



ANALIZA DANYCH ANKIETOWYCH – PRZEGLĄD WYBRANYCH TECHNIK NA PRZYKŁADZIE RYNKU MOTORYZACYJNEGO

*Mariusz Łapczyński, Uniwersytet Ekonomiczny w Krakowie,
Katedra Analizy Rynku i Badań Marketingowych*

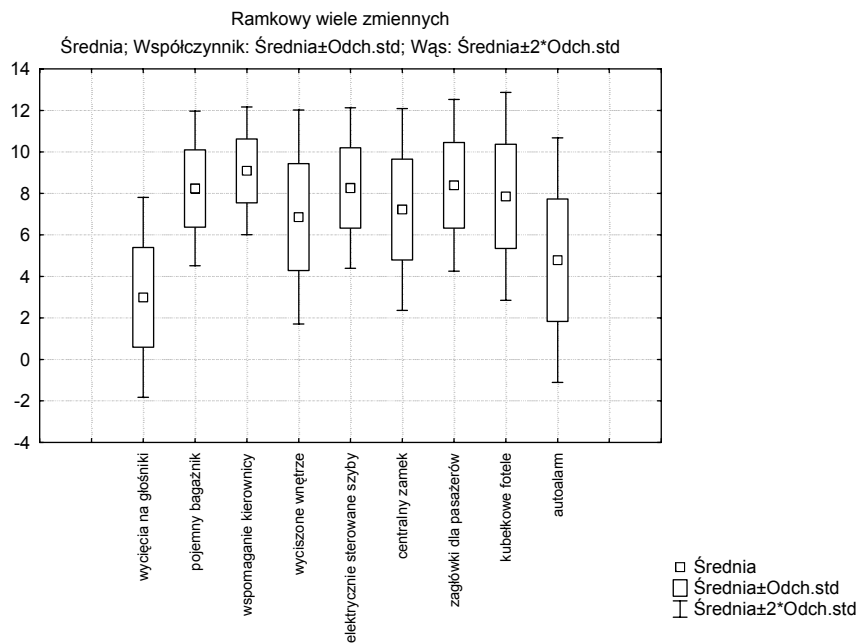
Poznanie odpowiedzi na pytania – wstępna analiza przekrojowa

Punktem wyjścia w analizie danych ankietowych jest pogrupowanie surowych danych za pomocą prostych technik i narzędzi wizualizacji danych. Grupowanie statystyczne jest próbą usystematyzowania materiału statystycznego (danych liczbowych) w celu odzwierciedlenia struktury badanej zbiorowości. Wyróżnić można grupowanie typologiczne, grupowanie wariacyjne oraz grupowanie analityczne. Grupowanie typologiczne stosujemy wówczas, gdy interesująca nas cecha jest zmienną jakościową, zaś grupowanie wariacyjne wówczas, gdy jest to cecha ilościowa. W grupowaniu analitycznym zestawia się dwie lub więcej cech w celu zbadania współzależności między nimi.

W poniższej tabeli przedstawiono rozkład procentowy odpowiedzi na pytanie o fazę cyklu życia rodziny respondentów. Jest to przykład grupowania typologicznego, gdzie badacz znajdzie informacje o liczbie i procencie badanych należących do poszczególnych kategorii tej zmiennej.

Faza cyklu życia rodziny	Tabela licznosci: faza cyklu zycia rodziny			
	Liczba	Skumulow. Liczba	Procent	Skumulow. Procent
młode osoby stanu wolnego	117	117	25	25
młode małżeństwa bez dzieci	33	150	7	32
małżeństwa z dziećmi do 18 roku życia	175	325	38	70
małżeństwa z dziećmi > 18 roku życia	63	388	14	83
puste gniazda	63	451	14	97
starsze osoby stanu wolnego	15	466	3	100

W przypadku zmiennych ilościowych (np. pozycji ze skali struktury korzyści, pozycji ze skali semantycznej czy skali Likerta) możliwe jest zastąpienie grupowania wariacyjnego (tu: szeregu rozdzielczego z cechą ilościową) stosownym wykresem dla danych ilościowych. Na poniższym wykresie typu „ramka-wąsy” przedstawiono odpowiedzi respondentów na pytanie o preferowane cechy nabywanego przez nich pojazdu. Każdej cesze samochodu przypisywano ocenę z przedziału od 1 do 10, będącą odzwierciedleniem stopnia pożądania danego atrybutu. Widać wyraźnie, że największym zainteresowaniem badanych cieszyły się: wspomaganie kierownicy, pojemny bagażnik i elektrycznie sterowane szyby.



