

# ローン未払い者の経歴の特徴についての分析

厚木 麻里<sup>\*1</sup>, 朝日 弓未<sup>\*2</sup>

## Analysis of the career of bankrupts of loans

by

Atsuki Mari<sup>\*1</sup> and Asahi Yumi<sup>\*2</sup>

(received on Sep.28, 2018 & accepted on Jan.10, 2019)

### あらまし

リーマンショックなどの国際金融情勢の悪化の影響から主要中南米通貨は大幅に下落し、中南米での日本製品の価値は上昇傾向となった。それに伴い分割払いを選択する顧客も増加したが、中にはローンに対する意識の低さなどから支払期間内に返済額を払わない顧客も存在する。この事象は国際的にも問題となっており、各機関では格付けシステムなどへの対策が必要となっている。本研究では中南米A国での二輪車販売ローンデータに基づいて、ロジスティック回帰分析とニューラル・ネットワークにおいてローン返済が未払いになる顧客の経歴の傾向を分析し、結果の違いについて比較を行う。

### Abstract

Due to destabilization of the global economy around 2011, the Latin American currency fell and prices of Japanese goods rose. And the number of customers who make loans together has increased. However, many customers are not accustomed to the loan system, many loan delinquencies occur, which is a serious problem as a vehicle manufacturer. In this analysis, we analyze the sales loan data of vehicles and analyze the characteristics of the customer's history of delinquent loans using logistic regression analysis and neural network analysis. And compare the difference between the results of the two analyzes.

**キーワード:** 中南米, ローン, ローン破産, 経歴, ロジスティック回帰分析

**Keywords:** Latin America, loan, Loan bankruptcy, career, Logistic regression analysis

## 1. はじめに

中南米諸国の実質 GDP 成長率は、2011 年通年で 4.5% となり、2010 年の 6.2% から減速した。A 国ではインフレ抑制のための段階的な政策金利の引き上げを行った。2011 年 8 月以降の国際金融情勢の悪化により、同年後半の成長率は大きく減速<sup>1)</sup>した。主要中南米通貨は大幅に下落し、中南米での日本製品の価値は上昇傾向となった。その数年後に開催された世界選手権大会へ向けて GDP が回復した。その影響からも分割払いを選択する顧客が増加した。中には分割払いにしているにもかかわらず規定の支払期間に払えない顧客も多く存在する。更に、ローン未払いにも関わらず新たなローンを重ね、頭金が高くなってもなおローンを組む事象も起り、国際問題にもなっている。この現状が深刻化すると製品メーカーは、製造費回収が難しくなり経営悪化につながると見込まれる<sup>2)</sup>。

本研究では、GDP が先進国と比較すると依然低い、中南米の A 国で実用的な交通手段として、通勤や農業な

ど経済の発展に欠かせない二輪車の販売ローンデータを基に未払いになる顧客の傾向を分析する。登録された経歴情報からその顧客がどの程度、未払いになる可能性があるのかを求めることを目標とする。

Luigi<sup>3)</sup>らは人種や年齢がローン破産に影響を与える要素だと示している。また、三好<sup>4)</sup>は地域の貸金率、治安、若者層の割合など経歴と地域の特徴が関係すると示している。そこで地域による経歴の差<sup>5)</sup>にも着目しながら、ローン未払いにつながる経歴の差について分析を進める。

## 2. 使用データ

### 2.1 データ概要

本研究で使用するデータは、2010 年 9 月 1 日～2012 年 6 月 30 日までの二輪車販売の中南米 A 国における 14,304 件の顧客データである。

データ項目は申込日、消費者金融からの照会件数、信頼機関のスコア(格付けスコア)、税金未納の履歴、生年月日、就労年、婚歴、性別、就労州、在住州、主収入、副収入、ディーラー査定、分割数、頭金、借入金、利子金額、銀行口座種類、職業、学歴、住居種類、商品種類、商品価格、排気量、サイズ、6 ヶ月で未払い、12 ヶ月で未払い、18 ヶ月で未払い等である。

(1) 年齢

販売ローンを申請した時点の年齢とする。申請日と生年月日の変数から求めた。

\*1 情報通信学部 経営システム工学科 学部課程  
School of Information and Telecommunication  
Engineering, Department of Management  
Systems Engineering, Undergraduate course

\*2 情報通信学部 経営システム工学科 教授  
School of Information and Telecommunication  
Engineering, Department of Management  
Systems Engineering, Professor

## (2) 州

1人の顧客ごとに、在住州、就労州、住民登録州、初回信用調査州の4つの州に関する変数が登録されている。しかし1人の顧客に関して、この4つの州に関する変数が全て同じ州で登録されているデータが全体の81.8%を占める。更に3つの変数が同一の州、残りの1つの変数のみが異なる州で登録されたデータは17.1%であり、この2つ累積割合は98.9%となる。

