

SALE PRICE PREDICTION OF USED TOYOTA CARS

BY:

242062017 KIRAN PANHALKAR

COURSE: ADVANCE MACHINE LEARNING

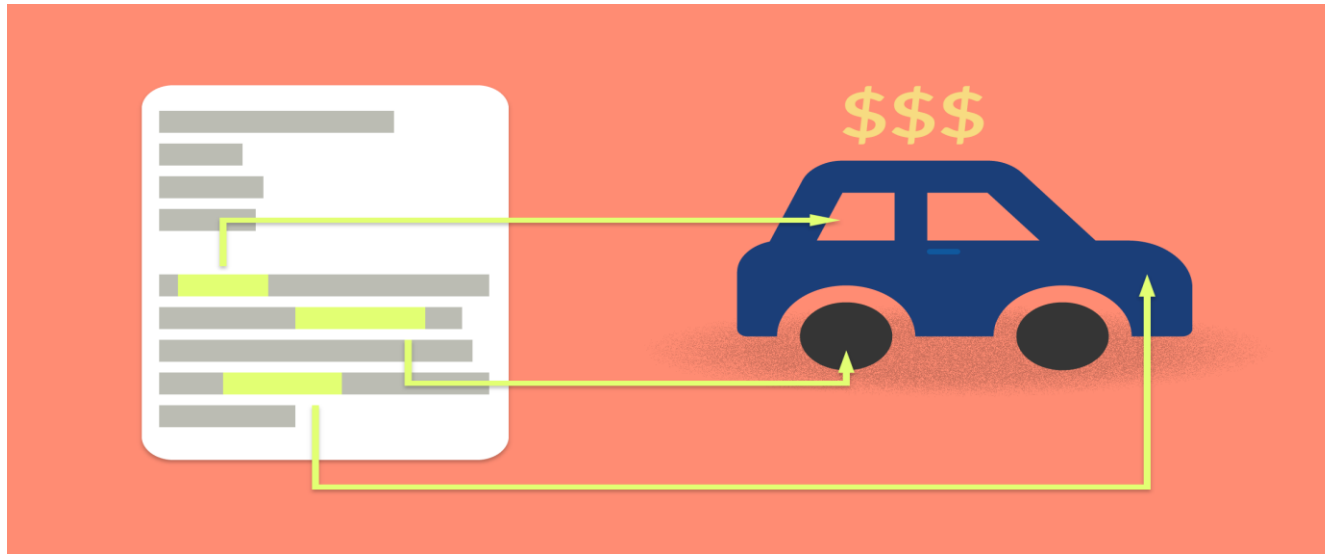


CONTENTS:

- Objective
- Description of Dataset
- Challenges
- Modelling Flow
- Methodology
- Models
- Results
- Graphs
- Deployment
- Conclusion
- Future Scope

OBJECTIVE:

- Build a machine learning model to predict the sale price of a used automobile and deploy the best model.



DESCRIPTION OF DATASET:

- **Description:** The data set includes sale prices and vehicle characteristics of 1436 used Toyota Corollas.
- **Independent variables:**
 - i. **Age:** Age in months
 - ii. **KM:** Accumulated Kilometers on odometer
 - iii. **Fuel Type:** Fuel Type (Petrol, Diesel, CNG)
 - iv. **HP:** Horse Power
 - v. **Met Color:** Metallic Color (Yes=1, No=0)
 - vi. **Automatic:** Automatic ((Yes=1, No=0)
 - vii. **CC:** Cylinder Volume in cubic centimeters
 - viii. **Doors:** Number of doors
 - ix. **Weight:** Weight in Kilograms
- **Dependent Variable:**
 - i. **Price:** Offer Price in Euros

CHALLENGES:

1. Checking for Correlation:

- We need to check for correlation between the features in the dataset.
- We need to find which are highly correlated and with are less correlated features in the dataset.
- Based on the analysis we need to go further in mode building.

2. Dealing with missing values:

- First we need to check for all the missing values in the dataset.
- If there are missing values we need to perform the required processing to deal with those values.
- After processing we again need to check are there any more missing values in the dataset.

3. Changing the format of the variables:

- In the dataset, there are some variables which are not in the proper format i.e. In doors column, the number of doors are in figures as well as in words.
- Our task here is to make it in one proper format before implementing the model.



