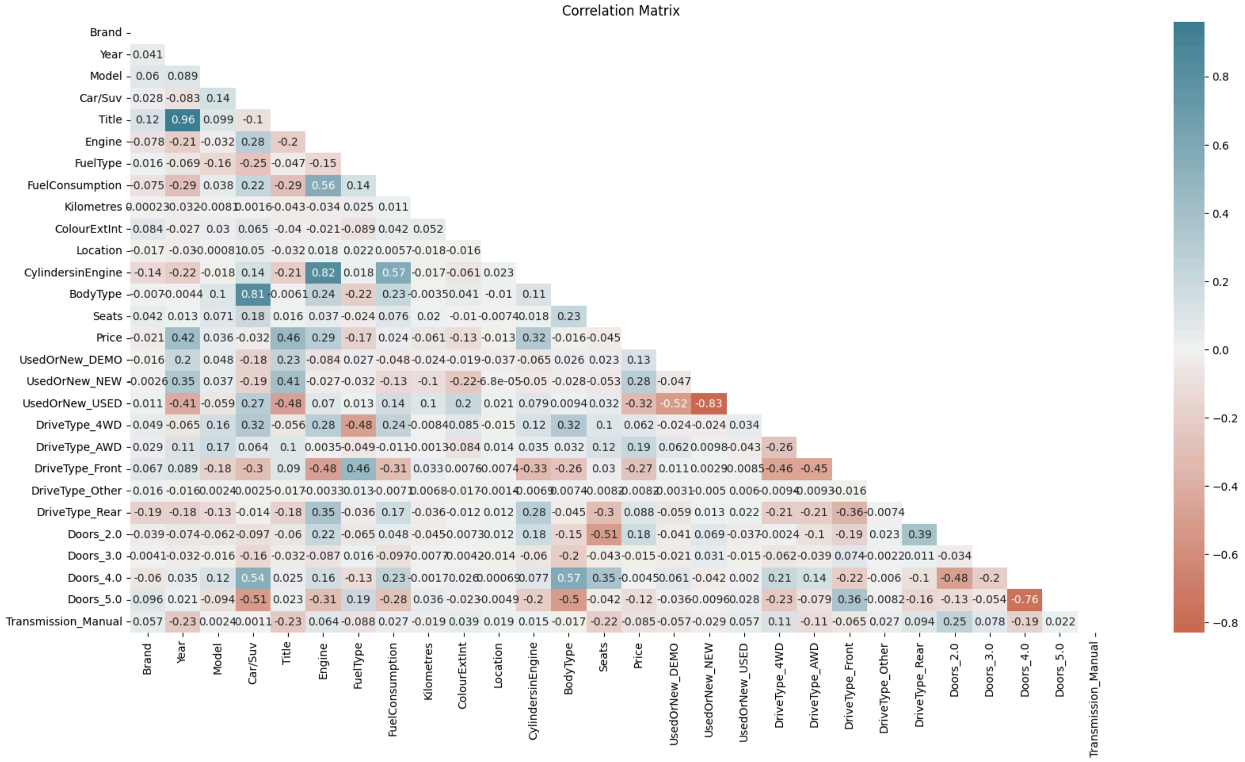


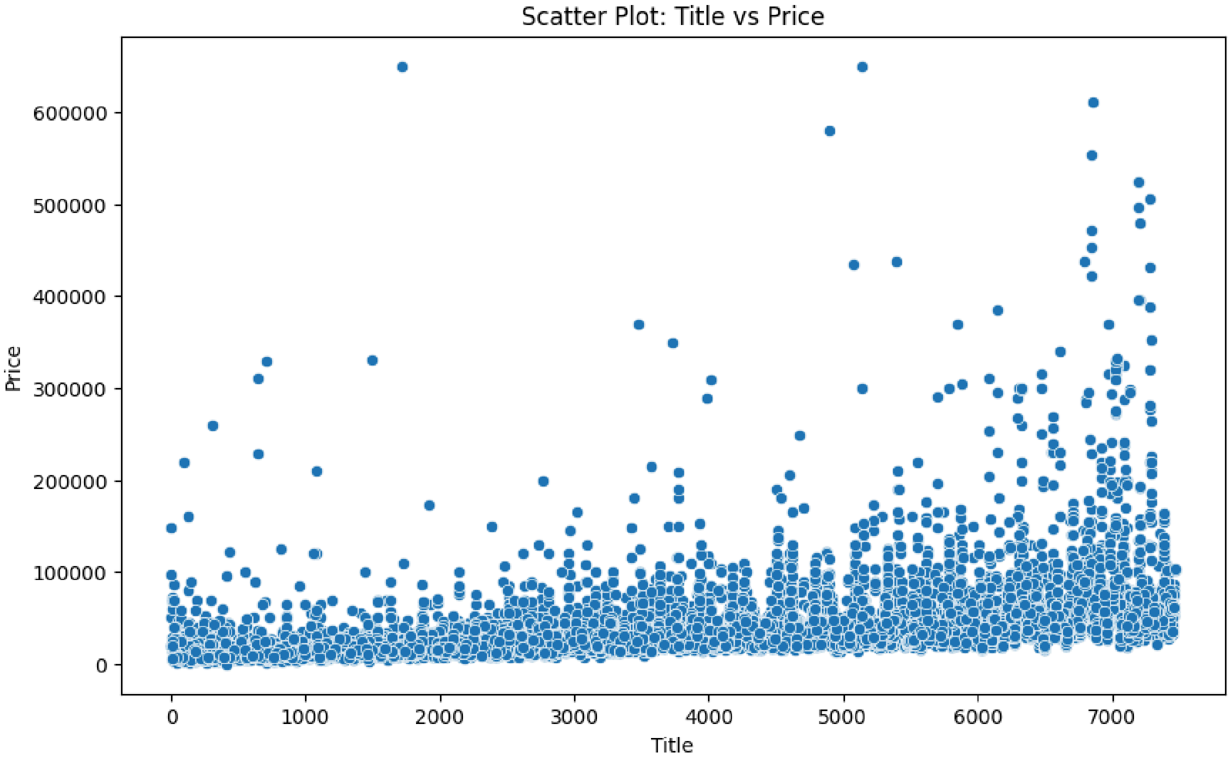
Then we have used Root Mean Square Error as an evaluation metric to compare the results of all three models. In Dataset 2, XGBoost model has performed best.

