最終課題

新規事業提案

最終課題　目次

- 市場分析
- A社課題
- 損失額の根拠
- 特徴量の選定
- EDA（探索的データ分析）
- クラスター分析
- モデルの選定・評価
- 効果の定量評価
- 事業案の提示

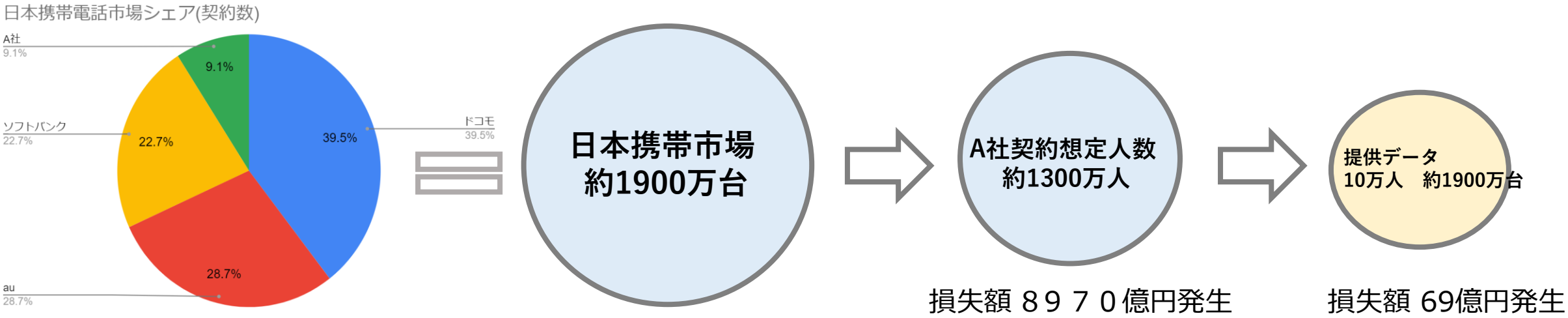
参考資料
https://www.kaggle.com/datasets/abhinav89/telecom-customer
https://speakerdeck.com/tom_uchida/gci-2020-winter-zui-zhong-ke-ti
https://medium.com/analytics-vidhya/using-machine-learning-models-to-predict-customer-turnover-7d52db91d459
GCI教材

市場分析　通信業界

- A社は、電気通信業で、電気通信サービスを提供していると想定
- 通信業界には、固定電話やパソコンの通信サービスを行う固定通信、携帯やスマートフォンなどのモバイル通信サービスを行う移動通信、インターネット接続サービスを提供するインターネットサービスプロバイダーがある。
- 移動系通信契約数における事業者別のシェア推移によると、日本の移動系通信の事業者別シェアは、ＮＴＴドコモ、ＫＤＤＩグループ、ソフトバンクグループの順で、トップ数社で独占。
- 2019年、改正電気通信法の施行により、携帯電話の販売方法にメスを入れ、**料金の価格競争**を促す
- 総務省のICT市場の動向によると、2020年の**電気通信事業の売上高は約15兆円前年比2.5%増と増加**。
- 固定通信の契約数は落ち込んでいるが、**移動通信契約は、固定契約数の約12倍に伸びている。**
- 固定通信よりも、**移動通信契約にチャンス**がある。
- 業界動向リサーチによると、テレワークに伴うクラウド需要の増加など、**業務効率改善やDXの推進の需要が大きく高まって**おり、Saas業界の売上高は急成長している（Saas：Softweare　as a Service）
- 通信業界の課題は、Iotの拡大、通信需要の増加に対するセキュリティ対策、光回線が通っていない地域への対応、**カスタマーサービスの需要の増加**によるカスターマーサービスの質の向上
- 日本においては、**生産年齢の人口減少が毎年進んでおり、業務効率化などデジタルを利用した生産性向上は、すべての企業の課題であり、今後も需要が高まる**と考えられる。

参考
https://www.soumu.go.jp/johotsusintokei/whitepaper/index.html
https://www.digital-transformation-real.com/blog/current-status-and-challenges-of-the-telecommunications-industry.html
https://iroots-search.jp/14609

