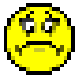
A Light Intro To Boosting

Machine Learning

- Not as cool as it sounds
  - Not iRobot
  - Not Screamers (no Peter Weller )
- Really just a form of
  - Statistics
  - Optimization
  - Probability
  - Control theory
  - ...
- We focus on classification

Classification

- A subset of machine learning & statistics
- Classifier takes input and predicts the output
- Make a classifier from a *training* dataset
- Use the classifier on a *test* dataset (different from the training dataset) to make sure you didn't just memorize the training set
- A good classifier will have low *test* error

Classification and Learning

- Learning classifier learns how to predict after being shown many input-output examples
- Weak classifier is slightly correlated with correct output
- Strong classifier is highly correlated with correct output
- (See the PAC learning model for more info)

Methods for Learning Classifiers

- Many methods available
  - Boosting
  - Bayesian networks
  - Clustering
  - Support Vector Machines (SVMs)
  - Decision Trees
  - . . .
- We focus on boosting

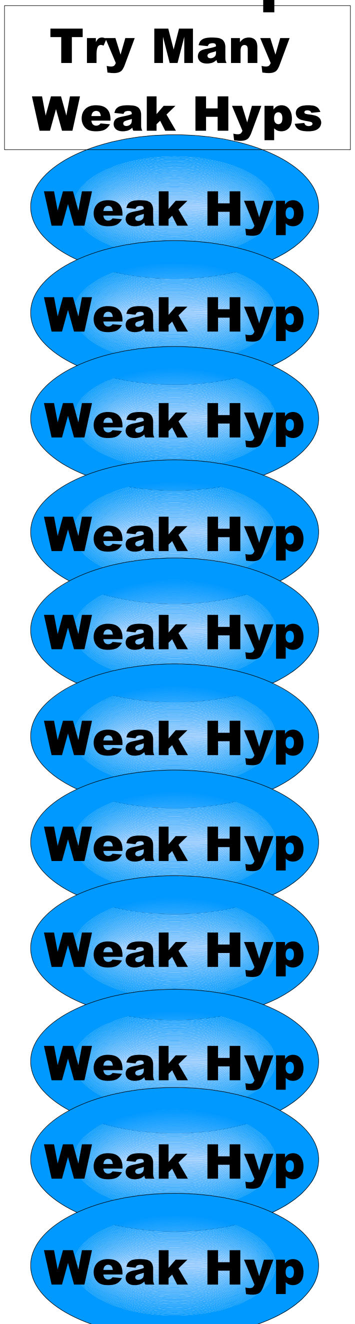
Boosting

- Question: Can we take a bunch of weak hypotheses and create a very good hypothesis?
- Answer: Yes!

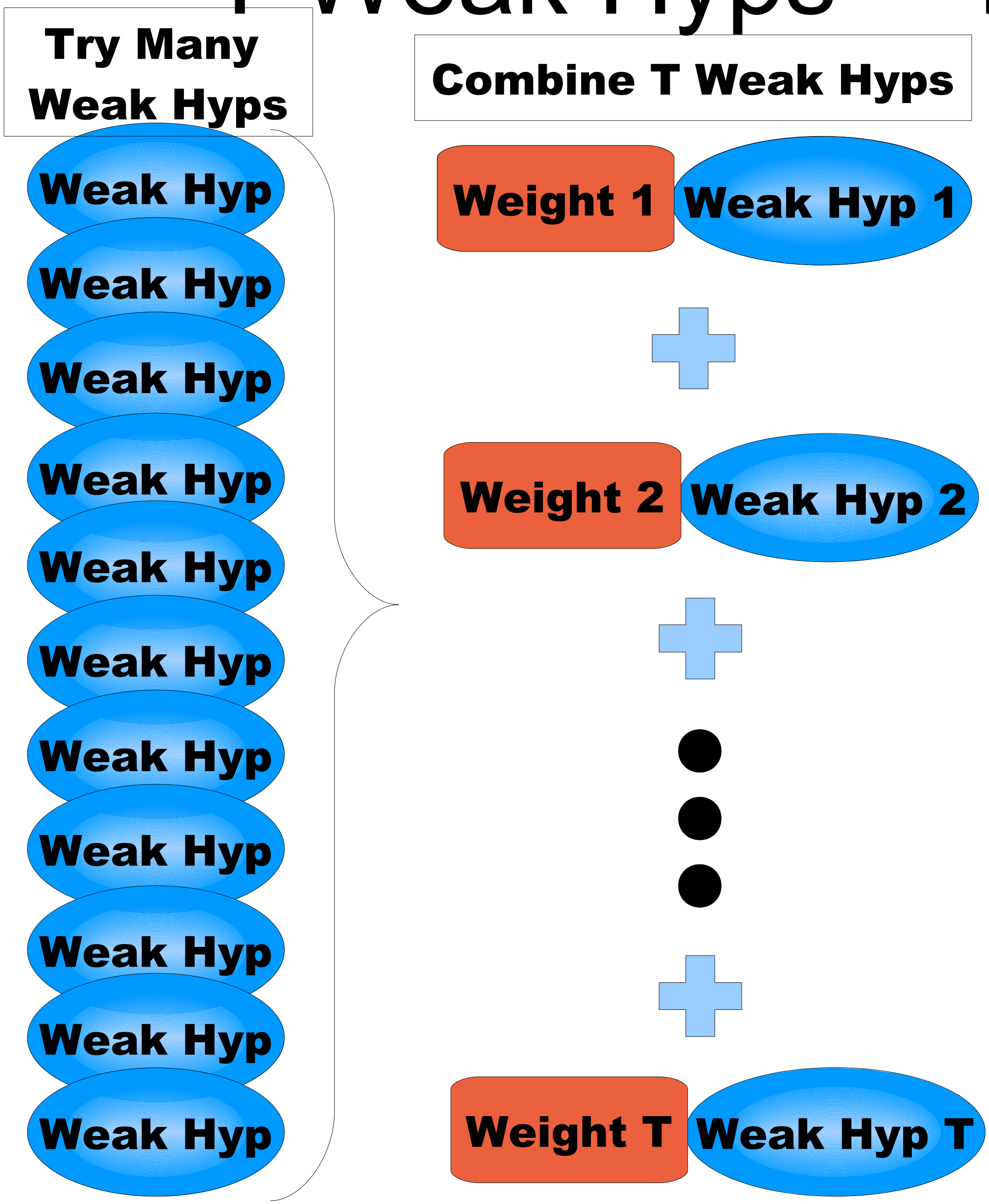
Brief History of Boosting

- 1984 - Framework developed by Valiant
  - Probably approximately correct (PAC)
- 1988 - Problem proposed by Michael Kearns
  - Machine learning class taught by Ron Rivest
- 1990 - Boosting problem solved (in theory)
  - Schapire, recursive majority gates of hypotheses
  - Freund, simple majority vote over hypotheses
- 1995 - Boosting problem solved (in practice)
  - Freund & Schapire, AdaBoost adapts to error of hypotheses

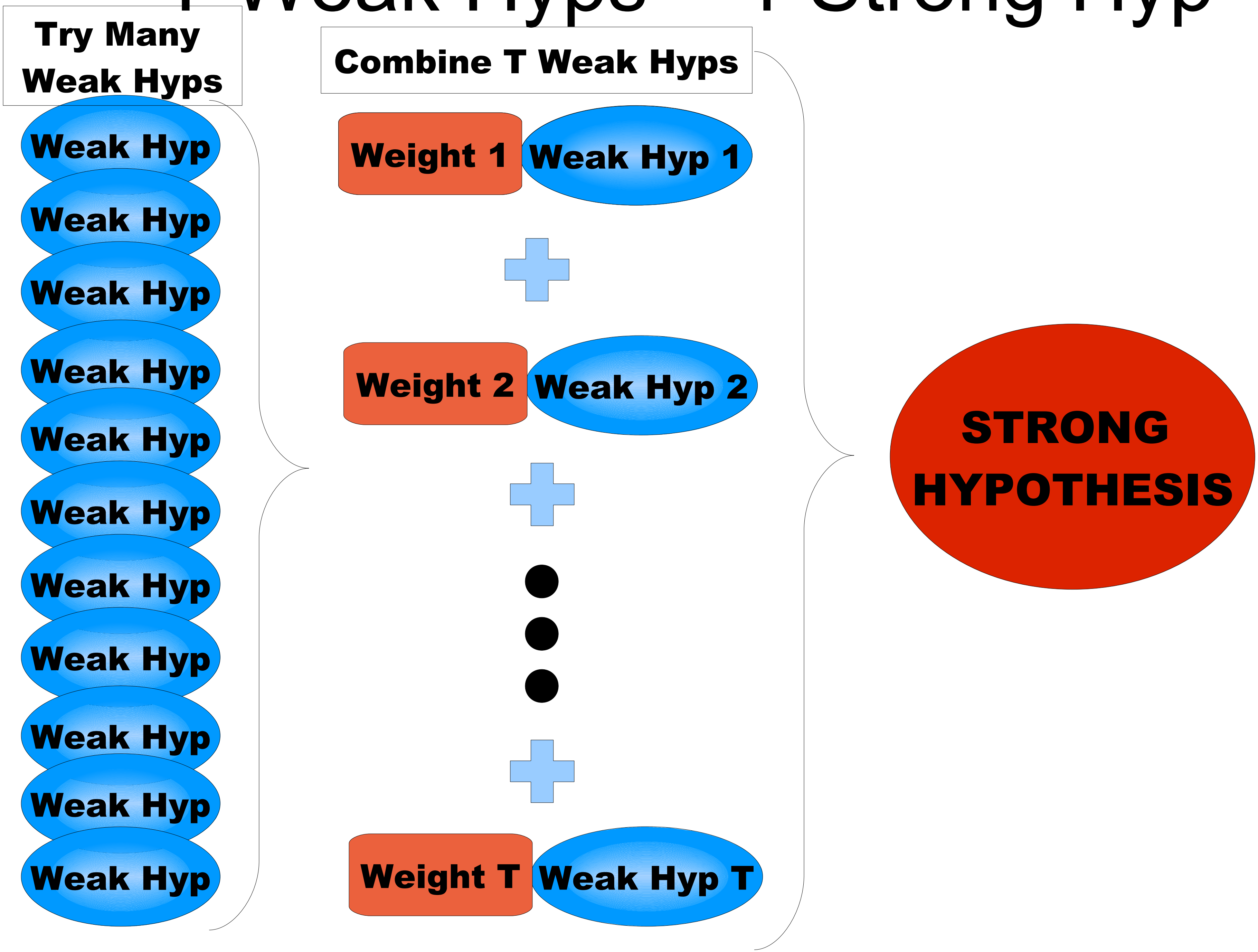
T Weak Hyps = 1 Strong Hyp



T Weak Hyps = 1 Strong Hyp

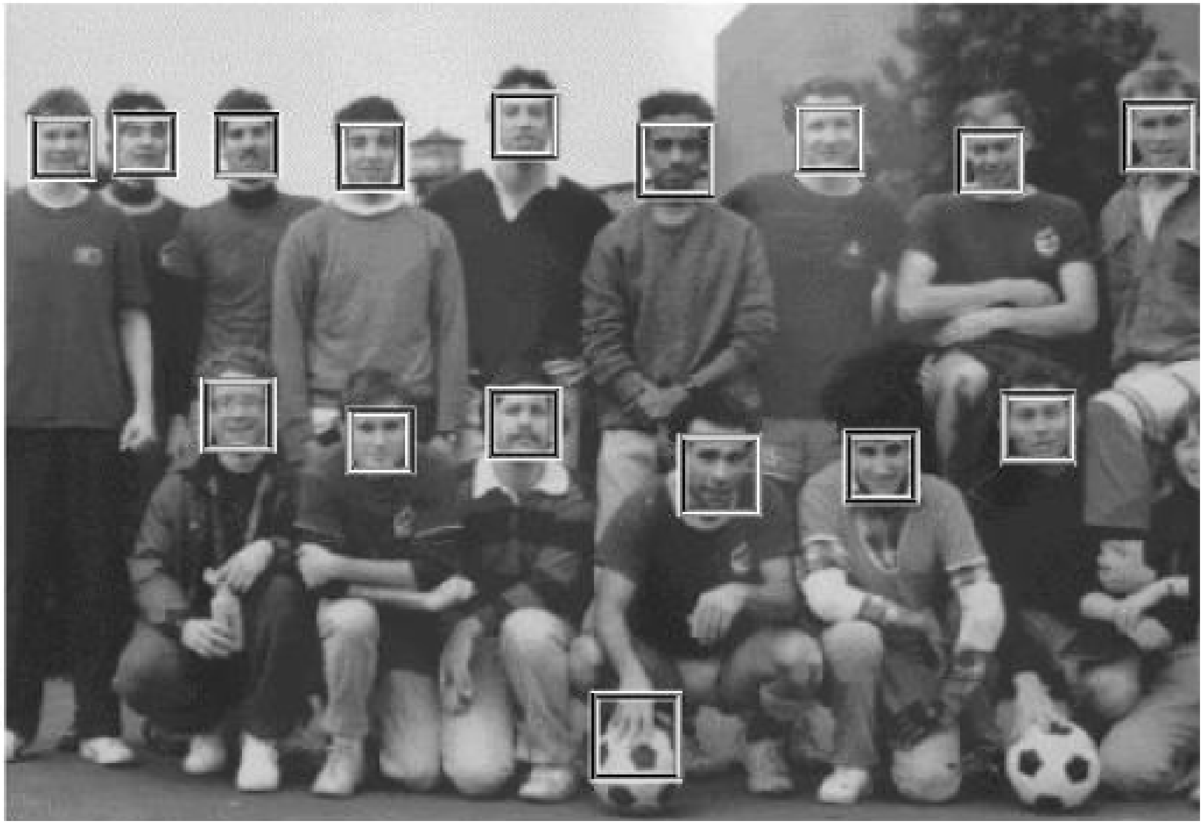


T Weak Hyps = 1 Strong Hyp



Example: Face Detection

- We are given a dataset of images
- We need to determine if there are faces in the images



Example: Face Detection

- Go through each possible rectangle
- Some weak hypotheses might be:
  - Is there a round object in the rectangle?
  - Does the rectangle have darker spots where the eyes should be?
  - Etc.
- Classifier = 2.1 * (Is Round) + 1.2 * (Has Eyes)
- Viola & Jones 2001 solved face detection problem in similar manner

Algorithms

- Many boosting algorithms have two sets of weights
  - Weights on all the training examples
  - Weights for each of the weak hypotheses used
- It is usually clear from context which set of weights is being discussed

Basic Boosting Algorithm

- Initial Conditions:
  - Training dataset $\{ (x_1 , y_1), \ldots (x_i , y_i) \ldots, (x_n , y_n) \}$
  - Each x is an example with a label y
- Learn a pattern
  - Use T weak hypotheses
  - Combine them in an “intelligent” manner
- See how well we learned the pattern
  - Did we just memorize training set?

An Iterative Learning Algorithm

Let w_i^t be the weight of example i on round t

$w_i^0 = 1/n$

For t = 1 to T:

1) Try many weak hyps, compute error $\Sigma_i\ w_i^t\ [[\ h(x_i) \neq y_i]]$

2) Pick the best hypothesis: h_t

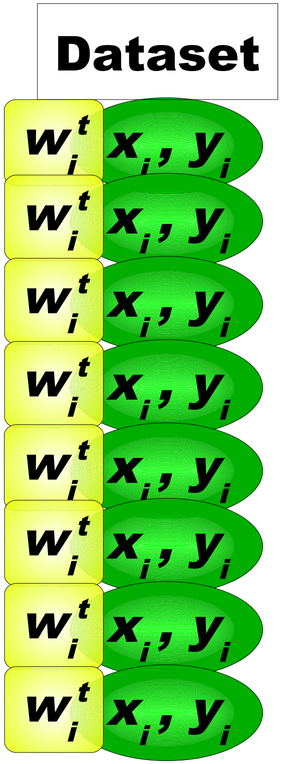
3) Give h_t a weight α_t

4) More weight to examples that h_t misclassified

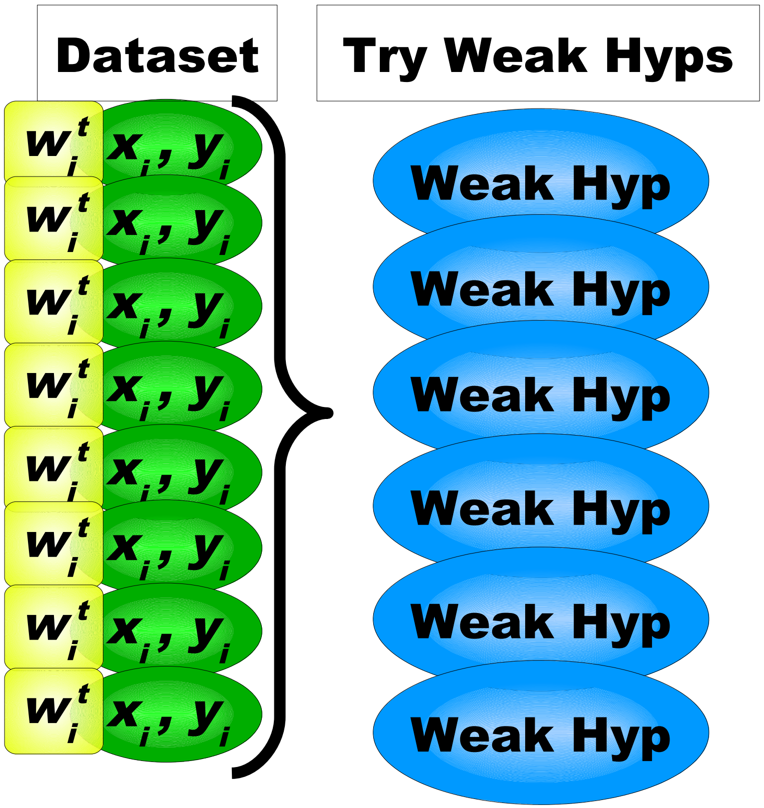
5) Less weight to examples that h_t classified correctly

