施策効果を考慮した各特徴量の影響度指標に基づく顧客セグメンテーション手法の提案

1X19C067-7　清水成
指導教員　後藤正幸

1. 研究背景と目的

顧客ニーズの多様化に伴い，企業は顧客全体に対し一律の施策を講じるのではなく，顧客グループ毎に適切な施策を講じることが重要となっている．そのためには，特徴に応じて顧客を分類する顧客セグメンテーションが必要である．従来，顧客セグメンテーションには k-means 法などのクラスタリング手法や決定木などの分類モデルが用いられてきた．しかし，これらの手法は顧客ごとに異なる施策の効果を決める要因を十分に考慮しているとはいえない．

顧客セグメンテーションの本来の目的は，セグメントごとに適切な施策を講じることで施策効果を向上させることである．従って，施策効果を決める要因が類似したセグメントを形成することが求められる．そこで本研究では，機械学習モデルの解釈手法で知られる SHAP 値 [1] ベクトルを用いたクラスタリングを行うことで，施策によって期待される効果（アウトカム指標）への特徴量の影響の仕方が類似した顧客同士をセグメント化する手法を提案する．これにより，アウトカム指標の類似性だけでなく施策効果を決める要因の類似性も考慮することが可能となるため，顧客セグメント毎に最も効果的な施策を講じることが可能になる．本研究では，人工データを用いた実験を行い，提案手法の有効性を示す．

2. 準備

2.1. 顧客セグメンテーション手法

顧客セグメンテーション手法は中村ら [2] の分類によると，教師なしデータによるものと教師ありデータによるものに分けられる．教師なしセグメンテーション手法は，顧客属性や購入商品の類似性に基づいて顧客セグメントを形成する手法であり，k-means や潜在クラスモデルなどのクラスタリング手法が用いられる．しかし，これらの手法の多くでは施策で期待される効果を考慮していない．一方，教師ありセグメンテーション手法は，ビジネスのアウトカム指標を目的変数とした分類モデルを用いて，各顧客の特徴量から属するセグメントを判別する手法であり，判別分析や決定木などが用いられる．しかし，これらの手法を適用すると，目的変数の値の大小をもとにセグメントが形成されるため，目的変数への特徴量の影響度を考慮しているとはいえない．

2.2. SHapley Additive exPlanation (SHAP)

SHAP [1] は複雑な機械学習モデルを説明する手法である．p 個の特徴量から目的変数の値を当てはめるように学習された予測モデルを f としたとき，n 番目のインスタンス $\boldsymbol{x}_n = (x_{n1}, x_{n2}, \cdots, x_{np})$ の予測値 $f(\boldsymbol{x}_n)$ と予測平均 $E[f(\boldsymbol{x}_n)]$ の差は，式 (1) のように入力特徴量 x_{ni} の貢献度 φ_{ni} の和で表される．

$$f(\boldsymbol{x}_n) - E[f(\boldsymbol{x}_n)] = \sum_{i=1}^{p} \varphi_{ni} \tag{1}$$

この特徴量貢献度 φ_{ni} が SHAP 値となる．特徴量の SHAP 値への変換は，すべての特徴量がアウトカム指標の予測値の単位に変換され，かつ予測値への影響度に応じて自然に重み付けされることを意味する．SHAP 値を用いたクラスタリングのアイディアは文献 [3] でも述べられており，階層型クラスタリングによってアウトカム変数の予測理由が類似したインスタンスを樹形図にまとめ，全体傾向が分析できることが示されている [3]．しかし，このクラスタリング分析では予測理由のパターンや規則性の解釈が主目的であり，顧客セグメンテーションへの適用については論じられていない．本研究では，アウトカム指標に対する施策効果の類似性が高いグループを構成するために SHAP 値を用いることで，顧客セグメンテーション手法への適用を行う．

