# チーム「五反田鉱業」における データ解析コンペティションへの挑戦の軌跡

紺谷　幸弘

## 1. 序

弊社，株式会社ブレインパッドは，2004 年に設立されたデータ分析を主な業務の一つとしているベンチャー企業である．今回，平成 24 年度データ解析コンペティション，一般フリー部門において，社内有志による参加メンバーを募り，参加当時オフィスの所在地であった五反田にちなみ，チーム「五反田鉱業」として参加をさせていただいた[1]．結果，光栄にも優秀賞を頂くことができたが，設定したテーマそのものについてはもちろん，特に分析の進め方においては，業務外での作業であるゆえの作業効率最大化に向けての大きな挑戦であったように思う．

本稿では，特に以下の二点に論の主眼を置き，チームとしての取り組みの内容を報告させていただきたい．

- 分析テーマ設定のプロセス
- チームにおける人的資源の最大活用を目指した作業プロセス

上記以外の詳細についても付録に記載している．必要に応じて都度参照されたい．

弊社における実際の分析業務では，分析の各ステップで分析を依頼した顧客クライアントとのコミュニケーションが発生し，その結果として，大小さまざまな方向修正が生じる．したがって，それをも見越したより柔軟的でかつ作業効率の良いプロジェクト遂行が常に求められており，本稿の進め方にはさらなる改善が必要だと考えている．弊社における分析業務の一端が，読者の方々と共有できれば望外の喜びである．

## 2. 分析テーマ設定のプロセス

### 2.1 コンペティション分析課題

分析用として（株）リクルート・テクノロジーズから同グループの共同購入型クーポンサイト「ポンパレ[2]」の加工された一部のトランザクションデータおよびクーポン情報，サイト訪問者（以降「カスタマー」と呼ぶ）のデータが提供された．データの概要は付録に記載しているので参照されたい．

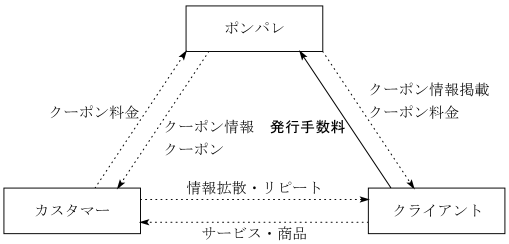


**図 1**　ポンパレビジネスの利害関係者と価値の流れ

課題フリー一般部門ではこのデータを利用して，学術的な新規性や結果の信頼性，適用可能性に加え，ビジネス上のアクションにつながる「実用的なメッセージ」を抽出することが求められた．

### 2.2 大枠となる分析テーマの設定

実際の分析業務と同様，取り組む分析テーマの洗い出しから始めた．本質的な課題・テーマを探索するために，分析テーマを洗い出す段階では，データの詳細な分析を敢えて避け，自由な発想で議論することにした．

われわれはまず，ポンパレの利益の向上に貢献できるようなテーマを探索するために，ポンパレビジネスの利害関係者とそれらの間の価値の流れについて確認した（図 1）．この図から，ポンパレに直接流入している価値は発行手数料であることがわかる．

発行手数料の総量
＝一枚当たりの発行手数料 × 購入枚数

上の式に従えば，ポンパレの売り上げを向上させる方法として，「一枚当たりの発行手数料」か「購入枚数」，

こんや　ゆきひろ
株式会社ブレインパッド
〒 108–0071　東京都港区白金台 3–2–10 白金台ビル

[1] メンバーは以下の 11 名に著者を加えた合計 12 名である；矢島安敏，丹沢良太，齋藤宗香，秋田谷孝俊，池田裕章，北條健太，松田慎太郎，橘信幸，Toan Nguyen, 中村崇，駒井隼人

[2] http://ponpare.jp/

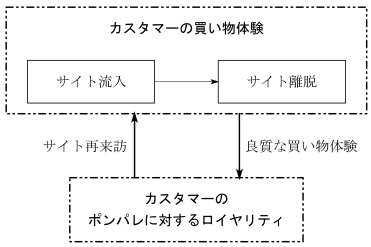


**図 2**　カスタマーの買い物体験とポンパレに対するロイヤリティ

あるいはその両方を向上させることが考えられる．同時に考えるべき制約条件は，特定の利害関係者が不利益を被らないことと，ポンパレビジネスの市場の健全性が失われないことである．