4. Dealing with the outliers in the dataset:

- Finding the outliers from the dataset and removing those outliers.

5. Dealing with Categorical variables:

- There were many categorical variables in the dataset such as 'MetColor','Automatic','CC','Fuel Type'
- For all these variables used Column Transformer and One Hot encoding to make them into numeric datatype.

6. Features Selection from the dataset:

- Using Random Forest Regressor finding the best features from our dataset
- RFE

MODELLING FLOW:

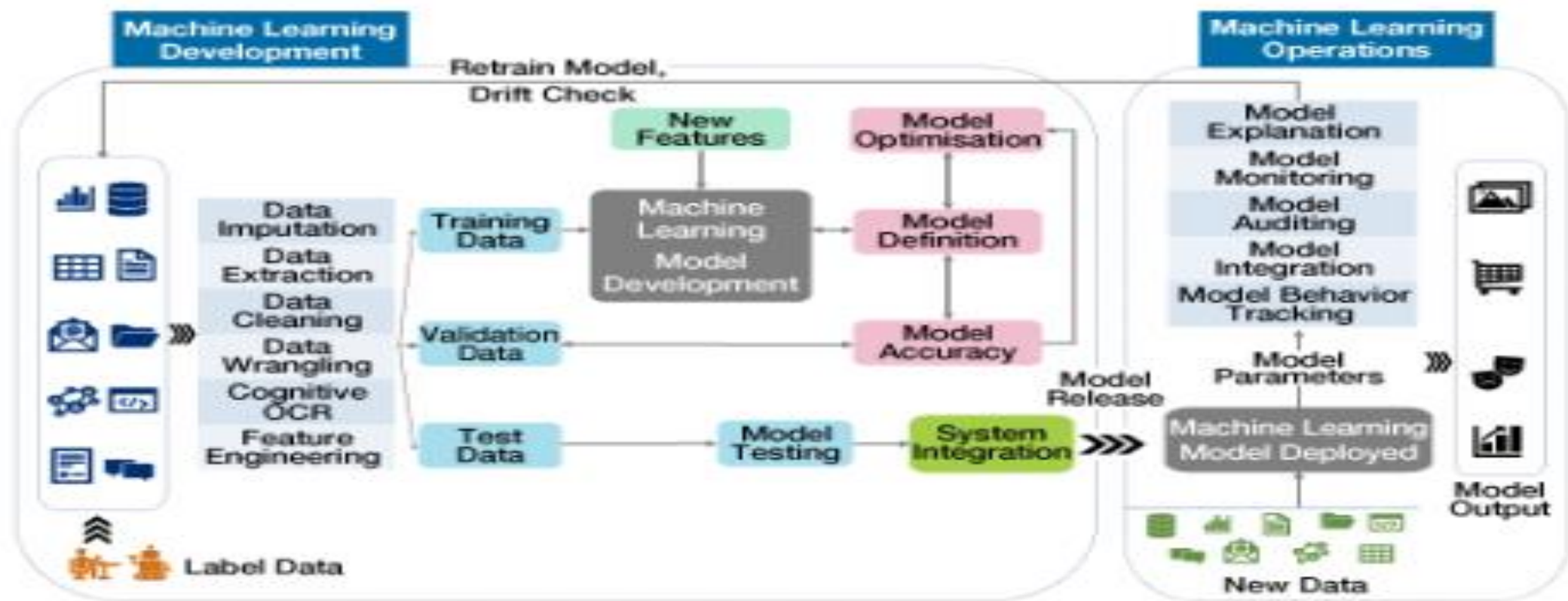


Fig 6 Basic ML modelling Flow

- For our modelling flow is as follows:

