경영전문석사학위 논문

퇴사자 예측 분류 모델을 활용한 기업 지속가능가치평가에 대한 연구

2022 년 8 월

서울과학종합대학원대학교

임 휘 종

경영전문석사학위 논문

퇴사자 예측 분류 모델을 활용한 기업 지속가능가치평가에 대한 연구

2022 년 8 월

서울과학종합대학원대학교

임 휘 종

퇴사자 예측 분류 모델을 활용한 기업 지속가능가치평가에 대한 연구

지도교수 박 정 열

이 논문을 경영학 석사 학위논문으로 제출함

2022 년 8 월

서울과학종합대학원대학교

임 휘 종

임휘종의 석사 학위논문을 인준함

2022 년 7 월

위 원 장 _____ 김 보 영 _____ (인)

위 원 _____ 임 효 숙 _____ (인)

위 원 _____ 박 정 열 _____ (인)

초 록

본 연구는 이직 및 퇴사자 증가가 기업의 지속가능가치에 부정적 영향을 야기함을 고찰하고, 이직 및 퇴사자 데이터의 정량적 독립 요인을 분석하여 분류 예측 모델을 설계한다. 정량적 데이터 분석을 통한 독립 요인 탐색 및 예측 분류 모델 설계는 추후 정성적 데이터 분석을 통한 정량적, 정성적 데이터를 모두 활용할 수 있는 예측 분류 모델을 설계하기 위한 초석이며, 본 연구에서는 정량적 데이터와 Logistic regression, RandomForest, Multi Layer Perceptron 을 사용하여 각 정량적 독립 요인의 특성 중요도를 파악하고 상관관계를 분석한다. 이는 모두 기업의 인사적 관점에서 이직 및 퇴사로 인한 인력 손실은 장기적 차원에서 지속가능가치의 저하를 전제로서, 방지 및 개선을 위한 목적으로 본 연구를 진행하였다.

목 차

표 목 차

그 림 목 차

제 I 장 서 론

제 1 절 연구의 배경 및 필요성

(1) 데이터 과학의 인문학적 접근

인문학이란 인간과 인간의 근원 문제, 인간 본연의 가치와 자기표현 능력을 바르게 이해하기 위한 학문 분야로서, 인간의 사상과 문화를 기초로 탐구하는 분야이다. 인간의 감정은 인문학적 관점에서 분석될 수 있지만 데이터 분야 관점에서 분석하기 위해서는 감정의 수식화 및 지표화를 통해 기준을 정립해야 하며, 그 기준에 의해 데이터로서 정량화될 수 있다. 통상 인문학과 연관된 인간의 감정과 생각, 가치를 내포한 글은 비정형 데이터로 분류될 수 있으며 데이터 과학의 관점에서 위 비정형 데이터는 목적에 따라 정량화될 수도, 비정형화된 날 것 그대로의 데이터로서도 활용될 수 있다. 하지만 이번 연구 활용에 사용되는 정량적 데이터를 활용한 예측 분류 모델은 감성적 요인에 의한 이직 및 퇴사가 아닌 정량적 요인에 의한 이직 및 퇴사율 평균과 예측치를 비교하는 데 의의가 있다. 실제값이 평균 혹은 예측값에서 가시적인 변동 폭이 존재한다면 독립 변수 외의 다른 외부 정량적 요인 혹은 감성 요인에 의한 변동이라 유추할 수 있다. 위 연구에서 인문학은 감성적 요인과 연관 있으며, 감성적 요인을 분류하고 분석하기 위해 정량적 데이터를 통해 이직 및 퇴사자를 예측 분류하는 것이다. 기업

퇴사자의 요인을 파악하는 데 있어서 정량화된 요인과 정성적인 요인 두 가지가 공존하므로, 정량적 데이터를 우선으로 기업 퇴사자를 예측하여 기업의 지속가능가치평가의 요인으로 활용하고, 그 외에 발생하는 특이치 혹은 이상치를 정성적 요인에 의한 요인으로 분석하기 위함이다.

(2) 상황 인식

위 연구의 목적은 인적자원 중요성을 전제로 한 기업 지속가능가치의 향상이며, 이를 위해 퇴사자 분류 예측기 모델을 설계하여 기업의 필요 인적 자원을 지속적으로 보유하는 것이 본 연구의 목적 중 하나이며, 선행연구를 통해 인사적 관점에서의 기업 지속가능가치 필요성을 전제할 것이다. 현 취업난과 구인난이 공존하는 경제적 상황 속, 구인 시장은 시장 자본주의에 입각하여, 유능하며 기업이 선호하는 기술들 보유한 인력일수록 수요가 공급보다 앞서며 가치가 더해진다. 위 상황적 인식에 따라 기업은 기업의 이윤 증대와 지속가능가치의 존속을 위하여 유능한 인력은 유입되어야 하며 기업 내 존재하는 우수한 인력은 보존되어야 한다. 위와 같이 이러한 상황 인식 속에서 기업은 기업의 지속가능가치 향상을 위하여 인적 자원의 필요성과 중요성을 인식해야 하며 현재 자사의 사내 직원 관점에서의 종합적인 현 상황과 동향을 파악해야 한다.

(3) 기업의 지속가능가치 평가에 대한 기준

기업은 국가와 마찬가지로 실체가 없는, 실존적 이념이 입각하지 않는 무형의 자산으로서 기업의 존재와 가치는 기업의 구성인 건물, 제품, 사옥 그리고 구성원과 같은 유형 자산으로 상징될 수 있다. 이러한 형태가 없는 무형적 자산이 지속가능한 가치를 보존하기 위해서는 유형 자산의 지속적인 발전이 이루어져야 하며, 유형자산의 가치 평가를 통해 기업의 지속가능가치 평가가 예측될 수 있다. 나아가, 기업유형자산의 지속가능성과 가치 평가에는 다양한 관점과 기준이 존재하며 여러 요인에 의해 결정되는데, 본 연구에서는 기업의 신제품 R&D 와 기업 혁신, 구조 개편, 조직 문화 등 여러 요인에 의한 전제보다 인적 자원의 중요성을 전제로 진행할 것이다.

(4) 지속가능성에 의한 인적 자원의 중요성

위 (3) 기업의 지속가능가치 평가에 대한 기준에서 언급한 인적자원의 중요성의 이유는 다음과 같다. 회사의 이윤 증대를 야기하는 신제품 R&D, 기술과 노하우, 서비스 등 모두 매출의 증대를 위해 끊임없이 발전되고 개선되는 요인이며, 그 기저에는 인적 자원 존재한다. 구성원 개개인 역량의 발전은 위에서 언급한 이윤 증대를 야기하는 요인들과 직결되는 기업의 성과 증대와 연관성이 높다. 하지만 구성원의 역량이 발전되지 않고 정체되어 있고, 우수한 직원의 유입보다 유출이 더 커진다면 기업은 점진적으로 쇠퇴하며 유형 가치가 감소함에 따라 기업의 본질과 그 무형의 가치 또한 하락한다. 다시 말해, 기업의 지속가능가치의 존속과 발전은 우수한 역량

유입과 성장의 선순환적 구조가 이루어져야 하며 구성원들의 자사에 대한 관점과 생각, 그리고 현 실태를 파악하기 위해서는 정량적 요인과 정성적 요인 모두가 고려되어야 한다.

제 2 절 연구의 목적

본 연구의 목적은 퇴사자 예측 분류 모델을 활용하여 정량적 데이터를 통해 사내 만족도를 파악하고, 위의 객관적 지표를 통해 인사 관점에서 기업 지속가능가치를 향상시키는 것이다. 사내 만족도는 통상 설문 조사나 사내 건의사항 수렴을 통해 파악될 수 있는 지표이지만, 본 연구에서 활용하는 정량적 데이터, 가령 도시 발전도, 업무 경험, 교육 수준, 회사 규모와 같이 전형적인 퇴사 및 이직 요인 등 여러 요인들을 통해 퇴사 및 이직률을 예측하는데, 이는 사내 만족도의 객관적 지표가 될 수 있다고 사료된다. 이러한 전형적인 퇴사 및 이직 요인들을 통해 예측 분류 모델을 설계하여 임직원의 퇴사 및 이직 예측률과 실제값을 비교하였을 때, 독립 요인에 의해 추산된 예측은 타 기업의 퇴사 및 이직자의 데이터에 의해 평균치의 퇴사 및 이직률을 나타내므로 해당 기업의 실제값이 더 높을 경우, 이는 다른 요인, 즉 사내 만족도와 같은 감정적 요인이 고려된 것으로 추론할 수 있다. 기업의 인력과 매출액은 우수한 인력을 보유 및 훈련시키기 위해 교육 훈련, 직무 능력 향상, 사내 복지 개선 등 여러 재정적 비용을 포괄하고 있고, 그 둘은 유의미한 구조적 관계를 갖고 있으므로, 기업 입장에서 사내 만족도와 같은

감정적 요인에 의한 이직 및 퇴사는 기업의 지속가능가치에 부정적 영향을 초래한다고 판단된다.

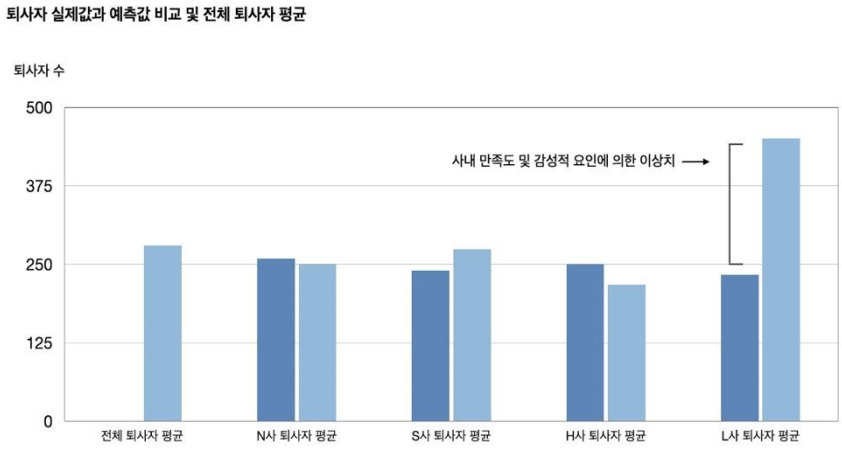


그림 1. 퇴사자 실제값과 예측값 비교 및 전체 퇴사자 평균과 이상치 예시

위 그림은 이전에 설명한 전체 이직 및 퇴사자 평균과 자사의 이직 및 퇴사자 평균을 기준으로, 예측 분류 모델을 활용한 예측값이 가시적인 이상치로 나타난다면, 현재 활용하는 정량적 데이터의 독립 변수 외, 외부 요인 혹은 감성적 요인에 의한 수치로 판단될 수 있음을 나타내는 자료이다. 본 연구는 보다 정확한 정량적 요인과 감성적 요인에 의한 이직 및 퇴사자 예측 분류 모델 설계에 초석이 되는 연구이며, 인사적 관점에서의 기업의 지속가능가치의 향상을 목표로 한다.

