



Enginius Segmentation

Sai Hemanth Puli, University

Copyright (c) 2024, DecisionPro Inc.

Table of Contents

Data transformation

Segment solution

- [4-segment solution](#)
- [Dendrogram](#)
- [Scree plot](#)

Segment description

- [Segment size](#)
- [Segment description](#)
- [Segmentation space](#)
- [Segment membership](#)

Segment profiles

- [Spider chart](#)
- [Segment 1 profile](#)
- [Segment 2 profile](#)
- [Segment 3 profile](#)
- [Segment 4 profile](#)

Descriptor analysis

- [Descriptors](#)
- [Descriptor space](#)

Classification model

- [Introduction](#)
- [Model coefficients](#)
- [P-values](#)
- [Confusion matrix](#)
- [Model predictions](#)

Data transformation

Standardization has not been performed.

Segment solution

4-segment solution

The ideal number of segments is a function of statistical fit (what the data say), managerial relevance (what makes the most sense from a managerial point of view), and targetability (can the segments be easily targeted).

When the three criteria do not perfectly converge, selecting the right number of segments becomes a judgment call.

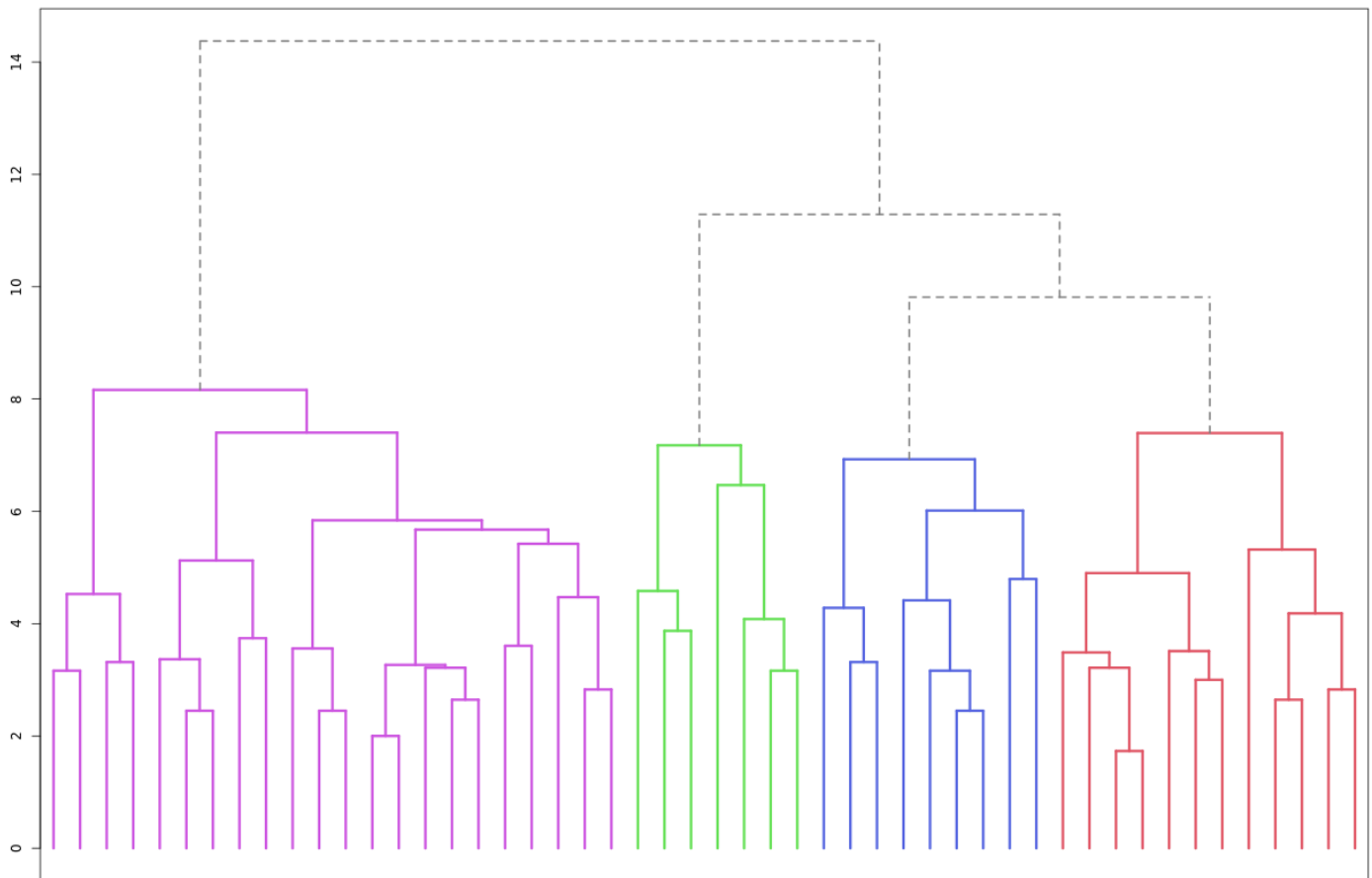
You have decided to perform the analysis with 4 segments.

The segmentation method relies on the hierarchical clustering approach. This approach generates a dendrogram that we display next.

Dendrogram

The dendrogram represents the grouping process of observations into clusters. The chart reads from bottom (all initial observations are separated) to top (all observations are clustered into one unique segment).

The height represents the distance between the two groups of observations being merged at each step. If two very distant groups are being merged, this will create a 'jump' in the dendrogram, indicating that it might be wise to stop the clustering process before.



