

기업 리뷰 데이터를 활용한 이직률 예측 모델 제안 및 기업 분석

팀: 열간이

팀원: 성균관대학교 인공지능융합전공 2020312845 김명섭

성균관대학교 인공지능융합전공 2022315160 이수형

성균관대학교 인공지능융합전공 2022314309 조정환

목차

I. 연구 소개

- 연구 배경
- 연구 목적
- 연구 flow chart

II. 데이터 수집

- 크롤링
-

III. 데이터 처리

- 토픽 모델링

IV. 모델링

V. 결론

- 분석
 - 활용방안과 기대효과
 - 한계
-

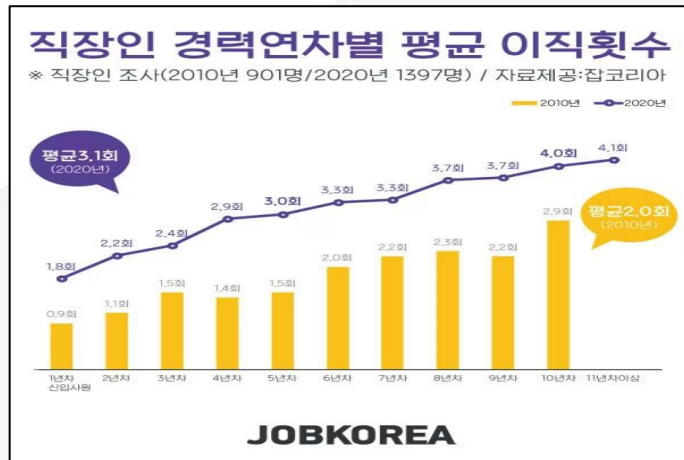
I . 연구 소개

연구 배경



출처: LX인터내셔널

조직문화의 변화



이직 횟수 증가 및 기업 대응
필요성

연구 목적

실근무자의 평가가 반영된
리뷰데이터를 활용한 이직률
예측 모델 개발

이직률 예측을 통해
기업 운영 평가와 방향 수정

*Journal of the Korean Data &
Information Science Society*
2021, 32(5), 1035-1047

<http://dx.doi.org/10.7465/jkdi.2021.32.5.1>
한국데이터정보과학회

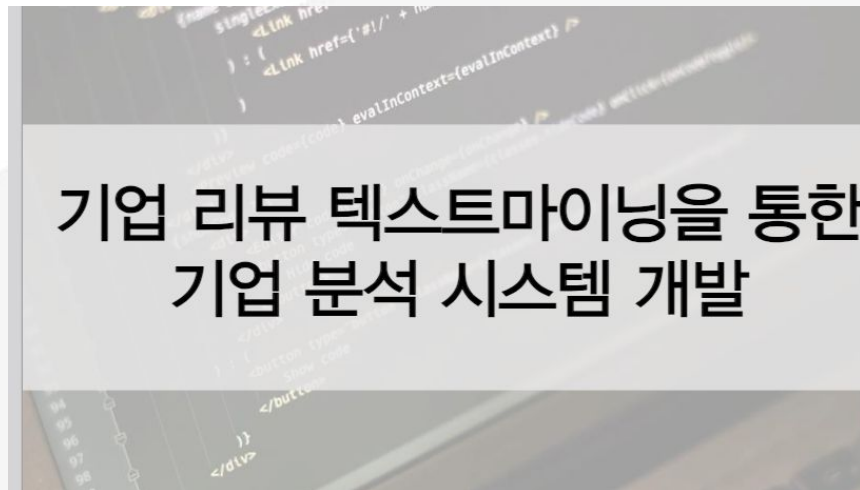
IT 기업 직원의 만족 및 불만족 요인에 따른 이직률 예측: 토픽모델링과 머신러닝을 활용하여[†]

최진욱¹ · 신동원² · 이한준³

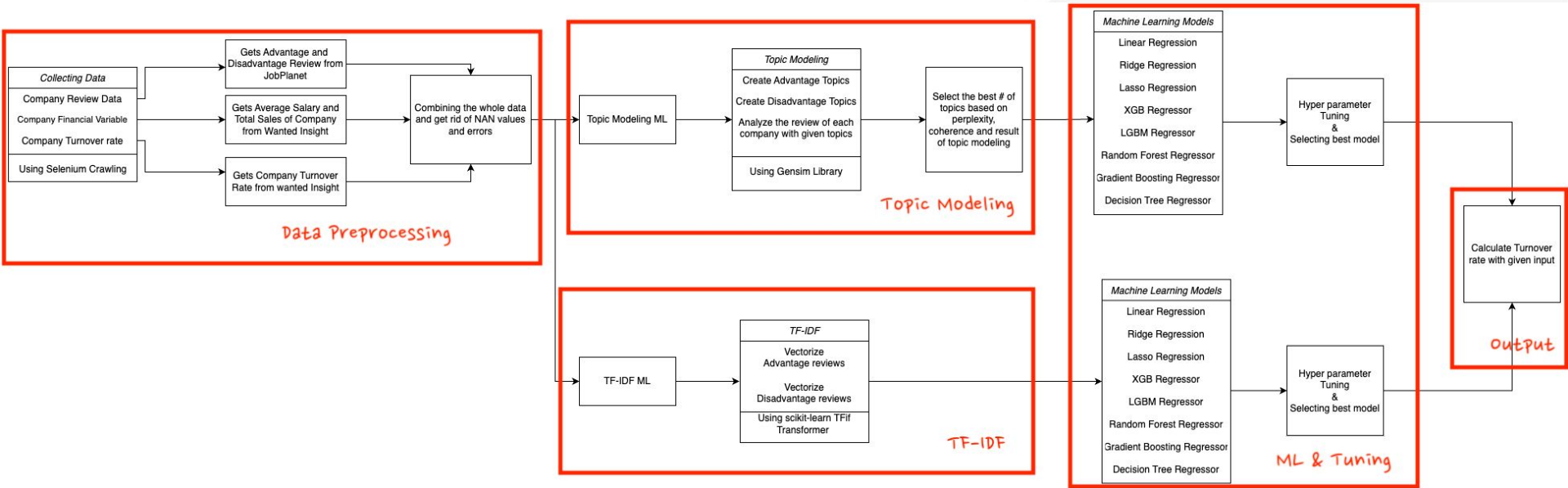
¹고려대학교 스마트미디어 서비스 연구센터 · ²명지대학교 산학협력단 · ³명지대학교 경영정보학과