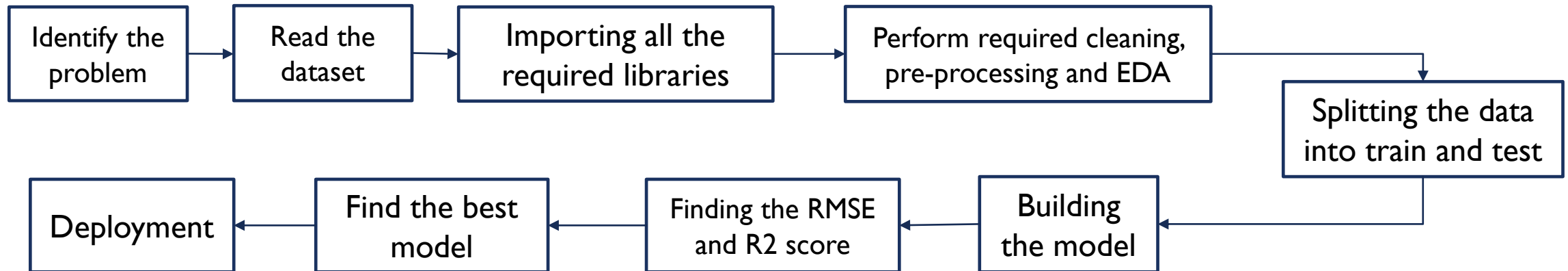




Fig 7 Our Modelling Flow

METHODOLOGY:

1. Identifying the problem Statement.
2. Reading the dataset in google colab
3. Importing all the required libraries for implementing the model.
4. Performing required preprocessing on the dataset which includes:
 - EDA:
 - i. Finding the correlation
 - ii. Finding the best features
 - iii. Statistical Analysis
 - Preprocessing:
 - i. Finding the missing values
 - ii. Dealing with categorical variables
 - iii. Dealing with variables which are in words and in digits
 - iv. Making the dataset in the proper format before building up the model.



5. Splitting the dataset in train data and test data

6. Building a model on various algorithms:

- i. Linear Regression
 - ii. Linear Regression with Less features
 - iii. Decision Tree
 - iv. Decision Tree Prune
 - v. Bagging
 - vi. Random Forest
 - vii. Random Forest Tuned
 - viii. Ridge
 - ix. Lasso
7. Finding the best RMSE and R2 score for train and test data.
8. Plotting the graph of those results
9. Deployment
- i. Creating a pickle file
 - ii. Testing new data on that pickle file.

MODELS:

■ Linear Regression: -

- i. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.
- ii. Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable.

■ Decision Tree: -

- i. Top Down Approach.
- ii. Each node in the decision tree represents a test case for an attribute and each descent (branch) to a new node corresponds to one of the possible answers to that test case.
- iii. In this way, with multiple iterations, the decision tree predicts a value for the regression task or classifies the object in a classification task.

■ Decision Tree Prune:

- i. Is used to overcome our problem of Overfitting.
- ii. It reduces the size of a Decision Tree which might slightly increase your training error but drastically decrease your testing error, hence making it more adaptable.

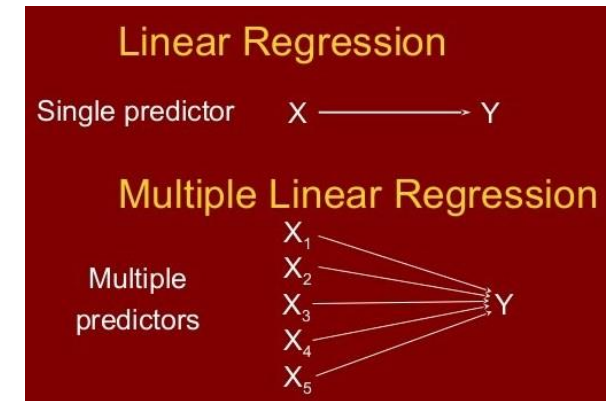


Fig 1 Linear Regression

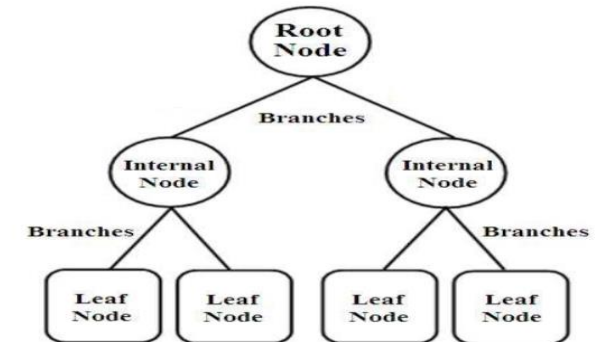


Fig 2 Decision Tree

- **Bagging:** -

- i. To improve the stability and accuracy of machine learning algorithms used in statistical classification and regression.
- ii. It also reduces variance and helps to avoid overfitting.

