Random Forests: Toyota Corolla Dataset

Group 9

Yabi Yirga

Wanying Xu

Adriana Choy S.

Favour Nnadi

Overview

1. Introduction to the Assignment
2. Introduction to the Toyota Corolla Dataset
3. Understanding Random Forest
4. Exploratory Data Analysis (EDA)
5. Random Forest Model Building and Performance
6. Feature Importance Analysis

Introduction to Assignment

Choose a dataset and a predictive task, preprocess the data as needed, and train a random forest to predict the target variable

Discuss the performance of the random forest

About the dataset

The dataset consists of 1434 Observations and 39 Variables

Variable	Description
Id	Record_ID
Model	Model Description
Price	Offer Price in EUROs
Age_08_04	Age in months as in August 2004
Mfg_Month	Manufacturing month (1-12)
Mfg_Year	Manufacturing Year
KM	Accumulated Kilometers on odometer
Fuel_Type	Fuel Type (Petrol, Diesel, CNG)
HP	Horse Power
Met_Color	Metallic Color? (Yes=1, No=0)
Color	Color (Blue, Red, Grey, Silver, Black, etc.)
Automatic	Automatic ((Yes=1, No=0)
CC	Cylinder Volume in cubic centimeters
Doors	Number of doors
Cylinders	Number of cylinders
Gears	Number of gear positions
Quarterly_Tax	Quarterly road tax in EUROs
Weight	Weight in Kilograms
Mfr_Guarantee	Within Manufacturer's Guarantee period (Yes=1, No=0)
BOVAG_Guarantee	BOVAG (Dutch dealer network) Guarantee (Yes=1, No=0)
Guarantee_Period	Guarantee period in months
ABS	Anti-Lock Brake System (Yes=1, No=0)
Airbag_1	Driver_Airbag (Yes=1, No=0)
Airbag_2	Passenger Airbag (Yes=1, No=0)
Airco	Airconditioning (Yes=1, No=0)
Automatic_airco	Automatic Airconditioning (Yes=1, No=0)
Boardcomputer	Boardcomputer (Yes=1, No=0)
CD_Player	CD Player (Yes=1, No=0)
Central_Lock	Central Lock (Yes=1, No=0)
Powered_Windows	Powered Windows (Yes=1, No=0)
Power_Steering	Power Steering (Yes=1, No=0)
Radio	Radio (Yes=1, No=0)
Mistlamps	Mistlamps (Yes=1, No=0)
Sport_Model	Sport Model (Yes=1, No=0)
Backseat_Divider	Backseat Divider (Yes=1, No=0)
Metallic_Rim	Metallic Rim (Yes=1, No=0)
Radio_cassette	Radio Cassette (Yes=1, No=0)
Parking_Assistant	Parking assistance system (Yes=1, No=0)
Tow_Bar	Tow Bar (Yes=1, No=0)

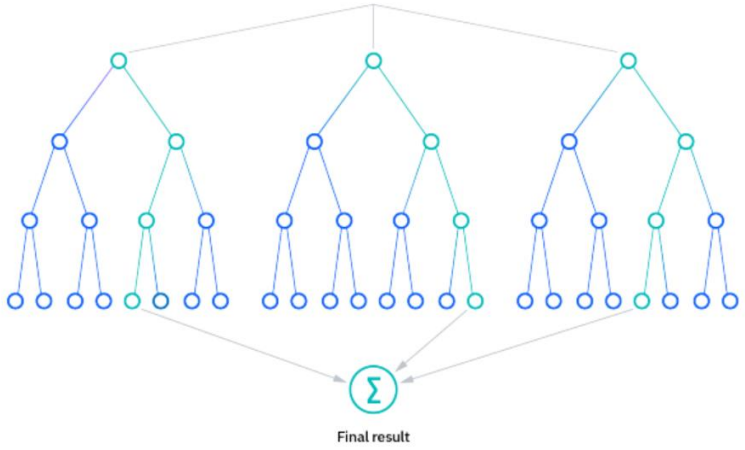
Why this dataset?