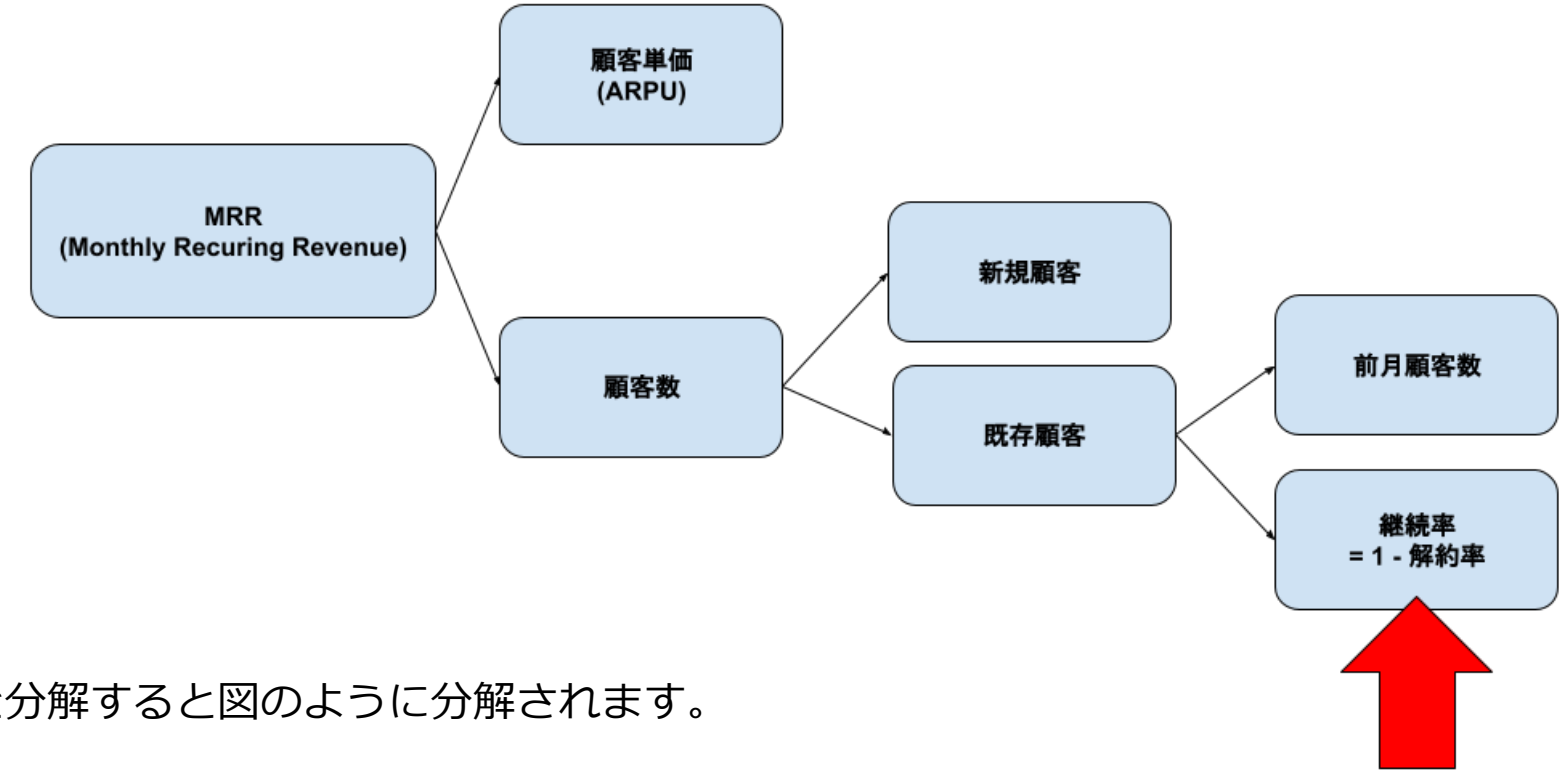
A社課題



- 市場分析により、移動系通信にチャンスがあると考えるため、日本の携帯電話市場をターゲットにします。
- 日本の携帯電話契約数台197,650,000台　https://www.tca.or.jp/database/
- 契約台数による日本市場シェアをドコモ40%, au30%, ソフトバンク20%, A社 1%とします。
- A社の契約台数 19,765,000台(197650000 × 1%)
- A社の契約数；1人1.5台契約しているとすると、1900万 ÷ 1.5 = 1300万人
- 提供データ10万人は1300万人から無作為抽出されたとします。
- A社の損失額が**53,663,564ドル**。130円/1ドルとすると**6,976,263,349円(69億円)。**
- **A社の契約者全体8970億円の損失があります。**
- **解約者を1%減らすことができれば、89億円の損失を防ぐことができます。**

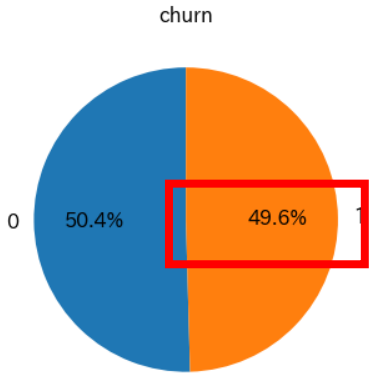
参考 https://speakerdeck.com/tom_uchida/gci-2020-winter-zui-zhong-ke-ti

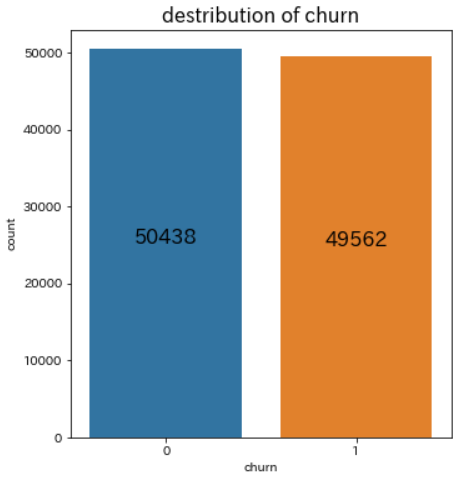
損失額の根拠



- 収益＝売上を分解すると図のように分解されます。
- 売上は、顧客単価と顧客数に分解されます。
  顧客数は、新規顧客、既存顧客に分解され、既存顧客は前月顧客数に継続率を掛けた値です。
  継続率は、１－解約率で計算されます。
- よって、解約率を下げることで既存顧客が上がり、顧客数が上がり
  売上が上がるため、収益と解約率に注目します。

損失額の根拠





A社の解約非解約別の収益

churn	total_revenue
0	55854382
1	53663564

機械学習モデル

	予測確率	pred
0	0.509348	1.0
1	0.445280	0.0
2	0.548358	1.0
3	0.445179	0.0
4	0.455829	0.0
...	...	...
19995	0.371046	0.0
19996	0.485515	0.0
19997	0.444285	0.0
19998	0.457509	0.0
19999	0.444235	0.0

20000 rows × 2 columns

予測確率が50%より大きいと１(解約)と判定

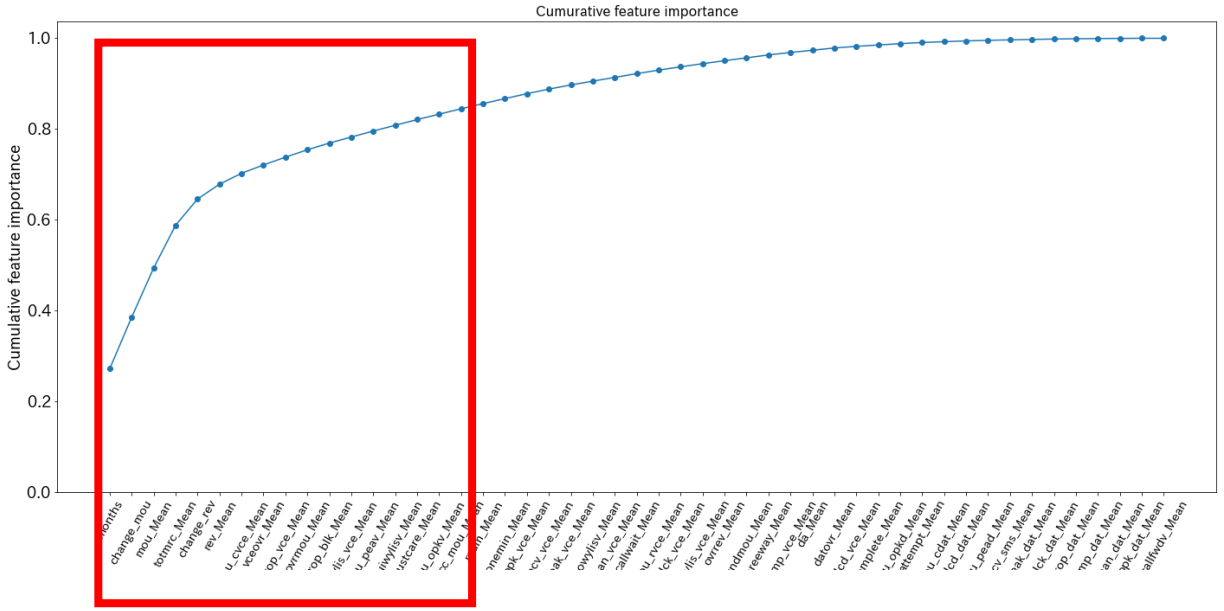
- Rev_Mean(1カ月平均収益)×months(サービス契約月数) = Total_revenue(総収益)と考えると、
  解約非解約別の収益より、**53,663,564ドル（69億円）**の損失があることが分かりました。

