博士論文

機械学習を用いた健診結果予測モデルの構築および検証

市川太祐

機械学習を用いた健診結果予測モデルの構築および検証

東京大学大学院　医学系研究科

社会医学専攻 臨床情報工学教室

指導教員：小山博史教授

市川太祐

要旨

本研究では、機械学習の手法を用いた健診結果予測モデルの構築とその検証を目的とする。健康診査の非必須項目から判定される疾患の罹患の有無を必須項目から予測するモデルを構築し、性能評価を行うこととした。非必須項目としては血清尿酸検査を、そこから判定される疾患は高尿酸血症を、必須項目としては特定健診の検査項目を用いた。機械学習の手法としてはL1 正則化ロジスティック回帰、ランダムフォレスト、Gradient Boosting Decision TreeおよびStacking法を適用した高尿酸血症予測モデルを構築し、比較検証した。結果として、L1 正則化ロジスティック回帰を適用した予測モデルは他の機械学習の手法より性能が良いことを明らかにした。今後は本研究を通して得られた予測モデル構築のアプローチを用いて他の検査項目を対象とした予測モデルを構築する。

目次

第一章　　緒言

第一節　　健康診断・健康診査と予測モデル

日本においては第二次世界大戦後の結核対策を契機に、疾病スクリーニングを中心とした疾病予防施策がとられてきた(厚生労働省, 2014)。当初、疾病スクリーニングは結核等の感染症を対象としていたが、衛生状況の改善に伴い、生活習慣病を代表とする慢性疾患を対象とする健康診断・健康診査（以下、健診）へと変わっていった。この流れにおいて、医療保険各法に基づく医療保険者が行う一般健診や、労働安全衛生法（1972 年法律第 57 号）に基づく事業者が行う定期健診等、老人保健法（1982 年法律第 80 号）に基づく健診が市町村で実施されてきた。さらに近年においては「医療制度改革大綱」（2005 年 12 月 1 日 政府・与党医療改革協議会）の中で、中長期的な医療費の伸びの適正化を図る目的で糖尿病等の生活習慣病有病者・予備群を、2015 年度には 2008 年度と比較して 25 %減少させることが政策目標として掲げられた。この目標を踏まえ、2008 年 4 月から、高齢者の医療の確保に関する法律（1982 年法律第 80 号。以下「高齢者医療確保法」という。）により、医療保険者に対して、内臓脂肪の蓄積等に着目し

た生活習慣病に関する健康診査（以下、特定健診）及び特定健診の結果により健康の保持に努める必要がある者に対する保健指導（以下、特定保健指導）の実施、健診結果の電子化及び保存が義務付けられた(厚生労働省, 2013)。これにより日本においては 2008 年以降莫大な健診結果データ（以下、ヘルスデータ）が蓄積されている (Fujimori, 2016)。国策としてこれほど大規模に蓄積されている例は世界に類を見ない (Dubey et al., 2006; Kobayashi, 2008; Liu et al., 2010)。

一方、健診項目は保険者によって異なる。具体的には健診は法制度に基づき受診が必須となっている健診項目（以下、必須項目）と保険者の判断により提供している健診項目（以下、非必須項目）で構成されている。臨床的有意性をもった項目であっても、必須項目に含まれていない場合もある。

そこで本研究は非必須項目から得られる罹患の有無を必須項目から予測するモデルを構築することを考えた。非必須項目の検査を行っている保険者から得られた健診データを用いて予測モデルを構築し、構築したモデルを非必須項目の検査を実施していない協会けんぽ及び国民健康保険等の保

険者が利用することで、最低限の健診項目しか受診できない保険者の加入者に対してもコストを抑えた形で現状の疾病スクリーニングを補完できることが期待される。具体的には、予測モデルにおいて実際の健診項目の受診が必要か否かを判定し、受診勧奨を行う。

今まで、医療分野における予測モデルは前向きコホートから得られたデータから構築されることが多かった。代表的なものにはフラミンガム研究で得られたフラミンガムリスクスコアや NIPPON DATA 研究の動脈硬化リスクスコアがある (Nippon Data80 Research Group, 2006; Wilson et al., 1998)。これらは 5 年から 10 年程度の一定期間、データを収集し特定の疾病の発症リスクを統計的手法を用い算出するものであり、予測手法としては生存時間分析の手法が用いられてきた。しかし、本研究においては、多くの保険者が長期の対象集団の追跡は困難であることから、健診の非必須項目で判定される疾病の罹患有無を必須項目から予測するモデルを構築することとした。

このような予測モデル構築について、近年めざましい成果を挙げているのが機械学習である (Murphy, 2012)。医療分野においても医用画像を用いた

診断補助 (Wu et al., 2016)、血糖値のリアルタイム予測 (Zeevi et al., 2015)、タンパク質の構造解析 (Panwar et al., 2013; Panwar et al., 2014) 等様々な適用例が報告されている。

機械学習の健診データへの応用は海外では報告されている (Baba et al., 2015)。この報告ではバングラデシュにおいて 2 年にわたる巡回訪問健診を実施し収集した健診データから疾病に特化しない形で受診者の将来の健康状態を予測している。しかし、この研究は健診の非必須項目による疾病の罹患の有無の予測を目的としておらず、1 年後の受診者の総合的な健康状態を予測することで行動変容を促す意識啓発を与えるにとどまっている。また日本国内における健診データに対する機械学習の応用例も同様数年後のリスク状態を予測する研究となっている (Uematsu et al., 2017)。このように、数年後のリスク状態ではなく、健診の非必須項目から判定される疾病罹患を同年の必須項目から予測する手法の開発研究はなされていない。そこで本研究では機械学習を用いた、非必須項目から判定される疾病の罹患有無の予測を提案し、罹患予測モデルの構築とその検証を行うこととした。

本研究では罹患予測の対象を高尿酸血症とした。次節では、高尿酸血症を対象とした理由として高尿酸血症の予測の必要性、臨床的有意性について述べる。

第二節　高尿酸血症の臨床的有意性

高尿酸血症がもたらす代表的な疾患として痛風がある。痛風の本態は析出した尿酸塩結晶が起こす炎症反応であり、これに起因する関節痛（痛風発作）が代表的な症状である (Yamanaka et al., 2011)。痛風発作は患者に強い疼痛をもたらすため、労働生産性に影響するという報告もある (Kleinman et al., 2007; Smith et al., 2014)。

また高尿酸血症は高血圧、糖尿病、慢性腎臓病の独立した危険因子として報告されている (Bhole et al., 2010; Edwards, 2009; Feig, 2009; Ohta, Tsuchihashi, Kiyohara, & Oniki, 2013)。特に慢性腎臓病の新規発症に高尿酸血症が関与している (Li et al., 2014)。

以上より予測モデルを用いて高尿酸血症か否かを把握し、医療機関受診を促すことには臨床的有意性がある。

第三節　　現状の健診における血清尿酸検査の占める位置

高尿酸血症を判定するための血清尿酸検査は特定健康診査や労働安全衛生法に基づく定期健診の必須項目には含まれてはおらず、現状では任意の測定項目として提供されている。これは血清尿酸検査の臨床的有意性を否定するものではなく、従来の健診が心血管系疾患の予防を重視しており、血清尿酸検査がそぐわないことにある（厚生労働省, 2007a; 厚生労働省, 2007b）。実際に特定健診制度に向けた検討会においても「尿酸は、メタボリックシンドロームのリスクマーカーとして重要ではある」（厚生労働省, 2007a）と述べられている。また、同時期（2007 年 3 月）に行なわれた「労働安全衛生法における定期的健康診断等に関する検討会報告書」においても「血清尿酸は内臓脂肪蓄積に伴う代謝状況を反映し、内臓脂肪が蓄積した場合には尿酸合成が亢進するため、内臓脂肪症候群のリスクマーカーとして重要であるとともに、最近の知見では動脈硬化性疾患の独立したリスクファクターとしても指摘されている。このため、他の健診項目から得られる情報と併せて、脳・心臓疾患のリスクファクターの状況をより把握す

ることが可能となる。」（厚生労働省, 2007b）と報告されており、国としても血清尿酸検査の臨床的有意性を認めながらも、必須項目としては組み入れていない。

第四節　高尿酸血症を予測することの意義

このように血清尿酸値はその臨床的有意性が認められながらも、従来の健診においてはその目的から外れるとされた結果、必須項目として組み入れられることはなかった。検査項目の追加は必然的にコストの増加を伴うため非必須項目に対する費用補助の提供は加入する保険者の財政に依存する。したがって、疾病スクリーニングの機会に不平等が生じているのが現状である。しかし国が制度上定める必須項目を増やすことでこの現状への対策とするのは社会保障費が増大しつつある日本の現状を鑑みても適切な解決策とはいえない。

そこで本研究では非必須項目として血清尿酸検査を提供している保険者において、現在の高尿酸血症の罹患有無を必須項目から予測するモデルを構築する。この予測モデルを必須項目のみの提供としている保険者におい

て提供することで健診受診者における高尿酸血症のリスク層別化を可能にし、現状よりも疾病スクリーニング精度の向上が期待できる。

第二章　本研究の目的

本研究では、機械学習の手法を用いた健診結果予測モデルの構築とその性能評価を目的とする。複数の機械学習の手法を用いて現在の高尿酸血症の罹患有無を必須項目から予測するモデルを構築し、性能評価を行った上で最適な予測モデルの選択を目指す。

第三章　機械学習を用いた予測モデル

第一節　　はじめに

まず、本研究で用いる機械学習の概要について述べる。機械学習とはデータからパターンを抽出する手法である (Murphy, 2012)。機械学習研究は、歴史上は人工知能を研究する流れにおいて、機械に自動的にパターンを認識させ知識として獲得させる手段として発展してきた。データマイニングとはそれぞれの分野で用いられるアルゴリズムにおいて共通するものも多い。しかしデータマイニングはそれを用いるのが人間であり、出力結果の解釈可能性を重視するのに対し、機械学習はその歴史的経緯からその出力である予測結果の精度を重視するという違いがある。

機械学習の手法は教師あり学習、教師なし学習に大別される。データを与えてパターンを抽出する際に、あらかじめパターンについての情報が与えられている場合を教師あり学習、与えられていない場合を教師なし学習と呼ぶ。高尿酸血症の罹患有無を予測する場合で考えると、高尿酸血症の罹患有無の情報が与えられている場合は教師あり学習、与えられていない

場合が教師なし学習となる。本研究は血清尿酸検査のデータを取得を前提としているため、前者に該当する

教師あり学習もまた予測対象である目的変数と予測する根拠となる説明変数において線形的な関係を仮定する手法と、非線形的な関係を仮定する手法に大別される (Murphy, 2012)。前者の代表的な手法としてはL1 正則化ロジスティック回帰（Tibshirani, 2011)、後者には木構造を利用した決定木(Breiman et al., 1984)が挙げられる。なお、決定木は健診データのような説明変数間に交互作用があるような場合もそれを考慮した予測を行える一方、与えられたデータに過剰に適合してしまう状態、すなわち過学習を起こしてしまうため予測精度は高くない。したがって、決定木を用いた予測モデルを構築する際は決定木を発展させたランダムフォレスト (Breiman, 2001)、Gradient Boosting Decision Tree (Ye et al., 2009)が通常用いられる。

L1 正則化ロジスティック回帰は医学分野でも頻繁に用いられてきたロジスティック回帰に対してL1 正則化という機械学習のアルゴリズムを加えることで予測精度を高めた手法である。ランダムフォレストは集団学習と呼ばれる機械学習の手法の一種である。シンプルなアルゴリズムでありなが

