不均衡データに適した学習モデルのアンサンブルによる生活習慣病の発症予測

○恒川充　岡夏樹　田中一晶　荒木雅弘（京都工芸繊維大学）
新谷元司（SG ホールディングスグループ健康保険組合）
吉川昌孝（日本システム技術株式会社）

Prediction of Onset of Lifestyle-related Diseases with Ensemble of Learning Models Suitable for Imbalanced Data

∗M. Tsunekawa, N. Oka, K. Tanaka and M. Araki (Kyoto Institute of Technology)
M. Shintani (SG Holdings Group Health Insurance Association)
M. Yoshikawa (Japan System Techniques Co. Ltd.)

Abstract− If it is possible to predict whether or not a person will develop lifestyle-related diseases within one year from periodical health checkup data, it will be possible to provide efficient health guidance. This study investigated a method for constructing a prediction model from health checkup data with an imbalanced class balance of positive and negative cases. First, as a baseline, we constructed 1) a model in which decision trees are used as weak learners for bagging after undersampling. Besides, we constructed 2) a model using lightGBMs as the weak learner, 3) a neural network model using focal loss as the loss function, and 4) an ensemble model of 2 and 3. To evaluate whether these models can suppress the number of targets for health guidance without increasingly neglecting those who are likely to become severely ill, we compared them with g-mean and PR-AUC. As a result, model 4 showed significantly higher classification performance than model 1.

Key Words: medical informatics, class imbalance, undersampling, bagging, lightGBM, focal loss

1　はじめに

日増しに健康志向が高まっている昨今，現在健康な人であっても将来起こりうる健康上の問題を知り，前もってそれを解決しようと行動することが予防医療の観点上重要であると考えられている．あらゆる病院で電子カルテシステムが導入されたことで，蓄積した医療情報の解析が盛んに行われており，医療情報を利用して健康者の将来の疾病リスクを見通せるようになることが期待されている．

実際，ある健康保険組合では病気を持たずに社会活動を行っている人の健診データを使って生活習慣病の発症リスクが高い人を抽出し，発症を予防するための保健指導を行っている．具体的には，保健師が設定した血圧やコレステロール値などの閾値を確認し，基準値を上回った人に対して重点的に保健指導を行うという方法である．しかし，健診データ全体を見て総合的に判断できているとは言い難く，機械学習を用いればデータからより網羅的に特徴を捉えることができると考えた．通常の社会活動を行っている人間に対して予防を促進することを目的とした研究としては，健診データから Lasso ロジスティック回帰により肺炎での入院を予測するモデルを提案したもの [1] がある．しかし，この研究は肺炎という単一の病気のみを対象にしたモデルを作成しており，より広範な生活習慣病を対象とした予測モデルを構築する研究はなされていない．

本研究の目的は，ある企業の労働者の健康診断データを機械学習手法を用いて分類し，生活習慣病を発症するであろう人を高い精度で抽出するモデルを構築することである．本研究は病気にかかって入院した後のデータを使うのではなく，ごく普通に社会活動を行っている人たちの中から生活習慣病を発症するであろう人を予測するという点で独自性の高いものである．学習データのクラスバランスが不均衡であることに対処するために，適切な機械学習モデルを採用した．また，抽出された生活習慣病発症予備軍の対象者に対して保健指導を施すことで，医療費の抑制や保険組合加入者の健康増進に貢献するといった実社会での実用化も見据えている．

2　関連研究

2.1　医療データを用いた研究

医療データを用いた疾病の予測に関する研究は古くから多く存在する．例えば，Hippisley-Cox らの研究 [2] では，特定の国や地域で統一されたデータベース上にある電子カルテの診察結果を利用して，大腸がんのリスクを推定するためにコックス比例ハザードモデルを使用してリスク方程式を作成した．検査結果の項目から心筋梗塞や脳梗塞の発症確率を予測するコホート研究には，統計分析手法が用いられている [3]．近年では，機械学習手法を用いた発病予測も行われるようになってきており，Weng らの研究 [4] では，定期的な臨床データを用いて心血管系のリスク予測を行う際に，機械学習技術を利用する優位性を示した．また，Lee らの研究 [5] では，突然の心臓死につながる恐れのある心室頻脈を防ぐために，ニューラルネットワークを使用して発症の 1 時間前に病状の急変を予測できるモデルを開発した．ここまでに挙げた研究では，病院にかかった際に生成されるデータや入院時に取得されているデータを用いて予測を行っていた．つまり，上記の研究では何らかの病状を持った人しか予測対象にされず，健康に社会活動を行っている人は利用できない．Ueyama らは，健診データから Lasso ロジスティック回帰によ

