李 聖 鍵 敎授指導

碩士學位 請求論文

영과잉-포아송분포(zero-inflated Poisson)를 이용한 신용평가모형에 관한 연구

2007

誠信女子大學校 大學院

統計學科

金 株 映

영과잉-포아송분포(zero-inflated Poisson)를 이용한 신용평가모형에 관한 연구

李 聖 鍵 敎授指導

이 論文을 碩士學位 論文으로 提出함

2006年 11月

誠信女子大學校 大學院

統計學科

金 株 映

認　准　書

金珠映**의** 碩士學位 論文**으로** 認准**함.**

審査委員_____印

審査委員_____印

審査委員_____印

誠信女子大學校 大學院

논 문 개 요

영과잉-포아송 분포는 이산형 확률분포에 있어서 정상적인 포아송 확률분포보다 영의 값이 과잉관측되는 경우에 적용할 수 있는 분포이다. 이러한 영의 값이 과잉관측되는 데이터는 여러 분야에서 현대문명의 발달과 기술의 발전 등으로 인해 나타나고 있으며, 그 중 한 분야인 금융분야에서도 금융기관의 체계적인 신용위험관리방법 도입으로 인한 연체율의 현저한 감소로 반응변수의 값에 영이 과잉관측되는 결과(정상고객은 0, 연체고객은 1이라 하면)를 가져왔다. 이를 기존의 포아송 회귀모형에 적합 시키는 것은 적절한 방법이라 할 수 없다. 따라서 본 논문에서는 영이 과잉관측된 국내 A은행의 연체정보데이터에 영과잉-포아송 분포를 적용한 영과잉-포아송 회귀모형을 구축하고 기존의 모형적합 방법인 로지스틱 회귀모형, 포아송 회귀모형, 의사결정 나무모형을 적합하여 적합된 모형을 예측력의 측면에서 비교하여 살펴본다. 또한 신용평가모형으로서, 본 논문에서 제안하는 영과잉-포아송 회귀모형의 적용 및 활용 가능성을 검토한다.

목 차

Ⅰ. 서 론

금융사업은 위험관리를 통해 수익을 창출하는 사업이다. 위험이란 금융상품의 가격을 결정하는 요소로서 고위험에는 고금리의 적용이 원칙이다. 따라서 위험은 이자율을 결정하는 중요한 요인으로서 금융사업에 있어 체계적이고 효율적인 위험관리는 은행의 경쟁력 및 수익성을 제고시키는 핵심적 역할을 하게 된다.

금융위험에는 시장위험(Market Risk), 신용위험(Credit Risk), 경영관리위험(Operational Risk) 등이 있으며 특히, 신용위험은 통제가 가능하면서 고객을 대상으로 계량이 가능한 위험이다. 따라서 금융기관의 거래고객 및 거래량의 증가로 인하여 신규고객유치 및 기존고객에 대한 효율적이고 체계적인 위험관리의 방법들이 필요하게 되었다.

이러한 신용위험관리방법으로는 CTI(Computer Telephony Integration), 고객관리(Customer Relationship Management), 개인신용평점제도(Credit Scoring System) 등이 있으며, 이 중 개인신용평점제도는 신청평점시스템과 행동평점시스템으로 구분되어 고객의 신용을 계량화한다(김학신, 2003).

개인신용평점제도 중 신청평점시스템은 신청서 상의 정보 및 외부정보기관의 정보를 이용하여 신용공여 가부를 결정하는 방법으로 신청서 처리의 효율 및 신용조사비용 절감, 일관된 의사결정, 연체율 및 신용손실 감소화의 효과를 이룰 수 있다. 이에 본 논문은 위험관리기법 중 하나인 신청평점시스템을 위한 새로운 영과잉-포아송 회귀모형(zero-inflated Poisson regression model)을 제안한다.

영과잉-포아송 회귀모형은 반응변수가 영과잉-포아송 분포(zero-inflated Poisson distribution)임을 가정한 회귀모형이다. 여기서 영과잉-포아송 분포란 이산형 확률분포에 있어 정상적인 포아송 확률분포보다 영의 값이 과

잉관측되는 경우를 위한 변형된 확률분포를 말한다. 만일 반응변수에 영이 과잉관측되는 경우 기존의 포아송 분포에 적용시켜 통계적 추정과 검정을 하게 되면 이는 제3종 오류를 범하는 결과를 초래하게 된다. 이러한 영과잉-포아송(zero-inflated Poisson) 분포는 Singh(1963)에 의해 처음 소개되었으나 수학적인 모형으로만 인식되어 응용분야가 다양하지 못하였다. 그러나 최근 Lambert(1992)는 Heilbron(1989)의 영변경(zero altered)-포아송 음이항 회귀모형과 유사한 공변량에 의존되는 반응변수가 영과잉-포아송 분포를 따르는 영과잉-포아송 회귀모형을 소개하였고 이를 실제의 자료에 응용하였다.

최근 금융기관의 체계적인 신용위험관리방법 도입으로 인한 연체율의 현저한 감소로 반응변수의 값에 영이 과잉관측되는 결과를 가져왔으며, 이를 기존의 포아송 회귀모형에 적합한다면 모형의 적합력의 문제를 해결할 수가 없다. 이에 본 논문은 영과잉-포아송 회귀모형과 기존의 모형적합 방법인 로지스틱 회귀모형, 포아송 회귀모형, 의사결정나무모형을 실제 금융데이터를 이용하여 모형을 적합하여 보고 적합된 모형을 예측력의 측면에서 비교하여 살펴보려고 한다.

Ⅱ. 신용평가모형의 이론적 배경

1. 신용평가모형(Credit Scoring Model)의 개념

신용이란 일반적으로 어떤 경제주체가 다른 경제주체에 미치는 믿음과 확실성의 정도를 의미한다. 좀 더 구체적으로 이는 경제활동에서 개인이나 기업의 금전이나 재화를 정해진 기간 내에 당초의 약속대로 상환지불 또는 변제할 수 있는 능력이라고 정의할 수 있다.

신용의 종류로는 개인신용(customer credit), 상품신용(merchandise credit), 투자신용(investment credit), 은행신용(bank credit), 공공신용(public credit)이 있다.

그런데 위와 같은 신용의 종류에 따라 필요로 하는 신용정보의 질과 양은 차이가 있다. 금융기관의 여신공여와 투자자의 투자판단을 위한 신용정보는 질적·양적으로 많은 정보를 요구하고 있으며, 개인의 신용과 관련한 정보는 개인의 사회적 지위, 과거의 대금지급성향, 재산상태 등의 정보를 요구한다.

거래대상이 다양화되고, 거래영역이 커지고, 거래형태가 복잡화됨에 따라 고도의 신용교환을 바탕으로 하는 사회가 도래하면서 거래활동에 있어 신용이 중요한 위치를 차지하게 되었다. 그러나 신용 공여자가 고객의 신용상태를 일일이 조사하여 파악한다는 것은 현실적으로 거의 불가능하고 이를 정확하게 판단할 수 있는 능력을 갖추기도 어렵다(이희만 · 박관수, 1999).

따라서 신용거래에 있어 신용의 가치를 측정할 수 있는 도구의 필요성을 느끼게 되었다. 이러한 도구의 하나로 사용되는 신용평점시스템(credit scoring system)은 신규 신용거래 신청자에 대한 현재의 사회적, 경제적 요소들을 분석함으로써 미래의 신용거래 행위를 예측하는 통계적 모형으로 과거 신용거래

자로부터 얻어진 경험적 정보를 바탕으로 우·불량 예측을 위한 확률을 계산하고 구조화함으로써 현재의 신청자를 우량집단 또는 불량집단으로 분류할 수 있도록 점수화하는 절차이다. 또한 고객집단별 특성을 파악하여 신용위험을 구체화시킨 평점표를 근거로 고객 개인의 신용상태를 점수로 평가하는 것을 말한다.

2. 로지스틱 회귀분석(Logistic Regression Analysis)

반응변수 Y가 2개의 가능한 값(0, 1)을 갖는 이항 반응이고 설명변수 X가 있을 때 반응변수 Y의 평균은 $Y=1$의 값을 가질 확률과 같다. 성공 또는 불량 등의 확률 P_x에 대해 선형 확률모형(linear probability model)은 $P_x=\alpha+\beta x$이다. 그런데 이러한 모형은 확률의 값이 0과 1사이의 값을 가져야 하므로 구조적으로 결함을 갖고 있고 또한 모수의 추정에 있어서도 최소제곱추정량(LSE)은 선형 불편추정량에 있어서 더 이상 최소 분산을 갖지 않는다. 따라서 이러한 식 대신 관심의 확률 P_x에 대해 설명변수 X와 비선형 식인

$$P_x=\frac{\exp(\beta_0+\beta x)}{1+\exp(\beta_0+\beta x)} \tag{2.1}$$

을 사용하게 되었다. k개의 설명변수가 있는 경우에 관심확률 P_x는

$$P_x = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)} \tag{2.2}$$

와 같이 모형화되며, 로짓모형은 아래의 식과 같다.

$$\ln \frac{P_x}{1 - P_x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \ . \tag{2.3}$$

로짓모형에서의 모수(parameter)들인 β_i의 최대우도추정은 각 표본들이 독립이라는 가정 아래 다음과 같이 주어지는 우도함수

$$\begin{aligned} & L(\beta_0, \beta_1, \cdots, \beta_k) \\ & = \prod_{i=1}^{n} f(Y_i | x_i) = \prod_{i=1}^{n} \pi(x_i)^{Y_i} [\, 1 - \pi(x_i) \,]^{1 - Y_i}, \qquad Y_i = 0, 1 \end{aligned} \tag{2.4}$$

의 최대화로부터 구해진다.

각 설명변수들이 반응 변수에 영향을 끼치는 정도는 회귀 계수 β_i를 추정함으로써 알 수 있다. 이 추정량들은 최대우도 추정량(MLE)을 통해 구해지며 근사적으로 정규분포를 따른다. 로짓모형에서의 모수(parameter)들인 β_i의 추정치를 b_i라고 하고 이의 표준오차를 s_{b_i}라고 하면 b_i / s_{b_i}는 귀무가설을 $H_0 : \beta_i = 0$으로 했을 때 근사적으로 표준정규분포를 따르게 되고 $\left(\frac{b_i}{s_{b_i}} \right)^2$는 근사적으로 자유도 1을 갖는 카이제곱 분포를 따르게 된다. 귀무가설과 대립가설을 다음과 같이 세웠을 때

$$H_0 : \beta_i = 0 \qquad \text{vs.} \qquad H_1 : \beta_i \neq 0 \ ,$$

이를 위한 검정 통계량은 왈드 통계량 $\chi_w^2 = \left(\frac{b_i}{s_{b_i}} \right)^2$으로 이 통계량이 자유도 1인 카이제곱분포의 임계치 $\chi_\alpha(1)$보다 클 때 귀무가설을 기각하게 된다.

모형의 유의성검정은 우도비검정(likelihood ratio test)을 통하여 구해지는데 귀무가설은 모형에 포함된 k개 독립변수들이 어떤 사건이 발생될 확률에 전혀 정보를 제공하지 못한다는 의미이고, 대립가설은 적어도 하나 이상의 유의한 영향을 미치는 독립변수가 모형에 포함되어 있다는 의미이다.

