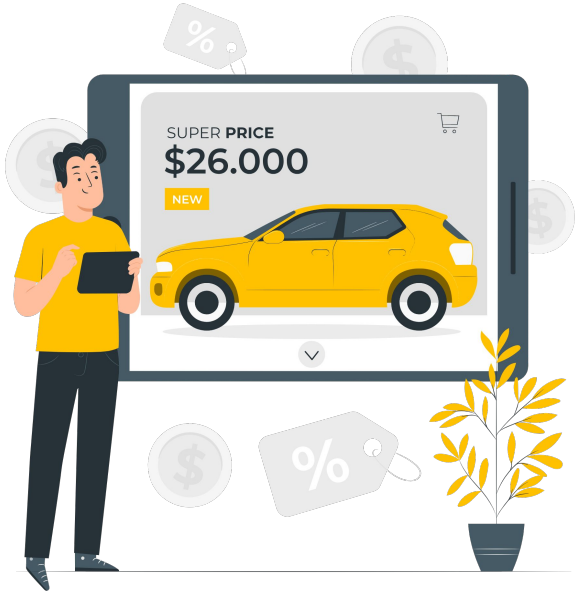
CAR PRICE PREDICTION

INTRODUCTION

- Predicting the price of a used cars has been studied extensively in various researches.
- Car price prediction is a somehow an interesting and popular problem.
- As per information from the Agency for Statistics of BiH, the percentage of personal car usage is increased by 2.7% since 2013 and it is likely that this trend will continue, and the number of cars will increase day by day.
- Accurate car price prediction involves expert knowledge, because price usually depends on many distinctive features and factors.
- Typically, the most significant ones are present price , brand and model, age, mileage etc. The fuel type used in the car as well as fuel consumption per mile highly affect the price of a car due to a frequent changes in the price of a fuel. Different features like exterior color, type of transmission, safety, air condition, etc. will also influence the car price.



BACKGROUND STUDY

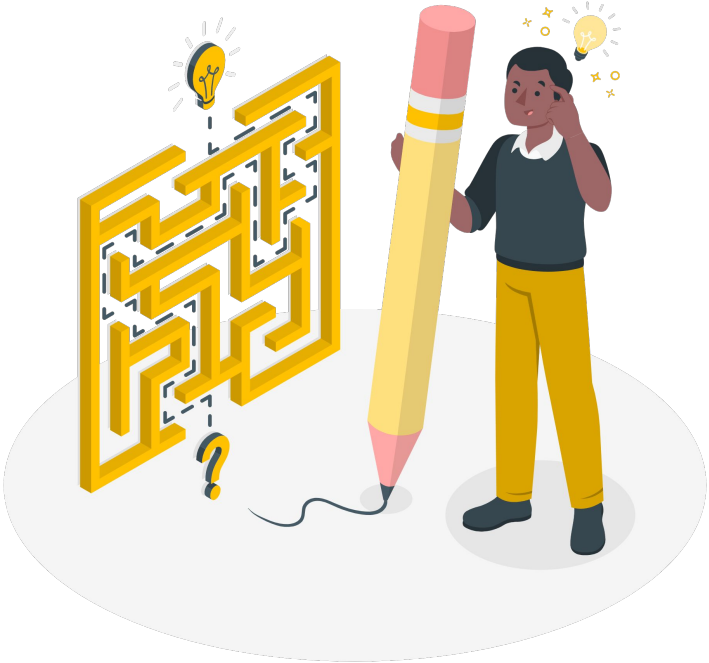
- Wu et al. conducted a car price prediction study, by using a neuro-fuzzy knowledge-based system. Their prediction model produced similar results as the simple regression model.
- Listian discussed, in her paper, that a regression model that was built using Support Vector Machines (SVM) can predict the price of a car that has been leased with better precision than multivariate regression.
- Another approach was given by Richardson in his thesis work . Richardson applied multiple regression analysis and demonstrated that hybrid cars retain their value for a longer time than traditional cars.
- Pudaruth applied various machine learning algorithms, namely: k-nearest neighbors, multiple linear regression analysis, decision trees and naïve bayes for car price prediction in Mauritius.
- Noor and Jan built a model for car price prediction by using multiple linear regression.

