

デシジョンツリーモデルの 個人与信モデルへの応用

2011/4/21

データマイニング株式会社
代表取締役 谷岡日出男

内容

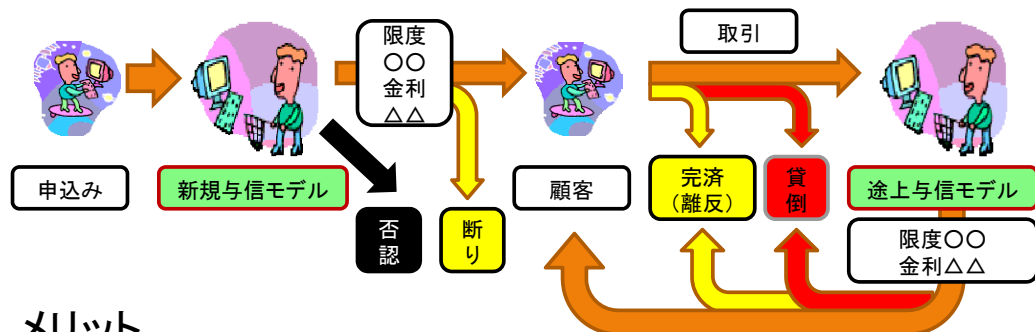
1. 与信モデルとは
与信モデル構築に必要な資源
2. 新規与信モデル
対象選択と分析データの作成
モデルの作成と検証
収益評価と限度設定
モニタリング
3. 途上与信モデル
4. 関連分析

無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business ¹

1. 与信モデルとは

- 金融機関の貸出審査業務で用いられる業務支援のための数式モデル。自動与信(自動審査)システムに組み込まれて利用されることが多い
- 顧客属性などにより顧客の信用度(貸倒確率)を推計し、貸倒確率に基づいて与信額などを決定する。ただし、法規制や商品スペックや会社の定める貸出制限などの範囲内となる
- 新規顧客に適用する与信(新規与信)と既存顧客に適用する与信(途上与信)がある



● メリット

- 与信モデルを用いることにより、業務の効率化、標準化、適正化が実現する
- 債権全体の現在と将来の状況を見通したり、長期の利益の最大化をもたらす
与信基準や限度額の策定が可能となる

無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business ²

与信モデル構築に必要な資源

①目的

- ・新規
 - ・承認率を上げたい
 - ・高額商品の開発
- ・途上
 - ・増額可能な対象
 - ・リスク対象

②データ

- ・新規申込時項目
- ・取引履歴データ(貸付・返済・延滞)
- ・外部信用情報等
- ・目的基準(移管・貸倒・代弁・リスクその他)の発生日

③人

- ・与信業務に精通した人
- ・データの所在, 項目の意味, 更新のタイミングなどに詳しい人
- ・分析手法や応用方法, 分析ツールに詳しい人

④分析ツール

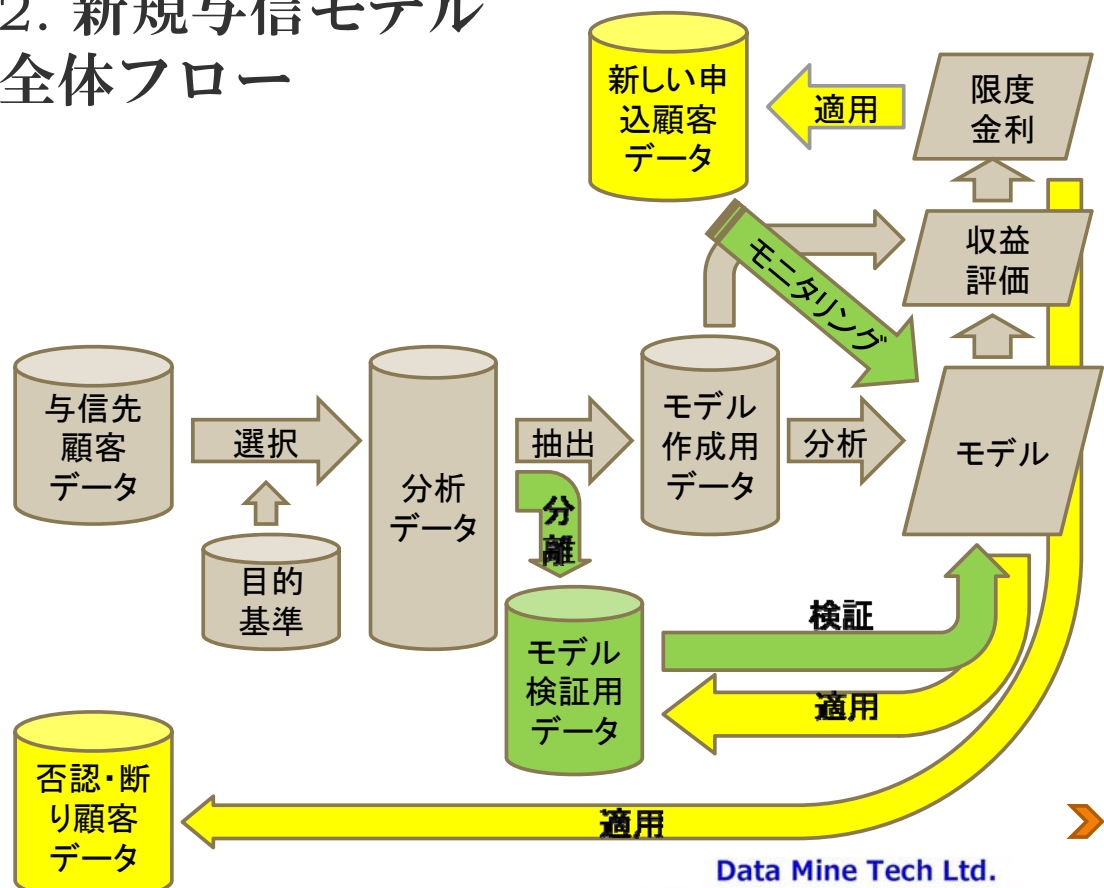
- ・データの切り出し・加工
- ・モデル作成・検証
- ・集計・レポート作成

無断転載を禁じます

Data Mine Tech Ltd.

Data Bring New Insight to Your Business 3

2. 新規与信モデル全体フロー

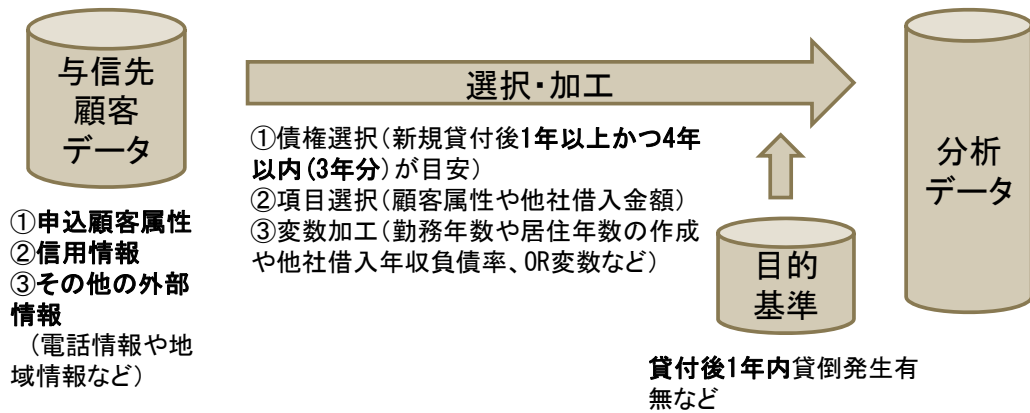


無断転載を禁じます

Data Mine Tech Ltd.

Data Bring New Insight to Your Business 4

対象選択と分析データの作成

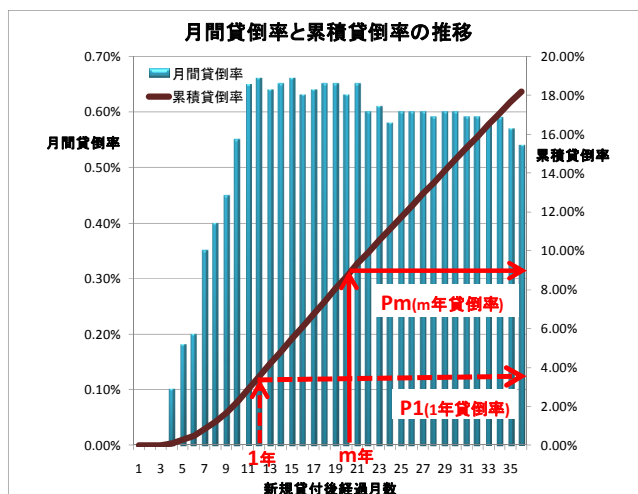


無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 5

分析データ件数が少ない場合

- 分析データ件数が少ない場合は、貸付後1年未満の債権データも分析データに含めることを検討する
- 貸付後1年以上経過した債権を含めて、現時点までの貸倒有無を基準としてモデル作成
- モデルの予測貸倒率を データの累積貸倒率曲線 とオッズ変換式 を用いて1年貸倒率に換算する



①データから累積貸倒率曲線を作成

②貸付後平均経過期間mの累積貸倒率 P_m と1年貸倒率 P_1 を読み取り、 P_1 と P_m のオッズ変換定数 $oddr$ を求める

$$oddr = \{P_1 / (1 - P_1)\} / \{P_m / (1 - P_m)\}$$

③上記オッズ変換定数を用いて、任意の貸倒予測値 p を1年換算貸倒率 q に変換する

$$q = oddr * p / (1 - p) / \{1 + oddr * p / (1 - p)\}$$

無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 6

累積貸倒曲線の作成

新規貸付月別、経過月別にデータ集計する方法

- クロス集計結果から新規貸付月別経過月別の件数情報が明確になる
- 集計が面倒、求める全体の累積貸倒曲線がギザギザになりやすい(経過月数が増えたのに累積貸倒率は減少するといった逆転現象が出現することもあり得る)

