スーパーマーケットにおける ヨーグルト商品の離反予測モデルの提案

中央大学大学院　理工学研究科　ビジネスデータサイエンス専攻

山口晴輝

目次

1. 研究背景・目的

背景

ヨーグルト商品には，腸内環境を整える商品のような機能性を持つ商品が多く存在し[1,2]，そのような機能性は継続して摂取することで効果を発揮する[3].
また，ヨーグルト商品を販売する小売店は，特定の商品の売上ではなく，商品カテゴリごとの売上を向上させたいと考えている.

メーカーの視点	小売店の視点
• 自社商品の売上を伸ばしたい • 商品の機能性を実感してもらいたい • クーポンなどの施策実施のための資金を効率的に運用したい	• 店舗の売上を伸ばしたい ⇒ 特定商品より，商品カテゴリの売上を伸ばしたい • 自社アプリによるパーソナライズしたマーケティングを行いたい

メーカー・小売店の双方の要望を全て満たす施策立案は難しいが，それぞれの要望の一部分ずつ満たした施策立案は可能ではないかと考えた.

⇒ **ヨーグルト商品の継続購入**は，顧客に商品の効能をより多くもたらし，店舗やメーカーにとっては安定した収入源につながる.

1. 研究背景・目的

目的	「商品の機能性を顧客に提供したい」メーカーの要望と，「ヨーグルト商品の売り上げを伸ばしたい」小売店の要望を同時に満たす施策提案に向けた「**顧客の離反の可能性**」の推定.

ヨーグルト商品の継続購入者を増やすためには，ヨーグルト商品における離反顧客を予測し，離反する可能性の高い顧客に継続購入を促す施策を実施すべきと考えた.

この施策を提案する上で，各顧客の離反する可能性を数値として把握することができれば，離反可能性の大きさによって実施する施策を顧客ごとにパーソナライズすることも可能となり，施策実施のための費用を効率的に使うことが期待できる.

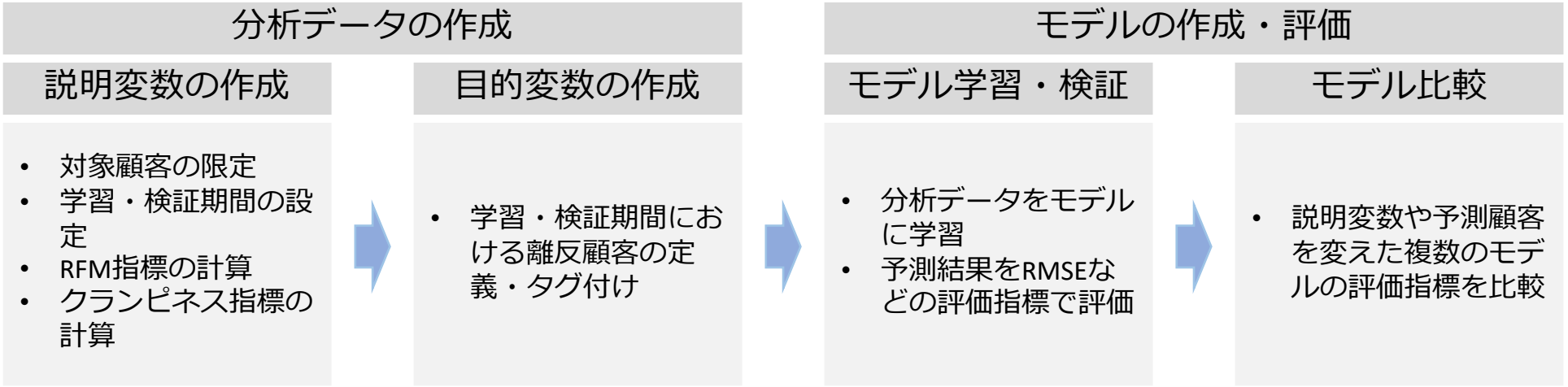
⇒近年，小売店ではセール情報の発信や割引クーポンの配布を，自社のスマホアプリで管理する動きが進んでいるため，パーソナライズされた施策提案が可能になりつつある.

施策例	• 離反可能性 0.8～1.0の顧客：ヨーグルト商品の２割引クーポンを配布 • 離反可能性 0.5～0.8の顧客：ヨーグルト商品の１割引クーポンを配布 • 離反可能性 0～0.5の顧客：ヨーグルト商品の機能性を特集した記事による宣伝

⇒メーカーはクーポンのために資金を投じる必要があるため，このように配布対象を限定できれば，資金を効率的に運用することが可能.

2. 本研究の概要

概要	あるスーパーマーケットのPOSデータを用い，ヨーグルト商品における顧客離反を予測するモデルを作成する． 顧客生涯価値の予測などに用いられるRFM指標に加えて，購買間隔の不均一性を評価するクランピネス指標を説明変数として，モデルの性能を評価する．



2. データ概要

あるスーパーマーケットのPOSデータ（購買履歴）データを使用.