Popularnym przykładem grupowania analitycznego są tabele kontyngencji. Najprostsza analiza danych zawartych w tabelach kontyngencji wiąże się z obliczeniem procentów w trzech kierunkach. Chodzi o wyliczenie proporcji liczebności w poszczególnych polach tabeli względem sumy z wiersza (pierwszy kierunek), względem sumy z kolumny (drugi kierunek) lub względem sumy z całości (trzeci kierunek). W pierwszym przypadku suma z wiersza stanowi 100%, a interpretacja dotyczy każdego wiersza z osobna – sprawdza się rozkład zmiennej niezależnej w każdym wariancie zmiennej zależnej. W drugim przypadku to suma z każdej kolumny wynosi 100%, a badacz sprawdza rozkład zmiennej zależnej w każdym wariancie zmiennej niezależnej. W trzecim przypadku suma proporcji z wszystkich pól tabeli stanowi 100%, a badacz wyciąga wnioski dotyczące całej populacji w oparciu o zmienne zestawione w danej tabeli.

Płeć	Podsumowująca tabela Procenty z wiersza ogółem					Wiersz Razem
	auta japońskie	auta francuskie	auta czeskie	auta niemieckie	auta włoskie	
Mężczyźni	22,87%	9,09%	42,23%	11,73%	14,08%	74,78%
Kobiety	27,83%	29,57%	23,48%	6,09%	13,04%	25,22%
Badani ogółem	24,12%	14,25%	37,50%	10,31%	13,82%	100,00%

W powyższej tabeli, zestawiającej płeć badanego z preferowanym krajem pochodzenia marki samochodu, przedstawiono przykład procentowania wg wierszy. Łatwo zauważyć, że wśród mężczyzn największym zainteresowaniem cieszą się auta czeskie oraz japońskie, zaś w grupie kobiet są to samochody francuskie i japońskie. Największe rozbieżności dotyczą samochodów francuskich – preferowanych przez kobiety i niemieckich – preferowanych przez mężczyzn.



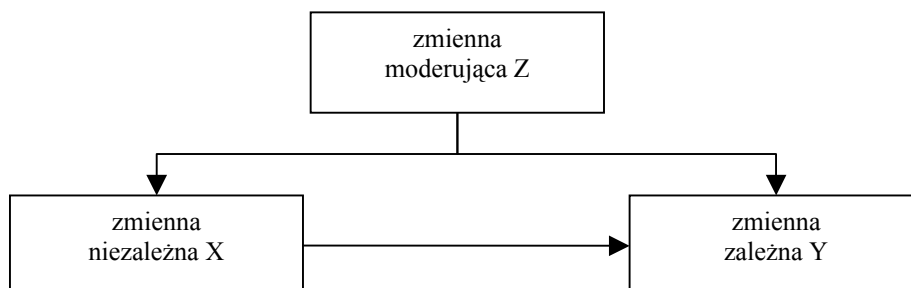
Poszukiwanie prostych zależności między zmiennymi – zmienne kontrolne w teście niezależności chi-kwadrat

Rozszerzeniem prostego obliczania procentów (wg wierszy, wg kolumn czy z całości) w tabelach kontyngencji jest sprawdzenie, czy między badanymi zmiennymi występuje statystycznie istotna zależność. Zwykle używa się w tym celu testu niezależności chi-kwadrat. Hipoteza zerowa zakłada, że nie ma zależności między zmiennymi, natomiast hipoteza alternatywna, że zależność taka występuje. Z informacji znajdujących się w poniższej tabeli wynika, że istnieje statystycznie istotny związek między płcią badanego a preferowanym krajem pochodzenia marki samochodu. Świadczy o tym niższa od 0,05 wartość prawdopodobieństwa testowego p.