제 II 장 이론적 고찰

제 1 절 선행 연구 고찰

본 연구는 이직 및 퇴사 요인에 대한 정량적 데이터 분석 및 예측 분류 모델을 기반으로 이직 및 퇴사자를 예측함과 동시에 예측치를 통한 평균 편차의 비교를 통해 정성적 요인과 해당 모델에 기인한 요인 외 외부적 요인에 대한 퇴사 원인을 유추한다. 하지만 그전에 선행 연구로서 기업의 이직 및 퇴사자의 예측 분류를 통한 상황 인지가 자사의 가치 및 제품, 서비스 향상에 기여하며, 기업의 지속가능가치를 지속 및 증가시킴을 전제한다. [1]이직 및 퇴사자의 기업 영향에 대한 논문 주제를 통해 분석한 결과, 퇴사 및 이직자 증가는 조직 성과 및 재무 성과에 부정적 영향을 미치며 그 관계 역시 역 U 자 모양을 띄고 있다.[2] 또한 기업에는 기업마다 특유의 적정 이직률이 존재하며 이 이직률의 편차에서 벗어난 수치는 이직의 비용 편익 관점에서 부정적 영향의 결과가 관측되었다.

---

[1] 나인강, “The Effect of Turnover on the Company Performance”, [한국고용노사관계학회], (2011), vol.35, no.1, pp. 23-48 (26 pages)
[2] 김영박,김형중, “Predicting Early Retirees Using Personality Data”, [한국디지털콘텐츠학회], (2018), p141-147

제 2 절 데이터 탐색

본 연구에서는 Kaggle 의 HR Analytics: predicting quitters dataset[3]을 사용하였다. 이 데이터는 사내 직원의 퇴사 및 이직 여부를 예측하기 위한 dataset 으로, 여러 퇴사 관련 요인들, 도시 발전도, 학력, 성별, 경력, 전공, 기업 크기, 직무 교육 시간, 이직 시기 등을 고려하여 모델을 설계하였다. 독립 변수는 총 10 개로 이루어져 있으며, Training dataset 은 19,158 개이다. Training dataset 이 많지 않은 관계로 훈련 세트에 중복을 허용하여 샘플링을 하는 Bagging 방식을 사용하여 모델에 적용할 것이다. 퇴사 및 이직 예측 데이터인 Target 또한 Training set 과 동일하게 19,158 개로 구성되어 있고, 본 연구에서는 Validation Set 을 Training dataset 중 500 개를 추출하여 생성하였다.

3 Val Bauman, “HR Analytics: predicting quitters”, [Kaggle], (2021)

항목	설명
enrollee_id	Personal user index
city	The city which lived in
city_development_index	Expressing urban development between zero and one
gender	Male, Female, Not wirte(others)
relevent_experience	Experience with the current position of work
enrolled_university	A graduate university
education_level	One's final education background
major_discipline	A graduate major
expeirence	Industry experience associated with one's current job
company_size	Current number of employees in the workpalce
last_new_job	The number of previous jobs
training_hours	Training hours associated with current work position
target	Dependent variable

표 1. Data demonstration

제 3 절 활용 모델 제시

본 연구에서는 Kaggle 의 HR Analytics: predicting quitters dataset 을 분석하여 사내 임직원 이직 및 퇴사율을 예측하며, 예측을 위해 Logistic Regression, RandomForest, 그리고 Multy Layer Perceptron 을 사용한다. 각 모델마다 앙상블 기법 및 hyperparameters 조정을 통해 Accuracy 를 향

상시킬 것이며, Logistic Regression 과 RandomForest 의 경우 특성 중요도와 가중치를 나열하여 특성의 중요도를 살펴볼 것이다. 우선, Log-odds 와 Sigmoid Function 을 활용한 Logistic Regression 모델을 통해 Training dataset 을 학습시키고 Validation dataset 에 넣어 Accuracy 와 AUC-ROC Curve 를 통해 모델의 분류 성능을 판단할 것이다. 두 번째로, RandomForest 의 기본 구성 요소인 Decision Tree 모델을 사용하여 각 독립 변수의 Cut-off 를 계산하여 이직 및 퇴사자를 분류할 것이며, 임계값(Cut-oof)은 CART 알고리즘을 통해 불순도는 지니 계수(Gini Index)로 계산한다. 마지막으로 Multy Layer Perceptron 은 Artifical Neural Network(인공 신경망) 알고리즘 중에서 여러 개의 층을 순차적으로 쌓아 만든 신경망을 의미한다. 한 층에는 여러 개의 노드(Node)로 이루어져 있으며 본 연구에서는 입력층과 3 개의 은닉층, 그리고 출력층으로 구성하여 하이퍼 파라미터 조정을 통해 Accuracy 를 향상시킬 것이다. hyperparameters 조정과 모델 설명은 3 장 2 절에서 상세히 다룰 것이다.

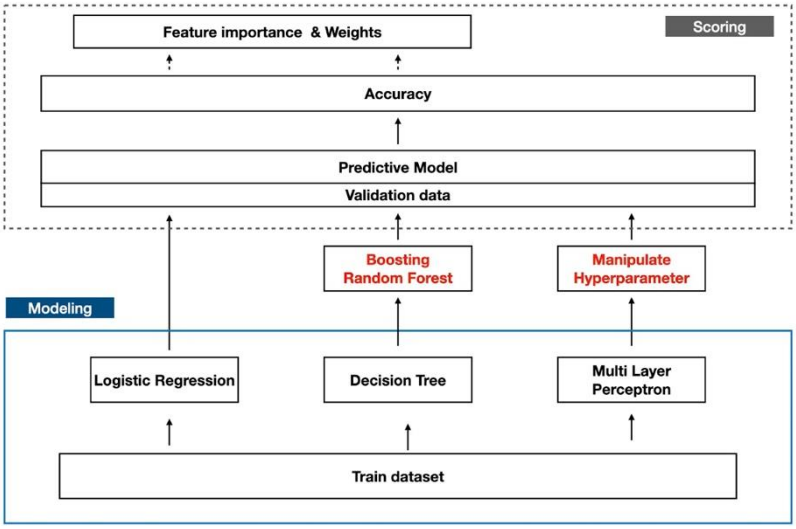


그림 2. 연구 모형

(1) Logistic Regression

Logistic Regression 모델은 이진 분류 모델로서 회귀분석의 유형 중 하나로, 종속 변수와 독립 변수들 간의 인과관계를 Log odds 와 Sigmoid Function 을 활용하여 추정하는 통계 기법이다. 일반 회귀 분석 모델과 다르게 Logistic Regression 은 범주형 데이터를 종속 변수로 설정하며, 이는 사건이 일어날 확률의 추정이 직접적으로 가능하다. Odds 는 로지스틱 회귀분석에서 임의의 이벤트에 대하여 어떤 요인에 의해 발생하지 않을 확률 대비 발생할 확률을 의미하며, Odds Ratio 는 특정 요인 여부에 따른 이벤트 발생 확률을 비교할 때 사용되는 척도이다.[4] 모형 추정에 사용하는 방법은 최대 우도 추정법으로

[4] Aurelien Geron, [Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow], (한빛미디어, 2018. 04), 130p

(Maximum Likelihood Estimation), 회귀 계수를 우도(likelihodd)가 최대가 되는 지점을 재귀적인 방식으로 추정한다.

$$odds = \frac{probability\ of\ event\ occurrence}{probability\ of\ event\ non-occurrence}$$
$$odds = \frac{p}{(1-p)} \quad (1)$$

$$odds\ ratio = \frac{odds\ of\ group\ one}{odds\ of\ group\ two}$$
$$odds\ ratio = \frac{p_1\ /\ (1-p_1)}{p_2\ /\ (1-p_2)} \quad (2)$$

그림 3. Odds formula

그림 4. Odds ratio formula

MLE: Maximum Likelihood Estimator

$$f(y_1, y_2, \cdots, y_n \mid \boldsymbol{x}) = f(y_1 \mid \boldsymbol{x}) f(y_2 \mid \boldsymbol{x}) \cdots f(y_n \mid \boldsymbol{x})$$
$$= Pr(y_1 = 1 \mid \boldsymbol{x}) Pr(y_2 = 0 \mid \boldsymbol{x}) \cdots Pr(y_n = 1 \mid \boldsymbol{x})$$
$$= \left(\frac{e^{\beta_0+\beta x_i}}{1+e^{\beta_0+\beta x_i}}\right)^k \left(\frac{1}{1+e^{\beta_0+\beta x_j}}\right)^{n-k} \quad 단,\ \Sigma y_i = k, \Sigma(1-y_j) = n-k$$

Find β by Recursive method

그림 5. Maximum Likelihood Estimator formula

(2) RandomForest

RandomForest는 기본 구성 요소인 Decision Tree 기반의 앙상블 학습 모델로서 여러 개의 Decision Tree 모델에 동일한 데이터를 삽입한 후, Training dataset 의 중복을 허용하여 무작위로 샘플링하는 Bagging 방법을 적용하여 최적의 분류 모델을 채택한다. 그리고 RandomForest 알고리즘은 트리의 노드를 분할할 때 전체 특성 중에서 최선의 특성을 찾는 대신, 무작위로 선택한 특성 후보 중에서 최적의 특성을 찾는 식으로 무작위성을 더 주입하여, 이를 통해 더욱이 다양한 Decision Tree 모델을 만들고, 편향을 손해보는 대신 분산을 낮추어 전체적으로 더 훌륭한 모델이 탄생된다.[5] 하기 그림과 같이, RandomForest 는 여러 개의 Decision Tree 모델 중 좋은 예측력을 가진 나무를 Majority voting 을 통해 선정하여 최종 예측기를 생성하는 단계를 거친다. 여기서 Decision Tree 란 각 독립 변수에 대해서 임계값을 통해 샘플을 분류하고, 불순도를 측정하여 특성에 따라 데이터를 분류하는 모델을 의미한다. 임계값(Cut off)는 RandomForest 에서 통상적으로 이용하는 Gini 계수의 미분을 통해 최적의 임계값을 계산할 수 있으며, 불순도가 0 이 될 경우 과적합이 발생되므로 Gini 불순도와 확률에 따른 최적의 임계값을 찾는다. [6]본 연

[5] 홍기혜, “A predictive Model for Suicidal Ideation of Adolescents Using RandomForests Machine Learning Algorithm”, [한국사회복지학회], (2020), p157-180

[6] Aurelien Geron, [Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow], (한빛미디어, 2018), p229-243

구에서 RandomForest 를 사용한 이유는 특성 간 상호작용을 고려하는 데 적합하며, 이를 활용하여 특성 중요도를 파악할 수 있기 때문이다.