概要	• 対象店舗数：1店舗（元データは18店舗のデータが混在していたため，売上高が最も高い1店舗を抽出） • 対象期間：2014年4月1日～2016年3月31日（２年間） • 対象顧客：4833人（学習データ），4877人（検証データ） 学習・検証データの期間において，ヨーグルト商品を3回以上購入した顧客（会員）を対象 ⇒ 学習・検証データの分割方法は後述

3. データ概要

元データから必要なデータを以下のように結合・抽出した.

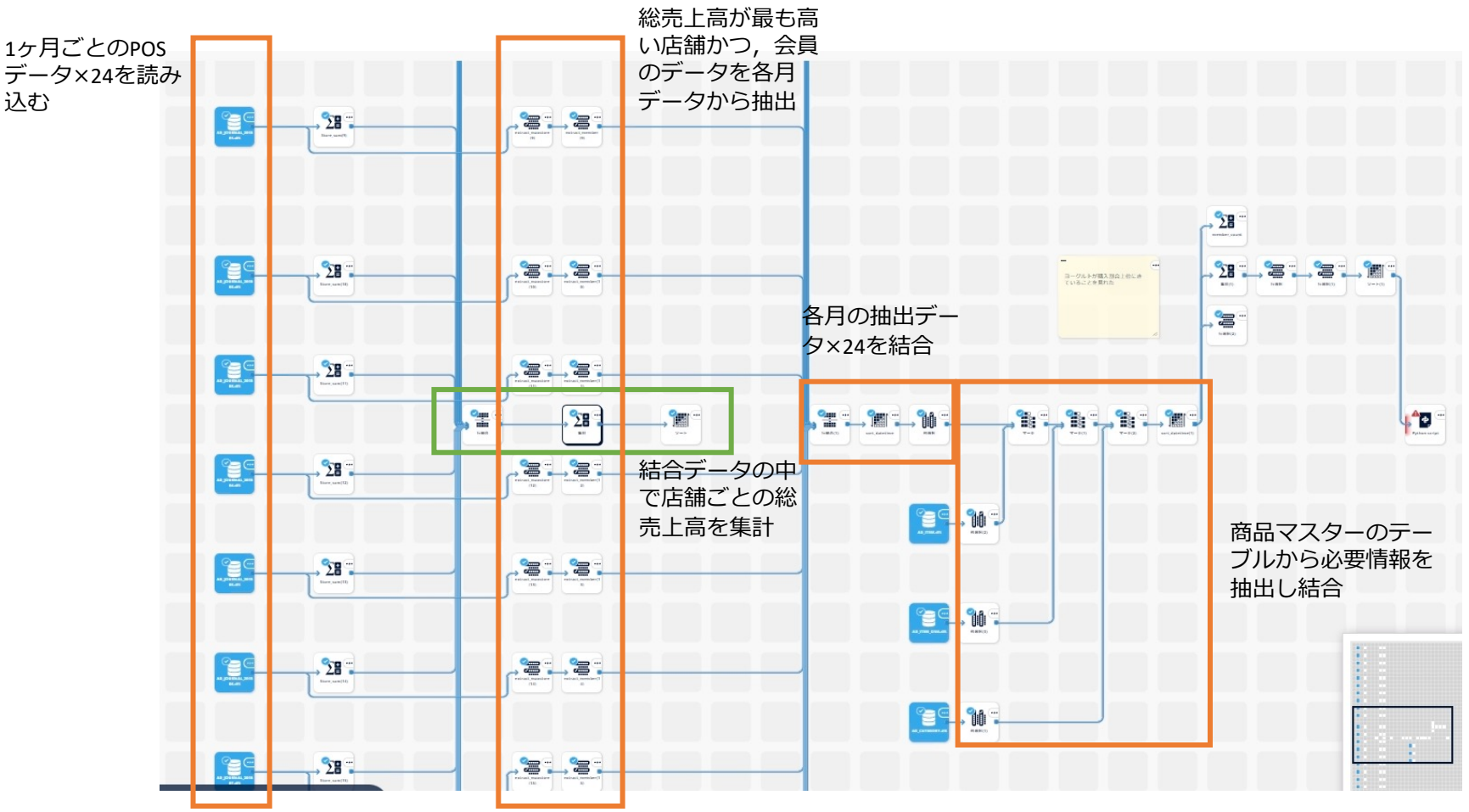


図１． Alkanoによる基本データの結合・抽出

3. データ概要

分析データに加工する前のデータの概要を以下に示す.

