Assignment 3 - Decision Tree Model Development using Cognos Analytics

Tanushree Kumar | tanushree.kumar@outlook.com

DATA 610 - Fall 2023

Dr. George Cross

The dataset that I will be using for decision tree model development using Cognos Analytics is the Used Car Prices (reddy, 2018). The dataset contains 1436 observations across 12 columns on various attributes of the vehicles. They are:

- Price: price of the car
- Age: age of the car
- KM: car mileage in kilometers
- FuelType: type of fuel used (petrol, diesel, or CNG)
- HP: Horsepower
- MetColor: is the car a metallic color (1 = yes, 0 = no); equivalent to 'MetColorType'
- Automatic: is the car automatic (1 = yes, 0 = no); equivalent to 'AutoType'
- CC: volume of the cylinder in cubic centimeters
- Doors: number of doors
- Weight: weight of the car

While the dataset has many attributes, some of them are unnecessary for the purpose of this assignment. I removed the following columns as they had no significant impact: HP, MetColor, Automatic, Doors, Autotype, and MetColorType. A snapshot of the original dataset can be seen in Figure 1 in Appendix A.

The objective of this assignment is to develop two predictive models and discuss their results. For the first predictive model, I chose the target variable to be the 'Price' of the used car based on 'Age', 'Weight', 'KM', and 'CC' variables. For the second predictive model, I chose the target variable to be 'Age' based on 'Price', 'CC', 'Weight', 'KM', and 'FuelType' variables.

I was able to identify these specific variables to predict the outcomes by first making a spiral visualization for both target variables, which highlighted the drivers.

With the pandemic and the current market, buying either a new or a used car is a significant financial decision that needs to account for many factors. Investing in a car requires a lot of thought and consideration, and the price and age of a car are at the top of the list. Used cars have already experienced the initial period of depreciation that occurs rapidly in the first few years of a car's life. By purchasing a used car, one can avoid the steepest part of the depreciation curve, leading to better long-term value for your investment. However, this trend has changed since the pandemic. Although prices have somewhat reduced since the height of the pandemic, "the average new-vehicle transaction price is still $48,763, according to Kelley Blue Book. Before the pandemic, the average new vehicle sold for $37,876" (Domonoske, 2023). Similarly, "the used-car market prices now average $26,510" (Domonoske, 2023). The price of a used car directly influences its affordability. It's essential to align the cost of the car with your budget to ensure financial stability.

While in the past it might have been more beneficial to buy a used car due to a lower rate of depreciation, it is unfortunately no longer the same case. Age is closely linked to the price as "cars that are only a year or two old will not have depreciated much, and may even cost close to what they did when they were new" (Preston, 2023). With that in mind, it might be better to purchase cars that are 3 to 5 years old, or maybe older as "many of them are just coming off lease and have been well-maintained" (Preston, 2023). Thus, it is imperative to look into the relations of 'Price' and 'Age' and its related variables to understand how they influence each other and other factors in an expensive market.

For the ‘Price’ prediction model, the combination of ‘Age’, ‘Weight’, ‘KM’, and ‘CC’ drivers had a predictive strength of 80%. I observed the same predictive strength of 80% for the ‘Price’, ‘CC’, ‘Weight’, ‘KM’, and ‘FuelType’ drivers for ‘Age’. Since predictors with less than 10% strength are omitted, I chose to edit the drivers for both models by removing ‘Weight’. I also didn’t see a logical relation between the cost of the car, its age, and its weight. This brought down the predictive strength of the ‘Price’ model to 75% and the ‘Age’ model to 79%. Although it affected the strength of the models, the ‘Weight’ variable wasn’t strong enough to be a driver by itself for both models which further reinforced my decision to remove it. The spiral visualizations for both models can be seen in Figure 2 and Figure 3 respectively in Appendix A.