접수 2021년 8월 27일, 수정 2021년 9월 7일, 게재확정 2021년 9월 7일

기업 리뷰 텍스트마이닝을 통한 기업 분석 시스템 개발



연구 flow chart



Ⅱ. 데이터 수집


리뷰 데이터 수집

잡플래닛에서 IT/웹/통신, 서비스업
등

총 10개 분야 중

100개 이상의 리뷰를 가진 회사의

2023, 2022년에 작성된 리뷰 크롤링



서울 | 2023. 09

★★★★★

승진 기회 및 가능성
■■■■■

복지 및 급여
■■■■■

업무와 삶의 균형
■■■■■

사내문화
■■■■■

경영진
■■■■■

"안정적인 통신산업군의 대기업으로 탄탄한 사업구조"

장점
재택(팀바탕), 기타 자잘한 복지, 통신비 지원, 복포, 의료비 지원 등
전반적으로 업무 강도가 그렇게 강하진 않음

단점
연봉이 너무 째
불필요한 보고, 광팔이들이 난무함, 일 처리가 답답함

경영진에 바라는 점
타사 대비 연봉 경쟁력이 약합니다.
너무 광팔이식 신사업에 휘둘리지 말고 기업을 탄탄하게 유지했으면 좋겠습니다.

이 기업은 1년 후 **성장**하고 있을 것이다.

이 기업을 추천 합니다!

재무제표 데이터 수집

원티드인사이드에서 리뷰 텍스트

데이터에 있는 회사들의

평균 연봉 및 총 매출 크롤링



예상평균연봉

12,848만원 상위 1%

월 세전 1,071만원

출처 : 금감원 2023.03



총 매출액

199.7조원 상위 1%

1인당 매출액 17.9억원

매출 대비 임금 3%

출처 : 금감원 2021.12

이직률 데이터 수집

원티드인사이드에서 리뷰 텍스트
데이터에 있는 회사들의 이직률 크롤링

이직률이 나와있지 않은 경우

이직률 = {(퇴사자 수) / (총인원)} x
100(%) 직접 계산



총 인원

35,448명 상위 1%

퇴사 3,864명 (11%)

입사 4,330명 (12%)

출처 : 국민연금 2023.09 ⓘ

Ⅲ. 데이터 처리

리뷰 데이터 전처리

결측치의 경우 아래의 상황에서 발생

- 잡플래닛 상에 기업 정보는 올라와 있으나, 2022년과 2023년에 리뷰데이터가 없는 경우
- 원티드 인사이트에 기업 정보를 찾을 수 없고 수기로도 못찾은 경우
- 원티드 인사이트에 퇴사율 혹은 퇴사자 정보가 없고 수기로도 못찾은 경우
- 원티드 인사이트에 재무제표가 없고, 수기로도 못찾은 경우

결측치는 모두 제거한 후 진행

- 기존 총 **3104**개의 기업에서 **2305**개의 기업을 대상으로 머신러닝 모델 생성

리뷰 데이터 전처리

불용어 처리

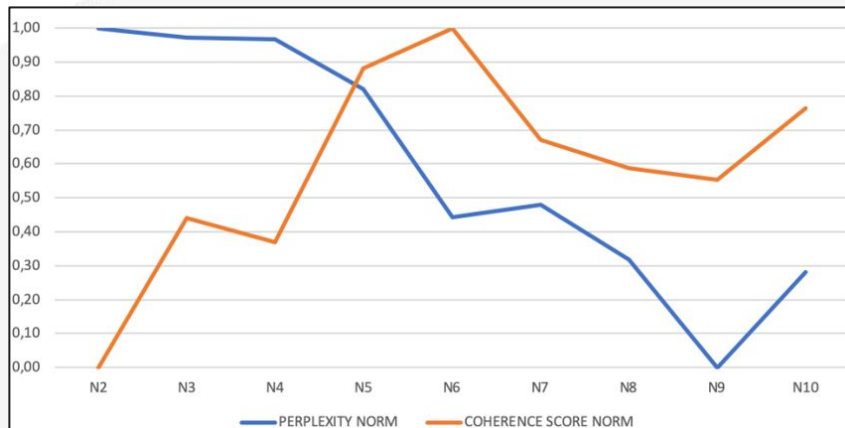
- 불용어 처리 전 단어 수
 - 장점 리뷰에서의 고유한 단어 수: **44973**
 - 단점 리뷰에서의 고유한 단어 수: **75796**
- 불용어의 경우 명사, 동사, 형용사를 제외한 단어들로 설정
- 단어의 수가 많아짐에 따라 빈도 수가 높은 **2000**개만 사용

