안녕하십니까~3조 전진경, 최은비 입니다.

회귀프로젝트 **V**여성인력 임금 예측 발표를 **V**시작하겠습니다. //

프로젝트 개요, 세부내용, 개선사항 순서로 **V**발표할 예정입니다.//

첫번째 프로젝트 개요입니다. //

저희 프로젝트를 시작하게 된 배경으로

남성근로자 대비 여성근로자의 임금은 **V**16년 17년을 비교했을 때 0.7%p 소폭 증가하였지만,

OECD 주요 회원국 중 **V**남녀임금격차가 가장 크고 **V**15년도 기준 남성대비 여성이 **V**약 37%정도 임금을 덜 받고 있었습니다.(빠르게)//

이에 저희는 **V**여성 임금에 관해 관심이 생겼고 이에 영향을 미치는 **V**여러 요인들이 무엇일까 라는 생각에서 이 프로젝트를 시작하게 되었습니다.\_

저희 목적은 연령, 학력, 결혼여부 등 다양한 feature들로 **V**여성인력 임금을 예측하는 것으로

데이터의 출처는 **V**여성가족 패널조사 사이트이며

조사되는 카테고리는 **V**크게 가구용, 여성개인용, 일자리용이 있습니다.//

화면은 상세 항목들이며 **V**임금에 영향을 주는 항목들을 선별하여 데이터를 얻었습니다. //

세부내용입니다. //

화면의 순서로 분석을 진행했습니다.

ˇ첫번째 데이터 전처리 부분에서 **V**결측치 확인 그리고 대체 및 제거, **V**노후데이터 제거, **V**경력컬럼 삽입, **V**범주형 데이터 변환.

두번째로 전처리된 데이터를 **V**인구통계학적 특성 및 응답 분포를 확인하였고 **V**각 변수간의 관계를 확인하였습니다.

사용한 모델들은 **V**하기와 같으며 회귀 관련된 **V**여러 모델들을 사용하였습니다.

다음으로 모델의 성능을 향상시키기 위해 **V**변수들의 조합을 변경하고 **V**gridsearch를 활용하여 최적의 옵션을 찾았습니다.//

다음은 데이터 전처리 관련 **V**세부내용입니다.//

결측치 관련하여 **V**종속변수인 급여에서의 결측치는 **V**데이터 활용이 어려워 삭제하였으며

다른 변수에서의 무응답은 0으로 대체하였습니다.

사용한 패널데이터는 **V**2006년도부터 2년간격으로 조사된 데이터로 **V**2012년도 이전 데이터는 제외하였으며, **V**첫직장 입직 기준 1990년도 이전데이터 **V**그리고 응답자 연령도 60세가 초과할 경우도 **V**노후데이터로 판단하여 제외하였습니다.

(참고첫입직이 1970~80년도, 첫 조사패널이 60세이상인 경우도 있어서 제외함)//

그리고 임금 예측에 있어서 **V**경력은 중요한 변수라 판단하여 **V**직장의 입직,퇴직 시기를 활용하여 경력을 산출하였습니다. 관련 세부기준은 이후 화면에서 설명드리겠습니다.

마지막으로 범주형 데이터는 one-hot인코딩을 활용하여 나눴습니다.//

화면을 보시면 왼쪽이 결측치 처리전, 오른쪽이 처리 후 입니다.(컬럼들이 사라진 이유는 경력산출용 컬럼이여서 산출 후 삭제함)//

경력데이터 산출 기준입니다. 얻을 수 있는 정보로 첫직장 입직/퇴직, 새로운직장 입직/퇴직, 이전직장 입직/퇴직, 그리고 첫직장 유지여부 였씁니다.

3가지 임의의 기준을 적용하여 경력을 산출하였습니다. 첫째 각 입직 시점이 없는 경우 삭제 : 이는 경력이 임금에 중요변수라 생각되어 내린 기준이었습니다. 둘째 첫직장 퇴직 시점이 없는 경우 ; 첫직장 유지 여부를 확인하여 유지 할 경우 마지막 조사차수를 년도로 환산하여 계산하였고 첫직장 유지를 하지않았을 경우 이전직장 퇴직시점으로 계산하였습니다. //

그리고 왼쪽의 다양한 범주형 변수들을 무응답의 경우 0으로 대체하였으며 one-hot인코딩으로 109개의 컬럼으로 분리하였습니다. //

데이터 탐색 부분입니다.//

연령대를 보시면 20대~30대 80%가 넘는 비율을 차지하였으며 혼인여부를 보시면 기혼 약 57%, 미혼 약 43%로 확인되었습니다. //(원그래프로 대체하는 것이 어떨까요?)

