# PREDICTING USED CAR PRICE

Ted Fitch
DATA610: Fall 2020
Assignment 3
Tedfitch4@gmail.com
Dr. Laila Moretto

## Introduction:

When attempting to purchase a vehicle, it's important to know what variables drive price in order to negotiate the best deal possible. Previous studies have predicted mileage being the single greatest factor for a used car's sales price (Engers et al., 2009). The dataset being explored here describes used car sales (**Figure 1**). It contains 12 columns with exactly 1,437 rows. Most variables are continuous categories including price, age, CC, KM, HP, and weight. In contrast, there are two binary variables: autotype (whether a vehicle is an automatic or manual transmission; 1 being manual and 0 being automatic) and metcolortype (whether a vehicle has a metallic color; 1 meaning that it has a metallic color and 0 meaning it has a non-metallic/matte finish). These variables are listed twice, once in binary fashion (0/1) and once using words (in order to describe which number pertains to which category). Lastly, there are 2 categorical variables: fueltype, and doors. There are no missing entries in this dataset. Once the binary variables were recorded (what 0 & 1 pertained to), the 2 columns written verbally were deleted from the dataset in order to consolidate the data. The unit of age is assumed to be months since this column ranges from 1 to 80. It is very unlikely a used car would be sold if it was 80 years old because cars rarely survive that long (unless it was a collector's item in which case the price would be much higher than is listed in this dataset). Furthermore, a new column was made by dividing KM/age in order to describe how much usage the automobile has had. This is important because a car may be 30 months old but only driven 1,000 KM per year thus preserving the value. Alternatively, a car may only be 5 years old but driven 50,000 KM per year. Thus, this is a ratio worth exploring. For these reasons, it is hypothesized that KM/age will be a predictor of sales price.

## Model 1- All Data:

The first model developed included all data with no filters and explored all variables as possible predictors of sales price. Cognos created a tree diagram with a predictive strength of 80.8% using age, weight, KM, and CC (**Figure 2**). The first obvious insight from this model is that age is the strongest predictor. When a car ages beyond 40 months, it drops off significantly in value. Cars under 40 months of age have an average sale price of $16,487 whereas all cars listed have an average sale price of $10,731. The next age grouping (40-55 months) is $11,308. This is a difference from the original node average of $6,000 and from the next age grouping of $5,000. This makes it a good rule because it highlights a sharp contrast in the data between younger cars and older cars. Intuitively, it makes sense that the older a car is, the less value it retains. Some cars increase in value with age due to being collector's items; however, there aren't any obvious signs of that in this dataset. In the real world, this rule would imply that a car less than 40 months with a sales price less than $16,487 would be less than average and thus a good deal for the buyer.

Subsequently, separating by weight appears to be the next best rule in this model. The node begins with an average sale price of $16,487 (for cars less than 40 months old) and then bifurcates into two nodes: less than 1105 kg (average price: $13,706) and more than 1105 kg (average price: $18,434). This shows the rule splits the group with a separation of $5,000. From a high-level view, this model shows that cars are most expensive when they are newest, have high weight, have low mileage, and have a high CC count. All of these would intuitively make sense. Weight may seem more elusive; however, it makes sense that a vehicle that weighs more, demands more material to build, and thus would cost more to buy. A pickup truck would be the best example where classic a Toyota Tundra will weigh more and cost more than a Toyota Tacoma (Car and Driver, n.d.). Larger trucks can haul more and need larger engines, and thus they cost more. This may be the reason why the Decision Tree chose CC to predict sales for high weight vehicles; because once cars are larger, it matters how larger of an engine it has.

Lastly, the sunburst diagram (**Figure 4**) gives a cleaner look at the same data from the tree diagram. It clusters the strongest variables together in a way that's more visually intuitive to see the sales price increases; however, it doesn't add the driver labels (so the only way to understand the sunburst diagram is with the ruleset **Figure 3**) which is a major limitation. The major plus of this diagram is it shows the relative number of entries in each category. This helps one ascertain there are no major outliers; each category contains many entries. The smallest two bins have 26 and 46 entries in each. When observing this diagram, it also became apparent that the two newest bins for cars each required more nuance to determine price. As cars senesce, it becomes easier to predict their price because age becomes the primary driver. For newer cars, other factors still have significant predictive strength (weight, CCs of the engine, mileage, etc.).

## **Model 2- Secondary Variables:**

The second model was developed by stripping away the 2 most weighted variables from the previous model: age and weight. This was in order to see what the next most powerful variables would be. Consequently, the next strongest variables were KM, CC, door number, and metcolortype. These predict sale price with a strength of 49.6%. It's important to note that KM has a predictive strength more than double any other factor in this model. Fascinatingly, door number and metallic coloring are the next strongest variables. When they split, each of the subsequent bins are separate from each other by approximately $2,000. Again, the model correctly predicts what happens in real life. Non-metallic cars tend to cost less because the paint is less expensive to create (Evans, 2013). Furthermore, it intuitively makes sense that cars with 5 doors would be more expensive than cars with 3 or 4 doors. Just like the variable weight, the more mass there is to the car, the more it will cost. As a note, a “5th door” or “3rd door” would be the trunk of a hatchback vehicle where there are already 4 doors or two doors respectively.

Model 2 is quite helpful in particular because it shows the stark difference the CC variable makes on the sales price. When the CC variable bifurcates, there is a range of $6,000, $2,000, and $4,000 for the three, subsequent nodes. We weren’t able to see as stark of a contrast in Model 1 because there was only a difference of $3,000. Thus far, it’s fair to conclude that age, weight, and KM are the 3 strongest variables, while CC is variable with moderate predictive strength.

**<u>Model 3- New Variables:</u>**

A final model was created again by removing the strongest two drivers in the previous model, KM and CC. This was in order to see what the resulting, next strongest predictive variables were. As a result, the drivers were door number, horsepower, and fuel type in that order. With a predictive strength of only 14.6% , it’s not a robust or powerful model but does help us see what variables still hold some weight. By far, the best rule in this model is separation based on fuel type. The average before the split is $11,000; but, after splitting the two categories are $11,000 for Petrol and CNG while Diesel climbs to $18,000. The sunburst diagram (**Figure** ) shows definitively that diesel cars are a small subset (2% of total). Thus, this isn’t enough data to make a statistically significant claim about all diesel cars; however, it does point us in the right direction to explore a hypothesis that diesels cars tend to cost more than petrol or CNG cars. Since diesel vehicles tend to get a much higher MPG rating than most petrol or CNG vehicles, it would make sense that they would sell at higher prices. With comparison to model 2, it’s interesting that metcolor didn’t becoming a defining variable in model 3 (even though doors remained a defining variable). Perhaps, this shows that metcolor was only useful to predict differences in price once the first three rules had been applied. Regardless, door number still remains a predictive variable and so our “variable hierarchy” is now: age, weight, KM, CC, and door number.

