2019年度 卒業論文

論文題目
学生の生活及び修学データを用いた
ロジスティック回帰分析による要注意学生の推定

指導教員
舟橋　健司　准教授
伊藤　宏隆　　助教

名古屋工業大学　工学部　情報工学科
2015年度入学　27115132番

名前　福田　太一

目 次

第1章　はじめに

近年, 情報通信技術の発達に伴い, 実験, 観測, 記録, 調査などの電子データが大量に保管されている. そのようなデータを媒体に規則, パターン, 知識を見つけ出す方法を「データの山」から有用な情報を「掘り出す」ことに基づき「データマイニング」と呼んでいる.

データマイニングの事例として, スーパーマーケットにおいて顧客の商品の購入データから購入した商品の組み合わせのパターンを抽出し商品の陳列を見直して利益を向上させる, 機械の故障データから故障の起こりやすい箇所と条件を調査する, オークションにおいて過去の利用データから不正利用者の行動をモデル化し不正が疑われる出品を検知するなどがある. 教育現場においてもデータマイニングは活用されており, ある女子短期大学の一つの科目における生徒の出席率, 宿題提出状況, 試験の採点データの相関から学生の学びの姿勢を分析したり [1], 授業アンケート結果から成績と生徒の授業態度および教育者の授業の進め方を考察する [2] などの事例が存在する.

ところで, 大学を含む多くの教育現場において問題となっているのが「将来的に留年および退学する学生」と定義される「要注意学生」[3] の存在である. 成績不振による留年および退学の対策として, 教員が学習面のアドバイスや相談を行う場を設けている大学も存在するが, 指導する学生一人当たりの指導量の多さや, 指導に要する時間が教員にとって負荷となってしまう.

我が研究室では要注意学生を早期に予測, 推定し指導する学生の絞り込みによる教員の負荷の軽減を目的にデータマイニングを用いた研究を行ってきた [3][4][5][6]. これらの研究において用いられるデータは成績の指標となる Grade Point Average(以下 GPA とする), IC カードリーダーによる講義の出欠席の打刻データが主に用いられており, 主成分分析, ベイジアンネットワークなどの分析手法を用いた研究が進め

られてきた.

本研究も過去のデータマイニングの研究同様に要注意学生の予測, 推定を目的にデータマイニングを行ったが, 本研究ではデータマイニングの対象となるデータにGPA と学生生活実態調査のデータを採用している. このデータには学生の就寝・起床時間, 睡眠時間に関するデータ (睡眠データ), 学生の通学時間・住所および通学手段を記録したデータ (住居・通学データ) が含まれている. また, 分析手法にはロジスティック回帰分析を採用した. ロジスティック回帰分析は目的となる 2 値データ (本研究では True:要注意学生である, False:要注意学生でない) の発生確率を出力する回帰関数をモデリングする分析手法である.

本研究ではモデリングした回帰関数の出力に基づく要注意学生の推定を行った. 採用するデータの組み合わせを変えてそれぞれの推定結果を比較したところ, GPA, 睡眠データ, 住居・通学データを採用したモデルよりも GPA, 睡眠データを採用したモデルの方が推定率が高かった. この結果に加えて, 後者は一部分を除いて GPA のみを採用したモデルよりも推定率が高かった. これらの結果から, GPA と睡眠データは要注意学生の推定に寄与し, 住居・通学データは推定結果にノイズをもたらすという結論に至った.

本論文では, 第 2 章では本研究で用いられたデータマイニングの手法について, 第 3 章では分析に用いるデータの詳細, 第 4 章ではロジスティック回帰分析を用いた要注意学生の推定結果の検証, 第 5 章では推定結果のまとめと今後の課題を述べる.

第2章　データマイニングの手法

本章では本研究のデータマイニングで実際に用いた研究手法の理論について述べる.

2.1　ロジスティック回帰分析

目的変数 (予測したい結果となる変数)y と説明変数 (結果を予測するための変数)x の関係性を探る回帰分析のうち, ロジスティック回帰分析は目的変数が 0（False）か 1（True）の 2 値データである場合に適している分析手法である. この手法によって作成されたモデルによる出力は 0〜1 であり, これは目的変数に含まれる事象が起こる（目的変数が True となる）確率を意味している [6]. 目的変数に含まれる事象は二項分布に従い, その事象が起こる確率を p とするとロジスティック回帰モデル L は次の通りに表される.

$$L = p = \frac{exp(s)}{1 + exp(s)} \tag{2.1}$$

$$s = a + \sum_{i=1}^{n} (b_i * x_i) \tag{2.2}$$

s のパラメータとなる a, b をそれぞれ定数項, 回帰係数と呼び, これらは最尤推定によって決定される [7].

2.2　firth の方法

ロジスティック回帰分析における最尤推定は,「完全分離」と呼ばれる状態が発生すると最尤推定量が定まらない [8]. 完全分離とは, 全ての回帰係数 b に対して

$$bx_i = \begin{cases} < 0 & (y_i = 0) \\ > 0 & (y_i = 1) \end{cases}$$

となる説明変数 x_i が存在する状態を表している. この状態は目的変数 y が 0 のデータ数と 1 のデータ数が不均衡な場合に発生しやすい. 本研究で実際にデータを投入してロジスティック回帰モデルの作成を試みたが, $y = 0$ のデータ数が 108 に対して $y = 1$ のデータ数が 2 と不均衡であったため, 完全分離が発生した. そこで, 完全分離への解決策として「firth の方法」が挙げられる. この方法は最尤推定で用いる対数尤度関数 $l(\theta)$ をフィッシャーの情報量行列 $i(\theta)$ を用いて次の通りに補正する [9].

$$l`(\theta) = l(\theta) + \frac{1}{2} log|i(\theta)| \tag{2.3}$$

この方法を用いてモデルを作成した結果, 完全分離状態でも最尤推定量を一意に定めることができた.

2.3　主成分分析

この方法は, データに含まれる多くの変数の相関関係を考慮してそれらを低い次元の合成変数に縮約する方法である [10]. これによって多くの変数を含むデータが有している情報を解釈しやすくすることができる. 変数 $x_1, x_2, ..., x_n$ をもとに合成変数 z_1 を作成する場合, 係数ベクトルを $a_1, a_2, ..., a_n$ とすると

$$z_1 = a_1x_1 + a_2x_2 + ... + a_nx_n \tag{2.4}$$

となる. ここで, z_1 の分散を最大にすることによって z_1 に含まれる情報量を多くできると考え, z_1 の分散を最大にし, かつ大きさが 1 のベクトル $a` = [a_1, a_2, ..., a_n]$ を導出する. このとき, z_1 を第 1 主成分とする. また, 主成分に変数を入力して得られ

る数値を主成分得点と呼ぶ.