Return a final hypothesis of $H_t(x) = \Sigma_t\ \alpha_t\ h_t(x)$

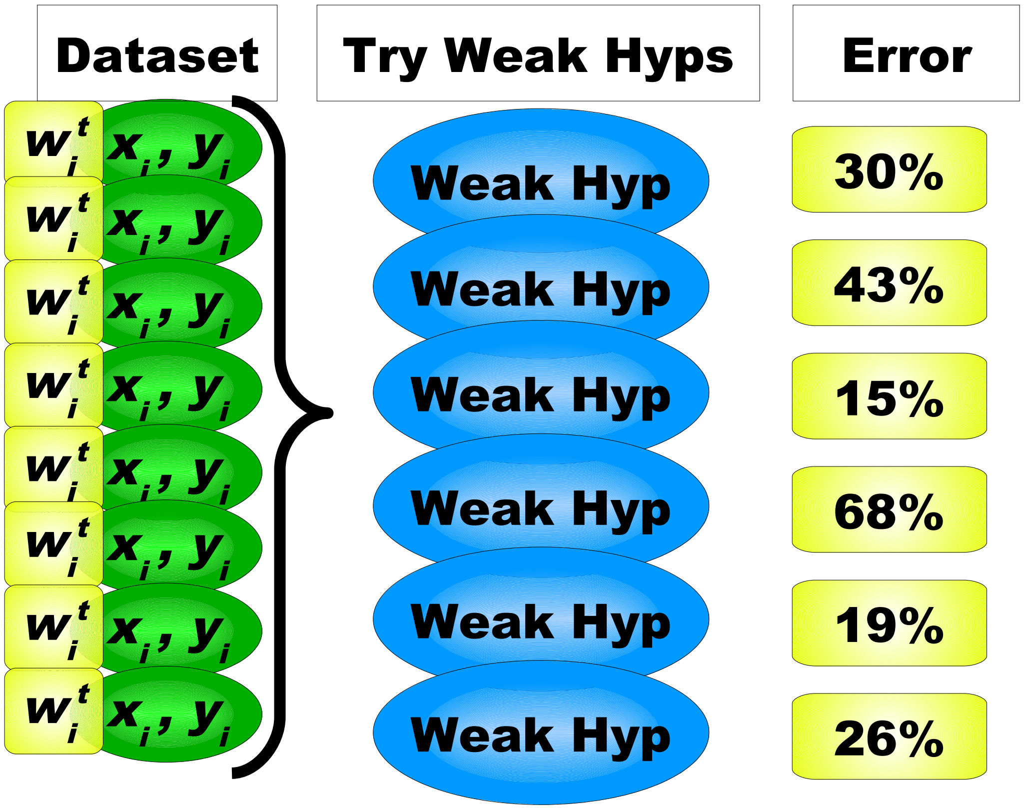
One Iteration



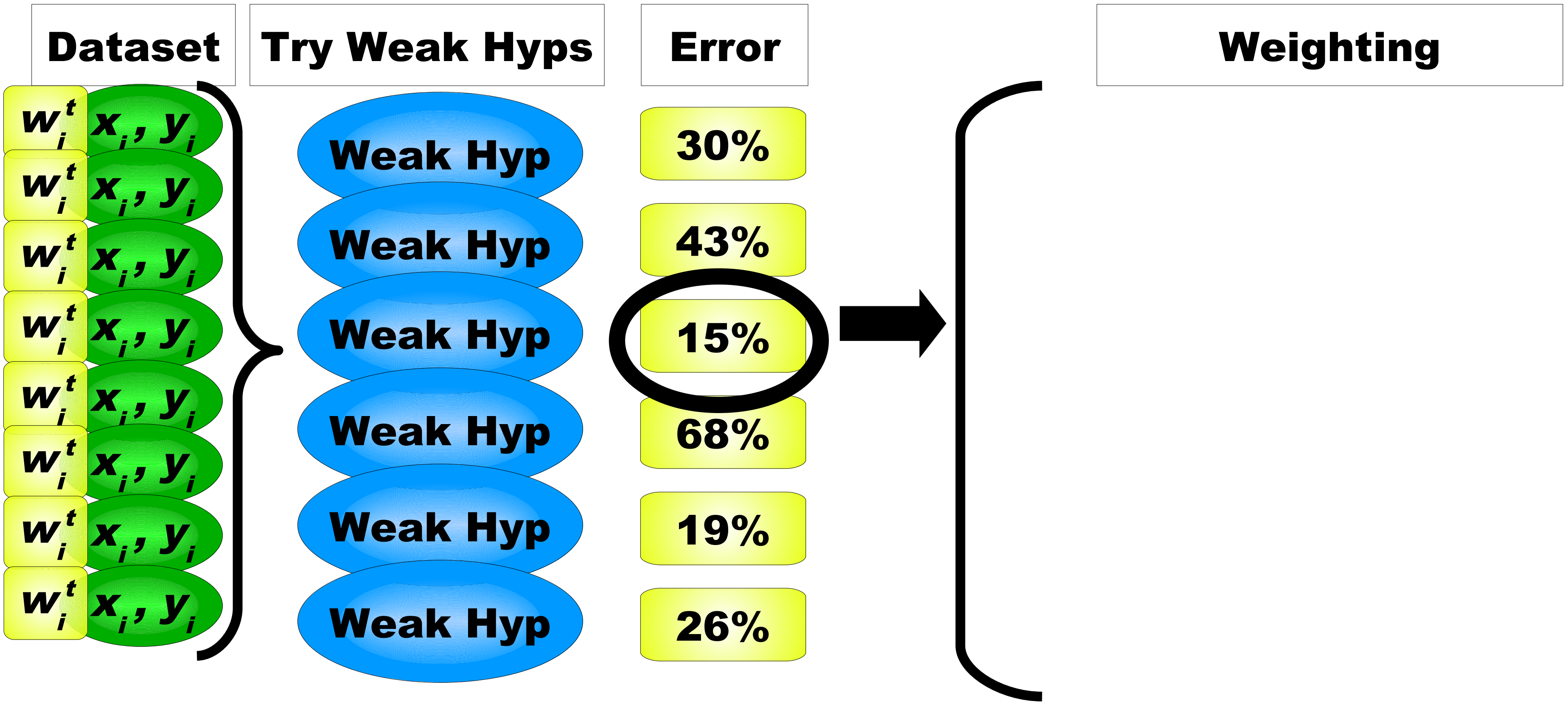
One Iteration



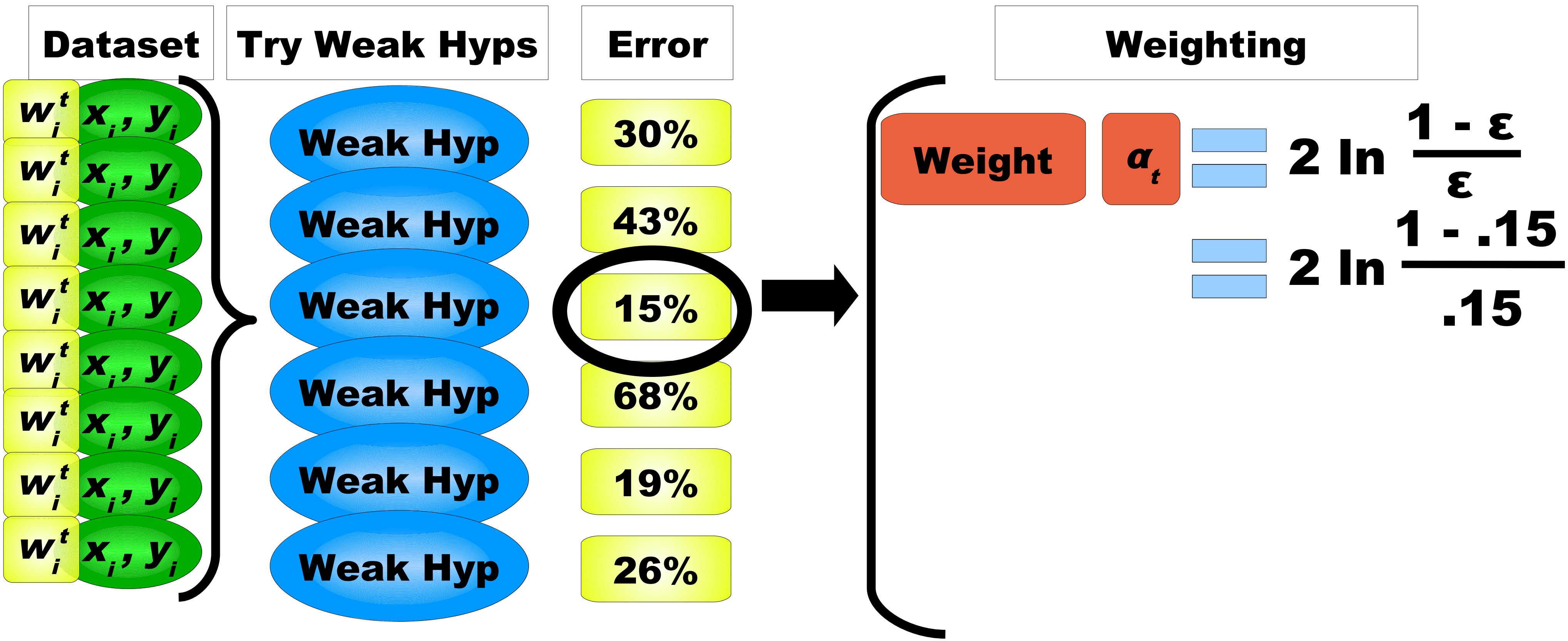
One Iteration



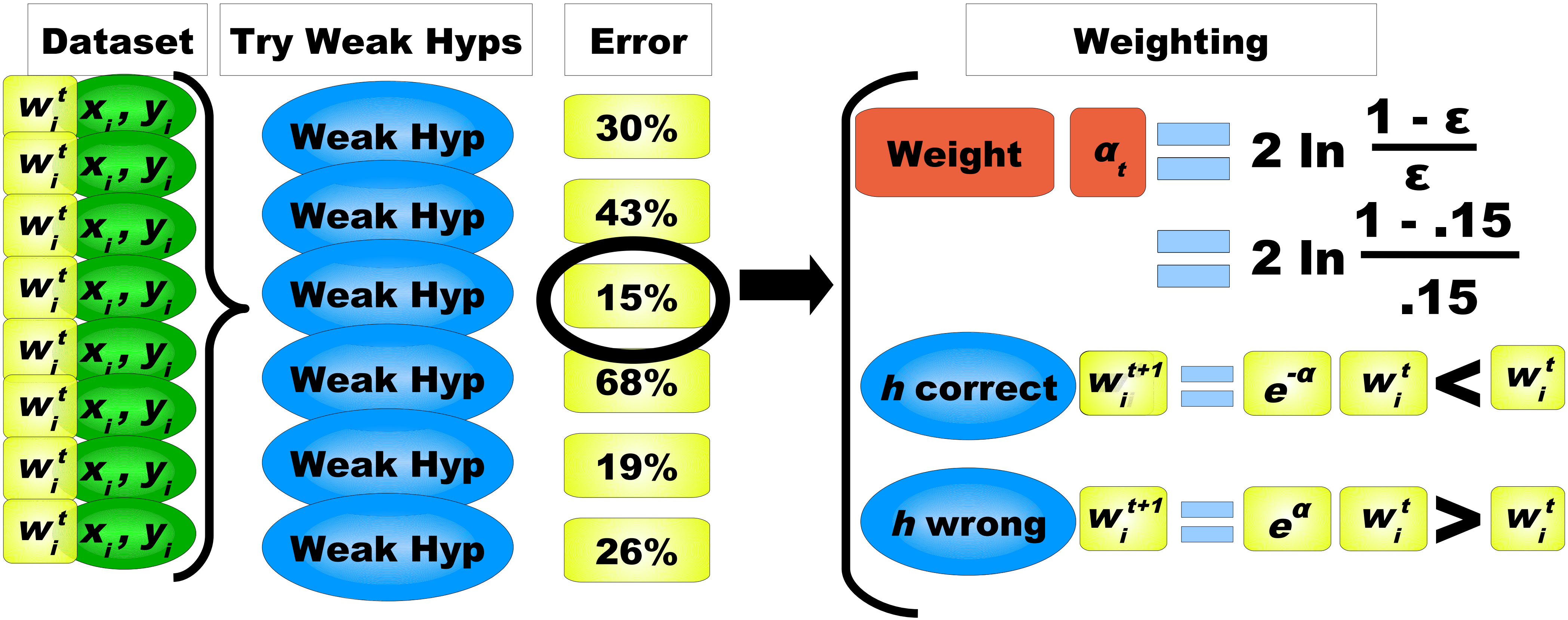
One Iteration



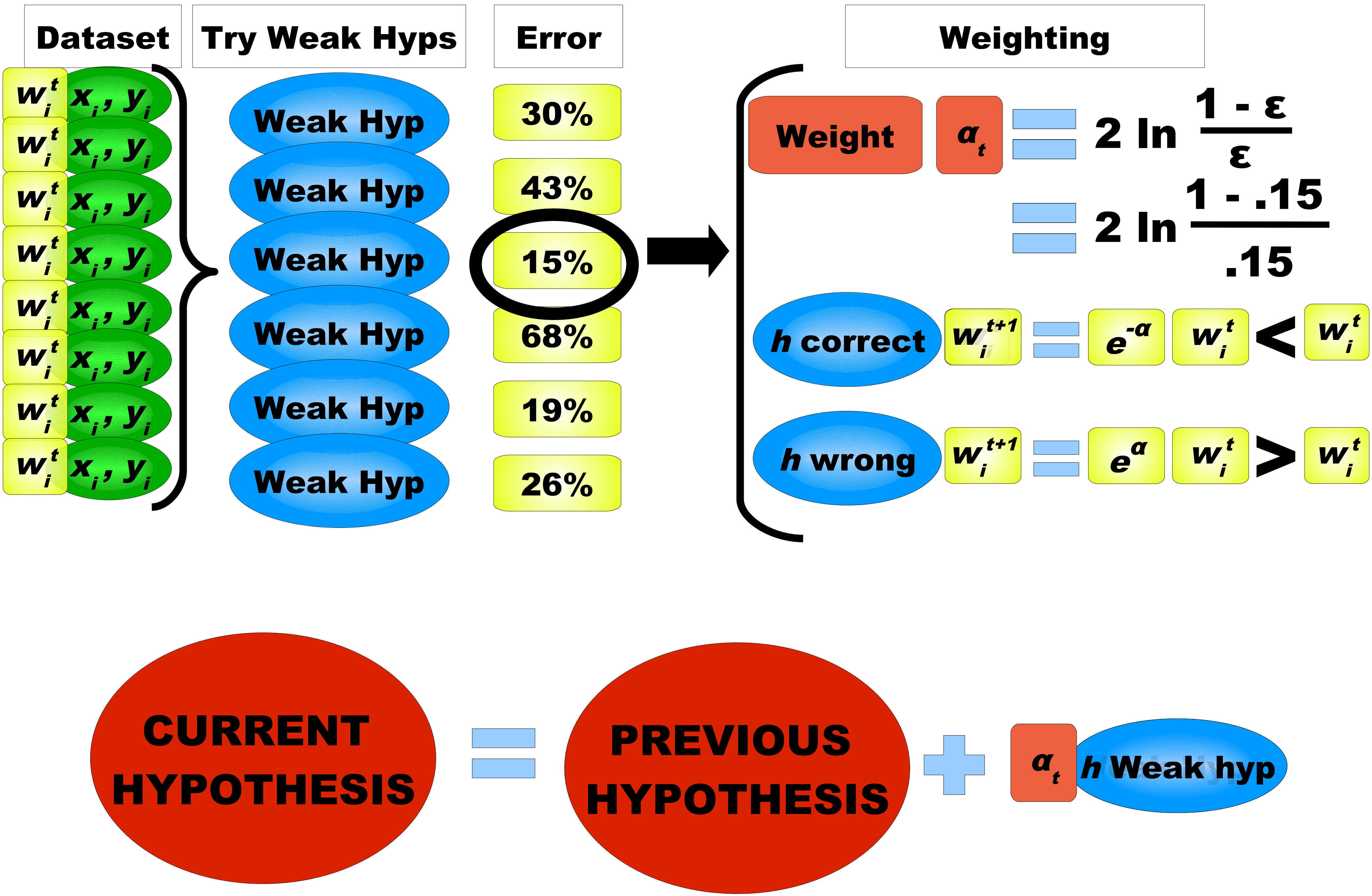
One Iteration



One Iteration

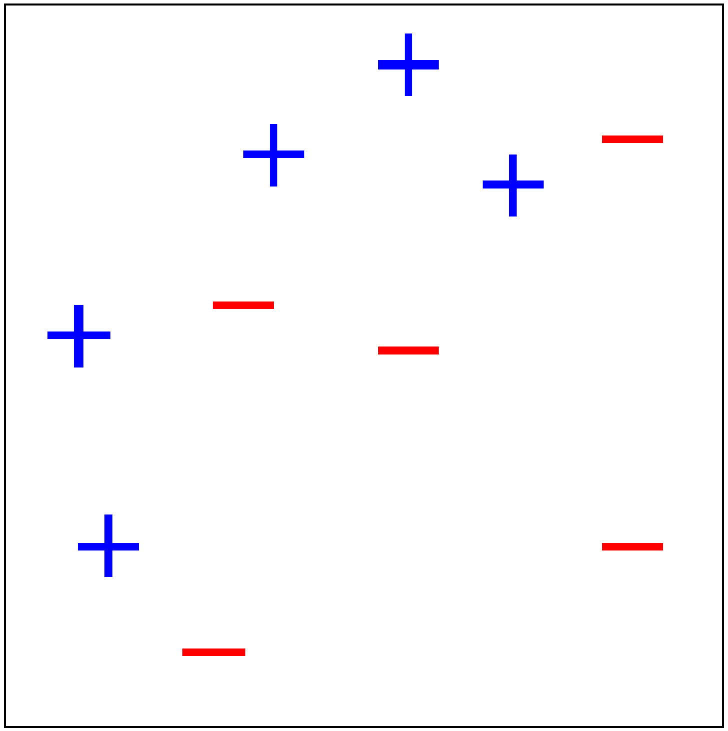


One Iteration



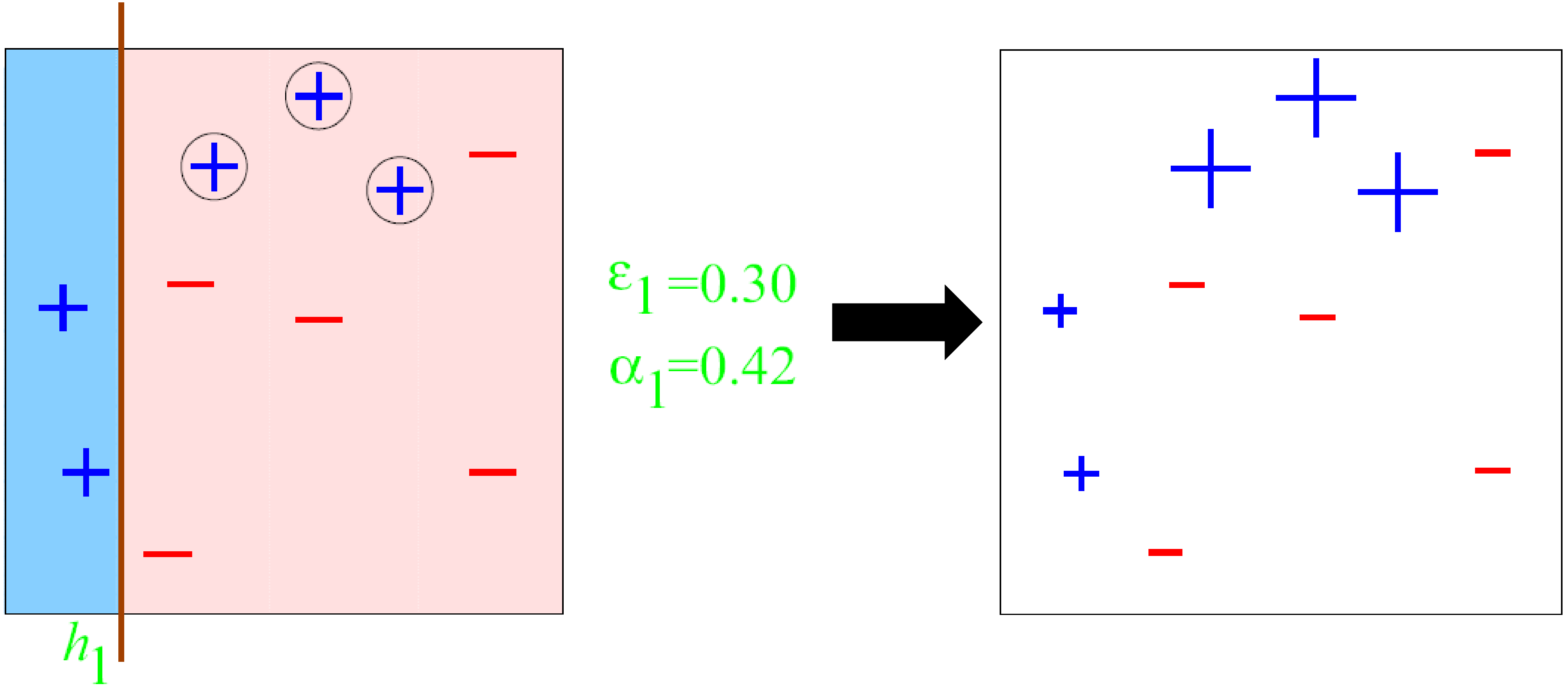
Toy Example

- Positive examples
- Negative examples
- 2-Dimensional plane
- Weak hyps: linear separators
- 3 iterations