ら予測性能が高く、並列計算であるため計算速度も速いといった理由から2005年の発表以降、機械学習分野では頻繁に用いられるようになった手法である。Gradient Boosting Decision Treeはランダムフォレストと同様に集団学習に属する手法である。近年有名になった手法であるため、ランダムフォレストよりも知名度は低い。しかし近年行われた機械学習のコンペティションで同等の予測精度であり、かつ計算速度において圧倒的に優位であるという結果を出している (Chen et al., 2015; Chen et al., 2016)。

また、ランダムフォレストやGradient Boosting Decision Tree とは異なる集団学習にStacking法がある (Wolpert, 1992)。これは事前に得られた予測モデルの予測結果を説明変数として用いて、算術平均やロジスティック回帰を適用することで新しい予測結果を得るという手法である。各予測モデルの欠点を補い合って予測性能が向上しやすいという長所があるため、複数の予測モデルを構築する際は同時に検討されている。

その他の手法としてサポートベクターマシン及び深層学習がある (Cortes et al., 1995; Hinton et al., 2006; LeCun et al., 2015)。サポートベクターマシンはランダムフォレスト以前にその予測性能の高さから機械学習分野で頻繁に

用いられた。しかし、データサイズに応じて多くのメモリを消費するというデメリットがある。実際に本研究の予備実験においてもメモリ不足により計算不能となった。そのため本研究では不適と判断した。また、深層学習は従来より予測モデルの構築に利用されてきたニューラルネットワークを大規模なもの（一般に５層以上）として設計したものである。近年の計算機の発展に伴って医学分野を含めたあらゆる分野においてその予測性能の高さで他手法を圧倒する結果を出している (LeCun et al., 2015)。健診データに対する深層学習の適用可能性について予備実験を行ったが、他手法を凌駕するような結果は得られなかった。深層学習は入力として与えた説明変数からその交互作用も加味した特徴量を生成し、それを用いて予測結果を算出するが、これは画像のように説明変数の数が数千から数万を超えるような場合にその力を発揮するといわれている。実際に医学分野においても医用画像処理において深層学習はめざましい成果を挙げている (Wu et al., 2016)。健診データの場合、経年の差分を考慮しても説明変数は 100 のオーダーであり、データの特性に深層学習が適合しなかったものと考えられ

る。以上の課題から本研究において深層学習は不適であると考え採用しなかった。

上記より予測モデル構築には Gradient Boosting Decision Tree (以下、GBDT) (Chen et al., 2016)、ランダムフォレスト (以下、RF) (Breiman, 2001)、L1 正則化ロジスティック回帰モデル (以下、LR) (Tibshirani, 2011)の 3 つ及び 3 つの予測結果を組み合わせた Stacking 法を用いることとした。次節では各手法のアルゴリズムについて説明する。

第二節　予測モデル構築に用いた機械学習のアルゴリズム

第一項　GBDT

GBDT は新しい手法ではあるものの、過去の報告より機械学習手法の中でも極めて良い結果を残しているため、本研究において採用した (Chen et al., 2016; Friedman, 2001; Ye, Chow et al., 2009)。GBDT と後述するランダムフォレストは集団学習に属する。集団学習とは決定木や線形回帰モデルのような予測力の低い予測モデルの結果を合成して予測能力の高い予測モデルを構築する機械学習の手法である。ここで用いる「予測力の低い予測モ

デル」のことを弱学習器と呼ぶ。GBDT は弱学習器として決定木を用いる。決定木は説明変数によって構成される特徴空間を再帰的に２分割することで、データを木の形で表現する。決定木は他の機械学習の手法に比べて高速にモデルを構築するのに有効であり、説明変数における外れ値の影響も受けないという特徴をもつ一方、予測精度は低い。GBDT はこの予測精度を集団学習の形で向上させている。

GBDT のアルゴリズムについて説明する（Hastie et al., 2014)。GBDT は集団学習の一種であるブースティングに属し、最終的に学習される予測モデル $f(x)$ は以下の加法モデルの形で表される。

$$f(x) = \sum_{m=0}^{M} \gamma_m I_m(x) \tag{2.1}$$

ここで、x は入力データである。M はアルゴリズムの更新回数の総数、m はアルゴリズムの更新時点のインデックス、γ_mは各更新で求められる弱学習器の重み、$I_m(x)$ は各更新でも求められる弱学習器（GBDT の場合、決定木）である。アルゴリズムを更新するたびに γ_mで重み付けされた弱学習器が $m-1$ 時点の予測モデルに加えられ、m 時点の予測モデルとして出力される（式（2.2））。

$$f_m(x) = f_{m-1}(x) + \gamma_m I_m(x) \tag{2.2}$$

ここで、重み γ_m については、式（2.3）を用いてデータ (x_i, y_i) に対する損失を最小化することで求める。

$$\gamma_m = \arg\min_{\gamma} \sum_{x_i} L(y_i, f_{m-1}(x_i) + \gamma I_m(x_i)) \tag{2.3}$$

ここでデータ (x_i, y_i) は訓練用データ集合（データ数 N）$\{(x_i, y_i)\}_{i=1,2,\ldots,N}$ に属する。損失関数 $L(y_i, f(x))$ は微分可能とし、ここではy_iと弱学習器の出力結果 $f(x_i)$ との二乗誤差 $\frac{1}{2}[y_i - f(x_i)]^2$ とする。

まず、式（2.2）において、$f_{-1}(x) = 0$、$I_0(x) = 1$として次式を用いて初期化を行う。

$$f_0(x) = \arg\min_{\gamma} \sum_{\mathrm{i}=1}^{\mathrm{N}} L(y_i, \gamma) \tag{2.4}$$

次に式（2.5）に示すように前回の更新時点（$m-1$）に求めた予測モデルで偏微分した結果 r_{im} を求め、この結果に適合するように学習した決定木 $I_m(x_i)$ を求める。

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}} \tag{2.5}$$

この決定木 $I_m(x_i)$と $m-1$ 時点での予測モデルを用いて、式（2.3）において全体の損失が最小となる γ_m を求める。ここでは直線探索（本研究で

は確率勾配法）が用いられる。γ_mで重み付けをした決定木 $I_m(x_i)$ を $m-1$ 回目時点の予測モデルに加えて、m回目の更新を終える（式（2.2））。既定の回数（M）の更新を終えた時点で学習を終了し、最終的な予測モデルを出力する。

弱学習器である個々の決定木の予測能力をコントロールするパラメータとして木に含まれるデータ数の最小値（min_child_weight）、木の深さ（max_depth）を設定する。min_child_weight の値を小さく、max_depth の値を大きくすることで決定木の表現力が上がり、個々の決定木の予測性能は上がる。一方で、データに過剰に適合してしまう弊害もあるため、最終的に良い結果が得られるとは限らない。したがって、このパラメータは複数の値を試行して最終的に予測性能が最も高くなるものを選ぶ必要がある。また、損失関数を更新していく際に、更新の程度をコントロールするパラメータとして shrinkage がある。shrinkage は 0 より大きく、1 より小さい範囲で設定でき、小さいほど更新の影響を小さく抑えられ、保守的に学習を進められる一方、学習時間は増加する。

第二項　　ランダムフォレスト

ランダムフォレストもまたGBDTと同様に性能の良い集団学習である(Breiman, 2001)。ランダムフォレストは集団学習の一種であるバギング（Breiman, 1996）を改良した手法である。バギングは訓練データから抽出したブートストラップ標本集合に対する弱学習器の構築を繰り返し、各弱学習器の結果の多数決をとることで最終的な予測結果を得る手法である。集団学習においては、統合後の予測モデルの予測性能を向上させる上で、各弱学習器の予測結果の相関が低いことが望ましい (Breiman, 2001; Hastie et al., 2014)。バギングの場合、ブートストラップ標本集合を繰り返し抽出することにより、この点に対応しているが、ランダムフォレストはブートストラップ標本集合からさらに複数の説明変数をランダムに選択するという改良を行っている。ランダムフォレストのアルゴリズムについて、0もしくは1に分類する2クラス分類器を最終的に得るものとして以下、説明する(Breiman, 2001)。

まず、構築する決定木の総数Bを設定する。

以下の手順1から3をB回繰り返す。ここでZは訓練データである。

1. Z からサイズ N のブートストラップ標本集合 Z^* を抽出する。

2. Z^*における説明変数のうち、m 個をランダムに選択する。

3. 2 で得られたデータに対して決定木を構築する。

最終的に得られた B 個の各決定木において得られる分類結果（0 もしくは 1）を平均し、ランダムフォレストによる予測結果としてクラスの所属確率を算出する。

第三項　　L1 正則化ロジスティック回帰

ロジスティック回帰は良く知られた古典的統計学アプローチである (Hosmer et al., 2005)。L1 正則化ロジスティック回帰はこのロジスティック回帰に L1 正則化項を加えたものである（Tibshirani, 2011）。

ロジスティック回帰は、以下で定義される。

$$\mathrm{P}(\mathbf{y}|\mathbf{x};\boldsymbol{\beta}) = \frac{1}{(1 + \exp(-\mathbf{y}\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}))} \tag{2.4}$$

ここで $\mathbf{y}$ は目的変数ベクトル、$\mathbf{x}$ は入力ベクトル、$\boldsymbol{\beta}$ は回帰係数ベクトル、$\mathrm{P}(\mathbf{y}|\mathbf{x};\boldsymbol{\beta})$ は $\mathbf{x}$ と $\boldsymbol{\beta}$ が与えられた際のクラス（高尿酸血症の罹患有無）への所属確率である。

$\boldsymbol{\beta}$ を求める際には以下の負の対数尤度 $f(\boldsymbol{\beta})$ が最小になるよう決定する。

$$f(\boldsymbol{\beta}) = -\sum_{n=1}^{N} \log \mathrm{P}(y_n | x_n; \boldsymbol{\beta}) \tag{2.5}$$

この時 L1 正則化ロジスティック回帰は、$f(\boldsymbol{\beta})$ に L1 正則化項を加えて以下のように最小化する。ここで N はデータ数である。

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (f(\boldsymbol{\beta}) + \lambda|\boldsymbol{\beta}|) \tag{2.6}$$

ここで λ はペナルティの強さで 0 から 1 まで値をとる。λ が 0 の場合、通常のロジスティック回帰と一致する。

L1 正則化を加えた場合、回帰係数が 0 に近いものは 0 になるという特性がある。これにより λ を設定して予測モデルを構築するとその値に応じて小さな回帰係数は 0 になるため、変数選択も同時に行えるという利点がある。通常、統計モデルにおいて変数選択を行う場合は、Akaike's Information Criterion (AIC) や Bayesian Information Criterion (BIC) といった指標に基づいて変数選択を行うが、L1 正則化ロジスティック回帰の場合は予測性能に着目してモデルを構築し、同時に変数選択も行うことができる (Chong et al., 2005)。

第四項　　Stacking 法について

Stacking 法とは集団学習の一種であり、複数の予測モデルから得られる予測結果を統合することにより新しい予測結果を算出する手法である(Murphy, 2012; Wolpert, 1992)。

Stacking 法を用いて得られる予測モデルの出力結果（予測対象とするクラスへの所属確率）$f(x)$ は以下の式から得られる。

$$f(x) = \sum_{m=1}^{M} w_m f_m(x) \tag{2.7}$$

ここでxは入力ベクトル、m は統合する予測モデルのインデックス、M は統合する予測モデルの総数、$f_m(x)$は統合する各予測モデルの出力結果、w_mは各予測モデルに対する重みである。

w_mを求める方法には算術平均、線形回帰、その他機械学習の手法等様々なものが用いられるが、必要な工数と得られる性能向上のバランスから算術平均または線形回帰を用いることが多い。高尿酸血症の罹患予測モデルにおいて算術平均を用いた場合、GBDT、ランダムフォレスト、L1 正則化ロジスティック回帰がそれぞれ出力した高尿酸血症の罹患確率の算術平均を Stacking 法による予測結果として出力する。

予測性能が求められる局面においては Stacking 法を用いた予測性能の向上が試みられることが多い (Bennett et al., 2007; Matlock et al., 2017)。

