# 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 문항의 답변을 만족도로 가정하고 서비스 개선의 방향을 제안한다.

## Goals

$1^{st}$ goal: 해석 가능한 모델을 사용하여 7번 문항 예측한다.

$2^{nd}$ goal: 서비스 개선 이후의 상황을 시뮬레이션한다.

# 1.2 DATA OVERVIEW

## Raw Data

**Data Resource:** 공공데이터포털(https://archive.ics.uci.edu/ml/datasets/SECOM)

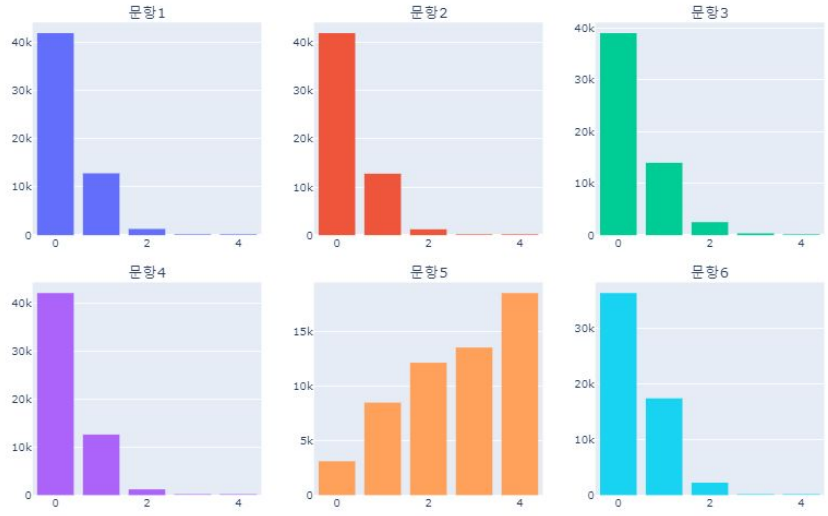
| | 기관유형 | 지역 | 서비스구분 | 제공기관 | 출생년도 | 성별 | 등록유형 | 문항1 | 문항2 | 문항3 | 문항4 | 문항5 | 문항6 | 문항7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 보건소 | 대전광역시 | 보건소 금연클리닉 | 대전 서구보건소 | 1970~1979 | 남 | 보건소 | 1 | 1 | 1 | 1 | 3 | 1 | 2 |
| 1 | 보건소 | 경기도 | 보건소 금연클리닉 | 경기 수원시 장안구보건소 | 1950~1959 | 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 Summary

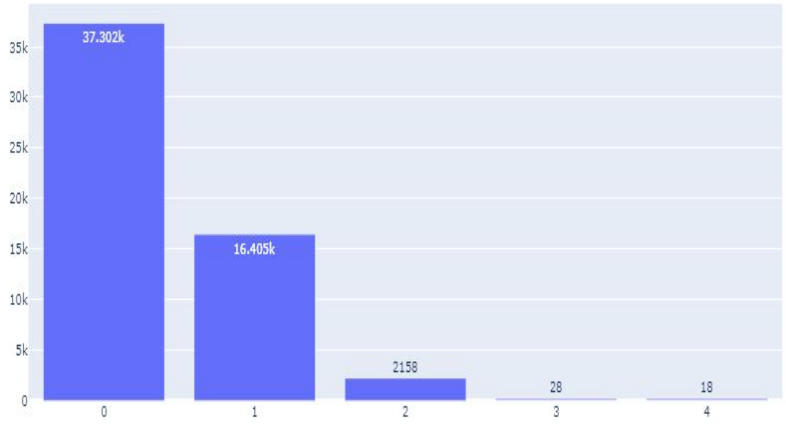
| | Types | Counts | Uniques | Nulls | Min | Max |
|---|---|---|---|---|---|---|
| 기관유형 | object | 55911 | 2 | 0 | 금연지원센터 | 보건소 |
| 등록유형 | object | 55911 | 11 | 0 | 기타 | 캠페인 |
| 문항1 | int64 | 55911 | 5 | 0 | 0 | 4 |
| 문항2 | int64 | 55911 | 5 | 0 | 0 | 4 |
| 문항3 | int64 | 55911 | 5 | 0 | 0 | 4 |
| 문항4 | int64 | 55911 | 5 | 0 | 0 | 4 |
| 문항5 | int64 | 55911 | 5 | 0 | 0 | 4 |
| 문항6 | int64 | 55911 | 5 | 0 | 0 | 4 |
| 문항7 | int64 | 55911 | 5 | 0 | 0 | 4 |
| 서비스구분 | object | 55911 | 3 | 0 | 단기금연캠프 | 찾아가는 금연서비스 |
| 성별 | object | 55911 | 2 | 0 | 남 | 여 |
| 제공기관 | object | 55911 | 229 | 0 | 강원 강릉시보건소 | 충북금연지원센터 |
| 지역 | object | 55911 | 17 | 0 | 강원도 | 충청북도 |
| 출생년도 | object | 55911 | 9 | 0 | 1920~1929 | 2000~2009 |