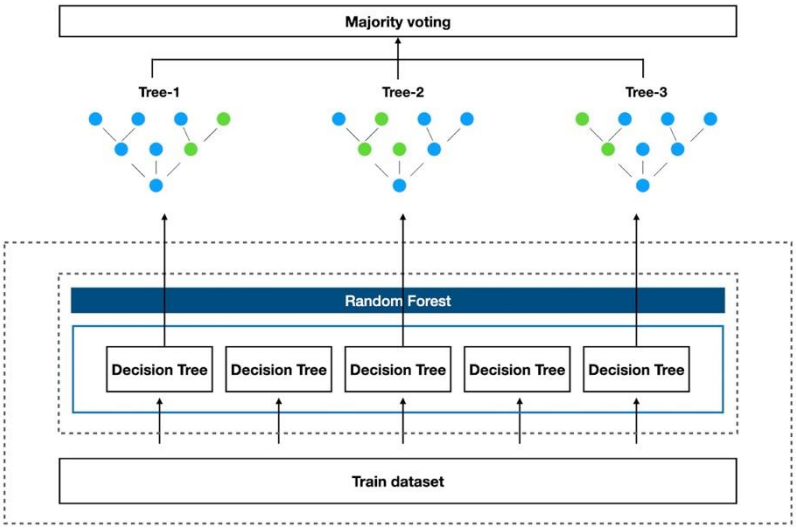


그림 6. RandomForest

(3) Mutli Layer Perceptron (MLP)

Multi Layer Perceptron 은 기본적인 인공 신경망 구조 중 하나로서 생물학적 뉴런에서 착안된 모델이다. 입력층(Input layer)에 학습 데이터를 삽입하면, 데이터가 층을 통과할 때마다 계산되는 가중치와 편향에 의해 출력층(Output Layer)에 유의미한 예측값을 계산한다. MLP 는 입력층, 은닉층, 그리고 출력층의 순서로 구성되어 있으며 층 수 및 Node 수, 활성화 함수 등을 조정하여 다른 층과 연결하여 가중치와 편향을 계산하고, 오차 역전파 방식을 사용하여 가중치와 편향이 최신화한다. 여기서 오차 역전파란, 순차적 방식에 의해 출

력층에 계산된 가중치와 편향에 의한 총 오차값을 미분하여 다시 반대 방향으로 가중치와 편향을 계산하는 방법이다.[7]

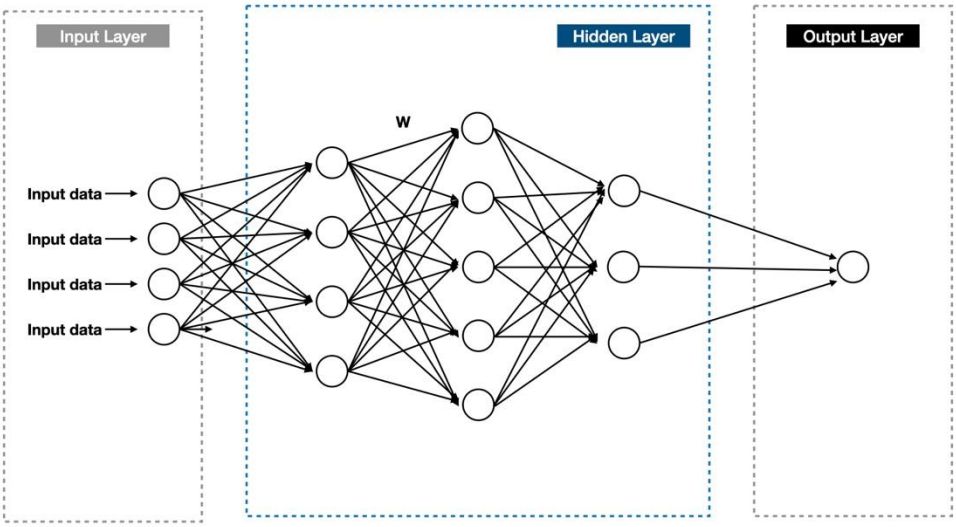


그림 7. Multi Layer Perceptron

[7] Aurelien Geron, [Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow], (한빛미디어, 2018), p412-459

제 III 장 연구 방법

제 1 절 데이터 분석 및 전처리

기업 퇴사자 및 이직자 분류를 위한 10 개의 독립 변수는 도시 발전도, 성별, 경력, 학력, 전공, 직무 경험, 기업 크기, 직무 훈련 시간 등으로 구성되어 있으며 각 데이터의 특징은 하기 표와 같다. 우선 City Development(도시 발전도)는 도시의 발전 정도를 나타내는 지표이며, 1 에 가까울수록 인프라, 산업, 교육 등 전반적인 도시 발전도가 타 도시에 비해 우수함을 나타낸다. 두 번째로 Gender(성별)은 0 은 남성 1 은 여성으로 분류되어 있으며, 기재 되지 않은 공란은 2로 구분하였다. 세 번째, Relevent experience(관련 경험)은 임직원이 담당하고 있는 해당 업무에 대해 관련 경험이나 업무 경험이 있을 경우 1 로 표시되며, 반대로 관련 경험이 전무할 경우 0 으로 표시한다. Enrolled university 는 전공 관련한 대학 강의를 들은 시간에 따라 분류된 데이터이며, 들은 적이 없을 경우 0, Part time 으로 청강하였을 경우 1, Full time 일 경우 2 로 정의 된다. Education level(학력)은 임직원의 학력을 나타내며 초등학교 졸업은 0, 고등학교 졸업은 1, 학사 졸업은 2, 석사 학위 취득은 3, 박사 학위 취득은 4, 그리고 기재 되어 있지 않은 경우 5 로 표기된다. Major discipline 은 학과 전공을 나타내며, 전공이 없는 경우 0, 과학 기술 공학 수학 관련 전공을 한 경우 1, 경영학 관련 전공은 2, 인문학 관련 전공은 3, 예체능 관련 전공은 4, 그 외 전공은 5 로 표기된다. Last new job 은 이전 직장

수를 의미하며 첫 직장인 경우 0 으로 나타내고 4 개 보다 초과된 이전 직장 수를 보유한 경우, 이상치로 판단하여 4 로 표기하였다. 마지막으로 Training hours 는 훈련 시간을 뜻하며 단위는 시간이다. Gender, Education level, Enrolled university, Company size, Major discipline 데이터들은 범주형 데이터로서 Scikit-learn package 의 OrdinalEncoder Preprocessing 기능을 사용하여 데이터 전처리 과정을 거쳤다.

항목	설명	범위 및 지정
City development index	Index	0 ~ 1
Gender	Male	0
	Female	1
	Other	2
Relevent experience	No relevent experience	0
	Has relevent experience	1
Enrolled university	No enrollment	0
	Part time course	1
	Full time course	2
Education level	Primary school	0
	High school	1
	Graduate	2
	Masters	3
	Phd	4
	Unknown	5
Major discipline	No major	0
	STEM	1
	Business degree	2
	Humanities	3
	Arts	4
	Others	5
Experience	< 20	20
Company size	The number of employees	< 10
		10 ~ 49
		100 ~ 500
		500 ~ 999
		1000 ~ 4999
		5000 ~ 9999
		> 10000
Last new job	> 4	4
	Never	0
Training hours	Continuous variable	

표 2. Data prepROCessing

Decision Tree 분석에 다중공선성을 확인하는 데에는 아직 여러 의견이 나뉘지만, 본 연구에는 Logistic Regression 분석 기법이 포함되어 있기에 독립변수간 다중공선성(Multicollinearity)을 확인한다. 다중공선성이란 회귀 분석에서 사용되는 독립 변수가 다른 독립 변수와 상관 정도가 높아, 데이터 분석 시 부정적인 영향을 미치는 현상을 의미하며 이는 회귀분석 전제 가정을 위배하는 것이므로 전처리가 필요하다.[8] 다중공선성 판단에는 산점도(Scatter plot)와 상관계수(Correlation coefficient), 그리고 분산팽창지수(Variance Inflation Factor) 등 여러 방법이 존재하지만 본 연구에서는 명목형 독립 변수가 3 개 이상이고, R-squared 값이 정확히 산출되지 않으므로, 상관계수 그래프를 통해 분석할 것이다.

8 Aurelien Geron, [Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow], (한빛미디어, 2018), p145

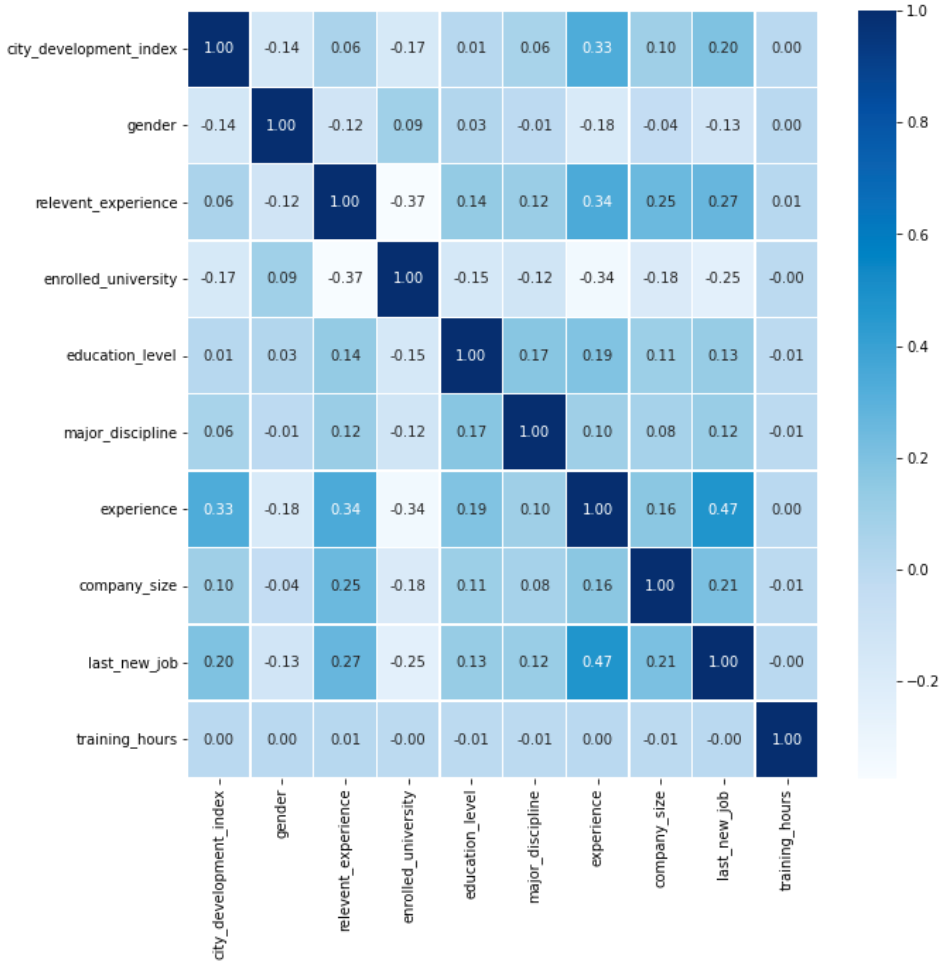


그림 8. Correlation coefficient

상관계수는 −1 에서 1 사이의 값을 가지며, 0 으로 수렴할수록 선형 관계가 전혀 없다는 것을 의미한다. 본 연구에서는 독립 변수 간 뚜렷한 양적 선형 관계는 0.7 에서 1.0 사이로 간주하나, 데이터의 특성과 샘플의 대표성 등 상황에 따라 상관계수 값 자체를 해석하는 데 있어 정확한 기준은 없으며, 명목형 변수가 3 개 이상 존재하므로 완벽한 다중공선성 판단의 기준에는 한계가 있

다. 본 분석의 목적은 독립 변수 간 선형성이 뚜렷한 변수를 파악하는 것이며, 분석 결과 뚜렷한 선형성이 나타나는 변수는 존재하지 않는다.