## **Comparison of Models:**

When comparing these three models, a few things become apparent. First, it’s interesting that the variable created, KM/age, was not a predictor in any of the three models. Age and KM were strong predictors individually, but they weren’t effective as the KM/age variable which was surprising. Assuming that this dataset isn’t too small or too niche to test this hypothesis on, then there could be a couple reasons why age is a better predictor than KM/age. It could be a psychological factor where people are willing to pay more based solely on the year (instead of thinking it through and basing it on KM/age). It sounds better to say your car is newer (even if it has been driven more roughly). More simply, it could be due to there not being large variance in KM/age where most people drive their cars at similar rates in this dataset.

Models 1 and 2 were far more complicated than Model 3; however, both Models 1 and 2 had much higher predictive strength than Model 3. Model 1 clearly was the best predictor and showed age, weight, and KM to be major predictive variables. In an attempt to improve this “best-case” model, a few simple outliers were removed. It was previously established that door number was a moderate predictor of sales price. Most data were either a 3-door or a 5-door car. So, both of these were explored, and the 5-door data was able to create a better model. Next, all diesel cars were removed. While diesel cars were a relatively small subset, it was established these were much more expensive and thus would be skewing the model. With these filters applied, Model 4 was made (Figure 11) with only 3 predictive variables (age, weight, and CC) and with a predictive strength of 84.4%. This is nearly 4% better than model 1 with one less variable. This is comparable to other real-life models (a machine learning model has predicted used-car sales price with a 87.38% accuracy)(Gegic et al., 2019). The sunburst diagram (**Figure 13**) shows there are no major outliers. Each rule applied with age, weight, and CC had defining impact as seen in comparing the difference from the internal node to each leaf node. Compared to the other models, Model 4 is simple (only 3 splits with 3 variables and 7 terminal leaves) and powerful.

## Organization Discussion:

My biotech organization is completely different from exploring the used car market; however, the rules-based approach discussed here is still just as applicable. In my forensic laboratory, one of the major decisions we have to make is identifying a type of fiber. We must make this decision based upon a series of measurements and features we observe. For a few examples, we have to check:

**Binary variables**: positive or negative sign of elongation; anisotropic or isotropic

**Categorical variables**: birefringence (split into high, medium, and low), interference order (1st, 2nd, 3rd, 4th), extinction (complete parallel, complete perpendicular, incomplete), morphological characteristics (lobing, scaling, hashing, twisting, tapering)

**Continuous variables**: width (measured in µm), RI value

Approximately 95% of fibers can be identified using these variables. There's currently no official workflow or operating procedure for this process of fiber identification; but, the analysts make this decision based purely on experience. It would be simple to predict fiber identity with a high level of integrity using a trained model. The first rule would rely on morphological characteristics; most of these single features are enough to identify a fiber (e.g. scaling only occurs on hairs; hashing only occurs on bast fibers). If there are no morphological characteristics, then the next nodes would separate based on extinction, sign of elongation, and a calculation (ratio of birefringence to width). This process would be quite helpful since the current process involves gathering all data, then checking if that data is congruent with what analysts have previously observed and with the data written in a reference atlas (McCrone Atlas, n.d.). Using a model, an analyst could type in the data where they normally would in their electronic lab notebooks, then the software would predict and autofill the fiber identity. This would be overwhelmingly accepted within my organization because each person is a technophile and early adopter. The only trepidation that people would need to be convinced the model is nuanced enough. Forensics requires a high degree of nuance because material compositions can be so complicated. Thus, a way to mitigate risk and allow analysts to add nuance would be to add a database field where analysts could write a summary describing any details so that the identity that the model produces is nuanced. This way, if an analyst disagreed with the predictive model, they would explain their reasoning and the final fiber identity.

## **Summary:**

Based upon this exploration, it appears the greatest predictive factors for this dataset were age, weight, KM, CC, and door number. While this investigation wasn’t able to show KM/age as a predictive variable, it did show age as being a stronger predictive variable than KM. This is interesting because other studies have previously show KM to be greater than age (Engers et al., 2009). It’s unclear why this result was found (though it’s possible this sample size was too small). In any case, the decision tree models developed show a clear “rules-based” approach towards dividing data such that predictions can be made. This approach can be applied in countless circumstance and can be just as helpful in a biotech forensic lab as in predicting used car sales prices.

## <u>References:</u>

Car and Driver (n.d.). *Car and Driver Research*. https://www.caranddriver.com/research/a33235783/tundra-vs-tacoma/

Engers, M., Hartmann, M., & Stern, S. (2009). Annual miles drive used car prices. Journal of Applied Econometrics, 24(1), 1-33.

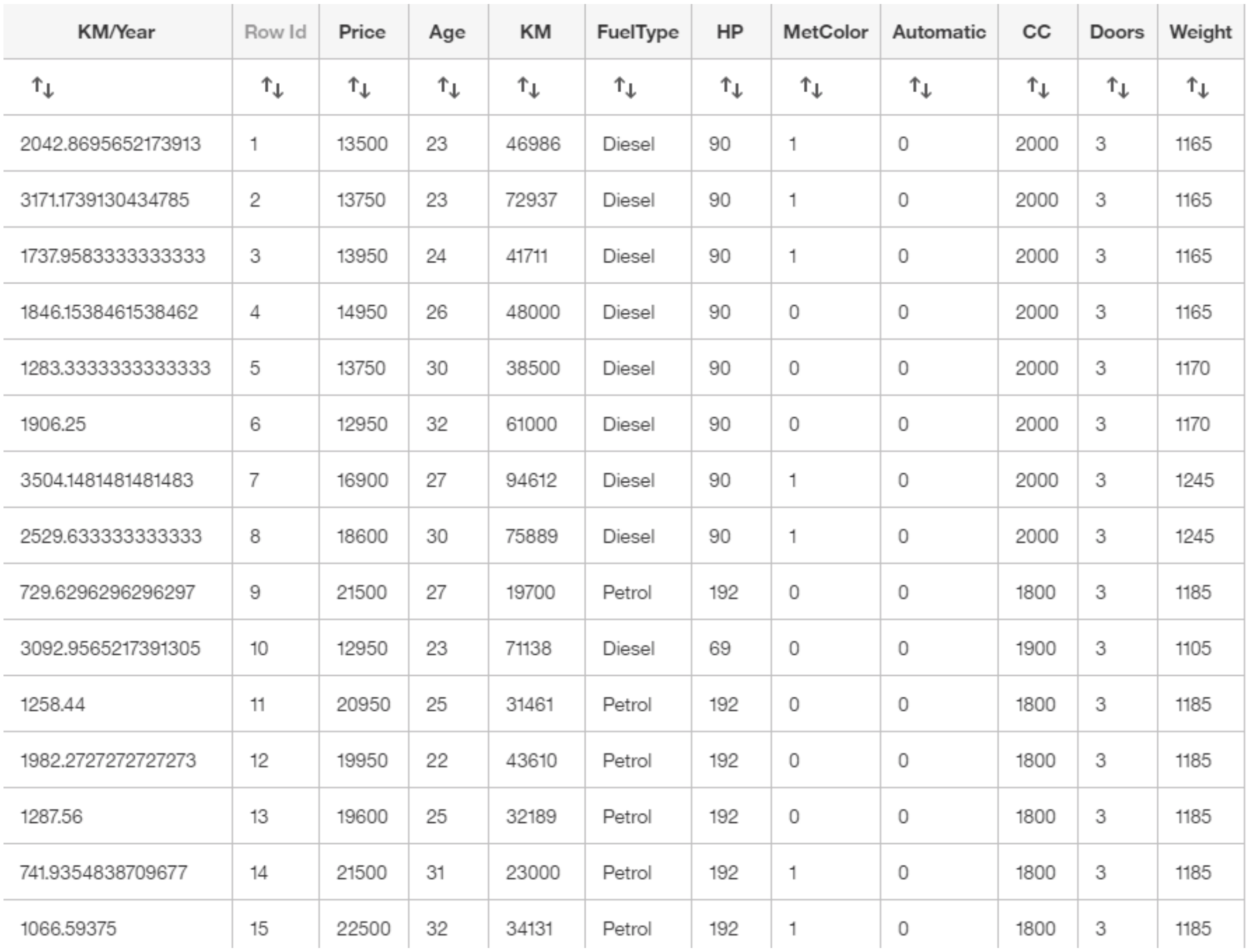
Evans, A. (2013). *Carwow*. https://www.carwow.co.uk/blog/metallic-paint-worth-it-393

Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car price prediction using machine learning techniques. TEM Journal, 8(1), 113.

McCrone (n.d.). *McCrone Atlas of Microscopic Particles*. http://www.mccroneatlas.com/

# Appendix:

**Figure 1. Screenshot of dataset of used cars.**



| KM/Year | Row Id | Price | Age | KM | FuelType | HP | MetColor | Automatic | CC | Doors | Weight |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ↑↓ | ↑↓ | ↑↓ | ↑↓ | ↑↓ | ↑↓ | ↑↓ | ↑↓ | ↑↓ | ↑↓ | ↑↓ | ↑↓ |
| 2042.8695652173913 | 1 | 13500 | 23 | 46986 | Diesel | 90 | 1 | 0 | 2000 | 3 | 1165 |
| 3171.1739130434785 | 2 | 13750 | 23 | 72937 | Diesel | 90 | 1 | 0 | 2000 | 3 | 1165 |
| 1737.9583333333333 | 3 | 13950 | 24 | 41711 | Diesel | 90 | 1 | 0 | 2000 | 3 | 1165 |
| 1846.1538461538462 | 4 | 14950 | 26 | 48000 | Diesel | 90 | 0 | 0 | 2000 | 3 | 1165 |
| 1283.3333333333333 | 5 | 13750 | 30 | 38500 | Diesel | 90 | 0 | 0 | 2000 | 3 | 1170 |
| 1906.25 | 6 | 12950 | 32 | 61000 | Diesel | 90 | 0 | 0 | 2000 | 3 | 1170 |
| 3504.1481481481483 | 7 | 16900 | 27 | 94612 | Diesel | 90 | 1 | 0 | 2000 | 3 | 1245 |
| 2529.633333333333 | 8 | 18600 | 30 | 75889 | Diesel | 90 | 1 | 0 | 2000 | 3 | 1245 |
| 729.6296296296297 | 9 | 21500 | 27 | 19700 | Petrol | 192 | 0 | 0 | 1800 | 3 | 1185 |
| 3092.9565217391305 | 10 | 12950 | 23 | 71138 | Diesel | 69 | 0 | 0 | 1900 | 3 | 1105 |
| 1258.44 | 11 | 20950 | 25 | 31461 | Petrol | 192 | 0 | 0 | 1800 | 3 | 1185 |
| 1982.2727272727273 | 12 | 19950 | 22 | 43610 | Petrol | 192 | 0 | 0 | 1800 | 3 | 1185 |
| 1287.56 | 13 | 19600 | 25 | 32189 | Petrol | 192 | 0 | 0 | 1800 | 3 | 1185 |
| 741.9354838709677 | 14 | 21500 | 31 | 23000 | Petrol | 192 | 1 | 0 | 1800 | 3 | 1185 |
| 1066.59375 | 15 | 22500 | 32 | 34131 | Petrol | 192 | 1 | 0 | 1800 | 3 | 1185 |

**Figure 2. Model 1: Tree Diagram showing drivers of sales price: age, weight, KM, and CC.**

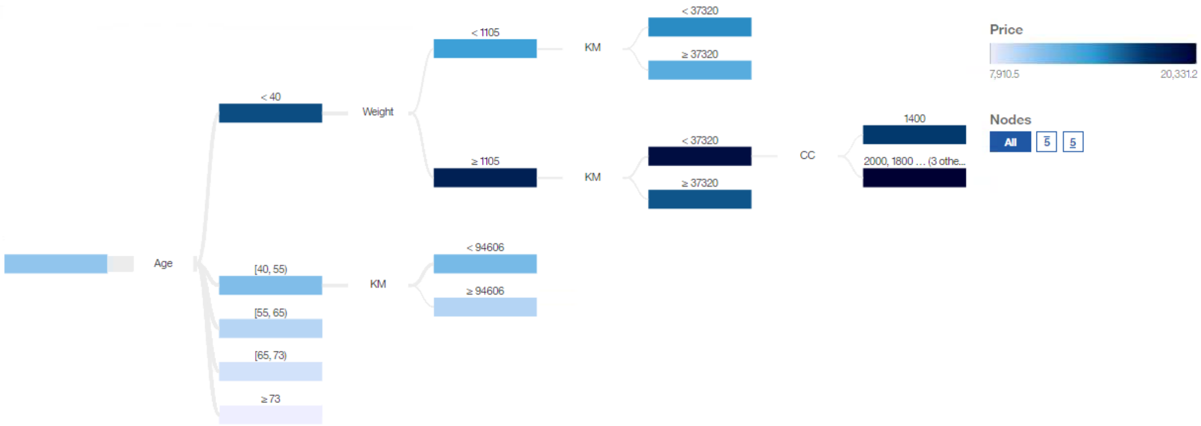
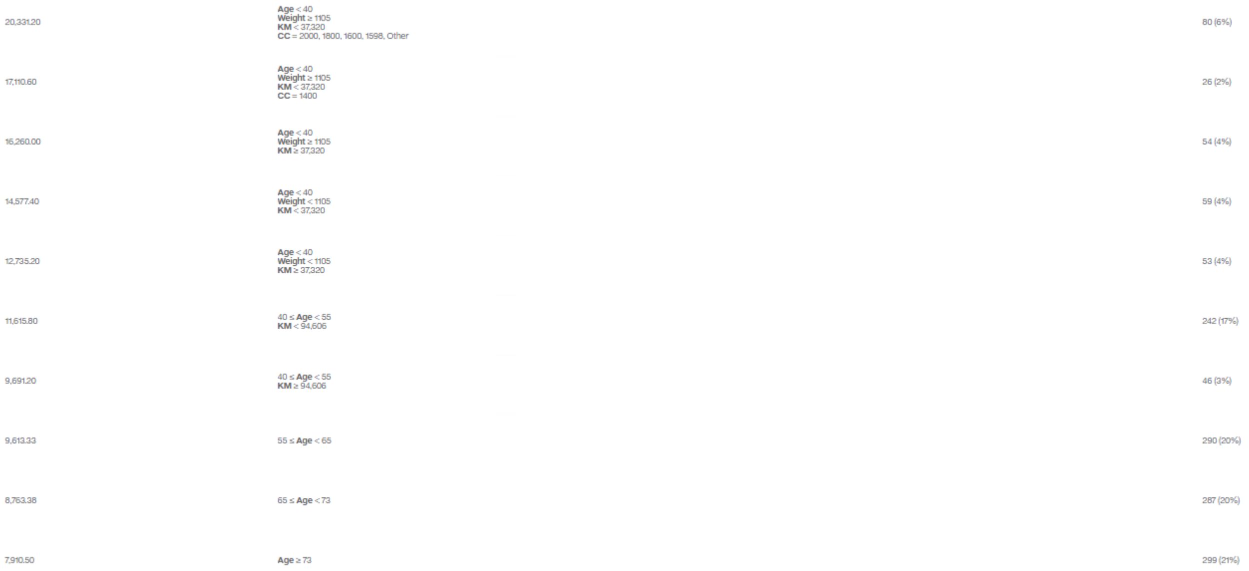