分析データ

ID	貸付月	貸倒月	貸倒経過月数	現在月	経過月数
0001	1			4	3
0003	1	3	2	4	3
0002	1			4	3
0005	1			4	3
0004	2	3	1	4	2
0006	2			4	2
0007	2			4	2
0008	3	4	1	4	1
0010	3			4	1
0011	3			4	1
0012	3			4	1
0009	3	4	1	4	1

貸付月	状態	貸付後経過月数		
		1	2	3
1	貸倒	0	1	1
	正常	4	3	3
	計	4	4	4
2	貸倒	1	1	
	正常	2	2	
	計	3	3	
3	貸倒	2		
	正常	3		
	計	5		
計	貸倒	3	2	1
	正常	9	5	3
	計	12	7	4
累積貸倒率		0.25	0.29	0.25

ギザギザ

生存時間分析を用いる方法

- 累積曲線を求めるのが非常に簡単、結果は常になめらかな曲線になり逆転現象は起こらない

分析データ

ID	貸付月	貸倒月	貸倒経過月数	現在月	新規貸付後経過月数	イベント発生時間(t)	イベントの種類(flag)
0001	1			4	3	3	(0)観測打ち切り
0003	1	3	2	4	3	2	(1)貸倒
0002	1			4	3	3	(0)観測打ち切り
0005	1			4	3	3	(0)観測打ち切り
0004	2	3	1	4	2	1	(1)貸倒
0006	2			4	2	2	(0)観測打ち切り
0007	2			4	2	2	(0)観測打ち切り
0008	3	4	1	4	1	1	(1)貸倒
0010	3			4	1	1	(0)観測打ち切り
0011	3			4	1	1	(0)観測打ち切り
0012	3			4	1	1	(0)観測打ち切り
0009	3	4	1	4	1	1	(1)貸倒

イベント発生時間 t ,
打ち切り $flag = 0$

累積貸倒率	貸付後経過月数		
	1	2	3
累積貸倒率	0.25	0.35	0.35

なめらか

無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business

7

予測貸倒率の1年換算

- (例) モデル作成に用いたデータの貸付後平均経過年数が 1.8年、累積貸倒曲線の1.8 年貸倒率が 5%、1年貸倒率が 2% だったとする

経過年数	累積貸倒率
1年(p1年)	2%
1.8年(p平均期間)	5%
オッズ変換定数(oddr) $= 0.02 / (1 - 0.02) / \{0.05 / (1 - 0.05)\}$	0.3878

- この場合、モデル予測貸倒率 p は、以下のように 1年貸倒率に換算できる

1.8年平均期間 モデル予測貸倒率 p	オッズ変換による 1年換算貸倒率 $\{oddr * p / (1 - p)\} / \{1 + oddr * p / (1 - p)\}$
1%	0.39%
2%	0.79%
3%	1.19%
4%	1.59%
5%	2.00%
6%	2.42%
7%	2.84%
8%	3.26%
9%	3.69%
10%	4.13%

無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business

8

目的基準該当件数が少ない場合

- 対応
 - 目的基準とする貸倒などの出現率が非常に小さい場合は、2か月超延滞など貸倒と関連が強いイベントを含めた基準を用いることを検討する
この場合は、モデルの出現率を本来の貸倒率に変換し、1年貸倒率の推計値とする
- 本来の貸倒率への変換方法
 - (1) 割増/割引抽出操作(Over Sampling)の考え方に基づく本来の貸倒率への変換
 - ・ 貸倒/非貸倒などの分類別に割増/割引抽出したデータにおける出現率を本来の出現率に戻す操作
 - ・ 変換後の平均出現率が本来の平均出現率に完全に一致することが確認できるという長所がある
 - (2) オッズ変換－出現率のみ変換する方法
 - ・ 得られた出現率の水準を目標出現率の水準に上げ下げするための操作
 - ・ オッズ変換式には出現率しか登場せず、全体平均値は実際には目標出現率に一致するが、件数の重み変化がわからないためそのまま平均をとると一致しない点に注意



Data Mine Tech Ltd.

Data Bring New Insight to Your Business 9

無断転載を禁じます

割増/割引抽出操作の考え方に基づく本来の貸倒率への変換

(例) 100件のデータのうち3件が貸倒、7件は非貸倒だが2か月超延滞だったため、これを含む10件を目的基準とした場合を考える

目的基準該当件数は、本来 3件だったのを 10件に割増抽出したものとみなす

(割増/割引抽出率 $XR=10/3$)

目的基準非該当件数は、本来 97件だったのを 90件に割引抽出したものとみなす

(割増/割引抽出率 $YR=90/97$)

割増/割引抽出後のデータにおける目的基準該当件数、非該当件数をそれぞれ X, Y とすると、

本来の出現率

= 本来の目的基準該当件数 / (本来の目的基準該当件数 + 本来の目的基準非該当件数)

= $(X/XR) / (X/XR + Y/YR)$

上記100件のデータを 2つのグループに分けたとき、以下の件数が得られたとすると、

- (1) $N=80$, 目的基準 4件 → 本来の出現率 = $(4 \times 3/10) / (4 \times 3/10 + (80-4) \times 97/90) = 1.2/83.111 = 0.0144385$
- (2) $N=20$, 目的基準 6件 → 本来の出現率 = $(6 \times 3/10) / (6 \times 3/10 + (20-6) \times 97/90) = 1.8/16.889 = 0.10657895$

ここで、(1)と(2)の本来の出現率の平均を求めると、全体平均と完全に一致する

(1)と(2)の平均 = $(83.111 \times 0.0144385 + 16.889 \times 0.10657895) / 100 = 0.03$



Data Mine Tech Ltd.

Data Bring New Insight to Your Business 10

無断転載を禁じます

オッズ変換－出現率のみ変換する方法

(同じ例) 100件のデータのうち3件が貸倒、7件は非貸倒だが2か月超延滞だったためこれを含む10件を目的基準とした場合を考える

目的基準出現率(y) $y=10/100=0.1$

貸倒出現率(目標出現率)(x) $x=3/100=0.03$

オッズとは貸倒率を p とすると $\text{オッズ}=p/(1-p)$ のこと。 $0 \leq p \leq 1$ が $0 \leq \text{オッズ} < \infty$ と変わる($p=0.5$ のときオッズ=1をとり、 p が大きいほどオッズも大きくなる)

オッズ比とは2つの出現率 x, y それぞれのオッズの比のこと。

出現率 x の出現率 y に対するオッズ比

$\text{odd_ratio}=x/(1-x)/\{y/(1-y)\}=0.1/0.9/(0.03/0.97)=0.278$

ここで、以下のように100件のデータを2つのグループに分けたとする

(1) $N=80$, 目的基準 4件 出現率(y_1) $y_1=0.05$

(2) $N=20$, 目的基準 6件 出現率(y_2) $y_2=0.3$

y_1 のオッズ変換による貸倒出現率 $x_1=\{y_1/(1-y_1)*\text{オッズ比}\}/\{1+y_1/(1-y_1)*\text{オッズ比}\}=0.0144385$

y_2 のオッズ変換による貸倒出現率 $x_2=\{y_2/(1-y_2)*\text{オッズ比}\}/\{1+y_2/(1-y_2)*\text{オッズ比}\}=0.10657895$

前ページの割増/割引抽出率による本来の出現率への変換結果と同じになる



無断転載を禁じます

Data Mine Tech Ltd.

Data Bring New Insight to Your Business 11

項目選択

①属性情報

- 申込属性情報のカテゴリー項目はそのまま用いる(組合せはツリーモデルが自動的に決定する)
- 年齢・居住年数・勤務年数などの経過期間を表す数値項目は申込時点または審査時点における情報とする(貸付時点ではなく、実際の審査のプロセスと合わせる)

②信用情報

- 信用情報項目は信用情報機関によって異なるが、基本的には信用情報照会時点における借入件数・借入金額、異動の有無、延滞日数や直近一定期間の照会件数など。必要であれば項目選択と加工を行った上で用いる

③その他の外部情報

- 信用リスクと関連があると思われる外部情報の利用も検討する

(1) **電話情報** 申込み顧客の携帯や勤務先の電話履歴情報(料金未納による利用停止など)

(2) **勤務先情報** 勤務先の信用度に関する情報(信用格付)

(3) **地域情報** 申込み顧客の住居エリアの平均年収などの統計情報



無断転載を禁じます

Data Mine Tech Ltd.

