한국빅데이터학회
제2권 제2호, 2017, pp. 17-31

공공데이터 기반 고용보험 가입 예측 모델 개발 연구*

A Development on a Predictive Model for Buying Unemployment Insurance Program Based on Public Data

조민수[1,2] · 김도현[1] · 송민석[1†] · 김광용[3] · 정충식[3] · 김기대[3]

포항공과대학교(POSTECH) 산업경영공학과[1]
울산과학기술원(UNIST) 경영공학과[2]
근로복지공단[3]

요 약

빅데이터의 중요성이 증가함에 따라 공공기관에서는 다양한 빅데이터 관련 인프라를 제공하고 있으며, 그 중 하나가 공공데이터이다. 공공데이터 기반의 다양한 활용 사례가 공유되고 있으며, 공공기관에서도 데이터 기반의 모델을 통해 공공의 문제를 해결하려는 움직임을 보이고 있다. 대표적으로 사회 보험 중 하나인 고용보험 케이스가 있다. 고용보험은 근로자의 권익 보호를 위해 근로자를 고용한 모든 사업주가 필수적으로 가입하여야 하는 보험이지만 가입누락의 경우가 많다. 가입누락을 막기 위한 데이터 기반의 접근이 필요하지만, 분산된 형태의 공공데이터, 수집 시기의 차이로 인해 데이터 통합이 어렵고, 체계적인 방법론이 부재한 상황이다. 본 논문에서는 공공데이터를 기반의 고용보험 가입 예측을 위한 모델 도출 방법론을 제시하고자 한다. 본 방법론은 데이터 수집, 데이터 통합 및 전처리, 데이터 탐색 및 이력 데이터 분석, 예측 모델 도출을 포함하며, 프로세스 마이닝 및 데이터 마이닝을 활용한다. 또한, 사례 연구를 통해 본 방법론의 유효성을 검증한다.

■ 중심어 : 공공데이터, 빅데이터, 고용보험, 데이터 마이닝, 프로세스 마이닝

Abstract

With the development of the big data environment, public institutions also have been providing big data infrastructures. Public data is one of the typical examples, and numerous applications using public data have been provided. One of the cases is related to the employment insurance. All employers have to make contracts for the employment insurance for all employees to protect the rights. However, there are abundant cases where employers avoid to buy insurances. To overcome these challenges, a data-driven approach is needed; however, there are lacks of methodologies to integrate, manage, and analyze the public data. In this paper, we propose a methodology to build a predictive model for identifying whether employers have made the contracts of employment insurance based on public data. The methodology includes collection, integration, pre-processing, analysis of data and generating prediction models based on process mining and data mining techniques. Also, we verify the methodology with case studies.

■ Keyword : Public Data, Big Data, Unemployment Insurance System, Data Mining, Process Mining

2017년 12월 1일 접수; 2017년 12월 19일 수정본 접수; 2017년 12월 26일 게재 확정
* 본 연구는 한국연구재단의 지원을 받아 수행되었습니다(과제 번호 NRF-2014K1A3A7A03073707).
† 교신저자 mssong@postech.ac.kr

Ⅰ. 서 론

‘정보화 시대의 원유’라 불리는 빅데이터는 모든 산업 내 문제 해결 및 의사 결정에 활용된다고 해도 과언이 아니다[5]. 빅데이터 시장의 규모는 나날이 확대되고 있고, 4차 산업혁명, 인공지능 등 데이터 분석이 중심이 되는 ICT 융합 분야 기술 또한 점점 늘어나고 있는 추세이다[5]. 이에 따라, 정부를 포함한 공공기관에서도 이러한 빅데이터 환경에 관한 인프라를 제공하고 있으며, 그 중 하나가 바로 대규모 공공데이터의 제공이다[3]. 공공기관에서 생성 또는 수집하여 관리하고 있는 공공데이터를 다양한 경로로 제공하여 많은 연구자 및 데이터 분석 전문가가 활용할 수 있는 가능성을 열어 주었고, 실제로 많은 국내외 공공데이터 활용 사례가 공유되고 있다. 특히, 공공데이터는 공공기관 내에서 겪고 있는 문제를 해결하기 위한 열쇠가 되기도 한다. 주요 문제 중 하나는 준법 감시에 대한 것이며, 데이터를 기반한 모델을 통해 해결 가능하다.

대표적으로 사회 보험 중 하나인 고용보험의 경우, 근로자의 권익 보호를 위해 근로자를 한 명 이상 고용한 모든 사업장의 사업주가 필수적으로 가입하여야 하는 보험이지만 실제로 그렇지 않는 경우가 많다. 따라서, 고용보험을 가입해야하는 사업장임에도 불구하고 미가입한 사업장, 즉 가입누락을 예측하기 위한 모델이 필요하다. 하지만, 공공데이터가 각 공공기관마다 수집되기 때문에 분산되어 있고, 수집 시점이 다양하여 통합의 어려움이 있다. 또한, 데이터 수집부터 예측 모델 도출까지의 체계적인 방법론이 부재한 상황이다.

본 논문에서는 공공기관에 수집된 데이터를 바탕으로 고용보험 가입 예측을 위한 모델 도출 방법론을 제시하고자 한다. 본 방법론은 데이터 수집, 데이터 통합 및 전처리, Exploratory Data Analysis & 프로세스 마이닝 기반 이력 데이터 분석, 예측 모델 도출을 포함한다. 프로세스 마이닝의 프로세스 모델 도출 알고리즘, 패턴 분석 방법을 통해 데이터의 특징 파악 및 지식을 도출하고, 분류 알고리즘인 Decision Tree, Random Forest, Support Vector Machine, Neural Network를 통해 고용보험 가입 예측 모델을 도출한다. 본 연구는 공공데이터의 활용성 증대, 데이터 기반의 고용보험 가입에 대한 준법 감시, 다양한 분야로의 확장 등의 주요한 기여점을 가질 것으로 판단된다.

본 논문의 이후 구성은 다음과 같다. 제Ⅱ장에서는 방법론 내에서 활용하는 주요 알고리즘을 소개하고, 제Ⅲ장에서는 고용보험 가입 예측 모델 개발 방법론을 제안한다. 제Ⅳ장에서는 사례연구를 통해 방법론을 검증하고, 제Ⅴ장에서는 본 연구의 기여점, 한계점을 포함한 시사점을 제시한다. 마지막으로, 제Ⅵ장에서는 결론 및 추후연구를 제시한다.

Ⅱ. 이론적 배경

본 장에서는 고용보험 가입 예측 모델 개발 방법론 소개에 앞서, 방법론 내에서 주요하게 활용되는 데이터 분석 기법에 대해 소개하고자 한다. 두 개의 항목으로 구성되어 있으며, 먼저 본 방법론에서 활용되는 네 가지 데이터 기반 분류 모델에 대해 소개한다. 이 후에, 프로세스 마이닝 연구에 대해 간단하게 소개한다.

2.1 데이터 기반 분류 모델 연구

본 연구에서 활용된 데이터 기반 분류 모델은 Decision Tree[2, 6], Random Forest[1, 4], Support Vector Machine[10, 13], Neural Network[7, 8] 네 가지이다.

