국가금연지원 서비스 만족도 설문분석:

주건재

OVERVIEW

1. INTRODUCTION
2. EDA
3. MODEL
4. RESULT

1. INTRODUCTION

1.1 TOPIC

Survey: 국가금연지원 서비스 만족도 조사

설문항목	매우 그렇다	그렇다	보통이다	그렇지 않다	전혀 그렇지 않다
1. 금연상담사는 상담 약속시간을 잘 지켰습니까?	○	○	○	○	○
2. 금연하는 동안 상담사로부터 도움을 충분히 받았습니까?	○	○	○	○	○
3. CO측정, 혈압, 체중 등을 충분히 체크를 받으셨습니까?	○	○	○	○	○
4. 금연상담사나 다른 직원들이 친절하게 잘 대해주었습니까?	○	○	○	○	○
5. 국가금연지원서비스를 정기적으로 방문하는 것이 불편하였습니까?	○	○	○	○	○
6. 국가금연지원서비스 이용이 금연성공에 얼마나 도움이 되었습니까?	○	○	○	○	○
7. 담배를 피우는 다른 사람에게도 국가금연지원서비스를 이용하도록 권유할 생각이 있습니까?	○	○	○	○	○

Research Question

설문지 7번 문항을 만족도로 가정하고 서비스 개선의 방향을 제안한다.

Goals

1st goal: 모델링 후 1~5문항의 중요도 순서를 나열한다.

2nd goal: 서비스 개선 이후의 상황을 시뮬레이션한다.

1.1 TOPIC

Survey Assumption

설문항목

1. 금연상담사는 상담 약속시간을 잘 지켰습니까?
2. 금연하는 동안 상담사로부터 도움을 충분히 받았습니까?
3. CO측정, 혈압, 체중 등을 충분히 체크를 받으셨습니까?
4. 금연상담사나 다른 직원들이 친절하게 잘 대해주었습니까?
5. 국가금연지원서비스를 정기적으로 방문하는 것이 불편하였습니까?
6. 국가금연지원서비스 이용이 금연성공에 얼마나 도움이 되었습니까?
7. 담배를 피우는 다른 사람에게도 국가금연지원서비스를 이용하도록 권유할 생각이 있습니까?

	대분류	소분류	개선방법
1	상담사	서비스	상담사 교육
2	상담사	인원	상담사 충원
3	장비	장비	장비 확충
4	상담사	서비스	상담사 교육
5	서비스	접근성	서비스 다양화(ex금연버스)
6	삭제		

Simulation Assumption

1.2 DATA OVERVIEW

Raw Data

Data Resource: 공공데이터포털()

	기관유형	지역	서비스구분	제공기관	출생년도	성별	등록유형	문항1	문항2	문항3	문항4	문항5	문항6	문항7
0	보건소	대전광역시	보건소 금연클리닉	대전 서구보건소	1970~1979	남	보건소	1	1	1	1	3	1	2
1	보건소	경기도	보건소 금연클리닉	경기 수원시 장안구보건소	1950~1959	남	보건소	0	0	0	0	4	0	0
2	보건소	광주광역시	보건소 금연클리닉	광주 광산구보건소	1980~1989	남	보건소	0	0	0	0	3	0	0
3	보건소	경기도	보건소 금연클리닉	경기 파주시보건소	1990~1999	남	보건소	1	1	1	0	3	1	1

Data Type

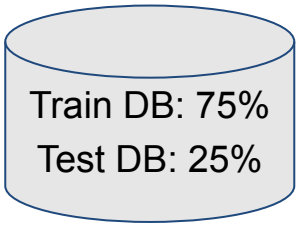
- 문항 1 ~ 문항 6 : 연속형
- 문항 7 (종속변수): 순서형
- 출생년도: 연속형
- 성별: 명목형

(Johnson & Creech, 1983; Norman, 2010; Sullivan & Artino, 2013; Zumbo & Zimmerman, 1993)