# 1.2 DATA OVERVIEW

## Survey Result




문항 1~6 히스토그램




문항7히스토그램

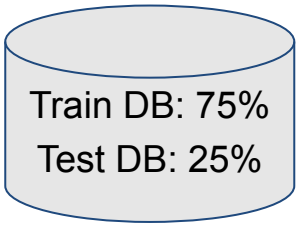
## Issue:

- 리커트척도의 불균형은 어떻게 처리해야하나?
- 종속변수 문항7의 2, 3, 4는 어떻게 처리해야하나?

# 1.3 Analysis Overview

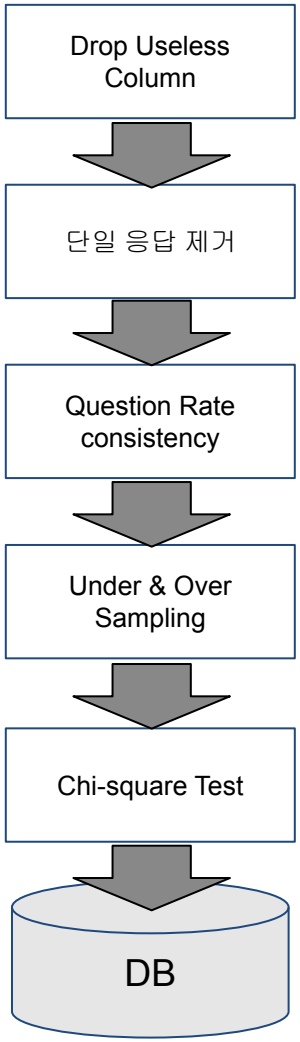
## Data Understanding

Research Goal:
중요한 문항 찾기




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

## Preprocessing & EDA

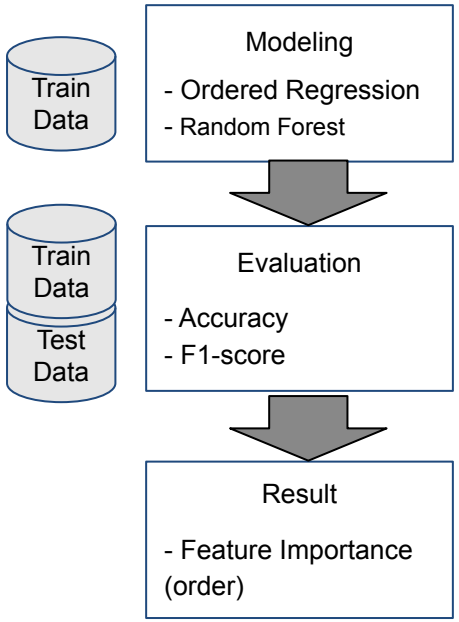



기관유형, 지역, 서비스 구분, 제공기관

6344개 샘플 제거
Cronbach Alpha Test:
0.65 -> 0.82

응답: 긍정=4 부정=0

## Modeling & Evaluation



# 2. EDA

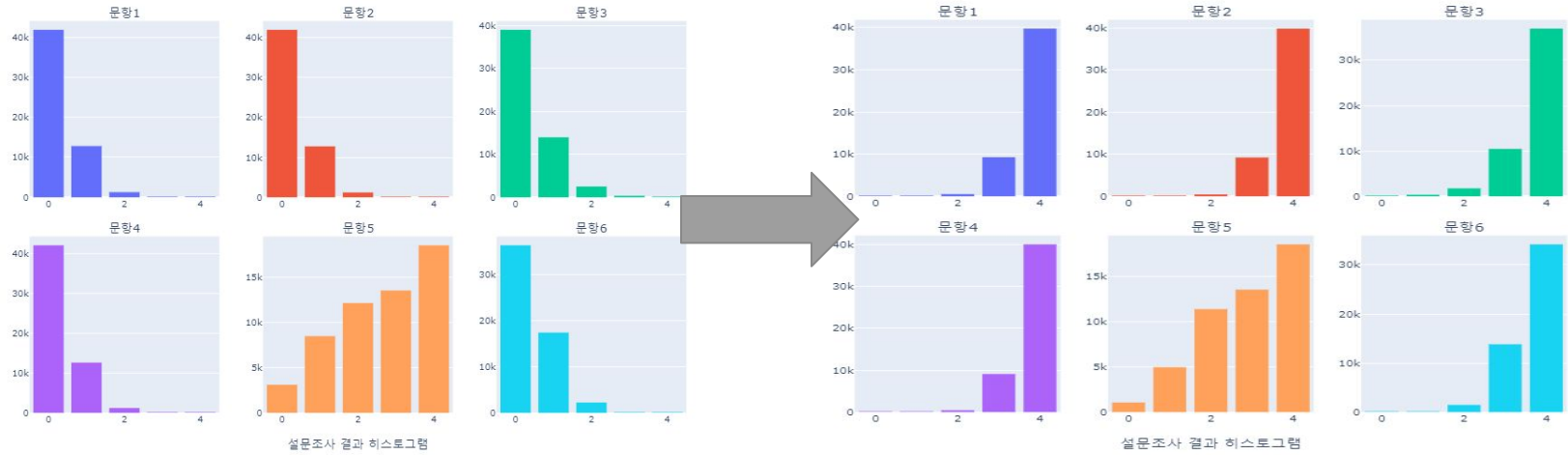
# 2.1 Drop False Answer & Question Rate Consistency

## 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 Chi-square Test

**P-value = 0 ?**

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$\chi^2$ = chi squared

$O_i$ = observed value

$E_i$ = expected value

**Ex) 문항1**

```
문항 1
chi 스퀘어 값: 18190.38
p-value (0.05): 0.0
자유도 수: 16
```

```
측정값:
문항7      4     3    2   1  0
문항1
0          0     0    0   0  1
1          3     3    4   1  1
2        106    78  192   3  2
3       1152  4933  472   3  3
4      23408  4014  301  11  5
```

<문항 1 vs 7 cross table>

```
기대값:
                0            1           2           3           4
0        0.711004     0.260203    0.027928    0.000519    0.000346
1        8.532050     3.122435    0.335139    0.006226    0.004150
2      270.892581    99.137307   10.640679    0.197660    0.131773
3     4666.320239  1707.711667  183.293377    3.404831    2.269887
4    19722.544126  7217.768388  774.702876   14.390766    9.593844
```

<문항 1 vs 7 기대값 cross table>

# 2.2 Chi-square Test

## P-value = 0 ?

- 종속성이 강해 0에 수렴하는것인가?
- 데이터 크기와 불균형으로 야기된 오류인가?
  - 추후 분석을 위해 데이터 불균형 처리 방안
