
홈 > 산업 > 모빌리티

고려아연 직원, 적대적 M&A 스트레스 심각..."이직 고려 60% 육박" 인력 유출 위기

뉴스 > #E fact > 전기·전자·통신

강지용 기자 | 입력 2024.12.02 11:17 | 댓글 0

SK하이닉스 '전 직원 핵심 기술 유출'... '법원 마이크론 이직 금지' 가처분

국정감사

한전기술 원전설계본부, 김천 이전 반대 전문 인력 이탈 심각

장철민 의원 "67명 휴직 및 연수 신청, 원전 설계 인력 유출 막아야"

이성철 기자 입력 2024.10.15 08:47

[1] [2] [3]

1. 서 론

1. 1 연구배경

오늘날 빠르게 변화하는 경영 환경 속에서 인적 자원의 중요성은 날로 커지고 있다. 특히 첨단 기술 산업을 포함한 다양한 산업 분야에서는 핵심 인력을 확보하고 유지하는 것이 기업의 경쟁력을 좌우하는 핵심 요소로 작용한다. 하지만 현실적으로 많은 기업이 우수한 인재의 이탈 문제에 직면하고 있으며, 이는 단순히 채용 및 교육에 들어간 비용 손실에 그치지 않는다.

선발된 인재가 이탈할 경우, 그 즉시 발생하는 자원의 손실뿐 아니라 장기적으로는 다음과 같은 심각한 문제로 이어질 수 있다. 첫째, 오랜 시간 동안 축적된 노하우와 경험의 유실로 인해 조직 전체의 역량이 저하될 수 있다. 둘째, 경쟁 기업으로의 인력 유출은 고급 기술이나 기밀 정보의 유출로 이어질 수 있으며, 이는 기업의 기술적 경쟁력을 위협하는 요소가 된다. 셋째, 조직 내부의 사기 저하 및 구성원 간 신뢰 관계의 약화로 인해 기업문화에도 부정적인 영향을 미칠 수 있다.

이와 같은 문제는 단순히 인사 관리 차원의 이슈를 넘어, 기업의 지속 가능성과 직결되는 전략적 과제로 인식되어야 한다. 특히 고용 시장의 유연화, MZ 세대를 중심으로 한 직무 만족도 및 워라밸(Work-Life Balance)에 대한 인식 변화, 원격 근무와 같은 새로운 근무 형태의 확산은 이탈 요인의 다양화와 예측의 어려움을 더욱 가중시키고 있다.

이에 따라 최근에는 데이터를 기반으로 한 인력 이탈 예측 시스템의 필요성이 대두되고 있다. 사내 인사 데이터, 업무 성과, 근태 기록, 조직 내 관계도 등 다양한 데이터를 활용하여 직원의 이

---

[1] https://www.todayenergy.kr/news/articleView.html?idxno=275469

[2] https://www.financialpost.co.kr/news/articleView.html?idxno=216731

[3] https://www.skyedaily.com/news/news_view.html?ID=223849https://www.todayenergy.kr/news/articleView.html?idxno=275469

탈 가능성을 사전에 예측하고, 이를 바탕으로 적절한 조직 개입과 맞춤형 대책을 마련할 수 있다면, 인재의 이탈을 최소화하고 조직의 안정성을 유지할 수 있을 것이다.

본 연구에서는 데이터를 활용한 퇴사 예측 모델을 구축함으로써, 기업이 인적 자원의 전략적 관리에 있어 보다 과학적이고 선제적인 접근을 할 수 있도록 하는 것을 목표로 한다.

1. 2 연구 목적

본 보고서에서는 직원에 대한 개인 정보, 근무 형태, 조직 내 위치, 성과 및 근태 관련 데이터 등을 기반으로 하여 직원의 퇴사 여부를 예측하고자 한다. 이를 위해 머신러닝 기법을 활용하여 퇴사 가능성이 높은 직원을 사전에 식별하고, 기업이 인재 이탈에 대해 보다 선제적이고 전략적인 대응을 할 수 있도록 지원하는 것을 목표로 한다.

또한 본 연구를 통해 어떤 변수가 퇴사 여부에 주요한 영향을 미치는지를 분석하고, 이를 기반으로 하여 퇴사에 큰 영향을 주는 특성을 가지는 직원들에 대하여 데이터 기반의 의사결정이 인사 관리에 미치는 긍정적인 영향을 분석하고, 실제 기업 환경에서 적용 가능한 예측 모델의 가능성을 제시함으로써, 효율적인 인적 자원 관리를 위한 실질적인 인사이트를 제공하고자 한다.

2. 선행 연구

연구를 진행하기 앞서 직원 이탈률과 관련하여 기존에 수행된 주요 연구들을 검토함으로써 연구의 방향성을 설정하고, 기존 연구에서의 한계점 및 개선 가능성을 도출하고자 한다. 아래에서는 두 편의 관련 논문을 중심으로 분석을 진행하였다.

2.1. 논문 1 : IT 기업 직원의 만족 및 불만족 요인에 따른 이직률 예측: 토픽모델링과 머신러닝을 활용하여[4]

해당 연구는 IT 기업 재직자들의 이직률을 예측하기 위해 잡플래닛 리뷰 데이터와 Dataguide DB의 재무 정보를 결합하여 분석을 수행하였다. 직원들의 만족/불만 요인을 추출하기 위해 토픽모델링 기법을 적용하였으며, 이를 통해 생성된 특징(feature)들과 재무 데이터를 기반으로 트리 기반의 알고리즘(Random Forest, XGBoost 등)과 서포트 벡터 머신(SVM) 등 총 6가지 머신러닝 모델을 비교 분석하였다. 본 연구는 비정형 텍스트 데이터와 정형 재무 데이터를 통합하여 예측 모델

[4] 최진욱, 신동원, and 이한준. "IT 기업 직원의 만족 및 불만족 요인에 따른 이직률 예측: 토픽모델링과 머신러닝을 활용하여." 한국데이터정보과학회지 32.5 (2021): 1035-1047.

을 구축했다는 점에서 기존의 단순 통계 분석을 넘어서는 시도를 보여주었다.

2.2 논문 2 : 머신러닝과 딥러닝 알고리즘을 활용한 간호사 이직 예측[5]

이 논문에서는 병원 간호사들의 근무 데이터를 기반으로 이직 가능성을 예측하고자 하였다. 주요 독립 변수로는 근무 환경, 업무 강도, 조직 만족도 등이 활용되었으며, 트리 기반 모델과 함께 심층 신경망(DNN)이 적용되었다. 특히 딥러닝 모델의 출력층에는 시그모이드(Sigmoid) 활성화 함수를 사용하여 이진 분류 형태로 퇴사 여부를 예측하였으며, 모델 성능 또한 비교적 우수하게 나타났다.

두 논문을 포함한 기존 연구들을 분석한 결과, 대부분의 선행 연구에서는 트리 기반 알고리즘을 중심으로 이진 분류 문제를 다루고 있으며, 일부 연구에서는 SVM이나 딥러닝 모델(DNN)을 적용하여 성능을 비교하는 경향이 확인되었다.

하지만 기존 연구들은 대부분 특정 산업(IT, 의료 등)에 국한되어 진행된 경향이 있으며, 이를 일반화하여 전 산업군에 적용 가능한 HR 이탈 예측 모델로 발전시키기에는 한계가 있다. 따라서 본 보고서에서는 다양한 산업 데이터를 포괄할 수 있는 일반화된 형태의 이탈 예측 모델을 구축함으로써, 산업 전반에서 활용 가능한 데이터 기반 HR 분석 모델의 기반을 마련하고자 한다.

