6K-3

Rによる楽天公開データに対するマイニング

劉 鎧誠　　山口 亨　　大楠 拓也　　徐 海燕

福岡工業大学

1. はじめに

近年, 蓄積された大量のデータからビジネスに活用できる有用な情報を取出すために, 「データマイニング」に関する研究が盛んに行われている. 本研究では, 楽天公開データ [2] 中のクチコミ情報に着目し, R 言語 [1] を利用したポジショニング分析と CS ポートフォリオによるデータマイニングを行っている. 50 代以上の顧客の評価が高い, 大都市がある都道府県の評価が低いという結果が得られている. 改善すべき項目として, 「距離の長さ」と「フェアウェイの広さ」があげられる結果も得られている.

2. 基本的事項

研究に当たって, 楽天公開データセット中の商品データセット中, 4つのデータベースの1つである golf データのクチコミ情報（合計 318389 件）を利用している. 項目としては1から5までの段階で評価された図1に示された, クチコミ ID, コース ID, 都道府県, 年齢平均, スコア, 利用回数, 総合評価, コストパフォーマンス, スタッフ接客, コース戦略性, 食事が美味しい, 設備が充実, フェアウェイが広い, 距離が長いという 14 個の項目である.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	クチコミID	コースID	都道府県	年齢	平均スコア	利用回数	総合評価	コストパフォ	スタッフ接客	コース戦略	食事が美味	設備が充実	フェアウェイ	距離が長い
2	535770	1148	13	40	0	4	3	3	4	4	5	2	3	2
3	625392	1148	13	40	80	0	5	5	5	5	5	2	3	1
4	677244	1148	4	30	100	6	3	2	4	2	3	2	3	3
5	683329	1148	6	30	100	2	4	4	5	3	5	2	3	3
6	689360	1148	4	50	0	1	4	2	4	4	4	3	2	2
7	702999	1148	13	40	90	13	4	4	5	4	5	2	2	2
8	704024	1148	13	40	90	13	4	4	5	4	5	2	2	2
9	813074	1148	4	0	87	0	4	3	4	3	2	2	3	3
10	582991	1407	14	40	95	1	4	2	3	4	4	3	3	2
11	588318	1407	14	40	100	118	3	3	3	3	3	3	3	3
12	510960	1407	14	40	105	2	4	5	5	4	5	4	4	3
13	538210	1407	14	50	84	8	3	2	4	3	3	3	3	2

図 1　クチコミ情報データ

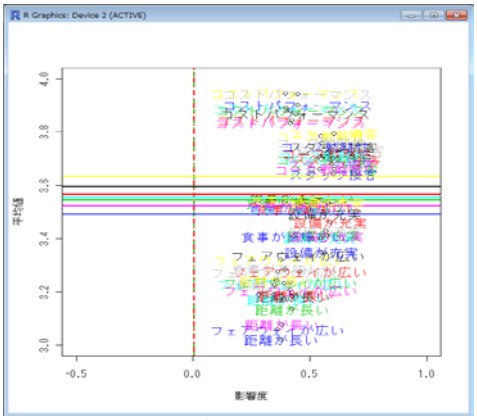
データマイニングには, CS ポートフォリオ分析とポジショニング分析という二つの手法を用いている. CS ポートフォリオは, 顧客の評価データから改善優先度の高い項目を抽出する大変便利なチャートである. ポジショニング分析によって, クチコミ情報に含めた評価データから, 顧客の感じるゴルフ場のポジションと目指す方向を明らかにして, 地域, 世代の相対的な位置からゴルフ場の強み・弱みを把握することができる.

3. クチコミ情報による CS ポートフォリオ分析

各地域, 各世代の顧客の評価データより, 総合評価への影響度を横軸, 各項目の評価の平均値を縦軸とした2次元マップを作成し, 分析を行っている.

3.1 各地域の CS ポートフォリオ分析

日本の八つの地域により, CS ポートフォリオマップは図2に示している. 右下にある評価項目「フェアウェイが広い」, 「距離が長い」が「要緊急改善」の項目である. 一方, 「コストパフォーマンス」, 「スタッフ接客」, 「コース戦略性」は満足度が高い項目である.



