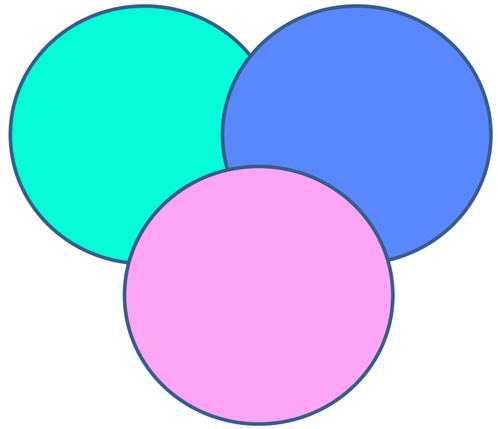
Presentation For KOREN Data



Bomi.Yeom

목차

목차

01 대회 개요

KOREN - Determination of Non-Disability

대회 목표

데이터 분석과 예측 모델 설계를 통해
이상 트래픽에 의해 발생하는 장애 티켓의 정확성을
판단하고, 장애 티켓 발행을 위한 인사이트를 도출

평가 기준

리더보드 점수 지표 : **정확도(Accuracy)**

01 대회 개요

Koren?

•소프트웨어 정의 네트워크, 클라우드, NFV로 결합된 통합 인프라로 산·학·연 등이 실험을 목적으로 무료로 이용할 수 있는 국내·외 선도기술 개발을 위한 통합연구시험망

•광대역, 고품질의 국내외 연구개발망을 산・학・연에 제공하여 미래네트워크 관련 기술의 시험검증을 지원함으로써 연구개발촉진 및 국제공동연구 협력기반을 조성하기 위한 비영리 통합연구시험망

•네트워크 인프라 연구개발 결과물의 사업화를 효율적으로 지원하는 ICT 연구개발 선순환생태계 기반 인프라

•초연결 지능형 연구개발망은 전국 10개 대도시 지역(서울, 수원, 판교, 대전, 전주, 광주, 대구, 부산, 제주, 춘천)을 10Gbps~660Gbps로 연결하는 백본망을 구축 운영

(단, 이 중 제주/춘천 접속점을 제외한 8개 지역에 백본 전송장비가 수용되어 있으며, 제주/춘천은 현재 10G 개통)

•해외접속점(TEIN홍콩, TEIN싱가폴)에서 국제연구망(TEIN)과 100Gbps로 직접 연동되어 GEANT/EU로 가는 연동점을 확보하고, 홍콩접속점에서 APAN-JP를 통해 Internet2/US로 가는 연동점 확보

ATT

02 데이터 개요

Feature 수(Column 수): 53개
데이터 개수(Row 수) : 340개

컬럼명	데이터 타입	결측치	결측비율 (%)	nunique 값
ticket_id	int	0	0	149
ticket_type	obj	0	0	1
strresnm	obj	0	0	15
stripaddr	float	340	100	0
strifname	obj	0	0	35
strifdesc	obj	0	0	28
strifspeed	int	0	0	4
striftype	obj	0	0	1
strifoperstatus	obj	0	0	1
nren_id	obj	139	40.9	39

컬럼명	데이터 타입	결측치	결측비율 (%)	nunique 값
nren_name	obj	139	40.9	39
node_id	obj	139	40.9	7
if_id	obj	139	40.9	17
strifid	float	0	0	13
strresid	float	0	0	3
strtypemin	int	0	0	1
inttimestamp	int	0	0	149
intyear	int	0	0	1
intmonth	int	0	0	4
intday	int	0	0	27

- 다수의 결측치
- Target 값
- 고유값 1개

02 데이터 개요

컬럼명	데이터 타입	결측치	결측비율(%)	nunique값
inthour	int	0	0	18
intmin	int	0	0	53
intweek	int	0	0	7
intbandwidth	int	0	0	1
fltbpsin	int	0	0	93
fltbpsout	int	0	0	94
fltppsin	int	0	0	200
fltppsout	int	0	0	204
flterrorin	int	0	0	1
flterrorout	int	0	0	1