1.3 Analysis Overview

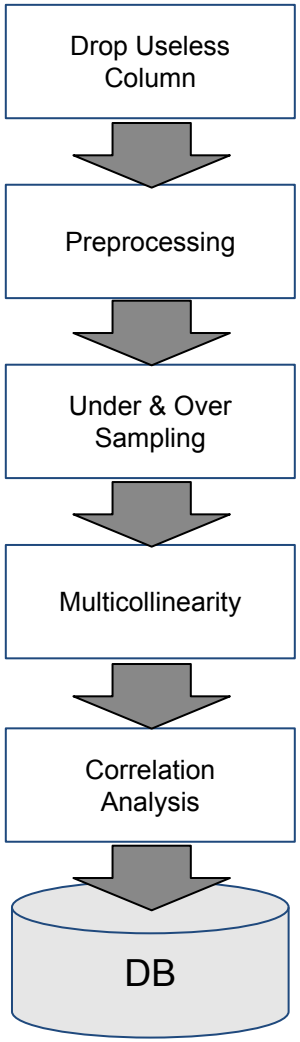
Data Understanding

Research Goal:
중요한 문항 찾기

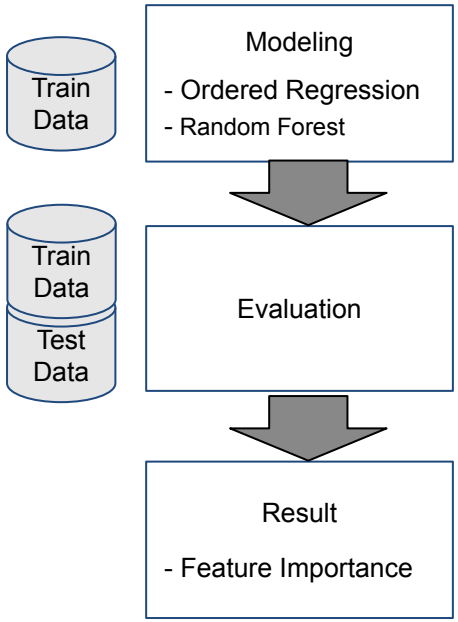


특징:
- shape: (55911,14)
- categorical data
- imbalanced data

Preprocessing & EDA



Modeling & Evaluation

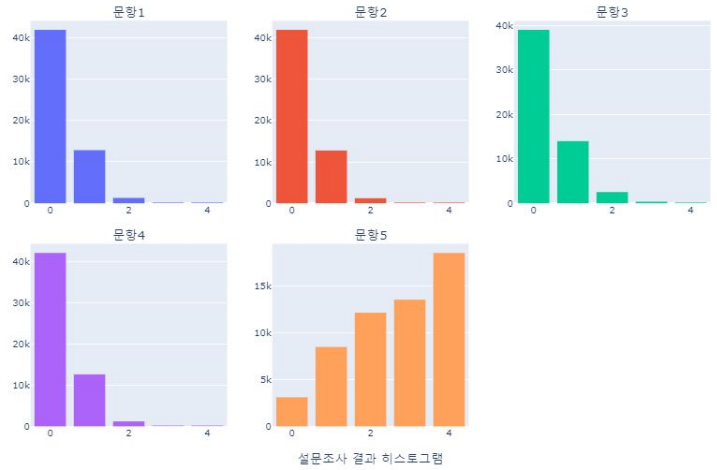


2. EDA

2.1 Drop False Answer & Question Rate Consistency

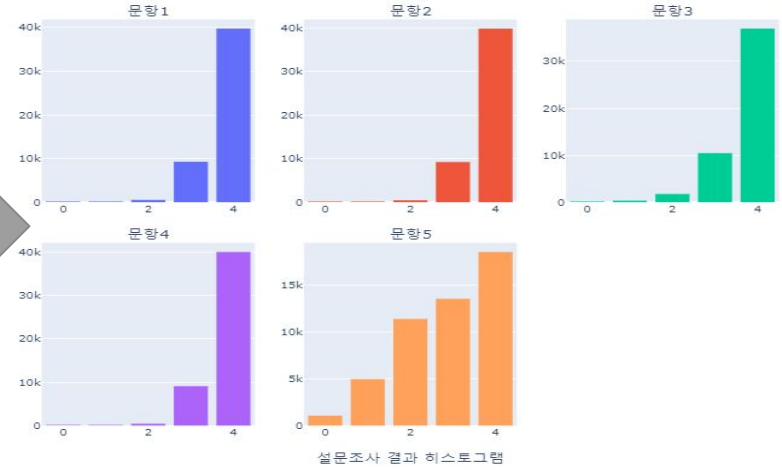
Drop False Answer

- 응답이 한가지로 되어있는 샘플
- 5번문항의 긍정 부정이 타문항과 반대
- Cronbach-alpha test: 0.65 → 0.82



Rate Consistency

- 매우그렇다: 0 →4, 그렇다: 1→3
