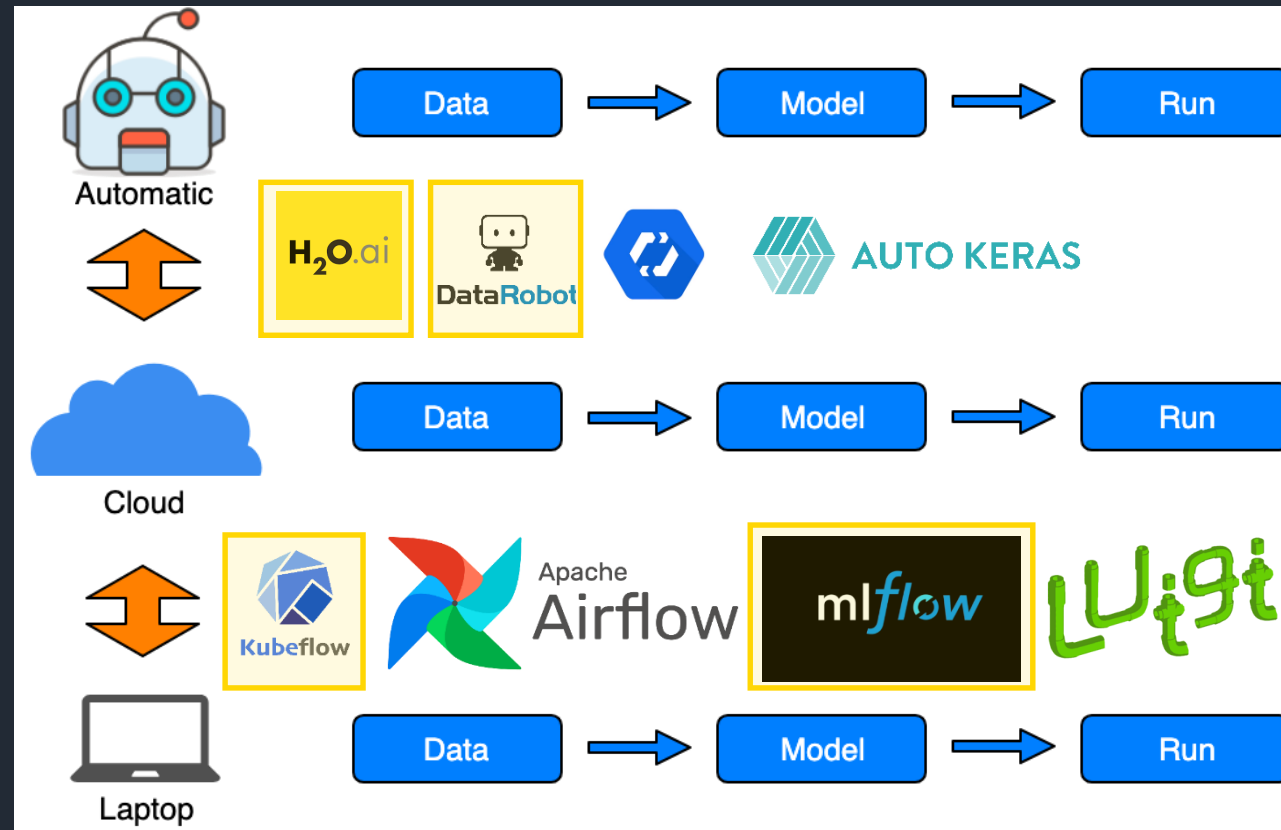
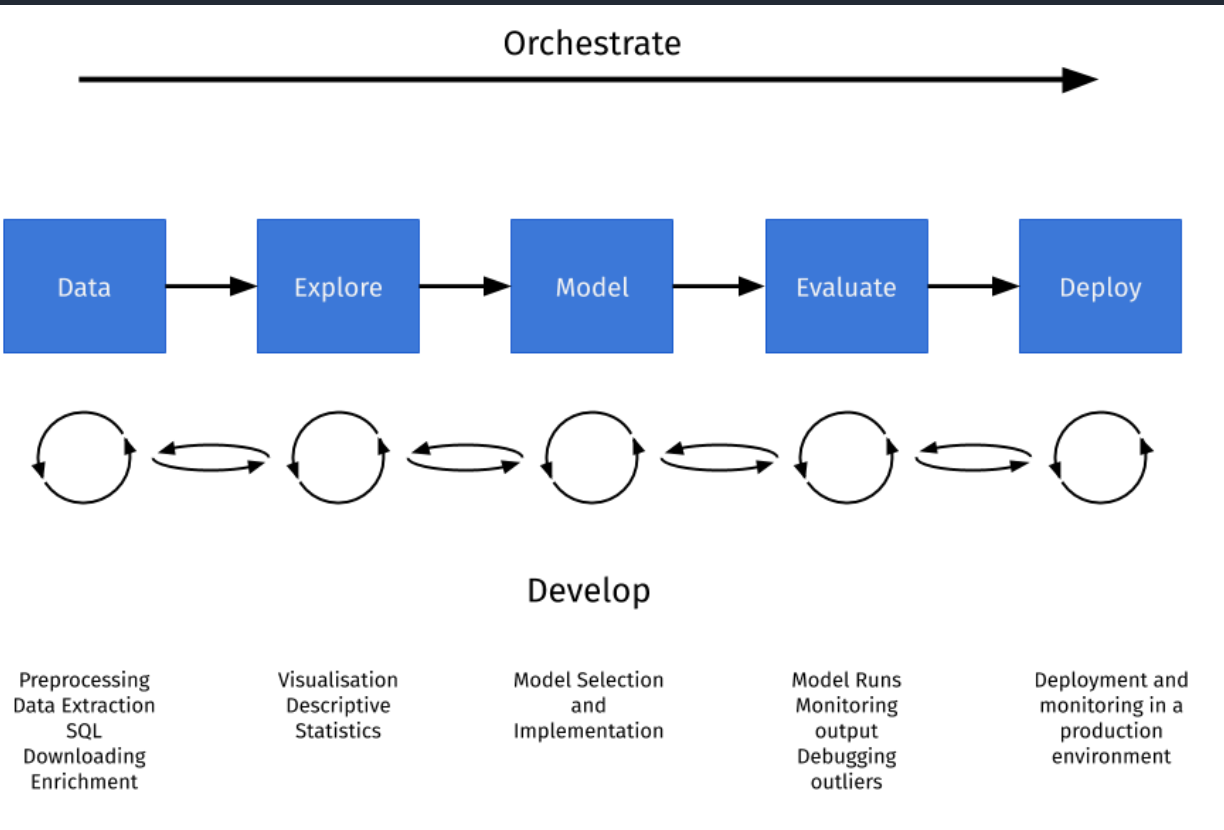


OMELET V2

- A ML Platform with AutoML / XAI / Model Mgmt.-

Presented by Eunjong Kim

Market Mega Trends



Market Mega Tree

1. Orchestrator (pipeline)

mlflow

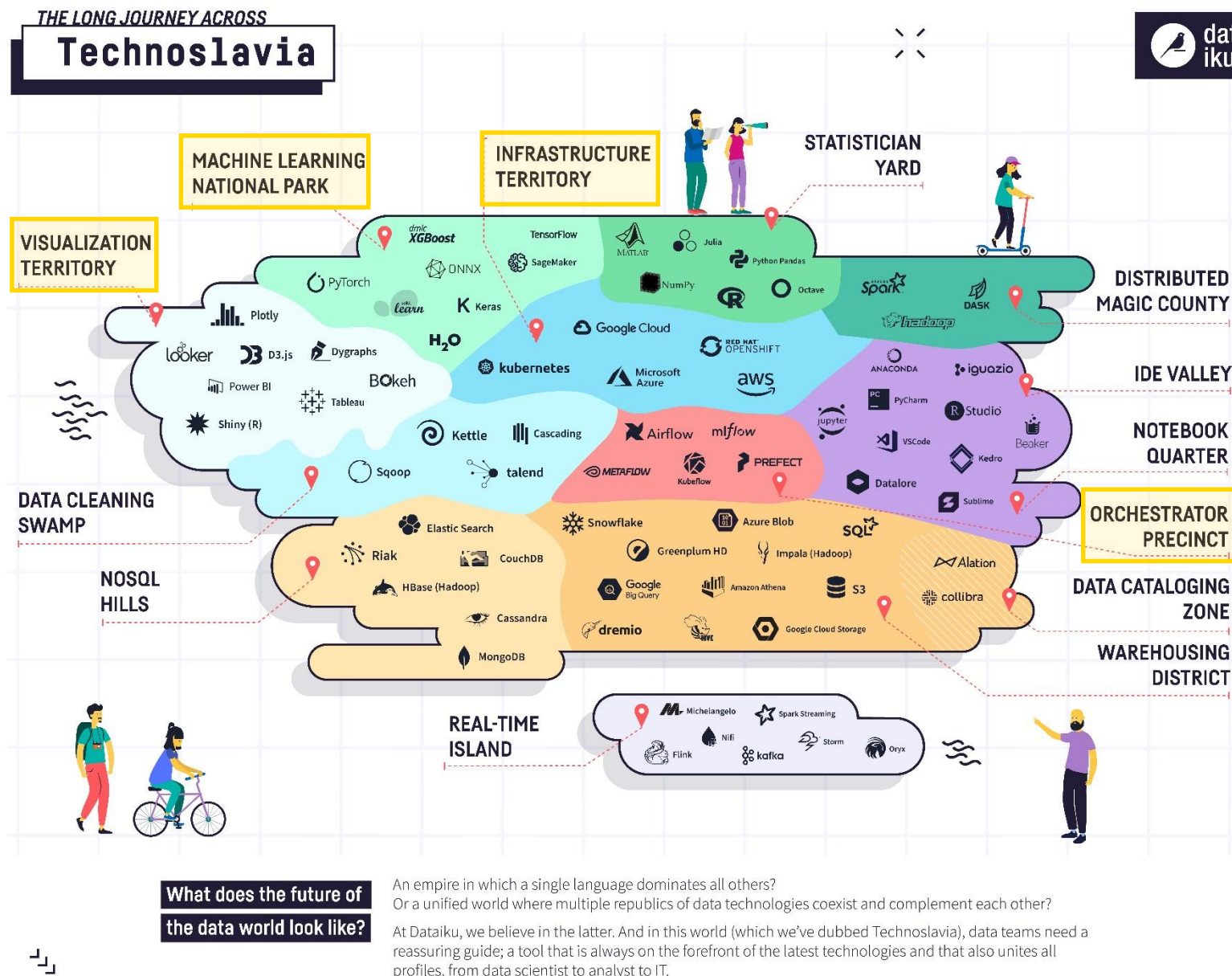
2. Automated ML

H₂O.ai

3. XAI (SHAP)

Shap

OMELET



Market Mega Trends

V3에서 봐



What is OMELET?

OMELET

mlflow

Shap

H₂O.ai

Omelet Service (**A**utomated **M**achine **L**earning **LET**)

누구나 쉽게 머신러닝을 적용할 수 있는 환경 제공

OMELET



Basic Concepts

Concept 1.