**Figure 3. Graph showing the ruleset for Model 1.**



| | | |
|---|---|---|
| 20,331.20 | **Age** < 40<br>**Weight** ≥ 1105<br>**KM** < 37,320<br>**CC** = 2000, 1800, 1600, 1598, Other | 80 (6%) |
| 17,110.60 | **Age** < 40<br>**Weight** ≥ 1105<br>**KM** < 37,320<br>**CC** = 1400 | 26 (2%) |
| 16,260.00 | **Age** < 40<br>**Weight** ≥ 1105<br>**KM** ≥ 37,320 | 54 (4%) |
| 14,577.40 | **Age** < 40<br>**Weight** < 1105<br>**KM** < 37,320 | 59 (4%) |
| 12,735.20 | **Age** < 40<br>**Weight** < 1105<br>**KM** ≥ 37,320 | 53 (4%) |
| 11,615.80 | 40 ≤ **Age** < 55<br>**KM** < 94,606 | 242 (17%) |
| 9,691.20 | 40 ≤ **Age** < 55<br>**KM** ≥ 94,606 | 46 (3%) |
| 9,613.33 | 55 ≤ **Age** < 65 | 290 (20%) |
| 8,763.38 | 65 ≤ **Age** < 73 | 287 (20%) |
| 7,910.50 | **Age** ≥ 73 | 299 (21%) |

**Figure 4. Tree sunburst diagram for Model 1 showing drivers for sales price.**

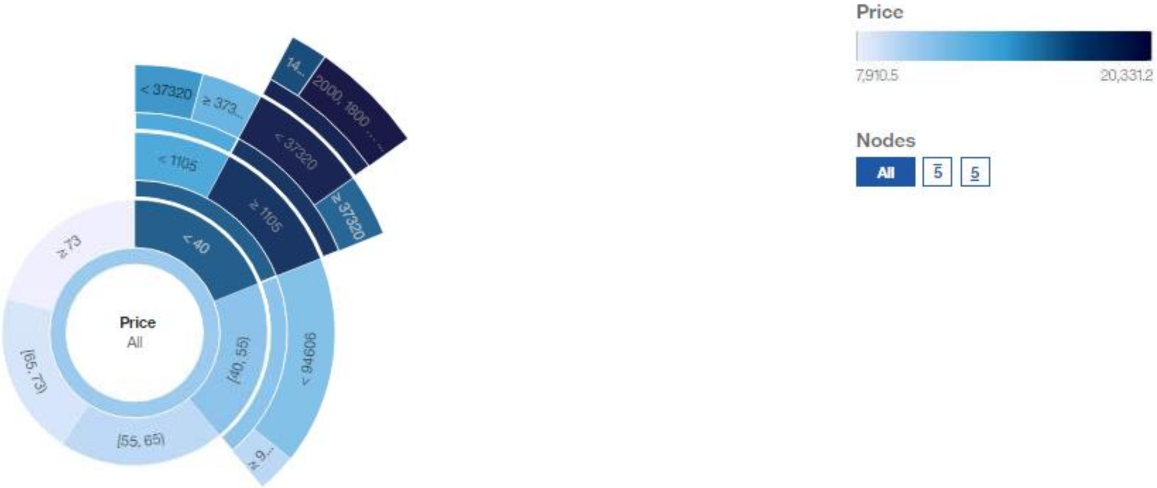


**Figure 5. Model 2: Tree Diagram showing drivers of sales price: KM, CC, door number, and whether the car has a metallic color.**

