

명세서

발명의 명칭: 2단계의 로지스틱 회귀분석을 이용한 신용평가모델 운영 방법 및 이를 수행하는 서버

기술분야

- [1] 본 발명은 2단계의 로지스틱 회귀분석을 이용한 신용평가모델 운영 방법 및 이를 수행하는 서버에 관한 것이다. 구체적으로, 본 발명은 사용자의 로그데이터를 이용한 신용평가모델의 성능 개선을 위해 2단계 로지스틱 회귀분석 모델을 운영하는 방법에 관한 것이다.

[2]

배경기술

- [3] 이 부분에 기술된 내용은 단순히 본 실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.
- [4] 최근 금융기관 또는 전자금융회사가 컴퓨팅 장치를 통하여 금융상품 및 서비스를 제공함에 따라, 사용자가 금융기관 또는 전자금융업자의 종사자와 직접 대면하지 않고, 온라인으로 수행되는 금융거래가 증가하고 있다. 다만, 금융 거래를 제공하는 채널이 다양화되고 거래량이 증가됨에 따라 금융거래의 부도비율도 함께 빠른 속도로 증가되고 있다. 이에 따라, 금융거래에서 사용자의 신용도를 정확하고 빠르게 평가하고, 예측할 수 있는 방법에 대한 중요성이 날로 높아지고 있다.
- [5] 국내외 대부분의 은행 등 소매금융회사들은 로지스틱 회귀분석(logistic regression)을 사용하여 신용평가모델을 개발하는데, 로지스틱 회귀분석 모델은 설명변수들이 상호 선형 독립적이어야 한다는 가정(multicollinearity)으로 인해서 최대 10개 내외의 설명변수만을 모델에 사용할 수 있다. 즉, 새로운 정보영역의 변수를 발굴하게 되더라도, 선형독립성의 제약으로 인해서 기존에 사용하던 변수들을 사용할 수 없게 되므로 해당 모델의 성능 개선에 한계가 존재하였다. 즉, 설명변수 간의 선형독립성이 감소하면 설명변수의 통계적 유의성이 과소평가되는 문제점이 존재하였다.
- [6] 한편, 최근에는 더 많은 변수를 사용하여 신용평가모델의 예측성능을 향상시킬 수 있는 머신러닝 또는 딥러닝 기술을 이용하려는 시도가 많아지고 있다. 다만, 이들 기술의 경우, 설명변수에 대한 제약사항이 없기 때문에, 가용한 모든 정보를 활용할 수 있다는 장점이 있으나, 설명변수와 예측결과 간의 함수관계를 파악할 수 없어, 설명력이 필요한 금융비즈니스 분야에서는 활용하기가 어렵다는 한계가 있었다.

[7]

발명의 상세한 설명

기술적 과제

- [8] 본 발명의 목적은, 신용평가모델의 설명력은 완벽하게 갖추면서 더 많은 설명변수를 사용하여 신용평가모델의 성능을 개선시킬 수 있는 2단계 신용평가모델을 제공하는 신용평가모델 운영 방법을 제공하는 것이다.
- [9] 또한, 본 발명의 목적은, 사용자의 로그데이터를 기초로 설명력 높은 중요변수들을 생성하고, 이를 기초로 1단계 로지스틱 회귀분석을 통해 선정된 파생변수를 이용하여 2단계 로지스틱 회귀분석을 수행하여 사용자의 신용도를 평가하는 신용평가모델 운영 방법을 제공하는 것이다.
- [10] 본 발명의 목적들은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 본 발명의 다른 목적 및 장점들은 하기의 설명에 의해서 이해될 수 있고, 본 발명의 실시예에 의해 보다 분명하게 이해될 것이다. 또한, 본 발명의 목적 및 장점들은 특허 청구 범위에 나타낸 수단 및 그 조합에 의해 실현될 수 있음을 쉽게 알 수 있을 것이다.

[11]

과제 해결 수단

- [12] 본 발명의 일 실시예에 따른 신용평가모델 운영 방법은, 사용자의 로그데이터를 수신하고, 상기 로그데이터에 포함된 변수기초항목을 선정하는 단계, 상기 로그데이터 내에서 상기 변수기초항목의 빈도(Frequency)를 산출함으로써 후보변수를 생성하는 단계, 상기 후보변수에 서로 다른 타임 윈도우(time-window) 또는 서로 다른 산출방법을 적용하여 복수의 제1 파생변수를 생성하는 단계, 상기 복수의 제1 파생변수와 관련된 값과 미리 정한 기준치를 비교하여 중요변수를 선별하는 단계, 상기 중요변수를 입력변수로 하고, 상기 사용자의 신용에 대한 정보를 종속변수로 하는 제1 단계 모델을 도출하는 단계, 상기 중요변수 중에서 상기 제1 단계 모델에 적용할 제1 최종변수를 선정하고, 상기 제1 최종변수에 대한 제1 가중치를 산출하는 단계, 상기 제1 최종변수 및 상기 제1 가중치를 이용하여 제2 파생변수를 생성하는 단계, 상기 제2 파생변수를 입력변수로 하고, 상기 사용자의 신용에 대한 정보를 종속변수로 하는 제2 단계 모델을 도출하는 단계, 및 상기 제1 파생변수 중에서 상기 제2 단계 모델에 적용할 제2 최종변수를 선정하고, 상기 제2 최종변수에 대한 제2 가중치를 산출하는 단계를 포함한다.
- [13] 또한, 상기 변수기초항목을 선정하는 단계는, 상기 로그데이터에 포함된 이벤트코드를 미리 정해진 카테고리를 이용하여 구분하고, 상기 카테고리에 속한 이벤트코드를 미리 정해진 복수의 피처를 이용하여 구분함으로써, 상기 이벤트코드에 대응되는 변수기초항목을 선정하는 것을 포함할 수 있다.
- [14] 또한, 상기 후보변수를 생성하는 단계는, 상기 변수기초항목의 단어빈도(TF)와 단어빈도-역문서빈도(TF-IDF)를 산출하여 상기 후보변수를 생성하는 것을 포함하되, 상기 단어빈도(TF)는, 단순빈도, 불린빈도, 증가빈도, 또는 로그빈도를 이용하여 산출되고, 상기 단어빈도-역문서빈도(TF-IDF)는, 단어빈도(TF)에

역문서빈도(IDF)를 곱함으로써 산출되는 것을 포함할 수 있다.

- [15] 또한, 상기 복수의 제1 파생변수를 생성하는 단계는, 상기 후보변수에 서로 다른 크기의 복수의 타임 윈도우 중 어느 하나와, 복수의 산출방법 중 어느 하나를 이용하여 상기 제1 파생변수를 생성하는 것을 포함하되, 상기 타임 윈도우는, 서로 다른 기간에 대해 설정될 수 있고, 상기 산출방법은, 평균, 합계, 최대값, 및 최소값을 포함할 수 있다.
- [16] 또한, 상기 중요변수를 선별하는 단계는, 상기 복수의 제1 파생변수 중, 단변량 로지스틱 회귀분석의 수행을 통한 P값(P-value)이 미리 정해진 기준값보다 작거나, 상기 복수의 제1 파생변수 중, IV값이 미리 정해진 기준치보다 큰 제1 파생변수를 상기 중요변수로 선별하는 것을 포함하되, 상기 IV값은, 하기 <수학식>에 의해 도출될 수 있다.
- [17] <수학식>
- [18]
$$IV = \sum_i (\% \text{ of Goods} - \% \text{ of Bads}) \times WOE_i$$
- [19] 여기서, '% of Goods'는 좋음(Good)으로 평가받은 집단에 대한 전체비율을 의미하고, '% of Bads'는 나쁨(Bad)으로 평가받은 집단에 대한 전체비율을 의미하고, WOE(Weights of Evidence; 이하, WOE)는 나쁨(Bad)으로 평가받은 집단에 대한 좋음(Good)으로 평가받은 모집단의 자연로그를 의미한다.
- [20] 또한, 상기 선별된 중요변수에 대하여 동일한 정보영역(F)에 속하는 변수들끼리 군집화(Grouping)하는 단계를 더 포함하되, 상기 제1 단계 모델을 도출하는 단계는, 특정 정보영역(F)에 포함된 상기 중요변수를 대상으로 상기 제1 최종변수를 선정하는 것을 포함할 수 있다.
- [21] 또한, 상기 제1 단계 모델 및 상기 제2 단계 모델은, 로지스틱 회귀분석 모델(logistic regression model)로 구성될 수 있다.
- [22] 또한, 상기 제1 단계 모델은, 단계적 선택법을 이용하여, 상기 중요변수 중에서 상기 제1 단계 모델에 적용할 상기 제1 최종변수를 선정하고, 상기 제2 단계 모델은, 단계적 선택법을 이용하여, 상기 제2 파생변수 중에서 상기 제2 단계 모델에 적용할 상기 제2 최종변수를 선정하는 것을 포함할 수 있다.
- [23] 또한, 상기 제1 최종변수가 적용된 상기 제1 단계 모델과, 상기 제2 최종변수가 적용된 상기 제2 단계 모델을 이용하여, 새로운 사용자의 로그데이터를 기초로 상기 새로운 사용자의 신용평가를 수행하는 단계를 더 포함할 수 있다.
- [24] 본 발명의 다른 실시예에 따른 신용평가모델 운영 방법은, 사용자의 로그데이터를 수신하고, 상기 로그데이터에 포함된 이벤트코드에 대한 빈도와, 상기 빈도에 대한 하나 이상의 전처리 과정을 통하여 중요변수를 선별하는 단계, 상기 중요변수를 입력변수로 하고, 상기 사용자의 신용에 대한 정보를 종속변수로 하는 제1 단계 로지스틱 회귀분석 모델을 도출하는 단계, 상기 중요변수 중에서 상기 제1 단계 모델에 적용할 제1 최종변수를 선정하고, 상기 제1 최종변수에 대한 제1 가중치를 산출하는 단계, 상기 제1 최종변수 및 상기 제1 가중치를 이용하여 파생변수를 생성하는 단계, 상기 파생변수를 입력변수로

하고, 상기 사용자의 신용에 대한 정보를 종속변수로 하는 제2 단계 로지스틱 회귀분석 모델을 도출하는 단계, 및 상기 파생변수 중에서 상기 제2 단계 모델에 적용할 제2 최종변수를 선정하고, 상기 제2 최종변수에 대한 제2 가중치를 산출하는 단계를 포함한다.

[25] 또한, 상기 제1 단계 모델은, 단계적 선택법을 이용하여, 상기 중요변수 중에서 상기 제1 단계 모델에 적용할 상기 제1 최종변수를 선정하고, 상기 제2 단계 모델은, 단계적 선택법을 이용하여, 상기 파생변수 중에서 상기 제2 단계 모델에 적용할 상기 제2 최종변수를 선정하는 것을 포함할 수 있다.

[26] 또한, 상기 중요변수를 선별하는 단계는, 상기 복수의 제1 파생변수 중, 단변량 로지스틱 회귀분석의 수행을 통한 P값(P-value)이 미리 정해진 기준값보다 작거나, 상기 복수의 제1 파생변수 중, IV값이 미리 정해진 기준치보다 큰 제1 파생변수를 상기 중요변수로 선별하는 것을 포함하되, 상기 IV값은, 하기 <수학식>에 의해 도출될 수 있다.

