회귀 프로젝트

여성인력 임금 예측하기

전진경, 최은비

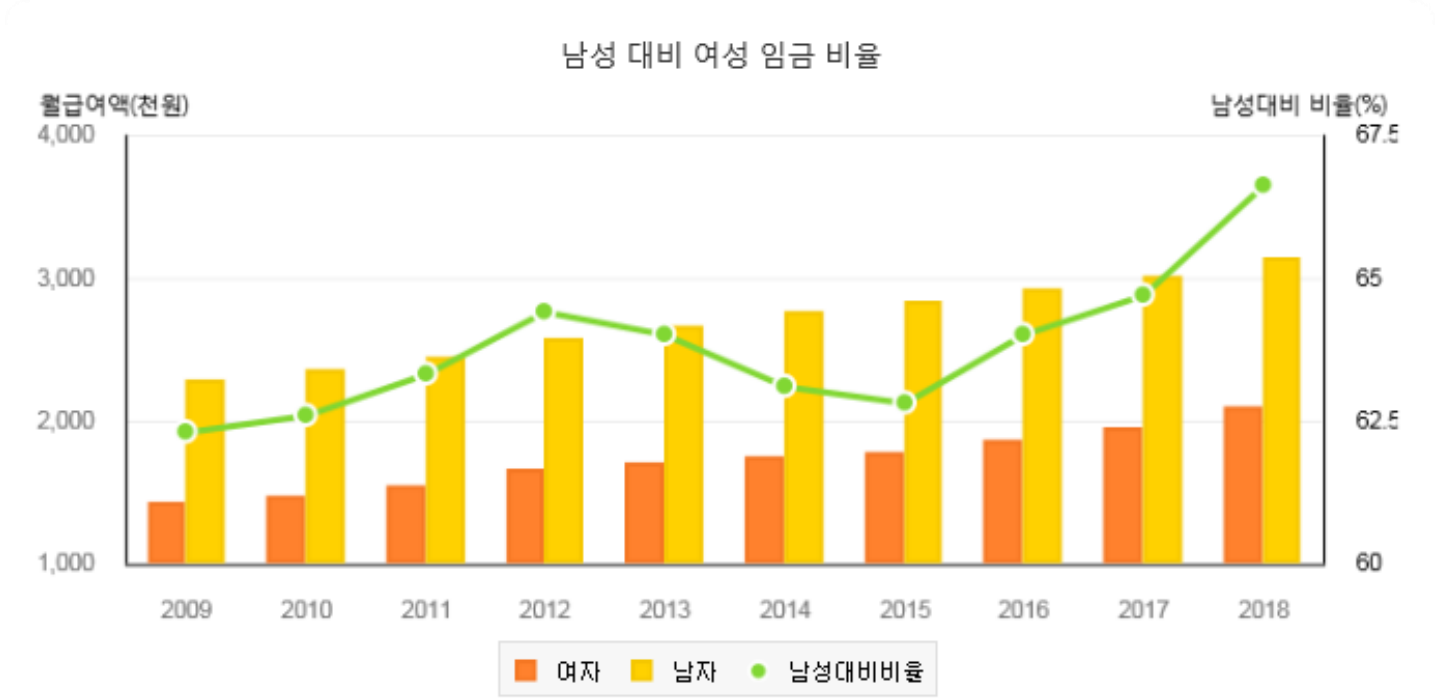
Contents

Ⅰ. 프로젝트 개요

1. 프로젝트 배경
2. 프로젝트 목적 및 개요

1. 프로젝트 배경

남성 근로자의 임금 대비 여성근로자의 임금 수준은 '16년 64.0%에서 '17년 64.7%로 0.7%p 증가하였지만, OECD 주요회원국 중 남녀임금격차가 가장 크며, 우리나라 여성은 남성보다 37% 정도 임금을 덜 받음.



연도	OECD 평균	호주	캐나다	덴마크	핀란드	프랑스	독일	일본	**한국**	영국	미국
'15	14.5	13.0	18.6	5.8	18.1	9.9	17.1	25.7	**37.2**	17.1	18.9

출처 : 고용노동부 고용형태별근로실태조사(1인 이상 기준)

2. 프로젝트 목적 및 개요

연령, 학력, 결혼여부, 고등학교 종류, 대학 전공, 대학 소재지 등 다양한 feature로 여성인력의 임금을 예측해보고자 함

조사된 데이터를 활용하여
여성인력의 임금 예측해보기

데이터 구하기

여성가족 패널조사

2006년부터 여성 패널에 대한 종단자료 구축을 위해
2년에 한 번 주기적으로 진행되는 조사

데이터 구성

조사 대상

전국 일반 가구 중 만19세 이상 만 64세 이하의
여성가구원이 있는 가구를 대상으로 추출된 9,068가구

가구용

여성개인용
(가족관계)

일자리용

참고) 활용 데이터 구성 상세

가구용

- 가구원 및 가족사항
- 주거상태
- 가구 소비
- 가구 소득
- 자산과 부채

여성개인용 (가족관계)

- 성장과정 및 학교생활
- 혼인 상태
- 혼인 상태에 따른 설문
- 첫 직장의 경험
- 첫 결혼 당시 직장경험
- 남편 일자리
- 임신 및 출산경험과 자녀
- 출산 당시 직장 경험
- 장애인 및 환자
- 개인의 특성/여가생활
- 가사노동
- 결혼과 부부생활
- 자녀 및 부모님과의 관계
- 형제와의 관계
- 가족 관련 가치관
- 건강과 여성·생활만족도
- 노후생활

일자리용

- 현재의 경제활동
- 임금근로자
- 비임금근로자
- 특수고용형태근로자
- 부가적 일자리
- 미취업 상태
- 이전 일자리
- 구직 경험
- 일 만족도
- 교육 및 훈련
- 사회 보험
- 직장생활과 가정생활
- 차별 사항
- 모성보호제도

Ⅰ. 프로젝트 세부내용

1. 분석 프로세스
2. 분석 내용
   1) 데이터 전처리
   2) 데이터 탐색
   3) 모델 fit 및 성능확인
   4) 예측모델 성능 향상

1. 분석 프로세스

다음 프로세스에 따라 분석을 진행함.

1) 데이터 전처리

결측치 확인 및 제거 & 대체

노후 데이터 제거

경력 데이터 삽입 (컬럼생성)

범주형 데이터 변환

2) 데이터 탐색

응답자 주요 특성 분석

인구통계학적 특성 및 응답의 분포 확인

변수간 관계 확인

임금과 관련된 다양한 변수와의 관계 확인

3) 모델 Fit & 성능확인

사용 모델

- OLS (Ordinary Least Squares)
- LinearRegression
- DecisionTreeRegressor
- RandomForestRegressor
- GradientBoostingRegressor
- XGBRegressor

4) 예측모델 성능 향상

변수 변경 및 조정

모델 성능 향상을 위해 독립변수 조합 변경

최적 모델 도출

GridsearchCV를 통한 최적모델 도출

1) 데이터 전처리

1. 결측치 확인 및 제거 & 대체
2. 노후데이터 제거
3. 경력데이터 삽입
4. 범주형 변수 변환

2. 분석 내용 _ 1) 데이터 전처리

원본 데이터에서 전처리를 위해 크게 결측치 확인 후 제거 및 대체, 노후 데이터 제거, 경력 데이터 생성 및 삽입, 범주형 데이터 변환의 네 가지 활동을 진행함.