第三節　機械学習を用いた予測モデルの評価方法

機械学習を用いた予測モデルを評価する際には、データを訓練用データセットと検証用データセットの 2 つに分割する。訓練用データセットで予測モデルを構築し、検証用データセットを構築した予測モデルの性能検証に用いる。

また予測モデルの性能については ROC（Receiver Operating Characteristic）曲線を用いて評価する。ROC 曲線は機械学習の分野において予測モデルの性能評価に用いられており、特にその曲線下面積である AUC（Area Under Curve）を用いて評価することが多い (Fawcett, 2006)。

罹患有無のような二値分類を行う予測モデルの場合、結果は二値のどちらに属するかの確率として出力される。確率は連続値として示されるため、閾値を設定して二値に分類する必要がある。ROC 曲線はこの閾値を変えながら各閾値において二値分類した結果の感度及び特異度を算出することで

描出される。AUC は 0 から 1 までの値をとり、1 に近づくほど良い予測モデルとされる。

第四節　機械学習のパラメータの決定

一般に機械学習の手法は設定すべきパラメータを持つ。パラメータは予測性能が最も高くなるようなパラメータを選ぶ。パラメータの選択には通常、グリッドサーチを用いて決定する。グリッドサーチとは、パラメータ間の組み合わせを設定し、しらみつぶしに評価を行うパラメータ探索手法である (Murphy, 2012)。例えば、GBDT の予測性能を決定するパラメータとしては nrounds、shrinkage、max_depth、min_child_weight がある。この中で nrounds については多いほど、shrinkage については小さいほど、予測モデルの性能が向上することが知られている。他のパラメータについてはデータによって、その最適なパラメータが異なる。したがって、例えば max_depth については 1 から 5 までと一定の範囲を設定した上で予測モデルを構築し、その性能を比較し、最も高い性能を持つパラメータの組み合わせを採用す

る、といった対応になる。この場合の予測性能の評価については一般に AUC を用いる (Murphy, 2012)。

グリッドサーチを行う際は 10 フォールドクロスバリデーションを用いる (Murphy, 2012)。これは、一定のパラメータのもと訓練用データを 10 に等分割して、90 %のデータを用いて予測モデルを構築し、残りの 10 %のデータを用いてその予測モデルの性能を検証する手法である。

第五節　欠損値への対応

予測モデルを構築する際の欠測への対応としては一般に以下の 2 つの方針がある (Hastie et al., 2014)。

1) 訓練用データ及び検証用データから欠測値を持つデータを除いて予測モデルを構築し、予測を実行する

2) 訓練用データ及び検証用データにおいて欠測値を補完した上で予測モデルを構築する

1) の方針を採用した場合、全ての健診項目においてデータが揃っている場合にのみ予測が実行できることになり、健診受診者間で機会不平等が生

じることになる。したがって、本研究においては 2) の方針を採用し、欠測値を補完して対応することとした。予測モデルにおける欠測値の補完手法の代表的なものとしては平均値・中央値による置換、他項目を用いた予測が挙げられる (Hastie et al., 2014)。後者は欠測がある項目とその他の項目について一定の相関が見込まれる際に有効である。

本研究では健診データを扱っており、項目間の相関は見込まれる。したがって他項目を用いた予測による補完が適していると考える。なお、この場合の予測手法には、本来目的としている予測モデルとはアルゴリズム面で異なる手法を用いる必要がある。本研究ではこれまで述べた L1 正則化ロジスティック回帰、ランダムフォレスト、Gradient Boosting Decision Tree とは大きくアルゴリズムが異なる knnImpute 法を用いることとした (Troyanskaya et al., 2001)。knnImpute 法は欠測項目以外の項目を用いて各データ間のユークリッド距離を算出し、その距離が近いデータを用いて欠測値を補完する方法であり、先行研究において性能及び実行速度に優れたと報告されている (Troyanskaya et al., 2001)。

第六節　　まとめ

本章では L1 正則化ロジスティック回帰、ランダムフォレスト、Gradient Boosting Decision Tree の３種類の手法及びこの３種を組み合わせた Stacking 法について説明してきた。次章では各手法に実際のデータを適用することで高尿酸血症罹患予測モデルを構築し、性能評価を行う。

第四章　機械学習を用いた高尿酸血症罹患予測モデルの構築

第一節　　はじめに

本章では、機械学習を用いた。高尿酸血症罹患予測モデルの構築について述べる。健診データとしては健康保険組合で提供された健診メニューのうち、特定健診の検査項目及び血清尿酸値を利用する。機械学習の手法については前章で述べた 3 つの手法及び Stacking 法を比較し、性能評価を行う。

第二節　　方法

第一項　　健診データの取得

今回、予測モデルを構築するにあたり、日本の 1 つの健康保険組合（以下、健保）から検査データを取得した。本健保の被保険者は自動車販売を主体とする業務に従事する会社員で構成されている。まず 2014 年 6 月に健保との交渉を開始し、研究計画を説明した。数度の議論を経て 2015 年 4 月に研究計画について合意を得た。この際、今回取得するデータの利用目的

を本研究内のみの利用に限定した。合意を得た研究計画について東京大学医学部倫理委員会に申請を行い、2015 年 6 月 29 日に承認を受けた（承認番号 10831）。2015 年 7 月にデータ授受について協議を開始し、合意が得られた 2015 年 12 月に著者が所属する臨床情報工学教室と健保において研究計画の外部公表を実施した。最終的に 2016 年 1 月に健保より本研究で用いるデータを取得した。

被保険者で 40 歳から 60 歳の男性で服薬歴が無く、かつ 2011 年から 2013 年の３年間にわたって毎年健診を受診した集団を本研究の対象とした。女性は高尿酸血症の有病率が非常に少ないため、研究対象からは除外した。最終的に本集団の全データ件数は 61,313 人となった。

第二項　　データセットの作成

予測モデルを構築及び検証するために、本集団から得られた健診データを訓練用データセットと検証用データセットの 2 つに分割した。分割にはランダムサンプリングを用いた。それぞれ、訓練用データセットは予測モデルの構築に利用し、検証用データセットは構築した予測モデルの性能の

検証に用いた。訓練用データセットは 2011 から 2012 年度の健診データとし、検証用データセットは 2012 から 2013 年度の健診データとした。両データセットのデータにオーバーラップは無く、それぞれ独立したデータセットである。最終的に訓練用データセットには 43,524 人、検証用データセットは 17,789 人のデータとなった。全てのデータは匿名化され対応表を持たない状態で入手した。

第三項　　健診のデータ項目

健診で取得したデータ項目は年齢、Body Mass Index（以下、BMI; 単位：kg/m^2）、高血圧関連項目として収縮期血圧（単位：mmHg）と拡張期血圧（単位：mmHg）、糖尿病関連項目として空腹時血糖（単位：mg/dl）とヘモグロビン A1c（以下、HbA1c; 単位：%）、脂質障害関連項目として中性脂肪、高密度リポタンパク質コレステロール（以下、HDL コレステロール; 単位：mg/dl）、低密度リポタンパク質コレステロール（以下、LDL コレステロール; 単位：mg/dl）、肝機能障害関連項目としてγ-グルタミルトランスペプチターゼ（以下、γ-GTP; 単位：U/l）、グルタミン酸オキサロ酢酸ト

ランスアミナーゼ（以下、GOT; 単位：U/l）、グルタミン酸ピルビン酸トランスアミナーゼ（以下、GPT; 単位 : U/l）、そして血清尿酸（単位 : mg/dl）であった。

BMI を算出するための身長、体重測定、血圧測定及び血液検査については特定健診で定められた測定方法に基づき測定された (厚生労働省 , 2013）。血圧は収縮期血圧、拡張期血圧共に２回測定を行い、その算術平均値を採用した。血液検査として中性脂肪は酵素比色法・グリセロール消去を、HDL コレステロール、LDL コレステロールは直接法（非沈殿法）を、空腹時血糖はブドウ糖酸化酵素電極法、ブドウ糖酸化酵素法、ヘキソキナーゼ法、グルコキナーゼ法、ブドウ糖脱水素酵素法、HbA1c はラテックス凝集比重法もしくは不安定分画除去 HPLC 法を、GOT、GPT、γ-GTP は JSCC 標準化対応法を、血清尿酸はウリカーゼ・ペルオキシダーゼ法を用いて測定された。

血清尿酸以外の項目については 2 年分の値（前年、後年）及びその差を予測モデルの説明変数として利用した。予測モデルの目的変数は後年の血清尿酸を用いた。前年の血清尿酸は本研究の予測モデルには利用していな

い。また検査結果から計算できる脈圧（収縮期血圧と拡張期血圧の差）、GOT/GPT 比、LDL コレステロール/HDL コレステロール比も説明変数として利用した (Kawamoto et al., 2012; Kjeldsen, 2017; März et al., 2017)。腹囲についてはその測定方法については標準化されているものの、測定者による測定誤差の違いが無視できないと考え、本研究では用いないこととした (Agarwal et al., 2009; Verweij et al., 2013)。問診については他保険者と異なる独自のフォーマットであったため、予測モデルの汎用性を考慮し、本研究では用いないこととした。検査項目の一覧を表 4-1 に示す。健診における検査方法は外部機関により品質評価を受けている (Kawano, 2001)。

表 4-1 健診検査項目一覧

番号	健診検査項目	内容
1	BMI	モデルの説明変数として2年分の値を利用した。
2	収縮期血圧	モデルの説明変数として2年分の値、収縮期血圧と拡張期血圧の差、各項目の２年間の差を用いた。
3	拡張期血圧	
4	空腹時血糖	モデルの説明変数として2年分の値、各項目の２年間の差を用いた。
5	HbA1c	
6	中性脂肪	モデルの説明変数として2年分の値及び２年間の差を用いた。
7	HDLコレステロール(HDL-C)	モデルの説明変数として2年分の値、HDL-CとLDL-Cの比、各項目の２年間の差を用いた。
8	LDLコレステロール(LDL-C)	
9	γ-GTP	モデルの説明変数として2年分の値及び２年間の差を用いた。
10	GOT	モデルの説明変数として2年分の値、GOTとGPTの比、各項目の２年間の差を用いた。
11	GPT	
12	血清尿酸	モデルの目的変数として利用した。

第四項　　高尿酸血症の定義

ここで予測対象としている高尿酸血症の定義について述べる。高尿酸血症の学会基準による定義は男性では 7.0 mg/dl となっている。一方、薬物療法が開始する基準は 9.0 mg/dl となっている (Yamanaka et al., 2011)。したがって、本研究においては 7.0 mg/dl と 9.0 mg/dl の両基準を用いて予測モデルを構築し、その性能がどのように変化するか比較を行った。

高尿酸血症の定義を 7.0mg/dl とした場合、該当者は訓練用データでは 10,960 人（総数に対して 25.2 %）、検証用データでは 3,469 人（総数に対して 22.0 %）となった。高尿酸血症の定義を 9.0mg/dl とした場合、該当者は訓練用データでは 614 人（総数に対して 1.4 %）、検証用データでは 215 人（総数に対して 1.2 %）となった。

第五項　　予測モデルの構築

予測モデル構築には Gradient Boosting Decision Tree (GBDT) (Chen et al., 2016)、ランダムフォレスト (RF) (Breiman, 2001)、L1 正則化ロジスティック回帰モデル (LR) (Tibshirani, 2011)及び Stacking 法を用いることとした。

全ての予測モデルはRを用いて構築した（R Core Team, 2014）。構築に際して、GBDTに対してはxgboostパッケージ（Chen et al., 2016）を、RFに対してはrangerパッケージ（Marvin et al., 2017）を、LRに対してはglmnetパッケージを用いた（Friedman et al., 2010）。また、Stacking法はGBDT、RF、LRで得られた予測値を算術平均する形で予測を実施した。

