머신러닝을 이용한 대졸자 조기 이직의도 예측 및 요인 분석
: 로지스틱, GBM, 랜덤 포레스트를 중심으로

1828013 송혜준, 1833011 손승현, 1790019 김민정

Table of Contents

서론

1. 연구 배경
2. 연구 목적 및 문제
3. 선행 연구 고찰
4. 연구 방법 및 구성

1-1. 연구 배경

"신입사원의
높은 이직률"

- 2021년 10월, 잡코리아가 20대와 30대 남녀 직장인 343명을 대상으로 '첫 이직 경험'에 대해 실시한 결과, 75.5%가 이직을 해본 것으로 드러남. 이들 중 상당수가 입사 1년이 채 되지 않아 퇴사를 결정한 것으로 나타남.

- 이직 경험이 있다고 밝힌 2030세대 직장인들 중 첫 이직 시기를 '1년 미만'으로 선택한 이들이 37.5%로 가장 많았음.

한경닷컴, 2021.12.6.
"옆팀 막내 또 관뒀대"…떠나는 신입들, 돈 때문이 아니었다 [곽용희의 인사노무노트]」

1-1. 연구 배경

“기업에
부정적 영향 초래”

◇뽑았는데 바로 나가면 기업에도 큰 부담

- 한국경영자총협회가 2013년 발표한 대졸 신입사원 1인당 교육 비용은 평균 5960만원. 대기업은 8630만원, 중소기업은 3474만원.
  업계에서는 대략 6000만~1억2000만원이 든다고 봄.

  따라서, 조기 퇴사자가 늘어나면 그만큼 기업의 부담도 커짐.

- 조기 퇴사자의 증가가 교육비용 손실만 초래하는 것이 아니다. 취업포털 ‘사람인’이 진행한 조사를 보면 기업들은 조기 퇴사자로 인해 ‘추가 채용으로 시간과 비용 손실’, ‘기존 직원의 업무량 증가’, ‘기존 직원의 사기 저하’, ‘잦은 채용으로 기업 이미지 실추’ 등을 우려.

1-2. 연구 목적 및 연구 문제

연구 목적

1. 다양한 머신러닝 기법을 이용한 대졸 신입사원 조기 이직의도 예측 모형 구축
2. 조기 이직의도 결정 요인 중요도 파악
3. 대졸 청년층 신입사원의 조기이직 문제 개선을 위한 기초자료 마련

연구 문제

1. 대졸 청년층의 조기 이직을 예측하는 주요 결정요인 변수는 무엇인가?
2. 대졸 청년층의 조기 이직 예측 중요도 상위 변수와 조기 이직 확률의 영향력의 패턴은 어떻게 나타나는가?

1-3. 선행연구 고찰

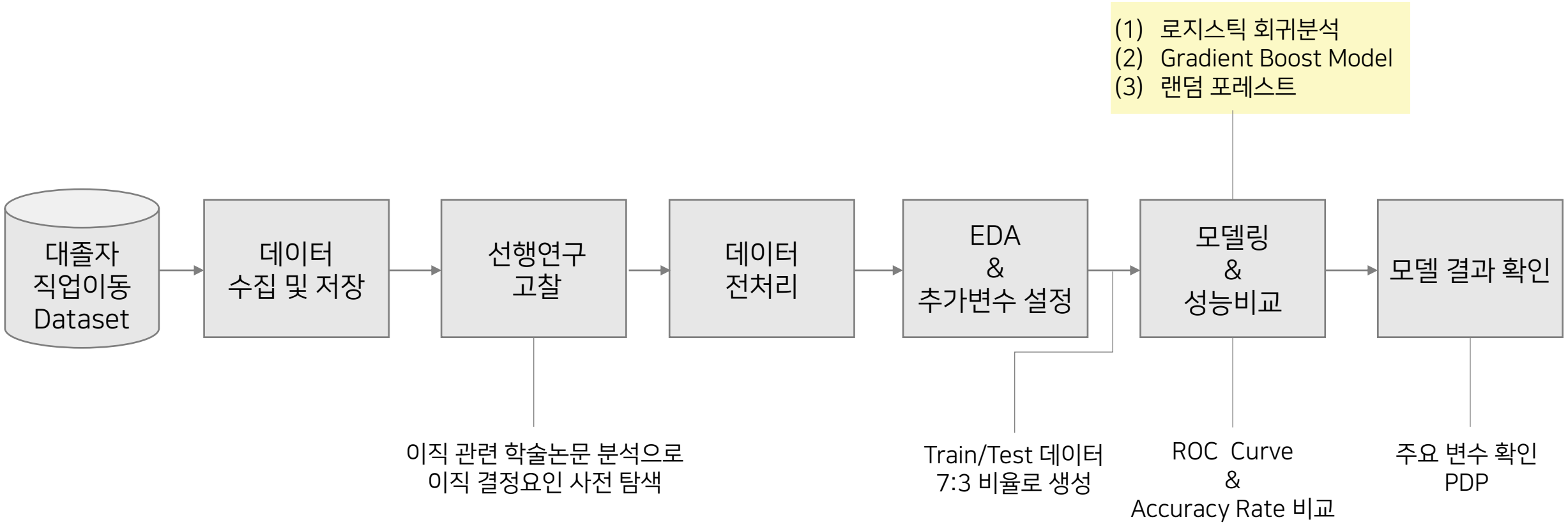
- **머신러닝 기법을 활용한 대졸 구직자 취업 예측모델에 관한 연구 (이동훈 외, 2020.6)**
  - 대졸자들의 취업 여부를 예측해 각 기법 별 성능비교, 취업 여부에 영향을 미치는 중요 변수 도출
  - 의사결정나무, 랜덤 포레스트 기법, 인공신경망 기법 활용
- **랜덤 포레스트를 활용한 대졸 신입사원 조기이직 예측 결정요인 탐색 (이은정 외, 2020.3)**
  - 조기이직 예측에 머신러닝 기법을 활용한 논문
  - 랜덤 포레스트 기법 1가지만을 활용