statystyka	Chi-kwadr.	df	p
Chi kwadrat Pearso	36,95101	df=4	p=,00000
Chi ² NW	34,52611	df=4	p=,00000
Fi	,2846628		
Wsp. kontyngencji	,2737860		
V Craméra	,2846628		

Kolejnym etapem analizy (po stwierdzeniu zależności między zmiennymi) jest obliczenie siły związku między zmiennymi. Istnieje wiele współczynników służących do pomiaru siły zależności, np. współczynnik ϕ Yule'a, współczynnik T Czuprowa czy współczynnik kontyngencji C Pearsona. Pewną niedogodnością w ich stosowaniu jest brak stałej górnej granicy. Stwarza to trudność przy interpretacji, gdyż za każdym razem trzeba oszacować tę wartość (zależną od liczby kolumn i wierszy tabeli kontyngencji). Dlatego też bardzo dobrym wyjściem jest wykorzystanie współczynnika V Cramera, który dla dowolnych tabel przyjmuje wartości z przedziału [0,1], gdzie 1 oznacza bardzo silny związek między zmiennymi. W naszym wypadku, jego wartość (w ostatnim wierszu powyższej tabeli) wynosi 0,285, co oznacza, że zależność między ww. zmiennymi jest względnie słaba.

Często prosta analiza związków między zmiennymi nie pozwala na „wykrycie” silnych zależności. W sytuacjach takich zaleca się zastosowanie tzw. zmiennych moderujących, nazywanych również zewnętrznymi zmiennymi kontrolnymi. Rolę zmiennej moderującej przedstawiono na poniższym schemacie:



Jeśli nie stwierdzi się istnienia związku między zmiennymi X i Y (bądź jeśli związek ten jest bardzo słaby), to można wówczas przeprowadzić analizę z uwzględnieniem dodatkowej zmiennej moderującej Z. Wciąż bada się związek między X i Y, jednak tym razem analizę przeprowadza się dla każdej kategorii zmiennej Z z osobna. Liczba możliwych



zmiennych moderujących jest tak duża, że przeprowadzenie pełnej analizy bez koncepcji modelu jest zadaniem trudnym do wykonania. W tym wypadku na zmienną moderującą wybrano wiek badanego.

Podtabela dla: wiek 18-24 lata			
statystyka	Chi-kwadr.	df	p
Chi kwadrat Pearso	7,062916	df=4	p=,13261
Chi ² NW	7,037049	df=4	p=,13395
Fi	,2971303		
Wsp. kontyngencji	,2848232		
V Craméra	,2971303		

Podtabela dla: wiek 25-34 lata			
statystyka	Chi-kwadr.	df	p
Chi kwadrat Pearso	16,38345	df=4	p=,00255
Chi ² NW	14,53548	df=4	p=,00577
Fi	,3726162		
Wsp. kontyngencji	,3491643		
V Craméra	,3726162		

Zmienna „wiek” przyjmowała pięć wariantów: 18-24 lata, 25-34 lata, 35-49 lat, 50-64 lata i 65+, a zatem dodatkowej analizie poddanych zostaje pięć nowych tabel kontyngencji z wyliczonymi dla nich statystykami. Z zamieszczonych powyżej dwóch przykładowych tabel wynika, że wśród osób do 24 roku życia nie stwierdzono zależności między płcią a preferowanym krajem pochodzenia marki samochodu. Z kolei w innym przedziale wiekowym – 25-34 lata – zaobserwowano istnienie związku, którego siła jest wyższa niż wtedy, gdy analizie poddano wszystkich respondentów.

Tworzenie map percepcji – skalowanie wielowymiarowe

Jedno z pytań w ankiecie dotyczyło postrzeganej jakości technicznej samochodu. Pytanie brzmiało: Który kraj kojarzy się Panu/Pani z samochodami o wysokiej jakości technicznej? W każdej z par krajów podanych poniżej proszę wybrać tylko jeden – ten, z którego Pana/Pani zdaniem pochodzą lepsze technicznie samochody.

