機械学習モデルを用いたシェアサイクルシステム利用目的の推定

Estimating the purpose of the shared bike system usage by machine learning models

室蘭工業大学理工学部創造工学科　○学生員　丹度彪雅(Hyoga Tando)
室蘭工業大学大学院工学研究科　学生員　大井啓史(Hirofumi Oi)
室蘭工業大学大学院工学研究科　正会員　有村幹治(Mikiharu Arimura)

1. はじめに

現在，我が国では，コロナ禍における生活様式の変化，地球温暖化対策，都市環境の在り方，国民の健康増進，観光地域づくり等様々な観点から自転車活用が促進されている．その中でも，生活利便性の向上，モビリティ向上による地域の活性化等を目的とした [1] シェアサイクルが注目されている．シェアサイクルの発展や利用環境の改善には，利用者ニーズの把握は不可欠である．しかし，無人管理型シェアサイクルの場合，移動記録データから利用動向の把握が可能である一方で，主観的な情報である利用目的を自動的に収集することは困難である．

大井ら [2] の先行研究では札幌市内のシェアサイクル「ポロクル」を対象として，アンケートデータと利用実績データにランダムフォレストを適用することで，利用目的推定するモデルを構築している．しかし，利用目的によっては精度が十分でないこと，構築したモデルを用いて複数年度にかけての利用行動変化を把握できていない等の課題がある．そこで本研究では，利用目的をデータ数が多い通勤とその他に分けてモデルを構築し，予測精度を高めること，構築したモデルを用いて 2019 年度から 2022 年度にかけての全利用データに対し予測をすることにより，利用行動の変化を把握することを目的とする．また利用目的の推定に寄与する特徴量について機械学習モデルを解釈する SHAP 指標を用いて分析する．

札幌のシェアサイクルであるポロクルは，都市内部のスムーズな移動，札幌都心部の活性化，環境負荷低減に関する啓蒙活動などへの寄与を目的 [3] として 2011 年にスタートし札幌中心部での通勤・ショッピング・観光等で利用されている．2022 年度には，過年度からのポートの新設・変更を経て，53 カ所のポートで自転車の貸し借りを行うことができ，利用回数も年々増加している．

2. データの概要

2.1 ポロクル全利用データ

ポロクルの 2020 年度全利用データは 323,766 回に利用記録からなり，多数の情報が含まれる．具体的には，ユーザーID，ユーザー種別，料金プラン，利用開始日時，返却日時，利用時間（秒），利用開始ポート名，返却ポート名，返却種別，車両情報等が含まれる．なお冬季は積雪により営業が困難なため 4 月~10 月までのデータである．本研究ではその中のユーザーID，料金プラン，利用開始日時，返却日時，利用時間（秒），利用開始ポート名，返却ポート名のデータを使用する．

表-1 建物分類

小分類	細分類
商業施設	-
官公庁施設	裁判所,道庁など
業務施設	会社,事務所など
宿泊施設	ホテル,旅館など
風俗娯楽施設	料理店,クラブなど
遊技施設	カラオケ,パチンコなど
住宅	-
店舗併用住宅	-
共同住宅	公営
作業所併用住宅	-
文教厚生施設	学校,神社など
軽工業施設	-
サービス工業施設	-
運輸・倉庫施設	-
その他の施設	-

2.2 ポロクル 2020 年度アンケートデータ

ポロクル 2020 年度アンケートデータは，ポロクル公式ホームページにて 2020 年度にポロクルを利用した人を対象に行なったアンケートデータであり，アンケート件数は 1194 件である．アンケートデータには，年代・性別・住まいなどの個人属性・利用目的・公共交通機関との接続に関してなど計 40 個の回答データが含まれている．本研究では，まず「ポロクルの利用目的は？（2 つまで選択可能）」という質問の回答で無回答と 2 つ回答しているデータを除いた 343 件のデータを抽出した．その中で通勤と通勤以外の回答を対象にして利用目的を推定する．抽出したデータの中で，通勤目的であると回答した利用者は 118 人，通勤目的以外であると回答した利用者は 225 人であった．