우도비 카이제곱 검정통계량은 설정된 모형의 우도와 모든 회귀계수가 0으로 제약된 모형의 우도를 비교하는 것이다. 전자는 k개 독립변수를 포함한 모형에서 최우추정법에 의해서 구해진 우도로 $L(\hat{\beta})$로 표시하자. 여기서 $\hat{\beta}$는 절편을 포함한 추정된 회귀계수벡터이다.

$$\begin{aligned} L(\hat{\beta}) &= \prod_{i=1}^{n} \pi(x_i)^{Y_i} [\ 1 - \pi(x_i)\ ]^{1-Y_i} \\ &= \prod_{i=1}^{n} \left(\frac{\exp(x_i^{'} \hat{\beta})}{1+\exp(x_i^{'} \hat{\beta})} \right)^{Y_i} \left(\frac{1}{1+\exp(x_i^{'} \hat{\beta})} \right)^{1-Y_i} . \end{aligned} \qquad (2.5)$$

후자는 귀무가설이 참인(모든 k개 기울기 계수들이 동시에 0일 때) 제약된 모형에서 최우추정법에 의해서 구해진 우도 $L(\hat{\beta_0})$는 다음과 같이 구해진다.

$$L(\widehat{\beta_0}) = \prod_{i=1}^{n}\left(\frac{\exp(\widehat{\beta_0})}{1+\exp(\widehat{\beta_0})}\right)^{Y_i}\left(\frac{1}{1+\exp(\widehat{\beta_0})}\right)^{1-Y_i}$$

$$= \prod_{i=1}^{n}\left(\exp(\widehat{\beta_0})\right)^{Y_i}\left(\frac{1}{1+\exp(\widehat{\beta_0})}\right) \quad . \tag{2.6}$$

우도비 카이제곱 검정통계량 LR은 두 가지 우도의 비율에 자연대수를 취하여 다음과 같다.

$$LR = -2\ln\left(\frac{L(\widehat{\beta_0})}{L(\widehat{\beta})}\right)$$

$$= -2\ln L(\widehat{\beta_0}) + 2\ln(\widehat{\beta}). \tag{2.7}$$

카이제곱 검정통계량 LR은 귀무가설이 참일 때 자유도 k인 카이제곱분포 χ_k^2에 따르게 된다(성웅현, 2001).

3. 포아송 회귀분석(Poisson Regression Analysis)

반응변수 Y가 계수형 데이터로 음이 아닌 어떤 정수값이라도 가질 수 있는 도수형 변수이다. 포아송변량 Y에 대한 기대값을 u라 하고 설명변수 X가 있을 때 포아송 로그선형모형은 $\log u = \alpha + \beta x$와 같다. 이 모형에서 평균은 관계식

$$u = \exp(\alpha+\beta x) = e^{\alpha}(e^{\beta})^{x} \tag{2.8}$$

을 만족한다.

포아송 모형에서의 모수(parameter)들인 β_i의 최대우도추정은 각 표본들이 독립이라는 가정 아래 다음과 같이 주어지는 우도함수

$$L(\beta_0,\beta_1,\cdots,\beta_k)=\prod_{i=1}^{n}\frac{\lambda^{Y_i}e^{-\lambda}}{Y_i!},\quad Y_i=0,1,2\cdots \tag{2.9}$$

의 최대화로부터 구해진다.

각 설명변수들이 반응 변수에 영향을 끼치는 정도는 회귀 계수 β_i를 추정함으로써 알 수 있다. 이 추정량들은 최대우도 추정량(MLE)을 통해 구해지며 근사적으로 정규분포를 따른다. 포아송 모형에서의 모수(parameter)들인 β_i의 유의성 검정은 왈드검정, 우도비검정, 스코어검정 등이 있으며, 가장 단순한 것은 최대우도추정치의 대표본 정규성을 이용한 검정통계량으로 β_i의 추정치를 b_i라고 하고 이의 표준오차를 s_{b_i}라고 하면 b_i/s_{b_i}는 귀무가설을 $H_0:\beta_i=0$으로 했을 때 근사적으로 표준정규분포를 따르게 되고 $\left(\frac{b_i}{s_{b_i}}\right)^2$는 근사적으로 자유도 1을 갖는 카이제곱 분포를 따르게 된다. 귀무가설과 대립가설을 다음과 같이 세웠을 때

$$H_0:\beta_i=0 \qquad \text{vs.} \qquad H_1:\beta_i\neq 0\ ,$$

이를 위한 검정 통계량은 왈드 통계량 $\chi_w^2=\left(\frac{b_i}{s_{b_i}}\right)^2$으로 이 통계량이 자유도 1인 카이제곱 분포의 임계치 $\chi_\alpha(1)$보다 클 때 귀무가설을 기각하게 된다.

모형의 유의성검정은 포아송 랜덤성분을 갖는 일반화선형모형이 성립한다는 귀무가설을 검정하는 것으로 설명변수 N가지 수준 중 i번째에서 관측도

수를 y_i, 기대도수를 $\hat{u_i}$라 하자. 적합도검정(goodness-of-fit test)에 이용되는 피어슨통계량 χ^2과 우도비통계량 G^2는 각각 다음과 같다.

$$\chi^2 = \sum \frac{(y_i - \hat{u_i})^2}{\hat{u_i}}, \qquad G^2 = 2\sum y_i \log\left(\frac{y_i}{\hat{u_i}}\right). \tag{2.10}$$

적합값 $\{\hat{u_i}\}$가 상대적으로 크고(대략 5이상) 수준수 N이 고정되어 있을 때 두 통계량은 근사적으로 카이제곱 분포를 따른다. 모형의 자유도를 잔차 자유도(residual df)라 부르는데, 이 때 df는 반응도수의 총 경우의 수에서 모형에 포함된 모수 개수를 뺀 값이다(정광모 · 최용석, 1999).

4. 영과잉-포아송 회귀분석 (Zero-Inflated Poisson Regression Analysis)

포아송 분포가 적용될 수 있는 실제 문제에서 영이 과다하게 포함되어 있는 경우에 적용되는 분석방법으로 반응값이 영인 부분과 아닌 부분으로 나누어 베르누이 분포와 포아송 분포와의 혼합된 영과잉-포아송 분포를 이용한 회귀분석방법이다. 영과잉-포아송 분포는 확률변수 Y가 일정 단위당 나타나는 계수형 자료(count data)로 영만 나타나는 상태(perfect state)의 확률값이 따로 정해진다. 반응벡터 $Y=(Y_1, Y_2, \cdots, Y_n)'$ 들이 독립이고 그리고

$$Y \sim 0 \quad p \text{ 의 확률로,}$$

$$\sim Poisson(\lambda) \quad 1-p \text{ 의 확률.}$$

여기에서 $0 \leq p < 1$는 영의 값에서 주어지는 임의의 확률이며 $\lambda > 0$는 포아송 분포의 평균이다. 이때 확률질량함수(pmf)는 아래와 같다.

$$P(Y=0) = p + (1-p)\exp(-\lambda)$$

$$P(Y=k) = (1-p)\exp(-\lambda)\lambda^k/k! \ , \quad k=1,2,\cdots.$$

반응벡터 $Y=(Y_1, Y_2, \cdots, Y_n)$가 k개의 설명변수가 있는 경우 모수벡터 $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_n)$ 그리고 영에 대한 확률벡터 $p=(p_1, p_2, \cdots, p_n)$는 다음을 만족하는 로그연결함수 λ와 로짓연결함수(logit link function) p로 모형화된다.

$$\log(\hat{\lambda}) = B\beta,$$

$$\text{logit}(P) = \log(\frac{P}{1-P}) = G\gamma. \tag{2.11}$$

여기에서 B와 G는 공변량들의 모형행렬(model matrix)이다. 위 모형은 많은 모수들을 포함하기 때문에 다음과 같이 몇 가지로 분류하여 생각할 수 있다. 첫째, λ와 p가 함수적인 관계가 없고 $B=G$ 일 때 모형에 포함되는 모수의 수는 포아송 회귀모형보다 두 배 많게 된다. 둘째, p가 공변량에 의존되지 않을 때 G는 원소가 1인 벡터가 되어 포아송 회귀모형보다 모수가

단 한 개 많게 된다. 셋째, λ와 p가 함수적인 관계가 있고 $B=G$ 때는 모수의 수가 많이 감소하게 된다. 일반적으로 영만이 완전하게 나타날(perfect state) 확률 p는 포아송 평균인 λ에 반비례한다. λ가 커질수록 p가 작아지는 함수는 여러 가지로 생각할 수 있으나, 두 모수에 대한 사전정보가 $p_i = 1/(1+\lambda_i^{\tau})$ 같이 알려져 있다면 위의 식은

$$\frac{\log(p/(1-p))}{\log\lambda} = \frac{B\gamma}{B\beta} = -\tau \ , \tag{2.12}$$

가 되므로 $B\gamma = -\tau B\beta$가 된다. 여기에서 τ는 형태모수(shape parameter)로서 이 값이 커지면 p는 기하급수적으로 감소하게 된다. 따라서 위의 영과잉 포아송 회귀모형은 다음과 같이 된다.

$$\log(\lambda) = B\beta,$$
$$\log\left(\frac{p}{1-p}\right) = -\tau B\beta. \tag{2.13}$$

위 모형은 일반화 선형모형을 만들기 위하여 포아송 평균의 로그 연결함수 그리고 베르누이 분포의 성공의 확률에 대한 로짓 연결함수가 이용되었다.

영과잉-포아송 회귀모형에서의 회귀계수의 추정은 λ와 p가 함수적인 관계 여부에 따라 다르다. λ와 p가 함수적인 관계가 없을 때 회귀계수 벡터 β와 γ에 대한 우도함수는 다음과 같이 구할 수 있다.

$$L(\gamma,\beta;y)=\sum_{y_i=0}\log(e^{G_i\gamma}+\exp(-e^{B_i\beta}))+\sum_{y_i>0}(y_iB_i\beta-e^{B_i\beta})-\sum_{i=1}^{n}\log(1+e^{G_i\gamma})-\sum_{y_i>0}\log(y_i!). \quad (2.14)$$

λ와 p가 함수적인 관계가 있을 때 회귀계수 벡터 β와 형태모수 τ에 대한 우도함수는 다음과 같이 구할 수 있다.

$$L(\beta,\tau;y)=\sum_{y_i=0}\log(e^{-\tau B_i\beta}+\exp(-e^{B_i\beta}))+\sum_{y_i>0}(y_iB_i\beta-e^{B_i\beta})-\sum_{i=1}^{n}\log(1+e^{-\tau B_{i\beta}}). \quad (2.15)$$

모형의 유의성검정은 설정된 로지스틱 회귀모형이 완전정보를 가진 모형과 비교해서 유의한 차이가 있는지 여부를 평가하는 방법으로 다음과 같다.

$$2[\,L(y;y)-L(\hat{\beta},\hat{\tau};y)\,]. \quad (2.16)$$

이는 점근적으로 자유도가 k인 카이제곱 분포 χ_k^2의 분포를 따르는 것으로 알려져 있다. 여기에서 k는 미지의 모수 개수이다(Lambert, 1992).

5. 의사결정나무분석(Decision Tree Analysis)

1) 의사결정나무의 개념

의사결정나무는 의사결정규칙(decision rule)을 나무구조로 도표화하여 분류(classification)와 예측(prediction)을 수행하는 분석방법으로 분류 또는 예측의 과정이 나무구조에 의한 추론규칙(induction rule)에 의해서 표현되기 때문에 연구자가 그 과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다. 의사결정나무는 판별분석(discrimination analysis) 또는 회귀분석(regression analysis) 등과 같은 모수적(parametric) 모형을 분석하기 위해서 사전에 이상치(outlier)를 검색하거나 분석에 필요한 변수를 찾아내고 모형에 포함되어야 할 교호효과를 찾아내는 데에 사용될 수도 있으며, 그 자체가 분류 또는 예측 모형으로 사용될 수도 있다.

