クーポン割り当て最適化実験

背景

クーポン A は 1000 円、クーポン B は 1500 円、クーポン C は 2000 円の割引を提供する。
ここではこれまでクーポン 2 が全ユーザーに配布されていた状況を仮定する。

目的

クーポンの割り当てを最適化することにより、CV 数を最大化すること。

実験 1: 2 種類のクーポンを用いたコスト最小化

背景・目的

クーポン A とクーポン B のみを用いて、CV 数の棄損を抑えた上でのコスト最小化を行う。ここで浮いたコストが、より高い割引額のクーポンを配布する機会を生む。そのため、この実験を通して CV 数最大化に向けた改善の余地があるか確認する。

検討ロジック

検討したロジックは大きく次の 2 段階に分かれる。

1. CVR とポイント利用率（PUR; Point Utilization Rate）を予測する。
2. 予測値を用いて CV 数を最大化する数理最適化問題をソルバーで解き、ユーザーへの最適なクーポン配布を行う。

まず、今回解きたい問題を以下のように定式化した。

$$
\begin{aligned}
\text{Minimize} \quad & \sum_{s \in S} \sum_{a \in A} x_{s,a} N_s r_{s,a} p_{s,a} C_a \\
\text{s.t.} \quad & x_{s,a} \in [0, 1] \\
& \sum_{a \in A} x_{s,a} = 1, \ \forall s \in S \\
& \sum_{s \in S} \sum_{a \in A} x_{s,a} N_s r_{s,a} \geq R
\end{aligned}
$$

表記は以下の通りである。

- S: ユーザーセグメント集合
- A: クーポン集合
- $x_{s,a}$: セグメントsへのクーポンaの配布率
- $r_{s,a}$: セグメントsにクーポンaを配布した時に CV する確率（CVR）
- $p_{s,a}$: セグメントsにクーポンをaを配布して CV した際にクーポンを利用する確率（PUR）
- N_s: セグメントsに属するユーザー数
- C_a: クーポンaを使って CV された際に発生するコスト
- R: 許容できる CV 数の下限

この最小化問題を解くことで、セグメントsにクーポンaを何%配布するかを決めることができる。最適な$x_{s,a}$を得た後のユーザーに対するクーポン割り当ては、配布割合を用いて確率的に行なった。

実験方法

ダミーデータを生成し、64%を学習データ、16%を検証データ、20%をテストデータとした。CV したかと CV した上でポイント利用したかの両方において分布に偏りを生じさせたくなかったため、データの分割には iterstrat を用いた。

CVR 予測モデルは、CV したかを目的変数とし、LightGBM を Binary Logloss で学習させることで得た。PUR 予測モデルは、CV した上でクーポンを利用したかを目的変数とし、LightGBM を Binary Logloss で学習させることで得た。

ユーザーセグメントは、次の 2 段階で作成した。

1. CVR 予測モデルによって得られたクーポン A 付与時 CVR をもとに、pandas の qcut で 10 分割。
2. 1.で得られたセグメントごとに、クーポン B 付与時 CVR をもとに pandas の qcut で 10 分割。

セグメントごとのユーザー数に偏りが出ないこと、なるべく CVR が近いユーザーを同じグループにまとめることを狙った。

予測値を用いた数理最適化は、pulp を用いて解いた。最適な$x_{s,a}$を得た後のユーザーに対するクーポン割り当ては、配布割合を用いて確率的に行なった。

最適な割り当てを行なった際に生じる CV 数とコストの推定は、次のように行なった。

1. 最適な割り当てを行なった場合のクーポンと、ログ上のクーポンが一致しているデータを抽出。
2. 1.のデータを用いて、クーポンごとに CVR と PUR を計算。
3. クーポンごとの付与人数に 2.で求めた CVR をかけることで推定 CV 数を計算。
4. 3.で求めた推定 CV 数に PUR とクーポンの割引額をかけることで推定コストを計算。

結果

CVR 予測モデルと PUR 予測モデル

CVR 予測モデルの学習曲線を図 1、PUR 予測モデルの学習曲線を図 2 に示す。CVR 予測モデルは損失が下がっていく傾向が見られたが、PUR 予測モデルは損失が上がる傾向が見られた。PUR 予測モデルに利用できるデータは CV したデータのみであり、サンプル数が少ないため CVR 予測モデルよりも学習が難しかった可能性や、説明変数と目的変数の間で関連が弱かった可能性が考えられる。

