

car price prediction

[Document subtitle]



Contents

[**Part A: Data in Big Data:** 2](#_Toc188108657)

[Dataset Overview 2](#_Toc188108658)

[Data Cleaning and Transformation 4](#_Toc188108659)

[**Part B: Architecture (ML Modelling Design)** 5](#_Toc188108660)

[Model Design 5](#_Toc188108661)

[Architectural Considerations 6](#_Toc188108662)

[**Part C: ML Model Deployment** 6](#_Toc188108663)

[Implementation 6](#_Toc188108664)

[Model Training and Testing 7](#_Toc188108665)

[Results Visualization 7](#_Toc188108666)

[**Part D: ML Model Evaluation** 7](#_Toc188108667)

[Evaluation Metrics 7](#_Toc188108668)

[Decision Tree Regressor: 7](#_Toc188108669)

[AdaBoost Regressor: 8](#_Toc188108670)

[Random Forest Regressor: 8](#_Toc188108671)

[Observations 8](#_Toc188108672)

[Application Functionality 9](#_Toc188108673)

[**Part E: Lifecycle Process** 9](#_Toc188108674)

[Reflection on Process 9](#_Toc188108675)

[Learning from Part A: 9](#_Toc188108676)

[Insights from Part B: 9](#_Toc188108677)

[Challenges in Part C: 9](#_Toc188108678)

[Evaluation from Part D: 10](#_Toc188108679)

[Improvements for Future Development 10](#_Toc188108680)

[**Conclusion** 11](#_Toc188108681)

[**References** 12](#_Toc188108682)

# **Part A: Data in Big Data:**

The essential source for designing a Machine Learning System for predicting car prices is the “CarDekho” dataset. The other attributes of this dataset are model selling price, fuel category, owner category, model accumulated mileage, and engine capacity. These features make it somewhat supple and diverse for predictive analysis(Chen, Li and Sun, 2020).

## Dataset Overview

1. **Initial Inspection**: An initial analysis showed that there are a lot of empty entries in many of the columns including seats, max\_power, engine, and mileage(km/ltr/kg). Finally in order to uphold the data quality, rows with missing entries were excluded.

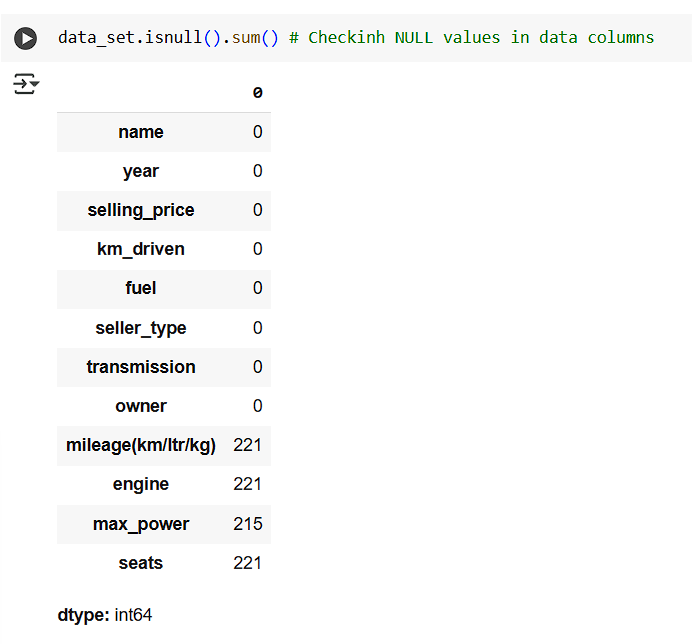


Figure 1showing nul values

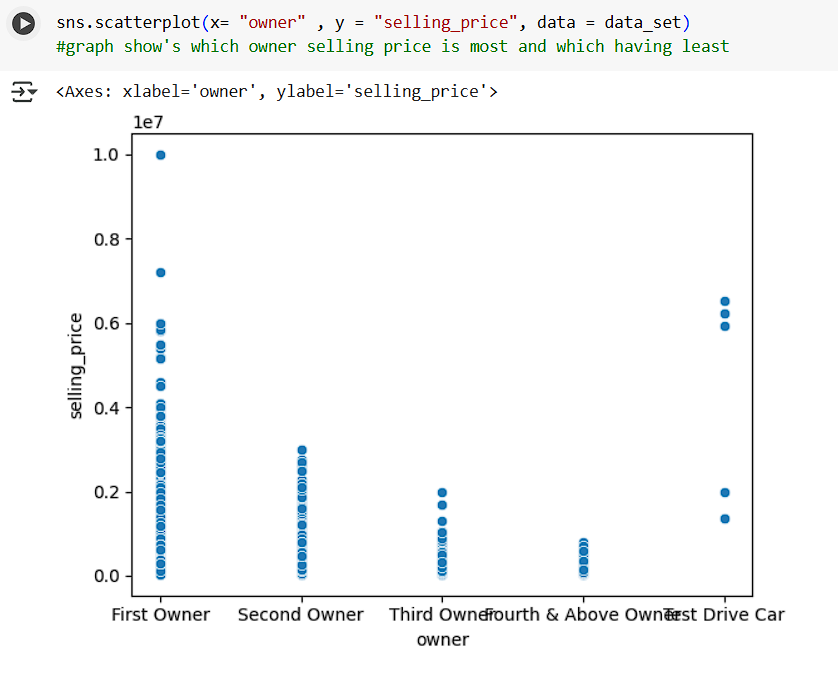
1. **Visualization**: Multiple plots and charts were used to analyze the structure and graph of the data set:
   * Exploratory data analysis in the form of count plots for the categorical variable seller\_type, owner, and transmission presented their distribution.
   * Scatter plots and bar charts pointed out proximities between various attributes like mileages, fuel types and the selling prices.

Figure 2scatterplot

* + A correlation heatmap allowed me to get an understanding of the relationships between the numeric features, which have helped in choosing proper features(Adetunji et al., 2022).