Data Bring New Insight to Your Business 12

説明変数の分布の確認

- 分布範囲の確認
 - カテゴリカル変数
 - ・ 存在するはずの無いカテゴリは無い、逆に存在するはずのカテゴリが存在しないということは無いか？
 - ・ バランスが悪い変数は無いか？(例：全体の99%はAで残りの1%がB)
 - ・ カテゴリ数が非常に多い変数はないか？
 - 数値変数
 - ・ 異常値、外れ値は無いか？(例：年齢17歳, 年収1億円)
 - ・ 金額の単位、日付の入力の仕方の確認
- 欠損値の確認
 - 欠損値の有効性
 - ・ 許される欠損値か、欠損に意味があるか？
 - ・ その数値変数の欠損値は0と同じとみなして良いか？
 - 欠損値の割合
 - ・ 欠損割合が多い変数を分析に用いて良いか？



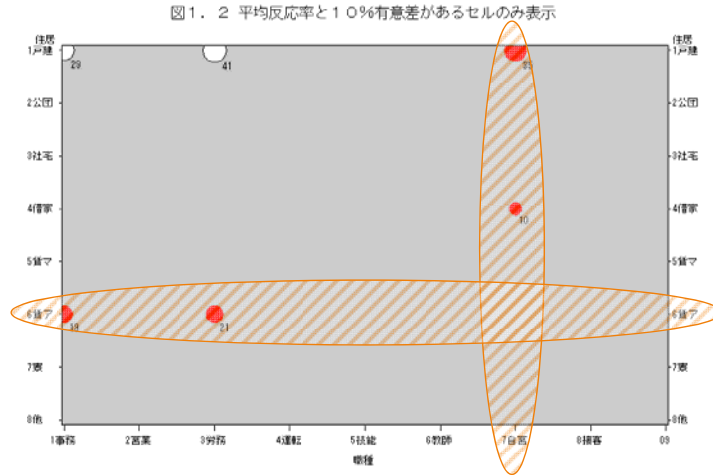
説明変数の加工

- カテゴリ数の多い変数
 - ・ おおむね20を超えるカテゴリカル変数は少数のカテゴリにまとめた変数を作って説明変数に追加しておく。まとめ方は“意味”が似ているカテゴリを併合する方法や“階層化”の方法(大分類、中分類といったまとめ方、例えば地域)などいろいろある
- 0(ゼロ)が多い変数
 - ・ 0の割合がおおむね2/3を超える数値変数は0とそれ以外という変数を作って説明変数に追加しておく
 - ・ 欠損が0と同じ意味なら“0または欠損”と“それ以外”とする
- 複数の変数にまたがる説明変数
 - 借入額年収負担率
 - ・ 申込時点の他社借入総額の年収に対する比率
 - ・ 貸金業法適用の場合は、当社貸付限度額を加算したとき総量規制に抵触しないようにする必要がある
 - OR変数
 - ・ デシジョンツリーモデルは判別に寄与する説明変数同士のAND条件を自動探索するが、複数の説明変数間のOR条件は仕組み的に発見できない
 - ・ そこで、説明力があると思われる、複数の説明変数にまたがるOR変数をあらかじめ作成できないかどうかを検討する
 - ・ 相互に代替的(例えば、キャッシング利用とカードローン利用)の関係にある項目はOR条件説明変数の候補となる。数量や金額や比率は合算した上で計算する



OR変数の探索

- 2変数間のカテゴリ組合せから探索する方法
 - クロスセルの貸倒率が平均より高いセル(●)と低いセル(○)から見つける例



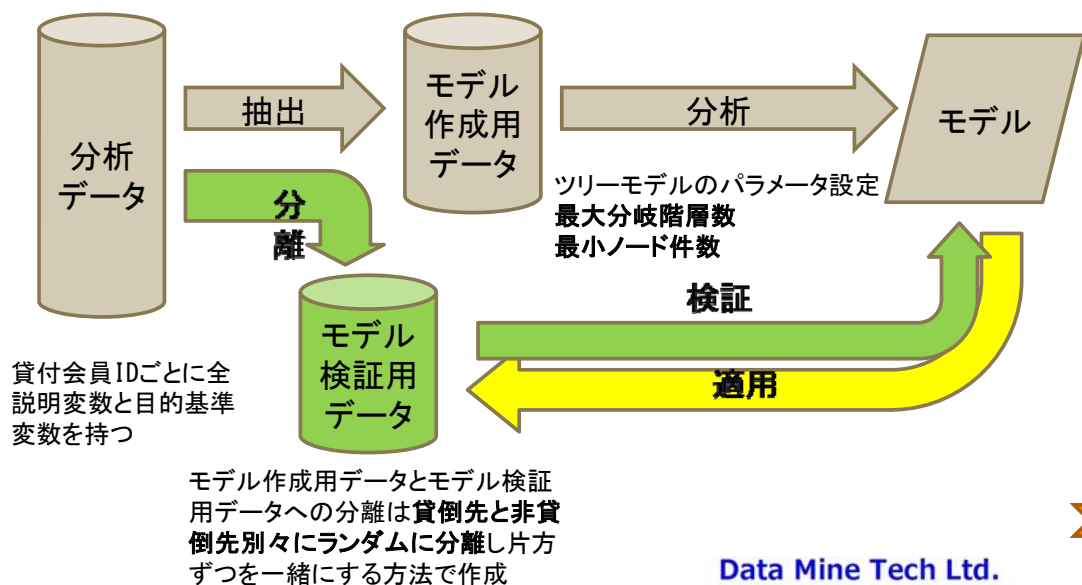
谷岡 (2007) "ターゲット判別に有効な2変数間の選言的合成変数の自動抽出"
(SASユーザー会学術総会 2007) より掲載

無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 15

モデルの作成と検証

- 分析データをモデル作成用データとモデル検証用データにランダムに分離する
- デシジョンツリーモデルのパラメータ(最大分岐階層数や最小ノード件数)を設定
- 作成したモデルを検証用データに適用し、当てはまり具合(精度)を評価

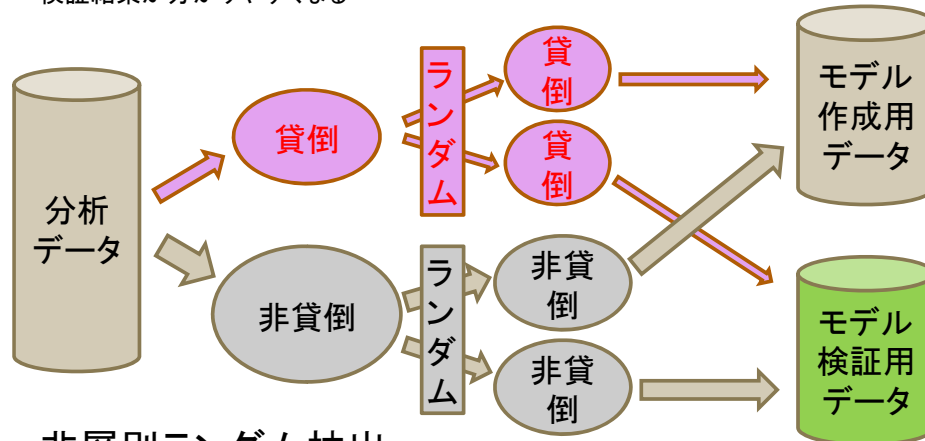


無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 16

層別ランダム抽出

- 分析データを貸倒債権と非貸倒債権に分けておき、それぞれのグループから同じ抽出率で分析データからランダム抽出した債権を一緒にしたものをモデル作成データとし、残りをモデル検証用データとする(あらかじめグループ別に分けておくため“層別”と呼ばれる)
- この方法により、モデル作成用データとモデル検証用データの貸倒率が等しくなり、モデルの検証結果が分かりやすくなる



非層別ランダム抽出

- 分析データを乱数によってランダムにモデル作成用データとモデル検証用データに分ける方法
- 会員番号といったランダム性が保証できると考えられる番号の奇数・偶数でモデル作成用データとモデル検証用データに分ける、といった方法でも実現できる
- モデル作成用データの貸倒率とモデル検証用データの貸倒率が微妙に違ってしまいうという欠点がある

Data Mine Tech Ltd.

Data Bring New Insight to Your Business 17

無断転載を禁じます

ツリーモデルで設定するパラメータ

最大分岐階層数 (maxlvl)

- 説明変数の複合レベル数の上限を決める(最も複雑な審査ルールの項目複合数)
 - (例) maxlvl=2の場合、例えば、年齢と性別の2項目しか見ないことを意味する。
- 予測値の種類数(終端セグメント数)の上限が決まる(2分木の場合、2の(最大分岐階層数)乗)
 - (例) maxlvl=5の場合、 $2^5=32$ 、maxlvl=8の場合、 $2^8=256$
- 新規与信モデルでは、それほど顧客の信用リスクの細分化が必要というわけではない場合が多い。その場合は5~8個の審査属性を見る、30~100個程度の終端セグメントを持つデジモンツリーモデルで十分と思われる

最小ノード件数 (mincnt)

- 中間ノード、終端ノードを含むすべてのノードに含まれるデータ件数の最小値の指定
- 統計的誤差と関係する。ある終端ノードの件数を $N (> \text{mincnt})$ 、観測貸倒率 p を母集団貸倒率の推計値とすると、真の母集団貸倒率が推計値からどれだけ離れた範囲に存在するかの程度を表す標準誤差(推計値のパラツキは“標準偏差”と呼ばず“標準誤差”と呼ぶならわしがある)は以下の式により計算できる

$$p \text{ の標準誤差} = \sqrt{\frac{p(1-p)}{N}}$$

- (例) $p=0.01$ とし、 N をさまざまに変化させたときの標準誤差と95%上下信頼限界 ($p \pm 1.96 \times \text{標準誤差}$) を計算すると、右表のようになる。
- $N=100$ の場合、95%上側信頼限界は2.95% と大きく、1%の観測貸倒率の信頼性は低い

p	N	標準誤差	95%下側 信頼限界	95%上側 信頼限界
1.00%	100	0.99%	0.00%	2.95%
	200	0.70%	0.00%	2.38%
	400	0.50%	0.02%	1.98%
	900	0.33%	0.35%	1.65%
	1,600	0.25%	0.51%	1.49%
	2,000	0.22%	0.56%	1.44%
	5,000	0.14%	0.72%	1.28%
	10,000	0.10%	0.80%	1.20%

Data Mine Tech Ltd.