→ **3가지 머신러닝 기법**으로 최적의 **이직의도** 예측 모형을 찾고,

이직의도의 **주요 결정요인을 분석**하고자 함

1-4. 연구 방법 및 구성

연구 데이터: 「2019년 대졸자 직업이동경로조사」



본론

1. 데이터 수집 및 저장
2. 데이터 전처리
3. EDA
4. 추가 변수 설정
5. 모델링

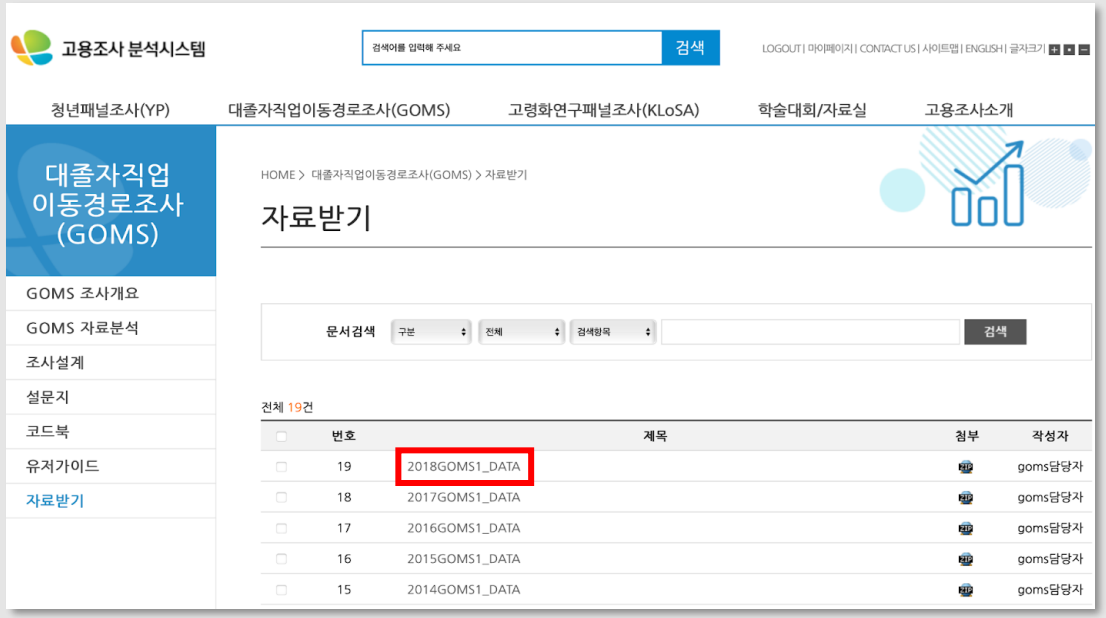
2-1. 데이터 수집 및 저장

고용조사 분석시스템

「2019년 대졸자 직업 이동 경로조사 (2018GOMS)」

한국고용정보원 주관

대졸자 경력개발 및 직업이동 횡단면 조사



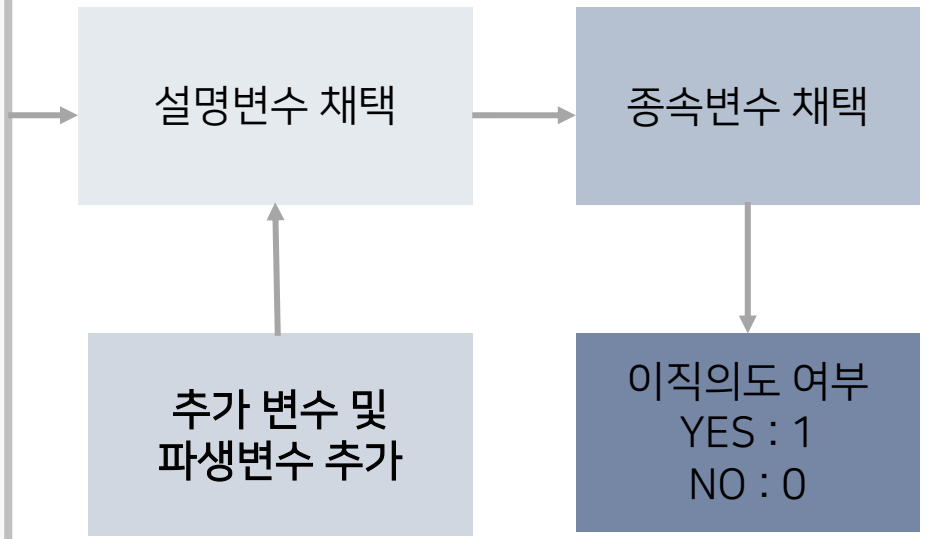
고용조사 분석시스템(https://survey.keis.or.kr/index.jsp)에서 자료 수집

2018년도에 2~3년제 대학 이상 고등교육과정을 이수한 졸업자 대상

(졸업 후 18개월 이내)

2-2. 데이터 전처리

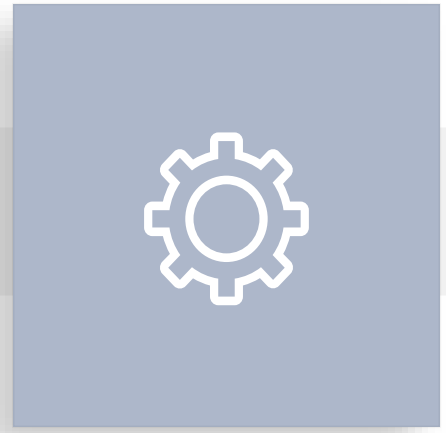
개인 특성 요인	직무 관련 요인	조직 구조적 요인	사회 심리적 요인
성별	직무 만족도	조직형태	개인의 발전 가능성 만족도
혼인상태	전공-직무 일치도	조직규모	직무사회적 평판 만족도
연령	직무-교육수준 일치도	직장 만족도	일자리 사회적 평판 만족도
전공계열	직무-적성 일치도	월평균 초임	인간관계 만족도
부모님 자산 규모	직무 관련 교육 훈련 만족도	임금만족도	
졸업평점	일의 자율성과 권한 만족도	근로시간 만족도	
재학중 일자리 경험 여부	일-자신 기술수준 일치도	복리후생제도 만족도	
		법정퇴직금 혜택 여부	
		근무환경 만족도	
		인사체계 만족도	
		고용안정성 만족도	
		노동조합유무	