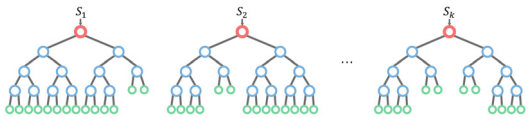
Decision Tree는 가장 간단한 데이터 분류 기법 중 하나로, 일련의 규칙을 기반으로 데이터

를 세분화하는 것이다[2, 6]. 분석 결과는 나무형태의 모델로 도출되며, 해당 모델은 반복적인 분할을 통해 예측으로 활용될 수 있는 일련의 규칙을 생성하게 된다. 가장 대표적인 Decision Tree 방법에는 C4.5, C5.0, Classification and Regression Trees(CART), Chi-squared Automatic Interaction Detection(CHAID) 등이 있으며, 각각 데이터 아웃풋, 인풋의 형태 차이, 분할 측정기준, 분할 방법 등의 따른 차이가 존재한다. 분할 측정 기준과 관련하여, 대표적인 방법은 Information Gain(IG), Gini Impurity(GI)이며, 이에 대한 수식은 다음과 같다. 먼저, 식 (1)에 정의된 Information Gain은 전체 개체 S를 attribute a를 바탕으로 분할하였을 때, Entropy(e.g., $H(S)=-\sum_{i=1}^{C} p_i \log_2 p_i$, where C: 클래스의 종류)가 얼마나 감소하는지를 나타낸다. 식 (2)에 정의된 Gini Impurity는 데이터에서 임의로 두 개체를 선택했을 때 동일 클래스로 구분될 확률을 나타낸다. 즉, 그룹 내의 동일 클래스를 가진 개체가 많을수록 수치가 1에 가까워진다. Decision Tree의 가장 큰 장점은 해석 가능한 규칙 혹은 논리를 제공한다는 점이며, 또한, 복잡한 계산과정 없이 쉽게 예측 모델을 생성한다는 장점을 가진다.

$$IG(S,a)=H(S)-H(S|a) \quad (1)$$

$$GI(p)=1-\sum_{i=1}^{C} p_i^2 \quad (2)$$

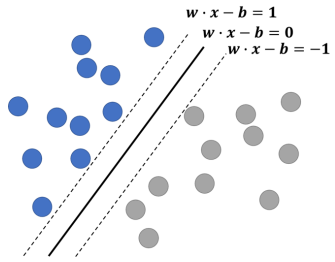
Random Forest는 <그림 1>과 같이 다수의 결정 나무를 기반으로 데이터를 분류하는 방법으로, Decision Tree부터 Bagging까지의 이론적 배경을 포함하고 있다[1, 4]. Random Forest는 여러 Decision Tree에서 도출된 결과를 기반으로 가장 빈번하게 발생한 결과 값을 최종 결과로 판단하는 방식을 가진다. 즉, 개별 Decision Tree로부터 도출된 결과가 완벽한 모델이 아니더라도 통합적인 관점에서는 확률적으로 올바른 결과가 나타난다는 개념이다. Random Forest의 장점은 Decision Tree에 비해 다양성이 극대화되고 이를 통해 예측 결과를 종합함으로써 예측력 및 안정성이 우수해진다는 점이다.



〈그림 1〉 Random Forest 도식화

Support Vector Machine은 <그림 2>와 같이 서로 다른 클래스를 가지는 개체 간 간격이 최대가 되는 초평면을 파악하여 데이터를 분류하는 모델이다[10, 13]. 초 평면의 마진($\frac{2}{\|w\|}$), 즉 두 서포트 벡터의 거리를 최대화하기 위한 방법은 $\|w\|$를 최소화하는 것이며, 이와 관련하여 식 (3)과 같이 최적화 문제로 표현가능하다.

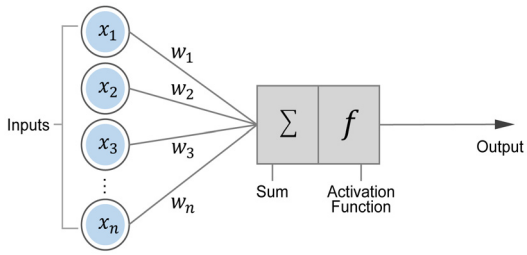
$$\begin{aligned} &Minimize\ \|w\| \\ &s.t.\ \ y_i(w \bullet x_i - b) \geq 1\ \forall i=1,\ \cdots,\ n \end{aligned} \quad (3)$$



〈그림 2〉 Support Vector Machine 알고리즘

Neural Network는 인간의 뇌와 연관된 신경 네트워크 구조에 대한 모델링을 기초로 개발되었으며, <그림 3>과 같은 기본 구조를 가진다[7, 8]. 각 뉴런(x_1, x_2, x_3, x_4)은 정보를 전달하기 위해 가중치와 결합되어 Activation Function을 통해 새로운 값으로 도출된다. 해당 구조는 식 (4)로 구체화될 수 있다. Neural Network는 예측력이 좋다는 강점을 가지고 있지만, Hidden Layer로 인해 결과 해석이 불가능하다는 단점을 포함한다.

$$y = f(\sum_i w_i x_i) \quad (4)$$



〈그림 3〉 신경망을 이용한 정보 전달 과정

2.2 프로세스 마이닝 연구

프로세스 마이닝은 정보시스템에 기록된 이벤트 로그로부터 프로세스 관점에서의 의미있는 지식을 도출하는 연구 분야이다[12]. 프로세스 마이닝은 비즈니스 프로세스 관리 연구(e.g., 프로세스 모델링 및 분석)와 데이터 과학 연구(e.g., 데이터 마이닝, 머신 러닝 등)를 연결하는 주요한 역할을 한다. 프로세스 마이닝은 Discovery (프로세스 모델 도출), Conformance(적합도 검사), Enhancement(프로세스 모델 확장)의 세 가지 주요 방향으로 구성되어 있다[11]. 본 연구 방법론에서는 Discovery를 중심으로 프로세스 마이닝을 활용하고 있으며, 이와 관련된 몇 가지 주요 알고리즘을 소개하고자 한다.

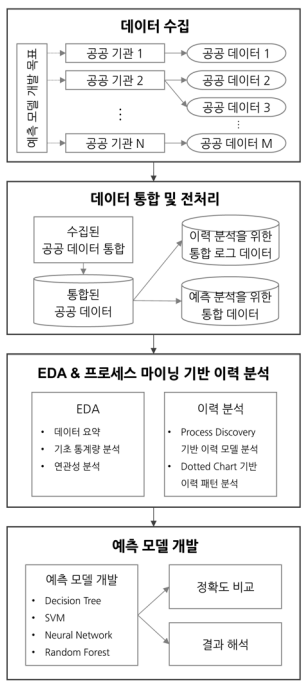
α -algorithm은 로그로부터 모델(Petri-net)을 도출하는 대표적인 방법이다[12]. transition(작업) 확인, 시작/종료 transition 확인, 인과 관계가 있는 transition 사이의 place 추가, 비최대 쌍 제거, 시작/종료 place 삽입, flow 관계 생성 및 모든 항목 조합의 8가지 단계로 구성되어 있다. 하지만, 본 알고리즘은 노이즈 및 불완전성의 한계점을 가진다.