  - over & under sampling을 진행해야하나?

## Solution

*Information Systems Research*
Vol. 24, No. 4, December 2013, pp. 906–917
ISSN 1047-7047 (print) | ISSN 1526-5536 (online)

http://dx.doi.org/10.1287/isre.2013.0480


**Research Commentary**

### Too Big to Fail: Large Samples and the *p*-Value Problem

Mingfeng Lin
Eller College of Management, University of Arizona, Tucson, Arizona 85721, mingfeng@eller.arizona.edu

Henry C. Lucas, Jr.
Robert Smith School of Business, University of Maryland, College Park, Maryland 20742, hlucas@rhsmith.umd.edu

Galit Shmueli
Srini Raju Centre for IT & the Networked Economy, Indian School of Business, Hyderabad 500 032, India, galit_shmueli@isb.edu

The Internet has provided IS researchers with the opportunity to conduct studies with extremely large samples, frequently well over 10,000 observations. There are many advantages to large samples, but researchers using statistical inference must be aware of the *p*-value problem associated with them. In very large samples, *p*-values go quickly to zero, and solely relying on *p*-values can lead the researcher to claim support for results of no practical significance. In a survey of large sample IS research, we found that a significant number of papers rely on a low *p*-value and the sign of a regression coefficient alone to support their hypotheses. This research commentary recommends a series of actions the researcher can take to mitigate the *p*-value problem in large samples and illustrates them with an example of over 300,000 camera sales on eBay. We believe that addressing the *p*-value problem will increase the credibility of large sample IS research as well as provide more insights for readers.