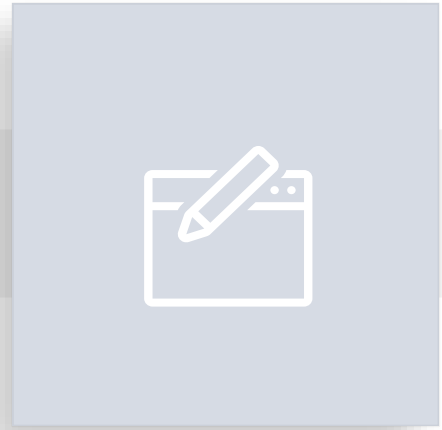
2-2. 데이터 전처리

데이터 탐색 및 조회

결측치 제거



이상치 제거



범주형 변수
척도 변경

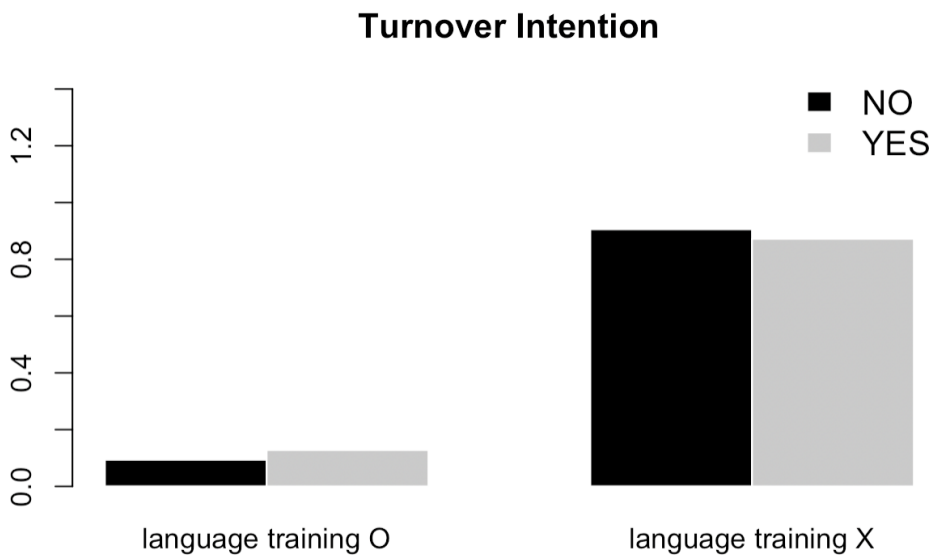
GPA 기준	변경 척도
GPA ≥ 3.7 (A대)	1
2.7 ≤ GPA <3.7 (B대)	2
GPA < 2.7 (C대)	3

2-3. EDA

구분		전체 분석 대상자	이직의도 없는 자	이직의도 있는 자
전체		11328(100)	8530(100)	2798(100)
성별	남	6406(56.5)	4976(58.3)	1430(51.1)
	여	4922(43.5)	3554(41.7)	1368(48.9)
혼인여부	혼인	433(3.9)	354(4.2)	79(2.9)
	비혼인	10895(96.1)	8176(95.8)	2719(97.1)
전공계열	인문	1343(11.9)	942(11.1)	401(14.3)
	사회	2247(19.8)	1631(19.1)	616(22.0)
	교육	806(7.1)	668(7.8)	138(4.9)
	공학	3495(30.9)	2707(31.7)	788(28.2)
	자연	1448(12.8)	1116(13.1)	332(11.9)
	의약	843(7.4)	663(7.8)	180(6.4)
	예체능	1146(10.1)	803(9.4)	343(12.3)
기업규모	구성원 100명 미만	1427(12.6)	1052(12.3)	375(13.4)
	구성원 100명 ~ 300명 미만	1608(14.2)	1239(14.5)	369(13.2)
	구성원 300원 ~ 1000명 미만	638(5.6)	495(5.8)	143(5.1)
	구성원 1000명 이상	7655(67.6)	5744(67.3)	1911(68.3)
기업형태	국내 사기업	7476(66.0)	5483(64.3)	1993(71.3)
	외국계 사기업	335(3.0)	228(2.7)	107(3.8)
	공공정부기관	1457(12.9)	1168(13.7)	289(10.3)
	교육기관	1265(11.1)	1049(12.3)	216(7.7)
	연구기관	116(1.0)	83(1.0)	33(1.2)
	기타	679(6.0)	519(6.0)	160(5.7)

2-4. 추가변수 설정

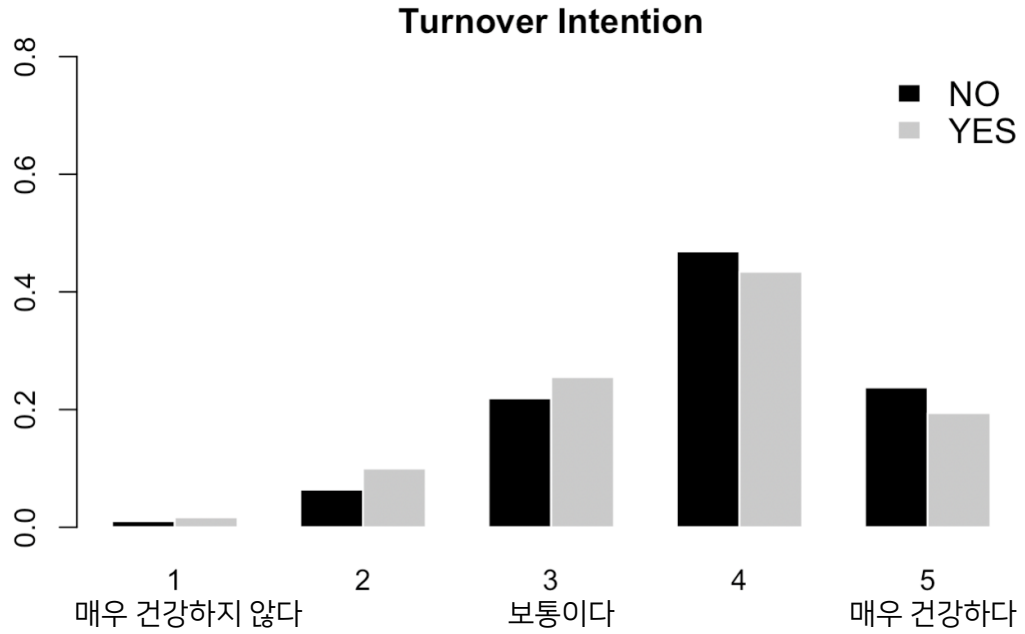
① 어학연수 여부(G181I001)



영어능력이 채용 승진 기준인 일자리에 근무하면서 영어 능력에 대한 투자 ↑

→ 영어 실력 투자에 대한 보상을 높이기 위하여 다른 직장으로 이직 ↑

② 현재 건강상태(G181Q001)

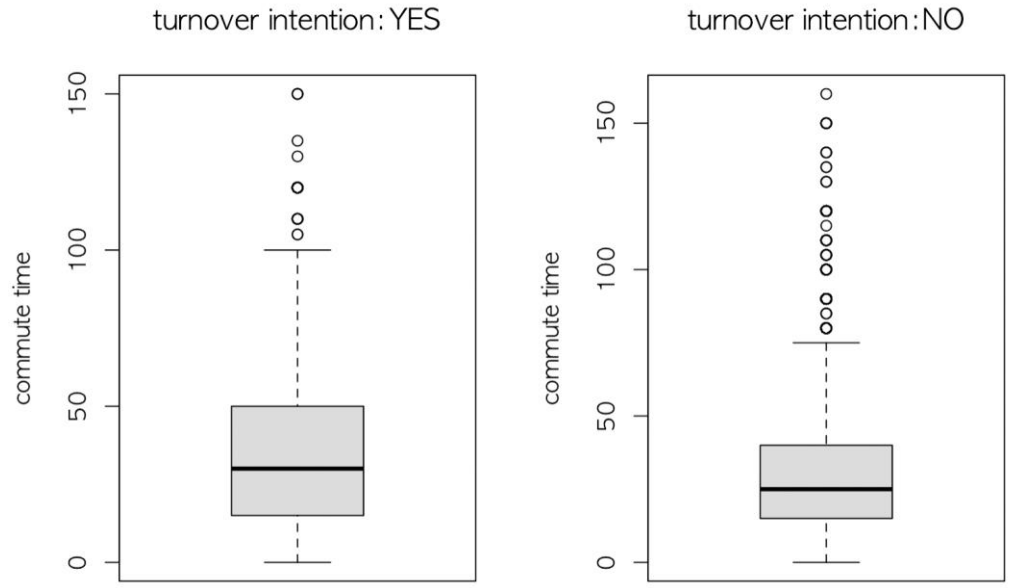


직무소진(job burnout)은 근로자의 정신 건강과 신체 건강 등 직업 건강을 측정하는 데 있어 중요한 진단적인 역할.