Data Bring New Insight to Your Business 18

無断転載を禁じます

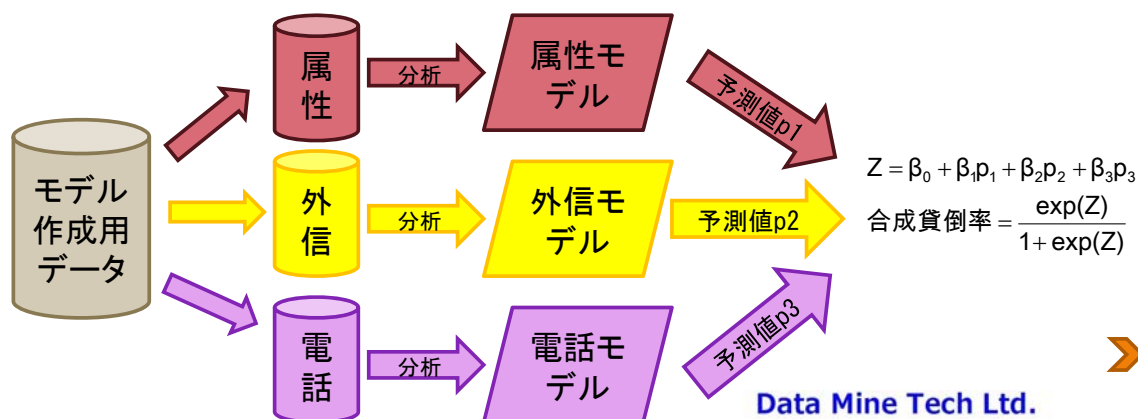
単一モデルの作成

- 全項目を説明変数として入力し、単一のツリーモデルを作成
- モデルに採用される審査項目数は比較的少数になる
 - これは良い点でもあり良くない点でもある
 - 少数の審査項目しか見ないで済む単純なモデルは精度が十分あるとすれば良いモデルと言える
 - しかし、重要と思われる項目を見ないで審査するのは危険かもしれない



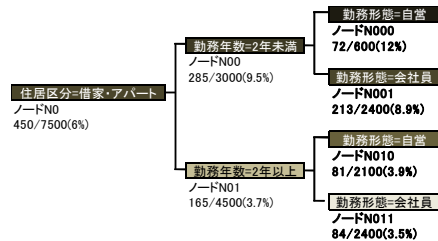
複数モデルの予測値を合成するモデル

- 各モデルにおける重要審査項目のチェックを必ず受ける審査モデルが得られるため、単一モデルの難点である重要な審査項目を見ないで判断する危険を減らすことができる
 - 各モデルの再作成などのメンテナンスが容易
 - 相互に独立性が高い複数モデルの合成の場合に特に有効
 - 一般的に精度が単一モデルより高くなる



シンプソンのパラドクス

- 作成したツリーモデルの内容を調べていくと、**納得しにくい部分にぶつかる**ときがある
- (例) 以下の**ノードN0(7500件中450件が貸倒)**に注目する



- 勤務年数別にみると**、2年未満のグループ、2年以上のグループ、いずれにおいても、**会社員より自営の方が貸倒率が高くなっている**
- しかし、**ノードN0全体で集計してみると**、こんどは**自営より会社員の方が貸倒率が高くなっている**

勤務年数	勤務形態						計		
	自営			会社員			ノード	貸倒	貸倒率
	ノード	件数	貸倒率	ノード	件数	貸倒率	件数	件数	貸倒率
勤務2年未満	600	72	12.0%	2400	213	8.9%	3000	285	9.5%
勤務2年以上	2100	81	3.9%	2400	84	3.5%	4500	165	3.7%
計	2700	153	5.7%	4800	297	6.2%	7500	450	6.0%

- これは、“**シンプソンのパラドクス(矛盾)**”と呼ばれる現象であるが、実は**パラドクス(矛盾)**ではなく、**事実**を表しているに過ぎない
- 例えば、年収と貸倒には関連が見られないのに、年齢別に年収と貸倒との関連を調べると関連が出てくるような場合も同じである。出てきた事実が、われわれには**パラドクス**と思われる場合にそう呼んでいるわけである

Data Mine Tech Ltd.

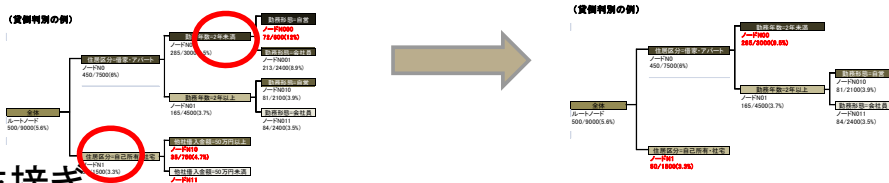
Data Bring New Insight to Your Business 21

無断転載を禁じます

枝刈りと枝接ぎ

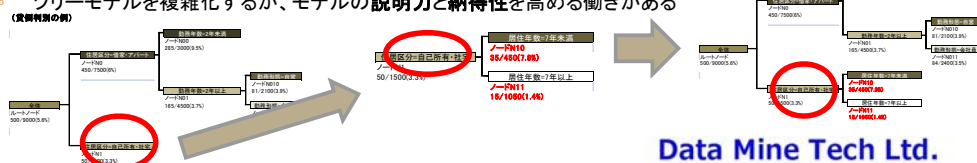
枝刈り

- ツリーモデルにおける**特定の中間ノードから下の部分を取り除き、その中間ノードを終端ノードにする**操作
- 末端に近いノードほど該当件数が少なくなるため、モデル作成用データに含まれるランダムな変動を普遍的な変動とみなしてモデルに取り込んでしまう傾向(これをモデル作成用データへの“**過剰適合**”と呼ぶ)が生じやすくなる
- ツリーモデルを簡素化し、モデルのモデル作成用データへの過剰適合を防ぎ、モデルを新しいデータに適用した場合の誤差(これを“**汎化誤差**”と呼ぶ)を小さくする働きがある



枝接ぎ

- ツリーモデルにおける**特定の終端ノードに別のツリーモデルを追加し、その終端ノードを中間ノードにする**操作
- 件数のバランスが悪いなどの理由で作成したツリーモデルには出現しないが、特定の終端ノードのデータの基準判別には有効と考えられる説明変数がある場合、その終端ノード該当データに対し、**最小ノード件数条件を緩和**したツリーモデルを作成し追加する
- ツリーモデルを複雑化するが、モデルの**説明力と納得性**を高める働きがある



Data Mine Tech Ltd.

Data Bring New Insight to Your Business 22

無断転載を禁じます

適合率と再現率

- 貸倒/非貸倒といった2つのクラスを判別する場合の予測精度に関連する2つの基本的指標
 - 適合率** = [貸倒と予測しかつ正しかった件数] / [貸倒予測総件数] (信頼度)
 - 再現率** = [貸倒と予測しかつ正しかった件数] / [貸倒事例総件数] (カバー率)
 - どちらも重要な予測の評価指標
 - 得られたモデルの**予測貸倒率があるしきい値以上/未満のとき貸倒/非貸倒と予測することとすると**、以下のような予測{貸倒, 非貸倒}と事例{貸倒, 非貸倒}の2*2のクロス表(正誤表)が得られる

正誤表

予測	事例		計
	貸倒	非貸倒	
貸倒	6	9	15
非貸倒	24	61	85
計	30	70	100

貸倒の予測に関する精度
適合率=6/15=0.4
再現率=6/30=0.2

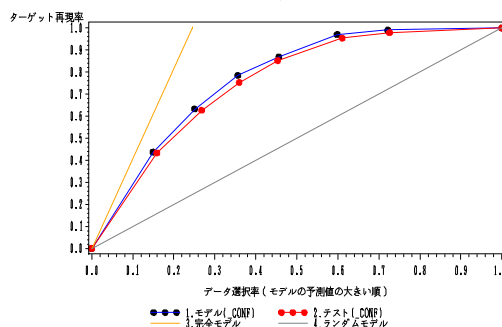
- しきい値をツリーモデルの最も大きい予測貸倒率から最も小さい予測貸倒率まで変化させていくと、
 - 適合率は下がる(予測貸倒率最大のノードのみ貸倒と予測するときが適合率最大)
 - 再現率は上がる(全ノードを貸倒と予測すると、再現率100%、適合率は全体平均貸倒率となる)

ゲインチャートとROC曲線

- ゲインチャート**(CAP曲線ともいう)は、**終端ノードを予測貸倒率の大きい順に並べておき、縦軸に貸倒再現率(貸倒と予測しかつ正しかった件数/貸倒事例総件数)、横軸に累積件数百分率をとった点を結んだ曲線**
- ROC曲線**は、ゲインチャートと同じく、**終端ノードを予測貸倒率の大きい順に並べておき、縦軸に貸倒再現率(ROC曲線の用語では貸倒再現率のことを“感度(Sensitivity)”と呼ぶ)、横軸に(1-非貸倒再現率)(ROC曲線の用語では非貸倒再現率のことを“特異度(Specificity)”と呼ぶ)をとった点を結んだ曲線**
- モデル作成用データ、モデル検証用データそれぞれに対してゲインチャートやROC曲線を作成し、1つの図に重ねて表示する**
- 一般にモデルより検証の方が若干当てはまりが落ちるものである。しかし、その差が大きい場合は、作成したモデルがモデル作成用データに含まれるランダムな変動を普遍的な変動とみなしてモデルに取り込んでしまっている(これを“**過剰適合**”と呼ぶ)可能性があり、予測精度が確保できない恐れがあるため、**枝刈りなどの対策を行う必要がある**

