📚🏥 Insurance 회귀 분석 캡스톤

아이티윌 데이터분석 부트캠프 52기 이광호 강사 (leekh4232@gmail.com)

“왜 어떤 사람의 의료보험 청구 비용은 높고, 어떤 사람의 비용은 낮을까?”

의료 비용은 개인의 건강 상태를 반영하기도 하지만, **나이, 성별, 거주 지역, 생활습관**에 따라 체계적으로 달라집니다.

특히 흡연 여부는 의료비를 얼마나 크게 변화시킬까? 젊은 사람도 흡연자라면 높은 비용을 지불할까?

이번 캡스톤 과제에서는 **Insurance 데이터셋**을 활용하여 개인의 인구학적·건강 특성이 **의료보험 청구 비용을 어떤 구조로 설명하는지**를 데이터와 회귀모형을 통해 단계적으로 탐구합니다.

이 과제의 목표는 예측 정확도를 높이는 것이 아니라, **“의료 비용 불평등은 어떤 논리로 형성되는가”를 수치와 언어로 설명하는 것**입니다.

※ 본 과제는 팀 / 개인 단위 모두 수행 가능합니다.

📘 재현성·환경 안내

- 데이터 로딩은 다음 헬퍼를 사용합니다: `load_data("insurance")`
- 분석 환경(패키지 버전)을 보고서 머리말에 명시하세요: `python`, `pandas`, `numpy`, `seaborn`, `statsmodels`, `scikit-learn`
- 임의성 통제를 위해 `random_state=42`를 일관되게 사용합니다. 학습/검증 분리와 교차검증에서 동일 시드로 재현 가능하게 설정하세요.
- 수치/그래프만 나열하지 말고, “왜 이 선택을 했는가”를 짧은 문장으로 남깁니다.

📘 데이터 불러오기

```
load_data("insurance")
```

📘 데이터 설명

개인의 기본 건강·인구학적 정보를 바탕으로 **의료보험 청구 비용**을 설명하기 위해 수집된 데이터입니다.

- 관측치: 약 1,338명

변수	설명
charges	의료보험 청구 비용 (종속변수, USD)
age	개인의 나이 (년)
sex	성별 (male, female)
bmi	체질량지수 (Body Mass Index)
smoker	흡연 여부 (yes, no)
children	부양 자녀 수
region	거주 지역 (southwest, southeast, northwest, northeast)

📘 과제 수행 미션

📝 미션 1. “이 데이터는 믿을 만할까?”

- 결측·이상치·편향을 점검하고, 처리 기준을 제시한다.
- charges가 음수이거나 극단값인지, age/bmi 등의 범위가 현실적인지 확인한다.
- 범주형 변수(sex, smoker, region)가 몇 개의 범주로 구성되어 있고, 특정 범주에 데이터가 지나치게 몰려 있지는 않은지 확인한다.
- 전처리 전·후가 어떻게 달라졌는지 한눈에 비교하는 표나 요약을 만든다.
- 단위·해석 주의: `charges`는 USD입니다. 현실적인 의료비 수준인지 평가하세요.

> 📌 **출제 의도** “이 값이 말이 되나?”를 먼저 묻고, 어떻게 처리했는지를 기록해 **나중 해석에 근거를 남기는 연습**입니다.

📝 미션 2. “의료비와 핵심 변수의 첫인상”

- charges, age, bmi, children 분포를 히스토그램/KDE로 확인하고 알 수 있는 객관적 사실을 서술한다.
- 왜도/이상치가 회귀에 줄 수 있는 영향과 변환할 필요가 있는지 서술하시오.
- 분포 비교는 동일 축 스케일로 제시하고, 평균/중앙값/꼬리의 차이를 문장으로 요약하세요.
- 의료비의 long-tail 분포(극단적 고액 청구)가 해석에 미치는 영향도 짧게 언급하세요.

> 📌 **출제 의도** 목표변수·핵심 변수의 생김새를 먼저 읽고 **변환 필요성을 스스로 판단**하게 합니다. 숫자보다 해석 문장이 중요합니다.