3. 데이터

본 보고서에서는 Kaggle의 "Employee Attrition Classification Dataset" 데이터 셋을 사용했다. 직원의 이탈 여부를 나타내는 "Attrition"을 포함하여 24개의 column으로 되어있으며, train 데이터에는 약 60,000개의 데이터, test 데이터에는 약 15,000개의 데이터로 총 75,000명의 직원에 대한 데이터를 포함하고 있다.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Employee	Age	Gender	Years at C	Job Role	Monthly I	Work-Life	Job Satisf	Performar	Number c	Overtime	Distance f	Educat	Marital St	Number c	Job Level	Company	Company	Remote V	Leadershi	Innovatio	Company	Employee	Attrition
2	52685	36	Male	13	Healthcar	8029	Excellent	High	Average	1	Yes	83	Maste	Married	1	Mid	Large	22	No	No	No	Poor	Medium	Stayed
3	30585	35	Male	7	Education	4563	Good	High	Average	1	Yes	55	Associ	Single	4	Entry	Medium	27	No	No	No	Good	High	Left
4	54656	50	Male	7	Education	5583	Fair	High	Average	3	Yes	14	Associ	Divorced	2	Senior	Medium	76	No	No	Yes	Good	Low	Stayed
5	33442	58	Male	44	Media	5525	Fair	Very High	High	0	Yes	43	Maste	Single	4	Entry	Medium	96	No	No	No	Poor	Low	Left
6	15667	39	Male	24	Education	4604	Good	High	Average	0	Yes	47	Maste	Married	6	Mid	Large	45	Yes	No	No	Good	High	Stayed
7	3496	45	Female	30	Healthcar	8104	Fair	High	Average	0	No	38	Associ	Divorced	0	Senior	Large	75	No	No	No	Good	Low	Stayed
8	46775	22	Female	5	Healthcar	8700	Good	High	Average	0	No	2	High S	Married	0	Mid	Small	48	No	No	No	Poor	High	Stayed
9	72645	34	Female	15	Technolo	11025	Fair	Medium	High	1	No	9	Maste	Single	4	Entry	Large	16	No	No	No	Good	Low	Left
10	4941	48	Female	40	Technolo	11452	Good	Medium	Below Av	0	No	65	Associ	Divorced	1	Mid	Large	52	No	No	No	Good	Medium	Stayed
11	65181	55	Female	16	Media	5939	Poor	High	Average	0	No	31	Associ	Divorced	1	Entry	Small	46	No	No	No	Good	High	Stayed
12	49522	32	Female	12	Healthcar	8144	Good	High	Below Av	0	Yes	28	High S	Married	1	Mid	Medium	57	No	No	No	Fair	Medium	Stayed
13	8195	26	Female	15	Finance	4758	Good	High	High	2	Yes	35	Bachel	Married	1	Senior	Medium	91	Yes	No	No	Good	Medium	Stayed
14	52948	45	Female	3	Media	6370	Fair	Medium	High	3	No	30	Bachel	Married	3	Senior	Medium	60	No	No	No	Good	High	Stayed
15	35067	42	Male	3	Healthcar	8344	Good	Very High	Average	0	No	88	Maste	Married	0	Mid	Small	14	No	No	No	Good	High	Stayed
16	53650	35	Female	5	Finance	10449	Good	Very High	Average	1	No	51	Bachel	Single	0	Entry	Small	69	No	No	No	Good	Low	Left
17	49803	52	Female	44	Technolo	12514	Fair	Very High	Average	1	Yes	38	High S	Married	4	Mid	Small	117	No	No	No	Good	Low	Stayed
18	2213	42	Male	14	Education	5076	Fair	High	High	0	Yes	53	Bachel	Single	5	Entry	Medium	68	No	No	No	Good	High	Stayed
19	48091	37	Male	28	Technolo	8287	Fair	Medium	High	0	No	10	Bachel	Married	5	Entry	Large	100	No	No	No	Good	Low	Left
20	49462	56	Male	24	Healthcar	8963	Excellent	High	Average	1	No	43	Bachel	Married	0	Mid	Medium	81	No	No	No	Good	High	Stayed
21	16073	21	Female	1	Healthcar	7869	Fair	High	High	0	No	23	Maste	Married	0	Senior	Large	64	No	No	Yes	Good	Low	Stayed
22	54967	47	Female	11	Media	6664	Fair	High	Average	1	No	77	Associ	Single	4	Mid	Small	41	No	No	No	Good	High	Left

<그림 1. 전처리 전 데이터>

[5] 노미진. "머신러닝과 딥러닝 알고리즘을 활용한 간호사 이직 예측." *재활복지공학회논문지* 17.2 (2023): 78-85.

Column Name	Data Type	Additional Info	Example value
Employee ID	Numerical		8410, 64756, 30257
Age	Numerical		31, 59, 24, 36, 56
Gender	Categorical	2 categories	Male, Female
Years at Company	Numerical		19, 4, 10, 7, 41
Job Role	Categorical	5 categories	Education, Media
Monthly Income	Numerical		5390, 5534, 8159,
Work-Life Balance	Categorical	4 categories	Excellent, Poor
Job Satisfactio	Categorical	4 categories	Medium, High
Performance Rating	Categorical	4 categories	Average, Low, High
Number of Promotions	Categorical	5 categories	2, 3, 0, 1, 4
Overtime	Categorical	2 categories	No, Yes
Distance from Home	Numerical		22, 21, 11, 27, 71
Education Level	Categorical	5 categories	Associate Degree
Marital Status	Categorical	3 categories	Married, Divorced, Single
Number of Dependents	Categorical	7 categories	0, 3, 2, 4, 1
Job Level	Categorical	3 categories	Mid, Senior, Entry
Company Size	Categorical	3 categories	Medium, Small, Large
Company Tenure	Numerical		89, 21, 74, 50, 68
Remote Work	Categorical	2 categories	No, Yes
Leadership Opportunities	Categorical	2 categories	No, Yes
Innovation Opportunities	Categorical	2 categories	No, Yes
Company Reputation	Categorical	4 categories	Excellent, Fair, Poor, Good
Employee Recognition	Categorical	4 categories	Medium, Low, High
Attrition	Categorical	2 categories	Stayed, Left