이러한 한계점의 해결을 위해 나타난 개선된 형태의 대표적인 알고리즘 중 하나는 Heuristic Algorithm이다. Heuristic Algorithm은 작업 간 causal dependency를 기반으로 프로세스 모델을 도출한다[14]. 작업 간 causal dependency를 파악하는 방법은 식 (5)와 같다. $|a \Rightarrow_L b|$의 값이 +1에 근접하면 a작업에서부터 b작업으로 갈 확률이 높으며, -1에 근접하면 반대의 경우를 나타낸다. 모든 작업 관계에 대해 causal dependency를 계산하고 특정 기준치보다 높을 경우 이를 연결함으로서 프로세스 모델을 도출한다.

$$|a \Rightarrow_L b| = \begin{cases} \dfrac{|a \gg b| - |b \gg a|}{|a \gg b| + |b \gg a| + 1} & \text{if } a \neq b \\ \dfrac{|a \gg b|}{|a \gg b| + 1} & \text{if } a = b \end{cases} \quad (5)$$

Ⅲ. 고용보험 가입 예측 모델 개발 방법론

본 논문에서 제시하는 고용보험 가입 예측 모델 개발 방법론은 데이터 수집, 데이터 통합 및 구성, 탐색적 데이터 분석(EDA) & 프로세스 마이닝 기반 이력 분석, 예측 모델 개발을 포함하여 총 네 가지 단계로 구성되어 있다. <그림 4>는 도식화된 고용보험 가입 예측 모델 방법론을 나타낸다.



〈그림 4〉 고용보험 가입 예측 모델 개발 방법론

3.1 데이터 수집

고용보험 가입 예측 모델을 도출하기 위한 가장 첫 단계는 관련 데이터를 수집하는 것이다. 공공데이터는 특정 조직 내 데이터와 다르게, 여러 공공기관에 데이터가 각각 흩어져 있다. 따라서, 예측 모델 개발 목표에 맞게 다수의 공공데이터를 수집해야한다. 이 때, 두 번째 단계의 데이터 통합을 위해서 각 공공데이터 내 통합 가능한 식별자를 기반으로 수집할 필요가 있다. 또한, 통합 데이터 분석을 위해 수집 당시 시기를 고려하여 특정 기간을 기준으로 수집해야 한다. 식별자 기반 및 수집 시기를 고려하지 않는 데이터 수집을 수행할 경우, 통합 데이터 기반의 모델 생성이 불가능할 수 있다.

3.2 데이터 통합 및 전처리

분석 방법론 내 두 번째 단계는 식별자를 기반으로 통합된 데이터를 생성하고 프로세스 마이닝 기반 이력 분석 및 예측 모델 개발용 데이터 생성 즉, 전처리를 수행하는 것이다. 프로세스 마이닝 기반의 이력 분석과 예측 분석을 위한 데이터의 형태는 서로 다르다. 먼저, 이력 분석의 경우, 이벤트 로그 형태의 데이터가 필요하다. 이벤트 로그는 프로세스 마이닝 분석을 위한 데이터 형태이며, 정보시스템으로부터 추출 가능한 데이터이다[11]. 이벤트 로그는 프로세스 인스턴스(케이스)와 관련된 이벤트들의 집합체이며, 각 이벤트는 작업, 작업자, 시간기록 등의 속성을 갖는다. 본 연구에서는 각 공공데이터 내 포함된 이벤트를 모두 통합하여 이벤트 로그로 정의한다. <표 1>은 공공데이터와 관련된 이벤트 로그의 예시를 나타낸다. 수집된 각 공공데이터 내 정보와 이와 관련된 작업, 작업자, 시간 기록이 한 개의 이벤트를 구성한다.

예측 분석을 위한 데이터는 일반적인 데이터 마이닝을 위한 데이터 형태를 요구한다[15]. 각 식별자의 보험 가입 여부가 종속 변수를 구성하며, 이외에 식별자와 연관된 예측을 위한 기타 변수들이 독립변수를 구성한다. 예를 들면, 식별자(사업장)의 세금 납부 여부, 해당 사업장의 건축물 정보 등이 독립변수를 구성할 수 있다.

〈표 1〉 이벤트 로그 예시

이벤트 No.	케이스	작업	작업자	시간기록
1	C1	사업개시	R1	2016-01-01
2	C1	보험가입	R2	2016-01-04
3	C1	업종변경	R3	2016-11-01
4	C1	보험소멸	R4	2016-12-31
5	C2	사업개시	R1	2016-01-05
6	C2	보험가입	R2	2016-01-07
7	C2	업종변경	R3	2016-05-01

3.3 EDA & 프로세스 마이닝 기반 이력 분석

통합된 두 개의 데이터에 대하여 먼저 탐색적 데이터 분석을 통해, 정보를 파악하는 과정이 필요하다. 데이터 요약, 기초 통계량 분석, 변수 간 연관성 분석 등이 포함될 수 있다. 이 중 데이터 요약은 데이터와 관련된 전반적인 개요를 도출하는 것이다. 도출된 데이터의 전체 기간, 이벤트의 수, 관련 사업자의 수 등이 포함된다. 기초 통계량 분석은 이력 및 예측 분석에 앞서, 데이터 자체에서 나타나는 특징을 파악하기 위함이다. 예를 들어, 보험 가입/미가입 빈도수, 사업장 내 고용인원의 수, 사업장별 매출액의 특징 등이 해당된다. 기초 통계량 분석을 통해 특징을 파악한 후, 변수 간 연관성 분석을 통해, 고용보험과 상관관계를 가지는 특징적인 변수를 선별할 수 있다. 이러한 정보는 파생변수로 분류되어 예측 모델 도출시 기존의 독립변수와 함께 활용될 수 있다.

탐색적 데이터 분석 후에 프로세스 마이닝 기반의 예측 분석이 필요하다. 해당 분석의 목적은 이력과 관련된 프로세스 모델 및 패턴을 도출하여 보험 가입과의 연관성을 파악하는 것이다.

제3.2절에서 소개한 두 가지 데이터 중 이벤트 로그의 형태를 활용하며, 앞서 제Ⅱ장에서 소개한 alpha mining[12], heuristic mining[14] 외에 다양한 프로세스 모델 도출 방법을 통해 보험 가입과 연관된 이력 기반의 프로세스 모델을 도출할 수 있다. 또한, 이력 데이터를 기반으로 dotted chart[9]를 활용하여 프로세스 패턴을 도출함으로써, 보험 가입 이력에 관한 흐름을 정형화 할 수 있다.

3.4 예측 모델 도출

제시된 예측 모델 도출 방법론의 마지막 단계는 데이터를 기반으로 보험 가입을 예측하기 위한 모델을 도출하는 것이다. 제3.2절에서 제시한 두 가지의 통합 데이터 중 두 번째 형태의 데이터를 활용한다. 제Ⅱ장에서 소개한 것처럼 Decision Tree, Support Vector Machine, Neural Network, Random Forest를 포함한 모델 도출을 위한 다양한 기법들이 활용될 수 있다. 각 기법을 기반하여 도출된 예측 모델에 대해 정확도 측정을 통해 모델을 평가할 수 있는 단계가 필요하다. 이를 위해, 사전에 데이터 분할 방법을 통해 데이터를 Training 및 Test 형태로 구분하고, 모델 도출 후, 정오 분류표[15]를 활용하여 모델의 정확성을 파악한다. Classification 문제에서 모델의 성능을 파악하기 위한 주요 방법 중 하나이며, 본 연구에서는 Precision 수치를 활용하여 평가한다.

Ⅳ. 사례 연구

본 장에서는 앞서 제시한 고용보험 가입 예측 모델 방법론의 검증을 위한 사례 연구 적용 결과를 소개하고자 한다. 본 사례 연구를 위해, 실제 공공데이터를 활용하였으며, 데이터 수집, 데이터 통합 및 전처리 과정, EDA 및 프로세스 마이닝 분석, 예측 모델 도출, 결과 해석의 전체적인 과정에 대해 연구 결과를 제공한다. 본 방법론 검증을 위해, 울산 내 사업장을 분류하여 편의점, 음식점, 의류점의 세 카테고리로 사례연구를 수행하였다. 본 논문에서는 울산 내 편의점 사업장에 관한 분석 결과를 소개한다.