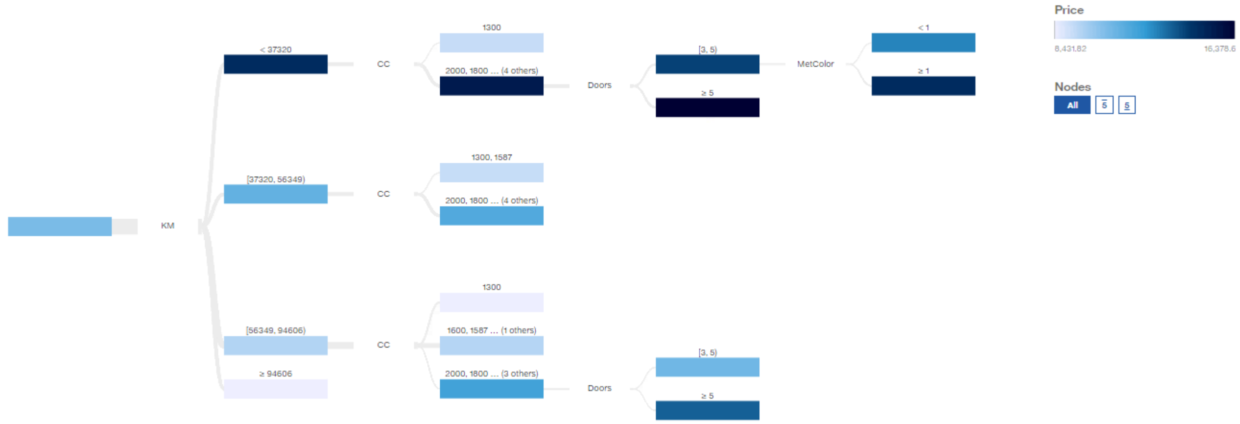


**Figure 6. Tree sunburst diagram for Model 2 showing drivers for sales price.**

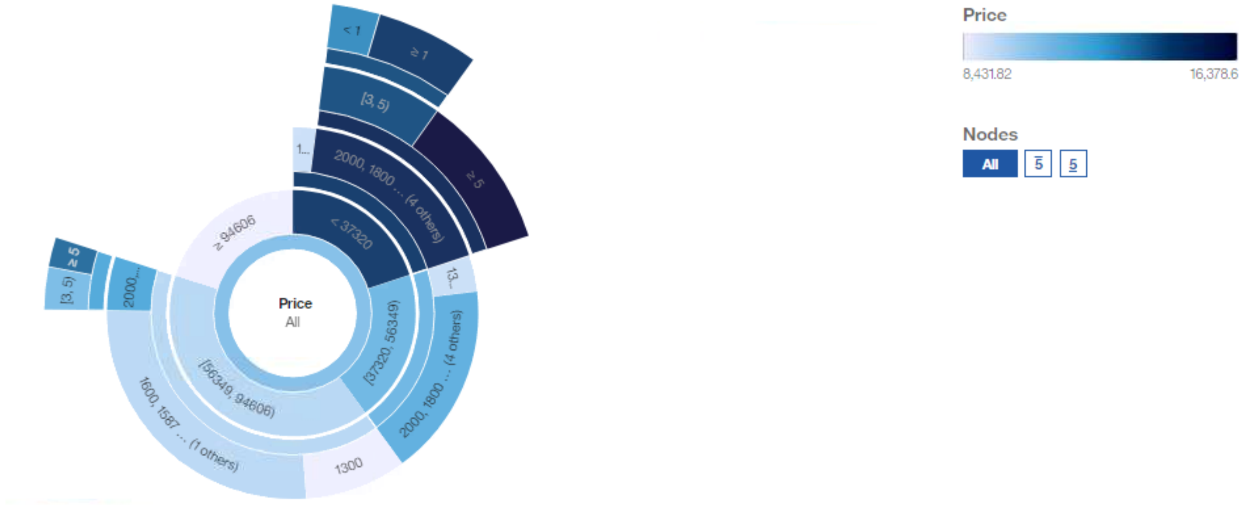


**Figure 4. Tree sunburst diagram for Model 1 showing drivers for sales price.**

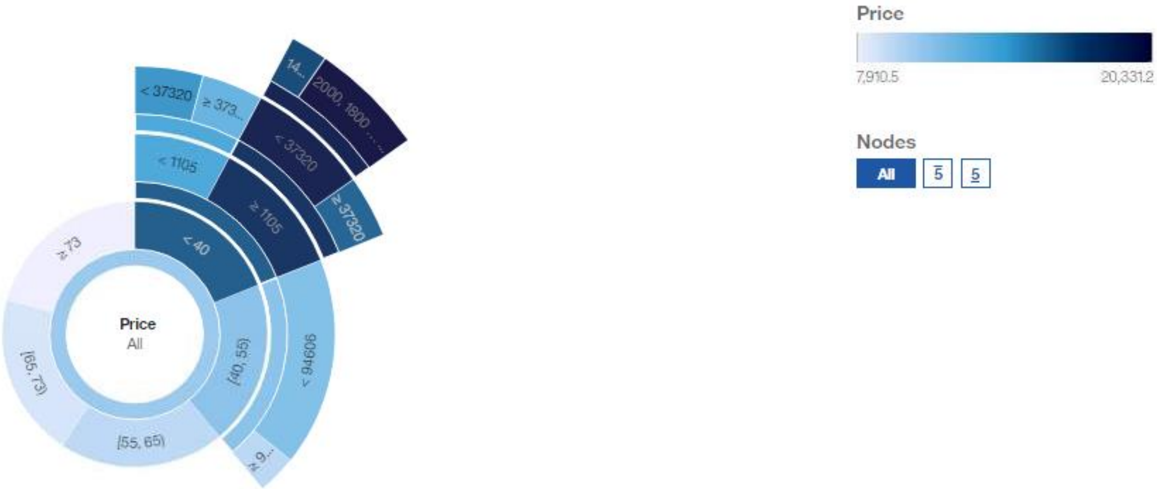
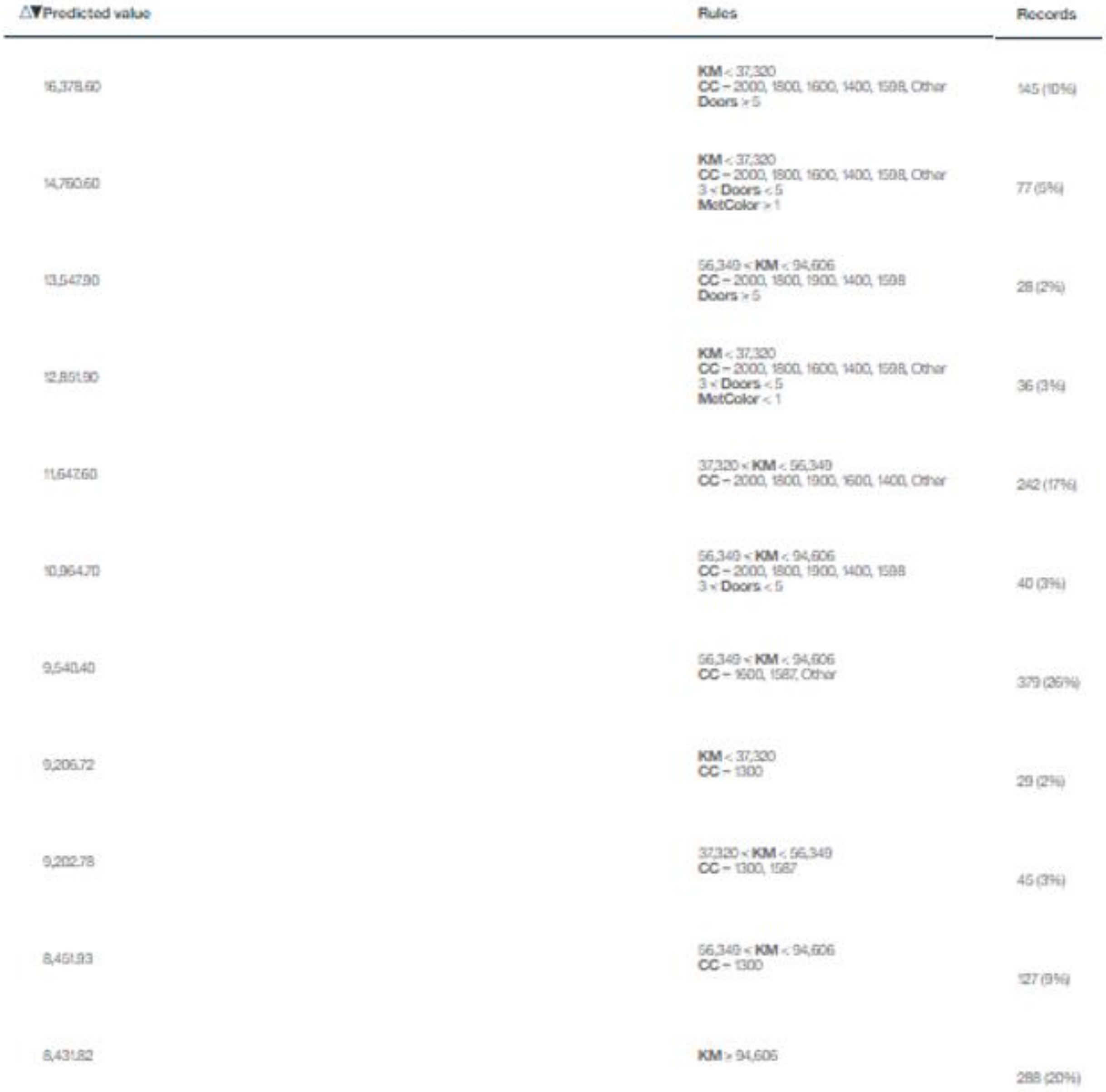