receipt_num	receipt_row	date	time	member_class	member_info	product_info	taxin_sumprice	taxin_price	amount	category_info	product_name	brand_name	volume	maker_name	category3_name	category4_name	datetime
20140403000050000010252182065	9	2014/04/03	18	1	b0de297b29c4ebae4c7588f682101e3e28076eb1ab01fa...	1068552065	91	91	1	1073.0	キリン大人のキリンレモンペット５００ｍｌ	キリン	500ml	キリンビバレッジ	清涼飲料	炭酸飲料	2014/04/03-18
20140403000050000010252182065	20	2014/04/03	18	1	b0de297b29c4ebae4c7588f682101e3e28076eb1ab01fa...	1152852065	597	597	1	565.0	日本ハムシャウエッセンスペシャルロング２５０ｇ	日本ハム	250g	日本ハム	加工肉	ウィンナー	2014/04/03-18
20140403000050000010252182065	35	2014/04/03	18	1	b0de297b29c4ebae4c7588f682101e3e28076eb1ab01fa...	660532065	152	152	1	626.0	モランボンワンタンスープの素７０ｇ	モランボン	70g	モランボン	加工調味料	スープベース	2014/04/03-18
20140403000050000010252182065	10	2014/04/03	18	1	b0de297b29c4ebae4c7588f682101e3e28076eb1ab01fa...	1030072065	101	101	1	1071.0	キリン午後の紅茶ストレートティーペット５００ｍｌ	キリン	500ml	キリンビバレッジ	清涼飲料	茶系飲料	2014/04/03-18
20140403000050000010252182065	30	2014/04/03	18	1	b0de297b29c4ebae4c7588f682101e3e28076eb1ab01fa...	641692065	152	152	1	636.0	中華用餃子の具味付け用たれ３５ｇ×２	モランボン	70g	モランボン	加工調味料	中華加工調味料	2014/04/03-18

図２．加工前の抽出・結合されたデータ

- receipt_num, receipt_row：レシート情報，レシート上の行番号
- date, time：販売日，時刻
- member_class, member_info：会員区分（0: 非会員，1:会員），会員情報
- product_info：商品情報（商品マスターと照合することでカテゴリ情報，商品名を参照可能）
- taxin_sumprice, taxin_price, amount：税込合計価格，税込商品単価，購入点数
- category_info：カテゴリー情報（カテゴリーマスターと照合することでカテゴリ名を参照可能）
- product_name：商品名
- brand_name：ブランド名
- volume：内容量
- maker_name：メーカー名
- category3_name, category4_name：商品カテゴリ（category1~5まであり，末尾番号が大きくなるほどカテゴリ粒度が細かくなる）

