

**Price Prediction of Used Cars Using Machine Learning**

**By**

Taufiqul Alam

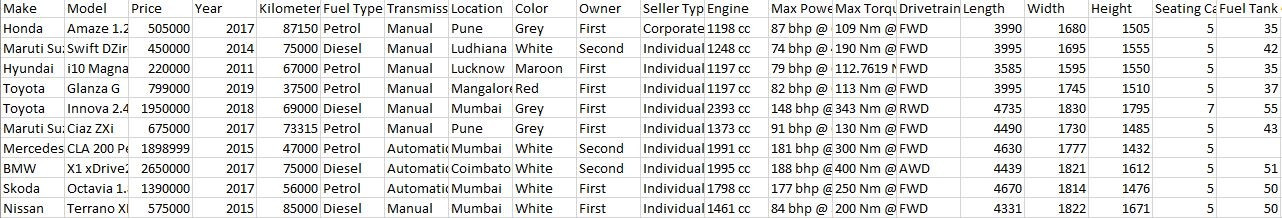
Md. Tamjid Hosain

Summiya Sunjida Kashpia

# 01

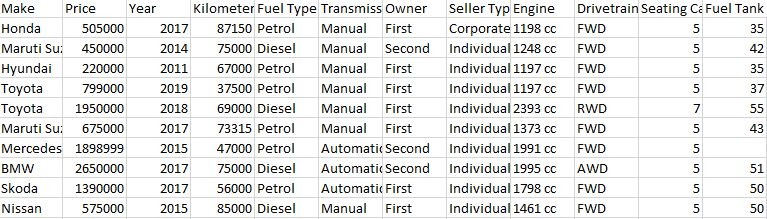
## Description of Dataset

* The dataset that we used in our model is taken from Kaggle.
* In our dataset there were 2059 rows and 20 attributes.
* The attributes of the cars that were given in our dataset were:

Make, Model, Price, Year, Kilometer Driven, Fuel type, Transmission, Location, Color, Number of Owners, Seller Type, Engine Capacity, Max Power, Max Torque, Length, Width, Height, Drivetrain, Seating Capacity and Fuel Tank Capacity.

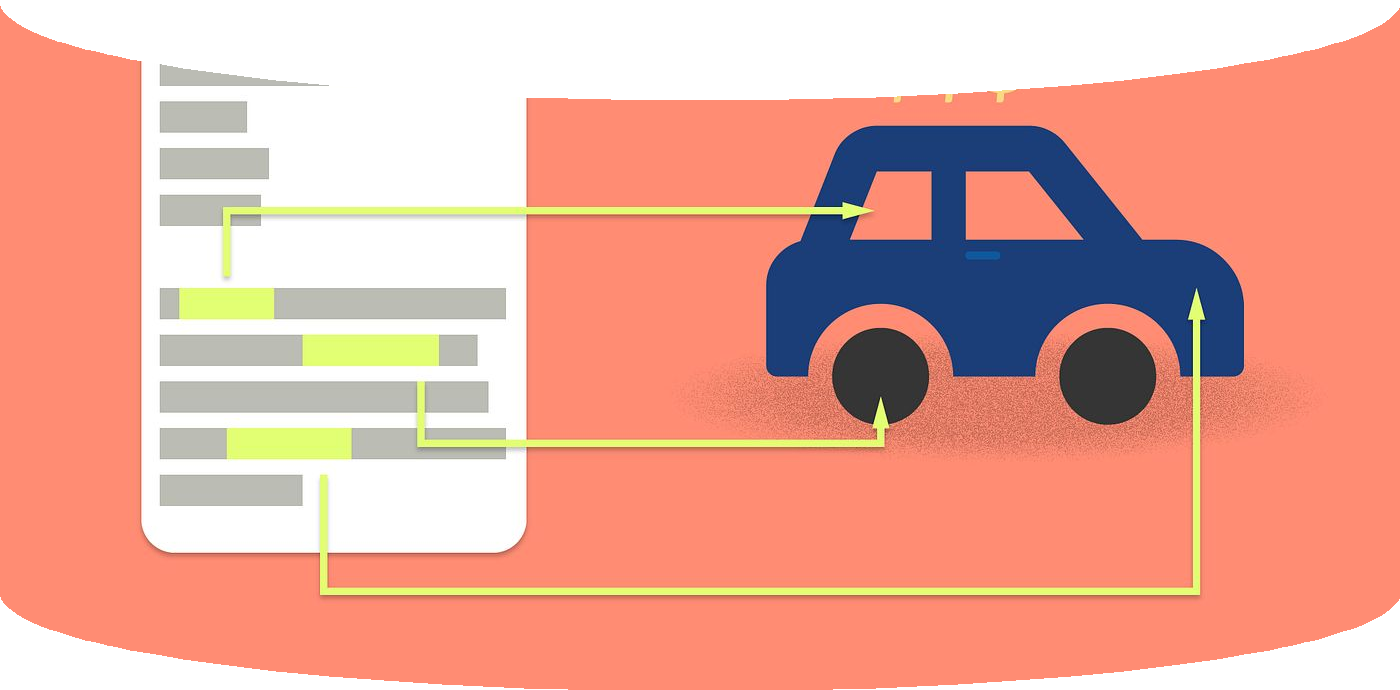
* We used recursive feature elimination to discard all the irrelevant features. After using this algorithm, the remaining 12 attributes of the car were:

Make, Price, Year, Kilometer Driven, Fuel type, Transmission, Number of Owners, Seller Type, Engine Capacity, Drivetrain, Seating Capacity, Fuel Tank Capacity, and Price.



### Preprocessing the Dataset

* In the dataset we had 2059 rows. But out of those rows there were 185 rows with null values. So the rows with null values were removed and at the end we had 1874 rows in our dataset.
* All the numerical values in the engine attribute ended with the string ‘cc’. To treat it as numerical data, the string ‘cc’ had to be removed and after removing it the values were converted to integers. So, in the end, the values in the engine attribute were all numerical data.
* The pricing values for the cars listed in the dataset were excessively high (from thousands to crores). Due to this large range of target values, a scaling method called MinMaxScaler was used to bring the values to a comparable range which is between 0 and 1.
* The categorical variables in the dataset were Make, Fuel Type, Transmission, Owner, Seller Type, Drivetrain, Seating Capacity, and Fuel Tank Capacity. All the categorical values were encoded into a binary vector representation using one-hot encoding.

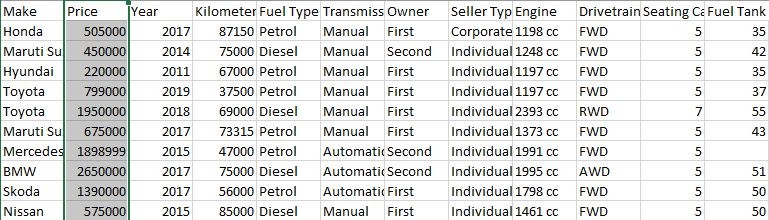


## 02 Target and Motivation

### Target

Based on the given attributes the target variable that our model aims to predict is the selling price of the cars.

The features of the car that will be utilized to predict the selling price of the cars are: Make, Year, Kilometer Driven, Fuel type, Transmission, Number of Owners, Seller Type, Engine Capacity, Drivetrain, Seating Capacity and Fuel Tank Capacity.



### Motivation

The price of brand-new cars has signiﬁcantly increased in recent years as a result of the global economic crisis. As a result, used car purchases are becoming more and more popular. However, ﬁguring out a used car's fair market value can be difﬁcult because it depends on a number of things, including the overall condition brand, model, year, mileage, transmission, etc. Uninformed buyers often fall prey to inﬂated price and end up paying a price which is not worth the car’s value.

To address this issue and help buyers and sellers make informed decisions, we aim to develop a system that accurately predicts the prices of used cars based on their different attributes. This will empower buyers to assess the value of a used car and negotiate a fair price, ensuring they get the best deal possible. Likewise, sellers can set realistic prices based on the provided information, fostering transparency and fairness in the used car market.