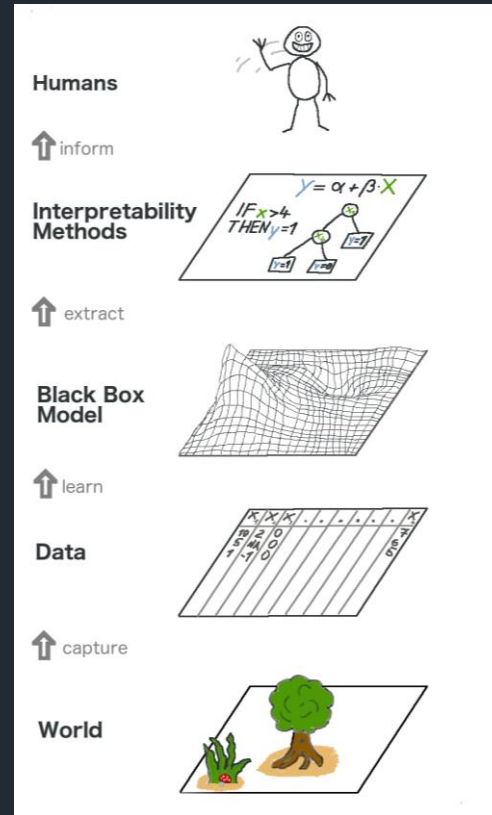
Simple is better than complex.
Complex is better than complicated.
Flat is better than nested.
<The Zen of Python>

Concept 2.

고객이 이해할 수 있는 언어로 설명하라
<대전에 계신 어느 분>

Concept 3.

Easy to learn and Difficult to master
<Nolan's Law>



Core Ideas

직관적인 편의성 제공

“같은 현상을 설명하는 두 개의 주장이 있다면, 간단한 쪽을 선택하라” <오컴의 면도날>

모델 관리의 용이성

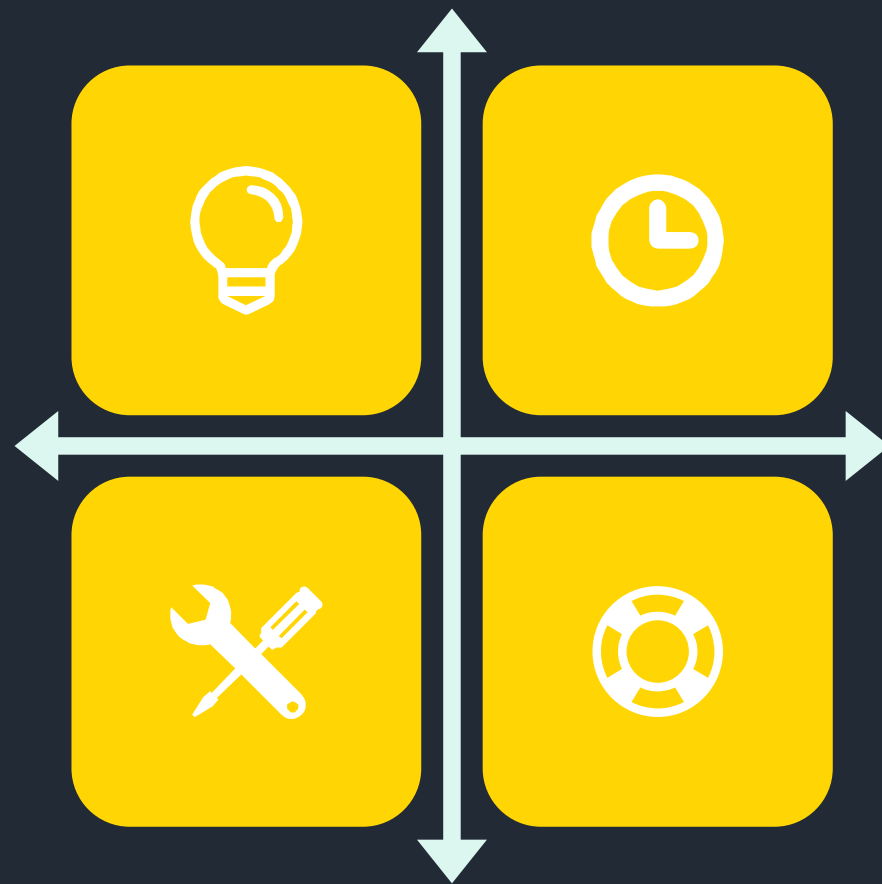
모델 실행 / 모델 비교 /
파라미터 저장 / 결과 저장 /
모델 배포 등 모델 관리에
필요한 모든 기능 통합

시간 절약

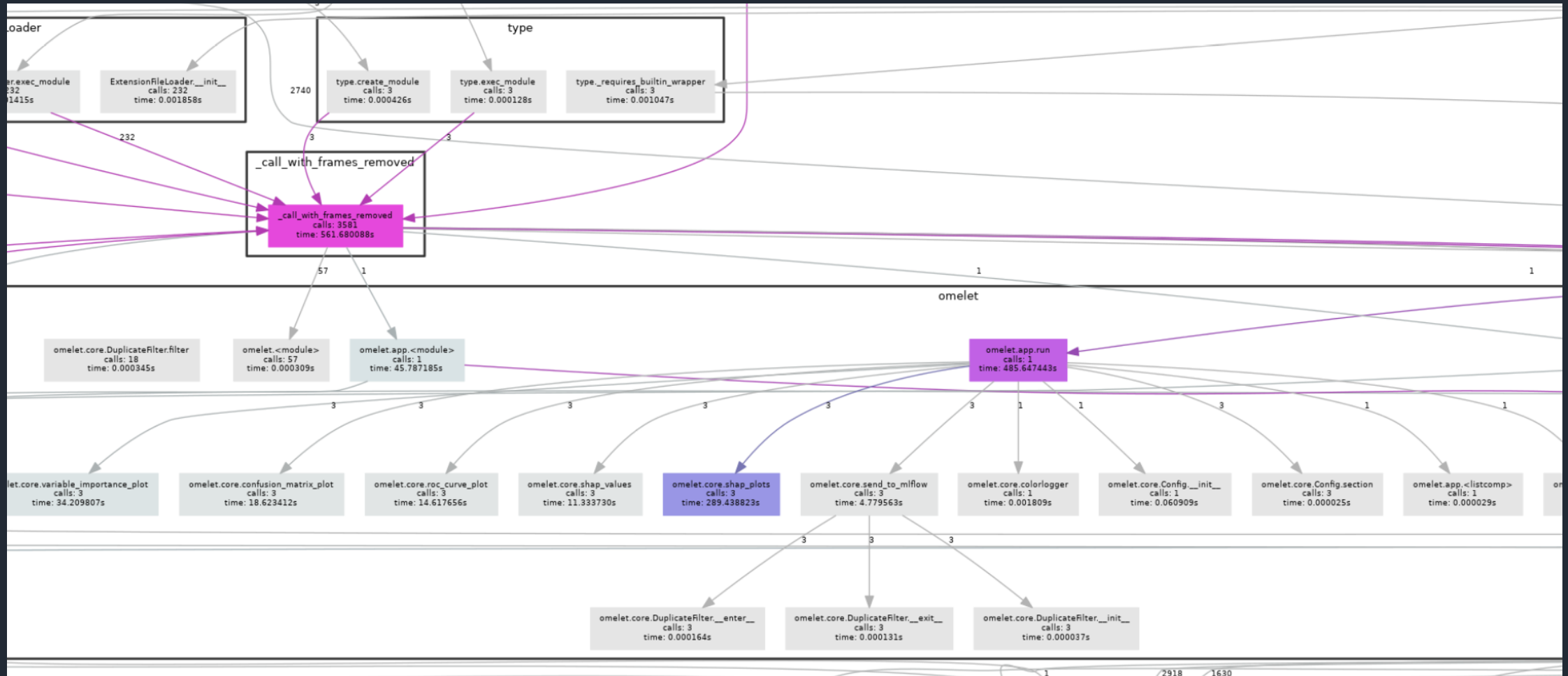
플랫폼 설치 시간 단축 /
프로필 설정 기능을 통해
One Click 실행

모델 배포 및 관리

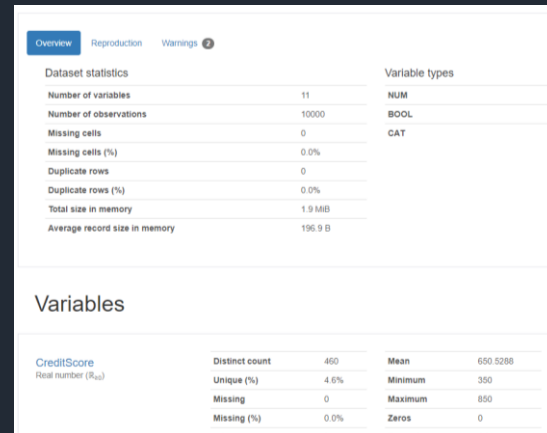
모델 파일을 통해
REST API를 제공하여
새로운 데이터에 대한
예측 결과 제공



OMELET Core



OMELET Features



Overview | Reproduction | Warnings 2

Dataset statistics

Number of variables	11
Number of observations	10000
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.9 MB
Average record size in memory	195.9 B

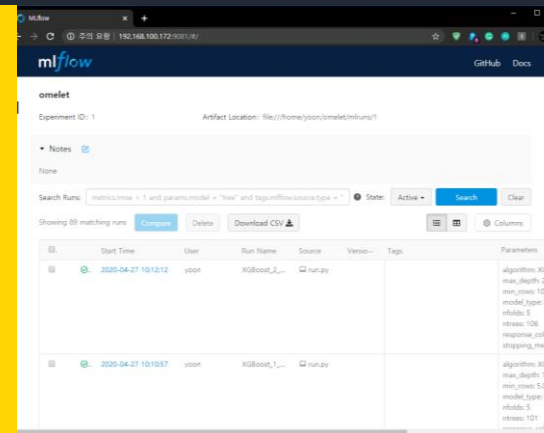
Variable types

NUM
BOOL
CAT

Variables

CreditScore	Distinct count	460	Mean	650.5288
Real number (R ₆₄)	Unique (%)	4.6%	Minimum	350
	Missing	0	Maximum	850
	Missing (%)	0.0%	Zeros	0
	Infinite	0	NaNs (%)	0.0%

Automated ML



omelet

Experiment ID: 1

Artifact Location: file:///home/you/omelet/runs/1

Notes

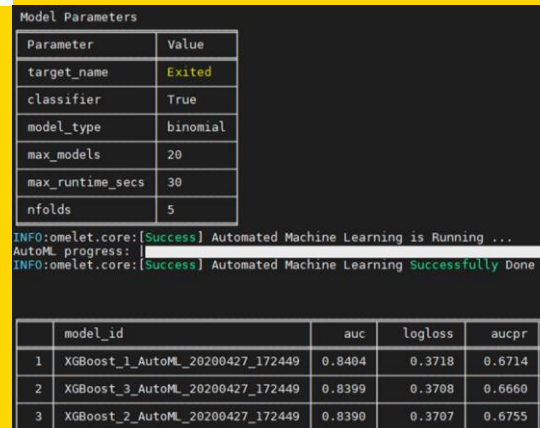
Search Runs: metricscore = 1 and parammodel = "True" and loglossscoretype = 1

