의료 인력 이탈 예측 및 해석

SOTA(State of the Amateur)

2023 Fall – Data Mining

김주혁 송준현 오승준 윤서환

TABLE OF CONTENTS

01. INTRODUCTION
문제 배경 및 중요성

02. EDA & PREPROCESSING
데이터 EDA 및 전처리

03. MODELING
모델 구축 및 분석

04. RESULT & SUGGESTION
결과 및 모델 활용방안

01.

INTRODUCTION

문제 배경 및 중요성

문제 배경 및 중요성 – 개인 / 사회적 측면

여전한 인력부족…“의료서비스 질마저 떨어져”

양영구 기자 | 입력 2016.06.24 06:01 | 댓글 0

특히 이 같은 인력부족 현상은 환자 안전에도 심각한 영향을 미치는 것으로 나타났다.

조사결과에 따르면 인력부족으로 인해 환자에게 적절한 의료서비스를 제공하지 못한다고 답한 응답자는 76.6%였고, 환자에게 친절하게 대응하지 못한다는 답변도 82.8%에 달했다.

아울러 병원에서의 인력부족이 의료서비스의 질을 저하시킨다는 응답도 79.8%로 높게 나타났고, 심지어 응답자 중 33.6%는 의료사고 발생을 초래하는 경우가 있었다고 답했다.

인력부족으로 인해 병원 노동자들의 근무시간도 덩달아 증가했다.

병원 노동자들의 하루 평균 시간외 근로시간은 112.3분으로 나타났고, 특히 간호사 직능에서는 116.9분으로 조사됐다. 하루에 두 시간 이상 초과근무를 하고 있는 것이다.

또 병원 노동자들의 하루 평균 휴식 및 식사시간은 39.2분에 불과했고, 응답자의 10명 중 7명(75.8%)은 40분 이하였다.

직종별로 살펴볼 때 간호사의 하루 평균 휴식 및 식사시간은 29.7분에 불과했고, 과도한 업무로 인해 식사를 거르는 경우도 한 달 평균 5회나 됐다.

이에 보건의료노조는 열악한 근무여건 개선을 위해 보다 적극적인 노력이 필요하다고 강조했다.

http://www.monews.co.kr/news/articleView.html?idxno=92203

급작스러운 인력 이탈

↓

의료서비스 품질 저하

문제 배경 및 중요성 – 경영적 측면

부산대병원, 신입간호사 이탈 '최다'

최근 5년간 501명 퇴직...국립대 14곳 중 1위
"교육전담간호사, 프리셉터쉽 기간 연장 등 노력"

유시온 기자 | 기사입력 2023/10/20 [13:19]

반면 부산대병원의 위치와 여건 특성상 의미를 부여하는 주장도 있다. 한 간호계 관계자는 "부산대병원이 여타 국립대 병원보다 여건이 우수한데도 이탈자가 가장 많다는 건 굉장히 심각하게 받아들여야 한다"고 우려했다.

현재 부산대병원 측은 신입 간호사 이탈을 줄이기 위해 사직 사유별로 다양한 대책을 강구하고 있다는 입장이다.

대표적으로 업무 부적응 간호사를 위한 교육전담간호사 프로그램 기간을 기존 3개월에서 6개월로 확대 적용하고, 프리셉터쉽(부서배치 후 현장교육) 기간 연장도 검토 중이다.

http://www.whosaeng.com/147540

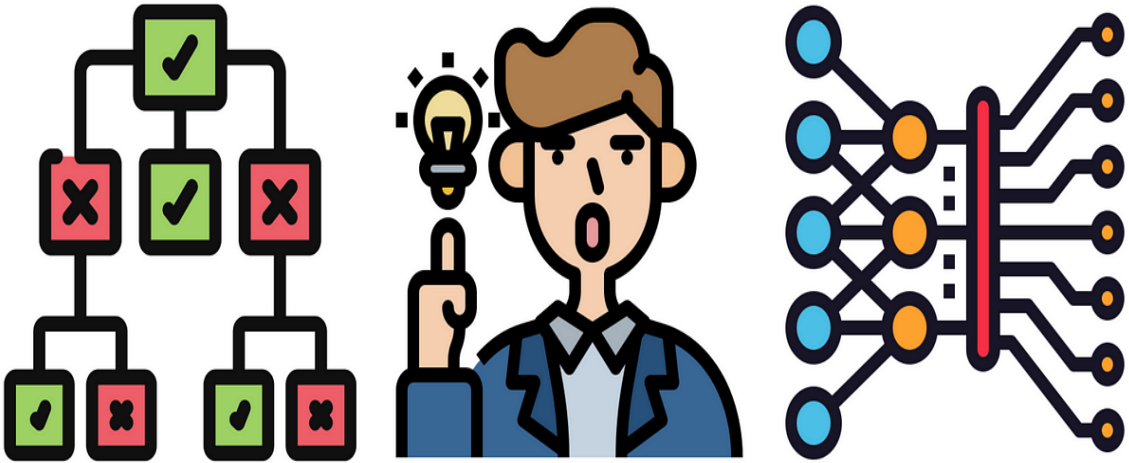
급작스러운 인력 이탈

↓

경영 리스크 발생

문제 배경 및 중요성

해석력이 높은 분류 모델 구축



→ 인력 이탈 원인 파악

→ 모델 활용 방안 제시

02.

