# Dataset/problem description

Accurately predicting the market value of a vehicle can save dealerships a lot of time in auctions. For example, a used car dealership who wants to purchase vehicles at an auction may not be 100% certain as to what price they should pay for a vehicle. In this example, Toyota Corollas can have many different features and specs, making it difficult to accurately assess the market value of different models in a live auction (when decisions need to be made quickly). Therefore, if predictive models can be used to give a pretty good estimate of a vehicle’s worth, dealers (and other buyers) can use such predictions to select if they want to bid, and if so, at which price they should open their bid at.

You have been hired as a machine learning consultant by ACB Auctions, and your task is to train and test three different machine learning models of your choice. The company has asked you to compare the performance of these three models in predicting the market values (i.e., prices) of Toyota Corollas. They would like you to recommend one of these three models, which they will then continue engineering “in-house” to optimize its performance before deploying the model for real-world decision making.

ACB Auctions has given you the file *ToyotaCorolla.csv* to develop your models, which contains data on used cars (Toyota Corollas) that are for sale in the Netherlands. This dataset contains 1436 observations and 23 potential predictors. The goal will be to predict the price of a used Toyota Corolla based on its specifications. The data dictionary is provided below:

1. Id: Unique identifier in the database
2. Price: Market price of the vehicle (in thousands USD)
3. Age 08 22: Age of the vehicle (in months) as of August 2022
4. KM: Accumulated Kilometers on odometer
5. Fuel Type: Fuel type (Petrol, Diesel, CNG)
6. HP: Horsepower
7. Color: Color (Blue, Red, Grey, Silver, Black, etc.)
8. Automatic: Automatic transmission? (Yes=1, No=0)
9. CC: Cylinder volume in cubic centimeters
10. Doors: Number of doors
11. Cylinders: Number of cylinders
12. Gears: Number of gear positions
13. Mfr Guarantee: Within manufacturer’s guarantee period? (Yes=1, No=0)
14. ABS: Anti-lock brake system? (Yes=1, No=0)
15. Airbag 1: Driver airbag? (Yes=1, No=0)
16. Airbag 2: Passenger airbag? (Yes=1, No=0)
17. Airco: Air conditioning? (Yes=1, No=0)
18. CD Player: CD player? (Yes=1, No=0)
19. Powered Windows: Powered windows? (Yes=1, No=0)
20. Radio: Radio? (Yes=1, No=0)
21. Mistlamps: Mistlamps? (Yes=1, No=0)
22. Sport Model: Spors Model? (Yes=1, No=0)
23. Metallic Rim: Metallic rims? (Yes=1, No=0)
24. Power Steering: Power steering? (Yes=1, No=0)

# Tasks

To successfully complete this assignment, you will need to address the following tasks:

1. Select 3 machine learning models (that we covered in class) that you would like to use to predict the price of Toyota Corollas.
2. Preprocess the data following the steps outlined in the first half of the semester.
3. Train and test all three models. Change the default value for *at least one* hyperparameter in each model.
4. Select one model that you will recommend to ACB Auctions for further testing and development.

# Questions

1. Is this a classification or regression (prediction) problem?
2. Which 3 machine learning models did you select? Are there any particular reasons why?
3. Outline how you preprocessed the data. Be sure to discuss (i) predictor selection, (ii) dealing with NA values, (iii) dealing with categorical predictors, and (iv) scaling.
4. For each model that you selected, which hyperparameter(s) did you change?
5. What are the training and testing performances for each model?
6. How would you judge the fit of each of the 3 models? Why?
7. Which model would you recommend to ACB Auctions? Why?