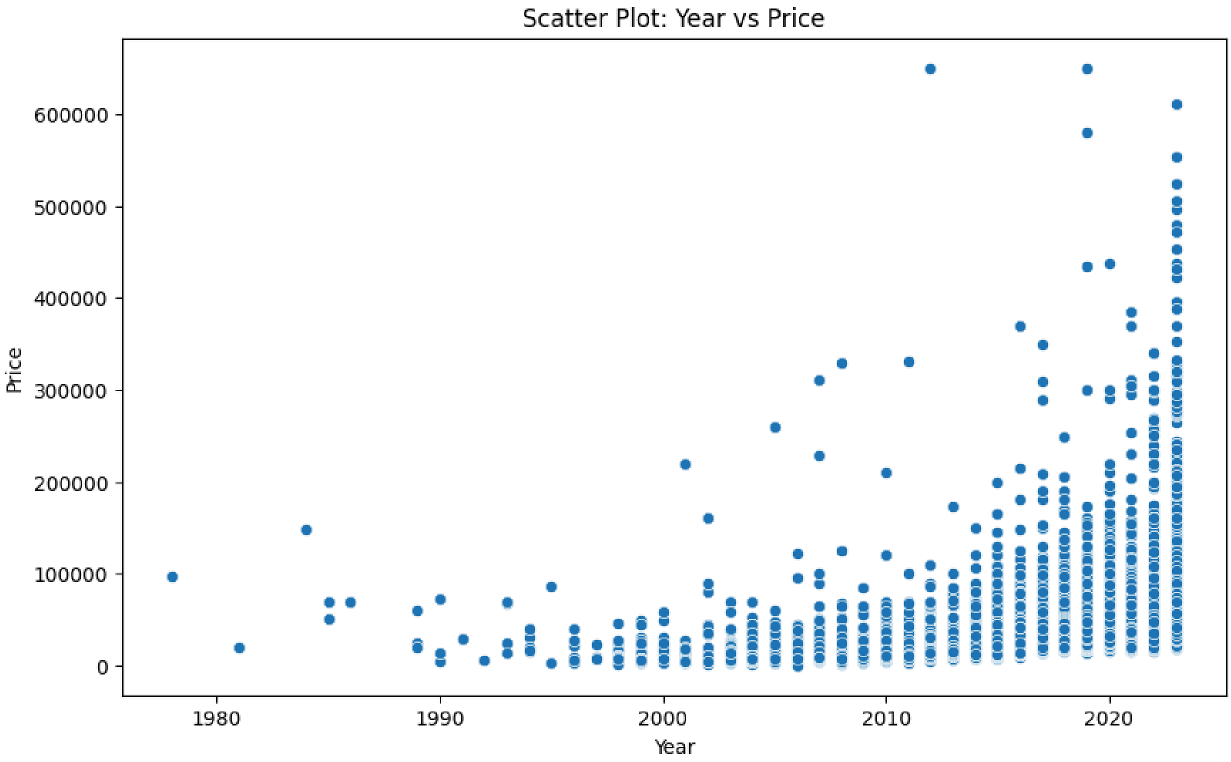


Dataset-3

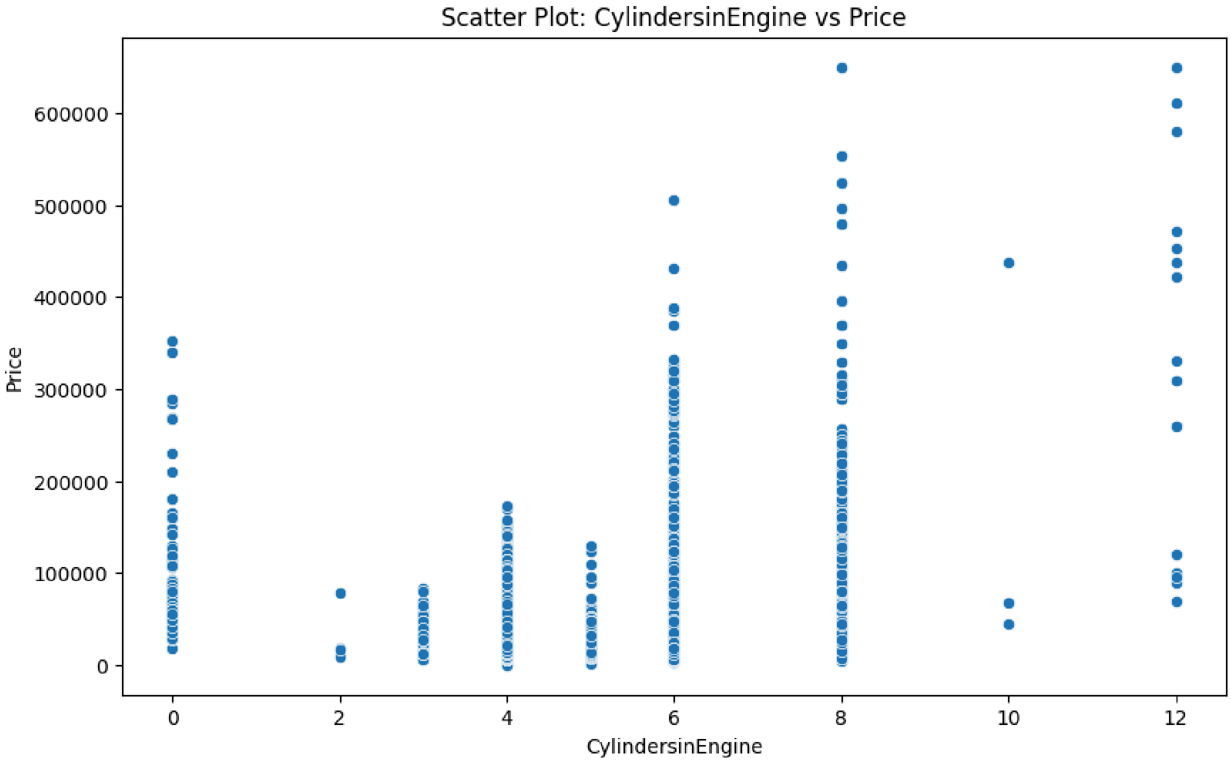
After cross checking with the Correlation matrix and the feature importance graph, we can see that the features cylinders in engine, title and engine dominate.