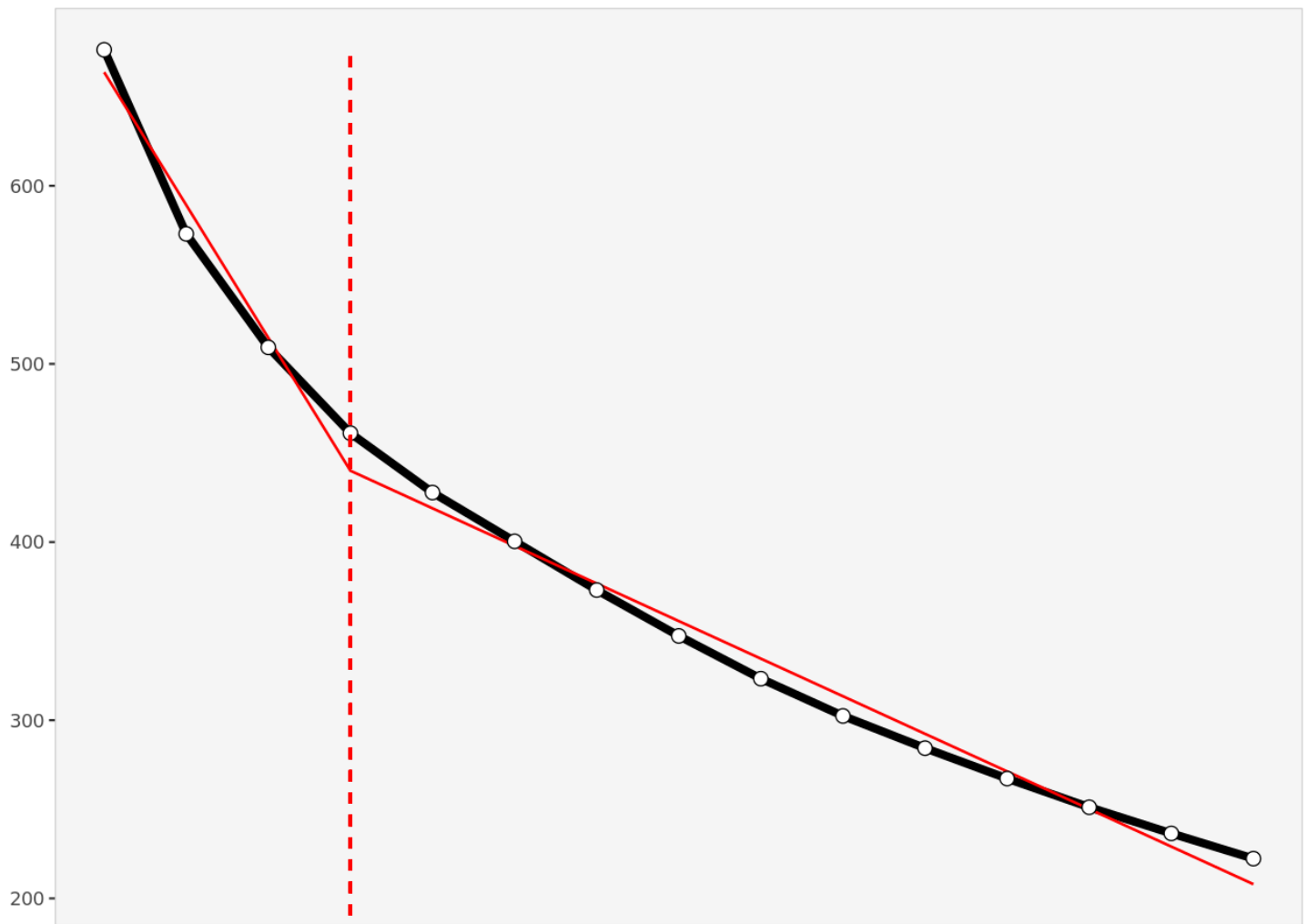
Dendrogram. The dendrogram is a tree diagram to illustrate the arrangement of clusters produced by hierarchical clustering, and how the observations are incrementally clustered together.

Scree plot

The screeplot displays, for each cluster solution, a measure of within-cluster heterogeneity. If clusters group observations that are widely different (which will happen if the number of clusters is too small to capture the variability in the data), the value will be high.

A good cluster solution might be where the screeplot displays an 'elbow', that is, where increasing the number of clusters beyond a certain point does not dramatically decreases within-cluster heterogeneity.

The measure displayed in the screeplot is related, but not equivalent, to the distance reported in the dendrogram.



Scree plot. The scree plot compares the sum of squared error (SSE) for each cluster solution. A good cluster solution might be when the SSE slows dramatically, creating an 'elbow'. Such elbow does not always exist.

From a statistical point of view, the SSE reported in the screeplot is computed as the sum of squared error between each observation and its cluster centroid (or center), summed over all the observations.

Segment description

Segment size

	Population	Segment 1	Segment 2	Segment 3	Segment 4
Size	50	9	22	12	7
Relative size	100%	18%	44%	24%	14%

Segment size.

Segment description

	Population	Segment 1	Segment 2	Segment 3	Segment 4
What type of content do you prefer to watch on Netflix?	1.42	1.56	1.32	1.58	1.29
How many hours per week do you typically spend watching content on streaming platforms? = 0-5	0.540	0.667	0.364	0.583	0.857
How many hours per week do you typically spend watching content on streaming platforms? = 6-10	0.240	0.111	0.455	0.083	0.000
How many hours per week do you typically spend watching content on streaming platforms? = 11-15	0.180	0.222	0.182	0.167	0.143
How important is the availability of local (Indian) content in your decision to subscribe to a streaming service?	3.80	4.33	3.91	4.50	1.57
Action	3.60	3.00	4.09	3.17	3.57
Comedy	4.30	5.00	4.45	3.58	4.14
Documentary	3.56	3.22	4.09	2.58	4.00
Drama	4.06	4.00	4.41	3.25	4.43
Thriller	3.96	3.78	4.73	3.00	3.43
Rate value for Price for each on scale of 1-5 [Netflix]	3.36	2.22	3.64	3.42	3.86
Rate Service Quality for each on scale of 1-5 [Netflix]	4.08	3.78	4.45	4.00	3.43
Rate content availability for each on scale of 1-5 [Netflix]	3.82	2.78	4.45	3.75	3.29
Rate device compatibility (usage on any/multiple device) for each on scale of 1-5 [Netflix]	3.48	2.22	4.05	3.75	2.86
Rate Content freshness (usage on any/multiple device) for each on scale of 1-5 [Netflix]	3.84	3.00	4.32	3.83	3.43
Bengali	0.020	0.000	0.045	0.000	0.000
English	0.960	1.000	0.955	0.917	1.000
Hindi	0.860	0.667	0.909	0.917	0.857
Kannada	0.080	0.000	0.136	0.083	0.000
Korean	0.160	0.222	0.136	0.000	0.429
Malayalam	0.220	0.111	0.455	0.000	0.000
Marathi	0.180	0.000	0.227	0.083	0.429
Others	0.060	0.000	0.136	0.000	0.000
Tamil	0.260	0.222	0.409	0.167	0.000
Telugu	0.520	0.667	0.636	0.333	0.286
spanish	0.080	0.000	0.091	0.083	0.143

Segment description. Average value of each segmentation variable, overall for each segment (centroid). Segmentation variables that are statistically different from the rest of the population are highlighted in red (lower) or green (higher).