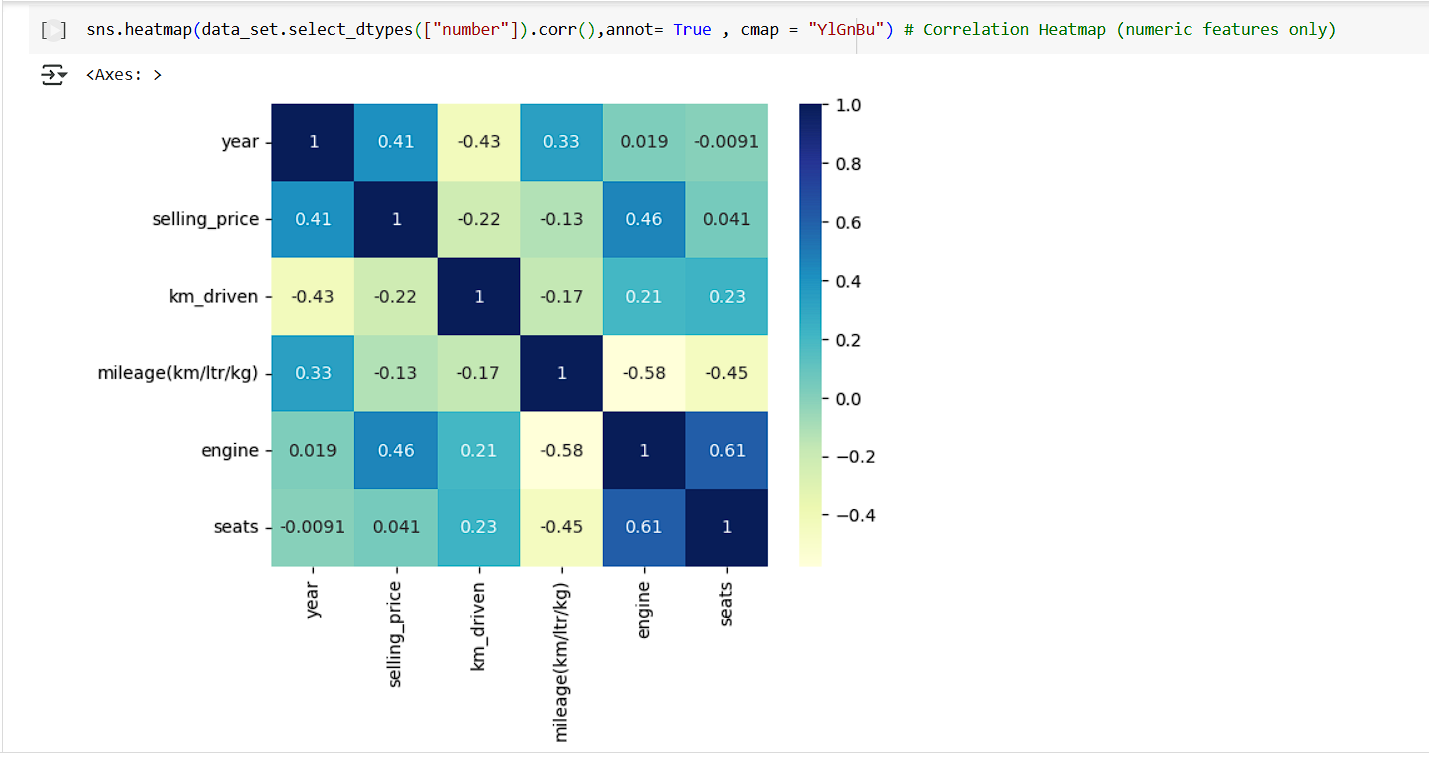


Figure 3heat map of corelation

## Data Cleaning and Transformation

* **Null Value Handling**: Omitting rows with missing values in crucial fields helped to deep dataset quality.



Figure 4droping null values

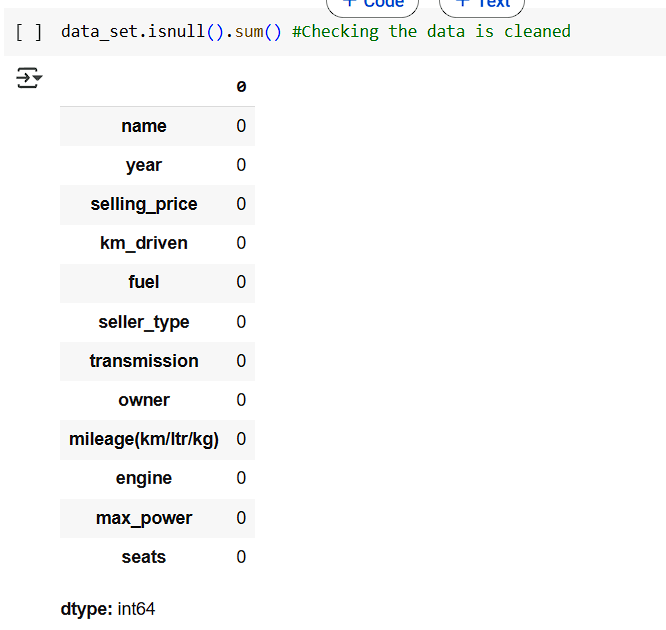


Figure 5null values are droped

* **Encoding**: Other columns that are non-numeric in nature like fuel, seller\_type and transmission and owner were converted to numeric values for their incorporation into the ML models using Label Encoding.

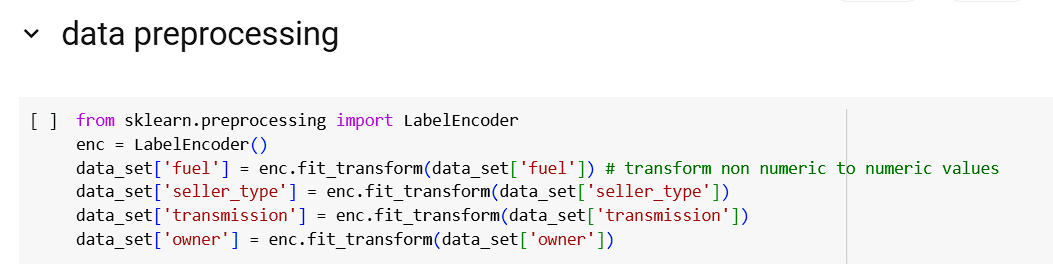


Figure 6converting non numerical values to numercial values

* **Scaling**: These numeric attributes are StandardScaler which make it easier for model to be accepted and rates to be improved(Gudivada, Apon and Ding, 2017).