**Figure 57. Graph showing the ruleset for Model 2.**



| △▼Predicted value | Rules | Records |
|---|---|---|
| 16,378.60 | **KM** < 37,320<br>**CC** – 2000, 1800, 1600, 1400, 1598, Other<br>**Doors** > 5 | 145 (10%) |
| 14,760.60 | **KM** < 37,320<br>**CC** – 2000, 1800, 1600, 1400, 1598, Other<br>3 < **Doors** < 5<br>**MetColor** > 1 | 77 (5%) |
| 13,547.90 | 56,349 < **KM** < 94,606<br>**CC** – 2000, 1800, 1900, 1400, 1598<br>**Doors** > 5 | 28 (2%) |
| 12,851.90 | **KM** < 37,320<br>**CC** – 2000, 1800, 1600, 1400, 1598, Other<br>3 < **Doors** < 5<br>**MetColor** < 1 | 36 (3%) |
| 11,647.60 | 37,320 < **KM** < 56,349<br>**CC** – 2000, 1800, 1900, 1600, 1400, Other | 242 (17%) |
| 10,964.70 | 56,349 < **KM** < 94,606<br>**CC** – 2000, 1800, 1900, 1400, 1598<br>3 < **Doors** < 5 | 40 (3%) |
| 9,540.40 | 56,349 < **KM** < 94,606<br>**CC** – 1600, 1587, Other | 379 (26%) |
| 9,206.72 | **KM** < 37,320<br>**CC** – 1300 | 29 (2%) |
| 9,202.78 | 37,320 < **KM** < 56,349<br>**CC** – 1300, 1587 | 45 (3%) |
| 8,451.93 | 56,349 < **KM** < 94,606<br>**CC** – 1300 | 127 (9%) |
| 8,431.82 | **KM** > 94,606 | 288 (20%) |

**Figure 8. Model 3: Tree Diagram showing drivers of sales price: door number, HP and fuel type.**

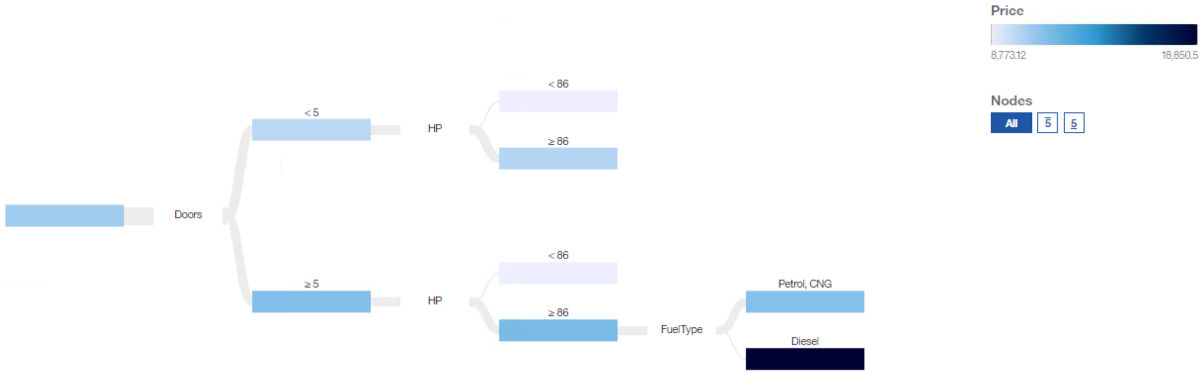


**Figure 4. Tree sunburst diagram for Model 1 showing drivers for sales price.**

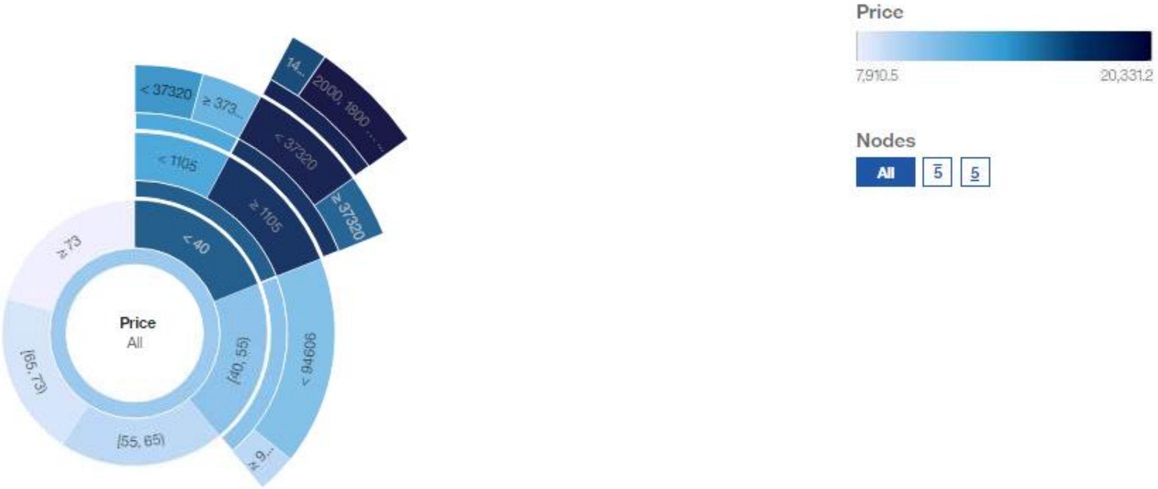
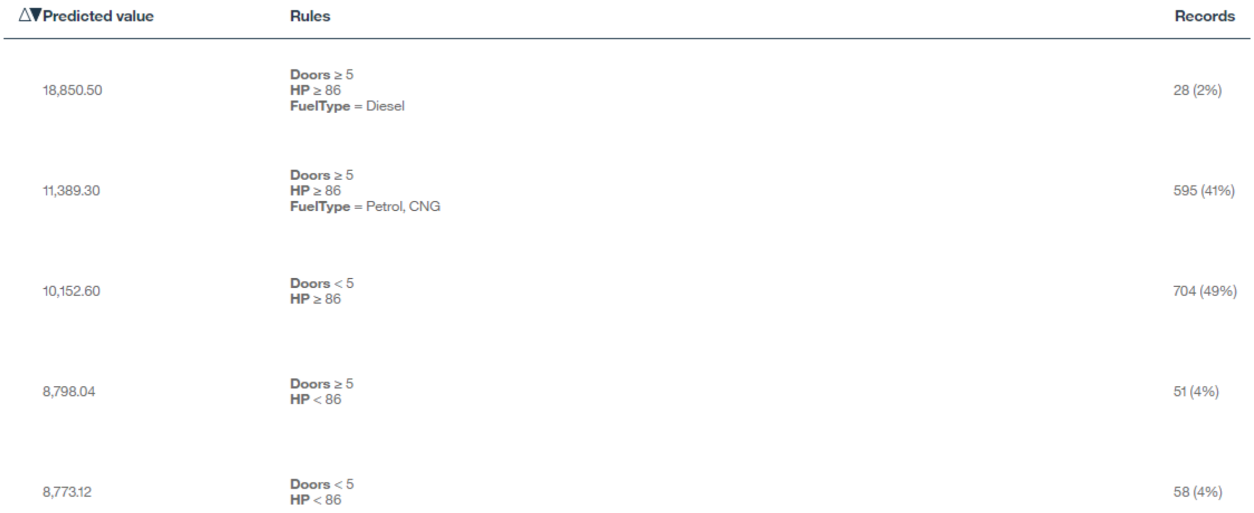