り肺炎での入院を予測するモデルを提案しており，通常の社会活動を行っている人間の健診データを利用している [1]．しかし，本研究では複数の病名を予測対象にしていることや，データの特徴に基づいた機械学習手法を採用している点が異なる．

2.2 不均衡データからの学習

本研究では，正例（今後 1 年以内に対象病名と診断される人の健診データ）が 1 %以下しかないという不均衡データを扱っている．データ数の多い負例（今後 1 年以内には対象病名と診断されない健康な人の健診データ）の特徴が識別結果に強く影響してしまう恐れがあるため，一般的に機械学習手法で不均衡データを扱うのは難しいと言われている．不均衡データを学習するために，cost-sensitive learning[6]，アンダーサンプリング，オーバーサンプリングなど多くの手法が提案されている．cost-sensitive learning は，許容し難い少数派クラスの誤りには大きなコストを, 許容できる多数派クラスの誤りには小さなコストを課す損失関数を用いて学習する手法である．いくつかの応用領域における不均衡データの学習において，cost-sensitive learning がアンダーサンプリングやオーバーサンプリングを行うよりも優れていることが様々な実証研究によって示されている [7]．アンダーサンプリングは，多数派クラスの訓練データを破棄することで，正･負両方のクラスから同数の例を考慮するのに対し，オーバーサンプリングは少数派クラスから重複した例を複製することでこれを実現している．データの拡張によってデータに重複が生まれると，過学習を引き起こす傾向があるため，隣接する例を組み合わせて少数派クラスの合成例を作成する Synthetic Minority Over-sampling Technique (SMOTE)[8] が提案されている．

Wallace らは,「アンダーサンプリング＋バギング」が不均衡を処理するための最良の戦略であると主張した [9]．ほとんどの分野で不均衡データを処理するためにはアンダーサンプリングが使用されるべきであり，バギングはアンダーサンプリングによって生まれた識別結果の分散を減らす効果があると述べている．また，先行研究 [10] において，健診データから生活習慣病の発症を予測すると,「アンダーサンプリング＋バギング」が「SMOTE ＋バギング」と比較して高い識別精度を示したため，本研究においてもモデルの一部に採用した．

2.3 アンサンブル学習

本研究では，不均衡データに適した学習モデルである「アンダーサンプリング＋バギング」と cost-sensitive learning モデルのアンサンブル学習を行うことで識別精度を向上させることができるかを検証した．アンサンブル学習は，複数の識別器を組み合わせ，それらの結果を統合することで個々の識別器よりも性能を向上させる方法である．近年，Kaggle や KDD-Cups などの国際的な機械学習競技会でも積極的に用いられて高い性能を発揮している．また，アンサンブル学習は，精度，安定性，一般化の面で優位性があるため，あらゆる問題を解決する際に広く採用されている [11]．医療分野においても，アンサンブルモデルを用いて腎移植生存率を予測するモデルを構築する研究 [12] がある．アンサンブル学習によって識別精度を向上させるためには，できるだけ独立な識別器を組み合わせることが必要である.「アンダーサンプリング＋バギング」では決定木ベースの機械学習モデルを利用するため，それとは機序の異なるニューラルネットワークを用意してアンサンブルを行った．