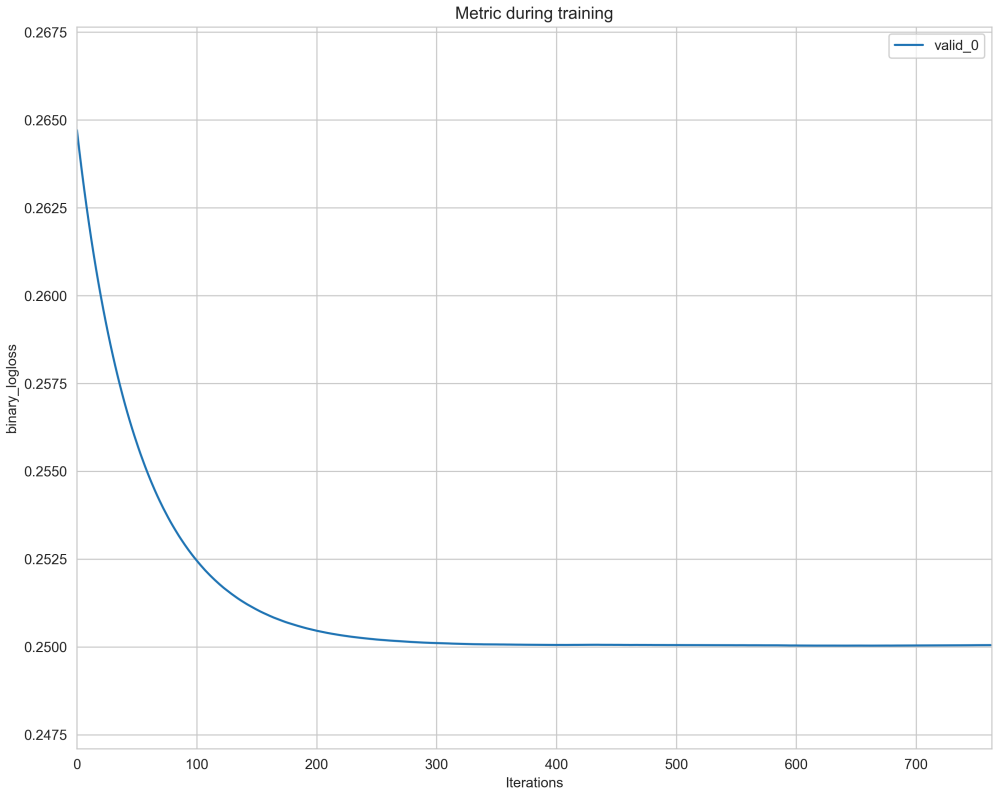

図1: CVR予測モデルの学習曲線

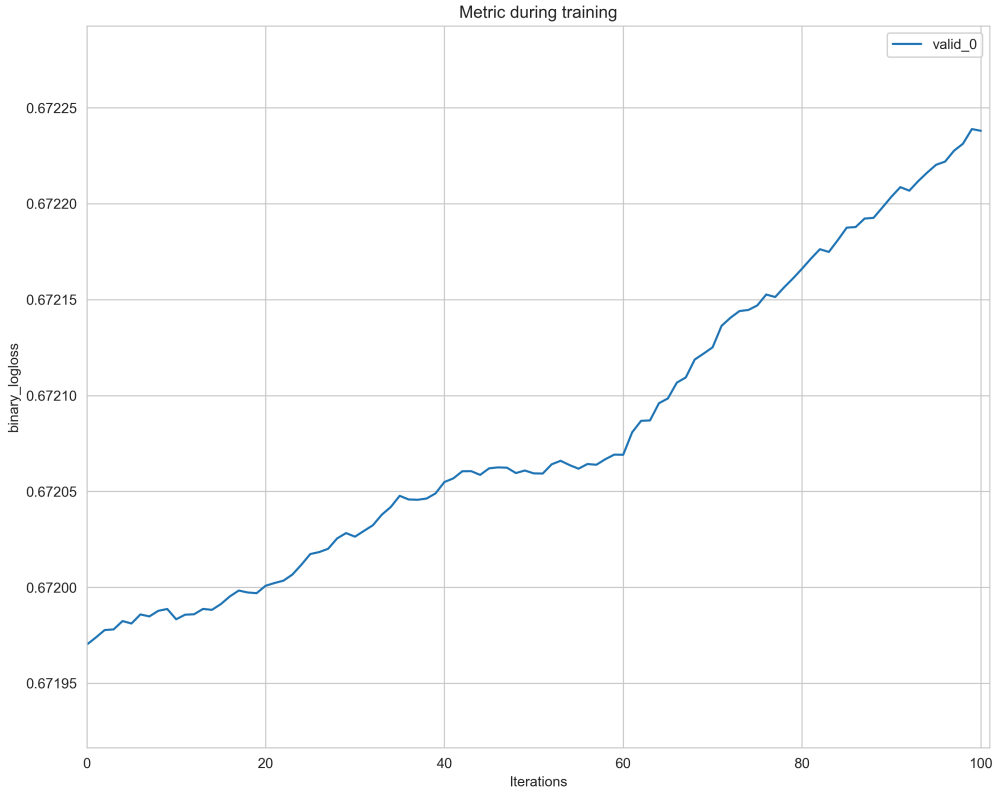

図2: PUR予測モデルの学習曲線

機械学習モデルの確率値が適切にキャリブレーションされているかを確かめるため、横軸に予測 CVR、縦軸に実測 CVR をとった散布図を図 3 と図 4 に示す。図 3 はクーポン A、図 4 はクーポン B の結果である。予測値と実測値に大きな乖離はないことがわかる。