Looking at the driver analysis gives a great insight into how each driver works for the first predictive model, whether it's a combination of drivers or single drivers. The driver analysis gives an overarching statement that “‘Age’ is the most important predictor of ‘Price’ as ‘Age’ is more than five times as important as any other field” This makes ‘Age’ the single best predictor variable. ‘Age’, as we have seen, by itself as a single driver is a very strong variable as it has a strength of 68%. By combining it with other variables, we can see a stronger relation. While ‘Age’, ‘KM’, ‘CC’, and ‘HP’ predict ‘Price’ with the strongest predictive strength of 75.3%, the next highest predictive strength of 74% is given by a combination of ‘Age’ and ‘CC’ whereas as ‘Age’ and ‘KM” have a strength of 73%. Although backed up by numbers, logically speaking, age has a powerful impact on the price of a used car as the value of a vehicle depreciates rapidly in the first few years. However, the combination of variables that gives a strength of 75% seems to give the best answer as the price of a car can not be assessed accurately from just one factor. A good combination of variables will give a comprehensive and overarching understanding of what makes up the final price.

The decision tree generated from using a combination driver gave 9 rules in descending order, starting from the highest price value. Each rule shows the natural language explanation for each predicted value and the number of data rows. The rules can be seen in Figure 6 in Appendix A. I noticed that no rule included all four variables. Taking a look at the first rule, we can see that only 108 rows, 8% of the total number of cases, satisfy the first rule which is: Age < 40, KM > 37,320, and HP ≥ 110, which gave a predicted value for 'Price' of 18,501.60. Whereas, the second rule had the same conditions for 'Age' and 'KM' but included CC = 2000, 1800 instead of 'HP which gave a predicted value of 17,177.80. I will discuss how the idea of these rules can be implemented in organizations later.

The driver analysis for the second predictive model gave a similar observation to the first predictive model:'"'Price' is the most important predictor of 'Age' as 'Price' is more than five times as important as any other field". This makes 'Price' the single best predictor variable. A lot of similarities can be observed here between the 'Price' and 'Age' models as ' Price' by itself as a single driver is a very strong variable with a strength of 74%. While 'Price', 'CC', 'KM', and 'FuelType' predict 'Age' with a strength of 79.2%, the next highest predictive strength of 78% is given by the combination of 'Price' and 'CC'. Even though the predictive strength of 'Price' and 'CC' alone is quite strong, the addition of the 'KM' and 'FuelType' makes a difference when it comes to predicting age. I was surprised to see that 'FuelType' was a combined driver for 'Age' but not 'Price' since "diesel fuel and diesel vehicles often cost more, diesel vehicles may be more prone to hold their resale value and get better mileage compared to gas vehicles" (*Diesel vs. Gas Engines*, n.d.).

The decision tree generated from using the combination driver gave 8 rules that were in descending order starting from the highest age value. The rules can be seen in Figure 7 in

Appendix A. The first and third rule allows us to compare how ‘FuelType’ can potentially help in predicting age if ‘Price’ is held as a constant value. The first rule shows that only 396 rows, 28% of the total cases, satisfy this condition: Price < 8950, FuelType = Petrol which gave a predicted age of 72.06 years. Comparatively, the third rule shows that only 73 rows, 5% of the total cases, satisfy this condition: Price < 8950, FuelType = Diesel, CNG which gave a predicted age of 64.78 years. Since the age of diesel and CNG cars is younger, a used car dealership could potentially use this model to reevaluate its inventory and choose to display more diesel and CNG-based cars since they are younger than their petrol counterparts. A younger car would allow them to price it adequately, or maybe even higher to break even in profit.

The decision tree models provide a set of rules derived from data patterns that offer valuable insights for strategic decision-making in an organization. The rules serve as guidelines for predicting outcomes based on input variables which make them useful in various business contexts. Let’s consider the context of a used car dealership where they can develop a pricing strategy by using the rules derived from a decision tree model predicting the 'Price'. This can inform the establishment of competitive and dynamic pricing strategies. The rules guide the organization on how different factors such as 'Age,' 'Weight,' 'KM,' and 'CC' contribute to the overall pricing structure since they’re able to see these variables play a larger role than other factors that affect the price minimally, such as ‘MetColor’ for example. The dealership would be able to manage their inventory accordingly since they’ll understand the factors influencing the 'Age' of used cars, as predicted by the decision tree model. The rules will help them to identify the key variables, such as 'Price,' 'CC,' 'Weight,' 'KM,' and 'FuelType,' that impact the age of cars in the inventory. If they’re able to leverage this and can secure cars that haven’t aged as much, they’ll be able to take advantage of the current high prices and sell more of their inventory for a

good price. The decision tree rules can also be used for marketing purposes. For instance, since 'Age' is primarily driven by 'Price,' the organization can tailor their marketing campaigns that emphasize its competitive pricing or unique financing options. All these tactics can boost their sales and total profit and potentially help them expand their business.