3. 提案

3.1. 提案への着想

施策効果への各特徴量の影響度を考慮したセグメンテーションを行うことで，顧客セグメントごとに最も効果的な施策を講じることができると考えられる．しかし，従来の顧客セグメンテーション手法は各特徴量の類似性などによりクラスタがつくられるため，アウトカム指標への特徴量の影響度が異なるデータが同じクラスタにまとめられてしまう可能性がある．そこで，SHAP 値を用いたクラスタリングを行うことで，施策で期待される効果に関連したアウトカム指標への各特徴量の貢献度が類似した顧客同士をセグメント化する手法を提案する．

3.2. 提案手法

提案手法について，顧客セグメントの形成から施策考案までの流れを 3 つのステップに沿って説明する．

STEP1) モデルの学習

SHAP はモデルの予測値を特徴量の貢献度で説明する手法であるため，まずはアウトカム指標を目的変数とした予測モデルを構築する必要がある．本研究では，顧客の特徴量を説明変数とし，施策で期待される効果に関連したビジネス上のアウトカム指標を目的変数とする予測モデルを構築する．

STEP2) SHAP 値を用いたクラスタリング

学習した予測モデルを用いて顧客の特徴量から SHAP 値を算出し，SHAP 値ベクトル $\boldsymbol{\varphi}_n = (\varphi_{n1}, ..., \varphi_{np})$ を用いたクラスタリングを行う．SHAP 値は顧客の特徴量を施策で期待される効果の影響度で重み付けした値となる．このとき得られた各クラスタ，すなわち顧客セグメントは，予測に対し類似した特徴量の影響をもつインスタンスで構成される．

STEP3) 得られたセグメントの解釈

各顧客セグメント内の SHAP 値に基づき有効な施策を検討する．SHAP 値を分析することにより，施策で期待される効果に大きく影響していると考えられる特徴量に絞って施策を検討することが可能となる．

4. 人工データを用いた実験

4.1. 実験条件

提案手法の有効性を検証するため，人工データを用いた実験を行う．ここで，人工データはセグメント毎に異なる特徴量が目的変数に影響を与えるように構成した．具体的には，真のセグメント数を 10，セグメント毎の顧客データ数を 100，顧客特徴量の数を 10 とした．また，データはすべて μ を設定パラメータとして正規分布 $\mathcal{N}(\mu, 3^2)$ から生成する．第 i セグメントのデータは，同じ添え字の特徴量 x_i に対し，100 個中 50 個を $\mu = 10$，残りの 50 個を $\mu = -10$ とし，それ以外の特徴量はすべて $\mu = 0$ として特徴ベクトルを生成した．対応する目的変数は，μ を顧客セグメント毎にセグメント番号の小さいほうからそれぞれ $\pm 100, \pm 110, \pm 120, \pm 130, \pm 140$ とした．作成した人工データは各セグメントでセグメント番号と同じ添え字の特徴量のみが 0 以外の平均を持つ正規分布に従うため，その特徴量が目的変数に影響を与える変数となる．また，各セグメント内の目的変数は同じ分布に従うため，その値（μ が 10 か-10）によらず目的変数への影響度の大きさも同じであるといえる．

提案手法で扱う予測モデルにはランダムフォレストを使用し，SHAP 値の計算には treeSHAP [3] を使用した．また，SHAP 値ベクトルのクラスタリングには k-means 法を使用した．比較手法である教師なしセグメンテーション手法には k-means 法を，教師ありセグメンテーション手法には決定木モデルを使用した．

4.2. 実験結果

表 1 にクラスタ数 10 のときの提案手法で得られた各セグメントの特徴量毎の SHAP 値の平均値を示す．作成した人工データでのセグメント番号と実験で得られたセグメント番号は一致している．また，濃い灰色部分は各セグメントで SHAP 値の絶対値が最大の値を表している．