일반적으로 의사결정나무 분석은 다음과 같은 단계를 거친다.

① 의사결정나무의 형성 : 분석의 목적과 자료구조에 따라서 적절한 분리기준(split criterion)과 정지규칙(stopping rule)을 지정하여 의사결정나무를 얻는다.

② 가지치기 : 분류오류(classification error)를 크게 할 위험(risk)이 높거나 부적절한 추론규칙을 가지고 있는 가지(branch)를 제거한다.

③ 타당성 평가 : 이익도표(gains chart)나 위험도표(risk chart) 또는 검증용 자료(test data)에 의한 교차타당성(cross validation) 등을 이용하여 의사결정나무를 평가한다.

④ 해석 및 예측 : 의사결정나무를 해석하고 분류 및 예측모형을 설정한다. 이상과 같은 과정에서 분리기준, 정지규칙, 평가기준 등을 어떻게

지정하느냐에 따라 서로 다른 의사결정나무가 형성된다. 연구자는 연구의 목적이나 자료구조 등을 파악하여 가장 적절한 의사결정나무를 얻기 위한 과정을 반복적으로 수행해야 한다(강현철・한상태・최종후・김은석・김미경, 2001).

2) 분리기준

분리기준은 하나의 부모마디로부터 자식마디들이 형성될 때, 입력변수(input variable)의 선택과 범주(category)의 병합이 이루어 질 기준을 의미한다. 즉, 어떤 입력변수를 이용하여 어떻게 분리하는 것이 목표변수의 분포를 가장 잘 구별해 주는지를 파악하여 자식마디가 형성되는데, 목표변수의 분포를 구별하는 정도를 순수도(purity) 또는 불순도(impurity)에 의해서 측정하는 것이다. 이 때 순수도란 목표변수의 특정 범주에 개체들이 포함되어 있는 정도를 의미한다.

① 이산형 목표변수에 대한 분리 기준

목표변수(target variable)가 이산형인 경우에는 목표변수의 각 범주에 속하는 빈도(frequency)에 기초하여 분리가 일어나며, 이러한 분리 기준으로 형성된 의사결정나무를 분류나무(classification tree)라고 한다. 분리기준으로는 다음과 같은 것들이 있다.

- 카이제곱 통계량(Chi-Square statistic)

  목표변수와 설명변수의 관측도수로 이루어진 $r \times c$ 분할표로부터 계산되며, 이때 카이제곱 통계량값은 다음과 같다.

$$\chi^2 = \sum_{i,j} \frac{(f_{ij} - E_{ij})^2}{E_{ij}} \ , \tag{2.17}$$

단, $E_{ij} = \frac{f_{i.} - f_{.j}}{f_{..}}$.

분리기준을 카이제곱 통계량으로 한다는 것은 p-값이 가장 작은 설명변수와 그때의 최적분리에 의해서 자식마디가 형성되게 한다는 것을 의미한다.

• 지니 지수(Gini index)

지니 지수는 n개의 원소 중에서 임으로 2개를 추출하였을 때, 2개가 서로 다른 그룹에 속해있을 확률로, 다음과 같이 표현된다.

$$\begin{aligned} G &= \sum_{j=1}^{c} P(j)(1 - P(j)) \\ &= 1 - \sum_{j=1}^{c} P(j)^2 = 1 - \sum_{j=1}^{c} (n_i/n)^2 . \end{aligned} \tag{2.18}$$

즉, 지니 지수는 각 마디에서의 불순도(impurity)를 재는 측도인데 이 지니 지수를 가장 감소시키는 설명변수와 그 변수의 최적분리를 자식마디로 선택한다.

• 엔트로피 지수(Entropy index)

지니 지수와 유사한 분리기준으로 다항분포(multinomial distribution)에서의 우도비 검정통계량(likelihood ratio test statistic)을 사용하는 것과 같다.

$$E = -\sum_{i=1}^{c} P(i) \ln P(i) \ . \tag{2.19}$$

② 연속형 목표변수에 대한 분리 기준

목표변수가 연속형인 경우에는 목표변수의 평균(mean)에 기초하여 분리가 일어나며, 이러한 분리 기준으로 형성된 의사결정나무를 회귀나무(regression tree)라 한다. 분리 기준으로는 다음과 같은 것들이 있다.

• F 통계량

y_{ij}를 i번째 설명변수의 범주에 속하는 j번째 관측개체의 목표변수의 값이라고 하고, $\bar{y}$를 i번째 범주의 평균 $\bar{y}$를 전체평균이라고 할 때 F 통계량은 다음과 같다.

$$F = \frac{\sum_{i=1}^{r} n_i(\bar{y_i} - \bar{y})^2/(r-1)}{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_i})^2/(n-r)}. \qquad (2.20)$$

이 통계량은 자유도$(r-1, n-r)$인 F-분포를 따르며, F통계량이 매우 작다는 것은 설명변수에 따른 목표변수의 평균차이가 유의하지 않다는 것을 의미한다.

• 분산의 감소량(variance reduction)

각 마디의 다양도(diversity)를 재는 측도로 다음과 같은 분산을 고려할 수 있다.

$$V = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2. \qquad (2.21)$$

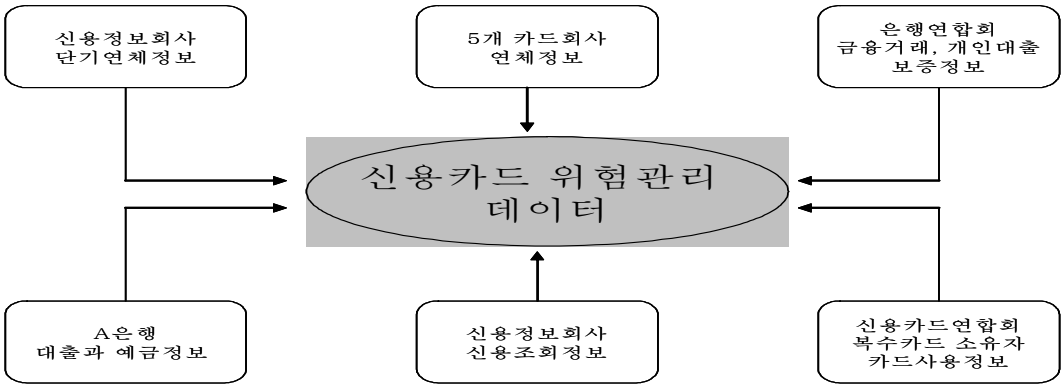
3) 의사결정나무의 유용성

의사결정나무는 나무구조에 의해 모형이 표현되기 때문에 모형의 이해가 쉬우며 두 개 이상의 변수가 결합하여 목표변수에 어떻게 영향을 주는지를 쉽게 알 수 있어 교호효과에 대한 해석 또한 가능하며, 모수적인 방법과는 달리 선형성(linearity)이나 정규성(normality) 또는 등분산성(equal variance) 등의 가정을 필요로 하지 않는 장점이 있다. 뿐만 아니라 순서형 또는 연속형 변수에 대해 단지 순위(rank)만이 분석에 영향을 주기 때문에 이상치(outlier)에도 민감하지 않다. 또한 변수에 대한 결측값도 하나의 범주로 간주하여 이를 모형화 과정에서 처리할 수 있다는 장점들을 가지고 있다.

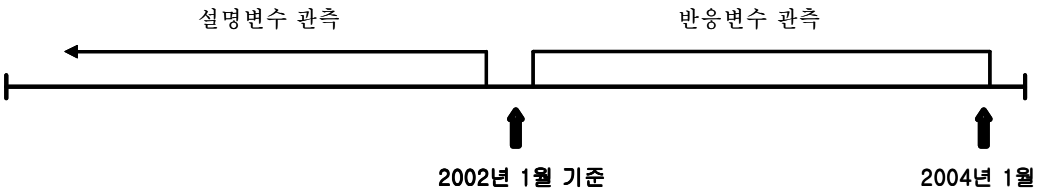
Ⅲ. 사례분석

1. 분석데이터 소개

본 연구를 위해 이용된 데이터는 국내 A은행의 데이터로 2년 동안 고객들의 연체회수(3개월 기준) 및 고객들에 대한 정보이다. 이 데이터는 총 490개의 변수와 53,190명의 관측치로 이루어져 있다.



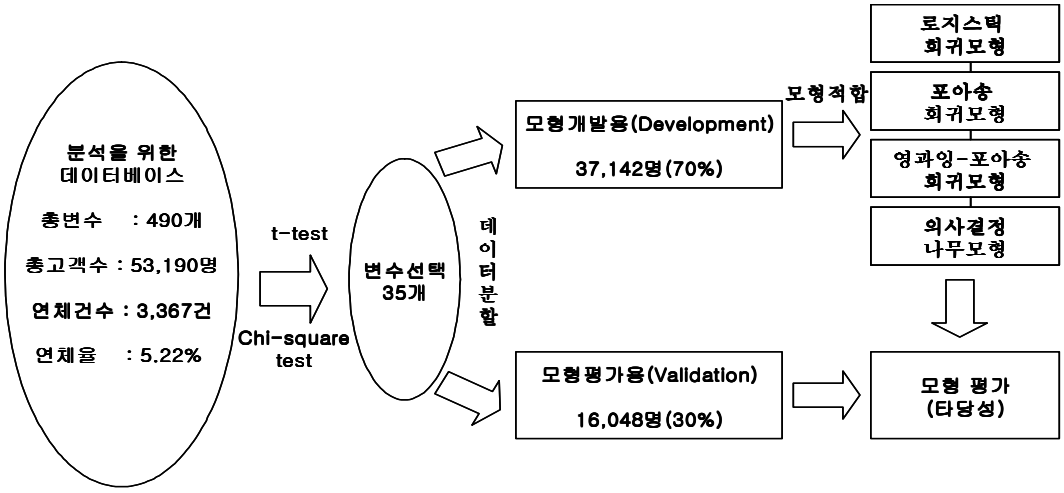
<그림 3-1> 분석자료의 원천 DB



<그림 3-2> 분석데이터의 관측시점 및 기간

2. 분석을 위한 데이터 준비

신용평가모형을 위한 금융데이터 중 모형개발용으로 37,142(70%)개와 모형평가용으로 16,048(30%)개를 이용하였으며, 분석에 이용할 변수의 선택은 변수의 수가 많아 우선 연체여부를 반응변수로 t-test를 실시하여 반응변수에 크게 영향을 주는 35개 변수를 우선 선택한 후 SAS/INSIGHT를 이용하여 설명변수들의 분포와 기술 통계량을 살펴보고 모형적합에 필요한 변수변환을 하였으며, 최종 변수선택은 여러 변수선택방법들을 이용하여 최적모형을 통해 선택되었다. 그 결과 설명변수로는 10개의 변수가 선택되었고 반응변수로는 로지스틱 회귀모형과 의사결정나무모형을 위해서는 연체여부가 포아송 회귀모형과 영과잉-포아송 회귀모형을 위해서는 연체건수가 모형적합 및 예측력의 비교를 위해 선택되었다. 선택된 변수에 대한 내용은 <표 3-1>에 요약정리 하였다.