Implementing a rule-based approach involves integrating the derived rules into the organization's decision-making processes. This can be achieved through strategic steps such as:

- Automated systems: embedding decision tree models into automated systems within the organization. For example, integrating the rules into the pricing system to automatically adjust prices based on the identified influential variables.
- Training and education: ensuring that relevant stakeholders are trained to interpret and apply decision tree rules.
- Continuous monitoring: establishing a system for continuous monitoring and updating of decision tree rules. As the market conditions evolve, regular recalibration of the models ensures that the rules remain relevant and effective.

The acceptance of these decision tree models depends on factors such as their accuracy, interpretability, and alignment with organizational goals. Having an accurate model would mean regularly validating the accuracy of the decision tree models against real-world outcomes. Demonstrating the models' predictive power would increase confidence among stakeholders. It's also important that the decision tree rules are presented in an interpretable manner. Making visualizations such as sunburst diagrams, can simplify complex rules, making them accessible to a broader audience. It will be more beneficial if feedback from stakeholders can be incorporated as this will help further refine the model. Seeking feedback would also increase acceptance from the stakeholders and make them feel more included in this process. The feedback would also

ensure that the models seamlessly blend into the existing organizational processes. The ideal scenario is that the models would complement, rather than disrupt, the established workflows to facilitate smoother adoption. The decision tree models and rules can become valuable tools for strategic decision-making by aligning them with the organizational objectives.

I next chose to make a sunburst visualization as the decision tree diagram doesn't clearly show the proportion of data that each node represents (Bytenskaya, 2021). Thus, the usage of the sunburst diagram can allow us to see the goal of what drives the target variable concisely. The visualizations for both models can be seen in Figure 8 and Figure 9 in Appendix A. The size of each segment in the Sunburst diagram corresponds to the importance of the variable at a particular decision node. In the 'Price' prediction model, we can see that 'KM', 'HP, and 'CC' influence the price of the car more when the age is under 40 years. In the 'Age' prediction mode, we can see that 'KM' and 'CC' affect the age when the price is above 12,500 units, whereas 'FuelType' affects the age when the price is between the range of <8950 and 'CC' when the price is between [8950, 10,500] units.

To further explore this dataset in the future, I would make a third decision tree model based on the 'KM' target variable. I would choose this variable as it's important to consider how many kilometers the car has traveled to gauge its condition. While mileage won't be the sole factor that affects the purchase of a used car, other factors such as price, age, and overall condition also come into play. It will also be interesting to see if there is a relation between 'KM' and 'FuelType' and differentiate between petrol, diesel, and CNG cars. This will give the customer a comprehensive view and understanding of what they should look for specifically when investing in a used car and what combination of factors will yield the most desirable result.

References

Bytenskaya, Y. (2021). *Cognos Analytics - Sunburst decision rules visualization*. Www.youtube.com. https://www.youtube.com/watch?v=67hHz_zAQ3M&list=PL2r2WGYKOnJWvwjzgazpro6MzdmvLZqBj&index=19

Diesel vs. Gas Engines. (n.d.). Www.progressive.com. https://www.progressive.com/answers/gas-vs-diesel-cars/

Domonoske, C. (2023, March 18). *Why car prices are still so high — and why they are unlikely to fall anytime soon*. NPR. https://www.npr.org/2023/03/18/1163278082/car-prices-used-cars-electric-vehicles-pandemic#:~:text=July%209%2C%202021.-

Preston, B. (2023, August 25). *How to Buy a Used Car*. Consumer Reports. https://www.consumerreports.org/cars/buying-a-car/how-to-buy-a-used-car-a5221672417/

Reddy, N. (2018). usedcarprices. Www.kaggle.com. https://www.kaggle.com/datasets/nitinreddy98/usedcarprices/data

Appendix A

All mentioned tables and figures throughout the paper can be found here.