4.1 데이터 수집

본 사례 연구는 울산 지역 내 편의점 사업장의 연관 정보를 통해 고용 보험 가입 여부를 예측하는 모델 도출을 목표로 하였다. 이를 위해, <표 2>에 해당하는 데이터가 수집되었다. 먼저, A 기관에서 보유하고 있는 사업장 고용보험 가입여부를 나타내는 가입이력 데이터를 활용하였다. 또한, B 기관에서 보유하고 있는 사업장등록, 업종변경, 휴업/폐업, 일용근로 정보, 매출액 등의 사업자 관련 다양한 정보들이 수집되었다. 이와 더불어, C 기관에서 보유하고 있는 건축물 정보 데이터를 통해 해당 사업장의 건축물 정보를 수집하였다. 울산 지역 내 편의점 사업장과 관련하여 세 개의 기관에서 보유하고 있는 총 12개의 자료를 수집하여 본 연구를 수행하였다.

〈표 2〉 데이터 수집 내용

수집 기관	수집된 데이터 항목
A	고용보험 가입 이력
B	사업장 가입/업종변경/휴폐업/거소이전/대표자정정/일용근로/사회보험정보/매출액/원천징수/유관정보
C	건축물 정보

4.2 데이터 통합 및 전처리

수집된 데이터를 기반으로 앞서 제Ⅲ장에서 소개한 바와 같이 두 가지 다른 방향의 데이터 분석을 위해 두 종류의 통합 데이터를 생성하였다. 먼저, 보험 가입 여부 현황 및 패턴 분석을

위해 각 사업장별 가입 이력 데이터를 생성하였다. 가입이력 및 사업장 정보에 대하여, 시간에 따른 이벤트화를 통해 관련 데이터를 생성하였다. 고용보험의 경우, 종업원의 유무에 따라 동일 사업장 내에서도 시기에 따른 고용보험 가입 여부가 달라지기 때문에, 미가입/가입 이벤트를 가입이력 데이터에서 도출하였다. 또한, B 기관 데이터로부터는 사업개시, 대표자 정정, 업종 변경 등의 정보를 추출하여 사업장과 관련된 이벤트를 생성하였다.

고용보험 가입 예측 모델 도출을 위해서 사업장 관련 모든 정보를 동일 열로 구성하는 과정이 필요하였다. 하지만, 앞서 설명한 것처럼 동일 사업장이라도 종업원 유무에 따라 미가입/가입이 시기에 따라 발생할 수 있기 때문에, 이를 구분하는 과정을 수행하였다. 즉, 특정 사업장에 대해서 가입이력 데이터로부터 미가입/가입 정보를 가져온 후, 해당 정보가 발생하기 전까지의 B 기관, C 기관 내 다른 사업장 관련 정보를 삽입하여 데이터를 구성하였다. 또한, 각 기관에 저장된 기본 데이터 외에 개인/법인 구분, 고용상시 인원 수, 사업장 대표 나이, 성별 등의 다양한 파생변수를 생성하여 데이터에 포함하였다. 이와 더불어, 반대로, 모델링에 있어서 불필요한 변수는 삭제하였다. 예를 들어, 파생변수로 대체 가능한 데이터, 결측치가 90% 이상인 데이터, 날짜 관련 데이터 등은 삭제하였다. 이와 같은 전처리 과정을 통해 예측 모델 개발을 위한 데이터를 구성하였고, 이러한 프로세스 및 포함된 데이터의 수 <표 3>에 나타나 있다.

〈표 3〉 예측 모델 개발을 위한 통합 테이블 구성 과정

과정 No.	수행 과정 내용	테이블 내 열 수
1	전체 테이블 수집	522
2	무의미한 열 제외(식별자 관련)	478
3	테이블 통합 및 중복 열 제거	191
4	파생변수 추가	214
5	불필요 변수 삭제	37

4.3 EDA & 프로세스 마이닝 기반 이력 분석

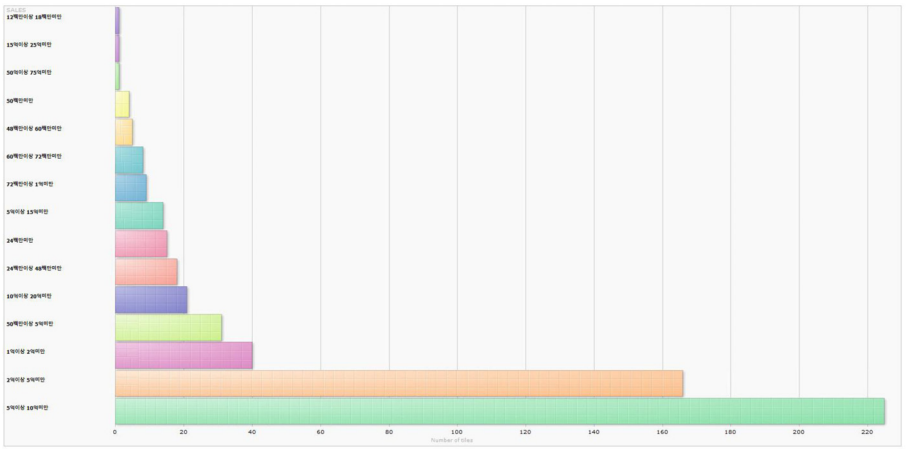
도출된 두 가지의 데이터 중 먼저 이력 데이터를 기반으로 데이터 탐색 및 프로세스 마이닝 분석을 수행하였다. 먼저 도출된 데이터 개요는 <표 4>에 나타나 있다. 울산 내 편의점 사업장 수는 총 836개로 분석되었으며, 고용보험을 가입한 사업장 수는 338개, 미가입한 사업장은 137개로 분석되었다. 사업장의 대표 연령대는 20대에서 70대로 다양하였고, 성별 빈도는 여성이 남성보다 100여 명 많은 것으로 나타났다.

〈표 4〉 데이터 개요

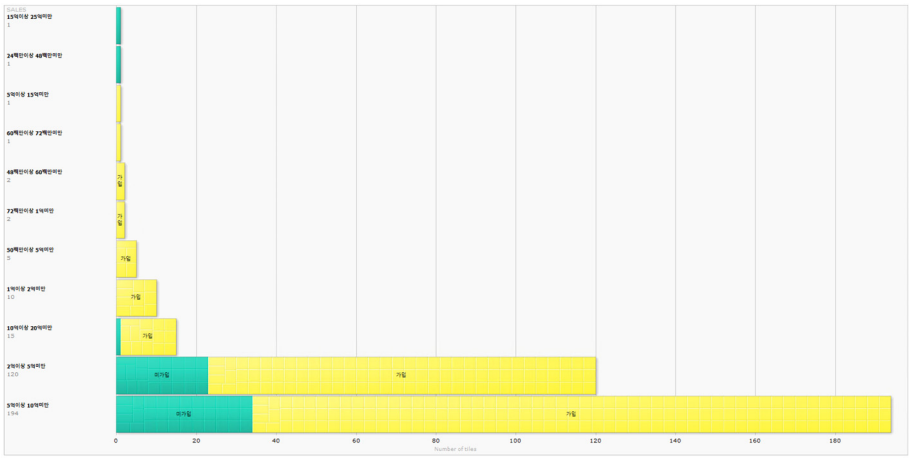
항목	데이터 개요 결과
사업장 수	836(가입 : 338, 미가입 : 137, 미확인 : 361)
편의점 프랜차이즈 수	7
울산 내 구/군 수	5(남/동/북/중구, 울주군)
사업장 대표 연령대	20대~70대
사업장 대표 성별 빈도	남 : 337명, 여 : 499명

다른 EDA 분석으로는 매출액 정보, 매출액 구간별 고용보험 가입여부, 가입여부에 따른 구/군별 사업장 종류의 빈도수, 가입여부에 따른 고용 상시 인원수별 사업장 종류의 빈도수 등이다. 먼저, 매출액 분석 결과(<그림 5>) 5억~10억의 구간 빈도수(<그림 5> 내 마지막 막대)가 가장 높으며, 2억~5억의 구간 빈도수(<그림 5> 내 마지막에서 두 번째 막대)가 뒤를 이었다. 또한, 일부 사업장의 경우, 1,800만 이하의 매출액을 가진 사업장 및 15억 이상의 매출액을 가진 사업장도 있는 것으로 분석되었다.