📝 미션 3. “로그/비선형 변환을 고민해 보자”

- charges 혹은 주요 변수(age, bmi, children)에 로그/제곱근 등 변환을 적용해 전후 분포를 **나란히** 비교한다.
- 변환이 해석과 모델 적합에 주는 장단점, 해석이 어떻게 달라지는지 예상한다.
- “이 변환이 없으면 어떤 함정에 빠질까?”를 한 줄로 정리한다.
- 선택 기준을 명시하세요: 왜 `log(charges)`인지, 왜 특정 변수에 변환을 적용하는지 데이터 분포 근거로 설명합니다.
- 필요한 경우 Box-Cox 등의 변환을 비교해 가장 해석 친화적인 옵션을 선택하세요.

> 📌 **출제 의도** 단순히 변환을 쓰는 것이 아니라 **“왜 필요했는가, 해석이 어떻게 달라지는가”**를 설명하는 연습입니다.

📝 미션 4. “성별과 거주 지역은 의료비를 결정할까?”

- 성별(sex)과 지역(region)별 의료비 분포를 시각화(박스플롯, 바이올린 플롯)한다.
- 중앙값·분포 겹침을 근거로 “어느 집단이 비싼가?”, “차이가 얼마나 뚜렷한가?”를 문장으로 적으세요.

- “왜 이런 차이가 생겼을까?”를 건강보험 체계·지역 의료 인프라·생활 비용 차이 등으로 추정해 보세요.

> 📌 **출제 의도** 범주형 요인이 의료비를 **어디서 가르고 어디서 겹치는지**를 이야기로 풀어내게 합니다.

📝 미션 5. “흡연은 정말로 의료비를 크게 높일까?”

- smoker(흡연 여부)에 따라 charges가 다른지 시각화하고, 두 집단 평균 차이를 가설검정(예: t-test)으로 확인한다.
- 효과 크기(차이의 크기)를 함께 제시하고, “실제로 의미 있는 차이인가?”를 말로 해석하세요.
- 정규성/등분산 가정 점검 후 필요 시 Welch’s t-test나 비모수 검정을 선택하세요.
- 효과 크기는 `Cohen's d` 또는 `Hedges' g`로 제시하고, 의료 정책 관점에서 의미를 한 줄로 번역합니다.

> 📌 **출제 의도** 두 집단 비교에서 **방법 선택과 효과 크기 해석**을 연습하고, 숫자를 의미로 번역하게 합니다.

📝 미션 6. “나이대별로 의료비 차이가 뚜렷할까?”

- age를 여러 구간으로 나누어(예: 18~30, 31~50, 51+) 각 연령대별 charges 분포를 시각화한다.
- 분산분석(ANOVA)으로 전체 차이를 확인하고, 사후검정으로 어느 연령대 사이에서 차이가 나는지 정리한다.
- 사후검정은 Tukey HSD 또는 Games-Howell(등분산 위반 시)을 사용하고, “의료비 연령 서열표” 형태로 요약하세요.

> 📌 **출제 의도** 여러 범주를 **전체→어디서 차이** 순서로 해석하며, 결과를 서열/지도처럼 정리하는 훈련입니다.

📝 미션 7. “변수들은 서로 섞여 있을까?”

- 주요 연속형 변수 간 상관행렬(age, bmi, children, charges)을 계산한다.
- 상관행렬을 히트맵으로 시각화하고, Variance Inflation Factor(VIF)로 다중공선성을 점검한다.
- age와 bmi 같이 약한 상관을 보이는 변수들이 모델에 주는 정보를 논의하세요.
- 공선성 문제가 없다면 그 이유를, 있다면 완화 전략(변수 제거, 결합지표, 정규화 회귀)을 비교하세요.

> 📌 **출제 의도 중복 정보와 독립성**을 의식하며, 회귀에 넣을 변수를 깔끔히 설계하게 합니다.

📝 미션 8. “어떤 변수가 의료비와 가장 가까울까?”

- charges와 age, bmi, children의 상관을 계산한다.
- Pearson과 Spearman을 모두 계산하고, 비교해 무엇이 더 적합한지 이유를 적으세요.
- 각 변수의 관계를 의료학·보건학적 언어로 해석하고(예: “나이가 많을수록 의료비 증가”, “과체중(높은 BMI)은 의료비와 강한 관계”), 인과성을 말할 수 있는지 비판적으로 평가합니다.