1 **결측치 확인 및 제거 & 대체**	• 종속변수인 급여에서의 결측치는 행데이터 삭제 • 독립변수에서의 무응답 항목은 0으로 대체
2 **노후 데이터 제거**	• 5차 이전의 조사는 분석 대상에서 제외 (2012년 이전) • 첫 직장 입직 기준 1990년 이전 패널 데이터는 분석 대상에서 제외 • 응답자 연령 60세 초과 응답데이터 삭제
3 **경력 데이터 삽입**	• 임금예측에 가장 핵심적인 특징을 경력으로 설정하여, 경력 변수 생성 • 첫 직장 입직, 퇴직, 유지 및 다음 직장, 이전 직장 입퇴직 시기를 기준으로 경력 산출
4 **범주형 변수 변환**	• One-hot 인코딩 활용

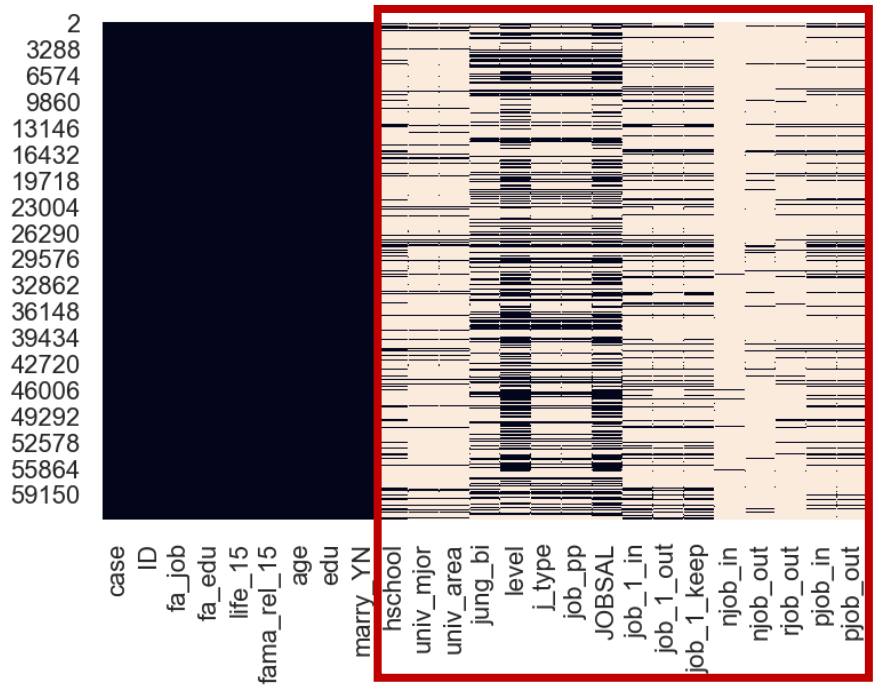
2. 분석 내용 _ 1) 데이터 전처리

전체 dataset에서 노후데이터를 제외하고 결측치를 처리하여 1245개의 최종 dataset을 도출함.

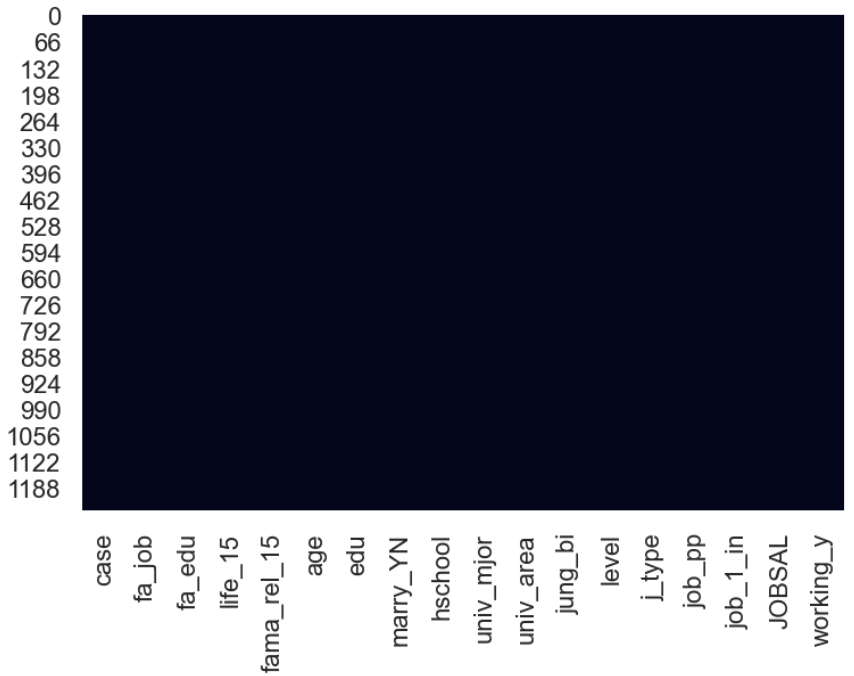
1) 데이터 전처리 | 2) 데이터 탐색 | 3) 모델 Fit & 성능확인 | 4) 예측모델 성능 향상

결측치 확인 및 제거 & 대체 | 노후 데이터 제거

결측치 처리 전



결측치 처리 후



2. 분석 내용 _ 1) 데이터 전처리

경력컬럼 추가를 위해 기존 보유정보를 최대한 활용하여 경력을 산출함.

1) 데이터 전처리 | 2) 데이터 탐색 | 3) 모델 Fit & 성능확인 | 4) 예측모델 성능 향상

경력데이터 삽입

보유 정보

첫 직장 입직 시기	첫 직장 퇴직 시기
새로운 직장 입직 시기	새로운 직장 퇴직 시기
이전 직장 입직 시기	이전 직장 퇴직 시기
첫 직장 유지 여부	