そこで本研究で扱う州に関する変数を単一化することにした。4つの州に関する変数のうち3つが同じ州、1つがその3つとは異なる州で登録されていた場合、その異なる1つの州が在住州に登録されているデータが最も少ないことが少ない。つまり在住州の信用度が最も高い。以上のことから以降は「在住州」を州のデータとして利用する。

## (3) 変数「nヶ月で未払い」

本変数を説明する上でn=6, 12, 18とする。nカ月で未払いの変数には6ヶ月で未払い(6 months unpaid)、12ヶ月で未払い(12 months unpaid)、18ヶ月で未払い(18 months unpaid)の3種類がある。これはnヶ月時点で支払い遅延が確認されていた場合、1が入る変数として登録されている。尚、何ヶ月時点からローン遅延が発生したかはこの変数からしか推測できない。

nヶ月で未払いに1が入ると自動的に同顧客のそれ以降の月の変数にも1が入り、翌月に不足分まで全額を支払ったとしても再度0が入ることはない。よって登録されるパターンはTable 1の4種類のみである。一度もローン遅延をせずに完済した場合はパターンA(Full repayment)の様にnヶ月で未払いには全て0が入る。また、仮に10ヶ月の時点で初めてローンの支払いが遅れた場合、パターンCのように6ヶ月で未払いには0が入り、12ヶ月で未払いと18ヶ月で未払いにはそれぞれ1が入っている。18ヶ月で未払いに着目するとローン遅延が発生した顧客の場合、必ず1が入っていると言える。

支払い遅延は企業にとって高いリスクであり、更に支払い遅延が発生した顧客はローン破産につながる可能性が非常に高いことから注意が必要である。しかし貧困層の顧客の中にはローンの制度に慣れておらず、返済義務について理解しないままローンの返済を忘れてしまう場合も含まれる。この場合、必ずしもローン破産につながるとは言い切れない。特にA国では景気回復によってローンを組めるような元貧困層も多いため、意図的にローン滞納する顧客とは違う傾向を示す可能性が高い。これによって傾向が定まらず、分析の精度が下がる可能性も考えられる。

「18ヶ月で未払い」の変数には1が入っている割合が高いこと、そして「6ヶ月で未払い」と「12ヶ月で未払い」に1が入っているものは「18ヶ月で未払い」にも1が入っているため、包括的な対策を考えられることから、以降は「18ヶ月未払い者」を「未払い者」として分析を進める。

6ヶ月未払い者：1,206件（全データの9%）

12ヶ月未払い者：2,219件（全データの17%）

18ヶ月未払い者：3,043件（全データの23%）

Table 1 n months unpaid

Pattern		6 months Unpaid	12 months Unpaid	18 months Unpaid
A	Full payment	0	0	0
B	Loan delay Within 6 months	1	1	1
C	Loan delay in 6 to 12 months	0	1	1
D	Loan delay in 12 to 18 months	0	0	1

## (4) 主収入によるグルーピング

A国の経済基盤によって発行された所得階級ランクに基づいて主収入をA/B層、C層、D層、E層の4つに分けた。A/B層は裕福、C層は中間層、D層とE層は貧困層である。A/B層の平均主収入額は145,594.12なのに対し、E層の平均主収入額は836.46と経済格差が見て取れる。データ数はTable 2に示したとおりで、A/B層の割合がかなり低く、全体の1%ほどである。反対に貧困層の割合は高く、D層とE層の合計で8割を占める。

Table 2 State data after grouping

Group name	Number	Percentage of Data
A/B layer	147	1.1%
C layer	2425	18.4%
D layer	5516	41.7%
E layer	5126	38.8%

## 2.2 データクリーニング

就労年数や年齢などのデータにおいて、極端に高い又は低い異常値がある変数を含むデータを削除した。その結果、データ数は14,304件から13,214件となった。

更に主収入の変数に欠損値が239件(全体の1.8%)あるため、データ補完を行なった。方法は以下の通りである。主収入が欠損値であるデータの信頼機関のスコアと同じ値の他のデータを抽出する。そのデータの主収入の平均金額を求め、欠損値に補完する。

## 3. 基礎統計

## 3.1 州ごとの収入格差

州データを農林水産省の地域区分<sup>6)</sup>に基づき、北部(North)、北東部(Northeastern)、中西部(Midwest)、南部(South)、南東部(Southeastern)の5つの区分に分類し、仮の名前をつけた。その州名を平均主収入順で並べたものをFig.1に示した。左から右に行くにつれて収入額は減少する。棒グラフは販売台数(データ数)、折れ線グラフは州別の未払い者の割合を示す。

