修 士 論 文 概 要 書

Master's Thesis Summary

Date of submission: 01/11/2022 (MM/DD/YYYY)

専攻名(専門分野) Department	経営システム 工学専攻	氏 名 Name	石倉 滉大 Kodai Ishikura	指 導 教 員 Advisor	後藤正幸 印 Seal
研究指導名 Research guidance	情報数理応用研究	学籍番号 Student ID number	CD 5221C004-7		
研究題目 Title	SHAP 値を活用した店舗販売データに基づく商品間の関係性分析モデルに関する研究 A Relationship Analysis Model between Products based on Store Sales Data with SHAP Values				

1. はじめに

一般的なスーパーマーケットでは何らかの方法で商品需要の予測を行い, 廃棄数が最も少なくなるような仕入戦略・品出し戦略を立案している. スーパーマーケットでは多数の商品が同時に販売されているため, 各商品の売上はその商品単体の需要量だけでなく, 他の商品の需要量および店頭在庫数などに左右される. このような商品間の関係には, ある商品が売り切れの場合に代わりに購入される代替関係や, 2 商品間で需要を奪い合う競合関係, 高確率で同時に購入される同時購買関係などが考えられる. しかしながら, このような商品間の関係性は定量的に分析が行われていない場合が多い.

商品間の関係性を特定する関連研究として, 納豆市場における商品間の関係性を特定した大石の研究[1]がある. 大石の研究[1]は Almost Ideal Demand System[2] (以下, AIDS モデル) を利用して経済学の観点から市場の価格弾力性を推定し商品間の関係性を分析しているが, 消費者は理想的な環境で理想的な購買を行うという強い仮定を置いた予測モデルに基づく手法である. そのため, 実際の購買行動とモデルの推定結果が乖離している可能性がある.

そこで本研究では, スーパーマーケットにおける購買履歴データから, 実際の顧客の購買行動を反映した商品間の関係性を特定し, 小売店の経営戦略における有用な知見を得ることを可能とした分析手法の開発を目的とする. 本研究では, 商品間の関係性を有する予測モデルを構築し, その予測モデルを説明可能 AI 手法により解釈することで商品間の関係性を特定する. 具体的には, はじめにある時点での全商品の店頭在庫の状態を特徴量として入力し, 閉店時の廃棄数を予測するモデルを構築する. この予測モデルは, 入力された特徴量の中で尤もらしいパラメータの推定を行うため, 特徴量間の関係性がモデルに反映されていると考えられる. そのため, このモデルは分析者が想定している特徴量 (商品) 間の関係性だけでなく, 事前に想像していないような商品間の関係性をも表現している可能性がある. しかし, 多くの機械学習に基づく予測モデルは内部がブラックボックスであり, 直接的にモデルの表現内容を分析することは難しいとされている. そこで, 本研究では説明可能 AI 手法の一種である SHapley Additive exPlanations[3](以下, SHAP)を適用し, モデルの解釈を行うことを考える. SHAP は, 特定の予測に対し, 各特徴量がどの程度予測に影響を与えるかを定量的に示した Shapley 値を算出することで, 予測値の理由説明という観点からモデルの解釈を行う手法である. したがって, はじめに構築した予測モデルに SHAP を適用するにより, 商品間の関係性を特定することが期待できる.

本研究では, スーパーマーケットの実際の購買履歴データに提案手法を適用し, 商品間の関係性を特定するとともに考察を行うことで, 提案手法の有用性を示す.

2. 関連研究

本章では, 購買履歴データから商品間の関係性を分析する関連研究および説明可能 AI の関連研究について説明する.