Taken from Freund 1996

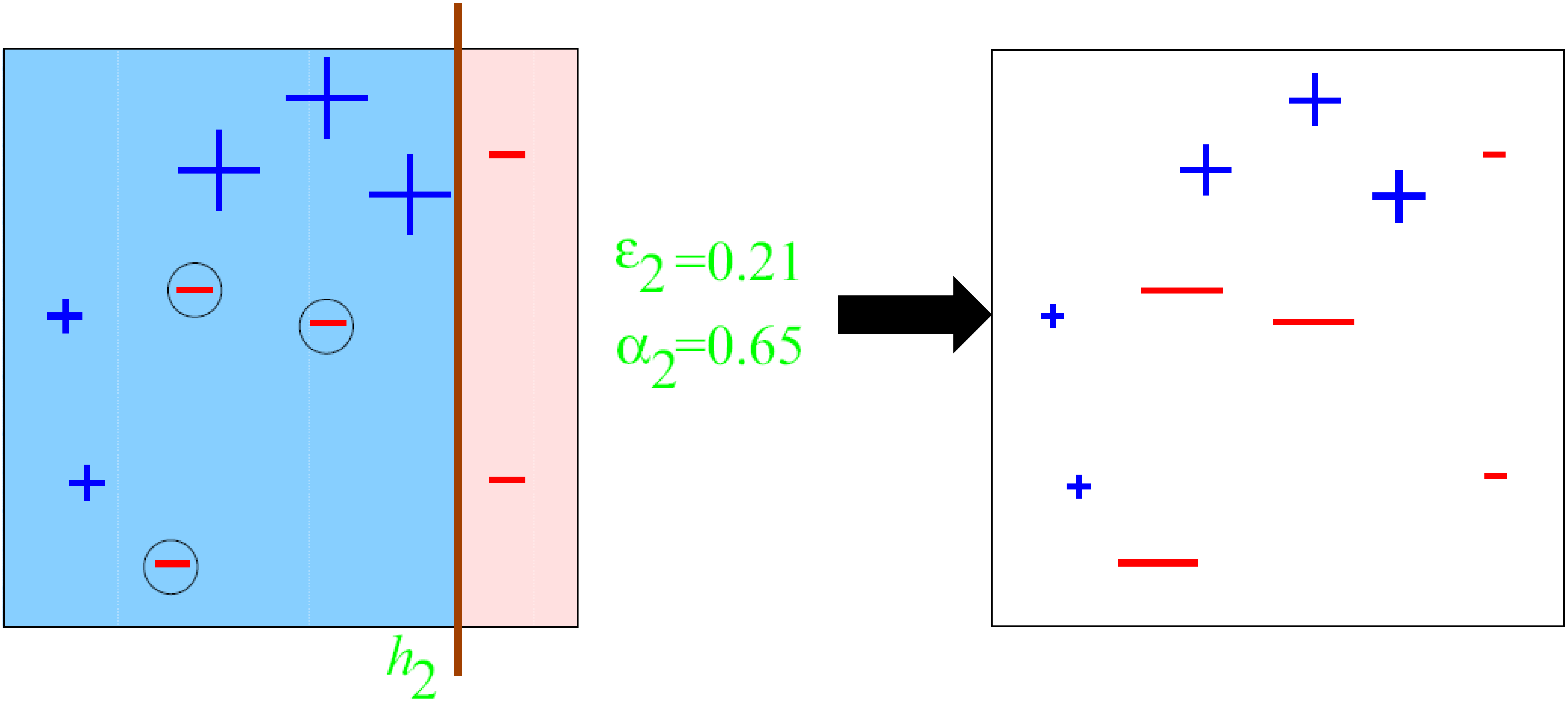
Toy Example: Iteration 1



Misclassified examples are circled, given more weight

Taken from Freund 1996

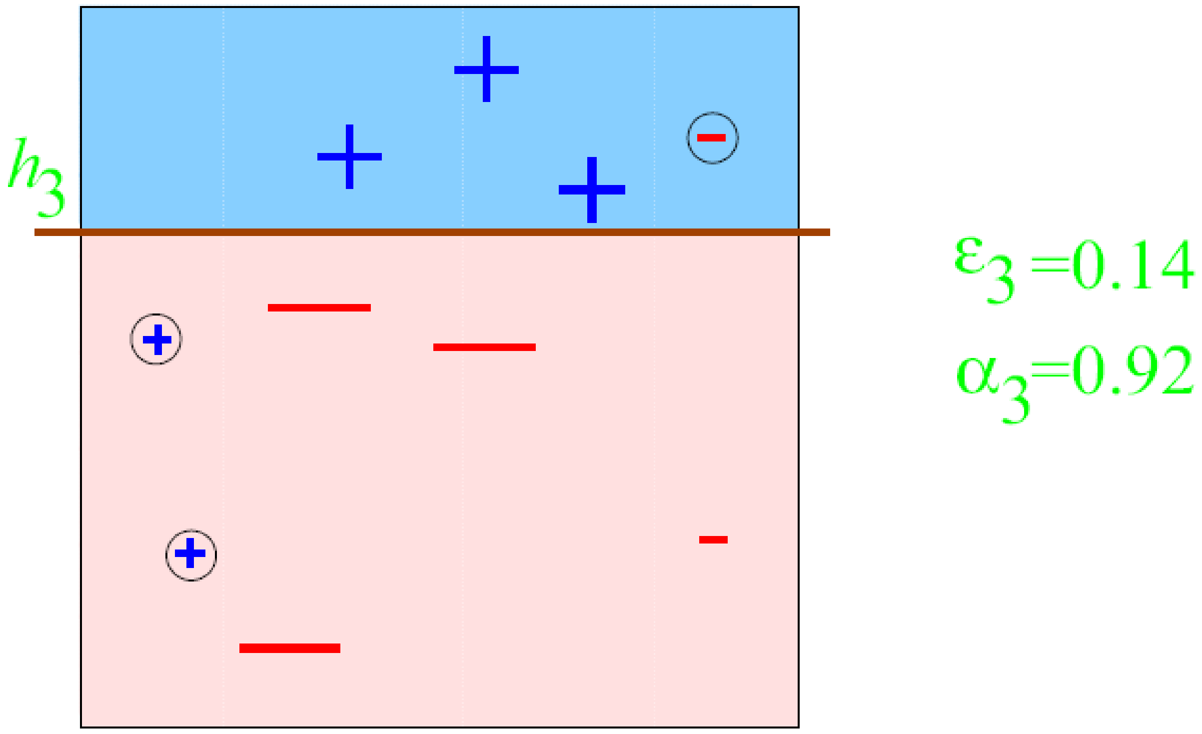
Toy Example: Iteration 2



Misclassified examples are circled, given more weight

Taken from Freund 1996

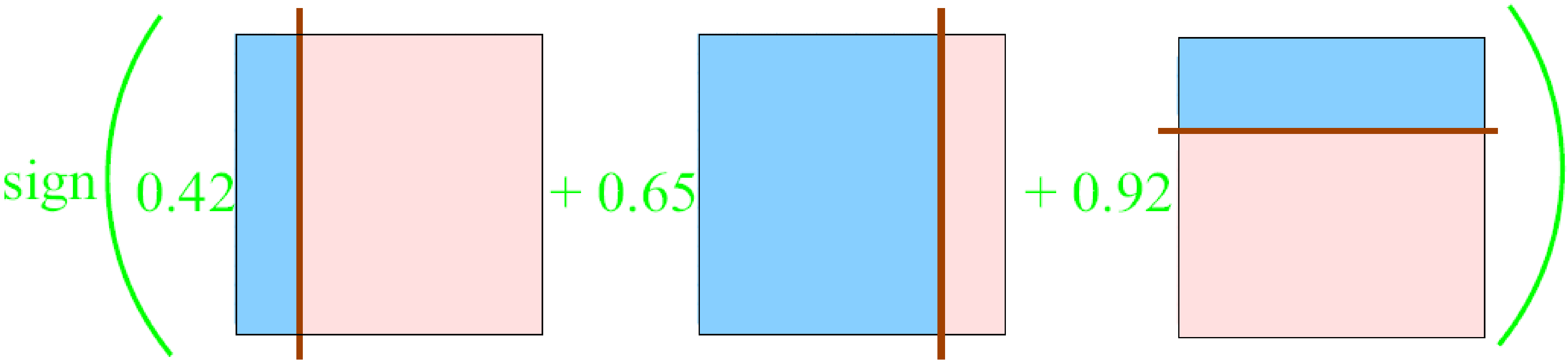
Toy Example: Iteration 3

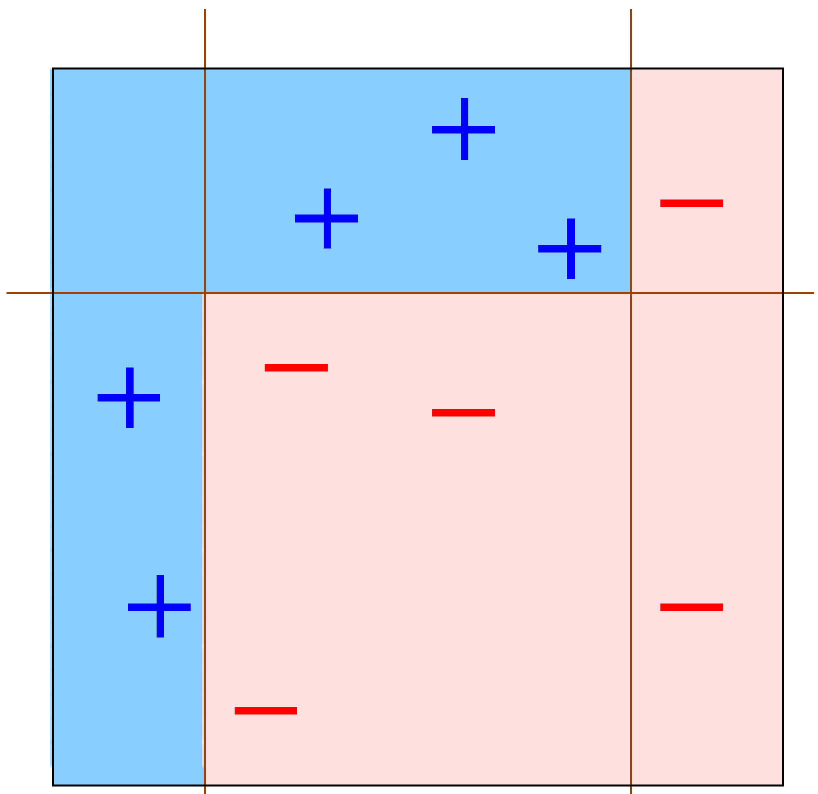


Finished boosting

Taken from Freund 1996

Toy Example: Final Classifier





Taken from Freund 1996

Questions

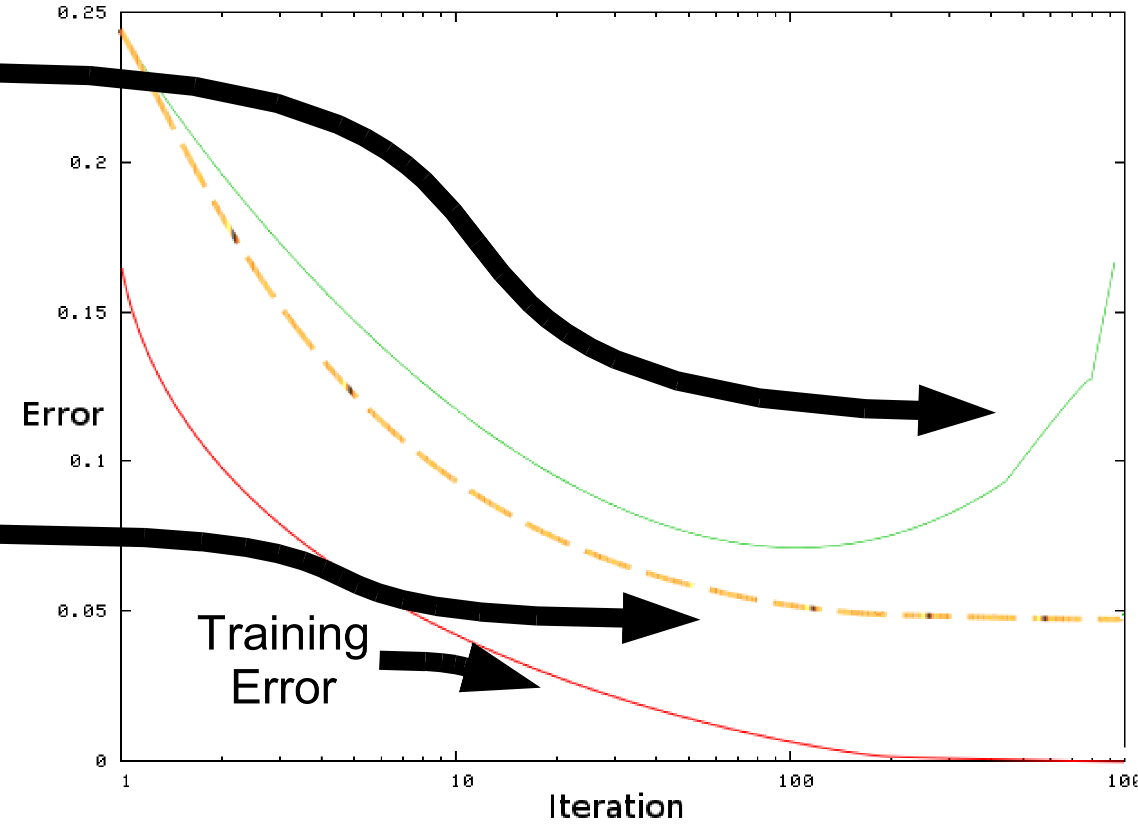
- How should we weight the hypotheses?
- How should we weight the examples?
- How should we choose the “best” hypothesis?
- How should we add the new (this iteration) hypothesis to the set of old hypotheses
- Should we consider old hypotheses when adding new ones?

Answers

- There are many answers to these questions
- Freund & Schapire 1997 – AdaBoost
- Schapire & Singer 1999 – Confidence rated AdaBoost
- Freund 1995, 2000 – Noise resistant via binomial weights
- Friedman *et al* 1998 and Collins *et al* 2000 – Connections to logistic regression and Bregman divergences
- Warmuth et al 2006 – “Totally corrective” boosting
- Freund & Arvey 2008 – Asymmetric cost, boosting the normalized margin

What's the big deal?

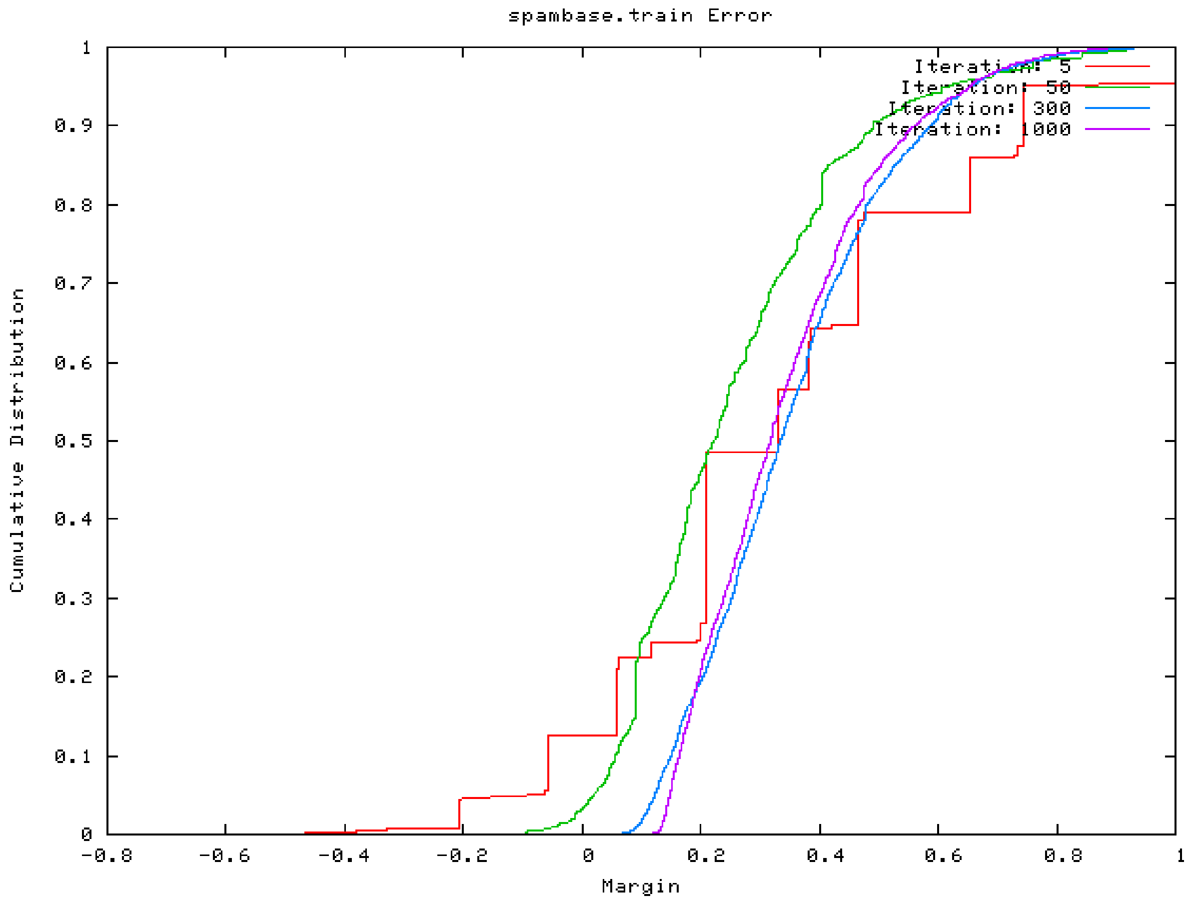
- Most algorithms start to memorize the data instead of learning patterns
- Most test error curves
    - Train decreases
    - Test starts to increase
    - Increase in test is due to "overfitting"
- Boosting continues to learn
    - Test error plateaus
- Explanation: margin



What's the big deal?

- One goal in machine learning is “margin”
  - “Margin” is a measure of how correct an example is
  - If all hypotheses get an example right, we'll probably get a similar example right in the future
  - If 1 out of 1000 hypotheses get an example right, then we'll probably get it wrong in the future
  - Boosting gives us a good margin

Margin Plot



- Margin frequently converges to some cumulative distribution function (CDF)
- Rudin et al. show that CDF may <u>*not*</u> always converge

End Boosting Section

Start Final Classifier Section

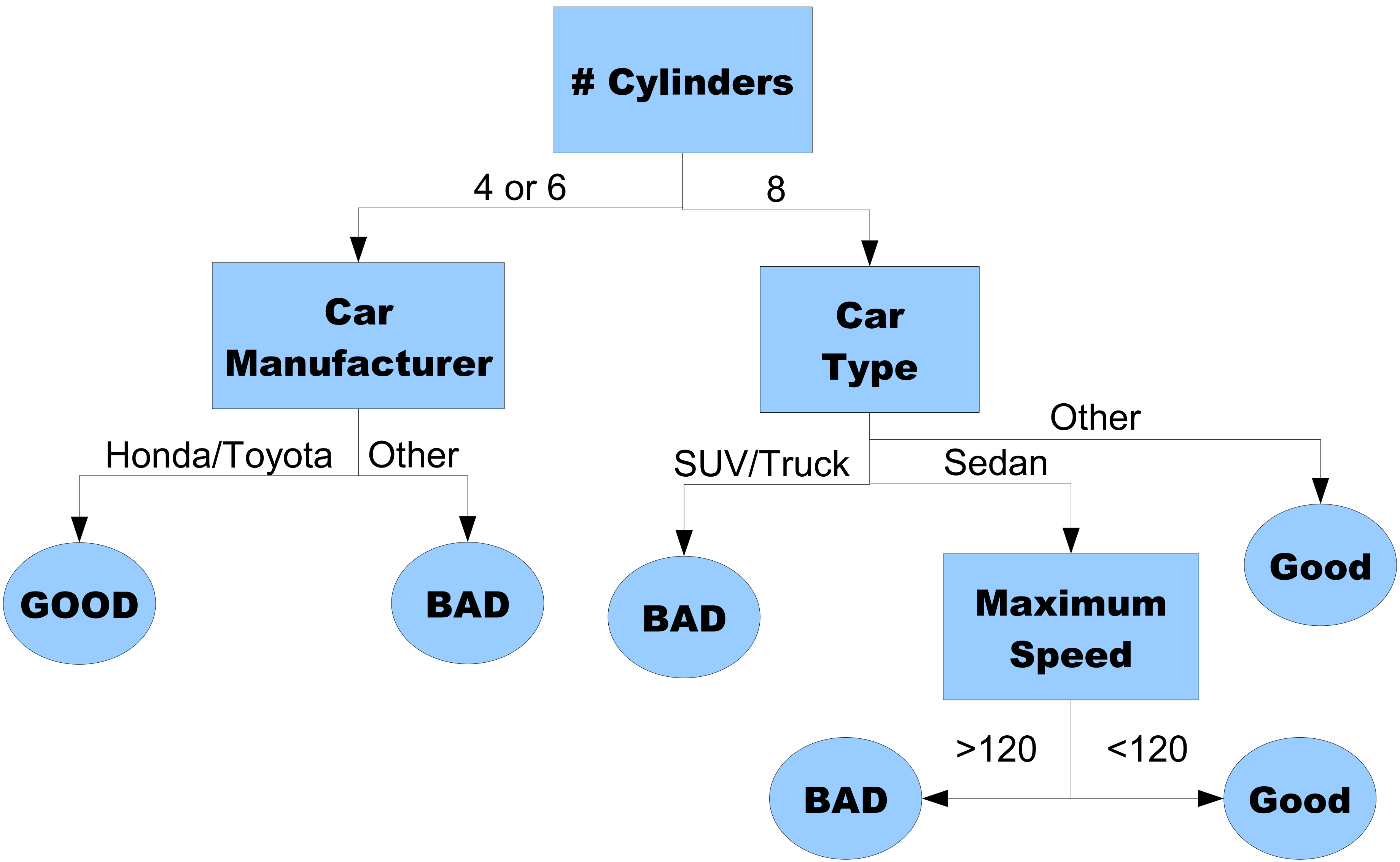
Final Classifier: Combination of Weak Hypotheses

- Original usage of boosting was just adding many weak hypotheses
- Adding weak hyps could be improved
  - Some of the weak hypotheses may be correlated
  - If there are a lot of weak hypotheses, the decision can be very hard to visualize
- Why can't boosting be more like decision trees
  - Easy to understand and visualize
  - A classic approach used by many fields

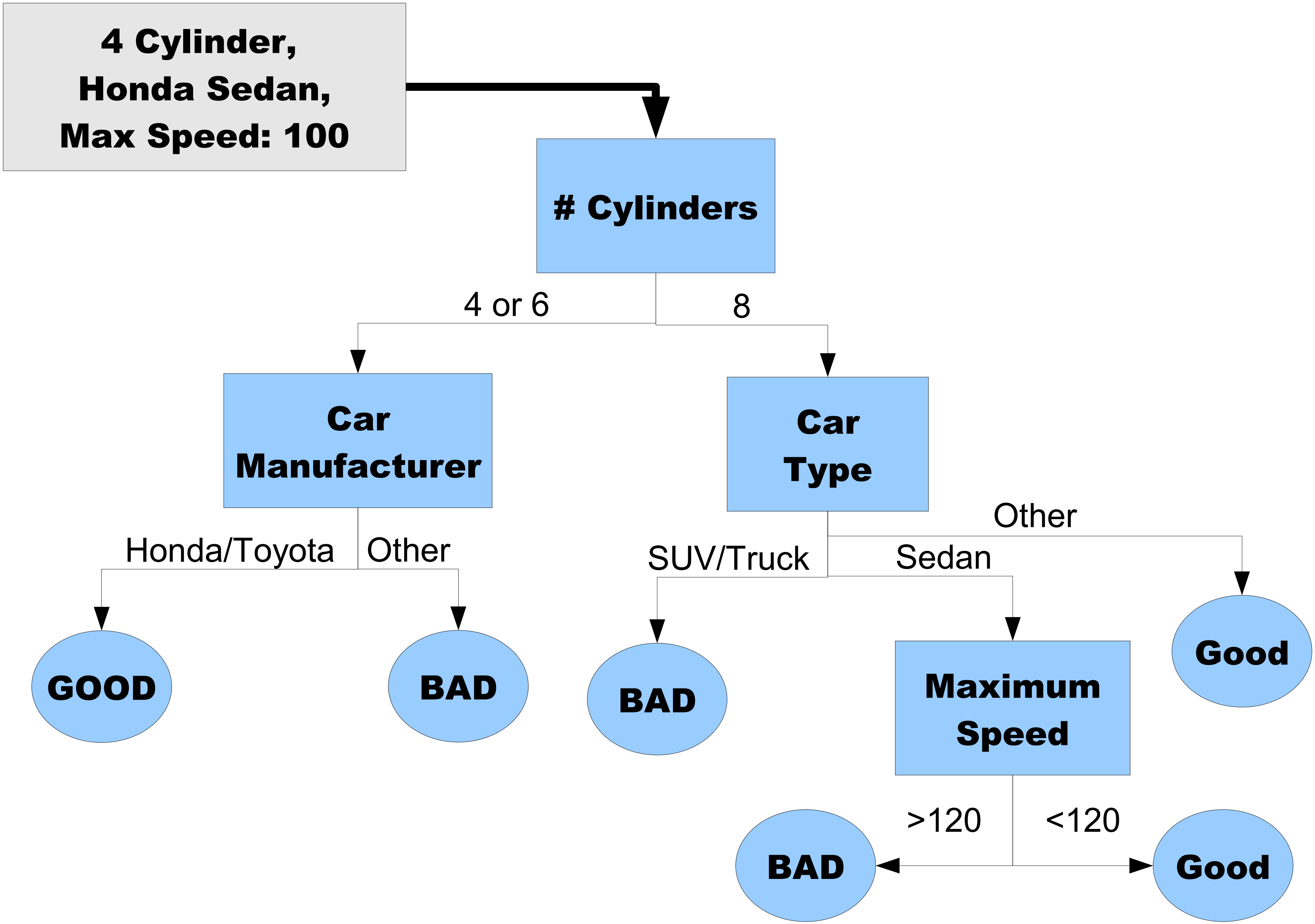
Final Classifier: Decision Trees

- Follow a series of questions to a single answer
- Does the car have 4 or 8 cylinders?
  - If #cylinders=4 or 8, then was the car made in Asia?
    - If Yes then you get good gas mileage
    - If no then you get bad gas mileage
  - If #cylinders=3,5,6, or 7 then poor gas mileage