컬럼명	데이터 타입	결측치	결측비율(%)	nunique값
fltdiscardin	int	0	0	1
fltdiscardout	int	0	0	1
fltunknown	int	0	0	1
fltusage	int	0	0	1
flhcinoctets	int	0	0	1
flhcoutoctets	int	0	0	1
flhcinucastpkts	int	0	0	1
flhcoutucastpkts	int	0	0	1
flhcinmulticastpkts	int	0	0	1
flhcoutmulticastpkts	int	0	0	1

- 다수의 결측치
- Target 값
- 고유값 1개

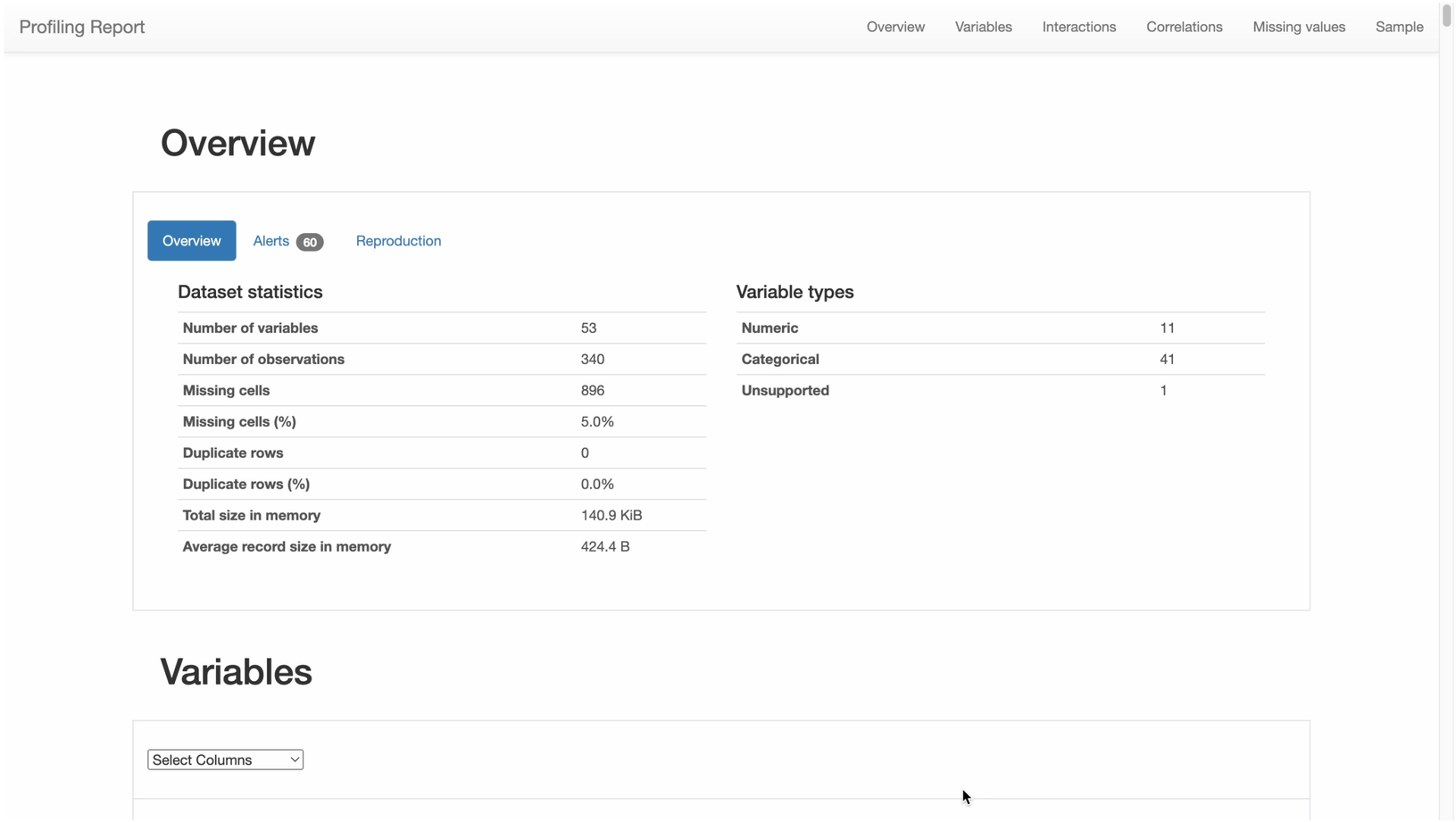
02 데이터 개요

컬럼명	데이터 타입	결측치	결측비율(%)	nunique값
flhcinbroadcastpkts	int	0	0	1
flhcoutbroadcastpkts	int	0	0	1
fltbpsinmax	int	0	0	1
fltbpsoutmax	int	0	0	1
fltppsinmax	int	0	0	1
fltppsoutmax	Int	0	0	1
flterrorinmax	int	0	0	1
flterroroutmax	int	0	0	1
fltdiscardinmax	int	0	0	1
fltdiscardoutmax	int	0	0	1

컬럼명	데이터 타입	결측치	결측비율(%)	nunique값
fltunknownmax	int	0	0	1
fltusagemax	int	0	0	1
ai_accuracy	int	0	0	2

- 다수의 결측치
- Target 값
- 고유값 1개

03 데이터 EDA


Profiling Report
Overview
Variables
Interactions
Correlations
Missing values
Sample
Overview
Overview
Alerts 60
Reproduction
Dataset statistics
Number of variables 53
Number of observations 340
Missing cells 896
Missing cells (%) 5.0%
Duplicate rows 0
Duplicate rows (%) 0.0%
Total size in memory 140.9 KiB
Average record size in memory 424.4 B
Variable types
Numeric 11
Categorical 41
Unsupported 1
Variables
Select Columns

04 데이터 CDA

범주형 데이터

Target : y 에 대한 독립성 검증

P-value < 0.05 독립성을 띄지 않는다

P-value >= 0.05 독립성을 띈다