Showing 89 matching runs

Run ID	Start Time	User	Run Name	Source	Version	Tags	Parameters
2020-04-27 10:12:12	you	XGBoost_2_...	runpy				algorithm: XGBoost, max_depth: 20, min_rows: 100, model_type: binomial, n_folds: 5, n_jobs: 100, response_column: shipping_metrics, shipping_metrics: 1
2020-04-27 10:10:57	you	XGBoost_1_...	runpy				algorithm: XGBoost, max_depth: 10, min_rows: 50, model_type: binomial, n_folds: 5, n_jobs: 100, response_column: shipping_metrics, shipping_metrics: 1

Model Serving

Data Profiling



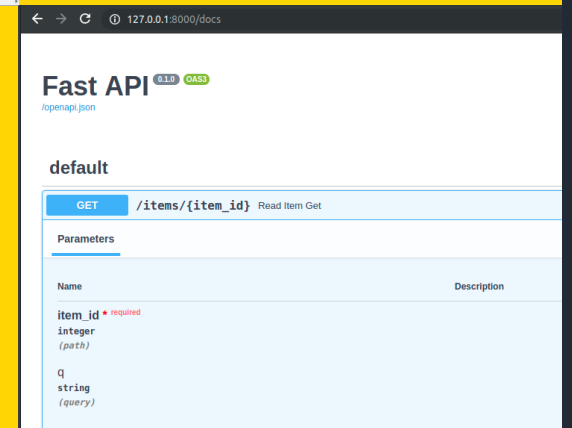
Model Parameters

Parameter	Value
target_name	Exited
classifier	True
model_type	binomial
max_models	20
max_runtime_secs	30
n_folds	5

INFO:omelet.core:[Success] Automated Machine Learning is Running ...
AutoML progress: [Progress Bar]
INFO:omelet.core:[Success] Automated Machine Learning Successfully Done !

	model_id	auc	logloss	aucpr
1	XGBoost_1_AutoML_20200427_172449	0.8404	0.3718	0.6714
2	XGBoost_3_AutoML_20200427_172449	0.8399	0.3708	0.6660
3	XGBoost_2_AutoML_20200427_172449	0.8390	0.3707	0.6755

Model Management



Fast API 0.10.0

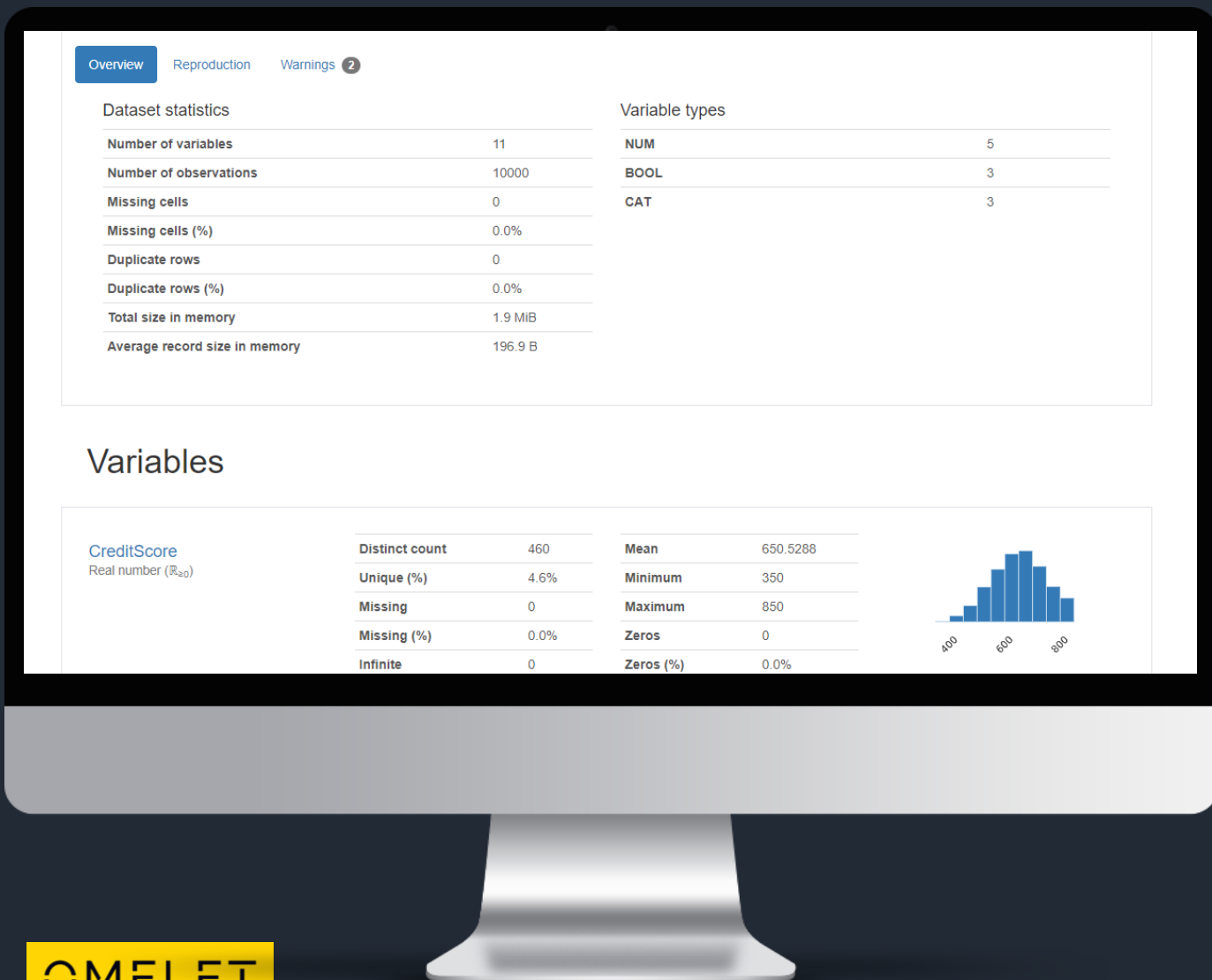
default

GET /items/{item_id} Read Item Get

Parameters

Name	Description
item_id	integer (path)
q	string (query)