合成変数 z_2 も同様に, 係数ベクトルを $b_1, b_2, ..., b_n$ とすると次の通りに定まる.

$$z_2 = b_1x_1 + b_2x_2 + ... + b_nx_n \tag{2.5}$$

z_2 の分散は z_1 の次に最大にする必要があるが, z_2 には z_1 に不足した情報を補足する役割があるため z_1 とは無相関になるように定める. z_2 の分散を z_1 の次に最大にする大きさ 1 のベクトルを $b' = [b_1, b_2, ..., b_n]$ とすると, z_1, z_2 が無相関になるためには a', b' が垂直であることが条件となっている.

$$\sum_{i=1}^{n} a_ib_i = 0 \tag{2.6}$$

こうして導出される z_2 を第 2 主成分とする.

これらの方法を繰り返して第 n 主成分まで作成するが, 主成分を作成するほどその主成分に含まれる情報量は小さくなっていくため, 情報量が十分に小さい主成分は切り捨てる必要がある. そのための指標として寄与率を用いる. 寄与率とは主成分が全体に対して占める情報量の大きさを表している. 第 m 主成分に対応する固有値を λ_m, 主成分の分散を V とおくと第 m 寄与率 C_m は次の通りに定まる.

$$C_m = \frac{\lambda_m}{\sum_{i=1}^{n} V_i} \tag{2.7}$$

また, 第 m 主成分までの寄与率の合計を第 m までの累積寄与率と呼ぶ. 主成分を採用する際は基本的に累積寄与率が 70〜80 パーセントになるように主成分を採用する.

2.4　変数選択

ロジスティック回帰モデルの作成にあたり, 用いる説明変数の中にはモデルの出力結果にノイズをもたらす説明変数が存在する可能性があるため, 説明変数の取捨選択が必要となる. 本節では本研究で採用した変数選択の方法について述べる.

2.4.1　強制投入法

全ての説明変数を投入してモデルを作成する方法．目的変数の予測に関して各説明変数がどれだけ寄与しているかを調べるために利用されることが多い [6].

2.4.2　ステップワイズ法

モデルの推定率が最も高くなる説明変数の組み合わせを探る方法．この方法には説明変数を増して行う変数増加法, 説明変数を減らして行う変数減少法の 2 つが存在し, 本研究では後者を採用した.

2.5　2 値分類問題

本研究で作成したモデルの評価方法に 2 値分類問題を用いる．2 値分類問題とは, データをあるクラスに属しているデータ (正例) と属していないデータ (負例) に分類する問題である．これにより, データは TP(True Positive), FN(False Negative), FP(False Positive), TN(True Negative) の 4 つのクラスに分類される (表 2.1)．これらのクラスは次の通りに定義される.

TP: 実際の正例を正例と予測したデータ
FN: 実際の正例を負例と予測したデータ
FP: 実際の負例を正例と予測したデータ
TN: 実際の負例を負例と予測したデータ

表 2.1: 2 値分類表

	モデルが正例と予測した	モデルが負例と予測した
実際に正例であった	TP	FN
実際に負例であった	FP	TN

また, それぞれ 4 つのクラスのデータ件数を用いて正解率 (accuracy), 再現率, 適合率, F 値 (F-measure) を次の通りに算出する.

$$\text{正解率} = \frac{TP + TN}{TP + FN + FP + TN} \tag{2.8}$$

$$\text{再現率} = \frac{TP}{TP + FP} \tag{2.9}$$

$$\text{適合率} = \frac{TP}{TP + FN} \tag{2.10}$$

$$F\text{ 値} = \frac{2 * \text{再現率} * \text{適合率}}{\text{再現率} + \text{適合率}} \tag{2.11}$$

正解率はクラスを正しく予測できた割合, 再現率は正例クラスと予測したデータに対する実際の正例クラスの割合, 適合率は実際の正例クラスに対する正例クラスと予測したデータの割合, F 値は再現率と適合率の調和平均を表している. 作成したモデルの評価は正解率と F 値を指標として行う.

第3章　データの詳細

本章では本研究で分析対象となるデータについて述べる．分析するデータにはある年度の名古屋工業大学に在学していた 4 年次の学生 110 人のデータが含まれている．なお, これらのデータには個人を特定する情報は含まれていない．

3.1　分析対象となるデータ

本節では, 分析対象となるデータの種類と詳細について述べる．

3.1.1　留年判定データ

学生が実際に留年したかどうかの判定を記録しているデータ．本来の要注意学生の定義は「将来的に留年および退学する学生」であるが, 4 年次の学生のうち実際に退学した学生は不明であるため, 本研究では便宜的に要注意学生を「将来的に留年する学生」と定義する．結果, 4 年次の学生 110 人に対して要注意学生の人数は 2 人となった．

3.1.2　GPA

学生の成績の指標となる GPA を記録しているデータ．名古屋工業大学では, 受講した科目の成績を秀・優・良・可・不可の 5 段階で評価しており, GPA は以下の方法で算出される．

$$GPA = \frac{4 * \text{秀の単位数} + 3 * \text{優の単位数} + 2 * \text{良の単位数} + \text{可の単位数}}{\text{総履修登録単位数}}$$

また, 算出された GPA を次の方法で偏差値として算出し, 全体の平均値が 50, 標準偏差が 10 になるように標準化している．

$$\text{偏差値} \quad = \frac{(\text{生徒の } GPA - GPA \text{ の平均値})}{\text{標準偏差}} \quad * 10 + 50$$

3.1.3　睡眠データ

学生の平日と休日それぞれの就寝時間, 起床時間, 睡眠時間を記録しているデータ. 就寝時間および起床時間は 24 時制で表記されている. 本研究では就寝時間は 12 時を起点（最小値）に 24 時をまたぎ 11 時を終点（最大値）としているため, 変換前の就寝時間 h を次の方法で H へと変換している.

$$H = \begin{cases} h & (12 \leq h \leq 24) \\ h + 24 & (0 \leq h \leq 11) \end{cases}$$

3.1.4　住居・通学データ

学生の出身高校所在地, 住居, 住所, 通学時間, 通学手段, 入構手段, 同居人に関する学生生活実態調査のアンケート結果を記録しているデータ.

3.2　データの総括

本節では, 節 3.1 で述べたデータを, 第 2 章で述べたロジスティック回帰分析で使用する目的変数, 説明変数に分割する.(表 3.1, 表 3.2)

表 3.1: 目的変数

変数名	意味
留年判定	0(False):留年していない, 1(True):留年した

表 3.2: 説明変数

データの種類	変数名	意味
GPA	GPA	生徒の成績の指標
睡眠データ	平日就寝時間	平日における就寝する時刻
	平日就寝時間	平日における起床する時刻
	平日睡眠時間	平日における睡眠時間の長さ
	休日就寝時間	休日における就寝する時刻
	休日就寝時間	休日における起床する時刻
	休日睡眠時間	休日における睡眠時間の長さ
住居・通学データ	出身校所在地	1:愛知県, 2:岐阜県, 3:三重県, 4:静岡県 5:関西, 6:北陸, 7:関東・甲信越, 8:東北・北海道 9:中国・四国, 10:九州・沖縄, 11:日本国外
	住居	1:自宅から（独立した家庭を持つ人も含む） 2:自宅以外から
	住所	1:大学からの距離が 1km まで 2:大学からの距離が 1km から 5km まで 3:1,2 を除く名古屋市内 4:名古屋市内を除く愛知県内 5:岐阜県または三重県 6:滋賀県
	通学時間	1:30 分未満, 2:30〜60 分, 3:60〜90 分 4:90 分以上
	通学手段	1:徒歩, 2:自転車, 3:原付・自動二輪, 4:自動車 5:JR, 6:地下鉄 7:JR, 地下鉄以外の鉄道 8:バス, 9:その他
	入構手段	1:徒歩, 2:自転車, 3:原付・自動二輪, 4:自動車 5:JR, 6:地下鉄 7:JR, 地下鉄以外の鉄道 8:バス, 9:その他
	同居人	1:ひとりで暮らしている（恒和寮および国際学生寮を含む） 2:家族（親、兄弟、祖父母、親戚など）と同居している 3:友人と同居している（ルームシェアを含む）

第4章　ロジスティック回帰分析による要注意学生の推定

本章ではロジスティック回帰分析モデルを作成し, そのモデルによる要注意学生の推定結果を検証する.