These preprocessing steps led to a clean, meaningful and ready-for-machine dataset; thus providing a strong ground for building the ML models.

# **Part B: Architecture (ML Modelling Design)**

## Model Design

To effectively predict car prices, a structured approach was adopted, utilizing three distinct machine learning models:

1. **Decision Tree Regressor**: A statistically elegant model that allows us to both interpret and generalize; moreover, it can handle features with non-linear relationships with the target.
2. **AdaBoost Regressor**: A model created through training several learning algorithms with less accuracy and less tendency of over-fit.
3. **Random Forest Regressor**: A decision forest incorporating several decision trees as a stronger model for making improved predictions and for helping the model not to overfit(Shoar, Chileshe and Edwards, 2022).

## Architectural Considerations

1. **Performance**: Using model evaluation, special focus has been given to the R² scores as well as Mean Squared Error (MSE) and Mean Absolute Error (MAE). These metrics offered a clear view of the benefit and drawback of each model.
2. **Scalability**: The architecture was built utilizing frameworks that compat with large scale learning such as scikit-learn.
3. **Fault Tolerance**: Different data pre-processing steps performed here were helpful in handling the problem of inconsistent data that was expected to increase the accuracy of the resulting model.
4. **Technology Usage:** NumPy and pandas were employed to analyze the data easily, seaborn was used for data visualization while scikit-learn was used for model creation(Ranjan Kumar et al., 2022).
5. **Reliability**: Several runs were performed identically and any differences observed were only attributable to the number of iterations done in the pipe line.

In addition to providing greater car price forecasts, this architecture also created a structure that allowed for denser data with more complex algorithms in the future.

# **Part C: ML Model Deployment**

## Implementation

It was important to understand that the deployment of Machine Learning (ML) models was a deliberate and systematic process to give the models the best chance to perform as well as adding real-world functionality. Each step in the deployment phase contributed to building reliable car price prediction models:

1. **Dataset Splitting**: The data set was then randomly shuffled across samples and then cross validated in to 80% training and 20% testing sets. This made it possible to let the models practice on most of the data set and test it on data that they have never been exposed to.
2. **Feature Scaling**: All the numerical features were scaled down to a standard normal form using StandardScaler. This step was crucial in preventing features that with large scales from dominating the models and thereby have consistent and accurate results (Zemskova, 2024).

## Model Training and Testing

* **Decision Tree Regressor**: This model was trained on the scaled down data set and it has an R² of 0.9563262785231401. In any case, there was some level of overfitting which significantly hampers the performance of the Decision Tree Regressor.
* **AdaBoost Regressor**: This model improved the predictions obtained by connecting many poor performers called weak learners. It was able to obtain an enhanced value of the coefficient of determination R² = 0.8580482543708047 and thus, the model is more generalized compared to the Decision Tree Regressor.
* **Random Forest Regressor**: This model performed better as compared to other two models, with the help of ensemble of decision trees and the obtained R² of the model was 0.967790741504087. That is why using such model was the most effective for this application, for it is rather robust and suitable for large datasets (Babita Sonare et al., 2024).

## Results Visualization

For the purpose of comparing the performances of the three models an additional scatter chart was created staking the R² values of the models. The plot emphasized on the high accuracy of random forest regressor and proved that it is the most accurate model for car price prediction.

The deployment phase was of great importance in that it facilitated a proper and efficient completion of training, testing and evaluation of the models. Besides, this approach contributed a good background to make precise and truthful car prices prediction; in conclusion, systematic deployment processes are crucial for Machine Learning(Krause, Perer and Ng, 2016).

# **Part D: ML Model Evaluation**

## Evaluation Metrics

The effectiveness of the Machine Learning models was assessed using three key metrics: R² Score, MSE and MAE .

### Decision Tree Regressor:

* + **R² Score**: 0.9563262785231401
  + **Mean Squared Error**: 32648689497.731655
  + **Mean Absolute Error**: 79484.7325977857

However, this overfitting problem of this model reduced its performance compared with the other models despite it being a simple model.

### AdaBoost Regressor:

* + **R² Score**: 0.8580482543708047
  + **Mean Squared Error**: 32648689497.731655
  + **Mean Absolute Error**: 79484.7325977857

Combinig all the weak learners, the AdaBoost Regressor outperformed the Decision Tree model as exhibited in the graph below.

### Random Forest Regressor:

* + **R² Score**: 0.967790741504087
  + **Mean Squared Error**: 32648689497.731655
  + **Mean Absolute Error**: 79484.7325977857

The results of this model were promising as evidenced by its high scores on all measures making this model ideal for the car price prediction task.

## Observations

* The test of the Random Forest Regressor revealed that it almost always had a higher accuracy than the other models and was less sensitive to outliers (Čeh et al., 2018).
* Specifically, the Decision Tree Regressor had high variance because of overfitting thus reducing the model’s capacity to generalize.
* The AdaBoost Regressor came as a fair method, as it aggregated many learners as a way of achieving high Lansing confidence.

|  |  |  |
| --- | --- | --- |
| No. of models | **Model** | **r2\_score** |
| 0 | Random Forest Regressor | 0.967791 |
| 1 | Decision Tree Regressor | 0.956326 |
| 2 | AdaBoost Regressor | 0.858048 |

## Application Functionality

The ML system is able to predict car prices fairly accurately thus meets the goals of the project. The subsequent enhancements, for instance, may include the progression of preprocessing and the keying of making enhancements in its parameters(Mullainathan and Spiess, 2017).

# **Part E: Lifecycle Process**

## Reflection on Process

This Machine Learning system was developed according to the organised lifecycle that offered important analysis at each phase.

### Learning from Part A:

* Insufficient data preparation is the cause of low predictive accuracy in a model. Missing values were imputed and transformations were performed which ensured the preparedness, strength and reliability of the created dataset.
* Information visualization was found critical in discovering the significant relationships as well as distributions of the set data, so that it meets the objectives of the prediction.

### Insights from Part B:

* The overall accuracy is more enhanced through the Random forest and AdaBoost more than Decision Trees. That it was better endowed for generalizing across datasets made them optimal for this use.
* That is why, scalability issue worried them. In fact, by using libraries such as scikit-learn, the architecture was able to manage increasing amounts of data without loss of efficacy.