上記の制約条件や施策への落し込みやすさから，「購入枚数の向上」に焦点を当て，分析テーマを洗い出すことにした．さまざまなアイデアがメンバーにより出されたが，結果としてレコメンデーションにより図 2 のように，「カスタマーの買い物体験」の質を向上させることで，カスタマーのポンパレサイトへのロイヤリティの強化を狙い，その結果として購入数が増加するのが良いのではないか，という方向でまとまった．

### 2.3　分析テーマの絞り込み

次にわれわれは，レコメンデーションによる訴求対象となる，「カスタマーにとっての良質な買い物体験とは何か」について検討した．

欲しいものを購買することだけが買い物の楽しみ方のすべてではない．特に目的を設けないで商品を見て回るウィンドウショッピングも楽しみ方の一つである．「見ているだけで楽しい」といった体験は，カスタマーを買い物に向かわせる原動力となり得る．このようなカスタマーにとって，ポンパレのサイトがウィンドウショッピングを楽しむ場になれば，サイト来訪が促進されて，定期的な購買に結び付くことも十分期待される．

すなわち，カスタマーにとって買い物に求める体験は単一ではなく，彼らに広く「良質な買い物体験」を提供するには，カスタマーそれぞれが求める買い物体験に応えるようなレコメンドを行う必要がある．われわれはカスタマーが買い物に求める要素として「**購買**」と「**回遊**」の二つを取り上げることにした．

**回遊**とは，ウィンドウショッピングのような，購買というよりむしろ商品の閲覧に主眼があるような要素である．求めるものが異なれば，レコメンドの訴求対象となる要素も異なるはずで，**購買**では**購買確率が高いクーポン**がレコメンド対象になるのに対し，**回遊**ではその後の**閲覧回数を増加させるようなクーポン**がレコメンド対象になる．また，例えば，ウィンドウショッピングを楽しみにきたカスタマーであっても，商品閲覧を通して「購買」に気分が傾いていくというのは十分考えられることである．われわれはそうしたカスタマーの経時的な欲求の変化を考慮することで，「買い物体験」を充実させることができるのではと考えた．

このようなメンバーによる議論の結果，「良質な買い物体験」を提供するアプローチとして，以下を実現するレコメンデーションロジックの構築を試みることでまとまった．「刻一刻と変わるカスタマーの気分を捉え，クーポン配信を通じてカスタマーの気分を高揚させる」．ここで言う「気分」とは，カスタマーの買い物に対する欲求が「購買」寄りであるか，「回遊」寄りであるか，ということである．ビジネスの新規性といった観点からも，「回遊」要素を同時に考慮したレコメンデーションという点は十分期待できると考えた．

### 2.4　基礎集計による分析テーマの価値の確認

次にわれわれは，「回遊」という要素が現状のポンパレビジネスにおいてどの程度大きな影響を持っているかを基礎集計により確認した．「回遊要素も考慮したレコメンド」という着想にどれほど新規性があっても，期待できるインパクトが小さければ別の分析テーマに取り組むことも検討せねばならないからである．

具体的には，以下の二つの基礎集計から回遊の影響の大きさを検討した．

(a) 1 セッションあたりの PV (Page View) 数と全体売上の関係
(b) 1 セッションあたりの最大 PV 数と来訪回数の関係

基礎集計 (a) の結果，売り上げの 1/4 が 6PV 以上のセッションにより構成されていること，(b) の結果最大 PV 数に比例し，来訪回数が増えていることが読み取れた．(b) の結果から，閲覧数の向上がポンパレに対するロイヤリティの醸成に貢献している可能性が示唆される．以上二つの基礎集計の結果から，われわれは「カスタマーの買い物体験の向上を考えるうえで，購買だけでなく回遊も重要な要素である」と結論づけ[3]，これを実現するレコメンドロジックの具体案として次のようなものを考えた．