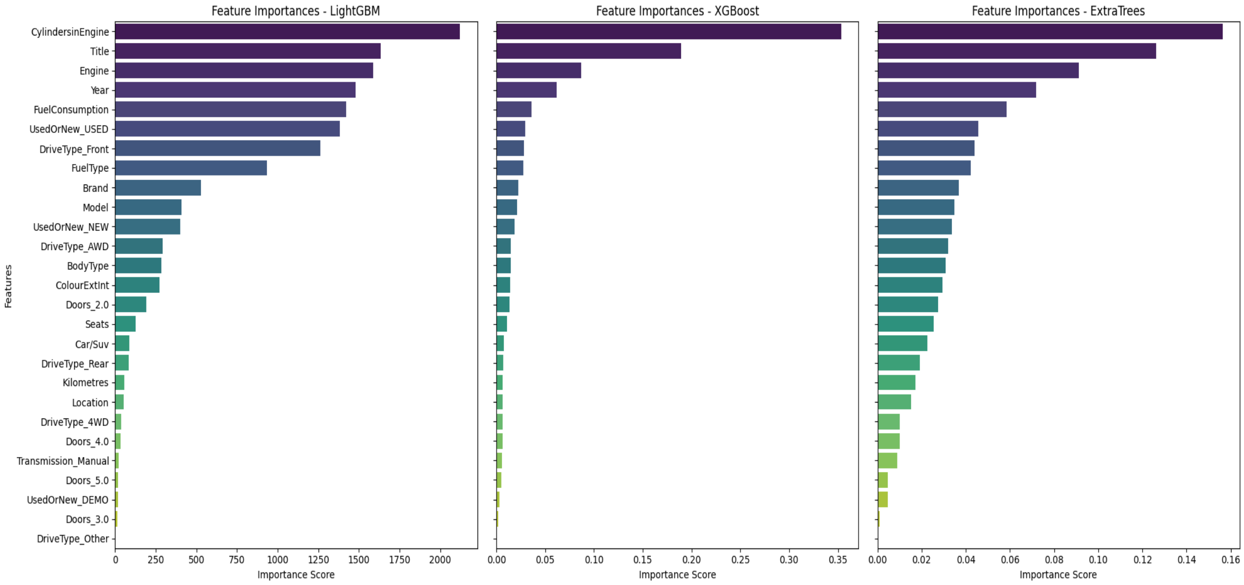


Scatter Plot: Year vs Price

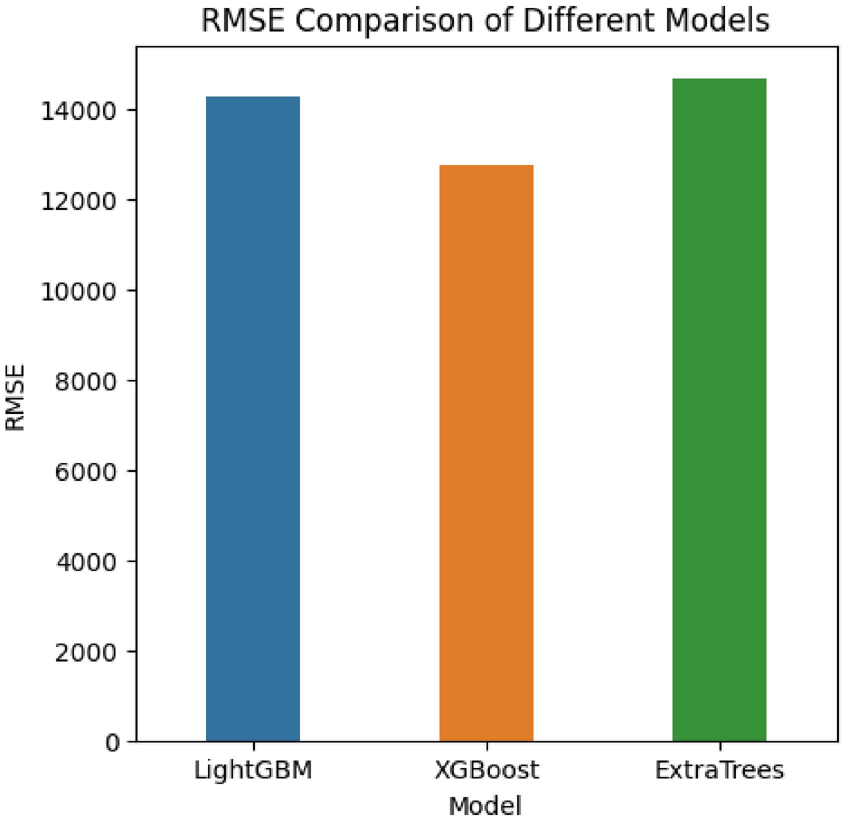


Scatter Plot: CylindersinEngine vs Price

Feature importance:



Then we have used Root Mean Square Error as an evaluation metric to compare the results of all three models. In Dataset 3, XGBoost model has performed best.



VIII. Conclusion

To sum up, the work we carried out referred to a delicate mindset towards the pre-processing of data, the training of models, hyperparameter tuning and the evaluation of the model to make sure that the forecasting power is up to the level.

The stage of pre-processing was indeed crucial. We solved the problem of missing values using K-Nearest Neighbors (KNN) mode imputation techniques. One-hot as well as label encoding were used for transforming categorical features revealing the robustness of the data model.

We utilized various models, such as LightGBM, XGBoost, and ExtraTrees, and we improved their performance by the grid search-based hyperparameter tuning. The metric of estimation, the Root Mean Squared Error (RMSE), was the one which provided the definite measure of the model accuracy that was done by capturing the magnitude of the errors in prediction.

Finally we could see that the best performer on dataset 1 was the ExtraTrees regressor, and on dataset 2 and dataset 3 it was the XGBoost model. The comprehensive approach taken in this project, from data preprocessing to detailed model evaluation through hyperparameter tuning using grid search and RMSE evaluation metric, ensured that our models were both accurate and reliable.