Niemcy	<input type="checkbox"/>	czy	Francja	<input type="checkbox"/>
Japonia	<input type="checkbox"/>	czy	Niemcy	<input type="checkbox"/>
Korea Płd.	<input type="checkbox"/>	czy	Niemcy	<input type="checkbox"/>
Japonia	<input type="checkbox"/>	czy	Włochy	<input type="checkbox"/>
Niemcy	<input type="checkbox"/>	czy	Czechy	<input type="checkbox"/>

itd ...

Zebrane dane można przedstawiać w postaci symetrycznej macierzy, w której liczba kolumn (i wierszy) jest równa liczbie analizowanych krajów (tabela poniżej). Wartości w polach tabeli oznaczają odsetek badanych, który twierdzi, że z kraju w kolumnie pochodzą lepsze technicznie samochody niż z kraju w wierszu. Przykładowo: wartość 0,4587 na przecięciu się kolumny „Niemcy” i wiersza „Japonia” oznacza, że 45,87% badanych uznało, że auta niemieckie są lepsze od aut japońskich.

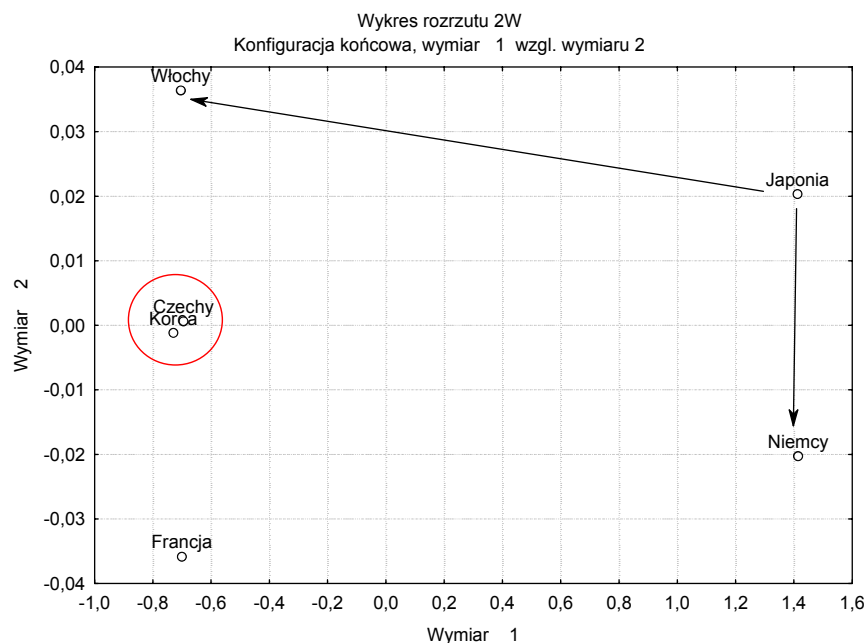


	1	2	3	4	5	6
	Niemcy	Japonia	Korea	Francja	Włochy	Czechy
Niemcy	0	0,5413	0,0046	0,0619	0,0138	0,0275
Japonia	0,4587	0	0,0046	0,1537	0,0849	0,0528
Korea	0,9954	0,9954	0	0,9151	0,8188	0,75
Francja	0,9381	0,8463	0,0849	0	0,1743	0,2018
Włochy	0,9862	0,9151	0,1812	0,8257	0	0,3853
Czechy	0,9725	0,9472	0,25	0,7982	0,6147	0

Dane te są punktem wyjścia do zbudowania mapy percepcji za pomocą techniki zwanej skalowaniem wielowymiarowym. Celem skalowania wielowymiarowego jest przedstawienie punktów reprezentujących obiekty (marki, produkty, tutaj: kraje) w zredukowanej – najczęściej 2-wymiarowej przestrzeni. Ponieważ dane wejściowe nie zawierają informacji o cechach produktu, mamy tutaj do czynienia z podejściem beztrybutowym i odległościowym.

Miarą, która informuje o stopniu odwzorowania punktów w przestrzeni o zredukowanej liczbie wymiarów, jest wskaźnik dopasowania STRESS. Wartość STRESS równa 0 wskazuje na idealne dopasowanie konfiguracji punktów do danych wejściowych. Zazwyczaj przyjmuje się z góry wartość tego współczynnika na poziomie 0,05 lub 0,01.