Data Profiling



1. 변수별 통계량 제공

수치형 변수일 경우 통계량
범주형 변수일 경우 빈도수

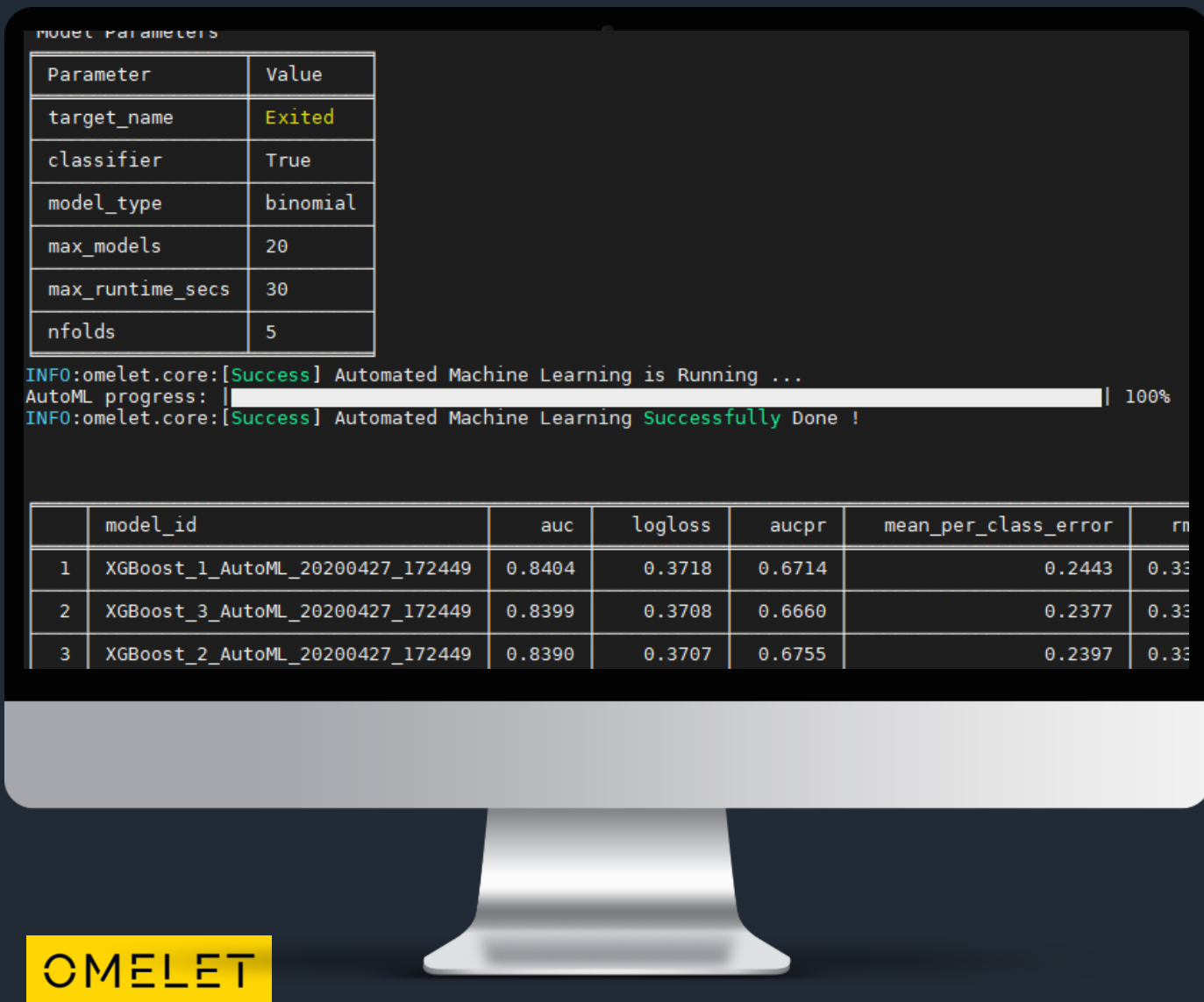
2. 통계량 관련 히스토그램

“TOGGLE DETAIL” 버튼 클릭 시
변수별 히스토그램, Outlier 확인 가능

3. 이상치 데이터에 대한 알림

상단의 “WARNINGS” 탭에서 변수별
왜도/첨도 혹은 Missing Value와 같은
이상치 데이터에 대한 알림 확인 가능

Automated ML



1. 다양한 머신러닝 알고리즘 지원

XGBOOST / GBM / XRT / DRF
머신러닝 알고리즘 지원

2. 유저 프로필 기능

유저/그룹 별 프로필 관리 기능을 통해
미리 설정해둔 값을 저장 및 로드

3. One Click 실행

데이터 및 프로필 설정만 하면
누구나 손쉽게 사용하도록
사용자 친화적 인터페이스 설계

Model Management

1. 검색기능 지원

알고리즘 및 파라미터 조건 검색 지원
원하는 모델 빠르게 검색 가능

2. 모델 Metric 비교기능

모델 Metric 비교를 통해
최적의 Best Model을 선택가능

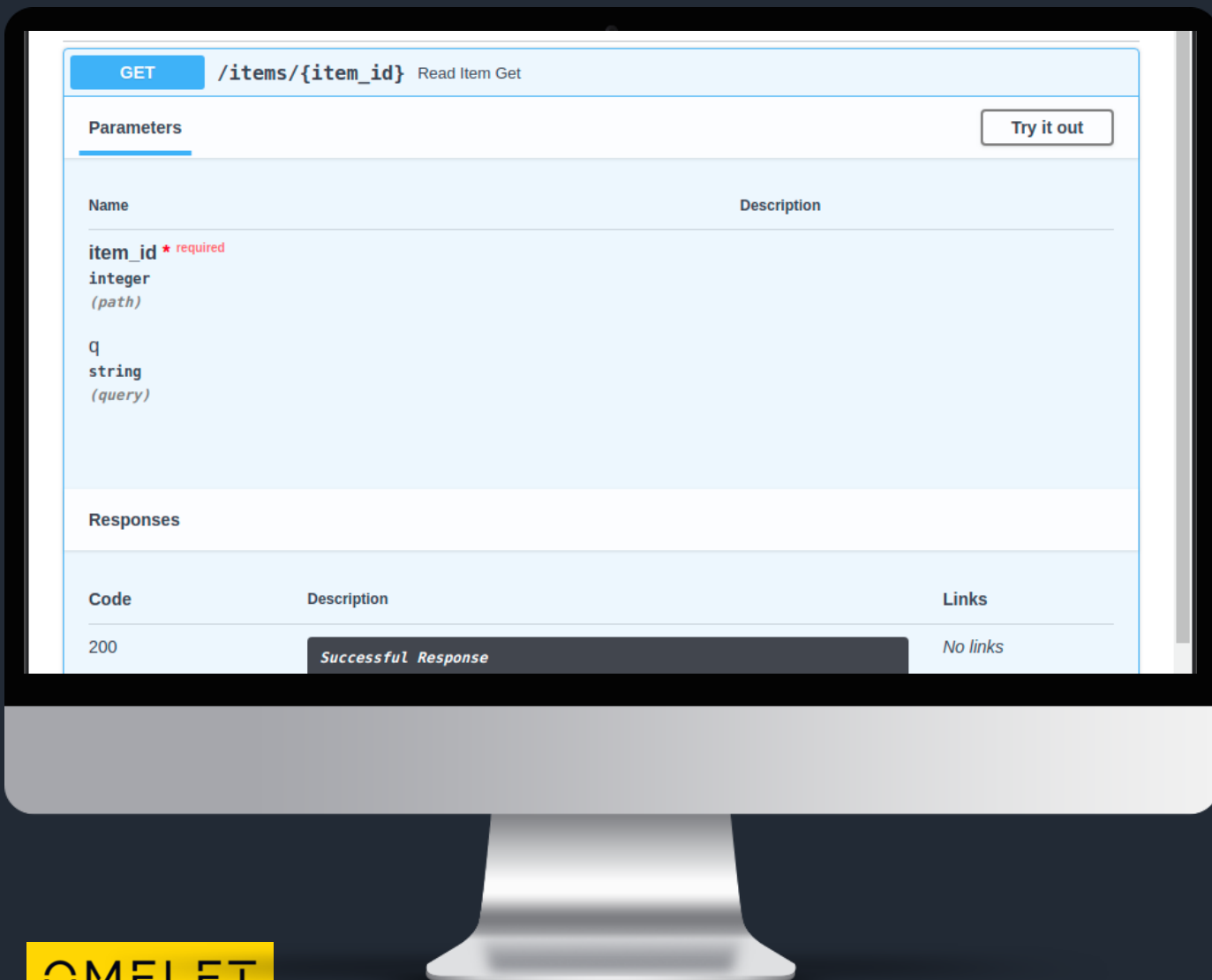
3. 모델 및 결과파일 다운로드

사용자들이 웹페이지에서 결과파일들을
쉽게 다운로드 할 수 있도록 저장기능 지원

The screenshot shows the mlflow Omelet web interface. The left sidebar has a search bar with 'omelet' entered. The main area shows 'Experiment ID: 1' and 'Artifact Location: file:///home/yoan/omelet/mlruns/1'. Below this, there's a search bar with the query 'metrics.rmse < 1 and params.model = "tree" and tags.mlflow.source.type = "LOCAL"'. The search results show 92 matching runs. A table displays the first three runs with columns for Start Time, User, Run Name, Source, Version, Tags, Parameters, and Metrics.

Start Time	User	Run Name	Source	Version	Tags	Parameters	Metrics
2020-04-27 17:29:06	yoan	XGBoost_2_A...	run.py			algorithm: XGBoost max_depth: 20 min_rows: 10.0 model_type: binomial nfold: 5 ntrees: 108 response_column: Exited stopping_metric: logloss	auc: 0.839 logloss: 0.371 mean_per_class_... 0.23 mse: 0.113 r2: 0.113 rmse: 0.336
2020-04-27 17:27:52	yoan	XGBoost_3_A...	run.py			algorithm: XGBoost max_depth: 5 min_rows: 3.0 model_type: binomial nfold: 5 ntrees: 76 response_column: Exited stopping_metric: logloss	auc: 0.84 logloss: 0.371 mean_per_class_... 0.226 mse: 0.114 r2: 0.114 rmse: 0.337
2020-04-27 17:26:39	yoan	XGBoost_1_A...	run.py			algorithm: XGBoost max_depth: 10 min_rows: 5.0 model_type: binomial nfold: 5 ntrees: 103 response_column: Exited stopping_metric: AUTO	auc: 0.84 logloss: 0.372 mean_per_class_... 0.229 mse: 0.114 r2: 0.114 rmse: 0.338

Model Serving



1. 모델 배포 기능

최적 모델을 배포하는 기능을 통해
사용자들이 손쉽게 서비스에 활용할
수 있는 모델 배포 기능 제공

2. 모델 결과 예측 API 제공

새로운 데이터 입력 시 모델 결과를
즉각적으로 예측하여 받아볼 수 있게
API 제공

More Details...



유저 프로필 관리



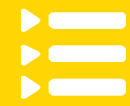
모델 테깅



모델 설정



모델 저장



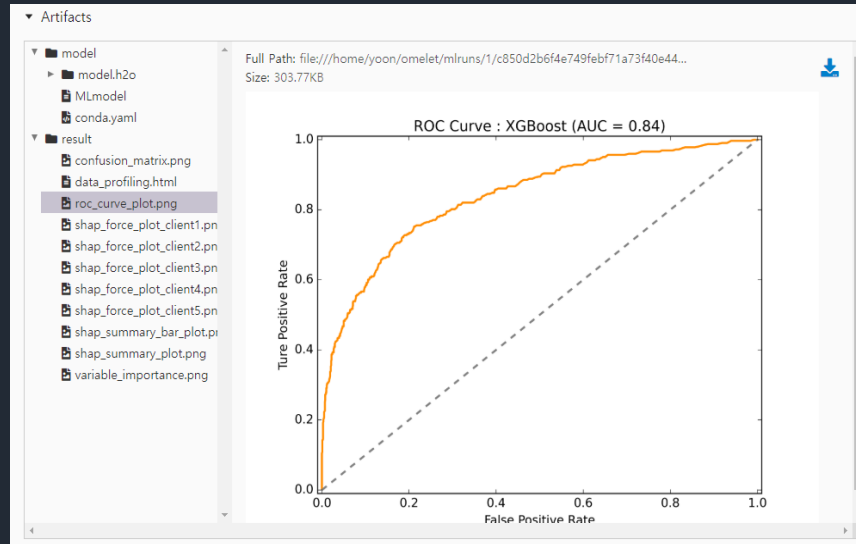
로그 관리



Automated Machine Learning + XAI + Model Management

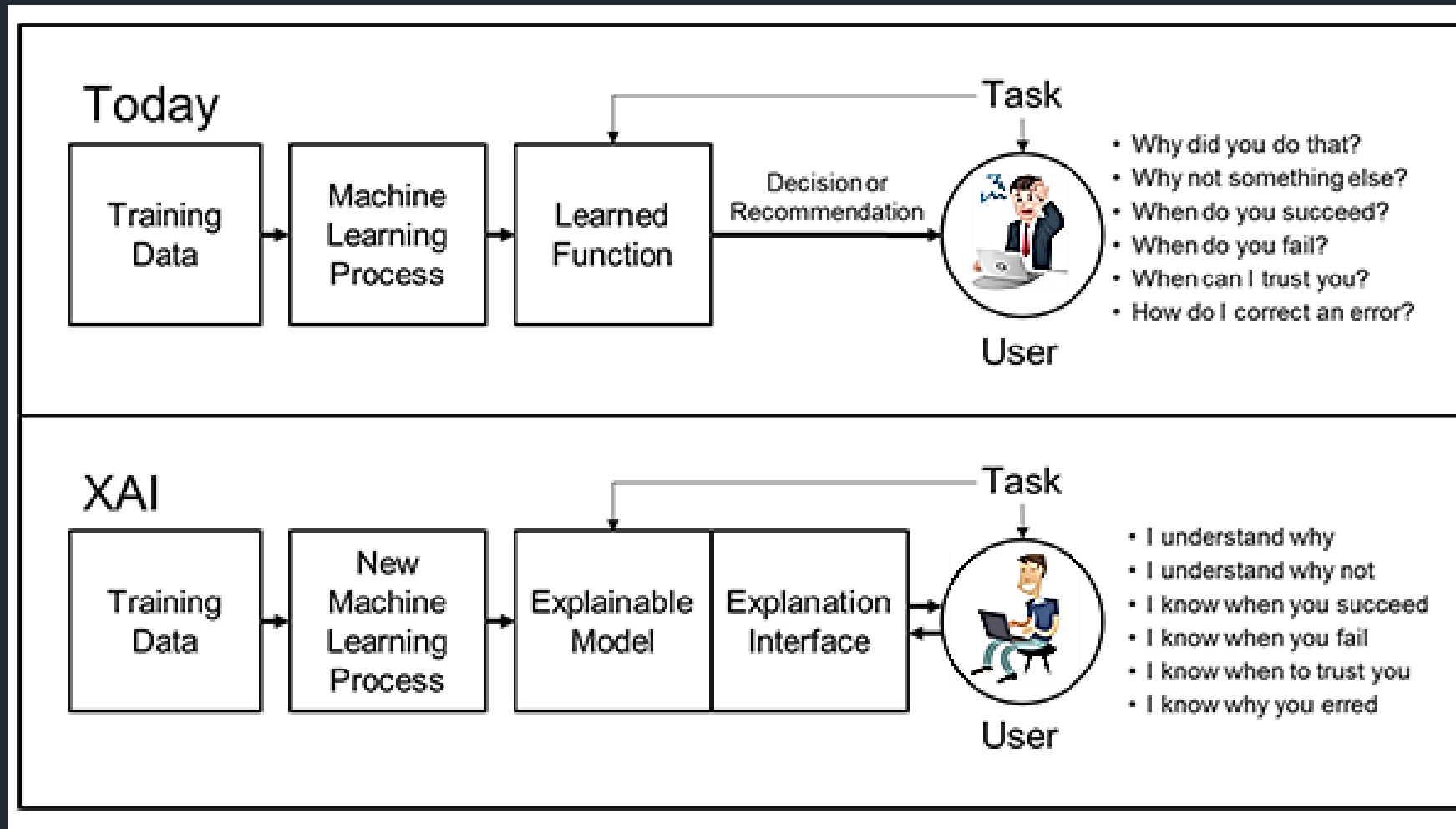
usage: run.py [-h] -f DATA [-s SAMPLE] [-c] [-p PROFILE] [-d]