PROBLEM STATEMENT

A model to predict the price of a used car should be developed in order to assess its value based on a variety of characteristics. Several factors affect the price of a used car, such as company, model, year, transmission, distance driven, fuel type, seller type, and owner type. As a result, it is crucial to know the car's actual market value before purchasing or selling it.

CHALLENGES

- Finding the best regression algorithm, among Linear Regression, Lasso Regression, Random Forest Regression, Ridge regression, etc., for our problem was a challenge.
- We had to learn about random forest regression since we weren't familiar with it.
- Reconstructing the given dataset. Changing the categorical values into numerical form. And removing the unnecessary features.
- The choice between *randomizedsearchCV* and *gridsearchCV* for finding hyperparameters of random forest regression presented a challenge.
- We had a challenge in choosing dominant features for which we used heat maps using pearson coefficient correlation and extra tree regressor and found the feature importances.
- We developed a webpage to demonstrate the workings of our model for which we had to learn HTML and CSS.



SOLUTION APPROACH

- **Step 1:** "Car_Data.csv" dataset was taken from kaggle.com and was reconfigured to reflect the important features.

  Data frames in which we loaded the dataset now include the Selling price , Present price, Kms driven, owner, age, Fuel Type, Seller type, Transmission type.

- **Step 2:** Categorical features were then converted to numerical values.

  We remove multiple columns of the dataset using the get_dummies function as some columns contain the same information because the original column could assume a binary value, The "CNG" fuel type, the "Automatic" transmission type, and the "Dealer" seller type were removed.

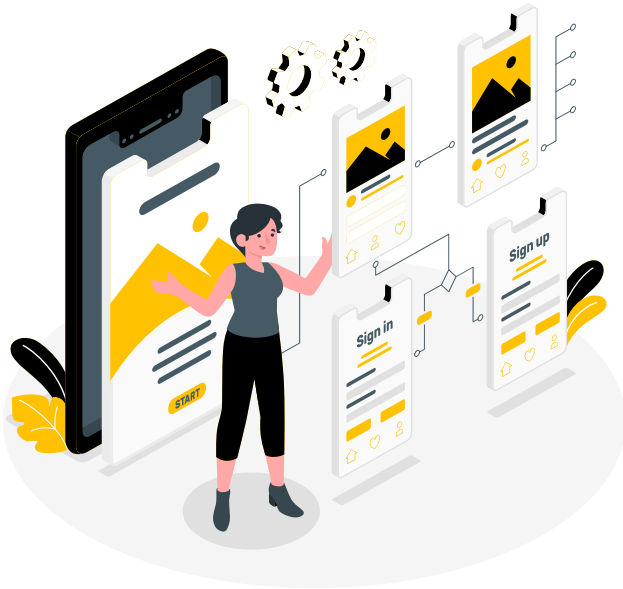
- **Step 3:** We created a heat map using Pearson correlation to illustrate how the features are related.

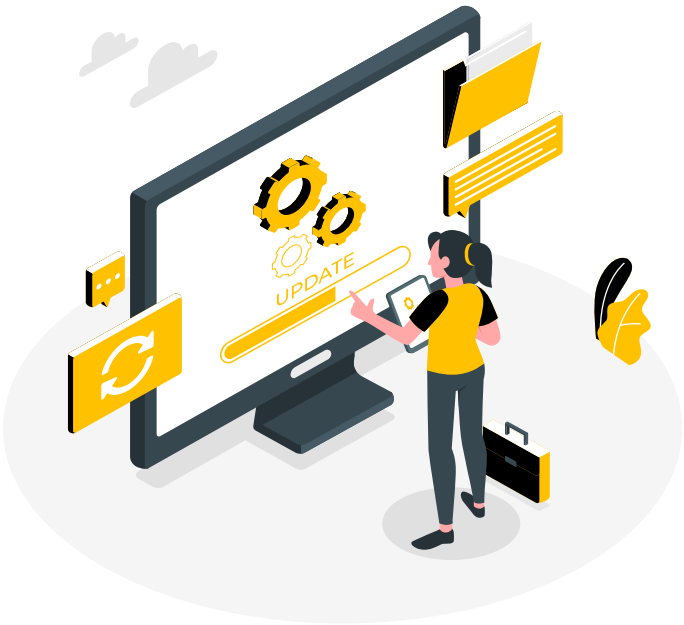
SOLUTION APPROACH

- **Step 4:** Using *ExtraTreesRegressor* the feature importances were obtained. And *Present_Price* turned out to be the feature with most importance.
- We used Random Forest Regression to solve the problem and created a model based on this.
- **Step 5:** By using *randomizedsearchCV*, we performed hyper parameter tuning

  (n_estimators, max_features, max_depth, min_samples_split, min_samples_leaves).