직무소진은 이직의도에 영향을 미침.

2-4. 추가변수 설정

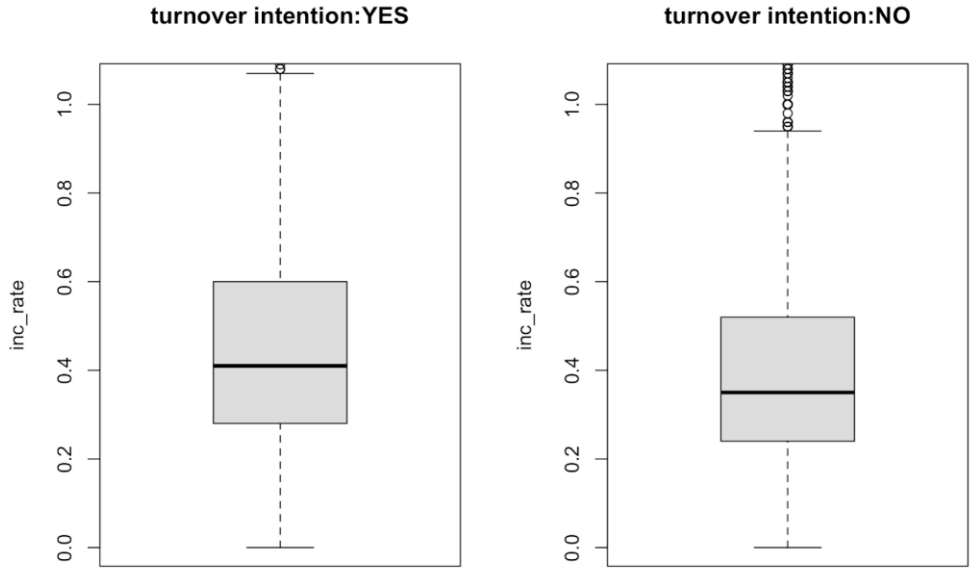
③ 출근 소요 시간(commute)



출퇴근 시간이 길어지면

직장-개인생활 균형에 부담 가중.

④ 생활비/월평균 초임(inc_rate) : 파생변수



개인의 경제적 상황을 반영한

절대적인 소득 관련 지표 생성.

2-5. 모델링

Machine Learning Model

- **로지스틱 회귀분석**

  범주형 종속변수의 Classification과 예측을 동시에 구할 수 있는 기법.
  Odds Ratio로 해석하기 쉬움.

- **Gradient Boost Model (GBM)**

  경사하강법(Gradient Descent)을 이용하여 가중치를 업데이트하여 오류를 개선해 나가는 알고리즘으로 뛰어난 예측성능을 보임.

- **랜덤 포레스트**

  배깅과 부스팅보다 더 많은 무작위성을 주어 예측력이 매우 높은 방법.
  특히 입력변수의 개수가 많을 때에는 Bagging이나 Boosting과 비슷하거나 더 좋은 예측력을 보이는 경우가 많음.
  또한 Tuning Parameter가 적어서 실제 자료분석에 쉽게 사용될 수 있음.

2-5. 모델링

(1) 로지스틱 회귀분석

① 모든 37개 설명변수를 포함한 Full Model로 시작

② Stepwise AIC를 활용하여 **Reduced Model** 도출

③ 변수간 다중공산성 확인

→ 모든 VIF값 < 10, ∴ **Reduced Model** 사용

```
# Final Model
fit_aic<-stepAIC(fit)
summary(fit_aic)
```

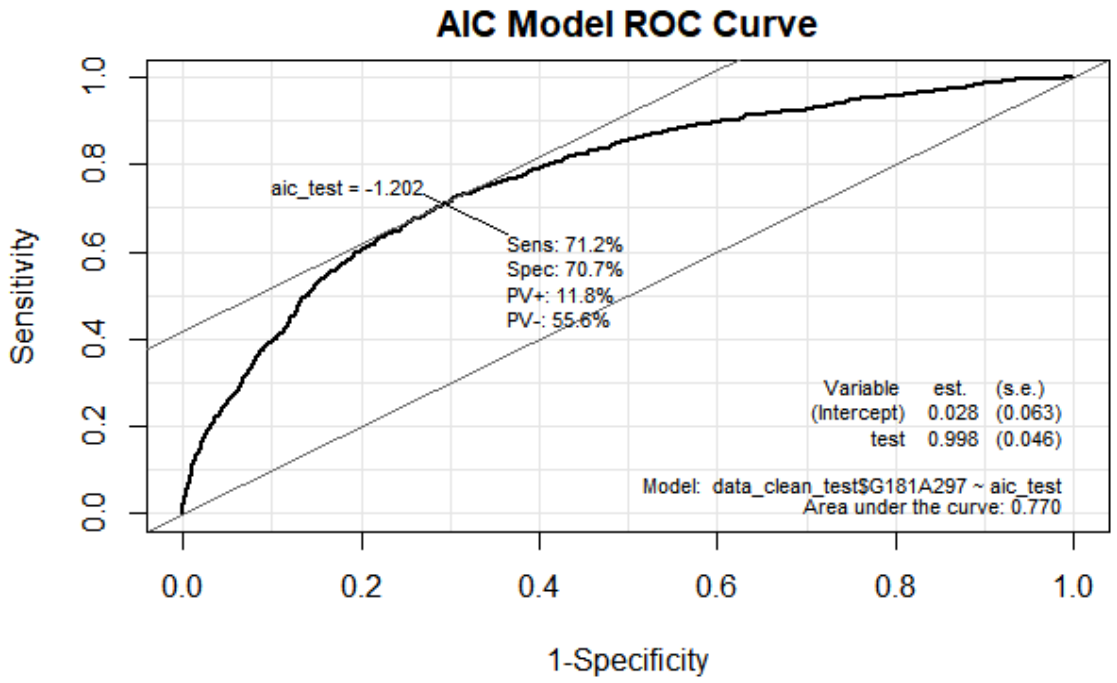
```
Call:
glm(formula = G181A297 ~ G181P001 + G181H001 + G181A141 + G181A144 +
    G181A142 + G181A136 + G181A140 + G181A125 + G181A126 + G181A038 +
    G181A129 + G181A134 + G181A127 + G181A166 + G181A131 + G181A135 +
    G181A137 + G181A132 + G181M001 + G181R023 + G181I001 + commute +
    grade + org_form, family = binomial, data = data_clean_train)
```