토픽모델링

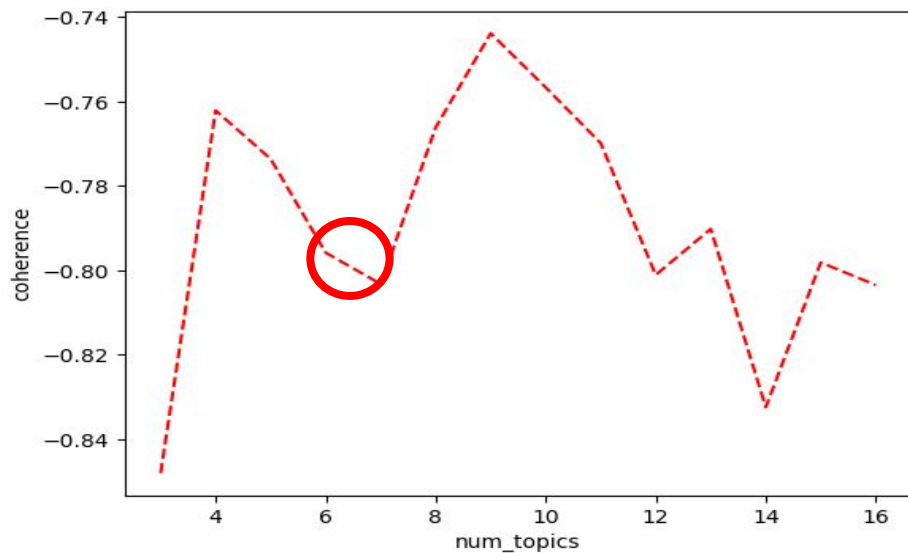
Gensim 라이브러리 활용



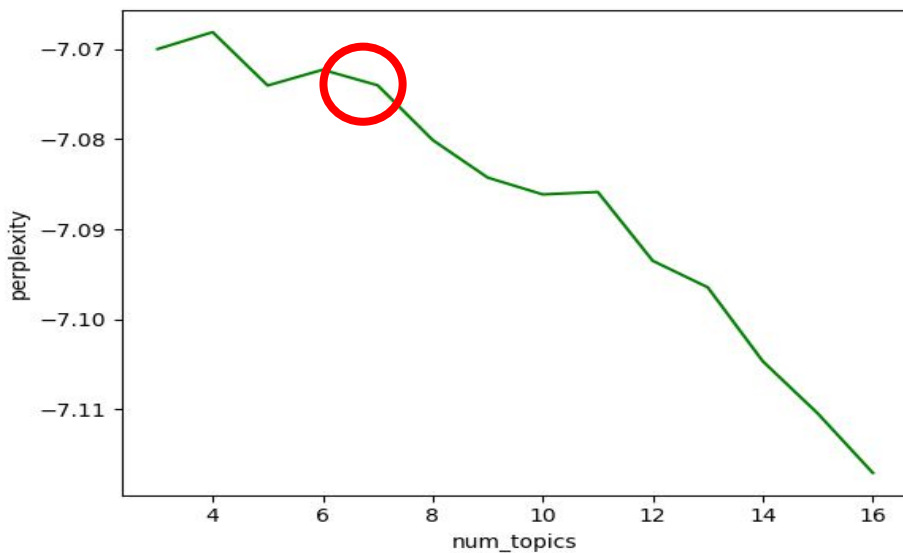
토픽 수에 따른 단어들의 **perplexity**와 **coherence** 값과
토픽 모델링 결과물을 비교하여 결정