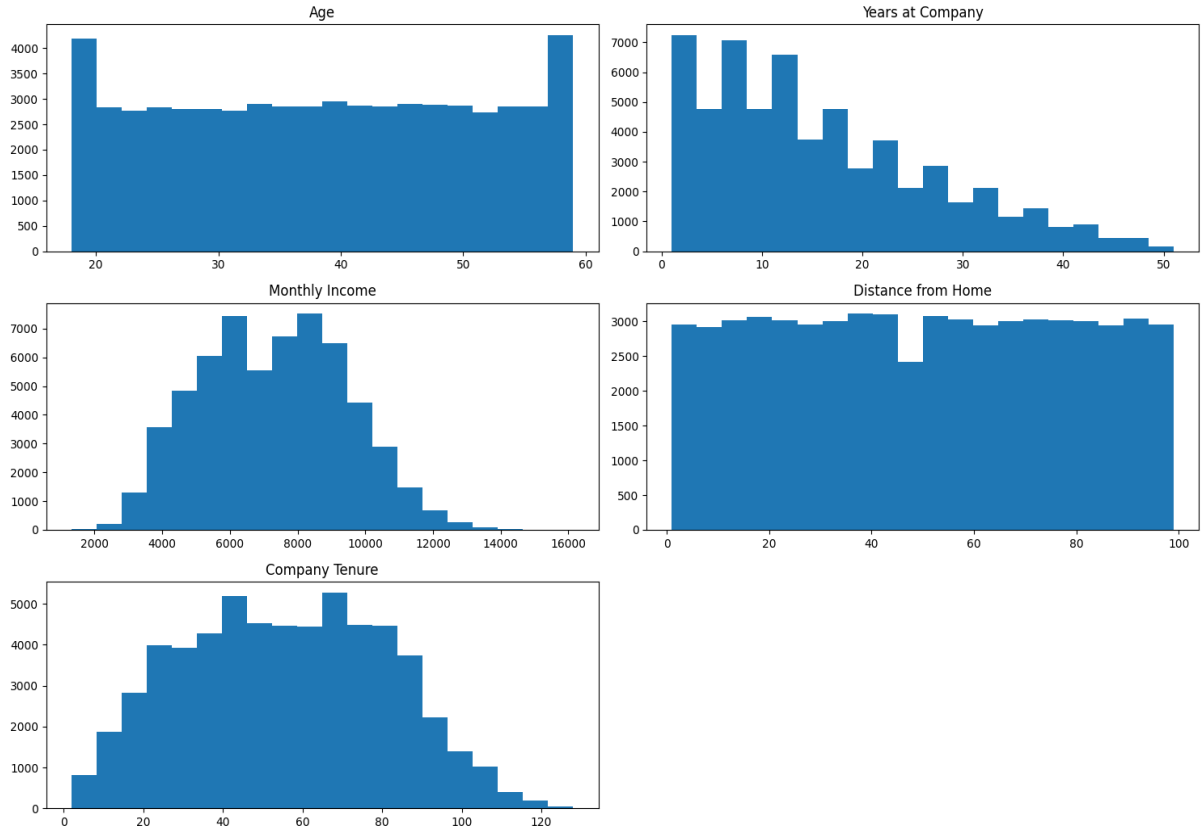
<표 1. 데이터 자료형 및 세부 범주>

3.1 데이터 EDA

주어진 데이터는 총 24개의 컬럼으로 되어있으며, 이 중 예측하고자 하는 변수인 Attrition 변수를 제외하면 23개의 컬럼으로 이탈 여부를 예측하는 모델을 만들고자 한다.

3.1.1 수치형 자료

먼저 수치형 자료에 대해 히스토그램을 그려 각 컬럼의 분포를 살펴보았다.



<그림 2. 연속형 자료 train 데이터 분포>

수치형 자료인 age, years at company, monthly income, distance from home, company tenure에 대한 히스토그램을 살펴보았을 때, 균등한 분포, 왼쪽으로 치우친 분포, 정규분포 등의 다양한 분포를 보이는 것으로 나타났다.

3.1.2 범주형 자료

24개의 컬럼 중 5개의 수치형 자료와 예측하고자 하는 종속변수 1개를 제외한 나머지 18개의 변수는 범주형 자료로 각 컬럼마다 다양한 범주를 가진다. 이는 앞으로 이탈 예측 모델을 만들 때, 어떤 ML 기법을 활용할지에 따라 주의 깊게 살펴보아야 할 사항으로 생각된다.

4. 모델 선정

본 보고서의 목적과 데이터의 형태를 보았을 때, Attrition 변수를 제외한 나머지 변수로 이탈 여부를 예측하는 이진 분류 모델을 만드는 것을 목표로 하였다. 또한, 퇴사 위험이 있는 직원을 예측하는 것에 그치는 것이 아닌, 주요한 변수를 찾아 주요한 변수를 미리 관리하는 것에 목적을 두기 때문에, 예측력과 해석력이 적절히 조합된 모델을 찾고자 하였고, 다음 두가지 모델을 선정하였다.

4.1 SVM (support vector machine)

서포트 벡터 머신(SVM)은 입력 변수 공간에서 클래스 간의 결정 경계를 정의하는 초평면(hyperplane)을 학습하는 분류 모델이다. 특히 선형 SVM에서는 두 클래스 간의 마진(margin)을 최대화하는 결정 경계를 설정함으로써, 분류 성능을 높이면서도 과적합을 방지하는 견고한 분류 기준을 구축한다. 이때 학습된 가중치 벡터 w는 각 입력 변수의 결정경계 형성에 대한 기여도를 의미하며, 이를 통해 변수 중요도를 해석할 수 있는 장점이 있다.

SVM은 특히 결정경계가 데이터를 나누는 방식 자체를 학습하는 구조이기 때문에, 예측 결과가 단순한 확률 추정이 아닌 "두 클래스 간의 명확한 분리 경계"를 중심으로 이루어진다. 특히 선형 SVM에서 각 변수에 대한 가중치가 분류 기준에 미치는 영향을 수치적으로 파악할 수 있다는 점에서, 일정 수준의 해석 가능성을 가진다.

예를 들어, Overtime_Yes 변수에 대해 학습된 가중치가 양의 값을 가진다면, 초과근무를 하는 사원이 이탈할 가능성이 높다는 해석이 가능하다. 반면, Job Satisfaction_High 변수의 가중치가 음수라면 해당 특성이 이탈 가능성을 낮추는 방향으로 작용한다고 판단할 수 있다.

이처럼, 선형 SVM은 단순한 구조 속에서도 예측 성능과 해석 가능성 간의 균형을 제공하며, 특히 변수 해석이 중요한 데이터 분석에서 활용될 수 있다. 다만, 범주형 변수는 전처리 과정에서 원-핫 인코딩 되어 여러 개의 더미 변수로 나뉘며, 이때 해석은 기준 범주 대비 상대적 영향력을 중심으로 이루어진다는 점에서 직관성이 다소 떨어질 수 있다.

4.2 의사결정나무(Decision tree)

의사결정나무는 설명변수로부터 반응변수를 어떻게 예측 또는 설명할 수 있는지를 트리 구조로 표현하는 대표적인 지도학습 기법이다. 본 프로젝트에서는 사원들의 이탈 여부(Attrition)를 예측하는 이진 분류 문제를 다루고 있으며, 문제에 모델이 적합한지에 대하여 아래와 같이 생각하였다.

의사결정나무는 데이터를 재귀적으로 분할하여, 각 노드에서 특정 변수의 조건(임계값)을 기준으로 데이터를 이진 분기한다. 이 과정에서 사용되는 분할 기준은 주로 지니 불순도(Gini Impurity) 또는 엔트로피(Entropy)로, 각 분할 후 생성된 자식 노드들이 부모 노드보다 더 순수하게(한 클래스에 더 집중되도록) 구성되도록 최적화된다.

본 데이터셋에는 사원들의 근속 연수(Years at Company), 월 소득(Monthly Income), 초과근무 여부(Overtime), 직무 만족도(Job Satisfaction) 등과 같은 다양한 수치형 및 범주형 변수들이 포함되어 있으며, 이는 Decision Tree 모델이 다양한 타입의 변수를 효과적으로 처리할 수 있다는 특성과 잘 부합한다.

예를 들어, 모델의 첫 번째 분할 기준이 월 소득(Monthly Income)이라면, 해당 값이 특정 임계값보다 낮은 경우 한쪽 가지로, 높은 경우 다른 가지로 나뉜다. 이후 근속 연수, 초과근무 여부, 직무 만족도 등의 변수들이 순차적으로 조건으로 활용되며, 최종적으로 도달한 리프 노드(leaf node)는 해당 조건에 부합하는 사원들이 과거에 이탈했는지 여부를 바탕으로 이탈 확률을 예측한다.

이처럼 의사결정나무는 데이터의 분포를 기반으로 의미 있는 조건 분기를 생성하며, 예측 과정 전체가 시각화 가능하고 직관적으로 해석 가능하다는 장점이 있다. 특히, 모델 내부의 분기 규칙을 통해 이탈에 영향을 주는 주요 요인을 도출할 수 있어, 향후 사전적 인사관리 전략 수립에도 유용하게 활용될 수 있다.