Statistical differences i	
What type of content do you prefer to watch on Netflix? -	
How many hours per week do you typically spend watching content on streaming platforms? = 0-5 -	
How many hours per week do you typically spend watching content on streaming platforms? = 6-10 -	
How many hours per week do you typically spend watching content on streaming platforms? = 11-15 -	
How important is the availability of local (Indian) content in your decision to subscribe to a streaming service? -	
Action -	
Comedy -	
Documentary -	
Drama -	
Thriller -	
Rate value for Price for each on scale of 1-5 [Netflix] -	
Rate Service Quality for each on scale of 1-5 [Netflix] -	
Rate content availability for each on scale of 1-5 [Netflix] -	
Rate device compatibility (usage on any/multiple device) for each on scale of 1-5 [Netflix] -	
Rate Content freshness (usage on any/multiple device) for each on scale of 1-5 [Netflix] -	
Bengali -	
English -	
Hindi -	
Kannada -	
Korean -	
Malayalam -	
Marathi -	
Others -	
Tamil -	
Telugu -	
spanish -	

■ Lower (p<.05)
■ Lower (p<.10)
■ n.s.
■ Higher (p<.10)
■ Higher (p<.05)

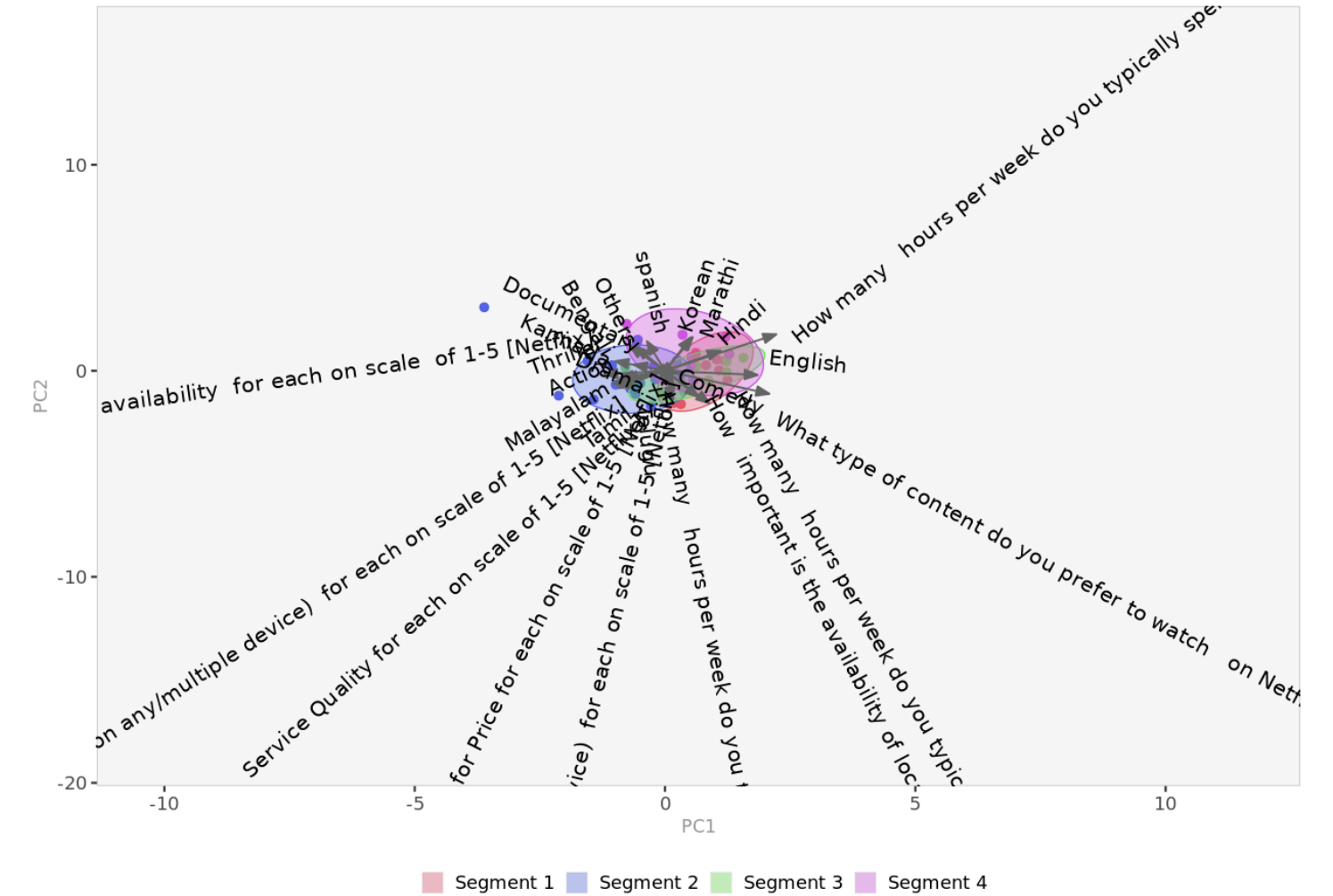
Segment differences per segment. Cell colors indicate to what extent a segment is statistically different from the rest of the population on each segmentation variable.

Segmentation space

The chart below is a graphical representation of the various segments, segment members, and segmentation variables. It is obtained by plotting the first two dimensions of a principal component analysis performed on the (standardized) segmentation data, on top of which segment information has been overlaid.