第六項　　欠測値補完

本研究において利用した特定健診データは制度上、HbA1cと空腹時血糖を除くすべての項目の検査実施が必須となっている。HbA1cと空腹時血糖については最低限どちらか一方の項目のみ実施すればよいとされているため、欠測が生じる。欠測数は訓練用データでは、HbA1c（前年）が210件（0.5 %）、HbA1c（後年）が182件（0.4 %）、空腹時血糖（前年）が1,182件（2.7 %）、空腹時血糖（後年）が1,048件（2.4 %）、検証用データでは、HbA1c（前年）が46件　（0.4 %）、HbA1c（後年）が34件（0.3 %）、空腹時血糖（前年）が627件（5.3 %）、空腹時血糖（後年）が620件（5.3 %）となっていた。

欠測値補完は先行研究において性能及び実行速度に優れたと報告されている knnImpute 法を採用した (Troyanskaya et al., 2001)。knnImpute 法は R の caret パッケージを用いて実行した (Kuhn et al, 2013)。

第七項　　性能評価方法

予測モデルの性能については ROC 曲線を描いて評価した。ROC 曲線は機械学習の分野において予測モデルの性能評価に用いられており、特にその曲線下面積である AUC（Area Under Curve）を用いて評価することが多い (Fawcett, 2006)。ROC 曲線の各軸は研究の目的によって異なるが、本研究における ROC 曲線の縦軸は感度、横軸は 1-特異度と設定した。本研究で構築する予測モデルはそれぞれ高尿酸血症である確率を連続値として算出し、さらに高尿酸血症であるか否かの二値化を行う。その際に二値化の閾値を設定する必要がある。ROC 曲線はこの閾値を変えながら各閾値において二値分類した結果の感度及び特異度を算出することで描出される。本研究において感度は実際に高尿酸血症である者のうち高尿酸血症である予測とされた者の割合を示し、特異度は実際に高尿酸血症ではない者のうち高尿酸

血症ではないと予測とされた者の割合を示す。AUC の 95 %信頼区間は DeLong 法を採用し、R の pROC パッケージを用いることで算出した (DeLong et al., 1988; Robin et al., 2011)。なお各予測モデルを代表する感度、特異度については感度 1 特異度 1（ROC 曲線を描画する際の左上の点）にもっとも近い ROC 曲線上の点を用いた (Hajian-Tilaki, 2013)。また、構築した予測モデルを検証用データに適用した際の予測に要した時間も処理速度として測定した（単位：秒）。

第八項　計算機環境

本研究は全て MacBook（プロセッサ：1.3GHZ Intel Core m7、メモリ：8GB 1867 MHz）(Apple Inc.,Cupertino,CA)上で実行した。

第三節　　結果

第一項　　健診データの基本属性

表 4-2 には今回の対象とした健診データのデモグラフィックデータおよび臨床的特性を示す。

表 4-2　利用した健診データにおける基本属性

検査項目	訓練用データ (n=43,524)		検証用データ (n=17,789)	
	前年 (2011年) 平均値(標準偏差)	後年(2012年) 平均値(標準偏差)	前年 (2012年) 平均値(標準偏差)	後年(2013年) 平均値(標準偏差)
年齢(歳)	47.9(5.5)	-	48.1(5.2)	-
BMI (kg/m^2)	23.5(3)	23.2(3.8)	22.7(4.8)	22.9(4.7)
収縮期血圧 (mmHg)	119.5(14.2)	118.8(18.5)	115.8(24.1)	117(23.6)
拡張期血圧 (mmHg)	76.7(10.5)	76.5(13.1)	74.5(16.5)	75.1(16.1)
空腹時血糖 (mg/dl)	95.1(20.9)	94.5(22.1)	89.7(26)	90.8(26)
HbA1c (%)	5.4(0.6)	5.4(0.8)	5.2(1)	5.3(1)
中性脂肪 (mg/dl)	127.1(89.5)	123.9(88.7)	117.9(85.1)	119.1(90.7)
HDLコレステロール (mg/dl)	58.3(14.9)	57.3(15.5)	58.6(17.4)	58.4(16.8)
LDLコレステロール (mg/dl)	126(29.8)	125(31.5)	124.8(34.5)	126.1(33.5)
GOT (U/l)	23.5(11.2)	23.1(11.2)	22.4(12.5)	22.6(10.5)
GPT (U/l)	27.3(18.5)	26.6(18.6)	25.5(19)	25.3(17.1)
γ-GTP (U/l)	49(50.3)	48.3(51.1)	47(49.5)	46.9(50.2)

BMI: Body Mass Index; HbA1c: Hemoglobin A1c; HDL コレステロール: High Density Lipoprotein コレステロール; LDL コレステロール: Low Density Lipoprotein コレステロール; GOT: Glutamic Oxaloacetic Transaminase; GPT: Glutamic Pyruvate Transaminase; γ-GTP: γ-glutamyl transpeptidase;

今回予測モデルを構築する際に用いた健診データにおいて、訓練用データと検証用データに大きな偏りは認められず、予測モデル構築・検証に支障はないものと考えた。

第二項　　予測モデルのパラメータ

グリッドサーチを用いて決定した予測モデルのパラメータを表 4-3 に示す。 GBDT のパラメータとしては、木の深さとして max_depth を、繰り返しの数として nrounds を、損失関数更新をコントロールするパラメータとして eta を、木に含まれるデータ数の最小値を min_child_weight としてそれぞれ設定した。なお、本研究で用いた xgboost パッケージでは GBDT における shrinkage を eta と呼んでいる。各パラメータは max_depth においては 1 から 1 ずつ増加させて 7 まで、nrounds においては 100 から 100 ずつ増加させて 500 まで、eta においては 0.01 と 0.1 を、min_child_weight においては 1 から 1 ずつ増加させて 3 までの範囲でグリッドサーチを実施した。RF のパラメータとしては木の数として num.trees を、木の構築に用いる説明変数の数として mtry を設定した。各パラメータは num.trees においては 500 と 1000 を、mtry においては 1 から 1 ずつ増加させて 5 までの範囲でグリッドサーチを実施した。LR のパラメータとしては罰則パラメータとして lambda を

設定した。lambda は 0.001 から 0.02 まで 0.001 ずつ増加させてグリッドサーチを実施した。

表 4-3. 最終的に決定した予測モデルのパラメータ

予測モデル	パラメータ	高尿酸血症の定義	
		7mg/dl	9mg/dl
GBDT	max_depth	5	3
	nrounds	500	500
	eta	0.1	0.1
	min_child_weight	1	1
RF	num.trees	1000	1000
	mtry	2	2
LR	lambda	0.001	0.006

GBDT: Gradient Boosting Decision Tree; RF: ランダムフォレスト; LR: L1 正則化ロジスティック回帰

第三項　　予測モデルの性能評価

AUC、感度、特異度、処理速度（単位：秒）で評価した 5 つの予測モデルの性能を表 4-4、4-5 に示す。

表 4-4 予測モデルの性能比較（高尿酸血症の定義 7mg/dl 以上の場合）

データセット	手法	AUC	95%CI (lower)	95%CI (upper)	特異度	感度	処理速度（秒）
訓練用	GBDT	0.91	0.91	0.91	0.84	0.81	−
	RF	0.66	0.66	0.67	0.69	0.54	−
	LR	0.70	0.69	0.70	0.61	0.68	−
	STACK	0.85	0.85	0.85	0.76	0.77	−
検証用	GBDT	0.70	0.69	0.71	0.66	0.63	0.70
	RF	0.66	0.65	0.67	0.69	0.55	7.24
	LR	0.69	0.69	0.70	0.67	0.62	0.10
	STACK	0.70	0.69	0.71	0.68	0.61	9.21

GBDT: Gradient Boosting Decision Tree; RF: ランダムフォレスト; LR: L1 正則化ロジスティック回帰; STACK: Stacking 法;

表 4-5 予測モデルの性能比較（高尿酸血症の定義 9mg/dl 以上の場合）

データセット	手法	AUC			特異度	感度	処理速度（秒）
			95%CI (lower)	95%CI (upper)			
訓練用	GBDT	0.98	0.97	0.98	0.91	0.93	−
	RF	1.00	1.00	1.00	1.00	1.00	−
	LR	0.75	0.73	0.77	0.82	0.57	−
	STACK	1.00	1.00	1.00	1.00	1.00	−
検証用	GBDT	0.76	0.72	0.79	0.67	0.72	0.38
	RF	0.78	0.75	0.81	0.75	0.65	7.31
	LR	0.80	0.77	0.83	0.78	0.68	0.10
	STACK	0.78	0.74	0.81	0.70	0.71	8.57

GBDT: Gradient Boosting Decision Tree; RF: ランダムフォレスト; LR: L1 正則化ロジスティック回帰; STACK: Stacking 法;

図 4-1、図 4-2 に高尿酸血症の定義を 7mg/dl 及び 9mg/dl に設定した場合に検証用データセットで算出した予測モデルの ROC 曲線を示す。

高尿酸血症の定義を 7mg/dl とした場合、各予測モデルの AUC は検証用データセットにおいて GBDT で 0.70[95 % CI: 0.69–0.71]、RF で 0.66 [95 % CI: 0.65–0.67]、LR で 0.69 [95 % CI: 0.69–0.70]、STACK で 0.70 [95 % CI: 0.69–0.71] であった。特異度は RF が最も高く、次いで STACK、LR、GBDT の順であった。感度は GBDT、LR、STACK がほぼ同等で、RF がそれに次いだ。

高尿酸血症の定義を 9mg/dl とした場合、各予測モデルの AUC は検証用データセットにおいて GBDT で 0.76 [95 % CI: 0.72–0.79]、RF で 0.78 [95 % CI: 0.75–0.81]、LR で 0.80 [95 % CI: 0.77–0.83]、STACK で 0.78 [95 % CI: 0.74–0.81]であった。特異度は LR が最も高く、ついで RF、STACK、GBDT の順であった。感度は GBDT と STACK がほぼ同等で、LR、RF の順に高い値となっていた。

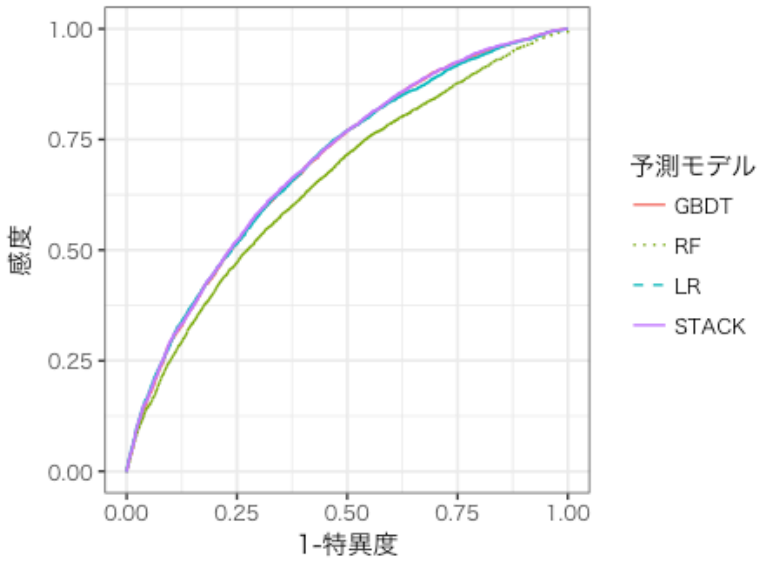

図 4-1. 予測モデルの ROC 曲線（血清尿酸の閾値 7mg/dl の場合）

GBDT: Gradient Boosting Decision Tree; RF: ランダムフォレスト; LR: L1 正則化ロジスティック回帰;

STACK: Stacking 法;