따라서 본 프로젝트에서는 이해 가능성, 변수 해석력, 예측의 직관성을 고려하여 의사결정나무를 주요 분류 모델 중 하나로 채택하였다.

위 같은 이유로 두 모델을 선정하여 변수의 해석 및 예측력을 비교하고자 하였다.

5. 모델적합

5.1 SVM (support vector machine)

5.1.1 데이터 전처리

주어진 데이터에는 표 1에서 볼 수 있듯이 다양한 범주형 변수들이 포함되어 있으며, 해당 범주형 변수들을 SVM 모델에 적합하기 위해 전처리 단계에서 원-핫 인코딩(one-hot encoding)을 적용하였다. 이때, 다중공선성을 방지하기 위해 각 범주형 변수의 첫 번째 범주를 기준으로 제외하는 방식을 사용하였다.

SVM은 입력 변수와 가중치 벡터의 내적을 통해 결정 경계를 정의하는 모델이기 때문에, 모든 입력 변수는 수치형 형식으로 표현되어야 한다. 그러나 범주형 변수는 고유한 값을 가지는 비수치형 변수로서, 이를 단순히 정수로 치환할 경우 SVM은 변수 간의 순서와 크기 차이에 의미를 부여하게 되어 잘못된 해석을 초래할 수 있다.

이에 따라 각 범주형 변수는 원-핫 인코딩을 통해 다차원 벡터로 확장되었으며, 이 중 첫번째 범주를 제거하여 선형 종속성을 해소하였다. 이로 인해 SVM은 각 범주가 구성하는 차원에서의 가중치를 학습하게 되며, 이를 통해 각 범주가 분류 결정에 얼마나 기여하는지를 가중치 벡터를 통해 해석할 수 있는 구조가 마련되었다.

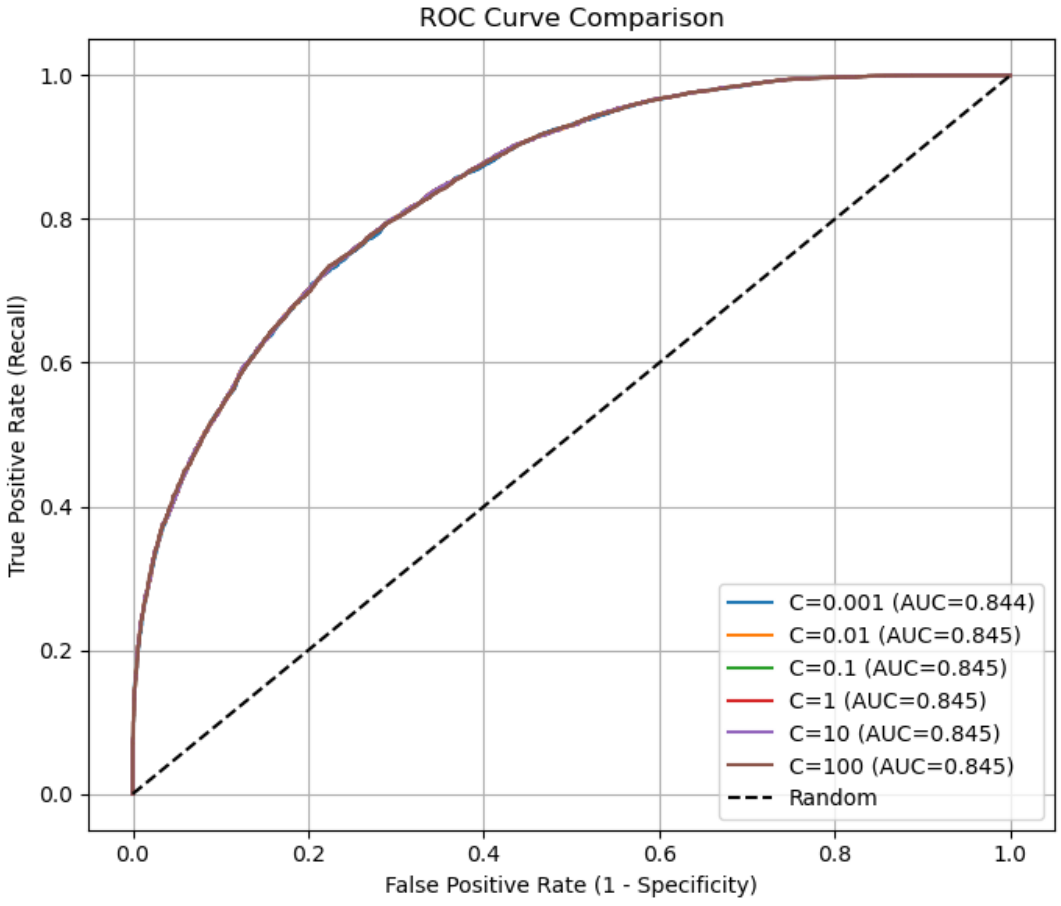
수치형 변수의 경우, 변수 간 단위의 차이로 인해 모델 학습의 왜곡이 생기므로, 평균이 0이고 분산이 1이 되도록 표준화(Standardization)를 수행하였다.

결과적으로 수치화된 범주형 변수와 정규화된 수치형 변수들은 SVM의 해석 가능성을 높이고 분류 성능 향상에 기여한다.

5.1.2 모델 적합

SVM은 오분류의 허용 범위와 마진 사이의 균형을 조절하는 하이퍼파라미터 C 값을 포함한다. C값이 작을수록 마진을 넓게 설정하고 오분류를 어느정도 허용하여 해석력이 우수한 단순한 모델이다. C값이 클수록 마진 폭 감소를 감수하되 오분류를 최소화하려는 방향으로 예측력이 우수한 복잡한 모델이다.

본 분석에서는 C값을 [0.001, 0.01, 0.1, 1, 10, 100]으로 설정한 후, 5-겹 교차검증(five-fold cross-validation)을 수행하고, 각 후보 C값에 대해 F1-score를 기준으로 평가하였다. 그 결과, F1-score가 가장 높은 C값인 0.1이 최적의 하이퍼파라미터로 선정되었으며, 이후 모델 학습에 해당 값을 적용하였다. ROC-curve를 그려본 결과 해당 데이터 셋에서 C값에 따른 AUC(곡선 아래 면적)가 유사하고 1에 가까우므로, C에 따른 큰 성능 차이가 없고 안정적인 분류 성능을 보인다고 할 수 있다.

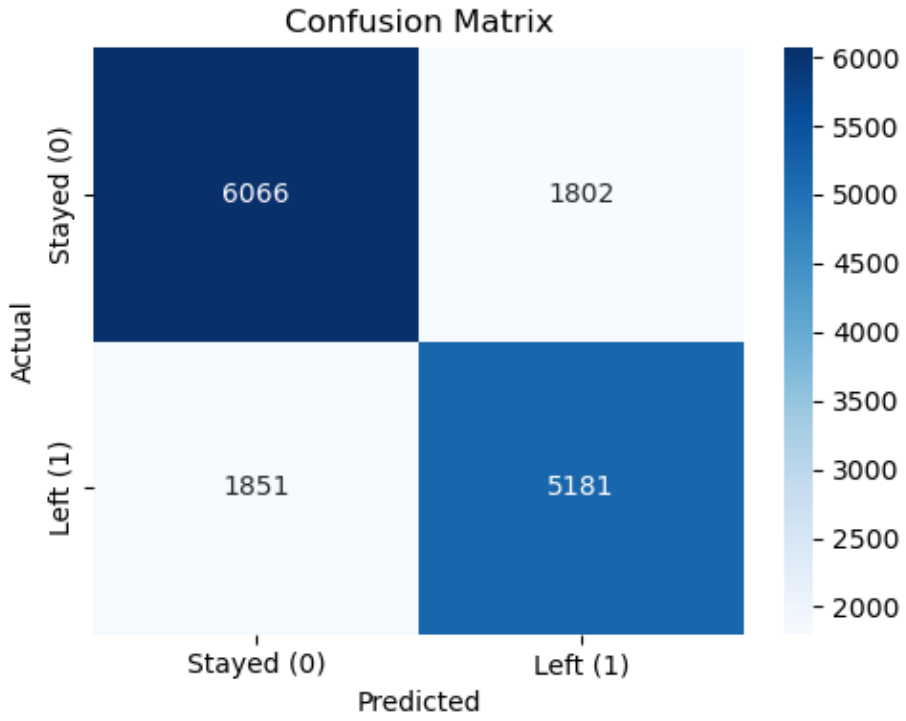


<그림 3. C값에 따른 ROC Curve>

C가 0.1인 모델을 학습하고 예측했다. 평가 지표와 혼동 행렬(Confusion Matrix)을 출력하면 다음과 같다.

