

過去企画へのオーバーフィッティングを回避した成約率予測モデルの構築 Construction of conversion rate prediction model which avoids overfitting to past events

福本 信吾[†] 毛利 研[‡] 深澤 信也[‡] 伊藤 悠希[†] 池田 千枝[†] 菅野 敬明[†]
Fukumoto Shogo Ken Mohri Shinya Fukasawa Yuki Ito Ikeda Chie Kano Takaaki

1. はじめに

医療保険やがん保険は、疾病率が加齢とともに上昇するため、あらかじめ生涯の保障をベースに加入する終身契約が多くなっている。ただ、ライフステージの変化、医療制度の変化、医療技術の進化によって備えるべき保障が変化するため、それに合わせた保険のニーズに対応するべく適時最新化を図ることが望ましい。

このため、最新化の必要性を契約者様の方に幅広くお知らせすべく、各種メディアを用いてご案内をお送りしているが、そのメディアの中でも、ダイレクトメール（以下、DMという）は、①住所などの情報を元に対象に直接かつ確実に届くことや②消費者からのメディアとしての信頼性も比較的高く開封閲読率も高いというリーチする力の強いメディアであるという調査結果もあり [1]、③CMやネット広告のように全ての対象に1つのメッセージしか伝えることのできないマスメディアとは一線を画し、情報量も多いといった機能を持つ。

DMを活用した既契約者様向けのクロスセル（他商品紹介）は、ダイレクトマーケティングにおける顧客活性化（リテンション）プロセスであり、既存顧客により多く、より長期の購買を働きかけて「顧客生涯価値（Life Time Value, LTV）」を高める活動、保険の場合には、より適切な契約への見直し（最新化）を働きかける活動と言い換えることができる [2,3]。

しかし、消費者の多様化と共に、顧客のライフステージの変化の仕方や価値観もより複雑なものになっており、ここにお客様の属性や傾向を理解するためのデータアナリティクスが有効に機能する余地がある。実際のところは、保険契約に対するニーズを共通した単一のモデルで測ることは簡単なことではなく、複雑な因果関係を表現するような分析とモデル構築が必要とされている。

これを実現するために、近年、機械学習が積極的に活用されるようになってきており、複雑で膨大なデータから顧客の過去における契約変更状況を探り、1人ひとりのニーズを高精度で予測、影響度が高い要素を的確に見つけ出すことで、顧客理解による契約最新化の勧奨といった活動に寄与できることが期待されている。

本論文では、四半期ごとに実施された過去約1年間分のDM企画で商品を案内した顧客データを利用して、商品案内後のある一定期間（Delayed feedback; 以下、DF期間という）にその内容に紐づく保険契約の成約（以下、コンバージョン CV という）が期待される確率を算出する、成約率予測モデルを提案する。この予測モデルは、以下の課題を解決したことが新規性として挙げられる。

- DM を発送する企画内容は毎回同様のものではなく、送付先の性質や件数が一定程度異なるため、単純にモデル

を構築すると、過去における特定の企画にオーバーフィッティングする。

- 事前の調査にて、企画毎に繰り返し DM が送付される対象者については成約率の低下が判明していたため、企画を独立に考えると想定通りに成約率が向上しない可能性がある。

なお、本論文にて構築した予測モデルは保険業界に限らず DM メールを送付する多くの企業にも参考となる。また、先に挙げた課題は DM に限らず e-mail やアプリのプッシュ通知にも適応できるものである。

2. 基礎分析

本論文で利用するデータは、表 2.1 に示す 4 つのカテゴリ（デモグラフィックデータ、ライフスタイルデータ、サイコグラフィックデータ、行動データ）に大別できる。

なお、本論文において過去 1 年分のデータを分析したが、顧客属性について過去時点の情報に戻す作業を実施せず、誤差の範囲と見なして分析時点で得られる情報で分析を進めた。

