再購入者予測に関するデータ分析と特徴量についての考察

Data Analysis and Investigation of Features for Re-purchase Prediction

董子華 Zihua Dong　柴田祐樹 Hiroki Shibata　高間康史 Yasufumi Takama

東京都立大学
Tokyo Metropolitan University

Abstract: This paper reports on the results of analyzing a dataset for repurchase prediction competition. This paper sets two research questions: identifying important time periods during which actions taken contribute to the prediction and effective features in terms of acquisition costs and privacy protection. Experimental results show that actions taken within one month from Singles' day is important for the prediction. It is also confirmed that the prediction accuracy do not decrease without product-level features.

1 はじめに

本稿では，ユーザが同じ販売者から再度購入するかを予測するタスク（再購入者予測タスク）を対象としたデータセットを分析した結果について報告する．

近年，Amazon[1]や天猫（Tmall）[2]に代表される電子商取引（E コマース）プラットフォームが発展し，多数の販売者がオンライン・チャネルを通じて商品を販売したり，サービスを提供したりするようになった．販売者にとって，ユーザのロイヤリティを推定することで，ターゲットを絞った販売促進活動を行うことが重要になっており，機械学習の適用が研究されている．関連するタスクの一つである再購入者予測は，ユーザが販売者から商品を購入した後に，同じ商品を再び購入したり，再びこの販売者から購入したりするかどうかを予測するタスクである [1,2]．このタスクに関して，Tmall のデータに基づくコンペティションが 2015 年に開催されている[3]．このコンペティションでは，匿名ユーザの 11 月 11 日（独身の日）に開催される大規模セールでの購入者が，6 か月以内に同じ販売者から再度購入するかを予測する．11 月 11 日当日から過去 6ヶ月のログデータが公開されており，分類モデルの学習に用いることができる．優勝者を含む数人の研究者による論文が発表されているが，6 か月ものデータが予測に必要かについては検討されていない．また，公開されているデータは匿名データではあるものの，各ユーザがアクセスした商品 ID までが記録されおり，かなり詳細なデータといえる．このうち，予測のためにはユーザに関するデータがどの程度詳細に必要なのかについても議論されていない．

本稿では，再購入者予測タスクに用いる特徴に関する知見を得ることを目的として，Tmall のデータセットを分析する．具体的には，以下の 2 つのリサーチクエスチョンを設定する．

- **RQ1**：どの時期の行動が，11/11 当日の行動に影響しているか
- **RQ2**：取得コストやプライバシー保護の観点から，再購入予測に有効な特徴は何か

RQ1 については，特徴量の計算に用いるデータの収集期間を変更して複数の分類モデルを構築し，その予測精度を比較する．実験結果より，11/11 に近い時期（1 か月以内）の情報は重要であること，それ以前の情報は性能にあまり影響しないことを示す．

RQ2 については，予測対象販売者に最後にアクセスしてからの経過時間など，商品レベルのデータを必要としない特徴を用いてモデルを構築し，予測精度を比較する．実験結果より，商品レベルの特徴を利用しなくても予測精度は低下しないことを示す．

2 関連研究

2.1 再購入者予測の研究

再購入者予測に関する研究は，多くの領域で注目されているが，それらはタスクによって分類できる．再購入者予測タスクでは，一定期間内に対象の販売者から購入したユーザが，将来の一定期間内にその販売者から再購入するか否かを予測することを目的とする．同

連絡先：高間 康史，東京都立大学システムデザイン研究科，〒191-0065 東京都日野市旭が丘 6-6，ytakama@tmu.ac.jp

[1]https://www.amazon.co.jp/
[2]https://www.tmall.com/
[3]https://tianchi.aliyun.com/competition/

じ商品を再度購入するかを予測するか，あるいは異なる商品を購入する場合も対象とするかは研究によって異なる．Liu らは E コマースを対象として，同じ販売者から将来再度購入するかを予測する手法を提案しており，同じ商品でない場合も対象としている [1]．Zhang らも E コマースを対象としているが，同じ商品を再度購入するかを予測対象としている [2]．