	precision	recall	F1-score	Support
0(Stayed)	0.77	0.77	0.77	7868
1(Left))	0.74	0.74	0.74	7032
Accuracy			0.75	14900
Marco avg	0.75	0.75	0.75	14900
Weighted avg	0.75	0.75	0.75	14900

<표 2. SVM 모델 성능지표>



<그림 4. SVM 모델 혼동행렬>

모델 전체의 정확도(accuracy)는 75%으로 baseline 이상의 성능을 보였다. 클래스 0(Stayed)와 클래스1(Left)에 대해 각각 F1-score(precision과 recall의 균형)은 0.77, 0.74로 비교적 균형 잡힌 분류 성능을 나타냈다. 두 클래스 간 성능 차이가 거의 없는 균형 잡힌 모델이라고 할 수 있다. 본 분석에서 퇴사자를 예측하고 적절한 조치를 통해 인력 유출을 막는 것이 핵심이므로 클래스 1 분류 성능에 집중하여 살펴보겠다. 퇴사자라고 예측한 것 중 74%가 실제 퇴사자였고(precision), 실제 퇴사자 중 74%를 맞췄다(recall). 이와 같은 결과를 통해, SVM모델은 본 데이터에서 퇴사자를 비교적 안정적으로 예측할 수 있는 수준의 성능을 보였다고 평가할 수 있다.

선형 SVM에서는 각 변수의 계수(coefficient)를 통해 분류 결정에 대한 기여도를 해석할 수 있다. 다만, 수치형 변수와 범주형 변수는 해석 단위가 다르기 때문에 계수 크기를 서로 비교하는 것은 적절하지 않다. 수치형 변수는 계수의 크기와 부호를 그대로 해석할 수 있지만, 범주형 변수는 기준 범주와의 상대적인 비교를 통해서만 해석할 수 있으며, 서로 다른 범주형 변수 간에도 직접적인 비교는 불가능하다. 이것이 SVM의 해석의 한계다.

수치형 변수와 계수는 다음과 같고 영향력이 큰 순서대로 나타냈다.

변수	계수
Distance from Home	0.105725
Number of Promotions	-0.092320
Number of Dependents	-0.90644
Years at Company	0.056848
Age	0.027534

<표 3. 수치형 변수의 가중치 계수>

집이 멀수록 퇴사 확률이 증가하는 경향을 보였다. 반면, 승진 횟수, 부양 가족 수, 재직 기간, 나이와 같은 변수들은 계수가 음수이다. 따라서 해당 값이 증가할수록 퇴사 가능성이 낮아지는 방향으로 작용하였다.

범주형 변수들 중 퇴사 가능성을 크게 높인 상위 3개만 예시로 해석해보자. 이혼자보다(기준범주) 미혼자의 퇴사 가능성이 크게 높다. 워라밸은 기준 범주인 매우 좋음(Excellent)보다 워라밸이 매우 나쁜 경우(poor) 퇴사 가능성이 급증한다. 직업 만족도에서는 기준 범주인 높음(High)보다 오히려 매우 높음(Very High)이 퇴사 가능성이 높았다.

또한 주목할 점은 직무 중 교육 직무(기준 직무)에 비해 모든 직무 범주들이 퇴사 가능성이 낮은 방향으로 작용하였다. 이는 교육 직무에서 상대적으로 퇴사율이 높게 나타냈을 가능성을 시사하지만 계수 크기는 범주 간의 우열을 나타내지 않으며, 직무들 간 직접 비교는 불가능하다.

교육 수준에 대해서는 기준 범주인 준학사(Associate Degree)에 비해 박사 학위 보유자는 퇴사 가능성이 뚜렷하게 낮은 방향으로 작용하였다.

전체 더미형 변수들의 계수와 해석의 표는 다음과 같다.

변수	범주	계수	기준 대비 해석
Gender	Gender_Male	-0.232956	기준(Female) 대비 남성은 퇴사 가능성이 낮음
Job Role	Healthcare	-0.028815	기준 직무(Educaion) 대비 퇴사 가능성이 낮음

Job Role	Technology	-0.031944	기준 직무(Education) 대비 퇴사 가능성이 낮음
Job Role	Finance	-0.036682	기준 직무(Educaaion) 대비 퇴사 가능성이 낮음
Job Role	Media	-0.037729	기준 직무(Education)대비 퇴사 가능성이 낮음
Work-Life Balance	Poor	0.559803	기준(Excellent) 대비 퇴사 가능성이 매우 높음
Work-Life Balance	Fair	0.494305	기준(Excellent) 대비 퇴사 가능성이 높음
Work-Life Balance	Good	0.107901	기준(Excellent) 대비 약간 높음
Job Satisfaction	Very High	0.184012	기준(High) 대비 퇴사 가능성이 높음
Job Satisfaction	Low	0.179339	기준(High) 대비 퇴사 가능성이 높음
Job Satisfaction	Medium	0.003298	기준(High) 대비 거의 차이 없음
Performance Rating	Low	0.221216	기준(Average)대비 퇴사 가능성이 높음
Performance Rating	Below Average	0.124333	기준(Average)대비 퇴사 가능성이 다소 높음
Performance Rating	High	0.001585	기준(Average)대비 거의 차이 없음
Overtime	Yes	0.131678	기준(No) 대비 퇴사 가능성이 높음
Education Level	Bachelor's Degree	0.017076	기준(Associate Degree) 대비 소폭 높음
Education Level	High School	0.011254	기준(Associate Degree)대비 소폭 높음
Education Level	Master's Degree	0.01047	기준(Associate Degree)대비 소폭 높음
Education Level	PhD	-0.575054	기준(Associate Degree)대비 퇴사 가능성이 매우 낮음
Marital Status	Single	0.582276	기준(Divoced) 대비 퇴사 가능성이 높음
Marital Status	Married	-0.096713	기준(Divorced)대비 퇴사 가능성이 낮음

Job Level	Mid	-0.380925	기준(Entry)대비 퇴사 가능성이 낮음
Job Level	Senior	-0.967643	기준(Entry) 대비 퇴사 가능성이 매우 낮음
Company Size	Small	0.076355	기준(Large) 대비 퇴사 가능성이 약간 높음
Company Size	Medium	0.002374	기준(Large) 대비 거의 차이 없음
Remote Work	Yes	-0.656681	기준(No) 대비 퇴사 가능성이 낮음
Leadership Opportunities	Yes	-0.060789	기준(No) 대비 퇴사 가능성이 낮음
Innovation Opportunities	Yes	-0.051319	기준(No) 대비 퇴사 가능성이 낮음
Company Reputation	Poor	0.280861	기준(Excellent)대비 퇴사 가능성이 높음
Company Reputation	Fair	0.173185	기준(Excellent) 대비 퇴사 가능성이 다소 높음
Company Reputation	Good	-0.0236	기준(Excellent) 대비 퇴사 가능성이 낮음
Employee Recognition	Low	0.014678	기준(High) 대비 약간 높음
Employee Recognition	Medium	0.016629	기준(High) 대비 약간 높음
Employee Recognition	Very High	-0.029661	기준(High) 대비 퇴사 가능성이 낮음

<표 4. SVM 모델 변수 해석>

5.1.3 C값에 따르는 차이

C값은 SVM 모델이 오분류를 얼마나 허용할 것인지에 대한 규제 파라미터로, 일반적으로 C값에 따라 모델의 결정 경계가 달라지고, 이에 따라 성능 지표에도 유의미한 차이가 발생할 수 있다. 그러나 본 분석에서는 서로 다른 5개의 C값을 사용하였음에도 불구하고, 모든 경우에서 성능 지표가 거의 동일하게 나타났다.