4.1　ロジスティック回帰モデルの定義

この検証ではあらかじめ設定したデータの4通りの組み合わせに基づくロジスティック回帰モデルの検証をそれぞれ検証1, 検証2, 検証3, 検証4と定義する (表4.1). また, 各検証それぞれにおいて説明変数の実データと説明変数を主成分分析して得られる主成分得点を入力とし, 強制投入法とステップワイズ法による変数選択を採用した計4通りのモデルを作成する (表4.2).

表4.1: 検証の種類

	採用するデータ
検証1	GPA, 睡眠データ, 住居・通学データ
検証2	GPA, 睡眠データ
検証3	GPA, 住居・通学データ
検証4	睡眠データ, 住居・通学データ

表4.2: ロジスティック回帰モデルの種類

変数選択 / 入力	強制投入法	ステップワイズ法
実データ	モデル1	モデル2
主成分得点	モデル3	モデル4

4.2　推定モデルの評価方法

モデルの評価方法である 2 値分類問題を解くにあたり, モデルの出力は「その生徒が要注意学生である確率」であるため, その出力があらかじめ設定した閾値 (確率) を超えた場合に正例と予測する. 閾値は 50%〜事前確率 (データに含まれる要注意学生の割合) を設定する. なお, この検証における事前確率は

$$\begin{aligned} \text{事前確率} &= \frac{(\text{要注意学生})}{(\text{対称の学生の人数})} * 100 \\ &= \frac{2}{110} * 100 \\ &\fallingdotseq 1.8(\%). \end{aligned}$$

である. こうして 2 値分類問題を解き正解率, 再現率, 適合率, F 値を算出する. 各検証において再現率と適合率の調和平均である F 値, クラスを正しく予測できた割合である正解率の 2 つを指標にモデルを評価する.

4.3　実験環境

変数の生成およびデータの分析にはそれぞれ Microsoft 社の Excel 2013[11], R version 3.60[13] を利用した.

4.4　推定結果

本節では作成したロジスティック回帰モデルの各検証結果を示す.

4.4.1　検証 1: GPA, 睡眠, 住居・通学データを用いた推定結果

モデル 1,2 のパラメータ, 主成分の固有ベクトル, 主成分の意味付け, モデル 3,4 のパラメータをそれぞれ表 4.3〜表 4.6 に示す. 主成分は累積寄与率が 82.9 パーセントとなる第 6 主成分までを採用した.

表 4.3: モデル 1,2 のパラメータ (検証 1)

変数名	回帰係数	
	モデル 1	モデル 2
定数項	39.1662866	21.11995374
GPA	0.097574676	-0.202581414
平日就寝時間	-0.345702842	不採用
平日起床時間	0.156191657	不採用
平日睡眠時間	0.237862513	不採用
休日就寝時間	-1.000858211	-0.374496075
休日起床時間	0.783385907	不採用
休日睡眠時間	-1.368871016	-0.640855592
出身校所在地	-0.021216325	不採用
住居	-0.885623593	不採用
住所	-0.973946105	-0.735658584
通学時間	0.237229862	不採用
通学手段	0.051335627	不採用
入構手段	0.212640609	不採用
同居人	-0.095380806	不採用

表 4.4: 主成分の係数ベクトル (検証 1)

変数名	係数ベクトル					
	第 1 主成分	第 2 主成分	第 3 主成分	第 4 主成分	第 5 主成分	第 6 主成分
GPA	-0.1961579	0.30338067	-0.14698333	0.745433417	-0.014854218	0.16277502
平日就寝時間	0.4336700	-0.02715254	-0.83171436	-0.069342085	0.065340728	0.07318361
平日起床時間	0.5064597	-0.67063221	-0.21322464	0.005379923	0.432879769	0.04613374
平日睡眠時間	0.1338995	-0.69007608	0.46491001	0.032188430	0.447371442	-0.01724413
休日就寝時間	0.4894760	-0.04834651	-0.78669965	-0.077959777	0.041690757	-0.03118361
休日起床時間	0.4935036	-0.63652740	-0.34521897	0.177258765	-0.380091901	0.10274034
休日睡眠時間	0.1437100	-0.76371115	0.32800506	0.254570456	-0.373902431	0.18038525
出身校所在地	0.4672384	0.48669208	0.10622906	-0.034229251	0.177885888	0.63252614
住居	0.7884073	0.03313755	0.28427982	-0.185986299	0.009364461	0.22537892
住所	-0.8818310	-0.19880490	-0.17307412	-0.013458473	0.119891863	0.20619020
通学時間	-0.8278986	-0.12879937	-0.13408984	0.026109633	0.040193701	0.32522110
通学手段	-0.7994048	-0.24584932	-0.14612762	-0.026789771	0.075708278	0.14557290
入構手段	0.4333609	0.20503292	0.02591604	0.597382655	0.287321398	-0.19678585
同居人	-0.8625877	-0.15073490	-0.25557089	0.028480780	0.121697986	-0.11276414

表 4.5: 各主成分の意味付け (検証 1)

主成分	意味
第 1 主成分	大学の近くに住んでいる
第 2 主成分	早起きで睡眠時間が短い
第 3 主成分	遅い時間に就寝する
第 4 主成分	GPA の高さ
第 5 主成分	平日に遅寝遅起きしている
第 6 主成分	出身校が愛知県より遠い

表 4.6: モデル 3, 4 のパラメータ (検証 1)

変数名	回帰係数	
	モデル 3	モデル 4
定数項	-5.6370960	-6.6539615
第 1 主成分	0.3650795	不採用
第 2 主成分	0.2433448	不採用
第 3 主成分	0.1940164	0.2423449
第 4 主成分	-1.8598503	-3.2566270
第 5 主成分	1.5538136	不採用
第 6 主成分	-1.3619286	-0.7786874

また, 各モデルの精度評価を表 4.7, 正解率 (accuracy) の比較, F 値 (F-measure) の比較を図 4.1, 図 4.2 に示す. 各モデルの閾値ごとの正解率, F 値を比較した結果

$$\text{モデル}\,4 \geq \text{モデル}\,2 \geq \text{モデル}\,3 \geq \text{モデル}\,1$$

となった. これにより, この検証においては入力よりも変数選択の方がモデルの優劣が付きやすく, 変数選択はステップワイズ法, 入力は主成分得点の方が良いモデルを作成できることが分かった. ステップワイズ法で採用された変数に着目すると, モデル 2 は GPA, 休日就寝時間, 休日睡眠時間, 住所, モデル 4 では平日・休日睡眠時間に関わる第 3 主成分, GPA に関わる第 4 主成分, 出身校所在地に関わる第 6 主成分が採用されており, これらの説明変数および主成分は要注意学生である学生の推定に大きく関わると考えられる.