optional arguments:
-h, --help show this help message and exit
-f DATA, --file DATA data only supports csv type [required]
-s SAMPLE, --sample SAMPLE sample number
-c, --clf turn on if classification
-p PROFILE, --profile PROFILE set user profile
-d, --disabled data profiling off



```
3. pb3.penta.co.kr
Re-attach Fullscreen Stay on top Duplicate
2020-04-27 17:16:18,945 INFO [Success] Data Load Successfully Done
2020-04-27 17:16:20,938 INFO [Success] Data Profiling Successfully Done
2020-04-27 17:16:29,694 INFO [Success] Get Model Type Successfully Done
2020-04-27 17:16:30,360 INFO [Success] Data Split Successfully Done
2020-04-27 17:16:30,362 INFO [Success] Automated Machine Learning is Running ...
2020-04-27 17:17:02,639 INFO [Success] Automated Machine Learning Successfully Done !
2020-04-27 17:17:02,840 INFO [XGBoost_3 AutoML_20200427_171630] Running | Saving Result ...
2020-04-27 17:17:48,981 INFO [Success] H2O / mlflow Client Connected.
2020-04-27 17:17:50,500 INFO [Success] Data Load Successfully Done
2020-04-27 17:24:48,723 INFO [Success] Data Profiling Successfully Done
2020-04-27 17:24:49,408 INFO [Success] Get Model Type Successfully Done
2020-04-27 17:24:49,943 INFO [Success] Data Split Successfully Done
2020-04-27 17:24:49,945 INFO [Success] Automated Machine Learning is Running ...
2020-04-27 17:25:21,788 INFO [Success] Automated Machine Learning Successfully Done !
2020-04-27 17:25:22,033 INFO [XGBoost_1 AutoML_20200427_172449] Running | Saving Result ...
2020-04-27 17:26:39,272 INFO [XGBoost_1 AutoML_20200427_172449] Running | Model Parameter Save
2020-04-27 17:26:39,293 INFO [XGBoost_1 AutoML_20200427_172449] Running | Model Metrics Save
2020-04-27 17:26:39,441 INFO [XGBoost_1 AutoML_20200427_172449] Running | Model Save (Remote)
2020-04-27 17:26:39,466 INFO [XGBoost_1 AutoML_20200427_172449] Running | Model Save (Local)
2020-04-27 17:26:39,477 INFO [XGBoost_1 AutoML_20200427_172449] Running | Artifacts Save (Remote)
2020-04-27 17:26:39,493 INFO [XGBoost_1 AutoML_20200427_172449] Finished | All Works Successfully Done
2020-04-27 17:26:39,562 INFO [XGBoost_3 AutoML_20200427_172449] Running | Saving Result ...
2020-04-27 17:27:52,362 INFO [XGBoost_3 AutoML_20200427_172449] Running | Model Parameter Save
2020-04-27 17:27:52,383 INFO [XGBoost_3 AutoML_20200427_172449] Running | Model Metrics Save
2020-04-27 17:27:52,542 INFO [XGBoost_3 AutoML_20200427_172449] Running | Model Save (Remote)
2020-04-27 17:27:52,563 INFO [XGBoost_3 AutoML_20200427_172449] Running | Model Save (Local)
2020-04-27 17:27:52,573 INFO [XGBoost_3 AutoML_20200427_172449] Running | Artifacts Save (Remote)
2020-04-27 17:27:52,588 INFO [XGBoost_3 AutoML_20200427_172449] Finished | All Works Successfully Done
2020-04-27 17:27:52,654 INFO [XGBoost_2 AutoML_20200427_172449] Running | Saving Result ...
2020-04-27 17:29:06,280 INFO [XGBoost_2 AutoML_20200427_172449] Running | Model Parameter Save
2020-04-27 17:29:06,293 INFO [XGBoost_2 AutoML_20200427_172449] Running | Model Metrics Save
2020-04-27 17:29:06,420 INFO [XGBoost_2 AutoML_20200427_172449] Running | Model Save (Remote)
2020-04-27 17:29:06,441 INFO [XGBoost_2 AutoML_20200427_172449] Running | Model Save (Local)
2020-04-27 17:29:06,453 INFO [XGBoost_2 AutoML_20200427_172449] Running | Artifacts Save (Remote)
2020-04-27 17:29:06,468 INFO [XGBoost_2 AutoML_20200427_172449] Finished | All Works Successfully Done
(base) [yoony@pb3 log] $
```

Furthermore



DEMO

Analysis

with

CMELET

Model Life Cycle

분석 프로젝트 계획

필요 데이터 정의
데이터 수집
프로젝트 일정 계획

데이터 분석 수요 발생

비즈니스 필요사항 정의
분석 요건 정의
분석 시나리오 정의

데이터 분석

탐색적 데이터 분석
모델 선택
모델 평가
모델 튜닝
...

모델 배포

최적모델 적용
비즈니스 인사이트 도출
분석 모델 배포



Overview

Overview

Overview

Reproduction

Warnings 2

Dataset statistics

Number of variables	11
Number of observations	10000
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.9 MiB
Average record size in memory	196.9 B

Variable types

NUM	5
CAT	3
BOOL	3

Overview

Reproduction

Warnings 2

Tenure has 413 (4.1%) zeros

Zeros

Balance has 3617 (36.2%) zeros

Zeros

Overview

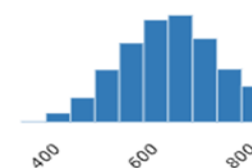
	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0
5	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
6	822	France	Male	50	7	0.00	2	1	1	10062.80	0
7	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
8	501	France	Male	44	4	142051.07	2	0	1	74940.50	0
9	684	France	Male	27	2	134603.88	1	1	1	71725.73	0

CreditScore

Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	460
Unique (%)	4.6%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	650.5288
Minimum	350
Maximum	850
Zeros	0
Zeros (%)	0.0%
Memory size	78.2 KiB

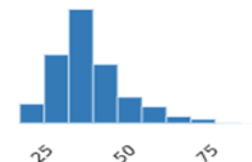


Age

Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	70
Unique (%)	0.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	38.9218
Minimum	18
Maximum	92
Zeros	0
Zeros (%)	0.0%
Memory size	78.2 KiB



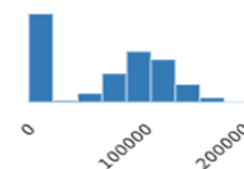
Balance

Real number ($\mathbb{R}_{\geq 0}$)

ZEROS

Distinct count	6382
Unique (%)	63.8%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

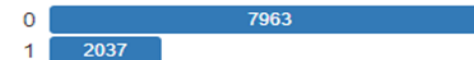
Mean	76485.88929
Minimum	0
Maximum	250898.09
Zeros	3617
Zeros (%)	36.2%
Memory size	78.2 KiB



Exited

Boolean

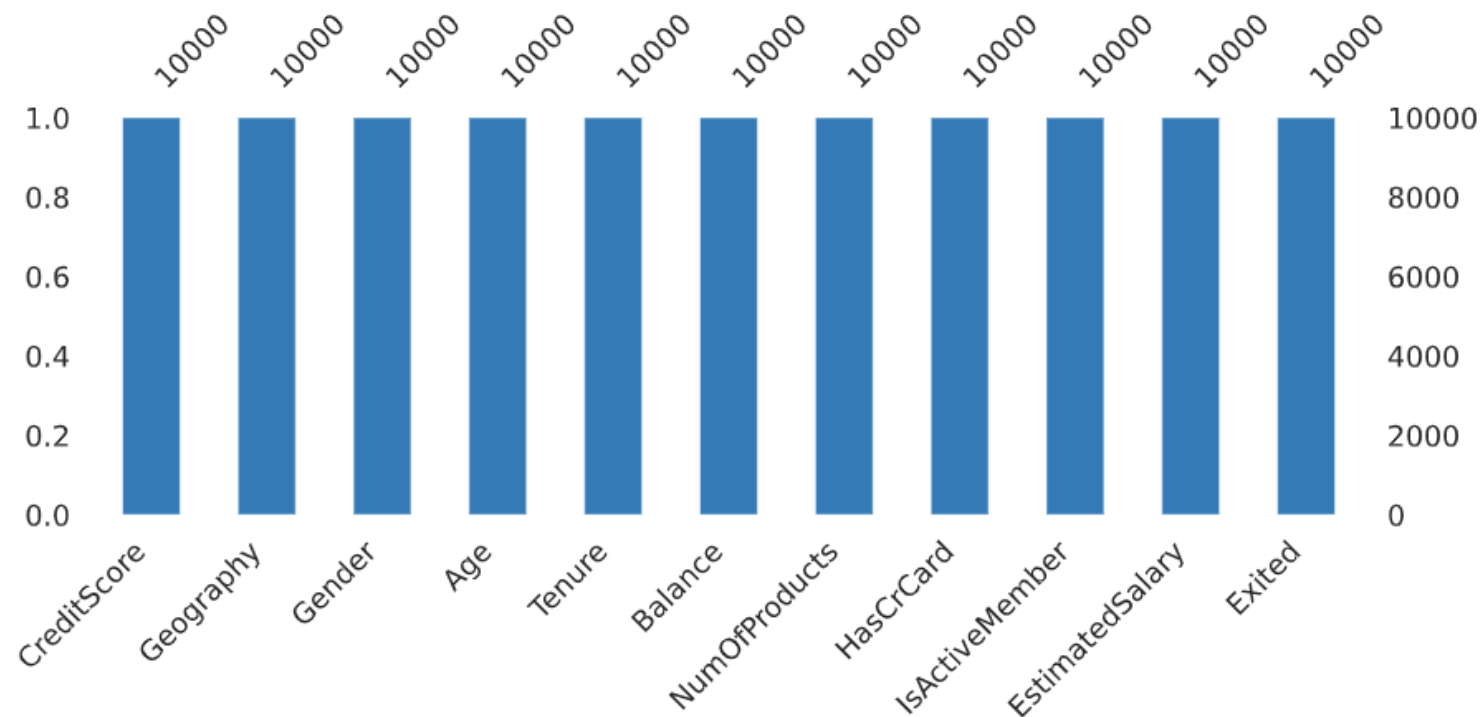
Distinct count	2
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	78.2 KiB