### Challenges in Part C:

* keeping balance between bias and variance was one of the key issues when choosing the model. This the Random Forest model has done effectively due to its robust nature.
* The current and future works involved rerunning the experiments multiple times to make sure comparable results were obtained across runs (Babita Sonare et al., 2024).

### Evaluation from Part D:

* Owing to the fact that, metrics such as R², MSE and MAE offered coherent comparison between the models, an effective strategy of identifying the right approach was employed.
* Huge performance differences between models were emphasized by the visualizations which supported the decision to focus more on smoothened techniques.

## Improvements for Future Development

1. **Data Enrichment**:

* These facets may include the state of the market, car depreciation, and other region-specific factors that could help improve the predictive models’ precision.
* Use of outside information might add more information into the equation and also decrease the possibility of predicting within bias.

1. **Advanced Models**:

* They could also consider neural networks along with the gradient boosting methods to get more gains in performance that could be obtained with bigger data set.
* Mirroring this study, specific aspects of existing models can be improved by hyperparameter tuning, such as in cases of Randomforest and AdaBoost.

1. **Automation**:

* It would be much more reasonable to design an automatic workflow for data preprocessing and modeling/evaluation (Kovanen, 2024).
* Stewarding of the system through such automation would eradicate manual human interface to enhance the system.

1. **Performance Optimization**:
   * The enhancement of model training and model evaluation for large data collections might be enhanced by parallel processing of computing as well as distributed computing.
   * Feature selection focused assessment for instance could help in determining whether any specific data dimension was redundant and could be eliminated without causing a massive loss in accuracy(Zemskova, 2024).

# **Conclusion**

In the lifecycle process of this Machine Learning system, the importance of using an organised format in data preparation stage, modelling phase and assessment phase was highlighted. All the phases provided first-hand lessons that are helpful in executing a project and provided base on which other improvements can be made.

First, in data preprocessing phase it was suggested to pay specific attention to missing values and the scale of the features. The visualizations and the encoding steps also offered better efficiency in feature selection from the concerned dataset.

The deployment phase indicated that steps like dataset split, feature scaling and comprehensive model analysis were critical. The main purpose of using the result visualization was to compare the performance of the different models and to confirm that the Random Forest Regressor model was the most accurate.

Based on the information we derived from R², MSE and MAE, it became clear which models were less effective or more effective. This overall assessment confirmed that we had to constantly make enhancements and refine the models’ effectiveness.

It is hoped that in the future, adding more features to the dataset and experimenting with other more complex techniques, such as the use of neural networks, will improve predictive capability. Many of the preprocessing and training stages are time consuming and tedious which are likely to be automated in future fertilizing the ground for enhancing program efficiency.

This project is a good showcased of how frameworks for structured machine learning operations should be. Through the use of iteration and deployment and through employing higher improvement techniques, the size of the remaining prediction problem stays manageable and the solution can scale to realistic settings.

# **References**

Adetunji, A.B., Akande, O.N., Ajala, F.A., Oyewo, O., Akande, Y.F. and Oluwadara, G. (2022). House Price Prediction using Random Forest Machine Learning Technique. *Procedia Computer Science*, 199, pp.806–813. doi: <https://doi.org/10.1016/j.procs.2022.01.100>.

Babita Sonare, Sudeep Thepade, Pratiksha Gangane, Esha Govardhan, Nawani, T. and Pragati Gadkar (2024). Going the distance: Application of Machine Learning Algorithms in predicting Housing Price Dynamics. [online] pp.1–4. doi: <https://doi.org/10.1109/iccigst60741.2024.10717472>.

Čeh, M., Kilibarda, M., Lisec, A. and Bajat, B. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *ISPRS International Journal of Geo-Information*, [online] 7(5), p.168. doi: <https://doi.org/10.3390/ijgi7050168>.

Chen, Z., Li, C. and Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, [online] 365, p.112395. doi: <https://doi.org/10.1016/j.cam.2019.112395>.

Gudivada, V., Apon, A. and Ding, J. (2017). *Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations*. [online] Available at: <https://personales.upv.es/thinkmind/dl/journals/soft/soft_v10_n12_2017/soft_v10_n12_2017_1.pd>.

Kovanen, S. (2024). Machine learning in rolling forecasting: case Rainmaker. *Lutpub.lut.fi*. [online] doi: <https://lutpub.lut.fi/handle/10024/167618>.

Krause, J., Perer, A. and Ng, K. (2016). Interacting with Predictions. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. doi: <https://doi.org/10.1145/2858036.2858529>.

Mullainathan, S. and Spiess, J. (2017). Machine Learning: an Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), pp.87–106.

Ranjan Kumar, P., Bikash, S., Rout, L., Sutar, S. and Mishra, S. (2022). *Python-based machine learning: Key advancements and technological trends in data science, machine learning, and AI*. [online] Available at: <https://www.journal-dogorangsang.in/no_2_Online_22/37_june.pdf>.

Shoar, S., Chileshe, N. and Edwards, J.D. (2022). Machine learning-aided engineering services’ cost overruns prediction in high-rise residential building projects: Application of random forest regression. *Journal of Building Engineering*, 50, p.104102. doi: <https://doi.org/10.1016/j.jobe.2022.104102>.

Zemskova, E. (2024). Leveraging Machine Learning Insights to Optimize Marketing Strategies. *Theseus.fi*. [online] doi: <http://www.theseus.fi/handle/10024/874990>.