表 4.7: 各モデルの精度評価 (検証 1: GPA, 睡眠データ, 住居・通学データ採用)

モデル	入力	変数選択	閾値	TP	FN	FP	TN	正解率	再現率	適合率	F 値
モデル 1	実データ	強制投入法	50%	2	0	0	108	1.00	1.00	1.00	1.00
			20%	2	0	5	103	0.954	0.286	1.00	0.444
			1.8%	2	0	64	44	0.418	0.030	1.00	0.059
モデル 2	実データ	ステップワイズ法	50%	2	0	0	108	1.00	1.00	1.00	1.00
			20%	2	0	1	107	0.991	0.667	1.00	0.80
			1.8%	2	0	23	85	0.791	0.080	1.00	0.148
モデル 3	主成分得点	強制投入法	50%	2	0	0	108	1.00	1.00	1.00	1.00
			20%	2	0	2	106	0.982	0.50	1.00	0.667
			1.8%	2	0	31	77	0.718	0.061	1.00	0.114
モデル 4	主成分得点	ステップワイズ法	50%	2	0	0	108	1.00	1.00	1.00	1.00
			20%	2	0	0	108	1.00	1.00	1.00	1.00
			1.8%	2	0	19	89	0.827	0.095	1.00	0.174

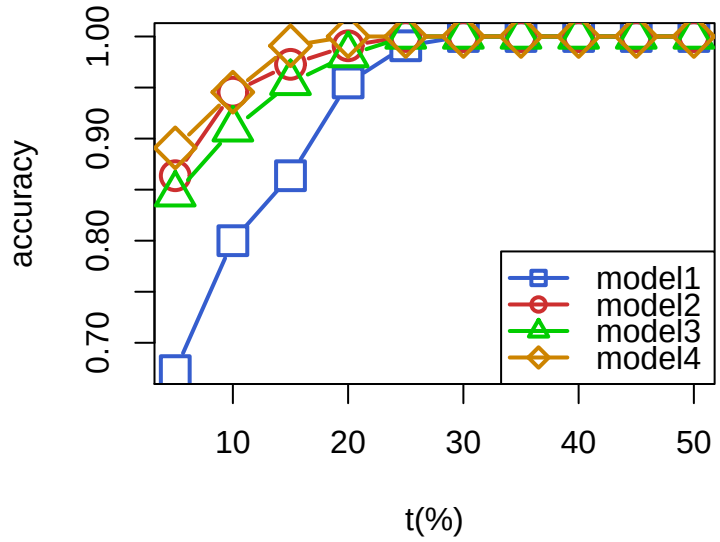

図 4.1: 閾値 t%に対する各モデルの正解率 (accuracy) の比較 (検証 1)

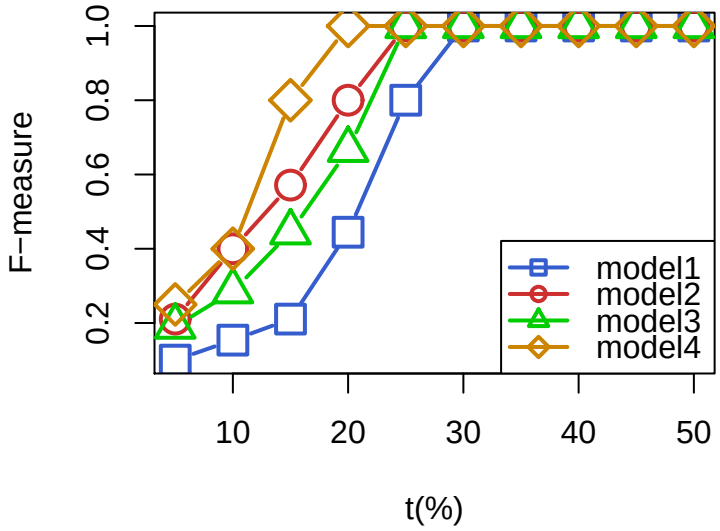

図 4.2: 閾値 t%に対する各モデルの F 値 (F-measure) の比較 (検証 1)

4.4.2　検証 2: GPA, 睡眠データを用いた推定結果

モデル 1,2 のパラメータ, 主成分の固有ベクトル, 主成分の意味付け, モデル 3,4 のパラメータをそれぞれ表 4.8〜表 4.11 に示す. 主成分は累積寄与率が 83.0 パーセントとなる第 3 主成分までを採用した.

表 4.8: モデル 1,2 のパラメータ (検証 2)

変数名	回帰係数	
	モデル 1	モデル 2
定数項	15.6611	14.92662
GPA	-0.21028	-0.26488
平日就寝時間	1.825694	不採用
平日起床時間	0.221003	不採用
平日睡眠時間	-0.27402	不採用
休日就寝時間	-2.02925	-0.16099
休日起床時間	0.153388	不採用
休日睡眠時間	-0.75584	0.48606

表 4.9: 主成分の係数ベクトル (検証 2)

変数名	係数ベクトル		
	第 1 主成分	第 2 主成分	第 3 主成分
GPA	-0.27473	0.245515	-0.85784
平日就寝時間	0.628284	0.70009	0.027444
平日起床時間	0.860941	-0.17133	0.098549
平日睡眠時間	0.367131	-0.76905	0.08977
休日就寝時間	0.653575	0.662462	0.048862
休日起床時間	0.869457	-0.04709	-0.28581
休日睡眠時間	0.472017	-0.70884	-0.32658

表 4.10: 各主成分の意味付け (検証 2)

主成分	意味
第 1 主成分	遅寝遅起き
第 2 主成分	睡眠時間の短さ
第 3 主成分	GPA が低く休日は早起きで 休日の睡眠時間が短い

表 4.11: モデル 3, 4 のパラメータ (検証 2)

変数名	回帰係数	
	モデル 3	モデル 4
定数項	-6.3301	-8.50597
第 1 主成分	0.310975	不採用
第 2 主成分	-0.17408	不採用
第 3 主成分	2.944	4.011669

表 4.12: 各モデルの精度評価 (検証 2: GPA, 睡眠データ採用)