2.3 都市計画基礎調査データ

都市計画基礎調査データは市街化区域内における全建物（430,884 棟）について位置情報，建物用途，階数，構造および延床面積などが含まれている．本研究では 2019 年度の札幌市における「都市計画基礎調査」を使用した．その中でポートから半径 100m 以内にある合計 15 種類に区分される建物用途（**表-1**）と延床面積を用いる．

2.4 データセットの概要

本研究では，2.1 で示した 2020 年度全利用データと 2.2 で示した 2020 年度アンケートデータをユーザーIDで紐づけして算出した 10838 件のデータ（以下，アンケート利用データと記す）を作成した．アンケート利用データを，料金プランの月額会員 7707 件，1 回会員 3031 件に分けて利用する．アンケート利用データ月額会員の中で通勤目的であると回答した利用パターンは 7138 件，通勤目的以外と回答した利用パターンは 669 件であった．アンケート利用データ 1 回会員の中で通勤目的であると回答した利用パターン 1798 件，通勤目的以外と回答した利用パターンは 1233 件であった．それぞれの利用目的と平日か休日であるかダミー変数を用いて分類し，全てのパターンで利用開始ポートと返却ポートの経度・緯度からポート間距離を算出する．これらのデータに利用開始ポートの半径 100m 以内の建物用途別延床面積と返却ポートの半径 100m 以内の建物用途別延床面積を利用開始ポートと返却ポートで紐づけすることで，データセットを構築した．

3. 機械学習モデルによる推定と解釈

3.1 目的変数と説明変数

本研究では，目的変数，説明変数を**表-2** のように設定することで，2020 年度のアンケートデータから，2020 年の全利用データにはない，各利用目的の割合を推定する．さらに，各説明変数の目的変数に対する影響度を表す寄与度を求めることにより，要因分析も行う．この分析に適した手法を検討するため，3 つの手法を用い，その精度について比較を行う．

3.2 機械学習モデル

A）ランダムフォレスト

ランダムフォレストとは，バギング（全体からランダムに取り出した一部のデータをから独立に多数のモデルを作成し集約する手法）をベースとして，決定木を用いる手法である．特徴として，個々の決定木機は高い精度を持たないが，複数用いることで高い予測精度を得られること，目的変数を推定する際の説明変数の重要度を出力可能であること，決定木を使用するため非線形な事象に対応できることがあげられる [4]．ハイパーパラメータとして，決定木の個数，決定木の最大高さ，ノードの分割関数等がある．

B）XGBoost

XGBoost とは，ブースティング（一部のデータを繰り返し抽出し，逐次的にモデルを学習させる手法）をベースとして，決定木を用いる手法である．特徴として木の構造を複雑になりすぎないように正則化項をつけて形が固定された木構造の最適解を近似的に求めていく，木構造は木の分岐する前とした後の誤差を考慮して，構造を決めていくことなどがあげられる [5]．ハイパーパラメータとして，決定木の深さ，決定木の葉の重みに関する正則化項，ランダムに抽出される標本の割合等があげられる．

C）ロジスティック回帰

ロジスティック回帰とは線形回帰の出力にシグモイド関数を用いることで，説明変数から 2 値の目的変数がおこる確率を予測する手法である．特徴として説明変数間に強い相関がある場合には，回帰係数が安定せず信頼性が低くなる，一般に非線形な事象に対応できない等の問題がある [6]．

3.3 機械学習モデルの解釈手法（SHAP）

SHAP は，精度は高いが解釈性が低い深層学習等のモデルを解釈説明するための手法の一つである．機械学習で学習したモデルを単純なモデルで近似し，可読表現を用いて説明する [5]．協力ゲーム理論の Shapley Value を機械学習に応用したものである．具体には協力ゲームが報酬により限界貢献度が定義されるのに対し，SHAP では予測値を用いて貢献度を計算する．各特徴量がある時，ない時の予測値の差分をもって限界貢献度をとり，すべての順番に対し求めて平均を取ることにより，説明変数の重要度を計算する．