경력 산출 방법

1. 각 입직 시점이 없는 경우

→ drop : 계산 불가능

2. 첫직장 퇴직시점이 없는경우

case1) 첫직장 유지 O -〉 첫 직장 입직 ~ 마지막 조사 차수 시점으로 경력 계산

case2) 첫직장 유지 X -〉 첫 직장 입직 ~ 이전직장 퇴직 시점으로 계산

3. 이전직장 퇴직시점이 없는 경우

→ 첫 직장 입직에서 마지막 조사 차수 까지의 기간을 경력으로 계산

2. 분석 내용 _ 1) 데이터 전처리

무응답의 경우는 0으로 변경하고, 범주형 변수는 pandas의 pd.get_dummies를 이용하여 one-hot 인코딩 진행

1) 데이터 전처리 | 2) 데이터 탐색 | 3) 모델 Fit & 성능확인 | 4) 예측모델 성능 향상

범주형 변수 변환

범주형 변수

가정환경

- 아버지 직업
- 아버지 교육정도
- 가정 환경
- 어머니, 아버지 관계

개인 정보

- 교육 수준
- 결혼여부
- 고등학교 종류
- 대학전공
- 대학 지역

일자리

- 고용형태
- 직장내 지위
- 회사형태
- 회사인원

처리 방법

1. *무응답 -9, -8번*
   → 0으로 대체

2. *범주형 변수는 one-hot 인코딩 진행*
   → 109개 컬럼으로 분리

2) 데이터 탐색

1. 응답자 주요 특성 분석
2. 변수간 관계 분석

2. 분석 내용 _ 2) 데이터 탐색

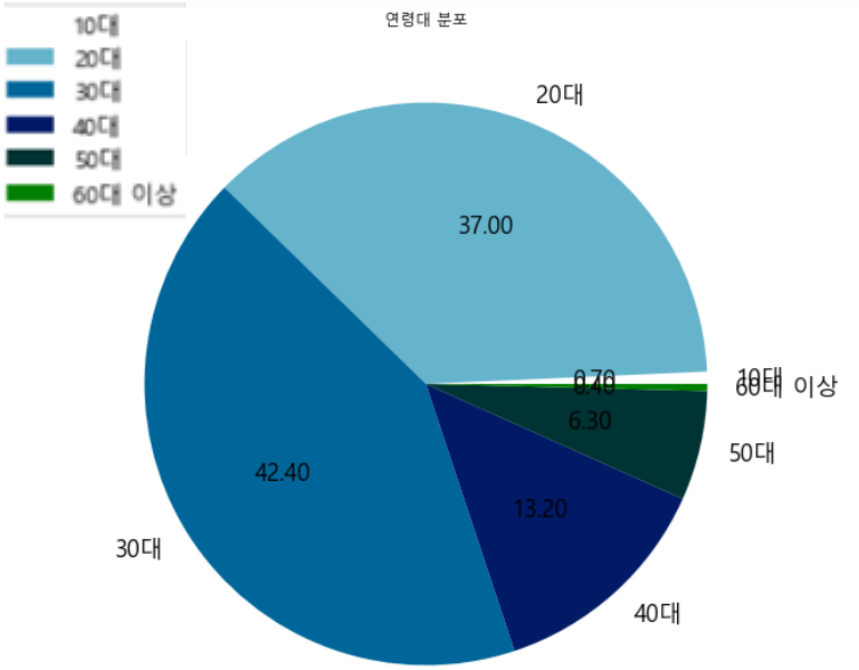
초기 62,426건의 데이터 중 전처리를 마친 1,877건의 데이터로 분석을 진행함.
연령에서 20-30대 쏠림 현상이 있으며, 혼인여부는 비교적 고르게 분포됨.

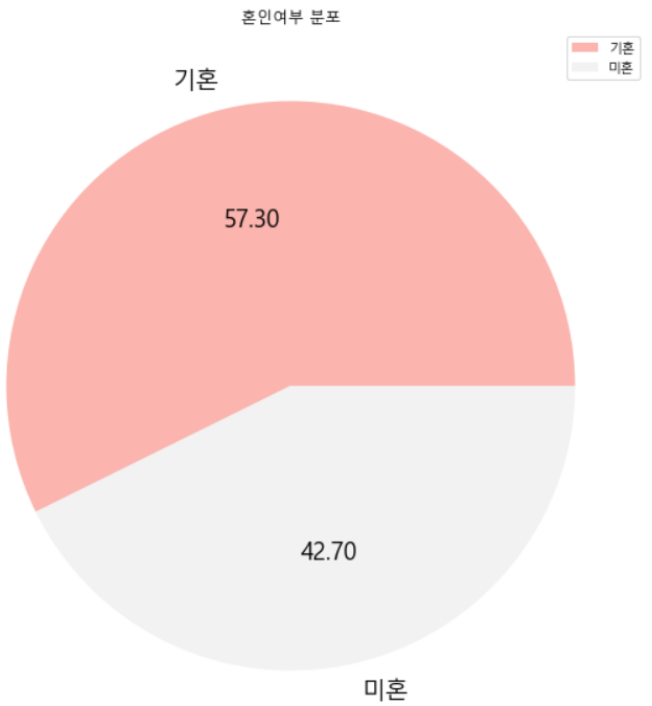
1) 데이터 전처리 | **2) 데이터 탐색** | 3) 모델 Fit & 성능확인 | 4) 예측모델 성능 향상

응답자 주요 특성 분석

분석데이터 요약

- 초기데이터: 62,426건
- 전처리 완료 후 데이터: 1,877건



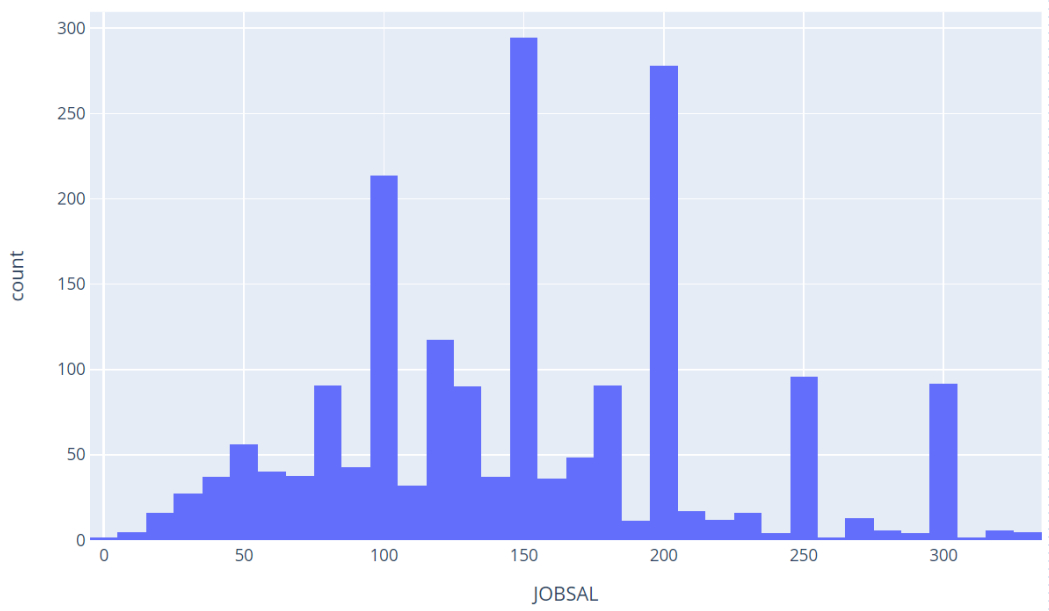


2. 분석 내용 _ 2) 데이터 탐색

응답자의 월급은 최저4만원부터 333만원 사이에 분포하였으며, 설문조사 특성상 50, 100, 150, 200 처럼 50단위 구간으로 응답이 집중되어 있음을 확인할 수 있음. 우러급의 평균값은 150만원이며, 중앙값과 같음.

응답자 주요 특성 분석

응답자 월급 분포




```
data_fin_1['JOBSAL'].describe()

count    1877.000000
mean      150.647576
std        67.649143
min         4.166667
25%       100.000000
50%       150.000000
75%       200.000000
max       333.000000
Name: JOBSAL, dtype: float64
```

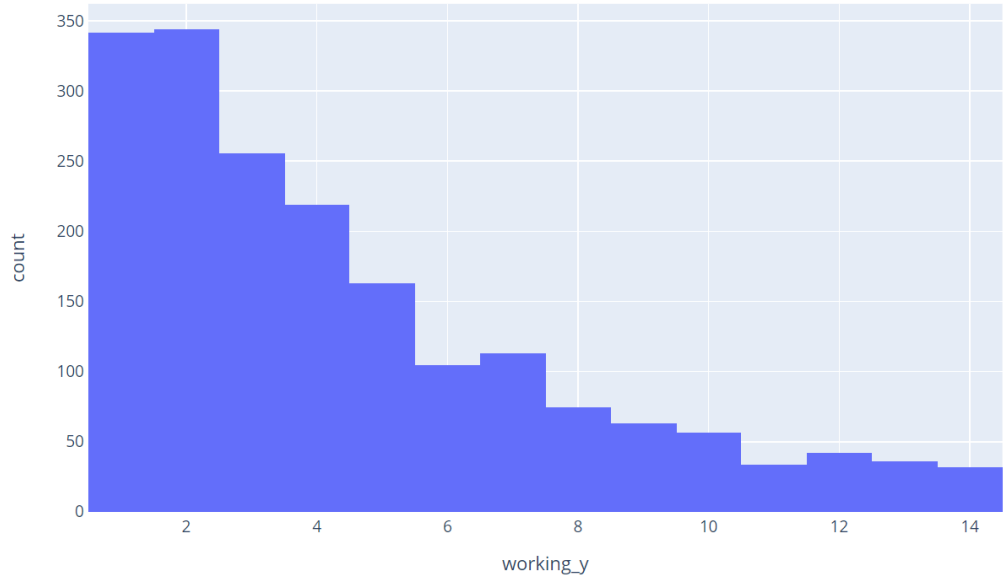
2. 분석 내용 _ 2) 데이터 탐색

응답자의 경력분포를 보면 최저 1년에서 최고 14년까지의 데이터로 분석을 진행하였고, 평균은 4년, 중앙값은 3년으로 분석됨.