  전혀 그렇지 않다: 5→0, 그렇지 않다 4→1
- 시각화에 도움이 되며 해석에 용이하다.



Cronbach-alpha Test

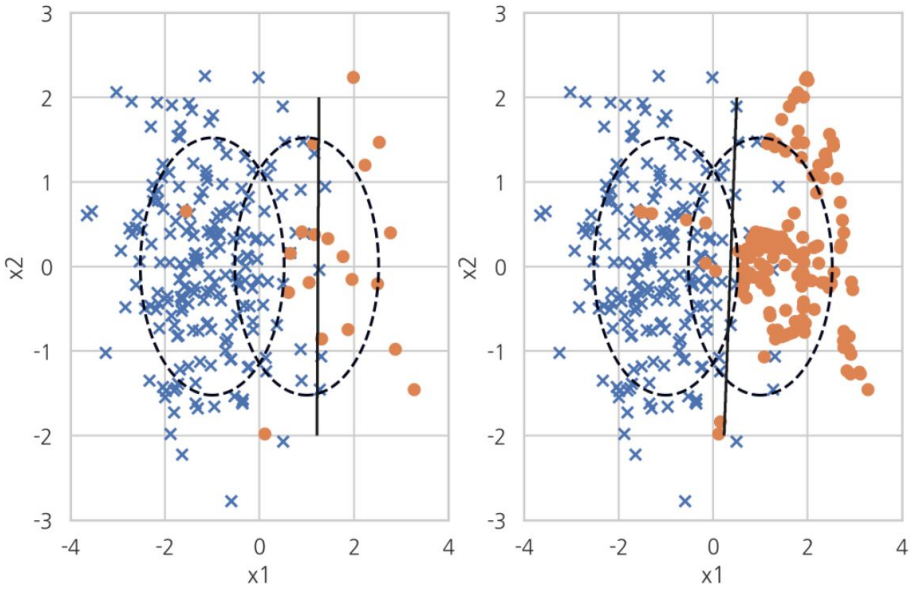
- 서베이의 internal consistency를 측정
- 값은 0과 1 사이에 위치하며
  1에 가까울수록 서베이가 reliable 하다.

Cronbach's Alpha	Internal consistency
0.9 ≤ α	Excellent
0.8 ≤ α < 0.9	Good
0.7 ≤ α < 0.8	Acceptable
0.6 ≤ α < 0.7	Questionable
0.5 ≤ α < 0.6	Poor
α < 0.5	Unacceptable

2.2 Resampling

Combining Oversampling Undersampling:

- SMOTE-Tomek
  - SMOTE: 소수 클래스의 데이터와 가까운 k개 클래스 데이터 사이 가상데이터 생성
  - Tomek: tomek link 에 있는 다수 클래스 제거
- Random Undersampling