Skalowanie wielowymiarowe jest procedurą iteracyjną, w której podczas każdego kroku tworzy się konfigurację punktów reprezentujących badane obiekty i oblicza dla nich wartość STRESS. Jeśli jej wysokość jest wyższa od przyjętego wcześniej poziomu (np. 0,05), to cała procedura rozpoczyna się od nowa. Powstaje kolejna – zmieniona konfiguracja punktów z nowym oszacowaniem wskaźnika STRESS. Procedura trwa do chwili, aż jakość dopasowania punktów w zredukowanej przestrzeni będzie zadowalająca. Dwuwymiarową mapę percepcji przedstawiono na poniższym rysunku.



Względna bliskość punktów „Japonia” i „Niemcy” oraz „Czechy” i „Korea” świadczy o tym, że samochody pochodzące z tych krajów są przez badanych podobnie postrzegane.



Zdaniem respondentów samochody marek niemieckich i japońskich charakteryzują się zbliżoną jakością techniczną, podobnie jak samochody marek czeskich i koreańskich. Przestrzenny układ punktów dostarcza menedżerowi dodatkowej informacji. Gdyby chciał posłużyć się wyłącznie prostym rankingiem, to okazałoby się, że kolejność krajów jest następująca (1 oznacza kraj, z którego pochodzą najlepsze technicznie samochody): 1. Niemcy, 2. Japonia, 3. Francja, 4. Włochy, 5. Czechy, 6. Korea Płd. Mapa percepcji dostarcza dodatkowych informacji o zjawisku substytucji (potencjalni nabywcy aut niemieckich mogą brać pod uwagę również samochody japońskie) oraz pozwala menedżerowi np. na kształtowanie spójnego wizerunku komisji. Jeśli przykładowo zdecyduje się na sprzedaż pojazdów uznawanych za te o wysokiej jakości technicznej, to nie powinien oferować aut koreańskich, nawet jeśli należą do klasy wyższej (Daewoo Leganza, Kia Opirus, Hyundai Santa Fe).

Budowa modeli – o regresji logistycznej i drzewach klasyfikacyjnych CART słów kilka

Regresja logistyczna jest matematycznym podejściem modelowym opisującym zależność między zestawem zmiennych niezależnych a jakościową – dychotomiczną zmienną zależną. Swą popularność zawdzięcza funkcji logistycznej, na której jest zbudowana. Funkcję tę można przedstawić za pomocą wzoru:

$$f(z) = \frac{1}{1 + e^{-z}}$$

Wartość z jest indeksem łączącym wkład poszczególnych zmiennych niezależnych, zaś wartość $f(z)$ jest prawdopodobieństwem dla danego z . Dla bardzo dużych dodatnich wartości z $f(z)$ jest bliskie jeden, natomiast dla bardzo dużych ujemnych $f(z)$ jest bliskie zero. Zakres wartości od 0 do 1 sprawia, że funkcja logistyczna stosowana jest często w sytuacji, gdy bada się prawdopodobieństwo wystąpienia jakiegoś zdarzenia Y . Prawdopodobieństwo takie również przyjmuje wartości od 0 (zdarzenie niemożliwe) do 1 (zdarzenie pewne).

Alternatywnym sposobem zapisu modelu logistycznego jest postać logitowa, którą otrzymuje się w wyniku transformacji logitowej:

$$\text{logit}P(\mathbf{X}) = \ln_e \left[\frac{P(\mathbf{X})}{1 - P(\mathbf{X})} \right]$$

Prawa strona modelu logitowego $\frac{P(\mathbf{X})}{1 - P(\mathbf{X})}$ może być w skrócie interpretowana jako iloraz prawdopodobieństwa tego, że zdarzenie wystąpi oraz prawdopodobieństwa tego, że zdarzenie nie wystąpi. Wskaźnik ten jest nazywany ilorazem szans (*odds ratio*, w skrócie OR). Przykładowo, jeśli prawdopodobieństwo dokonania zakupu produktu wynosi $P(\text{kupi}) = 0,25$, to znaczy, że prawdopodobieństwo zrezygnowania z zakupu będzie wynosić $P(\text{nie$



kupi) = 0,75, czyli iloraz szans będzie równy $0,25 / 0,75 = 1/3$. Interpretacja tego wskaźnika może być dwojaka:

- ◆ prawdopodobieństwo dokonania zakupu jest jedną trzecią prawdopodobieństwa zrezygnowania z zakupu;
- ◆ szansa, że klient zrezygnuje z zakupu wynosi 3 do 1.