컬럼명	P.Value	독립성 여부
Ticket_id	3.826700500173644e-17	독립X
Ticket_type	1.0	독립
strresnm	0.9935184358994206	독립
strifname	6.30820797820993e-05	독립X
strifdesc	0.0022770978518551954	독립X
striftype	1.0	독립
strifoperstatus	1.0	독립
nren_id	0.5388332725487257	독립
nren_name	0.5388332725487244	독립
node_id	0.9449002728738715	독립
lf_id	0.012272741427636602	독립X
strifid	0.055106585357850874	독립
strresid	0.13425760079306703	독립

- Y와 독립성을 가지지 않는 컬럼

04 데이터 CDA

연속형 데이터

Target : y 에 대한 상관관계 검증

P-value < 0.05 **상관관계가 있다**

P-value >= 0.05 **상관관계가 없다**

컬럼명	P.Value	상관관계 여부
strifspeed	0.7320662565633638	상관관계X
inttimestamp	0.20341788992828785	상관관계X
intmonth	0.6714674849133265	상관관계X
intday	0.0003716894503765013	상관관계O
inthour	0.6418345916782323	상관관계X
intmin	0.17610699705332994	상관관계X
intweek	0.39389127417271486	상관관계X
fltbspin	0.8643734721230046	상관관계X
fltbpsout	0.8643734721230046	상관관계X
fltppsin	0.24635081848188756	상관관계X
fltppsout	0.9492740040052362	상관관계X

- Y와 상관관계를 가지는 컬럼

04 데이터 CDA

- **Inttimestamp**
- **Intmonth**
- **Intday**
- **Inthour**
- **Intmin**
- **intweek**

inttimestamp 에 대한 상관관계 검증

P-value < 0.05 **상관관계가 있다**

P-value >= 0.05 **상관관계가 없다**

컬럼명	P.Value	상관관계 여부
intmonth	1.3670346197099758e-134	상관관계O
intday	5.597721790011813e-05	상관관계O
inthour	0.29768693266680424	상관관계X
intmin	0.18304691485299784	상관관계X
intweek	0.106110511910448	상관관계X