表 2.1 特徴量の分類

データ種別	主なデータ項目の内容
デモグラフィックデータ	顧客属性（年齢、性別等）、居住地
ライフスタイルデータ	居住地別の公的データから推計した特徴量
サイコグラフィックデータ	過去企画参加情報（今回、新しく定義した特徴量）
行動データ	契約内容、契約期間、異動

2.1 時間ウィンドウの定義

本論文における予測モデル構築に際して、時系列依存性を考慮した分析が必要だという前提に立つ。主な論点として、(1) 顧客が施策に参加してから成約するまである程度の DF 期間を要すること [2,3]、(2) DM が送付される前にも、DM のみならず他のチャネルからのアプローチによる影響を加味する必要があることが挙げられる。今回は、データが取得できる過去 8 ヶ月間を観測期間とし、電話コールを主体とした種々のアプローチを考慮に入れている。また、(3) 複数回 DM が送られているのにも関わらず、反応が無い状態が続く場合、どのぐらいの期間をおいて再度アプローチをすれば良いのか。これらを考慮した成約を評価するため、図 2-1 に示すような時間ウィンドウを設ける。

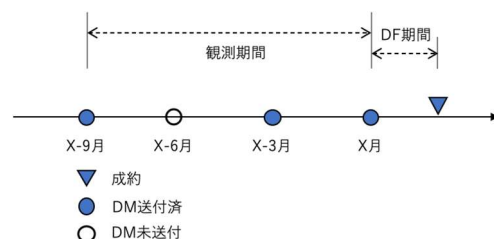


図 2.1 成約率予測モデルの時間ウィンドウの定義

[†]1 アフラック生命保険株式会社 デジタルイノベーション推進部

Aflac Life Insurance Japan Ltd., Digital Innovation Lab

[‡]2 有限責任監査法人トーマツ デロイトアナリティクス

Deloitte Touche Tohmatsu LLC, Deloitte Analytics

2.2 DF 期間の設定

前述したように DM を受け取ってから成約するまでの時間差, DF 期間を算出することが複数の企画を均一の条件で扱う上で重要となる. 時系列依存を考慮しない場合, 成約すれば正例, しなければ負例とするが, 直近の企画になるほど観測期間が短くなり, 成約率が低下しているように見えるためである. この DF 期間を設定するため, ラグ1ヶ月毎の累積成約率を計測した. その結果, その値が95%以上となる時点をも DF 期間としたとき, 本論文で扱う DM の場合は, 2ヶ月が適当であることが分かった (図 2.2 参照).

また, 過去8ヶ月以内に DM が再送付された場合には, その CV 率に統計的な優位な差があり, ある一定のラグ期間を置くと成約率が増加する傾向が見られた. そのため, これらの変数を特徴量として予測モデルに導入することとした.

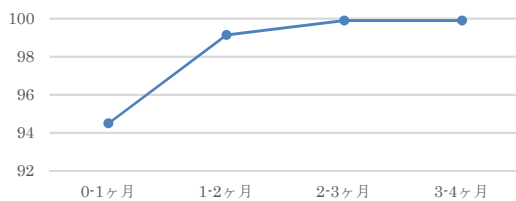


図 2.2 ラグ 1 ヶ月毎の累積成約率

3. モデリング

3.1 アルゴリズム

先行研究において, 予測するモデルとして多く提案されているのはロジスティック回帰モデルである [4,5]. 例えば, Lee (2012) [4] はディスプレイ広告における成約率予測の問題に対して, 過去の実績成約率を特徴量としてロジスティック回帰モデルを構築する手法を用いている. このように, ロジスティック回帰が多く採用される背景として, 計算面で高速であり学習の収束性も良いことや, 特徴量間の関係が判別しやすいため評価の根拠が理解しやすく業務改善につなげやすいという利点が挙げられる. 一方で, 説明変数間に多重共線性が存在する場合, 推定精度が低下する課題がある. またロジット変換した際, 説明変数と被説明変数にある程度の線形性がないと外れ値の推定精度が低くなる. 実際, 本論文において被説明変数(Target)と各説明変数との間でピアソンの相関係数を取るとほぼ無相関となっていた (図 3-1 参照). そのため, Kushal et al. (2010) [6] や Ilya et al. (2012) [7] と同様に非線形性を陽に考慮できる Gradient boosting decision trees の一つである XGBoost を基本アルゴリズムとして採用した.