Na potrzeby niniejszych badań ankietowych zdecydowano się wykorzystać regresję logistyczną w celu zbudowania modelu opisującego osoby preferujące zakup samochodu nowego oraz osoby preferujące zakup samochodu używanego. Zestaw zmiennych niezależnych obejmował pozycje ze skali struktury korzyści, pozycje z listy wartości LOV oraz zmienne demograficzne. Tabelę zestawiającą statystycznie istotne predyktory przedstawiono poniżej.

Model: Regr. logistyczna (logit)							
Zmn. zal.: P1nowy Strata: Największe prawd. bł.średnkw.skala.							
Całkowita strata: 124,23106998 Chi2(6)=38,359 p=.00000							
	Stała B0	P8OTplus	P8KEplus	P13kobieta	P19faza1	P19faza2	P19faza3
Ocena	0,6106	-0,9957	0,6727	0,6726	-1,8307	-1,5917	-1,0774
poziom p	0,0423	0,0069	0,0472	0,0776	0,0000	0,0192	0,0047
Iloraz szans z.jedn.	1,8415	0,3695	1,9595	1,9593	0,1603	0,2036	0,3405

W przypadku zmiennej „osiągi techniczne” (P8OTplus) iloraz szans jest równy 0,3695, co oznacza, że szansa na wybór auta nowego przez osoby kładące nacisk na osiągi techniczne samochodu jest 3-krotnie mniejsza od szansy wyboru auta używanego. Zależność tę potwierdza porównanie oferty na rynku pierwotnym i wtórnym. Kwota, jaką trzeba przeznaczyć na zakup nowego samochodu, wystarcza na zakup samochodu używanego z bogatszym wyposażeniem lub o zdecydowanie wyższych parametrach technicznych.

Iloraz szans dla zmiennej „koszty eksploatacji” (P8KEplus) wynosi 1,9595, co oznacza, że szansa zakupu auta nowego przez osoby kładące nacisk na niskie koszty eksploatacji jest prawie dwukrotnie wyższa niż dla osób, dla których niskie koszty eksploatacji nie są najważniejsze. Podobnie należy zinterpretować iloraz szans dla zmiennej „kobieta” (1,9593) – szansa zakupu nowego samochodu jest dwukrotnie wyższa wśród kobiet niż wśród mężczyzn.

Wartości ilorazu szans dla zmiennych „faza 1”, „faza 2” i „faza 3” wynoszą odpowiednio: 0,1603; 0,2036 i 0,3405. Oznacza to, że:

- ◆ prawdopodobieństwo zakupu nowego samochodu przez młode osoby stanu wolnego jest ponad 6-krotnie mniejsze niż prawdopodobieństwo zakupu przez osoby znajdujące się w pozostałych fazach cyklu życia rodziny,
- ◆ prawdopodobieństwo zakupu nowego samochodu przez młode małżeństwa bez dzieci jest 5-krotnie mniejsze niż prawdopodobieństwo zakupu przez osoby w pozostałych fazach cyklu życia,
- ◆ szansa zakupu nowego samochodu przez małżeństwa z dziećmi poniżej 18. roku życia jest 3-krotnie mniejsza od szansy zakupu przez osoby z pozostałych faz cyklu życia rodziny.