[27] <수학식>

[28]
$$IV = \sum_i (\% \text{ of Goods} - \% \text{ of Bads}) \times WOE_i$$

[29] 여기에서, '% of Goods'는 좋음(Good)으로 평가받은 집단에 대한 전체비율을 의미하고, '% of Bads'는 나쁨(Bad)으로 평가받은 집단에 대한 전체비율을 의미하고, WOE(Weights of Evidence; 이하, WOE)는 나쁨(Bad)으로 평가받은 집단에 대한 비율 대비 좋음(Good)으로 평가받은 집단의 전체비율 값에 자연로그를 취한 것을 의미한다. 신용평가모델 운영 방법.

[30] 또한, 상기 선별된 중요변수에 대하여 동일한 정보영역(F)에 속하는 변수들끼리 군집화(Grouping)하는 단계를 더 포함하되, 상기 제1 단계 모델을 도출하는 단계는, 특정 정보영역(F)에 포함된 상기 중요변수를 대상으로 상기 제1 최종변수를 선정하는 것을 포함할 수 있다.

[31] 또한, 상기 제1 최종변수가 적용된 상기 제1 단계 모델과, 상기 제2 최종변수가 적용된 상기 제2 단계 모델을 이용하여, 새로운 사용자의 로그데이터를 기초로 상기 새로운 사용자의 신용평가를 수행하는 단계를 더 포함할 수 있다.

[32] 한편, 본 발명의 또 다른 실시예에 따른 신용 평가 서버는, 프로세서, 상기 프로세서에 의해 실행되는 컴퓨터 프로그램을 로드하는 메모리, 및 상기 컴퓨터 프로그램의 실행과정에서 발생하는 데이터를 사용자 단말과 교환하는 인터페이스를 포함하되, 상기 컴퓨터 프로그램은, 상기 사용자 단말로부터 사용자의 로그데이터를 수신하고, 상기 로그데이터에 포함된 이벤트코드에 대한 빈도와, 상기 빈도에 대한 하나 이상의 전처리 과정을 통하여 중요변수를 선별하는 단계, 상기 중요변수를 입력변수로 하고, 상기 사용자의 신용에 대한 정보를 종속변수로 하는 제1 단계 로지스틱 회귀분석 모델을 도출하는 단계, 상기 중요변수 중에서 상기 제1 단계 모델에 적용할 제1 최종변수를 선정하고, 상기 제1 최종변수에 대한 제1 가중치를 산출하는 단계, 상기 제1 최종변수 및 상기 제1 가중치를 이용하여 파생변수를 생성하는 단계, 상기 파생변수를

입력변수로 하고, 상기 사용자의 신용에 대한 정보를 종속변수로 하는 제2 단계 로지스틱 회귀분석 모델을 도출하는 단계, 및 상기 파생변수 중에서 상기 제2 단계 모델에 적용할 제2 최종변수를 선정하고, 상기 제2 최종변수에 대한 제2 가중치를 산출하는 단계를 포함한다.

[33] 또한, 상기 제1 단계 모델은, 단계적 선택법을 이용하여, 상기 중요변수 중에서 상기 제1 단계 모델에 적용할 상기 제1 최종변수를 선정하고, 상기 제2 단계 모델은, 단계적 선택법을 이용하여, 상기 파생변수 중에서 상기 제2 단계 모델에 적용할 상기 제2 최종변수를 선정하는 것을 포함할 수 있다.

[34]

[35] 또한, 상기 중요변수를 선별하는 단계는, 상기 복수의 제1 파생변수 중, 단변량 로지스틱 회귀분석의 수행을 통한 P값(P-value)이 미리 정해진 기준값보다 작거나, 상기 복수의 제1 파생변수 중, IV값이 미리 정해진 기준치보다 큰 제1 파생변수를 상기 중요변수로 선별하는 것을 포함하되, 상기 IV값은, 하기 <수학식>에 의해 도출될 수 있다.

[36] <수학식>

[37]
$$IV = \sum_i (\% \text{ of Goods} - \% \text{ of Bads}) \times WOE_i$$

[38] 여기에서, '% of Goods'는 좋음(Good)으로 평가받은 집단에 대한 전체비율을 의미하고, '% of Bads'는 나쁨(Bad)으로 평가받은 집단에 대한 전체비율을 의미하고, WOE(Weights of Evidence; 이하, WOE)는 나쁨(Bad)으로 평가받은 집단에 대한 좋음(Good)으로 평가받은 모집단의 자연로그를 의미한다.

[39] 또한, 상기 선별된 중요변수에 대하여 동일한 정보영역(F)에 속하는 변수들끼리 군집화(Grouping)하는 단계를 더 포함하되, 상기 제1 단계 모델을 도출하는 단계는, 특정 정보영역(F)에 포함된 상기 중요변수를 대상으로 상기 제1 최종변수를 선정하는 것을 포함할 수 있다.

[40] 또한, 상기 제1 최종변수가 적용된 상기 제1 단계 모델과, 상기 제2 최종변수가 적용된 상기 제2 단계 모델을 이용하여, 새로운 사용자의 로그데이터를 기초로 상기 새로운 사용자의 신용평가를 수행하는 단계를 더 포함할 수 있다.

[41]

발명의 효과

[42] 본 발명의 신용평가모델 운영 방법은, 100개 이상의 많은 변수로 신용평가모델을 개발할 수 있으며, 동시에 해당 모델에 대해 완전한 설명이 가능한 신용평가모델을 제공할 수 있다. 즉, 많은 수의 변수를 이용하더라도 모델에 대한 최초변수값과 최종예측값이 선형관계로 표현되어 완벽한 설명이 가능하므로, 금융비즈니스 분야에 대한 활용성과 신용평가모델에 대한 신뢰성을 향상시킬 수 있다.

[43] 또한, 본 발명의 신용평가모델 운영 방법은, 사용자의 로그데이터를 기반으로 신용평가모델을 생성하되, 기존 로지스틱 회귀분석 모델을 2단계에 걸쳐

제공함으로써, 추가적인 비용을 들이지 않고 신용평가모델의 성능을 개선시킬 수 있다. 또한, 기존 신용평가모델에서 전혀 사용되지 않던 어플리케이션과 관련된 로그데이터를 추가적으로 활용할 수 있는 바, 신용평가에 대한 차별화된 성능지표의 개선을 기대할 수 있다.

[44] 상술한 내용과 더불어 본 발명의 구체적인 효과는 이하 발명을 실시하기 위한 구체적인 사항을 설명하면서 함께 기술한다.

[45]

도면의 간단한 설명

[46] 도 1은 본 발명의 몇몇 실시예에 따른 신용평가모델 운영 시스템을 설명하기 위한 도면이다.

[47] 도 2는 도 1의 신용 평가 서버를 설명하기 위한 도면이다.

[48] 도 3은 본 발명의 몇몇 실시예들에 따른 신용평가모델 운영 방법을 설명하기 위한 블록도이다.

[49] 도 4는 도 3의 변수기초항목을 선정하기 위한 카테고리화 피쳐의 일 예를 나타내는 테이블이다.

[50] 도 5는 도 3의 제1 과생변수를 생성하는 방법을 설명하기 위한 블록도이다.

[51] 도 6은 도 3의 S50 단계 내지 S70 단계에서 이용되는 2단계의 로지스틱 회귀분석 모델의 관계를 설명하기 위한 블록도이다.

[52] 도 7은 2단계의 로지스틱 회귀분석 모델을 이용하여 사용자의 신용을 평가하는 변수항목에 대한 일 예를 나타내는 블록도이다.

[53] 도 8은 본 발명의 몇몇 실시예에 따른 신용평가모델 운영 방법의 전처리 방법을 나타내는 순서도이다.

[54] 도 9는 본 발명의 몇몇 실시예에 따른 신용평가모델 운영 방법에서 2단계 로지스틱 회귀분석 모델을 이용한 신용평가방법에 관한 순서도이다.

[55] 도 10은 본 발명의 몇몇 실시예에 따른 신용평가모델과 종래의 신용평가모델 간의 성능지표 차이를 나타내는 테이블이다.

[56] 도 11은 본 발명의 몇몇 실시예에 따른 신용평가모델 운영 방법을 수행하는 시스템의 하드웨어 구현을 설명하기 위한 도면이다.

[57]

발명의 실시를 위한 형태

[58] 본 명세서 및 특허청구범위에서 사용된 용어나 단어는 일반적이거나 사전적인 의미로 한정하여 해석되어서는 아니된다. 발명자가 그 자신의 발명을 최선의 방법으로 설명하기 위해 용어나 단어의 개념을 정의할 수 있다는 원칙에 따라, 본 발명의 기술적 사상과 부합하는 의미와 개념으로 해석되어야 한다. 또한, 본 명세서에 기재된 실시예와 도면에 도시된 구성은 본 발명이 실현되는 하나의 실시예에 불과하고, 본 발명의 기술적 사상을 전부 대변하는 것이 아니므로, 본 출원시점에 있어서 이들을 대체할 수 있는 다양한 균등물과 변형 및 응용 가능한

예들이 있을 수 있음을 이해하여야 한다.

- [59] 본 명세서 및 특허청구범위에서 사용된 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. '및/또는'이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.
- [60] 본 명세서 및 특허청구범위에서 사용된 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서 "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [61] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해서 일반적으로 이해되는 것과 동일한 의미를 가지고 있다.
- [62] 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [63] 또한, 본 발명의 각 실시예에 포함된 각 구성, 과정, 공정 또는 방법 등은 기술적으로 상호 간 모순되지 않는 범위 내에서 공유될 수 있다.
- [64] 이하에서, 도 1 내지 도 11을 참조하여 본 발명의 몇몇 실시예에 따른 신용평가모델 운영 시스템 및 신용평가모델 운영 방법에 대해 살펴보도록 한다.
- [65]
- [66] 도 1은 본 발명의 몇몇 실시예에 따른 신용평가모델 운영 시스템을 설명하기 위한 도면이다.
- [67] 도 1을 참조하면, 신용평가모델 운영 시스템은, 신용 평가 서버(100), 금융 서버(200), 사용자 단말(300) 및 통신망(400)을 포함할 수 있다.
- [68] 사용자는 금융 서버(200)를 통해 다양한 금융 서비스를 이용할 수 있다. 이때, 사용자는 사용자 단말(300)을 통해 금융 서버(200)에 접속할 수 있고, 사용자 단말(300)에서 요청한 금융 서비스 및 금융 서버(200)에서 제공한 데이터는 로그데이터(log-data) 형태로 금융 서버(200)에 저장될 수 있다.
- [69] 금융 서버(200)와 사용자 단말(300)은 서버-클라이언트 시스템으로 구현될 수 있다. 금융 서버(200)는 각 고객계정에 사용자의 가입명의정보, 인증정보 및 활동정보를 저장 및 관리할 수 있고, 금융과 관련된 다양한 서비스를 사용자

단말(300)에 설치된 금융 어플리케이션을 통해 제공할 수 있다.