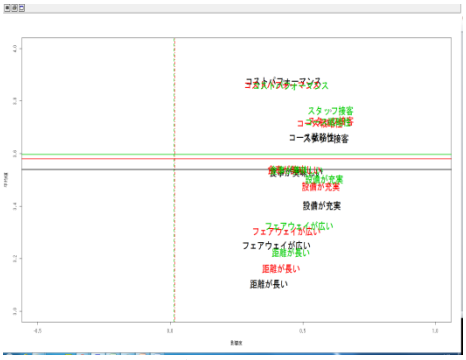
凡例
関東：黒
近畿：赤
九州：緑
四国：青
中国：ブル
中部：紫
東北：金
北海道：灰

図 2　地方より CS ポートフォリオ

地域に関しては, 各項目の得点を平均し, 合計した結果より, 評価の昇順は, 四国, 中部, 九州, 中国, 北海道, 近畿, 関東, 東北のという順番である.

3.2 各世代の CS ポートフォリオ分析

顧客の年齢情報には, 10 代から 100 代までの年齢層があるが. 10 代と 20 代, 30 代と 40 代, そして 50 代以上という三つの層に分けて. CS ポートフォリオマップを作成している（図 3）. 全体的には年齢が上がると, 評価が良くなる. さらに, 各年齢層とも, 「距離が長い」と「フェアウェイが広い」に関する評価が低く, 総合評価への影響し, 緊急改善項目である.



凡例
10 代, 20 代：黒
30 代, 40 代：赤
50 代以上：緑

図 3　3つ層別の CS ポートフォリオマップ

3.3 地域と世代による CS ポートフォリオ分析

人口多い都道府県（東京都, 大阪府, 愛知, 福岡）と一般的な都道府県（群馬, 奈良, 山梨, 佐賀）を4つずつ選んで. 30 代と 40 代を一つのグループ, 50 代以上の世代を一つのグループし, 分析を行った. 人口多い都道府県は一般な都道府県より平均評価が低く, 50 代以上のグループは 30, 40 代グループより評価が良くなる結果が得られている.

4. クチコミ情報によるポジショニング分析

ポジショニング分析で用いられる統計解析手法は，多変量解析の「因子分析」と「重回帰分析」である．

4.1 日本地域と都道府県のポジショニング分析

地域別と都道府県別にポジショニング分析を行っている．因子の数を決めるため，主成分分析手法を利用し，分析結果の固有値から，評価項目，すなわち，説明変数を２個の因子にすることと判断した．

日本地域ポジショニングマップは図 4 に示している．まず，因子分析手法を利用し，七つの評価項目を２個の因子に集約している．因子分析の結果より，因子１と因子２の意味付けは，「設備が充実」と「距離が長い」である．

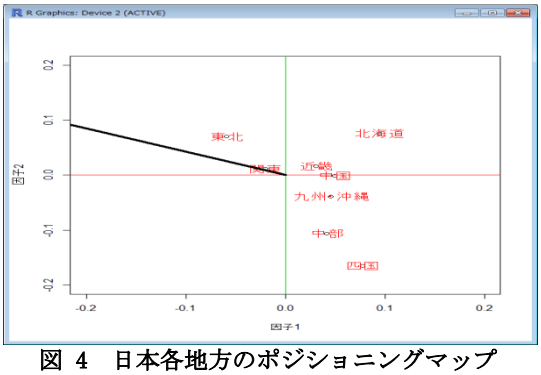


図 4　日本各地方のポジショニングマップ

理想ベクトルは２次元である．因子１と因子２を 0.60：0.25 の割合で重視していると見ることが可能である．選好方向は，因子１の負方向，因子２の正方向である．結果としては，東北と関東選好方向の次元にあり，九州，中部と四国は選好方向の逆の次元にある．

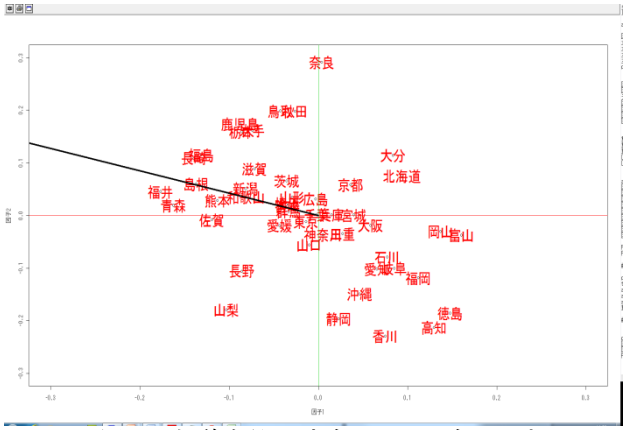


図 5　都道府県のポジショニングマップ

図 5 の都道府県のポジショニングマップにおいては，東北地方にある六つの都道府県のうち宮城県以外五つの都道府県が選好方向の次元にある．関東地方は選好方向である次元に入っているが，東京都は因子２の得点が低く，神奈川県は因子１と因子２の得点とも低い．