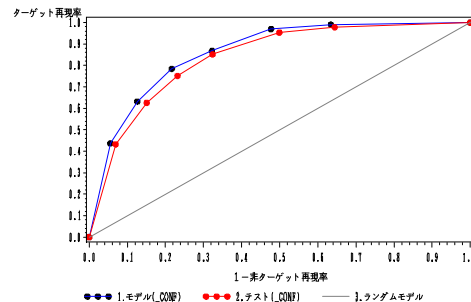
モデルのゲインチャート

AR=0.7222(モデル),0.6751(テスト)



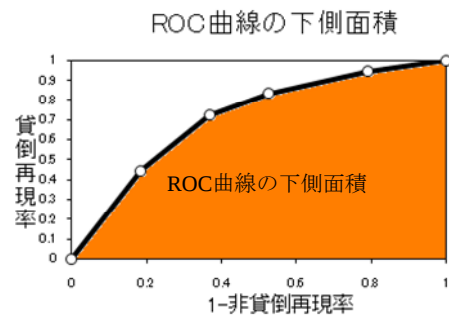
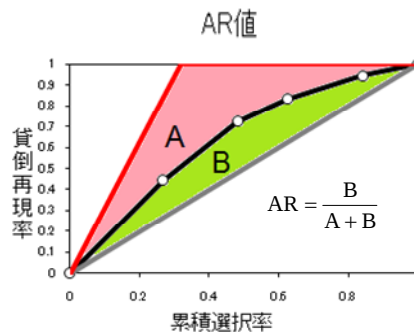
モデルのROCチャート

ROC_ARE A=0.6611(モデル),0.6316(テスト)



AR値とROC曲線の下側面積

- AR (Accuracy Ratio) 値はモデルのデータへの当てはまり度合いを示す1つの指標。0～1の範囲をとり、1に近いほどモデルのデータへの当てはまりが良いことを意味する。
- 計算方法は、ゲインチャートにおけるモデルのゲインチャート曲線とランダムモデルのゲインチャート曲線を意味する対角線に挟まれた領域面積の、完全(理想)モデルのゲインチャート曲線と対角線に挟まれた領域面積に対する割合として定義される
- ROC曲線の下側面積も AR値と同じ意味を持つ。医薬分野や計測分野で良く用いられる。ただし0.5～1の範囲をとる
- ROC曲線の下側面積 = $AR/2 + 0.5$ という関係がある



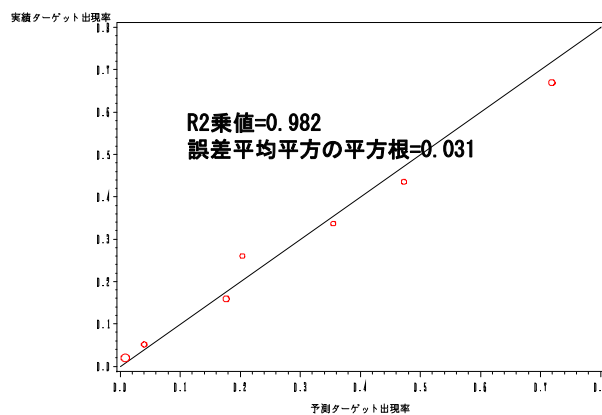
無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 25

比較プロットとR2乗値 および誤差平均平方の平方根

- ゲインチャートやROC曲線などが予測貸倒率の大きさそのものではなく、予測貸倒率の実際の貸倒出現率との対応順序関係を評価するのに対し、比較プロットは終端ノードごとの予測貸倒率と実際の貸倒率との差(誤差)を評価する

モデル予測値(samp_model) 対 実際値(test_model)の比較プロット
R2乗=0.982, 誤差平均平方の平方根=0.031



iをノードを表す添え字、
 W_i をi番目のノードの件数、
 Nodeをノード数として、
 誤差(E_i) = 実績値(Y_i) - 予測値(P_i)
 誤差平方和(ESS) = $\sum_{i=1}^{Node} (W_i E_i)^2$
 偏差(C_i) = 実績値(Y_i) - 実績値平均値($\bar{Y}$)
 偏差平方和(CSS) = $\sum_{i=1}^{Node} (W_i C_i)^2$
 $R^2 \text{乗値} = 1 - \frac{ESS}{CSS}$
 誤差平均平方の平方根 = $\sqrt{\frac{ESS}{\sum_{i=1}^{Node} W_i}}$

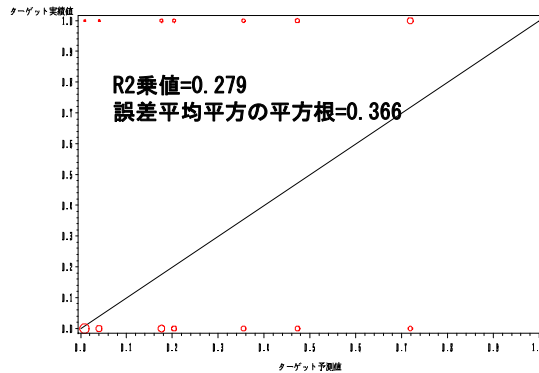
無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 26

誤差の計算方法について

- 予測値と実際値との違いを定義する誤差は、**ノード単位**として計算する場合と**個々の顧客単位**に計算する場合で異なり、それぞれに意味がある。なお、モデル式(実績値=予測値+誤差)から、常に、**誤差=実績値-予測値**と定義される
- 例えば、**年30%の貸倒率**と予測される**終端ノード**が得られたとし、検証データでもこのノードは**年29%の貸倒率**であったとする。この場合、**ノード単位**で見ると、このノードの**誤差=0.29-0.3=-0.01**と計算され、**誤差平方和**も小さな値になる。
- しかし、**個々の顧客**または**債権単位**に見ると、貸倒または非貸倒のいずれかの値しかとらない。貸倒は1、非貸倒は0で表現されるので、このノードの29%を占める個々の貸倒顧客の誤差は**1-0.3=0.7**、71%を占める非貸倒顧客の誤差は**0-0.3=-0.3**と計算され、**ノードの誤差平方和**もかなり大きな値となる

test_score におけるモデル予測値 対 実績値(ターゲット: income='50K')
R2乗=1.111, 誤差平均平方の平方根=1.366



顧客単位に計算した誤差に基づくモデルの当てはまりを表すR2乗値は、前ページのノード単位の計算結果と比較してかなり小さくなり、誤差平均平方の平方根は大きくなる

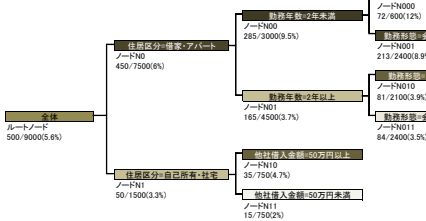
無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 27

モデルの精度の価値

- 作成したモデルを、既存のモデルや完全モデル(完全に正しい予測を行う仮想のモデル)と金額で比較する

(貸倒判別の例)

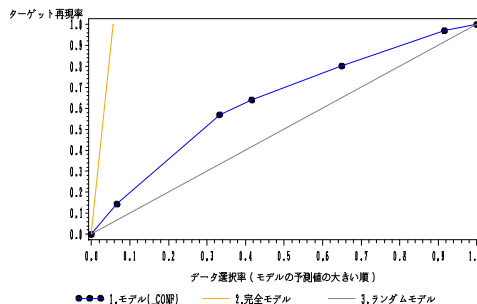


限度	10							
金利	0.1							
終端ノード	貸倒率	ノード該当件数	貸倒件数	非貸倒件数	期待収益	貸倒損失	期待利益	その貸倒率以下を貸付した場合の合計期待利益
N000	12.0%	600	72	528	528	-720	-192	3,500
N001	8.9%	2,400	213	2,187	2,187	-2,130	57	3,692
N10	4.7%	750	35	715	715	-350	365	3,635
N010	6.3%	2,100	81	2,019	2,019	-810	1,209	3,270
N011	3.5%	2,400	84	2,316	2,316	-840	1,476	2,061
N11	2.0%	750	15	735	735	-150	585	585

完全モデル=非貸倒先へのみ貸付	非貸倒件数	期待収益	貸倒損失	期待利益
	8,500	8,500	0	8,500

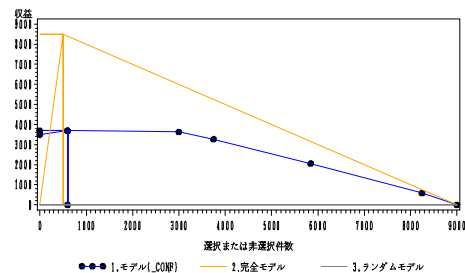
モデルのゲインチャート

AR=0.2111(モデル)



モデルの収益チャート

最大収益=8500(完全モデル), 1(ランダムモデル), 3333(モデル)
収益単位: 正予測=1, 正予測=0, 負予測=1, 負予測=0

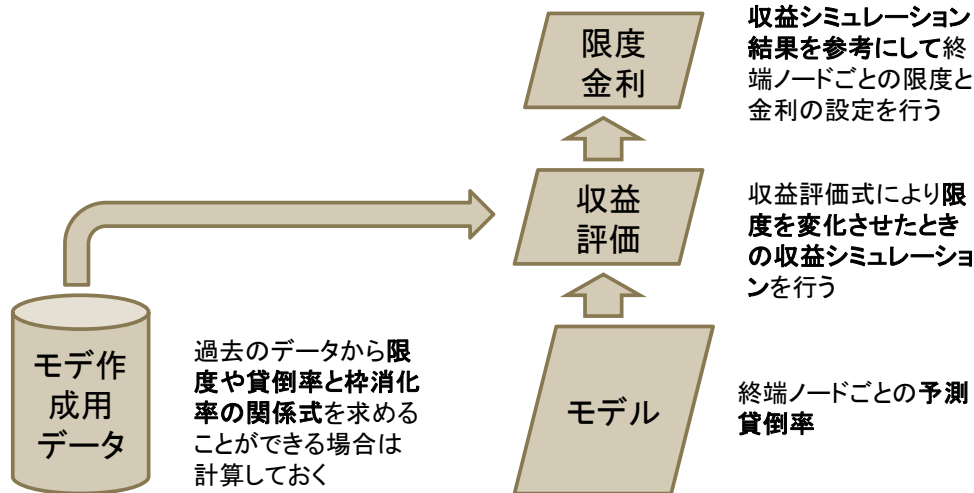


無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 28

収益評価と限度設定

- 予測貸倒率、平均残高(枠消化率)、貸付利率から見込み利益を計算
- 利益 ≤ 0 となる会員は貸付否認
- 限度と貸付利率をパラメータとして収益シミュレーションを行い終端ノードごとの限度と貸付金利を決定



