3つの生活習慣病(糖尿病・脂質異常症・高血圧)の3年以内発症確率の予測のための機械学習

背景

代表的な生活習慣病である糖尿病・脂質異常症・高血圧は、心筋梗塞や脳梗塞、末期腎不全や認知症などの重篤な病気を引き起こす危険因子であることが知られている。一方で生活習慣病は、個人の生活習慣の改善により発症を抑制することが可能である。

アスマイルでは、大阪府の国保加入者の健診データ等を活用して健康予測モデルを構築し、アスマイルユーザー個人の現在の健診データから近い未来に起こり得る生活習慣病の発症確率を正確に予測するAIを搭載する。生活習慣病等の発症確率の算出につき、当初は糖尿病、脂質異常症、高血圧を対象とする。これらの生活習慣病はそれぞれに診断基準が設定されている。今回のAIではそれぞれの病気で対象となる健康因子の数値が3年以内の健診結果において、診断基準に達することで病気発症とする。この発症確率は天気予報における降水確率と同じ指標であり、ユーザーにとってはわかりやすいのではないかと考えている。本稿では、アスマイルに搭載したAIの開発手順と評価方法について説明する。

大阪府国民健康保険データと対象者の選択

現在大阪府では年間で約200万人が国民健康保険に加入している。そのうちの約30%の54万人程度が特定健診を受診している。この健診データなどの詳細な分析を行い、3年以内の生活習慣病の発症確率の予測を行う。

AIの学習に用いる対象者の選択は次のように行う。健診データは個人が特定できないように加工された2013年度から2017年度までの5年分の大阪府国保データを用いる。健診受診日から3年間の健診データを追跡するために、2013年4月から2014年12月の期間に健診を受けた人を対象とする。このうち、対象期間の健診日から3年間で定期的な健診受診のない人を除くと、約25万人となった。そこから、初回健診時点で既に各疾病の診断基準を満たしている人や各疾病の治療薬を服用している人は解析対象者から除いた。なお、生活習慣病の診断基準は、各疾病のガイドラインに基づいて表1に示す基準を採用した。

表 1. 診断基準

生活習慣病	健診項目	診断基準数値
糖尿病	空腹時血糖	126 mg/dL 以上
	ヘモグロビン A1c (HbA1c)	6.5 % 以上
脂質異常症	LDL コレステロール (LDL-C)	140 mg/dL 以上
	HDL コレステロール (HDL-C)	40 mg/dL 未満
	中性脂肪	150 mg/dL 以上
高血圧	収縮期血圧	140 mmHg 以上
	拡張期血圧	90 mmHg 以上

次に、AIに入力する健診項目の選択を行う。はじめに、身体計測や血圧測定、血液検査などの検査値や質問票の回答結果など、予測に使えそうな健診項目を抜き出した。続いて、既往歴を除き、欠損の割合が全対象者の10%を超える変数は削除した。さらに、変数間の相関が非常に高い検査項目は、医学的見地に基づいて適切な変数を削除した。最後に必要項目に欠損値を持つ対象者を除

き、最終的に糖尿病で約 18 万人、脂質異常症で約 9.8 万、高血圧で約 13 万人の国保加入者が対象者となった。健診項目は、連続変数として年齢や BMI、収縮期血圧、中性脂肪、HDL-C、LDL-C、ALT(GPT)、HbA1c を使用し、カテゴリ変数として性別や喫煙状況、尿蛋白レベル、服薬歴や既往歴の有無を使用することにした。

これら対象者の初回健診から 3 年間の健診データを追跡し、初回健診以降、各疾患の診断基準を一度でも満たした人は病気の発症ありとし、それ以外を発症なしとして機械学習を行う。なお、3 年の追跡で新たに発症のあった人の割合は、糖尿病で約 5%、脂質異常で約 37%、高血圧で約 25%であった。

機械学習モデル

機械学習の手法として勾配ブースティング決定木を用いる。決定木は、「LDL コレステロール は 130 mg/dL 以上か？」、「喫煙をしているか？」 といった条件に基づいて分岐を繰り返していきモデルを構築する手法である。また決定木には欠損値やカテゴリ変数を扱いやすいという利点がある。勾配ブースティング決定木は、複数の決定木を組み合わせる手法であり、モデルが改良するように決定木を付け加えていくことでより強力なモデルを構築する。10 万を超えるビッグデータを扱うために、学習速度が速くメモリ効率にも優れる LightGBM という勾配ブースティング決定木のフレームワークを採用した。