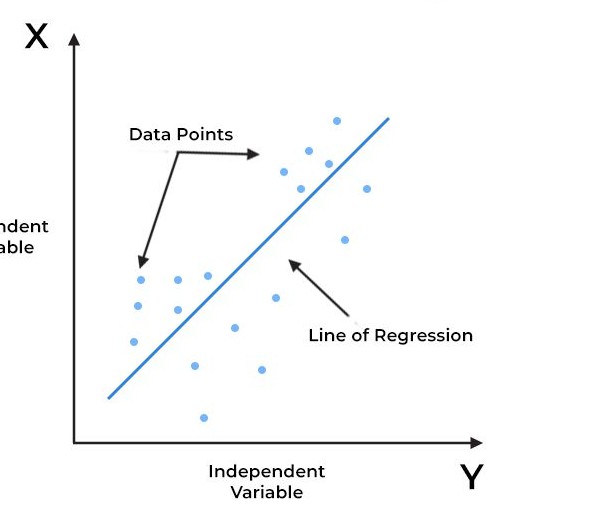
# 03

## Designing the Model

**Regression Algorithms Used**

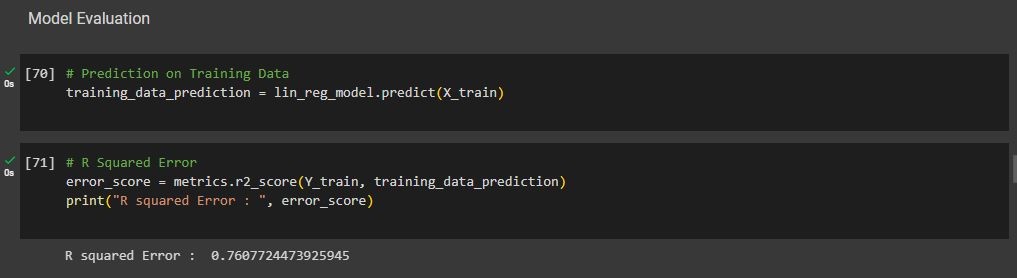
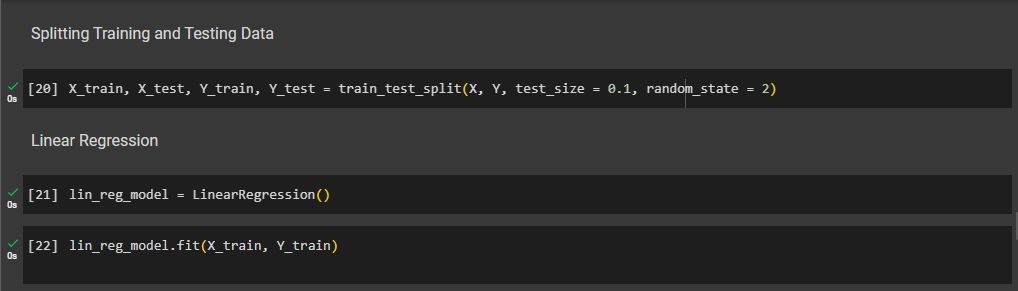