**Appendix**

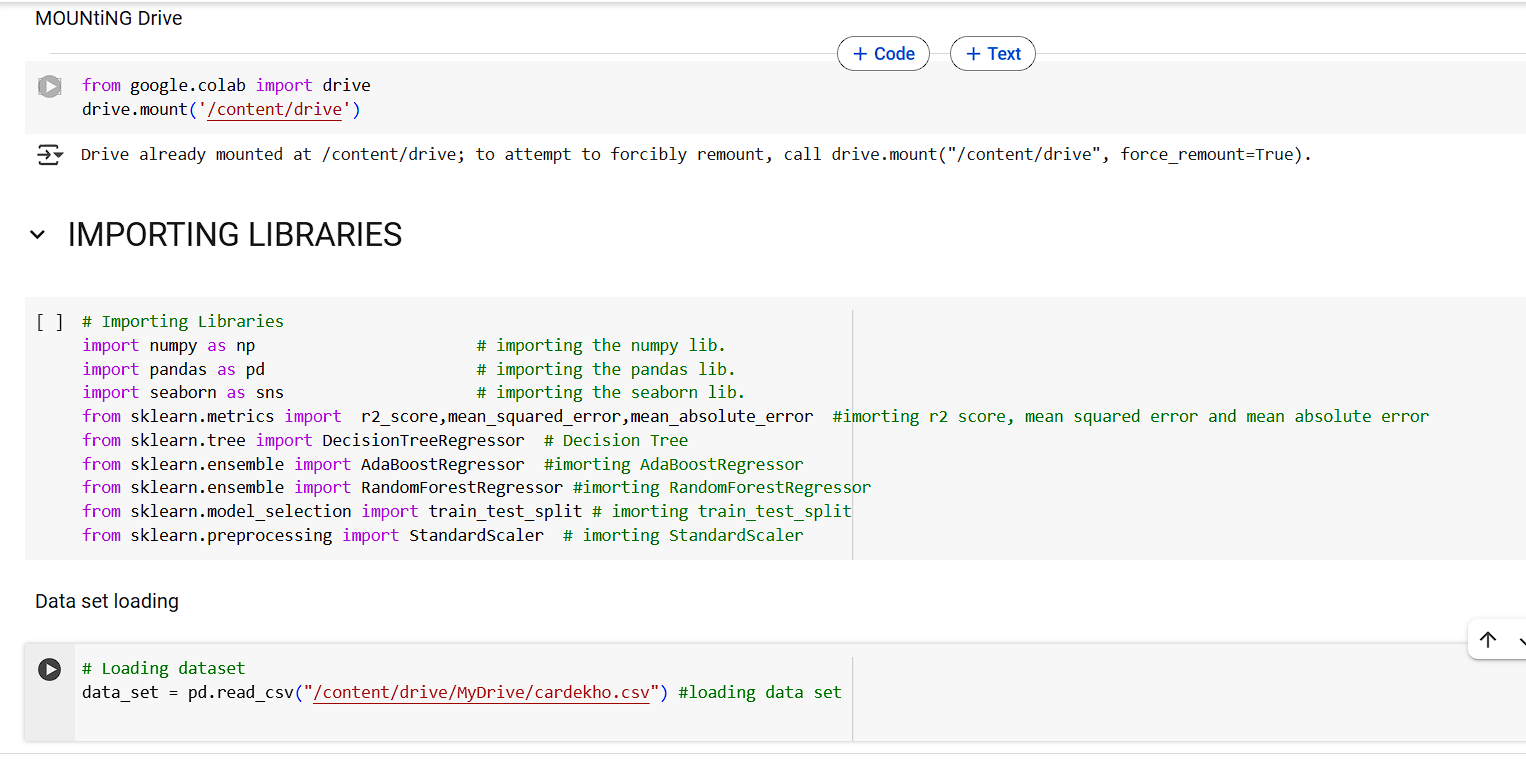


Figure 7loading data set and impoting lib.