토픽모델링

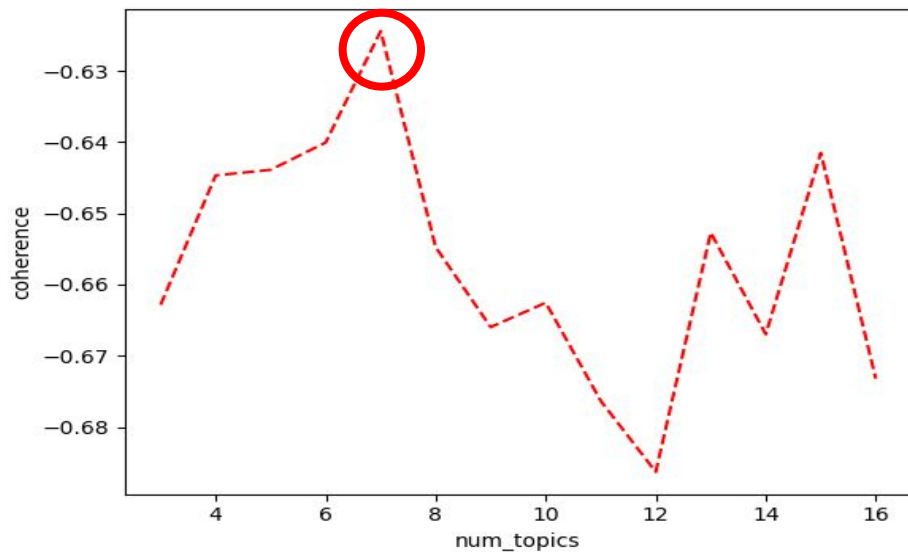


advantage coherence

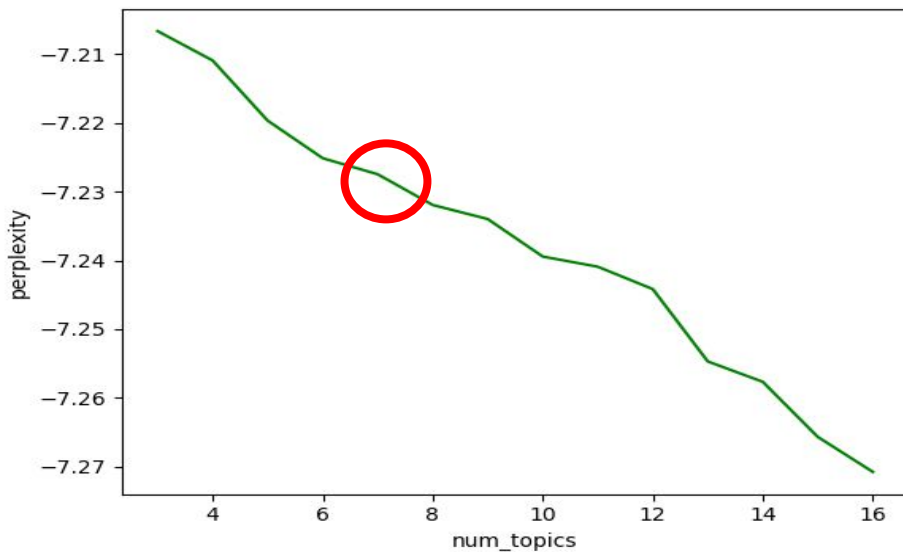


advantage perplexity

토픽모델링



disadvantage coherence



disadvantage perplexity

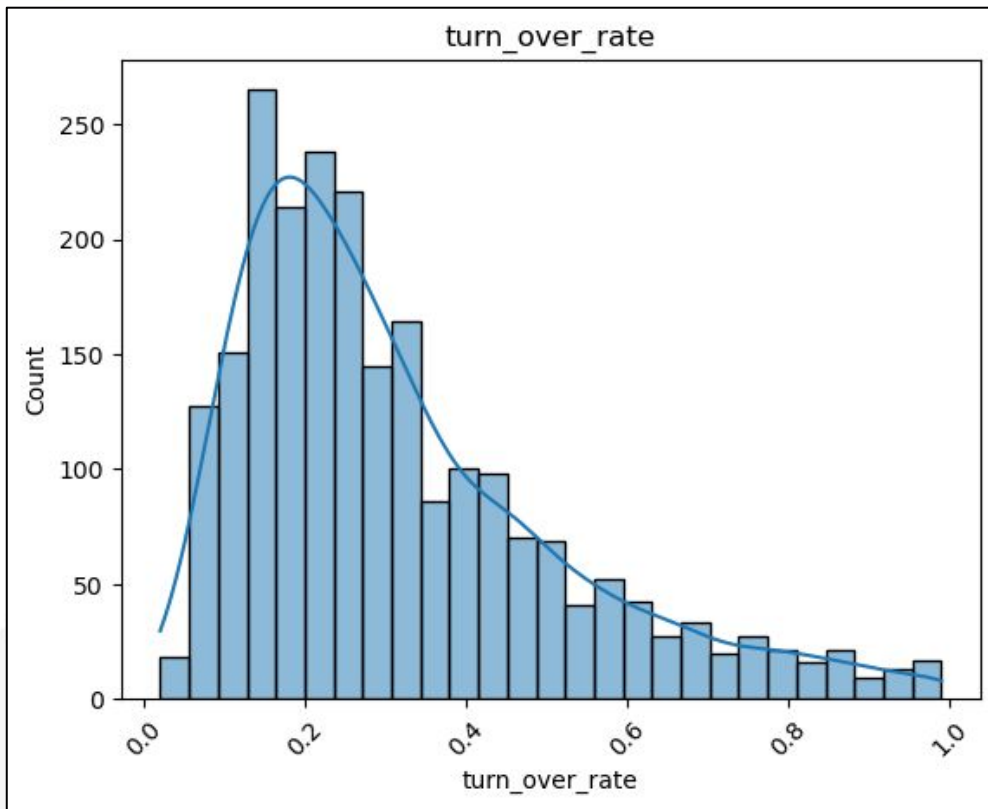
토픽모델링

장점 토픽	토픽모델링 결과물
직업 안정성 + 휴식	0.011*"안정" + 0.009*"공공기관" + 0.009*"서울" + 0.008*"부바" + 0.007*"유연근무제" + 0.007*"육아휴직" + 0.006*"정년" + 0.006*"강도" + 0.006*"높은" + 0.005*"공기업"
구내식당 + 제공서비스 만족	0.013*"점심" + 0.009*"저녁" + 0.009*"버스" + 0.009*"기숙사" + 0.008*"통근" + 0.008*"높은" + 0.008*"식당" + 0.007*"맛있음" + 0.006*"수당" + 0.006*"아침"
여가 지원	0.013*"데이" + 0.013*"금요일" + 0.012*"높은" + 0.008*"포인트" + 0.008*"제도" + 0.008*"리프트" + 0.007*"패밀리" + 0.007*"업계" + 0.006*"대비" + 0.006*"여름"
자율적 근무환경	0.013*"재택근무" + 0.008*"재택" + 0.008*"대기업" + 0.007*"동료" + 0.007*"포인트" + 0.007*"기회" + 0.006*"업계" + 0.006*"사내" + 0.005*"개인" + 0.005*"교육"
성과급 만족도	0.008*"수당" + 0.006*"강도" + 0.006*"대기업" + 0.005*"따라" + 0.005*"명절" + 0.005*"때문" + 0.005*"나눔" + 0.005*"지급" + 0.005*"사업" + 0.005*"없고"
회사 시설	0.012*"카페" + 0.012*"커피" + 0.011*"건물" + 0.010*"점심" + 0.009*"간식" + 0.009*"할인" + 0.009*"사내" + 0.009*"식대" + 0.008*"사옥" + 0.007*"사무실"

토픽모델링

단점 토픽	토픽모델링 결과물
진급의 어려움 + 수직적 문화	0.010*생산 + 0.009*진급 + 0.006*군대 + 0.005*사원 + 0.005*공장 + 0.005*수직 + 0.005*보수 + 0.005*출근 + 0.005*꼰대 + 0.005*승진
정규직으로의 전환	0.026*계약 + 0.015*정규직 + 0.008*현장 + 0.007*강도 + 0.006*사업 + 0.005*전환 + 0.005*수직 + 0.005*교대 + 0.005*군대 + 0.005*차이
평가체계부실	0.007*프로젝트 + 0.006*무량 + 0.006*평가 + 0.005*조직 + 0.004*소통 + 0.004*리더 + 0.004*팀바탕 + 0.004*부분 + 0.004*좋은 + 0.004*성과
근무지 불만	0.012*위치 + 0.006*행정 + 0.006*연구원 + 0.006*출퇴근 + 0.005*수당 + 0.005*식당 + 0.005*인상 + 0.005*교통 + 0.005*연구 + 0.005*주변
낮은 발전 가능성	0.009*임원 + 0.007*사업 + 0.007*영진 + 0.006*영업 + 0.005*경영 + 0.005*미래 + 0.005*매출 + 0.005*조직 + 0.004*성장 + 0.004*정치
고객 응대 + 성과에 대한 압박	0.011*고객 + 0.009*승진 + 0.008*영업 + 0.007*민원 + 0.007*본사 + 0.007*실적 + 0.007*스트레스 + 0.007*순환 + 0.006*매장 + 0.006*압박
성과금 + 사내복지 부족	0.007*수당 + 0.006*출근 + 0.006*사용 + 0.005*대표 + 0.005*관리자 + 0.004*교육 + 0.004*관리 + 0.004*진짜 + 0.004*주말 + 0.004*사원