1) 데이터 전처리 | **2) 데이터 탐색** | 3) 모델 Fit & 성능확인 | 4) 예측모델 성능 향상

응답자 주요 특성 분석

응답자 경력 분포




```
data_fin_1['working_y'].describe()
```

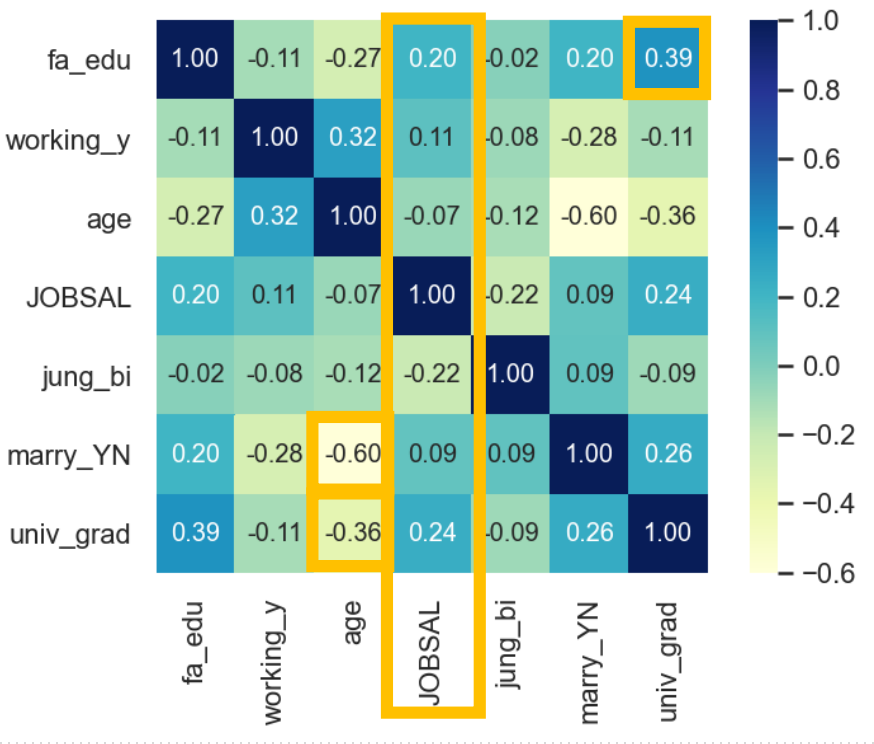
```
count    1877.000000
mean        4.483751
std         3.353924
min         1.000000
25%         2.000000
50%         3.000000
75%         6.000000
max        14.000000
Name: working_y, dtype: float64
```

2. 분석 내용 _ 2) 데이터 탐색

변수간 관계분석을 위해 급여와 개인적 특징의 상관을 살펴봄.

1) 데이터 전처리 | **2) 데이터 탐색** | 3) 모델 Fit & 성능확인 | 4) 예측모델 성능 향상

변수간 관계 분석



급여와 상관관계

- **급여**와 가장 상관이 높은 변수는 **대학 졸업 여부**(univ_grad, 0.24)
- 두 번째 상관이 높은 변수는 **아버지의 교육 수준**(fa_edu, 0.2)

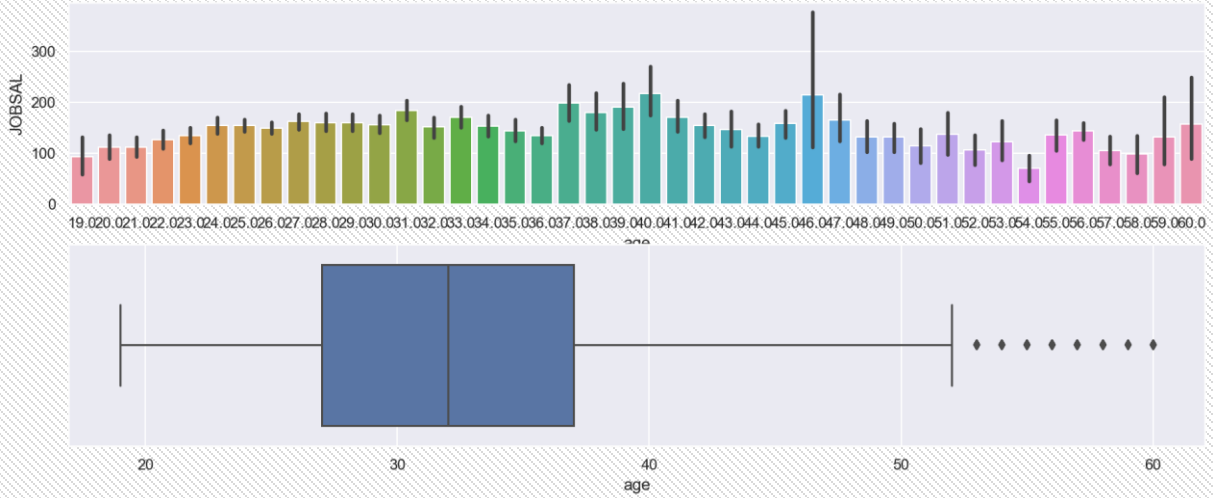
변수간 상관관계

- 나이와 결혼여부가 가장 높음 (−0.6)
- **딸의 대졸여부는 아버지의 학력과 양의 상관**을 보임 (0.39)
- 나이와 대졸여부는 음의 상관을 보임 (−0.36)