4. 分析　-基礎集計-

- 対象店舗における各商品カテゴリの売上高（上位15カテゴリ）

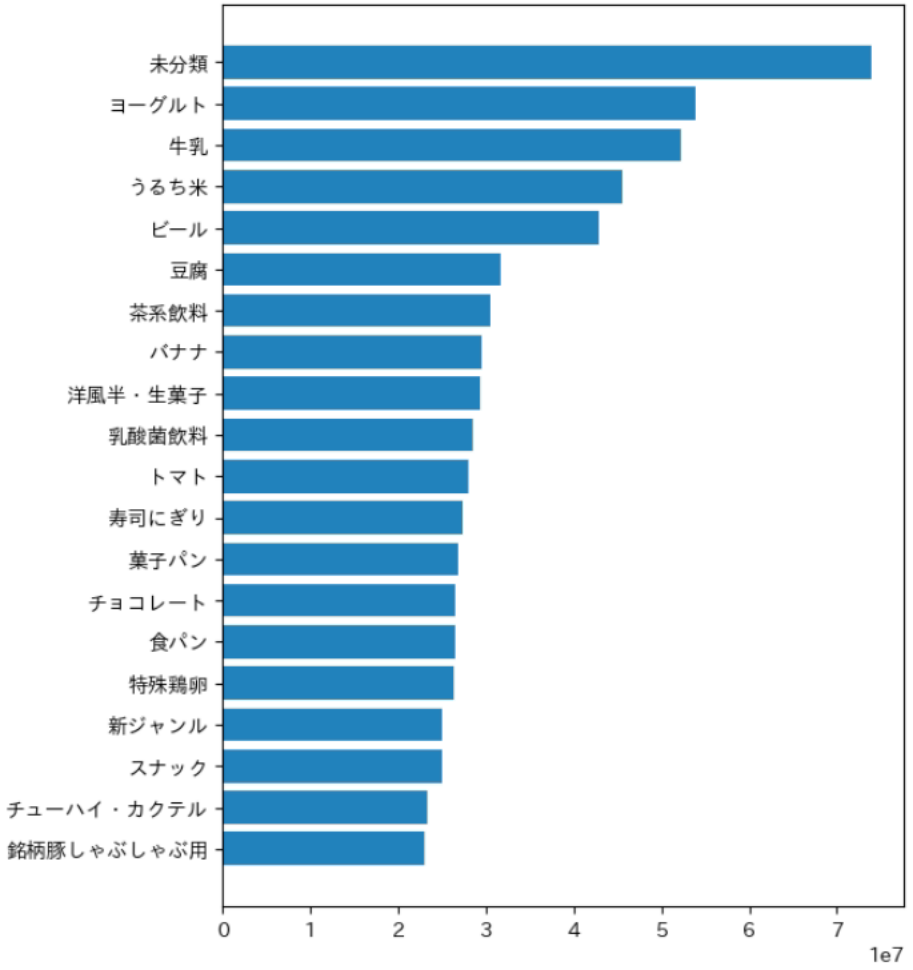


図３．対象店舗２年間のカテゴリ別総売上高

左図は，対象店舗の２年間のカテゴリ別合計売上高（会員・非会員含めて）の集計結果である．

未分類カテゴリにはcategory4の粒度では分類できなかった様々なカテゴリが分類されているので，実質ヨーグルト商品が売上高の１番高いカテゴリであることを示している．

4. 分析　-変数作成-

- **学習・検証データの分割**

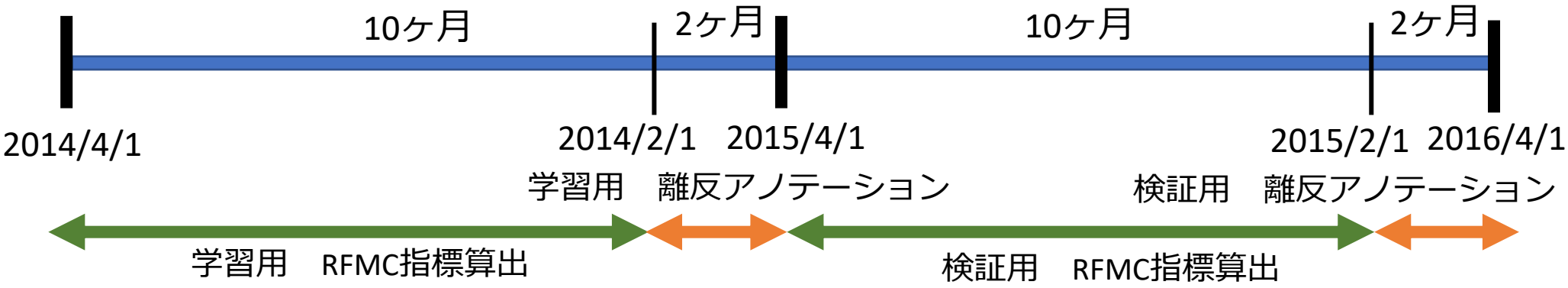


図４．学習・検証データの分割期間

- 学習データ期間を「2014年4月1日～2015年3月31日」，検証データ期間を「2015年4月1日～2016年3月31日」の１年間ずつで設定した.
- それぞれ期間では前10ヶ月間を説明変数とするRFMC指標（RFM＋クランピネス指標）を算出するための期間とし，残り2ヶ月を目的変数である各顧客の離反有無を判定する期間とした.
- 分析対象とした顧客は，それぞれの10ヶ月間において3回以上ヨーグルト商品を購入したことのある顧客とした.

4. 分析　-変数作成-

RFM指標

RFMとは以下の３つの指標の頭文字を組み合わせた名称であり，顧客生涯価値の算出などに用いられ，顧客の性質を把握することのできる指標であるとされている.

- 最終購入日（Recency）
- 購入頻度（Frequency）
- 購入金額（Monetary）

この３つの指標を，対象顧客の各10ヶ月間の期間内で算出した．その際に，全ての購買におけるRFM指標と，ヨーグルトを購入した日のみ抽出したデータにおけるRFM指標を計算した.

	date	member_info	product_info	taxin_sumprice	amount	ytag
0	2014-04-01	0035b560aff87d452918817ee3352c0213ef13f1680a30...	8	2522	11	0
1456	2014-04-02	0035b560aff87d452918817ee3352c0213ef13f1680a30...	12	4382	13	1
4321	2014-04-04	0035b560aff87d452918817ee3352c0213ef13f1680a30...	5	3396	7	0
5812	2014-04-05	0035b560aff87d452918817ee3352c0213ef13f1680a30...	10	4167	12	0
9231	2014-04-07	0035b560aff87d452918817ee3352c0213ef13f1680a30...	23	5775	31	0
...	...	...	...	...	...	...
529439	2015-03-20	0035b560aff87d452918817ee3352c0213ef13f1680a30...	6	1021	7	0
530989	2015-03-21	0035b560aff87d452918817ee3352c0213ef13f1680a30...	3	954	3	0
535780	2015-03-24	0035b560aff87d452918817ee3352c0213ef13f1680a30...	8	1374	9	0
542567	2015-03-28	0035b560aff87d452918817ee3352c0213ef13f1680a30...	20	6598	23	0
547362	2015-03-31	0035b560aff87d452918817ee3352c0213ef13f1680a30...	7	3201	8	0