- Y와 상관관계를 가지는 컬럼

05 데이터 전처리

01 →

Column 선별

UNIQUE 값 1개

- Ticket_type
- Striftype
- Strifoperstatus
- Strtypemin
- Intyear
- Strtypemin
- Flterrorin
- Flterrorout
- Fltdiscardin
- Fltdiscardout
- Fltunknown
- Fltusage
- Flhcinoctets
- Flhcoutoctets
- Flhcinucastpkts
- Flhcinucastpkts
- Flhcinmulticastpkts
- Flhcoutmulticastpkts
- Flhcinbroadcastpkts
- Flhcoutbroadcastpkts
- Fltbpsinmax
- Fltbpsoutmax
- Fltppsinmax
- Fltppsoutmax
- Flterrorinmax
- Flterroroutmax
- Fltdiscardinmax
- Fltdiscardoutmax
- Fltunknownmax
- fltusagemax

02 →

Column 선별

Id 값

- ticket id

Null 값 다수

- Stripaddr
- Nren_id
- Nren_name
- Node_id
- if_id

03 →

시간데이터 처리

- Inttimestamp
- Intmonth
- Intday
- Inthour
- Intmin
- intweek

위 컬럼중,
Inttimestamp와 intmonth, intday는
상관 관계가 있다고 판단되어 제거
Inthour, intmin, Intweek는
상관관계가 없다고 판단되어 남김

06 모델링

01 → 02 → 03 →

Train & Valid 데이터로 분리

Train : Valid

8 : 2

AutoML
Pycaret으로 모델 비교

	Model	Accuracy
dt	Decision Tree Classifier	0.9243
rf	Random Forest Classifier	0.9243
ada	Ada Boost Classifier	0.9243

모델 하이퍼파라미터

Decision Tree Classifier	
ccp_alpha	0.0
class_weight	None
criterion	'gini'
max_depth	None
max_features	None
max_leaf_nodes	None
min_impurity_decrease	0.0
min_samples_leaf	1
min_samples_split	2
min_weight_fraction_leaf	0.0
random_state	10
splitter	'best'