EDA & PREPROCESSING

데이터 EDA 및 전처리

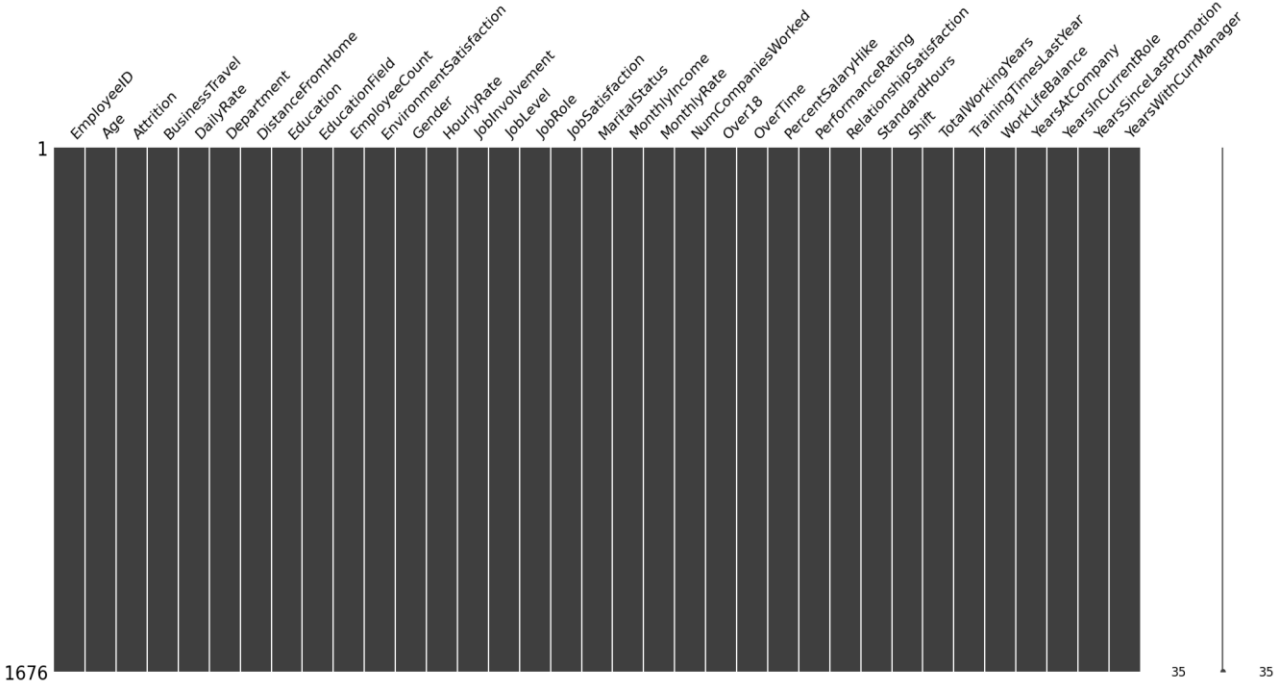
데이터 EDA 및 전처리

0. 데이터셋 설명

- 헬스케어 분야의 직원 이탈 여부 데이터셋
- 1,676개의 샘플, 35개의 특성 존재

데이터 EDA 및 전처리

1. 결측 유무 확인



결측치 존재 x

데이터 EDA 및 전처리

2.1 Categorical – 단일값 변수 제거

"Over18"

1개의 단일값 변수 제거

데이터 EDA 및 전처리

2.2 Categorical – Encoding

```
Attrition : ['No' 'Yes'].  총 2개

BusinessTravel : ['Travel_Rarely' 'Travel_Frequently' 'Non-Travel'].  총 3개

Department : ['Cardiology' 'Maternity' 'Neurology'].  총 3개

EducationField : ['Life Sciences' 'Other' 'Medical' 'Marketing' 'Technical Degree'
 'Human Resources'].  총 6개

Gender : ['Female' 'Male'].  총 2개

JobRole : ['Nurse' 'Other' 'Therapist' 'Administrative' 'Admin'].  총 5개

MaritalStatus : ['Single' 'Married' 'Divorced'].  총 3개

Over18 : ['Y'].  총 1개

OverTime : ['Yes' 'No'].  총 2개
```

Cardinality 고려한
Encoding

데이터 EDA 및 전처리

2.2 Categorical – Encoding

- Case 1) 모든 변수 Label Encoding
- Case 2) 모든 변수 OneHot Encoding
- Case 3) Cardinality 2개 이하 Label, 2~4개 OneHot, 5개 이상 HashEncoding
- Case 4) Cardinality 2개 이하 Label, 3개 이상 OneHot