*Key words*: empirical modeling; practical significance; effect size; *p*-value; statistical significance; inference
*History*: Alok Gupta, Senior Editor. This paper was received on August 15, 2012, and was with the authors 2 weeks for 1 revision. Published online in *Articles in Advance* April 12, 2013, and updated October 22, 2013.

### An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets

Bee Wah Yap[1], Khatijahhusna Abd Rani[2], Hezlin Aryani Abd Rahman[1], Simon Fong[3], Zuraida Khairudin[1], Nik Nairan Abdullah[4]

[1,2] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Selangor,Malaysia
[3]Faculty of Science and Technology, University of Macau, China
[4] Faculty of Medicine, Universiti Teknologi MARA, Selangor, Malaysia
[1]{beewah, hezlin, zuraida_k}@tmsk.uitm.edu.my, [2] ejahhusna@gmail.com, [3]ccfong@umac.mo, [4]niknairan@yahoo.com

**Abstract.** Most classifiers work well when the class distribution in the response variable of the dataset is well balanced. Problems arise when the dataset is imbalanced. This paper applied four methods: Oversampling, Undersampling, Bagging and Boosting in handling imbalanced datasets. The cardiac surgery dataset has a binary response variable (1=Died, 0=Alive). The sample size is 4976 cases with 4.2% (Died) and 95.8% (Alive) cases. CART, C5 and CHAID were chosen as the classifiers. In classification problems, the accuracy rate of the predictive model is not an appropriate measure when there is imbalanced problem due to the fact that it will be biased towards the majority class. Thus, the performance of the classifier is measured using sensitivity and precision Oversampling and undersampling are found to work well in improving the classification for the imbalanced dataset using decision tree. Meanwhile, boosting and bagging did not improve the Decision Tree performance.

**Keywords-** Bagging, Boosting, Oversampling, Undersampling, Imbalanced data

# Question

## 카이검정 및 불균형 처리를 안해도 모델링이 된다. 굳이 해야하나?

```
Optimization terminated successfully.
         Current function value: 0.266811
         Iterations: 157
         Function evaluations: 173
         Gradient evaluations: 173
```

OrderedModel Results

| | | | |
|---|---|---|---|
| Dep. Variable: | 문항7 | Log-Likelihood: | -9257.3 |
| Model: | OrderedModel | AIC: | 1.854e+04 |
| Method: | Maximum Likelihood | BIC: | 1.864e+04 |
| Date: | Tue, 03 May 2022 | | |
| Time: | 12:39:04 | | |
| No. Observations: | 34696 | | |
| Df Residuals: | 34684 | | |
| Df Model: | 12 | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| 출생년도 | 0.0070 | 0.012 | 0.566 | 0.571 | -0.017 | 0.031 |
| 성별 | -0.0868 | 0.060 | -1.450 | 0.147 | -0.204 | 0.031 |
| 문항1 | 0.0319 | 0.067 | 0.474 | 0.635 | -0.100 | 0.164 |
| 문항2 | 0.6260 | 0.077 | 8.122 | 0.000 | 0.475 | 0.777 |
| 문항3 | 0.7042 | 0.038 | 18.526 | 0.000 | 0.630 | 0.779 |
| 문항4 | 0.2049 | 0.066 | 3.107 | 0.002 | 0.076 | 0.334 |
| 문항5 | 0.1970 | 0.020 | 10.064 | 0.000 | 0.159 | 0.235 |
| 문항6 | 4.2164 | 0.043 | 98.817 | 0.000 | 4.133 | 4.300 |
| 0/1 | 7.2437 | 0.435 | 16.653 | 0.000 | 6.391 | 8.096 |
| 1/2 | 0.3819 | 0.242 | 1.579 | 0.114 | -0.092 | 0.856 |
| 2/3 | 1.7573 | 0.041 | 42.733 | 0.000 | 1.677 | 1.838 |
| 3/4 | 1.7303 | 0.011 | 154.114 | 0.000 | 1.708 | 1.752 |