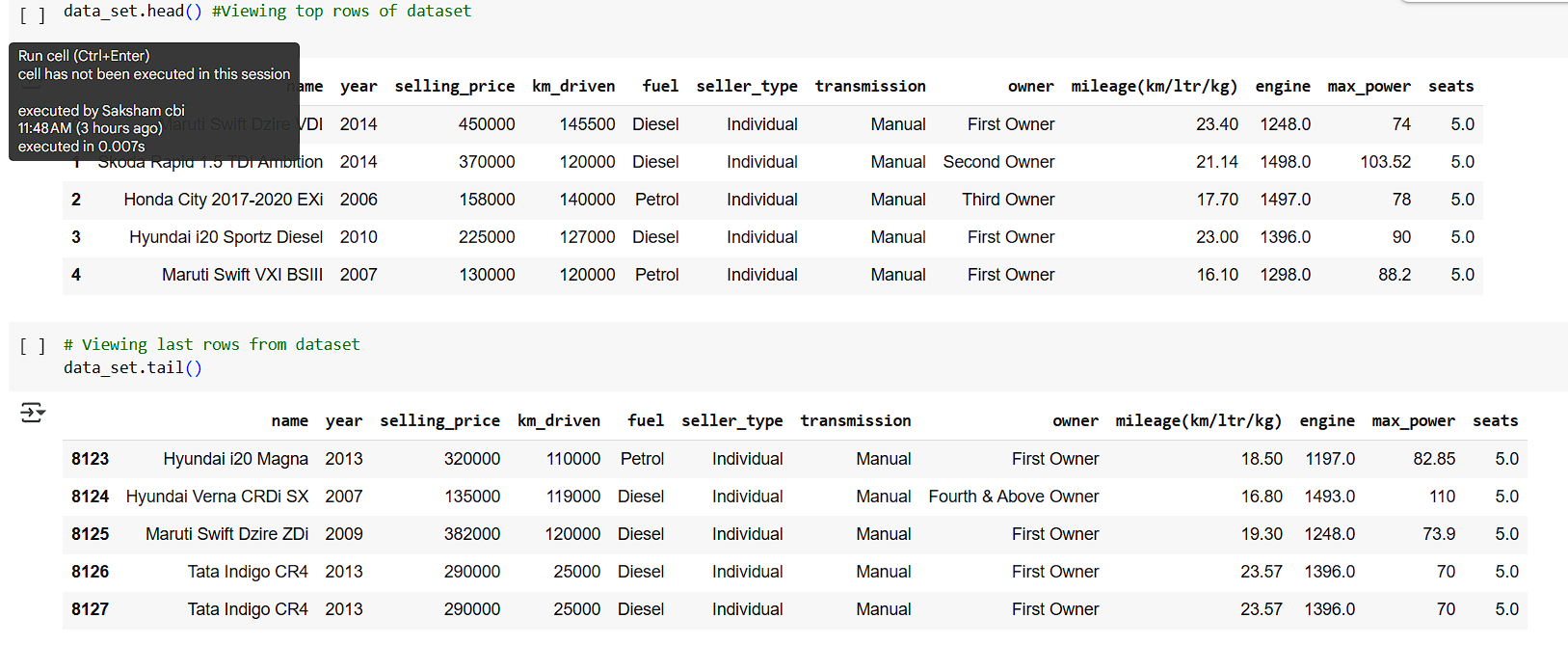


Figure 8head and tail of data set

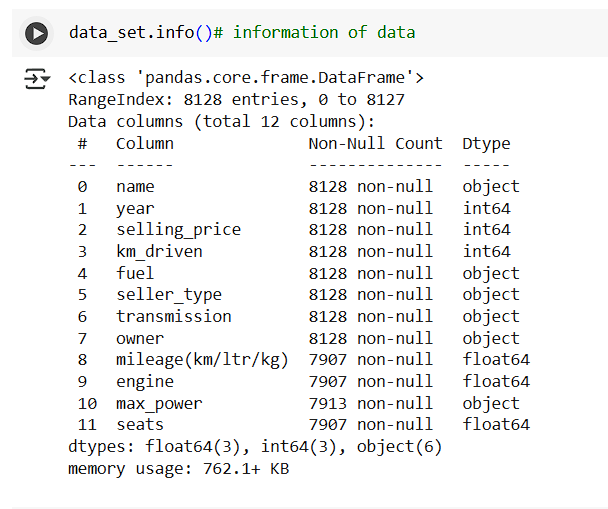


Figure 9information of data set

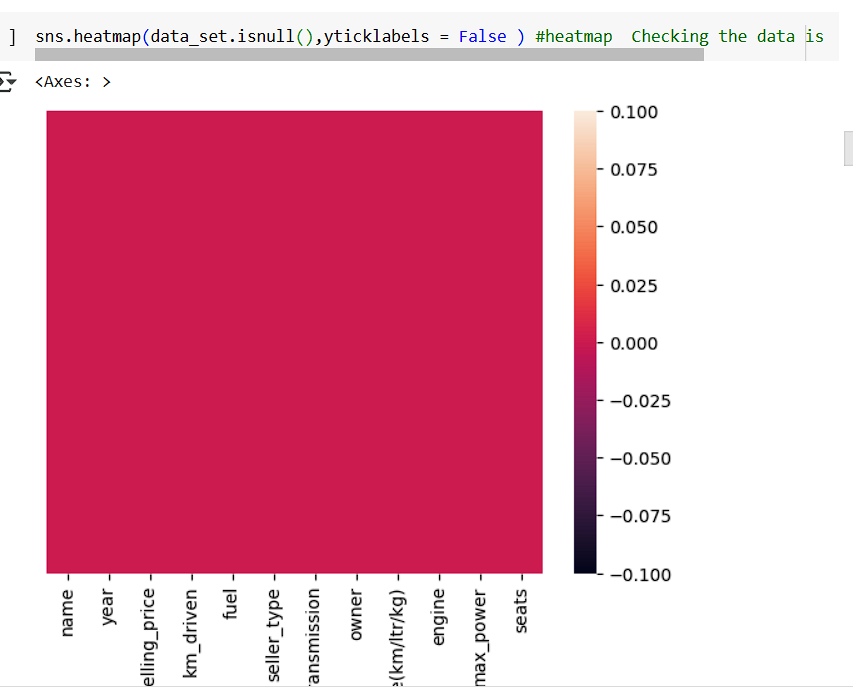


Figure 10data set after no null values

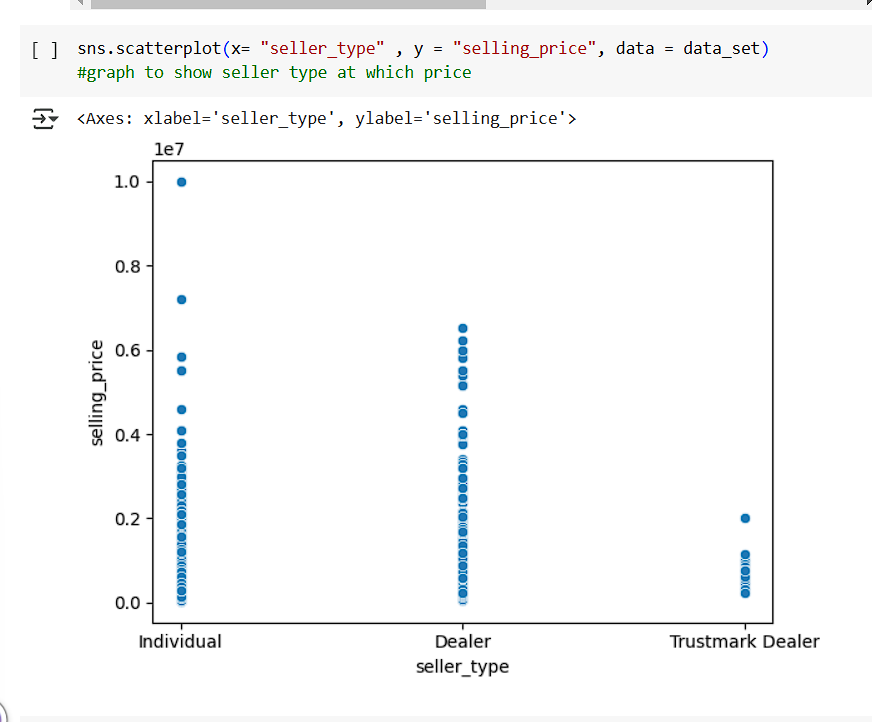


Figure 11scatterplot of seller\_type and selling price