우리 조는 이러한 결과에 대해 다음과 같은 두 가지 가설을 세웠다

1. 결정 경계 근처에 위치한 데이터가 거의 없고, 전체적으로 선형적으로 거의 완벽하게 분리 가능한 데이터셋일 수 있다.

2. 많은 범주형 변수를 원-핫 인코딩하면서 고차원 희소 벡터가 생성되었고, 이로 인해 대부분의 변수는 의미 있는 가중치를 갖지 않게 된다. 따라서 실제 결정 경계는 소수의 중요한 feature에 의해 주로 결정되며, C값 변화가 이들 feature의 영향력을 바꾸지 않기 때문에 성능 차이가 나타나지 않는다.

5.2 의사결정나무(Decision tree)

5.2.1 데이터 전처리

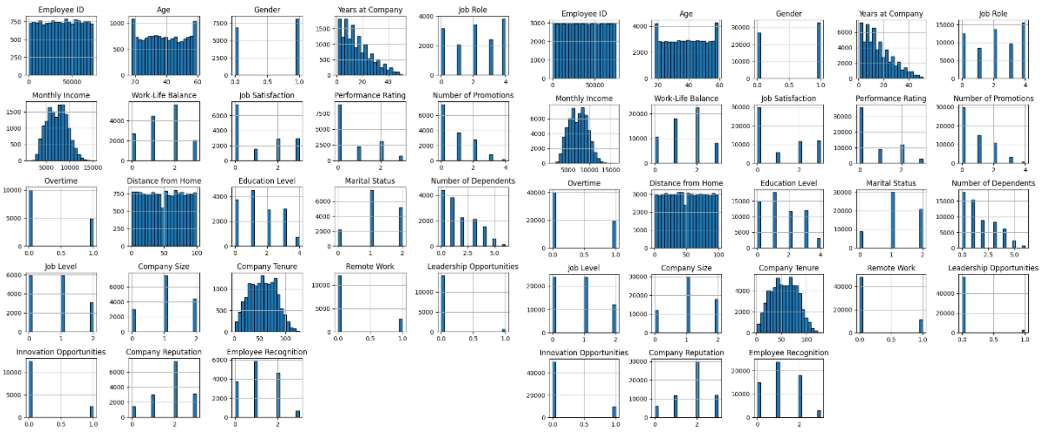
의사결정나무(Decision Tree) 모델에서는 범주형 변수에 대해 라벨 인코딩(label encoding)을 적용하였다. 이는 각 범주형 값을 정수로 매핑하는 방식으로, 예를 들어 "지역"이라는 변수의 경우 "서울", "부산", "대구"를 각각 0, 1, 2와 같이 숫자로 변환하는 방법이다.

의사결정나무는 모델 학습 과정에서 설명변수의 분할 기준을 불순도(impurity), 특히 지니지수(Gini index)나 엔트로피(Entropy)를 얼마나 감소시키는지를 기준으로 최적의 분할을 결정한다. 이때 의사결정나무는 단순히 값을 기준으로 하위 집단으로 나누는 구조를 가지므로, 라벨 인코딩을 통해 부여된 정수 값은 상대적 순서가 아닌 단순한 구분자로 기능하게 된다.

또한, 의사결정나무는 변수의 분할 기준을 실제 데이터 분포에 따라 자동으로 결정하기 때문에, 라벨 인코딩된 범주형 변수라 하더라도 불순도가 가장 크게 감소하는 방식으로 유연하게 분할을 수행할 수 있다. 예를 들어 "지역"이라는 변수에서 '서울(0)'과 '대구(2)'만을 하나의 그룹으로, '부산(1)'을 다른 그룹으로 나누는 식의 분할도 가능하다.

결과적으로, 의사결정나무에서는 라벨 인코딩을 통해 범주형 변수를 정수형으로 처리하는 것이 모델의 분류 규칙 학습에 문제없이 적용 가능하며, 불순도 기준에 따른 의사결정 경로 생성 및 해석력 유지에도 적절한 방법으로 평가된다.

라벨 인코딩을 적용했을 때, 데이터의 분포는 다음과 같다.



<그림 5. 전처리 후 train, test 데이터 분포>

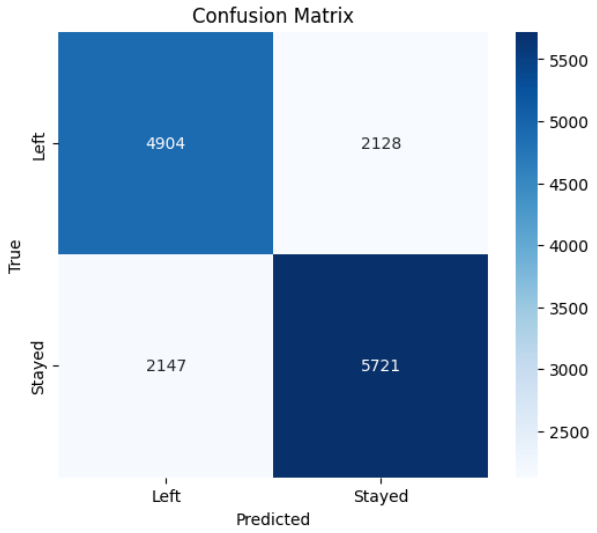
변수명	라벨 인코딩 매핑
Gender	Female → 0, Male → 1
Job Role	Education → 0, Finance → 1, Healthcare → 2, Media → 3, Technology → 4
Work-Life Balance	Excellent → 0, Fair → 1, Good → 2, Poor → 3
Job Satisfaction	High → 0, Low → 1, Medium → 2, Very High → 3
Performance Rating	Average → 0, Below Average → 1, High → 2, Low → 3
Overtime	No → 0, Yes → 1
Education Level	Associate → 0, Bachelor's → 1, High School → 2, Master's → 3, PhD → 4
Marital Status	Divorced → 0, Married → 1, Single → 2
Job Level	Entry → 0, Mid → 1, Senior → 2
Company Size	Large → 0, Medium → 1, Small → 2
Remote Work	No → 0, Yes → 1
Leadership Opportunities	No → 0, Yes → 1
Innovation Opportunities	No → 0, Yes → 1
Company Reputation	Excellent → 0, Fair → 1, Good → 2, Poor → 3
Employee Recognition	High → 0, Low → 1, Medium → 2, Very High → 3

<표 5. 라벨 인코딩 결과>

5.2.2 모델 적합

본 분석에서는 의사결정나무(Decision Tree) 분류 모델을 구성함에 있어 지니 불순도(Gini Impurity) 함수를 활용하여 각 분기에서의 노드 불순도를 효과적으로 줄이는 방향으로 분할이 이루어지도록 설정하였다. 이는 CART(Classification and Regression Tree) 알고리즘에서 널리 사용되는 표준적인 분할 방식이며, 이진 분류 문제에 적합한 기준으로 알려져 있다.