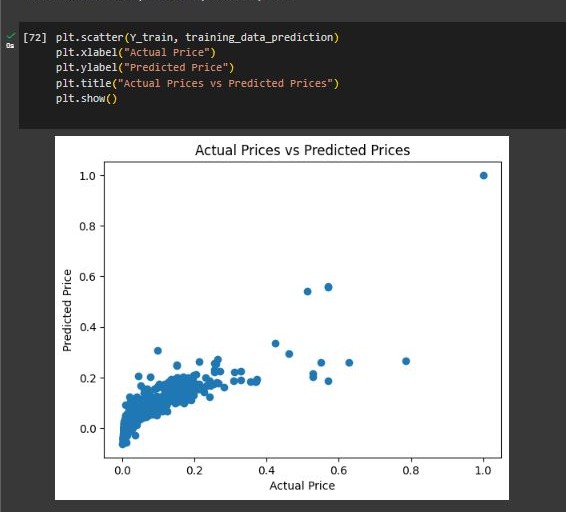
**Linear Regression**



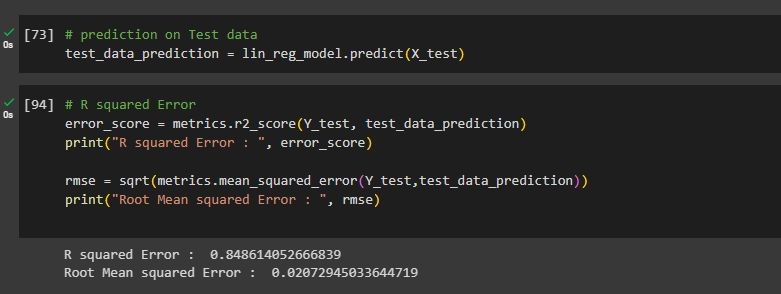
**Random Forest Regression**

**(Training Data)**



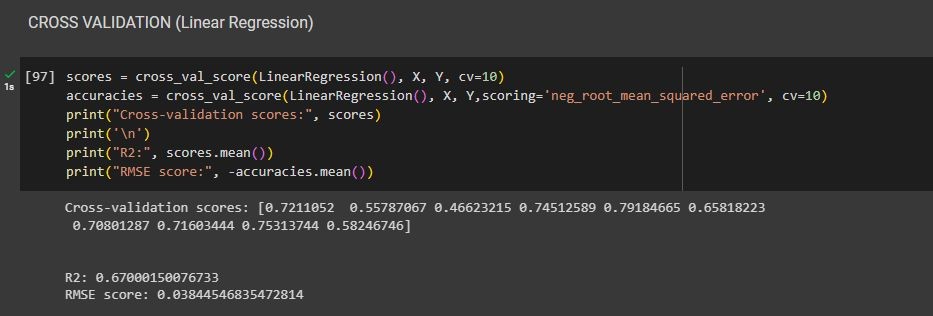
**(Scatter Plot for Training Data)**

**(Testing Data)**

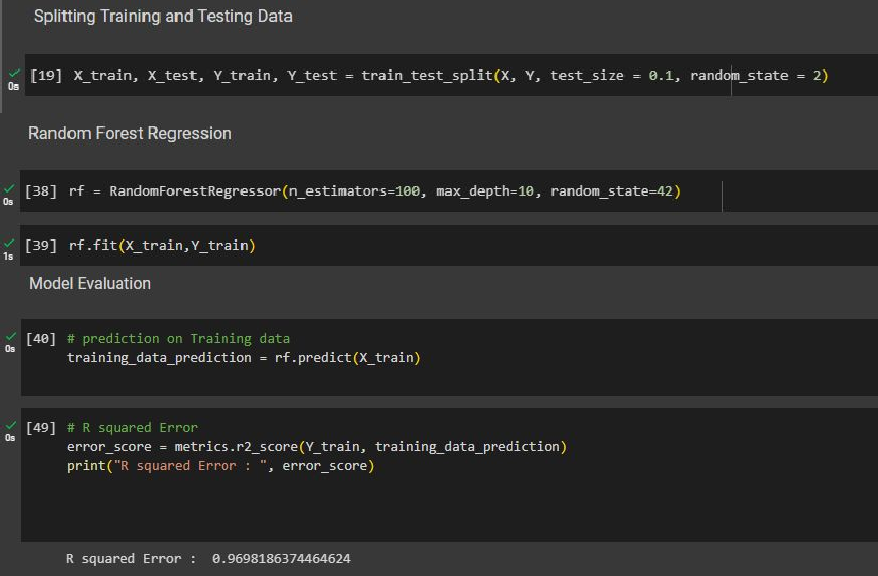


**(Scatter Plot for Testing Data)**

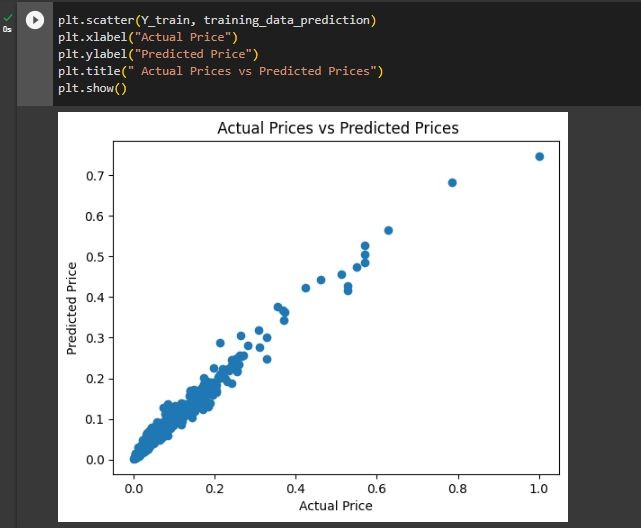
**(K-Fold Cross Validation)**



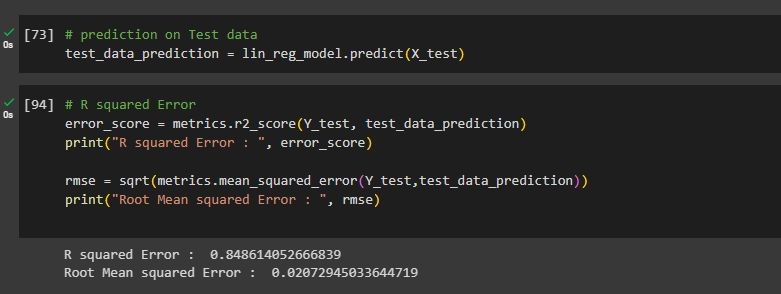