Because only the first two dimensions of the PCA are displayed, and these two dimensions capture only 29.7% of the variance in the data, some differences between segments might not appear here. Note that segmentation variables with no variance, if any, have been excluded.

Two clusters that appear to overlap on the first two dimensions might be distinct on other dimensions. Consequently, this chart is a useful guide, for checking which variables are correlated, but may be misleading if used to select the optimal number of segments.



Segment space. Spatial representation of segments and segmentation variables, using principal component analysis.

Segment membership

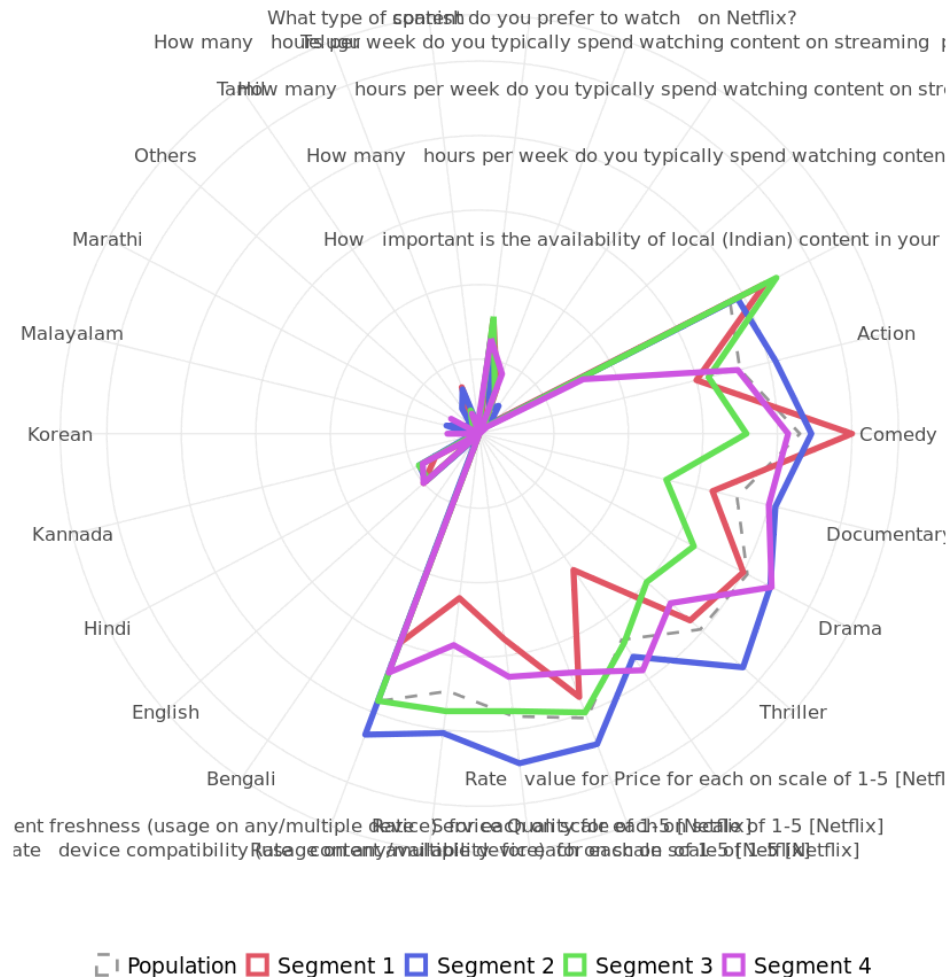
Segment	
1	1
2	2
3	1
4	3
5	3
6	3
7	2
8	1
9	1
10	3

Segment membership (excerpt). Segment to which each member of the population belongs to. The complete membership list is only available in the Excel formatted output.

Segment profiles

Spider chart

Spider chart comparing the averages of the segmentation variables across all segments.