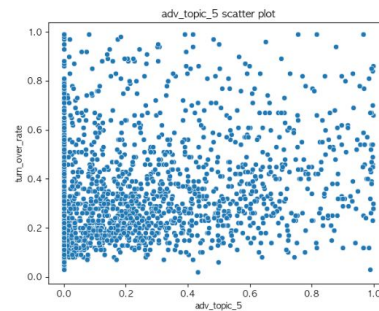
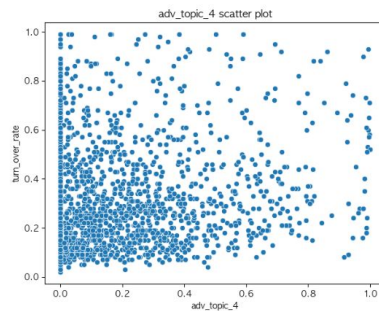
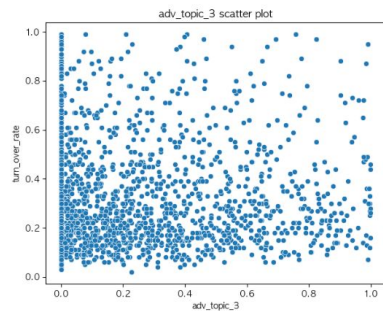
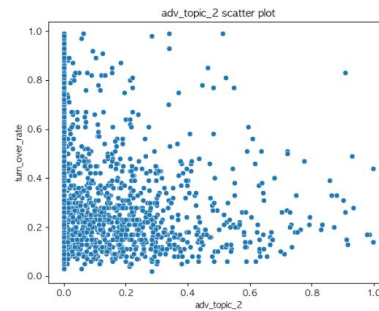
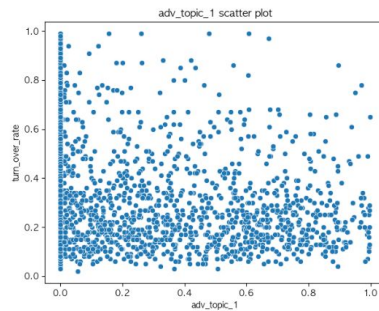
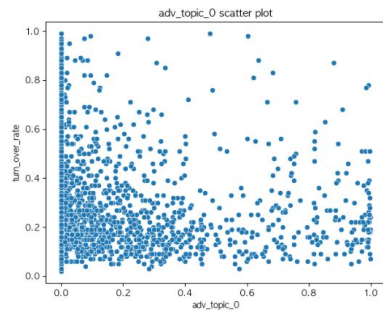
데이터 분포 형태 (이직률)