Data Leakage 고려 → Train data의 통계량으로 Scaling

데이터 EDA 및 전처리

2.3 Categorical – Scaling

- 각각의 Case → Standard, MinMax, Robust scaling 적용
- LogisticRegression → Accuracy, Recall, F1 score
- 최종 Dataset(Case4) + StandardScaler 선정

```
Dataset_3_StandardScaler_accuracy: 0.9360720130932897
Dataset_3_StandardScaler_recall: 0.6547619047619048
Dataset_3_StandardScaler_f1: 0.702495319454254

Dataset_3_MinMaxScaler_accuracy: 0.9292453173304238
Dataset_3_MinMaxScaler_recall: 0.5320105820105819
Dataset_3_MinMaxScaler_f1: 0.6362062339810232

Dataset_3_RobustScaler_accuracy: 0.9335115475541007
Dataset_3_RobustScaler_recall: 0.6113756613756614
Dataset_3_RobustScaler_f1: 0.6803025713430026
```

```
Dataset_0_StandardScaler_accuracy: 0.9326495726495725
Dataset_0_StandardScaler_recall: 0.6047619047619047
Dataset_0_StandardScaler_f1: 0.6727768030289039

Dataset_0_MinMaxScaler_accuracy: 0.9258374249863612
Dataset_0_MinMaxScaler_recall: 0.48994708994708996
Dataset_0_MinMaxScaler_f1: 0.6022124410404894

Dataset_0_RobustScaler_accuracy: 0.9343589743589742
Dataset_0_RobustScaler_recall: 0.5902116402116403
Dataset_0_RobustScaler_f1: 0.6733954011212612
```

```
Dataset_1_StandardScaler_accuracy: 0.9326677577741409
Dataset_1_StandardScaler_recall: 0.6333333333333333
Dataset_1_StandardScaler_f1: 0.6833343370988881

Dataset_1_MinMaxScaler_accuracy: 0.9283869794508093
Dataset_1_MinMaxScaler_recall: 0.5105820105820105
Dataset_1_MinMaxScaler_f1: 0.6252331213317082

Dataset_1_RobustScaler_accuracy: 0.933518821603928
Dataset_1_RobustScaler_recall: 0.6119047619047618
Dataset_1_RobustScaler_f1: 0.6787309795713158
```

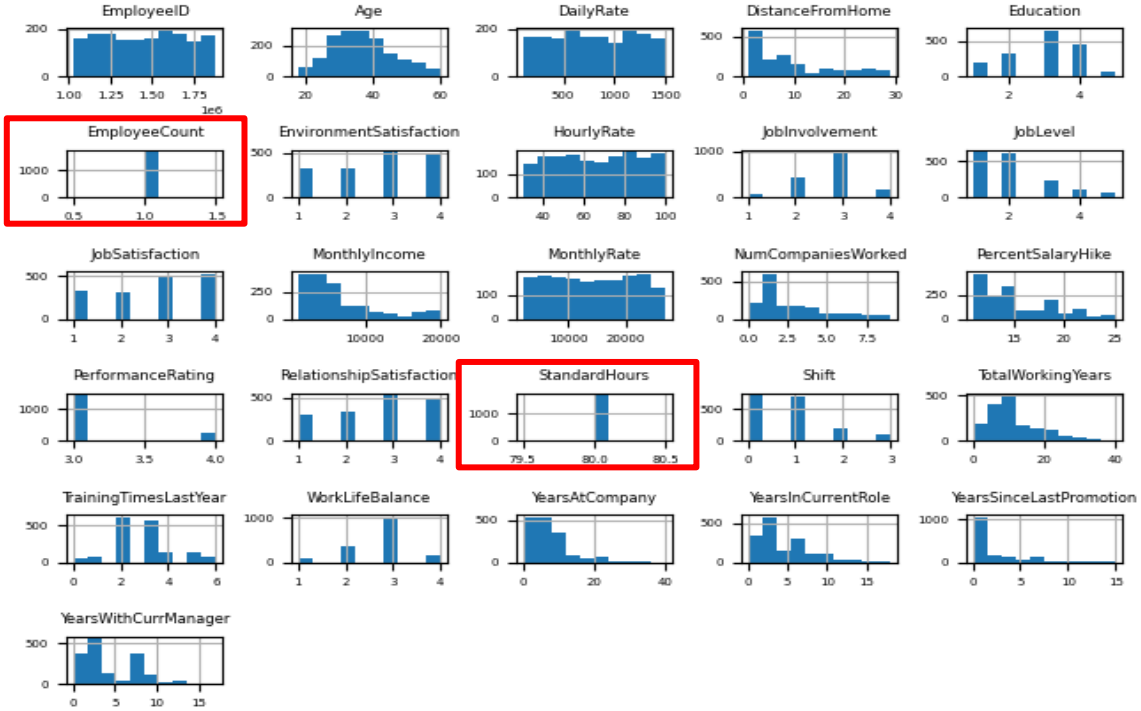
```
Dataset_2_StandardScaler_accuracy: 0.9326604837243135
Dataset_2_StandardScaler_recall: 0.6187830687830689
Dataset_2_StandardScaler_f1: 0.6807531213849776

Dataset_2_MinMaxScaler_accuracy: 0.9283942535006366
Dataset_2_MinMaxScaler_recall: 0.5251322751322751
Dataset_2_MinMaxScaler_f1: 0.6293453958050231

Dataset_2_RobustScaler_accuracy: 0.9335079105291871
Dataset_2_RobustScaler_recall: 0.6042328042328042
Dataset_2_RobustScaler_f1: 0.678900920699619
```