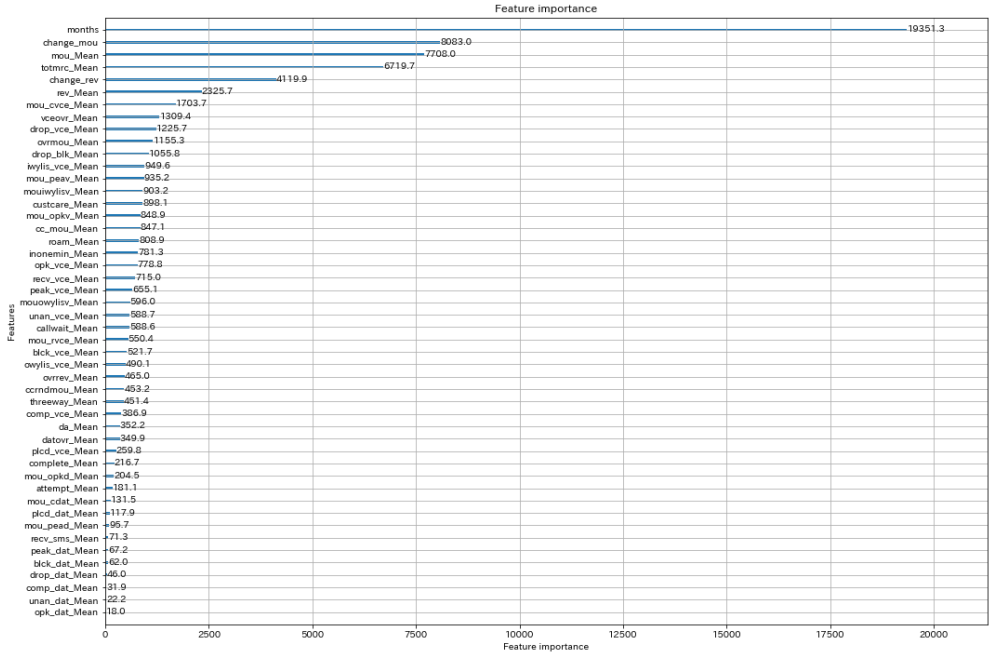
- Recordデータのみの損失額を考える場合、は以下です。
  機械学習で解約かどうかは、実際の解約の割合が50%程度であることを考え、予測確率が上位の50%より小さいかどうかで判定した場合、
  離反と予測した人の売上は**15,221,787ドル**でした。
  訓練データ・テストデータの分割(8割と2割)を考慮すると、これらがほぼランダムに分割できているとすると、
  19,027,233ドル（15221787÷0.8）(24億円)の損失があると言えます。機械学習を用いて離反を防ぐことができれば、これだけの損失を回避することができます。

- recordテーブルには、解約非解約の特徴量があり、目的変数をchurnカラム(解約か非解約)とし、
  解約か非解約を判定する機械学習モデルで解約有無を判別します。

- clientテーブルには、顧客の属性や行動規範などの特徴量があり、クラスター分析によって、
  ターゲットとする優良顧客の特徴の抽出を試みました。

特徴量の選定





- 機械学習モデルLightBGMより特徴量の重要度を見ると、上位20カラムで、約80%を占めています。
- よって重要度上位カラムの**months(顧客のサービス利用月数), change_mou(直近利用時間の変化率), mou_Mean(1カ月平均使用時間), change_rev(1カ月平均収益), rev_Mean(直近収益の変化率)**が顧客の解約か非解約の有無への影響が強いことが分かりました。

優良顧客(非解約顧客)の仮説

- 解約非解約顧客ごとに各特徴量の中央値を比較した表が右にあります。
- まず中央値で比較して、顧客の特徴をつかみました。下記が顧客の特徴です。

優良顧客(非解約顧客)は、
- サービスの契約月数の中央値が16カ月で、解約顧客よりも１カ月少ない
- 平均音声通話超過料金が0.42ドルで、解約顧客よりも0.5ドル少ない
- 3カ月平均使用時間に対する1カ月の使用時間の割合（change_mou）が-3で、解約顧客-10の約3倍大きい。直近の使用時間が長い。
- 平均使用時間(mou_Mean)が、380分で解約顧客329分よりも51分多い。
- 月額使用料金(totmrc_Mean)が45ドルで解約顧客43ドルより大きい。
  Total monthly recurring charge (MRC)：企業が顧客に毎月自動的に請求する金額
- 3カ月平均収益に対する1カ月平均の割合（change_rev）が-0.29で解約顧客-0.31よりも約0.03大きい。
- 平均収益(rev_Mean)48ドルは、解約顧客47ドルよりも約１ドル大きい。
- 完了した音声通話の平均使用時間 (分)(mou_cve_Mean)が多い。最後まで通話した時間が長い。

- 重要度の高い順で解約の有無に影響があると考え、右の仮説を念頭に置き基礎分析(EDA)を次のページで行いました。

解約非解約別の各特徴量の中央値の比較で非解約の値が小さい表

	churn_1_median	churn_0_median
ovrmou_Mean	3.500000	2.250000
ovrrev_Mean	1.200000	0.800000
vceovr_Mean	0.962500	0.437500
churn	1.000000	0.000000
months	17.000000	16.000000