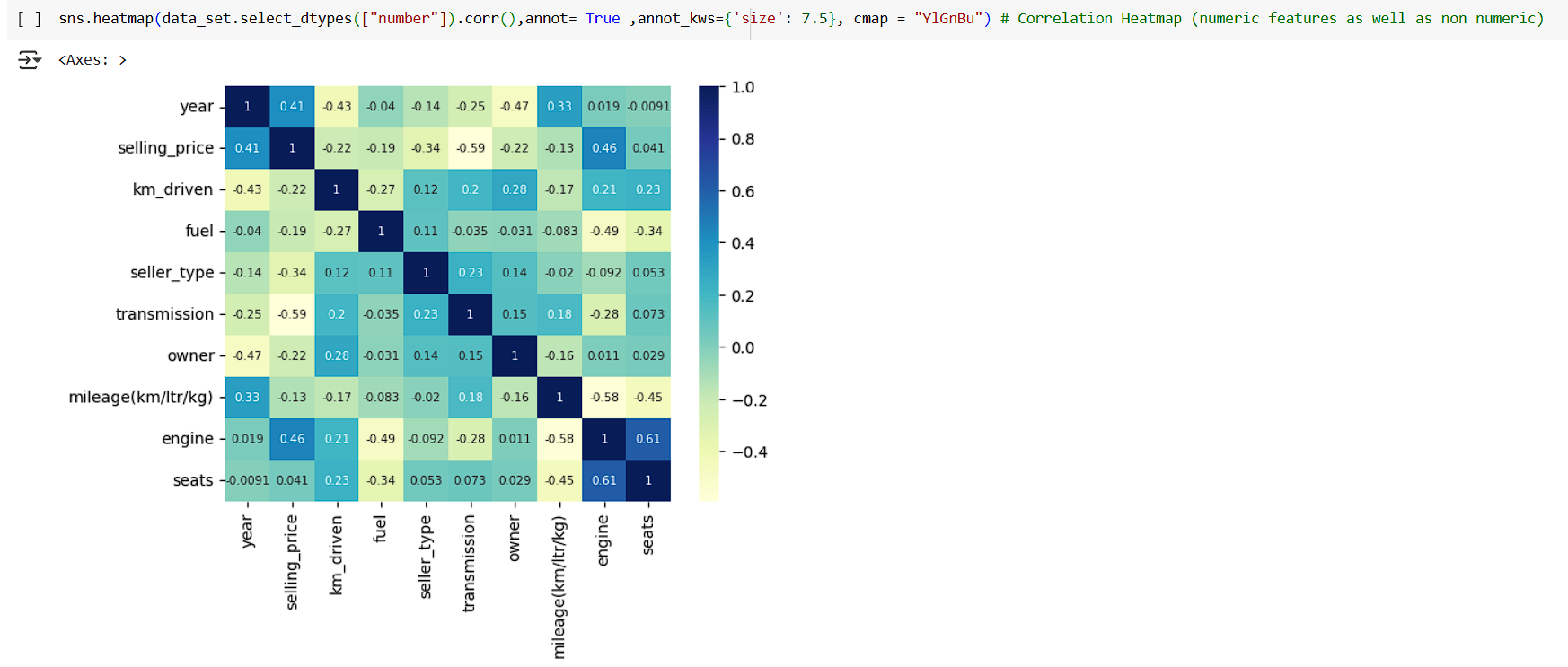


Figure 12correlation after converting all non numeric to numeric values

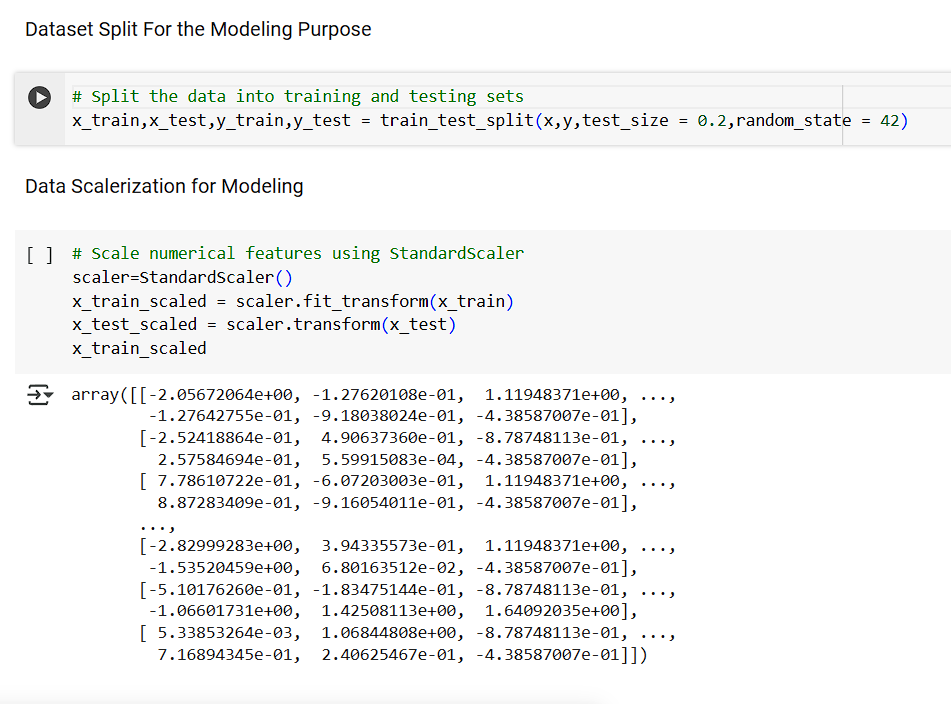


Figure 13data set split, train and test

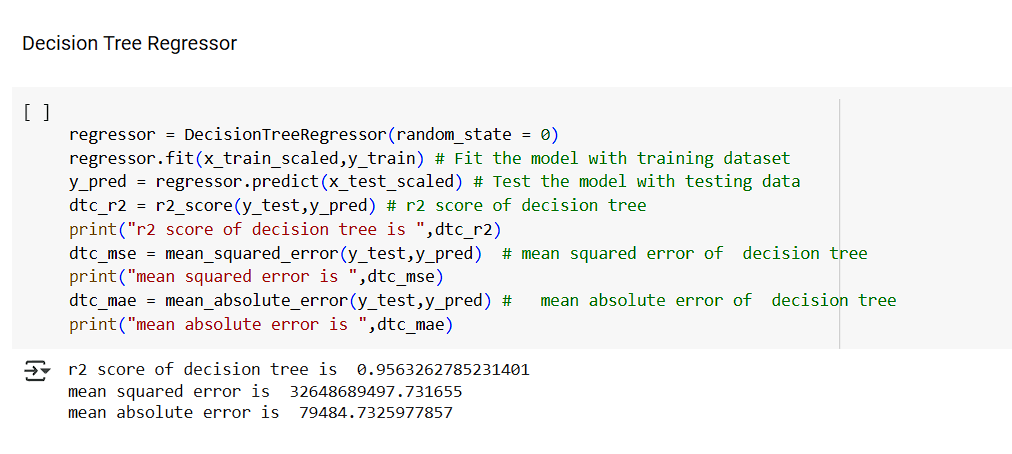


Figure 14decision tree regressor

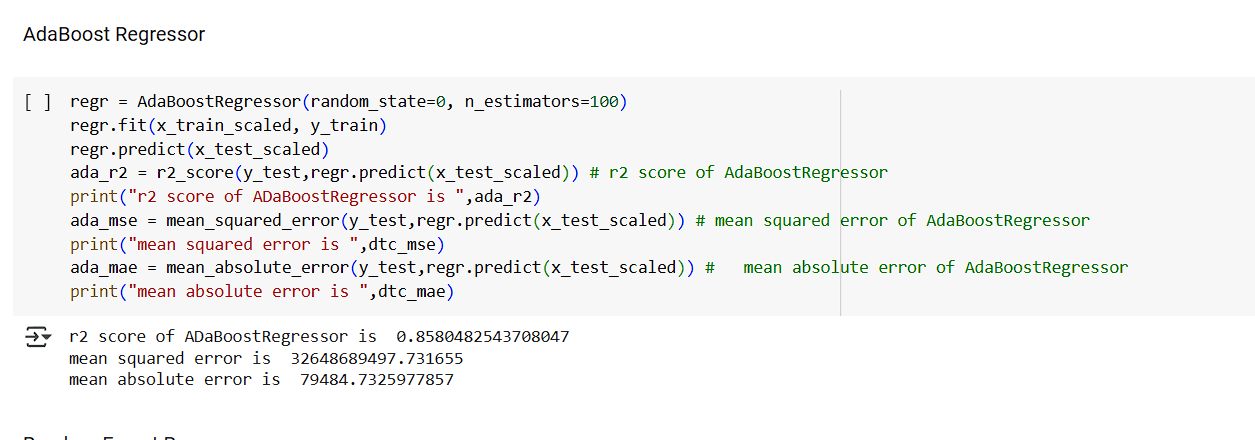