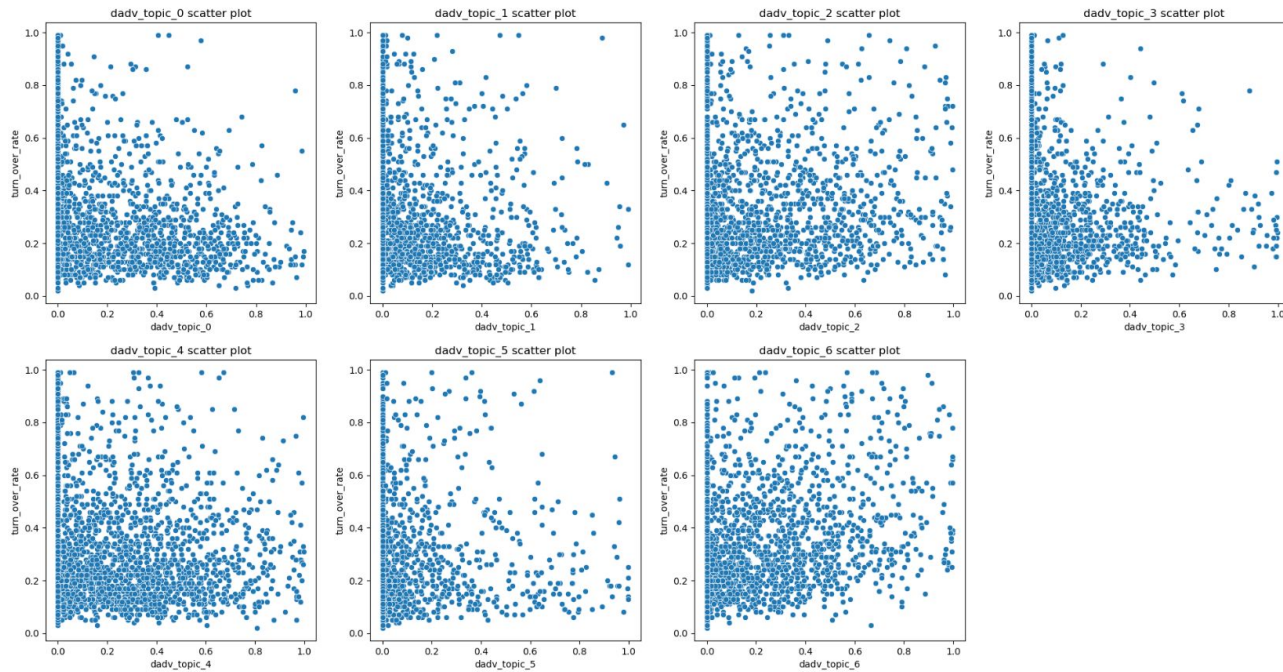
데이터 분포 형태



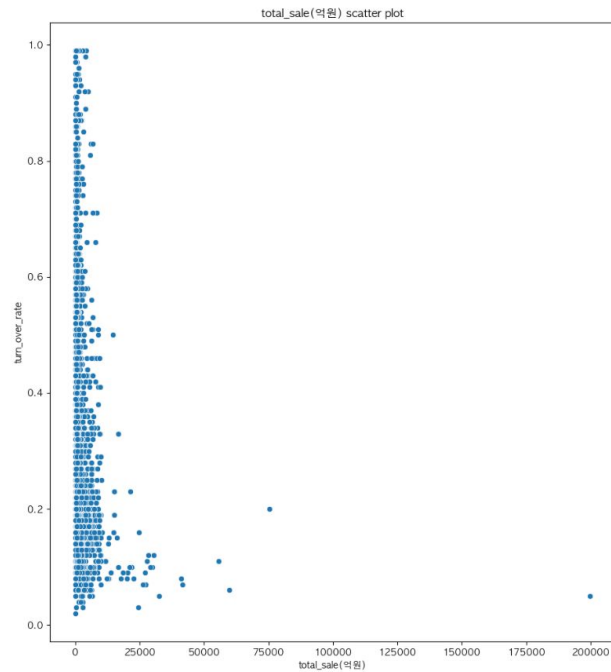
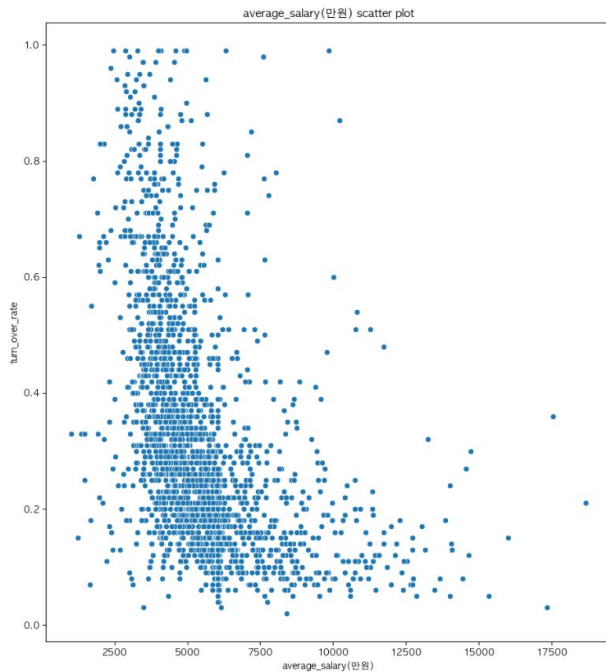
데이터 분포 형태 (토픽 모델링 장점 리뷰)



데이터 분포 형태 (토픽 모델링 단점 리뷰)



데이터 분포 형태 (재무제표)



데이터 분포 형태 (문제점)

거의 모든 피쳐 산점도에서
왜도가 있음을 발견

지나치게 왜곡된 피쳐의 존재는 회귀
예측 성능을 저하 할 수 있음



왜도가 1이상인 피쳐에서
모든 값에 0 다음으로 작은
값을 더한 후 log 변환

모든 피쳐에 0값이 존재하여 전체
피쳐에다가 0을 제외한 가장 작은 값을
더해준 후, log 변환

IV. 모델링

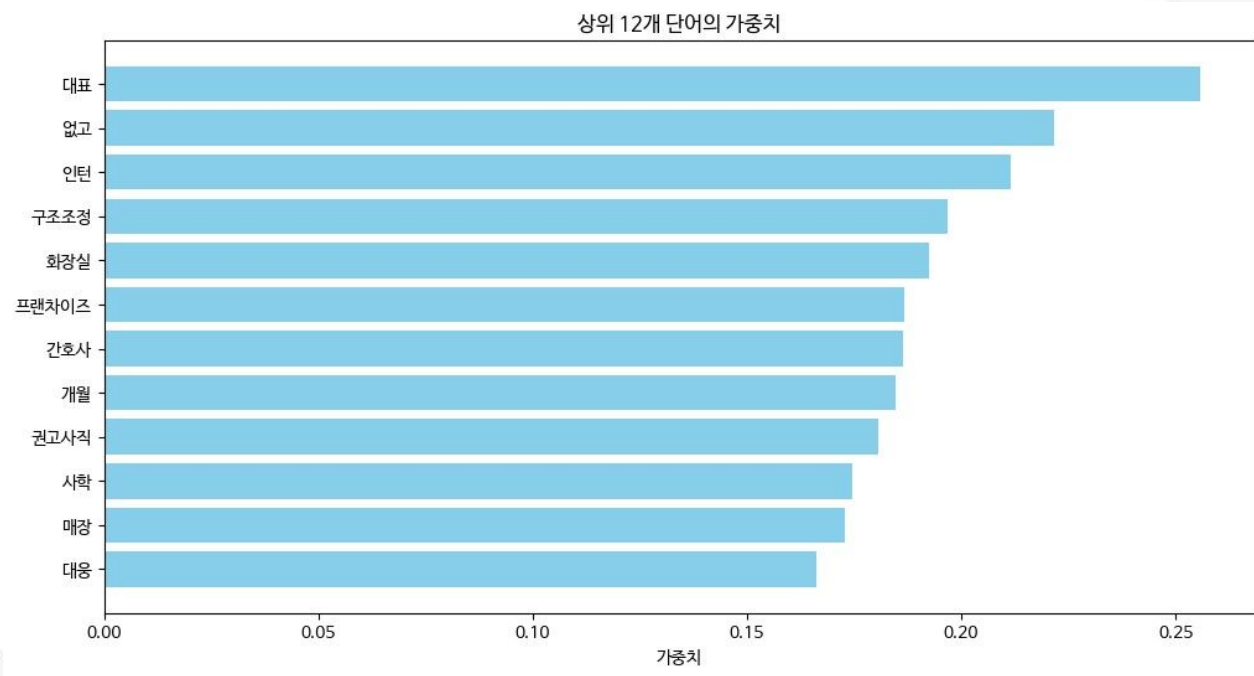
회귀 모델 사용

- Linear Regression
- Ridge Regression
- Lasso Regression
- XGB Regressor
- LGBM Regressor
- Random Forest Regressor
- Gradient Boosting Regressor
- Decision Tree Regressor