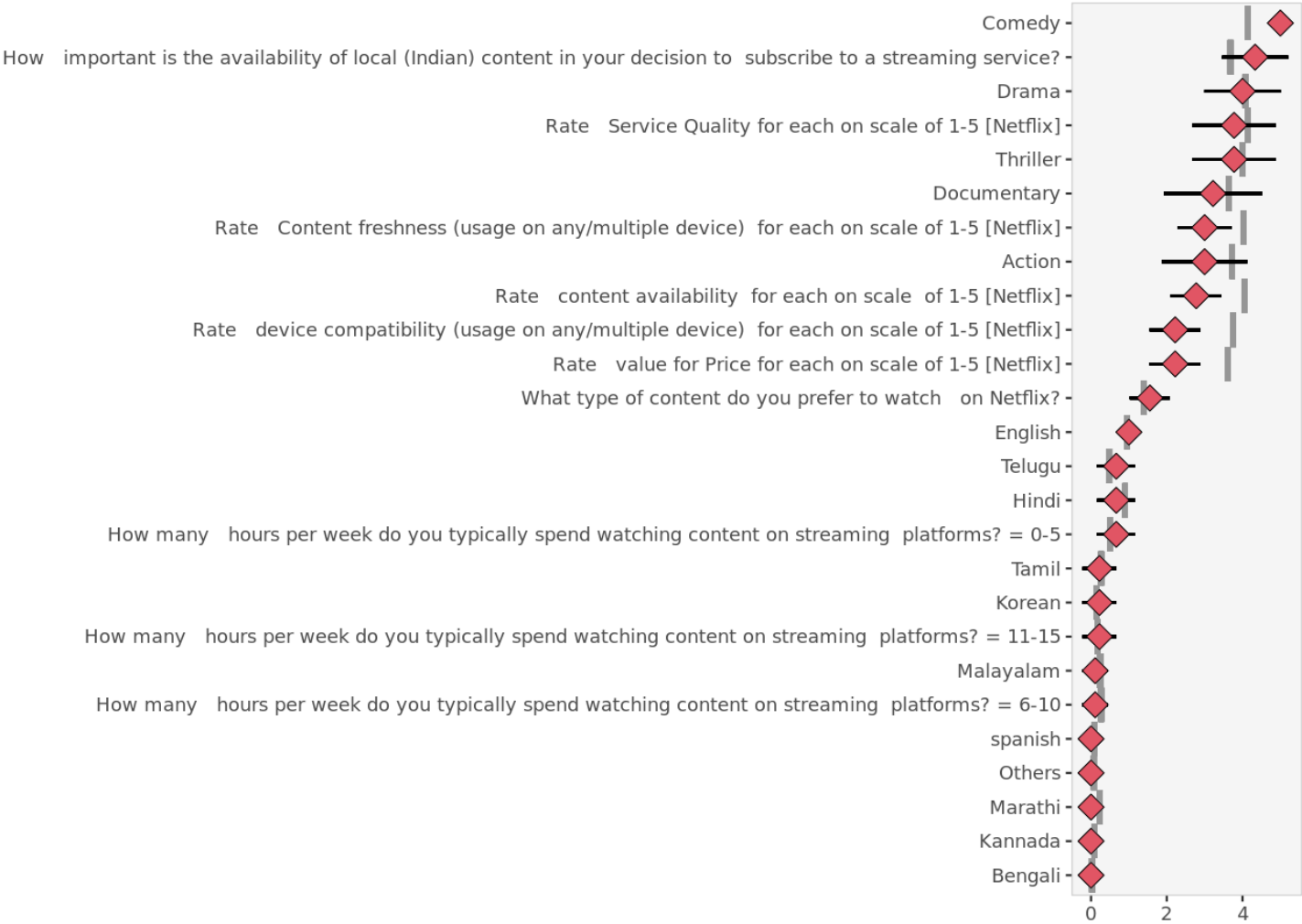


Spider chart.

Segment 1 profile

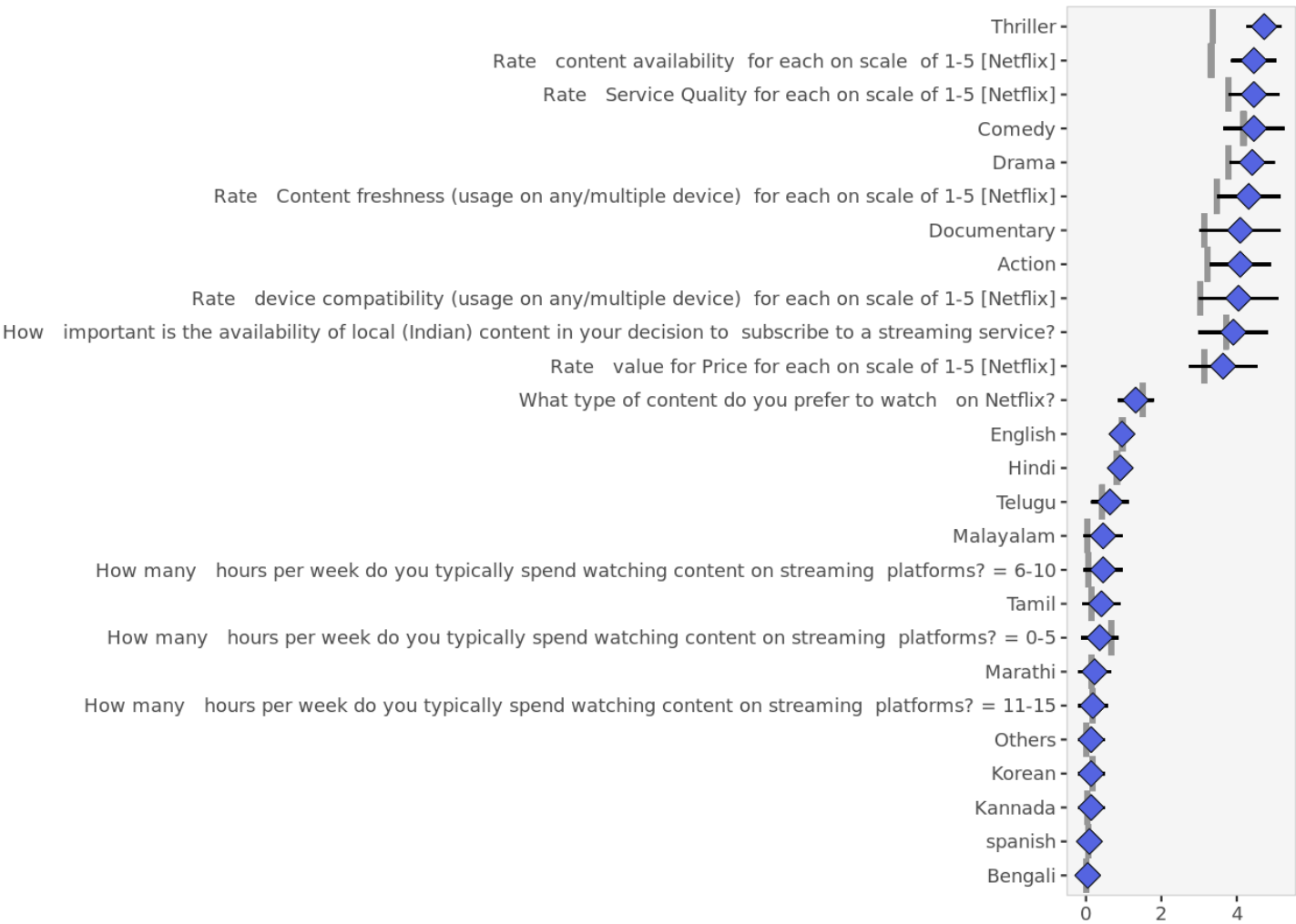
The following charts represent the profile of each segment. These charts are only available when the data are not standardized, hence the model assumes that all segmentation variables use the same scale.

- For each segment, the segmentation variables are ordered in decreasing order of magnitude.
- The colored dots represent the average of the segment.
- The horizontal lines represent the standard deviations within that segment.
- The vertical, gray lines represent the averages of the rest of the population, after excluding members of the segment under scrutiny.



Segment 1 profile.

Segment 2 profile



Segment 2 profile.

Segment 3 profile



Segment 3 profile.

Segment 4 profile



Segment 4 profile.