2.2 再購入者予測手法

機械学習を用いた再購入者予測に関する研究は，特徴の検討，予測モデルの改良の 2 つのアプローチに大別できる．特徴の検討に関して，Zhang らは本稿と同じ Tmall のデータセットを対象として，ユーザ，販売者，ユーザと販売者のインタラクションに関する特徴 147 種類の特徴を定義し，ランダムフォレストを用いた埋め込み法や ANOVA を用いたフィルター法などの特徴選択を適用し，予測に有効な特徴について考察している [3]．分析の結果，有効な特徴上位 10 種類のうち，6 種類はユーザの特徴であること，販売者やユーザと販売者の組み合わせに関する特徴の中では，購入クリック率と重複購入率に関する特徴が重要であることを報告している．Liu ら [4] も Zhang ら [3] と同様に，ユーザの特徴，アイテムの特徴やインタラクションに関する特徴について検討している．各販売者，ブランド，カテゴリごとにユーザが購入した日数などの平均値，分散，最大値を計算し，集約的特徴として定義している．これらの統計量に関する特徴だけでなく，各ユーザを，購入先販売者を単語とする文書と見なし，主成分分析（PCA）と潜在ディリクレ配分（LDA）を適用することで，上位 10 個の主成分座標と 40 個のトピックも特徴として定義している．また，販売者を文書，購入ユーザを単語とみなして同様の特徴抽出も行っている．合計 1364 種類の特徴を定義し，XG-Boost を用いた埋め込み法による特徴選択の結果，有効な特徴上位 20 種類のうち，10 種類はユーザの特徴であること，その中でも，販売者あたりの平均購入回数と平均クリック数が重要であること，販売者に関する特徴の中では，ユーザあたりの平均購入日数とその標準偏差に関する特徴が重要であることを報告している．

予測モデルの改良に関して，Zhu らは CNN(convolutional neural network) と LSTM(Long-Short Term Memory neural network) を組み合わせた予測手法を提案している [5]．CNN は短期間におけるクリックや購入の頻度などのユーザの行動に見られる局所的な特徴を抽出するために用いられ，LSTM はある商品の購入頻度が時間とともに増加するパターンなどといった，行動傾向の長期的な変化をモデル化するために用いられている．CNN と LSTM を単体で用いた場合と比較した結果，予測精度が向上することを報告している．Yang らはランダムフォレストと LightGBM（Light Gradient Boosting Machine）の予測結果を Soft voting により結合する手法を提案している [6]．過学習に強いランダムフォレストの特徴と，予測精度の高い LightGBM の特徴を組み合わせることで，それぞれ単体で用いた場合よりも予測精度が高くなることを報告している．

これらの研究では，予測精度を高めることが目的であり，データ収集におけるユーザのプライバシーに関する問題やデータ収集コストは議論されていない．

3 分析目的・方法

3.1 分析目的

本稿では，再購入予測に利用する特徴に関する知見を得ることを目的として，以下の 2 点をリサーチクエスチョンとする．

- **RQ1**：どの時期の行動が，11/11 当日の行動に影響しているか
- **RQ2**：取得コストやプライバシー保護の観点から，再購入予測に有効な特徴は何か

RQ1 について調査するために，本稿ではベースラインとなる分類モデルを定め，各特徴量の計算に用いるデータの収集期間を変更するなどして構築した各モデルの性能を比較する．ベースラインは，ユーザ，販売者，およびユーザと販売者のインタラクションに関して，基本的な統計量を計算し，特徴として用いる．また，特徴選択を適用して用いる特徴数を削減する．ベースラインの特徴については，4.1 節で詳細に説明する．

RQ1 は，以下の 2 種類の実験により調査する．

- **実験 1.1**：ベースラインの特徴のうち，データ収集期間を変更して計算可能なものについて，11/11 当日からどこまで遡ってデータを収集するとよいかについて調査する．
- **実験 1.2**：実験 1.1 と同じ特徴について，どの月のデータが予測に影響を与えているかについて調査する．具体的には，各月のデータを除いて特徴量を計算し，分類モデルを構築して性能を比較する．

RQ2 については，ユーザに関する詳細な情報を用いたほうが予測精度が高くなることが予想されるが，プライバシーの問題が懸念されたり，データ収集コストが高くなるなどの問題が発生する．そこで，RQ1 と同じベースラインから商品レベルの特徴を除去して分類モデルを構築し，除去前と性能を比較する．販売者レ