2.1. 大石の研究[1]

大石は, ミクロ経済学における需要関数モデルである AIDS モデル[2]を購買履歴データに適用し, モデルが示した価格弾力性から納豆市場における商品間の関係性を分析した. AIDS モデルは, 理想的環境における消費行動をモデル化した手法であり, 市場全体の需要方程式を推定するモデルと, 各商品のシェア率を推定するモデルの 2 つを組み合わせて, 商品間の関係性の分析を行う. 市場全体の需要方程式を推定するモデルのモデル式は式(1)で示される.

$$\ln\left(\frac{X}{P}\right) = \psi + \delta \ln P \tag{1}$$

ここでXは市場全体の売上総額を, $\ln P$は物価指数を意味するストーン価格指数$\sum_j w_j \ln p_j$で与えられる. ただし, p_jは商品jの価格を表す. $\ln(X/P)$は実質所得を表す. δは市場全体の需要弾力性ϵの推定値と考えられる. また, 商品sのシェア率ω_sを推定するモデルのモデル式は式(2)で表される.

$$\omega_s = \alpha_s + \sum_k \gamma_{sk} \ln p_k + \beta_s \ln\left(\frac{X}{P}\right) \tag{2}$$

$$\gamma_{sk} = \frac{1}{2}(\gamma^*_{sk} + \gamma^*_{ks}) = \gamma_{ks} \tag{3}$$

$$\sum_m \alpha_m = 1, \sum_m \gamma_{mk} = \sum_m \beta_m = 0 \tag{4}$$

このとき, γ_{sk}は(X/P)が一定の下での, 支出シェアω_jに対する価格p_jの弾力性の大きさを表す. 式(2)におけるαとβはパラメータであり, それぞれ商品sのシェア率の基準値, 商品sが市場全体の価格傾向に影響される度合いを表す. また, 式(4)はシェア推定モデルにおける制約条件である. 以上のもとで商品sと商品kとの価格弾力性ϵ_{sk}は式(5)

で与えられ, これを全商品に対して算出することで市場内の商品間の関係性を分析できる.

$$\epsilon_{sk} = \frac{\gamma_{sk} + \epsilon\beta_s\omega_k}{\omega_s}\omega_k + (1+\epsilon) \tag{5}$$

2.2. 説明可能 AI 手法

現在, 機械学習を用いたサービスは日々増加しており, 世の中へ対する機械学習の影響度が大きくなるにつれ, サービス提供側へ説明責任が求められるようになっている. しかし, 機械学習モデルの内部構造はブラックボックスである場合が多く, そのままでは説明責任を果たすことが困難であるため[4], モデルの解釈可能性に関する研究領域が注目されている[5]. 特に説明可能 AI という分野の研究が盛んに行われており, 高いパフォーマンスレベルを維持しながら, より説明可能なモデルを生成する一連の技術を確立することを目的としている[4]. 説明可能 AI 手法は, 知覚的解釈可能性と数学的構造による解釈可能性に分類できる[5].

2.3. SHapley Additive exPlanations[3]

SHAP は関数関係に対して解釈性を付与する説明可能 AI 手法の一種であり, 任意の学習済みモデルに対して SHAP を適用することで, 各インスタンスに対する特徴量重要度を算出し, 定量的にモデルを解釈することができる. SHAP はある特徴量を入力した場合におけるモデルの予測結果と, その特徴量を入力しない場合におけるモデルの予測結果の差分をもとに, その特徴量が予測へ寄与した度合いを定量的に示す Shapley 値を算出する. データ数NでM次元の特徴量を用いた予測モデルに対し, $\chi = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}, \boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{iM})^{\mathrm{T}}$と定義すると, SHAP は予測値$f(\boldsymbol{x}_i)$を各入力特徴量に対応する Shapley 値$\phi_{i,k}$の和で表現する手法であり, 式(6)で表される.

$$f(\boldsymbol{x}_i) = f(\boldsymbol{z}_i') = \phi_{i,0} + \sum_{k=1}^{M} \phi_{i,k} z_{ik}' \tag{6}$$

ただし, $\boldsymbol{z}_i'$は$\boldsymbol{z}_i' \in \{0,1\}^M$であり, その要素である$z_{ik}'$は特徴量$k$をモデルに含む場合は 1 を, 含まれない場合には 0 が格納される変数である. また, $\phi_{i,0}$は入力する特徴量が全て欠落している場合の基準値であり, 一般にはモデルの平均予測値が与えられる. また, インスタンスiの予測における特徴量kの Shapley 値$\phi_{i,k}$は式(7)で与えられる.