Missing values

Count

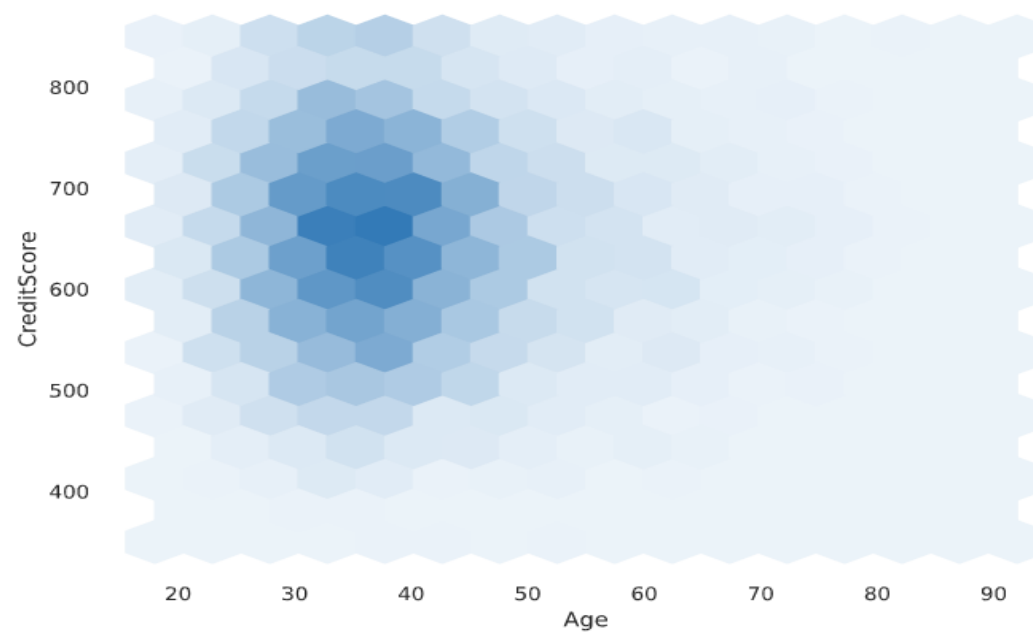
Matrix



Interactions

CreditScore Age Tenure Balance EstimatedSalary

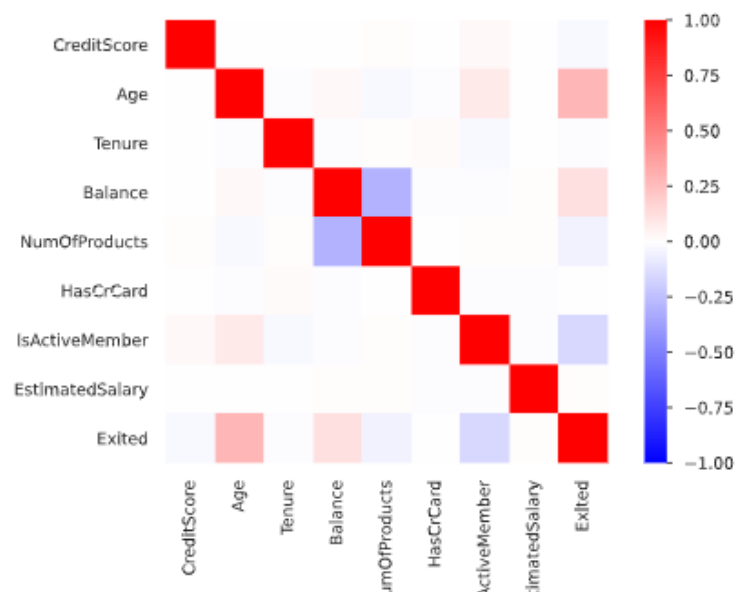
CreditScore Age Tenure Balance EstimatedSalary



Correlations

Pearson's r Spearman's ρ Kendall's τ Phik (ϕ_k)Cramér's V (ϕ_c)

Recoded

[Toggle correlation descriptions](#)

Pearson's r

The Pearson's correlation coefficient (r) is a measure of linear correlation between two variables. It's value lies between -1 and +1, -1 indicating total negative linear correlation, 0 indicating no linear correlation and 1 indicating total positive linear correlation. Furthermore, r is invariant under separate changes in location and scale of the two variables, implying that for a linear function the angle to the x-axis does not affect r .

To calculate r for two variables X and Y , one divides the covariance of X and Y by the product of their standard deviations.

Model Comparison







omelet > Comparing 9 Runs

Run ID:	b16f40080f41495293efa2f61e...	d2ba0feb98b14a06970a96f22...	344a7ef8739b4ad78d64a3c75...	d0cfbe4ccc4943da90861ed0c...	de3bbc4059fd4a9b986dc465...	01508c4f7c6e4ec0a10a4b9fc7...	65bf0ca3ef84440fbcc2ea4a31...	c1b587d927624:
Run Name:	GBM_grid__1_AutoML_20200...	XGBoost_grid__1_AutoML_20...	GBM_grid__1_AutoML_20200...	XRT_1_AutoML_20200507_09...	DRF_1_AutoML_20200507_09...	XGBoost_grid__1_AutoML_20...	XGBoost_grid__1_AutoML_20...	XGBoost_1_Auto
Start Time:	2020-05-07 09:53:13	2020-05-07 09:51:57	2020-05-07 09:50:46	2020-05-07 09:49:25	2020-05-07 09:46:49	2020-05-07 09:40:24	2020-05-07 09:37:51	2020-05-07 09:2

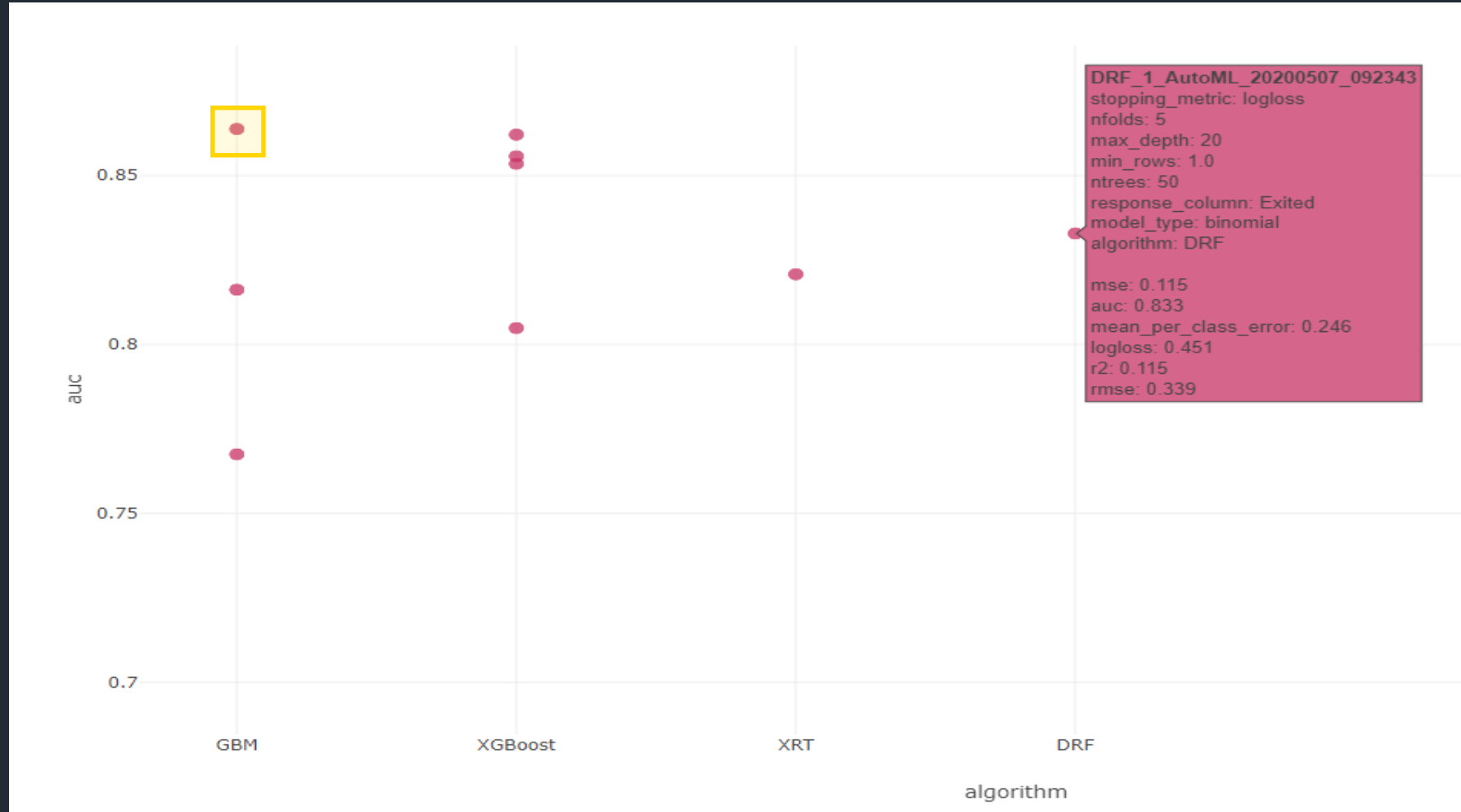
Parameters

algorithm	GBM	XGBoost	GBM	XRT	DRF	XGBoost	XGBoost	XGBoost
max_depth	6	20	12	20	20	10	10	10
min_rows	5.0	20.0	15.0	1.0	1.0	0.1	3.0	5.0
model_type	binomial	binomial	binomial	binomial	binomial	binomial	binomial	binomial
nfolds	5	5	5	5	5	5	5	5
ntrees	30	1	30	50	50	34	95	104
response_column	Exited	Exited	Exited	Exited	Exited	Exited	Exited	Exited
stopping_metric	logloss	logloss	logloss	logloss	logloss	logloss	logloss	AUTO

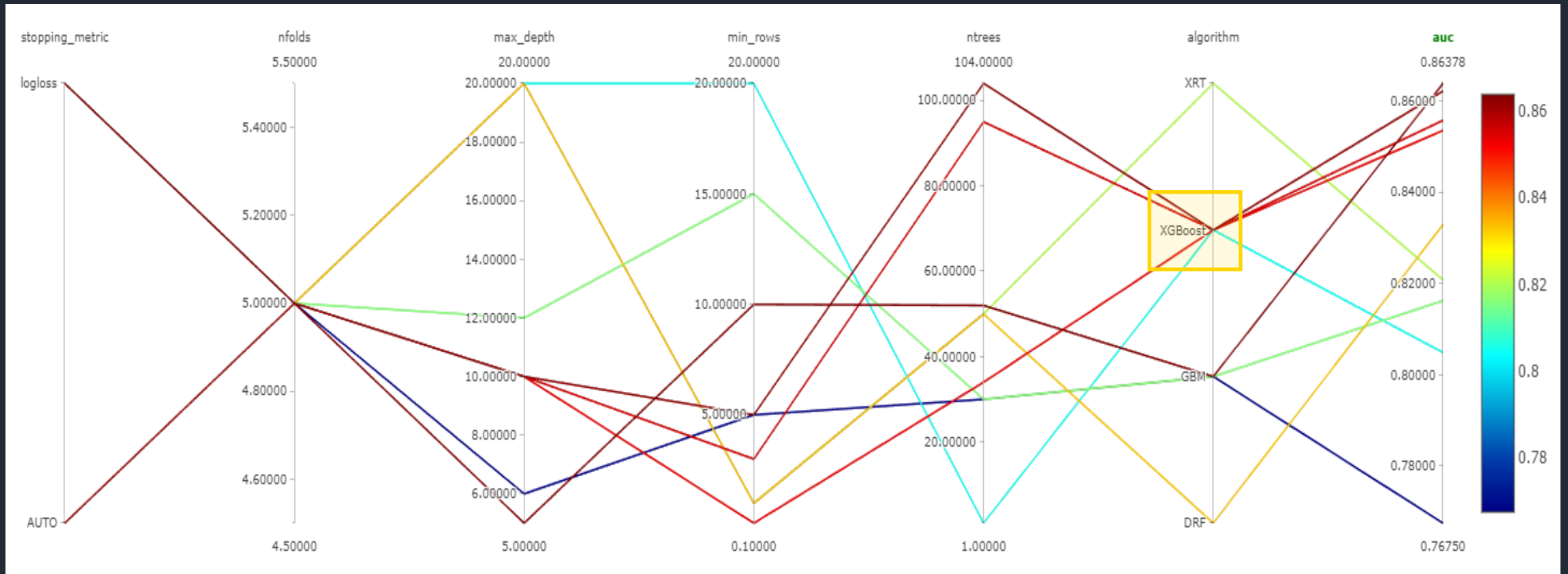
Metrics

auc 	0.768	0.805	0.816	0.821	0.833	0.853	0.856	0.862
logloss 	0.823	0.668	0.504	0.438	0.451	0.389	0.352	0.346
mean_per_class_error 	0.275	0.255	0.262	0.263	0.246	0.228	0.229	0.215
mse 	0.153	0.238	0.131	0.118	0.115	0.116	0.109	0.107
r2 	0.153	0.238	0.131	0.118	0.115	0.116	0.109	0.107
rmse 	0.391	0.487	0.362	0.343	0.339	0.341	0.33	0.327