더불어, 매출액 구간별 고용보험 가입여부의 경우(<그림 6>), 2억~5억 구간과 5억~10억 구간에서 많은 규모의 미가입 사업장이 발견되었다(<그림 6> 내 마지막 두 막대). 해당 사업장의 경우, 가족 규모로 사업장을 운영하고 다른 종업원을 고용하지 않는 것으로 집계되었다. 가족



〈그림 5〉 매출액 정보

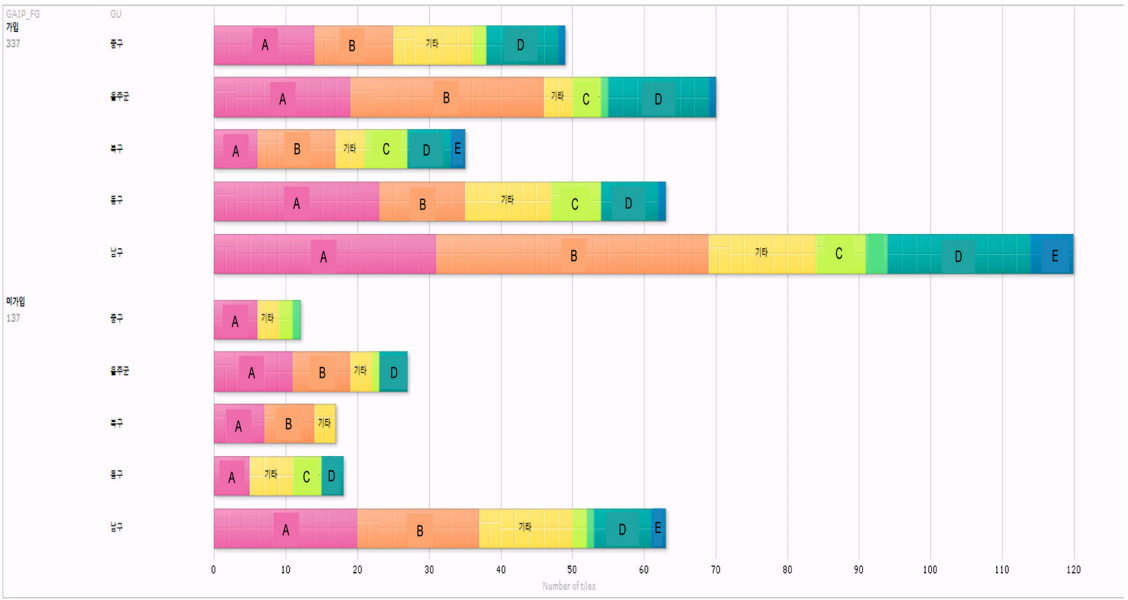


〈그림 6〉 매출액 구간별 가입여부

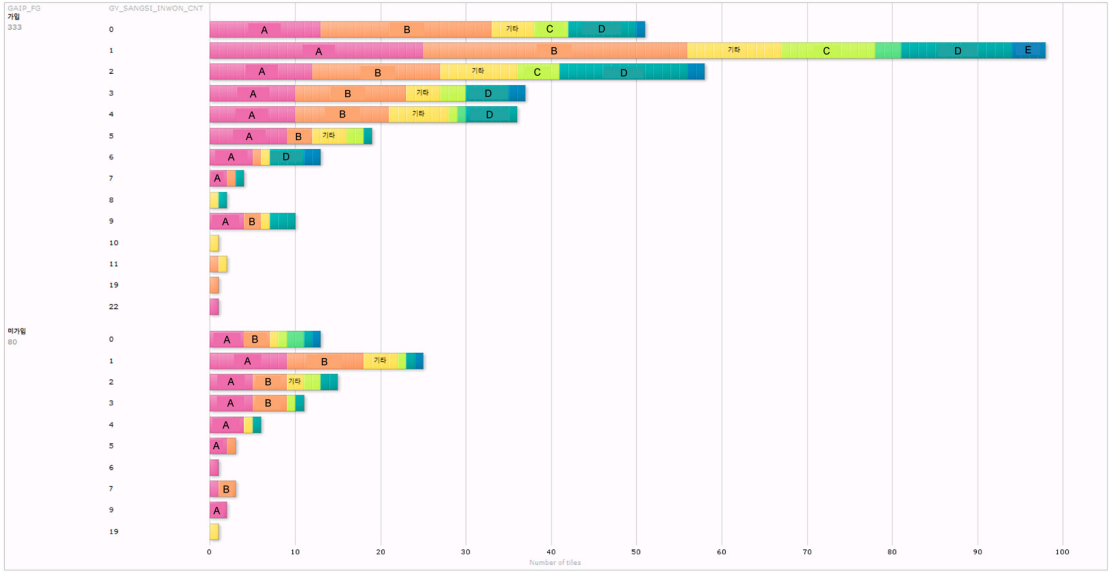
규모로 운영하는 경우, 고용보험을 따로 가입해야 할 필요가 없어서, 대다수 이러한 케이스가 발견되었지만 추가적인 고용에 대한 지속적인 확인이 필요한 것으로 사료된다.

다음은 가입여부에 따른 구/군별 사업장 종류의 빈도수 분석 결과(<그림 7>)이다. 편의점 프랜차이즈 B, D(<그림 7> 내 주황색, 녹색)의 경우, 다른 프랜차이즈 사업장에 비해 각 구/군별 고용보험 가입 빈도가 높은 것으로 분석되었다. 이와 반대로, 편의점 프랜차이즈 A(<그림 7> 내 붉은색)의 경우, 모든 지역에서 미가입 사업장이 발생하였다. 더불어, 편의점 사업장의 수가 가장 높은 남구의 경우, 가입 및 미가입 사업장 모두 많은 것으로 나타났다.

다음은 가입여부에 따른 고용 상시 인원수별 사업장 종류의 빈도수이다(<그림 8>). 전반적으로 사업장은 최소 0명부터 최대 22명까지 고용 상시 인원을 보유하여 운영하는 것으로 나타났다. 고용 상시 인원이 1명 이상임에도 불구하고 고용보험에 미가입된 사업장의 경우, 앞서 소개한 것처럼 가족이 운영하는 사업장으로 분석되었다. 하지만, 고용상시인원이 7명 이상인 경우도 나타나기 때문에, 가족 여부 확인을 위해 세부 조사가 필요한 것으로 사료된다.