3 データの概要

本研究では，SG ホールディングスグループ健康保険組合が持つ従業員のレセプトデータと定期健康診断データを利用した．レセプトとは，患者が受けた保険診療について医療機関が被保険者に請求する医療報酬の明細書のことである．例を挙げると，患者の性別や年齢，診療年月といった基本情報をはじめ，診断された病名や診療行為，処方された医薬品などがレセプトデータには含まれている．一方，健診データは，健康診断の結果をまとめてあり，身長，体重，血圧，赤血球数などが記されている．レセプトデータはある人が怪我もしくは病気にかかり，医療機関で受診した際に作成されるデータであるのに対し，健診データは概ね 1 年に 1 回，定期的に取られるデータである．この 2 つのデータは，従業員を一意に特定できる匿名のハッシュコードで紐づけされている．本研究では健診データを入力として病気の発症の有無を予測するが，レセプトデータから病気の発症とその時期を抽出し教師データとして用いた．

3.1 予測対象とする病名の同定

レセプトデータに含まれている病名コードを見て病名を判断した．病名コードには世界保健機関が作成した疾病及び関連保健問題の国際統計分類コードである ICD10 を用いた．本研究で予測対象とした疾病（以降，対象病名と呼ぶ）は，狭心症 (ICD10 コード:I20)，急性心筋梗塞 (I21，I22)，心筋症 (I42)，不整脈，伝導障害 (I44〜I49)，くも膜下出血 (I60，I690)，脳内出血 (I61，I691)，脳梗塞 (I63，I693) である．これは，健康保険組合の顧問医に提示された重症化しうる生活習慣病である．本研究では，一般的に生活習慣病の一種とされる糖尿病は対象病名から取り除いた．先行研究 [10] において，数ある対象病名のうち糖尿病は高い精度で識別できるが他の疾病の識別精度が低いという問題があったからである．糖尿病は他の疾病に比べてデータ数が多く，血圧や HbA1c といった糖尿病と診断を下す際に用いられる指標が学習データの特徴量に含まれていることから，他の疾病と比べて識別が容易であったと考えられる．正例の対象病名のうち，糖尿病識別のための特徴に合わないデータに対しては間違えた予測を行ってしまうモデルとなっていたと考えられるため，糖尿病のデータを取り除くことで糖尿病以外の生活習慣病の発症を予測する精度の向上を目指した．

3.2 予測に際して使用した特徴量

予測のための特徴量として利用した健診データの項目について説明する．まず，年齢，身長，血圧，コレステロール値などの検査結果の数値データが 22 項目存在する．また，尿蛋白判定，貧血判定，糖代謝判定，腎機能判定などの健康診断で測定したデータを用いて医療機関が導き出した六段階の判定結果が 13 項目存在する．さらに，既往症の有無や喫煙習慣，運動習慣，生活習慣の改善意思といった問診票のアンケート回答結果 22 項目についても特徴量として利用した．

3.3 学習データの選定

本研究では「1 年以内に対象病名を発症するか否か」を健診データから識別するという 2 クラス分類問題に取り組んだ．学習データの整形のための前処理も先行研究と同様のため，そちらを参照してほしい．初めて対象病名と診断されたかどうかを慎重に判断したうえで対象病名と診断された 1 年前までの健診データを抽出し，正例データとして用いた．また，1 年以内に対象病名を発症することがないことを担保したうえで負例データを生成した．

4 特徴量生成

特徴ベクトルを受け取って予測を行う機械学習モデルにおいて特徴量は非常に重要な要素であるため，提供された特徴量から新しい特徴を設計することが一般的に行われている．まず，健診データの変化量に注目し，先ほど取り出した健診データの 1 つ前の健診データとの差分，2 つ前の健診データとの差分を計算して特徴量に加えた．病気を発症する際には，健診データ上の何らかの項目に変化があると考えられるため，変化量を明示的に特徴量に加えることで識別精度が向上すると考えた．このように過去のデータとの差分を計算するためには，健診データが 3 年分存在する必要がある．3 年分存在しないデータは上記の特徴量を欠損値として，1 年分だけでも存在すればデータセットに取り入れた．本研究では，糖尿病を対象病名から取り除いた影響でより正例データの数が少なくなっているため，欠損値がノイズとなってしまうデメリットよりもデータ数増加のメリットの方が大きいと考えた．