데이터 EDA 및 전처리

3. Numerical - 단일 변수 제거

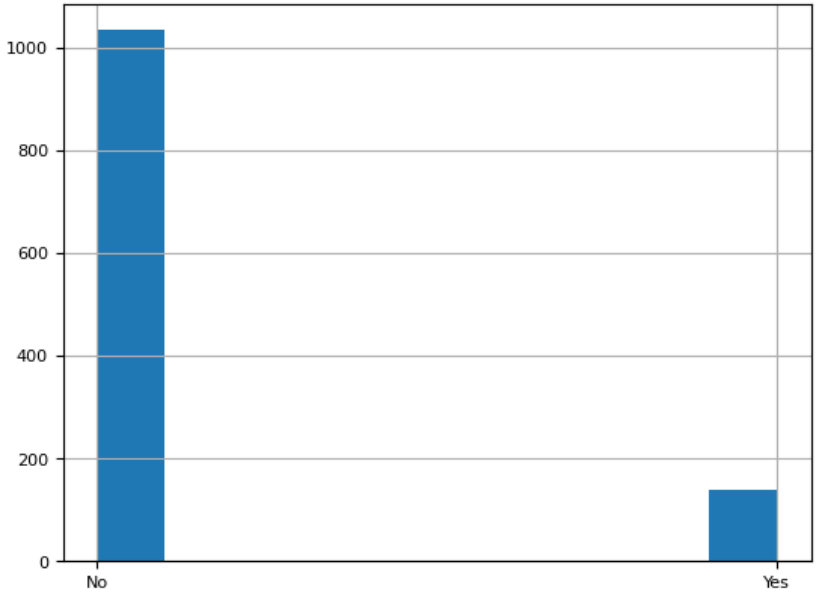


"EmployeeCount"
"StandardHours"

2개의 단일값 변수 제거

데이터 EDA 및 전처리

4. 데이터 불균형



Class 불균형

↓

Oversampling

데이터 EDA 및 전처리

4. 데이터 불균형

- Random Oversampling, SMOTE, ADASYN
- LogisticRegression → Accuracy, Recall, F1 score
- 유의미한 차이 없음 → 이후 Oversampling 적용하지 않고 진행

	Original	Random Oversampling	SMOTE	ADASYN
Accuracy	0.94	0.91	0.92	0.91
Recall	0.65	0.94	0.93	0.93
F1	0.70	0.91	0.92	0.92

데이터 EDA 및 전처리

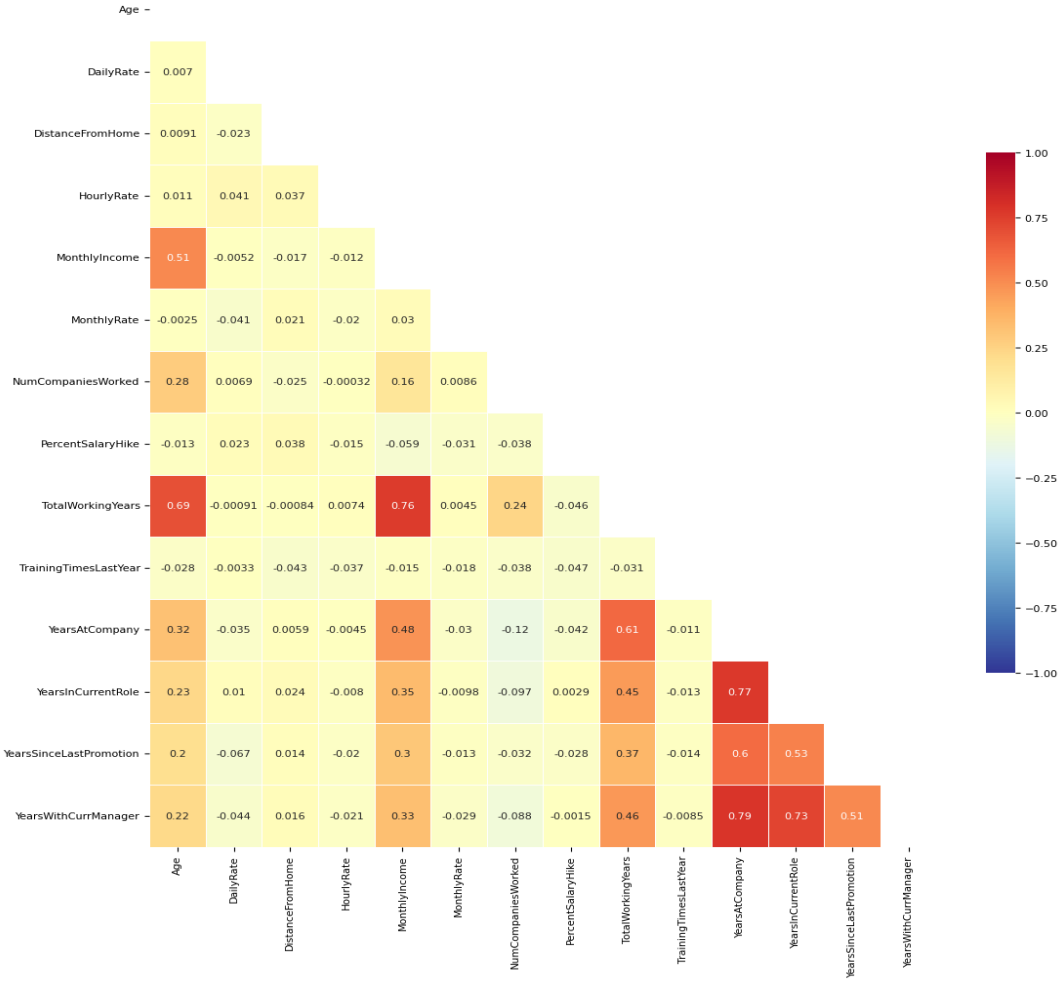
5. 최종 dataset

- 단일값 변수 제거
- Cardinality 2개 이하 Label, 3개 이상 OneHot Encoding → StandardScaler
- Oversampling 적용 x
- Train : 1173, Test : 503