예측 / 성능 (TF-IDF)

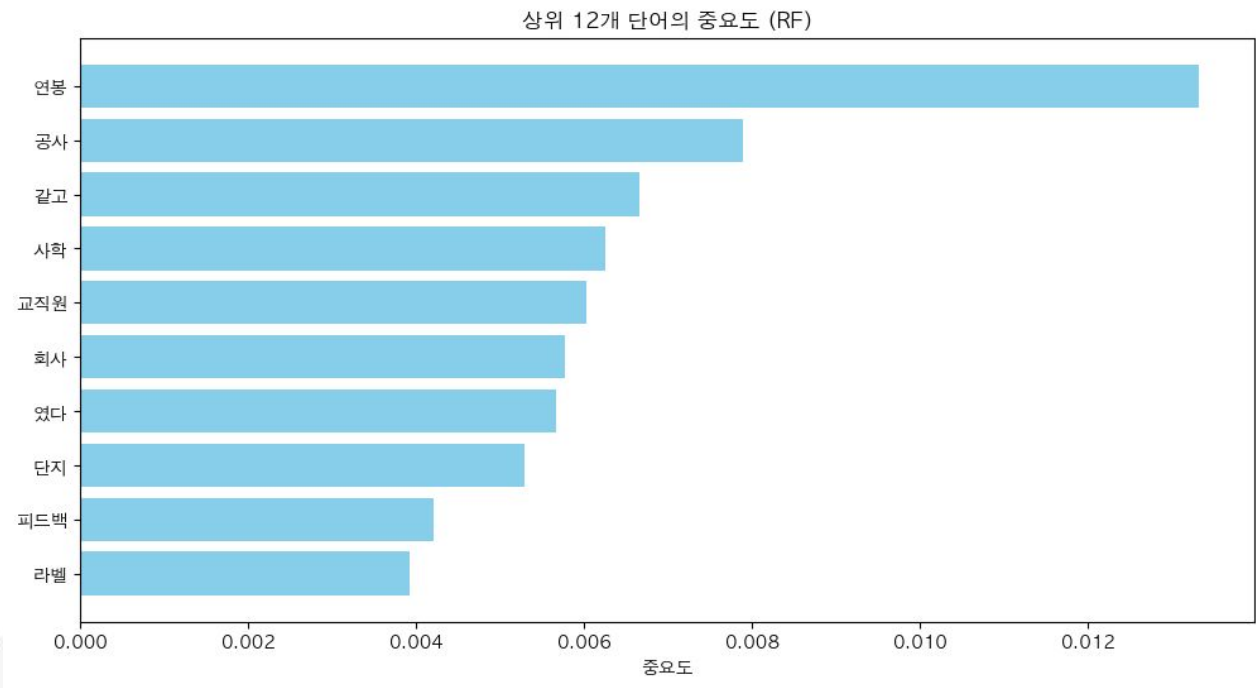
	MSE	RMSE	MAE
LR	0.026	0.161	0.121
Ridge	0.0241	0.1551	0.1130
Lasso	0.0316	0.1778	0.1354
XGB	0.0246	0.1567	0.1149
LGBM	NAN	NAN	NAN
GB	0.0301	0.1736	0.1227
RF	0.0236	0.153	0.1129
Decision Tree	0.0317	0.1781	0.1250

Ridge 모델의 피쳐 중요도 (TF-IDF)



ridge feature_importance

RF 모델의 피쳐 중요도 (TF-IDF)

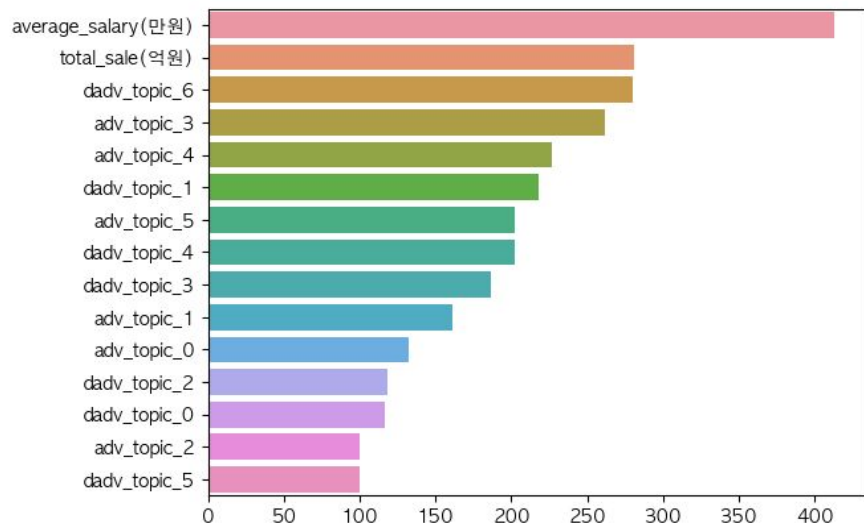


rf feature_importance

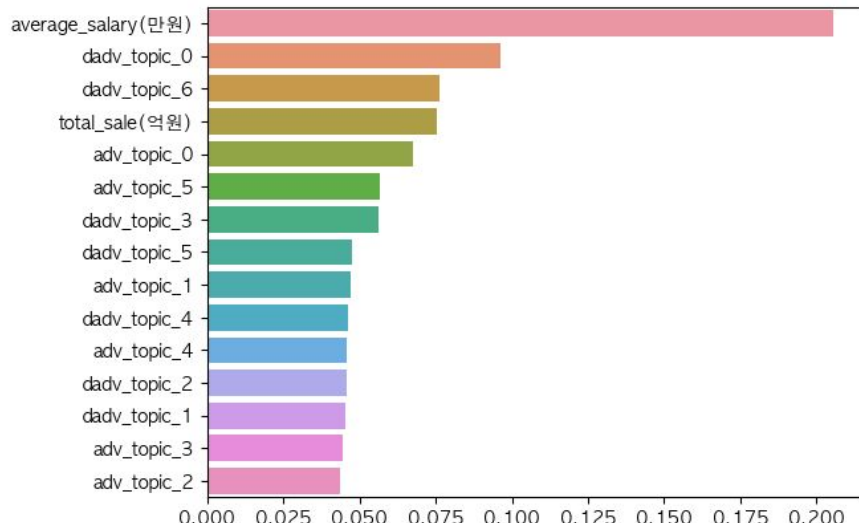
예측 / 성능 (Topic Modeling)