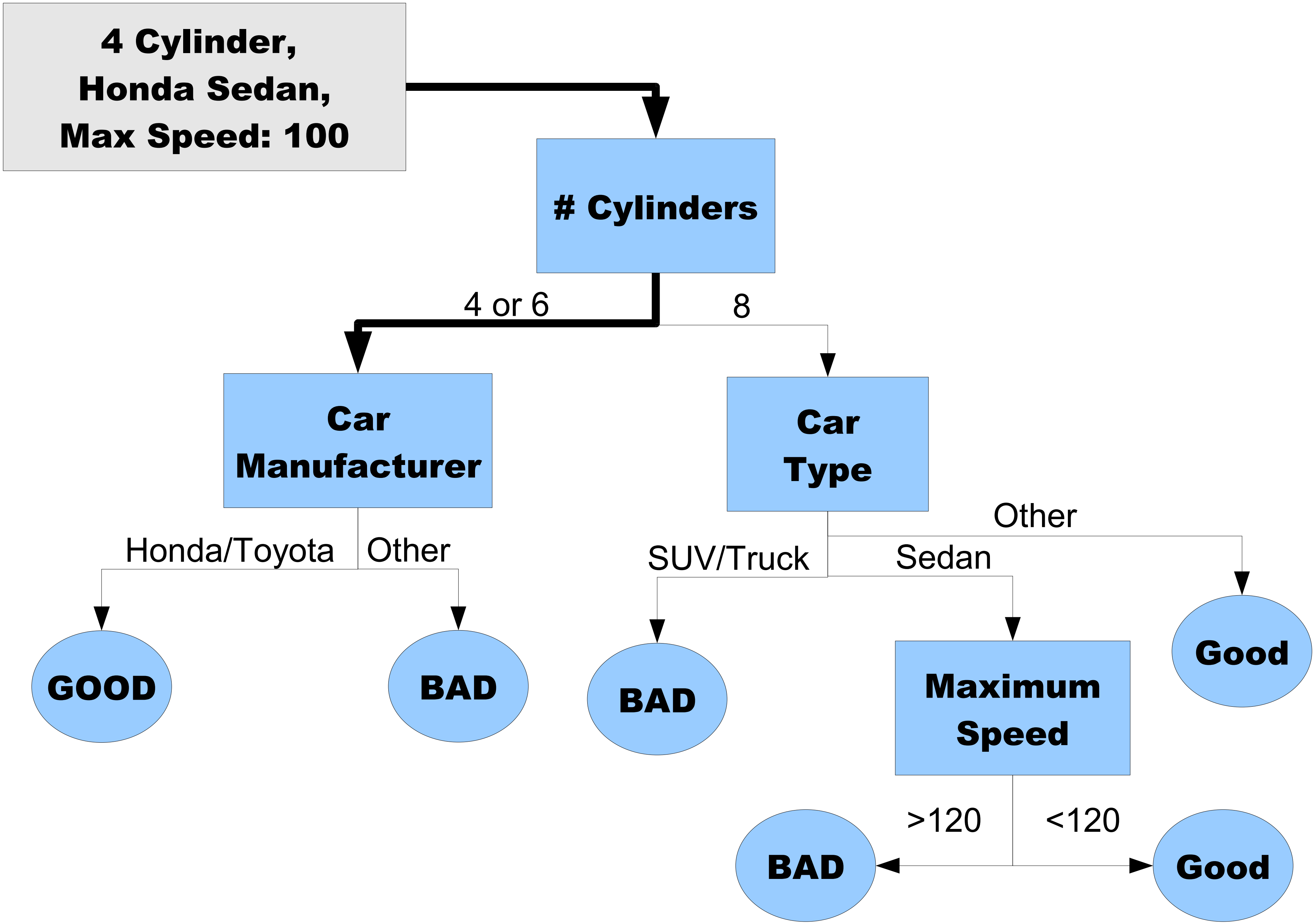
Decision Tree



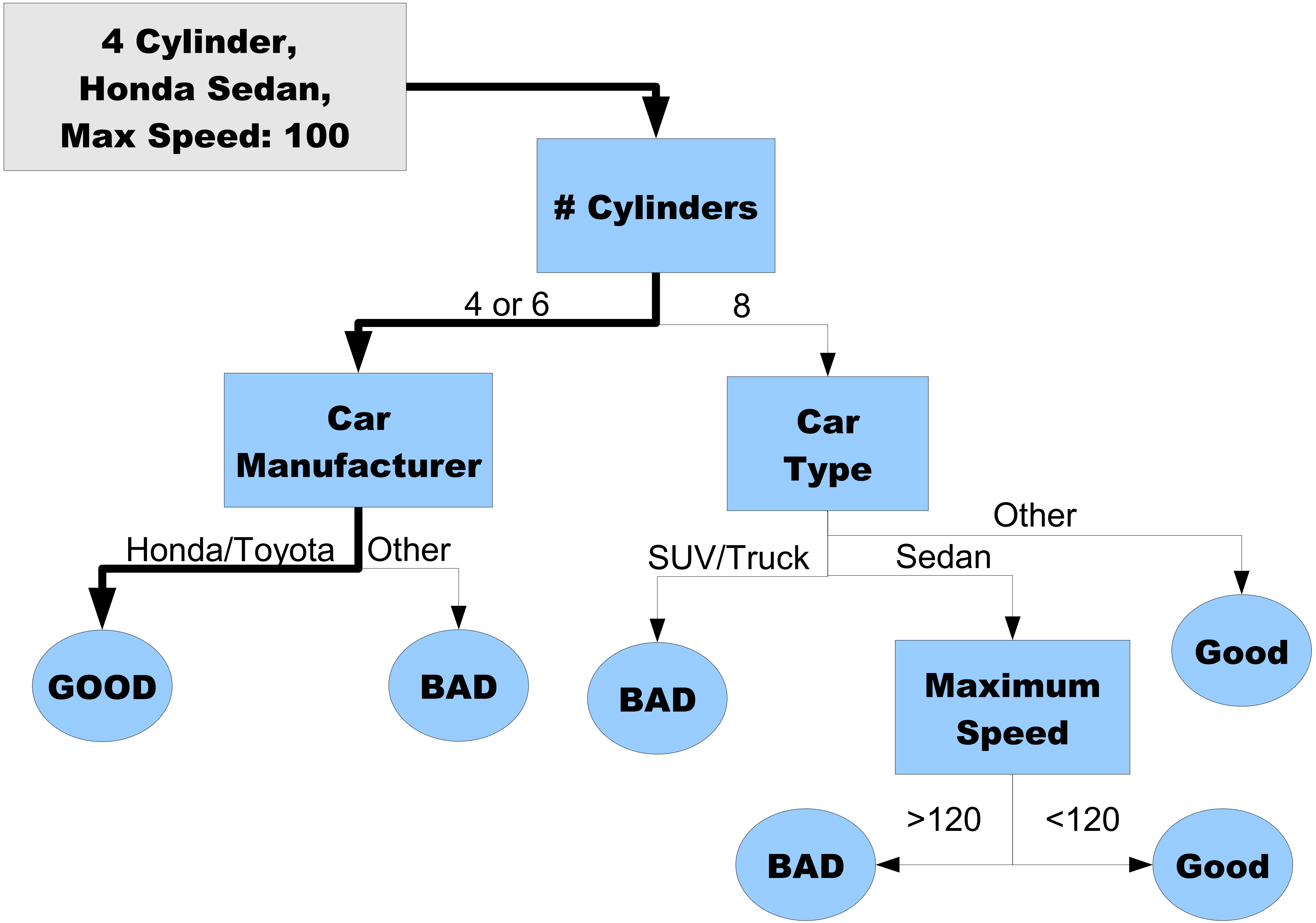
Decision Tree



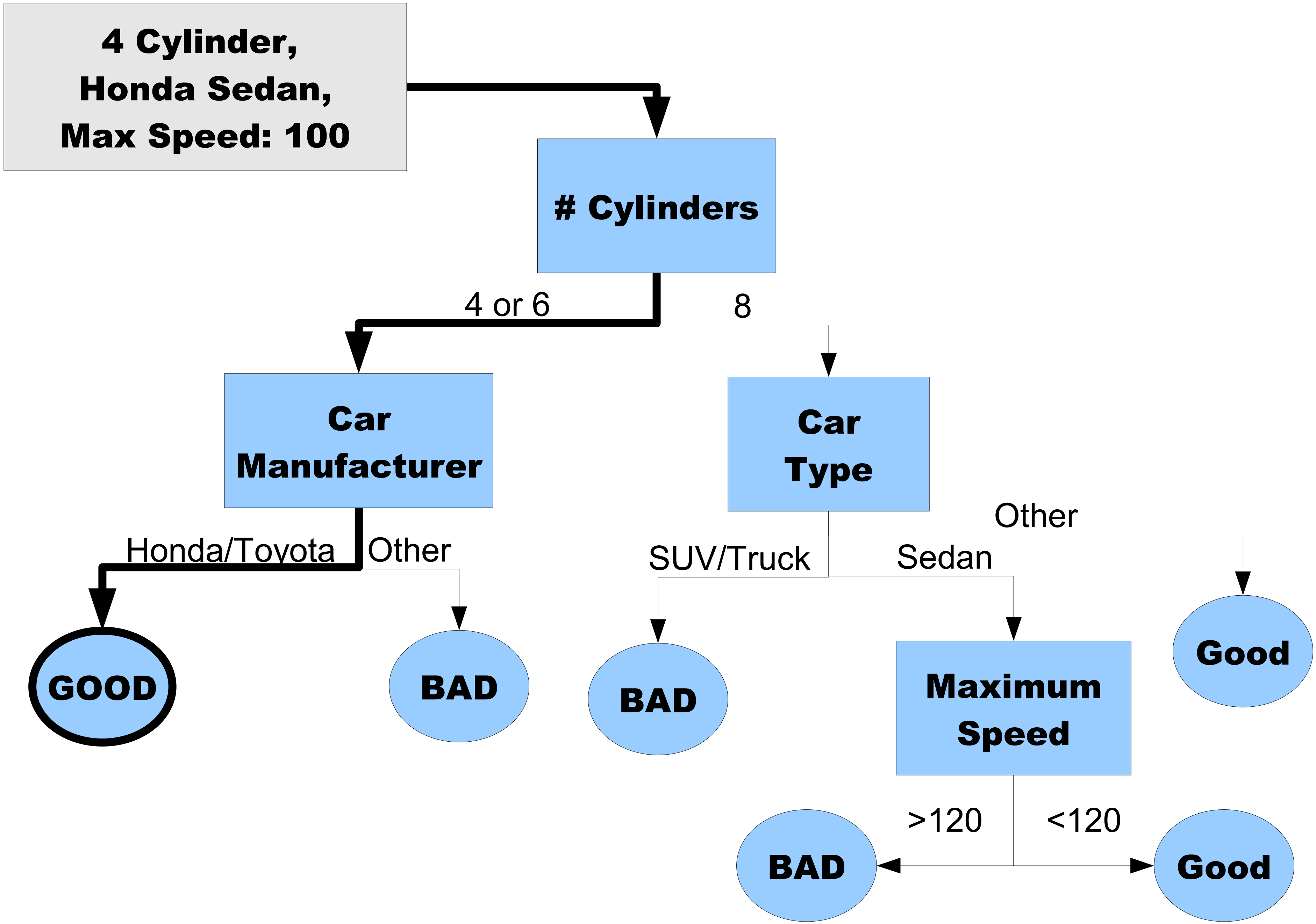
Decision Tree



Decision Tree

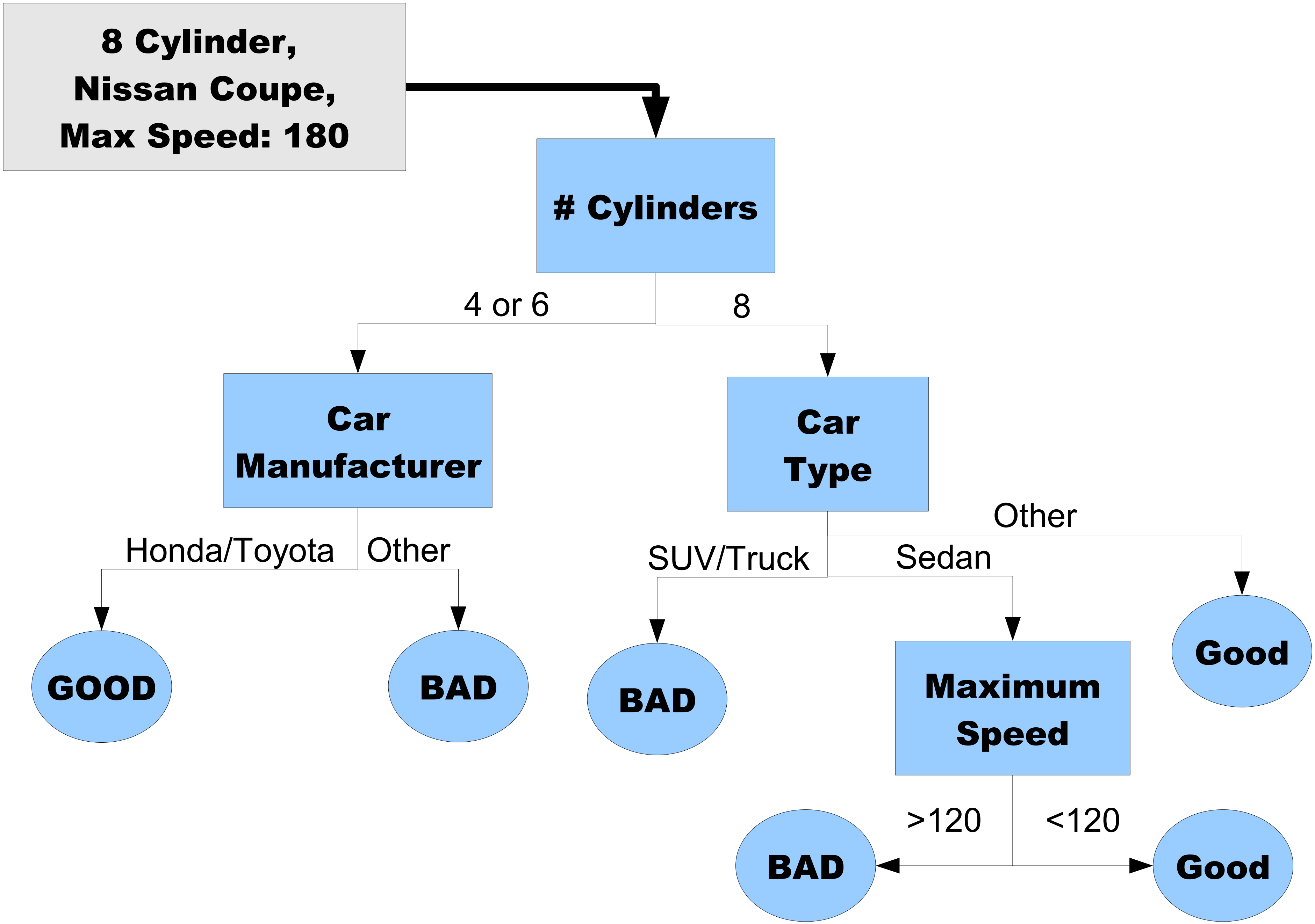


Decision Tree

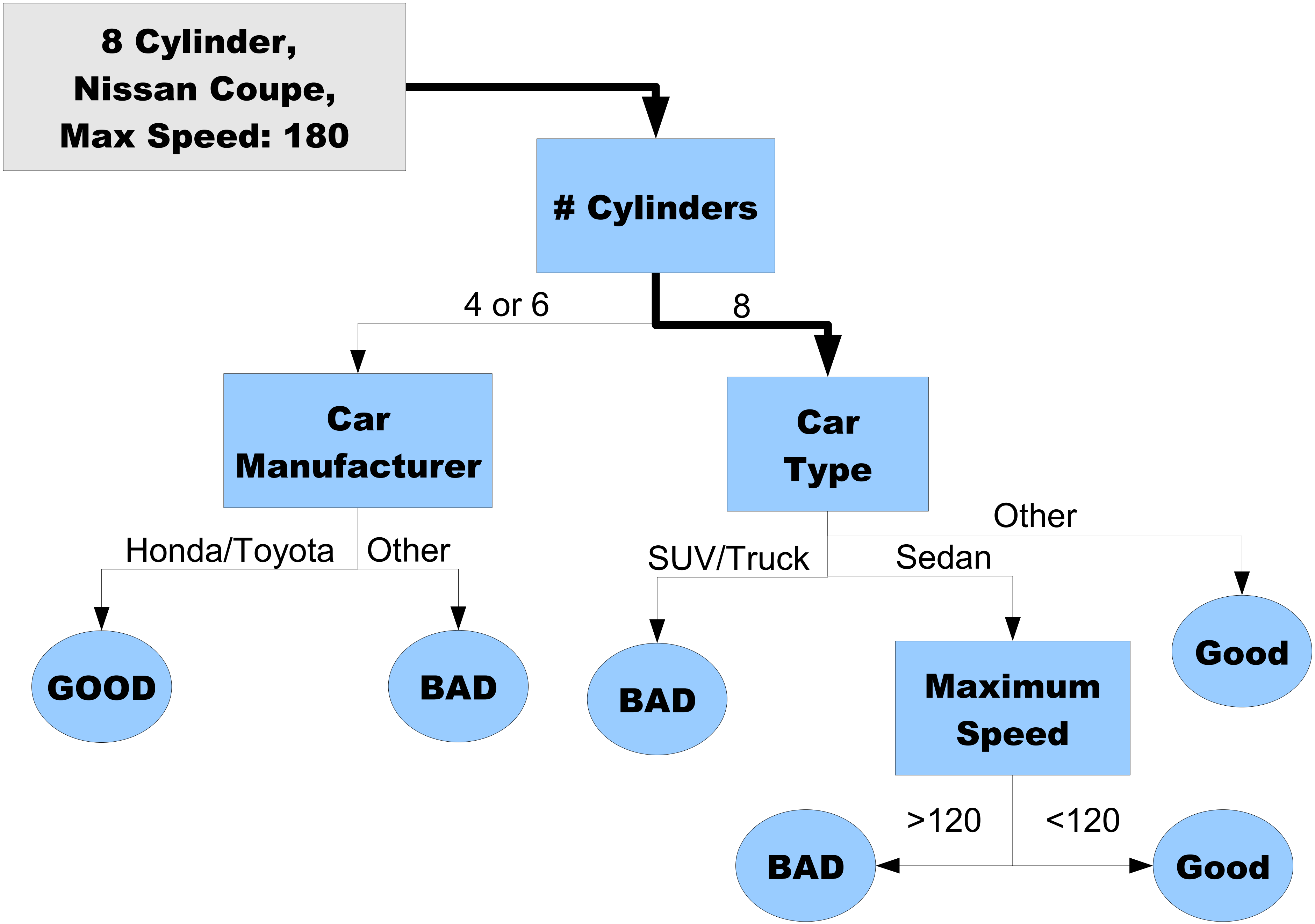


Another Example

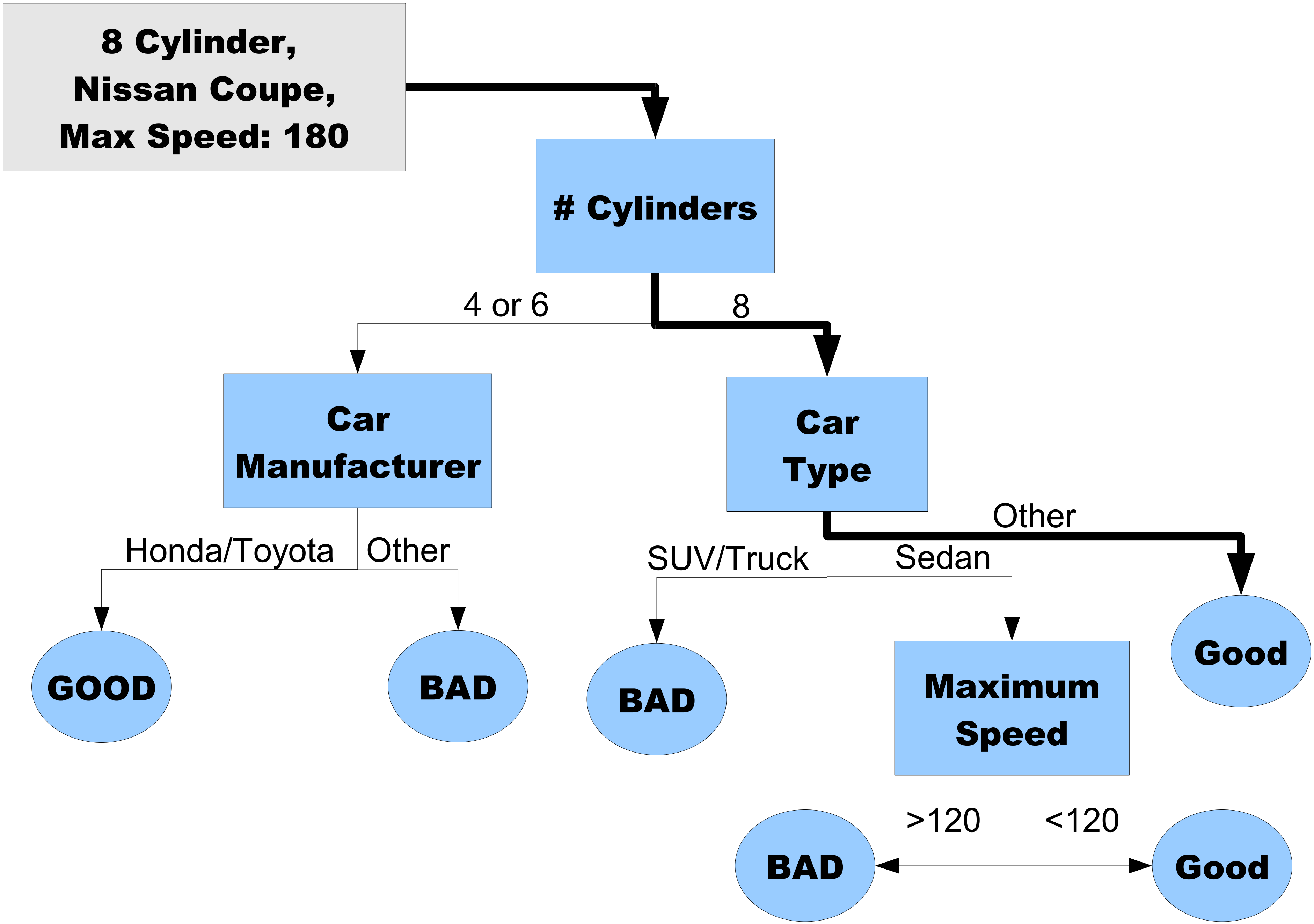
Decision Tree



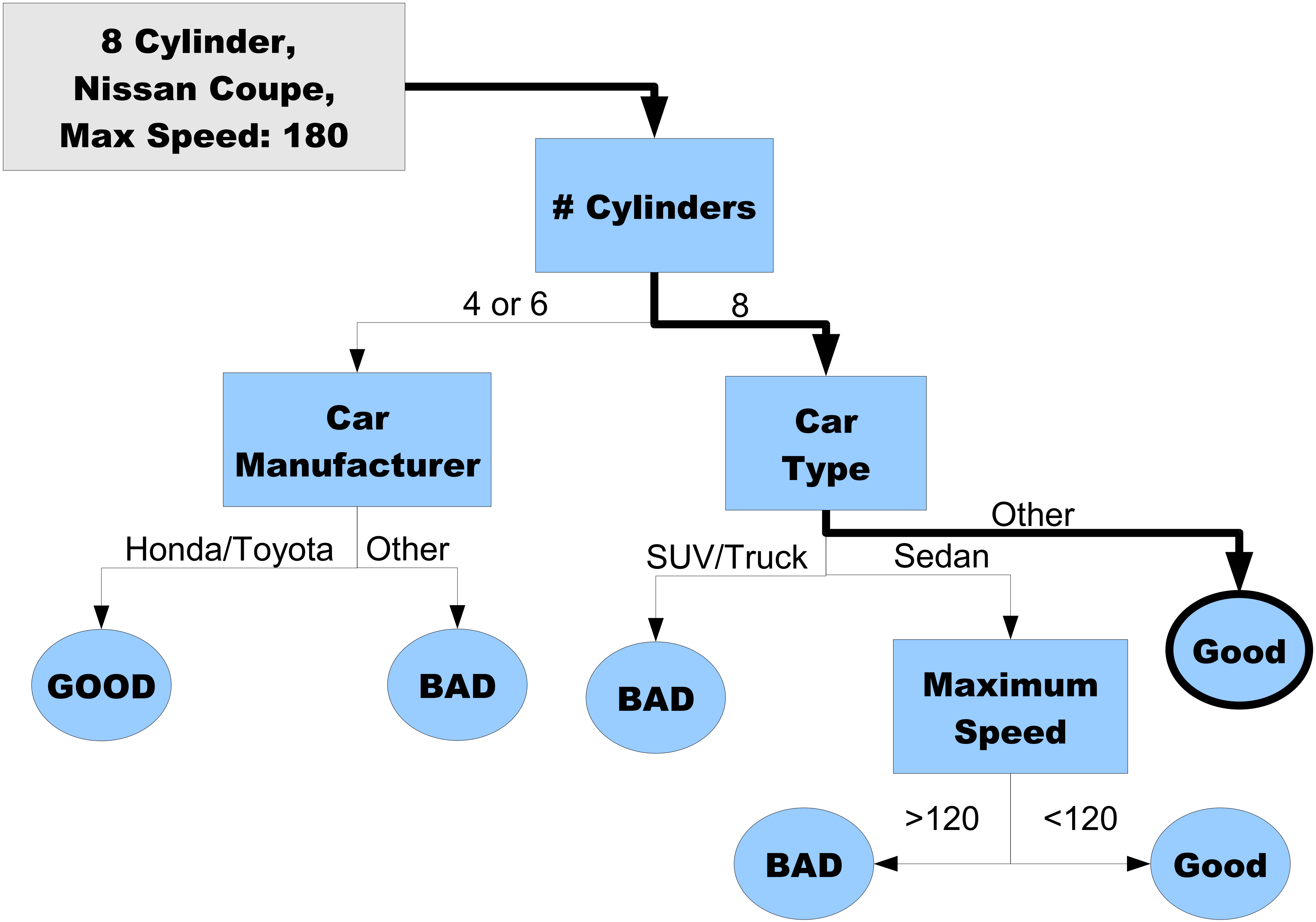
Decision Tree



Decision Tree



Decision Tree