販売台数と州別の平均主収入の相関は0.300の弱い相関であった。棒グラフからはあまり傾向が見られない。収入が高いから販売台数が高いと言えない訳ではないが、影響はかなり少ないと考察できる。最も販売台数の多い南東部D州は2番目に人口の多い州である。最も人口の多い南東部A州も同じく販売台数が多いが、鉄道など他の交通手段が発展していることからD州よりも販売台数が少ないと考えられる。

未払い者の割合と州別の平均主収入の相関は-0.542と負の相関を示していた。折れ線グラフからも、データ数が少なくて未払い者の割合が極端に上がっている北部A州を除いて全体を見れば、緩やかに右肩あがりになっていることがわかる。つまり、収入が低いほど未払いになりやすいと言える。

州名を見ると左に南東部・南部が多く見られ、右側に北部が偏っていることが分かる。これにより、南から北の方向に貧困層が増す経済格差があることが分かった。

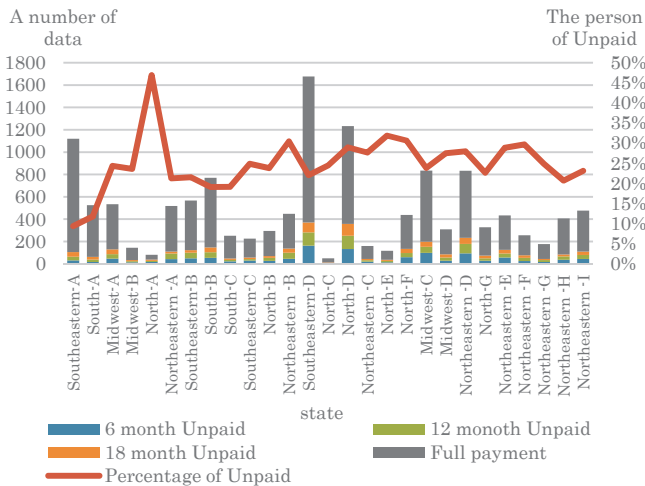


Fig. 1 State data in order of average income

州別の平均主収入額が1番高い南東部と1番低い北部を比較したのがTable 3である。主収入額が違う原因は自営業の割合だろう。自営業の割合は、北部は21%なのに対し、南東部は13%である。収入が少なく、安定しない仕事であるため、借入金額が南東部よりも低かったとしても未払いになる傾向にある。

Table 3 Regional comparison

Average	North	Southeast
Main income	1483.8	1784.807
Bad	28.5%	18.3%
Debt	6636.982	7529.249

### 3.2 学歴

A国の経済格差の原因の一つとして1,000万人の若者の13%が就学しておらず、また仕事にもついていないという現状がある。本データでは2.3%が「教育歴なし」のデータである。一般的に、教育を受ける余裕

がない者は二輪車を購入する余裕もないと考えられることから、本データでの割合は低いことに納得できる。

データ内の教育機関は4つあり、第一学位(First degree)は日本でいう小学校、第二学位(Second degree)は中学校、上級(High school)は高校に相当し、大学院(Graduate degree)も存在する。第二学位までは義務教育になっている。しかし欠席率、不登校率、自主退学率、落第率等が非常に高い。親の教育に対する意識の影響も大きく、高い教育を受けた富裕層の親は子供にも同じ水準の教育を求める。反対に貧しい家庭では子供も仕事をする必要があり、学校に通えなくなる。更に教育施設のインフラが整っていない地域もあり、教育格差が広がっている。

学歴ごとの未払い者の割合、平均主収入金額をFig.2に示した。学歴が上がるにつれて主収入額は上がり、未払い者率は減少していることが分かる。またどの教育課程においても退学よりも卒業している顧客の主収入の方が高い水準にあることがわかる。更に、第二学位を卒業した顧客の方が上級学校を中退した顧客よりも多くの収入を得ていることが分かる。