<그림 3-3> 분석을 위한 데이터 준비과정

<표 3-1> 사례분석에 이용된 변수설명

NO	변수명	평균값	최소값	최대값	변수변환 및 설명
1	신용카드 평균사용개월수 (avg_use_mnthcnt_ba)	46.13	2	90	
2	최근 12개월 신규대출건수 (c_12m_new_loan_cnt_BL2)	0.66	0	5	
3	최근 2년간 발행카드 사용금액 (in_24m_issue_use_a_p_me18)	466850.1	0	3000000	단위 : 십만단위 최대값14(변환)
4	최근 3년간 최장연체건수 (max_delq_days_3yr_cc3)	0.57	0	2	
5	NICE 금융관련 조회건수 (num_fi_inq_12m)	0.91	0	4	NICE에 금융관련 조회건수
6	평균 신용카드 사용금액 (tot_avg_use_a_p_me3_1)	291868.4	0	1417500	단위 : 십만단위 최대값 8(변환)
7	최근 1년간 전체현금서비스비율 (tot_cash_amt_r_p_mx12)	0.42	0	1	현금서비스한도대비
8	최근 6개월 최대사용금액비율 (tot_use_vs_max_amt_r_p_me6)	0.36	0	0.6	전체사용금액대비
9	상위직업군 (duty_ind)	0.57	0	1	관리/연구직=1, other=0
10	주거형태 (house_type_ind)	0.61	0	1	아파트거주=1, other=0
11	연체여부 (addbad30)	0.05	0	1	
12	연체건수 (addbad30_cnt)	0.07	0	5	

3. 로지스틱 회귀분석(Logistic Regression Analysis) 결과

연체여부를 반응변수로 모형의 예측력을 가장 좋게 하는 모형을 단계식(stepwise) 변수선택방법에 의해 다음과 같이 모형을 적합 시켰다.

1) 회귀계수의 추정

<표 3-2> 로지스틱 회귀 모수 추정값

Parameter	DF	Estimate	Standad Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.9730	0.1137	683.0949	<.0001
신용카드 평균사용개월수	1	-0.0275	0.0015	317.8883	<.0001
최근12개월 신규대출건수	1	0.1421	0.0250	32.2129	<.0001
최근 3년간 최장연체건수	1	0.4888	0.0316	239.6888	<.0001
NICE 금융관련 조회건수	1	0.1618	0.0219	54.6567	<.0001
평균 신용카드 사용금액	1	0.1869	0.0104	323.8262	<.0001
1년간 전체현금서비스비율	1	1.2873	0.0923	194.4447	<.0001
최근6개월 최대사용금액비율	1	-1.9071	0.1670	130.4144	<.0001
상위직업군	1	-0.3953	0.0516	58.6352	<.0001
주거형태	1	-0.4500	0.0522	74.2982	<.0001

<표 3-3> 호스머-렘쇼(Hosmer-Lemeshow) 검정

Group	Total	연체여부=1		연체여부=0	
		Observed	Expected	Observed	Expected
1	3715	11	8.30	3704	3706.70
2	3714	22	15.76	3692	3698.24
3	3716	17	23.98	3699	3692.02
4	3714	25	34.56	3689	3679.44
5	3714	47	49.94	3667	3664.06
6	3714	76	74.75	3638	3639.25
7	3714	103	118.96	3611	3595.04
8	3714	213	202.06	3501	3511.94
9	3714	380	383.74	3334	3330.26
10	3713	1025	1006.94	2688	2706.06

Chi-Square	DF	Pr > ChiSq
11.5987	8	0.1700

로지스틱 회귀분석 결과를 살펴보면 회귀계수 모두 유의수준 0.01에서 매우 유의하며 모형의 적합력은 호스머-렘쇼(Hosmer-Lemeshow) 검정결과 카이제곱통계량이 11.6이고 p-value는 0.17로 유의수준 0.05에서 모형이 적합하다는 귀무가설을 기각하지 못하므로 모형이 잘 적합함을 알 수 있으며, 추정된 회귀계수를 살펴보면 신용카드 평균사용개월수와 최근 6개월 최대 사용금액비율은 값이 클수록, 직업은 상위직업군이고, 거주 형태는 아파트인 경우가

그렇지 않은 경우보다 연체할 확률을 낮게 하는 변수이다. 반면에 최장연체건수와 신규대출건수, 금융과 관련된 신용조회건수, 1년 동안 전체 현금 서비스비율 등과 같은 대출과 관련된 변수는 변수의 값이 클수록 연체할 확률을 높이는 변수임을 알 수 있다.

2) 상관분석

<표 3-4> 상관분석(CORRELATION ANALYSIS)

	신용카드 평균사용개월수	최근12개월 신규대출건수	최근3년간 최장연체건수	NICE 금융관련 조회건수	평균 신용카드 사용금액	최근1년간 전체 현금서비스비율	최근6개월 최대 사용금액비율	상위직업군	주거형태
신용카드 평균사용개월수	1	-0.021	-0.085	-0.001	0.036	-0.060	0.073	0.025	0.190
최근12개월 신규대출건수	-0.021	1	0.158	0.460	0.245	0.226	-0.138	-0.038	0.033
최근3년간 최장연체건수	-0.085	0.158	1	0.139	0.178	0.294	-0.101	-0.046	-0.042
NICE 금융관련 조회건수	-0.001	0.460	0.139	1	0.213	0.203	-0.105	0.010	0.055
평균 신용카드 사용금액	0.036	0.245	0.178	0.213	1	0.524	-0.251	-0.089	-0.013
최근1년간 전체 현금서비스비율	-0.060	0.226	0.294	0.203	0.524	1	-0.148	-0.064	-0.106
최근6개월 최대 사용금액비율	0.073	-0.138	-0.101	-0.105	-0.251	-0.148	1	0.067	0.030
상위직업군	0.025	-0.038	-0.046	0.010	-0.089	-0.064	0.067	1	-0.043
주거형태	0.190	0.033	-0.042	0.055	-0.013	-0.106	0.030	-0.043	1

설명변수들 간의 연관성은 설명변수들의 회귀계수 추정에 있어 문제가 되므로 회귀분석에 앞서 설명변수들의 상호연관성을 살펴보고 상관이 높은 변수들은 제거하고 회귀분석을 실시하여야 한다. 하지만 실제 데이터에서 변수들의 상관관계가 없는 변수들을 선택하여 모형을 적합한다는 것은 매우 어려운 일이다. 그러므로 적절한 연관정도와 모형의 적합력 및 예측력을 고려하여 분석하는 것이 바람직 할 것이다. 위 결과를 살펴보면 설명변수들

간의 상관계수는 0.5보다 낮은 값들을 보이고 있어 모형적합에 문제는 없을 것으로 보인다. 다음은 다중공선성 진단을 위한 PROC REG의 옵션에서 지원되는 COLLIN의 결과이다.

3) 다중공선성 검토

<표 3-5> 상태지수(Condition Index)

No.	Eigen_ value	Condition Index	Intercept	신용카드평균 사용개월수	최근1년신규 대출건수	최근3년최장 연체건수	NICE 금융관련 조회건수	평균 신용카드 사용금액	1년간 전체 현금서비스 비율	최근6개월 최대사용 금액비율	상위직업군	주거형태
1	6.42	1.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01
2	0.94	2.61	0.00	0.01	0.10	0.04	0.05	0.04	0.04	0.02	0.06	0.03
3	0.66	3.12	0.00	0.00	0.19	0.23	0.18	0.04	0.07	0.00	0.00	0.01
4	0.52	3.51	0.00	0.00	0.01	0.62	0.01	0.20	0.07	0.00	0.00	0.00
5	0.44	3.84	0.00	0.01	0.00	0.01	0.01	0.00	0.01	0.00	0.57	0.31
6	0.34	4.37	0.00	0.00	0.68	0.00	0.74	0.00	0.00	0.00	0.00	0.00
7	0.28	4.81	0.01	0.02	0.00	0.03	0.00	0.18	0.11	0.15	0.27	0.31
8	0.23	5.28	0.00	0.05	0.00	0.03	0.00	0.39	0.65	0.02	0.03	0.26
9	0.14	6.84	0.00	0.63	0.00	0.00	0.00	0.11	0.02	0.40	0.00	0.04
10	0.05	11.56	0.98	0.28	0.01	0.03	0.00	0.03	0.02	0.40	0.06	0.03

<표 3-5>는 선택된 변수들의 다중공선성 정도를 나타내는 상태지수표이다. 설정된 회귀모형이 다중공선성이 존재하는지는 상태지수를 통해 그 상태지수가 일반적으로 30보다 크고 그 큰 값에 대응하는 회귀계수들에 대한 분산의 비율이 매우 크면 회귀계수를 추정하는 정확도에 영향을 주는 다중공선성이 존재한다고 한다. 위의 표에 나타난 상태지수는 최대값이 11.56으로 설명변수 상호 간의 다중공선성은 존재하지 않는다고 할 수 있어 로지스틱 회귀모형은 잘 적합 되었다고 할 수 있다.

4. 포아송 회귀분석(Poisson Regression Analysis) 결과

연체회수를 반응변수로 로지스틱 회귀모형에서 선택한 9개의 변수를 설명변수로 다음과 같이 모형을 적합 시켰다.

<표 3-6> Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	37E3	10301.1910	0.2774
Scaled Deviance	37E3	10301.1910	0.2774
Pearson Chi-square	37E3	40870.8261	1.1007
Scaled Pearson ×2	37E3	40870.8261	1.1007
Log Likelihood		-6782.3615	

<표 3-7> 포아송 회귀 모수 추정값

Parameter	DF	Estimate	Standad Error	Wald95% Confi_dence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-3.1481	0.0975	-3.3392	-2.9571	1042.9	<.0001
신용카드 평균사용개월수	1	-0.0211	0.0012	-0.0235	-0.0187	287.38	<.0001
최근12개월 신규대출건수	1	0.1172	0.0190	0.0799	0.1545	37.95	<.0001
최근 3년간 최장연체건수	1	0.5613	0.0258	0.5107	0.6120	472.52	<.0001
NICE 금융관련 조회건수	1	0.1253	0.0173	0.0915	0.1592	52.67	<.0001
평균 신용카드 사용금액	1	0.1355	0.0086	0.1187	0.1523	250.52	<.0001
최근1년간 전체현금서비스비율	1	1.3631	0.0806	1.2051	1.5210	286.04	<.0001
최근6개월 최대사용금액비율	1	-1.6672	0.1382	-1.9381	-1.3964	145.55	<.0001
상위직업군	1	-0.3508	0.0416	-0.4324	-0.2692	70.98	<.0001
주거형태	1	-0.3519	0.0420	-0.4342	-0.2696	70.22	<.0001

포아송 회귀분석 결과를 살펴보면 회귀계수의 추정값이 로지스틱 회귀모형에서 추정된 계수와 별 차이를 보이고 있지 않다. 이는 연체건수의 비율이 5%로 매우 작은 것과 연체건수가 0에서 5까지의 값으로 2번 이상의 연체 비율이 각각 0.6% 0.2% 0.06% 0.03% 로 비율이 매우 낮기 때문인 것으로 생각된다.