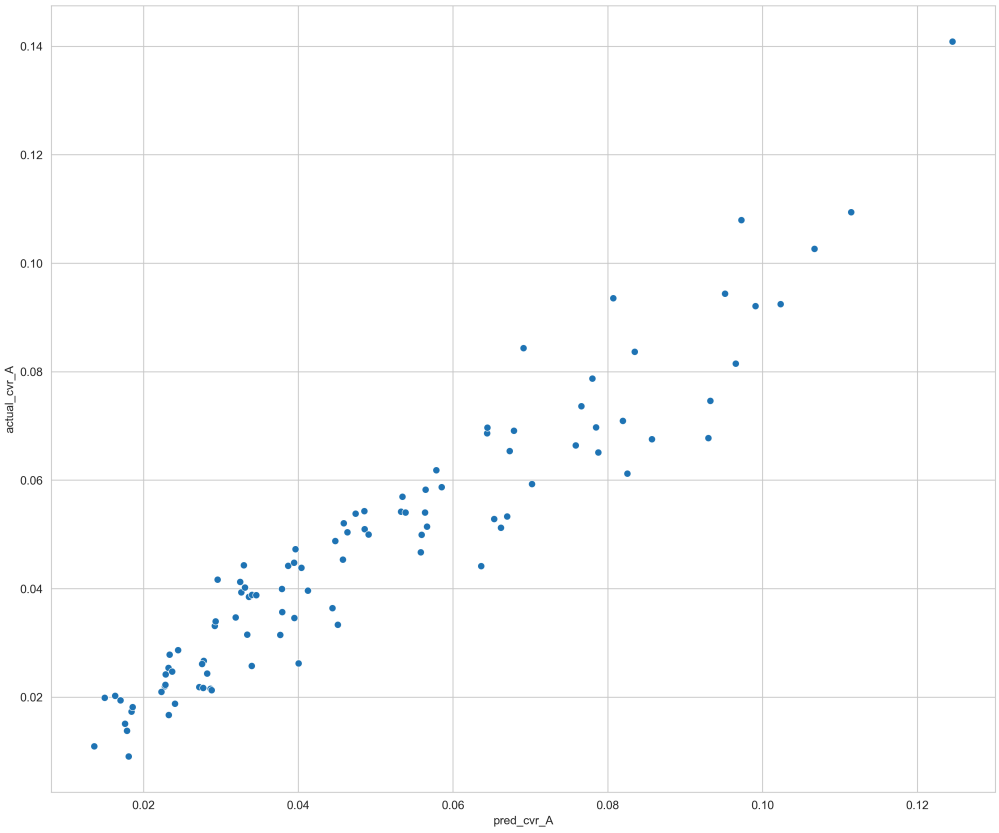

図3: クーポンAに対する予測CVRと実測CVR

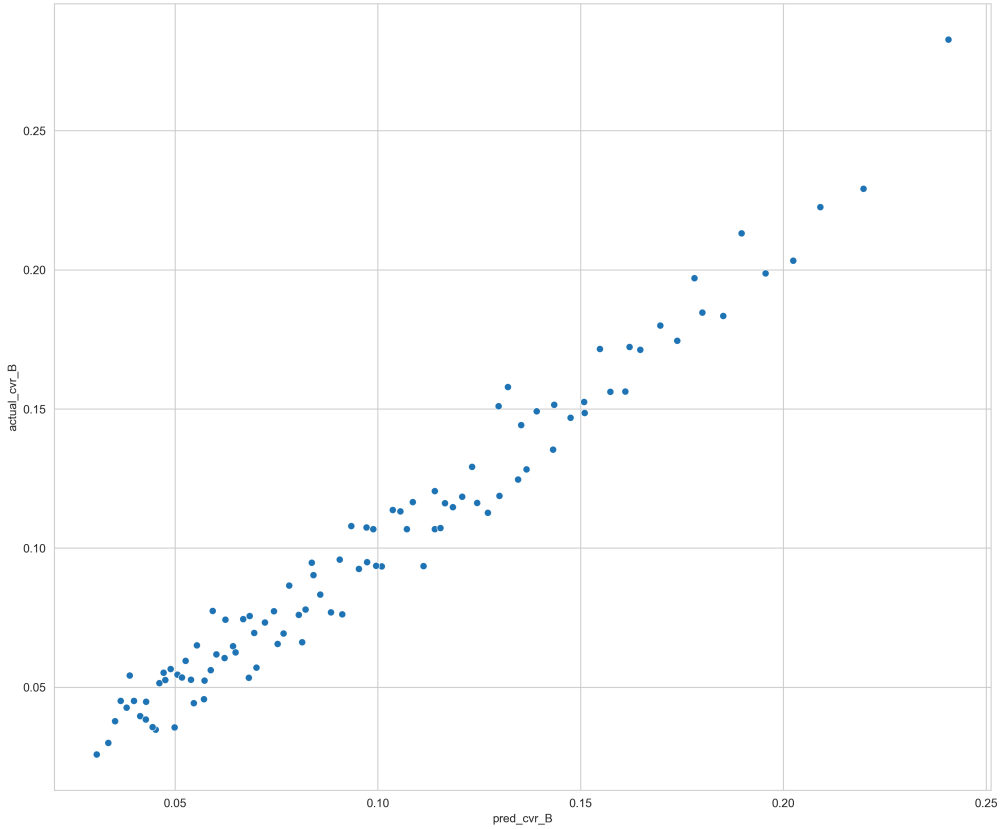

図4: クーポンBに対する予測CVRと実測CVR

PUR における同様の散布図を、図 5 と図 6 に示す。図 5 がクーポン A、図 6 がクーポン B に対する結果である。PUR は予測値が 0.6 付近に集中している一方、実測値は 0.4 から 0.8 程度までばらついている傾向が見える。PUR 予測モデルの学習がうまくいかなかったため、予測値と実測値に大きな乖離が生まれたと考えられる。