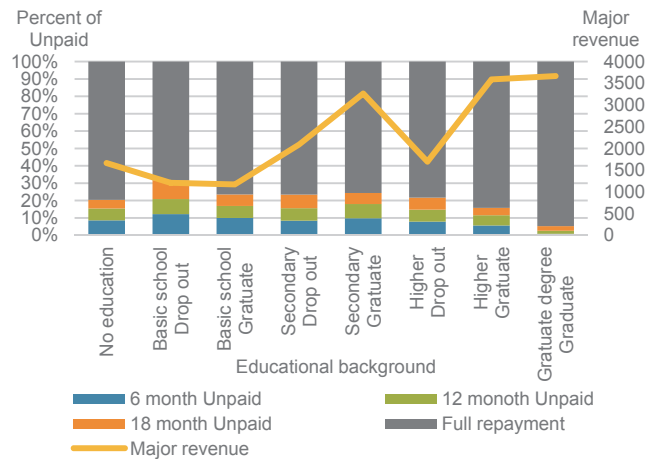


Fig. 2 Data by educational background

## 4. ロジスティック回帰分析

### 4.1 ロジスティック回帰分析とは

ロジスティック回帰分析は目的変数が1になる時の説明変数の影響力を調べることができる分析手法である。比率yと説明変数xの間には式(1)のような関係を想定する。b<sub>0</sub>は定数項(あるいは切片)、b<sub>1</sub>は回帰係数を表す。本研究においては、その人自身が未払い者になることに影響が大きい変数を求めることができる<sup>7)</sup>ようになる。

$$y = \frac{1}{1+e^{-(b_0+b_1x)}} = \frac{1}{1+Exp^{-(b_0+b_1x)}} \quad (1)$$

ロジスティック回帰分析で使用する変数のデータ形は質的変数である。重回帰分析の変数が量的変数のため、学歴や在住州などの数値では表せない変数は使

えないが、ロジスティック回帰分析ではそれらを利用した分析が可能である。

以上の二点から本研究の趣旨に、重回帰分析よりも沿っている<sup>8)</sup>と言えるだろう。

#### 4.2 ロジスティック回帰分析の手法

本研究では目的変数を未払い者であるか否かのダミー変数を作成した。1を未払い者、0をその他支払い者とし、それぞれの割合は1が23.03%、0が76.97%となっている。

説明変数は主収入、学歴、州の経歴に関わる変数を順に入れていく。これにより各変数同士の関係性の理解につながると考える。更にその後、全ての変数を説明変数にして分析し、総合的な分析を行う。モデルの精度についてはR2乗値(寄与率)を見る。0から1の値を示し、大きいほど精度が高いことを表す。2.1章のデータ概要において記した通り、データに傾向が無い為、低い精度であることが予測される。

この分析結果の判断基準は有意確率が0.05以下であることである。そして偏回帰係数(B)を指数変換したオッズ比(Exp(B))を確認する。オッズ比が1よりも大きい場合、その説明変数に1が付けば未払い者になりやすく、反対にオッズ比が1よりも小さい場合、その説明変数に1が付けば未払い者になりにくいと言える<sup>9)</sup>。

#### 4.3 主収入

目的変数を未払い者、説明変数をグルーピング後の主収入とした分析を行った。説明変数が質的変数である必要があるため、層ごとにダミー変数を作成して分析を行った。その際、A/B層のデータ数は147個(全データの12.9%)と極端に少ない為、除外したデータで分析を行った。結果をTable 4に示す。モデルの精度を示すR2乗値は0.027である。

有意確率が0.05以下で、オッズが最も高いのはD層であり、収入額がより少ないE層よりも約1.3倍未払いになりやすいことが分かる。その理由は、二輪車の企業が対策をとっているからである。従って貧困層の中でも、E層より対策不足のD層の方が、収入が多いとしても未払いになりやすい傾向がある。反対にC層は有意確率が高い為、確かではないが、オッズが他の変数よりも低く、この中では最も未払い者になりにくいと分かる。

Table 4 Main income of logistic regression analysis

	B	Standard error	Wald	df	Significance probability	Exp(B)
<b>E layer</b>	0.651	0.248	6.884	1	0.009	1.918
<b>D layer</b>	0.935	0.248	14.258	1	0.000	2.548
<b>C layer</b>	0.209	0.252	0.686	1	0.407	1.232
<b>constant</b>	<b>-1.908</b>	<b>.246</b>	<b>60.203</b>	<b>1</b>	<b>0.000</b>	<b>0.148</b>

#### 4.4 学歴

目的変数は未払い者、説明変数は学歴とした分析を行った。大学院卒業で登録されたデータ数は114件

(0.9%)と極端に少ないため、除外して分析を行った。結果をTable 5に示す。モデルの精度を示すR2乗値は0.005である。

有意確率が0.05以下の有意である変数の中でオッズが最も高いのは第一学位中退であった。他の変数と比べても明らかに高く、特に注意が必要な要因と言えるだろう。また、基礎統計の中では第二学位を卒業している顧客の方が上級学校を中退した顧客よりも高い収入を得ていることが分かったが、この結果を見ると第二学位卒業の顧客の方が約1.2倍未払い者になりやすいことが分かった。