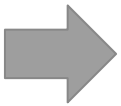
SMOTE Tomek 설명예시

2.3 Multicollinearity

VIF < 20

- 독립 변수: 연속형 데이터
- 문항4 제거

	Attribute	VIF Scores
0	문항1	21.10
1	문항2	26.09
2	문항3	9.08
3	문항4	20.34
4	문항5	6.46



	Attribute	VIF Scores
0	문항1	19.68
1	문항2	17.24
2	문항3	9.08
3	문항5	6.38

2.4 Correlation Analysis

Kendall Tau

- -1 에서 1의 값
- 무상관: $|r| < 0.25$
- 강한 상관: $|r| > 0.75$

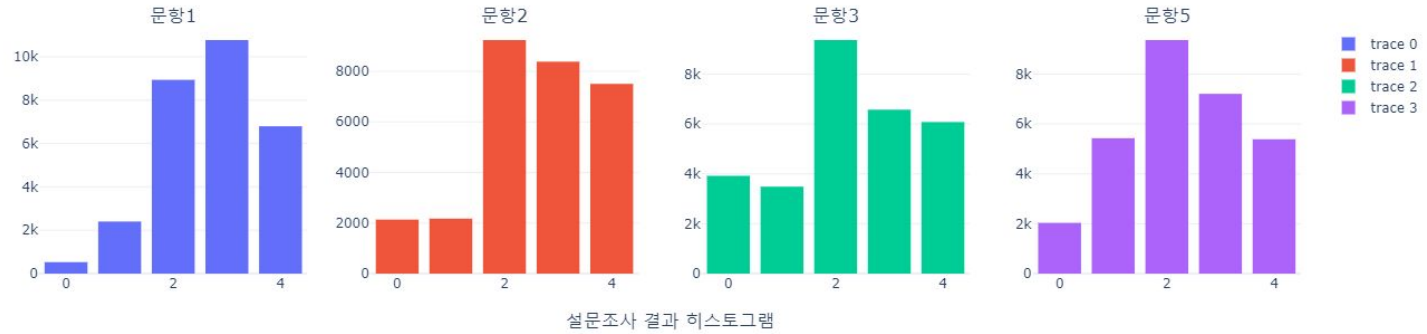
- 출생년도, 성별 제거

	문항7
~~출생년도~~	0.005
~~성별~~	-0.001
문항1	0.588
문항2	0.633
문항3	0.587
문항4	0.613
문항5	0.305

2.4 Preprocessed Data

Data shape: (29430 , 6)

Independent Variable



Cronbach-alpha(0.05) : 0.885 (CI [0.883, 0.887])