채택 변수	Pr(>|z|)
G181A140 (직장만족도)	< 2e-16 ***
G181A131 (개인 발전 가능성 만족도)	2.30e-15 ***
G181A142 (직무-교육수준 일치도)	1.20e-11 ***
G181A125 (월 평균 초임)	3.73e-09 ***
G181A038 (법정퇴직금 혜택 여부)	2.91e-07 ***
G181A127 (고용안정성 만족도)	1.18e-06 ***
G181H001 (재학중 일자리 경험 여부)	8.54e-05 ***
G181A144 (전공-직무 일치도)	0.000188 ***
G181I001 (어학연수 여부)	0.000537 ***
G181M001 (자격증 여부)	0.002809 **

2-5. 모델링

(1) 로지스틱 회귀분석 – Test data로 성능 확인

<ROC Curve>



Area Under Curve : 0.777

<Confusion Matrix>

Predicted	Observed 0	1
0	2408	151
1	609	231

Accuracy Rate :77.64%

2-5. 모델링

(2) Gradient Boost Model

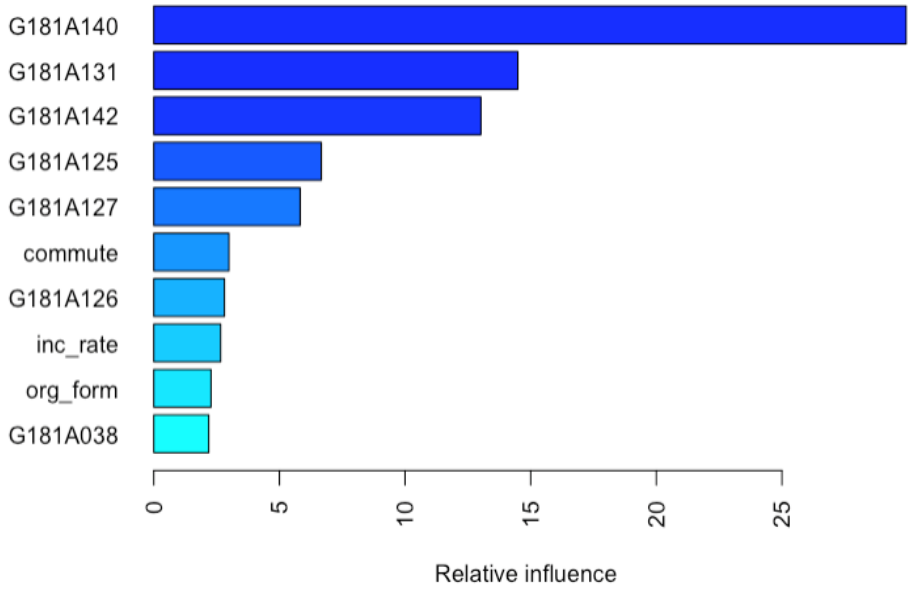
① 예측력을 높이기 위한 하이퍼 파라미터 튜닝

```
# create hyperparameter grid
hyper_grid <- expand.grid(
  shrinkage = c(.01, .1, .3),
  interaction.depth = c(1, 3, 5),
  n.minobsinnode = c(5, 10, 15),
  bag.fraction = c(.65, .8, 1),
  optimal_trees = 0,
  min_RMSE = 0
)
```

② 최종 모델 선정

```
# train GBM model
gbm.fit.final <- gbm(
  formula = G181A297 ~ .,
  data = data_clean_train,
  n.trees = 1385,
  interaction.depth = 3,
  shrinkage = 0.01,
  n.minobsinnode = 10,
  bag.fraction = .8,
  train.fraction = 1,
  n.cores = NULL, # will use all cores by default
  verbose = FALSE
  )
```