ベルの情報だけを利用することで，情報取得コストやプライバシー保護の点で利点があると考える．

3.2 評価指標

実験では，予測･分類タスクで一般的に使用される3つのモデル：LR，MLP，XG-Boostを学習して，予測性能を比較する．LRとMLPはscikit-learn，XG-BoostはXGBoost[4]を利用した．予測精度の評価にはscikit-learnを利用して，ROC（Receiver Operating Characteristics）曲線の下部分の面積であるAUC（Area Under the ROC Curve）を採用する．

3.3 データセット

本稿で使用するデータセットは，中国のオンラインショップTmall[5]のデータに基づくコンペティションで用いられたデータセット[6]である．データセット内の各データは，11月11日（独身の日）に商品を購入したユーザの属性情報及び，5月11日から11月11日までのTmall上での行動履歴から構成されている．データを取得した年については明らかにされていない．表1にTmallデータセットに含まれる属性を示す．ユーザの年齢（Age）は9段階に分類され，0は不明，1は18歳未満，2は18歳〜24歳，3は25歳〜29歳，4は30歳〜34歳，5は35歳〜39歳，6は40歳〜49歳，7，8は50歳以上をそれぞれ表す．本論文では50歳以上は7に統一して扱う．ユーザがとる行動（Action）は4種類に分類され，0はクリック，1はカートに追加，2は購入，3はお気に入りに追加をそれぞれ表す．

表 1: データセットに含まれる属性

属性	値	説明
Userid	数値（6桁）	ユーザのID
Itemid	数値（7桁）	商品のID
Catid	数値（4桁）	商品カテゴリのID
Merchantid	数値（4桁）	販売者のID
Brandid	数値（4桁）	商品ブランドのID
Age	1-8	年齢の範囲
Gender	0,1	男性（0）／女性（1）
Time	0511-1111	行動をとった日（mmdd）
Action	0,1,2,3	行動の種類

このコンペティションでは，11月11日にある販売者から初めて購入したユーザが，その後の半年間で同じ販売店から再度購入するかを予測するタスクを対象としている．説明変数（ラベル）は0，-1，NULL，1の4種類であり，1が再度購入，0が再度購入なし，-1は新規顧客ではない（11月11日の購入が初めてではない）ことを意味する[7]．コンペティションではラベルがNULLのデータについて予測した結果を提出すると，推薦精度が評価される．ラベルがNULLのデータに対する正解は公開されていないため，本稿では0，1のラベルを持つデータを訓練データ，テストデータに分割して用いる．

表 2: データセットの概要

ユーザ数	販売者数	商品数	カテゴリ数	ブランド数
424170	4995	1090390	1658	8444

表 3: 各行動をとった回数

Action	0	1	2	3
回数	48550713	76750	3292144	3005723

表2はデータセットに含まれるユーザ数などである．ここで，カテゴリとは商品を分類するために設定されたグループであり，1658種類のカテゴリが存在するが，idのみのためどの様なカテゴリであるかを具体的に知ることはできない．ブランドとは，特定の会社やメーカーが提供する商品の名称や商標を指す．これもカテゴリと同じく，idしか公開されていない．表3は，4種類の行動（Action）の回数を示す．表3より，クリック（0）はユーザの最も一般的な行動であるが，購入（2）に至る行動は少ないことがわかる．カートに追加（1）した回数は購入よりも大幅に少ないため，カートに入れずに購入している行動が多いことがわかる．これは，ECサイトで提供されている「即時購入」の機能を用いることで，カートに入れずに直接購入する場合が多いことを意味している．セール期間中は，大幅に値引きされた商品が数量限定で販売されているため，ユーザは商品の売り切れを避けるために，カートを通らずに直接購入することが多いと考える．

表 4: 訓練データとテストデータ

データ	データ数	ユーザ数	販売者数	再購入者の割合
訓練	208597	169649	1993	6.09%
テスト	52267	42413	1984	6.21%

分析において，0あるいは1のラベルがついたデータの80 %を訓練データ，20 %をテストデータにランダムに分割する．このとき，ラベルの比率が訓練データ，テストデータでほぼ同じなるように分割する．表4に訓練データとテストデータの規模を示す．データセットには，同じユーザのデータが複数個存在する場合もあるが，コンペティションでは訓練データに存在しないユーザがテストデータに用いられていた．これはデー