- [70] 이때, 금융 어플리케이션은 금융 서비스를 제공하기 위한 전용 어플리케이션이거나, 웹 브라우징 어플리케이션일 수 있다. 여기에서, 전용 어플리케이션은 사용자 단말(300)에 내장된 어플리케이션이거나, 어플리케이션 배포 서버로부터 다운로드 되어 사용자 단말(300)에 설치된 어플리케이션일 수 있다.
- [71] 한편, 금융 서버(200)는 사용자가 요청한 특정 금융 서비스를 제공함에 앞서, 고객의 신용도에 대한 체크가 요구될 수 있다. 이때, 금융 서버(200)는 사용자가 금융 어플리케이션을 이용하면서 남긴 흔적(즉, 로그데이터)을 신용 평가 서버(100)에 전달할 수 있으며, 신용 평가 서버(100)는 수신된 사용자의 로그데이터를 기초로 해당 사용자의 신용도를 예측하여 금융 서버(200)에 전달할 수 있다.
- [72] 신용 평가 서버(100)는 미리 학습된 신용평가모델을 이용하여 사용자의 신용도를 평가하거나 예측할 수 있다. 신용 평가 서버(100)는 복수의 사용자에 대한 데이터를 기초로 미리 학습된 신용평가모델을 이용할 수 있다.
- [73] 신용 평가 서버(100)는 신용평가모델의 정확도를 높일 수 있도록 가능한 많은 수의 변수들을 이용하여 신용평가모델을 운용할 수 있다. 신용 평가 서버(100)는 로지스틱 회귀분석을 이용한 신용평가모델을 운용할 수 있다.
- [74] 다만, 로지스틱 회귀분석 모델의 경우, 최대 10개 내외의 설명변수만을 모델에 사용할 수 있으므로, 새로운 정보영역의 변수를 발굴하게 되더라도 선형독립성의 제약으로 인하여 기존에 사용하던 변수들을 사용할 수 없게 되므로 모델 성능 개선의 한계가 존재할 수 있다.
- [75] 이에 따라, 본 발명의 신용 평가 서버(100)는 로지스틱 회귀분석 모델을 2단계로 신용평가모델을 구성함으로써, 더 많은 설명변수를 사용하여 신용평가모델의 성능은 향상시킴과 동시에, 설명력은 완벽하게 갖추도록 운용할 수 있다. 신용 평가 서버(100)에서 동작하는 본 발명의 신용평가모델 운영 방법에 대한 자세한 설명은 이하에서 후술하도록 한다.
- [76] 한편, 사용자 단말(300)은 유무선 통신 환경에서 복수의 어플리케이션을 동작시킬 수 있는 통신 단말기를 의미한다. 도 1에서 사용자 단말(300)은 휴대용 단말기의 일종인 스마트폰(smart phone)으로 도시되었지만, 본 발명이 이에 한정되는 것은 아니며, 상술한 바와 같이 금융 어플리케이션 또는 SNS 어플리케이션을 동작시킬 수 있는 장치에 제한없이 적용될 수 있다. 예를 들어, 사용자 단말(300)은 퍼스널 컴퓨터(PC), 노트북, 태블릿, 휴대폰, 스마트폰, 웨어러블 디바이스(예를 들어, 위치형 단말기) 등의 다양한 형태의 전자 장치를 포함할 수 있다.
- [77] 부가적으로, 사용자 단말(300)은 고객의 입력을 수신하는 입력부, 비주얼 정보를 디스플레이 하는 디스플레이부, 외부와 신호를 송수신하는 통신부, 고객의 얼굴을 촬영하는 카메라부, 고객의 음성을 디지털 데이터로 변환하는

마이크루, 및 데이터를 프로세싱하고 사용자 단말(300) 내부의 각 유닛들을 제어하며 유닛들 간의 데이터 송/수신을 제어하는 제어부를 포함할 수 있다. 이하, 고객의 명령에 따라 제어부가 사용자 단말(300) 내부에서 수행하는 명령은 사용자 단말(300)이 수행하는 것으로 통칭한다.

- [78] 한편, 통신망(400)은 신용 평가 서버(100), 사용자 단말(300) 및 금융 서버(200)를 연결하는 역할을 수행한다. 즉, 통신망(400)은 사용자 단말(300)이 신용 평가 서버(100) 및 금융 서버(200)에 접속한 후 데이터를 송수신할 수 있도록 접속 경로를 제공하는 통신망을 의미한다. 통신망(400)은 예컨대 LANs(Local Area Networks), WANs(Wide Area Networks), MANs(Metropolitan Area Networks), ISDNs(Integrated Service Digital Networks) 등의 유선 네트워크나, 무선 LANs, CDMA, 블루투스, 위성 통신 등의 무선 네트워크를 망라할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.
- [79]
- [80] 도 2는 도 1의 신용 평가 서버를 설명하기 위한 도면이다.
- [81] 도 2를 참조하면, 본 발명의 신용 평가 서버(100)는 로그 데이터 수집부(110), 변수 생성부(120), 변수 선정 및 관리부(130), 신용평가 모형 운영부(140), 데이터베이스부(150)를 포함한다.
- [82] 로그 데이터 수집부(110)는 금융 서버(200)와 사용자 단말(300) 간에 주고받는 데이터에 대한 로그파일(즉, 로그데이터)를 수집하는 기능을 수행한다.
- [83] 여기에서, 로그데이터는 사용자가 금융 어플리케이션을 통해 요청한 금융 서비스 및 이에 대한 응답으로 사용자에게 제공된 데이터를 포함할 수 있다. 로그데이터는 이벤트코드(Event code) 형태로 저장될 수 있으며, 로그데이터는 복수의 이벤트코드를 포함할 수 있다. 로그 데이터 수집부(110)는 금융 서버(200)로부터 사용자 계정에 저장된 로그데이터를 수신하거나, 사용자 단말(300)과 금융 서버(200) 간의 데이터 흐름을 모니터링함으로써 로그데이터를 수집할 수 있다.
- [84] 변수 생성부(120)는 수신된 로그데이터를 기초로 변수기초항목을 선정하고, 로그데이터 내에서 상기 변수기초항목의 빈도(Frequency)를 산출함으로써 후보변수를 생성할 수 있다.
- [85] 구체적으로, 변수 생성부(120)는 로그데이터에 포함된 이벤트코드를 미리 정해진 카테고리 및 피처를 이용하여 구분함으로써, 이벤트코드에 대응되는 복수의 변수기초항목을 선정할 수 있다.
- [86] 또한, 변수 생성부(120)는 선정된 복수의 변수기초항목에 대하여 단어빈도(TF)와 단어빈도-역문서빈도(TF-IDF)를 산출하여 복수의 후보변수를 생성할 수 있다. 단어빈도(TF)와 단어빈도-역문서빈도(TF-IDF)의 구체적인 산출방법은 이하에서 후술하도록 한다.
- [87] 또한, 변수 생성부(120)는 복수의 후보변수에 대하여 서로 다른 타임 윈도우(Time Windows) 및 서로 다른 산출방법을 각각 적용함으로써, 복수의 제1

과생변수를 생성할 수 있다. 제1 과생변수를 생성하는 방법에 대한 자세한 설명도 이하에서 후술하도록 한다.

- [88] 변수 선정 및 관리부(130)는 변수 생성부(120)에서 생성된 제1 과생변수에 대하여 미리 정해진 기준치와 비교하고, 이를 기초로 중요변수를 선별한다. 또한, 변수 선정 및 관리부(130)는 선별된 중요변수를 정보영역(F) 별로 군집화(또는, 그룹핑)하여 저장 및 관리할 수 있다. 이때, 중요변수는 1단계 로지스틱 회귀분석 모델의 입력변수로 이용될 수 있다.
- [89] 신용평가 모형 운영부(140)는 2-스테이지의 로지스틱 회귀분석 모델을 운영할 수 있는 1단계 모델 운영부(140a) 및 2단계 모델 운영부(140b)를 포함한다.
- [90] 1단계 모델 운영부(140a)는 첫번째 로지스틱 회귀분석 모델을 운영하는 운영주체가 될 수 있다. 1단계 모델 운영부(140a)는 동일한 정보영역(F)에 있는 중요변수들을 입력변수로 하고, 신용에 대한 우량/불량 정보(즉, 사용자 신용의 부도여부)를 종속변수로 하여 1단계 로지스틱 회귀분석 모델을 동작시킬 수 있다. 1단계 로지스틱 회귀분석 모델에 적용되는 입력변수 및 종속변수와, 이들 간의 관계를 나타내는 수식에 대한 자세한 설명은 이하에서 후술하도록 한다.
- [91] 2단계 모델 운영부(140b)는 두번째 로지스틱 회귀분석 모델을 운영하는 운영주체가 될 수 있다. 2단계 모델 운영부(140b)는 1단계 모델의 출력값(제1 최종변수의 우량/불량 확률)을 입력변수로 하고, 신용에 대한 우량/불량 정보(즉, 사용자 신용의 부도여부)를 종속변수로 하는 2단계 로지스틱 회귀분석 모델을 동작시킬 수 있다. 2단계 로지스틱 회귀분석 모델에 적용되는 입력변수 및 종속변수와, 이들 간의 관계를 나타내는 수식에 대한 자세한 설명도 이하에서 후술하도록 한다.
- [92] 데이터베이스부(150)는 신용평가 모형 운영부(140)에서 운영되는 각각의 로지스틱 회귀분석 모델에 적용되는 각종 변수들에 대한 정보를 저장 및 관리할 수 있다. 또한, 데이터베이스부(150)는 신용평가 모형 운영부(140)를 훈련시키거나, 신용평가를 수행하기 위한 다양한 사용자의 로그데이터 및 로그데이터의 전처리 과정에서 산출된 중간데이터도 함께 저장 및 관리할 수 있다. 다만, 이는 몇몇 예시에 불과하며 본 발명이 이에 한정되는 것은 아니다.
- [93] 이하에서는, 신용 평가 서버(100)에서 수행되는 본 발명의 몇몇 실시예에 따른 신용평가모형 운영 방법에 대해 자세히 설명하도록 한다.
- [94]
- [95] 도 3은 본 발명의 몇몇 실시예들에 따른 신용평가모형 운영 방법을 설명하기 위한 블록도이다. 도 4는 도 3의 변수기초항목을 선정하기 위한 카테고리화 피쳐의 일 예를 나타내는 테이블이다. 도 5는 도 3의 제1 과생변수를 생성하는 방법을 설명하기 위한 블록도이다. 도 6은 도 3의 S50 단계 내지 S70 단계에서 이용되는 2단계의 로지스틱 회귀분석 모델의 관계를 설명하기 위한 블록도이다. 도 7은 2단계의 로지스틱 회귀분석 모델을 이용하여 사용자의 신용을 평가하는 변수항목에 대한 일 예를 나타내는 블록도이다.