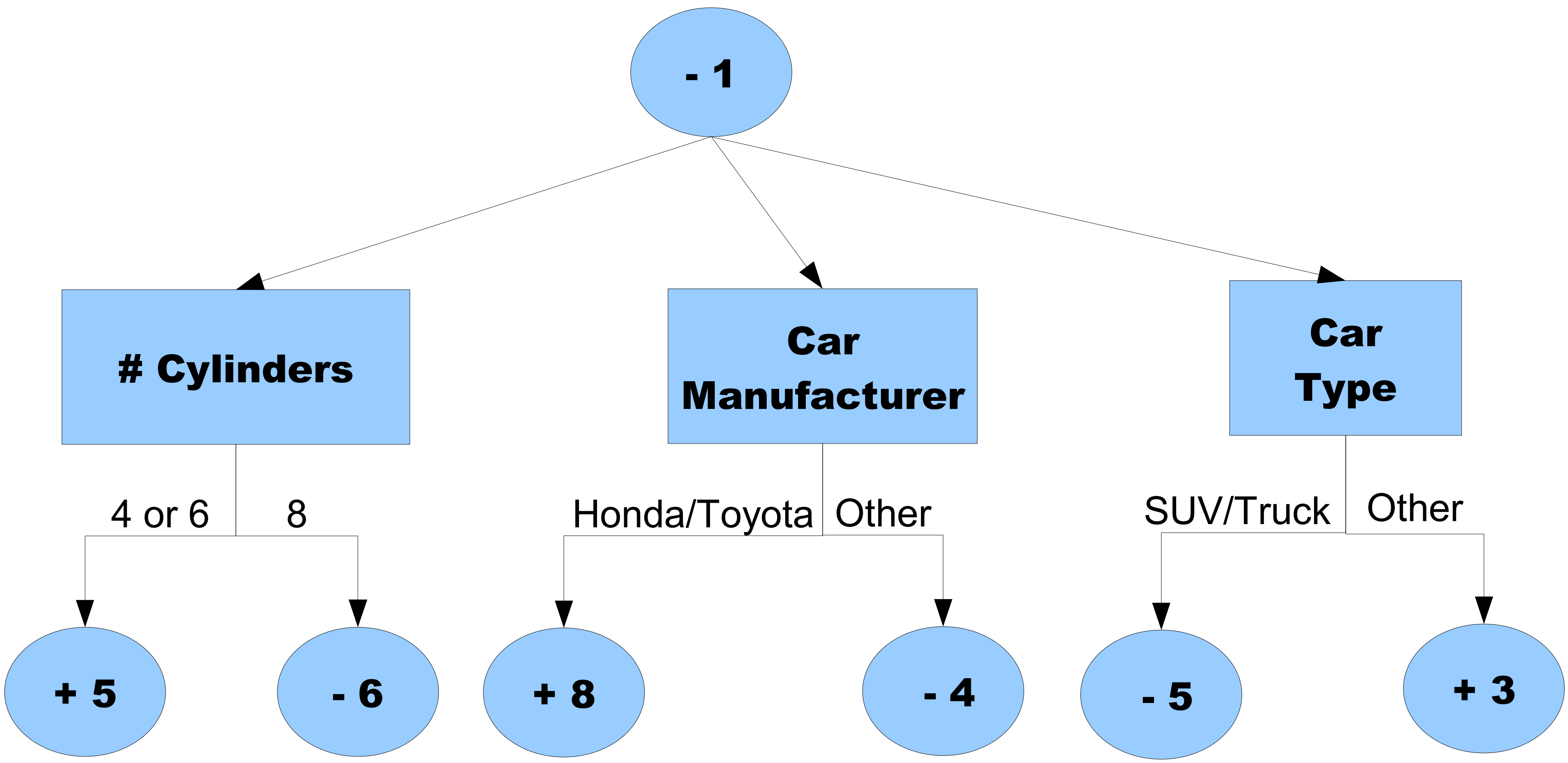
Decision Trees

- Follow a single path until reach decision
- No confidence levels
- Many criterion for growing decision trees

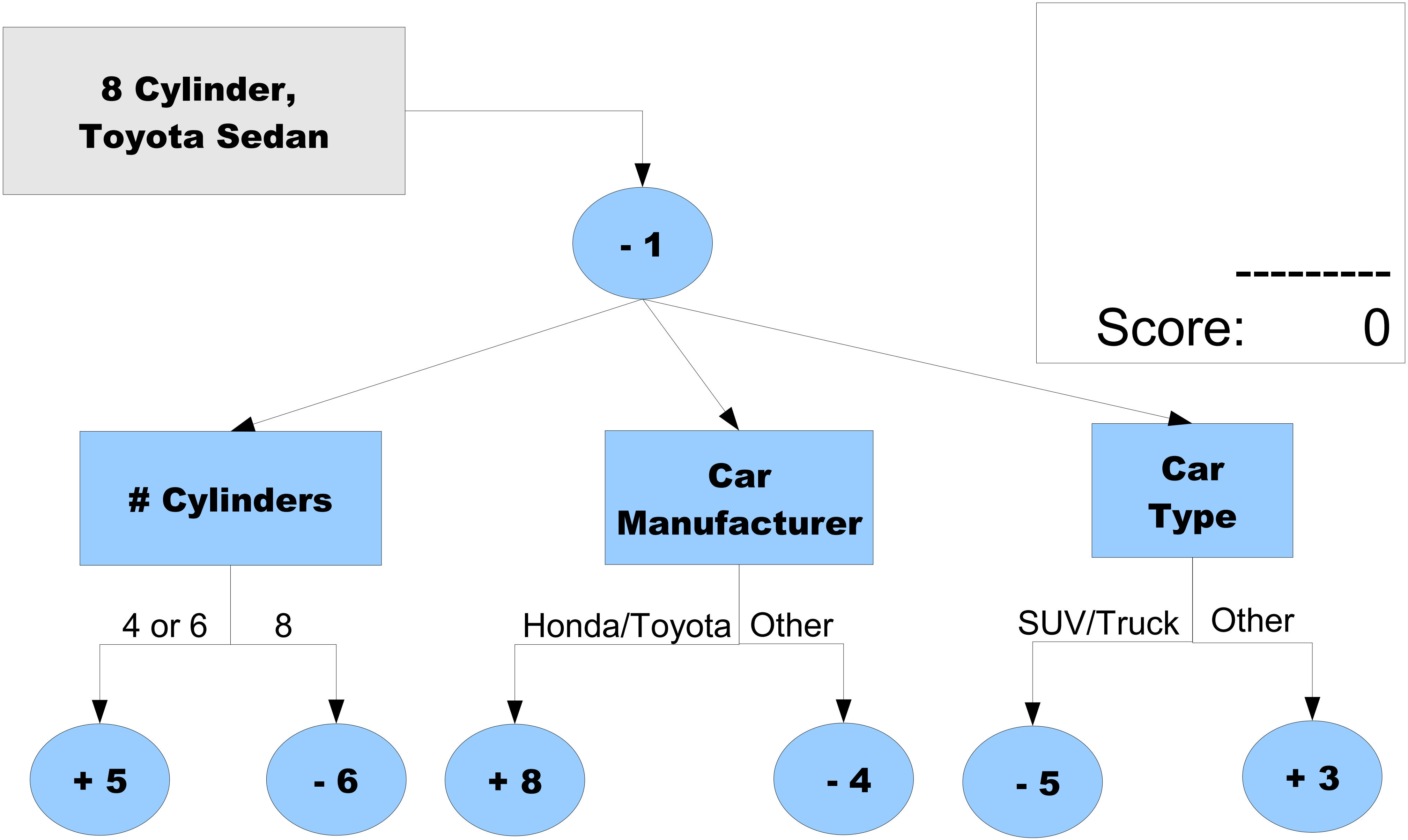
Final Classifier: Alternating Decision Tree

- Each path in tree is series of weak hypotheses
- Does the car have 4 or 6 cylinders?
  - Yes => +5,   No => -6
- Is the car a Toyota or Honda?
  - Yes =>+8,   No => -3
- A Honda with 8 cylinders => +2