Our objective is to predict the price of Toyota corolla price.

Numerous elements, including an automobile's age, mileage, fuel type, color, and other attributes, affect its price.

Random forests are great at capturing the relationships and interactions among these factors, they are a good choice for modeling the complex structure of Toyota Corolla pricing.

What is Random Forest?



- An ensemble learning method for classification and regression tasks
- Combines multiple individual models

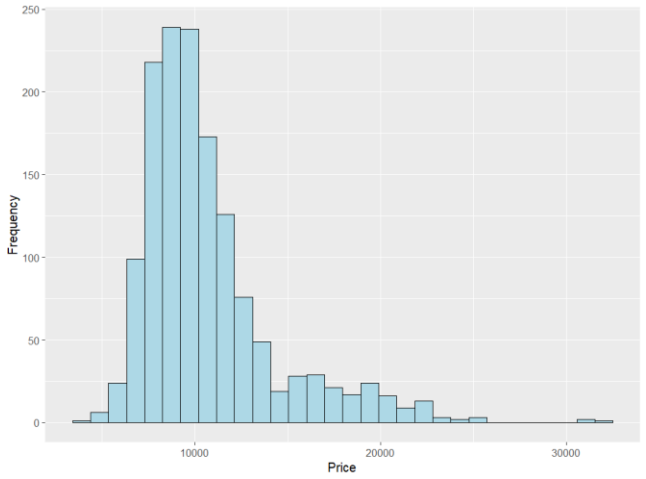
Here's how Random Forest works:

Bootstrap Sampling: Initially, multiple random samples are drawn from the data, with replacement.

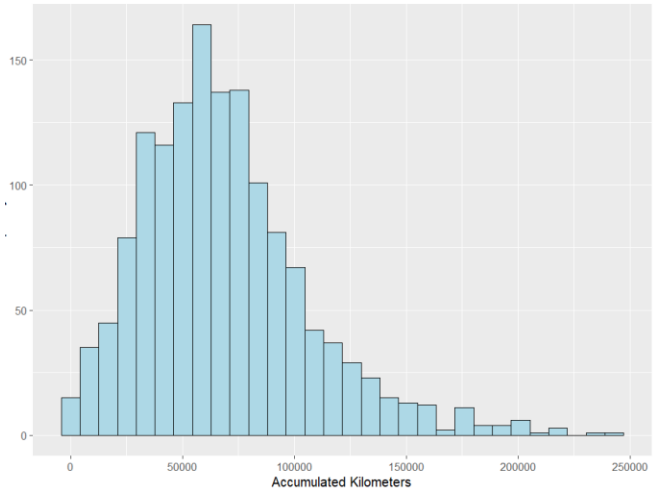
Tree Construction: Utilizing a random subset of predictors at each iteration, a classification or regression tree is fitted to each sample.

Aggregation: the predictions from the individual trees are aggregated to derive improved predictions. Voting for classification and averaging for regression.