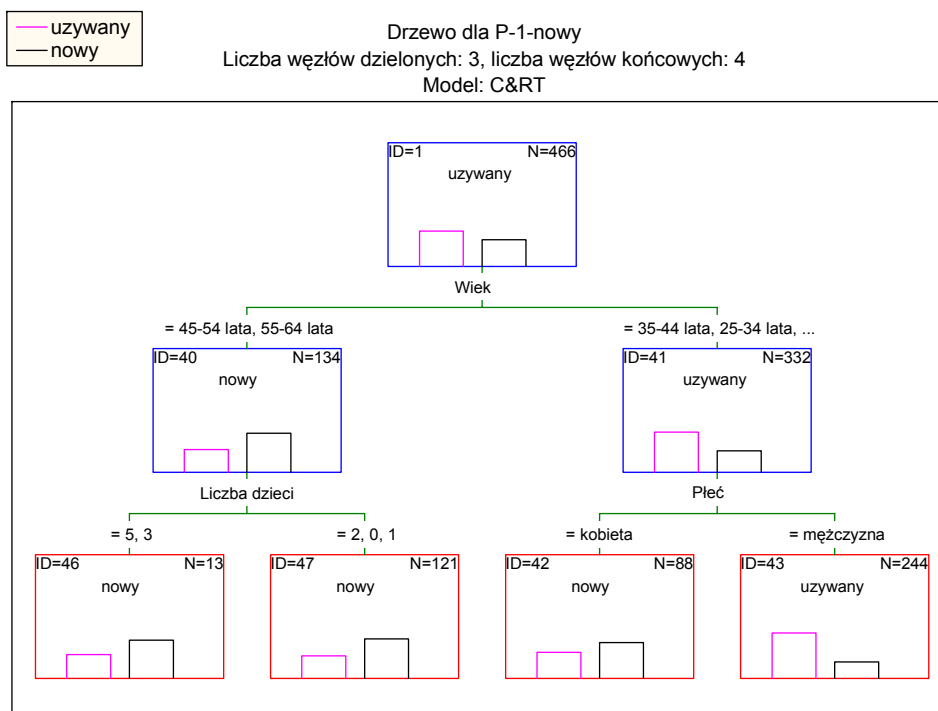


Reasumując, można powiedzieć, że respondenci preferujące nowe auta to przede wszystkim: a) kobiety, b) osoby kładące nacisk na niskie koszty eksploatacji, c) osoby, którym nie zależy na wysokich osiągnięciach technicznych auta oraz d) osoby znajdujące się w wyższych fazach cyklu życia rodziny (tj. małżeństwa z dorosłymi dziećmi, „puste gniazda” i starsze osoby stanu wolnego – dotyczy to, w dużym uproszczeniu, osób po 50. roku życia).

Drzewa klasyfikacyjne CART to narzędzie analityczne *data mining*, które jest uznawane za najbardziej zaawansowaną metodę podziału rekurencyjnego. Chociaż pierwotnym przeznaczeniem tego narzędzia była analiza danych zastanych, to jednak ze względu na łatwość i elastyczność obsługi, a także prostotę interpretacyjną i interesujący sposób wizualizacji wyników, metoda ta z powodzeniem może być wykorzystywana w analizie danych ankietowych.

Wynikiem analizy jest model drzewa klasyfikacyjnego (rysunek poniżej), który w tym wypadku można zamienić na zestaw czterech zdań warunkowych (reguł typu *if ... then ...*):