$$\phi_{i,k} = \sum_{\mathcal{S}\subseteq\mathcal{M}\setminus k} \frac{|\mathcal{S}|(M-|\mathcal{S}|-1)}{M!}\left[f_{x_i}(\mathcal{S}\cup\{k\}) - f_{x_i}(\mathcal{S})\right] \tag{7}$$

$$f_{x_i}(\mathcal{S}) = E\left[f(\boldsymbol{x}_i)\middle|\boldsymbol{x}_{i,s}, s\in\mathcal{S}\right] \tag{8}$$

ただし, $\mathcal{M}$は全ての特徴量集合を, 係数は全ての組合せ数を表しており, 式(7)において, 全ての組合せでの特徴量kを含む場合とそうでない場合のモデルの予測の期待値の差分を, 特徴量kにおける Shapley 値と定義している. モデルの予測の期待値は, 式(8)で与えられる.

3. 提案手法

3.1. 着想

AIDS モデルは式(2)のw_sと$\ln(X/P)$が共に確率変数で, w_sは$\ln P$の関数となっている. これによって, 誤差項と説明変数が相関を持つため, 通常の手法では推定結果に偏りが生じてしまうといった問題点を内包する[6]. また, AIDS モデルでは消費者が理想的な行動をとることを仮定しているため, 必ずしもモデルが実際の購買行動を正確に表現しているとは限らない. 一方で, 需要予測モデルのような直接的に消費者の行動を予測するモデルは, AIDS モデルのような強い仮定を置かずにパラメータを推定しており, さらに非線形なモデルを使うことで高い表現能力を実現できる. そのため, より実際の購買行動を反映した特徴量間の関係性を表現可能と考える. またスーパーマーケットを経営している企業では, 既に独自の予測モデルを構築している場合も想定されるため, 解釈性のあるモデルを再構築するのではなく既存の予測モデルにも適用できる説明可能 AI 手法が望ましい. そこで本研究では, 学習済みの消費者の行動を予測するモデルに対し, 説明可能 AI 手法によって解釈を与えることで, 下記に示す商品間の関係性を分析する手法を構築する.

- 代替関係：ある商品が売り切れの場合に代わりに購入される
- 競合関係：2 商品間で需要を奪い合う
- 同時購買関係：高確率で同時に購入される

モデルを解釈する際, 上に示した関係性の特定を目的とした場合, 日ごとや他の商品の在庫量によっても消費者の行動は変化するため, 1 つのインスタンスに対する特徴量ごとの正負を含む予測への寄与の度合いが必要と考える.

以上のことから, 購買履歴データから構築した多商品需要予測のようなモデルに対し, SHAP を適用することで, 商品間の関係性を分析する手法を提案する.

3.2. 提案手法

以下に分析プロセスに提案する分析手法を示す.

分析プロセス

Step1) ある時点の全商品の店頭在庫量を特徴量, 閉店時の廃棄量を目的変数とする廃棄数予測モデルを商品別に構築する.
Step2) 式(7)により Shapley 値を算出する.
Step3) 式(10)により算出される評価指標がθ以上の特徴量kを選択し, 予測に特に寄与した特徴量を抽出する.
Step4) 有向ネットワーク図などを作成し, Step3 で選択された特徴量を分析する.