〈그림 7〉 가입여부에 따른 구/군별 사업장 종류의 빈도수



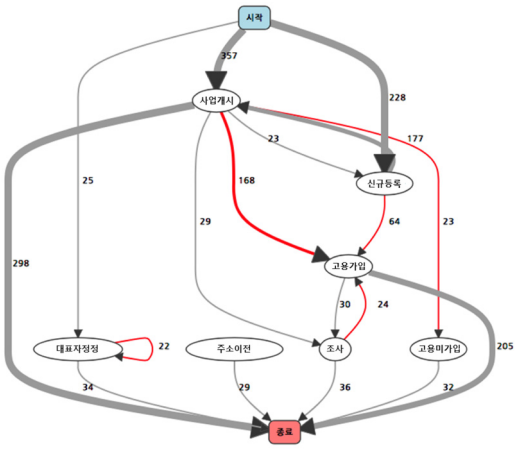
〈그림 8〉 가입여부에 따른 고용 상시 인원수별 사업장 종류

프로세스 마이닝 기반 가입이력 분석을 위해서 추출된 이벤트 로그 형태의 데이터 정보는 <표 5>에 나타나 있다. 총 836개의 사업장 중 사업개시일이 2000년 이후에 해당하는 809개의 데이터만 추출하였고, 총 7개의 상태 변경 이력 작업이 포함되었다. 또한, 총 1,744개의 이벤트가 추출되었고, 한 사업장 당 약 2,2개의 작업을 수행한 것으로 나타났다. 기간은 2000년 1월 1일부터 2016년 3월 1일까지의 데이터를 포함하였다.

〈표 5〉 가입이력 데이터 정보

항목	데이터 정보
사업장 수 (케이스 수)	809* (*총 836명의 사업자 중, 사업개시일이 2000년 이후에 해당하는 데이터만 추출)
이력 작업 수	7 (신규등록, 고용가입, 고용미가입, 조사, 사업개시, 대표자정정, 주소 이전)
이벤트 수	1,744
기간	2000. 01. 01.~2016. 03. 01.

본 가입이력 데이터를 기반으로 프로세스 모델 도출 분석 및 패턴 분석을 수행하였다. 먼저, <그림 9>는 도출된 고용 보험 가입이력 프로세스 모델을 나타낸다. 프로세스 시작 후, 신규 등록과 사업 개시를 초기에 수행하고, 일부는 그대로 종료하거나(i.e., 고용보험 없이 사업 지속 수행), 일부는 고용가입 후 종료하는 것(고용 보험 가입 후 사업 수행) 으로 분석되었다. 또한, 이 외에 사업 개시 후에, 조사 작업 수행 뒤 고용 가입이 되는 작업 흐름도 일부 발견되었다.



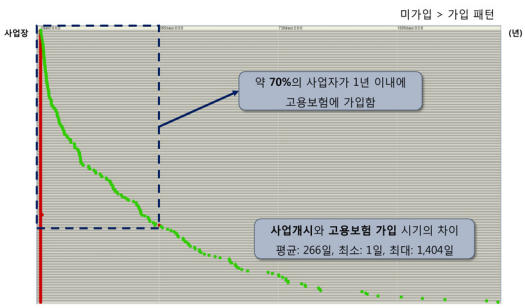
〈그림 9〉 가입이력 프로세스 모델

다음은 프로세스 패턴 분석 결과이다. <표 6>은 패턴 분석을 정리한 결과를 나타낸다. 분석 결과, 약 50% 이상의 사업장이 고용 보험 미가입이 지속된 상태로 사업을 수행하는 것으로 나타났고, 약 34%의 경우, 사업 개시 후 고용 보험을 가입한 것으로 나타났다. 약 5%의 사업장의 경우, 가입 후 고용 보험을 소멸한 패턴을 나타냈다.

〈표 6〉 패턴 분석 결과

패턴 No.	패턴	빈도(비율)
1	미가입 >	424(52.4%)
2	미가입 > 가입 >	274(33.4%)
3	가입 >	67(8.3%)
4	가입 > 미가입 >	42(5.2%)
5	미가입 > 가입 > 미가입 >	2(0.2%)

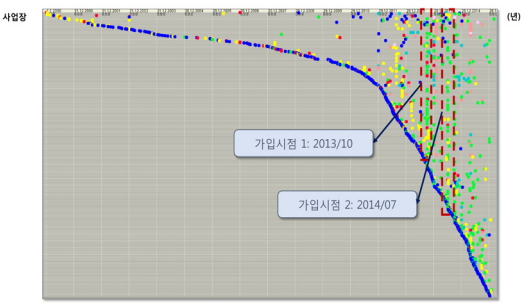
사업 개시 시점부터 일정 시간 후 가입한 2번 패턴을 대상으로 Dotted Chart를 통한 상세 분석을 수행하였다. <그림 10>은 Dotted Chart 기반의 본 분석의 결과를 나타낸다. Dotted Chart 내 붉은 색 점은 미가입 이벤트, 녹색 점은 가입 이벤트를 나타낸다. 분석 결과, 약 70%의 사업장이 1년 이내에 고용보험에 가입한 것으로 나타났다. 이 중 일부 사업장은 사업개시 다음날 고용보험을 가입한 경우도 있었고, 이와 반대로 약 1,400일 후에 고용보험을 가입한 사업장도 존재한 것으로 나타났다.



〈그림 10〉 패턴 2에 대한 상세 분석 결과

이와 더불어, 실제 시간에 따른 이력 작업에 관한 Dotted Chart를 통해 사업장들의 공통화된 패턴 파악 연관 분석을 수행하였다(<그림 11>).

Dotted Chart 내 푸른색 점은 사업개시 작업을 의미하며, 노란색 점은 공단 직원의 조사 작업, 녹색 점은 고용보험 가입 작업을 나타낸다. 분석 결과, 사업장에서 고용보험 가입 작업이 배치 형태로 나타나는 것을 확인할 수 있었다. 다시 말하면, 유사한 시기에 다수의 사업장이 고용보험을 가입했음을 의미한다. 녹색 점 주위에 노란색 점(조사 작업)이 다수 발견되었음이 분석되었고, 이와 더불어 실무진과의 토의를 통해, 해당 시점 주위에 표본조사가 있었음을 확인할 수 있었다.

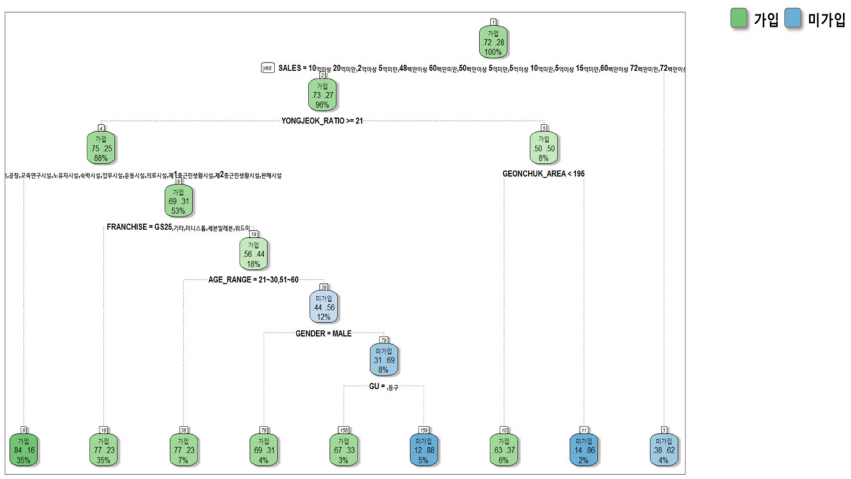


〈그림 11〉 실제 시간별 이력 작업 시각화 차트

4.4 예측 모델 도출

다음은 고용보험에 관한 가입여부 예측 모델 도출 결과이다. 먼저, 앞서 소개한 것처럼 데이터 분할 방법을 적용하여 70%의 데이터로 Training 셋, 30%의 데이터로 Test 셋을 구성하였다. 즉, 가입/미가입이 확실히 구분된 475개의 사업장 중 임의로 선택된 332개의 사업장이 Training 셋에, 143개의 사업장이 Test 셋에 포함되었다. Training 셋을 바탕으로 Decision Tree, Random Forest, Support Vector Machine, Neural Network 알고리즘 기반의 모델을 구성하였고, 결과는 다음과 같다.