Descriptor analysis

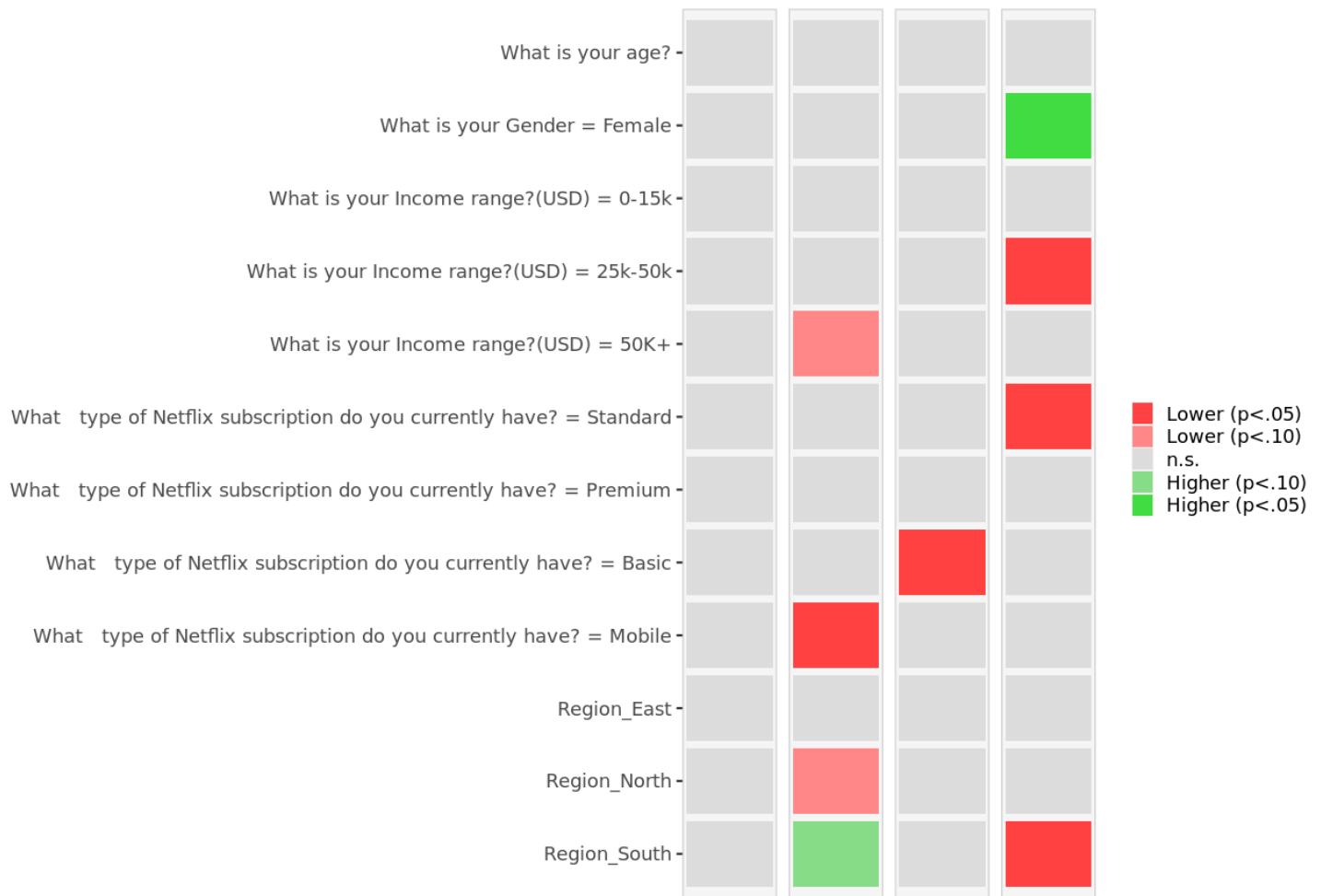
Descriptors

This table reports the descriptor averages of each segment. The more differences can be found, the easier it will be to predict segment membership based on descriptors alone.

	Population	Segment 1	Segment 2	Segment 3	Segment 4
What is your age?	24.6	24.7	24.4	24.3	25.3
What is your Gender = Female	0.620	0.667	0.591	0.417	1.000
What is your Income range?(USD) = 0-15k	0.700	0.778	0.773	0.500	0.714
What is your Income range?(USD) = 25k-50k	0.120	0.111	0.182	0.083	0.000
What is your Income range?(USD) = 50K+	0.140	0.111	0.045	0.333	0.143
What type of Netflix subscription do you currently have? = Standard	0.200	0.333	0.182	0.250	0.000
What type of Netflix subscription do you currently have? = Premium	0.340	0.222	0.364	0.500	0.143
What type of Netflix subscription do you currently have? = Basic	0.260	0.222	0.318	0.083	0.429
What type of Netflix subscription do you currently have? = Mobile	0.160	0.222	0.045	0.167	0.429
Region_East	0.020	0.000	0.000	0.083	0.000
Region_North	0.060	0.111	0.000	0.083	0.143
Region_South	0.480	0.556	0.636	0.333	0.143

Descriptor data per segment. Average value of each descriptor, overall and within each cluster. Descriptors that are statistically different from the rest of the population are highlighted in red (lower) or green (higher).

Statistical differences in descriptors



Descriptor differences per segment. Cell colors indicate to what extent the distribution of a descriptor in a segment is statistically different from the rest of the population.