데이터 EDA 및 전처리

6. 변수 간 상관관계 확인

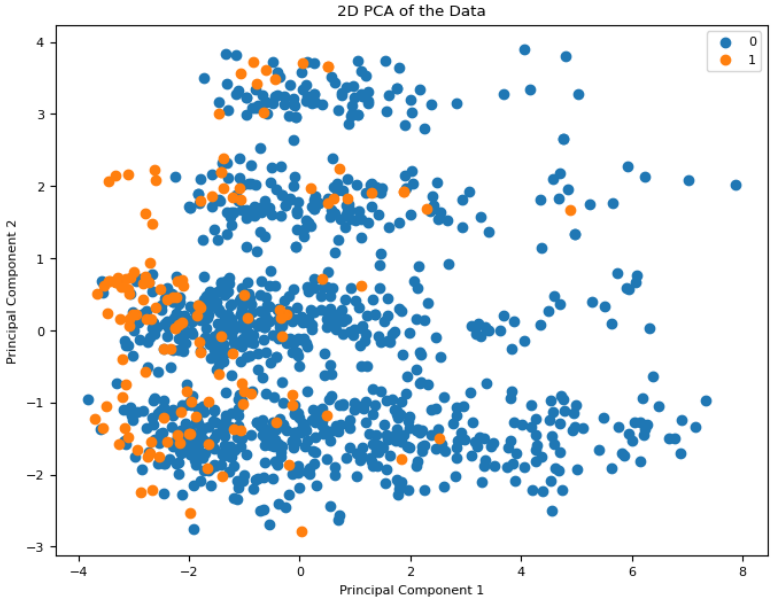
- TotalWorkingYears, MonthlyIncome 등 높은 상관관계를 가지는 조합 존재
- 다중공선성 문제 고려

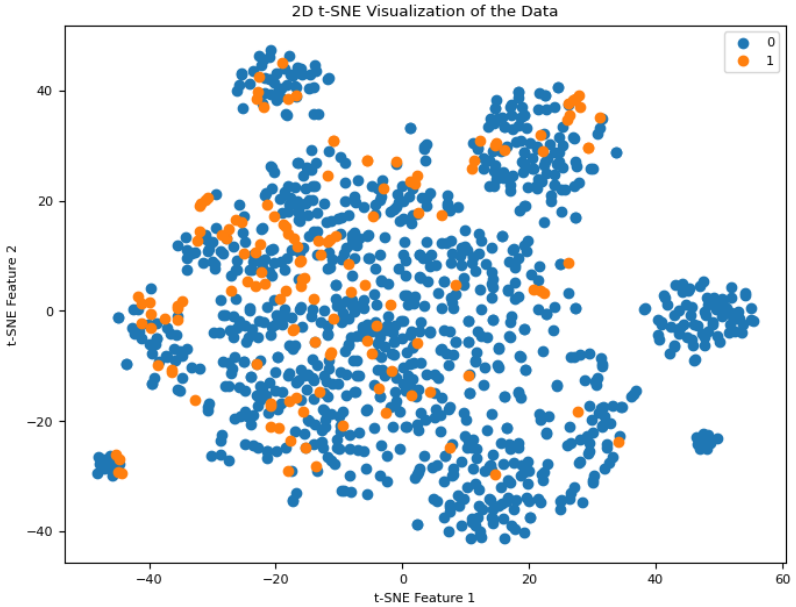


데이터 EDA 및 전처리

7. 차원 축소 기반 시각화 (PCA, t-SNE)

- 데이터 2차원 투영 결과 → 이탈한 경우 레이블 1





03.

MODELING

모델 구축 및 분석

모델 구축 및 분석

1. 모델 간 교차 검증

- 6개 모델의 기본 성능 확인
- LogisticRegression – Vanilla, Ridge, Lasso, ElasticNet
- SVM, NaiveBayes

	Logistic Regression	Ridge	Lasso	ElasticNet	SVM	Naïve Bayes
Accuracy	0.94	0.94	0.93	0.94	0.91	0.36
Recall	0.66	0.65	0.64	0.65	0.32	0.98
F1	0.71	0.70	0.69	0.70	0.45	0.27

모델 구축 및 분석

2. 변수 선택

- 다중공선성 문제를 고려하여 변수 선택 수행
- p-value, VIF, CI를 모두 고려하여 제거할 변수 선택 → 교차검증 기반 후진소거법 수행
- 기본 로지스틱 회귀 모델에서의 성능 향상 확인

