返信率予測モデル改修に伴う成果物説明書

2025 年 2 月 18 日

1　プロジェクトメンバー

TRANS-AM: 大石達也, Inchol Choi, 柳原啓美, 若月美樹
Panel Planning Group: 花立沙代子
BAU: 西村広之, 西尾洋平

ドキュメント作者: BAU 西尾洋平

2　概要

本プロジェクトは，回収予測業務の一つであるメール返信率の予測精度向上 (特に，提携パネル) を目的に発足したプロジェクトである．目的達成のために，新たに開発された成果物は下記 3 点である．

1. 各モニタにおける対象期間内の日次回答率を集計する SQL スクリプト
2. 各モニタにおける指数移動平均 (EMA) を算出する Python スクリプト
3. 最終アウトプットを指定の S3 へ保存する Python スクリプト

本ドキュメントは，それぞれのスクリプトの内容を記す．プロジェクト背景とモデル仕様，そして今後の課題は別資料 (「返信率予測モデル改修に伴うモデル設計書」) を参照すること．それぞれのスクリプトは，Databricks 上で動かし，使用クラスターは，EMA_Computation を想定している．また，Job として運用者がそれぞれの notebook を一貫して回せるようにも設定済みである．それぞれのスクリプトのため使用したパッケージの情報は，Databricks 上のディレクトリー

1. Volumes/respondents_reqruitment/model_millbox/package_infor/version_info.txt

へ保存した．パッケージに関するエラーが出る場合，パッケージ有無とバージョンの確認を行うこと．

3　ツールの利用手順書

3.1　各モニタにおける対象期間内の日次回答率を集計する SQL スクリプト

本 SQL は，4 つのブロックに分割される．また，この SQL にて操作されるデータベースは，SUNABA 内の transam テーブルを想定している．各ブロックでの処理は下記で示す．使用した AWS に関する情報 (i.e., credential information) は，別途ファイルに記載.

1. 対象期間のリスト作成

（a）対象期間は，変数 start_date および end_date として運用者が自由に設定できる仕様とした．検証時には，任意に 3 ヶ月間 を採用.

（b）例えば，運用者が”2024/12/1”の回答率の予測を行いたいとした時，start_date =’2024-08-30’, end_date = ’2024-11-30’ と 3 ヶ月間の日付を Databricks の Job 画面上で設定する

i. 予測したい日付の’ 前日’ から 3 ヶ月間の日付を設定する

2. **対象期間内に完了した調査 ID** に対し，回答したか否かの回答データを抽出

（a）予測値を安定させるため，使用するデータは調査がすでに完了し，回答が確定されたデータを使用する．

3. 2. で抽出したデータより，日次回答率 (= 回答数÷調査数) を集計し，下記データを整形する (表 1).

（a）モニタ ID

（b）配信日

（c）各モニタが日毎に受け取った調査数

（d）各モニタがに日毎に行った回答数

（e）日次回答率

4. 後ほど全モニタへデータを結合するため，このスクリプト内で，transam.panelist より全モニタ ID も抽出しておく．

カラム名	データタイプ	詳細
mm_panelist_id	bigint	モニタ ID
create_dt	timestamp	調査配信日
pjt_count	int	各モニタが，日毎に受信した調査数
ans_count	int	各モニタが，日毎に回答した調査数
ans_ratio	float	ans_count / pjt_count

表 1　本 SQL から抽出されるテーブル定義

3.2　各モニタにおける指数移動平均 (EMA) を算出する Python スクリプト

本 Python スクリプトは，下記 3 工程を行う．

1. データの読み込み
2. 指数移動平均 (自作) を算出する関数を適応

（a）指数移動平均 (EMA) の詳細は，別資料「返信率予測モデル改修に伴うモデル設計書」内の”スムージングについて”項目にて記載するので要参照.

3. 必要なカラムを選択し，最終アウトプットをするため整形を行う

特定のパッケージと関数をインストールした事を確認して，スクリプトを実行をする．想定アウトプットは表 2 にまとめた.

カラム名	データタイプ	詳細
mm_panelist_id	bigint	モニタ ID
ema	int	日毎の回答率へ指数移動平均によりスムージングを行った値

表 2　想定アウトプットテーブル

EMA の算出後は，先ほど抽出した全モニタ ID へデータへ左結合を行う．

3.3　アウトプットを指定の s3 へと格納する

最後の notebook では，EMA の処理がなされた後，csv ファイルを TRANS-AM チームの指定するディレクトリーへと保存を行う．

1. 開発用ディレクトリ：s3://transam-answer-rate-prediction-dev/model-change-2025/ (2025 年 2 月 18 日時点)