Figure 1

Snippet of the Used Car Prices dataset

Row Id	Price	Age	KM	FuelType	HP	MetColor	Automatic	CC	Doors	Weight	AutoType	MetColorType
1	13500	23	46986	Diesel	90	1	0	2000	3	1165	Manual	Metcolor
2	13750	23	72937	Diesel	90	1	0	2000	3	1165	Manual	Metcolor
3	13950	24	41711	Diesel	90	1	0	2000	3	1165	Manual	Metcolor
4	14950	26	48000	Diesel	90	0	0	2000	3	1165	Manual	NonMetcolor
5	13750	30	38500	Diesel	90	0	0	2000	3	1170	Manual	NonMetcolor
6	12950	32	61000	Diesel	90	0	0	2000	3	1170	Manual	NonMetcolor
7	16900	27	94612	Diesel	90	1	0	2000	3	1245	Manual	Metcolor
8	18600	30	75889	Diesel	90	1	0	2000	3	1245	Manual	Metcolor
9	21500	27	19700	Petrol	192	0	0	1800	3	1185	Manual	NonMetcolor
10	12950	23	71138	Diesel	69	0	0	1900	3	1105	Manual	NonMetcolor

Figure 2

Spiral visualization of the 'Price' prediction model

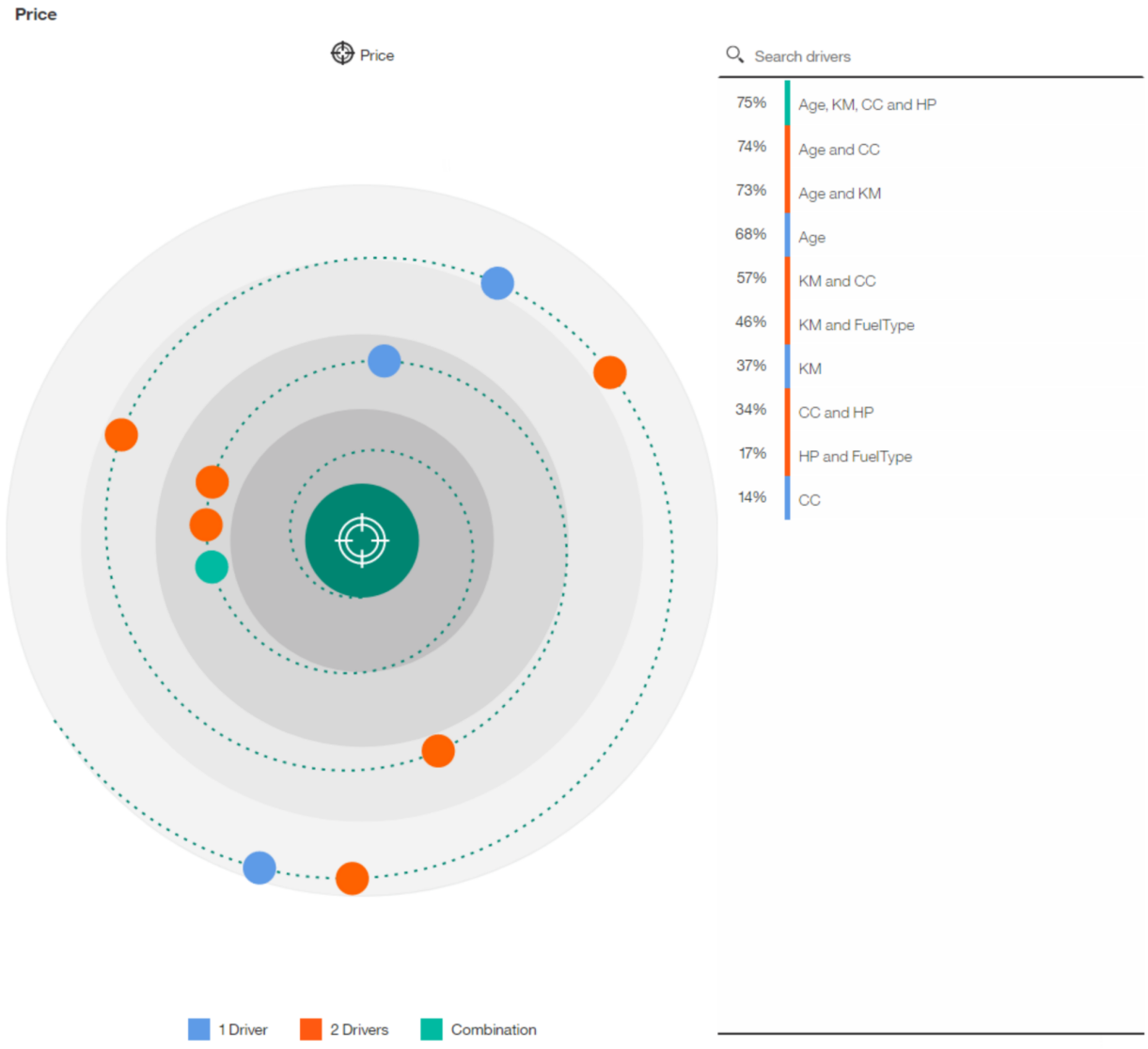

Figure 3