먼저 max_depth를 5로 설정하여 최대 나무를 만들었을 때, 모델의 예측결과와 시각화한 나무는 다음과 같았다. 추가로 퇴사(Left)한 직원을 예측하는 것에 초점을 두어 평가지표를 계산하였다.



Accuracy = 71.86%

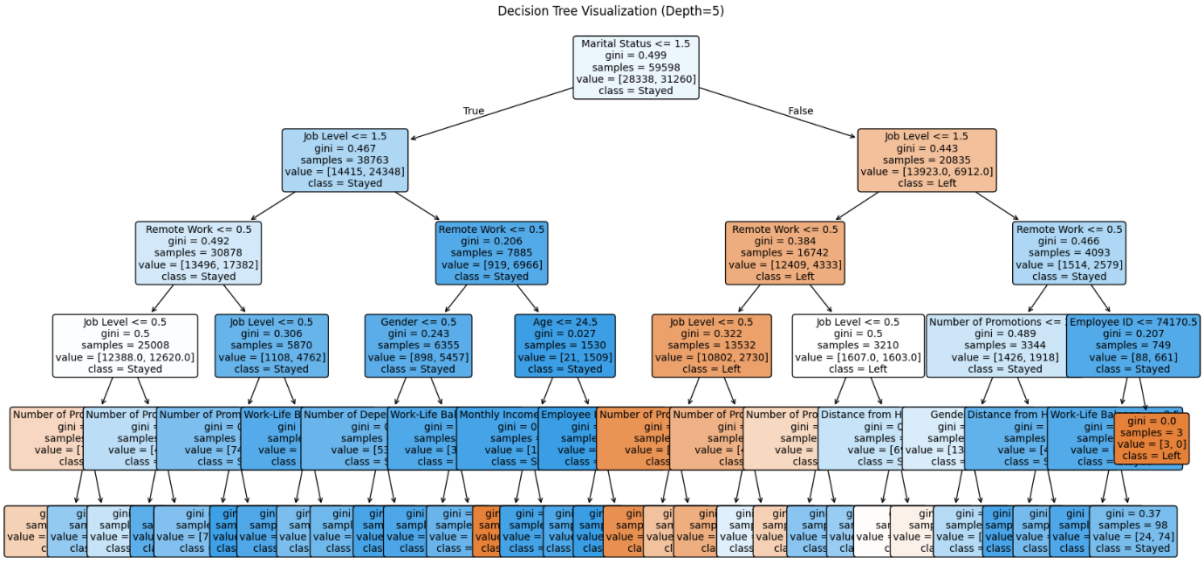
Precision (예측 Left 중 실제 Left)

= 69.55%

Recall (실제 Left 중 예측 Left)

= 69.75%

<그림 6. Max_depth = 5 분석 결과>



<그림 7. Max_depth = 5 나무 시각화>

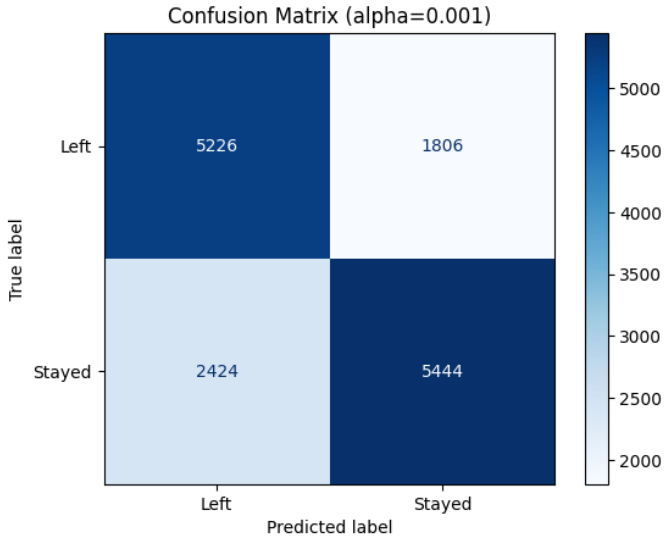
깊이에 제한을 두지 않거나, 가지치기를 하지 않은 나무모형은 나무모형의 장점인 해석의 어려움이 생기게 된다. 앞서 분석한 결과는 어느정도 준수한 성능과 해석을 할 순 있지만, 리프 노드의 개수가 불필요하게 많고, 더 중요한 변수들만 남겨 해석에 용이하게 하기 위하여 최대나무모형에서 a값에 따라 오분류 비용(misclassification cost)과 나무모형의 복잡도(complexity)를 동시 고려한 가지치기를 통해 나무를 간결하게 하면서 해석에 필요한 변수를 선정하기로 하였다.

a 값 = [0.000, 0.001, 0.005, 0.01, 0.02]에 대한 train data와 test data에 대한 accuracy는 다음과 같다.

alpha	Train Accuracy	Test Accuracy	Num Leaves
0.000	1.000000	0.667987	10465
0.001	0.718682	0.716107	18
0.005	0.699352	0.701141	7
0.010	0.660022	0.659463	5
0.020	0.642152	0.643221	3

<표 6. 선정된 alpha 값에 따른 성능비교>

교차검증을 통해 alpha 값에 따른 성능을 비교한 결과, alpha=0.005에서 가장 높은 테스트 정확도(0.701141)가 나타나 이를 최종 모델의 규제 계수로 채택하였다. 그때의 혼동행렬과 나무를 시각화 한 것은 다음과 같다.