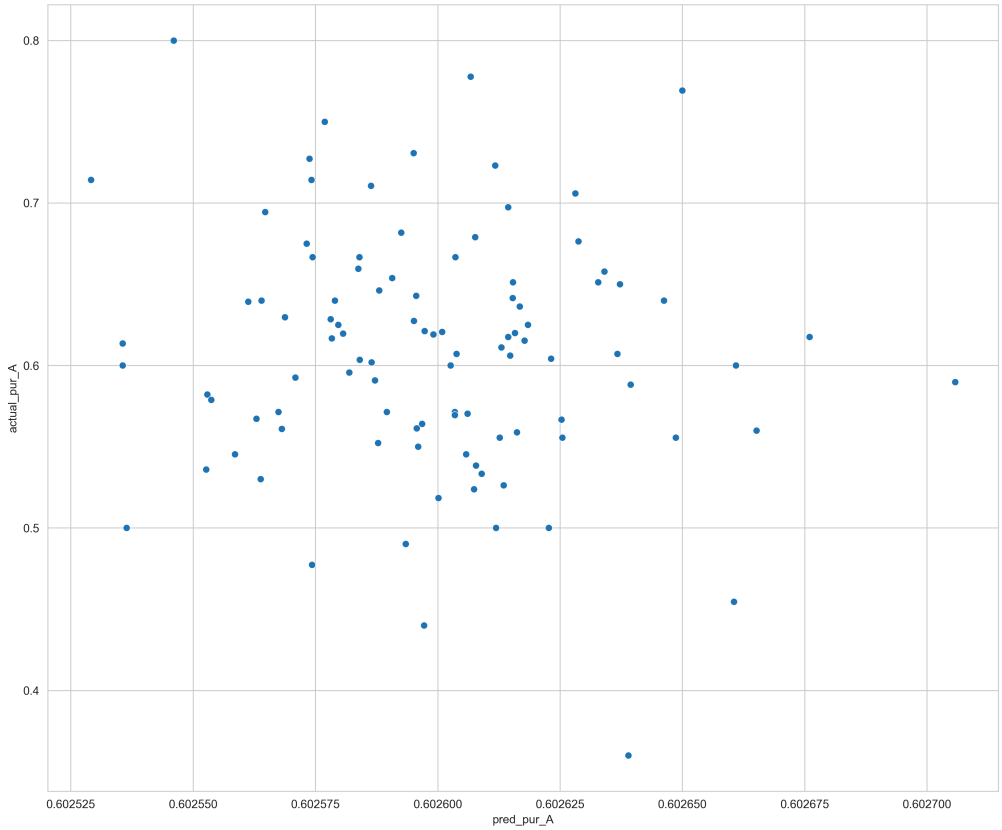

図5: クーポンAに対する予測PURと実測PUR

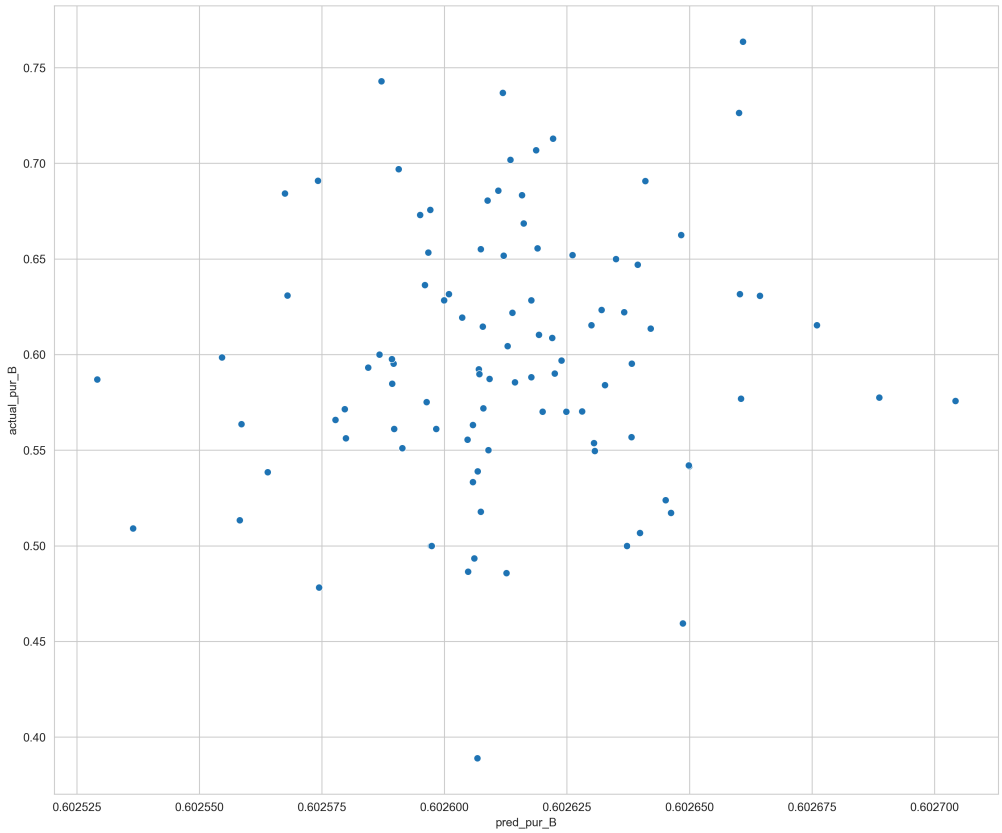

図6: クーポンBに対する予測PURと実測PUR

以上の結果より、CVR 予測モデルは採用し、PUR 予測モデルは不採用とした。PUR はログデータから求めた集計値をそのまま利用した。

コスト最小化を狙ったクーポン割り当て

横軸にクーポン B を全ユーザーに配布した場合の推定 CV 数からの棄損率、縦軸に CPA をとったグラフを図 7 に示す。クーポン B を全ユーザーに配布した時の CPA は 896 であり、CV 数の棄損を許容することによって CPA を小さくできることがわかる。
これより、クーポン 2 ではなくクーポン 1 を配布しても CV するユーザーが存在すると言える。

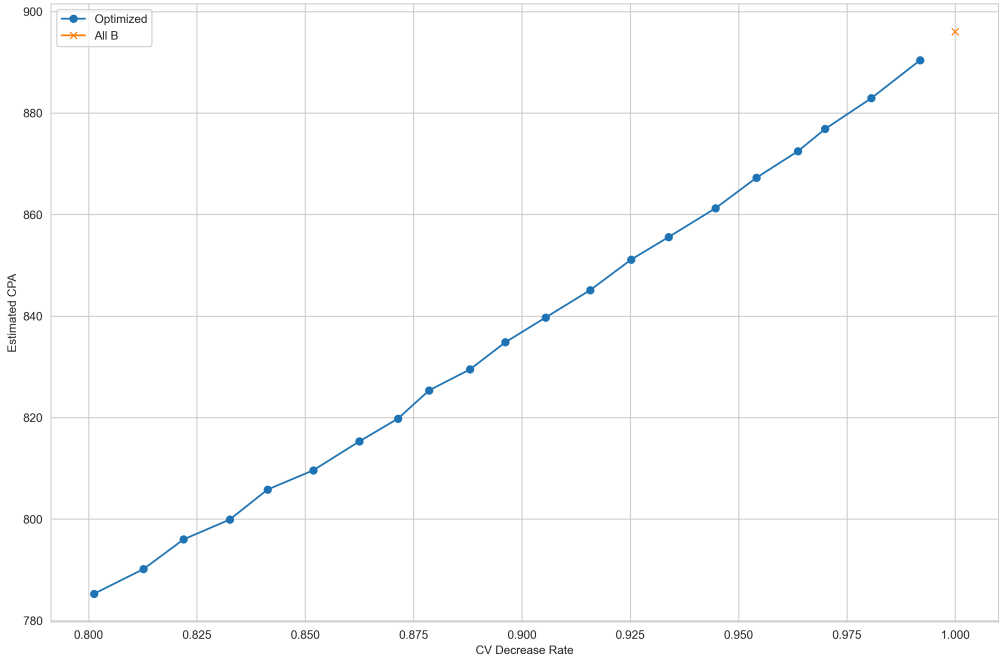

図7: 棄損率-CPA曲線

実験 2: 3 種類のクーポンを用いた CV 数最大化

背景・目的

実験 1 より、ある程度の CV 数の棄損を許容することで、CPA を下げられることがわかった。これより、クーポン 2 ではなくクーポン 1 を配布しても CV するユーザーが存在すると言える。