<Ordered Logit Regression 결과>

Train Accuracy: 0.919
Test Accuracy: 0.917

# Question

## 카이검정 및 불균형 처리를 안해도 모델링이 된다. 굳이 해야하나?

**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?

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| 출생년도 | 0.0070 | 0.012 | 0.566 | 0.571 | -0.017 | 0.031 |
| 성별 | -0.0868 | 0.060 | -1.450 | 0.147 | -0.204 | 0.031 |
| 문항1 | 0.0319 | 0.067 | 0.474 | 0.635 | -0.100 | 0.164 |
| 문항2 | 0.6260 | 0.077 | 8.122 | 0.000 | 0.475 | 0.777 |
| 문항3 | 0.7042 | 0.038 | 18.526 | 0.000 | 0.630 | 0.779 |
| 문항4 | 0.2049 | 0.066 | 3.107 | 0.002 | 0.076 | 0.334 |
| 문항5 | 0.1970 | 0.020 | 10.064 | 0.000 | 0.159 | 0.235 |
| 문항6 | 4.2164 | 0.043 | 98.817 | 0.000 | 4.133 | 4.300 |
| 0/1 | 7.2437 | 0.435 | 16.653 | 0.000 | 6.391 | 8.096 |
| 1/2 | 0.3819 | 0.242 | 1.579 | 0.114 | -0.092 | 0.856 |
| 2/3 | 1.7573 | 0.041 | 42.733 | 0.000 | 1.677 | 1.838 |
| 3/4 | 1.7303 | 0.011 | 154.114 | 0.000 | 1.708 | 1.752 |

<Ordered Logit Regression 결과>

Train Accuracy: 0.919
Test Accuracy: 0.917

# 1.4 Ordinal Regression

$$y_i^* = \mathbf{x}_i'\beta + u_i$$

$$y_i = j \quad \text{if} \quad \alpha_{j-1} < y_i^* \leq \alpha_j$$

$$p_{ij} = p(y_i = j) = p(\alpha_{j-1} < y_i^* \leq \alpha_j) = F(\alpha_j - \mathbf{x}_i'\beta) - F(\alpha_{j-1} - \mathbf{x}_i'\beta)$$

**Ordinal Logit Regression**

- For the ordered logit, $F$ is the logistic cdf $F(z) = e^z/(1+e^z)$.
- For ordered probit, F is the standard normal cdf.

**Ordinal Probit Regression**

**Marginal Effects**

# 3. MODEL

# 4. RESULT

## Research Result

- **1st goal: 중요한 독립변수를 찾아낸다.**

- 중요도가 낮은 변수를 제거하면 예측 성능이 크게 저하 한다.
- Model별 중요변수의 순위가 다소 다르다.

  분석: 모두 중요한 변수일것으로 예상되며 이를 정확히 설명하기 위해서는 데이터를 unmask하고 도메인 지식과 함께 보아야한다.

- **2nd goal: Imbalanced Data를 적절히 처리한다.**

- Oversampling 이론적 근거를 찾지 못했다.
- 여러 방법을 적용해 본 결과 예측 성능 저하로 이어졌다.

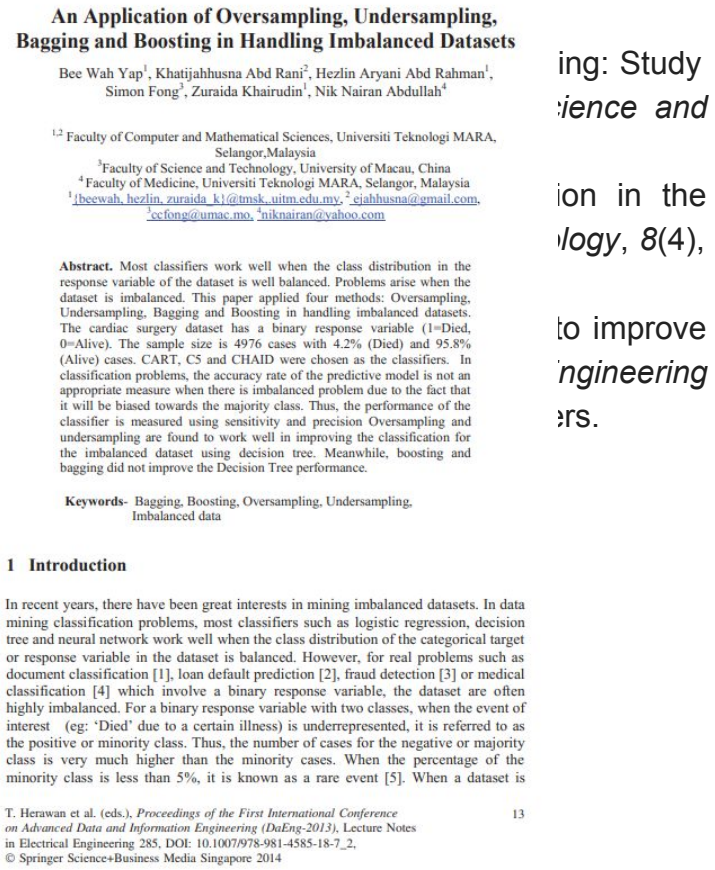
  분석: Oversampling에 적합하지 않은 데이터일 가능성이 농후하다.