- **Random Forest:** -

- i. The decision tree models over fit the data hence the need for Random Forest arises.
- ii. Decision tree models may be Low Bias but they are mostly high variance.
- iii. Hence to reduce this variance error on the test set, Random Forest is used.

- **Random Forest Tuned:**

- i. Hyper parameter tuning is very important as it helps us control bias and variance performance of our model.
- ii. Grid search is used after randomized search to narrow down the range to search the perfect hyper parameters.

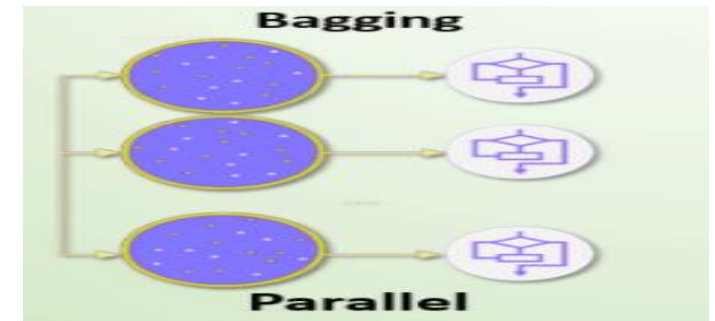


Fig 3 Bagging Model

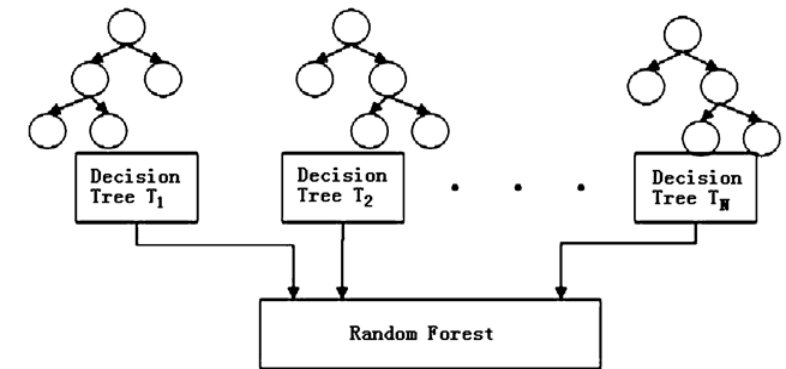


Fig 4 Random Forest Model

■ Ridge:

- i. In Ridge Regression, the loss function is modified with a shrinkage quantity corresponding to the summation of squared values of β . And the value of λ decides how much the model would be penalized.
- ii. The coefficient estimates in Ridge Regression are called the L2 norm. This regularization technique would come to your rescue when the independent variables in your data are highly correlated.

■ Lasso:

- i. In Lasso Regression, a penalty equaling the sum of absolute values of β (modulus of β) is added to the error function. It is further multiplied with parameter λ which controls the strength of the penalty. Only the high coefficients are penalized in this method.
- ii. The coefficient estimates produced by Lasso are referred to as the L1 norm. This method is particularly beneficial when there are a small number of observations with a large number of features.