2. 분석 내용 _ 2) 데이터 탐색

연령에 따른 급여의 변화를 보기 위해 두 변수의 상관 관계를 살펴봄

변수간 관계 분석



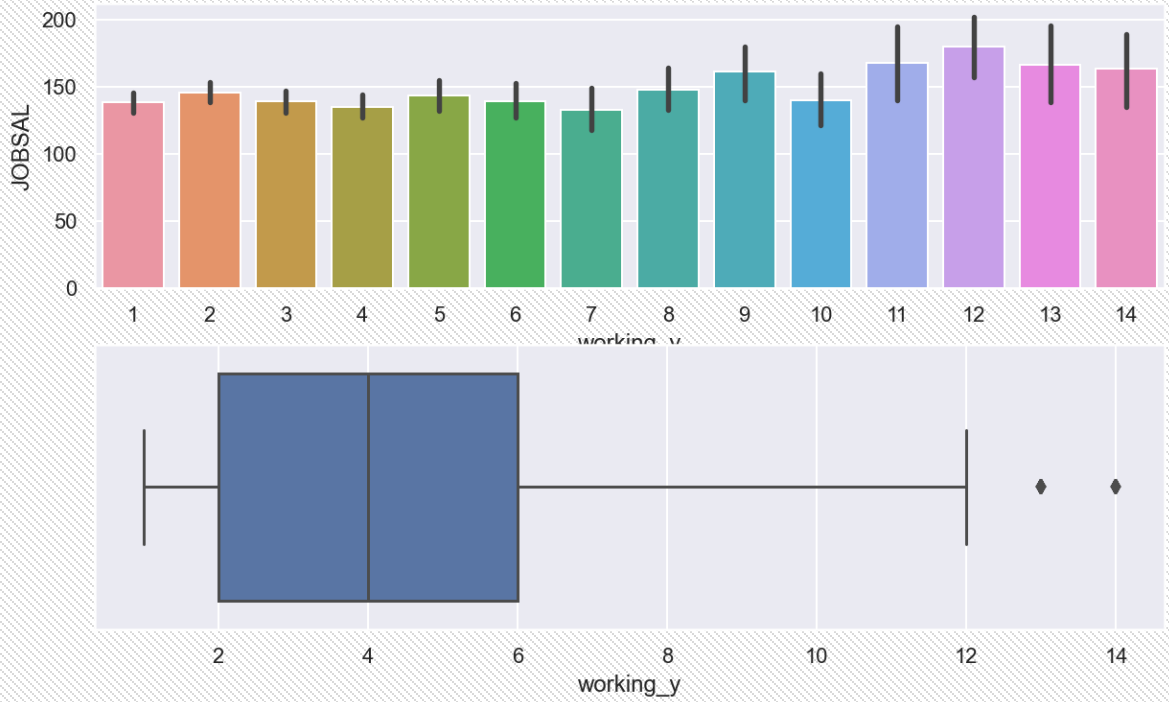
급여와 연령의 관계

- 여성의 임금은 30대까지는 연령 증가에 따라 증가하는 경향을 보이지만 30대에 들어서면서 정체 혹은 감소의 경향을 나타냄
- 30대 후반부터 급여 수준이 오르지만 40대로 진입하며 다시 감소추세를 나타냄

2. 분석 내용 _ 2) 데이터 탐색

경력에 따른 급여의 변화를 보기 위해 두 변수의 상관 관계를 살펴봄

변수간 관계 분석



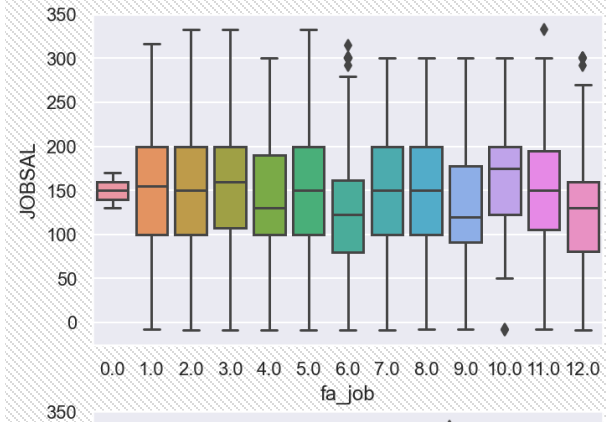
급여와 경력의 관계

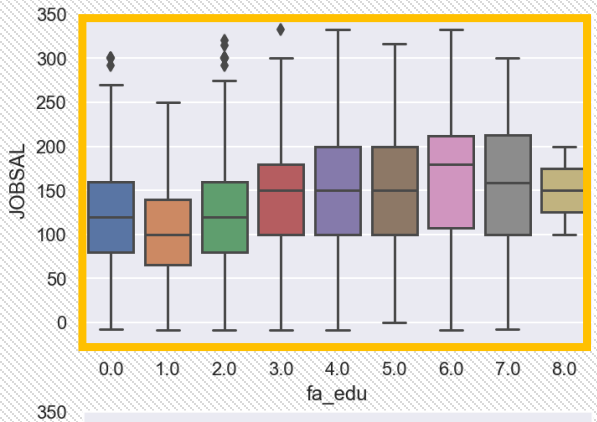
- 여성의 임금은 경력이 증가할수록 증가하는 경향을 보이지 않음
- 급여와 가장 강력한 선형 관계를 예측하였으나 그렇지 못함
- 경력변수 생성 과정에서 경력 단절이 반영되지 못한 것으로 추정
- 경력 단절 기간을 산입한다면 경력이 증가하더라도 급여가 증가하지 않을 수 있음

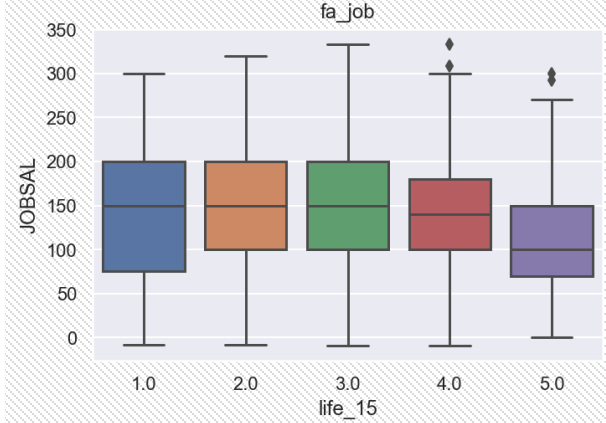
2. 분석 내용 _ 2) 데이터 탐색

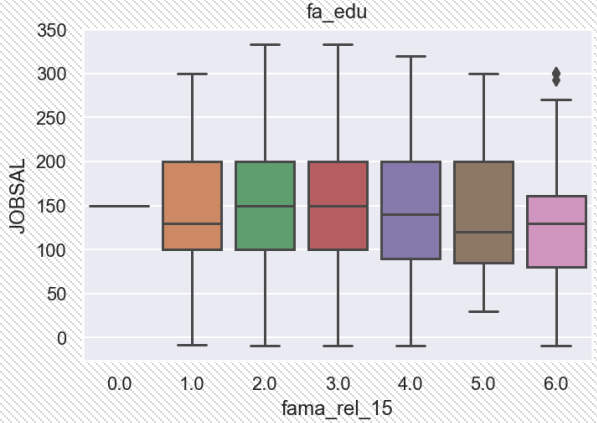
급여에 영향을 미치는 요인들을 확인하기 위해 각 변수와 급여를 비교함

변수간 관계 분석









급여와 다른 변수와의 관계

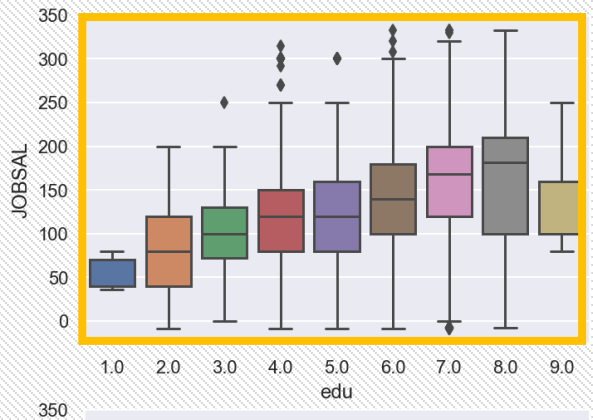
- 변수 간 가장 뚜렷한 차이를 보이는 것은 **아버지의 교육 수준과 딸의 급여**와의 관계
- 15세 무렵의 가정 형편, 부모님의 관계는 긍정적 응답에서는 급여와 관계가 적었지만 부정적 응답 시 급여에 부정적 영향을 주는 것을 확인
- 아버지의 직업에 따른 급여의 중앙값은 아버지가 군인일 때 (10번)가 가장 높았음