Table 5 Educational background

	B	Standard error	Wald	df	Significance probability	Exp(B)
<b>No education</b>	0.714	0.462	2.391	1	0.122	2.042
<b>First degree dropout</b>	1.323	0.473	7.829	1	0.005	3.755
<b>First degree graduate</b>	0.958	0.446	4.613	1	0.032	2.606
<b>Second degree dropout</b>	0.969	0.444	4.765	1	0.029	2.636
<b>Second degree graduate</b>	1.048	0.433	5.846	1	0.016	2.851
<b>Higher school dropout</b>	0.829	0.444	3.491	1	0.062	2.291
<b>Higher school graduate</b>	0.644	0.443	2.110	1	0.146	1.903
<b>constant</b>	<b>-2.100</b>	<b>0.433</b>	<b>23.575</b>	<b>1</b>	<b>0.000</b>	<b>0.122</b>

#### 4.5 在住州

目的変数を未払い者、説明変数を在住州とした分析を行った。結果を州別の平均主収入の順でTable 6に示す。モデルの精度を示すR2乗値は0.027である。

有意確率が0.05未満でオッズが1を超える未払い者になりやすい州は北部A州、北部E州の2州で、やはり貧困傾向にある北部に寄っていることが分かる。その中でも北部A州は、オッズが3.629と特に未払い者になりやすい州である。北部A州の未払い者の割合は46.9%で一番高い。全体の未払い者率が23.1%なのと比較すると非常に高いことがわかる。この要因は前述の通り、データ数が少ないこともあるが、地域的要因も考えられる。この州には国立公園が開設されるほど多くの自然が残っており、赤道直下型の高温多湿の気候の地方である。平均気温が27℃以上で、年間降雨量は3500ミリメートルと多い。それが川の氾濫の原因となり、自然災害の被害を受けることが負担にな



っている。木材を売って生計を立てる人が多いことから大きな打撃となる。これによりローンを返せなくなり、未払い者になりやすいのだろう。

反対にオッズが最も低く、未払い者になりにくい州だと言えるのは南東部 A 州である。実際に未払い者の割合も 9.4%と低い。平均主収入額は 15530.76 と格段に高いことから経済が安定していることが分かる。その理由はこの州が A 国における工業・商業・金融の中心地であることである。気候に恵まれていて、道路・鉄道・港などのインフラが整備されている。よって工業が発展し、人口も多い。サービス業も展開され、多くの職がある。安定した職を持った富裕層が多いことから、ローンを組んだとしても問題なく完済できる。

このように地域ごとの経済格差はローン未払い者のなりやすさにも影響していることを再確認することができた。

Table 6 State of logistic regression analysis

	B	Standard error	Wald	df	Significance probability	Exp(B)
Southeastern-A	-1.301	0.187	48.579	1	0.000	0.272
South-A	-0.944	0.194	23.593	1	0.000	0.389
Midwest-A	-0.230	0.161	2.044	1	0.153	0.795
Midwest-B	-0.338	0.309	1.191	1	0.275	0.713
North-A	1.287	0.338	14.503	1	0.000	3.623
Northeastern-A	-0.400	0.169	5.629	1	0.018	0.670
Southeastern-B	-0.392	0.179	4.822	1	0.028	0.676
South-B	-0.440	0.156	7.965	1	0.005	0.644
South-C	-0.463	0.219	4.462	1	0.035	0.630
Southeastern-C	-0.068	0.213	0.102	1	0.750	0.934
North-B	-0.244	0.191	1.624	1	0.203	0.784
Northeastern-B	-0.008	0.160	0.003	1	0.959	0.992
Southeastern-D	-0.273	0.135	4.103	1	0.043	0.761
North-C	-0.137	0.410	0.111	1	0.739	0.872
North-D	-0.023	0.137	0.027	1	0.869	0.978
Northeastern-C	-0.041	0.235	0.031	1	0.860	0.959
North-E	0.452	0.261	2.986	1	0.084	1.571
Midwest-C	-0.329	0.148	4.930	1	0.026	0.719
Midwest-D	-0.130	0.186	0.485	1	0.486	0.878
Northeastern-D	-0.014	0.142	0.010	1	0.921	0.986
North-G	-0.426	0.203	4.381	1	0.036	0.653
Northeastern-E	-0.058	0.165	0.124	1	0.725	0.944
Northeastern-F	0.051	0.187	0.076	1	0.783	1.053
Northeastern-G	-0.288	0.225	1.641	1	0.200	0.750
Northeastern-H	-0.409	0.176	5.386	1	0.020	0.664
Northeastern-I	-0.371	0.173	4.576	1	0.032	0.690
constant	<b>-0.802</b>	<b>0.115</b>	<b>48.365</b>	<b>1</b>	<b>0.000</b>	<b>0.449</b>

4.6 全体

最後に、目的変数を未払い者、説明変数は主収入層、

学歴、州を全て入れた分析行った。尚、前項でと同じく主収入の A/B 層と学歴の大学院卒業は除外する。結果は Table 7 の通りである。モデルの精度を示す R2 乗値は 0.051 となり、依然かなり低い値ではあるが、本研究の中では最も高い精度であった。これは変数の数が多いからだと推測する。