※黄色は緑より高い値を表しています

解約非解約別の各特徴量の中央値の比較で非解約の値が大きい表

	churn_1_median	churn_0_median
change_mou	-10.000000	-3.000000
mou_Mean	329.750000	380.500000
totmrc_Mean	43.417500	44.990000
change_rev	-0.315000	-0.292500
rev_Mean	47.490000	48.876250
mou_cvce_Mean	132.706667	159.876667
drop_blk_Mean	5.333333	5.666667
iwylis_vce_Mean	1.666667	2.333333
mou_peav_Mean	107.045000	123.201667
mouiwylisv_Mean	2.473333	4.060000
mou_opkv_Mean	66.401667	85.896667
inonemin_Mean	11.666667	13.333333
opk_vce_Mean	31.333333	37.666667

仮説

優良顧客は、
- 使用時間(分)が長い
- モデルチェンジを好んで、早期に新機種やモデルを好む

優良顧客(非解約顧客)の特徴の分布

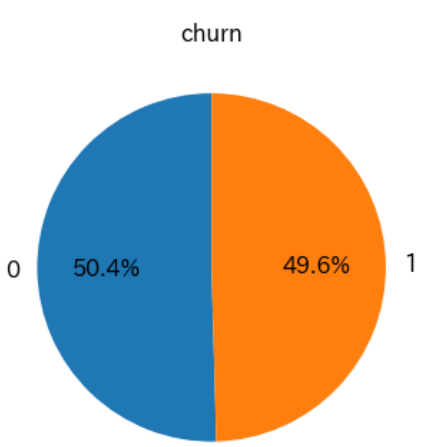
- recordテーブル概要
    1. データベース ：telecom(sqlite3ファイル)
    2. テーブル：record(カラム50件,100,000行)
    3. 目的変数をchurnとする。

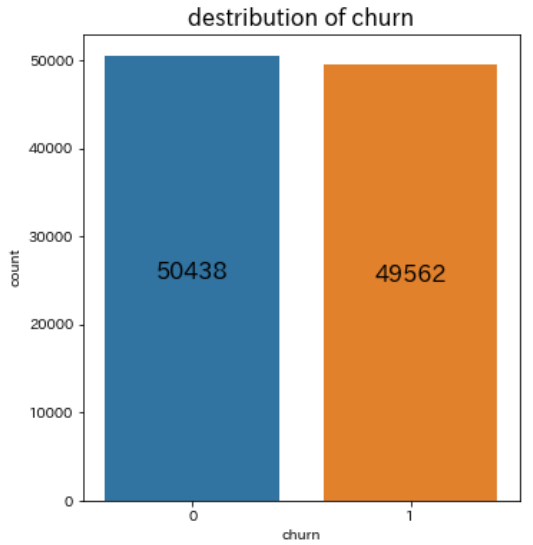
※機械学習モデルLGBMCClassifierの上位トップ5の重要度であるカラムについて記載

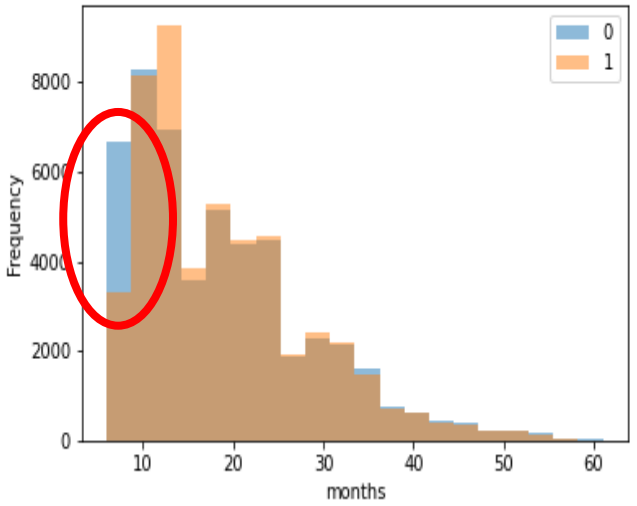
- カラム「churn」：顧客の離反有無
  - 離反1,離反していない0の割合は、50.4%、49.67%で、ほぼ均衡データである

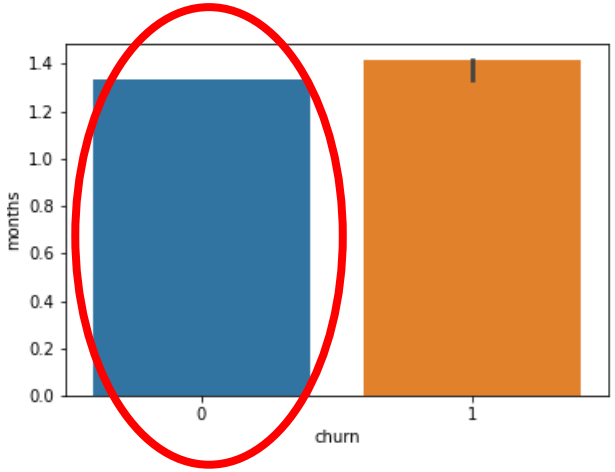
- カラム「months」：顧客の契約月数
  - 重要度1位で一番目的変数churnに寄与している変数
  - 右に裾が歪んでいる分布で、**非離反顧客は、離反顧客に比べて、10日以下の人数が多い。**

  - **優良顧客は、契約月数の中央値が1カ月短い。**
  - 解約顧客は、契約期間の月数が長い









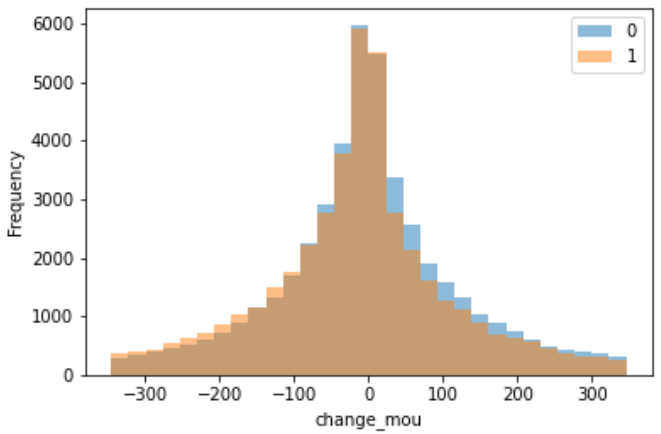
※ウィルコクソンの順位和検定（母集団の分布が正規分布を仮定しない検定）を有意水準を5%として検定し、2群の中央値に有意差がある。有意水準とはめったに発生しない確率のことでめったにないことが起きた、つまり偶然出ないことが起きた＝差があると考えます。