**Figure 59. Graph showing the ruleset for Model 3.**



| △▼Predicted value | Rules | Records |
|---|---|---|
| 18,850.50 | **Doors** ≥ 5<br>**HP** ≥ 86<br>**FuelType** = Diesel | 28 (2%) |
| 11,389.30 | **Doors** ≥ 5<br>**HP** ≥ 86<br>**FuelType** = Petrol, CNG | 595 (41%) |
| 10,152.60 | **Doors** < 5<br>**HP** ≥ 86 | 704 (49%) |
| 8,798.04 | **Doors** ≥ 5<br>**HP** < 86 | 51 (4%) |
| 8,773.12 | **Doors** < 5<br>**HP** < 86 | 58 (4%) |

**Figure 10. Tree sunburst diagram for Model 3 showing drivers for sales price.**

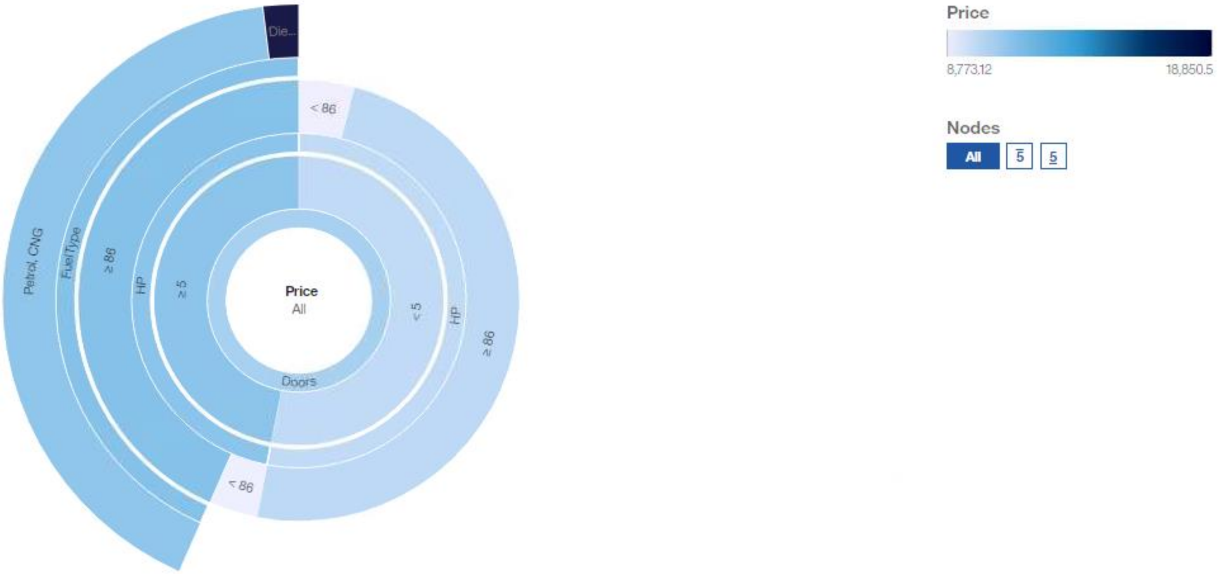


**Figure 11. Model 4: Tree Diagram showing drivers of sales price: age, weight, and CC.**

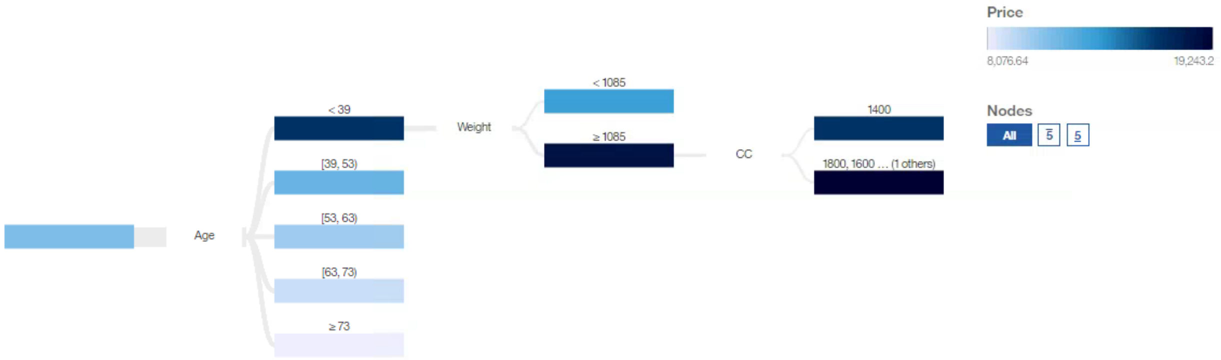
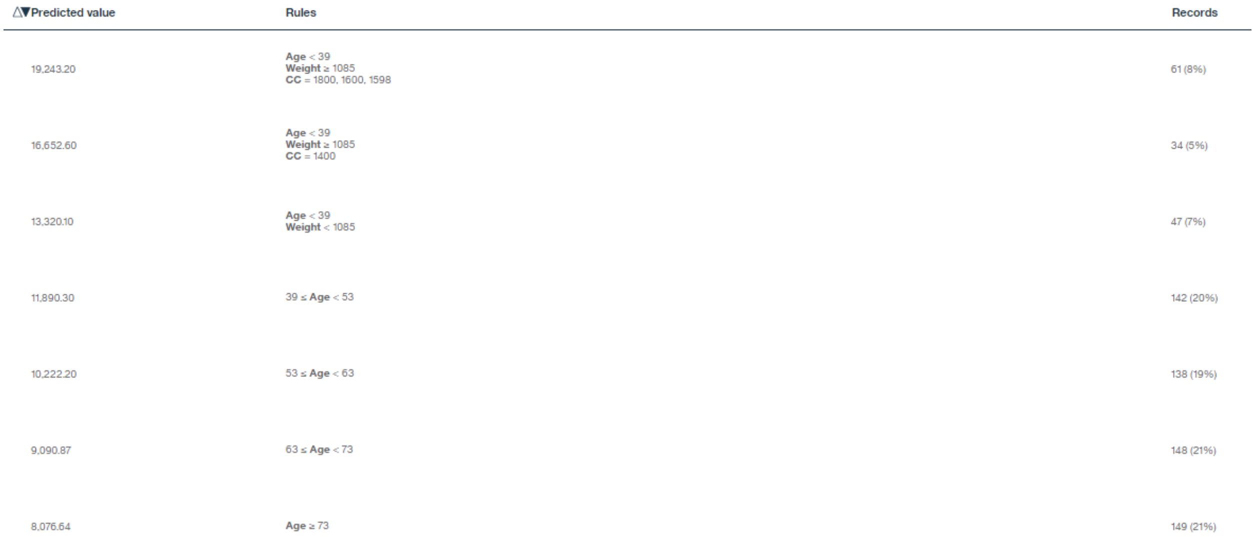