4.1 交互作用特徴量

交互作用特徴量とは，特徴量 A ×特徴量 B のように複数の特徴量の加減乗除をして新たに作られた特徴量のことを指す．まず，今回の問題設定において識別に有用な特徴量を見つけるために，ランダムフォレストを用いて特徴量重要度を可視化した．正例データの識別に重要な特徴量を把握したいため，アンダーサンプリングをして正例データ数と負例データ数を合わせたうえで特徴量の重要度を可視化している．

重要度が上位の特徴量の中で，上位 5 項目であった，「年齢，心電図判定，血圧を下げる薬を飲んでいるか，腹囲，心臓病の既往歴」と特徴量重要度の上位 10 位内に入っていた収縮期血圧と拡張期血圧に注目した．重要度が 1 位である年齢と 2〜5 位までの特徴量を掛け合わせて，4 つの新たな特徴量を作成した．また，収縮期血圧と拡張期血圧の差は脈圧と呼ばれ，脈圧が大きいと動脈硬化性の疾患が増えるといわれているため，拡張期血圧と収縮期血圧の差が意味のある特徴であると考え，特徴量化した．

4.2 フラグ特徴量

識別に有用な特徴量を生成するために，あらゆる特徴量においてクラスごとにデータの分布に違いがないかを確認した．全データを利用して散布図を描画してしうまうと正例と負例のデータ量の差が原因で特徴が現れづらいと考え，負例をアンダーサンプリングしたのちに散布図を描画した．すると，いくつかの特徴量の組み合わせにおいて，正例と負例を区別する識別境界が見つかった．

そこで，その条件を満たすと正例となるデータに対しては'1' とし，負例となるデータには'0' とする特徴量を作成した．このように 0,1 の二値で表される特徴量のためフラグ特徴量と呼んでいる．同様にほかの特徴量についてもデータの分布を可視化したところ以下のような特徴が見られたため，以下の条件を満たすデータに対してのみ'1' とする特徴量を新たに生成した．なお,「判定」とついている特徴量は 6 段階のカテゴリカル変数であり，数字が大きくなるほど不健康であることを表す．

- 腹囲が 100cm 以上かつ拡張期血圧が 90 以上
- 腹囲が 100cm 以上かつ年齢が 40 歳以上 60 歳以下
- 心電図判定が 4 以上かつ egfr が 60 以下
- 尿蛋白判定が 2 以上かつ egfr が 70 以下
- hba1c(ngsp) が 5.6 以上かつ egfr が 75 以下
- 0.3 × [拡張期血圧] + [心電図判定] >34
- 0.1 × [収縮期血圧] + [心電図判定] >18

4.3 target encoding

target encoding とは，目的変数を用いてカテゴリカル変数を数値に変換する方法である．本研究ではその中でも，カテゴリカル変数の各水準における目的変数の平均値を学習データを用いて集計し，その値で各水準を置換する手法である target mean encoding を行った．ここで，自身のレコードの目的変数をカテゴリカル変数に取り込めないことに注意しなければならない．誤ってこれを行ってしまうと，本来予測すべき答えを知った状態で学習してしまうため，未知のデータに対する予測がうまくいかなくなってしまう．そこで，学習データでは out-of-fold で各水準の平均値を計算したうえで置換し，評価データでは train データ全体で各水準の平均値を計算して置換を行った．今回，target encoding に利用した特徴量は,「心電図判定，代表判定，血圧判定，血圧を下げる薬を飲んでいるか，心臓病と診断されたことがあるか，脳卒中と診断されたことがあるか」である．これらは，ランダムフォレストで識別の際の特徴量重要度を可視化した際に，上位に現れたカテゴリカル変数である．

5 発症予測モデル

本章では，実験に用いた 5 つの機械学習モデルについて説明する．1 つ目のモデルは，弱識別器に lightGBM[13] を利用したアンダーサンプリング＋バギングモデルで，これ以降は LGB-bag と呼ぶ．2 つ目のモデルは，ベースラインとなる先行研究 [10] のモデルである，弱識別器に決定木を利用したアンダーサンプリング＋バギングモデルで，これ以降は DTC-bag と呼ぶ．3 つ目のモデルは，損失関数に focal loss[14] を用いたニューラルネットワークモデルで，これ以降は flNN と呼ぶ．4 つ目のモデルは，一般的な cost sensitive learning をニューラルネットワークに適用したモデルでこれ以降は csNN と呼ぶ．5 つ目のモデルは，LGB-bag と flNN の推定確率値の平均をとってアンサンブル学習をしたモデルで，これ以降は Ensemble と呼ぶ．提案モデルの概要図を Fig. 1 に記載する．