(1) 時点 $t$ までの情報から，購買，回遊のいずれの気分が相対的に優位かを判断（ここで優位と判断さ

[3] 上記推測については，本来因果の方向を含めた検討が必要であるが，その切り分けが困難であると思われたため，上記の検証にとどめている．

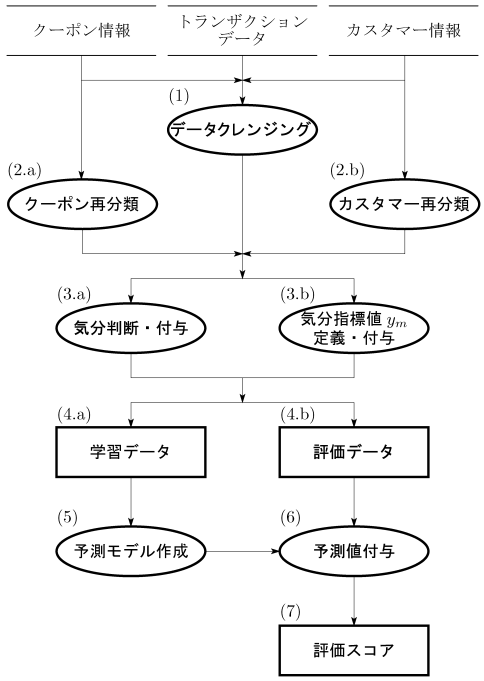


**図 3** 報告対象のレコメンドロジックのデータフローダイアグラム (DFD)

れた気分 (mood) を $m$ とする[4])

(2) (1) において優位と判断された気分 $m$ について，時点 $t$ で配信可能なすべてのクーポンに対して，時点 $t+1$ における気分 $m$ の指標となる値 $y_m$ を予測（以降この $y_m$ を気分指標値と呼ぶことにする）

(3) 予測スコアの降順で配信クーポンを決定

上記の手続きにそのほか必要と思われる処理を加え，データフローダイアグラム (DFD) の形でレコメンドロジックの内容を表したのが図 3 である．太字で示した箇所が具体的な定義が必要になる個所である．また実際には各処理の結果を逐次，データテーブルとして出力しているが表記を簡単にするため割愛している．

## 3. チームにおける人的資源の最大活用を目指した作業プロセス

### 3.1 二種類の作業プロセス

DFD（図 3）が構築されたばかりの段階では，各処理の内容は抽象的な処理内容にとどまっており，具体的な定義は与えられていない．例えば，気分指標値 $y_m$ が具体的に何であるか，などは全く定義されておらず，レコメンドの基準となる指標となる何か，である．

[4] 優位性の高い気分 $m$ を確率変数とするか否かは検討対象とした．

したがって，このレコメンドロジックを実現するためには各処理の内容を具体的に定義し，プログラムに落とし込む必要がある．

加えて，限られた時間の中でレコメンドロジックの質としてもある程度の水準を満たさなければならない．

この目的を達成するためには，チームの人的資源を最大限に活かす必要があることは容易に想像いただけるかと思う．

典型的な作業プロセスの一例として，一つ一つの処理を上流から逐次的に定義・実装していく**直列型プロセス**が考えられる．直列型プロセスの理想的な進め方の一つとして，以下のようなものが考えられるであろう．

**• 直列型プロセス •**

**(S-1)** メンバー全員で一つ，あるいは同列の処理を逐次的に定義・実装

**(S-2)** データフローの最終出力である評価スコアを算出

**(S-3)** レコメンド性能のボトルネックとなっている個所を選定

**(S-4)** **(S-3)** の箇所に対し精緻化を試み，結果をロジック全体に反映

**(S-5)** **(S-2)** – **(S-4)** のプロセスを繰り返し，レコメンド性能の向上を図る

(S-1) における「同列」とは，DFD（図 3）において同じ数値が付与されている（複数の）処理を指す．(S-1) から明らかなように，当該処理の定義・実装は，以前の処理の出力の形式に依存するため，当該処理以降の処理においてどのような情報が利用可能か，あるいは制約条件が何になるかは不明である．例えば，DFD（図 3）の処理 (2.a)「クーポン再分類」の定義・実装がなされている段階では，処理 (5)「予測モデル作成」において，どのようなデータが利用可能か，あるいは考慮すべき条件がどんなものかを明確に示すことはできない．