5. 영과잉-포아송 회귀분석 (Zero-Inflated Poisson Regression Analysis) 결과

반응변수가 연체건수와 같은 계수형 데이터는 일반적으로 포아송 분포를 따른다고 알려져 왔다. 하지만 정상적인 포아송 확률분포보다 영의 값이 과잉관측되는 경우 포아송 분포가 적합하지 않음은 기존의 여러 논문들에 의해 소개되어져 왔다. 사례분석을 위한 금융데이터도 반응변수인 연체건수의 영의 비율이 약 95%로 영이 과잉 관측되어진 경우에 해당되어 포아송 회귀모형을 적합할 수 없다. 따라서 분포의 적합성 검정을 카이제곱 검정을 통해 보이고 영과잉-포아송 분포인 경우 영과잉-포아송 분포를 적용한 영과잉-포아송 회귀분석의 결과를 함께 보이려 한다.

• 포아송분포 검정

연체건수의 분포가 포아송 분포를 따르는 지에 대한 검정방법으로는 스코어 검정, 우도비 검정, 카이제곱 검정 등이 있으며 여기서는 카이제곱 검정으로 분포의 적합성을 검정하여 보겠다.

<표 3-8> 연체건수 빈도표

연체건수	백분율	관측빈도	포아송기대빈도
0	94.78	50412	49823
1	4.34	2311	3258
2	0.57	301	107
3	0.22	117	2
4	0.06	33	0
5	0.03	16	0
합계	100.00	53190	53190

$$\chi^2 = \sum_{0}^{5} \frac{(\text{관측빈도} - \text{기대빈도})^2}{\text{기대빈도}} = 8526$$

카이제곱 통계량이 8,526으로 매우 크므로 유의수준 5%로 했을 때 자유도 5인 카이제곱의 임계치는 11.07이므로 포아송 분포를 따른다는 귀무가설을 기각시켜 반응변수가 포아송 분포를 따른다고 할 수 없으므로 포아송 회귀모형에 적합하기 어렵다. 따라서 이와 같이 반응변수의 관측 값에 0이 많이 나타나는 경우에 적합한 분포는 영과잉-포아송 분포이므로 다른 모형과 동일한 설명변수들로 영과잉-포아송 회귀모형을 적합하여 보았다. 그 결과는 다음과 같다.

<표 3-9> 적합 통계량(Fit Statistics)

-2 Log Likelihood	13790
AIC (smaller is better)	13826
AICC (smaller is better)	13826
BIC (smaller is better)	13979

<표 3-10> 영과잉-포아송 회귀 모수 추정값

Parameter ($\hat{p}$)	Estimate	Standad Error	t Value	Pr>|t|
Intercept	1.4646	0.2732	5.36	<.0001
신용카드 평균사용개월수	0.02863	0.0021	13.59	<.0001
최근12개월신규대출건수	-0.1912	0.0368	-5.20	<.0001
최근 3년간 최장연체건수	-0.2830	0.0731	-3.87	0.0001
NICE 금융관련 조회건수	-0.1317	0.0415	-3.17	0.0015
평균 신용카드 사용금액	-0.2947	0.0255	-11.56	<.0001
최근1년간 전체현금서비스비율	-0.5360	0.2350	-2.28	0.0226
최근6개월 최대사용금액비율	0.8385	0.3897	2.15	0.0314
상위직업군	0.2460	0.1133	2.17	0.0299
주거형태	0.6237	0.0721	8.65	<.0001

Parameter($\hat{\lambda}$)	Estimate	Standad Error	t Value	Pr>|t|
Intercept	-1.5194	0.2184	-6.96	<.0001
최근2년간 발행카드사용금액	0.0372	0.0056	6.60	<.0001
최근3년간 최장연체건수	0.3779	0.0484	7.80	<.0001
NICE 금융관련 조회건수	0.0529	0.0244	2.16	0.0304
평균 신용카드 사용금액	-0.0700	0.0174	-4.02	<.0001
최근1년간 전체현금서비스비율	0.7656	0.1933	3.96	<.0001
최근6개월 최대사용금액비율	-0.9782	0.2751	-3.56	0.0004
상위직업군	-0.2249	0.0750	-3.00	0.0027

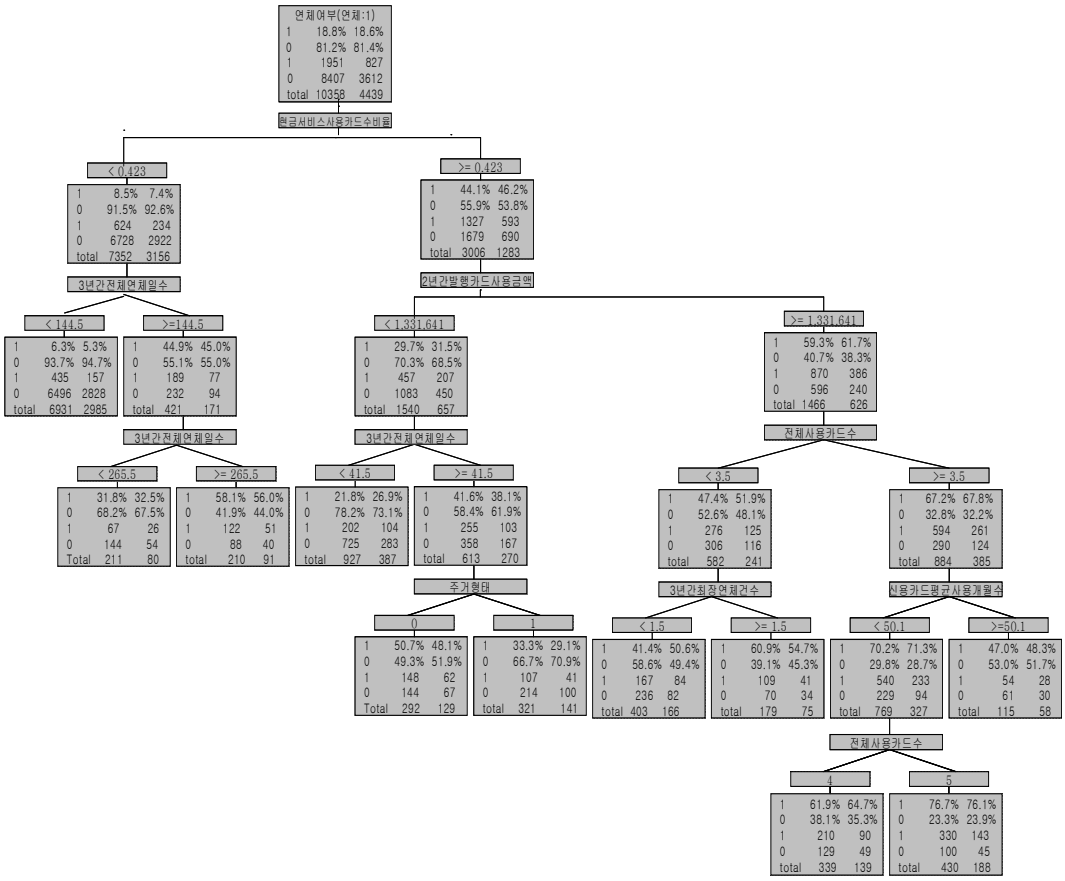
위의 결과표를 보면 다른 회귀모형에서 추정되는 모수들 보다 많음을 알 수 있다. 이는 연체가 없는 경우를 완전한 상태(perfect state)로 연체가 있는 경우를 불완전한 상태(imperfect state)로 나누어 각 각 p, λ를 추정한 결과이기 때문이다. 16개의 회귀계수는 연체가 없는 완전한 상태(perfect state)인 0의 확률 p와 연체 회수가 있는 불완전한 상태(imperfect state)인 λ의 회귀계수 추정값이다. 완전한 상태인 0의 확률(p)을 추정할 설명변수는 다른 모형에서 이용한 동일한 설명변수들로 계수의 부호만 바뀌었을 뿐 회귀계수 모두 유의확률 0.05에서 유의하게 나타났다.

불완전한 상태에서의 연체회수에 대한 평균(λ)을 추정할 설명변수들 역시 동일한 설명변수들로 그 중 추정된 계수의 값이 유의하지 않은 신용카드 평균 사용 개월수와 최근 12개월의 신규대출건수, 주거형태가 빠지고 최근 2년 안에 발행된 카드 사용금액의 변수가 새로이 설명변수로 추가되어 모형을 적합하였다. 그 결과 평균 신용카드 사용금액의 회귀계수의 부호가 이전의 로지스틱 회귀모형 및 포아송 회귀모형에서는 다르게 나타났고 다른 설명변수들은 부호는 같으나 회귀계수의 크기에서만 차이를 보이고 있다.

모형에 따라 모형의 적합력이나 예측력을 높일 수 있는 최적의 변수를 선택하는 과정은 매우 어려운 과정이다. 이러한 변수선택 과정은 통계적인 방법에만 의존할 것이 아니라 실제 금융기관의 실무담당자들이 경험상 유의하게 영향을 주는 변수들에 대해서도 모형의 적합 시 함께 고려해 모형을 구축한다면 적합력과 예측력을 높일 수 있는 최적의 모형을 구축할 수 있을 것이다. 최적 모형 구축을 위한 변수의 선택에 대한 다양한 방법들에 대해 향후 많은 연구가 이루어져야 하겠다.

6. 의사결정나무분석(Decision Tree Analysis) 결과

연체여부를 목표변수(target)로 입력변수(input)들은 다른 모형에서 이용한 35개의 변수로 분석을 하였으나 분리가 일어나지 않았다. 이는 연체율이 5.2%로 낮아 분리가 일어나지 않은 것으로 생각되었으며, 표본추출(over sampling)로 연체율(18.8%)을 높여 재분석을 시도하였다. 그 결과는 다음과 같다.



<그림 3-4> 의사결정나무분석에 의한 나무구조모형

의사결정나무분석의 결과 가지의 깊이(depth)는 5, 끝마디(leaf)는 11인 나무구조를 보이고 있다. 첫 번째 분류는 현금서비스 사용 카드수의 비율이 0.423보다 작고 3년간 총 연체일수가 265.5보다 많으면 연체확률이 58.1%로 높아진다. 두 번째 분류는 현금서비스 사용 카드수의 비율이 0.423보다 크고, 최근 2년간 발행카드의 사용금액이 1,331,641보다 작고, 3년간 총 연체일수가 41.5보다 크고, 아파트에 거주하지 않으면 연체확률이 50.7%로 높아진다. 세 번째 분류는 현금서비스 사용 카드수의 비율이 0.423보다 크고, 최근 2년간 발행카드의 사용금액이 1,331,641보다 많고, 신용카드 평균 사용 개월수가 50.1보다 작고 전체카드 사용수가 5이면 연체확률이 76.7%로 높아지는 것을 볼 수 있다. 이와 같은 분류결과는 모형의 예측력이 다른 모형에 비해 떨어지는 것을 알 수 있다.

7. 모형비교 및 검토

신청평점시스템을 위한 모형구축에 있어 반응변수에 따라 로지스틱 회귀분석, 포아송 회귀분석, 영과잉-포아송 회귀분석, 의사결정나무분석을 이용하여 모형을 적합하여 보았다. 이 중 로지스틱 회귀모형, 포아송 회귀모형, 영과잉-포아송 회귀모형의 예측력에 대한 모형비교를 위해 리프트 도표(Lift Chart)를 그려보았다. 또한 세 모형 비교에 앞서 개발모형에 대한 타당성평가를 위한 리프트 테이블(Lift Table)과 리프트 도표(Lift Chart)가 다음과 같이 제시되었다.