本研究でアンダーサンプリング＋バギングモデルの弱識別器に使用した lightGBM は，先発の Gradient Boosting decision tree モデルである XGBoost に，識別に使うデータを減らす仕組みである GOSS と特徴量

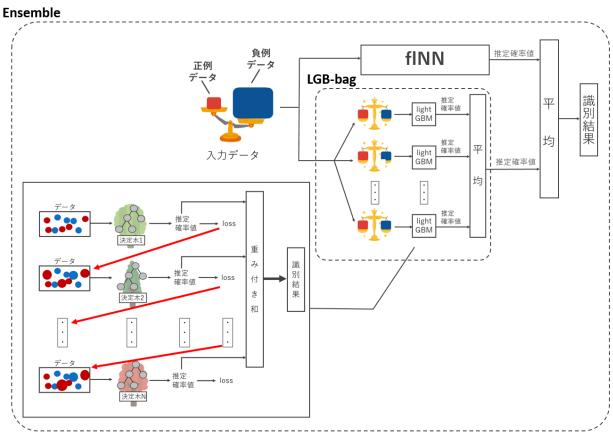

Fig. 1: 提案モデルの概要図

を減らす仕組みである EFB を組み合わせて高精度と高速な挙動を両立させたモデルである．近年，データ分析コンペである Kaggle などで頻繁に用いられている精度の高いモデルである．lightGBM は決定木よりも識別性能が高く，学習データに少し過剰に適合するようにパラメータを調整することができるため，バギングの弱識別器としても適していると考えた．

LGB-bag の弱識別器の数は 50 とした．弱識別器の数を 50 から 200 の範囲で変化させても分類精度にほとんど変化がなかったが，50 未満にすると分類精度が低下したからである．このモデルでは，交互作用特徴量，フラグ特徴量，target encoding により作成した特徴量を使用し，特徴量数は 209 であった．欠損値については，通常取りうる値の範囲外である-9999 で埋めた．決定木は，ある変数のある値でデータを二分することを繰り返すモデルであり，変数の値そのものではなく相対的な大小関係に依存してモデルが作成される．そのため，通常取りうる範囲外の値を与えることで欠損値か否かでデータを分離することができ，欠損値であることをモデルに教えることができるのである．同様の理由で，データのスケーリング処理は行っていない．

focal loss は，データセットのクラスバランスが不均衡であり，多数派クラスの大多数が容易に識別が可能である場合に有用である．focal loss を損失関数に用いることで，多数派クラスのデータのうち，識別が簡単なデータに対する重みを低減し，識別が難しいデータに対して学習を集中させることが可能となる．本研究においても，大量の負例データのうちのほとんどは何の異常もない健康体のはずである．そのようなデータが loss に与える影響を小さくすることで，軽い持病は持っているが対象病名は発症していない負例データや少数の正例データといった，より予測が難しいデータの特徴を学習できることを期待した．本研究では focal loss を 2 クラス分類問題に適用するため，損失関数を以下のように定義した．なお，p は $y=1$ に対するモデルの推定確率であり，α と γ はパラメーターである．

$$loss(p) = -\alpha(1-p)^{\gamma}\log(p) - (1-\alpha)p^{\gamma}\log(1-p) \quad ...(1)$$

$$(\gamma \geq 0, \alpha \in [0,1])$$

今回は 3 層のニューラルネットワークを用いて epoch 数を 100 として実験を行った．活性化関数は relu とし，勾配学習の際の最適化アルゴリズムには adam を用いている．パラメータのチューニングは，fold 数を 4 としたクロスバリデーションを行った．チューニングの際は後述する g-mean を最大化するようなパラメータを選択している．第一層のユニット数，第二層のユニット数，学習率，バッチサイズ，focal loss のパラメータ α と γ を以下の探索範囲でグリッドサーチを行った．