Table 7 All

	B	Standard error	Wald	df	Significance probability	Exp(B)
Southeastern-A	-0.206	0.169	1.490	1	0.222	0.813
South-A	1.090	0.170	41.126	1	0.000	2.975
Midwest-A	0.803	0.168	22.846	1	0.000	2.233
Midwest-B	0.938	0.240	15.325	1	0.000	2.554
North-A	1.880	0.264	50.773	1	0.000	6.553
Northeastern-A	0.652	0.172	14.380	1	0.000	1.920
Southeastern-B	0.729	0.169	18.599	1	0.000	2.073
South-B	0.513	0.163	9.947	1	0.002	1.671
South-C	0.513	0.210	5.969	1	0.015	1.670
Southeastern-C	0.873	0.205	18.109	1	0.000	2.394
North-B	0.797	0.192	17.156	1	0.000	2.218
Northeastern-B	1.093	0.170	41.559	1	0.000	2.984
Southeastern-D	0.700	0.146	22.847	1	0.000	2.014
North-C	0.908	0.360	6.352	1	0.012	2.479
North-D	1.029	0.148	48.216	1	0.000	2.799
Northeastern-C	1.045	0.223	21.894	1	0.000	2.844
North-E	1.201	0.241	24.758	1	0.000	3.322
Midwest-C	0.772	0.157	24.142	1	0.000	2.163
Midwest-D	0.964	0.186	27.014	1	0.000	2.623
Northeastern-D	1.007	0.155	42.096	1	0.000	2.736
North-G	0.716	0.189	14.364	1	0.000	2.046
Northeastern-E	1.018	0.172	35.236	1	0.000	2.768
Northeastern-F	1.057	0.192	30.242	1	0.000	2.877
Northeastern-G	0.846	0.221	14.669	1	0.000	2.329
Northeastern-H	0.617	0.182	11.469	1	0.001	1.854
Northeastern-I	0.761	0.173	19.286	1	0.000	2.141
No education	0.967	0.397	5.922	1	0.015	2.629
Basic school Drop out	1.414	0.413	11.711	1	0.001	4.113
Basic school Graduate	1.151	0.384	8.963	1	0.003	3.162
Secondary Drop out	1.212	0.381	10.115	1	0.001	3.360
Secondary Graduate	1.222	0.370	10.909	1	0.001	3.394
Higher Drop out	1.204	0.379	10.070	1	0.002	3.334
Higher Graduate	0.921	0.376	5.995	1	0.014	2.511
C layer	-0.023	0.258	0.008	1	0.929	0.977
D layer	0.562	0.255	4.847	1	0.028	1.755
E layer	0.249	0.256	0.945	1	0.331	1.283
constant	<b>-3.485</b>	<b>0.453</b>	<b>59.225</b>	<b>1</b>	<b>0.000</b>	<b>0.031</b>

有意確率が 0.05 以下のもので、オッズに着目すると北部 A 州が最も高い値を示した。理由は前項と同様だと考える。また、オッズが 1 未満で負の要素として示される変数は南東部 A 州と主収入の C 層のみであり、未払いになりにくい優良顧客の要素が強いと考える。

またオッズは 1 から離れれば離れるほど、その影響の大きさを示す。各変数間での影響の強さを表す為、1 を基準としたオッズの絶対値を主収入層、学歴、州の変数ごとに平均した。その結果、主収入層は 1.516、学歴は 2.215、州は 0.354 であった。よってこの変数の中で最も影響力の強い変数は学歴であると言える。

## 5. ニューラル・ネットワーク分析

### 5.1 ニューラル・ネットワーク分析とは

ニューラル・ネットワーク分析は要因の検出という観点から決定木分析や非線形回帰の一種である。Fig. 3 に示すようにニューラル・ネットワークで用いられる多層パーセプトロンモデルでは、入力層と出力層の間に中間層（隠れ層）を持つことが特徴である。それぞれの入力値を  $x$  とし、それぞれの内部に持っている値を  $w$  とすると、入力層から隠れ層への流れは式 (2) のように表すことができる。 $u$  は  $x$  の総入力値を表す。この  $u$  に活性化関数  $f(x)$  を通して式 (3) のような出力  $z$  を得ることができ、隠れ層の数が増えても同様の計算を繰り返す。活性化関数についてはシグモイド関数や双曲線正接関数などがある。本分析で利用した関数については後述する。