実験 2 では、クーポン 2 よりも割引額が高いクーポン 3 を導入し、予算制約下で CV 数の最大化を行う。

検討ロジック

検討したロジックは大きく次の 2 段階に分かれる。

1. CVR とポイント利用率（PUR; Point Utilization Rate）を予測する。
2. 予測値を用いて CV 数を最大化する数理最適化問題をソルバーで解き、ユーザーへの最適なクーポン配布を行う。

まず、今回解きたい問題を以下のように定式化した。

$$
\begin{aligned}
\text{Maximize} \quad & \sum_{s \in S} \sum_{a \in A} x_{s,a} N_s r_{s,a} \\
\text{s.t.} \quad & x_{s,a} \in [0, 1] \\
& \sum_{a \in A} x_{s,a} = 1,\ \forall s \in S \\
& \sum_{s \in S} \sum_{a \in A} x_{s,a} N_s r_{s,a} p_{s,a} C_a \leq B \\
& x_{s,a} \geq K,\ \forall s \in S, \forall a \in A
\end{aligned}
$$

表記は以下の通りである。

- S: ユーザーセグメント集合
- A: クーポン集合
- $x_{s,a}$: セグメントsへのクーポンaの配布率
- $r_{s,a}$: セグメントsにクーポンaを配布した時に CV する確率（CVR）
- $p_{s,a}$: セグメントsにクーポンをaを配布して CV した際にクーポンを利用する確率（PUR）
- N_s: セグメントsに属するユーザー数
- C_a: クーポンaを使って CV された際に発生するコスト
- B: 予算
- K: 最低送付率

この最大化問題を解くことで、セグメントsにクーポンaを何%配布するかを決めることができる。最適な$x_{s,a}$を得た後のユーザーに対するクーポン割り当ては、配布割合を用いて確率的に行なった。選択されたクーポンの配布確率をログに残しておくことで、そのログを用いたモデルの検討を行いやすくすることを狙った。制約条件に全てのセグメントにすべてのクーポンを少なくとも $100K$%割り当てる制約が入っているのも、この狙いのためである。

実験方法

ダミーデータを生成し、64%を学習データ、16%を検証データ、20%をテストデータとした。CV したかと CV した上でポイント利用したかの両方において分布に偏りを生じさせたくなかったため、データの分割には iterstrat を用いた。

CVR 予測モデルは、CV したかを目的変数とし、LightGBM を Binary Logloss で学習させることで得た。PUR 予測モデルは、実験 1 で有効性が確認されたなかったため、実験 2 では実装しなかった。

ユーザーセグメントは、次の 2 段階で作成した。

1. CVR 予測モデルによって得られたクーポン A 付与時 CVR をもとに、pandas の qcut で 5 分割。
2. 1.で得られたセグメントごとに、クーポン B 付与時 CVR をもとに pandas の qcut で 5 分割。
3. 2.で得られたセグメントごとに、クーポン C 付与時 CVR をもとに pandas の qcut で 5 分割。

セグメントごとのユーザー数に偏りが出ないこと、なるべく CVR が近いユーザーを同じグループにまとめることを狙った。また、実験 1 では各クーポンにおいて 10 分割していたが、実験 2 ではクーポンの数が増えるため、5 分割に抑えた。

予測値を用いた数理最適化は、pulp を用いて解いた。最低送付率は$K = 0.1$とした。最適な$x_{s,a}$を得た後のユーザーに対するクーポン割り当ては、配布割合を用いて確率的に行なった。

最適な割り当てを行なった際に生じる CV 数とコストの推定は、次のように行なった。

1. 最適な割り当てを行なった場合のクーポンと、ログ上のクーポンが一致しているデータを抽出。
2. 1.のデータを用いて、クーポンごとに CVR と PUR を計算。
3. クーポンごとの付与人数に 2.で求めた CVR をかけることで推定 CV 数を計算。
4. 3.で求めた推定 CV 数に PUR とクーポンの割引額をかけることで推定コストを計算。

結果

CVR 予測モデル

CVR 予測モデルの学習曲線を図 1 に示す。損失が下がっており、実験 1 と同様に学習が進んでいることがわかる。

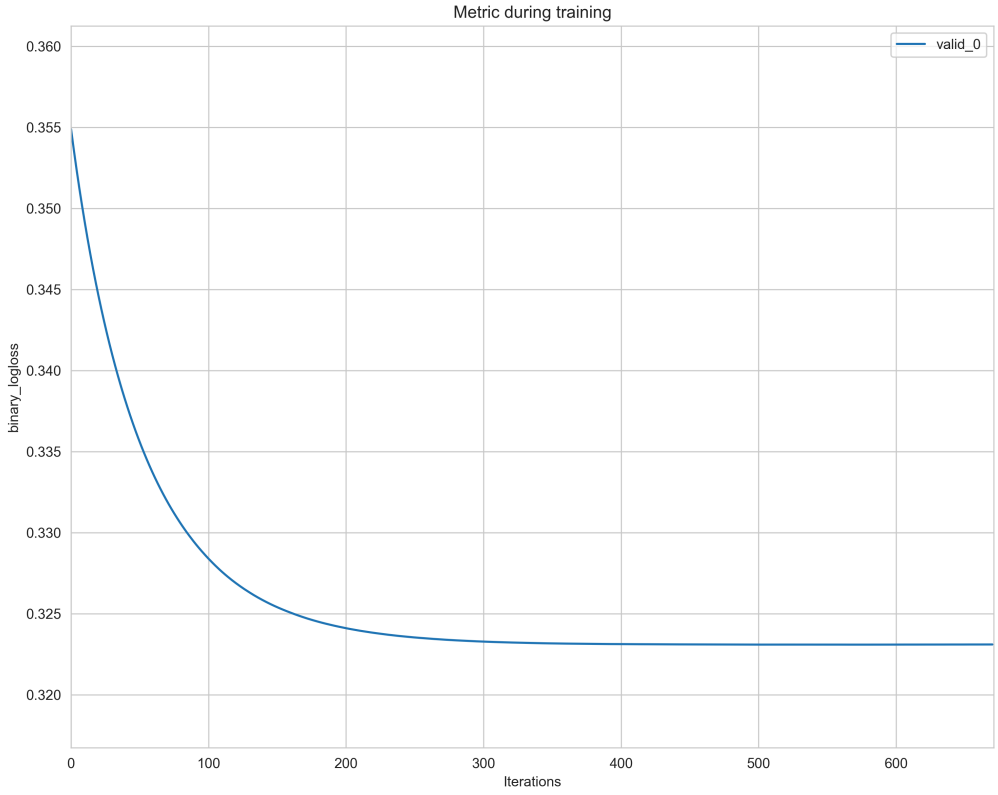

図1: CVR予測モデルの学習曲線

機械学習モデルの確率値が適切にキャリブレーションされているかを確かめるため、横軸に予測 CVR、縦軸に実測 CVR をとった散布図を図 8 と図 9 、図 10 に示す。図 8 はクーポン A、図 9 はクーポン B、図 10 はクーポン C の結果である。予測値と実測値に大きな乖離はないことがわかる。