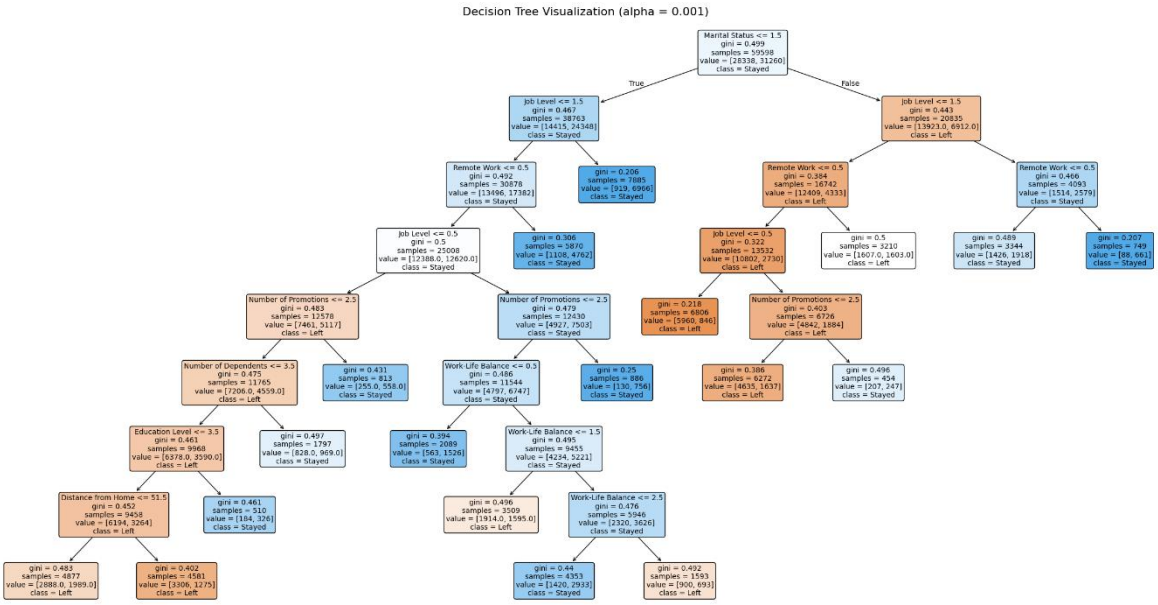


Accuracy = 71.68%

Precision (예측 Left 중 실제 Left) = 69.55%

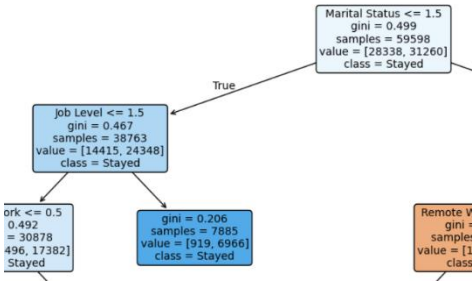
Recall (실제 Left 중 예측 Left) = 69.75%

<그림 8. alpha = 0.001 분석 결과>



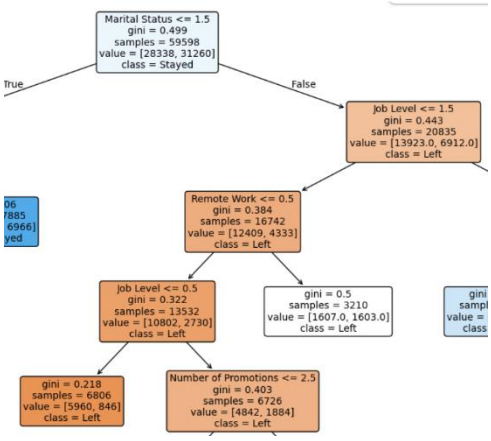
<그림 9. alpha = 0.001 나무 시각화>

학습된 나무모형에서 리프 노드의 지니 불순도가 낮은 노드에 대해 해석한 결과는 다음과 같다.



1. 이혼(0)이거나 기혼(1)인 직원이면서, 미혼(2)이 아닌

2. Job level이 entry(0), mid(1)인 직원이 아닌, senior(2)인 직원은 퇴사하지 않을 것이다. (stay)

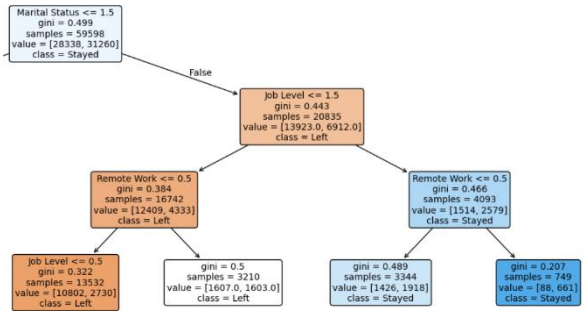


1. 미혼(2)이며,

2. Job level이 entry(0), mid(1)인 직원인,

3. 일에 동기부여가 없는(0)

4. Job level이 entry(0)인, 신입인 직원은 퇴사할 것이다. (Left)



1. 미혼(2)이며,

2. Job level이 senior(2)인

3. 일에 동기부여가 있는(1)

직원은 퇴사하지 않을 것이다. (Stay)

위처럼 의사결정나무를 해석하는 것은 로지스틱 회귀나 SVM처럼 선형 결정 경계를 사용하는 모델에 비해, 더 명확하고 해석 가능한 규칙을 제공함으로써 보다 현실적인 의사결정 지원에 유리함을 보여준다.

6. 모델 비교

지금까지의 보고서에서는 직원의 퇴사 여부를 종속변수로 하여 SVM과 의사결정나무 두 가지 모델을 학습시키고, 각 모델의 특성에 따라 결과를 해석하였다.

각 모델에서 직원의 퇴사여부에 큰 영향을 주는 주요한 변수는 다음과 같다.

SVM

1. Distance from Home
2. Number of Promotions
3. Number of Dependents

의사결정나무

1. Marital Status
2. Job Level
3. Remote Work

다양한 변수가 포함되었던 데이터에 두 모델을 적합하면서, 퇴사여부를 예측하는 예측력에는 큰 차이가 없었지만, 각 모델의 변수에 대한 중요도 해석에서 차이가 있었다. 이를 두고 다음과 같이 생각하였다.

우리에게 주어진 데이터는 직원의 개인정보와 회사에서 발생한 정보 등을 포함하고 있었다. 특히 범주형 자료가 많았고, 이 데이터들의 상관관계를 상관계수 같은 다중공선성을 확인할 수 있는 지표를 적용하기 어려웠다. 하지만 이 범주형 변수 사이의 다중공선성이 분명히 존재했을 것이며, 이러한 다중공선성은 선형 결합을 기반으로 한 모델에서는 계수 추정의 불안정성을 초래할 수 있

으며, 해석에 오류를 유발하거나 모델의 일반화 성능을 저하시킬 수 있다.

따라서 SVM에서 도출되는 변수의 가중치를 퇴사 여부에 대한 직접적인 영향력으로 해석하는 것은 다중공선성으로 인한 가중치 왜곡 가능성을 고려할 때 신중해야 한다고 생각하였고 이 결과와 비선형 결정경계를 가지는 의사결정나무에서의 변수들과 종합적으로 해석하며 앞으로의 인적 자원 관리에서 주요하게 다루어야 할 것이라고 생각하였다.

7. 결론 및 제언

SVM과 같은 선형 결정경계를 사용하는 모델에서는, 각 변수의 영향력을 가중치 벡터를 통해 정량적으로 파악할 수 있었고, 범주형 변수에 대해서도 원-핫 인코딩을 통해 해석이 간접적으로 나마 가능하였다. 예측력 또한 준수한 수준을 보였다.

반면 의사결정나무는 지니 불순도를 기반으로 비선형적인 결정 경계를 학습하며, 각 변수의 조건을 명시적으로 분기함으로써 더 세부적이고 직관적인 해석이 가능하였다. 특히 다양한 범주를 가진 변수들에 대해 개별 범주 수준에서 분류 기준을 설정하는 점이 인상적이었다.

위에서 다루었듯이 두 모델의 결정경계의 차이에서 다중공선성에 반응하는 정도가 다르다는 것을 언급한 바가 있다. 이는 앞으로 어떤 목적을 가지고 데이터를 다룰 때, 모델선정이나 해당 모델에 대한 해석을 할 때, 중요하게 다루어져야 할 부분이라고 생각하였다.

추가로 이혼(0), 기혼(1), 미혼(2)처럼 순서가 없는 범주형 변수에서, "미혼과 이혼"과 같이 특정 범주 조합에만 영향을 미치는 경우는 단순한 부등호 비교로는 표현하기 어렵다고 생각했다. 그러나 범주 간 단순한 선형 순서가 존재하지 않거나, 특정 범주의 조합(예: XOR 구조)이 의미 있는 영향을 주는 경우, 의사결정나무는 같은 변수를 반복적으로 분할 기준으로 선택함으로써 해당 비선형적인 조건도 효과적으로 포착할 수 있다. 이는 선형 모델에서 표현하기 어려운 비선형 관계를 탐지하는 데 유리할 것이다.

본 보고서는 이러한 분석을 바탕으로, SVM과 의사결정나무 각각의 결정 경계 해석 방식과 그 활용 가능성을 비교하였고, 이를 통해 퇴사 여부에 주요한 영향을 미치는 변수를 식별하고 선제적으로 관리할 수 있는 인사관리 방향을 제시하였다.

또한 모델 선택 측면에서도, 목적과 데이터의 특성에 따라 어떤 알고리즘을 사용할 것인지에 대한 실질적인 통찰을 얻을 수 있었으며, 이는 향후의 예측 및 정책 수립 과정에 있어 유의미한 기반이 될 것이다.