- 第一層のユニット数：[16, 32]
- 第二層のユニット数：[4, 8]
- 学習率：[0.0003, 0.0005, 0.0007，0.001]
- バッチサイズ：[32，64]
- α：[0.99, 0.992, 0.995, 0.994, 0.996]
- γ：[10, 11, 12, 13, 14, 15]

また，flNN と比較するため，一般的な cost sensitive learning をニューラルネットワークに適用したモデルを用意した．つまり，(1) 式において，γ の値を 1 としてクラス間の重みのみを考慮した損失関数を使用してニューラルネットワークで識別すればよい．α の値だが，探索範囲を [0.98, 0.985, 0.99] として，グリッドサーチを行った．

ニューラルネットワークの入力データに対してはアンダーサンプリングを行わないため，アンダーサンプリングを行って特徴を抽出しているフラグ特徴量は使用せず，交互作用特徴量と target encoding により作成した特徴量のみ追加した．その結果，特徴量数は 200 であった．欠損値は中央値で補完した．データのスケーリング処理として，標準化を行った．また，ニューラルネットワークは PCA を用いて特徴量の次元削減を行ったうえで学習した．これは，ニューラルネットワークは特徴量が多すぎるとデータ不足に陥って汎化性能が上がりにくい側面があるからである．学習時間も考慮したうえで試行錯誤の末，50 次元まで削減した．

6 評価指標

本研究では，評価指標として，geometric mean(g-mean)，PR-AUC を使用する．g-mean は，クラスごとの recall の積の平方根で表される．g-mean は正例の recall と負例の recall のどちらも正解できなければ数値が大きくならず，加えて混同行列の 4 つの値をすべて用いて計算するためデータセットのクラスバランスにも左右されない．こういった理由から，不均衡データの評価指標によく利用される [15]．precision と recall はトレードオフの関係であり，識別結果を直感的に評価するには不適であるため，この g-mean を見て識別結果の良し悪しを判断する．

PR-AUC とは，縦軸に適合率（precision）を，横軸に再現率（recall）をとる曲線である Precision-Recall 曲線の下側の面積にあたる値である．分類問題の際には ROC 曲線の下側の面積である ROC-AUC が評価に用いられることが多い．しかし，データセットのクラスバランスが著しく不均衡である場合は PR-AUC を用いるべきであることが報告されている [16]．また，PR 曲線はモデルが出力した推定確率値を利用して閾値を変化させながら描く．そのため，正解率や g-mean などのように予測確率に対してある特定の閾値を決めた際に得られる結果とは異なり，予測モデルの性能自体を評価することができる．

7　結果

3.3 章の前処理によって用意された，正例が 412 件，負例が 54401 件のデータを用いて実験を行った．10-fold クロスバリデーションを行い，出力された 10 個の評価指標に対して平均をとって比較する．Table 1 に各手法の recall，precision，g-mean，PR-AUC の平均値をまとめている．

Table 1: 各モデルの評価指標の平均値

モデル	recall	precision	g-mean	PR-AUC
LGB-bag	0.8494	0.0553	0.8687	0.3708
DTC-bag	0.8373	0.0547	0.8627	0.2861
flNN	0.8227	0.0517	0.8494	0.3418
csNN	0.7258	0.0730	0.8195	0.3229
Ensemble	0.8325	0.0598	0.8654	0.3835

LGB-bag と DTC-bag を比較すると，g-mean と PR-AUC 共に LGB-bag が上回っている．よって，弱識別器を lightGBM に変更したほうが弱識別器を決定木としたモデルより優れているといえる．flNN と csNN を比較すると，g-mean と PR-AUC 共に flNN が上回っており，損失関数に focal loss を用いたことによる識別精度の向上が見られる．focal loss の特徴によって，大量の負例データの中でも明らかに健康なデータの影響を低減し，識別が難しいデータに対して注力した学習ができたのではないかと考えている．また，提案モデルである Ensemble は PR-AUC において 4 つの中で最も高くなっている．よって，決定木ベースのモデルとニューラルネットワークという機序が異なるモデルを用いてそれぞれ不均衡データに適した学習モデルを作成し，それをアンサンブルすることによって学習モデルの性能が向上することが示された．