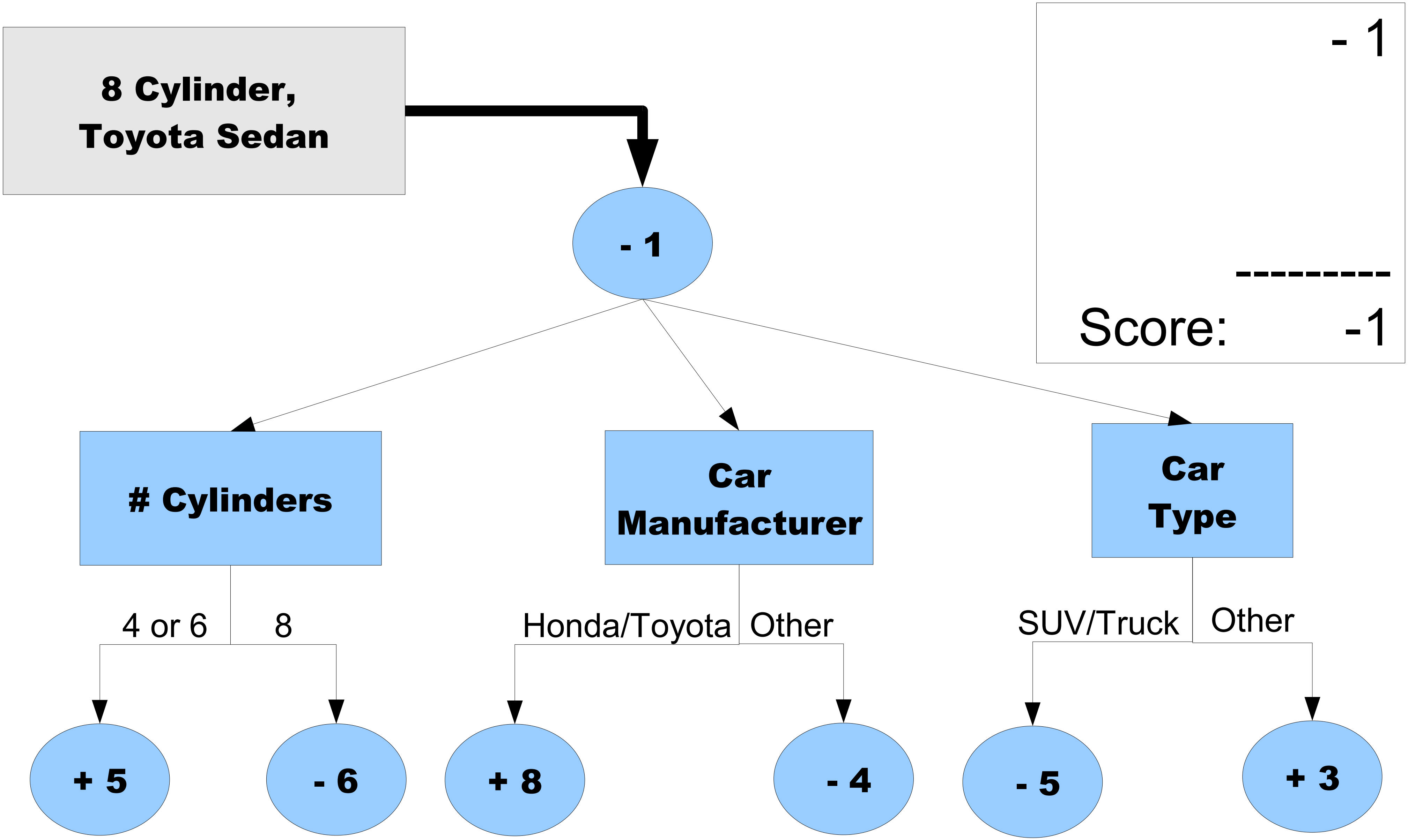
Alternating Decision Tree



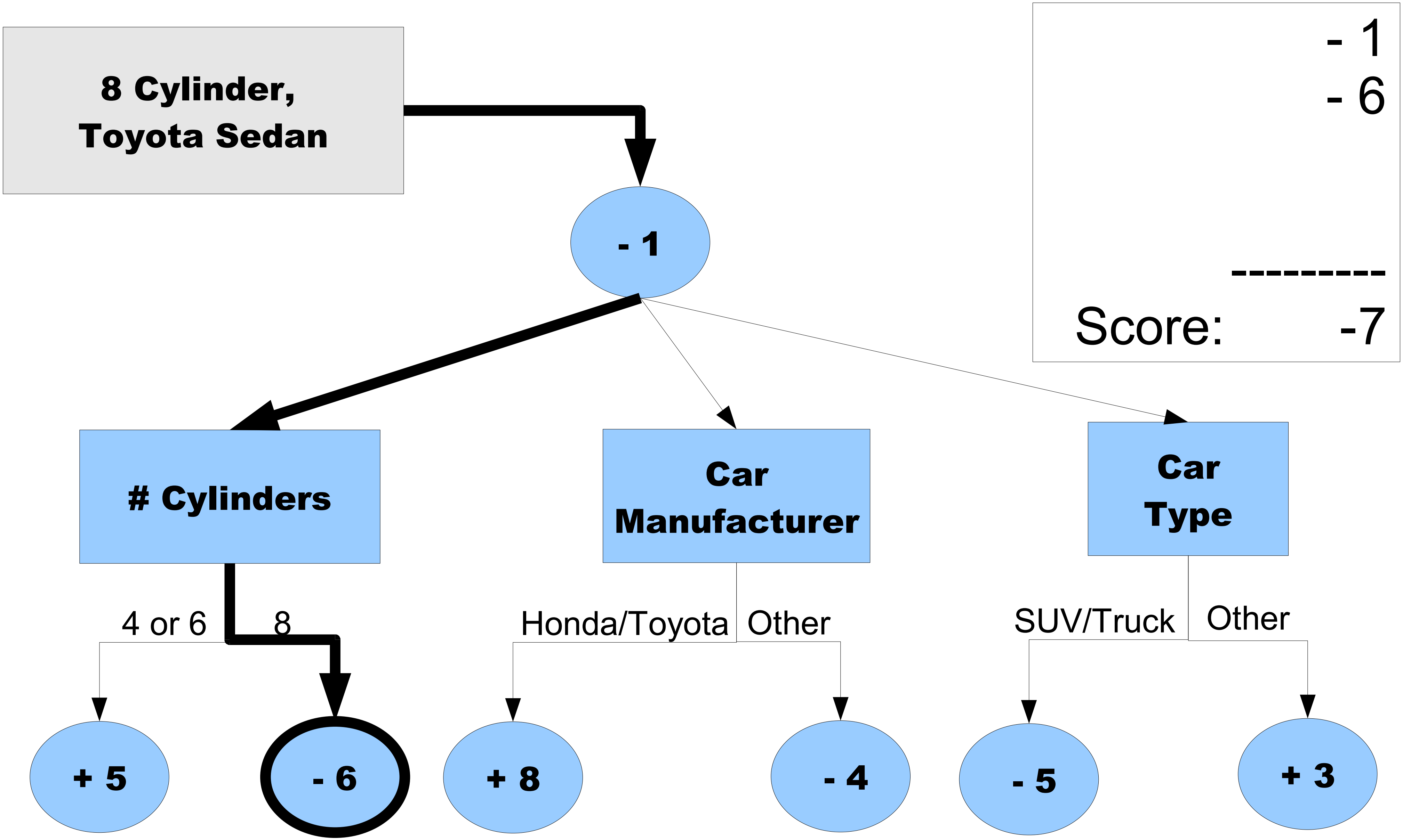
Alternating Decision Tree



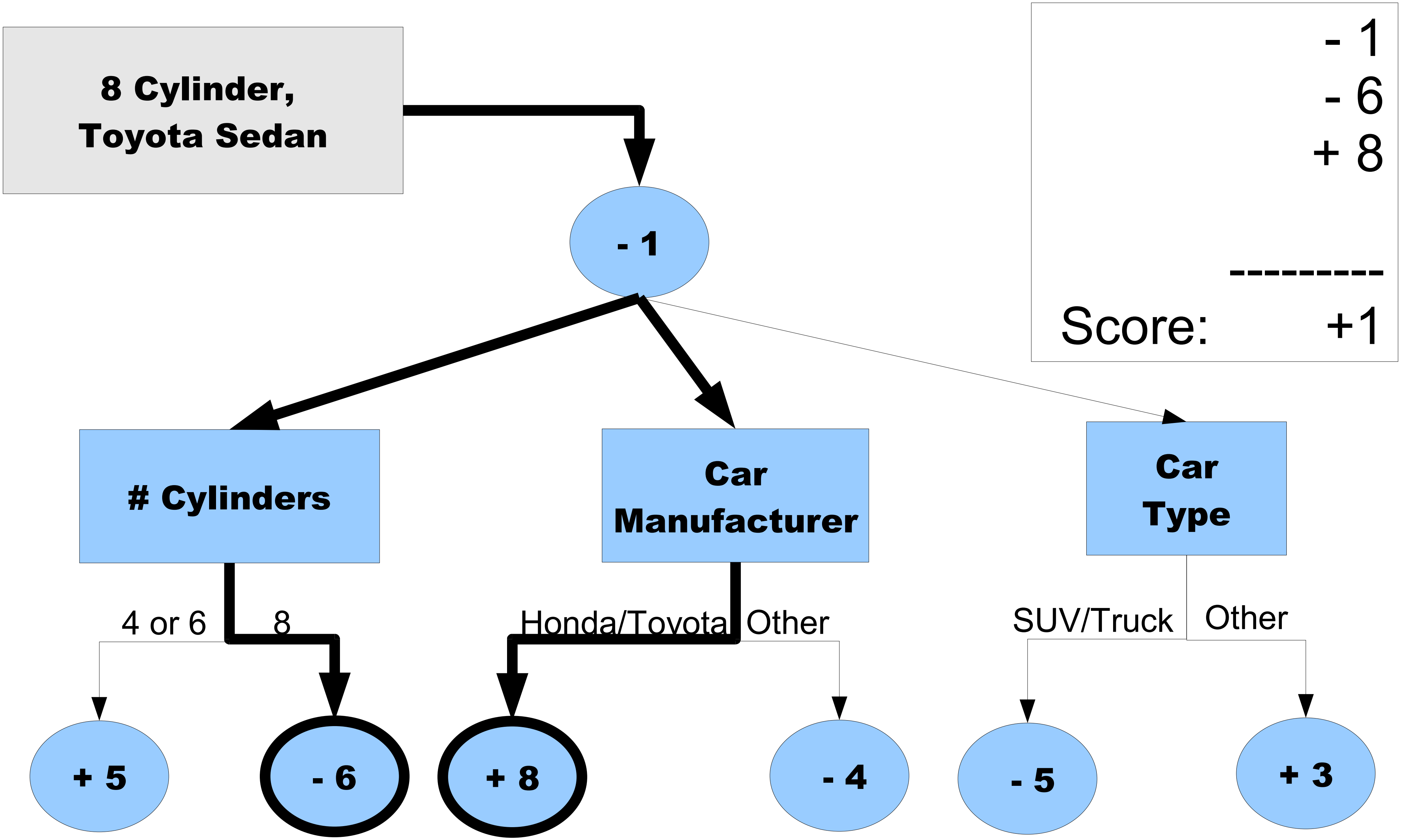
Alternating Decision Tree



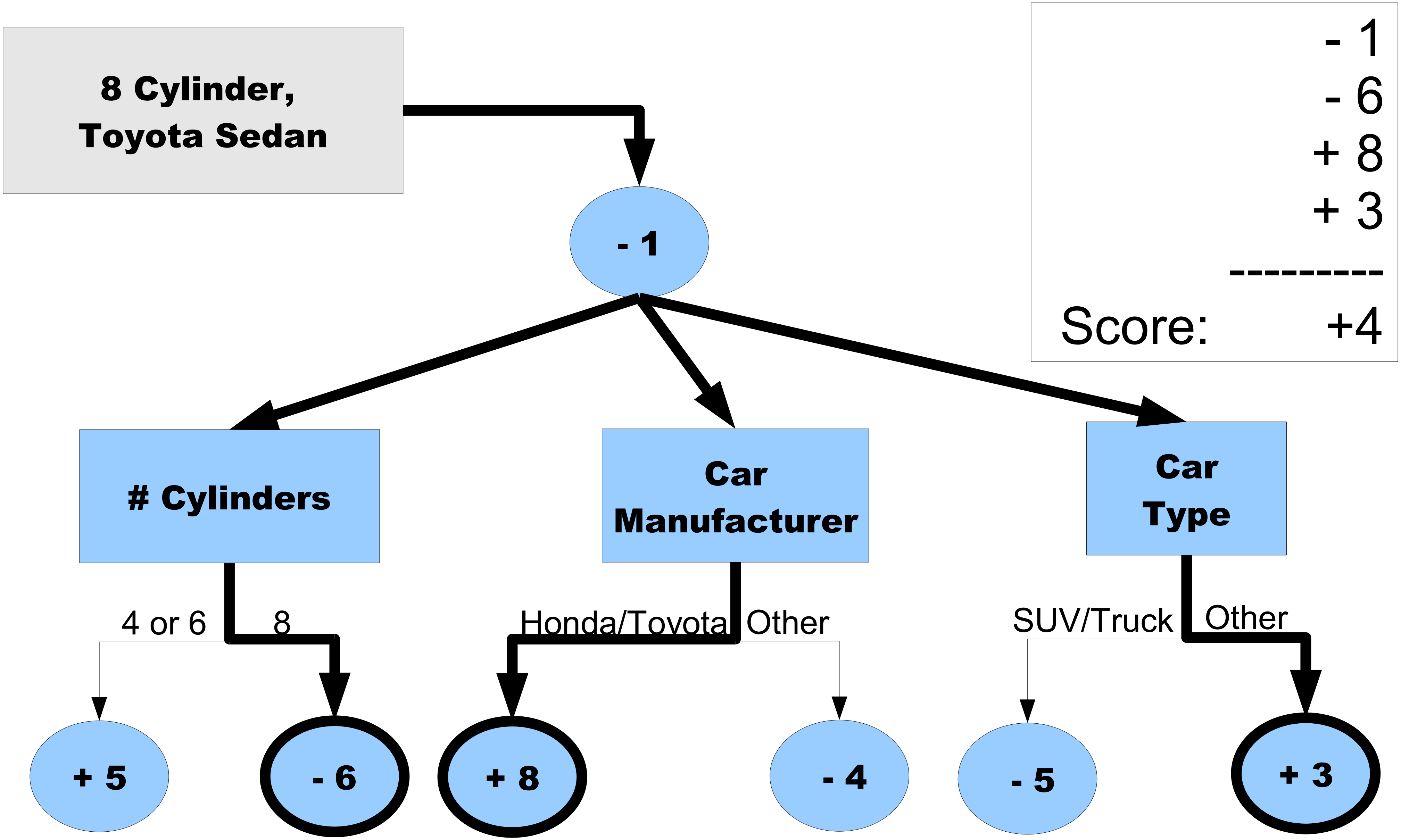
Alternating Decision Tree



Alternating Decision Tree

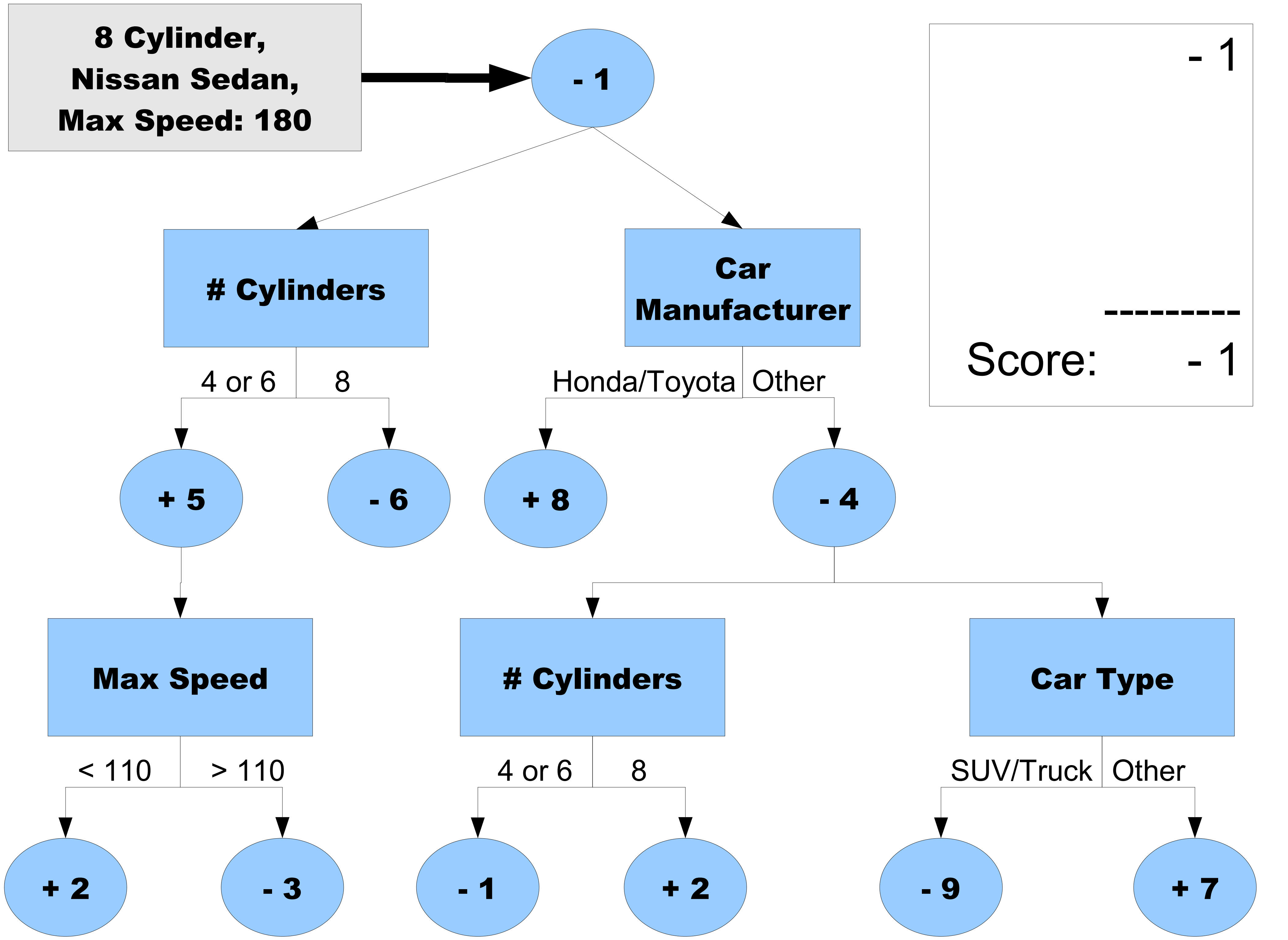


Alternating Decision Tree

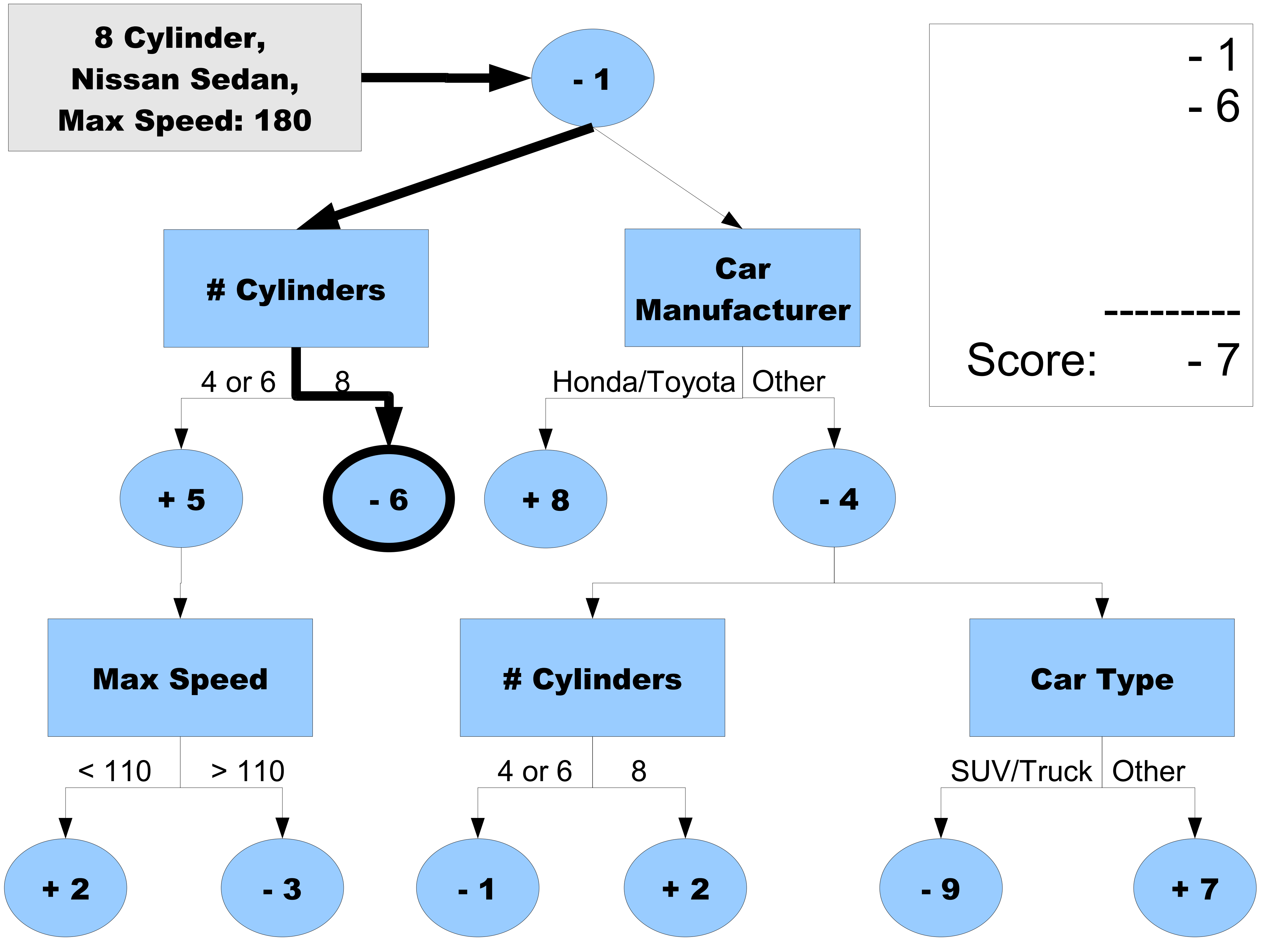


Another Example

- Previous example was pretty simple
  - Just a series of decisions with weights
  - A basic additive linear model
- Next example shows a more interesting ATree
  - Has greater depth
  - Some weak hypotheses abstain
- Two inputs are shown

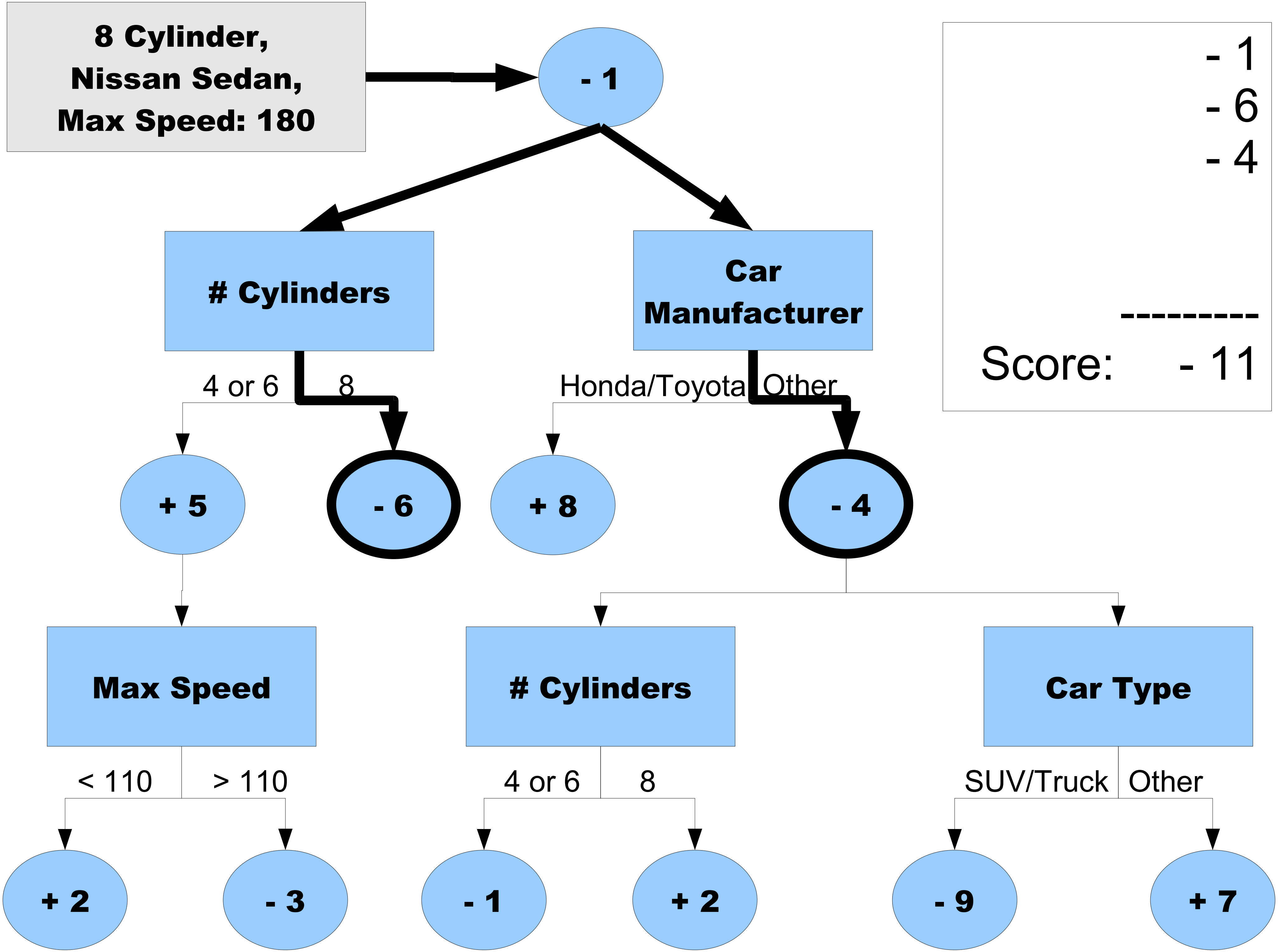


8 Cylinder,
Nissan Sedan,
Max Speed: 180
- 1
Cylinders
Car Manufacturer
4 or 6
8
Honda/Toyota
Other
+ 5
- 6
+ 8
- 4
Max Speed
Cylinders
Car Type
< 110
> 110
4 or 6
8
SUV/Truck
Other
+ 2
- 3
- 1
+ 2
- 9
+ 7
- 1
Score: - 1