제 2 절 연구 모델 설계

Logistic Regression

본 연구에서 Logistic Regression 분석은 총 두 가지 분석 방법을 활용하여 설계하였다. 하나는 로지스틱 회귀 분석의 대표적 머신러닝 알고리즘인 Sikit learn package 를 사용했으며, Hyperparameters 설정은 l2 규제에 값이 감소할수록 규제가 강해지는 C(Inverse of regularization strength)는 default 1.0 값에서 10.0 으로 상향시켰다. 규제를 완화한 이유는 Train dataset 의 데이터 량이 충분하지 않으므로, 과적합의 우려가 있기 때문이다. 또한, Train dataset 이 작으므로, 경사하강법을 선택하는 solver parameter 에서 소규모 dataset 에 유리한 liblinear 을 선택하였다. 마지막 다른 분석 방법은 통계 결과를 효율적으로 나타내주는 statsmodels package 를 활용하였다.

Hyperparameters	Figure size
Penalty	l2
C	10.0
solver	liblinear

표 3. Logistic Regression hyperparameters

RandomForest

RandomForest 또한 Logistic Regression 모델과 마찬가지로 전처리 과정에서 불필요한 city, enroll_id, company_type 데이터는 제외하고 명목형 데이터는 더미화 과정을 통해 범위를 분류하여 Train data 를 구성하였다. 모델을 설계함에 있어 여러 Hyperparameter 조정을 통한 Accuracy 를 비교하기 위해 각 Decision tree 모델 개수(n_estimators)와 리프 노드 개수(max_leaf_nodes)를 다르게 구성하였다.[9] n_estimators 는 각 100, 1000, 3000, 7000 단위로 구성하였으며 리프 노드 최대 개수는 각 50, 100, 150, 250 으로 구성하였다. 모델 함수에는 {n_estimators : max_leaf_nodes}로 쌍을 이루어 각 하나씩 순차적으로 대입하였다. 그중 최상의 Accuracy 가 나온 n_estimators 와 max_leaf_nodes 를 선택한다.

RandomForest 특징상 특성의 상대적 중요도를 측정하기 쉬우므로, Sikitlearn 패키지에서 지원하는 permutation_importance method 를 통해 이전에 구했던 최상의 모델을 대입하여 특성 중요도를 파악하고 시각화한다.

9 권수빈, 김종원, 전영빈, 차수진, 김부식 강태원, “Analyzing employee resignation through data mining”, [한국정보기술학회], (2021), p834-837

The best hyperparameters	Figure size
max_depth	2
max_leaf_nodes	250
min_samples_leaf	2
min_samples_split	2
n_estimators	7000

표 4. The best hyperparameters

Multi Layer Perceptron(MLP)

인공신경망인 다층 퍼셉트론을 구현하기 앞서, 가중치의 편향을 방지하기 위해 이상치를 조정하고, 각 데이터 열을 입력층에 넣기 위하여 Sikitlearn 패키지의 StandardScaler method 를 활용하여 전체적인 스케일을 조정하였다. 조정된 Train dataset 안에서 1000 개의 Validation dataset 을 추출하였으며, MLP 를 구성하는 데에 Accuracy 를 향상시키기 위하여 총 3 가지의 방법을 동원하였다. 기본적인 MLP 모델을 설계할 때에는 입력층, 은닉층, 출력층을 순차적으로 쌓아올리는 Sequential API 를 활용한 다층 퍼셉트론 구성 방식과 순차적이지 않은 다층 퍼셉트론이 존재한다. Wide & deep 신경망처럼 입력층을 두 개로 분할한 후, 하나는 은닉층을 거쳐 출력층으로 연결되고, 나머지 하나는 은닉층을 거치지 않고 바로 출력층으로 연결되어 입력층에서 출력층까지 가는 경로를 서로 다르게 작성하는 방식인 함수형 API 를 활용한 다층 퍼

셉트론 구성 방식을 예로 들 수 있다.[10] 본 연구에서는 기본적인 Sequentail API 를 활용한 MLP 모델과, Wide & deep 신경망을 사용하는 함수형 API 를 활용한 MLP 모델 둘 모두를 사용했으며, 마지막으로 둘 모델의 최적의 하이퍼 파라미터를 찾기 위해 Sikitlearn 패키지의 GridSearchCV method 를 동원하여 최적의 하이퍼 파라미터를 탐색하였다. 최적의 하이퍼 파라미터를 구하는 과정에서는 GridSearchCV 뿐만 아닌, 가중치와 편향의 스케일을 조정하는 활성화 함수와 최상의 모델을 기록하여 저장하는 Call back method, 손실함수 등 여러 매개 변수를 조정하여 Accuracy 향상에 기여하였다.

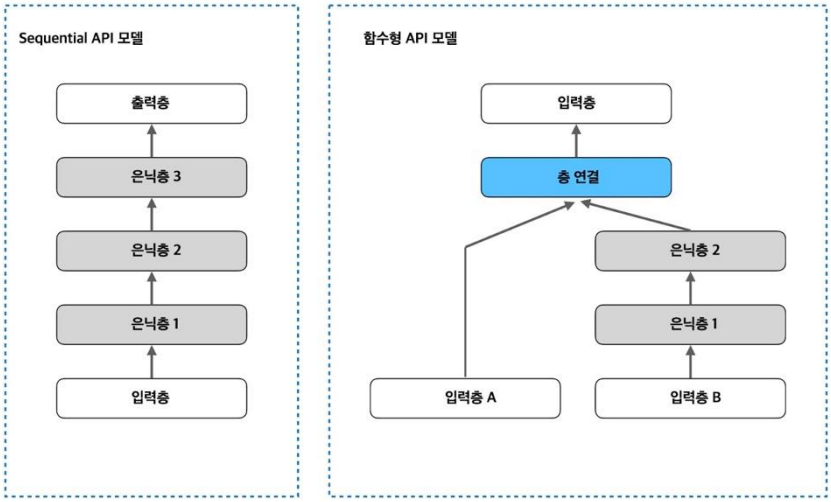


그림 9. Sequential API 모델과 함수형 API 모델

[10] Aurelien Geron, [Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow], (한빛미디어, 2018), p412-459

Hyperparameters	Figure size
Activation	relu
Output activation	sigmoid
hidden_layer_1	500
hidden_layer_2	300
hidden_layer_3	100
max_iter	3000
early_stopping	TRUE
solver	adam

표 5. The best hyperparameters of Multi Layer Perceptron

위의 표 4. The best hyperparameters of Multi Layer Perceptron 을 기준으로 Sequential API 모델과 함수형 API 모델에 적용시켰으며, hidden layer 의 활성화 함수로서는 relu 를 활용하였다[11]. 그 외 마지막 모델은 기본적인 Sequential 형식의 Multi Layer Perceptron 모델이지만, 심층 신경망의 활성화 함수와 옵티마이저를 조정하여 입력하였다. 활성화 함수로서는 현재 가중치 소실 및 폭주를 방지하는 데 유용한 Elu 활성화 함수와 초기화 전략으로서의 He 초기화 전략을 사용했으며, 훈련 속도를 높일 수 있는 고속 옵티마이저인 모멘텀 최적화(Momentum optimization)을 설정하여 훈련 속도를 향상시켰다[12].

[11] 김수주, 악푸도 우고추쿠 에지크, 허장욱, “A study on Fault Classification of Solenoid Pumps based on Multi-Layer Perceptron”, [한국신뢰성학회], (2021), p12-19

[12] Aurelien Geron, [Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow], (한빛미디어, 2018), p412-459

기본적인 은닉층 활성화 함수로 쓰이는 ReLU 함수는 f(x) = max(0, x) 형태의 간단한 수식으로 0 이하는 0 으로 고정하고 0 을 초과할 경우 해당 값을 그대로 출력하는 단순한 구조이다. [13]하지만 이러한 형태 때문에 층이 깊어질수록 가중치가 소실되거나 폭주하는 현상이 발생하기 때문에 이를 극복하고자 여러 활성화 함수가 연구되었으며, 본 연구에서는 ELU 활성화 함수를 활용하여 0 이하의 가중치 지수 함수를 활용하여 계산한다.

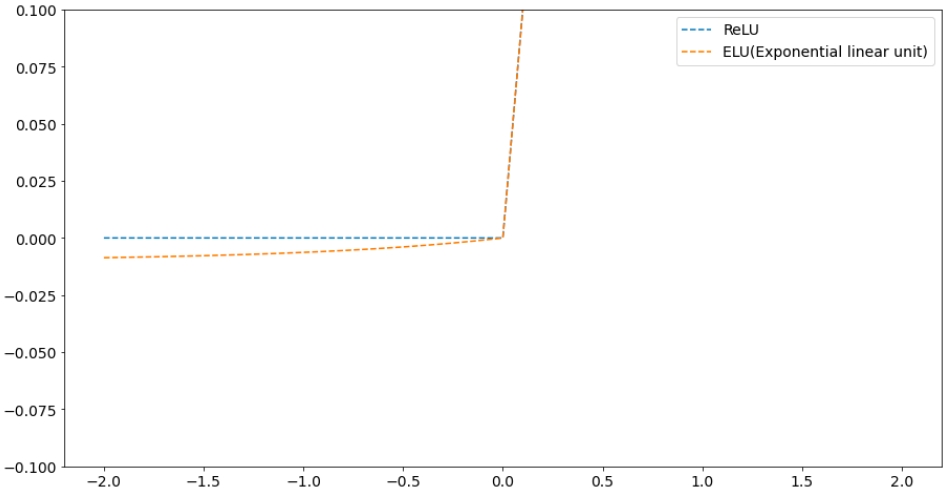


그림 10. relu & elu activatino functions

인공신경망(Artificial Neural Network) 학습에 주요 매개변수로서 중요한 가중치(Weigths)와 편향(bias)은 초기에 어떤 값을 갖고, 어떤 활성화 함수를 쓰느냐에 따라 최적화의 결과가 달라진다. 이에 따라, 학습을 통해 가중치를

13 Aurelien Geron, [Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow], (한빛미디어, 2018), p412

갱신할 때 가중치의 초기값이 중요한 변수로 작용한다. 하지만 특정 초기값을 알 수 없기에 랜덤한 특정 값으로 초기화를 하는 것이 합리적이며 절대값이 너무 크다면 시그모이드 활성화 함수에서는 기울기 소실 문제(Gradient vanishing)가 발생하고 반면 음수가 나온다면 일반적인 ReLU 활성화 함수에서 가중치가 소실될 수 있다. 본 연구에서는 은닉층의 ReLU 활성화 함수를 적용했을 때 발생할 수 있는 기울기 소실 문제를 개선하고자, He 초기화 전략(He initialization)을 사용하여 층이 깊어져도 층의 초기값이 고르게 분포되도록 하였다.