IX. References

[1] Y. Li, Y. Li and Y. Liu, "Research on used car price prediction based on random forest and LightGBM," 2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA), Dalian, China, 2022, pp. 539-543, doi: 10.1109/ICDSCA56264.2022.9988116. Keywords: {Industries;Pricing; Predictive models;Data collection;Prediction algorithms;Data models; Robustness;LightGBM;machine learning;used car;price;prediction;ensemble learning}

[2] G. V. Saatwik Kumar and K. Jaisharma, "Improve the Accuracy for Flight Ticket Prediction using XGBRegressor Optimizer in Comparison with Extra TreeRegressor Performance," 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), Gautam Buddha Nagar, India, 2023, pp.

2558-2562, doi: 10.1109/IC3I59117.2023.10397633. keywords:{Training; Software algorithms;Prediction algorithms;Software;Informatics;Regression tree analysis;Novel XGBRegressor Optimizer;Extra Tree;Machine Learning;Flight Ticket;Regression;Transport}

[3] H. Zhang, "Prediction of Used Car Price Based on LightGBM," 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Wuhan, China, 2022, pp. 327-332, doi: 10.1109/AEMCSE55572.2022.00073. keywords: {Machine learning algorithms;Software algorithms; Filtering algorithms;Prediction algorithms;Market research;Automobiles; Prediction of Used Car Price;LightGBM;Feature Selection;Data cleaning; Grid search}