먼저, <그림 12>는 Decision Tree 알고리즘 적용 결과를 나타낸다. 전체 37개의 변수 중 총 5개의 결정 변수가 발견되었고 각각은 다음과 같았다: 매출액, 용적율, 건축 면적, 프랜차이즈, 주 사용목적. 고용보험에 가입해야하는 사업장의 결정 규칙은 다음과 같다(매출액: 4,800만~1억 또는 2억~20억, 용적율: 21% 이상, 주 사용목적: 근린생활시설, 프랜차이즈: B 또는 D). 더불어, 고용보험 미가입으로 예측되는 사업장의 결정 규칙은 다음과 같다(매출액: 4,800만 이하 또는 1억~2억, 용적율: 21% 이하, 건축면적: 195m^2 이상). 정리하면, 매출액, 용적율을 기반으로 고용보험 가입여부가 크게 구분되고 있음을 알 수 있다.



〈그림 12〉 Decision Tree 적용 결과

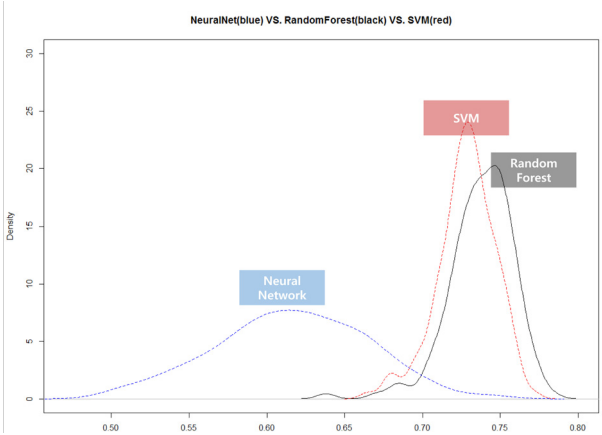
Decision Tree 적용에 따른 결과를 나타내는 정오 분류표는 <표 7>에 나타나 있다. 143개의 Test 셋을 도출된 모델에 적용하여 도출한 결과이다. 전체 개체 중 96개의 사업장에 대한 고용보험 가입 여부를 정확하게 예측하였고, 정확도는 약 0.671인 것으로 분석되었다. 이외에 예측이 어긋난 경우, 대부분 실제 가입이지만 미가입으로 예측된 개체가 많았다.

〈표 7〉 Decision Tree 기반 고용보험 가입 예측 모델 검증

		예측		
		가입	미가입	계
실제	가입	88	35	123
	미가입	12	8	20
	계	100	43	143

Decision Tree와 더불어, Random Forest, Support Vector Machine, Neural Network 알고리즘 기반의 모델 정확도 결과를 소개한다. <그림 13>은 해당 모델링 기법 적용에 따른 정확도 결과를 나타낸다(SVM: Support Vector Machine, NN: Neural Network, RF: Random Forest). 각 기법 별로 총 30번의 모델 생성 및 검증을 수행하였다. 먼저, Random Forest의 경우, 정확도 평균이 가장 높고, 표준편차도 가장 낮은 것으로 분석되었다(평균: 0.735, 표준편차: 0.025). 두 번째로 SVM이 정확도 평균이 높았으며(평균: 0.700, 표준편차: 0.026), Neural Network 기반의 모델이 가장 정확도 낮고, 편차가 큰 것으로 분석되었다(평균: 0.612, 표준편차: 0.042). 약 0.671의 정확도를 가진 Decision Tree 기반의 모델을 포함하여 총 정리하면, Random Forest 기반의 고용보험 가입여부 예측 모델이 가장 효과적인 것으로 판단된다.

고용보험 가입 예측 모델 도출과 더불어, 고용 상시 인원 유무 예측 모델 도출 분석을 수행하였다. 사업장에 고용 상시 인원이 있는 경우, 해당 사업장은 고용 보험 가입을 해야 하기 때문에 유사하게 고용 보험 가입 여부 예측을 할 수 있었기 때문이다. 본 예측 모델 도출을 위해서 Decision Tree 만을 적용하였고, 총 5개의 결정 변수가 발견되었고 각각은 다음과 같다: 건폐율, 총 주차 수, 프랜차이즈, 지역코드, 나이. 고용 상시 인원이 있는 것으로 예측되는 사업장의 결정 규칙은 다음과 같다(총 주차 수: 5대 이하, 건폐율: 58% 이상, 지역코드: 공업지역이 아닌 곳). 반대로, 고용 상시 인원이 없는 것으로 예측되는 사업장의 결정 규칙은 다음과 같다(건폐율 : 20% 이하, 나이: 30대, 프랜차이즈: A 또



〈그림 13〉 SVM, NN, RF 기반 모델 도출에 따른 정확도

는 D). 정리하면, 건폐율을 기반으로 고용 상시 인원 존재 여부가 크게 구분되고 있음을 알 수 있다.

도출된 모델에 대하여 Test 셋 기반의 검증 결과, 즉, 정오 분류표는 <표 8>에 나타나 있다. Test 셋 내 전체 개체 중 103개의 개체가 정확하게 예측 되었고, 모델의 정확도는 약 0.720으로 분석되었다. Decision Tree 기반 고용 보험 가입 예측 모델 보다는 더 정확도가 높은 모델이 도출되었지만, Random Forest 기반의 모델 보다는 정확도가 낮은 것으로 분석되었다.

〈표 8〉 고용 상시 인원 유무 예측 모델 검증

		예측		
		가입	미가입	계
실제	가입	97	26	123
	미가입	14	6	20
	계	111	32	143

Ⅴ. 시사점

본 논문에서 제시한 고용 보험 가입 여부 도출 모델은 기존에 A 기관이 가진 문제점에 대한 효과적인 해결 방안이 될 것으로 기대된다. 기존에 해당 기관은 특정 지역 내 사업장의 고용 보험 가입 여부를 파악하고 가입 누락 사업장에 대한 실태 조사를 지속적으로 수행해왔다. 이로 인해, 발생하는 시간과 비용이 누적되었고, 이러한 문제를 해결하기 위한 데이터 기반의 새로운 방법을 필요로 하였다. 본 연구에서 제시한 데이터 기반의 예측 모델은 큰 규모의 불필요한 비용을 줄이고 체계적인 사업장 관리가 가능하게 할 것으로 기대된다.

이와 더불어, 본 논문에서는 프로세스 마이닝 기법을 활용한 이력 분석 방법도 제시하였다. 데이터 마이닝의 예측 알고리즘을 바로 적용하는 것이 아니라, 프로세스 마이닝 내 모델 도출 및 패턴 분석을 통해, 프로세스 관점에서의 전반적인 흐름을 파악하는 등 도출 모델의 정확도를 높이기 위한 사전 분석 기능으로 충분히 효과적임을 보여주었다. 예를 들어, Dotted Chart를 통해, 실태 조사의 전후로 다수의 사업장이 고용 보험을 비슷한 시기에 가입하는 패턴을 파악하였다.