5 つのモデルの PR-AUC と g-mean について分散分析で比較した結果，有意な差が見られた（PR-AUC : $F(4,45) = 2.885, p < .05$，g-mean : $F(4,45) = 3.668, p < .05$）．そこで，Tukey-HSD 法で多重比較を行った結果，PR-AUC においては DTC-bag よりも Ensemble の方が有意に高いことが示された（$p < .05$）．また，DTC-bag よりも LGB-bag の方が高い傾向が見られた（$p = .084$）．g-mean において，csNN よりも LGB-bag，DTC-bag，Ensemble の 3 つのモデルのほうが有意に高かった（$p < .05$）．同じニューラルネットワークモデルである flNN と csNN の間には有意な差は見られなかったが，決定木ベースのモデルとアンサンブルするモデルとして同等の識別精度を有している flNN を用いたことは妥当であったと言える．

8　考察

まず，PR-AUC において有意差が現れた理由について考える．Fig.2 は，クロスバリデーションの各 fold の test データの推定確率値をすべて結合して描いた PR 曲線である．Ensemble が recall が 0.2〜0.6 付近のときわずかではあるが LGB-bag を上回っていることが分かる．DTC-bag は，recall が 0.1〜0.2 付近の時，他の 4 つのモデルと比べて precision が大きく劣っている．ここからもアンダーサンプリング＋バギングの弱識別器に lightGBM を用いた優位性が見て取れる．lightGBM と決定木の性能の違いにより，各弱識別器の識別性能と独立性が向上している．バギングの復元抽出の際には負例のみが異なるデータとなっているので，弱識別器の独立性は負例データ由来のものであり，LGB-bag のほうが多様な負例の特徴をうまく捉えることができていると言える．これが低 recall のときに LGB-bag が DTC-bag に比べて高い precision を示している理由である．そして，不均衡データであっても，適切なモデルである LGB-bag と flNN をアンサンブルすることで，お互いがうまく識別できていなかった部分を補い合い，より良い推定確率値を得ることに成功し，DTC-bag というベースラインモデルの PR-AUC を Ensemble が有意に上回ることが示された．

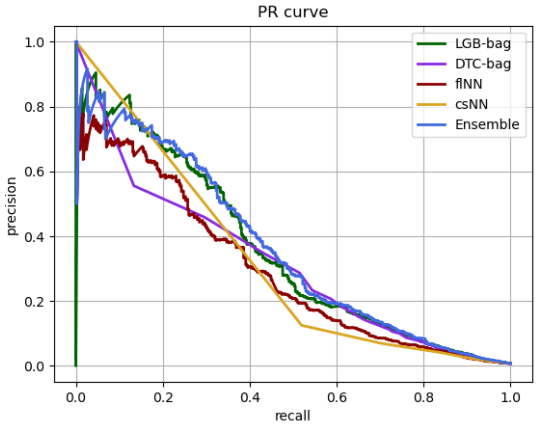

Fig. 2: 各モデルの PR 曲線

次に g-mean の有意差について考える．LGB-bag や DTC-bag はアンダーサンプリングを行っているため数少ない正例データの学習がうまく行えていると考えられる．対して，csNN は表 1 を見ても分かるようにほかの 4 つのモデルと比べて recall が低く，正例の予測がうまくできていない．csNN はデータ数が豊富にある負例に関してはうまく予測することができており，recall が低い領域では flNN よりも高い性能を示している．しかし，正例を重視する重みを与えるだけでは正例を精度高く予測することは難しく，recall が 0.5 以上になると，flNN のほうが高い precision を示すようになる．flNN は，focal loss の特徴によって識別が簡単な負例のデータの重みを小さくして，識別が難しい正例データを重視した学習ができているからであると考えている．g-mean は，予測確率に対してある特定の閾値を決めて得られる結果であるため，この高 recall の結果が反映されており，これが csNN と LGB-bag，DTC-bag，Ensemble 間で g-mean の有意差が現れた理由である．