앞서 언급한 ELU 활성화 함수를 통한 가중치 개선은 심층 신경망의 훈련 속도를 저하시킬 가능성이 있으므로, 표준적인 경사하강법 옵티마이저 대신 더 빠른 속도의 모멘텀 최적화(Momentum optimization)을 활용하였다. 모멘텀 최적화(Modentum optimization) 방식은 사전 훈련된 네트워크의 일부를 재사용하여 새로운 가중치의 최신화가 아닌 이전 가중치의 최신화를 통해 훈련 속도를 크게 증가시키고 옵티마이저가 최적값에 빠르게 수렴되도록 한다.

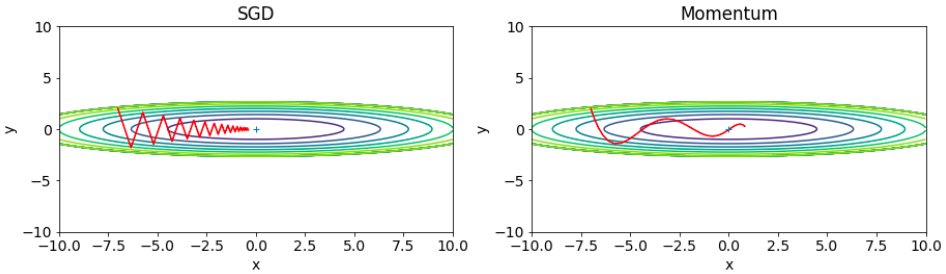


그림 11. Momentum optimization

제 IV 장 연구 결과

예측값과 실제값의 일치 또는 불일치 정도를 파악하고, 예측 성능을 평가하기 위해 본 연구에서는 모델을 평가하는 성과지표로서 정분류율(Accuracy), 민감도(Recall), 특이도(Specificity), 정밀도(Precision), 그리고 ROC curve(Receiver Operating Characteristic)를 통해 성능을 평가하고 해당 값이 클수록 모델의 예측력이 높음을 의미한다. 정분류율(Accuracy)란 전체 관측치 중에서 실제값과 예측값의 일치 정도를 나타내며, 민감도(Recall)는 모델의 실제값이 True 인 관측치 중에서 모델이 Positive 라 분류한 비율을 나타낸다. 정밀도(Precision)는 모델이 Positive 로 예측한 관측치 중에서 실제값이 Positive 인 비율을 의미한다. [14] 마지막으로 ROC curve(Receiver Operating Characteristic)는 실제값이 True 인 관측치 중 예측치가 적중한 정도를 나타내는 TPR(True Positive Rate)과 1 에서 실제값이 False 인 관측치 중 예측치가 적중한 정도를 나타낸 값을 뺀 값인 FPR(False Postivie Rate)의 비율을 나타낸 그래프로서, 그래프 상 나타나 있는 ROC curve 가 좌상단 즉, TPR 에 가까울수록 더 좋은 분류기임을 의미한다.

[14] 박연정, 이도길, “Development of a Resignation Prediction Model using HR data”, [한국정보통신학회], (2021), p100-103

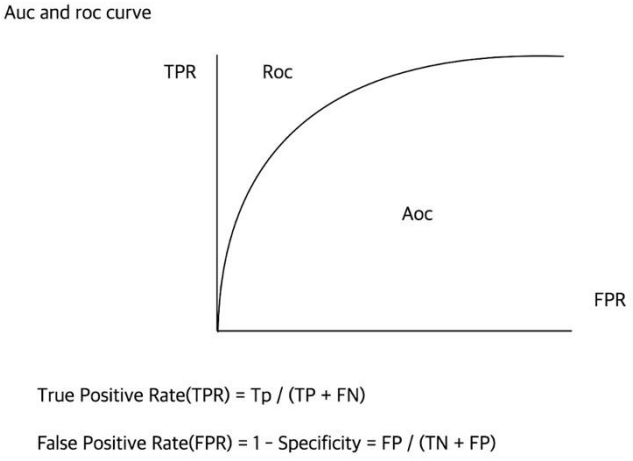
		Predicted	
		TRUE	FALSE
Actual	TRUE	True Postivie (TP)	False Negative (FN)
	FALSE	False Positive (FP)	True Negative (TN)

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

표 6. Confusion matrix



True Positive Rate(TPR) = Tp / (TP + FN)

False Positive Rate(FPR) = 1 - Specificity = FP / (TN + FP)

그림 12. Auc - ROC curve

제 1 절 Logistic regression

두 가지 모델링을 통해 분석한 결과, 결과값이 서로 상이하게 도출되었다. 우선 Sikit learn package 를 활용한 Logistic Regression 분석 결과, 정분류율(Accruacy): 0.78, 정밀도(Precision): 0.66, 민감도(Recall): 0.28 로 나타났다. 일반적으로 정밀도와 민감도는 반비례 관계로서 정밀도는 모델 예측 결과 중 실제값이 얼마나 포함되어 있는지를 나타내며, 민감도는 실제값 중에서 모델이 얼마나 예측을 잘 했는지 나타내는 지표가 된다. 두 지표 모두 모델의 성능과 연관되므로 어느 한 값을 기준으로 모델 자체를 평가하는 것은 올바른 방법은 아니나, 정밀도에 비해 민감도가 낮아 해당 모델의 실제값 분류 성능 및 유의성 검증 성능은 한계가 있다고 사료된다. AUC ROC curve 의 경우, AUC 값은 0.744 이며 분석 결과는 아래 자료와 같다.

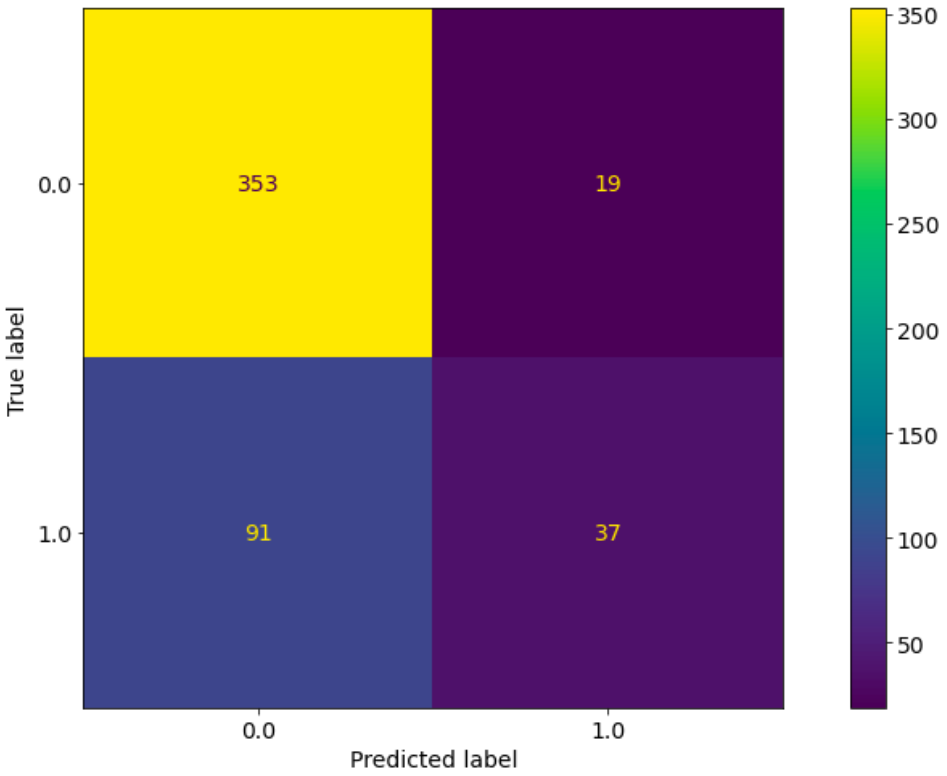


그림 13. Logistic Regression confusion matrix

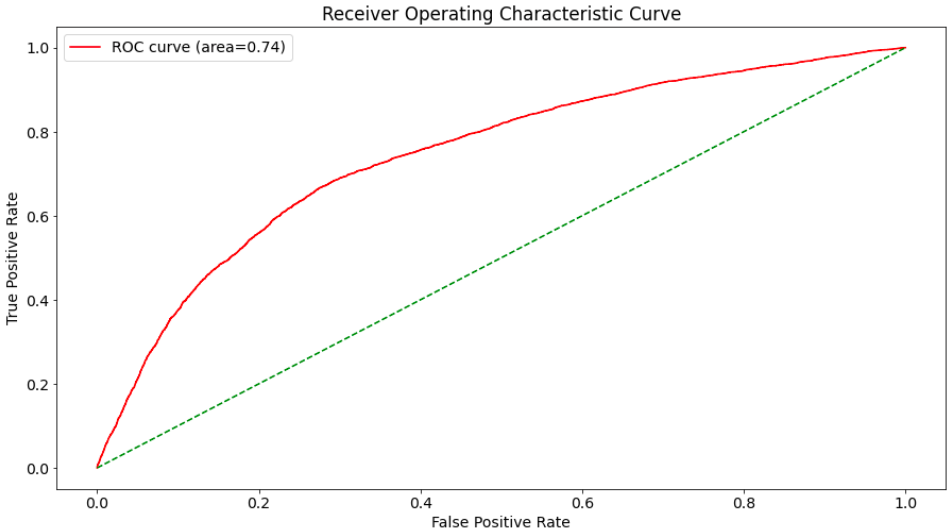


그림 14. Logistic Regression model ROC curve

Scikit learn package 모델링을 토대로 특성 중요도를 파악하기 위하여 각 독립 변수의 가중치를 추출하였다. 특성 중요도 가중치 분석 결과, 뚜렷한 상관관계를 나타내는 특성은 도시 발전도(city_development_index) 외에 관찰되지 않았다.