優良顧客(非解約顧客)の特徴の分布

- カラム「change_mou」：3カ月平均に対する1カ月平均使用時間（分）の割合（直近使用時間の変化率）
  - 3875, 31219の外れ値を除外すると、正規分布に近い

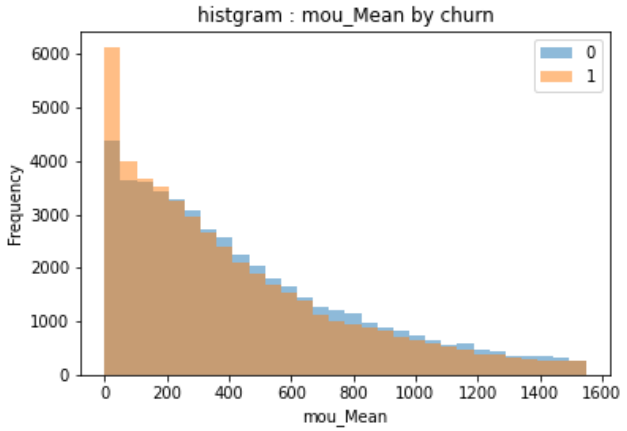
- 解約有無別の中央値は、解約している人の方が-10でそうでない人の約5倍低い。

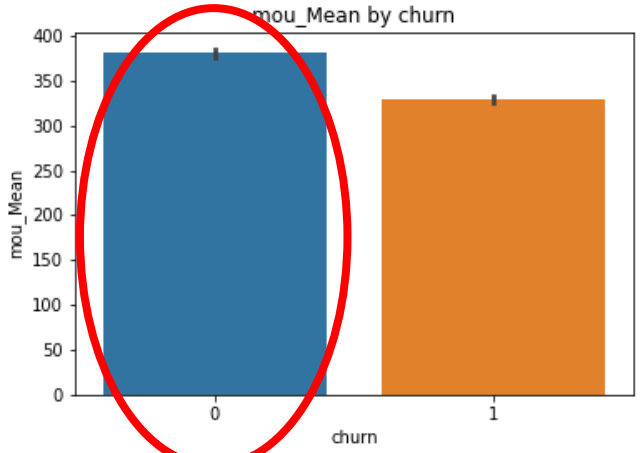
- **<u>優良顧客は直近使用時間の変化率が、解約顧客よりも小さい。＝直近での使用時間が多い。</u>**



- カラム「mou_Mean」：1カ月平均使用時間(分)
  - 12206の外れ値を除外

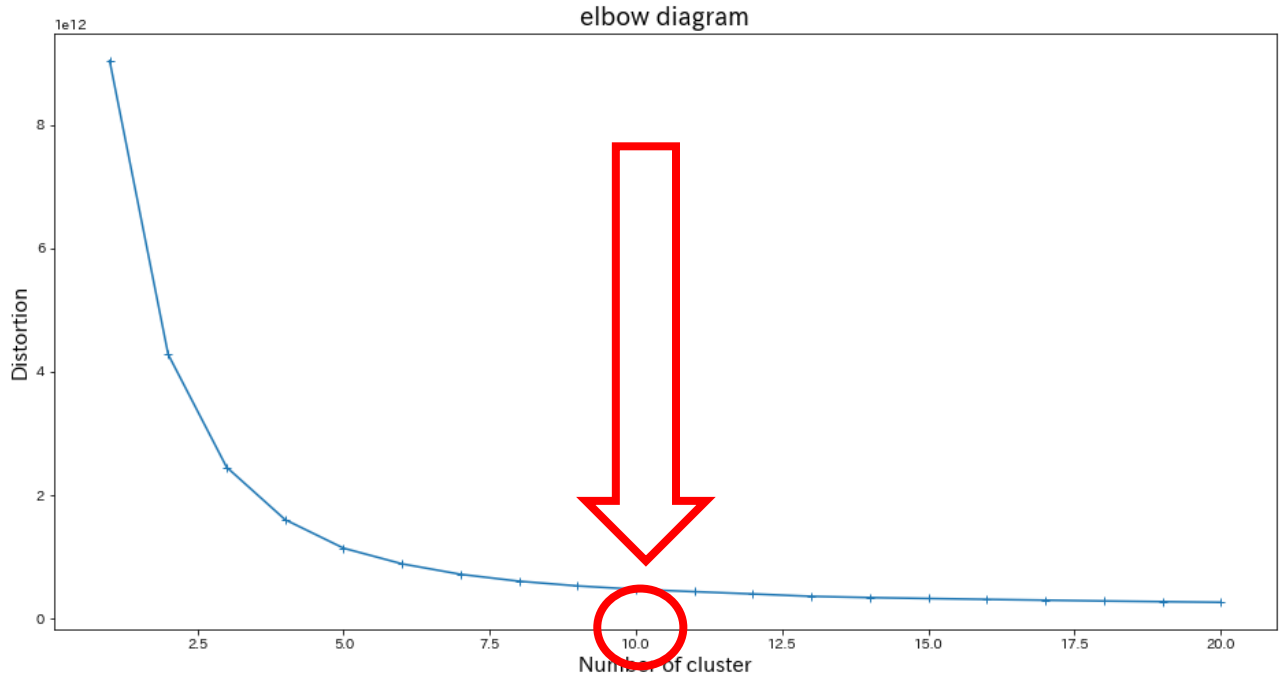
- 解約顧客は、１カ月平均使用時間が少ない
- **<u>優良顧客は、1カ月平均使用時間が解約顧客よりも長い</u>**