PR 曲線をもう一度見ると，recall が高くなっていくにつれて 5 つのグラフが重なっていき，recall が 0.8 以上となると precision にほとんど差がなくなってしまっている．これは，recall が高くなるほどデータ数が少ない正例を多く正解する必要があり，recall が低い時点で precision を高めるよりも難しい課題であることが理由として挙げられる．疾病予測においては，病気を発症しそうな対象者の見落としを少なくすることを重視した上で，実際は健康な人を誤って病気の発症者と分類してしまう数をより少なくすべきであるが，正例を重視して識別した際には，Ensemble でも他のモデルと同程度の負例の誤分類が起きてしまうという結果になった．

9 結言

本研究ではレセプトデータを根拠に対象病名の発症者を選定し，健診データを学習データとして用いて生活習慣病の発症を予測した．不均衡データに適した学習モデルであるアンダーサンプリング＋バギングモデルと focal loss を損失関数に用いたニューラルネットワークのアンサンブルモデルを構築したところ，ベースラインモデルと比較して有意に識別精度が向上した．しかし，提案モデルにおいて，recall が 0.8 以上の時の precision の値は他のモデルと大きく変わらなかった．recall が高いときでもさらに高い値で precision を維持することができれば，病気を発症する人を間違って予測する数を増やすことなく本当に病気を発症しそうな人にのみ保健指導を施すことができるため，社会的にもより意義のある成果となると考える．今後の課題としては，flNN のパラチューニングの際のクロスバリデーションの fold 数を増やすことで，チューニングに利用できる学習データが増加し，チューニングの精度が向上すると考えられる．flNN と LGB-bag のアンサンブルの比率もチューニングできる部分である．

参考文献

1) Hironori Uematsu, Kazuto Yamashita, Susumu Kunisawa, Tetsuya Otsubo, and Yuichi Imanaka. Prediction of pneumonia hospitalization in adults using health checkup data. *PLOS ONE*, 12(6):1–13, 06 2017.

2) Julia Hippisley-Cox and Carol Coupland. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. *British Journal of General Practice*, 62(594):e29–e37, 2012.

3) Hiroshi Yatsuya, Hiroyasu Iso, Yuanying Li, Kazumasa Yamagishi, Yoshihiro Kokubo, Isao Saito, Norie Sawada, Manami Inoue, and Shoichiro Tsugane. Development of a risk equation for the incidence of coronary artery disease and ischemic stroke for middle-aged japanese–japan public health center-based prospective study–. *Circulation Journal*, 80(6):1386–1395, 2016.

4) Stephen F Weng, Jenna Reps, Joe Kai, Jonathan M Garibaldi, and Nadeem Qureshi. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, 12(4):e0174944, 2017.

5) Hyojeong Lee, Soo-Yong Shin, Myeongsook Seo, Gi-Byoung Nam, and Segyeong Joo. Prediction of ventricular tachycardia one hour before occurrence using artificial neural networks. *Scientific reports*, 6:32390, 2016.

6) Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

7) Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

8) Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*, pages 107–119. Springer, 2003.

9) Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. Class imbalance, redux. In *2011 IEEE 11th international conference on data mining*, pages 754–763. IEEE, 2011.

10) Mitsuru Tsunekawa, Natsuki Oka, Masahiro Araki, Motoshi Shintani, Masataka Yoshikawa, and Takeshi Tanigawa. Prediction of onset of lifestyle-related diseases using regular health checkup data. In *Advances in Artificial Intelligence: Selected Papers from the Annual Conference of Japanese Society of Artificial Intelligence (JSAI 2019)*, volume 1128, pages 14–26. Springer Nature, 2020.

11) Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, pages 1–18, 2020.

12) Ethan Mark, David Goldsman, Brian Gurbaxani, Pinar Keskinocak, and Joel Sokol. Using machine learning and an ensemble of methods to predict kidney transplant survival. *PloS one*, 14(1):e0209068, 2019.

13) Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154, 2017.

14) Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

15) Jie Sun, Jie Lang, Hamido Fujita, and Hui Li. Imbalanced enterprise credit evaluation with dte-sbd: Decision tree ensemble based on smote and bagging with differentiated sampling rates. *Information Sciences*, 425:76–91, 2018.

16) Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):1–21, 03 2015.