- [96] 우선, 도 3을 참조하면, 본 발명의 몇몇 실시예에 따른 신용평가모델 운영 방법에서, 신용 평가 서버(100)는 신용평가모델에 입력되는 변수를 선정하기 위한 전처리과정(VS)을 수행할 수 있다.
- [97] 구체적으로, 신용 평가 서버(100)는 변수기초항목을 선정한다(S10). 이때, '변수기초항목'은 금융 서버(200)와 사용자 단말(300) 간에 주고받은 데이터의 기록(이하, 로그데이터(log-data))를 기초로 선정된다. 즉, 신용 평가 서버(100)는 금융 서버(200)로부터 수신한 로그데이터를 이용하여 변수기초항목을 선정할 수 있다.
- [98] 로그데이터는 사용자가 사용자 단말(300)에 설치된 어플리케이션을 통해 수행한 다양한 동작에 관련된 데이터를 포함하며, 사용자의 행동 특성을 포착하기 위한 기초 데이터로 이용될 수 있다. 이때, 금융 서버(200)는 로그데이터에 포함된 이벤트코드(Eventcode)를 미리 정해진 카테고리(Category)를 이용하여 구분하고, 해당 카테고리에 속한 이벤트코드를 다시 미리 정해진 복수의 피쳐(Features)를 이용하여 구분할 수 있다.
- [99] 예를 들어, 도 4를 참조하면, 로그데이터에 포함된 이벤트코드는 등록(Registration), 개인설정(Custom Setting), 메뉴클릭(menu click), 인증(Authentication), 계정활동(Account Activity), 송금활동(Transaction Activity), 체크카드(Check Card), 로그인/로그아웃(Login/Logout), 추천(Recommendation), OCR(Optical character recognition; 이하OCR)의 카테고리로 구분될 수 있다. 또한, 각각의 카테고리에 속한 이벤트코드는 복수의 피쳐로 구분되어 저장될 수 있다. 예를 들어, '계정등록' 카테고리에 속한 이벤트코드는 회원가입, 계좌개설, 중도해지, 추천인등록 등의 피쳐 아이템(Feature item; 이하, 피쳐)들로 구분되어 저장될 수 있다.
- [100] 이때, 로그데이터는 금융 서버(200)에 저장될 수 있으며, 신용 평가 서버(100)는 금융 서버(200)로부터 로그데이터를 수신하여 변수기초항목을 선정하는데 이용할 수 있다. 또한, 로그데이터는 금융 서버(200) 및 신용 평가 서버(100)에서 고객별로 구분되어 저장 및 이용될 수 있다.
- [101] 신용 평가 서버(100)는 로그데이터에 포함된 이벤트코드들의 집합을 일종의 문서로 보고, 자주 등장하는 이벤트코드가 고객 행동의 특징을 나타낸다고 가정한다. 이어서, 신용 평가 서버(100)는 고객 행동의 특징을 세부적으로 구분하기 위해 이벤트코드에 이벤트코드가 속한 카테고리(Category)와 피쳐에 대한 정보를 할당하여 이용할 수 있다. 신용 평가 서버(100)는 로그데이터에 포함된 복수의 이벤트코드를 카테고리 및 피쳐를 이용하여 구분함으로써 각각의 이벤트코드에 대응되는 복수의 변수기초항목을 선정할 수 있다.
- [102] 부가적으로, 이벤트코드에 대한 카테고리 및 피쳐에 대한 정보는 금융 서버(200)에서 미리 할당되어 신용 평가 서버(100)에 제공될 수 있음은 물론이다.
- [103]
- [104] 이어서, 신용 평가 서버(100)는 선정된 복수의 변수기초항목 또는 특정

이벤트코드에 대하여 각각의 빈도(Frequency)를 계산하여 복수의 후보변수(Candidate Variables; 이하, CV)를 생성하고, 생성된 후보변수(CV)에 대하여 미리 정해진 기간에 대한 복수의 산출방법을 이용하여 제1 파생변수를 생성한다(S20).

- [105] 여기에서, 후보변수(CV)는, 로그데이터에 포함된 각 이벤트코드 또는 이벤트코드가 속한 변수기초항목의 '단어빈도'(Term Frequency; 이하 TF)와 '역문서빈도'(Inverse Document Frequency; 이하, IDF), 및 단어빈도(TF)와 역문서빈도(IDF)를 곱한 '단어빈도-역문서빈도'(이하, TF-IDF)를 의미한다.
- [106] 구체적으로, 단어빈도(TF)는 특정 단어가 문서(또는, 로그데이터) 내에서 반복되는 빈도를 의미한다. 단어빈도(TF)는 단순빈도(Simple Frequency), 불린빈도(Boolean Frequency), 증가빈도(Increase Frequency), 로그빈도(Log Frequency)를 이용하여 다양하게 계산될 수 있다. 이때, 단순빈도는 특정 이벤트코드의 개수를 카운트한 값을 의미한다. 불린빈도는 특정 이벤트코드가 한번 이상 나타나면 '1', 그렇지 않으면 '0'으로 계산된 값을 의미한다. 증가빈도는 특정 이벤트코드의 단순빈도를 최빈 이벤트코드의 빈도값으로 나눈 값을 의미한다. 로그빈도는 특정 이벤트코드의 단순빈도에 1을 더한 후 자연로그를 취한 값을 의미한다.
- [107] 본 발명에서 단어빈도(TF)를 구하는 방법이 상기 기술한 4개의 방법으로 한정되는 것은 아니나, 이하에서는 설명의 편의를 위해 4개의 단어빈도를 산출하는 것을 예로 들어 설명하도록 한다.
- [108] 한편, 역문서빈도(IDF)는 다음 <수학식 1>에 의해 정의될 수 있다.
- [109] <수학식 1>
- [110]
$$IDF(t) = \log[n / \{1 + df(t)\}]$$
- [111]
- [112] 이때, t는 특정 이벤트코드를 의미하고, n은 전체 사용자(예를 들어, 고객)의 수를 의미하고, df(t)는 특정 이벤트코드 t가 발생한 사용자의 수를 의미한다.
- [113] 즉, 위 <수학식 1>에 따라, 역문서빈도(IDF)는 한 단어가 문서(또는, 로그데이터) 전체에서 얼마나 공통적으로 나타나는지를 나타내는 값을 의미한다. 즉, 역문서빈도(IDF)는 하나의 이벤트코드(또는, 변수기초항목)가 문서(또는, 로그데이터)에서 얼마나 공통적으로 나타나는지를 의미한다.
- [114] 이어서, 신용 평가 서버(100)는 각 이벤트코드(또는, 변수기초항목) 별로 단어빈도(TF)를 계산한 후, 역문서빈도(IDF)를 단어빈도(TF)에 곱하거나 곱하지 않음으로써 하나 이상의 후보변수(CV)를 생성할 수 있다. 예를 들어, 단어빈도(TF)가 4개의 방법으로 계산되고, 역문서빈도(IDF)를 곱하거나 곱하지 않음으로써 8개의 후보변수(CV)를 생성할 수 있다.
- [115] 만약, 특정한 이벤트코드(또는, 변수기초항목)의 기록이 소수의 사용자에게서 자주 등장할수록, 또는 전체 고객에게서 그 이벤트코드가 드물게 등장할수록, TF-IDF값은 커지게 된다. 따라서, 많은 고객에게 흔하게 발생하는 이벤트코드는

TF-IDF값이 작아지게 되므로, 결과적으로 이러한 이벤트코드는 모델 변수로 채택되기 어렵게 된다.

[116]

[117] 이어서, 신용 평가 서버(100)는 생성된 복수의 후보변수(CV)에 대하여 미리 정해진 기간에 대한 다양한 산출방법을 적용함으로써 복수의 제1 파생변수를 생성한다.

[118] 구체적으로, 도 5를 참조하면, 신용 평가 서버(100)는 복수의 후보변수(CV)에 대하여, 복수의 타임 윈도우(Time Windows) 중 어느 하나와, 복수의 산출방법 중 어느 하나를 선택하여, 복수의 제1 파생변수를 생성할 수 있다.

[119] 예를 들어, 복수의 타임 윈도우(Time Windows)는 최근 1개월, 최근 3개월, 최근 6개월, 최근 9개월, 최근 12개월을 포함할 수 있고, 복수의 산출방법은 특정 후보변수의 선택된 타임 윈도우에 대한 평균, 합계, 최대값, 최소값을 포함할 수 있다. 다만, 본 발명에서, 타임 윈도우가 전술한 내용에 한정되는 것은 아니며, 타임 윈도우의 시점 및 종점은 미리 정해진 기간으로 설정되어 이용될 수 있다. 산출방법 또한 전술한 내용에 한정되는 것은 아니다.

[120] 즉, 신용 평가 서버(100)는 복수의 후보변수(CV)에 대해 미리 정해진 복수의 타임 윈도우와 복수의 산출방법의 조합을 통하여, 다양한 제1 파생변수를 생성할 수 있다. 산술적으로 제1 파생변수는 타임 윈도우의 개수와 산출방법의 개수의 곱에 해당하는 개수만큼 산출되어 이용될 수 있다.

[121]

[122] 이어서, 다시 도 3을 참조하면, 신용 평가 서버(100)는 생성된 복수의 제1 파생변수 중에서 미리 설정된 기준을 만족하는 제1 파생변수를 중요변수로 선별한다(S30). 이때, 신용 평가 서버(100)는 중요변수를 선별하는 방법으로 1) 통계적 유의성을 나타내는 P값(probability value; p-value)과, 2) 정보가치(Information Value; 이하, IV 값)를 이용할 수 있다.

[123]

[124] 일반적으로, P값은, 특정 가설이 맞다는 전제 하에, 통계값(statistics)이 실제로 관측된 값 이상일 확률을 의미한다. 예를 들어, P값은 귀무 가설(null hypothesis)이 맞다는 전제 하에, 표본에서 실제로 관측된 통계치와 '같거나 더 극단적인' 통계치가 관측될 확률을 의미할 수 있다. P값에 대한 자세한 내용은 이미 공개되어 있으므로, 여기에서 자세한 설명은 생략하도록 한다.

[125] 본 발명에서 P값은 통계적 유의성에 대한 기준으로 이용된다. 신용 평가 서버(100)는 모든 제1 파생변수들을 대상으로 종속변수(예를 들어, 사용자 신용의 부도여부)와 단변량 로지스틱 회귀분석을 수행하여, 그 추정계수의 통계적 유의성을 의미하는 P값이 0.05 미만인 제1 파생변수들을 중요변수로 선택할 수 있다. 여기에서, P값이 0.05라는 의미는, 유의수준 5%일 때, 추정계수가 통계적으로 유의하다는 것을 의미한다.