> 📌 **출제 의도** 상관계수 선택 이유와 **수치→의미** 해석을 연습하며, 변수 설계의 근거를 쌓게 합니다.

📝 미션 9. “상관관계는 곧 원인일까?”

- 지금까지의 탐색 결과를 바탕으로 “상관이 높다 = 원인이다”라는 착각을 피하자.
- 예: age와 charges의 높은 상관이 정말 나이 자체가 비용을 결정하는가, 아니면 나이가 **건강 상태의 대리변수**인가?
- smoker 효과: 흡연자는 정말 더 많이 병에 걸려서인가, 아니면 보험사의 **위험 평가 기준**인가?
- 데이터 수집 시점·의료 접근성 차이·미측정 변수(직업, 교육, 소득)가 결론에 주는 한계를 서술하세요.

> 📌 **출제 의도** 데이터가 말해 주는 것과 말하지 못하는 것을 구분하며, 신중한 해석의 중요성을 체감하게 합니다.

📝 미션 10. “의료비를 설명하는 회귀모형 설계”

- charges(또는 변환값)를 종속변수로 하는 다중선형회귀를 설계한다.
- 변수 선택·인코딩(범주형 처리)·스케일/변환 이유를 “**설명하고 싶어서**” 관점으로 글로 남긴다.
- 학습/검증 데이터 분리 방식(예: train/test split 80/20)을 명시한다.
- 검증 전략을 구체화하세요: `train/test split`에 더해 `K-fold`(예: 5-fold)로 평균 성능과 분산을 함께 보고합니다.
- 과적합·공선성 위험 완화를 위해 Ridge/Lasso/Elastic Net을 후보로 비교하고, “해석-성능” 균형 관점에서 선택 이유를 기록합니다.

> 📌 **출제 의도** 설계 선택이 임의가 아닌 **설명 의도**에 근거하도록 훈련합니다.

📝 미션 11. “회귀계수는 무엇을 말해주나?”

- 계수(또는 표준화 계수)와 신뢰구간, 방향·크기를 해석한다.
- “나이가 1년 늘면 의료비가 어떻게 변하는가”, “흡연자는 비흡연자보다 평균 얼마나 더 높은 비용을 지불하는가” 처럼 **물리/의료적 의미**로 번역한다.
- 변환 변수가 있다면, 변환을 감안한 해석을 명확히 쓴다.
- 표준화 계수(베타)와 비표준화 계수를 병행 제시하고, 단위/변환을 고려한 해석 문장을 명확히 작성합니다.
- 범주형 변수(sex, smoker, region)의 계수는 기준 범주 대비 효과로 명확히 해석하세요.

> 📌 **출제 의도** 숫자를 **의료비 구조의 언어**로 바꾸는 연습입니다.

📝 미션 12. “모형 진단과 개선”

- 잔차 정규성/등분산/선형성, 영향력을 잔차플롯, Q-Q, Cook’s distance 등으로 점검한다.
- 문제 지점(예: 극단적 고액 청구, 특정 집단에서의 체계적 오류)과 개선 아이디어(변환, 변수 교체/제거, 강건 회귀 등)를 제안한다.
- R^2, RMSE/MAE 등 기본 지표를 보고하고, 실제 의료비 예측에서 이 오차가 얼마나 의미 있는지 해석하세요.
- 캘리브레이션(예측 vs. 실제) 산점도를 함께 제시하고, 체계적 과소/과대 예측 구간을 언급하세요.
- K-fold 평균 지표와 표준편차를 보고하여 안정성을 함께 판단합니다.

> 📌 **출제 의도** 점수보다 **가정·진단과 해석**을 통해 “얼마나 믿을 수 있는가”를 판단하게 합니다.

📝 미션 13. “같은 나이인데 왜 의료비는 다를까?”