	특성 선택 전	특성 선택 후
Accuracy	0.92	0.92
Recall	0.57	0.58
F1	0.64	0.65

12개의 변수 제거

"JobRole_Other", "MonthlyIncome", "EducationField_Medical"
"Department_Neurology", "MaritalStatus_Single", "PerformanceRating"
"HourlyRate", "MonthlyRate", "PercentSalaryHike"
"TrainingTimesLastYear", "Education", "Department_Maternity"

모델 구축 및 분석

3. 하이퍼파라미터 튜닝 전

- 선택된 변수를 기준으로 모델 성능 확인
- 기본 하이퍼파라미터 사용

	Logistic Regression	Ridge	Lasso	ElasticNet	SVM	Naïve Bayes
Accuracy	0.92	0.93	0.93	0.93	0.92	0.35
Recall	0.58	0.57	0.57	0.57	0.38	0.95
F1	0.65	0.65	0.65	0.65	0.52	0.26

모델 구축 및 분석

4. 하이퍼파라미터 튜닝 후

- 선택된 변수를 기준으로 하이퍼파라미터 튜닝 수행
- Sckit-learn optimize의 Bayesian optimization 사용
- 튜닝된 6개 모델의 성능 확인

  → 위음성, 위양성 수를 고려하여 최종 모델로 LogisticRegression - ElasticNet 사용

	Logistic Regression	Ridge	Lasso	ElasticNet	SVM	Naïve Bayes
Accuracy	0.92	0.92	0.92	0.92	0.93	0.37
Recall	0.58	0.57	0.57	0.57	0.55	0.92
F1	0.65	0.64	0.64	0.64	0.64	0.26

04.