[126]

- [127] 한편, IV값은 하기 <수학식 2>를 통해 산출할 수 있다.
- [128] <수학식 2>
- [129]
$$IV = \sum_i (\% \text{ of Goods} - \% \text{ of Bads}) \times WOE_i$$
- [130]
- [131] 이때, '% of Goods'는 좋음(Good)으로 평가받은 집단에 대한 전체비율을 의미하고, '% of Bads'는 나쁨(Bad)으로 평가받은 집단에 대한 전체비율을 의미하고, WOE(Weights of Evidence; 이하, WOE)는 각각의 변수(i)에 대해 아래 <수학식 3>에 의해 정의된다.
- [132] <수학식 3>
- [133]
$$WOE_i = \ln \left(\frac{\% \text{ of Goods}_i}{\% \text{ of Bads}_i} \right)$$
- [134]
- [135] 이때, WOE는 나쁨(Bad)으로 평가받은 집단의 비율 대비 좋음(Good)으로 평가받은 집단의 비율에 대해서 자연로그를 취한 값을 의미한다.
- [136] 여기에서, WOE값이 큰 양의 값(positive value)을 가질수록 위험도가 낮은 것을 의미하고, WOE값이 큰 음의 값(negative value)을 가질수록 위험도가 높은 것을 의미할 수 있다. 또한, 본 발명에서는 IV값이 0.02 미만이면 종속변수(예를 들어, 사용자 신용의 부도여부)에 대한 설명력이 부족하다고 판단할 수 있다.
- [137]
- [138] 다시 S30 단계에서, 신용 평가 서버(100)는 중요변수를 선별하기 위해, 복수의 제1 파생변수의 P값과, IV값을 계산한다. 예를 들어, 신용 평가 서버(100)는 계산된 P값이 0.05 미만인지 여부와, IV값이 0.02 이상인지 여부를 판단할 수 있다. 다만, 이는 하나의 예시에 불과하며 본 발명이 이에 한정되는 것은 아니다.
- [139] 이어서, 신용 평가 서버(100)는 P값이 0.05 미만이거나, IV값이 0.02 이상인 제1 파생변수를 중요변수로 선별하여, 1단계 로지스틱 회귀모델에서 이용할 수 있다.
- [140]
- [141] 이어서, 신용 평가 서버(100)는 선별된 중요변수에 대하여, 동일한(또는, 유사한) 정보영역(F)에 속하는 변수들끼리 군집화(Grouping; 즉, 그룹화)할 수 있다(S40).
- [142] 이때, 신용 평가 서버(100)는 전문가에 의한 휴리스틱(heuristic) 방법을 이용하거나, 데이터 드라이븐(data-driven; 예를 들어, k-means clustering 등) 방법을 이용하여 군집화를 수행할 수 있다. 휴리스틱 방법 및 데이터 드라이븐 방법은 이미 공개되어 있으므로, 여기에서 자세한 설명은 생략하도록 한다. 또한, 신용 평가 서버(100)는 제1 파생변수의 기초가 되는 후보변수가 속한 카테고리 및 피처를 이용하여 정보영역(Filtration; 이하, F) 별로 군집화를 수행할 수 있다. 다만, 이는 군집화의 하나의 예시일 뿐 본 발명이 이에 한정되어 해석되는 것은 아니다.

- [143] 부가적으로, 본 발명에서 S40 단계는 생략되어 실시될 수 있음은 통상의 기술자에게 자명하다. 다만, 설명의 편의를 위하여 군집화를 수행한 것을 예로 들어 설명하도록 한다.
- [144]
- [145] 이어서, 신용 평가 서버(100)는 정보영역(F) 별로 군집화된 중요변수를 제1 단계 로지스틱 회귀분석 모델(이하, 1단계 모델)에 적용한다(S50).
- [146] 이때, 1단계 모델은 서로 동일한 정보영역(F)에 있는 중요변수들을 입력변수로 하고, 신용에 대한 우량/불량 정보(즉, 사용자 신용의 부도여부)를 종속변수로 하는 신용평가 모델을 의미한다. 1단계 모델에는 통상적인 로지스틱 회귀분석(logistic regression) 모델이 채용될 수 있다.
- [147]
- [148] 구체적으로, 본 발명에서 로지스틱 회귀분석은 다음과 같은 가정을 기초로 동작한다.
- [149] (가정1) 특정 정보영역(F)으로부터 생성된 변수들에 대하여, 부도여부 예측에 관련된 정보(f_k)를 파악하여 가중치로 이용한다. 이를 구현하기 위해 1단계 모델에 의한 추정계수를 가중치로 사용할 수 있다.
- [150] (가정2) 관측데이터(s_{ki})들은 부도예측에 관련된 정보(f_{ki})와 그 이외의 노이즈(e_i)로 구성된다.
- [151] <수학식 4>
- [152]
$$s_{ki} = f_{ki} + e_{ki}, \text{ where } e_{ki} \sim \text{iid}(\mu_k, \sigma_k^2)$$
- [153]
- [154] 이때, k는 정보영역을 의미하고, i는 개별 변수를 의미한다. s_{ki} 는 정보영역 k에 속한 변수 i의 변수값을 의미한다. iid(independently and identically distributed)는 '독립적이고 동질적으로 분포됨'을 의미하며, where절은 노이즈(e_{ki})의 값이 (평균(μ_k), 분산(σ_k^2))인 임의의 분포로부터 독립적이며 동질적으로 생성됨을 의미한다.
- [155]
- [156] 구체적으로, 1단계 모델에서, 정보영역(F)에 속하는 관측데이터(즉, 관측변수 또는 설명변수)(s_{ki})들을 이용하여 조건부 부도예측모델을 만들면 아래 <수학식 5>와 같다.
- [157] <수학식 5>
- [158]
$$E(Y | s_{k1}, \dots, s_{kj}) = E(Y) + \sum_{i=1}^j w_{ki} s_{ki} \text{ where, } w_{ki} = \frac{\text{Cov}(Y, s_{ki})}{\text{Var}(s_{ki})}$$
- [159]
- [160] 이때, E()는 기대연산자를 의미하고, Y는 종속변수(즉, 사용자 신용의 부도여부)를 의미하고, X(즉, $s_{k1}, \dots, s_{kj}$)는 설명변수(또는, 입력변수)를 의미하고, E(Y|X)는 X를 이용하여 Y를 조건부 예측함을 의미하고, E(Y)는 X라는 정보를 이용하지 않을 때의 무조건부 기대값(unconditional mean)을 의미하고, Cov는

공분산(covariance)으로 두 변수의 공통된 움직임을 측정한 값을 의미하고, Var는 분산(variance)으로 변수 평균과의 차이를 제공한 값들의 평균을 의미한다.

- [161] 또한, 가중치(w_{ki})는 회귀추정계수(이하, 추정계수)를 의미하며, 종속변수(Y) 변동에 대한 시그널(s_{ki} ; 설명변수 또는 관측변수)의 설명력을 의미한다. 가중치(w_{ki})는 <수학식 6>을 통해 표현할 수 있다.

[162] <수학식 6>

$$[163] \quad w_{ki} = \frac{Cov(Y, s_{ki})}{Var(s_{ki})}$$

$$= \frac{Cov(Y, f_{ki} + e_{ki})}{Var(f_{ki} + e_{ki})} = \frac{Cov(Y, f_{ki}) + Cov(Y, e_{ki})}{Var(f_{ki}) + Var(e_{ki})} = \frac{Cov(Y, f_{ki})}{Var(f_{ki}) + \sigma_k^2}$$

[164]

- [165] 이때, 시그널(s_{ki} ; 설명변수 또는 관측변수)에 내포된 노이즈의 분산(σ_k^2)이 클수록(즉, 시그널의 신뢰도는 낮을수록), 해당 시그널에 대한 가중치(w_{ki})는 작아지게 된다.

- [166] 정리하면, 1단계 모델의 가중치(w_{ki})는 설명변수(X)와 종속변수(Y) 간의 공분산을 설명변수의 분산으로 나눈 것을 의미한다. 이때, 가중치(w_{ki})는 회귀분석을 수행했을 때의 추정계수의 정의와 동일하다.

[167]

- [168] 이어서, 1단계 모델은 단계적 선택법(step-wise method)을 통하여, 1단계 모델에 적용할 최종변수를 선정할 수 있다. 이때, 단계적 선택법은 변수를 모델에 추가하고 제거하는 과정을 수행하되, 모든 변수에 대하여 순차적으로 수행하면서 모델의 설명력이 높아지는 방향으로 변수들을 선택해 나가는 방법을 의미한다.

[169]

- [170] 도 6을 참조하면, 1단계 모델에 적용되는 종속변수(Y)는 사용자 신용의 부도여부이고, 설명변수(X; 즉, 시그널(s_{ki}))는 S30 단계에서 선별했던 중요변수에 해당한다.

- [171] 예를 들어, 1단계 모델은 정보영역(F)이 등록(registration) 카테고리인 하나 이상의 중요변수에 대하여 단계적 선택법(step-wise method)을 수행한다. 신용 평가 서버(100)는 특정 카테고리에 속한 각각의 중요변수에 대하여, 가중치(w_{ki})가 상대적으로 높은 중요변수들을 최종변수로 선택하고, 가중치(w_{ki})가 상대적으로 낮은 중요변수들을 최종변수에서 제외할 수 있다.

- [172] 부가적으로, 1단계 모델이 단계적 선택법만 이용하는 것은 아니며, 1단계 모델은 전진 선택법(Feedforward selection) 또는 후진 제거법(Backward Elimination)을 이용하여 최종변수를 선정할 수 있음은 물론이다. 여기에서, 전진 선택법은 변수를 추가해가며 성능지표를 비교해가며 해당 변수를 추가할지 말지 여부를 선택하는 방법을 의미하고, 후진 제거법은 전진 선택법의 반대 개념으로 변수를 제거해가며 성능지표를 비교해가는 방법을 의미한다.

[173]

[174] 이어서, 신용 평가 서버(100)는 선택된 최종변수(즉, 선택된 중요변수/시그널(s_{ki}))들과, 추정된 회귀계수(즉, 가중치(w_{ki}))들을 각각 곱한 후 합산한 값을 2단계 모델(즉, 2단계 로지스틱 회귀분석 모델)의 설명변수로 사용할 제2 파생변수($S_1, S_2, \dots, S_k$)로 생성한다(S60). 여기에서, 제2 파생변수는 각 정보영역(F) 별로 구분된 최종변수들의 우량/불량 확률을 나타낸다.

[175] 이어서, 신용 평가 서버(100)는 생성된 제2 파생변수를 2단계 로지스틱 회귀분석 모델(이하, 2단계 모델)에 적용한다(S70).

[176] 이때, 2단계 모델은 1단계 모델의 출력값(즉, 제2 파생변수; 최종변수의 우량/불량 확률)을 입력변수로 하고, 사용자의 신용에 대한 우량/불량 정보(즉, 사용자 신용의 부도여부)를 종속변수로 하는 신용평가 모델을 의미한다. 마찬가지로, 2단계 모델에도 통상적인 로지스틱 회귀분석(logistic regression) 모델이 채용될 수 있다.

[177] 구체적으로, 2단계 모델에서, 제2 파생변수($S_1, S_2, \dots, S_k$)들을 이용하여 조건부 부도예측모델을 만들면 아래 <수학식 7>와 같다.

[178] <수학식 7>

[179] $E(Y|S_1, \dots, S_k) = E(Y) + \sum_{k=1}^K \beta_k S_k$ where, $\beta_k = \frac{Cov(Y, S_k)}{Var(S_k)}$

[180]

[181] 이때, $E()$ 는 기대연산자를 의미하고, Y 는 종속변수를 의미하고, X (즉, $S_1, \dots, S_k$)는 설명변수(또는, 입력변수)를 의미하고, $E(Y|X)$ 는 X 를 이용하여 Y 를 조건부 예측함을 의미하고, $E(Y)$ 는 X 라는 정보를 이용하지 않을 때의 무조건부 기대값(unconditional mean)을 의미하고, Cov 는 공분산(covariance)으로 두 변수의 공통된 움직임을 측정한 값을 의미하고, Var 는 분산(variance)으로 변수 평균과의 차이를 제공한 값들의 평균을 의미한다. 또한, 가중치(β_k)는 회귀추정계수(이하, 추정계수)를 의미하며, 종속변수(Y) 변동에 대한 시그널(S_k ; 설명변수, 입력변수 또는 관측변수)의 설명력을 의미한다. 가중치(β_k)는 전술한 <수학식 6>과 실질적으로 동일한 방법으로 표현될 수 있으며, 중복되는 설명은 생략하도록 한다.