**(Training Data)**

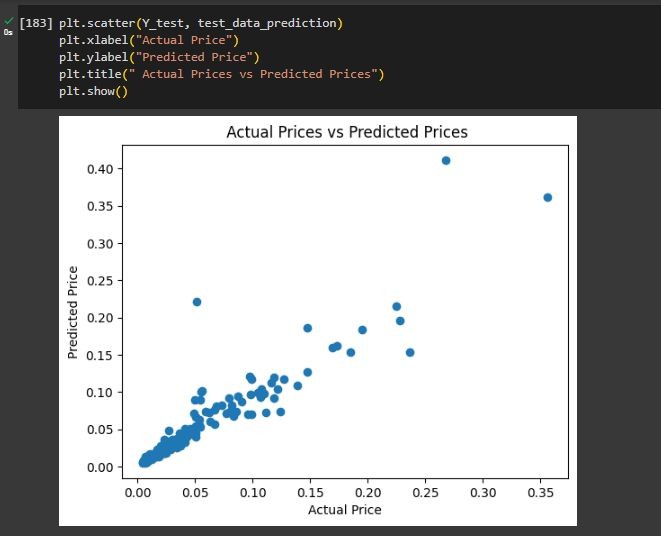


**(Scatter Plot for Training Data)**

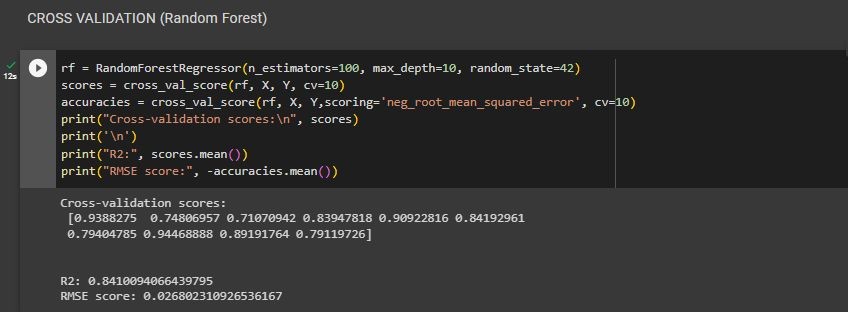


**(Testing Data)**



**(Scatter Plot for Testing Data)**

**(K-Fold Cross Validation)**

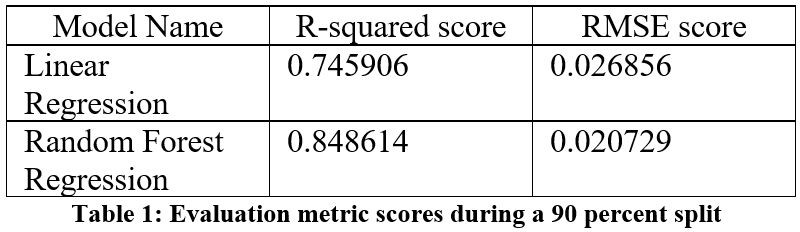


## 04 Result and Comments

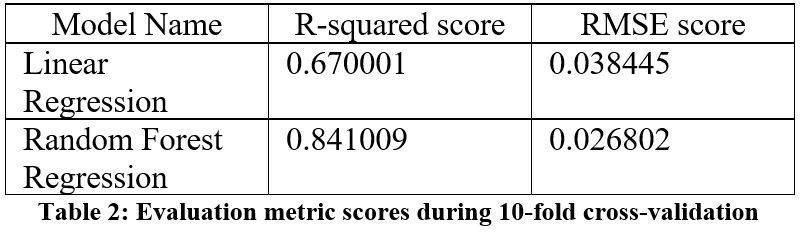
We used two evaluation metrics: **R-Squared score** and **RMSE value**.