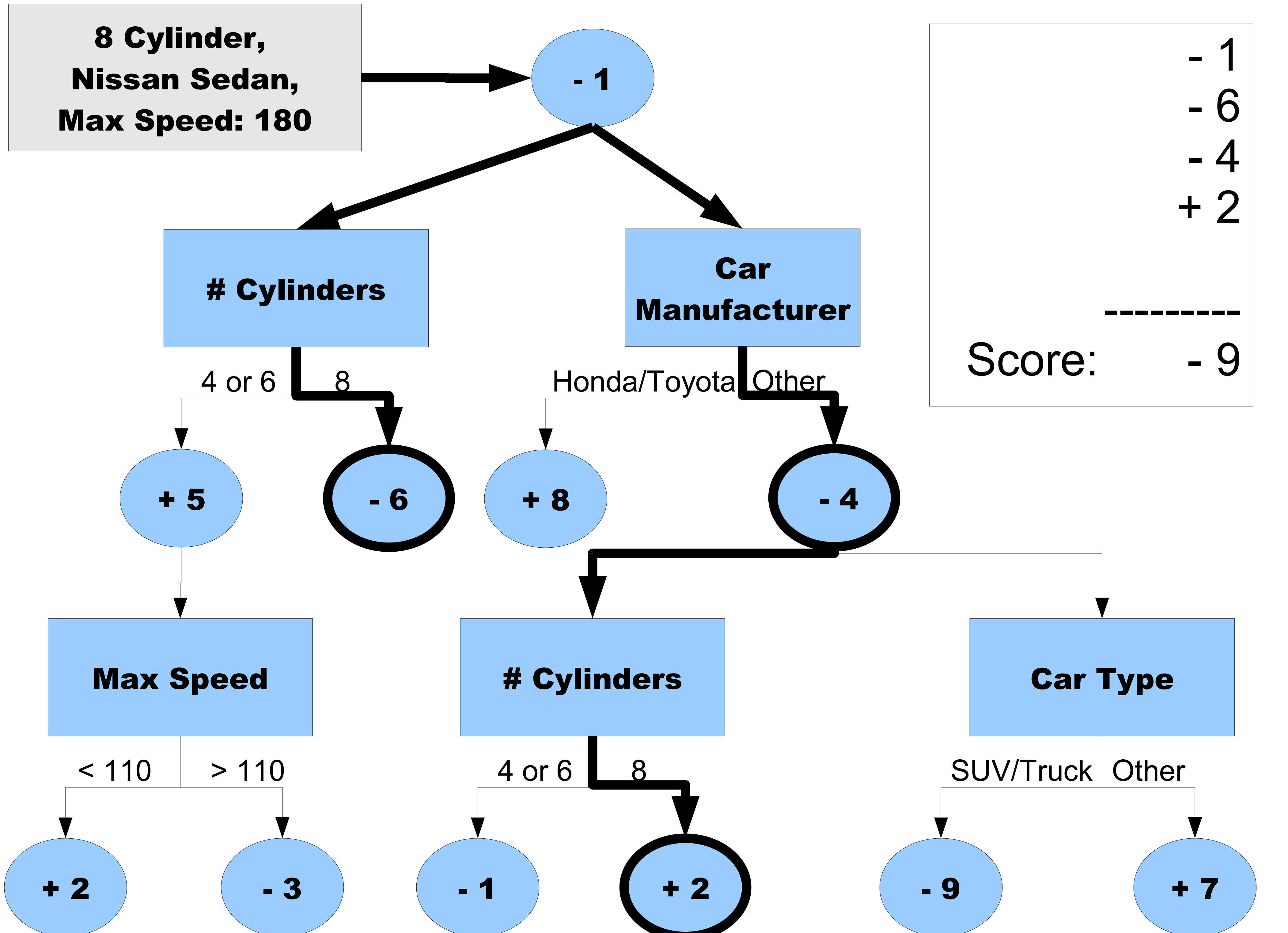


8 Cylinder,
Nissan Sedan,
Max Speed: 180
- 1
Cylinders
Car Manufacturer
4 or 6
8
Honda/Toyota
Other
+ 5
- 6
+ 8
- 4
Max Speed
Cylinders
Car Type
< 110
> 110
4 or 6
8
SUV/Truck
Other
+ 2
- 3
- 1
+ 2
- 9
+ 7
- 1
- 6

Score: - 7

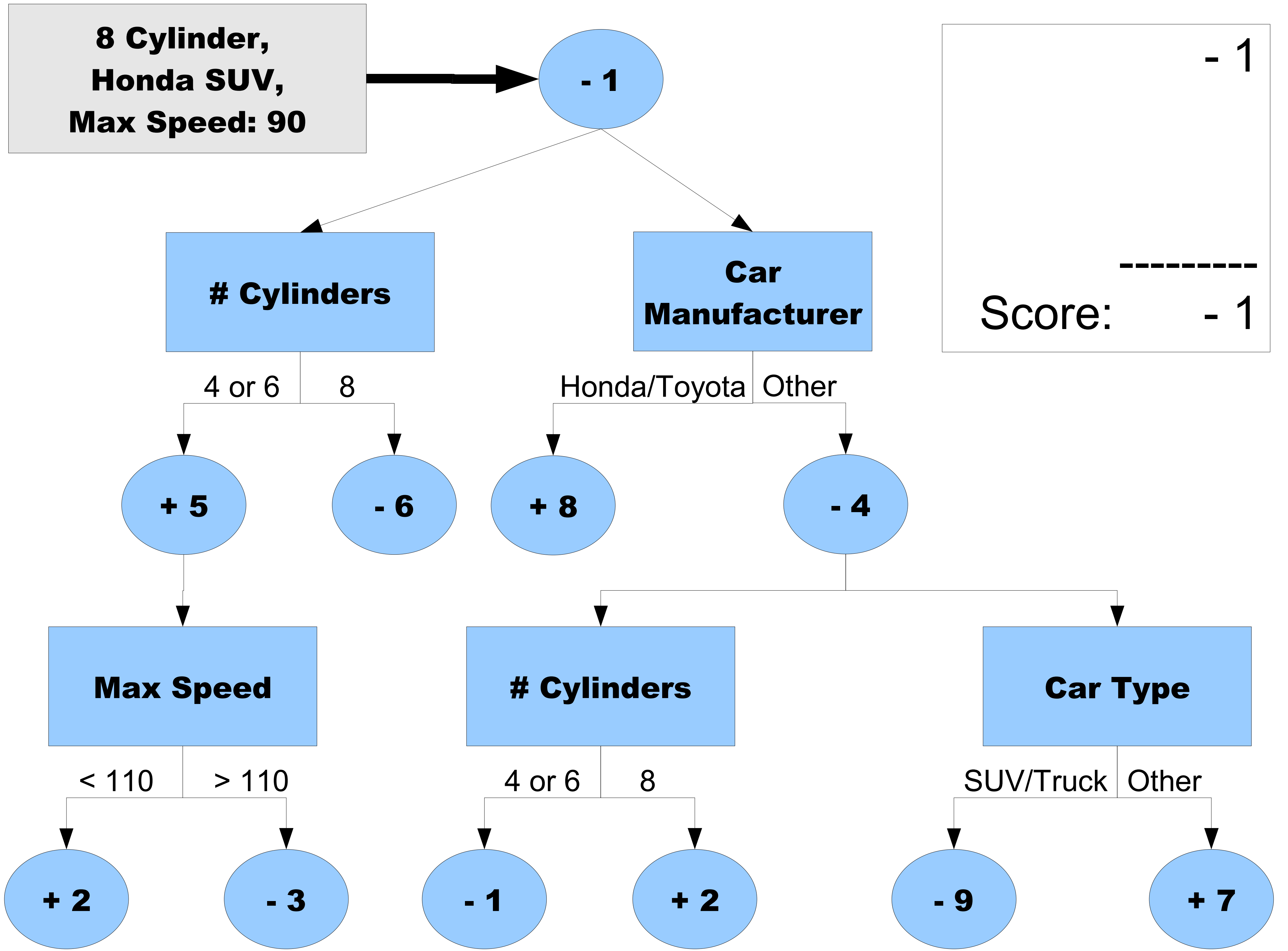

8 Cylinder,
Nissan Sedan,
Max Speed: 180
- 1
Cylinders
Car
Manufacturer
4 or 6
8
Honda/Toyota
Other
+ 5
- 6
+ 8
- 4
Max Speed
Cylinders
Car Type
< 110
> 110
4 or 6
8
SUV/Truck
Other
+ 2
- 3
- 1
+ 2
- 9
+ 7
- 1
- 6
- 4

Score: - 11

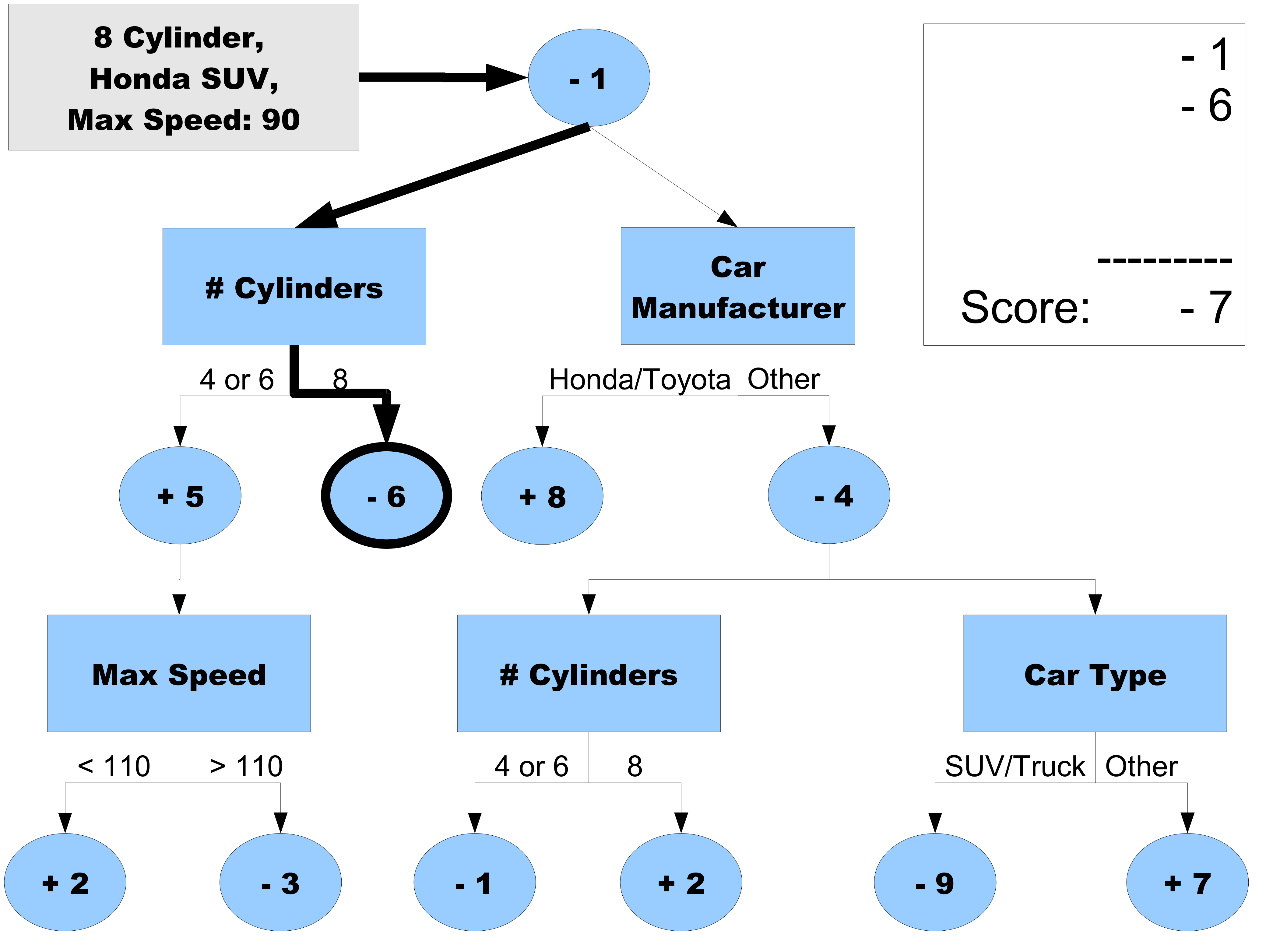

8 Cylinder,
Nissan Sedan,
Max Speed: 180
- 1
Cylinders
4 or 6
8
Car
Manufacturer
Honda/Toyota
Other
+ 5
- 6
+ 8
- 4
Max Speed
< 110
> 110
Cylinders
4 or 6
8
Car Type
SUV/Truck
Other
+ 2
- 3
- 1
+ 2
- 9
+ 7
- 1
- 6
- 4
+ 2

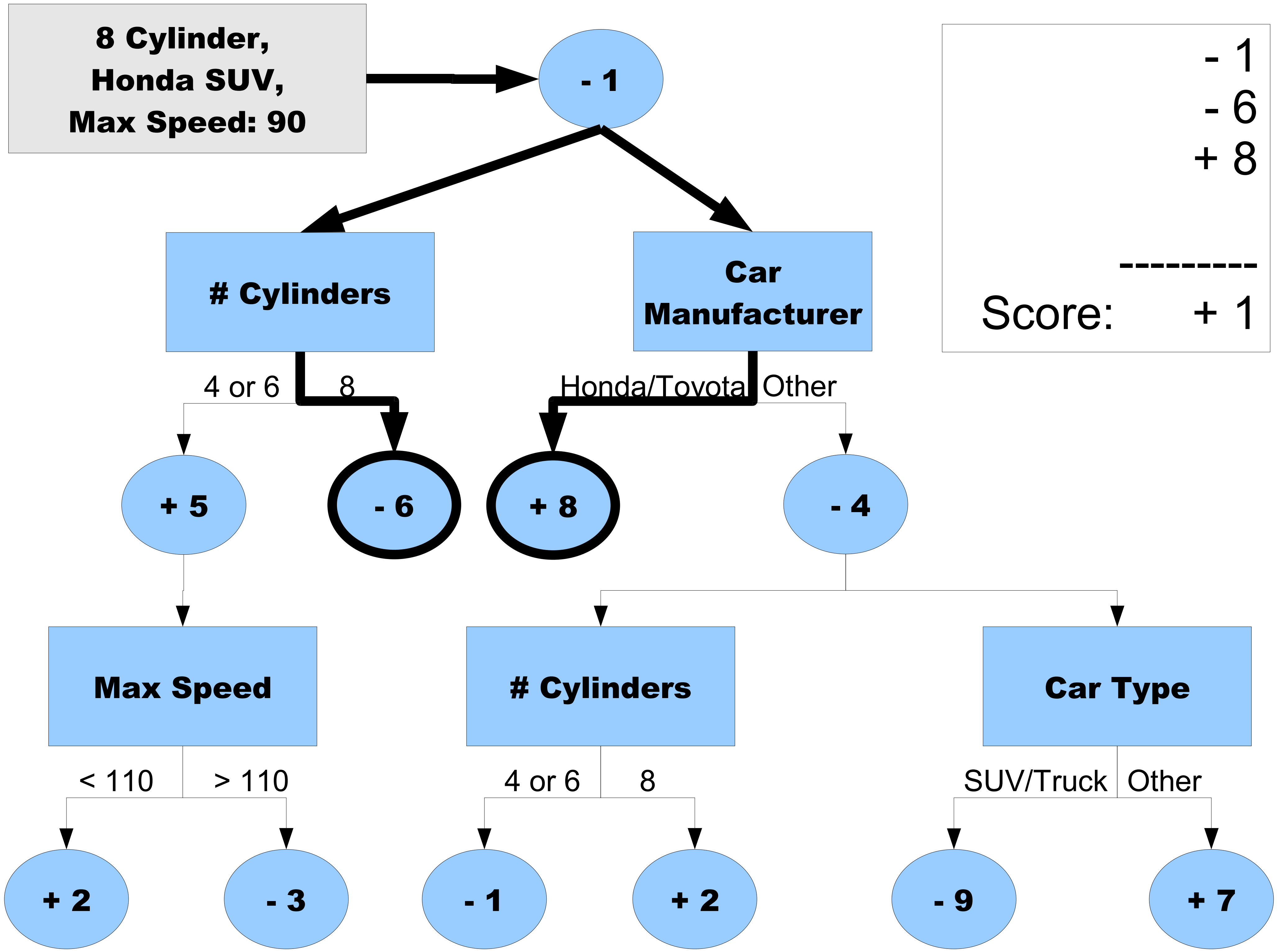
Score: - 9

Another Example

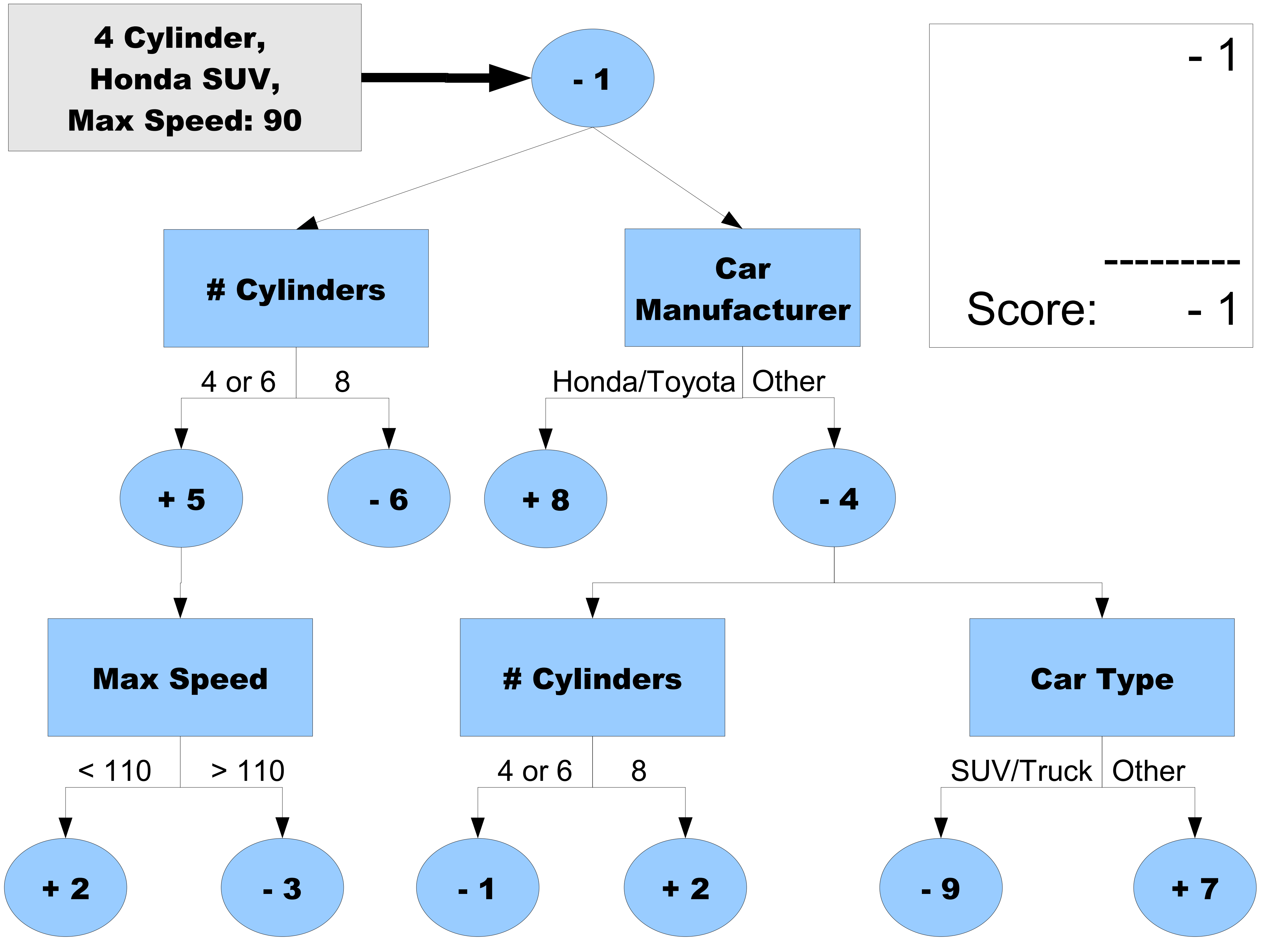

8 Cylinder,
Honda SUV,
Max Speed: 90
- 1
Cylinders
4 or 6
8
+ 5
- 6
Max Speed
< 110
> 110
+ 2
- 3
Car Manufacturer
Honda/Toyota
Other
+ 8
- 4
Cylinders
4 or 6
8
- 1
+ 2
Car Type
SUV/Truck
Other
- 9
+ 7
- 1

Score: - 1

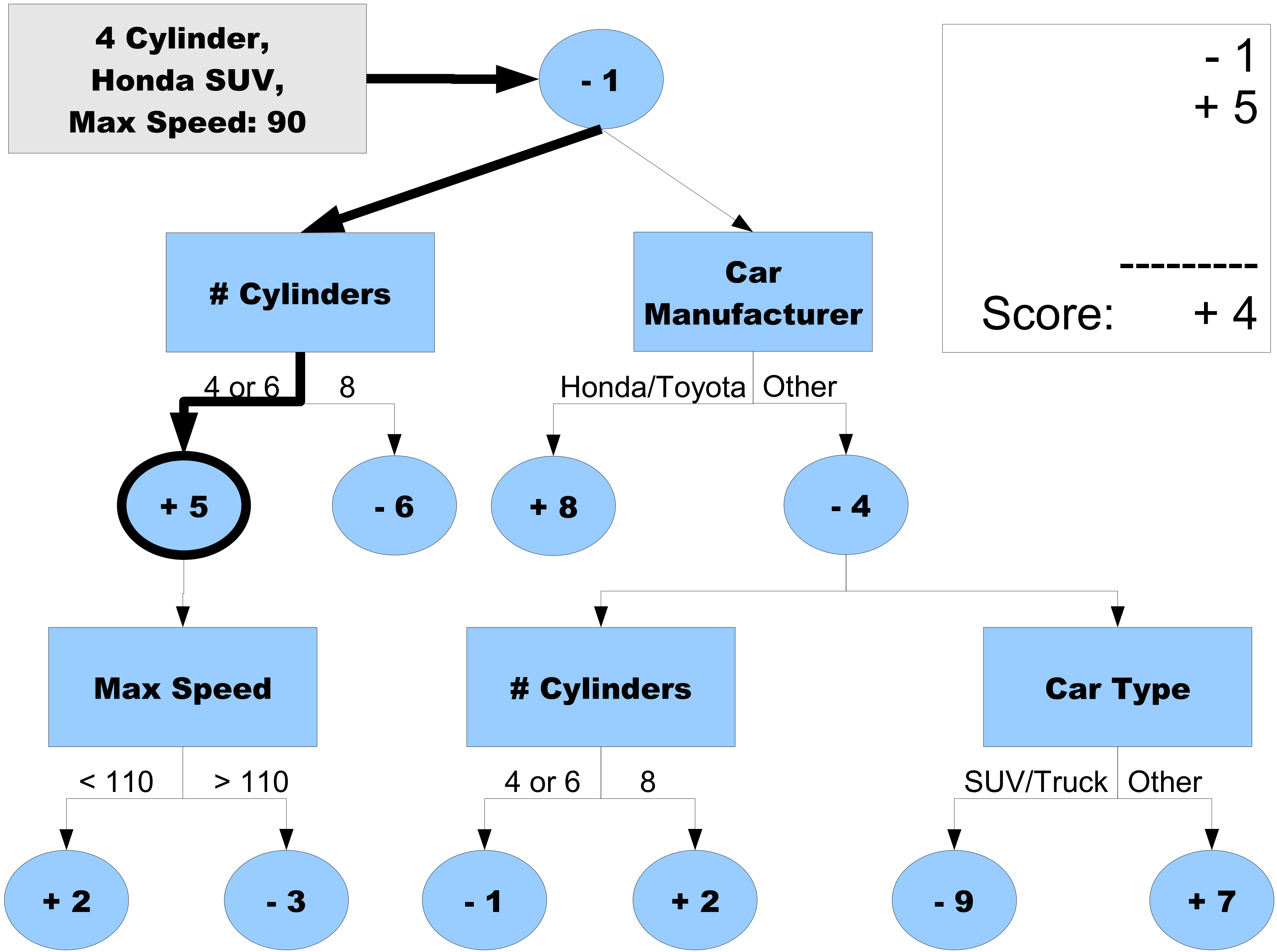

8 Cylinder,
Honda SUV,
Max Speed: 90
- 1
Cylinders
Car Manufacturer
4 or 6
8
Honda/Toyota
Other
+ 5
- 6
+ 8
- 4
Max Speed
Cylinders
Car Type
< 110
> 110
4 or 6
8
SUV/Truck
Other
+ 2
- 3
- 1
+ 2
- 9
+ 7
- 1
- 6
Score: - 7