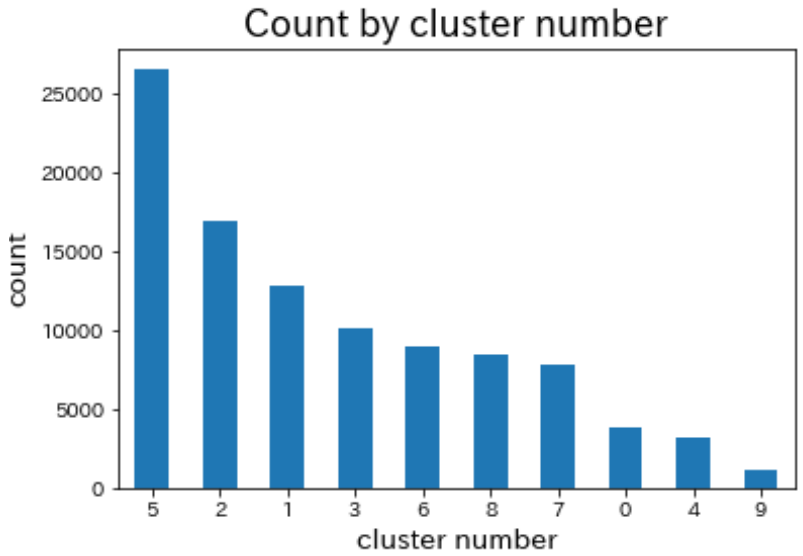




※ウィルコクソンの順位和検定を有意水準を5%として検定し、
2群の中央値に有意差がある（母集団の分布が正規分布を仮定しない検定）

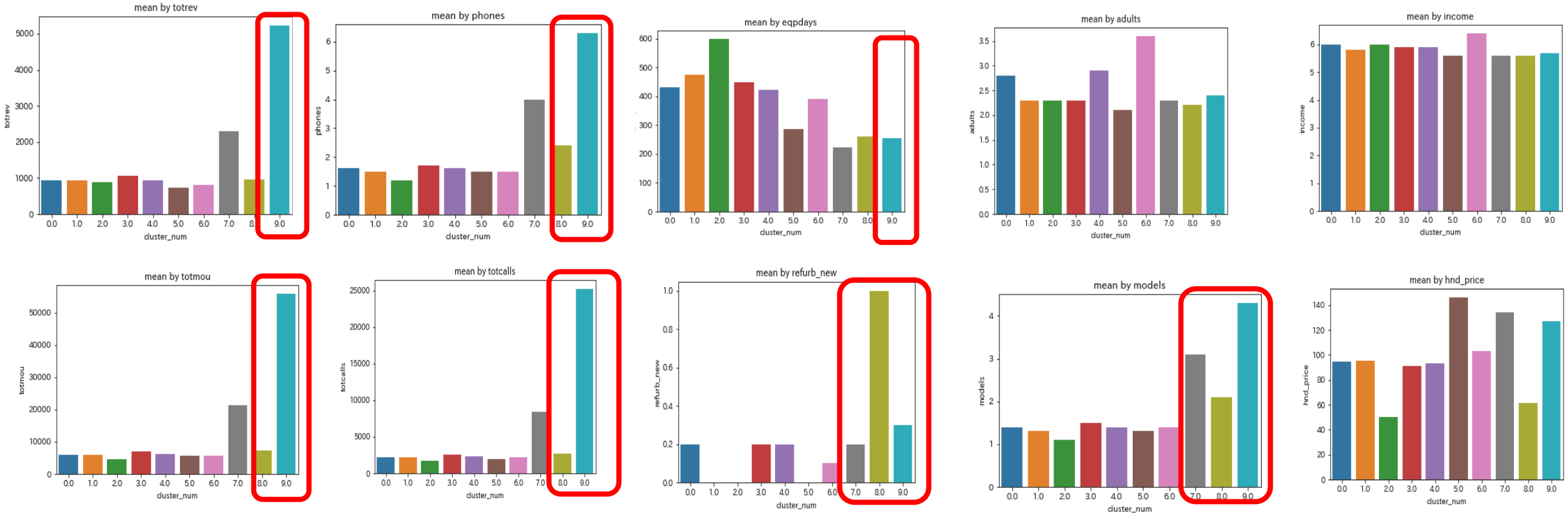
優良顧客(非解約顧客)の特徴　クラスター分析





- Kmeansというアルゴリズムで各変数間の距離を算出して距離の誤差を最適化して分類する方法です。
- エルボー法といって、複数のクラスター数を設定して、それぞれKmeansアルゴリズムで距離の総和を出力して
  最適なクラスター数を探しました。上記より10個で距離の総和が低くなりましたので、10個のクラスターを作成しました。

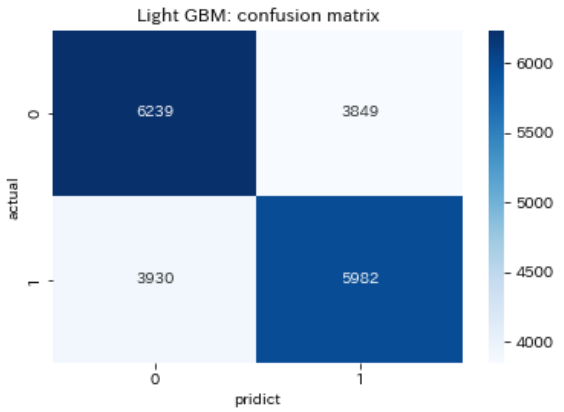
優良顧客(非解約顧客)の特徴　クラスター分析

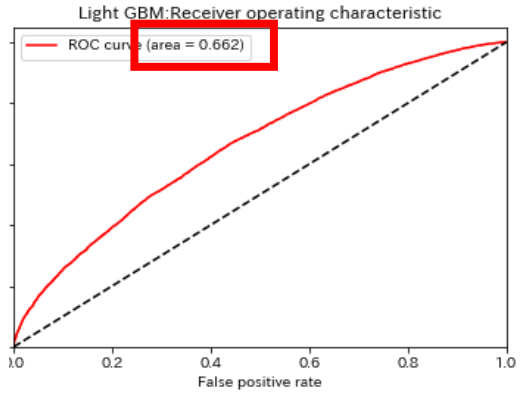


- 各クラスターと収益(total_price)と関係ある特徴量の平均値のグラフです。横軸にクラスター番号をとり、タテ軸に各変数の平均値をとります。
- **クラスター7,９は、収益、使用時間が大きいグループ。機器の価格、発行された携帯数、モデル数が大きい**一方で、**現在使用機器の期間は短くなっていました。**
  よって、**モデルチェンジや新機種志向で短期間で買い替える可能性の高い優良顧客グループ**と考えられます。また現在の使用機器価格も高く、価格が高い機種に対しても機種変更の可能性があると考えられます。
- クラスター4,6は、収益、使用時間が小さいグループ。リファービッシュ品（中古品）の携帯の顧客も若干存在しています。収入が一番高く、世帯の大人人数と子供の人数が高く、大家族グループと考えられます。価格は100ドル付近で、全体に対して中程度の高さです。よって、ファミリー層向けにも、モデルチェンジや新規携帯を試していただき、継続的に利用いただくことで、使用時間を引き上げる可能性が高まると考えます。