1. R-squared error measures the proportion of variance in the dependent variable that can be explained by the independent variables in regression model. The R-squared score ranges from 0 to 1. Higher values suggest a better ﬁt between the model and the data.
2. RMSE value is a measure of the average distance between the predicted and actual values of the target variable. Lower rmse value indicates better predictive accuracy, as they indicate smaller average prediction errors.

In application of both the regression algorithms, we got the following results of our evaluation metrics. The following table represents the R-squared score and RMSE value during a 90% split.



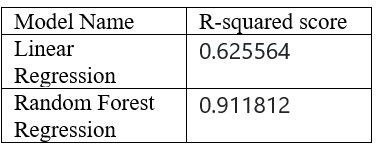
The following table represents the R-squared score and RMSE value during 10-fold cross validation.



**From Research papers**

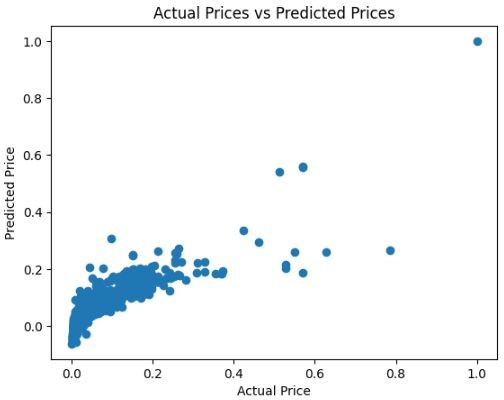
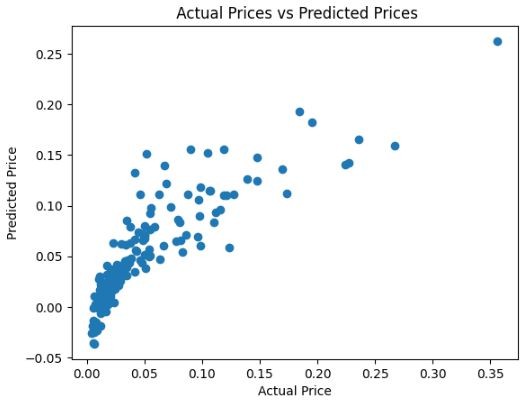
These are the R-squared values we obtained from research papers of car price prediction using linear regression and random forest regression. We can see that the results we achieved in our model are similar to the results they obtained in the research paper.

The following table represents the R-squared values of both models from [research paper 1](https://www.ijraset.com/research-paper/car-price-prediction-using-machine-learning-algorithms).

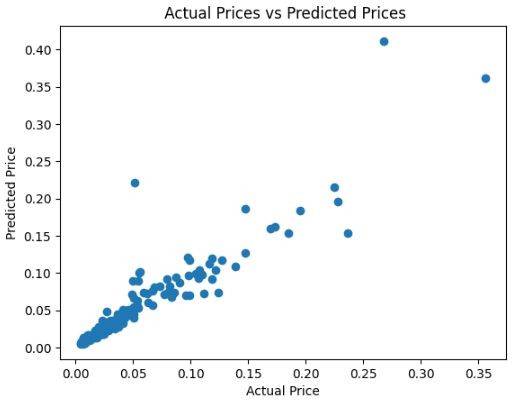


The following table represents the R-squared values of both models from [research paper 2](https://www.irjmets.com/uploadedfiles/paper/volume3/issue_3_march_2021/6681/1628083284.pdf).



**Training Data Testing Data**

**Training Data Testing Data**

### Comments

Based on our evaluation metric we can see that both the models performed with great accuracy in predicting car prices. But if we compare both the linear regression model and the random forest regression model, we can clearly observe that the random forest model performed much better than the linear regression model.

The R-squared score in forest regression model is higher than that of the linear regression model in both types of evaluation. In addition, the RMSE score in the random forest regression model is less than the linear regression model in both types of evaluation. This indicates that the random forest regression model performed with better accuracy than the linear regression model.

We also see that the results we achieved in our model are similar to the results they obtained in the research paper. So we obtained desirable results from our model.

In a car price prediction problem, the goal is to accurately predict car prices with minimal error, consistently providing predictions that closely match the actual prices. It would perform well across different car brands, models, and features, without being overly sensitive or biased. It would generalize effectively to new, unseen car instances and would identify and explain the most signiﬁcant features and their impact on the pricing of cars.

**THANK YOU**