無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 29

収益シミュレーション

● 収益計算式

- 簡単のため、貸倒が発生する場合は、すべて貸付1年後に発生すると仮定し、**貸付金利による収益と貸倒による損失のみを考慮**すると、個々の顧客から得られる期待利益額は以下の算式により求められる

$$\begin{aligned}\text{期待利益} &= \text{期待収益} - \text{期待損失} \\ &= \text{平均残高} * \text{貸付利率} - \text{限度} * \text{貸倒率} \\ &= \text{限度} * \text{枠消化率} * \text{貸付利率} - \text{限度} * \text{貸倒率} \\ &= \text{限度} * (\text{枠消化率} * \text{貸付利率} - \text{貸倒率})\end{aligned}$$

- したがって、期待利益の正負は以下の条件式によって決定される

$$\begin{aligned}\text{期待利益} &> 0 \text{ となる条件} \\ \text{枠消化率} * \text{貸付利率} - \text{貸倒率} &> 0 \\ \text{貸倒率} &< \text{枠消化率} * \text{貸付利率}\end{aligned}$$

● 枠消化率の推計

- 貸付した後の顧客ごとの**枠消化率**をどのように推計するかは、収益に関わる大きな問題
- **ノードごとの実績平均を用いる方法**や、**限度や貸倒率との関連に着目し、分析データを用いて、枠消化率を限度や貸倒率の関数として定式化を行い、これを上記期待利益計算式に用いる方法**がある

● 収益シミュレーション

- **政策変数である限度と貸付利率**を商品設計上の想定範囲の中で動かし、期待利益がどう変わるかを確認する
- 他社の競合商品や当社の方針および**途上与信との関係**なども考慮して、終端ノードごとの貸付限度や金利を決定する

無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 30

途上与信モデル限度との関係

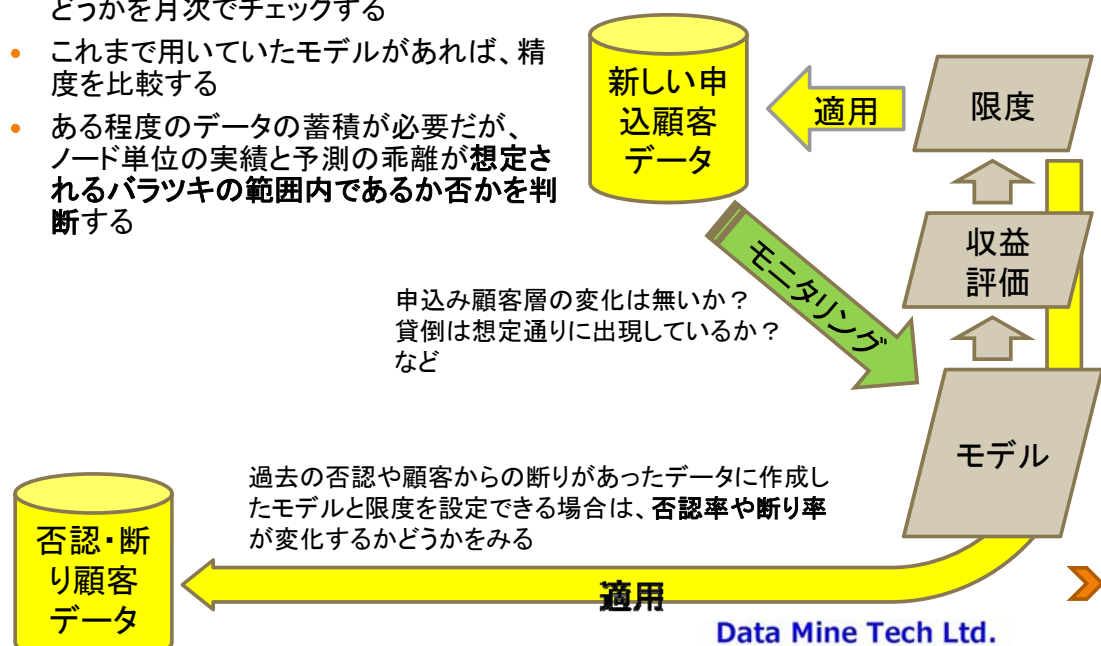
- 新規会員の1年貸倒率は、新規貸付後1年以上経過後の残あり正常会員に対する途上与信対象会員の1年貸倒率より低くなる
- 理由の1つに、新規与信対象会員は貸付時点では延滞などが発生していないのに対して、途上与信対象会員には延滞している場合も含まれることがある
- 前ページの収益シミュレーション式の利益 ≥ 0 基準で限度設定を行うと、**平均的な貸倒率の会員の途上与信額は新規与信額より低めに設定されることになり、実際上不都合となる**
- 実際には、新規会員には収益シミュレーション式で考慮していなかった**会員獲得コスト**や**審査コスト**などがかかっていることが多く、そのようなコストを考慮して収益計算を行う方が正しいと考えられる
- しかし、コストを適切に計算することはかなり困難であり、考慮した場合の1年間の収益計算を行うと、逆に**新規与信限度が低くなりすぎる可能性もある**。また、**会員期間を表す期待在籍年数 (TENURE)**を算出し、この期間内で収益を見るべきといった議論もある
- 現実的な対応**としては、新規と途上それぞれの平均貸倒率を持つ会員の**新規貸付時限度と途上限度がほぼ同じになるか、新規の方がやや低くなるような調整**を行うことがある

無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 31

モニタリング

- 新たな申込顧客に作成したモデルと限度を適用する。(テストもしくは本番運用)
- 運用結果がモデルの見込み通りであるかどうかを月次でチェックする
- これまで用いていたモデルがあれば、精度を比較する
- ある程度のデータの蓄積が必要だが、**ノード単位の実績と予測の乖離が想定されるバラツキの範囲内であるか否かを判断する**



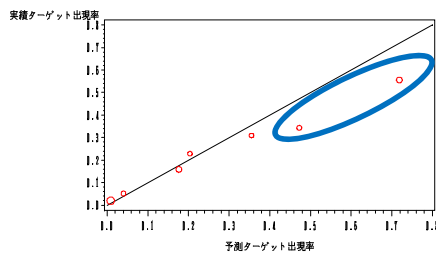
無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 32

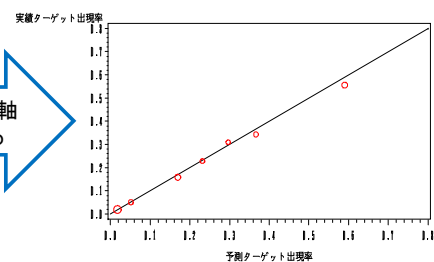
モデルの調整

- 特定の1個のノードのみ当てはまりが悪いような場合は、まず、モデルの適用プロセスなどをチェックする
- 貸倒率が予測より一律に上がる、または下がるといった現象が、全体的な傾向、または部分的な傾向(与信が高い/低い部分など)として見られる場合は、経済状況などの環境要因が急激に変化しているなどの原因が考えられる。
- 調整方法はさまざまだが、モニタリングデータと分析データの全部または最近の一部を一緒にしたデータセットを作成し、作成したツリーモデルを適用し分岐させ、生成された終端ノードごとの貸倒率を集計し、新たな予測貸倒率とする方法がある

モデル予測値(samp_model) 対 実測値(test2_model)の比較プロット
R²乗=0.019, 誤差平均平方の平方根=0.011



モデル予測値(samp2_model) 対 実測値(test2_model)の比較プロット
R²乗=0.999, 誤差平均平方の平方根=0.011



縦軸が横軸
に変わる

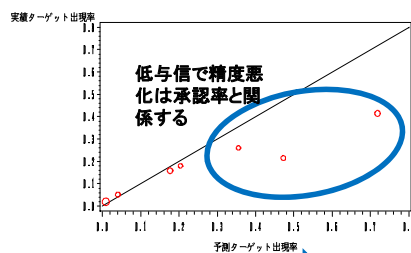
無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 33

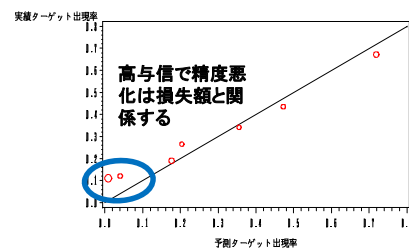
モデルの作り替え

- モデルの調整は、ノードごとの貸倒率のずれを調整するが、デシジョンツリーモデルそのものを改変するものではないため、モデルそのものの“切れ味”は時間とともに落ちていく可能性がある
- 一定の運用期間(おおむね1年~2年)、または一定数の貸付実績データ(おおむね、前回作成した分析データ件数の半分)が蓄積されたなら、前回用いたデータの全部または半分を加えた分析データを用いて、モデルの作り替えを試み、当てはまりを確認しつつ切れ味の良いモデルに作り替えることを検討する
- 切れ味が良い/悪いとは、例えば、あるグループを2つに分けるときの、貸倒率が90%と20%になるように2つに分けるモデルと貸倒率60%と30%になるように分けるモデルがあった場合、前者の方が切れ味が良いという意味である