図５．ある1人の顧客の全購買日データ

	date	member_info	category4_name	product_info	taxin_sumprice	amount
11553	2014-04-02	0035b560aff87d452918817ee3352c0213ef13f1680a30...	ヨーグルト	1	214	1
199023	2014-04-16	0035b560aff87d452918817ee3352c0213ef13f1680a30...	ヨーグルト	1	214	1
345288	2014-04-27	0035b560aff87d452918817ee3352c0213ef13f1680a30...	ヨーグルト	1	152	1
434662	2014-05-04	0035b560aff87d452918817ee3352c0213ef13f1680a30...	ヨーグルト	1	428	2
488946	2014-05-08	0035b560aff87d452918817ee3352c0213ef13f1680a30...	ヨーグルト	1	304	2
...	...	...	...	...	...	...
3619741	2015-01-30	0035b560aff87d452918817ee3352c0213ef13f1680a30...	ヨーグルト	1	204	1
3832502	2015-02-15	0035b560aff87d452918817ee3352c0213ef13f1680a30...	ヨーグルト	1	317	2
3848116	2015-02-16	0035b560aff87d452918817ee3352c0213ef13f1680a30...	ヨーグルト	1	518	3
3981490	2015-02-26	0035b560aff87d452918817ee3352c0213ef13f1680a30...	ヨーグルト	2	355	2
4204187	2015-03-15	0035b560aff87d452918817ee3352c0213ef13f1680a30...	ヨーグルト	1	183	1

図６．ある1人の顧客の全ヨーグルト購買日データ

⇒ 全顧客において，図5,6 のようにデータを抽出し，それぞれにおけるRFM指標を計算した.

4. 分析　-変数作成-

クランピネス指標

RFM指標は顧客関係管理(CRM)手法の１つして広く活用されてきた.

しかし，不定期のまとまった購買などはRFM指標だけでは識別できず，図7のようなケースはRFM指標だけでは同じ状態として扱われてしまう.

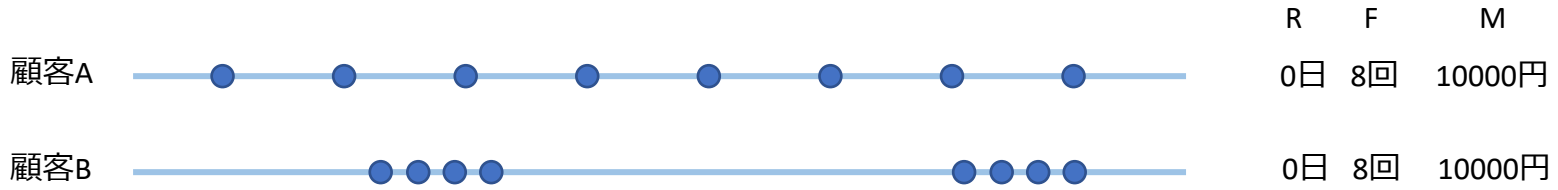


図７．RFM指標だけでは同一の状態として扱われるケース

クランピネス指標は，**イベントが均等な間隔に従わない度合い**を示し，イベント発生の不均一性を捉えることができるため，図7のような状況でもそれぞれにおいて違う値を記録する.

⇒ RFM指標に加えて，購買パターンの違いを数値データとして扱える.

4. 分析　-変数作成-

クランピネス指標

クランピネス指標H_pは，以下のように計算される[5].

t_i：購買機会N回のうちn回購買が行われた時，i回目に購買が行われたタイミング.

t_iを基準化したx_iを以下のように定義したとき，

$$x_i = \begin{cases} t_i & (i = 1) \\ t_i - t_{i-1} & (i = 2,3,\dots,n) \\ N + 1 - t_n & (i = n+1) \end{cases}$$

クランピネス指標は以下のように表せる.

$$H_p = 1 + \frac{\sum_{i=1}^{n+1} \log(x_i) x_i}{\log(n+1)}$$

4. 分析　-変数作成-

離反顧客のアノテーション

本研究では，RFMC指標の算出期間ではヨーグルトを購入していたが，離反アノテーション期間でヨーグルトを購入しなかった顧客を離反顧客として設定した.

学習・検証期間における継続顧客，離反顧客の人数は表１のようになった.

表１．離反顧客，継続顧客の人数集計結果

	学習データ	検証データ
継続顧客	3328	3303
離反顧客	1494 (373)	1574 (360)

離反顧客の括弧内の数値は，離反顧客内そもそもアノテーション期間中に来店をしていない顧客（店舗離反）の人数を示す.

アノテーション期間のヨーグルトの購買間隔が，前10ヶ月間のヨーグルト購買間隔より長くなっている顧客も離反とはならなくとも，購買頻度が下がっている傾向にあるため，離反タグづけを検討した．しかし，そのように離反の定義を広くした場合，離反顧客が全体の半数程度になってしまうことや，「ヨーグルトを買わなくなった顧客」と「ヨーグルトを買う頻度が下がった顧客」を同じ離反確率１のタグを付与することは適切ではないと考えた．そのため，本研究では上記定義で離反のタグ付けを行った.