表-2 目的変数と説明変数

変数	変数名	単位
目的変数1	利用目的ダミー(通勤)	無次元
目的変数2	利用目的ダミー(その他)	無次元
説明変数	利用時間	秒
	平日祝日ダミー	無次元
	利用開始時間	秒
	返却時間	秒
	ポート間距離	km
	利用開始ポート半径100m以内の建物用途別延べ床面積(表-1全て)	㎡
	返却ポート半径100m以内の建物用途別延べ床面積(表-1全て)	㎡

3.4 学習データとパラメータ調整

予測モデルの精度を検証するため 2.4 で示したデータセットをホールドアウト法により無作為に分割し，7 割を学習データ，3 割をテストデータとして用いる．ランダムフォレスト，XGBoost ではハイパーパラメータの調整が精度に大きく影響するため，ランダムフォレストではグリッドサーチ，XGBoost ではベイズ最適化を用いてパラメータの調整を行った．

4. 結果

4.1 精度検証

テストデータの各利用目的における推定精度検証結果を示す（**表-3**）．

テストデータの精度に関して，月額会員の場合はランダムフォレスト，XGBoost ともに 97%以上，1 回会員の場合は，ランダムフォレストが 89%以上，XGBoost が 90%以上と高い数値を示した．ランダムフォレスト，XGBoost を用いることで，利用目的を十分に推定できる学習モデルを構築できた．

4.2 2019 年度-2022 年度の利用目的推定

XGBoost で構築したモデルを用いて，2019 年度から 2022 年度のポロクル全利用データに対して，利用目的を推定した結果を示す．（**図-1**，**図-2**）．

図-1 より通勤目的の利用者に関して，月額会員の場合は 2020 年から 2022 年にかけて増加傾向にあることが分かる．また，1 回会員の場合は 2020 年から 2022 年にかけて利用者総数はほとんど変化がないものの，2022 年の利用者割合については減少という結果になった．

図-2 より通勤以外を目的とする利用者に関しては，月額会員の場合は 2020 年から 2022 年にかけて利用者総数，割合ともに減少傾向にあり，また，1 回会員の場合は 2022 年に利用者総数，割合ともに増加している結果となった．

4.3 利用目的推定の解釈

2020 年度データを用いた月額会員の通勤目的利用結果に関して，機械学習モデルの解釈のため Shapley Value を計算した．

各説明変数の寄与度を示す（**図-3**，**図-4**）．利用目的が通勤の場合，月額会員，1 回会員ともに，OD 距離が最も重要度が高くなり，返却時間，利用時間（秒），利用開始時間も重要な説明変数であると分かった．

また，各説明変数の特徴量の増減と目的変数の相関を示す（**図-5**，**図-6**）．横軸が目的変数の値で縦軸が説明変数の貢献度の高さ，色が各サンプルの特徴量の値（赤が正，青が負）である．**図-5** より，月額会員の利用目的の推定では，OD 距離は青（OD 距離が短いとき）が負の方向，赤（OD 距離が長いとき）が正の方向へ多く分布しており，OD 距離は距離が短いほど通勤目的が選択されない方向へ作用する正の相関がみられた．利用時間（秒）は青（利用時間が短い）ほど正の方向，赤（利用時間が長い）ほど負の方向へ多く分布しており，利用時間（秒）が長いほど通勤目的が選択されない方向へ作

表-3 テストデータの精度

機械学習モデル	テストデータの精度	
月額会員	通勤	その他
ランダムフォレスト	97.27%	97.27%
XGBoost	97.10%	97.23%
ロジスティック回帰	93.00%	93.00%
1回会員		
ランダムフォレスト	89.56%	89.56%
XGBoost	90.77%	90.44%
ロジスティック回帰	70.00%	70.00%