表 1: 各セグメントの特徴量毎の SHAP 値の平均値

pred_segment	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1.0	-75.58	4.70	5.16	6.06	5.51	-5.69	-7.13	-7.61	-6.66	-7.51
2.0	4.42	-83.09	5.17	4.11	6.61	-5.20	-6.43	-8.20	-7.45	-9.66
3.0	3.67	3.87	-90.58	4.73	6.64	-7.92	-7.64	-6.36	-7.27	-7.79
4.0	3.69	2.58	4.35	-97.55	5.14	-5.62	-7.33	-6.43	-7.30	-7.46
5.0	3.57	3.11	5.42	4.03	-101.63	-7.64	-6.24	-6.16	-9.26	-10.06
6.0	9.44	8.68	9.25	11.66	12.37	16.10	-2.51	-4.46	-4.55	-7.02
7.0	12.28	11.99	12.00	10.04	15.83	-2.10	46.64	-2.17	-2.63	-5.38
8.0	11.15	11.63	14.44	12.83	12.89	-1.80	-2.37	52.80	-2.91	-4.51
9.0	12.96	11.34	11.90	12.68	10.92	-1.94	-0.65	-1.55	63.74	-3.82
10.0	12.67	12.47	14.29	12.75	13.46	-1.09	-2.06	-1.99	-2.27	67.02

表 1 より，人工データで生成時に目的変数に影響を与えるように設定した特徴量の SHAP 値は，他の SHAP 値よりも大きな絶対値を取っていることが確認できる．

次に提案手法と教師なしセグメンテーション手法である k-means 法で形成されたセグメントを真のセグメントとの類似性を測るランド指数及び調整済みランド指数 [4] で評価する．20 回の試行の結果から得られた平均値と 95 %信頼区間を表 2 に示す。

表 2: 各手法のランド指数及び調整済ランド指数

手法	ランド指数	調整済ランド指数
提案手法	**0.948**($\pm$0.007)	**0.727**($\pm$0.031)
k-means 法	0.848($\pm$0.001)	0.181($\pm$0.005)

表 2 より，提案手法はランド指数及び調整済ランド指数のスコアで k-means 法を有意に上回っており，k-means 法よりも真のセグメントを再現できていることがわかる．

最後に人工データに対し教師ありセグメンテーション手法である決定木を適用した結果を示す．図 1 は目的変数の予測精度を示す R^2 スコアと決定木により生成された葉ノード数の関係を示している．

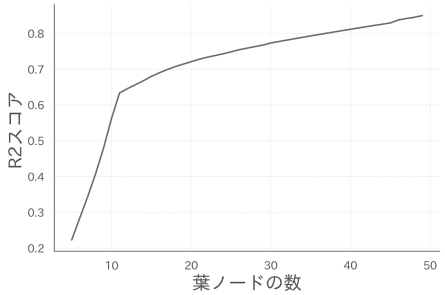

図 1: R^2 スコアと葉ノード数の関係

決定木では生成された葉ノードをセグメントとするが，図 1 より，真のセグメント数と同じ葉ノード数が 10 である場合の R^2 スコアは 0.56 と予測精度が低くなっており，真のセグメントを形成できていないことがわかる．一方，予測精度が高いところでは葉ノード数が多くなってしまっているため，生成されたセグメント及び特徴量の解釈が難しい．

5. 考察

実験より，提案手法は従来手法と比較して，目的変数への各特徴量の影響の仕方を考慮したセグメントを形成することができることを確認した．従って，提案手法は施策で期待される効果への特徴量の影響度が類似した顧客同士をセグメント化することができると考えられる．また，得られたセグメントの特徴的な SHAP 値を分析することで，顧客セグメント毎に最も効果的な施策を講じることが期待できる．

6. まとめと今後の課題

本研究では，施策で期待される効果への特徴量の影響の仕方が類似した顧客同士のセグメント化を行う手法を提案した．さらに，人工データを用いた実験を行うことで，提案手法の有効性を確認した．今後は実データに対し，提案手法を適用し実際にマーケティング施策に有用な情報を得ることができるか検証することが望まれる．

参考文献

[1] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, Vol. 30, , 2017.

[2] 中村博, 熊倉広志, 生田目崇. 顧客セグメンテーションのための分析手法とその効果:id-pos データをもとにした事例. 専修大学商学研究所報, Vol. 42, No. 5, pp. 1–28, 2011.

[3] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[4] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, Vol. 66, No. 336, pp. 846–850, 1971.