モデル	入力	変数選択	閾値	TP	FN	FP	TN	正解率	再現率	適合率	F 値
モデル 1	実データ	強制投入法	50%	2	0	0	108	1.00	1.00	1.00	1.00
			20%	2	0	5	103	0.954	0.286	1.00	0.444
			1.8%	2	0	32	76	0.709	0.059	1.00	0.111
モデル 2	実データ	ステップワイズ法	50%	1	1	0	108	0.991	1.00	0.50	0.667
			20%	2	0	1	107	0.991	0.667	1.00	0.80
			1.8%	2	0	23	85	0.791	0.080	1.00	0.148
モデル 3	主成分得点	強制投入法	50%	2	0	0	108	1.00	1.00	1.00	1.00
			20%	2	0	0	108	1.00	1.00	1.00	1.00
			1.8%	2	0	23	85	0.791	0.080	1.00	0.148
モデル 4	主成分得点	ステップワイズ法	50%	2	0	0	108	1.00	1.00	1.00	1.00
			20%	2	0	0	108	1.00	1.00	1.00	1.00
			1.8%	2	0	11	97	0.90	0.154	1.00	0.267

また, 各モデルの精度評価を表 4.12, 正解率 (accuracy) の比較, F 値 (F-measure) の比較を図 4.3, 図 4.4 に示す. 各モデルの正解率,F 値を比較したところ, 閾値 1.8〜40 パーセント区域では

$$\text{モデル } 4 \geq \text{モデル } 3 \geq \text{モデル } 2 \geq \text{モデル } 1$$

閾値 40〜50 パーセント区域では

$$\text{モデル } 4 = \text{モデル } 3 = \text{モデル } 1 > \text{モデル } 2$$

となり, 閾値に関わらずモデル 4 はモデル 3 よりも優秀だが, 閾値によってモデル 1 とモデル 2 の優劣が入れ替わる結果となった. これにより, 住居・通学データを採用しない場合は高閾値帯におけるモデル 2 による要注意学生の推定が難しくなることが分かった. また, モデル 4 は主成分の中でも GPA に関わる第 3 主成分のみが採用されており, 検証の結果モデル 3 よりも優秀なモデルであったため, GPA は要注意学生の推定に大きく寄与するデータであり, GPA に関係しない主成分はノイズになりやすいと考えられる.

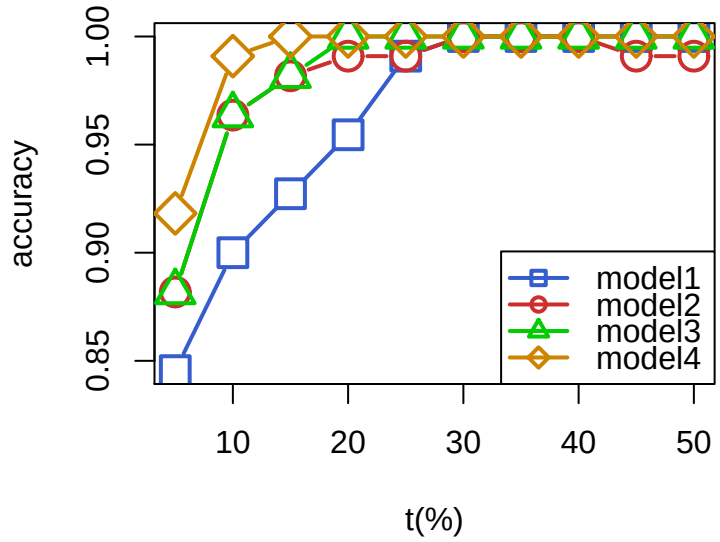

図 4.3: 閾値 t%に対する各モデルの正解率 (accuracy) の比較 (検証 2)

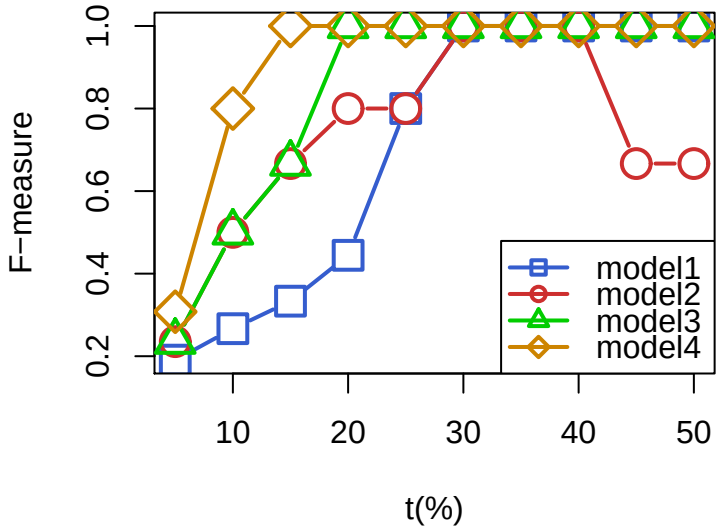

図 4.4: 閾値 t%に対する各モデルの F 値 (F-measure) の比較 (検証 2)

4.4.3　検証 3: GPA, 住居・通学データを用いた推定結果

モデル 1,2 のパラメータ, 主成分の固有ベクトル, 主成分の意味付け, モデル 3,4 のパラメータをそれぞれ表 4.13〜表 4.16 に示す. 主成分は累積寄与率が 78.8 パーセントとなる第 3 主成分までを採用した.

表 4.13: モデル 1,2 のパラメータ (検証 3)

変数名	回帰係数	
	モデル 1	モデル 2
定数項	0.28834604	7.459998
GPA	-0.14319139	-0.293408
出身校所在地	0.07539853	不採用
住居	-0.05040665	不採用
住所	-1.98548092	不採用
通学時間	-0.03284904	不採用
通学手段第 1 位	0.27368911	不採用
入構手段	0.39194652	不採用
同居人	3.73455584	不採用

表 4.14: 主成分の係数ベクトル (検証 3)

変数名	係数ベクトル		
	第 1 主成分	第 2 主成分	第 3 主成分
GPA	-0.1590498	-0.84532259	0.04834646
出身校所在地	0.5649528	-0.24394383	0.71535689
住居	0.8164342	0.21281062	0.21504967
住所	-0.9247263	-0.02532549	0.19907871
通学時間	-0.8489099	-0.06955038	0.29054320
通学手段	-0.8471171	0.04955239	0.09449957
入構手段	0.4605745	-0.57190348	-0.32281132
同居人	-0.9090292	-0.05780032	-0.09619257

表 4.15: 各主成分の意味付け (検証 3)

主成分	意味
第 1 主成分	大学の近くに下宿している
第 2 主成分	GPA が低く入構手段が徒歩に近い
第 3 主成分	出身校が愛知県から遠い

表 4.16: モデル 3, 4 のパラメータ (検証 3)

変数名	回帰係数	
	モデル 3	モデル 4
定数項	-5.33343470	-6.043561
第 1 主成分	-0.08503992	不採用
第 2 主成分	2.22983858	2.320236
第 3 主成分	0.76878424	不採用

表 4.17: 各モデルの精度評価 (検証 3: GPA, 住居・通学データ採用)