Spiral visualization of the 'Age' prediction model

Figure 4

Sunburst diagram for the 'Age' prediction model

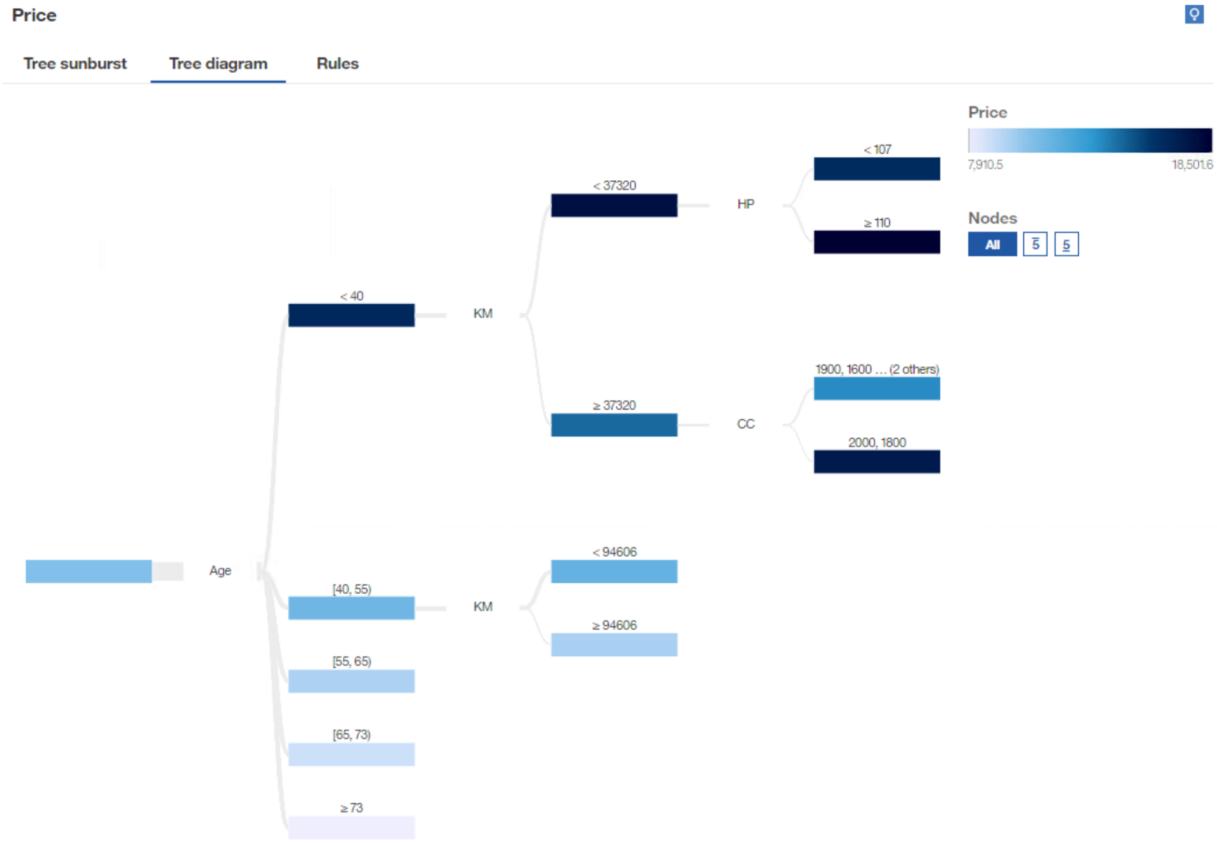

Figure 5

Decision tree diagram for the 'Age' prediction model

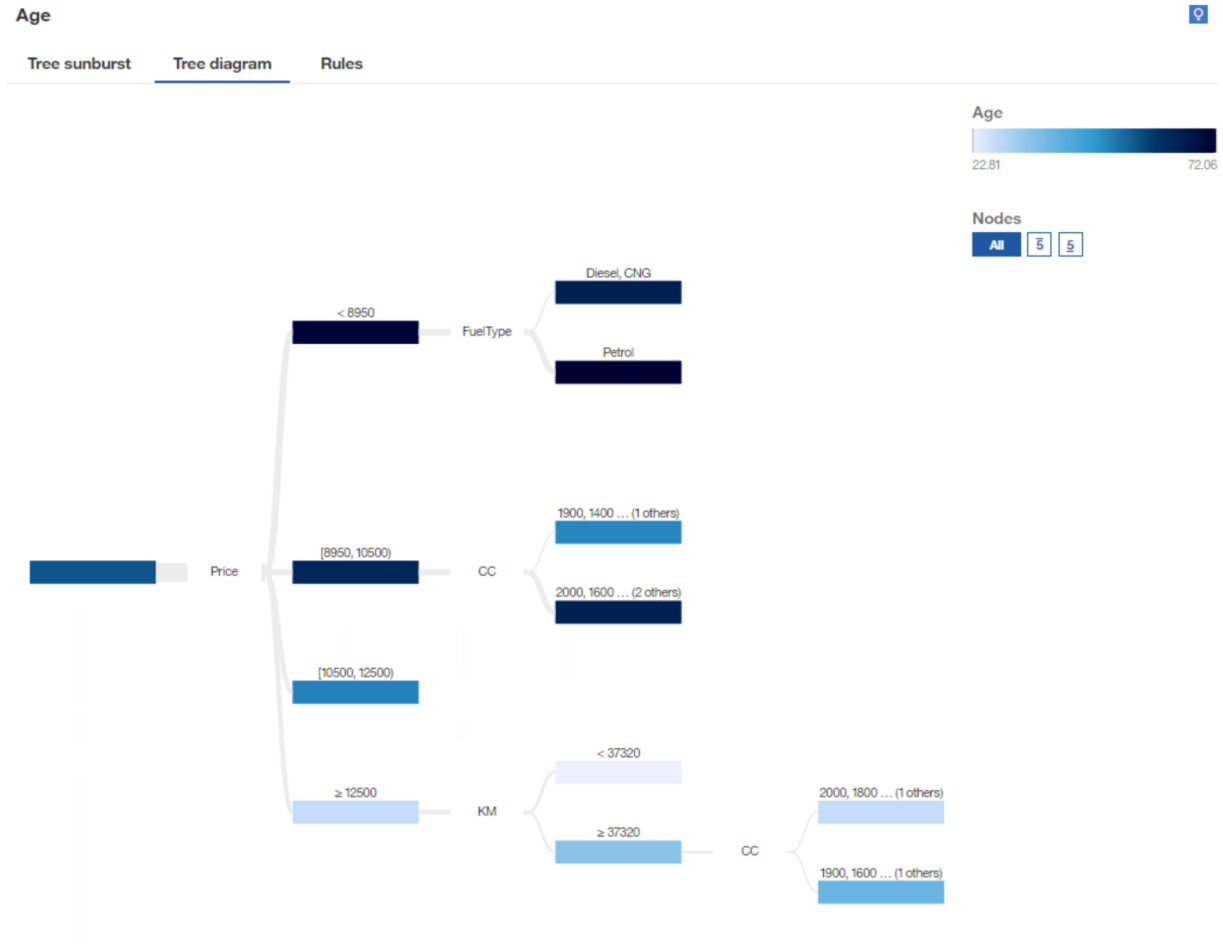

Figure 6