8 Cylinder,
Honda SUV,
Max Speed: 90
- 1
Cylinders
Car Manufacturer
4 or 6
8
Honda/Toyota
Other
+ 5
- 6
+ 8
- 4
Max Speed
Cylinders
Car Type
< 110
> 110
4 or 6
8
SUV/Truck
Other
+ 2
- 3
- 1
+ 2
- 9
+ 7
- 1
- 6
+ 8
Score: + 1

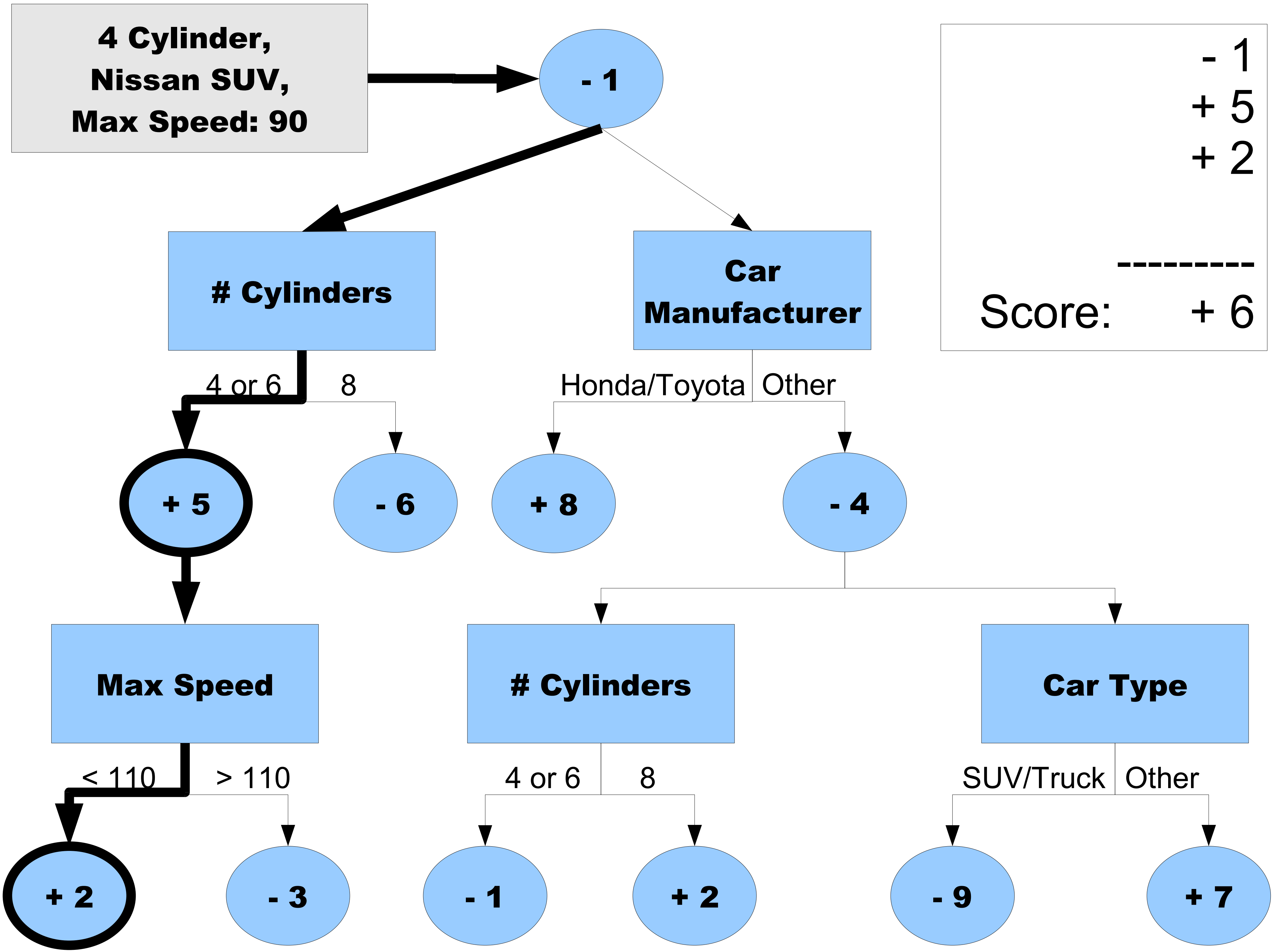
Another Example


4 Cylinder,
Honda SUV,
Max Speed: 90
- 1
Cylinders
Car Manufacturer
4 or 6
8
Honda/Toyota
Other
+ 5
- 6
+ 8
- 4
Max Speed
Cylinders
Car Type
< 110
> 110
4 or 6
8
SUV/Truck
Other
+ 2
- 3
- 1
+ 2
- 9
+ 7
- 1

Score: - 1

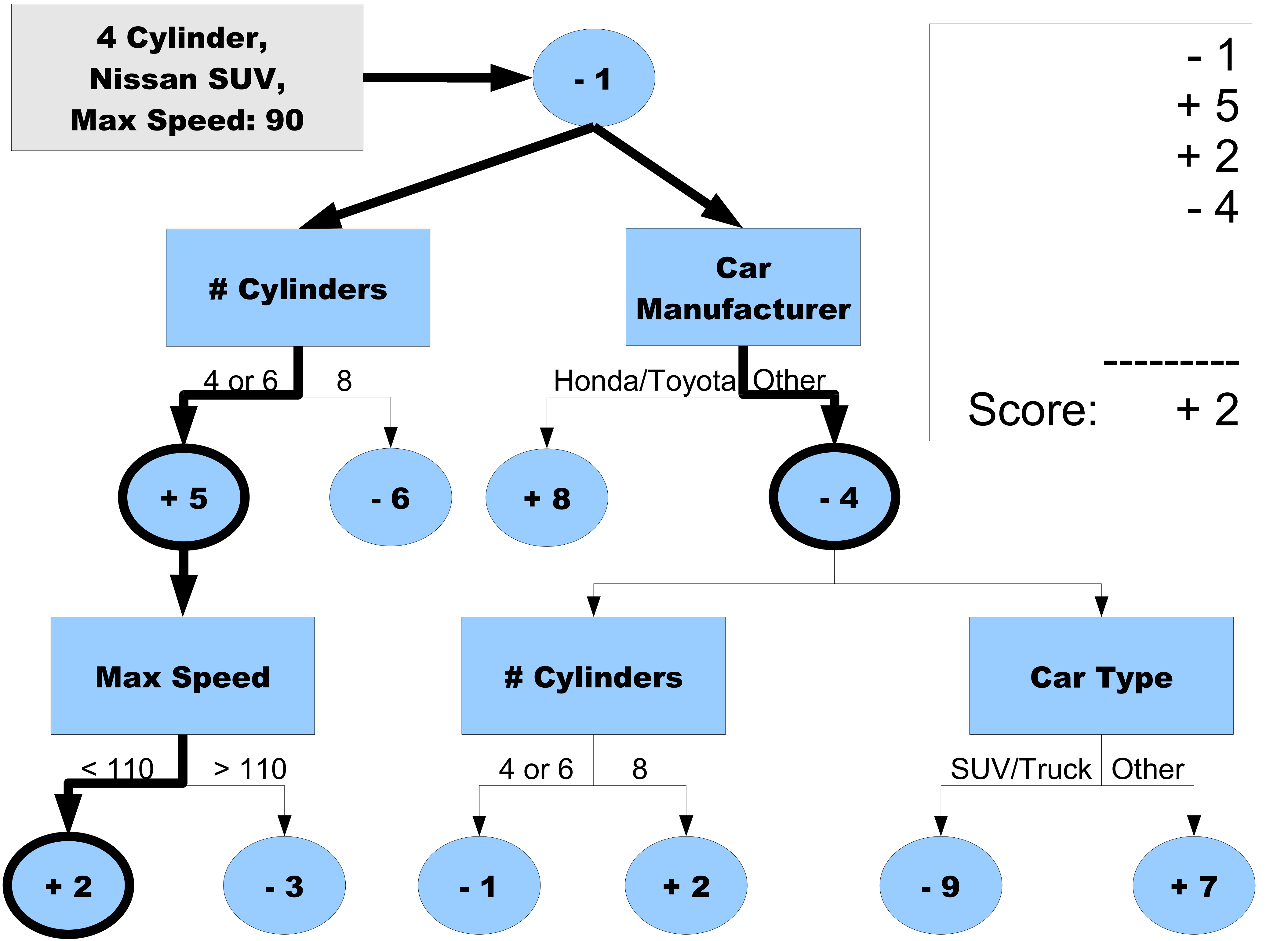

4 Cylinder,
Honda SUV,
Max Speed: 90
- 1
Cylinders
Car Manufacturer
4 or 6
8
Honda/Toyota
Other
+ 5
- 6
+ 8
- 4
Max Speed
Cylinders
Car Type
< 110
> 110
4 or 6
8
SUV/Truck
Other
+ 2
- 3
- 1
+ 2
- 9
+ 7
- 1
+ 5

Score: + 4



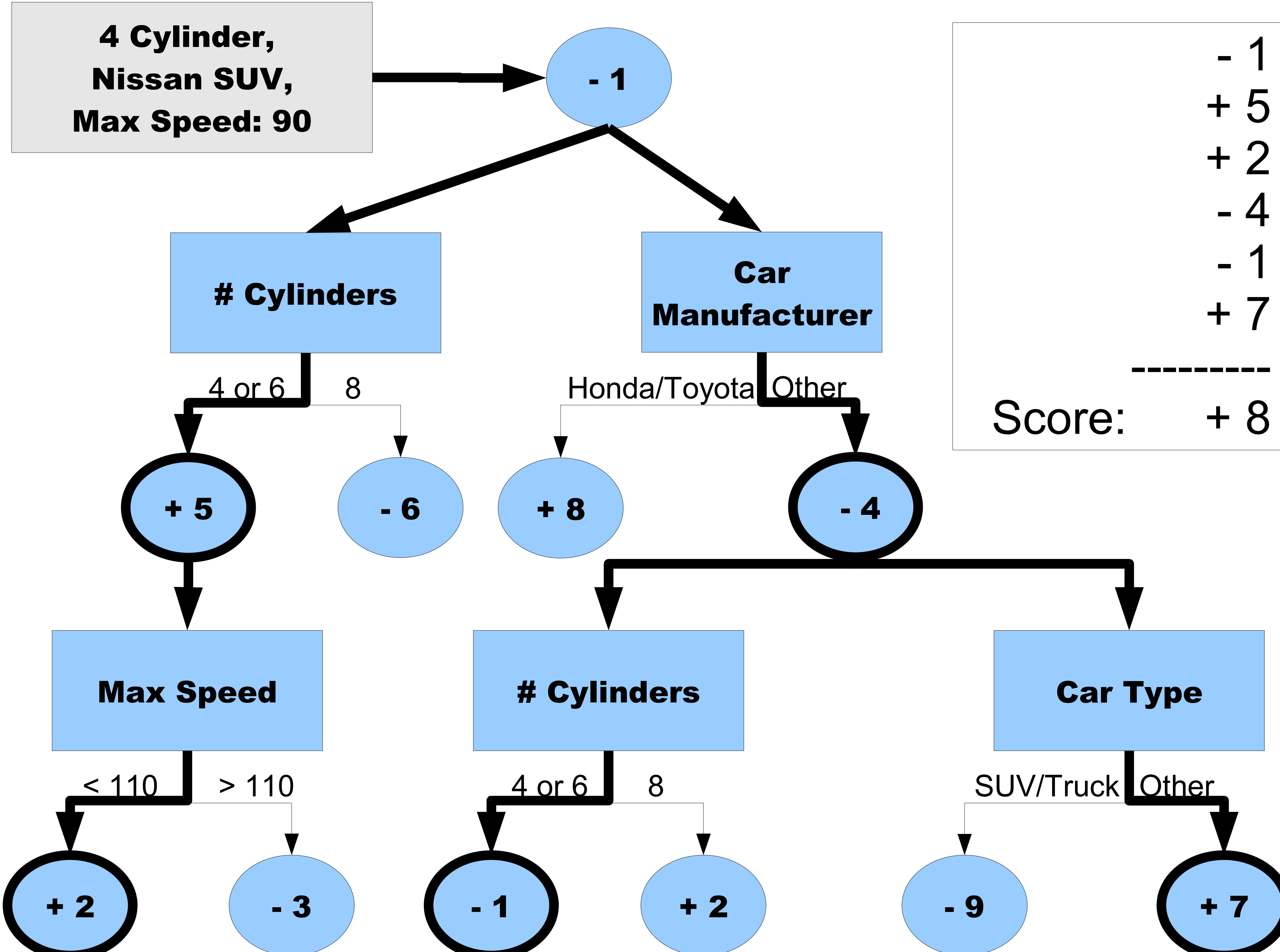
4 Cylinder,
Nissan SUV,
Max Speed: 90
- 1
Cylinders
Car Manufacturer
4 or 6
8
Honda/Toyota
Other
+ 5
- 6
+ 8
- 4
Max Speed
Cylinders
Car Type
< 110
> 110
4 or 6
8
SUV/Truck
Other
+ 2
- 3
- 1
+ 2
- 9
+ 7
- 1
+ 5
+ 2

Score: + 6



4 Cylinder,
Nissan SUV,
Max Speed: 90
- 1
Cylinders
Car Manufacturer
4 or 6
8
Honda/Toyota
Other
+ 5
- 6
+ 8
- 4
Max Speed
Cylinders
Car Type
< 110
> 110
4 or 6
8
SUV/Truck
Other
+ 2
- 3
- 1
+ 2
- 9
+ 7
- 1
+ 5
+ 2
- 4

Score: + 2


4 Cylinder,
Nissan SUV,
Max Speed: 90
- 1
Cylinders
Car Manufacturer
4 or 6
8
Honda/Toyota
Other
+ 5
- 6
+ 8
- 4
Max Speed
Cylinders
Car Type
< 110
> 110
4 or 6
8
SUV/Truck
Other
+ 2
- 3
- 1
+ 2
- 9
+ 7
- 1
+ 5
+ 2
- 4
- 1
+ 7

Score: + 8

ATree Pros and Cons

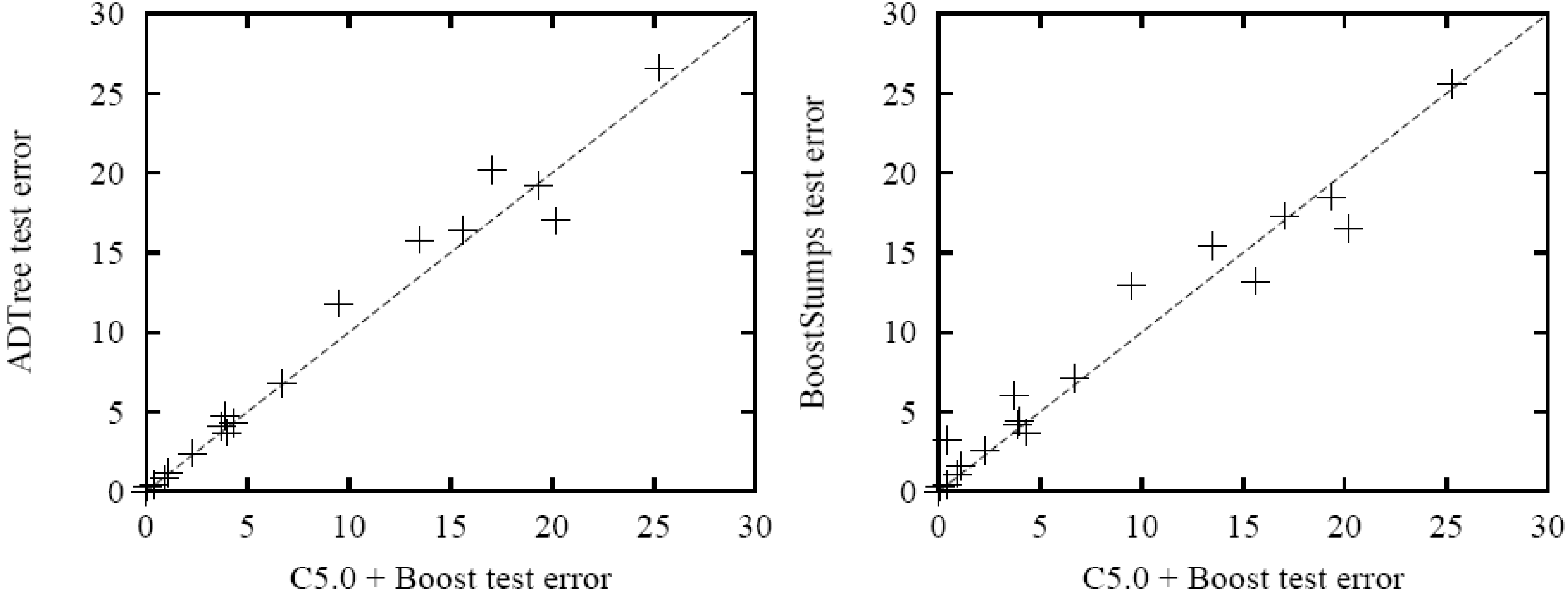
Pros

- Can focus on specific regions
- Similar test error to other boosting methods
- Requires far fewer iterations
- Easily visualizable

Cons

- Larger VC-dimension
  - Increased proclivity for overfitting

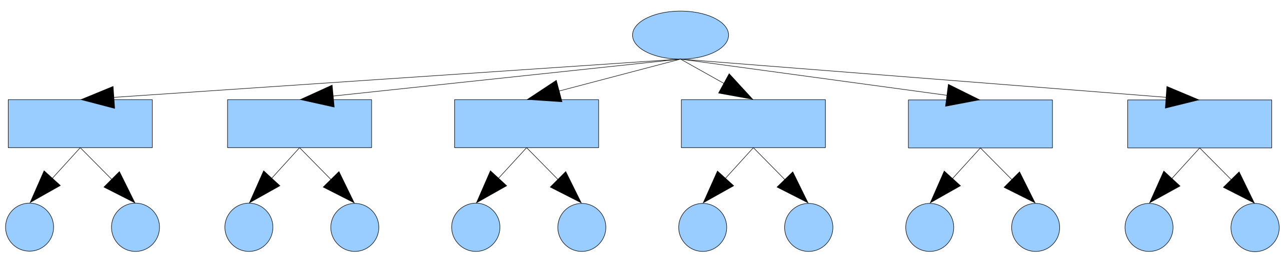
Error Rates



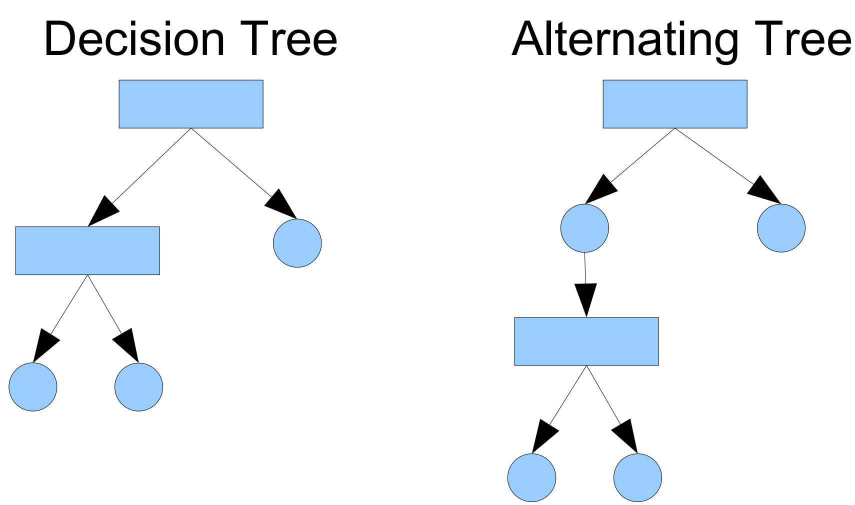
Taken from Freund & Mason 1997

Some Basic Properties

- ATrees can represent decision trees, boosted decision-stumps, and boosted decision trees
- ATrees for boosted decision stumps:



- ATrees for decision trees:



Resources

- Boosting.org
- JBoost software available at http://www.cs.ucsd.edu/users/aarvey/jboost/
  - Implementation of several boosting algorithms
  - Uses ATrees as final classifier
- Rob Schapire keeps a fairly complete list http://www.cs.princeton.edu/~schapire/boost.html
- Wikipedia