- age가 같은 두 가상의 개인을 설정하고(예: A는 비흡연 저BMI, B는 흡연 고BMI), 다른 변수 차이로 의료비 차이를 설명한다.
- 비전공자에게 이야기하듯, 모델이 설명하는 것과 못하는 것을 구분해 제시한다.
- 이야기의 흐름을 “데이터 관찰 → 회귀계수 → 의료비 차이 설명” 순서로 연결한다.
- 예시 틀: “두 사람 모두 40세이지만, A는 흡연하지 않고 BMI가 25인 반면 B는 흡연자이고 BMI가 35입니다. 모델에 따르면 B의 예상 의료비는 A보다 훨씬 높습니다. 왜냐하면 흡연이 [계수] 만큼 비용을 증가시키고, BMI 1 증가가 [계수] 만큼 증가시키기 때문입니다. 하지만 이 모델은 개인의 건강 이력, 유전 요인, 직업 스트레스 같은 정보를 담지 못합니다.”
- 모델이 설명하지 못한 요인(개인 건강 이력, 유전 요인, 사회경제적 지위, 의료 접근성 등)의 가능성도 덧붙여 설득력을 높입니다.

> 📌 **출제 의도** 지금까지의 분석을 **스토리로 엮어** 설득력 있게 전달하는 마무리입니다. 모델이 말하는 것과 말하지 못하는 것을 구분해 주세요.

📘 채점 기준 (총 100점)

미션	배점	평가 포인트	감점 기준
1. 데이터 신뢰도	7	결측·이상치·편향·단위 확인, 처리 기준·영향 서술	결측/이상치 언급 누락(-2), 처리 기준·영향 설명 없음(-2), 단위 고려 부족(-1)
2. 분포 첫인상	6	목표·핵심 변수 분포 해석, 변환 필요성·long-tail 논의	분포 시각화 부재(-2), 변환 필요성 언급 없음(-2), 해석 없이 그래프만(-1)
3. 변환 비교	6	변환 전후 비교, 변환 이유와 해석 변화 설명	변환 이유 미설명(-2), 전후 비교/해석 없음(-3)
4. 성별·지역 분석	7	범주별 분포·패턴·겹침 해석, 사회적 맥락	범주별 시각화 없이 수치만(-2), 패턴/겹침 해석 없음(-2), 맥락 논의 부재(-1)
5. 흡연 효과 검정	7	가설검정, 검정 방법·효과 크기 제시, 의료 정책 의미	검정/효과 크기 미제시(-2), 의미 해석 없음(-2), 시각화 부재(-1)
6. 연령대별 비교	7	ANOVA+사후검정, 연령대 구간 설정 논리, 서열표 정리	ANOVA만 하고 사후검정 없음(-2), 구간 설정 미설명(-1), 정리 부재(-1)
7. 다중공선성·상관	7	상관/VIF 점검, 변수 독립성 논의, 공선성 전략	VIF/상관 없이 주장만(-2), 공선성 대응 전략 없음(-2)
8. 연속형 상관	7	Pearson/Spearman 계산·선택 근거, 의료학적 해석, 인과성 비판	계수 선택 이유 미제시(-2), 해석 없이 수치만(-2), 인과성 비판 부재(-1)
9. 상관 vs 인과	6	상관의 한계, 대리변수·누락 변수 명시, 의료 정책 한계	한계/누락 요인 언급 없음(-2), 상관=원인 단정(-2), 정책 고려 부족(-1)
10. 회귀 설계	13	변수 선택·인코딩·스케일/변환·검증·정규화 전략 근거, 설명 의도 명시	설계 이유 미기재(-3), 검증 전략 부재(-2), 정규화 비교 미흡(-2)
11. 계수 해석	12	비표준화/표준화 계수, 신뢰구간, 의료적 의미 번역, 범주형 해석	계수 해석 없음(-3), 의료적 의미 번역 부재(-2), 범주형 효과 미설명(-2)
12. 모형 진단·성능	8	잔차/가정/영향력·캘리브레이션, K-fold 평균+분산, 실무 오차 의미, 개선안	진단 그래프 누락(-2), 안정성/캘리브레이션 언급 없음(-1), 오차 의미 부재(-1), 개선안 없음(-1)
13. 스토리텔링	12	가상 개인 사례 설명, 모델 설명/미설명 요인 구분, 비전공자 친화	사례 미제시(-3), 의료비 차이 이유 해석 없음(-3), 모델 한계 언급 없음(-2), 친화성 부재(-1)