$$\left. \begin{aligned} u_1 &= x_1w_{11} + x_2w_{12} + x_3w_{13} + b_1 \\ u_2 &= x_1w_{21} + x_2w_{22} + x_3w_{23} + b_2 \\ u_3 &= x_1w_{31} + x_2w_{32} + x_3w_{33} + b_3 \end{aligned} \right\} (2)$$

$$z = f(u) \quad (3)$$

ニューラル・ネットワーク分析は神経回路図を数式的なモデルにした機械学習の 1 つで、ディープラーニングに利用されている。母集団の分布型（母数）について一切の仮定を設けない、ノンパラメトリックの分析手法である。本分析は高い精度の推定性・説明力が検証され、近年予測手法として脚光を浴びていて<sup>10)</sup>、与信審査などでも利用される<sup>11)</sup>。高い推定力はあるが、中間層の値についてブラックボックス状態であるため、分析の経緯を見るためには他の分析と組み合わせる必要がある。

前項のロジスティック回帰分析では質的変数のみが利用できたのに対し、ニューラル・ネットワーク分析では変数の形に制限はなく、量的変数も扱うことができる。よって、学歴や在住州などだけでなく、グルーピングをしていない収入の変数や年齢などの変数も同時に利用することができる。量的変数をグルーピングした場合、その分け方によっては違う結果が出力される場合がある。しかし量的変数で分析を行えば、より正確な結果に期待できる。よって、総合的な判断を行うという観点において、ロジスティック回帰分析

よりも適していると考える。そして正統的な手法であるロジスティック回帰分析とニューラル・ネットワーク分析で比較を行っていく。

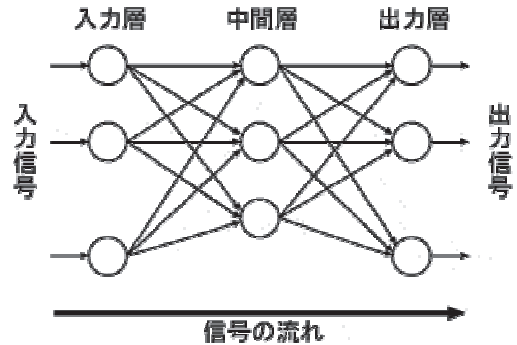


Fig.3 Neural network (multilayer perceptron) model

### 5.2 ニューラル・ネットワーク分析の手法

ニューラル・ネットワーク分析は機械学習であるため、同じ条件で分析を行ったとしても結果が毎回変わる。そのため、データを学習用のデータ (learning) とテスト用のデータ (test) に分けて分析を 5 回行う。事前にランダムに学習用データとテスト用データを 7:3 の割合<sup>8)</sup> で分け、毎回同じ学習データとテストデータで分析を行う。学習データで構築されたモデルを両データに適用させて行った推定作業がクロス集計で結果として出る。そのクロス集計において、予測値と実際のデータが同じ、つまり正解した時の割合を平均し、その分析の精度を見る<sup>12)</sup>。

本研究ではロジスティック回帰分析と同じく、未払い者のダミー変数を従属変数に置く。入力層については、因子は在住州と学歴、共変量に主収入額と年齢の計 39 ユニットとした。学習方法はバッチ法、アルゴリズムの最適化は調整された共役勾配である。隠れ層を 1 層から 50 層までに設定し、その他を自動構築で設定した結果、活性化関数  $f$  は隠れ層では式 (4) の双曲線正接、出力層では式 (5) のソフトマックス関数を利用した。誤差関数は交差エントロピーを利用している。隠れ層は 1 層で 39 ユニットである。

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (4)$$

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=0}^N \exp(x_j)} \quad (5)$$

### 5.3 ニューラル・ネットワーク分析の結果

結果は Table. 8 の通りである。学習用データの未払い者の割合は 22.8%、テストデータの未払い者の割合は 23.6% である。つまり全ての予測を 0 とした場合、それぞれの正解率は 77.2% と 76.4% となる。分析結果を見るとほとんどがその正答率に近かった。事実、クロス集計の結果を見ても、ほとんどの予測値が 0 で付けてしまい、予測値で 1 が付いた回数は正誤を問わず平均 3.1 回である。その中で予測値が 1 で、且つ正解したのは平均 1 回のみであることから、今回の分析はうまく機能しなかったと言える。

Table 8 Neural network analysis -accuracy

		Percentage of correct answers
1	learning	77.2%
	test	76.4%
2	learning	77.2%
	test	76.3%
3	learning	77.1%
	test	76.4%
4	learning	77.2%
	test	76.4%
5	learning	77.2%
	test	76.4%
Average	learning	77.2%
	test	76.4%

本分析では独立変数の重要度が出力される。各回で変数の合計は 100%となっている。あくまで重要度であるため、パーセンテージが高いからと言って未払い者になりやすいとは言えない。しかし未払いとの関係が深い変数を見ることで、重要視すべき変数を見ることができる。