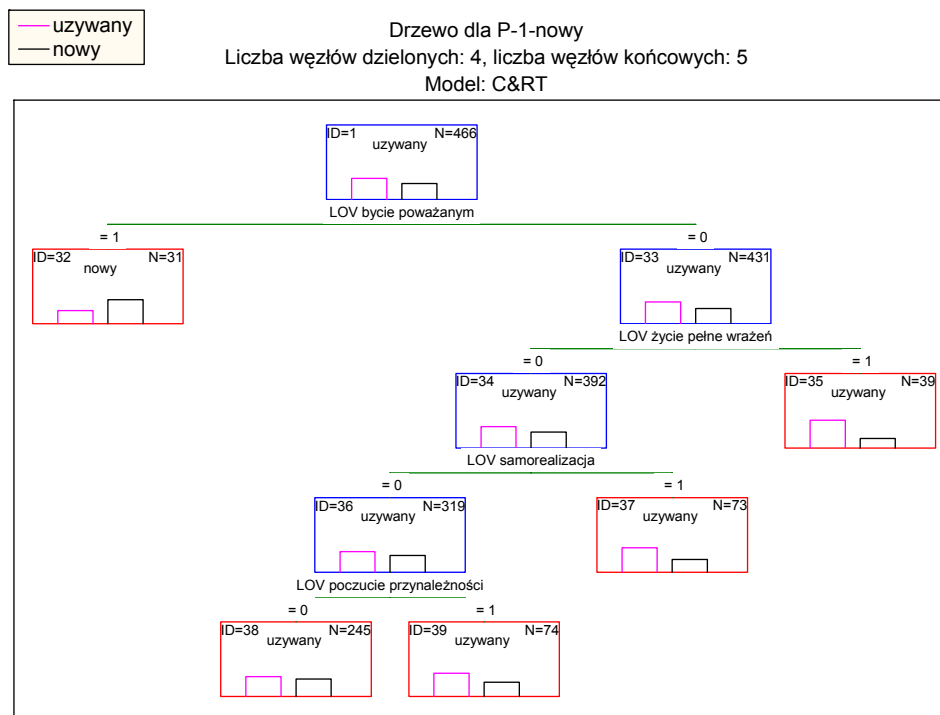
- ◆ jeżeli klient ma więcej niż 45 lat i ma 3 lub więcej dzieci to wybierze samochód nowy (z prawdopodobieństwem 69%);
- ◆ jeżeli klient ma więcej niż 45 lat i ma nie więcej niż dwoje dzieci, to wybierze samochód nowy (z prawdopodobieństwem 63%);
- ◆ jeśli klient to kobieta, która ma mniej niż 45 lat, to wybierze auto nowe (z prawdopodobieństwem 58%);
- ◆ jeśli klient to mężczyzna, która ma mniej niż 45 lat, to wybierze auto używane (z prawdopodobieństwem 73%).





Alternatywne modele drzew klasyfikacyjnych CART – modyfikowanie struktury drzew przy użyciu zmiennych konkurencyjnych

Interesującą i przydatną innowacją metody CART jest występowanie zmiennych konkurencyjnych (*competitors*) i zmiennych zastępczych (*surrogates*). Na każdym etapie podziału drzewa zestawiany jest ranking zmiennych niezależnych, które zapewniają najlepszy podział danego węzła. Pozycja w tym rankingu zależy od trafności predykcji zmiennej zależnej w wydzielanych węzłach potomnych. Najlepszy z predyktorów wykorzystywany jest do budowy modelu, a pozostałe pełnią funkcję bądź to zmiennych konkurencyjnych bądź zmiennych zastępczych (bądź obie te role jednocześnie). Kolejność w tych lokalnych rankingach zależy od wartości tzw. wskaźnika poprawy (*improvement*). Zmienna niezależna, dla której wartość ta jest najwyższa, zostaje uznana za najlepszy predyktor pierwotny, który dzieli dany węzeł. Informacja o pozostałych predyktorach może zostać wykorzystana m.in. w trakcie budowy alternatywnej struktury drzewa klasyfikacyjnego. Badacz ma bowiem możliwość zastąpienia podziałów pierwotnych własnymi podziałami opartymi na innych – subiektywnie dobranych zmiennych objaśniających.



Powyższy model można tym razem opisać za pomocą pięciu zdań warunkowych:

- ◆ jeśli klient kieruje się w życiu wartością „bycie poważanym”, to wybierze samochód nowy (z prawdopodobieństwem 65%);
- ◆ jeśli klient nie kieruje się w życiu wartością „bycie poważanym”, ale kładzie nacisk na „życie pełne wrażeń”, to kupi samochód używany (74%);
- ◆ jeśli klient nie kieruje się w życiu wartością „bycie poważanym”, ani wartością „życie pełne wrażeń”, ale kładzie nacisk na „samorealizację”, to kupi samochód używany (65%) itd.



Trafność predykcji mierzona odsetkiem poprawnie sklasyfikowanych przypadków wynosi dla tego modelu 59%, co oznacza, że ingerencja w strukturę modelu „kosztowała” badacza 9 punktów procentowych. O tyle bowiem obniżyła się jakość modelu, jeśli porównać go do rozwiązania pierwotnego (68%).