  Due to its speed, *RandomizedSearchCV* is better than *GridSearchCV*.
- **Step 6:** The data has been split into training data (80%) and test data (20%).





SOLUTION APPROACH

- **Step 7:** After getting the best parameters from the RandomizedSearchCV we train the model using the Random Forest Regression .

  In the random forest regressor, the decision tree will scale the input, so we do not have to scale the values.

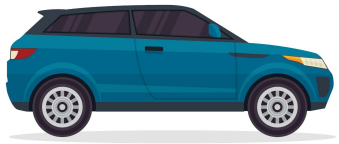
- The graph of predictions is plotted using displot/ distplot in which we observe that it resembles a normal distribution with mean 0.

- **Step 8 :** The file is put in a pickle file. Also the performances of the model are computed.

- **Step 9 :** The model is demonstrated by entering features in an HTML file and predicting the price by using a Python file from the set of values given as features.

  Lastly, the final output is displayed on the HTML page.

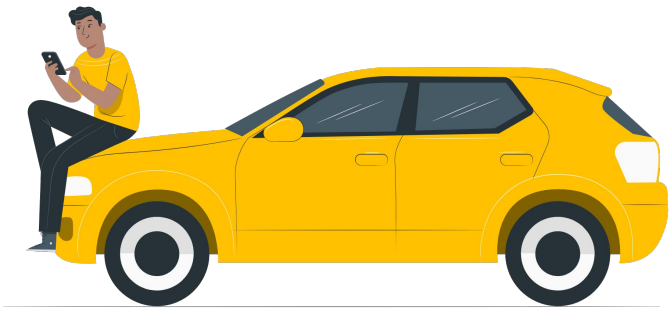
DATASET



The data we have used in this project was downloaded from Kaggle. It was uploaded from Cardekho.com . The dataset consists of 301 rows and 9 columns with no null values. Column data consist of independent Features. The independent features contain both categorical and numeric values.

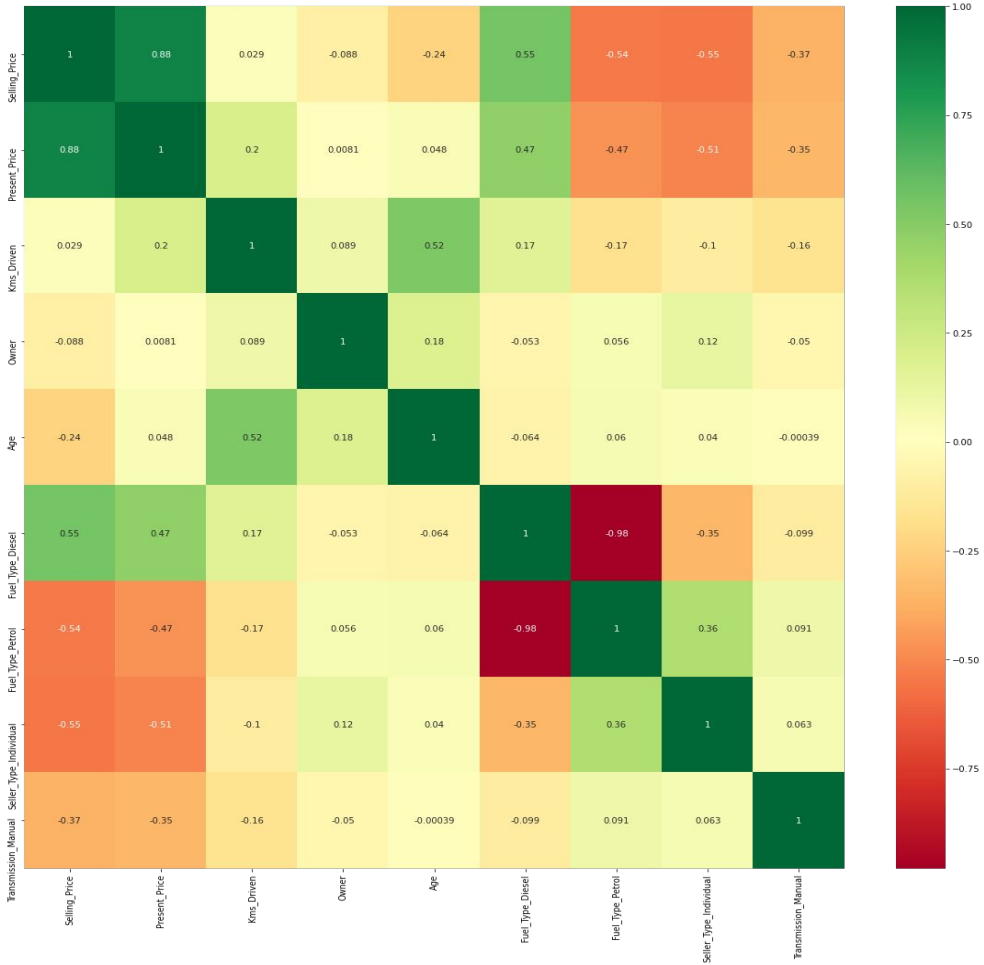
	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