Multicollinearity

	Attribute	VIF Scores
0	문항1	19.68
1	문항2	17.24
2	문항3	9.08
3	문항5	6.38

Correlation Analysis

	문항7
문항1	0.670
문항2	0.662
문항3	0.765
문항5	0.438

3. MODEL

3.1 Ordinal Regression

$$y^* = \mathbf{x}^\mathsf{T}\beta + \varepsilon$$

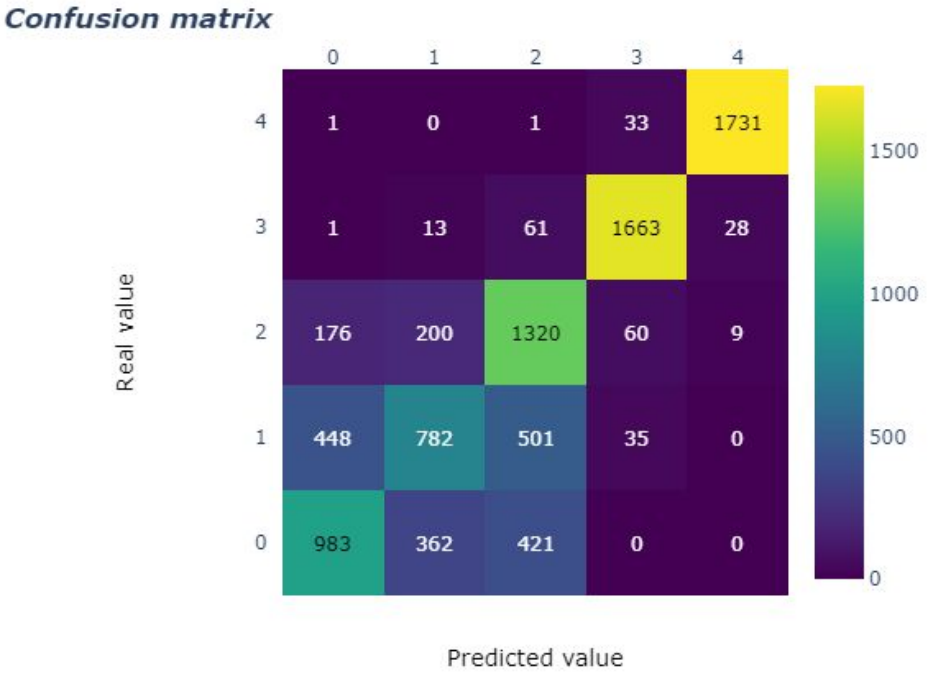
Ordinal Logit Regression

- 종속변수가 순서형인 회귀모형
- logistic regression과 같은 link function
- error term의 정규성 검정 불가때 사용
- 각 class의 threshold 또한 통계량

$$y = \begin{cases} 0 & \text{if } y^* \le \mu_1, \\ 1 & \text{if } \mu_1 < y^* \le \mu_2, \\ 2 & \text{if } \mu_2 < y^* \le \mu_3, \\ \vdots & \\ N & \text{if } \mu_N < y^* \end{cases}$$

Model Summary

	precision	recall	f1-score
0	0.61	0.56	0.58
1	0.58	0.44	0.50
2	0.57	0.75	0.65
3	0.93	0.94	0.94
4	0.98	0.98	0.98
accuracy			0.73
macro avg	0.73	0.73	0.73
weighted avg	0.73	0.73	0.73



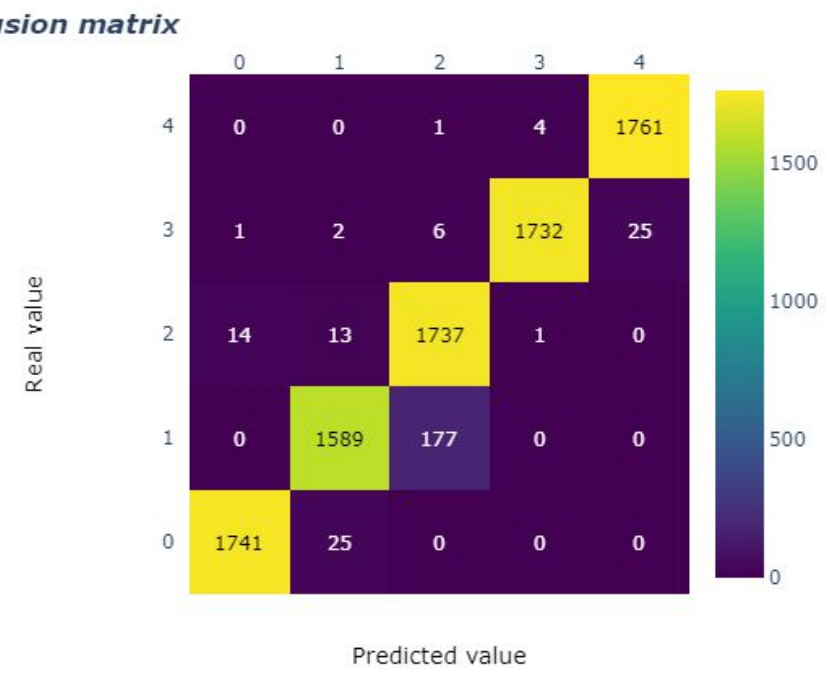
3.2 Random Forest

Feature Importance

- 통계량이 아닌 input과 output으로 각 feature의 중요도를 계산
- 통계적 검증이 어려움
- 설명력이 떨어짐
- 모든 모델에 적용 가능