4.2 世代のポジショニング分析

全世代よりポジショニングマップは，図 6 に示している．60 代から 100 代までは選好方向である次元に入っている．件数一番多かった 30 代と 40 代は原点付近であることが分かった．

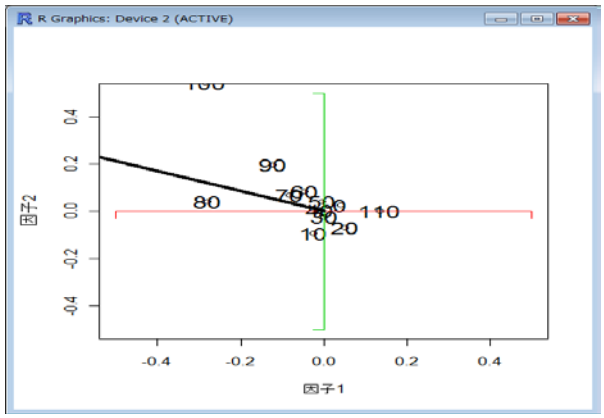


図 6　世代のポジショニングマップ

30 代と 40 代の顧客からのクチコミ情報は全部データ件数の３分の２を占用しており，大きなターゲット層として扱うことができる．まず，主成分分析の結果は，因子の数が３個となった．３個の因子は，「設備が充実」，「距離が長い」，「コストパフォーマンス」である．新たに出た第３因子は，30 代と 40 代は別の世代より経済的な面も重視していることが伺える．

4.3 地域と世代によるポジショニング分析

3.2 節のデータを用いて人口多い都道府県（東京都，大阪府，愛知，福岡）と一般的な都道府県（群馬，奈良，山梨，佐賀）の 30 代と 40 代グループ，50 代以上のグループに対するポジショニング分析を行った．

世代より，50 代以上グループは 30 代，40 代グループより因子２（距離が長い）の得点が高い．人口少ない都道府県は人口多い都道府県より因子２の得点が高いことが分かった．人口少ない都道府県（30 代，40 代）が理想ベクトル上であるが，因子１と因子２の得点が低く，原点に近い．人口多い都道府県（30 代，40 代）は理想ベクトルとは正反対にあるので評価が低くなっている．

5. まとめ

今回の研究では，R 言語を用いて楽天データセットの中ゴルフ場へのクチコミ情報 318389 件に対するポジショニング分析と CS ポートフォリオ及びポジショニングマップ分析及び視覚化を行った．総合評価に影響する因子，項目に着目し，地域別，都道府県別と世代別に分析を行った．全体的に東北地方と関東地方評判が高く，中部地方では顧客数が多いが，評価はあまり高くないことが分かった．世代においては，50 代以上は，評価が高かったが，全体の顧客数の３分の２を占める 30 代と 40 代は，「設備が充実」，「距離が長い」以外に，コストパフォーマンスも重視していることが判明した．

今回利用した CSV ファイルに対する前処理は，PHP 言語を通して，サーバー側から行っている．マイニングは，R 言語で行っている．データの前処理は，システム上の分析において一つ重要な課題である．楽天データセットの中，テキストデータが大量に存在しているが，それに対する前処理が今後の一つの課題である．

謝辞：楽天公開データを利用させてくださった楽天株式会社に感謝致します．

参考文献

1）石川 朋雄：商品企画のための統計分析―R によるヒット商品開発手法、オーム社（2009）

2）http://rit.rakuten.co.jp/rdr/index.html

Data Mining for Rakuten Data with R Programming Language

Kaicheng Liu　Fukuoka Institute of Technology
Toru Yamaguti　Fukuoka Institute of Technology
Takuya Ogusu　Fukuoka Institute of Technology
Haiyan Xu　Fukuoka Institute of Technology