4. 分析　-モデル学習・検証-

本研究では，**LightGBM**モデルを採用した.

LightGBM：決定木アルゴリズムに基づいた勾配ブースティングの機械学習フレームワーク.

複数の弱学習器（LightGBMでは決定木）をアンサンブル学習のブースティングという手法を用いてまとめ（図8），それぞれの弱学習器は前の弱学習器の誤差を学習している.

この決定木を用いた勾配ブースティングの機械学習フレームワークには，Xgboostというモデルもある.

Light GBMはXgboostと違い，Leaf-wiseという手法を用い，決定木の枝分かれの数を制限することで，学習時間を短縮するという利点がある.

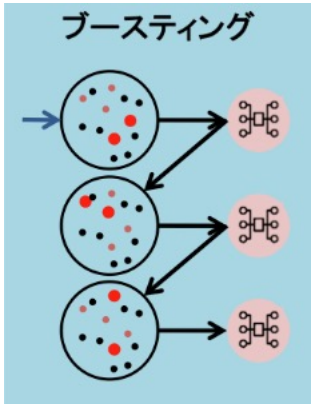


図8．ブースティングのイメージ図[6]

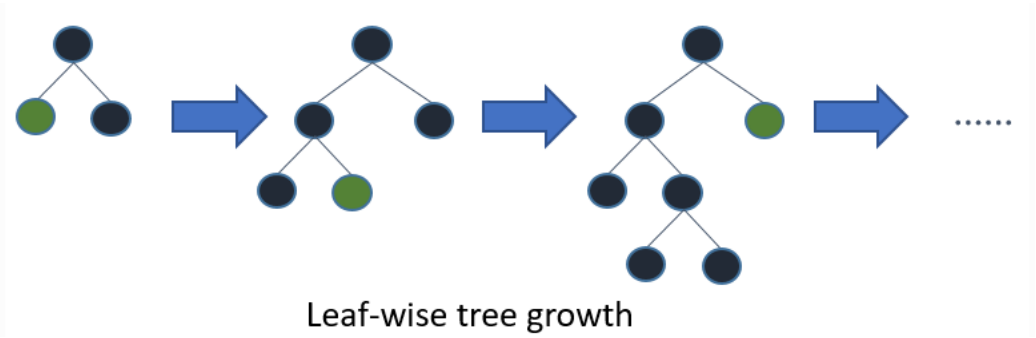


図9．Leaf-wiseのイメージ図[7]

4. 分析　-モデル学習・検証-

LightGBMモデルを採用した理由は以下の通りである.

採用理由	• Light GBMモデルは，特徴量のスケーリングを必要としないため，本研究のRFMC指標のようにスケールが異なるデータセットを入力することが可能である. • Kaggleなどのコンペで頻繁に使われており，精度の高い優秀なアルゴリズムとされている. • 似たアルゴリズムであるXgboostと比べ，決定木の葉の数を制限しているため，学習時間の短縮や過学習の軽減が見込める.

本研究では，分類モデルではなく**回帰モデル**を採用した.

研究目的でも述べたように離反可能性を推定することが目標であるため，0,1の分類問題として扱うより，回帰モデルを構築し，出力に任意の閾値を設け離反の線引きをする方が実用的なモデルであると考えた.

4. 分析　-モデル学習・検証-

本研究では次の４パターンのモデルを作成し，評価する.

- モデル１：RFMC指標を用いて，離反顧客を予測.
- モデル２：RFM指標を用いて，離反顧客を予測.
- モデル３：RFMC指標を用いて，離反顧客（店舗離反を除く）を予測.
- モデル４：RFM指標を用いて，離反顧客（店舗離反を除く）を予測.

入力データは図8のような構成となっている.