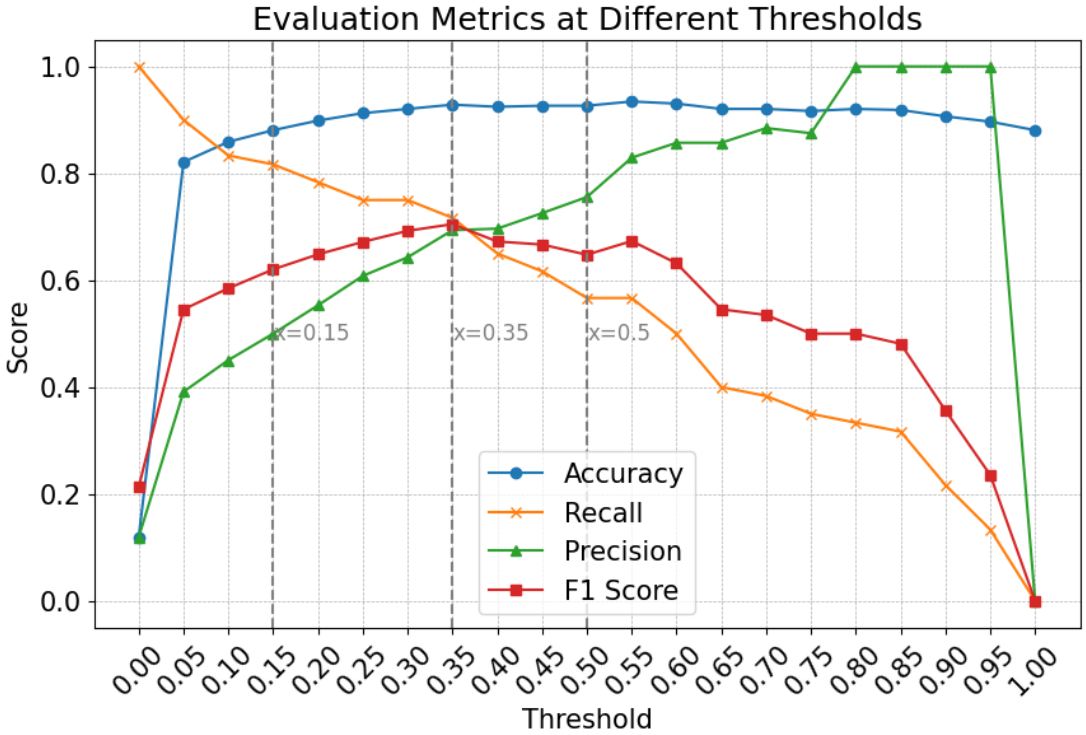
RESULT & SUGGESTION

결과 및 모델 활용방안

결과 및 모델 활용방안

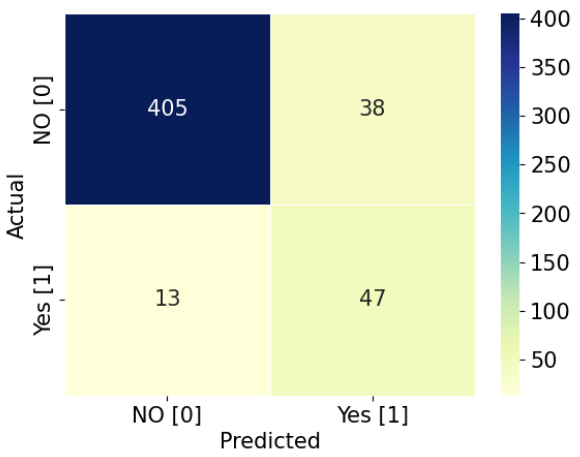
1. 최종 모델의 성능 확인

- 임계값 변화에 따른 지표 변화
- 각각의 구체적인 상황 고려



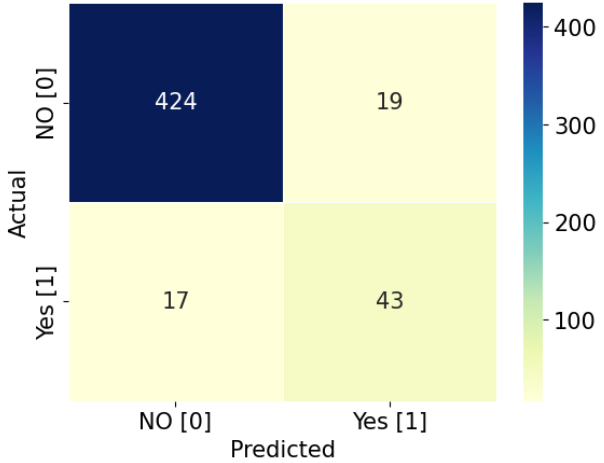
결과 및 모델 활용방안

1. 최종 모델의 성능 확인



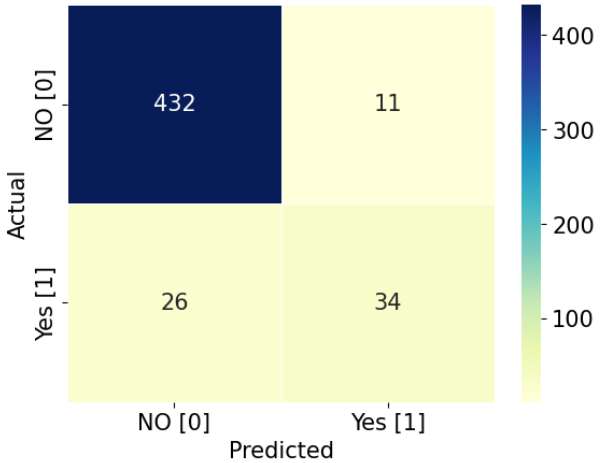
Threshold = 0.15

Test accuracy: 0.90
Test recall: 0.78
Test precision: 0.55
Test f1-score: 0.65



Threshold = 0.35

Test accuracy: 0.93
Test recall: 0.72
Test precision: 0.69
Test f1-score: 0.71



Threshold = 0.50

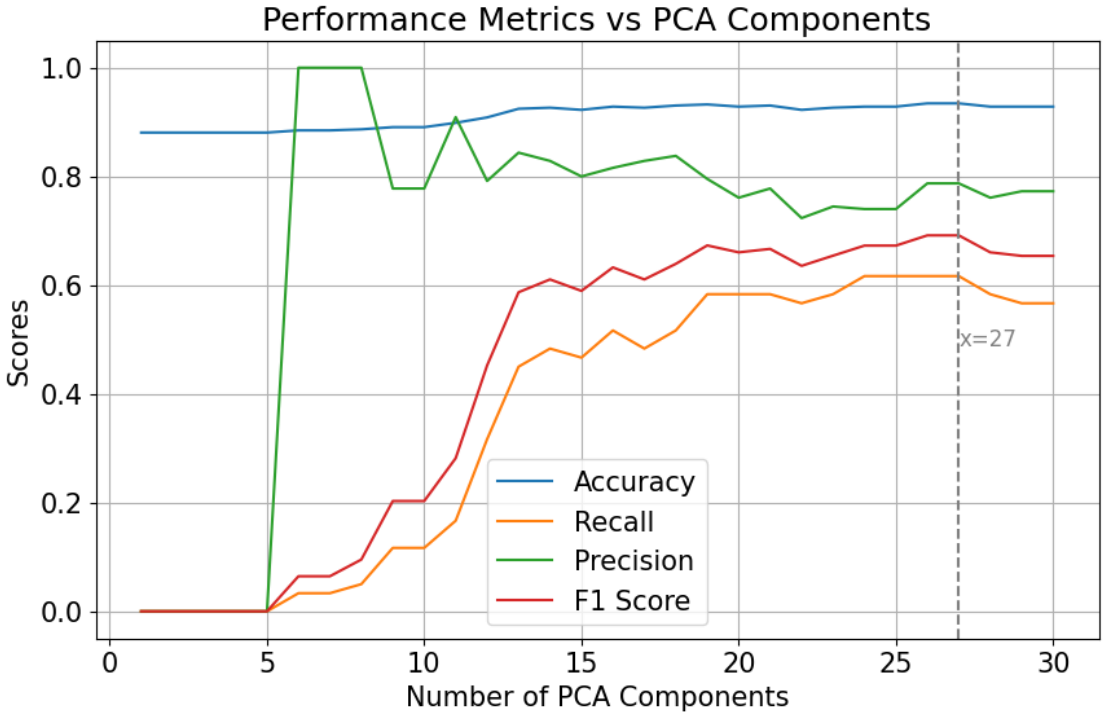
Test accuracy: 0.93
Test recall: 0.57
Test precision: 0.71
Test f1-score: 0.65