[4]https://xgboost.ai/
[5]https://www.tmall.com/
[6]https://tianchi.aliyun.com/
[7]新規顧客でない場合は予測対象ではないが，分類モデル構築に利用しても構わないとしてデータセットに含められている．

タリークを避けるためと考えられるため，本実験でもユーザの重複がないように分割した．販売者については両データで共通である．再購入者（ラベル 1）の割合は 6 %程度であり，ほとんどのユーザが再購入していないことがわかる．モデルの学習においては 10 分割クロスバリデーションを適用し，訓練データの 20 %をバリデーションデータとして用いる．

4 分析結果

4.1 ベースライン

ベースラインは，ユーザ，販売者，およびユーザと販売者のインタラクションに関して，基本的な統計量を計算し，特徴として用いる．また，実験結果の解釈を容易にするため，特徴選択を適用してベースラインに用いる特徴数を削減する．

4.1.1 特徴の説明

ユーザ u の属性を以下に示す．

- u_1: u の年齢区分．表 1 の Age の値を用いる．
- u_2: u の性別．Gender の値を用いる．
- u_3: u がインタラクションした回数．Action の回数をカウントする．
- u_4: u が行動をとった日数．Time に基づき求める．
- u_5，u_6，u_7，u_8: u がインタラクションした販売者数，商品数，カテゴリ数，ブランド数．Merchantid，Itemid，Catid，Brandid に基づき求める．
- u_9，u_{10}，u_{11}，u_{12}: u がクリックした回数，カートに追加した回数，お気に入りに追加した回数，購入した回数．同じ商品に対する同じ行動が複数回あった場合はそれぞれカウントする．Action の回数を行動の種類ごとにそれぞれカウントする．
- u_{13}: u の購入クリック比（式 (1)）．

$$u_{13} = \frac{u_{12}}{u_9} \tag{1}$$

- u_{14}: u の重複購入率（式 (2)）．

$$u_{14} = \frac{\text{repurchase\_merchants}(u)}{\text{purchase\_merchants}(u)} \tag{2}$$

式 (2) において，repurchase_merchants(u)，purchase_merchants(u) はそれぞれ u が複数回購入したことがある販売者数，一度でも購入したことのある販売者数を意味する．

販売者 m の属性を以下に示す．

- m_1: m に対するインタラクションの回数．
- m_2: m に対するインタラクションがあった日数．
- m_3，m_4，m_5，m_6: インタラクションがあった m の商品数，ユーザ数，商品カテゴリ数，ブランド数．
- m_7，m_8，m_9，m_{10}: m の商品がクリックされた回数，カートに追加された回数，お気に入りに追加された回数，購入された回数．同じユーザから同じ行動が複数回あった場合はそれぞれカウントする．
- m_{11}: m の購入クリック比（式 (3)）．

$$m_{11} = \frac{m_{10}}{m_7} \tag{3}$$

- m_{12}: m の重複購入率（式 (4)）．

$$m_{12} = \frac{\text{repurchase\_users}(m)}{\text{purchase\_users}(m)} \tag{4}$$

式 (4) において，repurchase_users(m)，purchase_users(m) はそれぞれ m から複数回購入したユーザ数，一度でも購入したユーザ数を意味する．

ユーザと販売者の組み合わせに関する属性 um を以下に示す．

- um_1：u が m の商品にインタラクションした回数．
- um_2：u が m の商品にインタラクションした日数．
- um_3，um_4，um_5：u がインタラクションした m の商品数，カテゴリ数，ブランド数．
- um_6，um_7，um_8，um_9：u が m の商品をクリックした回数，カートに追加した回数，お気に入りに追加した回数，購入した回数．同じ商品に対する同じ行動が複数回あった場合はそれぞれカウントする．
- um_{10}：u の m の商品に対する購入クリック比（式 (5)）．

$$um_{10} = \frac{um_9}{um_6} \tag{5}$$

4.1.2 特徴選択

本稿では，LR，MLP，XG-Boost を実験で用いるため，特定の分類モデルに有利な特徴選択とならないよう，推定器としてこれらとは異なる決定木を用いて RFE (Recursive Feature Elimination）により特徴選択を行う．重要度の計算には情報ゲインを採用し，特徴数が