計算結果

予測確率の評価には信頼度曲線を用いる。図 1 の横軸は, AI が予測した確率であり、縦軸はデータから計算した確率である。予測確率の評価には学習で用いたデータとは異なる検証用のデータを用いる。予測確率とデータから求めた確率が一致するというのは 45 度の直線に乗ることに対応する。図中のそれぞれの点の横に書かれてある数字は、データの数を表している。

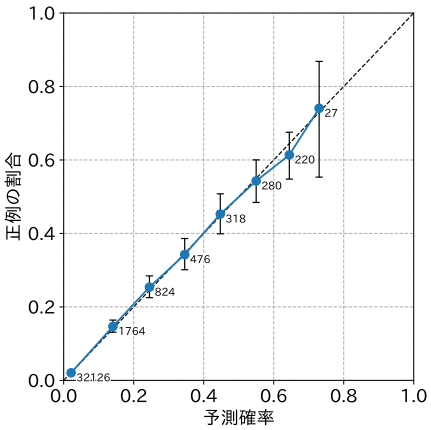

図 1. 糖尿病の信頼度曲線

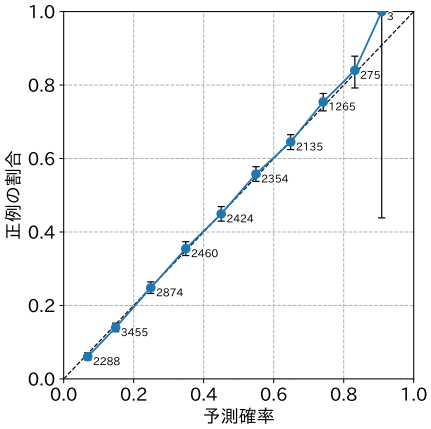

図 2. 脂質異常症の信頼度曲線

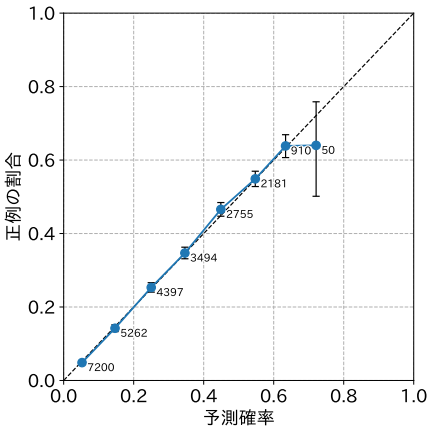

図 3. 高血圧の信頼度曲線

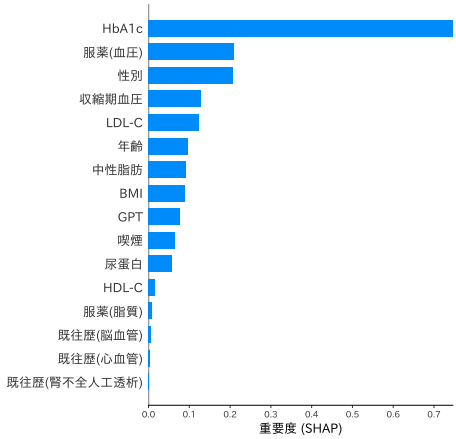

図 4. 糖尿病の発症確率の算出に寄与する健康因子の重要度。重要度の大きさを横棒の長さで示す。上位にくる変数ほど予測確率の推定に寄与しており、一番重要な因子は HbA1c である。