DATASET MANIPULATION



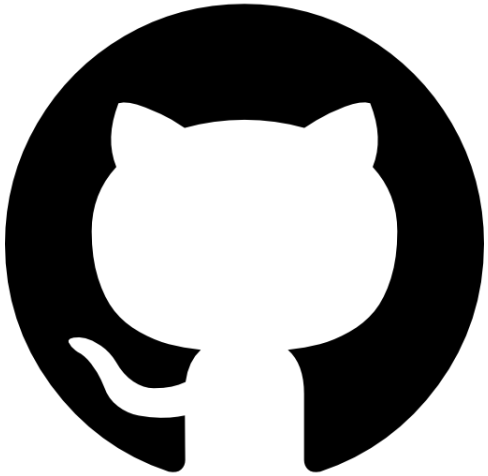
- From the original dataset we searched for null values using “isnull()”.
- Removed the “car_name” column from the dataset. We added a new feature “Age” by subtracting current year and purchase year. Then we dropped the “Year” column.
- As the dataset has numerical values and categorical values , We remove multiple columns of the dataset using the get_dummies function as some columns contain the same information because the original column could assume a binary value,
- The "CNG" fuel type, the "Automatic" transmission type, and the "Dealer" seller type were removed from the unique values.

	Selling_Price	Present_Price	Kms_Driven	Owner	Age	Fuel_Type_Diesel	Fuel_Type_Petrol	Seller_Type_Individual	Transmission_Manual
0	3.35	5.59	27000	0	7	0	1	0	1
1	4.75	9.54	43000	0	8	1	0	0	1
2	7.25	9.85	6900	0	4	0	1	0	1
3	2.85	4.15	5200	0	10	0	1	0	1
4	4.60	6.87	42450	0	7	1	0	0	1



FINAL DATASET

- Using the Extra Trees Regressor the feature importances are found.
- The Present Price is the feature with most importance
- And also we computed the correlation of all the features , we used “pearson” method.
- And from that the we found that Selling price and the present price are much correlated.
- And a heatmap is used to plot the correlation between 2 features.