Model Summary

	precision	recall	f1-score
0	0.99	0.99	0.99
1	0.98	0.90	0.94
2	0.90	0.98	0.94
3	1.00	0.98	0.99
4	0.99	1.00	0.99
accuracy			0.97
macro avg	0.97	0.97	0.97
weighted avg	0.97	0.97	0.97



4. RESULT

Feature Importance

- **Ordinal Regression**

  문항 3 > 1 > 5 > 2

	coef	std err	z	P>|z|	[0.025	0.975]
문항1	0.9210	0.027	34.425	0.000	0.869	0.973
문항2	0.6264	0.025	24.964	0.000	0.577	0.676
문항3	1.7854	0.023	76.448	0.000	1.740	1.831
문항5	0.8225	0.017	47.586	0.000	0.789	0.856

- **Random Forest**

  문항 3 > 2 > 1 > 5

index	feature importance
문항1	0.20
문항2	0.25
문항3	0.38
문항5	0.17

Discussion

- 2번 문항의 중요도가 쟁점
- Ordinal Regression성능의 문제가 있어 Feature Importance 신뢰도도 떨어짐
- Random Forest Feature Importance를 신뢰
- Further Study: 과년도 데이터로 재차 테스트

Research Result

- **1st goal: 모델링 후 1~5문항의 중요도 순서를 나열**
  - 문항3 > 문항2 > 문항1(4) > 문항5

- **2nd goal: 서비스 개선 이후의 상황을 시뮬레이션**
  - 시간 부족으로 인한 실패

설문항목

1. 금연상담사는 상담 약속시간을 잘 지켰습니까?
2. 금연하는 동안 상담사로부터 도움을 충분히 받았습니까?
3. CO측정, 혈압, 체중 등을 충분히 체크를 받으셨습니까?
4. 금연상담사나 다른 직원들이 친절하게 잘 대해주었습니까?
5. 국가금연지원서비스를 정기적으로 방문하는 것이 불편하였습니까?
6. 국가금연지원서비스 이용이 금연성공에 얼마나 도움이 되었습니까?
7. 담배를 피우는 다른 사람에게도 국가금연지원서비스를 이용하도록 권유할 생각이 있습니까?

	대분류	소분류	개선방법
1	상담사	서비스	상담사 교육
2	상담사	인원	상담사 충원
3	장비	장비	장비 확충
4	상담사	서비스	상담사 교육
5	서비스	접근성	서비스 다양화(ex금연버스)
6	삭제		

Research Result

- **Limitation**
  - 심한 불균형의 데이터
  - 문항이 너무 적어 공선성 제거의 어려움
  - 시간 부족
- **Further Study**
  - 과년도 데이터로 테스트
  - Tree 모델 적용

References

- Cao, C., Chicco, D., & Hoffman, M. M. (2020). The MCC-F1 curve: a performance evaluation technique for binary classification. *arXiv preprint arXiv:2006.11278*.
- Kim, J., Han, Y., & Lee, J. (2016). Data imbalance problem solving for smote based oversampling: Study on fault detection prediction model in semiconductor manufacturing process. *Advanced Science and Technology Letters*, *133*, 79-84.
- Munirathinam, S., & Ramadoss, B. (2016). Predictive models for equipment fault detection in the semiconductor manufacturing process. *IACSIT International Journal of Engineering and Technology*, *8*(4), 273-285.
- Kerdprasop, K., & Kerdprasop, N. (2010, March). Feature selection and boosting techniques to improve fault detection accuracy in the semiconductor manufacturing process. In *World Congress on Engineering 2012. July 4-6, 2012. London, UK.* (Vol. 2188, pp. 398-403). International Association of Engineers.
- Johnson, D.R., & Creech, J.C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, 48, 398-407.
- Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education*, *15*(5), pp. 625-632. Retrieved from: https://link.springer.com/article/10.1007%2Fs10459-010-9222-y#citeas.
- Sullivan, G. & Artino Jr., A. R. (2013). Analyzing and Interpreting Data From Likert-Type Scales. *Journal of Graduate Medical Education*. *5*(4), pp. 541-542.
- Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology*, 34, 390-400.

Q&A