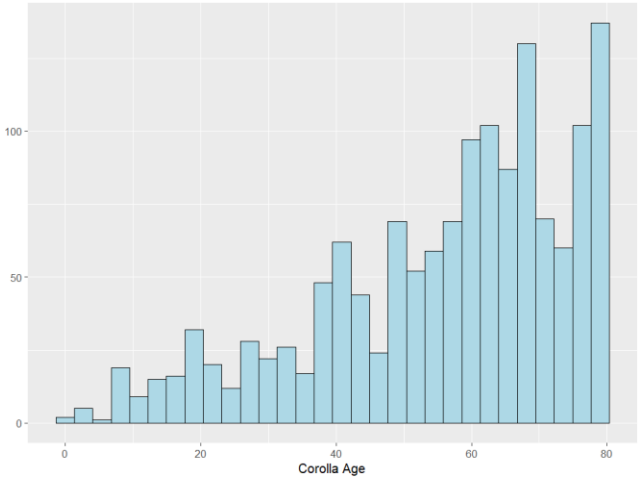
Exploratory Data Analysis



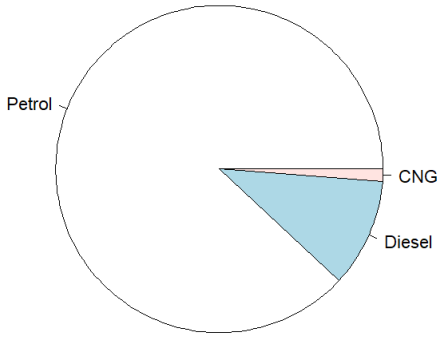
Toyota Corolla price Distribution



Accumulated Kilometers Distribution



Toyota Corolla Vehicle Age Distribution



Fuel Type

Random Forest Model Building

70% of the data was allocated for training purposes

30% was set aside for testing our model's performance

```
#Set Train set and Test set
set.seed(1)
index.train = createDataPartition(TC$Price, p = .7, list=FALSE)
TC.train = TC[index.train,]
TC.test = TC[-index.train,]
str(TC.train)
str(TC.test)
```

Random Forest Function in R

randomforest(formula, data, **ntree**, **mtry**, **nodesize**, **importance**)

ntree – Number of trees to grow

mtry - Number of variables randomly sampled as candidates at each split

Default value for classification: sqrt(p)

Default value for regression: p/3

nodesize - Minimum size of terminal nodes

Default value for classification: 1

Default value for regression: 5

importance - Should importance of predictors be assessed?

Finding an mtry for our model

```
#number of mtry
ctrl = trainControl(method = 'cv', number = 10) #10-fold cross validation
fit.mtry = train(Price ~ ., data = train.df, method='rf', trControl = ctrl,
                 tunelength = 32) #fit rf using cv
print(fit.mtry)
plot(fit.mtry)
```

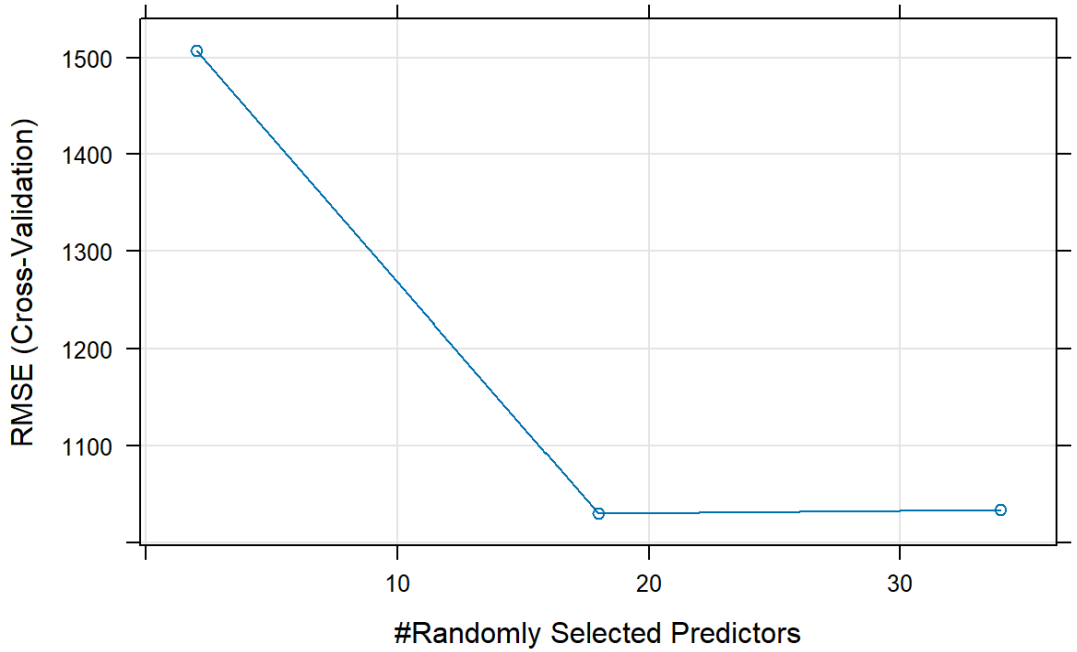
```
> print(fit.mtry)
Random Forest

1005 samples
  33 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 905, 904, 906, 904, 904, 904, ...
Resampling results across tuning parameters:
```