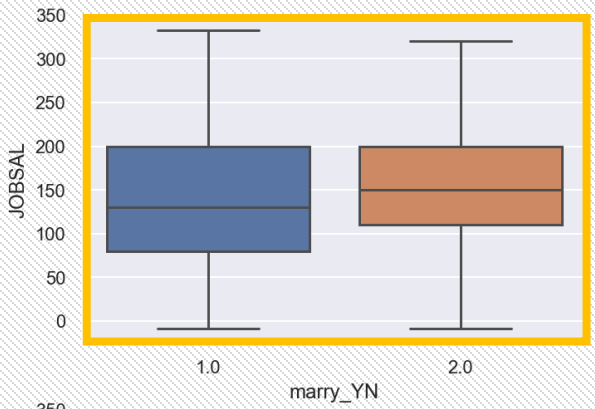
2. 분석 내용 _ 2) 데이터 탐색

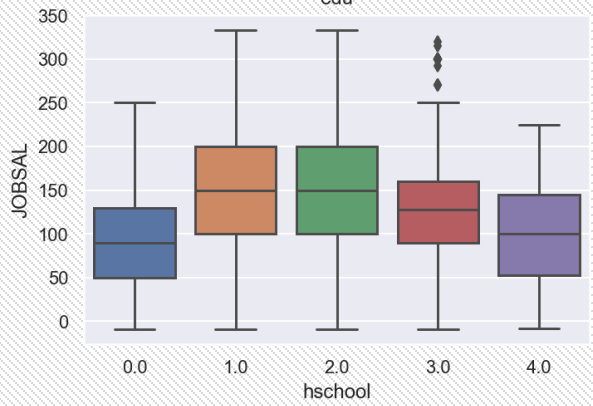
급여에 영향을 미치는 요인들을 확인하기 위해 각 변수와 급여를 비교함

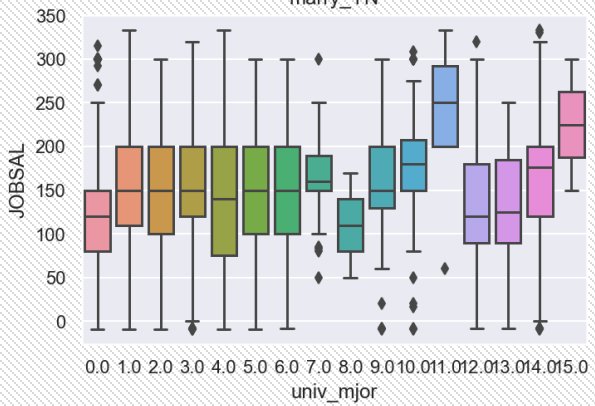
1) 데이터 전처리 | **2) 데이터 탐색** | 3) 모델 Fit & 성능확인 | 4) 예측모델 성능 향상

변수간 관계 분석









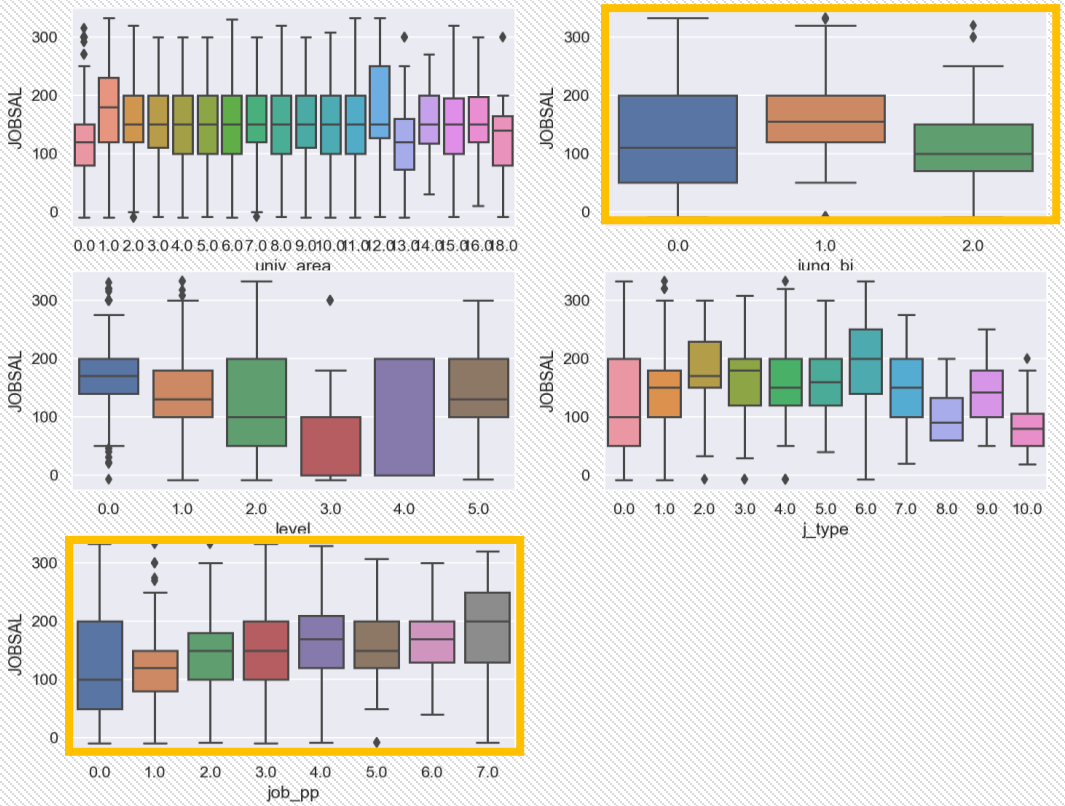
급여와 다른 변수와의 관계

- 급여와 가장 큰 상관을 보인 것은 본인의 교육 수준
  (6번 전문대, 7번 4년제(5-6년제 포함), 8번 대학원 석사, 9번 대학원 박사)
- 본 분석에서는 대학졸업 여부로 구분하여 진행 (초대졸 포함)
- 결혼여부는 1번이 유경험, 2번이 무경험. 중앙값은 미혼자가 조금 더 높음을 확인
- 대학전공과 급여에서 급격히 높아지는 전공은 (간호, 약학, 의학계열)

2. 분석 내용 _ 2) 데이터 탐색

급여에 영향을 미치는 요인들을 확인하기 위해 각 변수와 급여를 비교함

변수간 관계 분석



급여와 다른 변수와의 관계

- 대학교 소재지는 서울지역에서만 다소 높은 중간 값을 보임
- 고용형태가 정규직일 경우 비정규직보다 급여의 중간 값이 50만원정도 높음
- 회사의 규모(인원수)와 급여는 높은 상관을 보였지만, 예측하려는 변수로는 적합하지 않아 사용하지 않음

3) 모델 Fit & 성능확인

2. 분석 내용 _ 3) 모델 Fit & 성능확인

전처리 완료 후, 변수와 모델을 선정하고 fitting을 진행함.

1) 데이터 전처리 | 2) 데이터 탐색 | **3) 모델 Fit & 성능확인** | 4) 예측모델 성능 향상

최종 선정 변수

- 연령
- 경력
- 대학 소재지:서울
- 결혼여부: 미혼
- 교육수준: 4년제 대학졸업, 석/박사졸업
- 정규직 여부: 정규직

활용모델

- OLS
- LinearRegression()
- DecisionTreeRegressor(max_depth=3, random_state=13)))
- RandomForestRegressor(n_jobs=-1, n_estimators=100, max_depth=3)
- GradientBoostingRegressor()
- XGBRegressor(max_depth=3)

2. 분석 내용 _ 3) 모델 Fit & 성능확인

OLS분석결과 R^2값은 0.791로 모델의 설명력이 79%정도임을 확인할 수 있었음.