モデル予測値(samp_model) 対 実測値(test3_model)の比較プロット
R²乗=-0.149, 誤差平均平方の平方根=0.151



モデル予測値(samp_model) 対 実測値(test4_model)の比較プロット
R²乗=0.005, 誤差平均平方の平方根=0.007



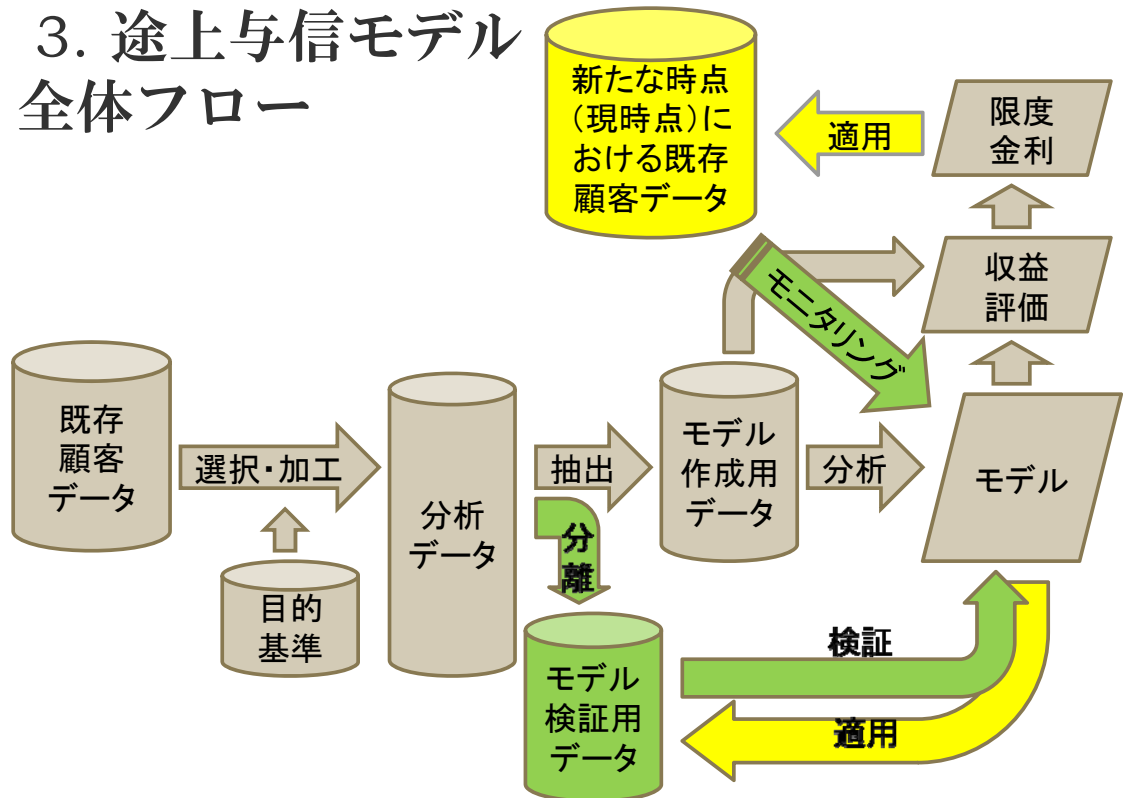
縦軸を横軸に変えてもモデルの切れ味が悪い

モデルの
作り替え

無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 34

3. 途上与信モデル 全体フロー

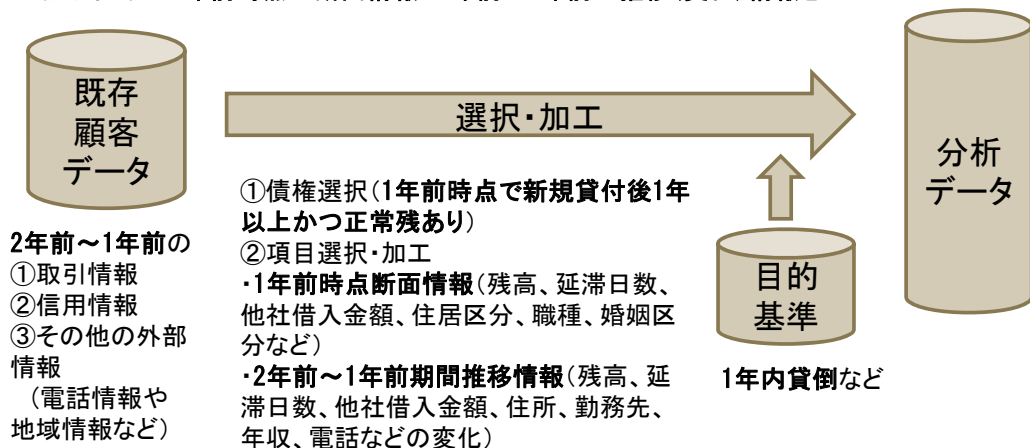


無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 35

対象選択と分析データの作成

- 1年前時点で新規貸付後1年以上経過かつ正常残あり顧客を分析対象とする
- 条件緩和、買取り等異質な債権については別途分析する場合が多い
- モデルの目的基準は1年以内の貸倒or移管とすることが一般的
- 説明変数は1年前時点の断面情報と2年前～1年前の推移(変化)情報を用いる



無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 36

項目選択

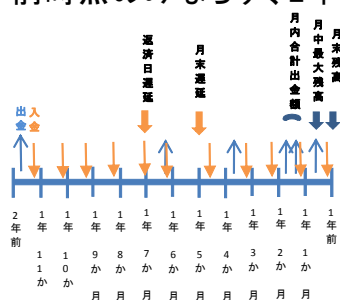
- ①取引情報
 - 取引履歴(入出金日付、入出金額、残高、次回約定返済日、不足金など)
- ②信用情報
 - 他社借入金額等の照会履歴
- ③属性情報
 - 属性変更履歴(住所、就業形態など)
- ④その他の外部情報
 - 信用リスクと関連があると思われる外部情報の利用も検討する
 - (1)電話情報 携帯や自宅や勤務先の電話履歴情報(料金未納による利用停止など)
 - (2)勤務先情報 勤務先の信用度に関する情報

無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 37

推移変数作成のための月次変数作成

- 取引履歴データから各説明変数の月次項目を作成する
1年前時点のみならず、2年前～1年1カ月前の各月の月次項目も作成



項目	月次変数			
月末残高	月末残高0	月末残高1	...	月末残高12
月末遅延日数	月末遅延日数0	月末遅延日数1	...	月末遅延日数12
月内最大残高	最大残高1	最大残高2	...	最大残高12
月内遅延日数	月内遅延日数1	月内遅延日数2	...	月内遅延日数12
月間入金額	月間入金額1	月間入金額2	...	月間入金額12
月間出金額	月間出金額1	月間出金額2	...	月間出金額12

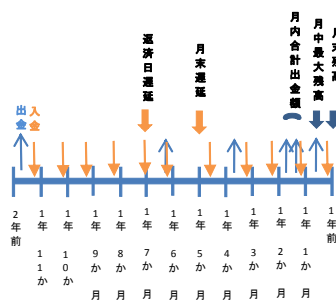
- 残高、延滞日数などは月末時点値とともに月内最大値を表す項目も作成
(例) 1か月前の月末残高と月内最大残高
- 入出金回数・金額は月内合計値を作成
- 属性変数の推移情報(変更履歴)を取ることができる場合は、2年前～1年前の各月次項目を作成
 - 年収、勤務年数、居住年数、就業形態など
- 加工変数も、同様に月次項目を作成
 - 貸付額に対する年収負担率など

無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 38

期間推移変数、統計変数、イベント変数の作成

- 月次変数から各変数の期間推移(変化)や期間内の平均やばらつきなどの統計情報を表す説明変数を作成
- 期間推移は常に1年前を終点とし、2年前～1年1カ月前の適当な時点(2年前、1年半前、1年3カ月前、1年1ヶ月前の4通り程度)を始点とした推移を作成
(例) 6か月残高推移=1年前の残高-1年6ヶ月前の残高
- 信用情報照会履歴などについては照会間隔に注意
 - 毎月照会の場合なら任意の期間の推移変数を作成可能
 - そうでない場合は、信用照会を行った会員と行わなかった会員が含まれる可能性がある短い期間の推移変数は作成しづらい
- 統計変数についても同様の終点、始点の期間範囲別に作成
(例) 3か月残高平均=(1年前の残高+1年1ヶ月前の残高+1年2ヶ月前の残高)
- 属性変数の月次履歴項目については、貸倒判別に寄与する推移パターン(イベント変数)があるかどうかを別に分析して、説明変数を作成



項目	月次変数		
月末残高	月末残高0	月末残高1	月末残高12

直近月末残高	=月末残高0
1か月月末残高差	=月末残高0-月末残高1
3か月月末残高差	=月末残高0-月末残高3
6か月月末残高差	=月末残高0-月末残高6
1年月末残高差	=月末残高0-月末残高12

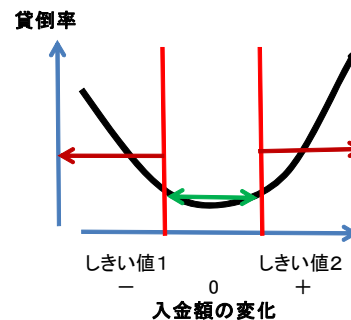
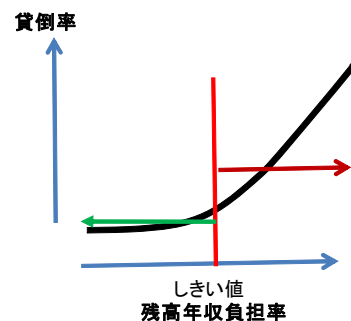
Data Mine Tech Ltd.