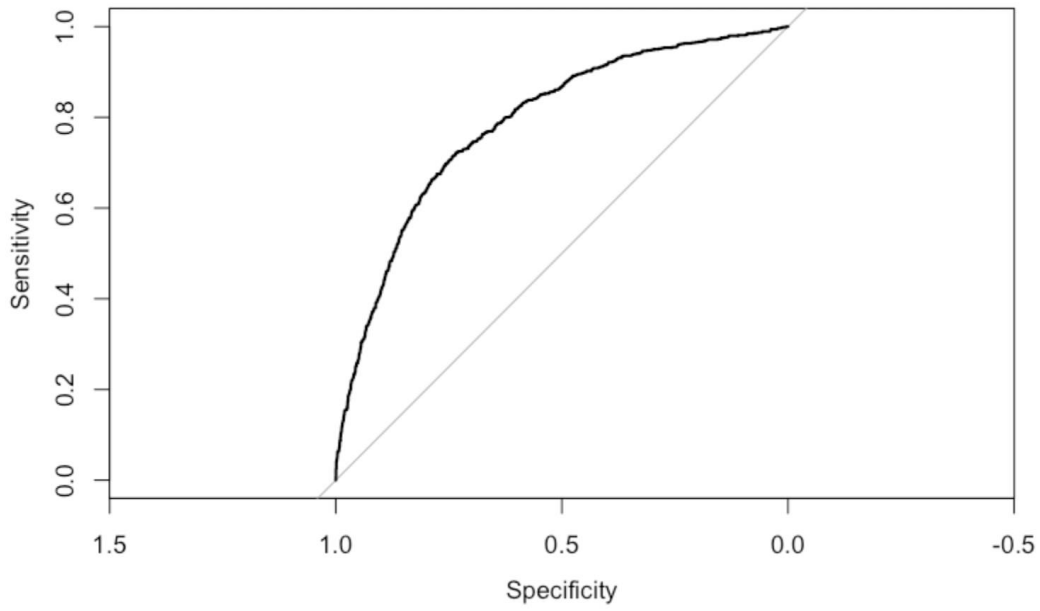
③ 변수 중요도 확인



2-5. 모델링

(2) Gradient Boost Model – Test data로 성능 확인

<ROC Curve>



<Confusion Matrix>

	0	1
0	2393	166
1	571	269

Area Under Curve : 0.787

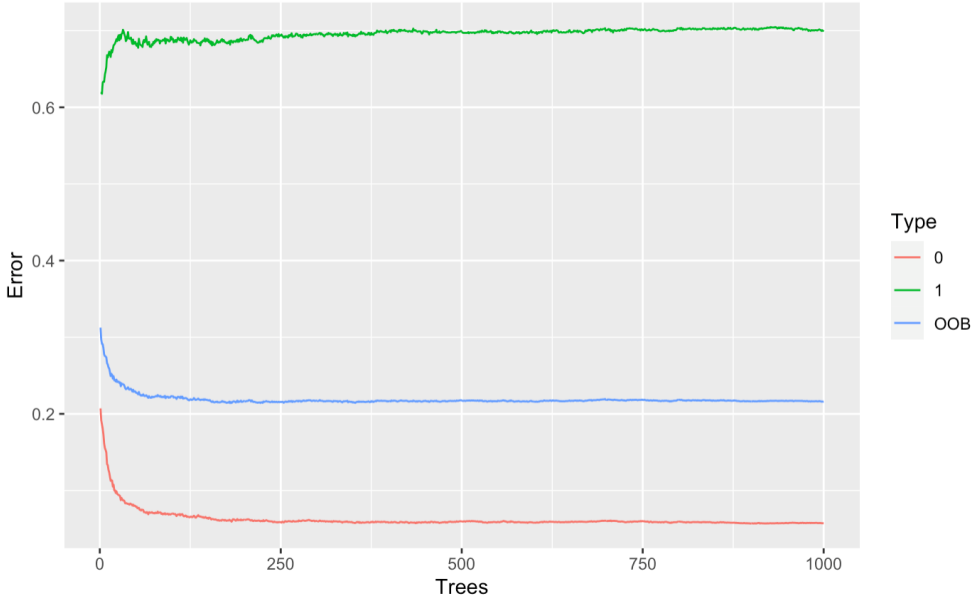
Accuracy Rate :78.31%

2-5. 모델링

(3) 랜덤 포레스트

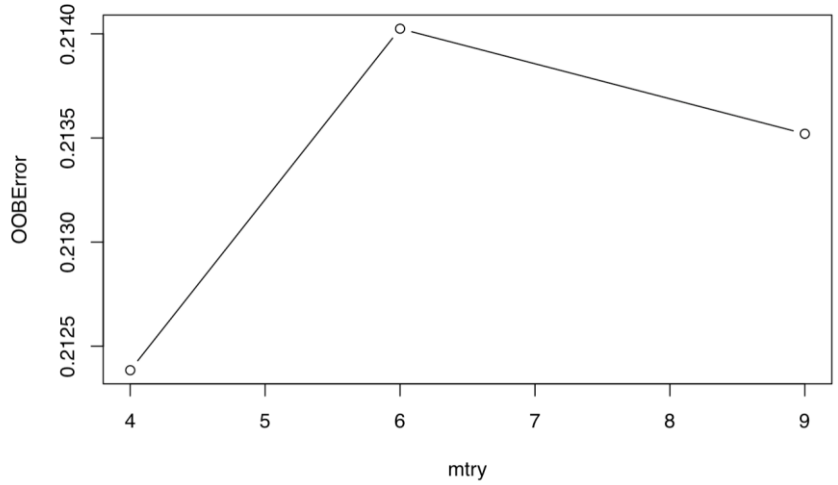
① 예측력을 높이기 위한 하이퍼 파라미터 튜닝

1) Number of Trees:



Tree가 500개 이상일 때 Error가 안정화

2) Number of mtry:



	mtry	OOBError
4.OOB	4	0.2123849
6.OOB	6	0.2140245
9.OOB	9	0.2135200

mtry=4 일때 OOB Error가 가장 작음

2-5. 모델링

(3) 랜덤 포레스트

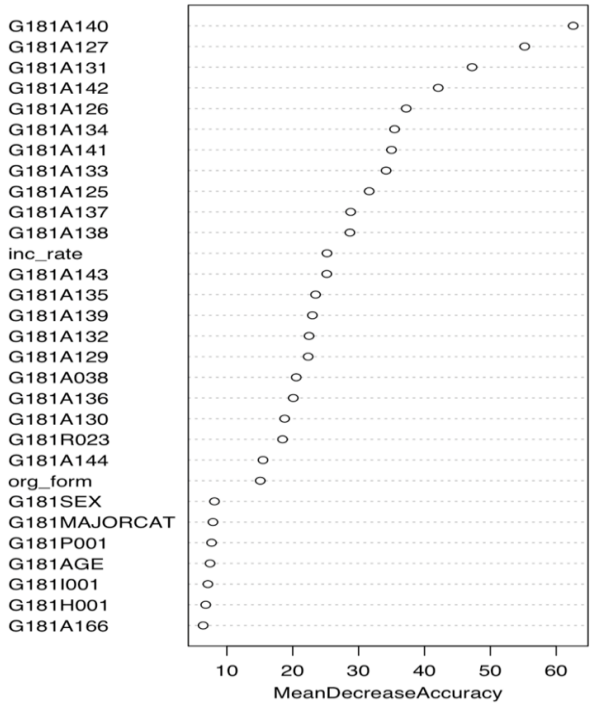
② 최종모델 선정

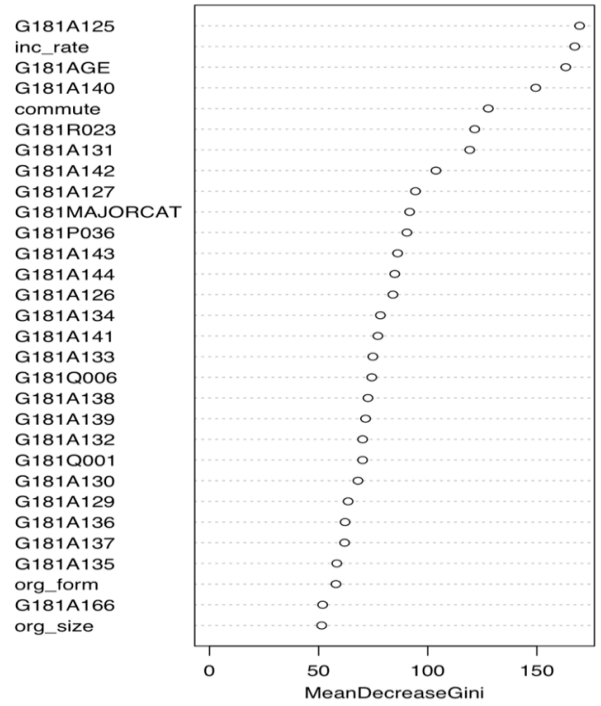
```
# Modeling
final = randomForest(G181A297 ~., data = data_clean_train, importance = TRUE,
                     ntree = 1000, mtry = 4)
```

```
Call:
 randomForest(formula = G181A297 ~ ., data = data_clean_train,      importance = TRUE, ntree =
1000, mtry = 4)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 4

        OOB estimate of  error rate: 20.91%
Confusion matrix:
     0   1 class.error
0 5669 302  0.05057779
1 1356 602  0.69254341
```