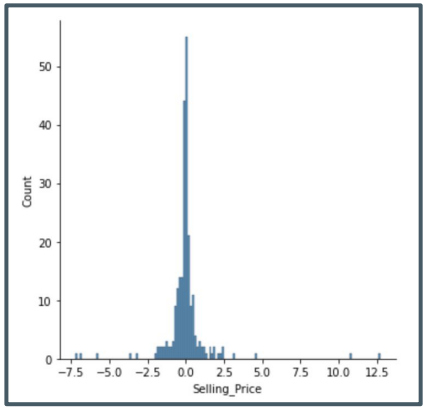


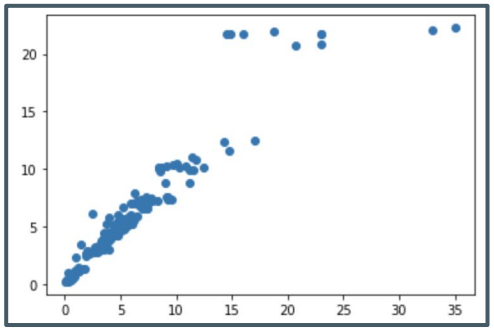
GITHUB REPO

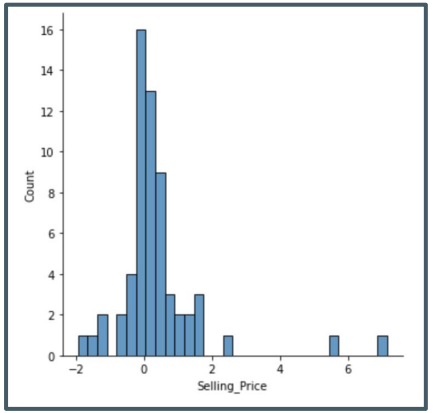
Link:
iyashk/Car-Price-Prediction(github.com)

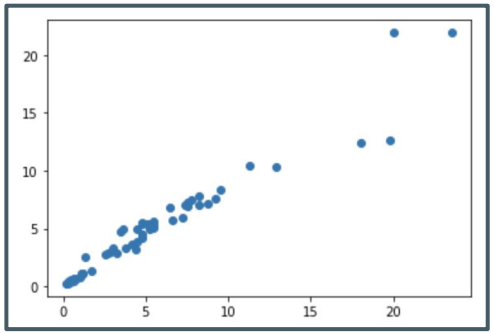
OBSERVATION

- After the model is trained , we plotted the difference between test and train data along with the predictions made using displot . This is shown in normal distribution with mean 0.
- This tells us that the error is not much. So the predicted value might actually be accurate.
- Also when we use the scatter plot the test points and the predicted points are along the line y=x , which means that the predicted values are equal to the test values using the model.





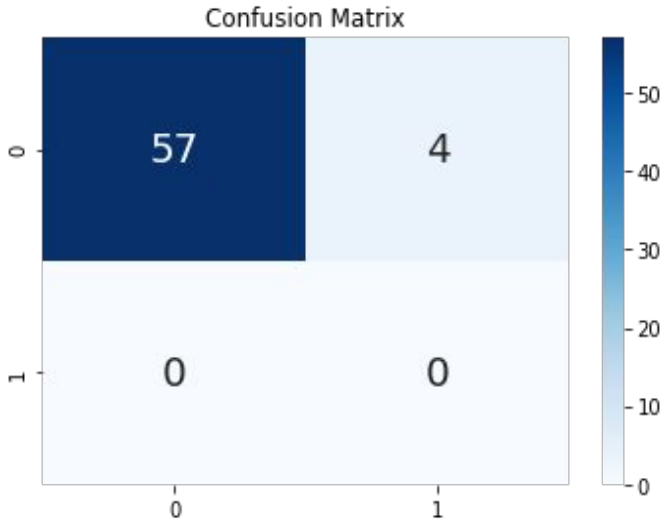




OBSERVATION

- The performance measures are found using the Root Mean Square Error , R2 Score , Mean Square Error.
- Also accuracy of the model is calculated from the sklearn's built in functions. As this is only done for classification we made our predicted output using a cutoff. If the difference between the test and predicted value is greater than cutoff (MSE) then we classify as wrongly predicted else correct prediction.
- Accuracy of the model calculated from the "accuracy_score" gave us a score of 93.4 percentage.
- Also calculating the Root Mean Square Error for the test and predicted test values is 1.36.
- Calculating the Mean Square Error for the test and predicted test values we get 1.97.

```
accuracy of the model : 0.9344262295081968
```