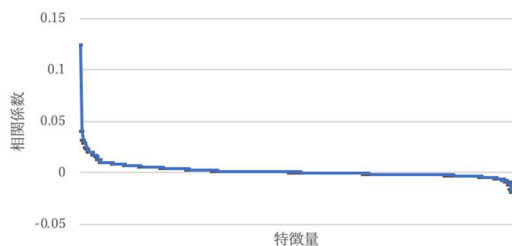


図 3.1 Target と説明変数間のピアソンの相関係数を降順でグラフ化 (特徴量は 300 個程度あるため, 具体的な名称は省略した)

3.2 正例・負例の定義

一般的に成約率予測は, 全データに対して正例 (成約した顧客) が少ないことが特徴である. そのため, Agarwal (2007) [8] に従い, 顧客の最終成約を起点として属性データやその他特徴量を抽出した. 彼らに従うと, 図 2-1 に示すように顧客が最後に DM を受け取った企画に対してのみ成約を定義される. 成約したと判定する期間 (DF 期間) は, 2.2 節で述べたように 2ヶ月間としており, それ以外の顧客は負例として扱った.

しかし負例に関して, 上記のように顧客が最後に DM を受け取った企画に対してのみ正例 (成約) を定義した場合, 過去の企画ほど ”見かけ” の成約率が高くなることが課題としてあげられた. そのため, 本論文において負例を各企画断面からサンプリングすることで正例の負例に対する割合を ”実際に” 近づけることで, 特定の企画に偏らない, 全般的な成約率の予測精度を向上させた (図 3.2 参照).

負例サンプリング前		負例サンプリング後	
実施月	割合の倍率※	実施月	割合の倍率※
X 月	13.29	X 月	2.19
X - 3 ヶ月	7.03	X - 3 ヶ月	1.41
X - 6 ヶ月	2.06	X - 6 ヶ月	1.23
X - 9 ヶ月	1.00	X - 9 ヶ月	1.00

※ 各企画の負例に対する正例の割合について, (X-9 カ月) を基準に倍率で表した場合の値

図 3.2 負例の補正

本論文で取り扱った DM における成約までの時間スケールは数ヶ月と, WEB への広告出稿サービスの数時間から数日よりも長いことが特徴である. そのため, 原ら (2015) [9] が行ったような, 統計モデルを構築, DF 期間を待たずして正例および負例を予測することで学習データ量を確保するアプローチは取らなかった. しかし, 今後 DM 以外のチャネル (例えば, email や SNS アプリによる通知) において成約を予想する場合には, 同手法の導入を検討する必要がある.

3.3 学習・テスト・検証用データの定義

本論文で構築する予測モデルは, 企画参加履歴を含む顧客属性から, 成約しやすい顧客を識別することを目的とするため, 成約率が高い企画に参加した顧客を予測するだけの汎化性能が低いモデル構築を避ける必要がある. 今回のデータにおける特徴として企画毎に企画特性が違うことや, 外部環境が変化していることに起因するため成約傾向が異なることが確認されている (前掲の図 3.2 参照). そのため, 単純に学習データを作成してモデルを構築した場合, ある特定の DM 企画にオーバーフィッティングした予測モデルとなることが判明している (図 3.3 参照).