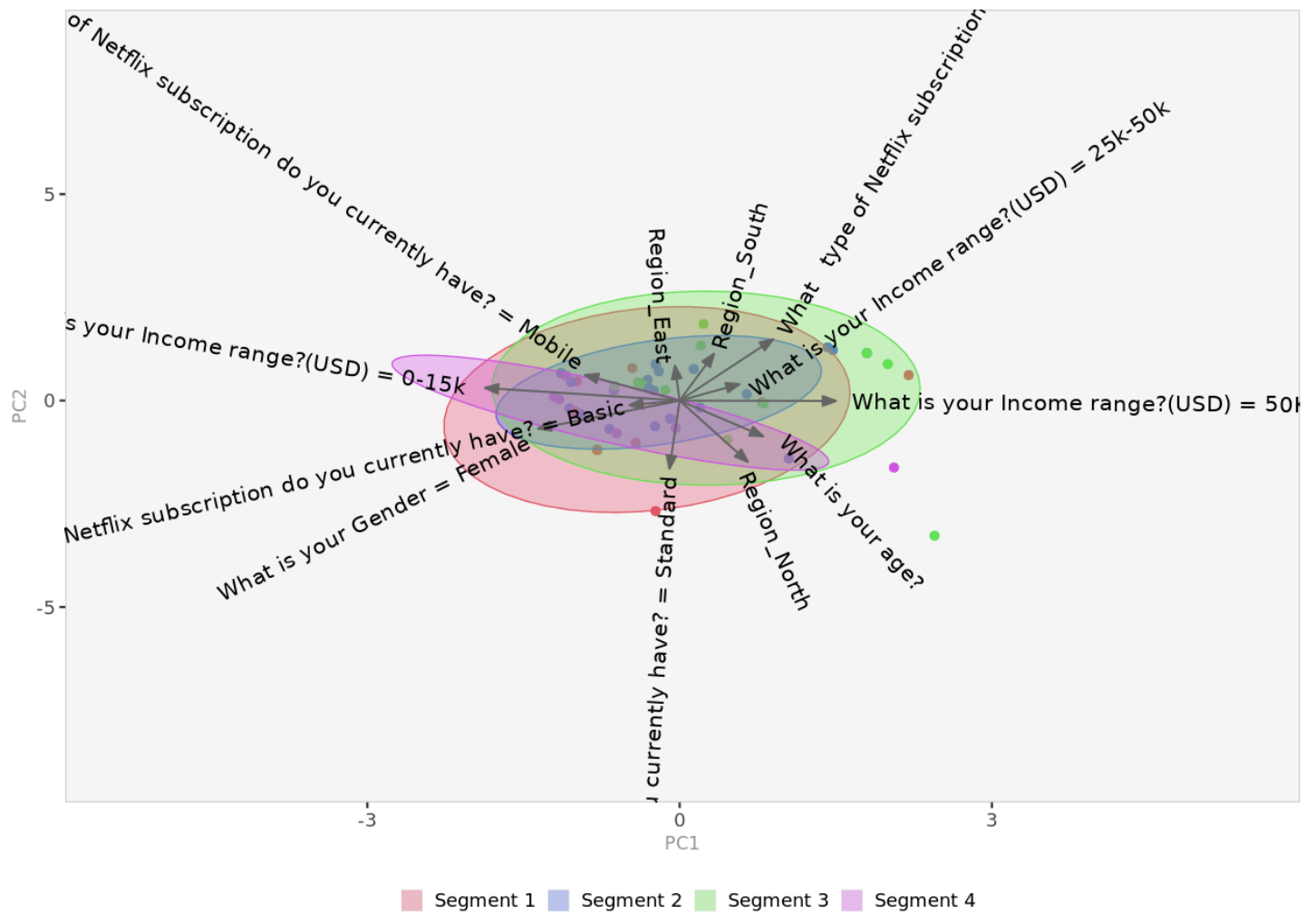
Descriptor space

The chart below is a graphical representation of the various segments, segment members, and descriptors. It is obtained by outputting the first two dimensions of a principal component analysis performed on the (standardized) descriptors, on top of which segment information has been overlayed.

Because only the first two dimensions of the PCA are displayed, and these two dimensions capture only 35.2% of the variance in the data, some differences between segments might not appear here. Note that descriptors with no variance, if any, have been excluded.

If two or more segments fully overlap, it is unlikely that they could be clearly separated based on descriptors alone.

However, two segments that seem to overlap on two dimensions may be more clearly separated on other dimensions. Consequently, the confusion matrix is a better guide to assess the quality of segment classification based on descriptors.



Descriptor space. Spatial representation of segments and their descriptors, using principal component analysis.

Classification model

Introduction

Often, segmentation (needs) variables for each customer may not be available to managers, but descriptors variables for customers may be available.

In this section, we explore whether descriptors alone can predict segment membership with sufficient accuracy. The confusion matrix and hit rates (reported below) indicate whether the model is accurate enough.

For member classification based on descriptors, Enginius uses a multinomial logit model (similar to the one used to predict 'choices between multiple alternatives (A/B/C)' in the predictive modeling module.

The largest segment is selected as the default option (dummy), and the model identifies which descriptors are the most significant for predicting cluster memberships. If a descriptor is highly predictive, its p-values will be close to zero, and the cells will appear in green (or red).

Model coefficients

	Segment 1	Segment 3	Segment 4
(Intercept)	-428	197	-515
What is your age?	0.029	-0.185	0.514
What is your Gender = Female	0.151	0.010	915.785
What is your Income range?(USD) = 0-15k	119	-330	-341
What is your Income range?(USD) = 25k-50k	118	-330	-1 026
What is your Income range?(USD) = 50K+	120	-325	-124
What type of Netflix subscription do you currently have? = Standard	309	136	-557
What type of Netflix subscription do you currently have? = Premium	308	136	-257
What type of Netflix subscription do you currently have? = Basic	307.9	134.8	-72.3
What type of Netflix subscription do you currently have? = Mobile	310	139	401
Region_East	-94.5	544.5	55.5
Region_North	236	235	703
Region_South	-0.364	-3.168	-472.182

Model parameters. Segment 2 is the model baseline.

P-values

	Segment 1	Segment 3	Segment 4
(Intercept)	0.000	0.000	0.000
What is your age?	0.825	0.455	0.311
What is your Gender = Female	0.885	0.994	0.000
What is your Income range?(USD) = 0-15k	0.000	0.000	0.000
What is your Income range?(USD) = 25k-50k	0.000	0.000	0.000
What is your Income range?(USD) = 50K+	0.000	0.000	0.000
What type of Netflix subscription do you currently have? = Standard	0.000	0.000	0.000

What type of Netflix subscription do you currently have? = Premium	0.000	0.000	0.000
What type of Netflix subscription do you currently have? = Basic	0.000	0.000	0.000
What type of Netflix subscription do you currently have? = Mobile	0.000	0.000	0.000
Region_East	NaN	0.000	0.000
Region_North	0.000	0.000	0.000
Region_South	0.721	0.065	0.000

p-values. Probabilities that parameter estimates are different from zero only by chance.