全購買日データのRFMC指標

全ヨーグルト購買日データのRFMC指標

目的変数離反タグ

店舗離反のタグ

⇒店舗離反を除く場合はこのタグを用いる

	member_info	R_s	F_s	M_s	C_s	R_y	F_y	M_y	C_y	churn_tag1	store_churn
0	0035b560aff87d452918817ee3352c020b0b4987f06137...	1	22	91002	0.086960	1	9	1607	0.108922	1	0
1	0035b560aff87d452918817ee3352c0213ef13f1680a30...	2	157	648707	0.068814	2	62	18865	0.080246	0	0
2	0035b560aff87d452918817ee3352c021f52236b7ed5b0...	3	127	253820	0.077511	106	7	1495	0.166804	1	0
3	0035b560aff87d452918817ee3352c0224da27e4040fb4...	1	87	207326	0.082593	12	8	1320	0.081767	0	0
4	0035b560aff87d452918817ee3352c022638f618c8c338...	4	66	138456	0.172285	153	4	466	0.264298	0	0
...	...	...	...	...	...	...	...	...	...	...	...
4817	ffbef574e816196cf6c7166c8782402eeafc820a7731f5...	26	5	45389	0.230783	26	4	1418	0.249191	1	0
4818	ffbef574e816196cf6c7166c8782402ef09decb55a2bc6...	5	14	21227	0.082304	5	4	400	0.189811	1	0
4819	ffbef574e816196cf6c7166c8782402ef9d2f368393558...	5	118	136817	0.071109	7	25	5157	0.106278	0	0
4820	ffbef574e816196cf6c7166c8782402efcbaec2b007702...	5	23	10552	0.114281	5	6	795	0.209222	1	0
4821	ffbef574e816196cf6c7166c8782402effb5db793dec9b...	23	20	97724	0.286364	40	15	2923	0.293968	0	0

図10．入力データの構成

5. 結果・考察

各モデルの予測精度を表2,3に示す.

表２．学習データの予測値評価

	RMSE	MAE	RMSLE
model1	0.318	0.253	0.223
model2	0.335	0.270	0.235
model3	0.321	0.253	0.224
model4	0.330	0.259	0.230

表３．検証データの予測値評価

	RMSE	MAE	RMSLE
model1	0.397	0.311	0.275
model2	0.399	0.318	0.277
model3	0.396	0.306	0.274
model4	0.398	0.310	0.276

- RFMC指標を説明変数としたmodel1,3と，RFM指標を説明変数としたmodel2,4を比べると，わずかにC指標が説明変数となっているmodel1,3の方が精度が良い結果となった.
- model3は，店舗離反の客を取り除きヨーグルト商品の離反を予測した．学習データにおける精度ではmodel1に劣るものの，検証データにおいては最も精度が良いことが示された.

⇒ この結果から，汎化性能を高める上では，店舗離反とカテゴリ離反は別にして予測すべきであると考えた.

5. 結果・考察

各モデルの変数重要度を図11～14に示す.