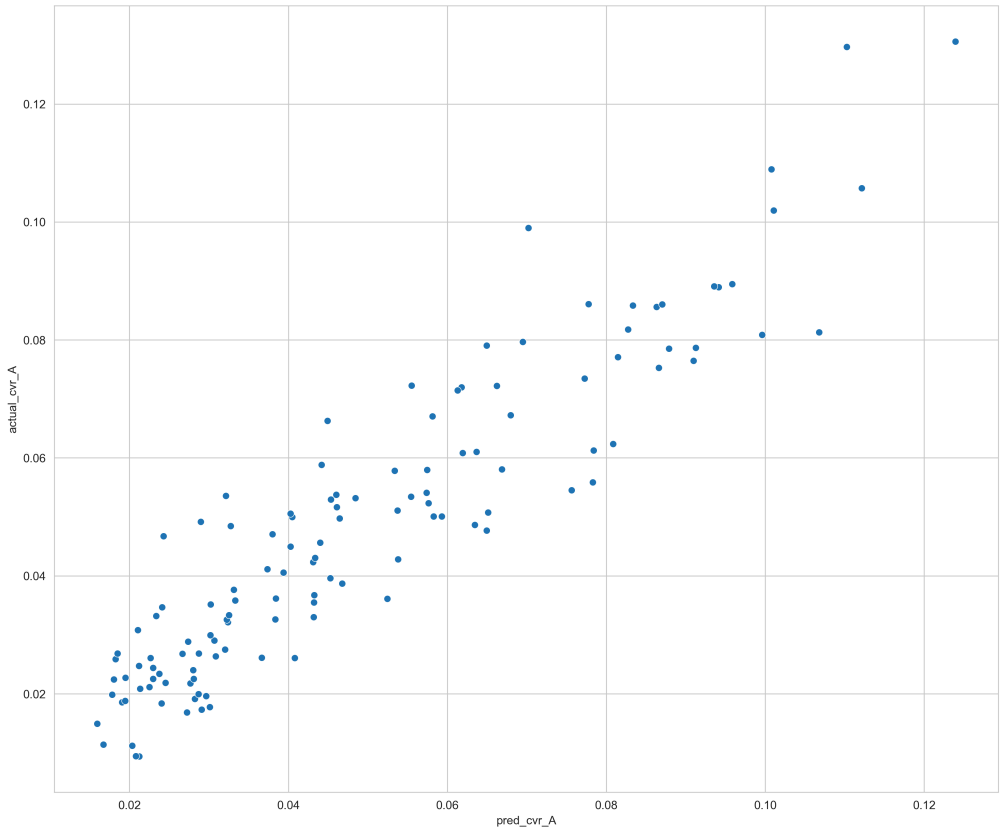

図8: クーポンAに対する予測CVRと実測CVR

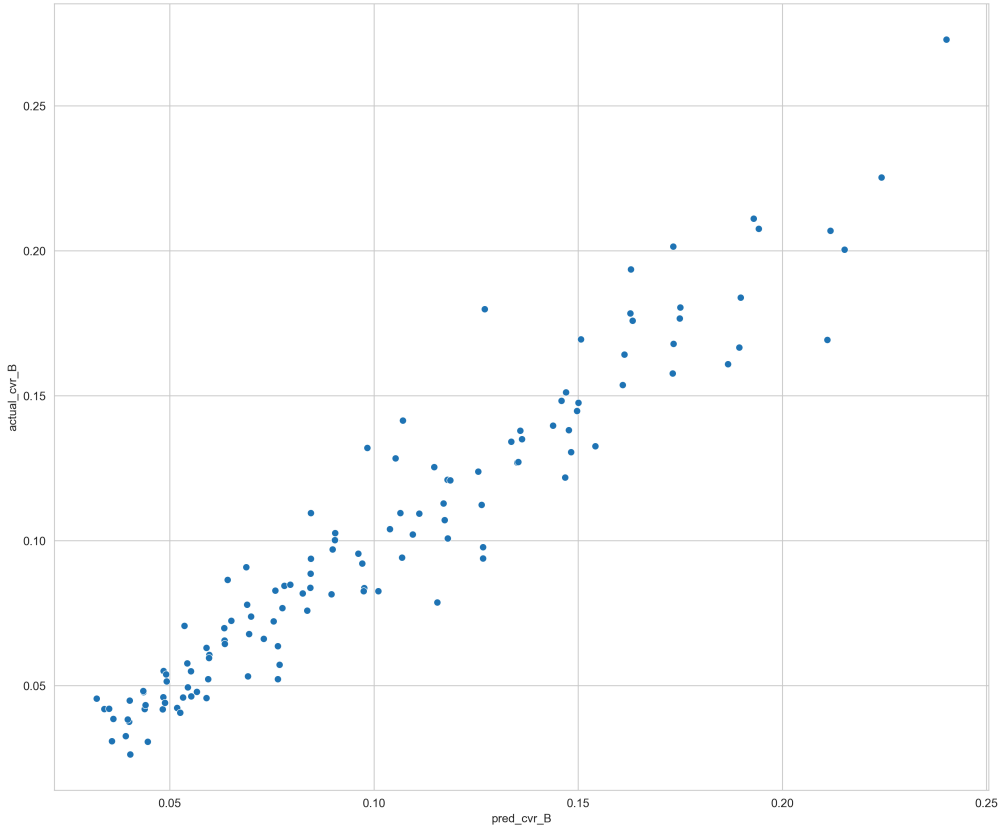

図9: クーポンBに対する予測CVRと実測CVR

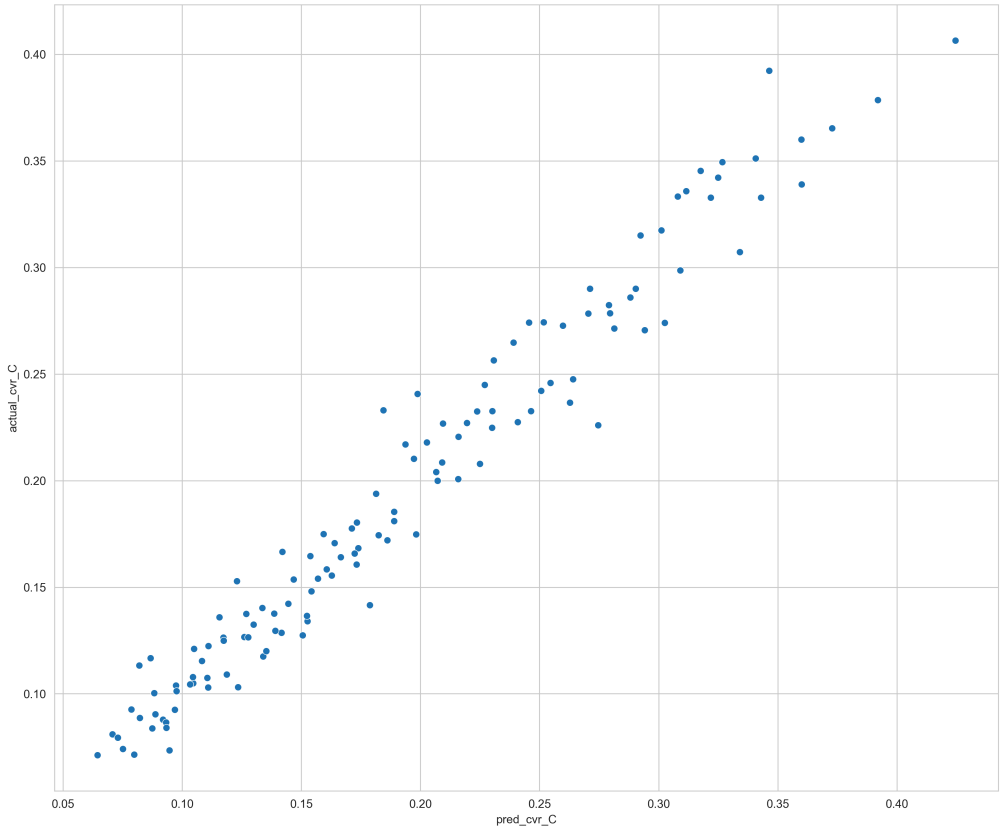

図10: クーポンCに対する予測CVRと実測CVR

CV 数最大化を狙ったクーポン割り当て

横軸にクーポン B を全ユーザーに配布した場合の推定 コストからのコスト増加率、縦軸に推定 CV 数をとったグラフを図 11 に示す。クーポン B を全ユーザーに配布した場合の推定コスト下で CV 数を増加させることができなかったことがわかる。

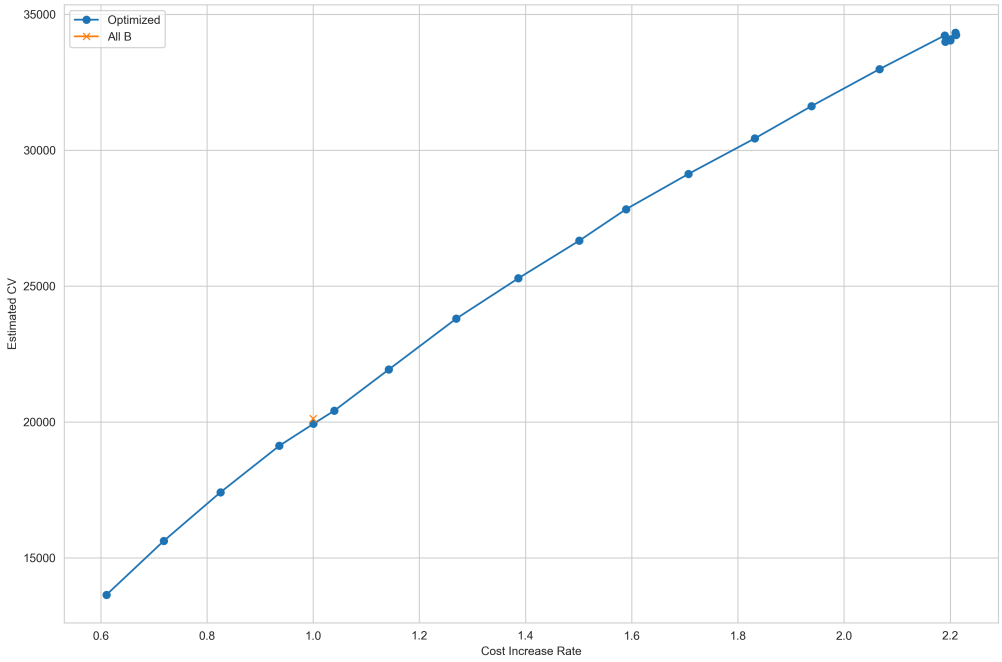

図11: コスト増加率-CV数曲線

この原因として以下のような点が考えられる。

1. 全てのセグメントにすべてのクーポンを 10%配布するという制約が厳しかった
   a. クーポン A を過剰に配布することで CV を取りこぼす
   b. クーポン C を過剰に配布することで余計なコストがかかる
2. クーポン C の効果が薄かった
   a. クーポン C にしても、クーポン B とさほど CV しやすさが変わらなかった可能性がある
3. クーポン A を割り当てて CV を獲得することが難しかった
   a. クーポン A を割り当てても CV するユーザーに正確にクーポン A を配布することができず、コストを浮かせることができなかったためにクーポン C の配布機会が少なくなった

原因 2 に関しては、クーポンの値段設定の見直しが必要なため本実験のスコープ外である。原因 1 については、制約条件を見直すことで解決できる可能性がある。原因 3 は、機械学習モデルの性能をあげたり、セグメント分割方法を見直したりすることで解決できる可能性がある。

追加実験 2-1: 制約条件の変更

最適化問題における最低送付率を$K = 0.01$とした結果を図 12 に示す。全ユーザーにクーポン B を配布した時の推定 CV 数が 20092 であるのに対し、最適化したクーポン配布はコスト増加率が