도출된 예측 모델과 관련해서도 약 70% 내외의 정확도를 가진 것으로 분석되었으며, 이와 더불어 정오 분류표 내 FP(False Positive), 즉, 예측은 가입으로 나타나지만 실제는 미가입으로 발생하는 사업장에 대한 상세 확인이 필요한 것으로 판단되었다. 실무진과의 협의를 통해, 해당 사업장들은 충분히 고용보험이 필요할 수도 있는 개체들로 지속적인 실태 조사 등의 관리를 수행하기로 하였다.

본 연구는 다양한 확장성을 가질 수 있다. 고용 보험의 가입과 관련성이 높은 외부 데이터를 추가적으로 활용하여 모델을 정교화 할 수 있다. 예를 들어, 사업장에 이용되는 교통수단의 수에 따라 피고용인의 고용 가능성을 추정할 수 있다. 또한, 본 사례연구는 울산 지역 및 특정 산업 범위 (편의점) 내에서만 수행하였다. 하지만, 이를 타 지역으로 확장하여 지역별 특성을 고려한 모델을 개발하거나, 혹은, 산업별 특징적인 사항을 추가 반영하여 더 정교화된 모델을 도출할 수 있다.

추가적으로, 모바일 앱 혹은 시스템 개발을 통해 실용화 추진이 가능하다. 이에 따라, 도출된 모델을 기반으로 모바일 앱 개발을 수행할 예정이며, 해당 앱 개발은 가입 가능성이 높은 미가입 사업장에 대한 정보를 제공하는 것을 목표로 할 것이다.

Ⅵ. 결론 및 추후연구

본 논문에서는 데이터 수집, 데이터 통합 및

구성, 탐색적 데이터 분석(EDA) & 프로세스 마이닝 기반 이력 분석, 예측 모델 개발을 포함한 고용보험 가입 예측 모델 개발 방법론을 제시하였다. 또한, 울산 내 편의점 사업장 연관 데이터를 기반의 사례연구를 통해 본 방법론을 검증하였다.

본 연구는 기존의 방식이 가지고 있던 불필요한 시간 및 비용의 소모를 줄이고, 데이터 기반으로 실제적인 고용 보험 가입 여부를 파악한데 있어 직접적인 기여점을 가진다. 이와 더불어, 본 연구는 특정 기관에 수집된 데이터가 아니라 다양한 공공데이터를 결합하여 활용함으로써, 공공데이터의 활용성 증대에 기여할 것으로 판단된다.

추후 연구로 앞서 소개한 것처럼 다른 지역, 다른 산업 범위에 적용하여 해당 특성을 고려한 Context 기반의 고용 보험 예측 모델을 도출할 계획이다. 또한, 고용 보험 외에 산재 보험 등의 다른 분야에 적용하고자 한다. 이외에 Test 셋 기반이 아닌 본 모델 기반의 실제 실태 조사를 통해 본 모델을 검증할 계획이며 이를 기반으로 한 정교화된 모델링 방법에 관한 연구를 수행할 계획이다. 마지막으로, 본 분석을 포함하여 고용보험 가입에 관한 사업장 모니터링이 가능하도록 관련 방법론 확장 및 시스템 구축을 수행할 계획이다.

참 고 문 헌

[1] Breiman, L., "Random Forests", *Machine Learning*, Vol.45, No.1, pp.5-32, 2001.

[2] Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases", *AI magazine*, Vol.17, No.3, pp.37-54, 1996.

[3] Kim, G.H., S. Trimi, and J.H. Chung, "Big-data applications in the government sector", *Communications of the ACM*, Vol.57, No.3, pp.78-85, 2014.

[4] Liaw, A. and M. Wiener, "Classification and regression by random forest", *R News*, Vol.2, No.3, pp.18-22, 2002.

[5] Provost, F. and T. Fawcett, "Data science and its relationship to big data and data-driven decision making", *Big Data*, Vol.1, No.1, pp.51-59, 2013.

[6] Quinlan, J.R., "Simplifying decision trees", *International Journal of Man-Machine Studies*, Vol.27, No.3, pp.221-234, 1987.

[7] Rosenblatt, F., "Principles of neurodynamics. perceptrons and the theory of brain mechanisms", *Brain Theory*, pp.245-248, 1961.

[8] Rowley, H.A., S. Baluja, and T. Kanade, "Neural network-based face detection", *IEEE Transactions on pattern analysis and machine intelligence*, Vol.20, No.1, pp.23-38, 1998.

[9] Song, M. and W.M.P. van der Aalst, "Supporting process mining by showing events at a glance", In Proceedings of the 17th Annual Workshop on Information Technologies and Systems(WITS), pp.139-145, 2007.

[10] Tong, S. and D. Koller, "Support vector machine active learning with applications to text classification", *Journal of Machine Learning Research*, Vol.2, pp.45-66, 2001.

[11] van der Aalst, W.M.P., "Process mining: data science in action", Springer, 2016.

[12] van der Aalst, W.M.P., A.J.M.M. Weijters, and L. Maruster, "Workflow Mining: Discovering process models from event logs", *IEEE Transactions on Knowledge and Data Engineering*, Vol.16, 2003.

[13] Vapnik, V., The nature of statistical learning theory, Springer, 2000.

[14] Weijters, A.J.M.M., W.M.P van der Aalst, and A.A. De Medeiros, "Process mining with heu-

ristics miner-algorithm", *Technische Universiteit Eindhoven, Tech. Rep. WP*, Vol.166, pp.1-34, 2006.

[15] Witten, I.H., E. Frank, M.A. Hall, and C.J. Pal, "Data Mining: Practical machine learning tools and techniques", *Morgan Kaufmann*, 2016.

저 자 소 개

조 민 수(Minsu Cho)

- 2013년 : 울산과학기술원 테크노경영학부 (학사)
- 2013년~현재 : 울산과학기술원 경영공학과 (석, 박사 통합과정)
- 관심분야 : 프로세스 마이닝, Data Analytics, BPM(Business Process Management)

김 도 현(Dohyeon Kim)

- 2016년 : 울산과학기술원 컴퓨터공학과 (학사)
- 2016년~현재 : 포항공과대학교 산업경영공학과 (석, 박사 통합과정)
- 관심분야 : 프로세스 마이닝, Data Analytics, Data Visualization

송 민 석(Minseok Song)

- 2006년 : 포항공과대학교 산업경영공학과 (공학박사)
- 2006년~2009년 : 아인트호벤 공과대학교 (박사후과정)
- 2010년~2015년 : 울산과학기술원 경영학부 조/부교수
- 2016년~현재 : 포항공과대학교 산업경영공학과 부교수
- 관심분야 : 프로세스 마이닝, BPM(Business Process Management), 비즈니스 분석(Business Analytics)

김 광 용(Kwangyong Kim)

- 1998년 : 동아대학교 경영대학원 경영학석사
- 현재 : 근로복지공단 산재심사위원회 상임위원
- 관심분야 : 공공데이터 분석, Data analytics, 데이터 마이닝

정 충 식(Chungsik Jeong)

- 1991년 : 경상대학교 경제학과 (학사)
- 현재 : 근로복지공단 부산업무상질병판정위원회 위원장
- 관심분야 : 공공데이터 분석, Data analytics, 데이터 마이닝

김 기 대(Kidae Kim)

- 2001년 : 청주대학교 호텔경영학과 (학사)
- 현재 : 근로복지공단 충주지사 재활보상부 과장
- 관심분야 : 공공데이터 분석, Data analytics, 데이터 마이닝