提案手法の Step1 では, まずある時点の全商品の店頭在庫量を特徴量, 閉店時の廃棄量を目的変数とする廃棄数予測モデルを商品ごとに構築する. このモデルの目的変数としては, 関係性の基準としたい購買行動を分析者が選択することができるが, 本研究では廃棄量を減少させることに寄与する商品間の関係性を分析するために, 廃棄量を目的変数とした. この予測モデルは, 入力された特徴量の中で尤度を極大化するようにパラメータを推定するため, 特徴量間（商品間）の関係性を反映していると考えられる. 続いて Step2 として, このモデルに, 説明可能 AI 手法の一種である SHAP[3]を適用し, 各特徴量の Shapley 値を

表 1. $r_{s,k}$ の解釈と商品間の関係性（相関の有無の基準を 0.5 とした場合）

$r_{s,k}$	$r_{k,s}$	予測に対する解釈	商品間の関係性
≥ 0.5		商品kの在庫量が多い場合， 目的変数である商品sの廃棄量を多くするような予測 （商品kの在庫量が多い場合，商品sは売れづらい）	商品sは商品kの 代替関係
≥ 0.5	≥ 0.5	両方向から代替関係が成り立っている	商品sと商品kは 競合関係
≥ 0.5	-	商品kの在庫量が多い場合， 商品sの廃棄数を少なくするような予測 （商品kの在庫と合わせて，商品sの廃棄量が減っている）	商品sと商品kは 同時購買関係
-	-	商品sは商品kをモデルに反映していない	商品sと商品kは 関係性がない

算出する．次に Step3 では，Step2 で算出した Shapley 値を用いて，商品間の関係性を定量的に算出する方法を考える．本研究では，競合関係，代替関係，同時購買関係，他の商品とは関係性ががないという計 4 つの商品間の関係性の特定・分析を目的とする．これらの商品間の関係性を特定するためには，2 商品間の在庫量の推移を考慮する必要がある．例えば，2 商品の在庫が同時に減る場合には，商品間の関係性は同時購買にある関係性といえる．また，ある商品が購入されると，もう一方の商品は購入されづらくなる場合には，代替関係があるといえる．このように商品間の関係性の特定には 2 商品の在庫量の推移は重要である．しかし，式(7)で与えられる Shapley 値$\phi_{i,k}$は 2 商品の在庫量の推移については評価できない．そこで，これら 4 つの関係性にある商品ペアを抽出するために，商品sの廃棄数予測の際に商品kが与える影響を示すShapley値と商品kの在庫量の相関係数を算出する．

$$r_{s,k} = \frac{SS_{X_k,\Phi_{s,k}}}{S_{X_k} S_{\Phi_{s,k}}} \tag{9}$$

ただし，SS_{XY}はXとYの共分散，S_XはXの標準偏差，X_Kは特徴量kの観測値つまりある時点における商品kの在庫量，$\Phi_{s,k}$は予測モデルf_sにおける特徴量kの全てのインスタンスの Shapley 値，f_Sは商品sに対する予測モデル，$r_{s,k}$は商品sに対する予測モデルにおける特徴量kと Shapley 値$\Phi_{s,k}$の相関係数を表す．また，表 1 に$r_{s,k}$の値の解釈について示す．表 1 の解釈のように式(9)で算出される$r_{s,k}$を評価する．ただし，表 1 では，例として相関係数の絶対値が 0.5 以上の場合を相関があるとしているが，その基準については分析者が選択することができる．しかしながら，強い正の相関を持っている場合であってもモデルの予測に大きく寄与していない場合には，これらの商品間の関係性はあまり重要ではない．そこで，予測に寄与した度合いと式(9)の相関係数$r_{s,k}$の両方を反映した評価指標を式(10)に示す．

$$L_{s,k} = \frac{\Phi_{s,k}|r_{s,k}|}{\sum_{k=1}^{M} \Phi_{s,k}} \tag{10}$$

ただし，|·|は絶対値を示す．式(10)で与えられる評価指標に閾値θを設け，θ以上の$L_{s,k}$を示す商品kを目的変数である商品sと関係性のある商品と定義する．最後に Step4 では，Step3 で特定された関係性のある商品を分析する．スーパーマーケットには多数の商品が販売されるため，1 商品ごとに関係性のある商品を確認することはコストがかかる．そこで，ネットワーク図を用いることでより豊かな分析をする方法を示す．この分析により，分析者が想像していないような商品間の関係性の発見が期待できる．