本分析の計 5 回の結果は Table 9 の通りである。全体の平均を見ると年齢の重要度は 31.5%と、未払いになるかどうかの影響が大きいことが分かった。

Table 9 Neural network analysis -importance

	State	Educational	Age	Main income
1	25.4%	14.4%	32.8%	27.4%
2	30.0%	26.5%	34.0%	9.5%
3	14.9%	11.4%	32.2%	41.5%
4	32.1%	11.6%	29.2%	27.0%
5	30.6%	18.0%	29.3%	22.1%
Average	26.6%	16.4%	31.5%	25.5%

#### 5.4 ロジスティック回帰分析の精度と比較

最も違いが大きかったのは学歴の変数である。4.5 章で行ったロジスティック回帰分析のオッズの絶対値の平均は他の 2 変数よりも高かったのに対し、ニューラル・ネットワーク分析での重要度は他の変数よりも低くなっている。これは変数の数の違いによるものだと考察する。州の変数は 26 種類あるのに対し、学歴では 8 種類である。その為、自然と重要度が下がるのではないかと考える。

### 6. まとめ

中南米 A 国の二輪車販売ローンデータから未払い者の経歴の傾向を知るために分析を行った。

地域・州に関しては北部から南部へ向けて経済格差があり、平均主収入額が低い北部・北東部では未払い者になりやすい傾向にあることが分かった。その中でも特に北部 A 州が未払い者になる傾向が強いとロジスティック回帰分析によって求められた。

主収入を 4 つの層に分けてから行ったロジスティック

回帰分析の結果では、最も収入額の低い E 層よりも D 層の方が未払い者になりやすい傾向を見ることができた。必ずしも、主収入額が低ければ低いほど未払い者になりやすいとは言いきれないのである。学歴に関しては、平均主収入額が多い第二学位卒業者の方が平均主収入額の低い上級中退者よりも未払い者になりやすいことが分かった。これらは二輪車メーカーが頭金や利子を上げてローンを組みにくくし、信用の薄い顧客のローンのリスクを回避した結果だと考える。

今後の課題は他の変数なども利用したニューラル・ネットワーク分析の実施である。これにより精度の向上が見込まれる。今回の分析においても、因子や共変量などの組合せを 20 パターン程度試みたが、結果に変化はなかった。これは学習データの不足が原因だと推測する。今後は学習データの割合を増加させることを考慮しつつ、隠れ層の層数の増加、パラメータの調節なども試みたい。また変数の種類が少なく、比較が難しいと考えて利用を控えた職種の変数に関しても利用していきたいと考える。A 国では公務員の給与水準が高く、民間企業の中でも優秀な社員は公的機関への転職を期待しているほどである。その公務員と最も割合の多いサラリーマンの違いに着目して分析を行いたい。

### 参考文献

- 1) 経済産業省：通商白書2012, 第1章 世界経済の動向, 第6節 中南米, ロシア経済, pp128-145
- 2) 書間文彦：アメリカにおける個人破産に関する実証研究サーベイ, 早稲田商学第400号, pp.171-214, 2004年
- 3) L.G.P.S Luigi Zingales, "The Determinants of Attitudes toward Strategic Default on Mortgages," The Journal of FINANCE Vol.68, No.4, 2013
- 4) 三好祐輔：消費者金融市場における逆選択の実証研究, 香川大学経済論叢 第88号pp.537-542, 2016
- 5) 中嶋健太：中南米における二輪車販売の顧客与信管理モデル, 東海大学紀要情報通信学部 Vol.10, No.2, pp26-32, 2017
- 6) 農林水産省：平成21年度カンントリーレポート <http://www.maff.go.jp/primaff/koho/seika/project/pdf/nikokukan12-3.pdf> (最終閲覧日：2017/12/18)
- 7) 内田治「SPSSによるロジスティック回帰分析」オーム社, 2011
- 8) 内田治：ロジスティック回帰分析におけるモデルの適合度指標に関する考察と提案, 東京情報大学研究論集, pp.9-14, 2004
- 9) 内閣府：平成22年度 高齢者の現状及び今後の動向分析についての調査報告書, pp.90-113, 2010
- 10) 田中克明・勝田英紀・萩原統宏：ニューラル・ネットワークによる格付付与構造の 解明のための説明変数の数について, 経営情報研究 第18巻第2号 (2011年2月), pp.25-38
- 11) 柴内康文：調査データ分析におけるニューラルネットワークモデルの適用, 評論・社会科学 88, 45-60, 2009
- 12) 樋口大輔：消費者金融顧客の自己破産—その特徴と原因—, 消費者金融サービス研究所, pp.1-51, 2004