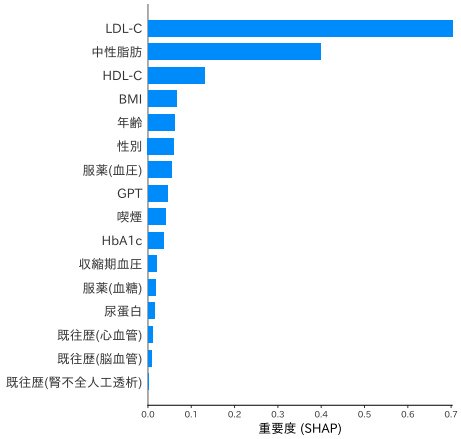

図 5. 脂質異常症の発症確率の算出に寄与する健康因子の重要度。一番重要な因子は LDL-C である。

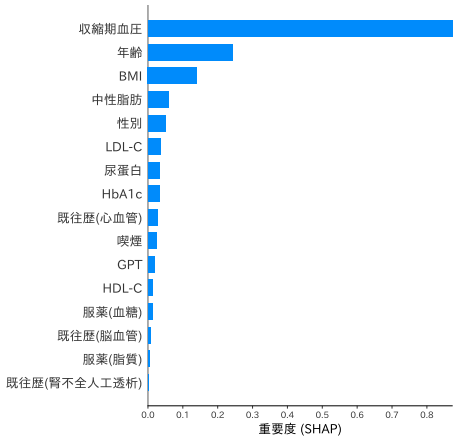

図 6. 高血圧の発症確率の算出に寄与する健康因子の重要度。一番重要な因子は収縮期血圧である。

図 1 では糖尿病の計算結果がプロットされている。結果は黒点線で示してある 45 度の直線によくのっている。病気の発症確率を予測する良いモデルになっていることがわかる。図 2 には脂質異常症、図 3 には高血圧の信頼度曲線がプロットされている。脂質異常症と高血圧では、データ数の少ない点を除いて、ほぼ 45 度の直線上に点が打たれており、信頼度が非常に高いことがわかる。

次は、発症確率にどの健康因子が寄与しているかを示す重要度について議論する。協力ゲーム理論のシャープレイ値を機械学習に応用した SHAP 値を全ての検証用データで計算し、その絶対値の平均値から重要度を計算する。糖尿病の場合が図 4 に示されている。一番重要な健康因子は HbA1c である。二番目から四番目には高血圧の薬服用、性別、収縮期血圧が重要な要因になっている。図 5 では脂質異常症の重要度が示されている。一番目は LDL-C であり、二番目から四番目は中性脂肪、HDL-C、BMI になっている。図 6 には高血圧の重要度が示されている。一番目は収縮期血圧であり、二番目から四番目は年齢、BMI、中性脂肪になっている。

アスマイルにおける予測確率の表示

過去の大阪府国保加入者の大規模な健診データを使って学習させた AI に、アスマイルユーザーの健診データを入力することで、3 年以内の生活習慣病の発症確率を予測する。結果は 42%などの 1 %刻みの確率で表示する。確率を 3 つの領域 (20%未満、20%以上 50%未満、50%以上) に分けて、それぞれの生活習慣病のガイドラインに沿って確率に応じたコメントを行う。コメントについて

は糖尿病では日本糖尿病学会の糖尿病診療ガイドライン、高血圧症では日本高血圧学会の高血圧治療ガイドライン、脂質異常症では日本動脈硬化学会の動脈硬化性疾患予防ガイドラインの推奨する生活改善の方法についての内容を表示する。

アスマイルユーザーのうち、直近の健診データが既に診断基準を満たしていた場合は「発症予測の対象外です」と表示する。さらに、予測するために必要となる健診データの必須項目に欠損や異常な数字が入っている場合がある。この場合には予測が出来ないので、コメント欄に「必須項目が全て揃っていない、あるいは著しい異常値を示す項目があるため、発症予測の対象外です。」と表示する。ユーザー自身で自らの健診結果を確認してほしい。なお、3 年以内の健診データが存在するにもかかわらず、いずれかの理由で解析対象外となってしまうユーザーの割合は、糖尿病で約 19%、脂質異常症で約 57%、高血圧で約 38%であると推定される。

さらには、アスマイルの予測結果の出力画面に、それぞれの生活習慣病で最も大きな役割を持っている健康因子のヒストグラムを表示し, ユーザーの値を図に表示する。そのことにより、自らの健康因子の数字は同じ性別･年齢層の人たちの中でどの位置にいるかを分かるようにする。糖尿病については HbA1c、高血圧については収縮期血圧値、脂質異常症については LDL-C を使う。脂質異常症については 3 種類の健康因子である LDL-C, HDL-C, 中性脂肪が病気の診断基準を与えるが、今回は重要度で LDL-C の順位が最も高かったことと、脂質異常症の診断基準を超過した方で LDL-C を原因とする方が最も多かったため LDL-C を用いた。

まとめ

アスマイルでは大阪府国保加入者の健康促進のために機械学習プロジェクトを立ち上げた。今回アスマイルに搭載する AI を十分に活用してくださることで、自らの健康を自らで知ることを経験していただき、行動変容につなげていただけれ嬉しい限りです。現段階では 3 年以内の発症確率を予測する AI を公開するが、さらに歩数の増加や体重の減少など、個人の努力に応じて、どの程度発症確率が改善するかを示す AI も、今後に搭載する予定です。(文責：大阪大学キャンパスライフ健康支援センター)