RESULTS

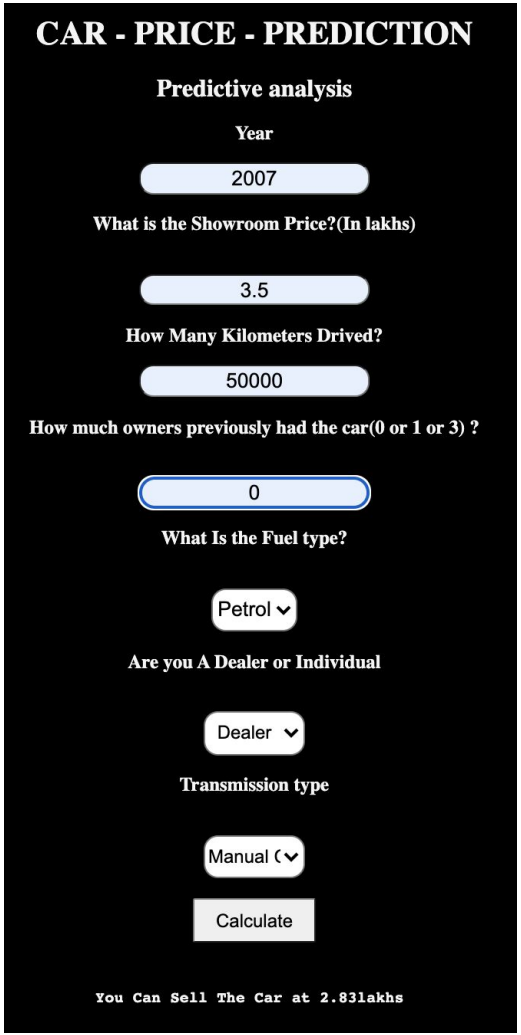
The results are demonstrated in html page. The features are entered and the selling price is calculated.

FEATURES

Year	:	2007
Present price	:	3.5 lakhs
Kms driven	:	50000
Owner	:	0
Fuel Type	:	Petrol
Seller Type	:	Dealer
Transmission type	:	Manual

PREDICTED PRICE

The price predicted is 2.81 lakhs



RESULTS

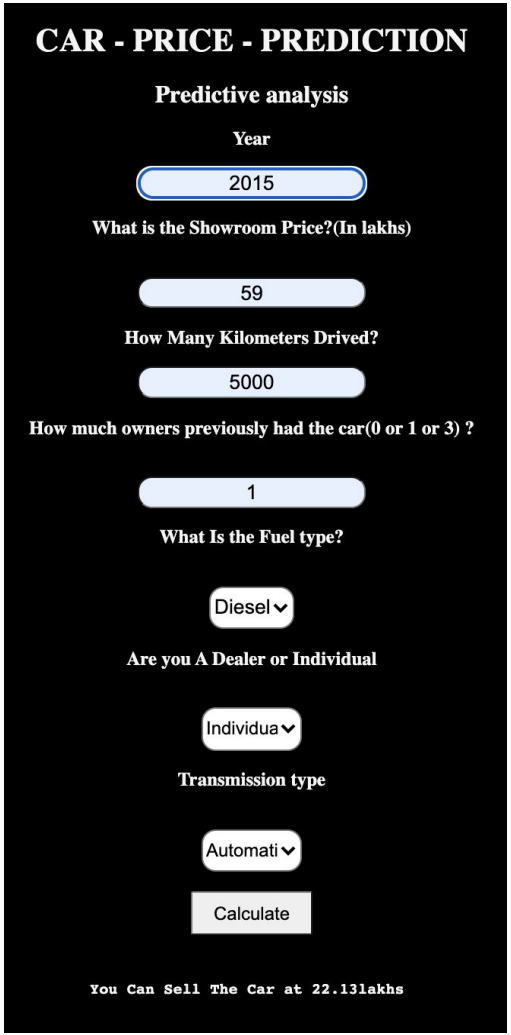
The results are demonstrated in html page. The features are entered and the selling price is calculated.