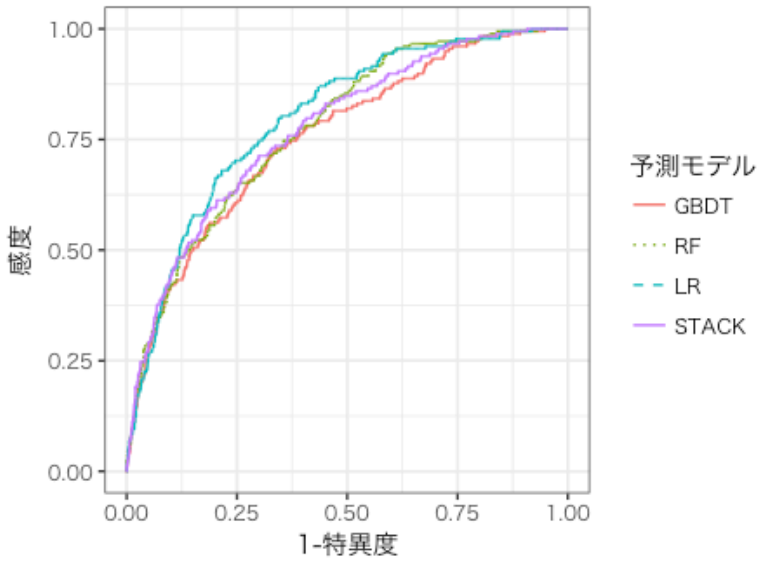

図 4-2. 予測モデルの ROC 曲線（高尿酸血症の定義 9mg/dl の場合）

GBDT: Gradient Boosting Decision Tree; RF: ランダムフォレスト; LR: L1 正則化ロジスティック回帰;

STACK: Stacking 法;

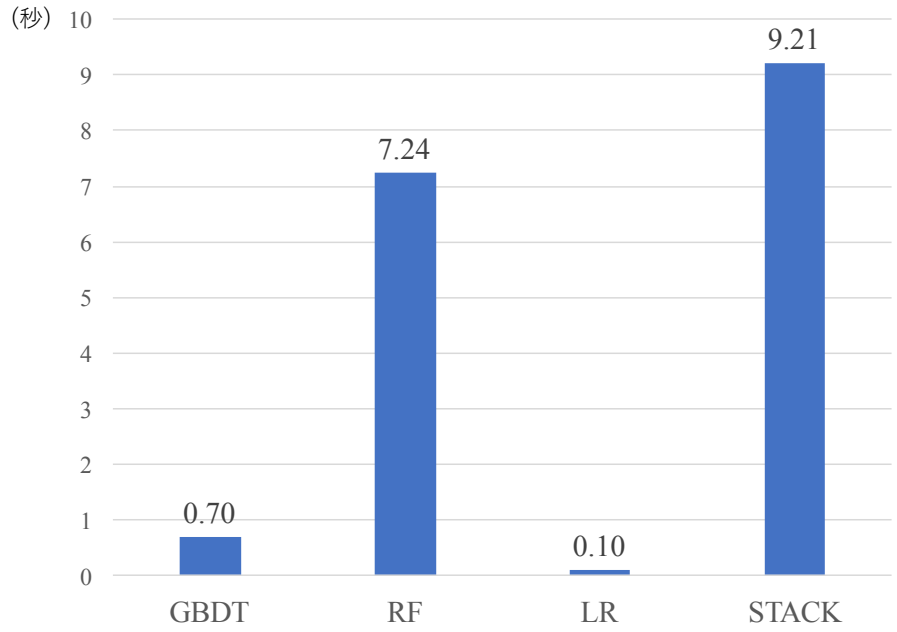

図 4-3. 予測モデルの処理速度（高尿酸血症の定義 7mg/dl の場合）

GBDT: Gradient Boosting Decision Tree; RF: ランダムフォレスト; LR: L1 正則化ロジスティック回帰;

STACK: Stacking 法;

註: MacBook（プロセッサ：1.3GHZ Intel Core m7、メモリ：8GB 1867 MHz）（Apple Inc.,Cupertino,CA）上で実行

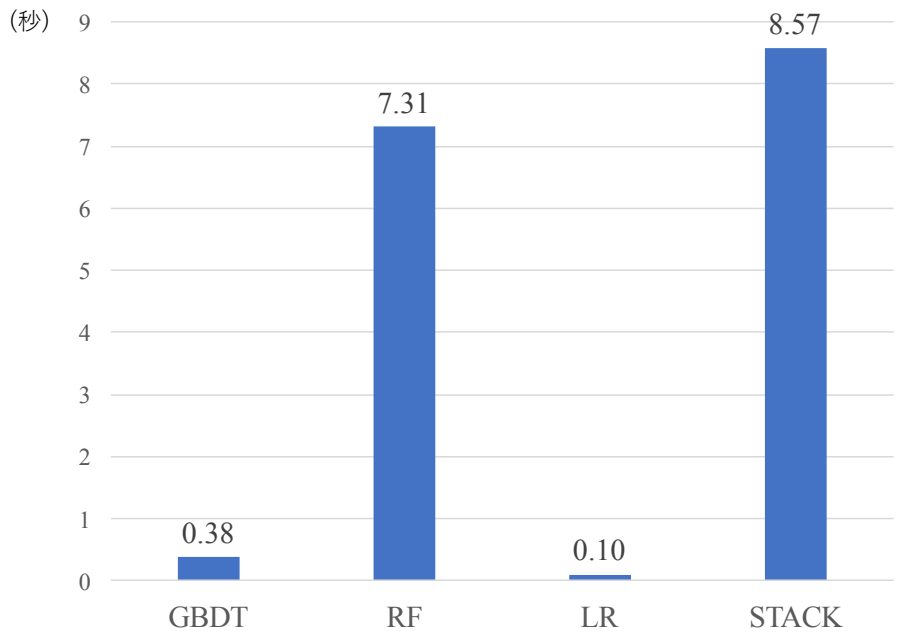

図 4-4. 予測モデルの処理速度（高尿酸血症の定義 9mg/dl の場合）

GBDT: Gradient Boosting Decision Tree; RF: ランダムフォレスト; LR: L1 正則化ロジスティック回帰;

STACK: Stacking 法;

註: MacBook（プロセッサ：1.3GHZ Intel Core m7、メモリ：8GB 1867 MHz）（Apple Inc.,Cupertino,CA）上で実行

図 4-3、図 4-4 に高尿酸血症の定義を 7mg/dl 及び 9mg/dl に設定した場合に検証用データセットで算出した予測モデルの予測実行時の処理速度（秒）を示す。

訓練用データを用いて構築した予測モデルを検証用データに適用した際の処理速度は、高尿酸血症の定義を 7mg/dl とした場合、9mg/dl とした場合のいずれにおいても LR が最も速く、GBDT、RF、STACK の順となった。

第四節　考察

高尿酸血症の定義を 7mg/dl 、9mg/dl とした場合の予測モデルにおいて AUC が最高になった手法は前者が GBDT 及び STACK（ともに AUC 0.7）で後者が LR（AUC 0.8）であった。医学分野において予測モデルの AUC の目安は 0.9 以上が high accuracy、0.7 以上 0.9 未満が moderate accuracy、0.5 以上 0.7 未満が low accuracy とされている (Fischer et al., 2003)。この基準に照合すると 7mg/dl の場合は low accuracy、9mg/dl は moderate accuracy となり、後者は予測モデルとして妥当な予測性能を有していると判断した。これは投薬治療の対象となる 9mg/dl 以上の高尿酸血症においては、他健診項目においても高血糖、高血圧等のリスク因子が認められることが予想され、結果として高尿酸血症か否かの識別境界が引きやすかったものと考える。

今回設定した高尿酸血症の定義に対し、機械学習を用いて構築した 4 つの予測モデルの間ではその性能に大きな差は認められなかった。一般に集

団学習を用いた予測モデルは仮説空間が多岐にわたる、つまり説明変数の数が膨大で、その説明変数間に複雑な交互作用が認められるような場合に他手法を凌駕する性能を示すことが報告されている (Yang et al., 2010)。今回、決定木を弱学習器として用いた RF と GBDT の 2 つの手法と LR との間に大きな差が認められなかったことから、健診データにはそのような特性が認められなかったものと考える。さらに、交互作用が認められる場合は、クロスバリデーションを用いて決定した木の深さのパラメータ（GBDT の場合は max_depth、RF の場合は mtry）に反映される。AUC が高かった 9mg/dl の場合で GBDT の木の深さは 3、RF の木の深さは 2 であったことから、複雑な交互作用を仮定しない予測モデルの方が予測性能が高かったことを示している。

訓練用データセットの AUC と検証用データセットの AUC を比較すると、高尿酸血症の定義を 7mg/dl 、9mg/dl とした場合のいずれにおいても前者の AUC が高い傾向が認められた。これは訓練用データセットに過剰に適合した予測モデルが構築された結果、そのモデルの汎用性が失われた状態と考えられ、機械学習における過学習と呼ばれる現象である (Murphy, 2012)。特

にGBDTは過学習傾向が強く、高尿酸血症が7 mg/dlの場合で0.91から0.70、9 mg/dlの場合で0.98から0.76へと他の予測モデルに比べて大きくAUCが低下している。GBDTはRF等の他の集団学習に比べて予測性能が高い一方、過学習しやすいとも報告されており、本結果とも一致している (Natekin et al., 2013; Sewell, 2011)。一方で、LRは7 mg/dlの場合で0.70から0.69、9 mg/dlの場合で0.75から0.80へとAUCにおいてわずかな低下、もしくは増加が認められており、過学習を起こしていない。この点からもLRが望ましい予測モデルといえる。

本研究における予測モデルは高尿酸血症の測定対象者を絞り込むためのスクリーニングとして位置付けており、その場合、特異度よりも感度を優先すべきものと考える。感度が最大となる予測モデルは7 mg/dl、9 mg/dlのいずれの場合もGBDTだった。ただし7 mg/dlの場合、GBDTとLRの感度の差は0.01と小さい。また、9 mg/dlの場合は、ROC曲線上でGBDTと同じ特異度0.67で比較すると、LRの感度はGBDTよりも高い。今回、予測モデルを代表する感度、特異度については、感度1特異度1にもっとも近いROC曲線上の点を利用したため、結果としてLRは特異度寄りの点が代

表値になったものと考える。以上より感度、特異度の面からは GBDT、LR が望ましいモデルと言える。なお、予測モデルの利用を考慮した評価という観点では意思決定時の閾値を考慮すべきという議論もある (Vickers et al., 2006; Steyerberg et al., 2010; Moons et al., 2015)。この議論の中では意思決定時の閾値を考慮した評価として Decision Curve Analysis が提唱されている (Vickers et al., 2006)。予測モデルが出力する確率値を医師が意思決定に利用する際、確率値が 80 %以上であれば治療し、それ未満であれば治療しない、というように意思決定に関する閾値を各医師が持っているはずである。この閾値に応じて、真陽性率及び偽陽性率から算出される Net benefit がどのように変化するか評価する感度分析が Decision Curve Analysis である。予測モデルを実際の施策に適用する段階ではこのような評価も必要であると考える。

予測の処理速度という面からは機械学習を用いた予測モデル間では LR が最も速いという結果になった。RF 及び GBDT は弱学習器で一旦予測を実行した上でその結果を統合する必要がある集団学習であることを考えると、

一つの学習器のみで予測実行が可能な LR の実行速度が速いことは妥当な結果といえる。

また、本章では予測モデルの予測性能を改良すべく Stacking 法を用いた。しかし、Stacking 法を適用した予測モデルは適用しなかった場合と比較して改善は認められなかった。これは各予測モデルが結果として相補的な関係に無かったことに起因することが考えられる。実際に ROC 曲線を確認すると各手法では ROC 曲線はほぼ一致していた。