[182] 또한, 2단계 모델은 1단계 모델과 마찬가지로, 단계적 선택법(step-wise method)을 통하여 2단계 모델에 적용할 최종변수들을 선택할 수 있다.

[183] 이어서, 신용 평가 서버(100)는 최종적으로 1단계 모델 및 2단계 모델에 적용될 최적의 최종변수들을 선별하기 위해, 복수의 사용자에게 대한 로그데이터를 기초로 전술한 S10 내지 S70 단계를 반복 수행한다.

[184] 이어서, 신용 평가 서버(100)는 전술한 단계를 통하여 최종변수가 특정된 1단계 모델(즉, 트레이닝된 1단계 모델) 및 최종변수가 특정된 2단계 모델(즉, 트레이닝된 2단계 모델)을 이용하여 새로운 사용자의 신용평가를 수행할 수 있다.

- [185]
- [186] 예를 들어, 도 7을 참조하면, 신용 평가 서버(100)는 로그데이터에 포함된 이벤트코드들을 미리 정해진 카테고리 및 미리 정해진 복수의 피처를 이용해 변수기초항목을 선정하고, 선정된 변수기초항목을 정보영역(F)(예를 들어, F1 내지 F9) 별로 군집화한다.
- [187] 이어서, 신용 평가 서버(100)는 트레이닝된 1단계 모델을 이용하여, 각 정보영역(F) 별로 구분된 최종변수(즉, 선택된 중요변수/시그널(s_{ki}))와, 최종변수의 미리 추정된 회귀계수(즉, 가중치(w_{ki}))의 곱을 산출함으로써, 제2 파생변수(예를 들어, S1 내지 S9)를 산출한다.
- [188] 이어서, 신용 평가 서버(100)는 트레이닝된 2단계 모델을 이용하여, 사용자의 신용에 대한 우불량에 대한 결과치로서 부도가 발생할 확률을 산출한다. 이때, 결과치는 0과 1 사이의 값으로 출력될 수 있으며, 신용 평가 서버(100)는 산출된 결과치를 기초로 사용자의 신용에 대한 우량 또는 불량 정도를 판단할 수 있다.
- [189] 이어서, 신용 평가 서버(100)에서 산출된 결과치는 금융 서버(200)에서 전달되어 사용자의 신용도에 따른 금융 서비스 제공여부를 결정하는데 이용될 수 있다.
- [190] 이하에서는, 본 발명의 몇몇 실시예에 따른 신용평가모델 운영 방법에 대해 살펴보도록 한다.
- [191]
- [192] 도 8은 본 발명의 몇몇 실시예에 따른 신용평가모델 운영 방법의 전처리 방법을 나타내는 순서도이다. 도 9는 본 발명의 몇몇 실시예에 따른 신용평가모델 운영 방법에서 2단계 로지스틱 회귀분석 모델을 이용한 신용평가방법에 관한 순서도이다.
- [193] 본 발명의 몇몇 실시예에 따른 신용평가모델 운영 방법은, 각각 단계가 신용 평가 서버(100) 및 금융 서버(200)에 분배되어 상호 보완적으로 실시되거나, 신용 평가 서버(100)에서만 실시될 수 있다. 다만, 이하에서는 설명의 편의를 위하여, 신용 평가 서버(100)에서 실시되는 것을 예로 들어 설명하도록 한다. 또한, 이하에서는, 전술한 내용과 중복되는 내용은 생략하고 차이점을 위주로 설명하도록 한다.
- [194] 도 8을 참조하면, 신용 평가 서버(100)는 우선 금융 서버(200)로부터 수신한 로그데이터를 기초로 변수기초항목을 선정한다(S110). 신용 평가 서버(100)는 로그데이터에 포함된 이벤트코드를 미리 정해진 카테고리 및 피처를 이용하여 구분함으로써, 이벤트코드에 대응되는 복수의 변수기초항목을 선정할 수 있다.
- [195] 이어서, 신용 평가 서버(100)는 선정된 복수의 변수기초항목에 대하여 단어빈도(TF)와 단어빈도-역문서빈도(TF-IDF)를 산출하여 복수의 후보변수를 생성한다(S120).

- [196] 예를 들어, 단어빈도(TF)는 단순빈도, 불린빈도, 증가빈도, 로그빈도를 포함하는 4개의 방법으로 산출될 수 있으며, 역문서빈도(IDF)는 특정 변수기초항목이 로그데이터 내에 공통적으로 나타나는지를 나타내는 값으로 전술한 <수학식 1>을 이용하여 산출될 수 있다. 단어빈도-역문서빈도(TF-IDF)는 4개의 단어빈도(TF)에 역문서빈도(IDF)를 곱하여 생성될 수 있다. 이를 통해, 4개의 단어빈도(TF)와 4개의 단어빈도-역문서빈도(TF-IDF)를 포함하는 8개의 후보변수(CV)가 생성될 수 있다. 다만, 이는 하나의 예시에 불과하며 단어빈도(TF)의 개수와 후보변수(CV)의 개수는 다르게 변형되어 실시될 수 있음은 물론이다.
- [197] 이어서, 신용 평가 서버(100)는 생성된 복수의 후보변수(CV)에 대하여 서로 다른 타임 윈도우(Time Windows) 및 서로 다른 산출방법을 각각 적용함으로써, 복수의 제1 파생변수를 생성한다(S130).
- [198] 예를 들어, 타임 윈도우(Time Windows)는 최근 1개월, 최근 3개월, 최근 6개월, 최근 9개월, 최근 12개월을 포함할 수 있고, 산출방법은 특정 후보변수의 선택된 타임 윈도우에 대한 평균, 합계, 최대값, 최소값을 포함할 수 있다. 따라서, 신용 평가 서버(100)는 각각의 후보변수(CV)에 대해 서로 다른 타임 윈도우와 서로 다른 산출방법의 조합을 통하여, 복수의 제1 파생변수를 생성할 수 있다.
- [199] 이어서, 신용 평가 서버(100)는 복수의 제1 파생변수에 대한 P값 및 IV값을 계산한다(S140). 여기에서, P값은 제1 파생변수를 대상으로 단변량 로지스틱 회귀분석을 수행함으로써 도출된 종속변수와 추정계수 간의 통계적 유의성을 나타내는 통계적 유의성을 의미한다. IV값은 전술한 <수학식 2>를 통해 산출될 수 있다.
- [200] 이어서, 신용 평가 서버(100)는 계산된 P값 및 IV값과 미리 정해진 기준치를 비교하여 중요변수를 선별한다(S150). 예를 들어, 신용 평가 서버(100)는 계산된 P값이 0.05 미만인지 여부와, IV값이 0.02 이상인지 여부를 판단함으로써, 중요변수를 선별할 수 있다. 이때, 신용 평가 서버(100)는 P값 또는 IV값만을 미리 정해진 기준치와 비교하여 중요변수를 선별할 수 있으며, P값 및 IV값이 모두 미리 정해진 기준치와 비교하여 정해진 조건을 만족하는 제1 파생변수만을 중요변수로 선별할 수 있음은 물론이다.
- [201] 이어서, 신용 평가 서버(100)는 선별된 중요변수를 정보영역(F) 별로 군집화할 수 있다(S160). 이때, 신용 평가 서버(100)는 전문가에 의한 휴리스틱(heuristic) 방법 또는 데이터 드라이븐(data-driven; 예를 들어, k-means clustering 등) 방법을 이용하여 정보영역(F) 별로 군집화를 수행할 수 있다. 또한, 신용 평가 서버(100)는 제1 파생변수의 기초가 되는 후보변수가 속한 카테고리 및 피처를 이용하여 정보영역(F) 별로 군집화를 수행할 수 있다.
- [202] 정보영역(F) 별로 군집화된 중요변수는 1단계 로지스틱 회귀분석 모델(즉, 1단계 모델)의 입력변수로 이용될 수 있다.
- [203]

- [204] 이어서, 도 9를 참조하면, 신용 평가 서버(100)는 정보영역(F) 별로 군집화된 중요변수를 제1 단계 모델에 적용한다(S210). 이때, 1단계 모델은 서로 동일한 정보영역(F)에 있는 중요변수들을 입력변수로 하고, 신용에 대한 우량/불량 정보(즉, 사용자 신용의 부도여부)를 종속변수로 하는 신용평가 모델을 의미한다. 1단계 모델에는 통상적인 로지스틱 회귀분석(logistic regression) 모델이 채용될 수 있다.
- [205] 이어서, 신용 평가 서버(100)는 단계적 선택법(step-wise method)을 통하여 1단계 모델에 적용할 제1 최종변수를 선정하고, 선정된 제1 최종변수에 대한 가중치를 산출한다(S220). 이때, 신용 평가 서버(100)는 특정 카테고리에 속한 각각의 중요변수에 대하여 가중치(w_{ki})를 계산하고, 가중치(w_{ki})가 상대적으로 높은 중요변수들을 제1 최종변수로 선택하고, 가중치(w_{ki})가 상대적으로 낮은 중요변수들을 제1 최종변수에서 제외할 수 있다. 중요변수에 대한 가중치(w_{ki})는 <수학식 6>를 이용하여 산출할 수 있다.
- [206] 이어서, 신용 평가 서버(100)는 선정된 제1 최종변수 및 가중치를 이용하여 제2 파생변수를 생성한다(S230). 여기에서, 제2 파생변수는 선정된 제1 최종변수와 이에 대한 가중치의 곱한 값이다. 제2 파생변수는 각 정보영역(F) 별로 구분된 제1 최종변수들의 신용불량확률(즉, 사용자 신용의 부도여부)을 나타낸다.
- [207] 이어서, 신용 평가 서버(100)는 제2 파생변수를 2단계 로지스틱 회귀분석 모델(이하, 2단계 모델)에 적용한다(S240). 이때, 2단계 모델은 1단계 모델의 출력값(즉, 제2 파생변수; 제1 최종변수의 우량/불량 확률)을 입력변수로 하고, 신용에 대한 우량/불량 정보(즉, 사용자 신용의 부도여부)를 종속변수로 하는 신용평가 모델을 의미한다. 마찬가지로, 2단계 모델에는 통상적인 로지스틱 회귀분석(logistic regression) 모델이 채용될 수 있다.
- [208] 이어서, 신용 평가 서버(100)는 단계적 선택법(step-wise method)을 통하여 2단계 모델에 적용할 제2 최종변수를 선정하고, 선정된 제2 최종변수에 대한 가중치를 산출한다(S250).
- [209] 이어서, 신용 평가 서버(100)는 선정된 제2 최종변수가 적용된 1단계 모델 및 2단계 모델을 이용하여, 새로운 사용자에게 대한 신용평가를 수행한다(S260).
- [210] 이때, 신용 평가 서버(100)는 새로운 사용자에게 대한 로그데이터를 금융 서버(200)로부터 수신하고, 수신된 로그데이터에서 선정된 제2 최종변수에 대응되는 변수기초항목을 추출한 뒤, 미리 생성된 1단계 모델 및 2단계 모델을 이용하여 사용자의 신용에 대한 우량/불량 여부를 산출할 수 있다. 이를 통해, 신용 평가 서버(100)는 새로운 사용자에게 대한 신용평가를 수행할 수 있다.
- [211]
- [212] 도 10은 본 발명의 몇몇 실시예에 따른 신용평가모델과 종래의 신용평가모델 간의 성능지표 차이를 나타내는 테이블이다.
- [213] 도 10을 참조하면, 2단계의 로지스틱 회귀분석 모델로 구성된 신용평가모델은, K-S 통계량(Kolmogorov-Smirnov Statistics) 및 AUROC(Area Under the Receiver

Operating Characteristics)의 성능지표를 이용하여 모델의 신뢰성을 평가할 수 있다. K-S 통계량과 AUROC에 대한 내용은 이미 공개되어 있으므로 여기에서 자세한 설명은 생략하도록 한다.