モデル	入力	変数選択	閾値	TP	FN	FP	TN	正解率	再現率	適合率	F 値
モデル 1	実データ	強制投入法	50%	1	1	0	108	0.991	1.00	0.50	0.667
			20%	2	0	6	102	0.945	0.250	1.00	0.40
			1.8%	2	0	46	62	0.582	0.042	1.00	0.080
モデル 2	実データ	ステップワイズ法	50%	1	1	0	108	0.991	1.00	0.50	0.667
			20%	2	0	2	106	0.982	0.50	1.00	0.667
			1.8%	2	0	16	92	0.855	0.111	1.00	0.20
モデル 3	主成分得点	強制投入法	50%	0	2	0	108	0.982	NA	0	NA
			20%	1	1	2	106	0.973	0.333	0.50	0.40
			1.8%	2	0	29	79	0.736	0.065	1.00	0.121
モデル 4	主成分得点	ステップワイズ法	50%	0	2	1	107	0.972	0	0	NA
			20%	1	1	2	106	0.973	0.333	0.500	0.40
			1.8%	2	0	21	87	0.809	0.087	1.00	0.160

また, 各モデルの精度評価を表 4.17, 正解率 (accuracy) の比較, F 値 (F-measure) の比較を図 4.5, 図 4.6 に示す. 正解率, F 値は閾値 1.8〜20 パーセント区域においては

$$\text{モデル } 2, \text{モデル } 3, \text{モデル } 4 > \text{モデル } 1$$

となり, 閾値 25〜50 パーセント区域では

$$\text{モデル } 1 \geq \text{モデル } 2 > \text{モデル } 3, \text{モデル } 4$$

となった. モデル 3, モデル 4 は閾値 25〜50 パーセント区域の正解率と F 値はモデル 1, モデル 2 を下回り, 閾値 50 パーセントにおいて F 値は欠損値 (NA) を記録した. 検証 2 におけるモデル 3, モデル 4 はモデル 1, モデル 2 よりも優秀なモデルであったが, この検証において GPA と住居・通学データの主成分得点を入力として作成したモデルは実データを入力して作成したモデルよりも劣る結果となった. また, モデル 1 は検証 2 同様高閾値区域ではモデル 2 よりも優秀であったため, 高い閾値においてはステップワイズ法よりも強制投入法を採用した方がモデルの推定率が上昇する場合もあると考えられる.

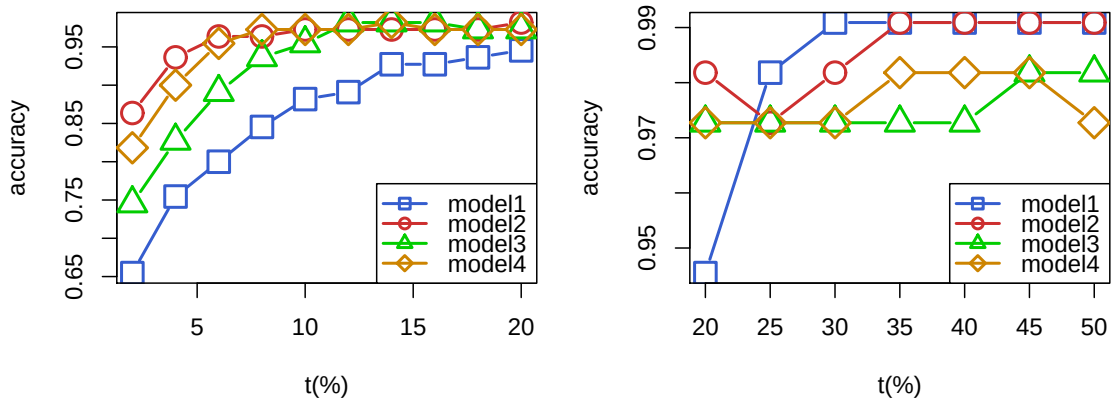

図 4.5: 閾値 t%に対する各モデルの正解率 (accuracy) の比較 (検証 3)

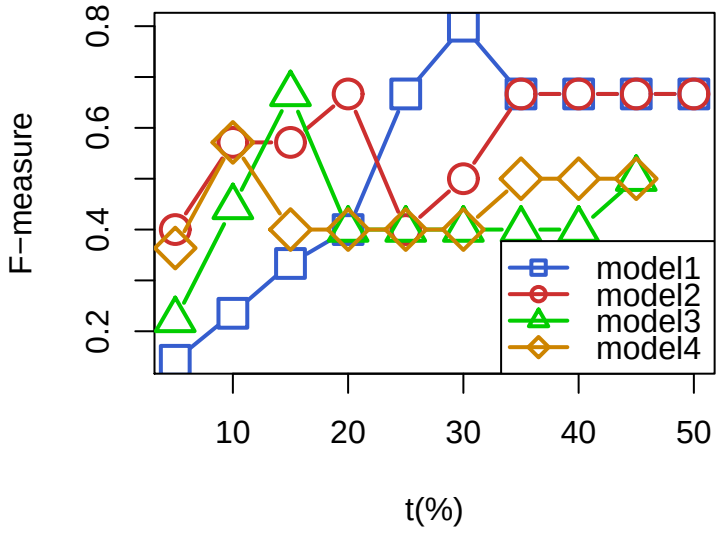

図 4.6: 閾値 t%に対する各モデルの F 値 (F-measure) の比較 (検証 3)

4.4.4　検証 4: 睡眠データ, 住居・通学データを用いた推定結果

モデル 1,2 のパラメータ, 主成分の固有ベクトル, 主成分の意味付け, モデル 3,4 のパラメータをそれぞれ表 4.18〜表 4.21 に示す. 主成分は累積寄与率が 77.8 パーセントとなる第 4 主成分までを採用した.

表 4.18: モデル 1,2 のパラメータ (検証 4)

変数名	回帰係数	
	モデル 1	モデル 2
定数項	25.03320058	5.8118400
平日就寝時間	-0.98496857	不採用
平日起床時間	1.08715727	不採用
平日睡眠時間	-0.20905116	不採用
休日就寝時間	0.16255726	不採用
休日起床時間	-1.11562497	-0.8911583
休日睡眠時間	0.19715656	不採用
出身校所在地	-0.02405460	不採用
住居	-1.43135286	不採用
住所	-1.89355570	-0.7406547
通学時間	0.08034408	不採用
通学手段	0.40552807	不採用
入構手段	-0.22546624	不採用
同居人	-0.48146120	不採用

表 4.19: 主成分の係数ベクトル (検証 4)

変数名	係数ベクトル			
	第 1 主成分	第 2 主成分	第 3 主成分	第 4 主成分
平日就寝時間	-0.4378364	-0.09552294	0.82613148	-0.04549249
平日起床時間	-0.4966391	-0.69397816	0.15611460	-0.05363066
平日睡眠時間	-0.1206505	-0.65878206	-0.52141352	-0.09153411
休日就寝時間	-0.4926809	-0.11316579	0.77977585	-0.06219767
休日起床時間	-0.4880528	-0.68114666	0.28230364	0.10214804
休日睡眠時間	-0.1323795	-0.75281106	-0.39825413	0.15909870
出身校所在地	-0.4745688	0.48635139	-0.07134563	-0.05903056
住居	-0.7850641	0.05046989	-0.28076170	-0.30642316
住所	0.8835436	-0.19820686	0.16346352	-0.03033247
通学時間	0.8284949	-0.12742868	0.12865124	0.02413534
通学手段	0.8045298	-0.23661667	0.13753063	0.04847228
入構手段	-0.4436174	0.17042034	-0.03513663	0.84023478
同居人	0.8626259	-0.15848721	0.24813788	0.06335077