	MSE	RMSE	MAE
LR	0.0135	0.116	0.088
Ridge	0.0135	0.116	0.088
Lasso	0.0137	0.117	0.091
XGB	0.0112	0.106	0.081
LGBM	0.0108	0.104	0.079
GB	0.0121	0.110	0.085
RF	0.0121	0.110	0.086
Decision Tree	0.0121	0.110	0.084

LGBM / XGB 모델의 피쳐 중요도 (Topic Modeling)



Light GBM Feature coef.



XGB Feature coef.

모델 이용한 이직률 예측

	company_name	adv	dadv	average_salary	total_sale	turn_over_rate
0	현대카드	일단 오피스 환경이 쾌적한것은 장점대기업만의 성과급이나 복지의 혜택등이 좋다생각보다...	곧대문화 부서에따라 존재. 급신급신 매일 퇴근은 정시에 포기 할일 다하면 다른업무...	8737	2300.0	0.35



	company_name	adv_topic_0	adv_topic_1	adv_topic_2	adv_topic_3	adv_topic_4	adv_topic_5	dadv_topic_0	dadv_topic_1	dadv_topic_2	dadv_topic_3	dadv_topic_4	dadv_topic_5	dadv_topic_6	average_salary(만원)	total_sale(억원)
0	현대카드	0.682214	0.0	0.010139	0.0	0.0	0.0	0.0	0.071102	0.0	0.013451	0.0	0.648406	0.0	9.075437	7.741099



Predicted turnover rate of the company using lgbm model is : [0.23826028]
Real turnover of the company is : 0.35

Predicted turnover rate of the company using xgb model is : [0.2879109]
Real turnover of the company is : 0.35

For topic: adv_topic_0
The topic modeling value is greater than mean and 75% of data
For topic: adv_topic_2
The topic modeling value is greater than 50% of data
For topic: dadv_topic_1
The topic modeling value is greater than 50% of data
For topic: dadv_topic_3
The topic modeling value is greater than 50% of data
For topic: dadv_topic_5
The topic modeling value is greater than mean and 75% of data

V. 결론

결론 (분석)

- TF-IDF 후 피쳐 중요도는 모델 별로 상이한 모습을 보여줌
 - Ridge:
 - 회사의 **이직과 직접적으로 연관된 키워드**들이 높은 중요도를 가지고 있음
 - ex) 대표, 인턴, 구조조정, 프랜차이즈, 권고 사직
 - RF:
 - 연봉, 피드백과 같이 이직률과 관련된 단어들도 존재했으나 대체로 **미흡한 불용어 처리**로 인한 쓸모 없는 단어들의 영향이 많은 것으로 보여짐

결론 (분석)

- Topic Modeling 후 나온 지표에서는 **사원의 평균 연봉**과 **기업의 총 매출 지표**가 가장 중요한 요소임을 확인
 - **LGBM:**
 - 성과금 + 사내복지 부족 (단점)
 - 자율적 근무 환경 (장점)
 - 성과급 만족도 (장점)
 - 정규직으로의 전환 (단점)
 - **XGB**
 - 진급의 어려움 + 수직적 문화 (단점)
 - 성과금 + 사내복지 부족 (단점)
 - 직업 안정성 + 휴식 (장점)
 - 회사 시설 (장점)

결론 (활용방안 & 기대 효과)

- 기업 경영자:
 - 활용 방안: 결과값을 이용하여 인사 전략을 수립 가능
 - 기대 효과: 양질의 인력 영입, 기업 이미지 개선
- 취업 및 이직 희망 인원:
 - 활용 방안: 복지 및 근무 환경 등의 요소들을 비교 분석해 자신에게 맞는 회사에 지원
 - 기대 효과: 근무 만족도 개선, 장기적 관점에서의 이직률 감소

결론 (한계)

- 기업 별 리뷰 개수의 차이로 인한 단어들의 불균형
- 웹사이트 별 보유 데이터의 차이로 인해 줄어든 데이터 수

참고문헌

- 박상언 강주영.(2022).파이썬 텍스트 마이닝 완벽 가이드
- 전영근, 이종태, 이승욱, 이수현, 송승연.(2017).기업 리뷰 텍스트마이닝을 통한 기업 분석 시스템 개발.대한산업공학회 추계학술대회 논문집,(1),3138-3152.
- 최진욱, 신동원, 이한준.(2021).IT 기업 직원의 만족 및 불만족 요인에 따른 이직률 예측: 토픽모델링과 머신러닝을 활용하여.한국데이터정보과학회지,32(5),1035-1047.
- Zhao, Yue & Hryniewicki, Maciej & Cheng, Francesca & Fu, Boyang & Zhu, Xiaoyu. (2018). Employee Turnover Prediction with Machine Learning: A Reliable Approach. 10.1007/978-3-030-01057-7.
- Rohit Punnoose and Pankaj Ajit, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms" International Journal of Advanced Research in Artificial Intelligence(IJARAI), 5(9), 2016.
<http://dx.doi.org/10.14569/IJARAI.2016.050904>



감사합니다

프로젝트 코드 링크

https://github.com/Sue-HyeongLee/SCAICO_turnover_rate