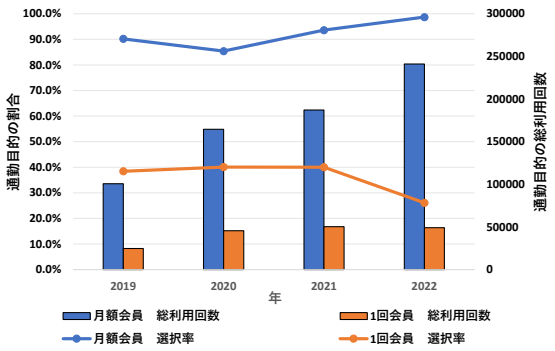

図-1 2019 年から 2022 年までの全利用データ
通勤利用目的の割合と総利用回数

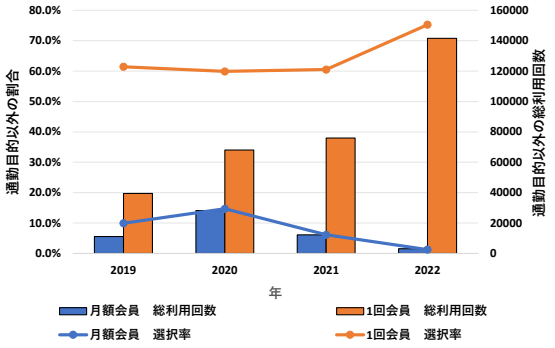

図-2 2019 年から 2022 年までの全利用データ
通勤利用目的以外の割合と総利用回数

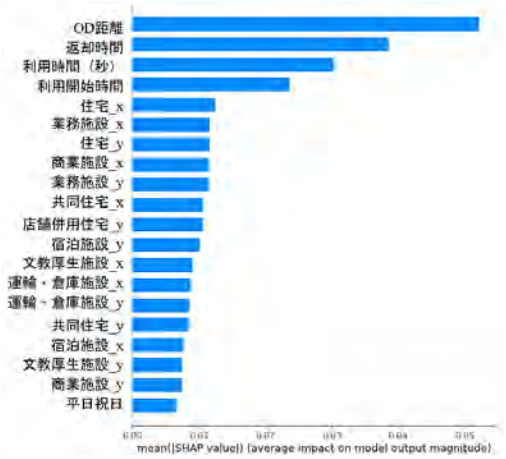

図-3 月額会員・通勤目的の Shapley Value

用する負の相関がある．図-6 より，1 回会員の推定では，OD 距離は月額会員の推定と同様に， OD 距離が長いほど通勤目的が選択される方向へ作用する正の相関があることを確認した．利用時間（秒）に関しては利用時間（秒）が長いほど通勤目的が選択されない方向へ作用する負の相関がみられた．

5. まとめ

本研究では，ランダムフォレスト，XGBoost，ロジスティック回帰を用いてアンケート利用データから全行動パターンの利用目的の推定を行った．その結果を以下に示す．

- ランダムフォレスト，XGBoost ともに通勤目的，通勤以外目的の予測精度が高く，概ね推定精度が得られたといえる．
- XGBoost を用いて，2019 年から 2022 年までの全利用データに対し，利用目的を推定することにより，年度を経ていくにつれて月額会員は通勤目的で利用し，1 回会員は通勤以外目的で利用する傾向がみられた．
- SHAP を用いることでそれぞれの予測モデルの特徴量をつかむことができ，OD 距離，返却時間，利用時間（秒），利用開始時間が重要な説明変数であることを示した．

課題としては，本研究では通勤とそれ以外に利用目的を分けてモデルを作成したため，それ以外の中に分類される観光やショッピング等の行動分析ができないことが挙げられる．今後 COVID-19 の感染拡大が収束し，通勤目的以外でポロクルが多く使われるようになったときに再び細かく利用項目ごとに予測モデルを作成し分析していきたい．