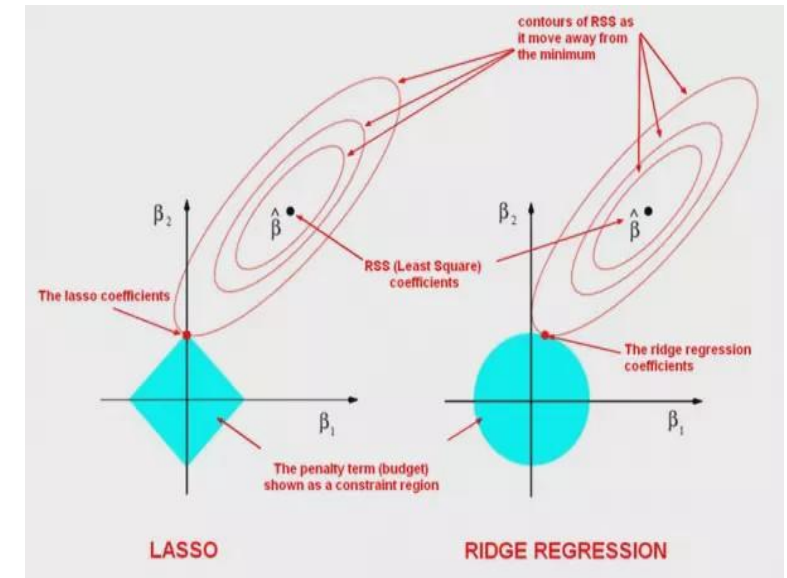


Fig 5 Lasso and Ridge Regression Model

RESULTS:

- Results of all the models implemented with their RMSE values, R2 on train Data, and R2 on test data.
- Looking at the results we can say that Random Forest tuned has highest R2 score on train data as well as on test data.

	Model	RMSE	Model Accuracy Score	R2_score_test
0	Linear Regression	1334.124101	0.872686	0.784689
1	LR with Features selection	1334.124101	0.810852	0.784689
2	LR with REF	1380.493837	0.852088	0.769462
3	Decision Tree Regression	1385.464580	0.871504	0.767799
4	Decision Tree Prune	1382.148279	0.914706	0.768909
5	Bagging	1299.360391	0.851540	0.797299
6	Random Forest	1226.758508	0.928317	0.817950
7	Random Forest Tuned	1165.452106	0.939334	0.835691
8	Ridge Regression	1326.886590	0.867084	0.787019
9	Lasso Regression	1334.122371	0.872686	0.784690

Table I Shows the results of each Algorithm implemented

GRAPHS:

■ Comparing Accuracy and R2_score_test:

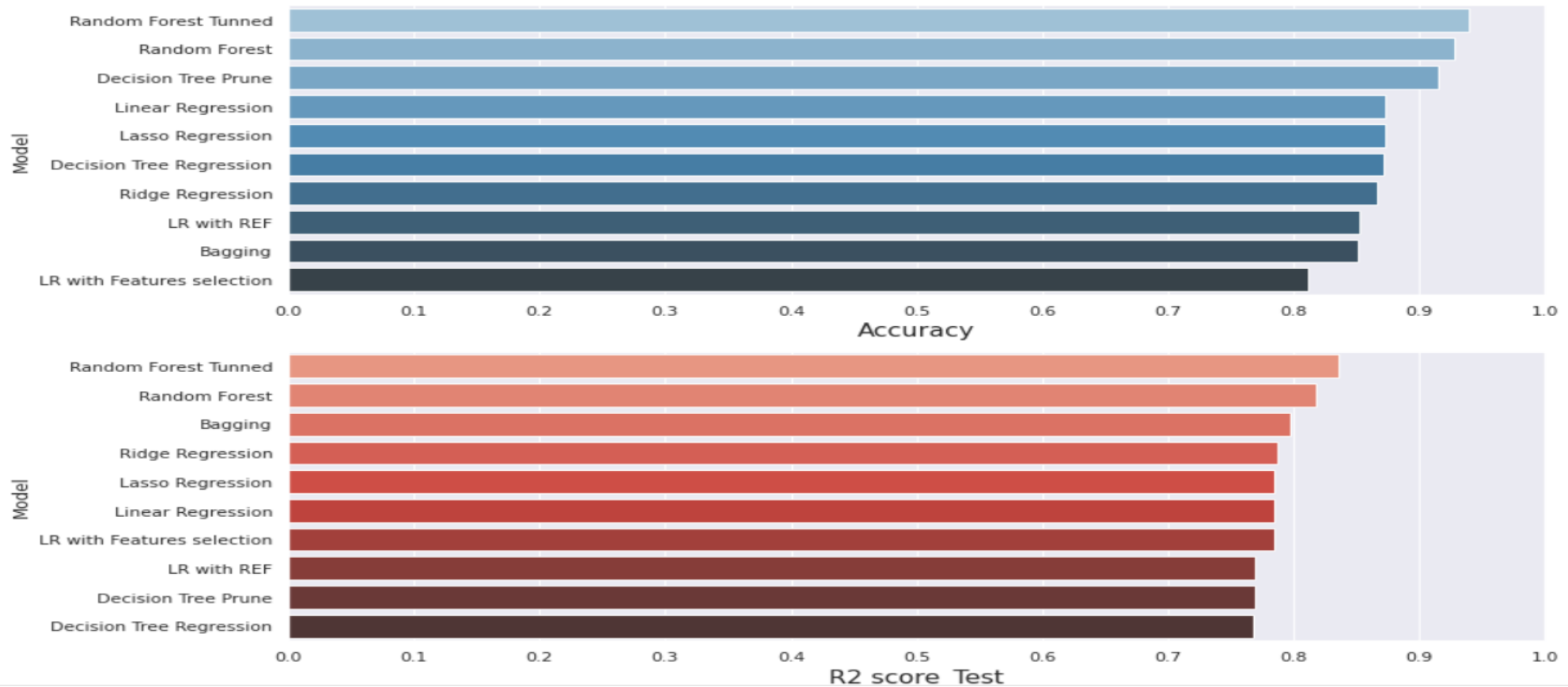


Fig 8 Comparing Accuracy and R2_score_test



■ RMSE for each algorithm:

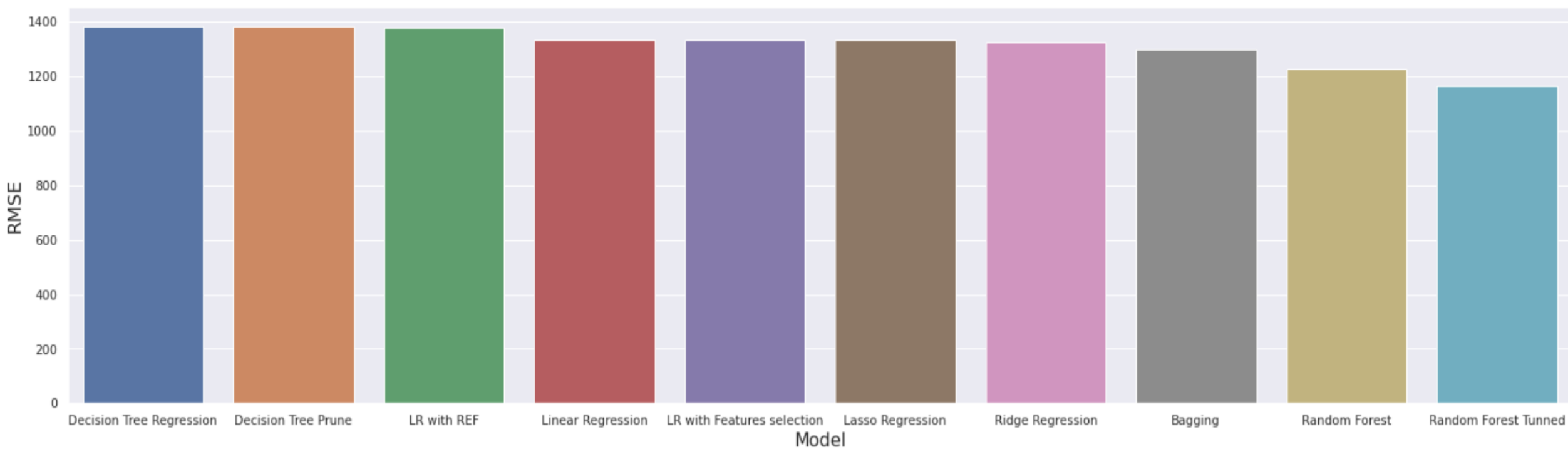


Fig 9 Comparing RMSE for each Algorithm implemented

DEPLOYMENT

- Based on the results obtained we have found that Random Forest tuned is best model for the deployment.
- Preparing a pickle file for the deployment.
- Testing that pickle file on new data.

CONCLUSION:

- In this way, we have implemented a Machine Learning model on various algorithms on Toyota used cars dataset .
- The best Algorithm based on R^2 is for Random Forest tuned, also it has the lowest RMSE amongst all the algorithms implemented.
- Finally, we have deployed Random forest tuned for further implementation.

FUTURE SCOPE

- We can further implement our deployed model on real time dataset to predict the price of used Toyota cars.



Thank You!