Confusion matrix

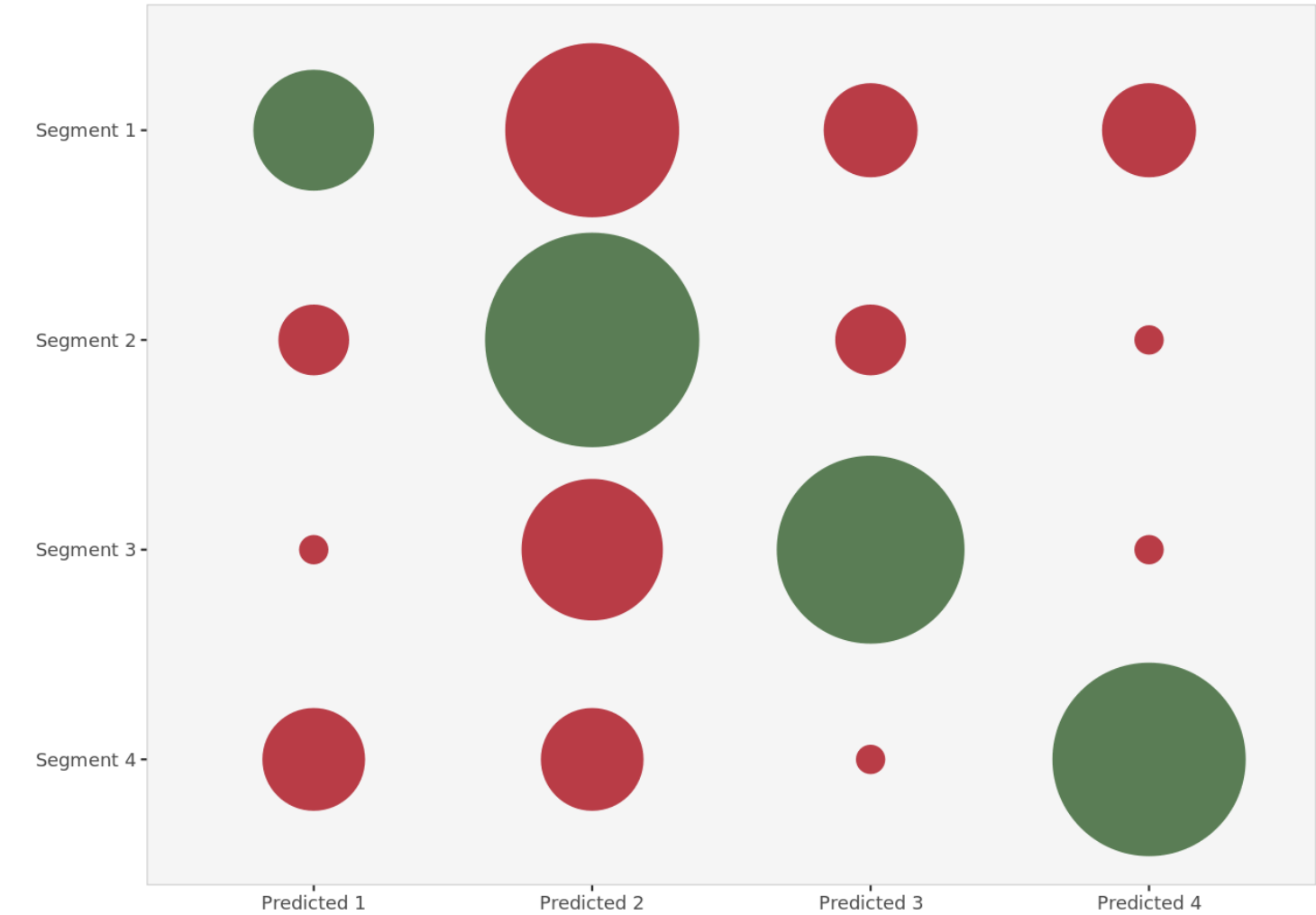
The confusion matrix compares actual segment membership (obtained from the segmentation analysis and the original segmentation variables) and predicted segment membership (obtained from the in-sample classification analysis and the descriptors alone). When actual and predicted segment memberships coincide, the diagonal elements will be comparatively large, indicating that the classification model based on available descriptors is accurate.

	Predicted 1	Predicted 2	Predicted 3	Predicted 4	Total
Segment 1	2	5	1	1	9
Segment 2	1	20	1	0	22
Segment 3	0	4	8	0	12
Segment 4	1	1	0	5	7
Total	4	30	10	6	50

Confusion matrix (count). The model has correctly classified 35 of the 50 observations. The off-diagonal elements are classification errors.

	Predicted 1	Predicted 2	Predicted 3	Predicted 4	Total
Segment 1	22%	56%	11%	11%	100%
Segment 2	5%	91%	5%	0%	100%
Segment 3	0%	33%	67%	0%	100%
Segment 4	14%	14%	0%	71%	100%

Confusion matrix (%). The global hit rate of the model is 70%. The diagonal elements represent segment-specific hit rates.



Confusion matrix (plot). Graphic representation of the confusion matrix: actual segment membership versus predicted segment membership. Bubbles in the diagonale represent correct classification.

Model predictions

	Prob(cluster 1)	Prob(cluster 2)	Prob(cluster 3)	Prob(cluster 4)	Predicted	Actual	Correct
1	69%	0%	31%	0%	1	1	1
2	11%	87%	2%	0%	2	2	1
3	18%	77%	6%	0%	2	1	0
4	31%	0%	69%	0%	3	3	1
5	13%	48%	39%	0%	2	3	0
6	27%	42%	31%	0%	2	3	0
7	0%	100%	0%	0%	2	2	1
8	16%	81%	3%	0%	2	1	0
9	28%	13%	6%	53%	4	1	0
10	12%	19%	70%	0%	3	3	1

Model predictions (in-sample) (excerpt). This table details the probabilities of each member of the segmentation dataset to belong to each cluster (as predicted by the in-sample classification model and the descriptors alone). The segment with the highest probability is retained, and is compared to the actual segment membership to measure model accuracy and classification errors.

Copyright (c) 2024, DecisionPro Inc.