FEATURES

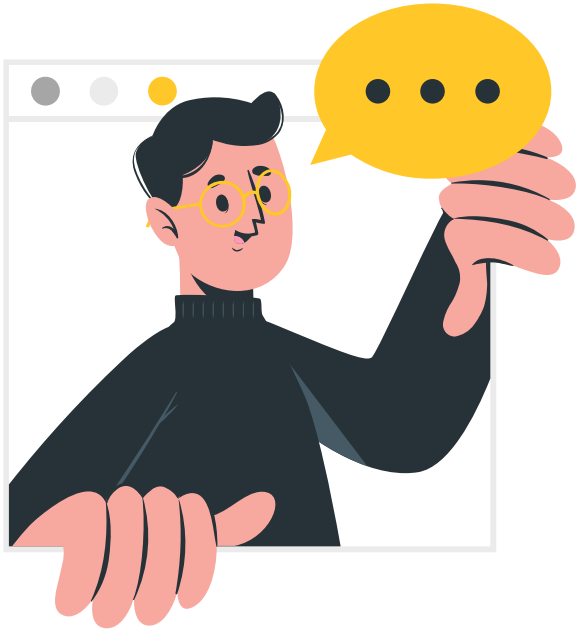
Year	:	2015
Present price	:	59 lakhs
Kms driven	:	5000
Owner	:	1
Fuel Type	:	Diesel
Seller Type	:	Individual
Transmission type	:	Automatic

PREDICTED PRICE

The price predicted is 22.13 lakhs



REFERENCES



[1] Gegic, E.; Isakovic, B.; Keco, D.; Masetic, Z.; Kevric, J. Car price prediction using machine learning techniques. TEM J. 2019, 8, 113.

[2]USA_CAR_SELLING_PRICE/Second-checkpoint.ipynb at master · harsh0703-harsh/USA_CAR_SELLING_PRICE (github.com)

[3] Capstone_Project_Machine_Learning/Capstone_Project_Report_with_Python.pdf at master · EnesGokceDS/Capstone_Project_Machine_Learning (github.com)

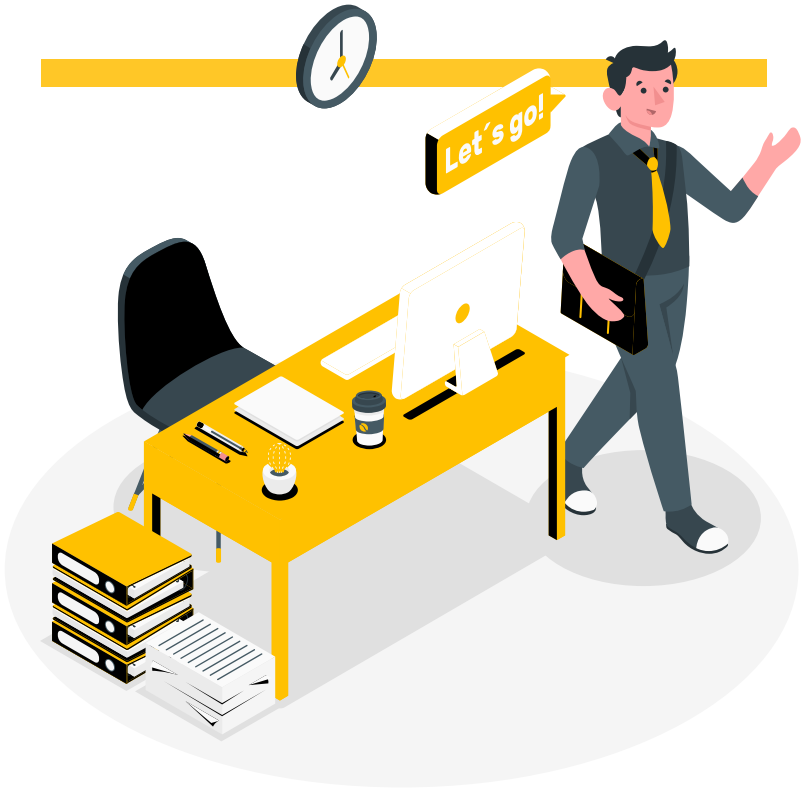
[4] sklearn.ensemble.ExtraTreesRegressor — scikit-learn 1.0.2 documentation

[5] sklearn.ensemble.RandomForestRegressor — scikit-learn 1.0.2 documentation

[6] sklearn.model_selection.RandomizedSearchCV — scikit-learn 1.0.2 documentation

[7] used-car-price-prediction/ml-models.ipynb at master · abhashpanwar/used-car-price-prediction (github.com)

[8] Predicting Used Car Prices.pdf (stanford.edu)



THANK YOU!

“Predicting the future isn’t magic
It’s Artificial Intelligence.”

~Dave Waters