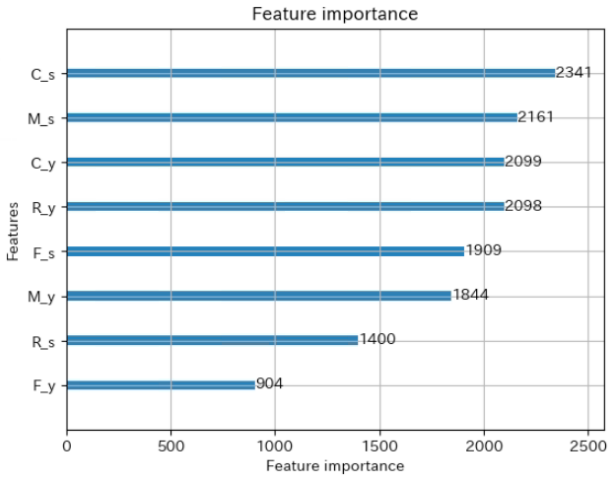


図11. Model1 変数重要度

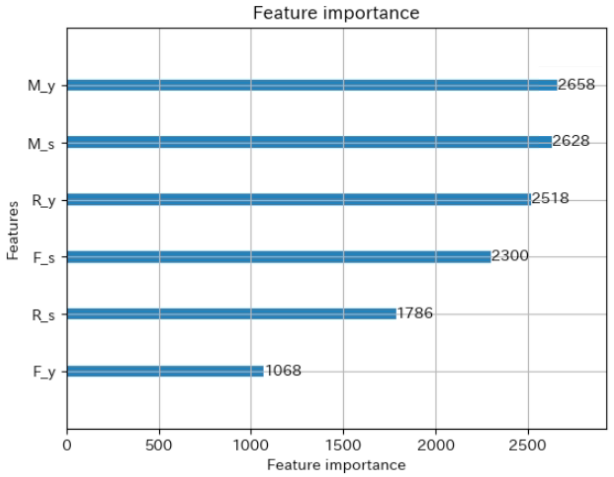


図12. Model2 変数重要度

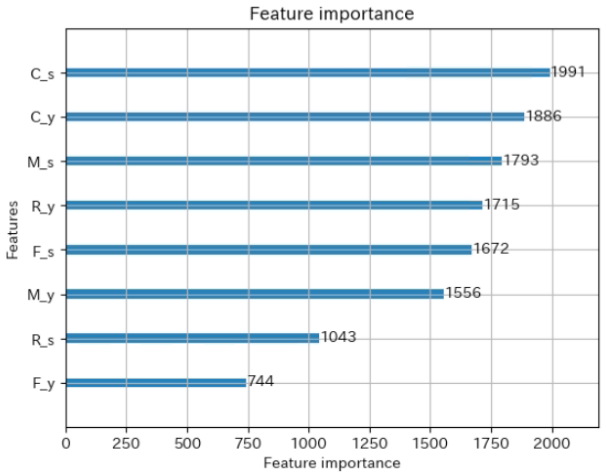


図13. Model3 変数重要度

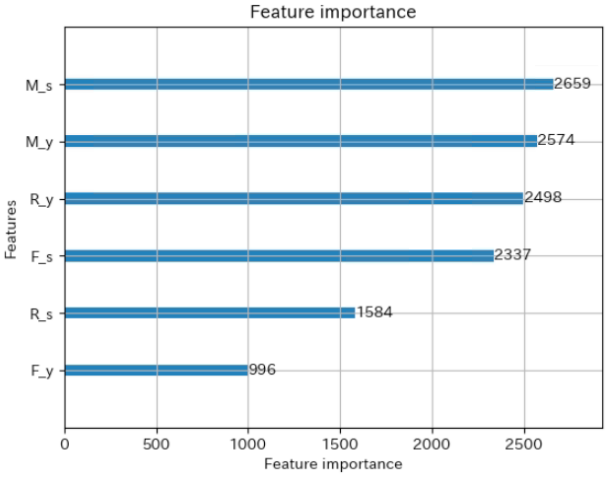


図14. Model4 変数重要度

- 図12,14を見ると，離反顧客が変わっても，RFM指標の変数重要度の順番は変化しないことが分かる.
- 図11,13を見ると，店舗離反を除いた離反顧客の予測の際には，ヨーグルト購買日データから算出したクランピネス指標(C_y)の順位が上がっていることが分かる.
- また，図11,13ともにクランピネス指標が変数重要度の上位を占めており，表2,3の結果からも，クランピネス指標が離反予測に寄与している可能性が高いと考えた.

6. まとめと今後の課題

まとめ

- 店舗離反とカテゴリ離反を別々に予測することで，ヨーグルト商品における離反予測モデルの汎化性能を高められる可能性が高い.
- モデルの比較や変数重要度から，クランピネス指標は離反予測に寄与している可能性が高い.
- ４つのモデルを比較したが，全てにおいて予測精度が低く，改善の余地が残る.

6. まとめと今後の課題

今後の課題

- **離反のアノテーションの定義の見直し**
  P14で述べたように，離反アノテーション期間において前の期間よりヨーグルト購買間隔が広くなっている場合についても，「機会損失の可能性」を考慮し，離反タグ0.5など重みを付けた離反タグを付与することを検討したい.
- **説明変数の追加**
  RFMC指標計算期間において，ヨーグルトを購入した日の間に何回ヨーグルトを買わない来店を行なっているか集計し，その最大値，最小値，平均値を説明変数に加えることは離反の予測精度向上につながるのではないかと考えた．.
- **モデルの出力値に閾値を設け，二値分類で評価**
  本研究ではRMSEなどの，回帰分析の一般的な評価指標のみでモデルの評価を行ったが，研究目的で述べた実用を視野に入れて，出力値に対して閾値を設け二値分類を行いPrecision，Recall，AUCなどで評価することも検討すべきであろう.
- **カテゴリ離反から商品ブランドの離反も予測可能なモデルへ**
  本研究では店舗側の視点で考え，カテゴリにおける離反予測モデルを考えたが，メーカー側の視点では商品ブランドの離反も把握したいと考えるであろう．今回提案したモデルに加えて，商品ブランドの離反予測モデルも検討したい.

参考文献

[1] 機能性表示食品　商品紹介　森永乳業株式会社
https://www.morinagamilk.co.jp/products/pickup_product/?id=5

[2] 雪印メグミルク　機能性表示食品　https://www.meg-snow.com/products/list.php?l=kinou

[3] 明治ヨーグルトライブラリー　乳酸菌研究最前線　乳酸菌による腸内環境と便秘の改善について
https://www.meiji.co.jp/yogurtlibrary/laboratory/report/lb81/01/

[4] Zhang Y, Bradloww ET, Small DS (2015) Predicting customer value using clumpiness : From RFM to RFMC. Marketing Science 34(2): 195208.

[5] 中山雄司 , “顧客関係管理研究の新動向　–来客/顧客間隔の不均一性を測るクランピネス指標–” ,甲南経営研究, Vol.57, No.2, pp161-181, 2016年12月

[6] LightGBM 徹底入門 – LightGBMの使い方や仕組み、XGBoostとの違いについて
https://www.codexa.net/lightgbm-beginner/

[7] LightGBM Features https://lightgbm.readthedocs.io/en/latest/Features.html

[8] 蓮本 恭輔, 雲居 玄道, 後藤 正幸, “非負値行列因子分解を用いたプラットフォームビジネスにおける顧客障害価値予測”, 情報処理学会論文誌, Vol.60, No.7, pp1283-1293, 2019年7月