モデルの選定・評価

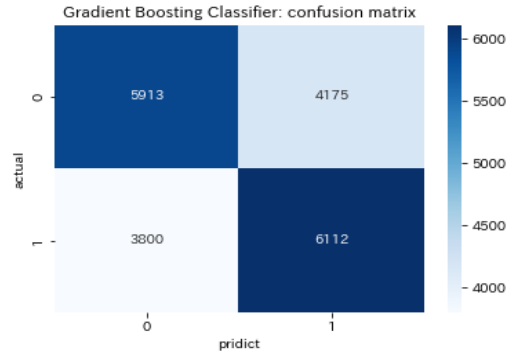
<採用した機械学習モデル>

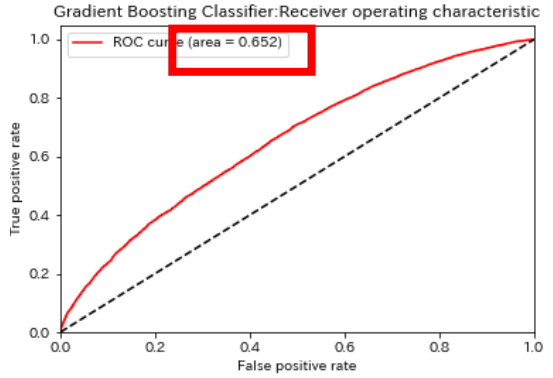


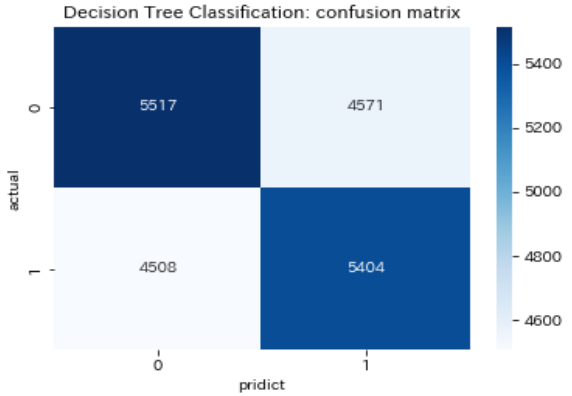


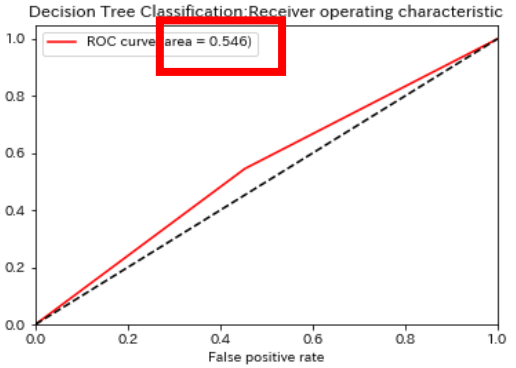
- A社の課題の顧客離反による売上損失について、顧客の離反有無の予測モデルを作成しました。
- 訓練データ、テストデータに対して、ホールドアウト法により右のように複数モデルを検討しました。
  上図のLightBGMのAUCスコア(0.662)が、一番高かったためこのモデルを採用しました。(LigbhtGBM,DecisionTree, AdaBoost, RandomForest, GradientBoostの5モデル)
- AUC(area under curve)とは、-1から１の値をとり、値が大きいとモデルの性能が高いと評価します。正解率だけだと、標本分布が負例や正例に偏っていると、正解率が高く出てしまい、モデルの性能がうまく図れないためです。