③ 변수 중요도 확인

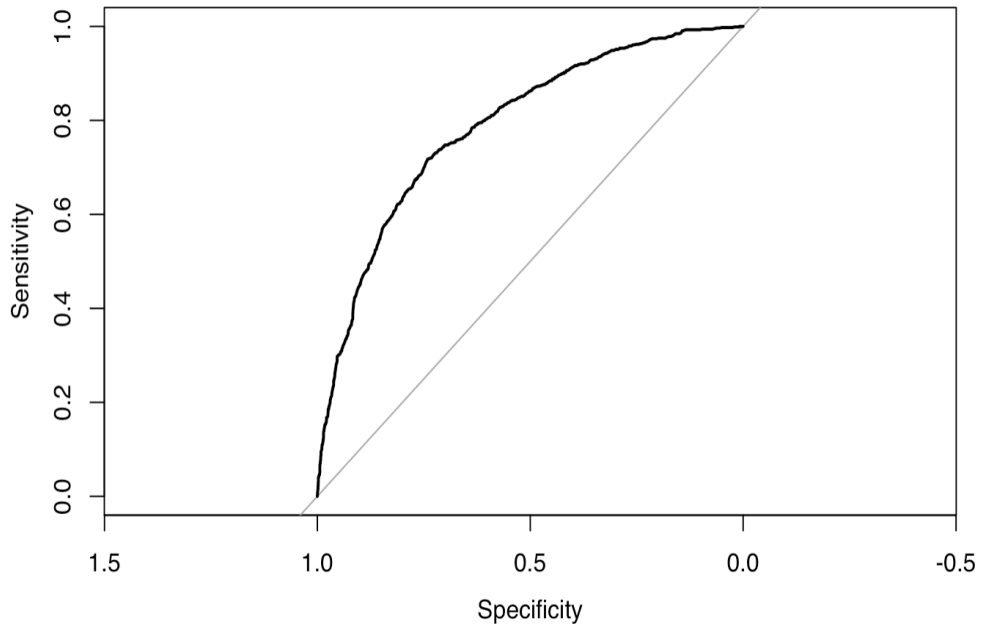




2-5. 모델링

(3) 랜덤 포레스트 – Test data로 성능 확인

<ROC Curve>



Area Under Curve : 0.7887

<Confusion Matrix>

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2445  606
         1  114  234
```

Accuracy Rate :78.82%

결론

1. 각 예측 모델들 간의 성능 비교

2. 중요 변수 확인

3-1. 각 예측 모델 간의 성능 비교

성능 비교를 통한 최적의 대졸자 이직의도 예측 모형 선택

Area Under Curve, Accuracy Rate

	로지스틱 회귀분석	GBM	**랜덤 포레스트**
AUC	0.777	0.787	**0.7887**
Accuracy Rate (%)	77.64	78.31	**78.82**

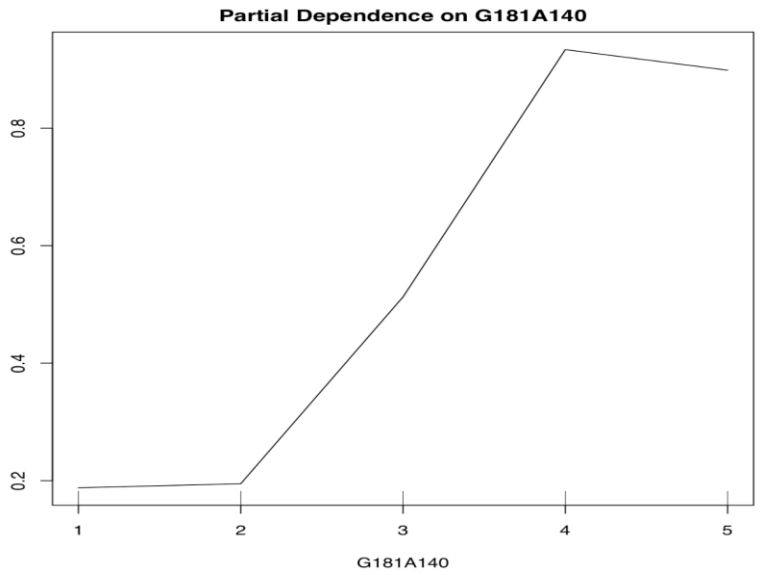
3-2. 중요 변수 확인

대졸자 이직이도에 영향을 미치는 중요도 상위 10개 변수

개인 특성 요인	inc_rate(월평균 생활비 지출액/월평균 초임) G181AGE(나이) commute(출근 시간)
조직 구조적 요인	G181A140(직장 만족도) G181A127(고용안정성 만족도) G181A126(임금 만족도) G181A125(월평균 초임)
직무 관련 요인	G181A142(직무-교육 수준 일치도)
사회 심리적 요인	G181A131(개인 발전 가능성 만족도)

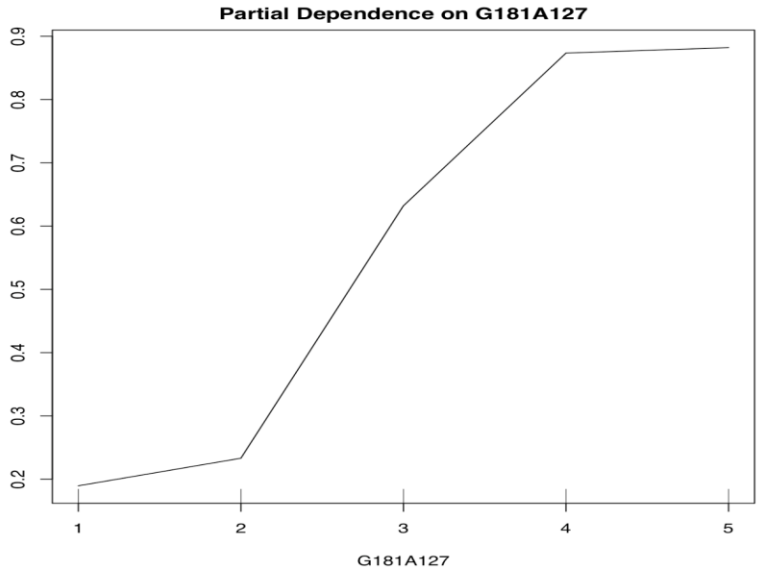
3-2. 중요 변수 확인

Partial Dependence Plot (PDP)



직장만족도가 2점에서 4점으로 증가할 수록
이직의도가 있을 확률이 평균적으로 증가.

4점에서 5점으로 증가하면 이직의도가 있을 확률이
평균적으로 줄어듦.

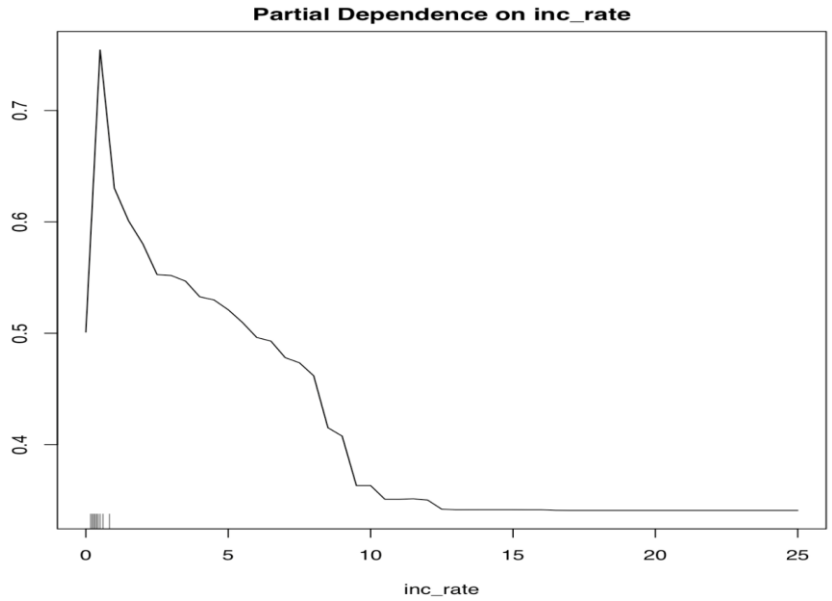


고용안정성 만족도는 전체적으로 증가할 수록
이직의도가 있을 확률이 평균적으로 증가.

특히 2점에서 4점으로 증가할수록 이직의도가 있을 확률이
평균적으로 증가.

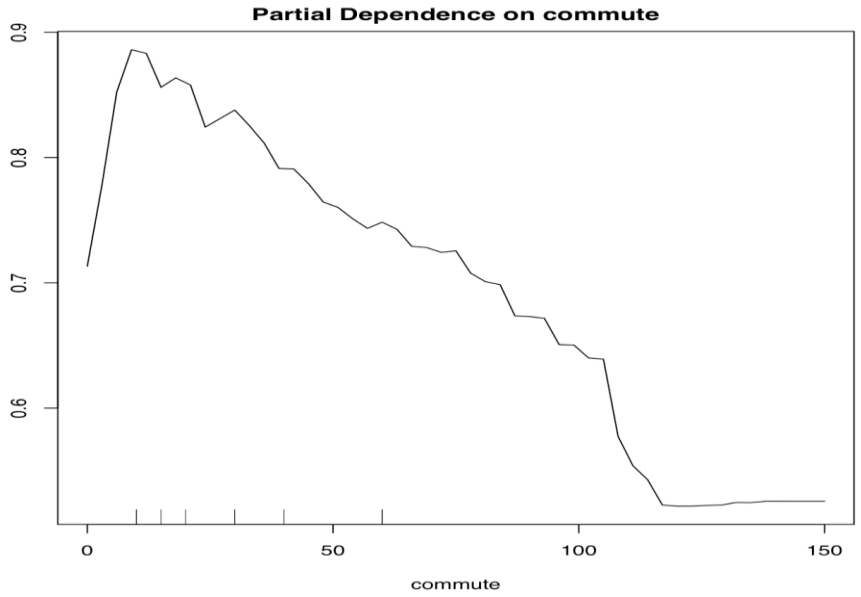
3-2. 중요 변수 확인

Partial Dependence Plot (PDP)



월평균 생활비 지출액/월평균초임 비율이 0에서 1까지 증가할 수록 이직의도가 있을 확률이 평균적으로 높아지지만

1 이후부터는 비율의 값이 커질수록 이직의도가 있을 확률이 평균적으로 감소.



출근 시간 (분)이 0에서 15분까지 증가할 수록 이직의도가 있을 확률이 평균적으로 증가하지만

15분 이후부터는 출근시간이 증가할 수록 이직의도가 있을 확률이 평균적으로 감소.

논의

1. 연구 의의

2. 연구 한계 및 연구 방향 제시

3. 참고 문헌

4-1. 연구 의의

여러가지 머신러닝 기법을 이용하여 대졸 청년층 신입사원의 이직의도를 예측하는 **다양한 요인** 탐색.