Data Bring New Insight to Your Business 39

無断転載を禁じます

数値説明変数と貸倒れ出現率との関連

- 途上与信モデルにおいては、顧客の取引情報などの数値タイプ説明変数が多数利用できる
- 数値タイプ説明変数の大きさや変化が貸倒れ発生と関連する場合
 - (例) 残高年収負担率...大きいほど貸倒率大、増加すると大
このような説明変数がデシジョンツリーモデルのノード分岐に採用される場合、ある1つのしきい値以上/未満によって2つのノードに分けられる
- 推移変数の中には、減少と増加、どちらの方向に変化しても、変化した場合が変化しない場合より貸倒れリスクが大きくなる場合がある
 - (例) 入金額...金額の大きさそのものはリスクとの関連小、増減幅が大きいとリスク大
このような説明変数がノード分岐に採用される場合、ある2つのしきい値を設定し、2つのしきい値の範囲内/範囲外によって2つのノードに分けられる



Data Mine Tech Ltd.

Data Bring New Insight to Your Business 40

無断転載を禁じます

4. 関連分析

ノード間推移確率に基づく多年度の予測 貸倒率の推計

- ある顧客集団の新規与信モデルの1年予測貸倒率を $P(1年)$ とすると、他に情報が無い場合の2年予測貸倒率は、以下の式で計算することが多い

$$P(1年)=0.01 \text{なら } P(2年)=1-(1-0.01)^2=0.0199$$

- 同様に、出発点の会員集団を固定した上で、時点に関わらず一様に貸倒が出現するとした場合、一般に、観測した期間 t の貸倒率を $P(t)$ とすると、任意の期間 u の貸倒率 $P(u)$ は、以下の式で計算できる

$$P(u)=1-(1-P(t))^{(u/t)}$$

- しかし、新規与信モデルと途上与信モデルの終端ノード間の推移確率行列を作成すると、貸付後1年経過時点の会員状態に関する情報が増え、さらに2年目の貸倒率を1年経過時点の状態(途上与信モデルの終端ノード)別に細かく推計できるため、新規与信モデルの終端ノードごとの貸付後2年予測貸倒率をより正確に計算できる

新規 与信ノード	途上与信ノード							計
	B.1	B.2	B.3	B.4	B.5	残なし	貸倒	
N.1	0.56	0.16	0.05	0.01	0.01	0.2	0.01	1.000
N.2	0.12	0.38	0.25	0.1	0.025	0.075	0.05	1.000
N.3	0.1	0.15	0.2	0.15	0.25	0.05	0.1	1.000

途上 与信ノード	途上与信ノード							計
	B.1	B.2	B.3	B.4	B.5	残なし	貸倒	
B.1	0.512	0.25	0.1	0.01	0.02	0.1	0.008	1.000
B.2	0.19	0.44	0.15	0.1	0.05	0.05	0.02	1.000
B.3	0.05	0.13	0.56	0.14	0.03	0.04	0.05	1.000
B.4	0.01	0.02	0.11	0.64	0.1	0.04	0.08	1.000
B.5	0.01	0.01	0.03	0.19	0.62	0.03	0.11	1.000
残なし	0.012	0.011	0.005	0.019	0.015	0.935	0.003	1.000
貸倒	0	0	0	0	0	0	1	1.000

新規与信ノードN.1の

2年予測貸倒率

$$=0.56 \times 0.008 + 0.16 \times 0.02$$

$$+ 0.01 \times 0.05 + 0.01 \times 0.08$$

$$+ 0.01 \times 0.11 + 0.2 \times 0.003$$

$$+ 0.01 \times 1$$

$$= 0.02268$$

- この仕組みを用いると直近の1年間の債権状態推移に基づき、全体債権の多年度の貸倒率の推計や収益の予測が可能になる

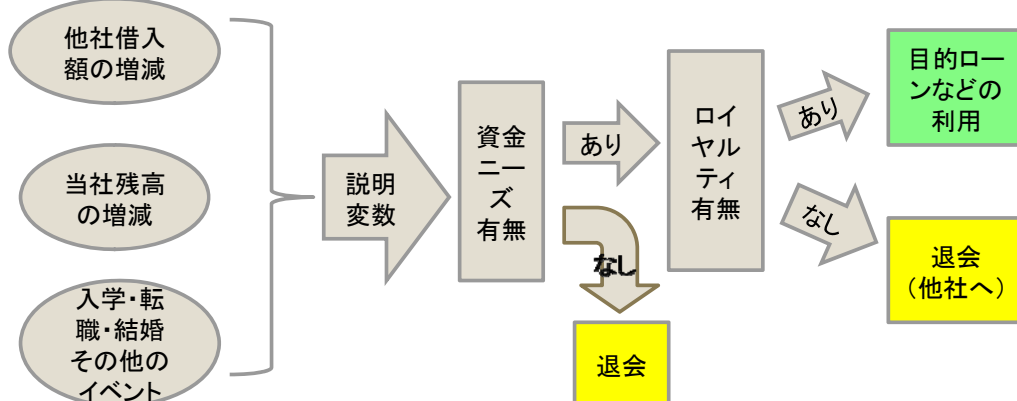
Data Mine Tech Ltd.

Data Bring New Insight to Your Business 41

無断転載を禁じます

マーケティング・退会防止

- 顧客の取引状況の推移(差分)や属性変数の変化(イベント)は、顧客の資金ニーズや他社との顧客獲得競争と関連づけられる目的変数(例えば他の証書貸付商品の利用や退会)とするマーケティングモデル構築のための説明変数として利用することができる
- このモデルは、既存顧客のさらなる資金ニーズの確信度や予測退会確率を求めるものであり、マーケティング施策の効果を計測するものではない



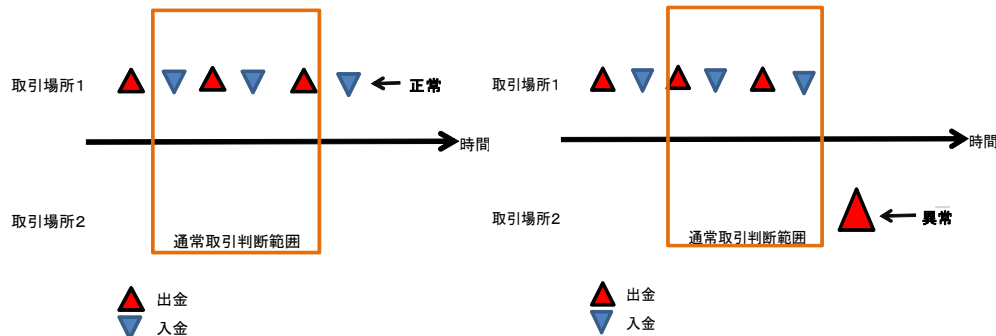
Data Mine Tech Ltd.

Data Bring New Insight to Your Business 42

無断転載を禁じます

不正利用の検知

- 顧客の取引状況、特に取引時間、取引場所および出金額の範囲などが通常取引範囲を逸脱しているとみなされる場合、他人によるカード不正利用の可能性があります、この確率を予測するモデルを事例データに基づき作成し、リアルタイムに検知する試みが行われている
- 途上与信モデルで用いる月次取引履歴の統計情報はこの通常取引範囲を規定するためのデータとして用いることができる

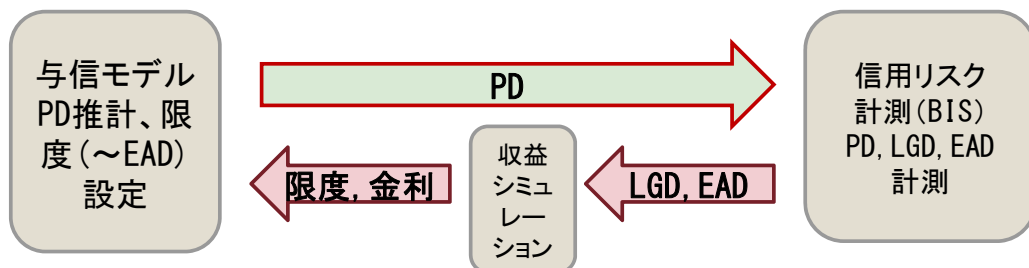


無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 43

信用リスクの計測(BIS)との関連

- 与信モデルによる個々の債権や債務者に対する予測貸倒率などはバーゼルIIの信用リスク計測手法(内部格付法)における格付プール定義に用いられ、PD(デフォルト率)計測と大きな関連がある
- LGD(デフォルト時損失率=1-回収率)、EAD(デフォルト時エクスポージャー=デフォルト時債権額+追加利用額)は与信モデルによる与信限度額設定のための収益計算シミュレーション上のパラメータに用いることができる
- このように、与信モデルは信用リスク計測(BIS)に役立ち、信用リスク計測結果は、与信モデルの限度や金利の設定に反映することにより、信用リスクの管理強化に役立つと考えられる



期待利益 = 期待収益 - 期待損失
 = 平均残高 * 貸付利率 - 限度 * 貸倒率 * (1 - 回収率)
 ～ = 平均残高 * 貸付利率 - EAD * PD * LGD

プール定義

(End)

無断転載を禁じます

Data Mine Tech Ltd.
Data Bring New Insight to Your Business 44