결과 및 모델 활용방안

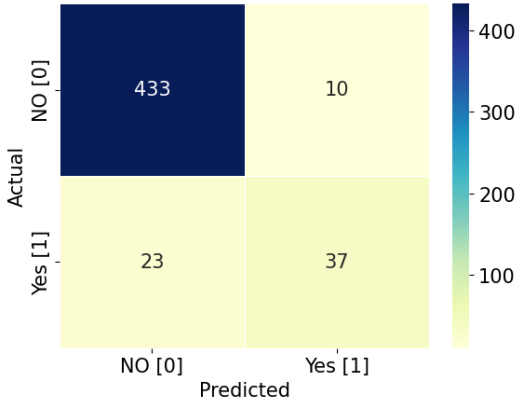
1. 최종 모델의 성능 확인

- 선택된 특성에 PCA 적용

  → 주성분 개수를 다르게 하며 성능 확인



27개의 주성분 사용 시 성능



Test accuracy: 0.93
Test recall: 0.61
Test precision: 0.79
Test f1-score: 0.69

결과 및 모델 활용방안

2. 변수 가중치 해석

- 이탈 확률에 대한 log-odds 증가량
- 상식적 이탈 원인 → 초과근무, 잦은 출장
- 상식에 반하는 이탈 원인 → 급여

-1.260289	YearsInCurrentRole
-1.115358	TotalWorkingYears
-1.003250	Age
-0.753258	JobInvolvement
-0.747590	EnvironmentSatisfaction
-0.706962	MaritalStatus_Divorced
-0.668331	MaritalStatus_Married
-0.564554	WorkLifeBalance
-0.532284	JobRole_Therapist
-0.457418	JobSatisfaction
-0.452682	JobRole_Administrative
-0.419004	JobLevel
-0.310051	Shift
-0.258726	RelationshipSatisfaction
-0.208758	JobRole_Admin
-0.081571	DailyRate
-0.074787	YearsAtCompany
-0.030237	JobRole_Nurse
0.000000	YearsWithCurrManager
0.040703	EducationField_Other
0.126755	EducationField_Marketing
0.132161	EducationField_Life Sciences
0.162011	EducationField_Human Resources
0.179332	Gender
0.184448	Department_Cardiology
0.301628	EducationField_Technical Degree
0.468850	BusinessTravel
0.603450	NumCompaniesWorked
0.740551	DistanceFromHome
0.939236	YearsSinceLastPromotion
1.785012	OverTime

결과 및 모델 활용방안

3. 이탈 비율에 대한 신뢰구간 계산 – 모비율의 구간추정

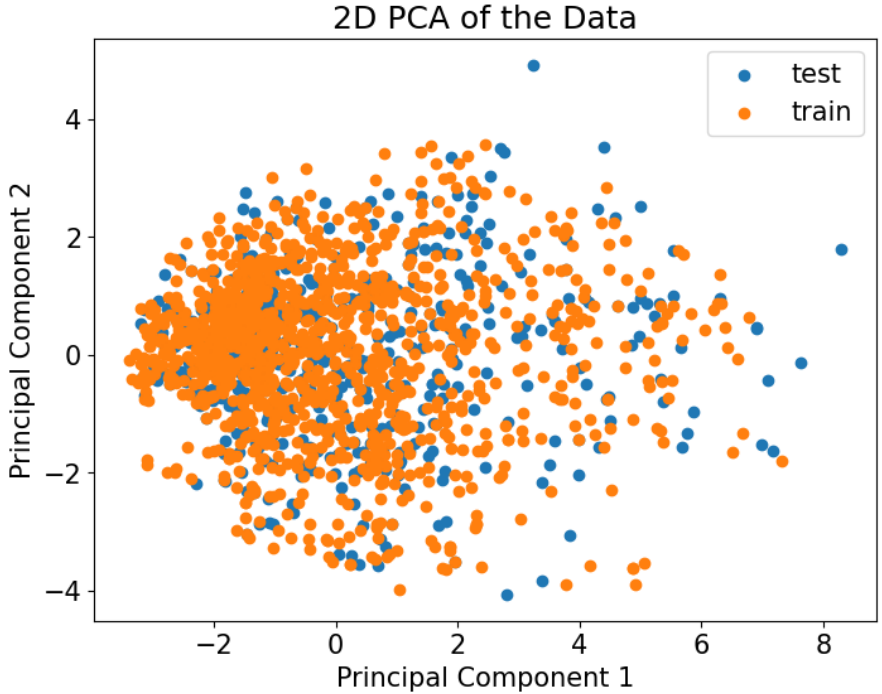
- 훈련 데이터의 이탈 비율을 사용하여 모비율의 구간 추정 수행

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- 이탈 비율의 신뢰구간 상한 : 0.14
- 이탈 비율의 신뢰구간 하한 : 0.10

결과 및 모델 활용방안

3. 이탈 비율에 대한 신뢰구간 계산 – unseen / train data의 분포 비교



- PCA를 이용해 서로 다른 데이터의 분포 비교 가능
- 훈련 데이터와 unseen 데이터의 2차원 투영 결과
- unseen 데이터의 이탈 비율 = 0.119
- 이탈 비율을 고려한 보수적인 채용 등

THANKS

2023 Fall - Data Mining

김주혁 송준현 오승준 윤서환