TEST DATA에 대한 결과

Accuracy 0.9069767441860465

NTT

07 데이터 개요

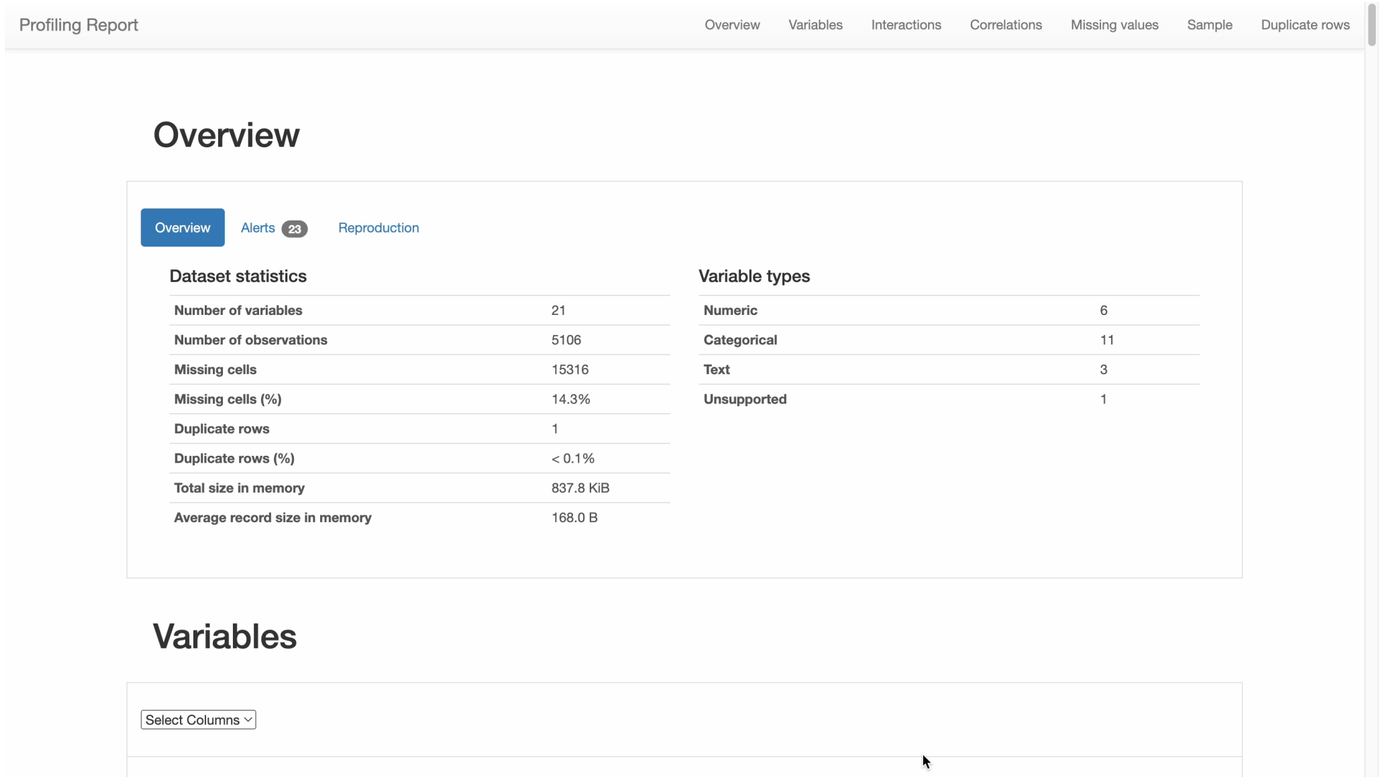
Feature 수(Column 수): 21개
데이터 개수(Row 수) : 5106개

컬럼명	데이터 타입	결측치	결측비율 (%)	nunique 값
ticket_id	int	0	0	5904
ticket_type	obj	0	0	1
strresnm	float	5105	99.9	1
strresip	obj	0	0	6
strresname	obj	0	0	6
strs_ip	obj	0	0	791
strs_port	int	0	0	2496
strd_ip	obj	0	0	985
strd_port	int	0	0	2235
strs_mac	obj	0	0	34

컬럼명	데이터 타입	결측치	결측비율 (%)	nunique 값
Strd_mac	obj	0	0	22
strprotocol	int	0	0	2
stripv4tos	int	0	0	16
strchannel	obj	0	0	14
strsenderip	obj	0	0	6
strin_interface	obj	5105	99.9	1
strout_interface	float	5106	100	0
Strbytes_col	int	0	0	1034
strcounts	int	0	0	481
dateregdate	obj	0	0	48
ai_accuracy	int	0	0	2

- 다수의 결측치
- Target 값
- 고유값 1개

08 데이터 EDA


Profiling Report
Overview
Variables
Interactions
Correlations
Missing values
Sample
Duplicate rows
Overview
Overview
Alerts 23
Reproduction
Dataset statistics
Number of variables 21
Number of observations 5106
Missing cells 15316
Missing cells (%) 14.3%
Duplicate rows 1
Duplicate rows (%) < 0.1%
Total size in memory 837.8 KiB
Average record size in memory 168.0 B
Variable types
Numeric 6
Categorical 11
Text 3
Unsupported 1
Variables
Select Columns

09 데이터 CDA

범주형 데이터

Target : y 에 대한 독립성 검증

P-value < 0.05 독립성을 띄지 않는다

P-value >= 0.05 독립성을 띈다

컬럼명	P.Value	독립성 여부
Ticket_id	0.4461864122465706	독립
Ticket_type	1.0	독립
strresip	3.6583526014256788e-270	독립X
strresname	3.6583526014256788e-270	독립X
strs_ip	1.5973812833681765e-139	독립X
strd_ip	2.6773996565107496e-123	독립X
strs_mac	3.483489377996286e-137	독립X
strd_mac	1.9412063432134064e-266	독립X
strchannel	1.9286577616655264e-65	독립X
strsenderip	3.6583526014256788e-270	독립X
strin_interface	1.0	독립
dateregdate	0.0	독립X