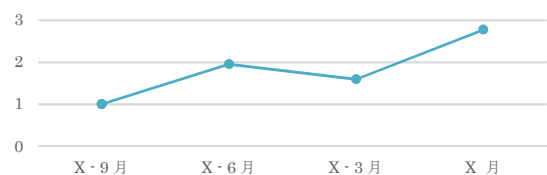


図 3.3 X-9 月の成約率を基準とした各企画の成約傾向 (学習データを全企画一律に扱った場合)

本論文において上記の課題を解決するため、学習データの作成方法とモデルの評価指標を工夫することで、企画バイアスに引きずられないモデル構築を行った。具体的には、学習データは様々な企画におけるデータを学習データとし、特定の企画における成約パターンにオーバーフィッティングすることを防いだ。また評価指標は Borkan et al. (2019) [10] に従い、(1) subgroup AUC : ある DM 企画 (企画 A) の AUC, (2) BNSP AUC : 企画 A の正例 + 企画 A 以外の負例の AUC, (3) BPSN AUC : 企画 A の負例 + 企画 A 以外の正例の AUC を算出し、企画バイアスを考慮した評価を行った。これらの評価指標を用いることで、成約率が高い企画に参加した顧客を予測するだけのモデルが構築された場合、上記(1)~(3)いずれかの指標が著しく悪くなる。従ってモデル構築時は上記(1)~(3)を常時モニタリング、いずれの評価指標も著しく悪い値とならないように注意を払った。最終的なモデルの評価指標としては、これらの指標が一定条件を満たすという条件を付けた上で通常の AUC を用いた。

学習データとテストデータは利用可能な全データを学習 : テスト = 8 : 2 となるようにランダムスプリットして作成した。また、検証用データは、k-分割交差検証 (今回は、k=3) によって実施した。

3.4 不均衡データの取り扱い

本分析ではモデルの評価方法と構築方法にそれぞれ AUC とアンダーサンプリングを使用する。アンダーサンプリングの際の調整可能なパラメータは正例と負例の件数比及び、バギングするサンプルデータセット数である。ここでバギングとは、学習データセットから複数のサンプルデータセットを作成し、サンプルデータセットから学習したモデルを作成、アンサンブルする手法である。これらの調整可能なパラメータをいくつか試行し、バリデーションデータにおける最も良い AUC のパラメータを採用することにする。

ダウンサンプリングを実施すると少数派クラスの事前確率が大きくなるという予測確率のバイアスが生じる。具体的には、正例のクラスが少数派クラスであるので、正例と予測される確率 (事後確率) が大きくなる [11,12]。そのため、3.4 式を用いて calibration を行うことで、バイアスの影響を除去した。calibration 後のモデル予測値分布は、図 4.1 (右) に示す。

$$p = \frac{p_s}{p_s + (1 - p_s)/\beta} \quad \dots\dots\dots 3.4 \text{ 式}$$

ここで、 p_s はダウンサンプリングを行ったデータセットで学習されたモデルの予測確率、 β はダウンサンプリング率である。

3.5 特徴量選択

特徴量選択には過学習しやすい特徴量の削除を目的として null importance [13] を用いた手法を採用した。これは、ターゲット変数をシャッフルしたデータセットを複数作成し、そのデータそれぞれに対して学習モデル (null モデルと呼ぶ) を構築したときに、ある特徴量の実際のモデルによる重要度 (actual importance) と複数の null モデルから得られる重要度 (null importance) の n パーセント値 (例えば $n=75$) の対数比がある閾値以上 (例えば 0) になれば選択するという手法である。これにより、対数比が小さい特徴量は過学習を引き起こしやすいと解釈できる (図 3.5 参照)。

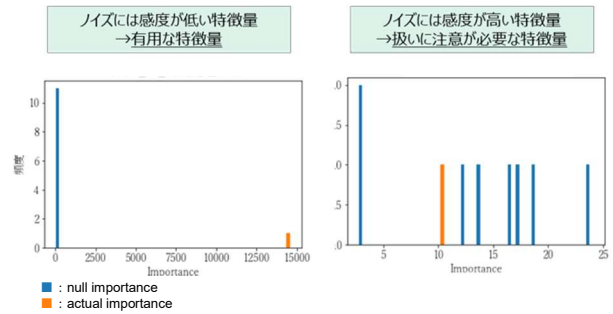


図 3.5 null importance による特徴量選択

4. モデル構築結果

構築したモデルの ROC 曲線および予測値 (確率キャリブレーション前) の分布を図 4.1 に示す。今回構築した予測モデルの AUC は 0.77 と、企画バイアスに引きずられていた際の 0.63 と比べ大幅な向上が見られた、また、Borkan et al. (2019) に従い算出した企画別の AUC 及び予測値の分布 (図 4.2-1 ~ 図 4.2-3) に示す通り、企画に依らず汎化性能が高い予測モデルが構築できたと結論できる。また、今回新たに導入した、過去 8 カ月以内において一定の期間で再送付する特徴量が予測に大きなインパクトを与えていることも明らかになった (表 4.1 参照)。