OLS Regression Results

Dep. Variable:	JOBSAL	R-squared (uncentered):	0.791
Model:	OLS	Adj. R-squared (uncentered):	0.791
Method:	Least Squares	F-statistic:	1064.
Date:	Wed, 26 Aug 2020	Prob (F-statistic):	0.00
Time:	18:27:19	Log-Likelihood:	-11275.
No. Observations:	1972	AIC:	2.256e+04
Df Residuals:	1965	BIC:	2.260e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	2.5858	0.095	27.286	0.000	2.400	2.772
working_y	3.0437	0.525	5.801	0.000	2.015	4.073
univ_area_1.0	13.0741	5.051	2.589	0.010	3.169	22.979
marry_YN_2.0	47.6072	3.105	15.334	0.000	41.518	53.696
edu_7.0	41.6965	3.593	11.605	0.000	34.650	48.743
edu_8.0	39.3204	9.264	4.244	0.000	21.152	57.489
edu_9.0	9.6672	26.216	0.369	0.712	-41.746	61.080

Omnibus:	5.357	Durbin-Watson:	1.931
Prob(Omnibus):	0.069	Jarque-Bera (JB):	6.038
Skew:	0.050	Prob(JB):	0.0488
Kurtosis:	3.252	Cond. No.	551.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

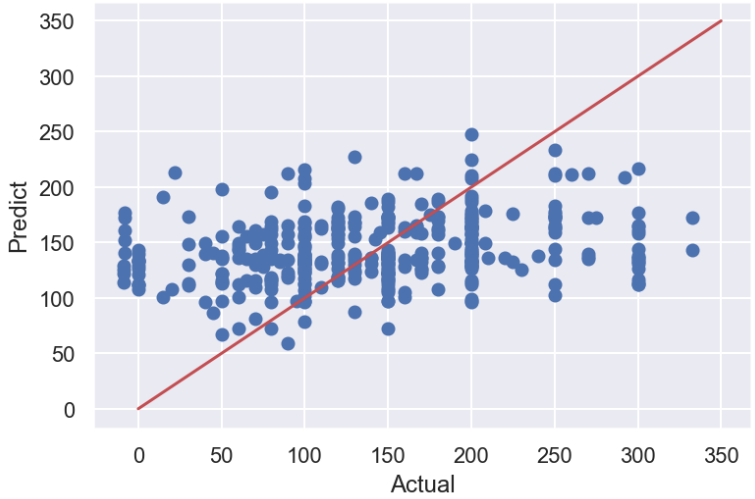
OLS분석 결과

- 분석 변수
  – 나이, 경력, 인서울 대학 여부, 미혼여부, 4년제 대학졸업, 석사, 박사 여부
- 모델의 R^2값: 0.791
- 변수별 유의수준을 보면 박사 졸업(edu_9.0)은 급여에 영향을 주지 못함

2. 분석 내용 _ 3) 모델 Fit & 성능확인

머신러닝 모델을 이용하여 동일 데이터를 분석하였을 때, R^2값은 0.09, rmse는 71만원 정도로 분석됨.
실제 값과 예측값의 차이를 확인하기 위해 그래프로 나타냄

	model_name	train_r2_score	test_r2_score	train_rmse	test_rmse
0	LinearRegression	0.08	0.09	70.72	71.68
1	DecisionTreeRegressor	0.11	0.05	69.52	73.25
2	RandomForestRegressor	0.13	0.06	68.90	72.63
3	GradientBoostingRegressor	0.21	0.05	65.28	73.09
4	XGBRegressor	0.20	0.06	65.87	72.96



Model Fit 결과

- 모델 분석 결과
  - Linear Regression가 가장 높은 R^2값과 가장 낮은 rmse를 보임
- 예측값과 실제값 비교 그래프
  - 기울기 1인 직선상에 있는 데이터가 별로 없음
  - 데이터의 퍼짐이 심함

4) 예측모델 성능 향상

2. 분석 내용 _ 4) 예측모델 성능 향상

모델의 개선을 위해 기존 최종학력 컬럼을 대학졸업 여부로만 두고, 변수를 조정하며 최적 모델을 도출함.
최종으로 선정된 변수는 연령, 경력, 미혼여부, 대졸여부, 정규직 여부임.

1) 데이터 전처리 | 2) 데이터 탐색 | 3) 모델 Fit & 성능확인 | **4) 예측모델 성능 향상**

변수 변경

(최종학력 컬럼을 대학졸업 컬럼으로 변경)
- 학력 변수 대학졸업 여부로만 구분

변수 조정

모델 성능 향상을 위해 독립변수 조합 변경

최적 모델 도출

GridsearchCV를 통한 최적모델 도출

최종 선정 변수

- 연령
- 경력
- 미혼여부
- 대졸 여부
- 정규직 여부

2. 분석 내용 _ 4) 예측모델 성능 향상

OLS분석 결과 R^2값은 이전보다 개선된 0.892로 나타났고, 모든 변수는 본 모델에서 유의한 것으로 확인됨.

OLS Regression Results

Dep. Variable:	JOBSAL	R-squared (uncentered):	0.892
Model:	OLS	Adj. R-squared (uncentered):	0.891
Method:	Least Squares	F-statistic:	1134.
Date:	Fri, 28 Aug 2020	Prob (F-statistic):	0.00
Time:	14:42:04	Log-Likelihood:	-3803.8
No. Observations:	690	AIC:	7618.
Df Residuals:	685	BIC:	7640.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	2.7194	0.130	20.840	0.000	2.463	2.976
working_y	2.1486	0.730	2.944	0.003	0.715	3.582
marry_YN_2.0	36.4140	4.864	7.487	0.000	26.864	45.964
jung_bi_1.0	46.9656	4.790	9.806	0.000	37.561	56.370
univ_grad	33.1661	4.900	6.768	0.000	23.545	42.787

Omnibus:	30.246	Durbin-Watson:	1.994
Prob(Omnibus):	0.000	Jarque-Bera (JB):	33.486
Skew:	0.496	Prob(JB):	5.35e-08
Kurtosis:	3.426	Cond. No.	93.4

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS분석 결과

- 분석 변수
  – 나이, 경력, 미혼여부, 정규직 여부, 대학졸업여부(전문대 포함)
- 모델의 R^2값: 0.892
- 모든 변수는 p값이 0.01 이하로 99% 수준에서 유의한 것으로 확인

2. 분석 내용 _ 4) 예측모델 성능 향상

MinMaxScaler를 적용하여 다시 분석한 결과 R^2값이 다소 떨어졌지만 개별 변수의 유의확률은 높아짐.