- Y와 독립성을 가지지 않는 컬럼

09 데이터 CDA

연속형 데이터

Target : y 에 대한 상관관계 검증

P-value < 0.05 **상관관계가 있다**

P-value >= 0.05 **상관관계가 없다**

컬럼명	P.Value	상관관계 여부
Strs_port	4.5457524088463747e-26	상관관계 O
Strd_port	1.0922831407810835e-24	상관관계 O
strprotocol	0.0022207917102954875	상관관계 O
stripv4tos	0.1794105067325901	상관관계 X
Strbytes_col	2.898934406166783e-100	상관관계 O
strcounts	5.008711805170854e-91	상관관계 O

- Y와 상관관계를 가지는 컬럼

10 데이터 전처리

01 →

Column 선별

UNIQUE 값 1개

- Ticket_type
- Strresnum
- Strin_interface

Null 값 다수

- Strout_interface

02 →

Column 선별

상관관계 있는 컬럼

- Strs_port
- Strd_port
- Strbytes_col
- Stripv4tos

독립적이지 않은 컬럼

- dateregdate

03 →

Column 변환

Strs_mac	smac1
	smac2
	smac3
	smac4
	smac5
	smac6

Strd_mac	dmac1
	dmac2
	dmac3
	dmac4
	dmac5
	dmac6

11 모델링

01 →

Train & Valid 데이터로 분리

Train : Valid

8 : 2

02 →

AutoML
Pycaret으로 모델 비교

	Model	Accuracy
gbc	Gradient Boosting Classifier	0.9698
dt	Decision Tree Classifier	0.9684
knn	K Neighbors Classifier	0.9653

Gradient Boosting Classifier			
ccp_alpha	0.0	min_weight_fraction_leaf	0.0
criterion	‘friedman_mse’	n_estimators	100
init	None	n_iter_no_change	None
Learning_rate	0.3	random_state	3751
loss	Log_loss	subsample	1.0
max_depth	3	tol	0.0001
max_features	None	Validation_fraction	0.1
max_leaf_nodes	None	verbose	0
min_impurity_decrease	0.0	Warm_start	False
min_samples_split	1		

TEST DATA에 대한 결과

Accuracy 0.9741581832419733

12 하이퍼파라미터 튜닝

01 → 02 →

Train & Valid 데이터로 분리

Train : Valid

8 : 2

AutoML
Pycaret으로 모델 비교

	Model	Accuracy
gbc	Gradient Boosting Classifier	0.9698
dt	Decision Tree Classifier	0.9684
knn	K Neighbors Classifier	0.9653

Gradient Boosting Classifier			
ccp_alpha	0.0	min_weight_fraction_leaf	0.0
criterion	'friedman_mse'	n_estimators	90
init	None	n_iter_no_change	None
Learning_rate	0.3	random_state	3751
loss	Log_loss	subsample	1.0
max_depth	3	tol	0.0001
max_features	1.0	Validation_fraction	0.1
max_leaf_nodes	None	verbose	0
min_impurity_decrease	0.0002	Warm_start	False
min_samples_split	4		

TEST DATA에 대한 결과

Accuracy 0.9741581832419733

12 결론 & 향후 개선사항

01 ATT 모델은 학습시킬 때보다 test data를 예측할 때 accuracy 값이 낮았음

02 NTT 모델에서 ticket_id를 제외하고 진행하였을 때의 모델의 수치가 0.2 정도 낮아지는 것을 보았음
-> ticket_id로 데이터의 중복을 줄일 수 있음(중복값을 구별할 수 있어서)

03 두가지 모델 수치를 조금 더 개선할 수 있도록 전처리 방법 수정

감사합니다