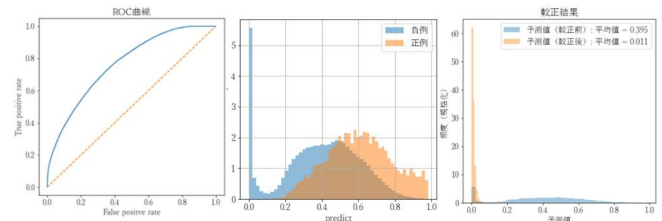


図 4.1 ROC 曲線 (左) と予測値の分布 (中)、calibration 前後のモデル予測値分布 (右)

表 4.1 特徴量重要度上位 5 個

特徴量	重要度 (※)
高度先進医療特約	15,115
居住地における人口密度	5,948
過去 8 ヶ月以内再送付期間	5,446
保有特約件数	3,767
手術に関する特約	3,624

※ total gain の値を記載

5. まとめ

本論文では、定期的な企画の内容が毎回同様のものではなく、対象顧客の性質も異にする過去データを用いて、契約者様が契約を最新化する成約率予測モデルを構築した。また、このモデルは前の企画からの期間を特徴量として利用することで、最後に DM が契約者様に届いてから次の DM を送付するまでの最適な再送付期間を判別することを可能とした。

本論文においては、契約者様とのコミュニケーションチャネルの中で、DM を取り上げた。今後は、コミュニケーション手段が多様化する中で、契約者様にとって一番納得できる形で案内を差し上げるべく、包括的に契約者様のニーズ、および複数あるチャネルを管理・最適化する、顧客一人ひとりに合わせてカスタマイズした One to One マーケティングの実現を、機械学習のひとつでもある強化学習なども活用して検討していく所存である。