表 4.20: 各主成分の意味付け (検証 4)

主成分	意味
第 1 主成分	大学から遠い場所に住んでいる
第 2 主成分	早起きで睡眠時間が短い
第 3 主成分	遅い時間に就寝している
第 4 主成分	徒歩より遠い手段で入構している

表 4.21: モデル 3, 4 のパラメータ (検証 4)

変数名	回帰係数	
	モデル 3	モデル 4
定数項	-4.20075635	-4.2990184
第 1 主成分	-0.08158273	不採用
第 2 主成分	0.42073630	-0.8849835
第 3 主成分	-0.93150644	不採用
第 4 主成分	-0.14385692	不採用

表 4.22: 各モデルの精度評価 (検証 4: 睡眠データ, 住居・通学データ採用)

モデル	入力	変数選択	閾値	TP	FN	FP	TN	正解率	再現率	適合率	F 値
モデル 1	実データ	強制投入法	50%	1	1	0	108	0.991	1.00	0.50	0.667
			20%	2	0	6	102	0.945	0.25	1.00	0.40
			1.8%	2	0	65	43	0.409	0.030	1.00	0.058
モデル 2	実データ	ステップワイズ法	50%	0	2	0	108	0.982	NA	0	NA
			20%	0	2	1	107	0.973	0	0	NA
			1.8%	2	0	35	73	0.682	0.054	1.00	0.103
モデル 3	主成分得点	強制投入法	50%	0	2	0	108	0.982	NA	0	NA
			20%	0	2	5	103	0.936	0	0	NA
			1.8%	2	0	50	58	0.545	0.038	1.00	0.074
モデル 4	主成分得点	ステップワイズ法	50%	0	2	0	108	0.982	NA	0	NA
			20%	0	2	1	107	0.973	0	0	NA
			1.8%	2	0	43	65	0.609	0.044	1.00	0.0851

また, 各モデルの精度評価を表 4.17, 正解率 (accuracy) の比較, F 値 (F-measure) の比較を図 4.7, 図 4.8 に示す. この検証ではモデル 1 以外のモデルは F 値の欠損値を多く記録したため, モデルの安定性を鑑みた結果モデル 1 を最良のモデルと判断した. 従って, GPA 以外のデータを採用した場合モデル 1 以外のモデルでは要注意学生の推定は難しいと考えられる. また, そのことは逆にモデル 1 以外の実用的な要注意学生推定モデルの作成には GPA のデータが不可欠であるとも言える.

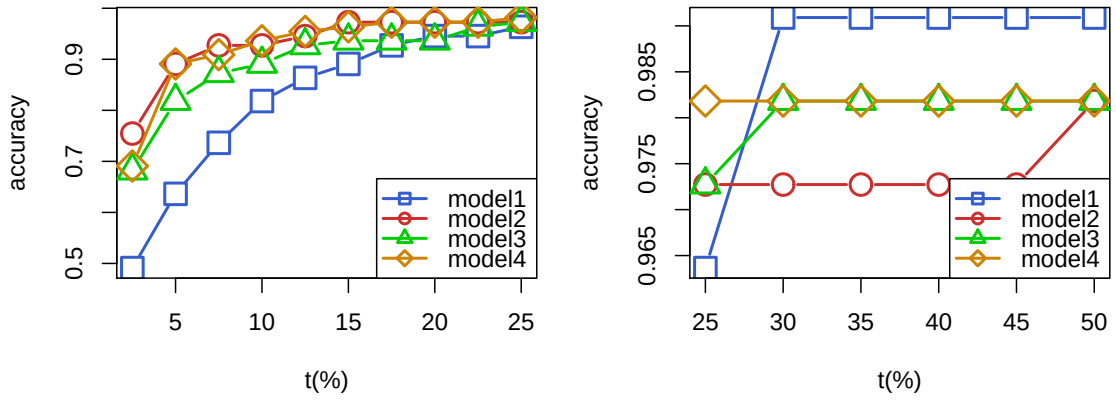

図 4.7: 閾値 t%に対する各モデルの正解率 (accuracy) の比較 (検証 4)

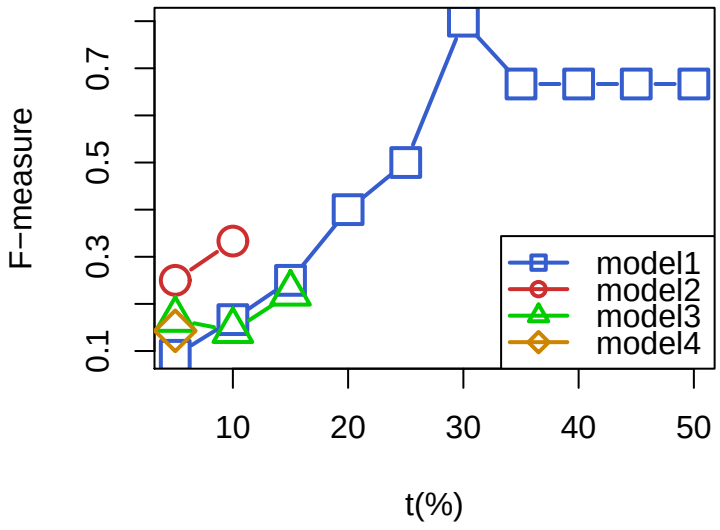

図 4.8: 閾値 t%に対する各モデルの F 値 (F-measure) の比較 (検証 4)

4.5　各検証結果の比較

検証 1, 検証 2, 検証 3, 検証 4 においてそれぞれ最も優秀であったモデルをケース 1, ケース 2, ケース 3, ケース 4 と定義する (表 4.23). これらのモデルの構築方法, パラメータ等は第 4 節を参照されたい. これらの正解率と F 値を比較した結果を図 4.9, 図 4.10 に示す. その結果閾値 5.0〜7.5 パーセント区域では

$$\text{ケース } 3 \geq \text{ケース } 2 \geq \text{ケース } 1 > \text{ケース } 4$$

閾値 1.8〜5.0 パーセント区域または 7.5〜50 パーセント区域では

$$\text{ケース } 2 \geq \text{ケース } 3, \text{ケース } 1 > \text{ケース } 4$$

となった. 閾値が 5.0〜7.5 パーセントの区域を除いてケース 2 が最良となっている. ケース 2 が同じモデル構築法であるケース 1 よりも優れている結果から, 検証 1 で採用した住居・通学データが要注意学生の推定においてノイズとなるデータであったと考えられる. また, 閾値が 1.8〜25 パーセントの場合 GPA のみを変数に持つケース 3 が広い閾値帯でケース 2 に劣る結果から, GPA のデータだけでは最良のモデルは作成できないと考えられる. これらから, 最良のモデル構築に必要な条件は入力が GPA と睡眠データの主成分得点, 変数選択法がステップワイズ法であることが言える. ただ, 一部の閾値を除いてケース 3 がケース 2 よりも優れていた結果を踏まえるとケース 2 とケース 3 の併用が要注意学生の正確な推定に有効であると考えられる.