0.972589 で推定 CV 数が 20205.5 となっており、より少ないコストでより多くの CV 数を獲得できたことがわかる。

これより、制約条件を緩めることで CV 数を増加させられることがわかった。

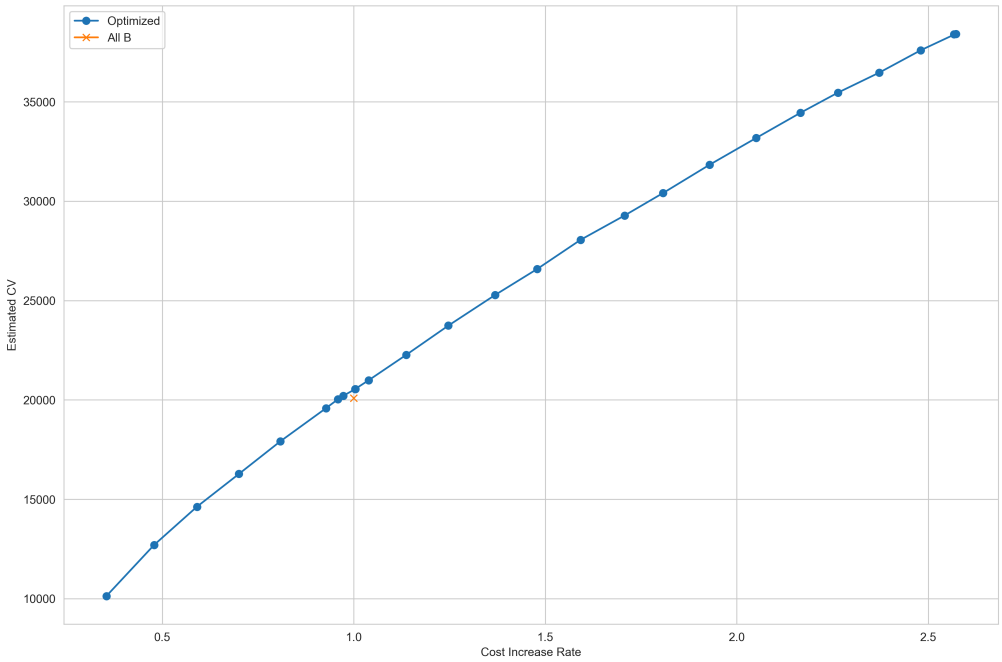

図12: コスト増加率-CV数曲線

追加実験 2-2: 機械学習モデルの性能向上

おそらく特徴量設計によってモデルの性能を向上するのが現実的だが、本実験ではダミーデータを用いているため実施しない。

追加実験 2-3: セグメント分割方法の見直し

セグメント数を 5 個から 6 個に増やした場合の結果を図 13 に示す。ここでは最適化問題における最低送付率を$K = 0.1$に戻した。全ユーザーにクーポン B を配布した時の推定 CV 数が 20092 であるのに対し、最適化したクーポン配布はコスト増加率が 0.988475 で推定 CV 数が 20161.4 となっており、より少ないコストでより多くの CV 数を獲得できたことがわかる。

これより、セグメント分割数を増やすことで CV 数を増加させられることがわかった。

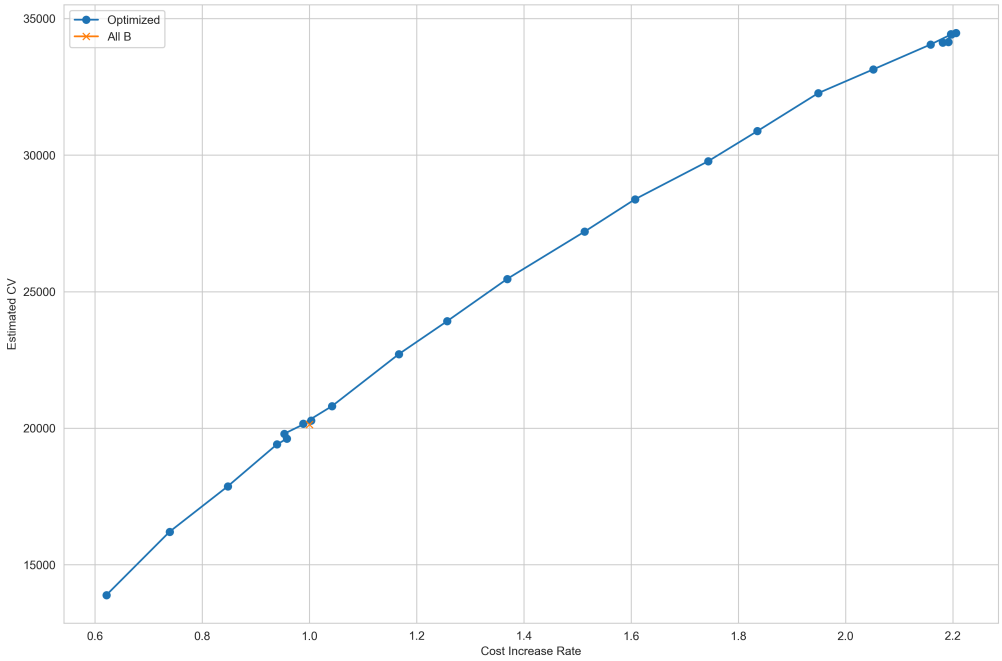

図13: コスト増加率-CV数曲線

まとめ

本稿では、クーポンの最適な割り当て問題について、2 パターンの場合におけるコスト最小化問題と、3 パターンの場合における CV 数最大化問題を検討した。

2 パターンの場合におけるコスト最小化問題において CPA を改善できたことから、3 パターンにした場合に同じ予算でより多くの CV 数を獲得できる可能性があると見立てて、実験を進めた。

3 パターンの場合における CV 数最大化問題では、クーポン 2 を一律配布した場合における推定コスト以内で推定 CV をより多く獲得する割り当て方を実現できるかに焦点を当てた。結果、初めに検討したアプローチでは CV 数を増やすことができなかったが、ログ収集を考慮した制約における最小配布率を小さくしたり、セグメント数を増やすことによって CV 数を増やすことができた。また、本稿ではダミーデータを用いたため実施しなかったが、特徴量設計による機械学習モデルの性能改善も有効な可能性があることを提案した。