Column name	Coefficient
enrolled_university	0.164359
major_discipline	0.163300
last_new_job	0.086243
gender	0.039291
education_level	0.037578
training_hours	-0.000925
experience	-0.018515
company_size	-0.136476
relevent_experience	-0.345226
city_development_index	-5.641044

표 7. Logistic Regression feature importance with Scikit learn

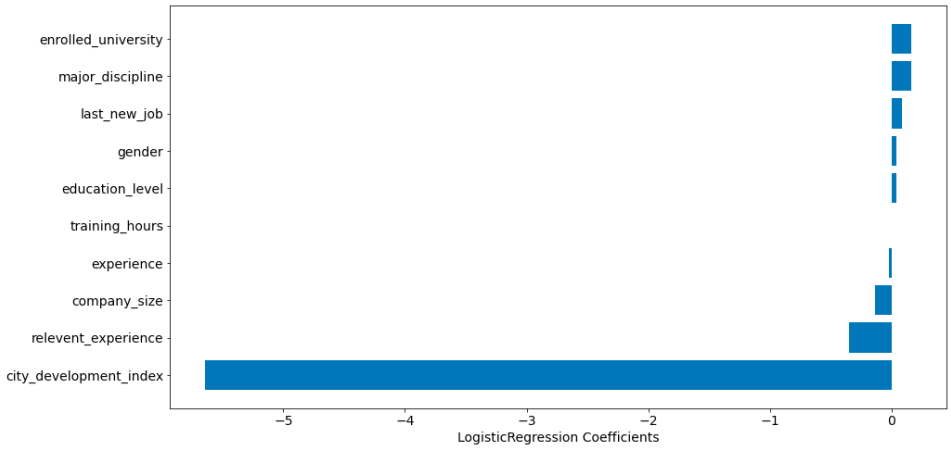


그림 15. Logistic Regression coefficients with Scikit learn

두 번째로 Statsmodel package 를 통해 분석한 결과, P 값이 유의 수준 0.05 보다 벗어난 relevent_experience 와 training_hours 를 제외하고 모두 유의 수준에 포함되므로 통계적으로 유의하다고 판단할 수 있다. 기존 Scikit learn 을 통해 모델링한 가중치 값과는 상이하나, 기존 Scikit learn package 의 모델링에서 하이퍼 파라미터를 조정하며 생긴 변화로 사료된다. 분석 결과는 아래 표와 같다.

```
Optimization terminated successfully.
         Current function value: 0.509954
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                19158
Model:                          Logit   Df Residuals:                    19148
Method:                           MLE   Df Model:                            9
Date:                Thu, 30 Jun 2022   Pseudo R-squ.:                 0.09199
Time:                        08:21:11   Log-Likelihood:                -9769.7
converged:                       True   LL-Null:                       -10759.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
city_development_index          -2.2338      0.079    -28.434      0.000      -2.388      -2.080
gender                           0.1396      0.020      7.028      0.000       0.101       0.179
relevent_experience             -0.0272      0.042     -0.646      0.518      -0.110       0.055
enrolled_university              0.3484      0.022     15.965      0.000       0.306       0.391
education_level                  0.2588      0.021     12.549      0.000       0.218       0.299
major_discipline                 0.1866      0.019      9.620      0.000       0.149       0.225
experience                      -0.0305      0.004     -8.659      0.000      -0.037      -0.024
company_size                    -0.1260      0.009    -14.757      0.000      -0.143      -0.109
last_new_job                     0.0906      0.015      6.005      0.000       0.061       0.120
training_hours                   0.0003      0.000      1.138      0.255      -0.000       0.001
================================================================================================
```

표 8. Logistic Regression result with Statsmodel

Column name	Coefficient
enrolled_university	0.3484
major_discipline	0.1866
last_new_job	0.0906
gender	0.1396
education_level	0.2588
training_hours	0.0003
experience	0.0305
company_size	-0.1260
relevent_experience	-0.0272
city_development_index	-2.2338

표 9. Logistic Regression feature importance with Statsmodel

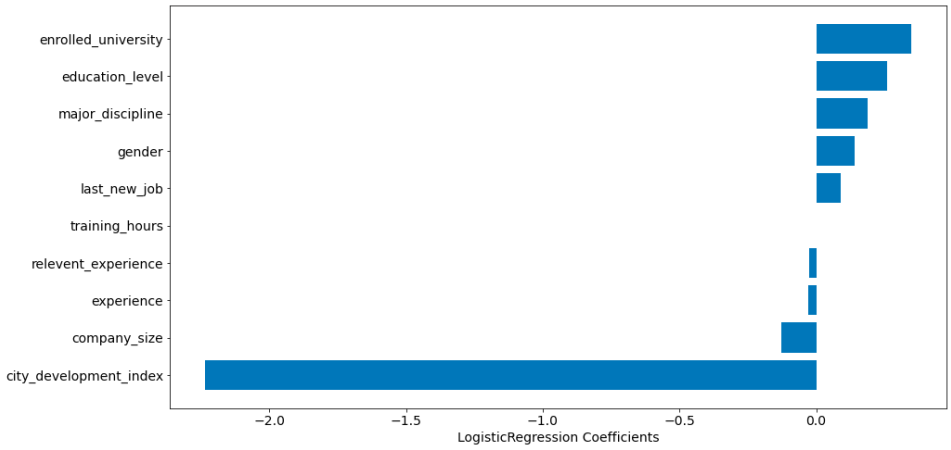


그림 16. Logistic Regression coefficients with Statsmodel

분석 결과 Scikit 을 활용한 모델과 Statsmodel 을 활용한 모델 모두 enrolled_university, education_level, major_discipline, gender, last_new_job 모두 양의 상관관계를 나타냈으며, 그중 enrolled_university 변수가 양의 상관관계 중 가장 영향력 있는 요인으로 나타났다. training_hours 는 양 모델 모두 상관계수가 0 이며 이는 종속 변수와 상관관계가 없음을 의미한다. 마지막으로 유의 수준을 벗어난 relevent_experience 를 제외하고, expereicne, company_size, city_development_index 모두 음의 상관관계가 나타났으며, city_development_index 독립 변수가 두 모델에 모두 가장 음의 상관관계가 높은 변수로 분석되었다.

제 2 절 RandomForest

최적의 Hypterparameters 를 탐색하기 위해 매개 변수 n_estimators 와 max_leaf_nodes 를 점진적으로 증가시켜 대입한 후 각 Accuracy 를 측정하였다. 기존 RandomForest 모델 설계에서 제시했던 표 3 의 The best hyperparameteres 의 값을 대입한 결과, 정분류율(Accuracy)은 0.82 로 예측 모형을 신뢰할 만한 수준이라고 나타났다. Confusion matrix method 를 활용하여 측정한 성능 지표는 민감도(Recall): 0.57, 정밀도(Precision): 0.67 로서 실제값을 틀리는 경향보다 예측값을 틀리는 경향이 더 높은 것으로 확인됐다. AUC ROC curve 분석 결과는 아래 자료와 같으며, AUC 값은 0.86 으로 분석됐다.

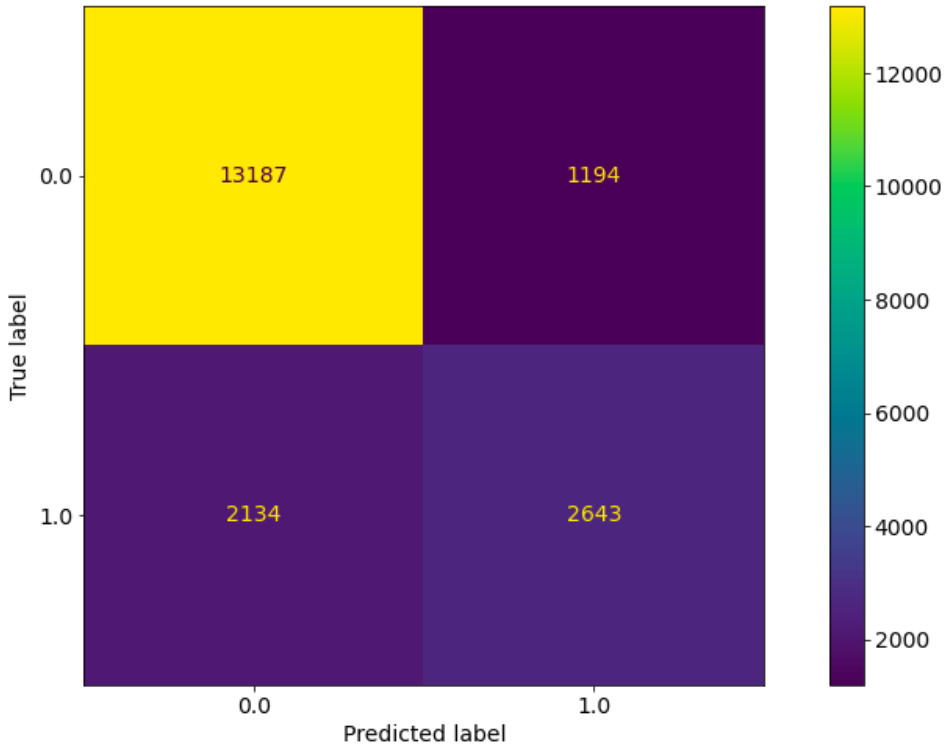


그림 17. RandomForest confusion matrix

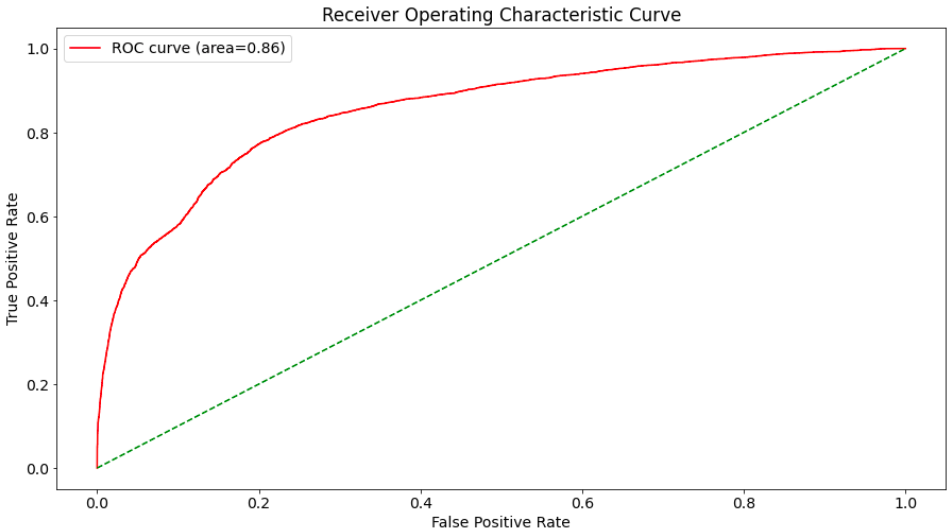


그림 18. RandomForest ROC curve

The best hyperparameters 를 대입한 RandomForest 특성 중요도 분석 결과 아래 자료와 같이 나타났다. RandomForest 의 특성 중요도 분석표는 Logistic Regression 의 상관관계 분석표와 다르게 독립 변수와 종속 변수의 상관관계를 나타내는 것이 아닌, 종속 변수에 대한 각 독립 변수의 중요도를 나타낸다. 그러므로, RandomForest 의 특성 중요도를 Logstic regression 의 가중치 즉, 상관관계에 반영하여 퇴사자 분류 요인을 분석해야 하는데, 아래 RandomForest 특성 중요도의 독립 변수 중요도 순서와 Logistic Regression 의 가중치에 따른 특성 중요도 순서가 상이한 것으로 나타난다. 이러한 불일치가 발생하는 이유는 RandomForest 의 특성 중요는 각 독립 변수가 분할될 때 불순도 감소분의 평균을 중요도로 정의되는데, 이는 각각의 node 에 관측치 개수를 고려하여 불순도 감소 정도가 계산되고, 이 값에 따라 중요도가 판단되는 것이다. (값이 클수록 중요도가 높다) 이러한 계산 방법은 빠르고 직관적이라는 장점이 있으나, high-cardinality 의 범주형 변수가 많을 경우 편향(bias)가 발생할 확률이 높다. 다시 말해, 명목형 변수끼리 계산이 될 경우 편향되어 과적합의 우려가 존재한다. 이러한 오류를 반영했을 때, RandomForest 의 특성 중요도는 city_development_index 와 company_size 와 같이 중요도 계산 값이 큰 변수의 경우, 편향의 우려를 위해 보다 낮은 가중치를 부여한다.