この直列型プロセスには次のような課題がある．

(a) 完成までにかなりの時間を要する

(b) 以下の理由により，処理単体の精緻化が困難

(b-i) すべての処理がそれ以前の処理の定義に依存する

(b-ii) 処理全体の性能評価指標が唯一存在し，個々の処理単体の性能評価指標が存在しない

また，実際に，(S-1) のように少数の処理に多くの人的資源を投入する場合，少数の限定的なメンバーに作業負荷が集中し，チームの人的資源を効率的に活かせているとは言いがたい状況になりやすい．

そこでわれわれは以下のような**並列型プロセス**による進め方を試みることにした；

● 並列型プロセス ●

**(P-1)** 各処理の入出力を予め厳密に定義

**(P-2)** メンバーを複数のチームに分割しチーム毎に担当処理を決定，並行して担当処理の「簡易版」を短期間で作成

**(P-3)** レコメンド性能のボトルネックとなっている個所を選定

**(P-4)** **(P-3)** で選定された個所について全メンバーにより精緻化を試みる

**(P-5)** **(P-3)**，**(P-4)** を繰り返す

(P-2) の**簡易版**とは，「各処理を最短期間で実現可能な内容・手順により実装したレコメンドロジック」である．

この並列型プロセスでは，直列型プロセスの課題に対し，それぞれ以下のように対応している；

**直列型プロセスの課題 (a)：時間的な問題**

直列型プロセスの課題 (a) の要因の一つは，すべての処理に対し，同時に人的資源を割けないことである．この「全処理への人的資源の同時配置」を困難にしているのは，各処理の入力が以前の処理の出力の累積によって規定されることにより生じる，入力情報の不明瞭性である．並列型プロセスでは，この問題に対し (P-1) により解決を試みている．各処理において入力の情報が明確に決定されることによって，処理の内容がイメージしやすくなり，個々の処理の性能を不問とすれば，各処理の定義および実装の同時進行が可能になる．

**直列型プロセスの課題 (b)：個々の処理単体の精緻化の問題**

並列型プロセスにおいても，この課題を根本的に解決するのは困難である．そこで，簡易版を基本とした精緻化のプロセス (P-5) を短期間で何度も循環させることで間接的な解決を試みた．

### 3.2 並列型プロセスの具体的内容

作業当初想定した，並列型プロセス (P-1)–(P-5) の具体的手順は以下のとおりである；

(p-i) 簡易版の完成期日を確定（原則厳守）

(p-ii) 各処理の入力と出力の形式（変数名および型）をメンバー内の協議により確定（原則として，以降この変数名と型の変更は行わない）

(p-iii) メンバーを複数のチームに分割，担当する処理を決定

(p-iv) 各チームで作業担当となった処理の具体的な内容や手順の候補について議論

(p-v) (p-iv) の内容について，メンバー全員による定例ミーティング内で共有およびレビュー，結果を踏まえ各チーム内で再度検討

(p-vi) 各処理において挙がった処理内容・手順のうち，工数の観点から最も容易に実装できるものを採用

(p-vii) (p-vi) の進捗や作業中に挙がった課題について，メンバー全員による定例ミーティングやメール，クラウドサービス上で適宜相談，結果を反映

(p-viii) 処理間の整合性を確認，レコメンド性能評価スコアを算出

開始当初，並列型プロセスによる時間削減の要は各処理の入出力を厳密に定義することであると思われた．しかしその一方で，入出力の厳密な定義を行うことは，レコメンドに利用する情報を限定することにもなり，レコメンドロジック全体の限界を規定してしまう側面もある．したがって，上記手続き (p-ii) 実施の際，予測モデルに投入する可能性のある説明変数の洗い出しに関しては，参加メンバー内である程度の時間をかけて綿密に行った．こうした説明変数の洗い出しを行う一方で，実際の作業時に得られるアイデアを予め網羅することには無理があると思われたため，拡張可能性を考慮し予備のカラムを複数個用意しておくことにした．

### 3.3 作業の実情と振り返り

当初は上記の (p-i)–(p-viii) に従い進めることを想定したが，実際には簡易版作成の段階で，各処理内で分析に差し支えるレコードの存在が明らかになるなどの理由から，データクレンジングの処理に変更が加わったり，メンバー間の議論を通じさまざまなアイデアが生まれ，(p-ii) の定義に変更が加わることもあった．そういった事態にも見舞われたが，簡易版は当初の想定どおりに完成した．簡易版の具体的内容や，簡易版完成以降の精緻化のプロセス，最終版の具体的な処理内容については付録を参照されたい．