- [214] 일반적인 1단계 로지스틱 회귀분석(Logistic regression) 또는 머신러닝(machine learning) 모듈만을 이용하는 1단계 신용평가모델(즉, 베이스라인 모델(M1))은, 신용평가모델에 적용할 최종변수를 도출하기 위한 트레이닝 데이터셋(training dataset)를 입력하는 경우, AUROC가 61.4, K-S 통계량이 18.4로 측정되었다. 또한, 최종변수가 도출된 1단계 신용평가모델에 테스트 데이터셋(test dataset)를 입력하는 경우, AUROC가 57.5, K-S 통계량이 12.2로 측정되었다.
- [215] 반면, 전술한 본 발명의 몇몇 실시예에 따른 2단계 로지스틱 회귀분석 모델(즉, 2단계 모델(M2))은, 트레이닝 데이터셋(training dataset)를 입력하는 경우, AUROC가 64.8로 측정되어 베이스라인 모델(M1)에 비해 약 5%의 성능지표가 개선되었고, K-S 통계량은 22.4로 측정되어 베이스라인 모델(M1)에 비해 약 18%의 성능지표가 개선되었다.
- [216] 또한, 2단계 모델(M2)은, 테스트 데이터셋(test dataset)를 입력하는 경우, AUROC가 62.0으로 측정되어 베이스라인 모델(M1)에 비해 약 7%의 성능지표가 개선되었고, K-S 통계량은 18.5로 측정되어 베이스라인 모델(M1)에 비해 약 34%의 성능지표가 개선되었다.
- [217] 따라서, 본 발명의 몇몇 실시예에 따른 2단계 로지스틱 회귀분석 모델로 구성된 신용평가모델은, 종래의 1단계 신용평가모델에 비하여 성능지표가 유의미하게 개선됨을 확인할 수 있었다.
- [218]
- [219] 도 11은 본 발명의 몇몇 실시예에 따른 신용평가모델 운영 방법을 수행하는 시스템의 하드웨어 구현을 설명하기 위한 도면이다.
- [220] 도 11을 참조하면, 본 발명의 몇몇 실시예들에 따른 신용평가모델 운영 방법을 수행하는 신용 평가 서버(100)는 전자 장치(1000)로 구현될 수 있다. 전자 장치(1000)는 컨트롤러(1010, controller), 입출력 장치(1020, I/O), 메모리 장치(1030, memory device), 인터페이스(1040, interface) 및 버스(1050, bus)를 포함할 수 있다. 컨트롤러(1010), 입출력 장치(1020), 메모리 장치(1030) 및/또는 인터페이스(1040)는 버스(1050)를 통하여 서로 결합될 수 있다. 이때, 버스(1050)는 데이터들이 이동되는 통로(path)에 해당한다.
- [221] 구체적으로, 컨트롤러(1010)는 CPU(Central Processing Unit), MPU(Micro Processor Unit), MCU(Micro Controller Unit), GPU(Graphic Processing Unit), 마이크로프로세서, 디지털 신호 프로세서, 마이크로컨트롤러, 어플리케이션 프로세서(AP, application processor) 및 이들과 유사한 기능을 수행할 수 있는 논리 소자들 중에서 적어도 하나를 포함할 수 있다.
- [222] 입출력 장치(1020)는 키패드(keypad), 키보드, 터치스크린 및 디스플레이 장치 중 적어도 하나를 포함할 수 있다.

- [223] 메모리 장치(1030)는 데이터 및/또는 컴퓨터 프로그램 등을 저장할 수 있다.
- [224] 인터페이스(1040)는 통신 네트워크로 데이터를 전송하거나 통신 네트워크로부터 데이터를 수신하는 기능을 수행할 수 있다. 인터페이스(1040)는 유선 또는 무선 형태일 수 있다. 예컨대, 인터페이스(1040)는 안테나 또는 유무선 트랜시버 등을 포함할 수 있으며, 사용자 단말(300)로부터 수집된 다양한 데이터를 수신하거나, 메모리 장치(1030)에서 로드된 컴퓨터 프로그램을 실행하는 과정에서 발생하는 데이터를 사용자 단말(300)과 교환할 수 있다.
- [225] 한편, 도시하지 않았지만, 메모리 장치(1030)는 컨트롤러(1010)의 동작을 향상시키기 위한 동작 메모리로서, 고속의 디램 및/또는 에스램 등을 더 포함할 수도 있다. 메모리 장치(1030)는 내부에 컴퓨터 프로그램 또는 어플리케이션을 저장할 수 있다.
- [226] 본 발명의 실시예들에 따른 신용 평가 서버(100) 및 금융 서버(200)는 각각 복수의 전자 장치(1000)가 네트워크를 통해서 서로 연결되어 형성된 시스템일 수 있다. 이러한 경우에는 각각의 모듈 또는 모듈의 조합들이 전자 장치(1000)로 구현될 수 있다. 단, 본 실시예가 이에 제한되는 것은 아니다.
- [227] 추가적으로, 신용 평가 서버(100)는 워크스테이션(workstation), 데이터 센터, 인터넷 데이터 센터(internet data center(IDC)), DAS(direct attached storage) 시스템, SAN(storage area network) 시스템, NAS(network attached storage) 시스템, RAID(redundant array of inexpensive disks, or redundant array of independent disks) 시스템, 및 EDMS(Electronic Document Management) 시스템 중 적어도 하나로 구현될 수 있으나, 본 실시예가 이에 제한되는 것은 아니다.
- [228] 또한, 신용 평가 서버(100)는 네트워크를 통해서 금융 서버(200)에 데이터를 전송할 수 있다. 네트워크는 유선 인터넷 기술, 무선 인터넷 기술 및 근거리 통신 기술에 의한 네트워크를 포함할 수 있다. 유선 인터넷 기술은 예를 들어, 근거리 통신망(LAN, Local area network) 및 광역 통신망(WAN, wide area network) 중 적어도 하나를 포함할 수 있다.
- [229] 무선 인터넷 기술은 예를 들어, 무선랜(Wireless LAN: WLAN), DMNA(Digital Living Network Alliance), 와이브로(Wireless Broadband: Wibro), 와이맥스(World Interoperability for Microwave Access: Wimax), HSDPA(High Speed Downlink Packet Access), HSUPA(High Speed Uplink Packet Access), IEEE 802.16, 롱 텀 에볼루션(Long Term Evolution: LTE), LTE-A(Long Term Evolution-Advanced), 광대역 무선 이동 통신 서비스(Wireless Mobile Broadband Service: WMBS) 및 5G NR(New Radio) 기술 중 적어도 하나를 포함할 수 있다. 단, 본 실시예가 이에 제한되는 것은 아니다.
- [230] 근거리 통신 기술은 예를 들어, 블루투스(Bluetooth), RFID(Radio Frequency Identification), 적외선 통신(Infrared Data Association: IrDA), UWB(Ultra-Wideband), 지그비(ZigBee), 인접 자장 통신(Near Field Communication: NFC), 초음파 통신(Ultra Sound Communication: USC), 가시광

통신(Visible Light Communication: VLC), 와이 파이(Wi-Fi), 와이 파이 다이렉트(Wi-Fi Direct), 5G NR (New Radio) 중 적어도 하나를 포함할 수 있다. 단, 본 실시예가 이에 제한되는 것은 아니다.

- [231] 네트워크를 통해서 통신하는 신용 평가 서버(100)는 이동통신을 위한 기술표준 및 표준 통신 방식을 준수할 수 있다. 예를 들어, 표준 통신 방식은 GSM(Global System for Mobile communication), CDMA(Code Division Multi Access), CDMA2000(Code Division Multi Access 2000), EV-DO(Enhanced Voice-Data Optimized or Enhanced Voice-Data Only), WCDMA(Wideband CDMA), HSDPA(High Speed Downlink Packet Access), HSUPA(High Speed Uplink Packet Access), LTE(Long Term Evolution), LTEA(Long Term Evolution-Advanced) 및 5G NR(New Radio) 중 적어도 하나를 포함할 수 있다. 단, 본 실시예가 이에 제한되는 것은 아니다.
- [232] 이상의 설명은 본 실시예의 기술 사상을 예시적으로 설명한 것에 불과한 것으로서, 본 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 실시예의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다. 따라서, 본 실시예들은 본 실시예의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

청구범위

- [청구항 1] 금융서버와 연계된 신용 평가 서버에서 수행되는 신용평가모델 운영 방법에 있어서,
 사용자의 로그데이터를 수신하고, 상기 로그데이터에 포함된 변수기초항목을 선정하는 단계;
 상기 로그데이터 내에서 상기 변수기초항목의 빈도(Frequency)를 산출함으로써 후보변수를 생성하는 단계;
 상기 후보변수에 서로 다른 타임 윈도우(time-window) 또는 서로 다른 산출방법을 적용하여 복수의 제1 파생변수를 생성하는 단계;
 상기 복수의 제1 파생변수와 관련된 값과 미리 정한 기준치를 비교하여 중요변수를 선별하는 단계;
 상기 중요변수를 입력변수로 하고, 상기 사용자의 신용에 대한 정보를 종속변수로 하는 제1 단계 모델을 도출하는 단계;
 상기 중요변수 중에서 상기 제1 단계 모델에 적용할 제1 최종변수를 선정하고, 상기 제1 최종변수에 대한 제1 가중치를 산출하는 단계;
 상기 제1 최종변수 및 상기 제1 가중치를 이용하여 제2 파생변수를 생성하는 단계;
 상기 제2 파생변수를 입력변수로 하고, 상기 사용자의 신용에 대한 정보를 종속변수로 하는 제2 단계 모델을 도출하는 단계; 및
 상기 제1 파생변수 중에서 상기 제2 단계 모델에 적용할 제2 최종변수를 선정하고, 상기 제2 최종변수에 대한 제2 가중치를 산출하는 단계를 포함하는
 신용평가모델 운영 방법.
- [청구항 2] 제1 항에 있어서,
 상기 변수기초항목을 선정하는 단계는,
 상기 로그데이터에 포함된 이벤트코드를 미리 정해진 카테고리를 이용하여 구분하고,
 상기 카테고리에 속한 이벤트코드를 미리 정해진 복수의 피처를 이용하여 구분함으로써, 상기 이벤트코드에 대응되는 변수기초항목을 선정하는 것을 포함하는 신용평가모델 운영 방법.
- [청구항 3] 제1 항에 있어서,
 상기 후보변수를 생성하는 단계는,
 상기 변수기초항목의 단어빈도(TF)와 단어빈도-역문서빈도(TF-IDF)를 산출하여 상기 후보변수를 생성하는 것을 포함하되,
 상기 단어빈도(TF)는, 단순빈도, 불린빈도, 증가빈도, 또는 로그빈도를 이용하여 산출되고,
 상기 단어빈도-역문서빈도(TF-IDF)는, 단어빈도(TF)에

역문서빈도(IDF)를 곱함으로써 산출되는 것을 포함하는 신용평가모델 운영 방법.