## Further Study

- Unique Value 100개 이하 독립변수 40개에 대한 각각의 전처리
- PCA 와 PCA loading을 활용한 Modeling 및 Feature Importance 방법
- Oversampling의 Convexity 등의 이론적 가능성을 먼저 검토후 적절한 처리

# 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 pr… ing: Study on fault detection prediction model in semiconduct… *…ience and Technology Letters*, *133*, 79-84.
- Munirathinam, S., & Ramadoss, B. (2016). Predic… …ion in the semiconductor manufacturing process. *IACSIT Interna… …logy*, *8*(4), 273-285.
- Kerdprasop, K., & Kerdprasop, N. (2010, March). Fe… …to improve fault detection accuracy in the semiconductor manufa… *…ngineering 2012. July 4-6, 2012. London, UK.* (Vol. 2188, pp. 398… …ers.

**An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets**


Bee Wah Yap[1], Khatijahhusna Abd Rani[2], Hezlin Aryani Abd Rahman[1], Simon Fong[3], Zuraida Khairudin[1], Nik Nairan Abdullah[4]

[1,2] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Selangor,Malaysia
[3]Faculty of Science and Technology, University of Macau, China
[4] Faculty of Medicine, Universiti Teknologi MARA, Selangor, Malaysia
[1]{beewah, hezlin, zuraida_k}@tmsk.uitm.edu.my, [2]ejahhusna@gmail.com, [3]ccfong@umac.mo, [4]niknairan@yahoo.com



**Abstract.** Most classifiers work well when the class distribution in the response variable of the dataset is well balanced. Problems arise when the dataset is imbalanced. This paper applied four methods: Oversampling, Undersampling, Bagging and Boosting in handling imbalanced datasets. The cardiac surgery dataset has a binary response variable (1=Died, 0=Alive). The sample size is 4976 cases with 4.2% (Died) and 95.8% (Alive) cases. CART, C5 and CHAID were chosen as the classifiers. In classification problems, the accuracy rate of the predictive model is not an appropriate measure when there is imbalanced problem due to the fact that it will be biased towards the majority class. Thus, the performance of the classifier is measured using sensitivity and precision Oversampling and undersampling are found to work well in improving the classification for the imbalanced dataset using decision tree. Meanwhile, boosting and bagging did not improve the Decision Tree performance.



**Keywords-** Bagging, Boosting, Oversampling, Undersampling, Imbalanced data


**1 Introduction**

In recent years, there have been great interests in mining imbalanced datasets. In data mining classification problems, most classifiers such as logistic regression, decision tree and neural network work well when the class distribution of the categorical target or response variable in the dataset is balanced. However, for real problems such as document classification [1], loan default prediction [2], fraud detection [3] or medical classification [4] which involve a binary response variable, the dataset are often highly imbalanced. For a binary response variable with two classes, when the event of interest (eg: 'Died' due to a certain illness) is underrepresented, it is referred to as the positive or minority class. Thus, the number of cases for the negative or majority class is very much higher than the minority cases. When the percentage of the minority class is less than 5%, it is known as a rare event [5]. When a dataset is


T. Herawan et al. (eds.), *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, Lecture Notes in Electrical Engineering 285, DOI: 10.1007/978-981-4585-18-7_2,

# Q&A