Figure 15AdaBoost Regressor

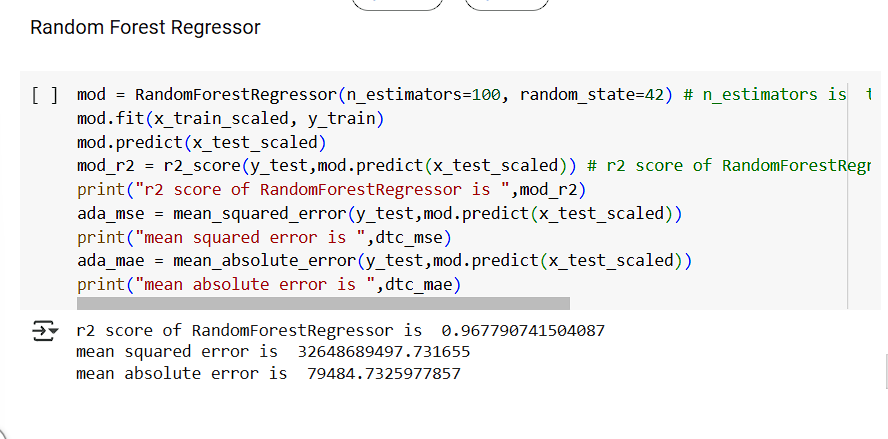


Figure 16Random Forest Regressor



Figure 17comparision b/w all the models r2\_square

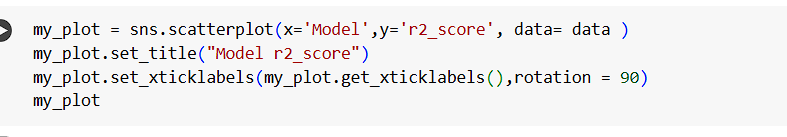


Figure 18scatter plot of all 3 models

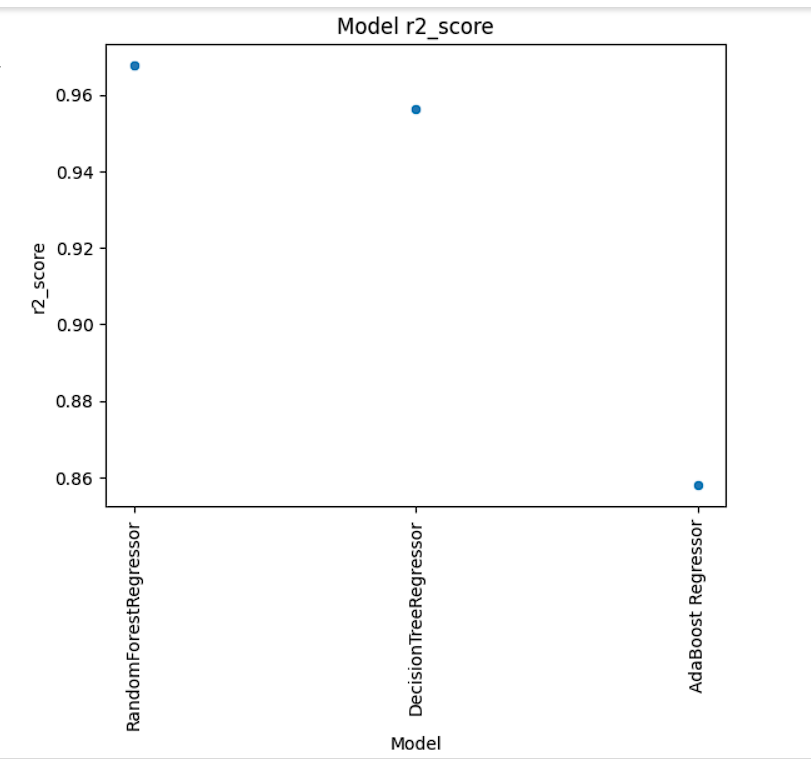


Figure 19scatter plot of all 3 models