謝辞：本研究では，NPO 法人ポロクルからの貴重なデータを頂きました．ここに記して感謝の意を表します．

参考文献

1) 第 4 回 シェアサイクルの在り方検討委員会 配付資料，国土交通省：https://www.mlit.go.jp/road/ir/ir-council/sharecycle/giji04.html（2022.12.06 閲覧）
2) 大井啓史，野崎脩人，坂本信，浅田拓海，有村幹治：ランダムフォレストを用いたシェアサイクルの利用目的別トリップの推定，第 63 回土木計画学研究発表会，2021
3) ポロクルホームページ：https://porocle.jp/（2022.12.05 閲覧）
4) 波部斉：ランダムフォレスト，情報処理学会研究報告，Vol.2012-CVIM-182 NO.31，2012
5) 板橋将之，本田あおい，大北剛：SHAP 値や重要度を用いたモデル解釈性:包除積分ネットワークと XGBoost の比較，火の国情報シンポジウム 2020 情報処理学会 九州支部， 2020
6) ロジスティック回帰分析と傾向スコア（propensityscore）解析，天理医学紀要 2016 年 19 巻 2 号 p.71-79，2016

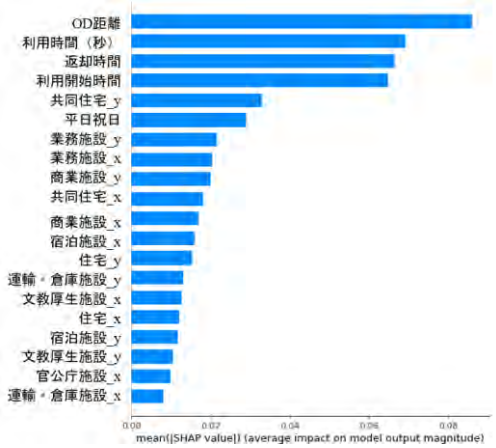

図-4 1 回会員・通勤目的の Shapley Value

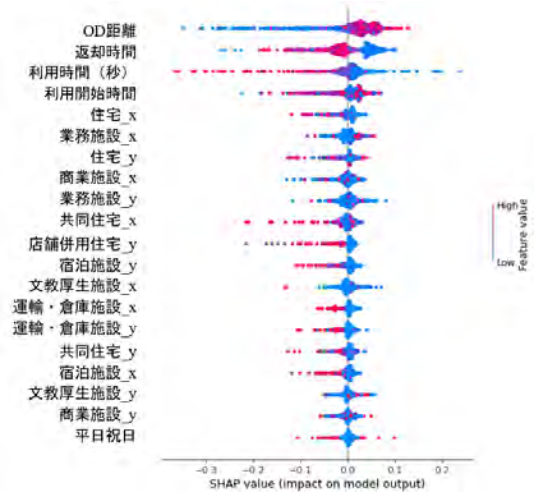

図-5 月額会員・通勤目的の Shapley Value

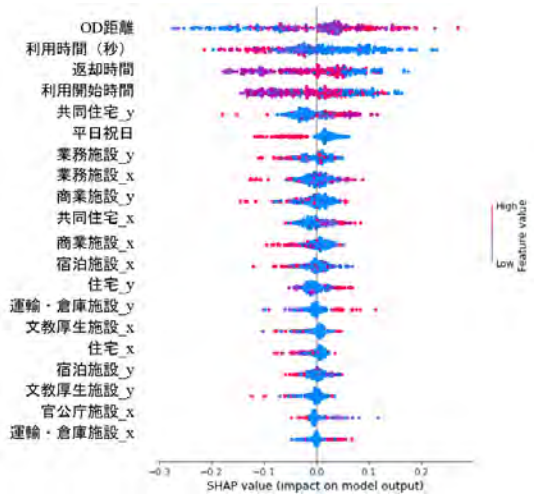

図-6 1 回会員・通勤目的の Shapley Value