**Figure 12. Graph showing the ruleset for Model 4.**



| △▼Predicted value | Rules | Records |
|---|---|---|
| 19,243.20 | **Age** < 39<br>**Weight** ≥ 1085<br>**CC** = 1800, 1600, 1598 | 61 (8%) |
| 16,652.60 | **Age** < 39<br>**Weight** ≥ 1085<br>**CC** = 1400 | 34 (5%) |
| 13,320.10 | **Age** < 39<br>**Weight** < 1085 | 47 (7%) |
| 11,890.30 | 39 ≤ **Age** < 53 | 142 (20%) |
| 10,222.20 | 53 ≤ **Age** < 63 | 138 (19%) |
| 9,090.87 | 63 ≤ **Age** < 73 | 148 (21%) |
| 8,076.64 | **Age** ≥ 73 | 149 (21%) |

**Figure 4. Tree sunburst diagram for Model 1 showing drivers for sales price.**

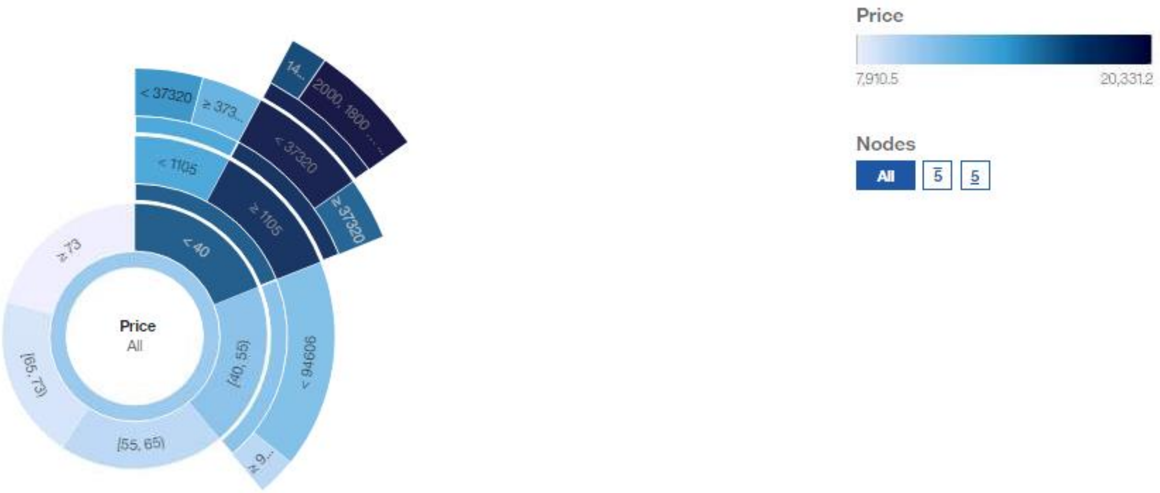
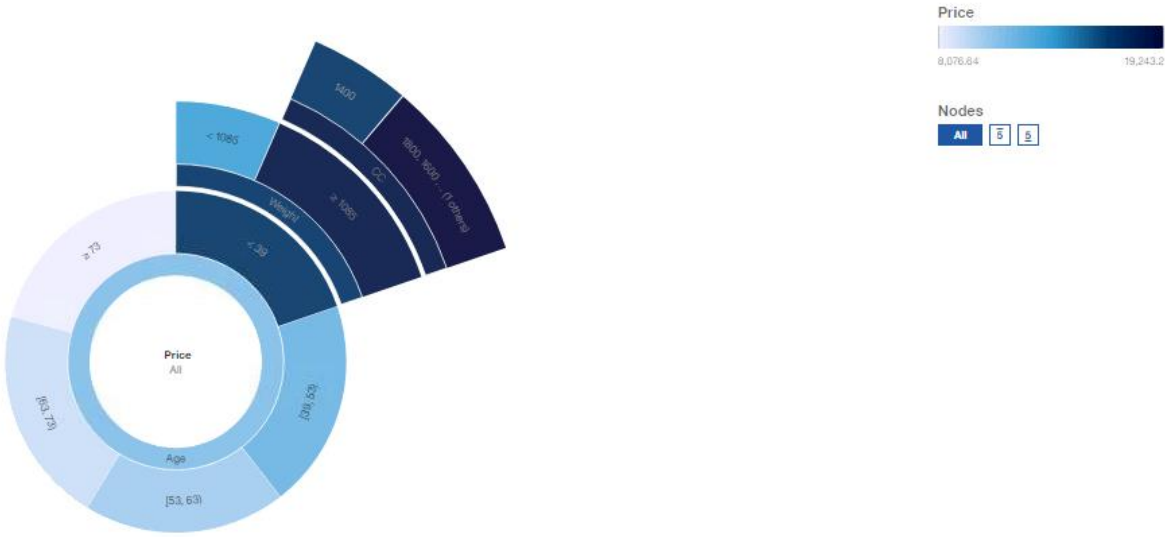


**Figure 513. Tree sunburst diagram for Model 4 showing drivers for sales price.**


Price
All
Age
< 39
[39, 53)
[53, 63)
[63, 73)
≥ 73
Weight
< 1085
≥ 1085
CC
1800, 1600, ... (1 others)
Price
Nodes
All