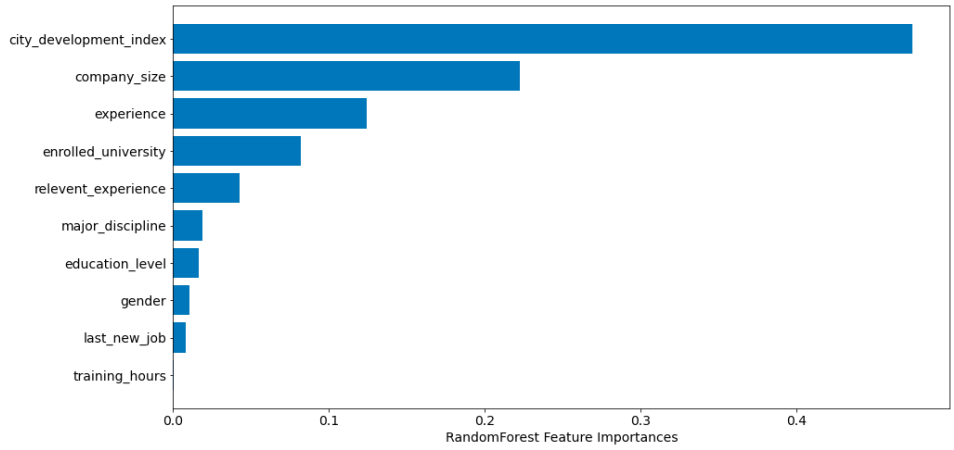


그림 19. RandomForest featrue importance

제 3 절 Multi Layer Perceptron

임직원 퇴사 및 이직 분류를 위한 모델 설계로서 총 3 가지의 방식으로 진행한 Multi Layer Perceptron 중 기본적인 Sequential MLP 는 활성화 함수 relu 와 sigmoid 를 사용하였으며, 3000 번의 epochs 설정 결과 1503 번의 반복 실행 후에 Callback method 에 의해 조기 종료되었다. Train dataset 의 accuracy 는 0.9611, 잔차는 0.092 로 나타났으며, validation dataset 의 경우 accuracy 는 0.967, 잔차는 0.0758 로 분석되었다. Callback 의 patience 매개변수를 100 으로 설정하여 100 번의 accuracy 가 반복될 경우 조기종료하도록 설정되었기 때문에 해당 accuracy 에서 더 향상될 가능성은 낮을 것으로 사료된다. 하기 자료는 해당 모델의 1503 번의 epoch 동안의 accuracy 와 loss 를 시각화한 자료이다.

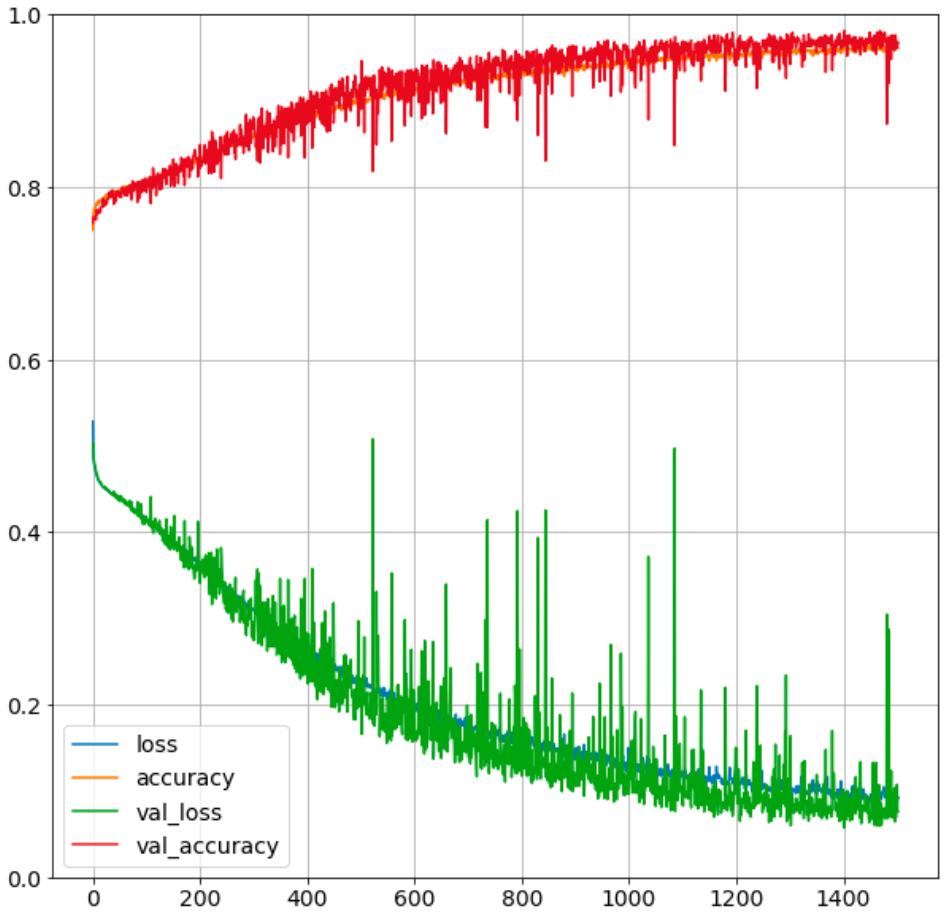


그림 20. Sequentail MLP loss and accuracy curve

다음 함수형 API 를 적용한 MLP 의 경우, 총 열두 개의 입력층을 6 개씩 두 그룹으로 분할하여 한 개의 그룹은 은닉층을 통과하여 가중치가 계산된 후 출력층에서 결과값이 도출되었으며, 나머지 그룹의 입력층은 은닉층 없이 바로 출력층에서 결과값이 도출된다. 분석 결과, 3000 번의 epochs 설정 후, callback method 에 의해 1129 번의 반복 학습 때 조기종료되었으며, train dataset 의 정확도(accuracy)는 0.7711, 잔차(loss)는 0.4719 로 나타났다.

또한, validation dataset 의 경우 정확도(accuracy) 0.7710, 잔차(loss) 0.4685 로 본 연구에서 활용한 모든 MLP 모델 중 가장 낮은 정확도(accuracy)와 잔차(loss)가 나타났다.

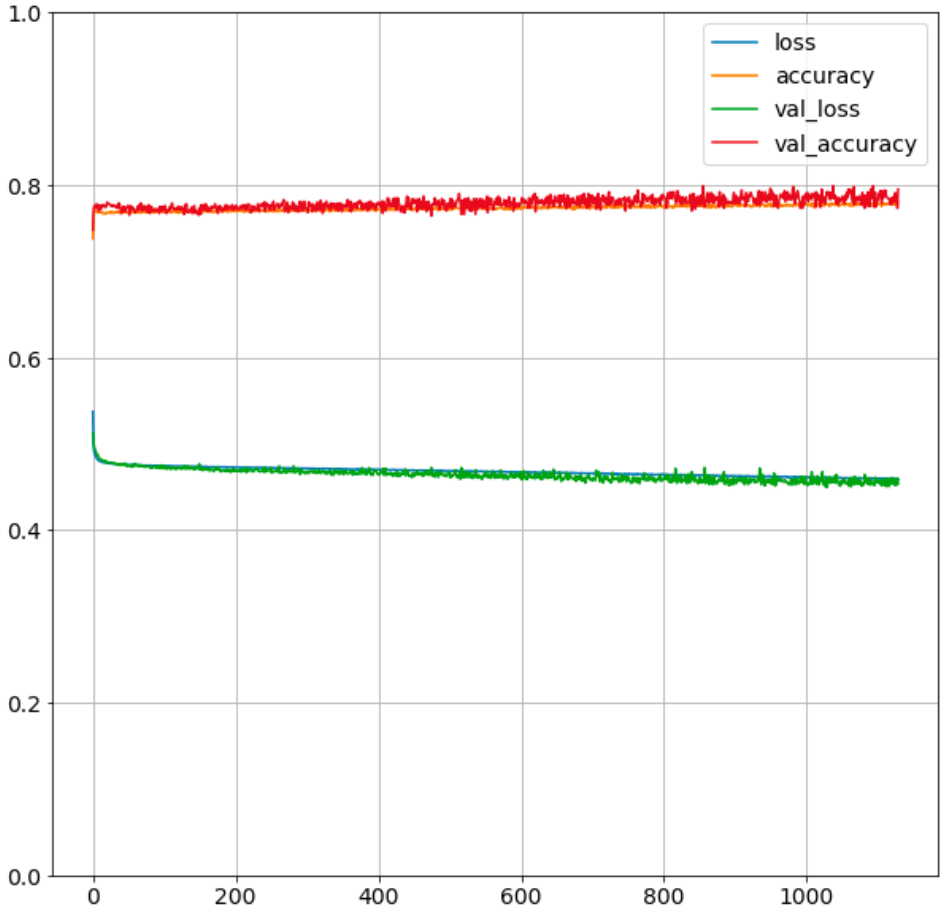


그림 21. Wide & deep MLP loss and accuracy curve

마지막으로 기본적인 Sequnetial 모델과 동일하나, 활성화 함수와 고속 옵티마이저를 조정한 모델의 결과값은 총 3000번의 epochs 중 691번 반복 학습 후 callback method에 의해 조기종료되었다. 해당 모델의 train dataset의 잔차(loss)는 0.1568, 정확도(accuracy)는 0.9263이며, validation dataset의 경우 잔차(loss)는 0.1289, 정확도(accuracy)는 0.9330으로 확인되었다. 모델 비교 분석 결과 정확도(accuracy)와 잔차(loss)는 미세하게 hyperparameter를 교정하지 않은 기본 sequentail MLP가 더 우세하였으나, validation dataset 학습 시 튀는 현상이 발생하며 이는 과적합(Over fitting)의 가능성이 있기 때문에 hyperparameters를 조정한 MLP 모델이 더 안정적인 모델로 사료된다.