第五節　まとめ

本章では機械学習を用いた高尿酸血症罹患予測モデルを構築した。今回用いた手法間において予測性能に差は無く、また少ない説明変数で予測が実行できる、過学習を起こしにくい、という点から予測モデルの構築に用いる手法としては L1 正則化を用いたロジスティック回帰（LR）が最適であることが示唆された。

なお、予測モデルの改善の方向性としては予測性能の改善の他に、現状の予測性能を保持したままモデル構築に必要なデータサイズを削減する手法がある。次章では本手法を用いた予測モデルの改善について述べる。

第五章 学習データサイズの削減

第一節　　はじめに

前章では機械学習を用いた高尿酸血症罹患予測モデルとしてL1正則化を用いたロジスティック回帰の優位性を示した。しかし、さらに改良の方向性として、現状の予測性能を保持したままモデル構築に必要なデータサイズを削減した効率的な予測モデルの構築が考えられる。

通常の健診データでは、時として欠測が発生し、説明変数において十分なデータが揃わない可能性も考えられる。第四章ではこのような場合の対処として欠測値の補完を実施した。一方で、予測性能は保持したままで必要な説明変数は削減することができれば補完する変数の数も減らせるため実運用上望ましい。

本章では、必要なデータサイズを削減した予測モデルという観点から、現状の性能を保持したデータ面でより効率的な予測モデルの構築を検討することとした。削減の方針としては 1) 説明変数の削減と 2) データ数の削減が挙げられる。データを行列として捉えると、1) は列方向のデータサイズ削減、2) は行方向のデータサイズ削減といえる。

次節では検証手法について述べる。

第二節　方法

第一項　説明変数の削減

あるデータを説明する際に用いる統計モデルにおいて、説明変数の削減として変数選択を行う場合は、AIC や BIC といった指標に基づいてデータに対するモデルの当てはまりを確認しながら変数選択を行う。一方、予測モデルの場合は、検証用データセットにおける予測性能が指標となる。本研究では予測性能の指標として AUC を用いており、この AUC が最も高くなるような予測モデルにおける説明変数が必要な説明変数といえる。前章で用いた LR においては予測モデル構築の際に小さな標準偏回帰係数は 0 に

なるという性質を持ち、結果として説明変数が削減される。つまり、LR の場合は予測性能に着目してモデルを構築し、同時に変数選択も行うことができる (Chong et al., 2005)。したがって、本章では説明変数の削減結果として前章で最終的に得られた LR の説明変数の内容及び個数について検討する。なお、高尿酸血症の閾値については前章において moderate accuracy の予測性能となった 9mg/dl を用いることとした。

第二項　　データ数の削減

必要データサイズの削減への対応としては、サンプリング法（以下、サンプリング）を用いる。サンプリングは対象データから一定の基準のもと、またはランダムにデータを抽出することである。予測モデルにおいて必要データサイズの削減を目的とする場合のサンプリングを行う際は、データを多数派データと少数派データの２つに分けて考える。少数派データは予測モデルの予測対象とする本研究の場合、少数派データは高尿酸血症患者のデータであり、多数派データは非高尿酸血症患者のデータである。多数

派データと少数派データのどちらに着目するかによって、サンプリングは以下の 3 つに分類される (Rahman et al., 2013)。

1) オーバーサンプリング

2) アンダーサンプリング

3) 1 と 2 を組み合わせた方法（以下コンビネーション法と呼ぶ）

1) のオーバーサンプリングは、訓練用データにおける少数派データを繰り返し抽出する方法である。本研究においても高尿酸血症患者（血清尿酸 9mg/dl 以上）は 1.4 %と少ない。したがって、オーバーサンプリングを行う際は、再抽出を許す形で抽出することになる。

2) のアンダーサンプリングは、1) とは逆に多数派データから目的する数までデータを繰り返し抽出する方法である。第四章で述べたように多数派データが占める割合が 98.6 %と大きいため、アンダーサンプリングを行う際に再抽出を許さない形の抽出も可能となる。

3) のコンビネーション法はオーバーサンプリングで少数派データを増やしつつ、アンダーサンプリングで多数派データを減らすという方法であり、代表的なものとして SMOTE 法、ROSE 法等が挙げられる (Blagus et al., 2012; Dang et al., 2013; Menardi et al., 2012; Ramezankhani et al., 2016)。

本研究においては訓練データに対してアンダーサンプリング法を適用した (Drammond et al., 2003)。その理由はこれまで本研究の様な医療データへの適用において、本手法はオーバーサンプリング法及びコンビネーション法より良い結果が得られるとされていることによる (Blagus et al., 2012; Domingos, 1999; Drammond et al., 2003)。アンダーサンプリングは、ランダムにアンダーサンプリングするランダムアンダーサンプリング法を採用した。

訓練用データにおいて高尿酸血症患者データが占める割合を 10 %から、非高尿酸血症患者データと同数の 50 %まで 10 %刻みで変化させ、予測性能がどのように変化するかを検証することとした。予測性能評価には AUC、F 値を用いた。F 値はトレードオフの関係にある感度と陽性的中率を統合して評価するものであり、２指標の調和平均として以下の式で算出した。

$$F\text{値} = \frac{2 \cdot \text{感度} \cdot \text{陽性的中率}}{\text{感度} + \text{陽性的中率}} \tag{5.1}$$

また高尿酸血症の閾値については前章において moderate accuracy の予測性能となった 9mg/dl を用い、予測モデルの構築には LR、GBDT、RF を用いてその結果を比較した。

第三節　結果

第一項　説明変数の削減

LF を用いた変数選択結果を表 5-1 に示す。全説明変数 39 のうち、標準偏回帰係数の絶対値が 0 以上となった説明変数の数は 5（BMI（後年）、中性脂肪（後年）、γ-GTP（後年）、中性脂肪（後年）、GOT（後年）、拡張期血圧（後年））であった。標準偏回帰係数の絶対値が最大だったのは BMI（後年）であり、最小は拡張期血圧（後年）であった（表 5-1）。

表 5-1　LR による変数選択結果

説明変数	標準偏回帰係数
BMI（後年）	0.23
中性脂肪（後年）	0.16
γ-GTP（後年）	0.09
GOT（後年）	0.03
拡張期血圧（後年）	0.01

第二項　　データ数の削減

ランダムアンダーサンプリングを用いて訓練用データにおける多数派データを削減した。その場合の LR、GBDT、RF における評価指標の変化を表 5-2、表 5-3、表 5-4 に示す。AUC の変化は各手法共に±0.01 の範囲の変化にとどまっていた。同様に F 値についても大きな変化は認められなかった。

表 5-2　予測モデルにおける評価指標の変化（LR）

総データ数に占める少数派データの割合（%）	AUC			感度	陽性的中率	F値	多数派データのデータ数
		95%CI 下限	95%CI 上限				
サンプリング前	0.80	0.77	0.83	0.68	0.07	0.12	42910
10	0.80	0.77	0.83	0.75	0.06	0.11	5526
20	0.80	0.77	0.83	0.76	0.07	0.13	2456
30	0.80	0.77	0.83	0.75	0.06	0.11	1432
40	0.79	0.76	0.83	0.75	0.06	0.11	921
50	0.79	0.75	0.82	0.78	0.06	0.11	614

表 5-3　予測モデルにおける評価指標の変化（GBDT）

総データ数に占める少数派データの割合（%）	AUC			感度	陽性的中率	F値	多数派データのデータ数
		95%CI 下限	95%CI 上限				
サンプリング前	0.76	0.72	0.79	0.72	0.05	0.09	42910
10	0.76	0.73	0.80	0.67	0.05	0.10	5526
20	0.76	0.72	0.79	0.76	0.05	0.09	2456
30	0.75	0.71	0.79	0.66	0.05	0.10	1432
40	0.75	0.72	0.79	0.70	0.05	0.09	921
50	0.75	0.72	0.79	0.67	0.05	0.10	614

表 5-4　予測モデルにおける評価指標の変化（RF）

総データ数に占める少数派データの割合（%）	AUC			感度	陽性的中率	F値	多数派データのデータ数
		95%CI 下限	95%CI 上限				
サンプリング前	0.80	0.77	0.83	0.84	0.05	0.09	42910
10	0.79	0.76	0.83	0.80	0.05	0.09	5526
20	0.80	0.77	0.83	0.80	0.05	0.10	2456
30	0.80	0.77	0.83	0.87	0.05	0.09	1432
40	0.81	0.78	0.83	0.74	0.06	0.11	921
50	0.80	0.77	0.83	0.84	0.05	0.09	614

第四節　　考察

LR による変数選択結果は全 39 の説明変数のうち、5 つのみが選択されるという結果であった。これらの選択された説明変数はいずれも高尿酸血症の罹患を予測したい年と同年の検査項目であり、前年の検査項目は選択されなかった。本研究においては健診結果の経年変化も考慮して予測モデルに説明変数として組み入れたが、高尿酸血症予測モデルにおいては単年度の健診結果のみで予測が可能という結果となった。

アンダーサンプリングを用いて多数派データを削減した結果の AUC の変化を確認した結果、LR、GBDT、RF 共に多数派データのデータ数を元の 42,910 から少数派データとの比が 1 対 1 になる 614 まで削減しても AUC、F 値に大きな変化は認められなかった。多数派データを元のデータの 1.4 %(= 614 / 42,910)まで削減しても予測性能には影響は無かったことになる。LR の基礎手法であるロジスティック回帰が統計モデルとして用いられることの多い症例対照研究では、多数派データと少数派データの比が 1:1 であることが望ましいとされており、今回得られた LR の結果はこれらの報告と一致する（Lewallen et al., 1998; Hennessy et al., 1999）。一方で、GBDT 及

び RF においても予測性能に大きな変化は認められなかった。これは多数派データの多くが予測性能には影響を及ぼさないデータであることが示唆される。なお、近年の報告ではアンダーサンプリングの適用は予測モデルの性能を改善するとされているが、本研究においては改善にまでは至らなかった (Lusa et al., 2012; Menardi et al., 2012)。

第五節　まとめ

本章では、必要なデータサイズの削減の観点から現状の性能を保持したデータ面でより効率的な予測モデルの構築を検討した。説明変数の削減については前年の健診結果は不要であるという結果になった。また、データ数の削減については多数派データの削減が有効であるという示唆が得られた。予測モデルを構築する際は、データのサイズばかりが着目されがちだが、真に解析に必要なデータとは何かに着目するアプローチとして本章の検討結果は実用に資するものと考えられる。

第六章　考察

第一節　　予測モデルの構築

本研究では疾病スクリーニングの受診機会を平等にする目的で、検診の必須項目から非必須項目から判定される疾患の罹患有無の予測方法について提案し、具体例として高尿酸血症罹患予測モデルを構築した。まず、機械学習の手法で GBDT、RF、LR、STACK を比較検証し、LR の優位性を示した。さらに予測性能を保持したままのデータ削減については多数派データの削減が有効であることも示した。

以上のアプローチを整理すると図 7-1 に示すような流れとなる。データ取得は自身でデータを取得する場合以外は保険者等とのデータ保有者との交渉から始まる。この際、後の節で述べるような予測モデルの知的財産権について整理が必要となる。次の段階としてモデル構築を行う。ここでは複数の手法を比較し、クロスバリデーションを用いて手法のパラメータを設定する。なお、本研究では予測モデルを構築した後に多数派データの削減等のモデル構築に必要なデータサイズの検討を行なったが、これはモデル構築の段階で実施しても問題ないものと考える。本研究で得られた知見に

基づくと、少数派データと同数まで多数派データを削減したところから始めて、予測性能を確認しながら多数派データを増やすというアプローチが適切といえる。最後は構築した予測モデルの効率化を行う。