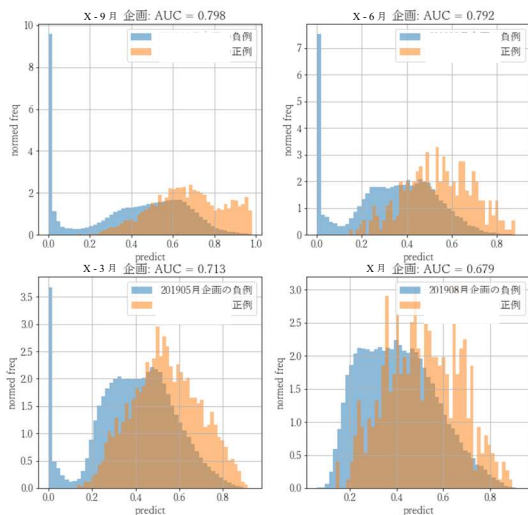


図 4.2-1 subgroup AUC

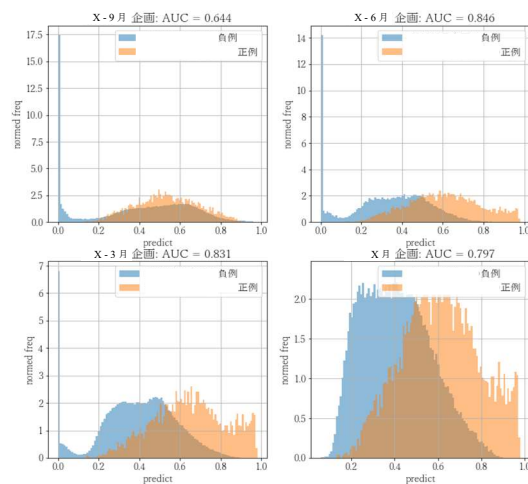


図 4.2-2 BPSN AUC

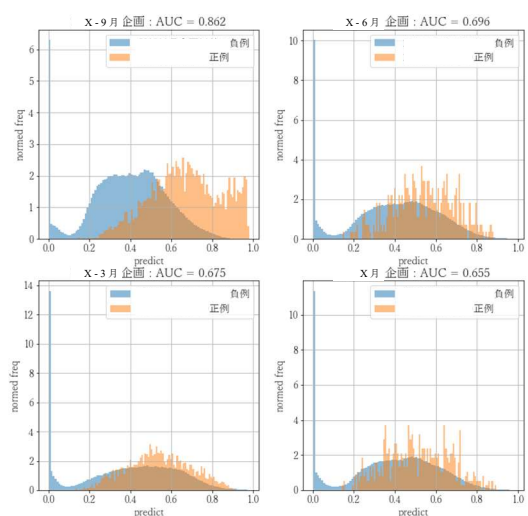


図 4.2-3 BNSP AUC

謝辞

本論文を推薦するにあたりビジネス上の示唆を与えてくださったアブラック生命保険株式会社の佐伯和則さんに、謹んで感謝の意を申し上げます。

また、Deloitte Analytics の研究開発部門をリードしている神津友武さんには、本プロジェクトの進行から本論文への推薦まで終始適切な御助言と御指導を賜りました。心から敬意と感謝の意を表します。

参考文献

- [1] 一般社団法人日本ダイレクトメール協会: 「DM メディア実態調査 2018」報告 (要約版, available from <https://www.jdma.or.jp/upload/research/20-2019-000013.pdf>) (accessed 2020-04-03).
- [2] Olivier Chapelle, “Modeling delayed feedback in display advertising.”, In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, (KDD '14)
- [3] Yoshikawa, Yuya and Yusaku Imai. “A Nonparametric Delayed Feedback Model for Conversion Rate Prediction.”, ArXiv abs/1802.00255 (2018)
- [4] Lee, K.-c., Orten, B., Dasdan, A., and Li, W.: Estimating Conversion Rate in Display Advertising from Past Performance Data, in Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD '12)
- [5] Rosales, R., Cheng, H., and Manavoglu, E.: Post-click Conversion Modeling and Analysis for Nonguaranteed Delivery Display Advertising, in Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, (WSDM '12)
- [6] Kushal S. Dave and Vasudeva Varma. Learning the clickthrough rate for rare/new ads from similar ads. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, 2010.
- [7] Ilya Trofimov, Anna Kornetova, and Valery Topinskiy. Using boosted trees for click-through rate prediction for sponsored search. In Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy, ADKDD '12, 2012
- [8] Agarwal, D., Broder, A. Z., Chakrabarti, D., Diklic, D., Josifovski, V., and Sayyadian, M.: Estimating Rates of Rare Events at Multiple Resolutions, in Proc. of KDD 2007, pp. 16–25 (2007)
- [9] 原 淳史, 高野 雅典, Roman Shtykh, 川端 貴幸, インターネット広告におけるコンバージョンに近いユーザの抽出方法の検討, 2015 年度人工知能学会全国大会 (第 29 回), (2015)
- [10] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, Lucy Vasserman: Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. WWW (Companion Volume) 2019: 491-500.
- [11] A. D. Pozzolo, O. Caelen, R. A. Johnson and G. Bontempi, “Calibrating Probability with Undersampling for Unbalanced Classification,” 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, 2015, pp. 159-166, doi: 10.1109/SSCI.2015.33.
- [12] Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, and Joaquin Quiñonero Candela. 2014. “Practical Lessons from Predicting Clicks on Ads at Facebook.”, In Proceedings of the Eighth International Workshop on Data Mining for Online Advertising (ADKDD'14).
- [13] André Altmann, Laura Toloşi, Oliver Sander, Thomas Lengauer, Permutation importance: a corrected feature importance measure, *Bioinformatics*, Volume 26, Issue 10, 15 May 2010