実際には並列型プロセスのみを行っており，直列型プロセスとの比較はできないが，結果としてこのプロセスは機能した．繰り返しになるが，分析作業開始当初は「並列型プロセスの要は各処理の入出力を厳密に定義すること」と思われた．しかし，実際の作業を振り返ってみると，基本となる土台（簡易版）を作成することにより，全体像の見通しが良くなったことが，効用としては大きかったように思う．その一方で，「全体像の見通しが良くなること」の効用だけでは説明として十分でないこともあった．具体的には，上記のように各処理の定義内容が頻繁に起こるような状況であったにもかかわらず，各チームの作業がチーム間で発散

  

せず収束に向かったことである．これは，メンバー間の密な連携が維持された賜物であり，それを可能にしたのはメンバー各人の動機が明確，かつ一致していたことであることは疑念の余地がない．

## 4. まとめ

本稿では，われわれの作業の進め方について紹介した．本文中や付録に記載した内容は，われわれが行った議論のうち，最終報告時まで残ったものが主になっており，実際にはさらに多くの着想や議題があった．そのすべてをここに記すことは紙面の制約から困難である．

実作業においては上でも触れたように試行錯誤の連続であり，定義の変更による手戻りなどは日常茶飯事であった．ようやく報告に漕ぎ着けたというのが事実である．

また，賢明な読者の方々はすでにおわかりかと思うが，今回報告したレコメンドロジックには依然として改善点が山積している．しかしながら，レコメンドロジックと作業プロセスの両面において挑戦的な試みであったこと，また，すべてのメンバーが通常業務を抱えていたことを鑑みれば，成功と言える最低限の水準は満たしているように思う．分析テーマの洗い出しからそのほかすべての議論，処理の実装に至るまで，まぎれもなくメンバー全員によるチームとしての働きが不可欠であったし，結果，賞が頂けたことはそれが結実した証左であると強く感じている．

## 5. 付録

### 5.1 提供されたデータ概要

（株）リクルート・テクノロジーズより提供されたデータの概要は以下のとおりである．

**トランザクションデータ**

カスタマーごとのセッション情報，すなわち各カスタマーが「ポンパレ」に接続し，離脱するまでにどんなクーポンを閲覧し，購買したかなどの情報を含む．レコード（オブザベーション）の最小単位はある時点におけるクーポン閲覧および購買有無情報（以降 PV）．クリックされたクーポンのログのみから成り，同時に配信されたクーポンのログは含まない．

**クーポン情報**

クーポンのタイトルやページの内容，価格（定価や割引率，割引後の価格），掲載期間，ジャンル，中ジャンル，売り切れの有無，クーポンの発行者（クライアント）の実店舗の住所などが含まれる．

**カスタマー情報**

カスタマーの登録日や性別，年齢や退会日，居住都道府県などが含まれる．

### 5.2 レコメンドロジックの前提

レコメンドロジックの前提として以下の三つの仮定を設定した．

- カスタマーの気分には購買気分と回遊気分のみが存在
- 購買気分の回遊気分に対する相対的な優位性は判別可能
- 購買気分の回遊気分に対する相対的な優位性は逐次変化する

### 5.3 DFD （図 3）の実現に係る作業内容

各処理の実現に係る具体的な作業内容を以下に示す；

**(1) データクレンジング**

分析に利用しやすいようにデータの選別を行う．レコメンド全体でどのようなものを定義するかによって，選別対象となるデータは異なる．レコメンドの各処理を定義するうえで，扱いが困難なレコードの存在が明らかになった場合，すべての処理において対応結果が反映される必要がある．

**(2.a) (2.b) クーポン・カスタマーの再分類**

特定のカテゴリに大きな偏りがないような再分類の方法を定義する．また実際に再分類を行う．

**(3.a) 気分判定・付与**

時点 $t$ において相対的に優位な気分 $m$ の判断手続きの具体的な処理内容を定義する．また，入力データに対して，上記処理内容に基づく判断結果をデータに付与するプログラムを作成する．