mtry	RMSE	Rsquared	MAE
2	1507.022	0.8757782	1107.6433
18	1030.039	0.9207880	782.4415
34	1033.529	0.9189136	784.9783

```
RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 18.
```



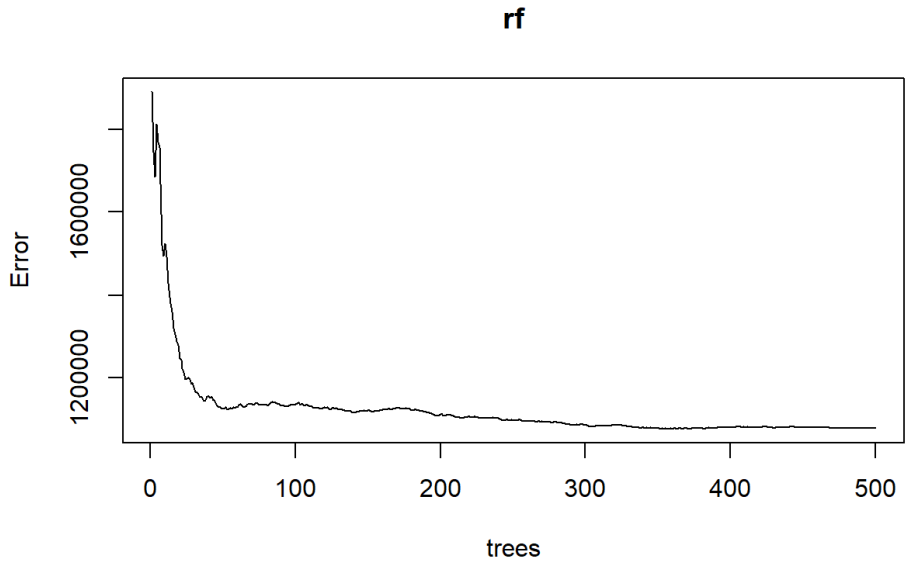
Random forest model fitting

```
#fit random forest using training data
rf = randomForest(Price ~ ., data = train.df, ntree=500, mtry=18, nodesize=5,
                  importance = TRUE)
plot(rf) #number of trees
print(rf)
```

```
> print(rf)

Call:
 randomForest(formula = Price ~ ., data = train.df, ntree = 500,      mtry = 18, nodes
ize = 5, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 18

          Mean of squared residuals: 1078339
                    % Var explained: 91.89
```