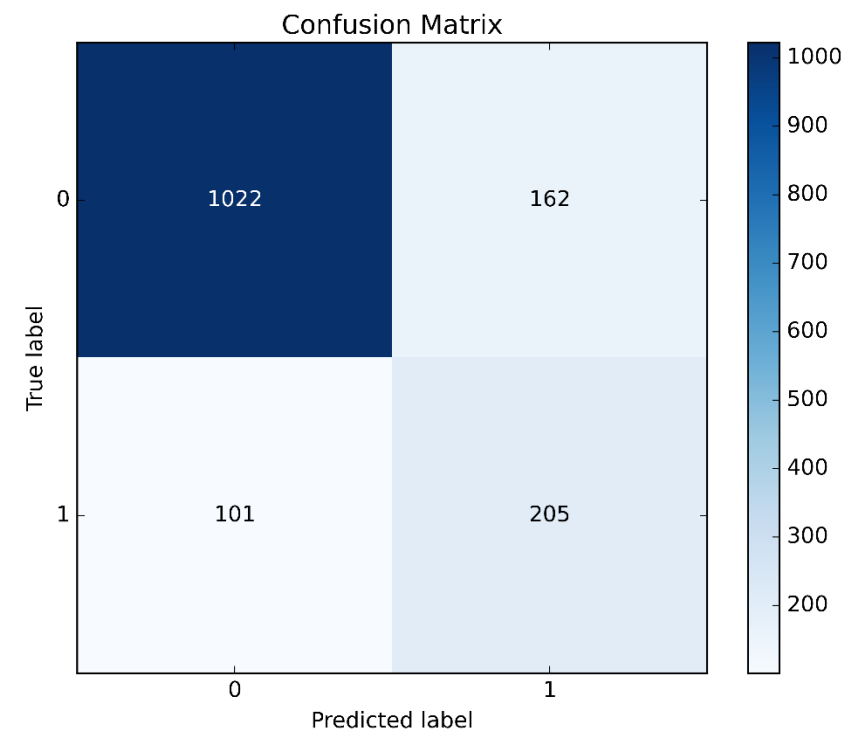
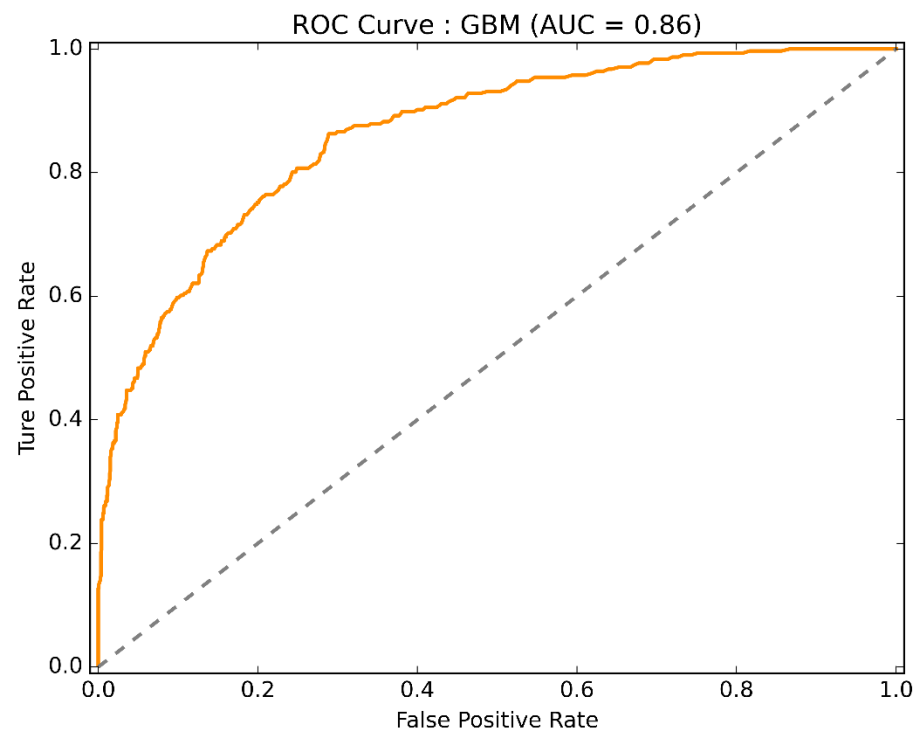
Model Comparison



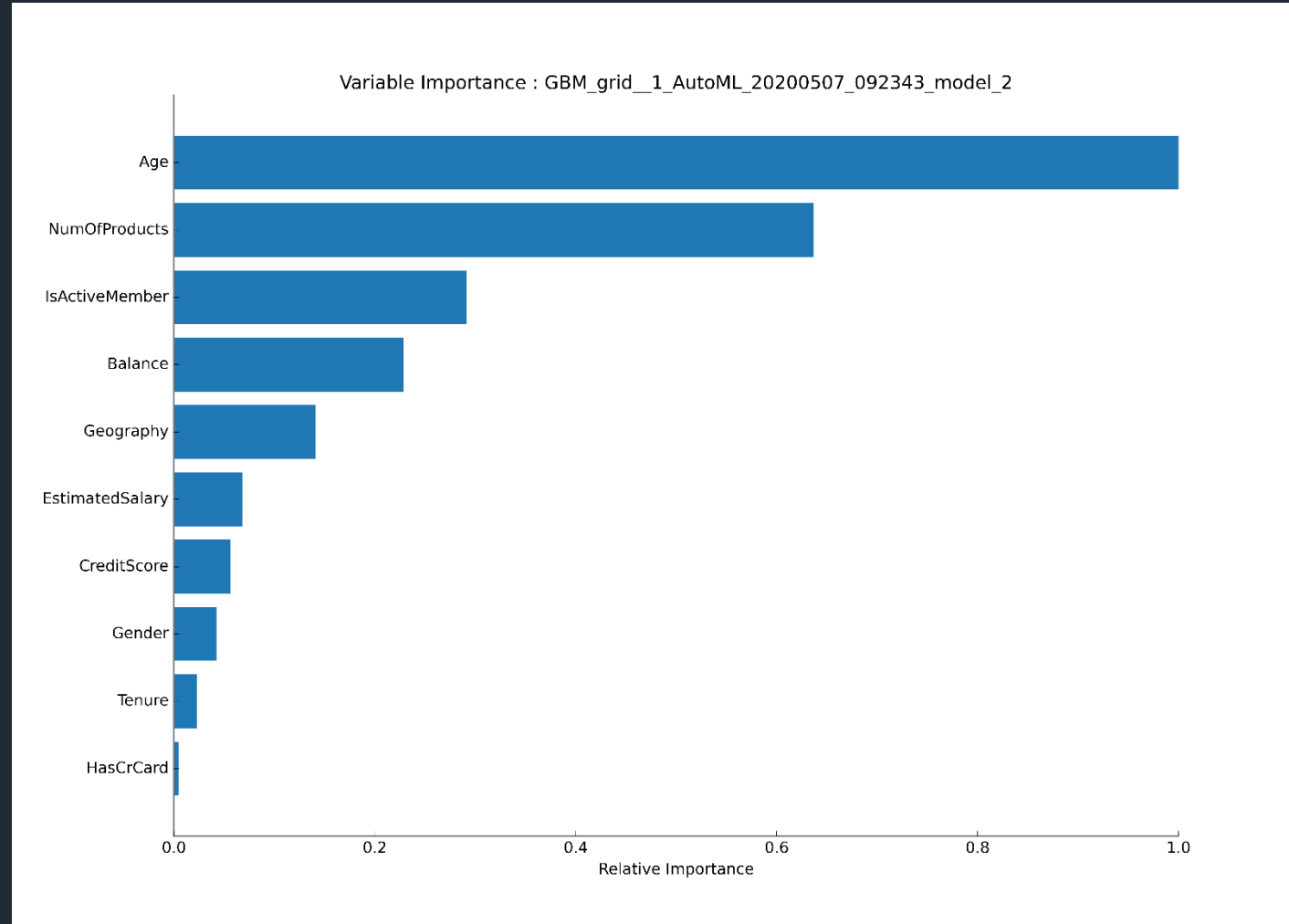
Model Comparison



Champion Model

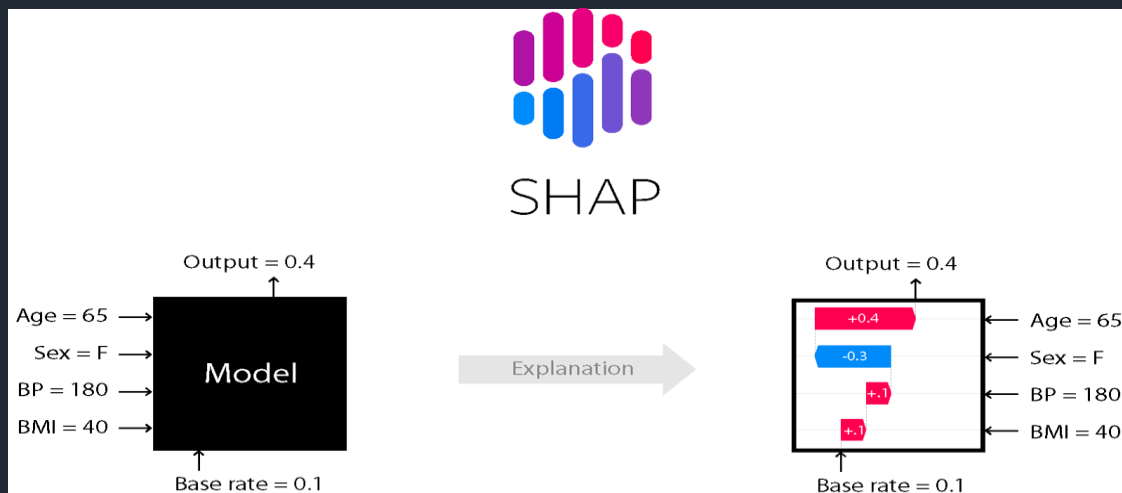


Variable Importance

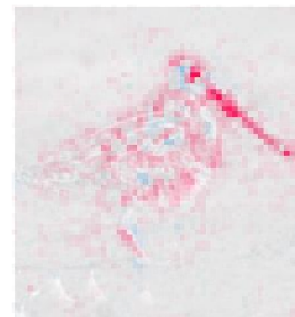


Model Explanation

항목	내용
잊혀질 권리 (right to be forgotten)	제17조 - 정보 주체가 본인의 개인정보 처리를 더 이상 원치 않거나 개인정보를 보유할 법적 근거가 없으면 해당 정보 삭제
자동화된 의사결정 제한	제22조 - 자동화된 처리 (프로파일링 포함)에만 근거한 결정의 대상이 되지 않을 권리
설명을 요구할 권리 (right to explanation)	제13-14조 - 알고리즘에 의해 행해진 결정에 대해 질문하고, 결정에 관여한 논리에 대해 의미있는 설명을 요구할 권리
EU 집행력	규정 위반시 해당 기업의 전세계 매출의 최대 4%까지 벌금 부과
발효	2018년 5월 28일