**(3.b) 気分指標値 $y_m$ 定義・付与**

各気分 $m$ の指標値 $y_m$ の具体的内容を定義する．また，入力データに対して，具体的内容に基づく指標値をデータに付与するプログラムを作成する[5]．

**(4.a) (4.b) 学習データ・評価データの作成**

学習データと評価データの切り分けおよびこれを実現するプログラムを作成する．

**(5) (6) 予測モデル作成・予測値の付与**

時点 $t$ において，$t+1$ の気分指標値 $y_m$ を目的変数とする予測モデルを作成する．また，評価データを入力として予測スコアを出力するプログラムを作成する．

[5] 実際には，上記の「気分判断」の処理内容によって，例えば $m$ =purchase となるレコードは変わるため，すべてのレコードについて $y_{\text{purchase}}$, $y_{\text{browse}}$ の両方を付与した．

**(7) 評価スコア算出**

評価スコアの定義，比較用モデルの設定，評価データ作成プログラムおよび評価スコア算出プログラムを作成する．

### 5.4 簡易版の定義内容

簡易版において実際に採用されたDFD（図3）の処理内容を以下に示す;

**(1) データクレンジング（簡易版）**

分析テーマ検討時の基礎集計の結果[6]から，6PV以上のセッションのみを分析対象に選定した．また，以下のようなデータについて稀な例であることを確認後，データから除外した．

- 複数のカスタマーを含むセッション
- 全く同時刻に複数の異なるクーポンの購買を含むセッション

**(2.a) クーポンの再分類（簡易版）**

クーポンのタイトルおよび説明文に対し，形態素解析を行い提供データに存在する中ジャンルに関連すると思われる単語を抽出，人手により再分類を実施．具体例としては，中ジャンル「ヘアサロン」に属するクーポンを「カット」「カラー」などのカテゴリに再分類した．

**(2.b) カスタマーの再分類（簡易版）**

性別，年代（10歳区切り），居住都道府県の組み合わせで再分類した．

**(3.a) 気分判定・付与（簡易版）**

当該PVを含む過去直近5PVの閲覧クーポンに大きな偏りが見られれば，購買前の細かい比較に入っている（$m$ =purchase），大きな偏りが見られなければ，さまざまなクーポンを見て回っている（$m$ =browse）と考え，以下の手続きにより気分を判断した；

$$m = \begin{cases} \text{purchase} & \text{if} \quad \frac{\max_{j=1,\ldots,J}\{\#j\}}{5} > 0.5 \\ \text{browse} & \text{otherwise} \end{cases}$$

ここに，$\{\#j\}$ は直近5PVにおける再分類されたクーポンカテゴリ $j$ の出現回数を表す．

**(3.b) 気分指標値 $y_m$ 定義・付与（簡易版）**

時点 $t$ における購買の気分指標値 $y_{\text{purchase}}$ を，当該クーポンが購買されていれば1，購買されていなければ0と定義，時点 $t$ における回遊の気分指標値 $y_{\text{browse}}$ を，時点 $t+1$ でセッションの離脱が起こっていなければ（当該PV以降少なくとも1PV以上存在すれば）1，離脱が起こっていれば0と定義した．

**(4.a) (4.b) 学習データ・評価データの作成（簡易版）**

以下の基準で評価データおよび未使用データ，学習データを設定した．

- 評価データ（予測対象データ）
  評価対象日のPVを含むセッションの第11PV以降のすべてのPV
- 未使用データ
  評価対象日のPVを含むセッションの第10PV以前のすべてのPV，および評価データに含まれないもののうちで評価対象日以降のすべてのPV
- 学習データ
  上記以外のすべてのデータ

ただし，評価データの各PVには，その時点で配信され得たすべてのクーポンのデータが追加されている．気分指標値 $y_m$ の予測はこれらすべてのクーポンに対して行われ，レコメンドされるクーポンは予測値 $\hat{y}_m$ の降順になる．