表 5: 選択された特徴と重要度

特徴	重要度
u_3	0.046
u_4	0.051
u_6	0.051
u_7	0.052
u_8	0.043
u_9	0.046
u_{13}	0.077
m_{11}	0.022
m_{12}	0.040
um_1	0.031

10 となるまで削減する．表 5 に，RFE によって選択された特徴と重要度を示す．

表 5 より，ユーザの特徴が多く選択されているが，これは先行研究 [3,4] の結果と一致する．一方，販売者に関する特徴では，比率に関する特徴（購入クリック比，重複購入率）のみが選択されていることから，ユーザと販売者では異なる性質を持つ特徴が選択されていると言える．また，ユーザと販売者の組み合わせに関する特徴は一つだけであり，特定のユーザ，販売者の組み合わせに限定されるため，データ数が少ないことが影響していると考えられる．

表 6 は実験で用いる各モデルについて，4.1 節で定義した全ての特徴を用いた場合と，表 5 に示した特徴のみを用いた場合の AUC を示す．表 6 より，どちらの場合も予測精度は低く，本タスクの難易度が高いことがわかるが，特徴数を減らしても性能の低下はほとんどないことがわかる．

表 6: 全ての特徴を使用した場合と特徴選択後の AUC

モデル	全ての特徴を使用	特徴選択後
LR	0.67	0.65
MLP	0.65	0.66
XG-Boost	0.68	0.67

4.2 行動時期に関する分析

3.1 節で述べた RQ1 について調査するために，4.1.2 節で選択した各特徴について，データの収集期間を変更して計算する．

4.2.1 実験 1.1

実験 1.1 では，11 月 11 日当日からどこまで遡ってデータを収集するとよいかについて調査する．11 月 11 日当日のみ，11 月 11 日の 1 週間前，2 週間前，1 か月前，6 週間前，7 週間前，2 か月前，10 週間前，11 週間前，3 か月前，15 週間前，16 週間前，4 か月前，20 週間前，21 週間前，5 か月前，24 週間前，25 週間前までの各期間をそれぞれデータ収集期間と設定して各特徴を計算する．ベースラインの特徴のうち，u_4 は 11 月 11 日当日のみの場合に全てのデータで 1 となるため，予測精度が低下することが考えられる．その場合，収集期間が与える影響の分析が正しく行えない可能性があるため，本実験では対象外とする．また，割合に基づく特徴である u_{13}，m_{11}，m_{12} も 11 月 11 日当日のみの場合に意味がないと判断し，対象外とした．図 1 に，LR，MLP，XG-Boost の各モデルについて，データ収集期間の違いによる AUC の変化を示す．

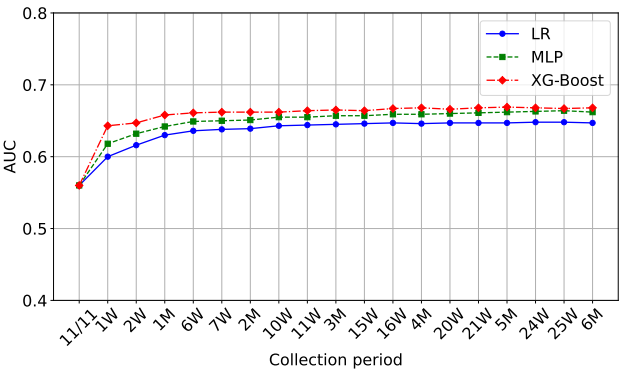

図 1: データ収集期間による AUC の変化

図 1 において，一番右の 6M はデータセットの全期間のデータを用いた場合であり，表 6 に示した結果（特徴選択後）と同一である．図 1 より，どのモデルでも 11 月 11 日当日のデータのみでは不十分であることがわかる．また，1 か月前までは AUC が向上するが，それ以前のデータは性能にあまり影響しないと言える．

4.2.2 実験 1.2

実験 1.2 では，どの月のデータが予測に影響を与えているかについて調査する．5 月から 11 月のうちの各月を除いてそれぞれ各特徴を計算する．図 2 に，各月を除去した場合の各モデルの AUC を示す．

図 2 において，一番左の None はデータセットの全期間のデータを用いた場合であり，表 6 に示した結果（特徴選択後）と同一である．図 2 より，11 月 11 日当日に近い月（11 月，10 月）のデータを除いた場合に AUC は低下するが，9 月以前のデータは除去しても性能にあまり影響しないと言える．