4. 実データを用いた分析

本章では，提案手法の有効性を示すため，提案手法を実データに適用し，分析を行う．

4.1. 分析条件

東海地方を中心にスーパーマーケットチェーンを展開する株式会社バローから提供された，スーパーマーケットにおける惣菜の販売，廃棄履歴データを対象とする．対象期間は 2018 年 1 月 2 日～2019 年 12 月 31 日とし，商品数は各店舗で年間 30 日以上売られていた 50 アイテムとした．また，予測モデルの精度を検証するため，2018 年を学習データ，2019 年をテストデータとした．推定精度の評価指標としてアイテムや店舗の全てに関して平均化した MAE とする．

また本分析では，説明変数をt時点での各商品の在庫量，目的変数を閉店時におけるある 1 つの商品の廃棄量とした回帰モデルを LightGBM[7]により惣菜アイテムの数だけ構築し，各商品における廃棄数を予測した．また，惣菜アイテムの需要は，季節や時間帯，気温によっても左右されるため，説明変数として気温，日射量，風速，風向，降水量，降雪量，相対湿度，天気といった気象データも併せて入力とする．説明変数に用いる各商品の在庫量を測る時刻tについてはいくつかのケースを分析したが，ここでは閉店 1 時間前の$t = 23$の結果を示す．

4.2. 分析結果と考察

4.2.1. 抽出された商品に関する定量的な分析

本節では，提案手法により抽出された商品に関する定量的な分析を行う．具体的には，全ての特徴量を入力として学習したモデルと各予測モデルf_sにおいて式(10)で算出される評価指標が 0.05 以上の特徴量kのみを入力とするモデルの比較を行う．表 2 に各モデルの MAE を示す．

表 2 より，LightGBM モデルでは，全ての特徴量を入力とするモデルよりも，提案手法によって選択された重要度の高い特徴量のみを入力とするモデルの方が，テストデータに対する推定精度が高い値を示した．これは予測に不要な特徴量を除外したことで，モデルが過学習するのを回

表 3. 商品ペアと各指標の値（左表：$|r_{s,k}|$上位 3 つの商品ペアを抜粋，右表：$L_{s,k}$上位 3 つの商品ペアを抜粋）

商品 1（商品s）	商品 2（商品k）	$r_{s,k}\downarrow$	$L_{s,k}$
デリシャス牛肉コロッケ	炭火焼つくねと野菜の甘辛丼	0.947	0.00317
12 種の鮨盛り合わせ	3 種のサーモン丼	-0.944	0.00702
ジャンボメンチカツ	デリシャスクリーミーコロッケ	0.942	0.135

商品 1（商品s）	商品 2（商品k）	$r_{s,k}$	$L_{s,k}\downarrow$
コク旨ソースの屋台風焼きそば	塩ゆで枝豆	0.607	0.256
ジャンボメンチカツ	デリシャスクリーミーコロッケ	0.942	0.135
お好み焼と焼そばセット(豚玉)	とろーり屋台風たこ焼	0.918	0.0777

表 2. 各モデルのテストデータに対する MAE

モデル名	MAE
LightGBM 全説明変数	0.899
LightGBM 変数選択	**0.487**

避し汎化性能が上昇したためと考えられる．

4.2.2. 抽出された商品に関する定性的な分析

本節では，式(10)によって導かれる商品間の関係性に関する分析を行う．図 1 に，評価指標の閾値を 0.02 と設定した際に抽出された商品間の関係性を可視化したネットワーク図を示す．また，エッジの色は式(9)によって導かれる商品間の関係性の種類を表し，青色は代替関係を，緑色は同時購買が起こりやすい関係を，赤色は競合関係を表す．