1) 타당성평가를 위한 리프트 테이블 및 도표

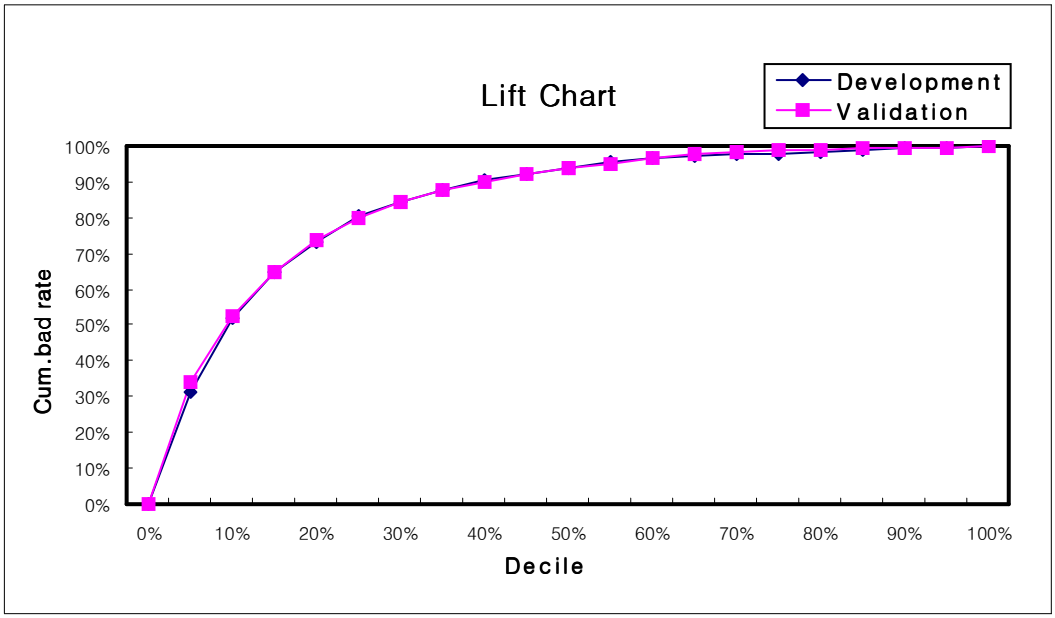
<표 3-11> 리프트 테이블(Lift Table)

(영과잉-포아송 회귀모형)

Development					Validation				
Decile	#	Actual bad	Simple bad rate	Cum.bad rate	Decile	#	Actual bad	Simple bad rate	Cum.bad rate
5%	1857	601	31.32%	31.32%	5%	802	294	34.23%	34.23%
10%	1857	398	20.74%	52.06%	10%	803	157	18.28%	52.50%
15%	1857	240	12.51%	64.56%	15%	802	106	12.34%	64.84%
20%	1857	164	8.55%	73.11%	20%	803	75	8.73%	73.57%
25%	1857	140	7.30%	80.41%	25%	802	55	6.40%	79.98%
30%	1857	76	3.96%	84.37%	30%	803	38	4.42%	84.40%
35%	1857	65	3.39%	87.75%	35%	802	27	3.14%	87.54%
40%	1857	52	2.71%	90.46%	40%	803	19	2.21%	89.76%
45%	1857	33	1.72%	92.18%	45%	802	22	2.56%	92.32%
50%	1857	37	1.93%	94.11%	50%	803	15	1.75%	94.06%
55%	1857	24	1.25%	95.36%	55%	802	8	0.93%	94.99%
60%	1857	20	1.04%	96.40%	60%	803	13	1.51%	96.51%
65%	1857	12	0.63%	97.03%	65%	802	9	1.05%	97.56%
70%	1857	12	0.63%	97.66%	70%	803	7	0.81%	98.37%
75%	1857	5	0.26%	97.92%	75%	802	5	0.58%	98.95%
80%	1857	10	0.52%	98.44%	80%	803	1	0.12%	99.07%
85%	1857	8	0.42%	98.85%	85%	802	5	0.58%	99.65%
90%	1857	11	0.57%	99.43%	90%	802	0	0.00%	99.65%
95%	1857	5	0.26%	99.69%	95%	802	0	0.00%	99.65%
100%	1859	6	0.31%	100.0%	100%	802	3	0.35%	100.0%
합계	37142	1919	100%		합계	16048	859	100%	

- Actual bad(실제연체) : 20등분의 각 등급에서 연체여부변수의 연체빈도
- Simple bad rate(연체비율) : 해당등급에서 연체빈도 / 전체연체빈도
- Cum. bad rate(누적연체비율) : 연체빈도를 누적해 가며 구한 연체비율

<표 3-11>은 모형에 의한 예측확률을 확률의 값이 높은 순서에 따라 데이터 세트를 정렬(sort)한 후 전체 데이터 세트를 균일하게 20등분한다. 20등분한 각 등급에서 실제연체(actual bad)의 빈도를 구한다. 빈도를 구한 다음 해당등급에서 실제연체빈도를 전체연체빈도로 나누어 각 등급에서의 연체비율(simple bad rate)과 누적연체비율(cum. bad rate)의 값을 구한다. 구해진 값으로 <그림 3-5>의 리프트 도표(Lift Chart)를 그려 타당성 여부를 시각적으로 판단해 볼 수 있다.



<그림 3-5> 리프트 도표(Lift Chart)

<그림 3-5>를 살펴보면 개발모형(Development)과 평가모형(Validation)의 그래프가 거의 일치하는 것을 볼 수 있다. 이는 설정된 모형의 타당성이 있음을 알 수 있으며, 모형의 예측력도 상위 30%에서 실제연체의 약 84%를 예측할 수 있어 모형이 잘 적합함을 알 수 있다.

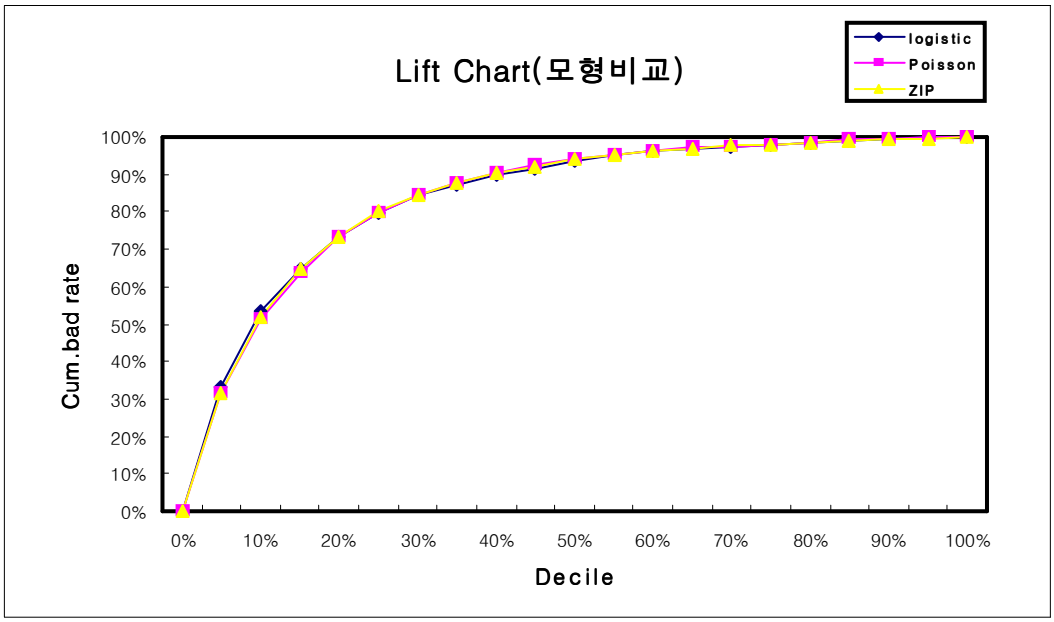
2) 모형비교를 위한 리프트 테이블 및 도표

로지스틱 회귀모형, 포아송 회귀모형, 영과잉-포아송 회귀모형의 예측력 비교를 위한 리프트 테이블(Lift Table)과 리프트 도표(Lift Chart) 이다.

<표 3-12> 리프트 테이블(Lift Table)

Logistic					Poisson					ZIP				
Decile	#	Actual bad	Simple bad rate	Cum.bad rate	Decile	#	Actual bad	Simple bad rate	Cum.bad rate	Decile	#	Actual bad	Simple bad rate	Cum.bad rate
5%	1857	638	33.25%	33.25%	5%	1857	606	31.58%	31.58%	5%	1857	601	31.32%	31.32%
10%	1857	387	20.17%	53.41%	10%	1857	383	19.96%	51.54%	10%	1857	398	20.74%	52.06%
15%	1857	217	11.31%	64.72%	15%	1857	237	12.35%	63.89%	15%	1857	240	12.51%	64.56%
20%	1857	163	8.49%	73.22%	20%	1857	179	9.33%	73.22%	20%	1857	164	8.55%	73.11%
25%	1857	121	6.31%	79.52%	25%	1857	125	6.51%	79.73%	25%	1857	140	7.30%	80.41%
30%	1857	92	4.79%	84.31%	30%	1857	93	4.85%	84.58%	30%	1857	76	3.96%	84.37%
35%	1857	59	3.07%	87.39%	35%	1857	62	3.23%	87.81%	35%	1857	65	3.39%	87.75%
40%	1857	44	2.29%	89.68%	40%	1857	52	2.71%	90.52%	40%	1857	52	2.71%	90.46%
45%	1857	37	1.93%	91.61%	45%	1857	34	1.77%	92.29%	45%	1857	33	1.72%	92.18%
50%	1857	39	2.03%	93.64%	50%	1857	31	1.62%	93.90%	50%	1857	37	1.93%	94.11%
55%	1857	26	1.35%	95.00%	55%	1857	22	1.15%	95.05%	55%	1857	24	1.25%	95.36%
60%	1857	21	1.09%	96.09%	60%	1857	28	1.46%	96.51%	60%	1857	20	1.04%	96.40%
65%	1857	12	0.63%	96.72%	65%	1857	11	0.57%	97.08%	65%	1857	12	0.63%	97.03%
70%	1857	13	0.68%	97.39%	70%	1857	8	0.42%	97.50%	70%	1857	12	0.63%	97.66%
75%	1857	8	0.42%	97.81%	75%	1857	10	0.52%	98.02%	75%	1857	5	0.26%	97.92%
80%	1857	9	0.47%	98.28%	80%	1857	10	0.52%	98.54%	80%	1857	10	0.52%	98.44%
85%	1857	11	0.57%	98.85%	85%	1857	13	0.68%	99.22%	85%	1857	8	0.42%	98.85%
90%	1857	11	0.57%	99.43%	90%	1857	4	0.21%	99.43%	90%	1857	11	0.57%	99.43%
95%	1857	6	0.31%	99.74%	95%	1857	7	0.36%	99.79%	95%	1857	5	0.26%	99.69%
100%	1859	5	0.26%	100.0%	100%	1859	4	0.21%	100.0%	100%	1859	6	0.31%	100.0%
합계	37142	1919	100.0%		합계	37142	1919	100.0%		합계	37142	1919	100.0%	

- Actual bad(실제연체) : 20등분의 각 등급에서 연체여부변수의 연체빈도
- Simple bad rate(연체비율) : 해당등급에서 연체빈도 / 전체연체빈도
- Cum. bad rate(누적연체비율) : 연체빈도를 누적해 가며 구한 연체비율



<그림 3-6> 리프트 도표(Lift Chart)

<그림 3-6>은 세 모형에 대한 리프트 도표(Lift Chart)의 결과이다. 세 모형의 그래프선이 거의 일치하고 있어 모형의 예측력은 차이가 없는 것을 알 수 있다.