- [청구항 4] 제1 항에 있어서,
 상기 복수의 제1 파생변수를 생성하는 단계는,
 상기 후보변수에 서로 다른 크기의 복수의 타임 윈도우 중 어느 하나와,
 복수의 산출방법 중 어느 하나를 이용하여 상기 제1 파생변수를 생성하는 것을 포함하되,
 상기 타임 윈도우는, 서로 다른 기간에 대해 설정될 수 있고,
 상기 산출방법은, 평균, 합계, 최대값, 및 최소값을 포함하는
 신용평가모델 운영 방법.

- [청구항 5] 제1 항에 있어서,
 상기 중요변수를 선별하는 단계는,
 상기 복수의 제1 파생변수 중, 단변량 로지스틱 회귀분석의 수행을 통한 P값(P-value)이 미리 정해진 기준값보다 작거나,
 상기 복수의 제1 파생변수 중, IV값이 미리 정해진 기준치보다 큰 제1 파생변수를 상기 중요변수로 선별하는 것을 포함하되,
 상기 IV값은, 하기 <수학식>에 의해 도출되는 신용평가모델 운영 방법.
 <수학식>

$$IV = \sum_i (\% \text{ of Goods} - \% \text{ of Bads}) \times WOE_i$$

여기에서, '% of Goods'는 좋음(Good)으로 평가받은 집단에 대한 전체비율을 의미하고, '% of Bads'는 나쁨(Bad)으로 평가받은 집단에 대한 전체비율을 의미하고, WOE(Weights of Evidence; 이하, WOE)는 나쁨(Bad)으로 평가받은 집단의 비율 대비 좋음(Good)으로 평가받은 집단 비율의 값에 자연로그를 취한 값을 의미한다.

- [청구항 6] 제1 항에 있어서,
 상기 선별된 중요변수에 대하여 동일한 정보영역(F)에 속하는 변수들끼리 군집화(Grouping)하는 단계를 더 포함하되,
 상기 제1 단계 모델을 도출하는 단계는, 특정 정보영역(F)에 포함된 상기 중요변수를 대상으로 상기 제1 최종변수를 선정하는 것을 포함하는
 신용평가모델 운영 방법.

- [청구항 7] 제1 항에 있어서,
 상기 제1 단계 모델 및 상기 제2 단계 모델은, 로지스틱 회귀분석 모델(logistic regression model)로 구성되는 신용평가모델 운영 방법.

- [청구항 8] 제7 항에 있어서,
 상기 제1 단계 모델은, 단계적 선택법을 이용하여, 상기 중요변수 중에서 상기 제1 단계 모델에 적용할 상기 제1 최종변수를 선정하고,
 상기 제2 단계 모델은, 단계적 선택법을 이용하여, 상기 제2 파생변수 중에서 상기 제2 단계 모델에 적용할 상기 제2 최종변수를 선정하는 것을

- 포함하는 신용평가모델 운영 방법.
- [청구항 9] 제1 항에 있어서,
 상기 제1 최종변수가 적용된 상기 제1 단계 모델과, 상기 제2 최종변수가 적용된 상기 제2 단계 모델을 이용하여, 새로운 사용자의 로그데이터를 기초로 상기 새로운 사용자의 신용평가를 수행하는 단계를 더 포함하는 신용평가모델 운영 방법.
- [청구항 10] 금융서버와 연계된 신용 평가 서버에서 수행되는 신용평가모델 운영 방법에 있어서,
 사용자의 로그데이터를 수신하고, 상기 로그데이터에 포함된 이벤트코드에 대한 빈도와, 상기 빈도에 대한 하나 이상의 전처리 과정을 통하여 중요변수를 선별하는 단계;
 상기 중요변수를 입력변수로 하고, 상기 사용자의 신용에 대한 정보를 종속변수로 하는 제1 단계 로지스틱 회귀분석 모델을 도출하는 단계;
 상기 중요변수 중에서 상기 제1 단계 모델에 적용할 제1 최종변수를 선정하고, 상기 제1 최종변수에 대한 제1 가중치를 산출하는 단계;
 상기 제1 최종변수 및 상기 제1 가중치를 이용하여 파생변수를 생성하는 단계;
 상기 파생변수를 입력변수로 하고, 상기 사용자의 신용에 대한 정보를 종속변수로 하는 제2 단계 로지스틱 회귀분석 모델을 도출하는 단계; 및
 상기 파생변수 중에서 상기 제2 단계 모델에 적용할 제2 최종변수를 선정하고, 상기 제2 최종변수에 대한 제2 가중치를 산출하는 단계를 포함하는
 신용평가모델 운영 방법.
- [청구항 11] 제10 항에 있어서,
 상기 제1 단계 모델은, 단계적 선택법을 이용하여, 상기 중요변수 중에서 상기 제1 단계 모델에 적용할 상기 제1 최종변수를 선정하고,
 상기 제2 단계 모델은, 단계적 선택법을 이용하여, 상기 파생변수 중에서 상기 제2 단계 모델에 적용할 상기 제2 최종변수를 선정하는 것을 포함하는 신용평가모델 운영 방법.
- [청구항 12] 제10 항에 있어서,
 상기 중요변수를 선별하는 단계는,
 상기 복수의 제1 파생변수 중, 단변량 로지스틱 회귀분석의 수행을 통한 P 값(P -value)이 미리 정해진 기준값보다 작거나,
 상기 복수의 제1 파생변수 중, IV 값이 미리 정해진 기준치보다 큰 제1 파생변수를 상기 중요변수로 선별하는 것을 포함하되,
 상기 IV 값은, 하기 <수학식>에 의해 도출되는 신용평가모델 운영 방법.
 <수학식>

$$IV = \sum_i (\% \text{ of Goods} - \% \text{ of Bads}) \times WOE_i$$

여기에서, '% of Goods'는 좋음(Good)으로 평가받은 집단에 대한 전체비율을 의미하고, '% of Bads'는 나쁨(Bad)으로 평가받은 집단에 대한 전체비율을 의미하고, WOE(Weights of Evidence; 이하, WOE)는 나쁨(Bad)으로 평가받은 집단에 대한 좋음(Good)으로 평가받은 모집단의 자연로그를 의미한다.

- [청구항 13] 제10 항에 있어서,
상기 선별된 중요변수에 대하여 동일한 정보영역(F)에 속하는 변수들끼리 군집화(Grouping)하는 단계를 더 포함하되,
상기 제1 단계 모델을 도출하는 단계는, 특정 정보영역(F)에 포함된 상기 중요변수를 대상으로 상기 제1 최종변수를 선정하는 것을 포함하는 신용평가모델 운영 방법.
- [청구항 14] 제10 항에 있어서,
상기 제1 최종변수가 적용된 상기 제1 단계 모델과, 상기 제2 최종변수가 적용된 상기 제2 단계 모델을 이용하여, 새로운 사용자의 로그데이터를 기초로 상기 새로운 사용자의 신용평가를 수행하는 단계를 더 포함하는 신용평가모델 운영 방법.
- [청구항 15] 프로세서;
상기 프로세서에 의해 실행되는 컴퓨터 프로그램을 로드하는 메모리; 및
상기 컴퓨터 프로그램의 실행과정에서 발생하는 데이터를 사용자 단말과 교환하는 인터페이스를 포함하되,
상기 컴퓨터 프로그램은,
상기 사용자 단말로부터 사용자의 로그데이터를 수신하고, 상기 로그데이터에 포함된 이벤트코드에 대한 빈도와, 상기 빈도에 대한 하나 이상의 전처리 과정을 통하여 중요변수를 선별하는 단계;
상기 중요변수를 입력변수로 하고, 상기 사용자의 신용에 대한 정보를 종속변수로 하는 제1 단계 로지스틱 회귀분석 모델을 도출하는 단계;
상기 중요변수 중에서 상기 제1 단계 모델에 적용할 제1 최종변수를 선정하고, 상기 제1 최종변수에 대한 제1 가중치를 산출하는 단계;
상기 제1 최종변수 및 상기 제1 가중치를 이용하여 파생변수를 생성하는 단계;
상기 파생변수를 입력변수로 하고, 상기 사용자의 신용에 대한 정보를 종속변수로 하는 제2 단계 로지스틱 회귀분석 모델을 도출하는 단계; 및
상기 파생변수 중에서 상기 제2 단계 모델에 적용할 제2 최종변수를 선정하고, 상기 제2 최종변수에 대한 제2 가중치를 산출하는 단계를 포함하는
신용 평가 서버.
- [청구항 16] 제15 항에 있어서,
상기 제1 단계 모델은, 단계적 선택법을 이용하여, 상기 중요변수 중에서

상기 제1 단계 모델에 적용할 상기 제1 최종변수를 선정하고,
상기 제2 단계 모델은, 단계적 선택법을 이용하여, 상기 파생변수 중에서
상기 제2 단계 모델에 적용할 상기 제2 최종변수를 선정하는 것을
포함하는 신용 평가 서버.

- [청구항 17] 제15 항에 있어서,
상기 중요변수를 선별하는 단계는,
상기 복수의 제1 파생변수 중, 단변량 로지스틱 회귀분석의 수행을 통한
P값(P-value)이 미리 정해진 기준값보다 작거나,
상기 복수의 제1 파생변수 중, IV값이 미리 정해진 기준치보다 큰 제1
파생변수를 상기 중요변수로 선별하는 것을 포함하되,
상기 IV값은, 하기 <수학식>에 의해 도출되는 신용 평가 서버.
<수학식>

$$IV = \sum_i (\% \text{ of Goods} - \% \text{ of Bads}) \times WOE_i$$

여기에서, '% of Goods'는 좋음(Good)으로 평가받은 집단에 대한
전체비율을 의미하고, '% of Bads'는 나쁨(Bad)으로 평가받은 집단에 대한
전체비율을 의미하고, WOE(Weights of Evidence; 이하, WOE)는
나쁨(Bad)으로 평가받은 집단에 대한 좋음(Good)으로 평가받은 모집단의
자연로그를 의미한다.

- [청구항 18] 제15 항에 있어서,
상기 선별된 중요변수에 대하여 동일한 정보영역(F)에 속하는
변수들끼리 군집화(Grouping)하는 단계를 더 포함하되,
상기 제1 단계 모델을 도출하는 단계는, 특정 정보영역(F)에 포함된 상기
중요변수를 대상으로 상기 제1 최종변수를 선정하는 것을 포함하는 신용
평가 서버.

- [청구항 19] 제15 항에 있어서,
상기 제1 최종변수가 적용된 상기 제1 단계 모델과, 상기 제2 최종변수가
적용된 상기 제2 단계 모델을 이용하여, 새로운 사용자의 로그데이터를
기초로 상기 새로운 사용자의 신용평가를 수행하는 단계를 더 포함하는
신용 평가 서버.

- [청구항 20] 제1 항 내지 제14 항 중 어느 하나에 기재된 방법을 실행할 수 있는
프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.