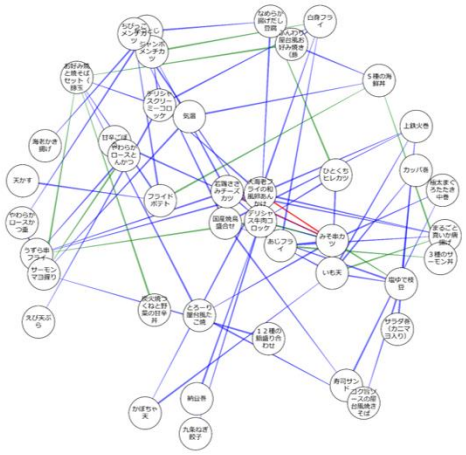

図 1. 商品間の関係性を示したネットワーク図

図 1 より，「みそ串カツ」と「大海老フライの和風卵あんかけ」が競合関係にあることを示している．これらの商品はどちらも揚げ物かつ一品料理の商品である．また，「うずら串フライ」と「デリシャス牛肉コロッケ」は同時購買が起こりやすい商品であることがわかる．これらの商品は酒のつまみとして購入されることが考えられ，実際の購買行動に即した分析結果といえる．また，「ジャンボメンチカツ」の代替として「デリシャスクリーミーコロッケ」が購入されていることを示す．これらの商品もどちらも揚げ物かつ一品料理の商品である．以上より，提案手法は実際の購買行動に即した商品間の関係性を検出可能といえる．

4.2.3. $r_{s,k}$, $L_{s,k}$に関する分析

本節では$r_{s,k}$および$L_{s,k}$に関する分析を行う．表 3 に$|r_{s,k}|$の上位 3 つの商品ペアと，$L_{s,k}$の上位 3 つの商品ペアを各指標の値と共に示す．

表 3 の商品ペアの中では，「ジャンボメンチカツ」と「デリシャスクリーミーコロッケ」および，「お好み焼と焼そばセット(豚玉)」と「とろーり屋台風たこ焼」の商品ペアは購買行動を容易に想像できる．これらの商品ペアは$r_{s,k}$と$L_{s,k}$がどちらも高水準な値を有する．そのため，$r_{s,k}$と$L_{s,k}$の 2 つの指標に対して閾値を設けて分析することにより，現実の購買行動に即した商品ペアの検出が期待される．

5. 結論と今後の課題

本研究では，スーパーマーケットの購買履歴データにおいて，実際の購買行動に即した商品間の関係性を特定する分析手法を提案した．具体的には，購買行動に対する予測モデルを構築し，モデルにおいて各特徴量が予測に与える影響を定量化した Shapley 値と特徴量の相関係数を計算することにより複数の商品間の関係性とその大きさを特定する手法を構築した．そして，実際の購買・廃棄履歴データに対して提案手法を適用し，提案手法の有用性を示した．今後の課題としては，より複雑な予測モデルを使用した際の提案手法の挙動の確認などが挙げられる．

謝辞

本研究は，株式会社バローおよび日本気象協会と早稲田大学の共同研究であり，データ提供元である株式会社バローおよび日本気象協会の皆様に深く感謝致します．

参考文献

[1] 大石敦志. ``日次 pos データによる納豆市場の AIDS 需要分析.” 食品経済研究, No. 36, pp. 54-72, 2008.

[2] Deaton Angus and Muellbauer John. ``An almost ideal demand system,” *The American economic review*, Vol. 70, No. 3, pp. 312–326, 1980.

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

[4] A. Adadi and M. Berrada. ``Peeking inside the blackbox: a survey on explainable artificial intelligence (xai),” *IEEE access*, No. 6, pp. 52138–52160, 2018.

[5] Tjoa and C. E., Guan. ``A survey on explainable artificial intelligence (xai): Toward medical xai,” *IEEE transactions on neural networks and learning systems*, No. 32(11), pp. 4793–4813, 2020.

[6] 溝渕健一, 谷崎久志. ``AI 需要システムによる弾力性の推定について: ブートストラップ法の応用,” 日本統計学会誌シリーズ J, No. 37(1), pp. 161–178, 2007.

[7] KE, Guolin, et al. ``Lightgbm: A highly efficient gradientboosting decision tree,” *Advances in neural information processing systems*, No. 30, 2017.