Sunburst diagram for the 'Price' prediction model

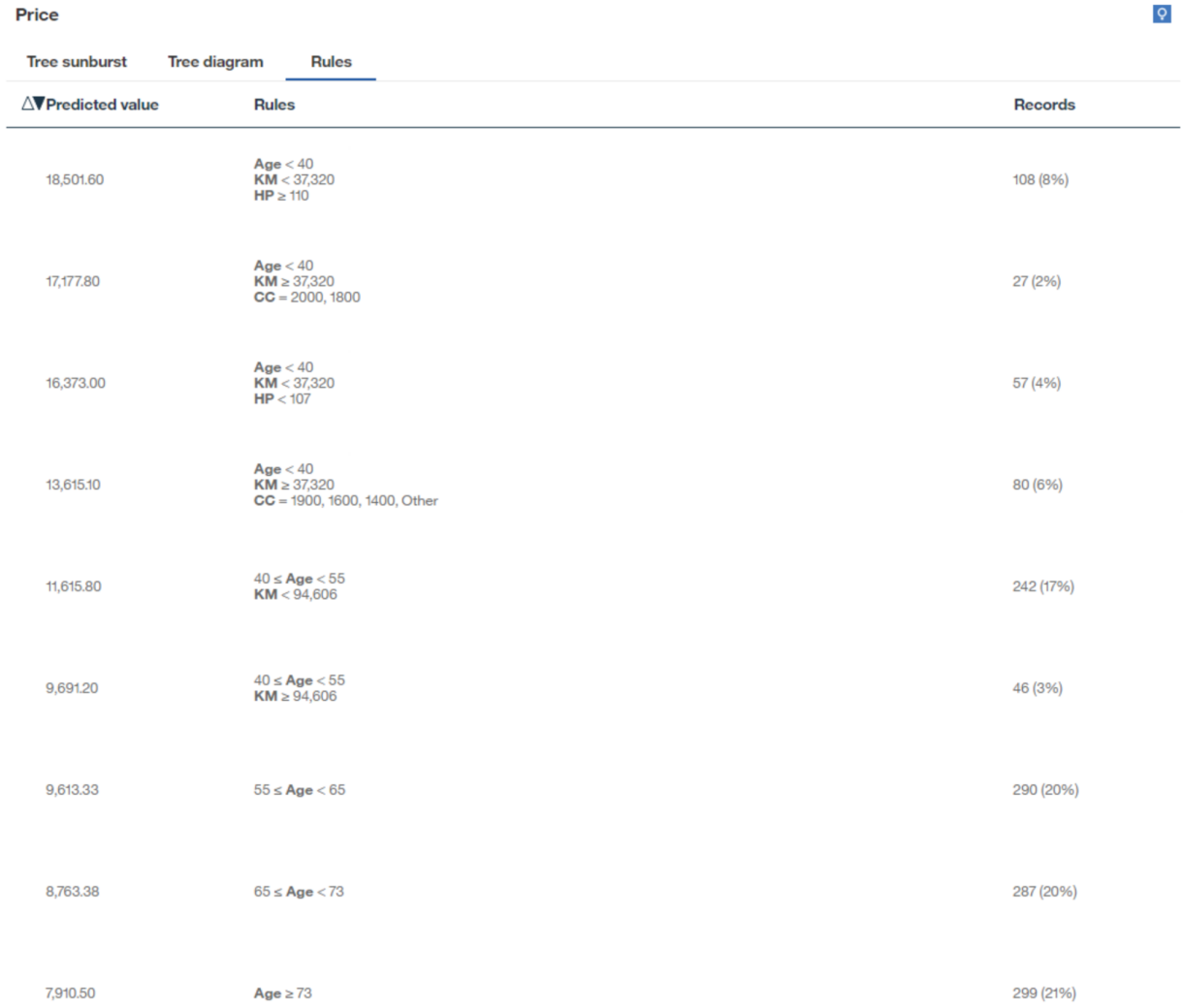

Figure 7

Rules for the 'Age' prediction model

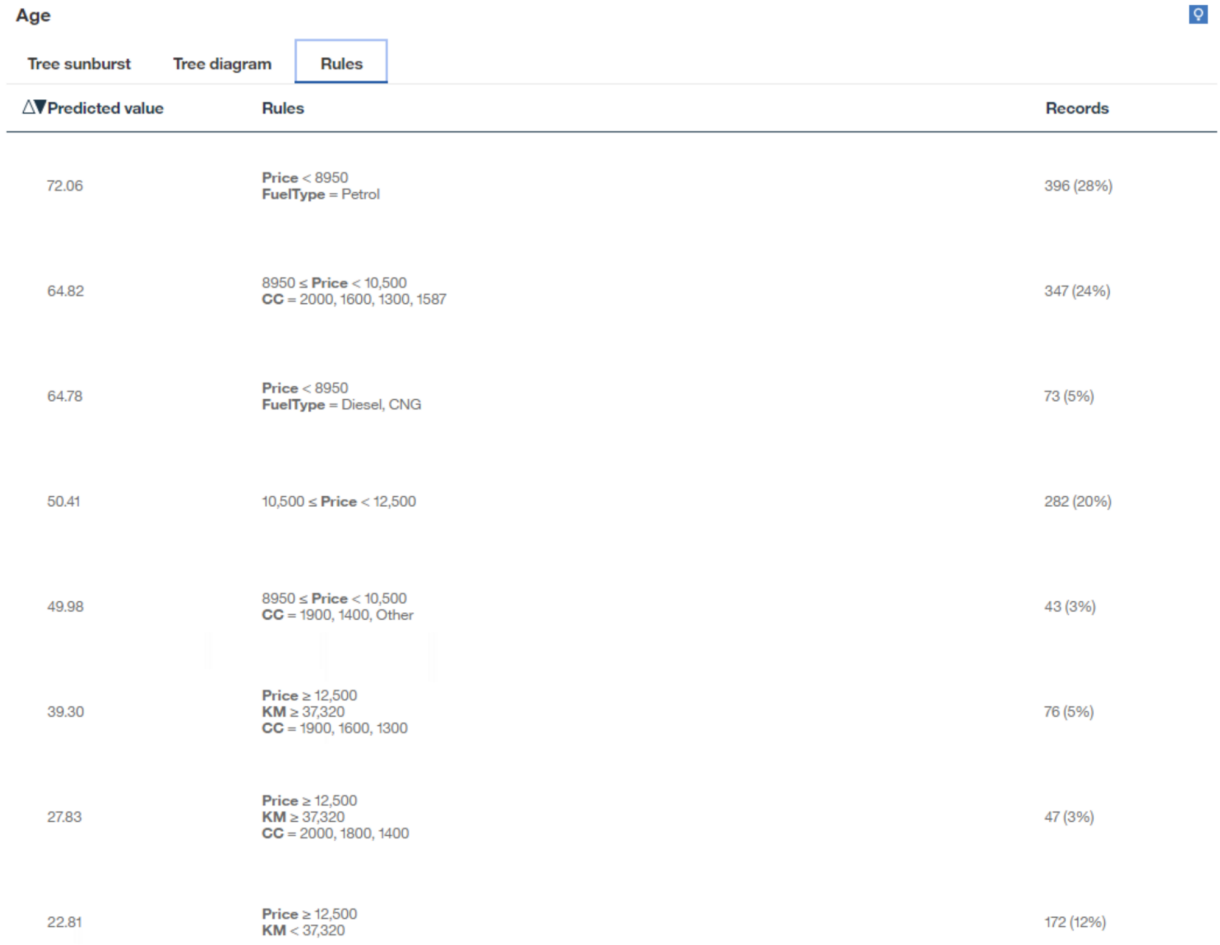

Figure 8

Sunburst diagram for the 'Price' prediction model

Figure 9

Sunburst diagram for the 'Age' prediction model