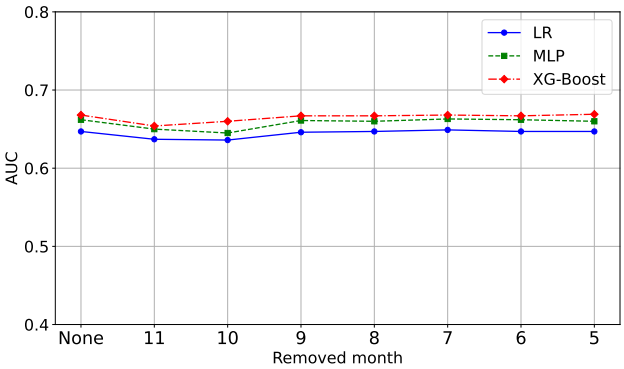

図 2: 各月を除去した場合 AUC の変化

実験 1.1 と実験 1.2 両方の結果から，11 月 11 日当日から 1 か月前くらいまでの情報が重要であると言える．

4.3 ユーザ情報に関する分析

実験 2 では，RQ1 と同じベースラインで用いているユーザに関する特徴のうち，商品レベルの情報を利用するものを除去して分類モデルを構築し，除去前と性能を比較する．具体的には，Itemid, Catid, Brandid をそれぞれ利用する u_6，u_7，u_8 を除去する．表 7 は実験で用いる各モデルについて，商品レベルの特徴を利用した場合（ベースライン）としない場合（商品レベル特徴不使用）の AUC を示す．

表 7: 商品レベルの特徴の有無による AUC の比較

特徴	LR	MLP	XG-Boost
ベースライン	0.65	0.66	0.67
商品レベル特徴不使用	0.65	0.66	0.67

表 7 において，モデルの予測精度に変化はなく，商品レベルの特徴は収集不要と言える．

5 結論

本稿では，再購入者予測タスクを対象として，Tmall のデータセットを分析し，どの時期の行動が，11/11 当日の行動に影響しているか，および取得コストやプライバシー保護の観点から有効な特徴は何かについて検討した．実験により，11 月 11 日当日に近い時期（1 か月以内）の情報は重要であること，それ以前の情報は性能にあまり影響しないことを示した．後者に関しては，商品レベルの特徴を利用しなくても予測精度は低下しないことを確認した．

今後の計画としては，11/11 当日のログデータを分析し，予測対象となる販売者のアイテムに対し連続して実行された行動系列に基づく特徴を導入する予定である．この特徴は 1 セッションの情報のみが得られる場合を想定したものであり，アカウントがないユーザにも適用できるため，プライバシー保護効果が高いと考える．また，予測精度の向上を目指すため，他の分類モデルでの実験も計画している．

謝辞

本研究は JSPS 科研費 JP22K19836, JP23K21724, and JP23K24953 の助成を受けたものである．

参考文献

[1] Liu, Y., Zhang, H., and Ren, H.: An integrated learning-based prediction model for purchasing propensity of jingdong visitors, *Highlights in Science, Engineering and Technology*, Vol. 70, pp. 60-66 (2023).

[2] Zhang, W. and Wang, M.: An improved deep forest model for prediction of e-commerce consumers' repurchase behavior, *Plos one*, Vol. 16, Issue 9 (2021).

[3] Zhang, M., Lu, J., Ma, N., Cheng, T.E., and Hua, G.: A feature engineering and ensemble learning based approach for repeated buyers prediction, *International Jouranal of Computers Communications & Control*, Vol. 17, No. 6 (2022).

[4] Liu, G., Nguyen, T.T., Zhao, G., Zha, W., Yang, J., Cao, J., Wu, M., Zhao, P., and Chen, W.: Repeat buyer prediction for e-commerce, KDD2016, pp. 155-164 (2016).

[5] Zhu, C., Wang, M., and Su, C.: Prediction of consumer repurchase behavior based on LSTM neural network model, *International Journal of System Assurance Engineering and Management*, Vol. 13, Suppl 3, pp. 1042-1053 (2022).

[6] Yang, L., Niu, X., and Wu, J.: RF-LighGBM: a probabilistic ensemble way to predict customer repurchase behaviour in community e-commerce, *arXiv preprint arXiv*. 2109.00724 (2021).