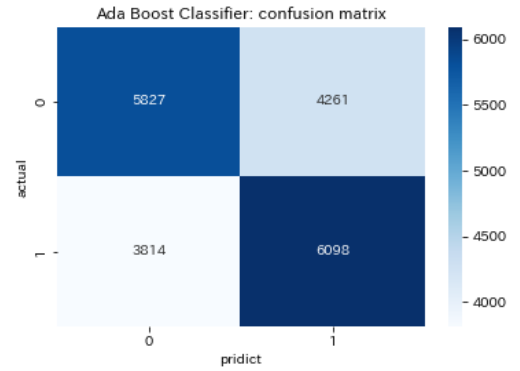
<採用しなかった機械学習モデル>

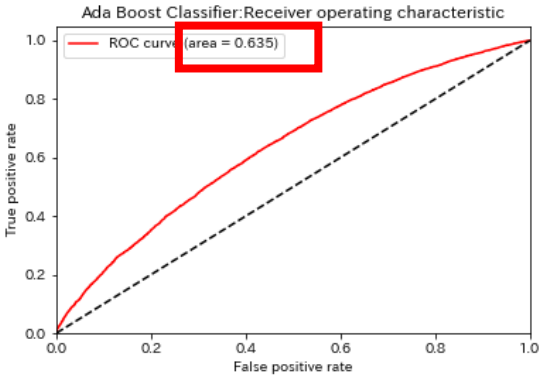




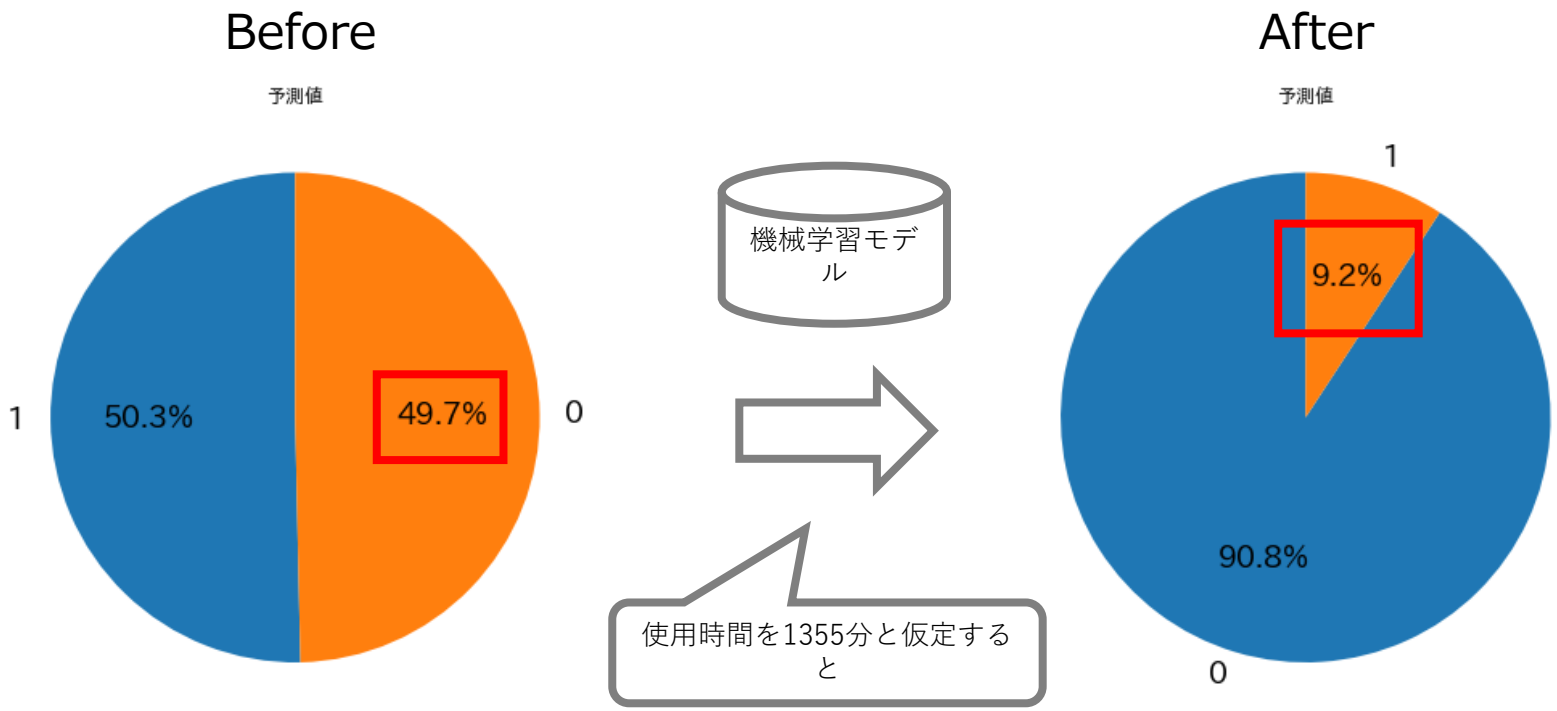








効果の定量評価



- 仮に解約顧客に対して、モデル数や発行された契約数アップによって、
  クラスター7，9(優良顧客グループ)の使用時間の平均値1355分に使用時間を引き上げれたと仮定すると、
  機械学習モデルによる予測値の分布をみると**解約率が49.7%から9.2%に減少**します。
- 解約比率を40.5%減ると1%あたり89億の損失を減らせることが冒頭よりわかりましたので、このモデルによると
  本施策を実施することで、**約3600億円の損失を回避する**ことができます。
- よって、クラスター７、９の優良顧客グループの使用時間をさらに引き上げ、解約率を下げて、収益の損失額を下げるために
  新機種、モデルチェンジ事業を提案します。

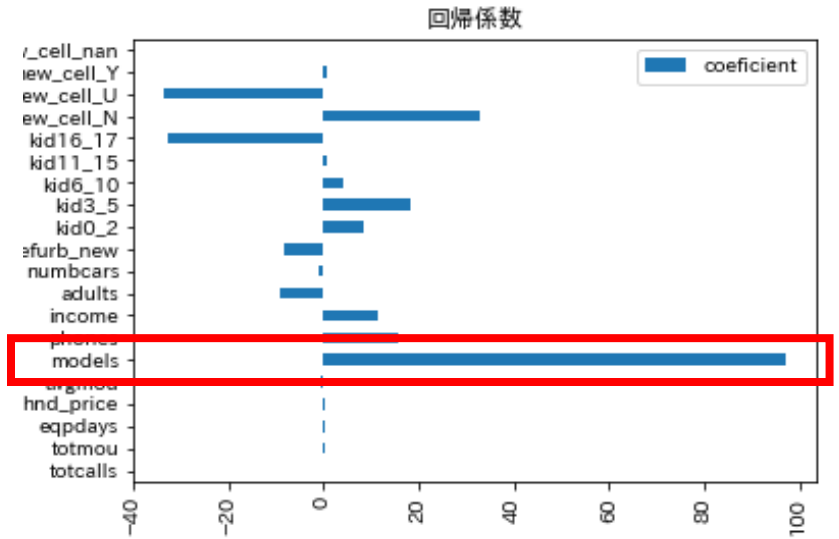
参考　https://speakerdeck.com/tom_uchida/gci-2020-winter-zui-zhong-ke-ti

提案

・クラスター7，9のグループ（使用時間、収益大のグループ）に、
モデルチェンジ、新機種の変更が多い顧客の使用時間が多いことより、
モデルチェンジ、新機種変更の1カ月間の間に、レンタルできるサブスクリプション型の事業を提案します。

・クラスター７，９の優良顧客グループは、モデルチェンジや新機種の感度が高いと考えられるため、
複数の携帯をお試しいただくことで、お客さまの満足につながり、収益アップにつながると考えます。

- 収益を目的変数にして、右図の特徴量を説明変数として
  重回帰分析を行った場合、モデル数の回帰係数が97でした。
  これは、発行されたモデル数が1増えると収益が97ドル増える
  ということを意味しており、モデル数が収益に関係があり、
  モデル数を引き上げることにチャンスがあると考えます。



R^2決定係数：76.1%
モデルの説明力を表し、100％になるほどモデルの妥当性が高いと判断します