表 4.23: ケース (各検証における最良モデル) の定義

ケース	定義	採用データ	入力	変数選択
ケース 1	検証 1 のモデル 4	GPA 睡眠データ 住居・通学データ	主成分得点	ステップワイズ法
ケース 2	検証 2 のモデル 4	GPA 睡眠データ	主成分得点	ステップワイズ法
ケース 3	検証 3 のモデル 2 (閾値 1.8〜25 パーセント)	GPA 住居・通学データ	実データ	ステップワイズ法
	検証 3 のモデル 1 (閾値 25〜50 パーセント)	GPA 住居・通学データ	実データ	強制投入法
ケース 4	検証 4 のモデル 1	睡眠データ 住居・通学データ	実データ	強制投入法

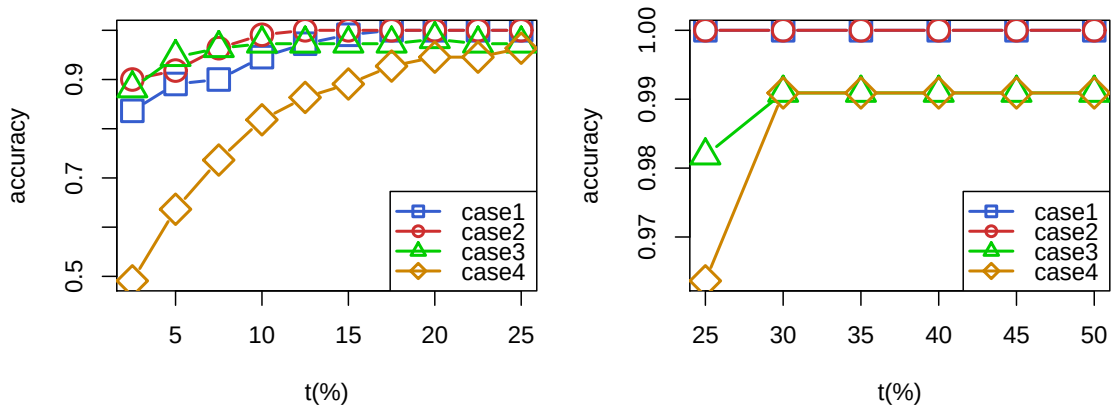

図 4.9: 閾値 t%に対する各ケースの正解率 (accuracy) の比較

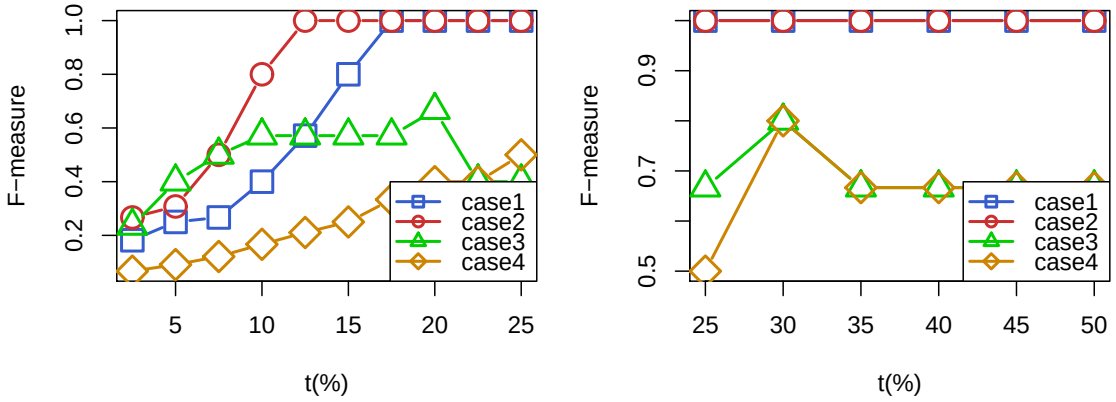

図 4.10: 閾値 t%に対する各ケースの F 値 (F-measure) の比較

第5章　むすび

今回のロジスティック回帰分析を通して, 今回用いたデータの中でも要注意学生の推定に大きく寄与しているのはGPAと睡眠データであり, 住居・通学データはノイズとなることが分かった. また, GPAと睡眠データを採用した場合, 入力は実データよりも主成分得点, 変数選択法は強制投入法よりステップワイズ法を用いることで推定率の優れたモデルを作成できることも明らかになった. 更なる推定率の向上に向けた今後の課題の一つとしてこれらのデータの関係性を様々な手法で分析し, 抽出された新たなる特徴量を入力に用いることが挙げられる. 他の課題としてGPA, 睡眠データ以外のデータを追加することも視野に入れている. 結果的にそのデータが要注意学生の特徴を裏付けるデータであればより精度の高いモデルの作成が期待できる. 将来的に, 実用的な要注意学生の推定モデルを作成し早期の要注意学生の絞り込みによって留年・退学する学生の削減に貢献したい.

謝辞

本研究を進めるにあたり, 日頃から多大な御指導, 御尽力, 御協力を賜りました名古屋工業大学, 舟橋　健司准教授, 伊藤　宏隆助教および舟橋研究室諸氏に心から感謝致します.

参考文献

[1] 南俊郎、大浦洋子:“ 授業データ解析による授業改善策 ”、九州情報大学研究論集第 15 巻、 March 2013.

[2] 原圭司、高橋健一、上田祐彰:“ ベイジアンネットワークを用いた授業アンケートからの学生行動モデルの構築と考察 ”、情報処理学会論文誌、Vol.51, No.4, pp.1215-1226, 2010.

[3] 伊藤圭佑:“ データマイニングによる『要注意学生』の発見に関する研究 ”、平成 25 年度名古屋工業大学修士論文、2013.

[4] 平田大智:“ ベイジアンネットワークによる要注意学生の半期毎の発見精度に関する検証実験 ”、平成 26 年度名古屋工業大学卒業研究論文、2014.

[5] 西脇雅弥:“ 教育支援を目的とした要注意学生の推定精度改善法 ”、平成 27 年度名古屋工業大学修士論文、2015.

[6]] 鈴木博也:“ 修学指導支援のためのロジスティック回帰分析を用いた要注意学生の推定 ”、平成 28 年度名古屋工業大学卒業論文、2016.

[7] D.W. Hosmer Jr, S. Lemeshow, R.X. Sturdivant:“ データ解析のためのロジスティック回帰モデル ”富岡悦良 (監訳)、共立出版、東京都、2013.

[8] 大倉 征幸、鎌倉 稔成:“ 精確ロジスティック回帰の近似推定値 ”、応用統計学 2007；36: pp.87-98.

[9] Firth D: ”Bias reduction of maximum likelihood estimates”, Biometrika 1993; 80: pp.27-38.

[10] 永田靖、棟近雅彦:“ 多変量解析入門 ”、サイエンス社、東京都、2012.

[11] Microsoft Corporation: ”Excel 2013”, https://products.oce.com/ja-jp/excel

[12] The R Project for Statistical Computing: ”R version 3.60”, https://www.r-project.org, 2019 年 12 月 12 日更新、2019 年 12 月 12 日参照.