モデル構築の段階で目指すべき予測精度としては AUC 0.8 が一つの目安となる。

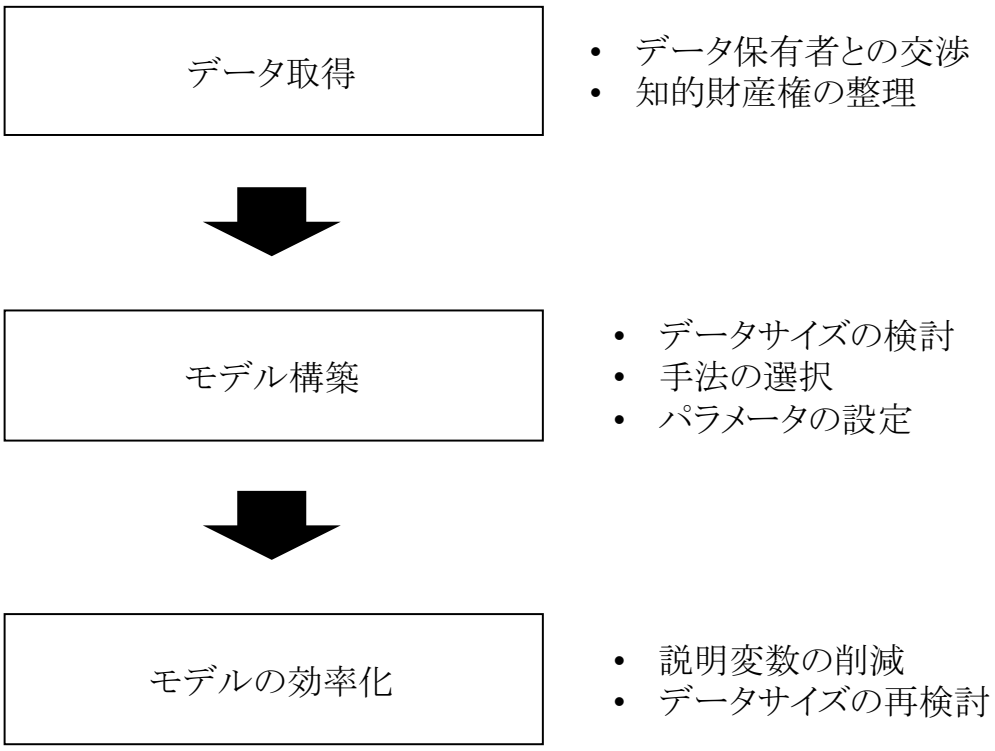

図 6-1　本研究で確立した健診結果予測モデル構築フロー

本研究の健診制度に対する貢献の可能性を述べる。健診はその有効性が議論されているが、これまで日本に根付いてきた制度でもあり大きな制度変更はむずかしく、ほぼ現状維持の形で制度は存続していくものと予想される。一方で、社会保障費は増大しており、これまでの個人への費用補助を前提とした制度の場合、費用補助の対象となる項目はより削減されることが予想される。費用補助が無い項目を個人が自主的に受診することは少

ないと考えられるので結果として受診機会の不平等が拡大することが予想される。したがって、従来の健診から低費用の効率的な健診へと移行して行く必要がある。2008 年度から開始した特定健診制度をきっかけに日本において健診データの電子化が進み、データ活用の機会が増えてきた。健診項目は互いに一定の相関があり、本研究では高尿酸血症罹患を別の項目から予測可能であることを示した。このような予測モデルを活用することで、限定された項目から他の非必須項目を予測し、その分の費用を抑えるという施策の実現可能性が高まる。また、罹患予測モデルの構築により健診項目の必要性についての判断にも利用できる。本研究における高尿酸血症のように他項目から罹患有無の予測が可能であれば非必須項目とする判断の根拠になりうる。

第二節　施策導入上の課題

本研究成果を実際の施策に導入する際には本研究における 3 つの課題が挙げられる。

1. 予測モデルの性能の向上

2. 費用対効果

3. 構築した予測モデルの知的財産権面での配慮

第一項　　予測モデルの性能の向上

本研究で得られた高尿酸血症の定義 9mg/dl の高尿酸血症予測モデルの性能は、最終的に採用した LR において、λ=0.006 のもと、AUC0.80、特異度 0.78、感度 0.68 だった。AUC については moderate accuracy という判定にはなったが感度、特異度という面からみると性能の向上の余地がある。一方、本研究では機械学習の手法間で予測性能に差はなく、データ面においては多数派データを全体の 1.4 %まで削減しても予測性能に影響を与えないことが確認された。したがって、性能向上の方向性としては少数派データ、つまり高尿酸血症該当者のデータを増やすことが考えられる。本研究では１つの健康保険組合のみからデータを取得したが、複数の健康保険組合からデータを取得することで moderate accuracy 以上の性能の向上が期待される。

第二項　　費用

本研究を施策に活用するにあたっては費用の詳細な試算が必要である。本研究は予測モデルの構築を主目的とし、費用についての精緻な解析は今後の検討課題とした。

通常、血清尿酸は痛風を診断する際に用いられる。日本の健康保険システムにおいては、血清尿酸の測定には保険点数 11 点のコストがかかる (Ministry of Labor and Welfare, 2015a)。また特定健康診査は毎年 5300 万人が受診している (Ministry of Labor and Welfare, 2015b; Ng et al., 2015)。したがって、健診受診時に全員に血清尿酸を測定した場合、約 58.3 億円のコストがかかることになるが、予測モデルを活用することで、このコストをかけずに疾病スクリーニングが可能となる。この際、予測モデルの構築・運用費用については別途検討が必要である。なお、算出した費用をもとに費用対効果を算出する場合は、血清尿酸の直接測定と予測モデルによる推定についてスクリーニングとしての精度面における比較検討が必要と考える。

第三項　　構築した予測モデルの法律面での配慮

3 つめの課題として一つの集団のデータを用いて構築した予測モデルを他の集団へ適用する際の法律面での配慮が挙げられる。この点について 1) 予測モデルに関する権利、2)予測モデルの利用に伴う法的責任の 2 つにおいて論じる。

まず予測モデルに関する権利について述べる。機械学習を用いた予測モデルは予測モデルを実装したプログラムとパラメータから構成されるが、これが法律上どのような権利を発生させるかについては明確ではない。国の検討会においても「学習済みモデルは、『AI のプログラムとパラメータの組み合わせ』であることから、現行知財制度上、著作権法の要件を満たせば「プログラムの著作物」として保護される可能性がある。しかし、パラメータが AI プログラムと別に保持されている場合に、『AI のプログラムとパラメータの組み合わせ』が著作権法上の『プログラム』に該当するのかどうか必ずしも明確ではないと考えられる」と報告されている (首相官邸, 2017)。

したがって、今後、本研究で用いたアプローチに基づき、予測モデルを構築する際は構築後の使用範囲まで考慮した上で、予測モデルに関わる権利関係について知的財産権同様にデータ提供者である保険者と予測モデル作成者間の契約の中で明確にする必要があるものと考える。例えば、保険者が自身で収集したデータを用いて予測モデルを構築し、自身の施策のために利用するのであれば問題ないが、外部業者が保険者よりデータを収集して、そこで構築したモデルを他の保険者に適用するような場合はデータ収集時点であらかじめ利用範囲についての合意形成が必要となる。

次に予測モデルの利用に伴う法的責任について述べる。本研究では非必須項目から判定される疾病の罹患有無について予測を行った。予測結果は健診受診者に対して必須項目の健診結果と併せて提示する形を想定している。これは診断に類するものであり、「医薬品、医療機器等の品質、有効性及び安全性の確保等に関する法律」（以下、薬機法）の規制対象となる医療機器に該当する可能性がある。2014 年の薬機法改正により、従来のカテーテルのような医療機器に加えて、予測モデルのようなプログラム単体も医療機器に該当するようになった。国は、本研究で扱った予測モデルが

Artificial Intelligence（以下、AI）に属するものとした上で、2018 年中までに診断・治療支援を行う AI の医師法上の取扱を明確化し、AI が医療機器に該当するのか基準を明確化する (厚生労働省, 2017a)としている。今後、予測モデルを実用化し、健診を補完するものとして利用する場合はこのような規制も考慮する必要がある。

第三節　本研究の限界

本研究の限界としてデータの問題がある。本研究では 40-60 歳の男性かつ非服薬者で 3 年間連続して健診を受診した健康保険被保険者のデータを用いて予測モデルを構築した。したがって、構築した予測モデルは女性や 60 歳を超える高齢者や 40 歳未満の若年者について適用できない。また本研究においては非服薬者のみを対象としている。これは本研究では服薬の有無については確認可能だったものの、具体的な薬剤名については確認できなかったという理由による。薬剤によっては血清尿酸の低下作用を持つものがある。例えば血圧降下薬のアンジオテンシン II 受容体拮抗薬（以下、ARB）の一種であるロサルタンは尿酸低下作用があると言われている（Yamanaka

et al, 2011）。今後、服薬の有無も予測モデルの説明変数として利用する際は、具体的な薬剤名も併せて利用すべきものと考える。また 3 年間連続受診者のデータを利用している点については、高尿酸血症の罹患者数を少なく見積もる方向へのサンプリングバイアスがある。実際、国民健康保険における健診未受診者には心血管疾患の高リスク者が多くを占めるという報告がある (Ikeda et al., 2005)。しかし、国民健康保険の場合、健診受診率は 36.3 %と低い一方（厚生労働省，2017b）、当該健康保険組合の被保険者の特定健診受診率は 99.2 %（健康保険組合非公開資料より）と非常に高いため、サンプリングバイアスの影響は小さいものと考える。

また、本研究においては予測モデルの説明変数として問診のデータは用いていない。ここで、説明変数の削減を行なった際に選択された 5 つの説明変数の回帰係数を確認すると BMI、中性脂肪、γGTP、GOT、拡張期血圧 という順であった。この中で BMI、中性脂肪は食習慣を、γGTP、GOT は飲酒習慣を強く反映するものと考えられ、これらに関連する問診項目を加えることで予測モデルの性能が向上する可能性がある (Adachi et al., 2013; Muramoto et al., 2014; Poobalan et al., 2004)。

第七章　結語

本研究では保険者より収集した健診データに機械学習の手法を適用し、高尿酸血症罹患を例に、健診の非必須項目から判定される疾患の罹患の有無を必須項目から予測するモデルを構築した。最良の手法は LR であり、その予測性能は AUC 0.8 で予測モデルとして妥当な結果が得られた。また、予測モデル構築に必要なデータは、説明変数の観点からは前年の健診結果は不要であり、データ数の観点からは総データ数の 2.8 %まで削減できることを示した。今後は本研究を通して得られた予測モデル構築のアプローチを用いて他の検査項目を対象とした予測モデルを構築する。

謝辞

本論文の執筆にあたってご指導いただきました東京大学大学院医学系研究科臨床情報工学教室小山博史先生、斎藤季先生に深く感謝申し上げます。

また、論文執筆途上にご助言くださった同教室員の皆様に感謝いたします。

引用文献

Agarwal SK, Misra A, Aggarwal P, Bardia A, Goel R, Vikram NK, Wasir JS, Hussain N, Ramachandran K, Pandey RM (2009) Waist Circumference Measurement by Site, Posture, Respiratory Phase, and Meal Time: Implications for Methodology. Obesity 17: 1056–1061

Baba Y, Hiramatsu, Kimura M, Shimizu S, Kobayashi K, Tsuda K, Sugiyama M, Blondel M, Ueda N, Kitsuregawa M, Nakashima N, Kashima H, Nohara Y, Kai E, Ghosh P, Islam R, Ahmed A, Kuroda M, Inoue S (2015) Predictive Approaches for Low-Cost Preventive Medicine Program in Developing Countries. KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining：1681–1690

Bennett J, Elkan C, Liu B, Smyth P, Tikk D (2007) KDD Cup and workshop 2007. ACM SIGKDD Explorations Newsletter 9: 51–52

Bhole V, Choi JW, Kim SW, Vera M & Choi H (2010) Serum Uric Acid Levels and the Risk of Type 2 Diabetes: A Prospective Study. The American Journal of Medicine 123: 957–961

Blagus R, Lusa L (2012) Evaluation of SMOTE for high-dimensional class-imbalanced microarray data. 2012 11th International Conference on Machine Learning and Applications 2: 89–94

Breiman L, Friedman J, Olshen RA, Stone CJ (1984) Classification and Regression Trees. Florida: CRC press

Breiman L (1996) Bagging Predictors. Machine Learning 24: 123-140.