**(5) 予測モデル作成（簡易版）**

利用可能なすべての説明変数を投入したロジスティック回帰モデルを作成した．

**(7) 評価スコア算出（簡易版）**

レコメンドの性能を精度および多様性の二面から評価することにした；

- 精度：predictionおよびrecallより算出されるF値
- 多様性：レコメンドされるクーポンのエントロピー

気分指標値 $y_m$ の予測精度について，実際には，投入される説明変数の粒度により，異なるクーポンであっても全く同じ予測値を返すことが起きるため，各PVについて「再分類したクーポンカテゴリ」と「クーポン提供者の店舗地域」の組み合わせの単位で一つのクーポンを選定することを想定した[7]．こうしてレコメンドされたクーポンのうち，上位11件に対して評価を行うことにした．

多様性は評価データの全PVのレコメンドされるクーポンから算出している．

[6] 6PV以上のセッションにより売上の1/4を占める．

[7] 「再分類したクーポンカテゴリ」「クーポン提供者の店舗地域」の組み合わせで複数の異なる予測値 $\hat{y}_f$ が存在する場合，その中で最大の予測値を持つものを採用することにしている．

比較用モデルには協調フィルタリングを採用することにした．

### 5.5 DFD（図 3）の精緻化のプロセス

購買と気分判断に利用する量

$$\frac{\max_{j=1,\ldots,J}\{\#j\}}{5}$$

の関連について，基礎集計を実施した．詳細な数値は紙面の都合上割愛するが，購買 PV を含むセッションでは，購買気分の回遊気分に対する相対的な優位性が高いほど，購買率が低くなるのに対し，購買 PV を含まないセッションではその逆，すなわち回遊気分の相対的な優位性が高いほど，購買確率は下がっていることが読み取れた[8]．以上の結果から，カスタマーおよびクーポンの再分類および気分判断までのプロセスはおおむね問題がないと判断し，予測モデルの精緻化に注力することにした．

**(3.b) 気分指標値 $y_m$ 定義・付与（最終版）**

回遊の気分指標値 $y_{\text{browse}}$ について，簡易版の定義では，評価データのサイズが小さすぎることが懸念されたため，時点 $t$ における回遊の気分指標値 $y_{\text{browse}}$ を，時点 $t+3$ でセッションの離脱が起こっていなければ（当該 PV 以降少なくとも 3PV 以上存在すれば）1，離脱が起こっていれば 0 と定義した．

**(4.a) (4.b) 学習データ・評価データの作成（最終版）**

評価データとして十分なサイズを用意するために，以下のように変更した．

- 評価データ（予測対象データ）
  評価対象日の PV を含むセッションの第 6PV 以降のすべての PV
- 未使用データ
  評価対象日の PV を含むセッションの第 5PV 以前のすべての PV，および評価データに含まれないものの内で評価対象日以降のすべての PV
- 学習データ
  上記以外のすべてのデータ

**(5) 予測モデル作成（最終版）**

以下のような手法について検討した．

- L1 正則化を行ったロジスティック回帰モデル
- L2 正則化を行ったロジスティック回帰モデル
- 判別木 (CART)

正則化パラメータはいずれもクロスバリデーションによるパラメータ選択を行っている．上記のモデルについて，モデルの作成単位（カスタマーの性別，居住地域）のバリエーションや，投入する説明変数の選択・追加を検討した．最終的には，購買気分指標値 $y_{\text{purchase}}$ 予測用モデルとして L2 正則化を行ったロジスティック回帰モデルを，回遊気分指標値 $y_{\text{browse}}$ 予測用モデルとして「カスタマーの性別」および「カスタマーの居住都道府県」の組み合わせを単位とするロジスティック回帰モデルを採用した．

**(7) 評価スコア算出（最終版）**

レコメンドの単位として，実際に配信される状況を考慮し，「クーポン提供クライアントの店舗地域」を追加し，「再分類されたクーポンカテゴリ」と「クーポン提供クライアントの店舗地域」の組み合わせから一つのクーポンをレコメンドすることを想定することにした．

実際には，予測モデルの精緻化以外の処理は早い段階で作業を終え，以降モデルの評価スコアの向上に注力している．また，これらの作業と並行して発表内容の精査を行った．

[8] ここでは購買の有無のみに着目しているが，本来は回遊についても同様の基礎調査を行うべきである．今回は時間的な制約の観点から，ここの箇所の調整および調査はここで打ち切りにした．