3) 모형의 비교 및 검토

금융데이터의 연체건수의 값들이 0~5로 각각 94.78% 4.34% 0.6% 0.2% 0.06% 0.03%로 각 연체건수의 비율이 너무 적어 연체 유무인 이진변수와 별 차이를 나타내지 못해 연체확률을 모형화한 로지스틱 회귀모형과 평균연체건수를 모형화한 영과잉-포아송 회귀모형과의 예측력의 차이를 보이지 못한 결과로 나타난 것으로 생각할 수 있겠다. 하지만 영과잉-포아송 회귀모형에서

는 로지스틱 회귀모형에서 연체확률만을 예측하는 것과는 다르게 연체확률뿐만 아니라 연체고객일 경우 몇 번 연체할 것인가에 대한 예측값을 동시에 구할 수 있다는 장점이 있다.

이와 같이 영의 값만이 나타나는 완전한 상태(perfect state)의 확률값(p)와 불완전한 상태(imperfect state)를 나타내는 연체건수에 대한 평균(λ)의 추정값 그리고 영과잉-포아송 회귀모형의 추정치인[$(1-p)\lambda$]를 결과로 주어지기 때문에 각각의 추정값들을 이용하여 데이터집합을 세분화할 수 있다. 예를 들면 p의 값이 높고 λ가 낮은 그룹과 p의 값이 낮고 λ가 높은 그룹 또는 p와 λ의 값이 동시에 높거나 낮은 그룹 등으로 나누어 그룹별 모형 해석을 할 수 있다.

이러한 그룹의 세분화는 신용카드 신청 시 인수거절 뿐만 아니라 인수 고객의 이자율의 차등적용 등을 통해 신용도가 높은 고객들을 다른 경쟁사보다 낮은 이자율 등을 제시하여 고객의 카드사용을 권장시키거나 잠재이탈을 사전에 방지할 수도 있다. 또한 연체확률은 매우 낮으나 연체건수가 높은 그룹은 그룹 내에서의 다시 세분화를 통한 신청거절이나 위험을 감수할 수 있는 적정범위에서 높은 이자율을 적용시켜 신청수락을 결정할 수도 있을 것이다. 이와 같이 영과잉-포아송 회귀모형을 위한 다양한 그룹화 방법들의 적용가능성은 금융기관의 경쟁력을 높일 수 있을 뿐만 아니라 이익을 최대화 시킬 수 있는 장점이라고 할 수 있다.

Ⅳ. 결 론

개인 신용평가제도란 개인의 신용상태에 영향을 미칠 수 있는 요소들을 통계적인 기법을 이용하여 개인의 신용점수와 연체할 확률 등을 계산해 내는 것을 말한다. 이와 같은 통계적 기법에 의한 시스템을 이용할 경우 금융기관들은 보다 객관적이고 합리적인 기준으로 우수한 양질의 회원을 확보할 수 있을 것이다. 따라서 본 논문에서는 좀 더 예측력이 우수하고, 정교한 신용평가모형을 위해 본 논문에서 제안한 영과잉-포아송 회귀모형과 기존에 이용하고 있는 통계적 신용평가 모형들을 적합하였다. 그리고 적합된 모형들의 예측력과 적용 및 활용가능성을 비교 검토하였다.

신용평가모형들을 적합하여 본 결과 의사결정나무분석은 연체율이 낮은 데이터에 잘 적합하지 않아 분리가 일어나지 않았으며, 반응변수가 연체여부인 로지스틱 회귀모형과 반응변수가 연체회수인 포아송 회귀모형, 영과잉-포아송 회귀모형의 예측력은 비슷한 결과를 보였다. 이는 연체회수의 값이 2번 이상 연체한 비율이 0.9%로 매우 낮아 연체여부인 반응변수와의 차이가 없어 연체확률을 모형화한 로지스틱 회귀모형과 평균 연체건수를 모형화한 영과잉-포아송 회귀모형이 비슷한 예측력을 보이게 된 것으로 판단된다. 그러나 영과잉-포아송 회귀모형에서는 로지스틱 회귀모형에서 연체확률만을 예측하는 것과는 다르게 연체확률 뿐만 아니라 연체고객일 경우 몇 번 연체할 것인가에 대한 예측값을 동시에 계산할 수 있다는 장점이 있다. 이는 금융기간의 수익을 증대시킬 수 있는 다양한 방법들을 모색할 수 있어 모형의 평가에 있어 다른 모형에 비해 우수하다고 할 수 있겠다.

끝으로 본 연구의 한계점으로는 모형개발에 이용된 데이터의 변수가 많아 각 변수의 특성을 파악하기 어려웠고, 신용카드발급 업무에 있어 경험적으

로 신용평가를 위한 중요한 변수들에 대해 사전에 고려하지 못한 점이다. 또한 최적모형을 위한 변수 선택 시 로지스틱 회귀모형에서 선택된 변수들을 이용하여 서로 다른 모형에 모형을 적합하였다는 문제점을 안고 있다. 이는 최적모형을 위한 적절한 방법이라 할 수 없으며, 향후 이러한 문제를 해결하기 위한 통계적인 방법들에 대해 연구가 이루어져야 하겠다.

참 고 문 헌

[1] 강현철・한상태・최종후・김은석・김미경 (2001). 『*SAS Enterprise Miner 4.0을 이용한 데이터마이닝 -방법론 및 활용-*』, 서울, 자유아카데미.

[2] 김학신 (2003). 개인신용위험관리에 대한 실증적 연구. 석사학위논문. 중앙대학교 대학원.

[3] 성웅현 (2001). 『*응용 로지스틱 회귀분석*』, 서울, 탐진.

[4] 이희만・박관수 (1999). 『*신용스코어링 소개*』, 서울, 피리.

[5] 정광모・최용석 (1999). 『*범주형 자료분석*』, 서울, 자유아카데미.

[6] Heilborn, D.C. (1989). Generalized Linear Models for Altered Zero Probabilities and Overdispersion in Count Data, *unpublished technical report*, university of California, San Francisco, Dept. of Epidemiology and Biostatistics.

[7] Lambert, Diane. (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, *Technometrics*, 34, pp.1-14.

[8] Singh, S.N. (1963). A Note on Inflated Poisson Distribution, *Journal of the Indian Statistical Association*, 1, pp.140-144.

ABSTRACT

A Study on the Credit Scoring Model with Zero-Inflated Poisson

Kim, Joo Young
Department of Statistics
The Graduate School
Sungshin Women's University

Zero-inflated Poisson distribution is more applicable distribution in case of excess zeros than regular Poisson probability distribution as a discrete type probability distribution.

The data of that has excess zeros come out in a variety field in company with development of modern technology. The finance, one of those field, produce results that excess zeros of response value with remarkable decrease of arrear rates from systematic credit risk management methods.

Therefore existing Poisson regression is not a suitable application for financial field, since they have too may zeros.

This thesis is a comparative study that compare with existing methods, such as logistic model, Poisson model, decision tree model and zero-inflated Poisson model using the data of Korean domestic A bank.