이를 통해 최적의 이직의도 예측 모형을 제시함으로써 기존 연구들과는 차별화된 정확도 높은 모형 제시.

이직의도를 예측하는 **중요도 상위 10개 설명변수**를 제시함으로써 신규 인력의 조직 안착을 위해
기업이 노력해야 할 부분에 대한 시사점 제시.

생활비 대비 월 평균 초임(inc_rate), 출근 시간(commute)등 **파생 변수**와 **추가변수**를 통해
대졸자 이직의도와 관련된 새로운 주요 변수 발견.

4-2. 연구 한계 및 연구 방향 제시

1) 연구 한계점

① 본 연구에서는 랜덤 포레스트 모델 성능이 가장 높게 산출되었지만
로지스틱 회귀 및 앙상블 기법 외 기타 사용하지 않은 분류기법을 적용한 최적의 모델로 산출될 가능성 존재.

② 본 연구에서는 종속 변수를 종사상 지위, 회사 종류, 직업 등으로 세분화하지 않고 예측함.

2) 연구 방향 제시

① 향후 보다 다양한 분류 기법을 활용하여 더 나은 성능과 안정성을 갖춘 모델을 제시해야 함.

② 향후 연구에서 종속 변수를 세분화 하여 각 계급별 중요한 변수들을 도출해낼 수 있을 것으로 기대함.

4-3. 참고문헌

[논문]

1. 이동훈, 김태형. "머신러닝 기법을 활용한 대졸 구직자 취업 예측모델에 관한 연구", 정보시스템연구, 29.2, (2020) : 287.
2. 이은정, 조희숙, 송영수. "랜덤 포레스트를 활용한 대졸 신입사원 조기이직 예측 결정요인 탐색", 기업교육과 인재연구, (2020): 163-193.
3. 황광훈. “청년층의 이직 결정요인 및 임금효과 분석”, 한국직업능력연구, 22.1, (2019): 144.

[기사]

1. 박용희, 「"옆팀 막내 또 관뒀대"…떠나는 신입들, 돈 때문이 아니었다 [곽용희의 인사노무노트]」, “한경닷컴", 2021.12.5., https://www.hankyung.com/society/article/202112045467i, 접속일 2021.12.6.
2. 송혜미, 「사라진 평생직장, 이직 꿈꾸는 직장인들…이직사유 1위는 ‘연봉’」, “동아닷컴”, 2021.11.08., https://www.donga.com/news/Economy/article/all/20211108/110136096/1, 접속일 2021.12.6
3. 송혜미, 「직장인 10명 중 8명 “직장 옮기겠다”… 경력 채용 늘며 커지는 이직 시장」, “동아닷컴”, 2021.12.10., https://www.donga.com/news/article/all/20211108/110143550/1, 접속일 2021.12.6.

[기타]

1. “대졸자 직업이동경로조사", https://survey.keis.or.kr/goms/goms01.jsp, 접속일 2021.11.9.
2. 김현동, “출퇴근 소요시간과 직장인의 태도", 2014 고용패널 학술대회 연구보고서, 한국노동패널조사, 2014.
3. 김현동, “일자리에서 영어능력 중요성과 개인의 영어능력이 재직기간에 미치는 영향을 생존분석”, 2017 노동패널 학술대회 연구보고서, 한국노동패널조사, 2017.

Q & A

Thank you