**우리 데이터에 있는 사람들의 특징들 소개 느낌**

(**합치기**)

앞의 연령대데이터를 그림 그래프이며 30대가 가장 높고 그 다음 20대, 40대 순으로 나타납니다. 최소연령 19세에서 최대 60세까지 포함하며 평균 33세입니다.//

다음은 종속변수인 월급의 분포도이며 최저 4만원에서 최대 333만원으로 확인되며 설문조사 특성상 100, 150, 200, 250 과 같은 대략적인 수치로 응답이 집중 되어있었습니다. //

다음은 경력데이터 분포로 최소 1년 최대 14년으로 그림과 같은 분포를 보였습니다.//

다음은 컬럼별 상관관계를 히트맵으로 표현하였습니다. 급여와의 관계를 먼저 보시면 상관계수가 가장 높은 변수는 대학졸업여부였습니다. 이는 학력에서 분리된 변수로 가장 높은 상관계수인 변수를 활용하였습니다. 다음 두번째로 높은 변수는 아버지의 학력이었습니다.

추가적으로 변수간 상관관계를 확인하였더니 컬럼 중 가장 높은 상관계수를 가진 것이 -6으로 나이와 결혼여부로 나이가 많을수록 결혼이 결혼을 많이 함... 이런.. 그리고 딸의 대졸 여부는 아버지의 학력와 0.39로 높은 상관관계를 보였습니다. // (기혼이 1 미혼이 2)

다음 그림은 경력과 임금의 관계를 표현한 것으로 저희가 중요한 변수로 생각한 경력인 것에 반해 경력이 높을수록 임금이 증가하는 경향은 보이지 않았습니다. 저의 자체적으로는 경력변수 산출과정에서 여러가지 경력 단절이나 기타 변동성을 반영하지 못했기 떄문이라고 추청하였습니다. 후다닥

이제 저희가 선정한 컬럼별 종속변수와의 관계를 분석한 표들 입니다.

4개의 그림은 15세 무렵 아버지의 직업, 아버지의 학력, 생활정도, 부모님과의 관계에 관한 표이며 변수간 변화가 보이는 것은 앞서 상관관계에서도 3번째로 높았던 아버지의 학력입니다.

그 외 그림들은 급여와의 관계는 적었으나 부정응답의 추가설명~ 예= 뒤쪽으로 갈수록 안좋은 부정응답을 언급할 것 경우 임금에도 부정적영향을 주는것으로 확인했습니다. //

(컬럼 종류들 15세아버지직업-1관리자 2전문직 3사무직 4서비스직 5판매직 6농림어업 7기능직 8기계조작조립직 9단순업무 10군인 11무직 12아버지x,

아버지교육-1무학 2초등 3중등 4고등 5전문대 6대학(4~5,6년) 7석사 8박사

생활정도-1아주잘삼 2대체로잘삼 3보통 4대체로어려움 5아주어려움

부모님관계-1아주좋 2대체로좋 3보통 4별로안좋 5매우안좋)

다음은 본인의 학력, 결혼유무, 고등종류, 대학전공이며 본인의 학력이 가장 관계가 뚜렷함이 보입니다. 숫자가 높아질수록 고학력자이며 실제 분석에서는 대졸여부로 구분하여 사용하였습니다. 그 외 그림들을 보시면 미혼자가 중앙값이 기혼자보다 조금 높고, 대학전공에서 급여가 높은 전공은 의학계열이었습니다. //

(컬럼 종류들 학력 -1무학 2초등 3중학 4전문고등 5고등 6전문대 7대학 8석사 9박사

결혼유무-1무학 2초등

고등종류-1문과 2이과 3전문계 4예체능계

대학전공-1어문계(언어학) 2인문계(사회,철학) 3사회계(법학,경제) 4이학계(수학,화학) 5공학계(전자,재료) 6가정계(식품영양,가정) 7농림계(농학,원예,축산) 8수산해양계(해양,양식) 9의학계 10간호계 11약학계 12예술계 13체육계 14사범계(교육학) 15기타)

다음은 대학소재지, 고용형태(정규직/비정규직), 종사상지위, 직장종류, 직장직원수 입니다.

눈에 띄는 변수로는 정규직/비정규직 구분과 직장직원수 였으며

정규직일경우 중앙값을 비교했을 때 약 50만원정도 차이가 있으며

직장의 인원수가 증가할수록 금여가 높아짐을 확인했으나 저희가 예측하려는 변수로는 적합하지않아 적용하진 않았습니다. **임금예측인데 회사의 인원 임금증가는 당연. 입직전? -> 회사크기상관없이 예측이 가능해야함. 컬럼 자체가 너무 큰 영향.. <- 왜 적합하지 않은지 정하기**

(컬럼 종류들 대학소재지 -

고용형태-1무학 2초등

종사상지위-1고용된 일자리 2내사업 3가족,친척 주18시간이상 돈x 4가족,친척 주18시간미만 돈x 5보험모집, 화물차량기사, 학습지교사, 골프장보조, 가전a/s 등… 명확하지않네…ㅡㅡ

대학전공-1어문계(언어학) 2인문계(사회,철학) 3사회계(법학,경제) 4이학계(수학,화학) 5공학계(전자,재료) 6가정계(식품영양,가정) 7농림계(농학,원예,축산) 8수산해양계(해양,양식) 9의학계 10간호계 11약학계 12예술계 13체육계 14사범계(교육학) 15기타)

다음은 모델 성능 확인입니다.