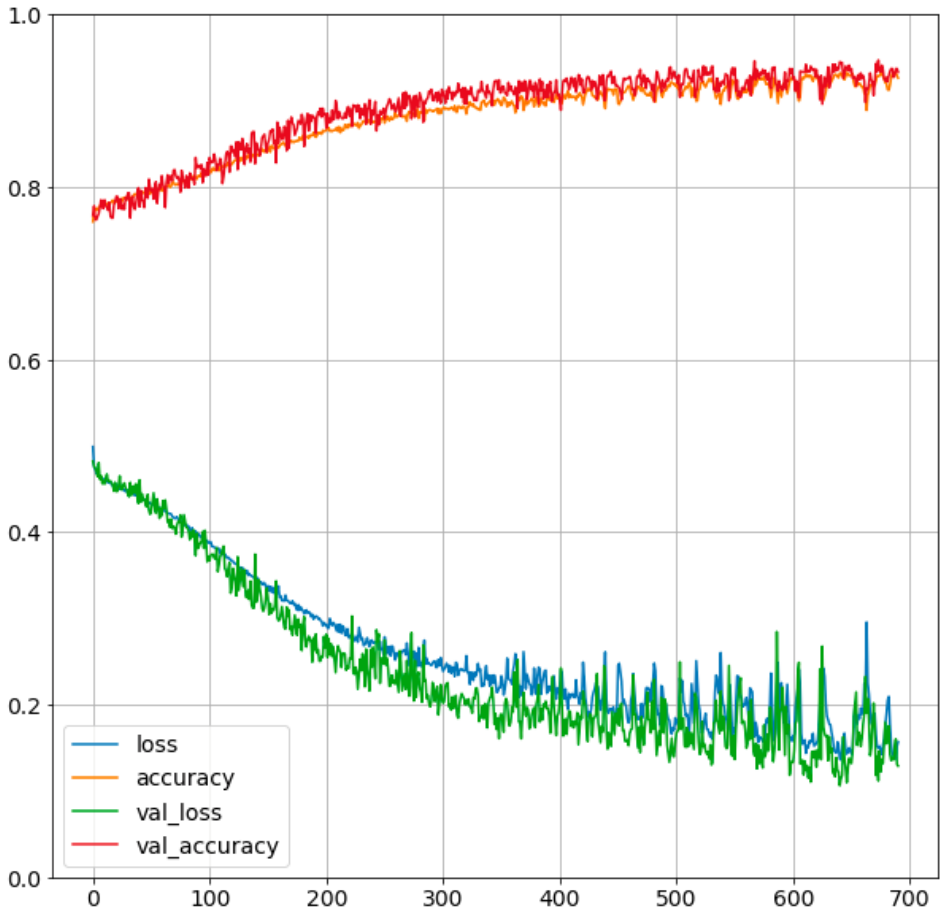


그림 22. Sequential MLP with manipulated hyperparameteres loss and accuracy curve

제 Ⅴ장 결론 및 시사점

제 1 절 요약 및 결론

본 연구는 퇴사자 예측 분류 모델 설계를 통해 퇴사 및 이직 가능성이 높은 직원을 예측 및 분류함으로써 현 기업의 인사적 관점에서의 기업지속가능가치를 평가하고 개선하는 데 의의가 있다. 우선 Logistic Regression 의 가중치(coefficients)와 RandomForest 를 활용한 특성 중요도를 반영하여 각 독립변수의 상관관계에 대한 분석은 다음과 같다. 우선 특성 중요도가 가장 높은 city_development_index 는 음의 상관관계로, 일반적으로 대도시에 속한 근무자일 경우 대도시의 특성 상 기업이 밀집되어 있으며 이에 따른 수요와 공급의 순환이 원활하여 일자리가 지방 지역에 비교하여 높으므로, 이직 및 퇴사 가능성이 높다고 판단할 수 있다. 하지만 모델 분석 결과 음의 상관관계가 강하고 뚜렷하게 나타났다는 의미는 다른 여러 요인을 고려해 볼 필요가 있다는 의미로 해석된다. 가령, 경제적 취업난에 의한 근무자들의 퇴사 혹은 이직 기피가 고조 되었을 경우와 경기 호황의 인력난으로 인한 기업의 우수 인재 확보를 위한 노력의 결과로도 예측할 수 있다. 두 번째 특성 중요도가 높은 company_size 또한 음의 상관관계로서, 이는 규모가 큰 기업에 근무하는 근무자들보다 규모가 작은 기업의 근무자들의 이직 및 퇴사가 더 활발하게 이루어진다는 의미로 유추할 수 있다. 그 다음 세 번째로 중요도가 높은 experience 을 살펴보면 위 변수들과 동일하게 음의 상관관계를 나타내며, 이

또한 타 기업의 스카우트 제의로 인한 이직 혹은 퇴사보다 장기근속에 의한 개인적 이윤이 크다고 해석할 수 있다. 또한 개인적 사유에 의한 요인도 배제할 수 없다. 다음 major_discipline 요인은 반대로 양의 상관관계를 나타내며 데이터를 살펴보았을 때 STEM 관련 전공이 majority 를 이루므로, STEM 관련 직종 종사자일수록 이직 및 퇴사 가능성이 높아진다는 의미이며 커리어 향상과 같은 개인적 사유에 의한 증가인지, 직업 환경 특성상의 원인인지는 더 깊은 요인 분석이 사료된다. 이 외에 gender, last_new_job 또한 미약하지만 양의 상관관계를 나타냈으며, training_hours 는 이직 및 퇴사 가능성에 상관관계를 미치지 않는 것으로 분석됐다.

본 연구에서 정량적 데이터의 독립 변수를 활용하여 얻은 이직 및 퇴사 예측률을 통해 이직자가 타 기업 평균 퇴사자보다 높거나 작년 자사 퇴사자 비율보다 높게 예측됐다면, 현재 기업의 독립 요인을 제외한 외부 요인 혹은 감성 요인에 의한 이직 및 퇴사율이 증가하였다는 것을 유추할 수 있다. 물론 모든 이직 요인이 포함되지 않았으며, 단순히 감성 요인에 의한 이직 및 퇴사율이 증가하였다고 단정 짓는 것은 한계가 있으나, 더 많은 데이터가 축적되고, 독립 요인이 추가될수록 더 정교한 예측 분류 모델로 발전될 가능성을 염두하여 추후 연구로 보류할 것이다. 인사적 관점에서 기업지속지능가치란 직원의 우수 역량과 업무의 노하우 등, 기업 가치 지속 및 제품 혹은 서비스 향상을 야기할 수 있다는 관점에서 인적 자원의 확보는 기업의 자산 가치 지속 및 증가에 긍정적인 영향을 미친다는 것을 선행 연구를 통해 전제할 수 있다. 결론적

으로 본 연구에서 설계한 예측 분류 모델은 세 가지 Logsitc regression, RandomForest, Multi Layer Perceptron 을 적용한 예측 분류 모델을 통해 정량적 데이터 관련 이직 및 퇴사자들의 수를 예측하여 작년 자사 이직 및 퇴사자 비율 또는 다른 기업 이직 및 퇴사자 평균 비율을 비교하여 예측률이 높다면, 현 상황을 인지하여 다음 인사 채용에 반영 혹은, 자사에 대한 임직원의 인식을 개선하고 복지 개선 및 다른 정량적 외부 요인의 개선의 노력을 통한 이직 및 퇴사율 감소를 기대할 수 있다. 마지막으로 이직 및 퇴사 데이터를 활용하여 얻은 예측 분류 모델 결과를 요약하면 다음과 같다.

(1)이직 및 퇴사 분류기를 구현하여 이직 관련 데이터를 통해 Logistic Regression, RandomForest, Multi Layer Perceptron을 기반으로 이직 및 퇴사자를 분류, 예측할 수 있다.

(2)이직 관련 데이터를 layer로 구성된 3개의 은닉층(Node 수 500, 300, 100)에 적용한 MLP 기반 예측 분류기에서 92.6%의 정확도를 나타내었다.

제 2 절 시사점

본 연구의 시사점은 여러 선행 연구에서 근거한 인사적 관점에서의 기업의 지속가능성 및 가치의 유지 및 향상을 기반으로 우수한 역량의 인재를 보존하며, 기업이 극복할 수 있는 사내의 감성적 요인 및 정량적 요인에 의한 이직 및 퇴사를 예방한다. 정량적 요인이란 연속적 자료 혹은 명목형 자료로 이루어진 범위 구분이 가능하고 수치화할 수 있는 데이터에 의한 독립 변수를 의미하며 본 연구에서는 보다 정확한 감성적 요인에 근거한 이직 및 퇴사자를 식별하기 위해, 정량적 요인에 의한 예측 분류 모델을 설계하며 다른 외부적 요인에 의한 정량적 독립 변수를 추가하며 점진적 발전이 가능한 모델 설계에 의의를 둔다.

제 3 절 연구의 한계 및 향후 연구

해당 연구에서 설계한 모델은 정량화할 수 있는 데이터를 전제로 독립변수를 구성하여 설계하였으며, 본 연구에서 지향하는 향후 연구는 정량적 데이터와 정성적 데이터 모두 독립 변수로 활용할 수 있는 모델의 설계를 목표한다. 정량화 가능한 외부 요인을 수집한 예측 분류 모델을 통해 보다 정확한 감성적 요인에 의한 이직 및 퇴사자 탐색을 목표로 하는 만큼, 정량화될 수 있는 모든 독립 변수 및 요인을 보충 및 개선에 의해 지속적인 데이터 수집

및 탐색이 요구되며, 정량적 데이터를 활용한 모델의 개선뿐만 아닌 정성적 데이터의 수집 방법과 객관화를 위한 연구가 지속되어야 한다. 본 연구는 Logistic regression, RandomForest, Multi Layer Perceptron 을 활용한 정량적 데이터의 예측 분류 모델을 설계하였으나, 감성적 데이터를 활용한 모델의 경우, 더 객관적 지표가 생성되는 데 한계가 있으며 주관이 가미된 데이터 경향성이 내포되어 있으므로, 보다 복잡한 모델 설계가 요구되며 향후 연구로서 예상하는 정성적 모형은 자사에 대한 평가 및 주관적 의견을 공유할 수 있는 Application 및 Platform 에서부터 Data crawling 을 통해 수집한 후, 긍부정 모델을 통해 자사에 대한 긍정적 의견과 부정적 의견을 분류하여 이직 및 퇴사의 독립 요인으로 활용하는 것이다.

참고문헌

[1] 김수주, 악푸도 우고추쿠 에지크, 허장욱, “A study on Fault Classification of Solenoid Pumps based on Multi-Layer Perceptron”, [한국신뢰성학회], (2021), p12-19

[2] 김영박,김형중, “Predicting Early Retirees Using Personality Data”, [한국디지털콘텐츠학회], (2018), p141-147

[3] 권수빈, 김종원, 전영빈, 차수진, 김부식 강태원, “Analyzing employee resignation through data mining”, [한국정보기술학회], (2021), p834-837

[4] 나인강, “The Effect of Turnover on the Company Performance”, [한국고용노사관계학회], (2011), vol.35, no.1, pp. 23-48 (26 pages)

[5] 박연정, 이도길, “Development of a Resignation Prediction Model using HR data”, [한국정보통신학회], (2021), p100-103

[6] 홍기혜, “A predictive Model for Suicidal Ideation of Adolescents Using RandomForests Machine Learning Algorithm”, [한국사회복지학회], (2020), p157-180

[7] Aurelien Geron, [Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow], (한빛미디어, 2018. 04)

[8] Val Bauman, “HR Analytics: predicting quitters”, [Kaggle], (2021)

Abstract

A Study on the Sustainable Value Evaluation of Enterprises Using the Predictive Classification Model of Retiree

Hwi Jong Im

Seoul School of Integrated Sciences and Technologies

Advisor: Cheong Yeul Park

The purpose of this study is to design a turnover and resignation classification model using quantitative data on resignation factors. Through previous studies, we can find that human resources are correlated with corporate sustainable value. Therefore, predicting and classifying turnover and resignation through the turnover classification models, the company can improve corporate value and enhance their products and services. Moreover, if the rates of turnover of employees is more than the deviation of mean of turnover rates, these facts infer that except for existing quantitative independent variables of turnover factors and emotional factors might be existed.