부록 : 모형개발프로그램

1. 데이터 분할 및 변수선택

```
/*DATA SET 설명
temp_thin_clean : 원데이터 (53190개)
credit_rm :  Development(70%)  Validation (30%) 데이터분할 및 변수변환
credit_rm_1 :  T-TEST와 CHISQ검정  결과 유의한 변수(37개)*/


/*모형개발 및 평가*/
/*Development(약70%) : 37142개  Validation(약30%):16048개 데이터분할 및 변수변환 */
DATA credit_rm;
   SET temp_thin_clean;
       tot_avg_use_a_p_me3_1 =  tot_avg_use_a_p_me3/100000;
       in_24m_issue_use_a_p_me18_1=in_24m_issue_use_a_p_me18/100000;
       sum_delq_amt_3yr_1=sum_delq_amt_3yr/1000000;
       whl_tot_use_a_p_me3_1=whl_tot_use_a_p_me3/1000000;
       IF in_24m_issue_use_a_p_me18_1=>14 THEN in_24m_issue_use_a_p_me18_1=14;
          ELSE  in_24m_issue_use_a_p_me18_1=in_24m_issue_use_a_p_me18_1;
       IF sum_delq_amt_3yr =>4 THEN sum_delq_amt_3yr_1=4;
            ELSE sum_delq_amt_3yr_1=sum_delq_amt_3yr;
       IF tot_avg_use_a_p_me3_1 =>8 THEN tot_avg_use_a_p_me3_1=8;
            ELSE  tot_avg_use_a_p_me3_1 = tot_avg_use_a_p_me3_1;
       IF whl_tot_use_a_p_me3_1 =>4 THEN whl_tot_use_a_p_me3_1=4;
            ELSE  whl_tot_use_a_p_me3_1 = whl_tot_use_a_p_me3;
       rannum=UNIFORM(161328064);
       IF rannum<0.7 THEN addbad30_d=addbad30;
            ELSE addbad30_d='';
       IF addbad30=0  THEN addbad30_cnt=0;
          ELSE IF addbad30=1 and card_co_by_dlnqncy_cnt_d>0
          THEN addbad30_cnt=card_co_by_dlnqncy_cnt_d;
             ELSE addbad30_cnt=1;
       IF rannum<0.7 THEN addbad30_cnt_d=addbad30_cnt;
            ELSE addbad30_cnt_d='' ;
RUN;
```

```
/*종속변수에 유의한 영향을 주는 독립변수 선정 Development(70%)*/
PROC TTEST DATA=credit_rm;
   CLASS addbad30_d;
   VAR actv_ltst_init_loan_amt_BL1--avg_use_mnthcnt_ba init_dlnqncy_amt_d
       ltst_dlnqncy_amt_d dlnqncy_amt_d--tot_use_card_r_p_dt62_ind  age;
RUN;

 /*종속변수에 유의한 영향을 주는 독립변수 선정 Development(70%)*/
PROC FREQ DATA=credit_rm;
   TABLE addbad30_d*(duty_ind gndr house_type_ind) / CHISQ;
RUN;


/*변수선택************************************************/
/* T-TEST와 CHISQ검정 결과 유의한 변수 선택 후 데이터셋 */
DATA credit_rm_1;
   SET  credit_rm;
      KEEP  addbad30 addbad30_cnt  ddbad30_d  addbad30_cnt_d
            avg_delq_amt_3yr  avg_delq_days_3yr_cc avg_use_mnthcnt_ba
            c_12m_new_loan_cnt_BL2 cash_use_tot_card_c_p_mx12
            cash_use_bc_card_c_p_1 in_12m_issue_use_a_p_me18
            in_24m_iss_use_amt_r_p_me18 in_24m_issue_use_a_p_me18_1
            in_6m_issue_use_a_p_me18 ltst_card_use_a_p_me12 max_cash_a_p_av18
            max_cash_tot_use_a_p_me3 max_crdt_tot_use_a_p_me12
            max_delq_amt_3yr  max_delq_days_3yr_cc3 max_use_a_p_me3
            num_fi_inq_12m num_fi_inq_12m_1 sum_delq_amt_3yr_1
            sum_delq_cnt_3yr_1 sum_delq_days_3yr tot_avg_cash_a_p_me3
            tot_avg_use_a_p_me3_1 tot_cash_amt_r_p_mx12
            tot_cash_use_card_r_p_mx18  tot_use_card_r_p_mx12
            tot_use_vs_max_amt_r_p_me6  use_tot_avg_use_a_p_me6
            use_tot_card_c_p_1  whl_tot_cash_a_p_mx12 whl_tot_use_a_p_me3_1
            duty_ind  gndr  house_type_ind;

RUN;
```

```
/* T-TEST 결과 유의한 변수 중 FORWARD방법으로 최종변수선택*/
PROC LOGISTIC DESCENDING DATA=credit_rm_1;
   MODEL addbad30_d = avg_delq_amt_3yr avg_delq_days_3yr_cc avg_use_mnthcnt_ba
                      c_12m_new_loan_cnt_BL2 cash_use_tot_card_c_p_mx1
                      cash_use_bc_card_c_p_1  in_12m_issue_use_a_p_me18
                      in_24m_iss_use_amt_r_p_me18 in_24m_issue_use_a_p_me18_1
                      in_6m_issue_use_a_p_me18 ltst_card_use_a_p_me1
                      max_cash_a_p_av18 max_cash_tot_use_a_p_me3
                      max_crdt_tot_use_a_p_me12  max_delq_amt_3yr
                      max_delq_days_3yr_cc3 max_use_a_p_me3 num_fi_inq_12m
                      num_fi_inq_12m_1 sum_delq_amt_3yr_1  sum_delq_cnt_3yr_1
                      sum_delq_days_3yr tot_avg_cash_a_p_me3
                      tot_avg_use_a_p_me3_1 tot_cash_amt_r_p_mx12
                      tot_cash_use_card_r_p_mx18 tot_use_card_r_p_mx12
                      tot_use_vs_max_amt_r_p_me6 use_tot_avg_use_a_p_me6
                      use_tot_card_c_p_1  whl_tot_cash_a_p_mx12
                      whl_tot_use_a_p_me3_1 duty_ind  gndr  house_type_ind /
                      SELECTION=FORWARD   SLENTRY =0.000001 ;
     OUTPUT OUT=forward P=phat;
RUN;

/* 정분류율(절단점0.5)*/
DATA forward;
   SET forward  (KEEP= addbad30 addbad30_d phat);
RUN;

DATA forward;
   SET forward;
      IF PHAT>0.5  THEN YHAT=1;
             ELSE YHAT=0;
RUN;

PROC FREQ DATA=forward;
  forward: TABLES addbad30_d*YHAT/NOCOL NOROW NOPERCENT;
RUN;
```

```
/* T-TEST 결과 유의한 변수 중 BACKWARD방법으로 최종변수선택*/
PROC LOGISTIC DESCENDING DATA=credit_rm_1;
   MODEL addbad30_d = avg_delq_amt_3yr avg_delq_days_3yr_cc avg_use_mnthcnt_ba
                      c_12m_new_loan_cnt_BL2 cash_use_tot_card_c_p_mx1
                      cash_use_bc_card_c_p_1  in_12m_issue_use_a_p_me18
                      in_24m_iss_use_amt_r_p_me18 in_24m_issue_use_a_p_me18_1
                      in_6m_issue_use_a_p_me18 ltst_card_use_a_p_me1
                      max_cash_a_p_av18 max_cash_tot_use_a_p_me3
                      max_crdt_tot_use_a_p_me12  max_delq_amt_3yr
                      max_delq_days_3yr_cc3 max_use_a_p_me3 num_fi_inq_12m
                      num_fi_inq_12m_1 sum_delq_amt_3yr_1  sum_delq_cnt_3yr_1
                      sum_delq_days_3yr tot_avg_cash_a_p_me3
                      tot_avg_use_a_p_me3_1 tot_cash_amt_r_p_mx12
                      tot_cash_use_card_r_p_mx18 tot_use_card_r_p_mx12
                      tot_use_vs_max_amt_r_p_me6 use_tot_avg_use_a_p_me6
                      use_tot_card_c_p_1  whl_tot_cash_a_p_mx12
                      whl_tot_use_a_p_me3_1  duty_ind  gndr  house_type_ind /
                      SELECTION=BACKWARD  SLENTRY =0.0000001 ;
     OUTPUT OUT=backward P=phat;
RUN;

/* 정분류율(절단점0.5)*/
DATA backward;
   SET backward  (KEEP = addbad30 addbad30_d phat);
RUN;

DATA backward;
   SET backward;
      IF PHAT>0.5  THEN YHAT=1;
             ELSE YHAT=0;
RUN;

PROC FREQ DATA=backward;
  backward : TABLES addbad30_d*YHAT/NOCOL NOROW NOPERCENT;
RUN;
```

```
/* T-TEST 결과 유의한 변수 중 STEPWISE방법으로 최종변수선택*/
PROC LOGISTIC DESCENDING DATA=credit_rm_1;
   MODEL addbad30_d = avg_delq_amt_3yr avg_delq_days_3yr_cc avg_use_mnthcnt_ba
                      c_12m_new_loan_cnt_BL2 cash_use_tot_card_c_p_mx1
                      cash_use_bc_card_c_p_1  in_12m_issue_use_a_p_me18
                      in_24m_iss_use_amt_r_p_me18 in_24m_issue_use_a_p_me18_1
                      in_6m_issue_use_a_p_me18 ltst_card_use_a_p_me1
                      max_cash_a_p_av18 max_cash_tot_use_a_p_me3
                      max_crdt_tot_use_a_p_me12  max_delq_amt_3yr
                      max_delq_days_3yr_cc3 max_use_a_p_me3 num_fi_inq_12m
                      num_fi_inq_12m_1 sum_delq_amt_3yr_1  sum_delq_cnt_3yr_1
                      sum_delq_days_3yr tot_avg_cash_a_p_me3
                      tot_avg_use_a_p_me3_1 tot_cash_amt_r_p_mx12
                      tot_cash_use_card_r_p_mx18 tot_use_card_r_p_mx12
                      tot_use_vs_max_amt_r_p_me6 use_tot_avg_use_a_p_me6
                      use_tot_card_c_p_1  whl_tot_cash_a_p_mx12
                      whl_tot_use_a_p_me3_1  duty_ind  gndr  house_type_ind /
                      SELECTION=STEPWISE  SLENTRY =0.000001
                                  SLSTAY =0.000001;
          OUTPUT OUT=stepwise P=phat;;
RUN;

/* 정분류율(절단점0.5)*/
DATA stepwise;
   SET stepwise  (KEEP = addbad30 addbad30_d phat);
RUN;

DATA stepwise;
   SET stepwise;
      IF PHAT>0.5  THEN YHAT=1;
             ELSE YHAT=0;
RUN;

PROC FREQ DATA=stepwise;
  stepwise : TABLES addbad30_d*YHAT/NOCOL NOROW NOPERCENT;
RUN;
```

2. 로지스틱 회귀모형

```
/*최종 로지스틱회귀모형 구축(상관관계높은 변수 제거)*/
PROC LOGISTIC DESCENDING DATA=credit_rm_1;
   MODEL addbad30_d =avg_use_mnthcnt_ba c_12m_new_loan_cnt_BL2
                     max_delq_days_3yr_cc3 num_fi_inq_12m
                     tot_avg_use_a_p_me3_1 tot_cash_amt_r_p_mx12
                     tot_use_vs_max_amt_r_p_me6  duty_ind
                     house_type_ind / Lackfit;
      OUTPUT OUT=logic P=phat;
RUN;

/* 최종 선택된 변수(stepwise) 기초통계량*/
PROC MEANS  N NMISS MEAN MIN MAX  DATA=credit_rm_1;
   VAR  avg_use_mnthcnt_ba  c_12m_new_loan_cnt_BL2 in_24m_issue_use_a_p_me18_1
        max_delq_days_3yr_cc3 num_fi_inq_12m  tot_avg_use_a_p_me3_1
        tot_cash_amt_r_p_mx12 tot_use_vs_max_amt_r_p_me6 duty_ind
        house_type_ind;
RUN;

/* 최종 선택된 변수(stepwise) 상관분석*/
PROC CORR DATA=credit_rm_1;
   VAR  avg_use_mnthcnt_ba  c_12m_new_loan_cnt_BL2 in_24m_issue_use_a_p_me18_1
        max_delq_days_3yr_cc3 num_fi_inq_12m  tot_avg_use_a_p_me3_1
        tot_cash_amt_r_p_mx12 tot_use_vs_max_amt_r_p_me6 duty_ind
        house_type_ind;
RUN;

/*다중공선성 진단 (stepwise에 의한 변수선택)*/
PROC REG DATA=credit_rm_1;
   MODEL addbad30_d =avg_use_mnthcnt_ba c_12m_new_loan_cnt_BL2
                     max_delq_days_3yr_cc3 num_fi_inq_12m
                     tot_avg_use_a_p_me3_1 tot_cash_amt_r_p_mx12
                     tot_use_vs_max_amt_r_p_me6  duty_ind
                     house_type_ind / VIF COLLIN;
RUN;
```

3. 포아송 회귀모형

```
/*최종 포아송 회귀모형 구축(상관관계높은 변수 제거)*/
PROC GENMOD DATA=credit_rm_1;
   MODEL addbad30_cnt_d = avg_use_mnthcnt_ba c_12m_new_loan_cnt_BL2
                          max_delq_days_3yr_cc3  num_fi_inq_12m
                          tot_avg_use_a_p_me3_1 tot_cash_amt_r_p_mx12
                          tot_use_vs_max_amt_r_p_me6  duty_ind
                          house_type_ind / link=log dist=poisson;
      OUTPUT OUT=Poi P=pred;
RUN;
```

4. 영과잉-포아송 회귀모형

```
/*the nlmixed procedure models a degenerate zero and a Poisson distribution the end
product giving you a probability of an observation being in the zero distribution */

proc nlmixed data= credit_rm_1;

/* a0 = intercept of the logistic model of the inflation prob, a1 is that slope,  b0-b1
are the regression coefficients for the Poisson mean */

parameters a0=0 a1=0 a2=0  a3=0 a4=0 a5=0 a6=0 a7=0 a8=0 a9=0
                b0=0 b1=0 b2=0  b3=0 b4=0 b5=0  b6=0 b7=0 ;

/* linear predictor for the inflation probability */

 linpinfl = a0 + a1*avg_use_mnthcnt_ba + a2*c_12m_new_loan_cnt_BL2
              + a3*max_delq_days_3yr_cc3 + a4*num_fi_inq_12m
              + a5*tot_avg_use_a_p_me3_1 + a6*tot_cash_amt_r_p_mx12
              + a7* tot_use_vs_max_amt_r_p_me6 + a8*duty_ind + a9*house_type_ind ;
```

```
/* infprob = inflation probability for zeros              */
/*         = logistic transform of the linear predictor */

 infprob   = 1/(1+ exp(-linpinfl));


/* Poisson mean */

lambda = exp(b0 +  b1*in_24m_issue_use_a_p_me18_1 +  b2*max_delq_days_3yr_cc3
                +  b3*num_fi_inq_12m +  b4*tot_avg_use_a_p_me3_1
                +  b5*tot_cash_amt_r_p_mx12 +  b6*tot_use_vs_max_amt_r_p_me6
                +  b7*duty_ind);



/* Build the ZIP log likelihood */

 if addbad30_cnt_d=0 then

       ll = log(infprob +  (1-infprob)*exp(-lambda));

   else ll = log((1-infprob)) +  addbad30_cnt_d*log(lambda)
              - lgamma(addbad30_cnt_d+ 1) - lambda;
 model addbad30_cnt_d   ~ general(ll);


/*  predict statement to get the predicted number of
 RESPONSES given the Poisson mean and the inflation probability */

predict 1-infprob out = pred_p;
predict (1-infprob)*lambda out = pred_z;
predict lambda out = pred_lambda;

RUN;
```