앞서 보여드렸던 종속변수와의 관계들을 고려하여 총 6가지의 변수를 선정하였으며 아래의 모델들을 사용했고 저희가 사용한 컬럼들은 범주형 변수들이 많아 선형회귀모델 외 다른 모델들도 활용하였습니다. **해당 옵션들은.. 임의로 한거일까요?? 아님 그리드서치로 최적인애들을 적용한거일까용?? (진경궁금쓰)**

연령에 대한 비교 -> **언니 추가해주신다고 함**

,그리고 상관계수가 높았던 아버지 직업은 급 사라짐 -> **간결하게 정리하기**

**이미 대학졸업 여부가 독립변수로 설정이 되었기 때문에 아버지 학력을 변수로 넣는게 의미가 없다. 중복이다. 예측하려는 본인의 학력이 더 중요한 변수이기 때문에.** **변수끼리의 연관성이 클떄 전부 넣으면 영향을 줌..::**

OLS 분석결과 이며 ols의 R값은 조건들을 변경할 때 좋고 나쁨의 의미로 해석하였고 변수들이 종속변수에 미치는 영향 위주로 확인 하였습니다. **(찾아 봤더니 train 만 취급한 R값.찾은 내용 언급하기!! 정확한 이론 부탁드림-강사님에게 은근슬쩍.)** 그림을 보시면 저희가 선정한 변수들 중 마지막 변수인 박사졸업여부는 급여에 영향을 주지않아 학력에 대한 구분을 다시 정하기로 했습니다.

동일 조건으로 모델을 학습시킨 결과 Linear Reg에서 가장 나은 결과를 얻었고 R값이 0.09, rmse는 약 72가 나왔습니다. R값이 낮아 실제와 예측값을 그래프로 그려봤으나 데이터의 퍼짐이 심했습니다.

다음은 위에서 얻은 모델들의 성능향상을 위한 변경사항들입니다.

앞서 말씀드린대로 학력컬럼을 대졸여부로 변경하였고 변수들의 조합들을 변경하여 아래릐 5가지 변수로 정리했으며 그후 최적의 모델을 찾기위해 gridsearch를 활용했습니다.

앞의 ols와 비교하여 r값이 0.89로 상승한것으로 보아 변경한 변수들이 더 나은 결과를 보인다고 생각되며 선정한 변수들은 모두 유의미한 영향을 주는 것으로 확인됐습니다.

R값이 떨어지지만 연령과 경력의 수치적 차이로 MinMax스케일을 적용한 데이터로 진행하였습니다. OLS결과를 보면 각 컬럼들이 전부 p값이 0.000으로 더 유의미하게 바뀐걸 확인했습니다.

모델 결과를 보시면 성능이 좋은 모델은 RandomForestReg로 변경되었고 R값은 0.23, rmse는 53으로 많이 향상되었습니다. 종속변수의 단위가 만원임을 고려한다면 53만원정도의 오차가 예상됩니다.

실제 예측 그래프를 비교해서 보시면 희미하게 선형적인 모습을 보일려고 하며

**(성능향상 후의 그래프 축의 범위가 다름, 옆데이터와 다른 데이터가 향상 후에 보임 – 질문 들어오면 대답하는것으로.)**

마무리를 어떻게 하는게 좋을찌… 흠… 모델성능은 높혔으나 실제 예측을 하였을 때 데이터를 보여주며.. 한계점을 말해야하나…

25세 예측, 편차 컷던 사람(양사이드) 250 이상의 조건을 안맞는거 하나 보여주고

우리가 사용한 패널 데이터의 한계였다..!!! 다양한 분포가 없어서 대표를 할 수 가 없다.

요기서 아웃라이어가 우리한테 맞는 모델이 아니다라는 말은 어떤거에용…??

고급인력자료는 아님쓰~ 기술직엔 포함안됨..;;

개선사항은 적는것도 좋을듯해용 거기까지도 분석이니꼥

프로젝트를 마지며…(한계점)

데이터의 한계 – 패널데이터가 모집단을 대표할 수 없음.

결측치가 발생할 가능성이 높음.->자체적으로 정의하기때문에 정확한 예측 어려움( 예)경력 )

실 데이터에서 중요한 종속변수가 정확한 데이터가 아님. 대략적인 수치(50만원단위로 높아짐)

주관적으로 판단하는 데이터들이 많이 있었음.