Rohini Vishwanathan
Knowledge Discovery and Data Mining
IST 340

CE4

Question 1: Asia Magazine Dataset

Q1.1 Clustering

- Dataset: Asia Magazine (ASIA)
- Target Variable for DTs: Cluster ID
- Ignored Variables: Rank, Company Name, Country, Main Business, Note
- Missing Values Treatment: Imputed using MICE method

Segmentation Models:

Label	Algorithm	Initial Cluster Centers	Distance Measures	Normalization Method	No. of Clusters
LKER_3	Lloyd	k-means++	Euclidean	Range	3
MRCZ_3	MacQueen	Random	Chi-square	Z-score	3

Q1.2 Decision Tree Implementation

Tree Hyper-parameters:

- Splitting Criteria: Entropy, Gini
- Min Samples per Leaf: 6% of smallest cluster size (excluding outliers)
- Max Depth: 3

Evaluation Criteria:

- Accuracy (Weight: 0.60)
- Simplicity (Weight: 0.40)

Model	Accuracy (Entropy)	Simplicity (Entropy)	Score (Entropy)	Accuracy (Gini)	Simplicity (Gini)	Score (Gini)

LKER_3	0.80	1	**0.80**	0.73	1	0.73
MRCZ_3	0.75	1	**0.75**	0.72	1	0.72

Best Decision Tree for LKER_3: Entropy (Accuracy: 0.80, Simplicity: 1)
Best Decision Tree for MRCZ_3: Entropy (Accuracy: 0.75, Simplicity: 1)

Q1.3 Summary of Results

- Winner: LKER_3 (Entropy-based DT) with Accuracy = 0.80 and Simplicity = 1.0
- Reason: Highest weighted score (0.80), clearer rule structures, more interpretable clusters

Q1.4 Compare Clustering Results Using DT Rules

LKER_3 DT Rules (Entropy):

- Cluster 0: Characterized by high revenue growth and moderate asset size
- Cluster 1: Dominated by firms with high profits and global presence
- Cluster 2: Smaller firms with low asset turnover and regional influence

MRCZ_3 DT Rules (Entropy):

- Cluster 0: Balanced profile but with low capital efficiency
- Cluster 1: Rapidly growing companies with lean operating models
- Cluster 2: Conservative, legacy companies with stable earnings

Question 2: DMAGECR Dataset

Q2.1 Clustering

- Dataset: DMAGECR
- Target Variable for DTs: good_bad (used only for DT evaluation, not for clustering)
- Sampling: Stratified; Train = 80%, Test = 20%

Segmentation Models:

Label	Algorithm	Init Centers	Distance Measure	Normalization	No. of Clusters
LKER_2	Lloyd	k-means++	Euclidean	Range	2
ERER_2	Elkan	Random	Euclidean	Z-score	2

Q2.2 Decision Trees

Tree Hyper-parameters:

- Splitting Criteria: Entropy, Gini
- Min Samples per Leaf: 6% of smallest cluster
- Max Depth: 3

Evaluation Criteria:

- Accuracy (Weight: 0.60, Threshold > 0.70)
- Lift @ 2nd Decile (Weight: 0.30, Threshold > 0)
- Stability (Weight: 0.10, Threshold > 0)

Segment	Cluster	Accuracy (Best DT)	Proportion	Lift	Stability
LKER_2	C1	0.77 (Entropy)	0.55	1.78	0.012
	C2	0.73 (Gini)	0.45	1.79	0.017
ERER_2	C1	0.72 (Entropy)	0.50	1.76	0.017
	C2	0.70 (Gini)	0.50	1.76	0.019

Q2.3 Summary of Results

Presentation Table for Overall Accuracy Calculation:

Segmentation	Cluster C1			Cluster C2			Overall Accuracy
	Accuracy	Best Acc (acc1)	prop1	Accuracy	Best Acc (acc2)	prop2	acc_{Seg}
LKER_2	Entropy	0.77	0.55	Gini	0.73	0.45	**0.752**
ERER_2	Entropy	0.72	0.50	Gini	0.70	0.50	**0.71**

Q2.4 Compare the Accuracy

Best DT from CE1 Q1 (from en_6): Accuracy = 0.74, Lift = 1.214

Comparison with CE1:

- LKER_2 outperforms CE1 in terms of accuracy (0.752 > 0.74)
- ERER_2 does not outperform CE1 (0.71 < 0.74)

Conclusion for Q2:

- Winner: LKER_2 segmentation with weighted DTs showing best performance (Accuracy > 0.70, Lift > 1.75, Stability > 0.01)
- Overall: LKER_2 is more effective than CE1's best model in this scenario

Final Notes

- Both parts of CE4 demonstrate the strength of well-parameterized clustering followed by DT modeling
- For Q1, LKER_3 (Entropy) is recommended
- For Q2, LKER_2 segmentation yields the best model performance and outperforms CE1

Google collab link -
https://colab.research.google.com/drive/1Px1tlG1tiwkBmZK_DpbDYmxLnvWzNOOI?usp=sharing