1) 데이터 전처리 | 2) 데이터 탐색 | 3) 모델 Fit & 성능확인 | **4) 예측모델 성능 향상**

OLS Regression Results

Dep. Variable:	JOBSAL	R-squared (uncentered):	0.876
Model:	OLS	Adj. R-squared (uncentered):	0.875
Method:	Least Squares	F-statistic:	970.9
Date:	Fri, 28 Aug 2020	Prob (F-statistic):	4.78e-308
Time:	14:53:43	Log-Likelihood:	-3851.1
No. Observations:	690	AIC:	7712.
Df Residuals:	685	BIC:	7735.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	168.2064	8.648	19.450	0.000	151.226	185.186
x2	39.0067	10.111	3.858	0.000	19.154	58.859
x3	58.9375	5.291	11.140	0.000	48.550	69.325
x4	55.9908	5.071	11.041	0.000	46.034	65.947
x5	46.9117	5.087	9.223	0.000	36.924	56.899

Omnibus:	29.138	Durbin-Watson:	1.953
Prob(Omnibus):	0.000	Jarque-Bera (JB):	32.854
Skew:	0.468	Prob(JB):	7.34e-08
Kurtosis:	3.517	Cond. No.	5.83

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS분석 결과 _ MinMaxScaler적용

- 더미변수는 별도의 스케일링 없이 사용
- R^2값은 0.876으로 scaler 적용 전보다 떨어짐
- 각 변수의 유의확률은 모두 0.01 이하로 99% 수준에서 유의한 것으로 확인

2. 분석 내용 _ 4) 예측모델 성능 향상

머신러닝 모델 분석 결과 RandomForest Regressor가 가장 좋은 성능을 나타냄. 실제값과 예측값 사이 53만원 정도의 오차가 예상됨.

	model_name	train_r2_score	test_r2_score	train_rmse	test_rmse
0	LinearRegression	0.13	0.22	59.18	53.42
1	DecisionTreeRegressor	0.22	0.18	56.11	54.85
2	RandomForestRegressor	0.25	0.23	55.23	53.01
3	GradientBoostingRegressor	0.42	0.12	48.44	56.73
4	XGBRegressor	0.51	0.02	44.45	59.87

```
best_model.best_estimator_
```

```
Pipeline(steps=[('clf', RandomForestRegressor(max_depth=5, random_state=13))])
```

Model Fit 결과

- 가장 성능이 좋은 모델은 RandomForest Regressor였으며, test의 R^2값이 0.23, rmse가 53.01로 분석됨
- 월급의 단위가 만원인 것을 고려할 때 예측값과 실제값 사이에 53만원 정도의 오차가 있을 수 있음을 의미함

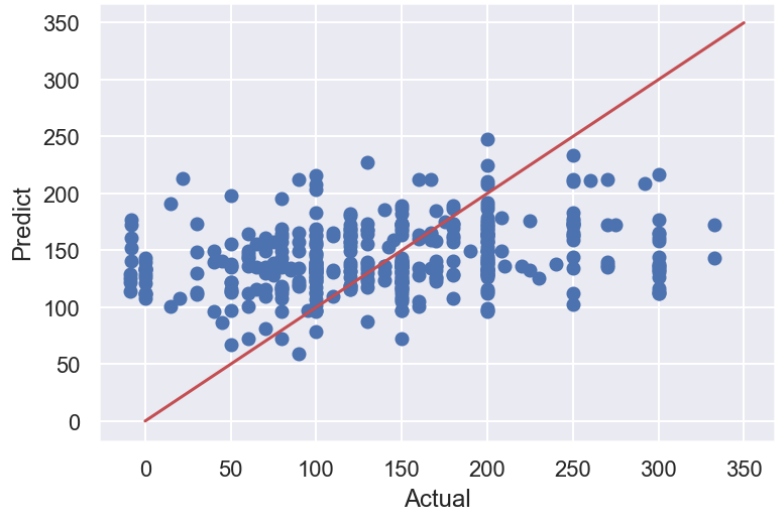
2. 분석 내용 _ 4) 예측모델 성능 향상

성능 향상 전, 후 예측값과 실제값의 비교 그래프 확인

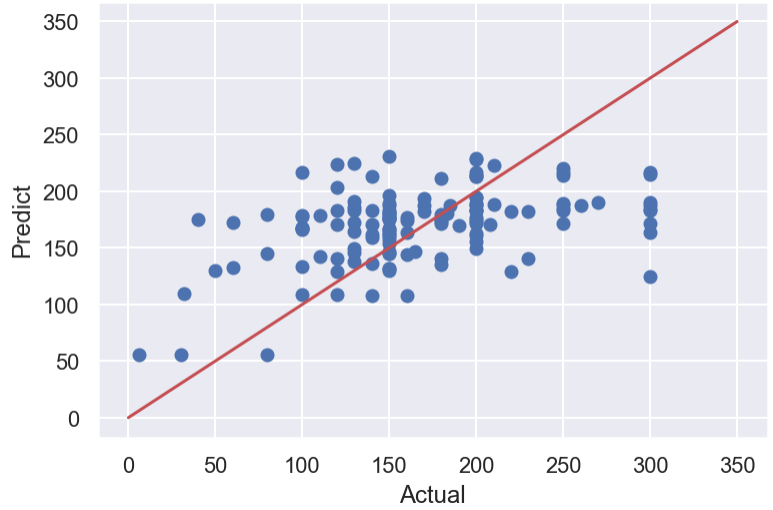
1) 데이터 전처리 | 2) 데이터 탐색 | 3) 모델 Fit & 성능확인 | 4) 예측모델 성능 향상

성능 향상 전후 비교

성능향상 전



성능향상 후



2. 분석 내용 _ 4) 예측모델 성능 향상

실제 월급 예측을 위해 가상의 인물을 선정하여 예측함.

실제 값 예측하기

- 나이 : 21세
- 경력 : 1년
- 결혼여부 : 미혼
- 정규직여부 : 비정규직
- 대졸여부 : 고졸

```
test_data = [[21, 1, 1, 0, 0]]
reg = RandomForestRegressor(max_depth=5, n_estimators=1000)
reg.fit(X_train, y_train)
reg.predict(test_data)
```

```
array([108.06521622])
```

예측임금 월 108만원

- 나이 : 30세
- 경력 : 5년
- 결혼여부 : 기혼
- 정규직여부 : 정규직
- 대졸여부 : 대졸

```
test_data = [[30, 5, 0, 1, 1]]
reg = RandomForestRegressor(max_depth=5, n_estimators=1000)
reg.fit(X_train, y_train)
reg.predict(test_data)
```

```
array([228.46811862])
```

예측임금 월 228만원

프로젝트를 마치며

본 분석은 다음의 한계점을 가짐

한계점

- 여성 패널데이터 활용. 패널데이터를 ***모집단을 대표하는 데이터로 보기는 어려움***
  - 전국 일반가구 중 만19세 이상 만64세 이하의 여성가구원이 있는 가구를 대상으로 추출된 9,068가구 (도서지역 제외, 제주도 포함)
  - 추출된 가구에 거주하는 만19세 이상 만64세 이하의 여성 9,997명이 원표본임.
- 초기부터 급여 예측을 목적으로 수집된 데이터가 아니기 때문에 ***변수 선정에 제한***이 있었음
- 설문을 통해 수집된 데이터로 ***수집 단계에서 오염***이 발생할 수 있음
  (급여 질문에서 50단위 응답이 많음을 확인)
- ***결측치*** 발생률이 높음 (초기 데이터 6만여 건 → 최종 데이터 2천건 이하)
- 경력 변수 산정에서 ***여성 인력의 경력 공백이 제대로 반영되지 못 함***

프로젝트를 마치며

분석 과정에서 느낀점


깜식이 누나와 별이 언니의 소감

- 다른 목적으로 수집된 데이터를 내 연구의 목적에 맞게 활용하는 일은 쉽지 않다
- 모델을 돌리는 일은 전처리, 다시 전처리, 또 전처리의 연속이다 (전처리 파티..)
- 기존 변수에서 새로운 변수를 도출하는 일은 매우 조심스러워야 하는 일이다
- 이상치 처리는 신중해야 한다 (우리는 350만원 이상 월급여를 받는 사람을 이상한 사람으로 취급했다) 성능은 좋아졌지만 ***우리와 같은 고급인력***의 임금 예측은 어려워졌다
- 작업파일은 분명히 잘 관리되어야 한다

Q&A

감사합니다 :-)