Breiman L (2001) Random Forests. Machine Learning 45: 5–32

Chen T, He T (2014) Higgs boson discovery with boosted trees. In Proceedings of the 2014 International Conference on High-Energy Physics and Machine Learning 42: 69-80.

ChenT, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining：785–794

Chong IG, Jun CH (2005) Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems 78: 103–112

Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20: 273–297

Dang XT, Hirose O, Bui DH, Saethang T, Tran VA, Nguyen LA, Le TK, Kubo M, Yamada Y, Satou K (2013) A Novel Over-Sampling Method and its Application to Cancer Classification from Gene Expression Data. Chem-Bio Informatics Journal 13: 19–29

DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics 44: 837

Domingos PM (1999) MetaCost: a general method for making classifiers cost-sensitive. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining 155-164.

Drammond C, Holte R (2003) C4. 5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In Workshop on learning from imbalanced datasets II 11

Dubey V, Mathew R, Iglar K, Moineddin R, Glazier R (2006) Improving preventive service delivery at adult complete health check-ups: the Preventive health

Evidence-based Recommendation Form (PERFORM) cluster randomized controlled trial. BMC Family Practice 7: 1–12

Edwards NL (2009) The role of hyperuricemia in vascular disorders. Current Opinion in Rheumatology 21: 132

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27: 861–874

Feig DI (2009) Uric acid: a novel mediator and marker of risk in chronic kidney disease? Current Opinion in Nephrology and Hypertension 18: 526-530

Fischer JE, Bachmann LM, Jaeschke R (2003) A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. Intensive Care Medicine 29: 1043–1051

Friedman J (2001) Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29: 1189–1232

Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 33: 1-22

Fujimori K (2016) Current Status and Issues of the National Database. 医療と社会 26: 15–24

Hajian-Tilaki K (2013) Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian journal of internal medicine 4: 627–35

Hastie T, Tibshirani R (2014) 統計的学習の基礎. 東京: 共立出版

Hennessy S, Bilker WB, Berlin JA, Strom BL (1999). Factors influencing the optimal control-to-case ratio in matched case-control studies. American journal of epidemiology, 149: 195-197.

Hinton GE, Salakhutdinov RR (2006) Reducing the Dimensionality of Data with Neural Networks. Science 313: 504–507

Hosmer DW, Lemeshow S (2005) Applied Logistic Regression, Second Edition. New York: Wiley

Ichikawa D, Saito T, Ujita W, Oyama H (2016) How can machine-learning methods assist in virtual screening for hyperuricemia? A healthcare machine-learning approach. Journal of Biomedical Informatics 64: 20-24

Ichikawa, Saito, Oyama H (2017) Impact of predicting health-guidance candidates using massive health check-up data: A data-driven analysis. International Journal of Medical Informatics 106: 32-36

Ikeda A, Iso H, Toyoshima H, Fujino Y, Mizoue T, Yoshimura T, Inaba Y, Tamakoshi A, JACC Study Group (2005) The relationships between interest for and participation in health screening and risk of mortality: The Japan Collaborative Cohort Study. Preventive Medicine 41: 767–771

Kawamoto R, Kohara K, Kusunoki T, Tabara Y, Abe M, Miki T (2012) Alanine aminotransferase/aspartate aminotransferase ratio is the best surrogate marker for insulin resistance in non-obese Japanese adults. Cardiovascular Diabetology 11: 1–8

Kawano K (2001) [Quality control survey implemented by Japan Medical Association and standardization of external quality assessment in Japan]. Rinsho byori. The Japanese journal of clinical pathology 49: 860–3

Kjeldsen SE (2017) Hypertension and cardiovascular risk: general aspects. Pharmacological Research Available at https://doi.org/10.1016/j.phrs.2017.11.003 [Accessed 2017/10/1]

Kleinman NL, Brook RA, Patel PA, Melkonian AK, Brizee TJ, Smeeding JE, Joseph-

Ridge N (2007) The Impact of Gout on Work Absence and Productivity. Value in Health 10: 231–237

Kobayashi A (2008) Launch of a National Mandatory Chronic Disease Prevention Program in Japan. Disease Management & Health Outcomes 16: 217–225

Kuhn M, Johnson K (2013) Applied Predictive Modeling. Berlin: Springer

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521: 436–444

Lewallen S, Courtright P (1998). Epidemiology in Practice: Case-Control Studies. Community Eye Health, 11: 57–58.

Li L, Yang C, Zhao Y, Zeng X, Liu F, Fu P (2014) Is hyperuricemia an independent risk factor for new-onset chronic kidney disease?: a systematic review and meta-analysis based on observational cohort studies. BMC Nephrology 15: 1–12

Liu H, Liu Y, Wang L, Xu D, Lin B, Zhong R, Gong S, Podda M, Invernizzi P (2010) Prevalence of primary biliary cirrhosis in adults referring hospital for annual health check-up in Southern China. BMC Gastroenterology 10: 1–5

Lusa L, Blagus R (2012) The class-imbalance problem for high-dimensional class prediction. 2012 11th International Conference on Machine Learning and Applications 2: 123–126

März W, Kleber ME, Scharnagl H, Speer T, Zewinger S, Ritsch A, Parhofer KG, Eckardstein A, Landmesser U, Laufs U (2017) HDL cholesterol: reappraisal of its clinical relevance. Clinical Research in Cardiology 106: 663–675

Wright MN, Ziegler A (2017) ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77: 1-17

Matlock K, Niz CD, Rahman R, Ghosh S, Pal R (2017) Investigation of Model Stacking

for Drug Sensitivity Prediction. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics: 772

Menardi G, Torelli N (2012) Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery 28: 92–122

Ministry of Labor and Welfare (2015a) Outline of FY 2015 revision of medical fees. Available at: http://www.mhlw.go.jp/english/policy/other/ budget/dl/1st-fy2014-e.pdf [Accessed 2017/10/1]

Ministry of Labor and Welfare (2015b) The data of specific health check-ups and specific health guidance. Available at: http://www.mhlw.go.jp/bunya/shakaihosho/iryouseido01/info02a-2.html [Accessed 2017/10/1]

Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Collins GS (2015) Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Annals of internal medicine 162; W1-W73.

Murphy K (2012) Machine Learning. Cambridge: The MIT Press

Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Frontiers in Neurorobotics 7: 21

Nippon DATA80 Research Group (2006) Risk Assessment Chart for Death From Cardiovascular Disease Based on a 19-Year Follow-up Study of a Japanese Representative Population. Circulation Journal 70: 1249–1255

Ng WWY, Hu J, Yeung DS, Yin S, Roli F (2015) Diversified Sensitivity-Based Undersampling for Imbalance Classification Problems. IEEE Transactions on Cybernetics 45: 2402–2412

Ohta Y, Tsuchihashi T, Kiyohara K, Oniki H (2013) Increased Uric Acid Promotes Decline of the Renal Function in Hypertensive Patients: A 10-year Observational Study. Internal Medicine 52: 1467–1472

Panwar B, Gupta S, Raghava GPS (2013) Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information. BMC Bioinformatics 14: 1–14

Panwar B, Arora A, Raghava GPS (2014) Prediction and classification of ncRNAs using structural information. BMC Genomics 15: 1–13

Rahman MM, Davis DN (2013) Addressing the Class Imbalance Problem in Medical Datasets. International Journal of Machine Learning and Computing: 224–228

Ramezankhani A, Pournik O, Shahrabi J, Azizi F, Hadaegh F, Khalili D (2016) The Impact of Oversampling with SMOTE on the Performance of 3 Classifiers in Prediction of Type 2 Diabetes. Medical Decision Making 36: 137–144

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12: 1–8

Sewell M (2011) Ensemble Learning. Available at: http://www.cs.ucl.ac.uk/fileadmin/UCL-CS/research/Research_Notes/RN_11_02.pdf [Accessed 2017/10/1]

Smith E, Hoy D, Cross M, Merriman TR, Vos T, Buchbinder R, Woolf A, March L (2014) The global burden of gout: estimates from the Global Burden of Disease 2010 study. Annals of the Rheumatic Diseases 73: 1470–1476

Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Kattan MW (2010) Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology 21: 128-138

Tibshirani R (2011) Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73: 273–282

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17: 520–525

Uematsu H, Yamashita K, Kunisawa S, Otsubo T, Imanaka Y (2017) Prediction of pneumonia hospitalization in adults using health checkup data. PLOS ONE 12: e0180159

Verweij LM, Terwee CB, Proper KI, Hulshof CT, Mechelen W (2013) Measurement error of waist circumference: gaps in knowledge. Public Health Nutrition 16: 281–288

Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making, 26: 565-574.

Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97: 1837–47

Wolpert DH (1992) Stacked generalization. Neural Networks 5: 241–259

Wu G, Shen D, Sabuncu M (2016) Machine Learning and Medical Imaging. Amsterdam: Elsevier

Yamanaka H, Japanese Society of Gout and Nucleic Acid Metabolism (2011) Japanese Guideline for the Management of Hyperuricemia and Gout: Second Edition. Nucleosides, Nucleotides and Nucleic Acids 30: 1018–1029

Yang P, Yang YH, Zhou BB, Zomaya AY (2010) A Review of Ensemble Methods in Bioinformatics. Current Bioinformatics 5: 296–308

Ye J, Chow JH, Chen J, Zheng Z (2009) Stochastic gradient boosted distributed decision

trees. In Proceedings of the 18th ACM conference on Information and knowledge management: 2061–2064

Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, Ben-Yacov O, Lador D, Avnit-Sagi T, Lotan-Pompan M, Suez J, Mahdi JA, Matot E, Malka G, Kosower N, Rein M, Zilberman-Schapira G, Dohnalová L, Pevsner-Fischer M, Bikovsky R, Halpern Z, Elinav E, Segal E (2015) Personalized Nutrition by Prediction of Glycemic Responses. Cell 163: 1079–1094

厚生労働省 (2007a) 標準的な健診・保健指導プログラム（暫定版）の見直しに係る論点. Available at: http://www.mhlw.go.jp/shingi/2007/02/s0219-4a.html [Accessed 2017/10/1]

厚生労働省.(2007b) 「労働安全衛生法における定期健康診断等に関する検討会」報告書. 厚生労働省

厚生労働省 (2013) 標準的な健診・保健指導プログラム（改訂版）. Available at: http://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/kenkou/seikatsu/dl/hoken-program1.pdf [Accessed 2017/10/1]

厚生労働省 (2014) 平成 26 年度版　厚生労働白書. 厚生労働省

厚生労働省 (2017a) 保健医療分野における AI 活用推進懇談会報告書. Available at: http://www.mhlw.go.jp/file/05-Shingikai-10601000-Daijinkanboukouseikagakuka-Kouseikagakuka/0000169230.pdf [Accessed 2017/10/1]

厚生労働省. (2017b) 平成 27 年度特定健康診査・特定保健指導の実施状況について. Available at: http://www.mhlw.go.jp/file/04-Houdouhappyou-12401000-Hokenkyoku-Soumuka/0000173093.pdf [Accessed 2017/10/1]

首相官邸 (2017) 新たな情報財検討委員会 報告書. Available at: https://www.kantei.go.jp/jp/singi/titeki2/tyousakai/kensho_hyoka_kikaku/2017/johozai/houkokusho.pdf [Accessed 2017/10/1]