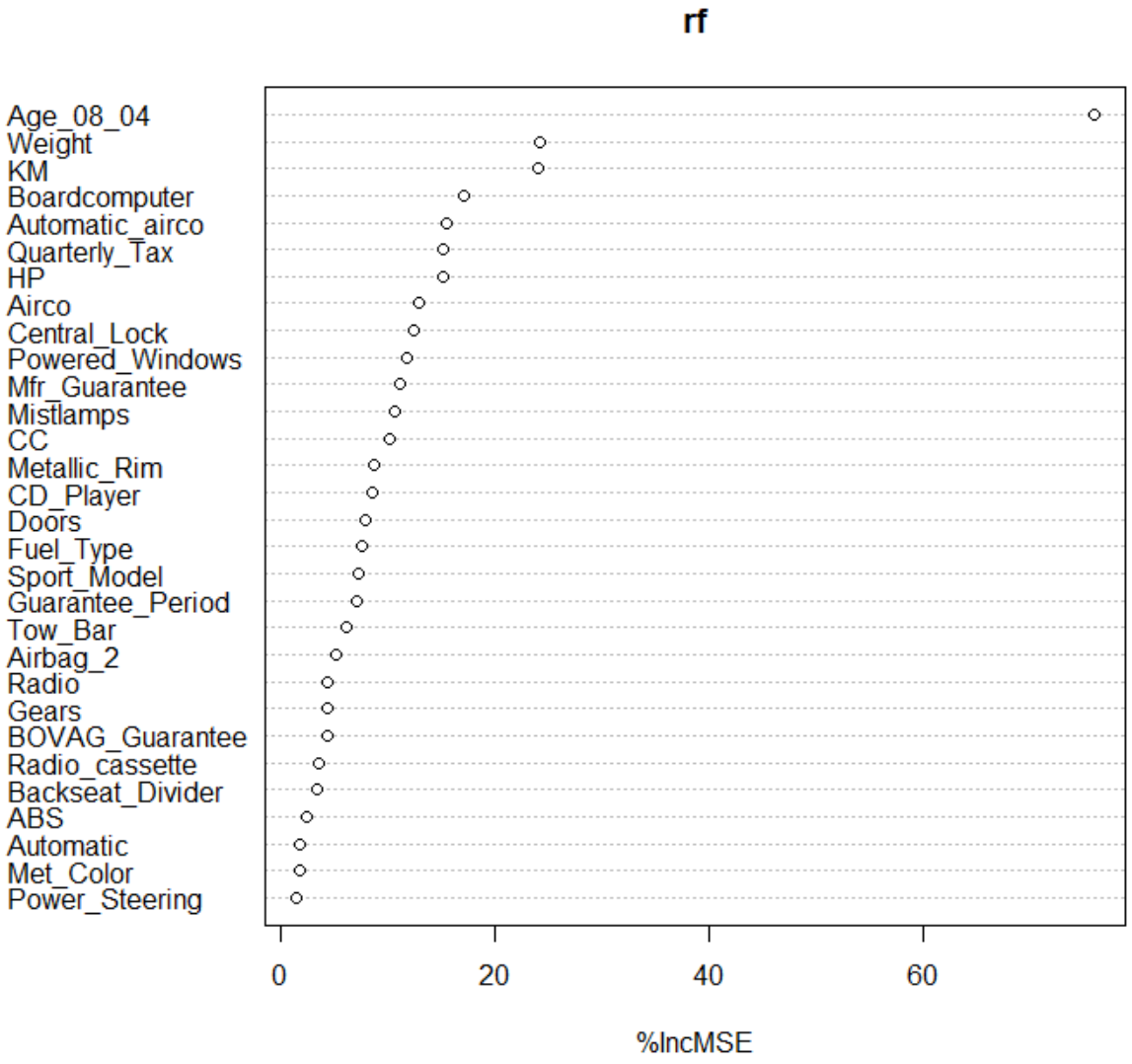


Model performance evaluation

randomForest(formula, data, **ntree = 500**, **mtry = 18**, **nodesize = 5**, importance = TRUE)

Metric	Current model	ntree = 400	Default mtry (p/3)	nodesize = 10
Mean of squared residuals	1078339	1056931	1093830	1097651
% of variance explained	91.89	92.05	91.78	91.75
RMSE	1026.646	1027.023	1020.874	1027.953
MAE	754.8749	753.5944	750.42	754.66
MSE	1054002	1054776	1042183	1056688

Feature importance (Concept + Application)



Feature importance: Measures the relative contribution of the different predictors.

Age is the continuing factor with the highest score of 79.

Relevance of Feature Importance for dataset

- Identify relevant features: e.g. Age of car being
- Feature Selection: We can use only relevant features for model training
- Understanding model behavior: Which ones are informative for predictions?
- Interpretability: Easy to understand for stakeholders and practitioners
- Improving Model Performance: By selecting only the most informative features, feature selection can help improve model accuracy.

Thank you