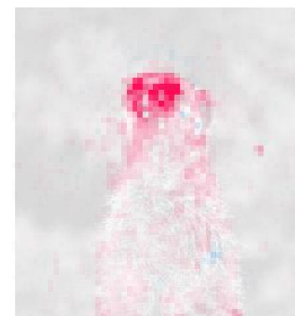
dowitcher



red-backed_sandpiper



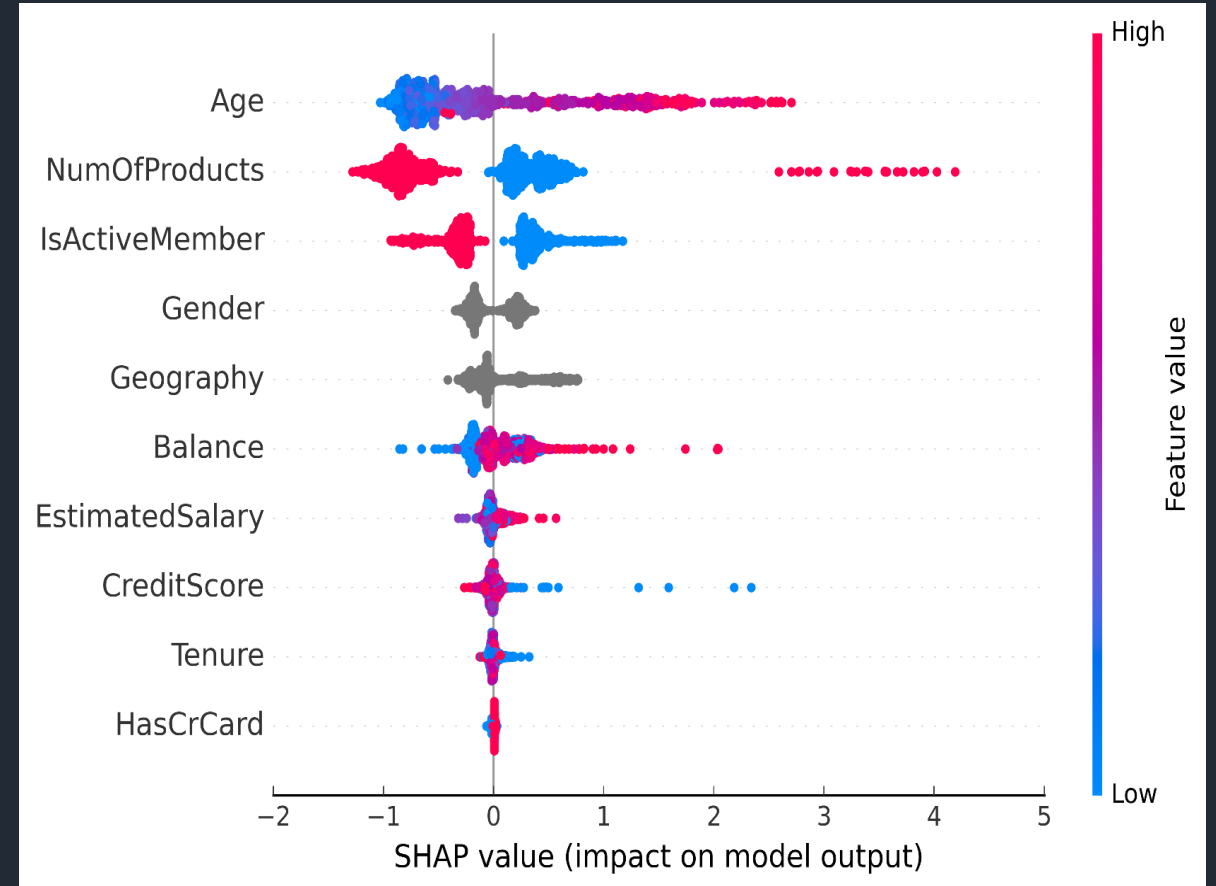
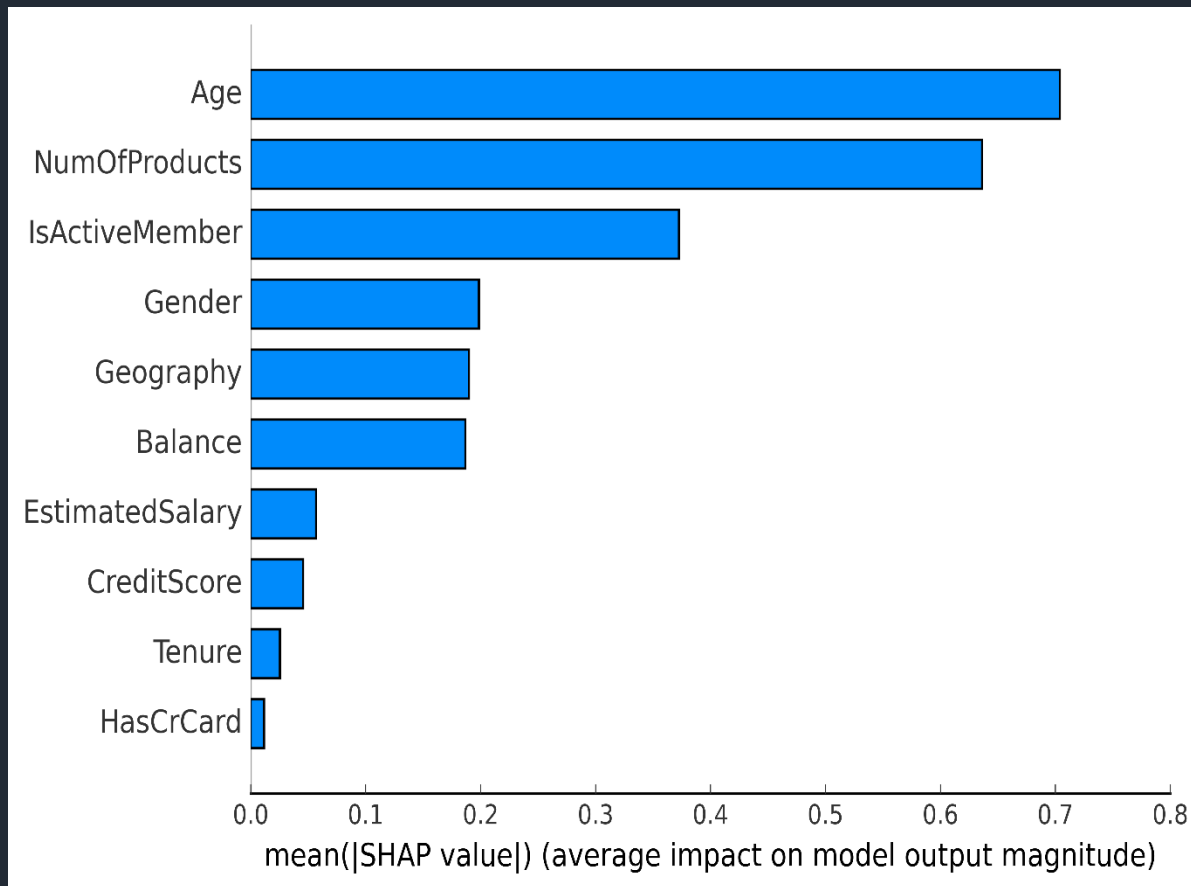
meerkat



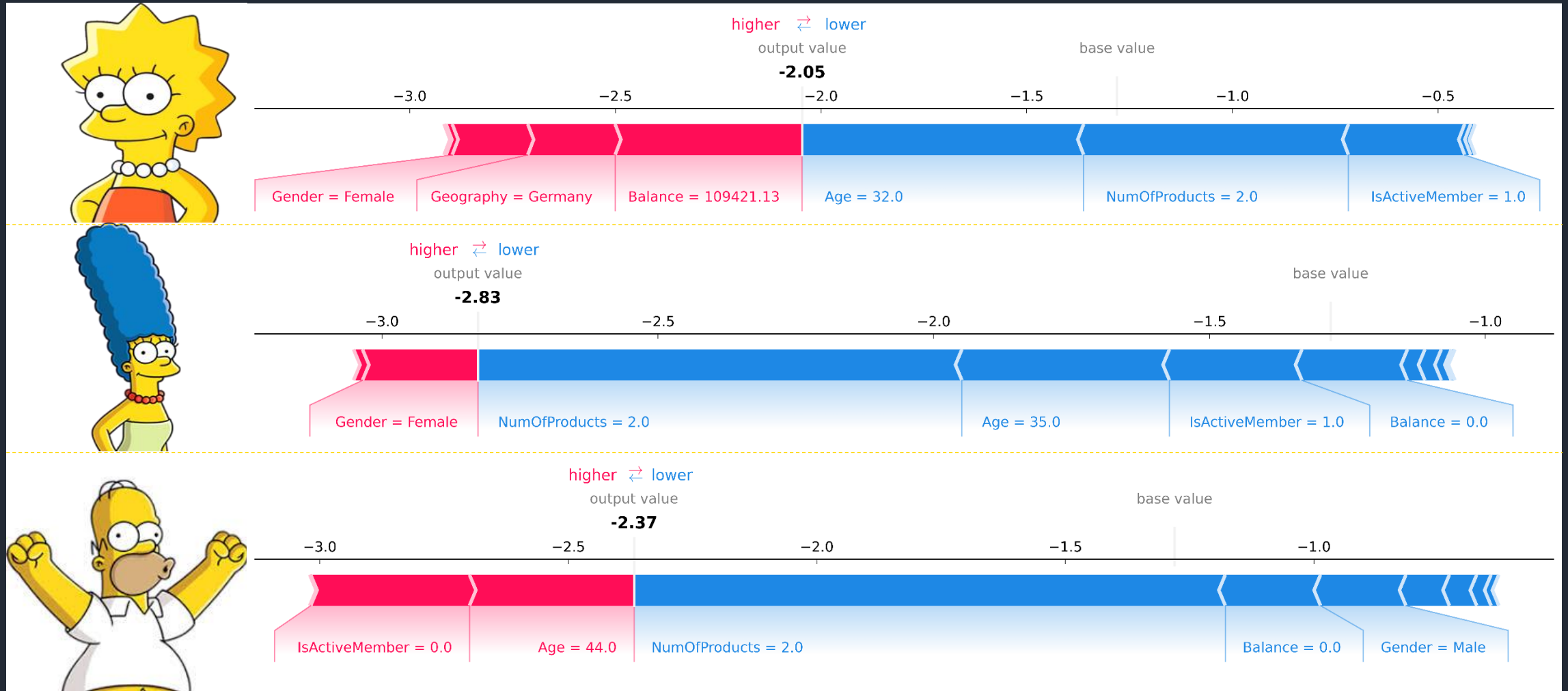
mongoose



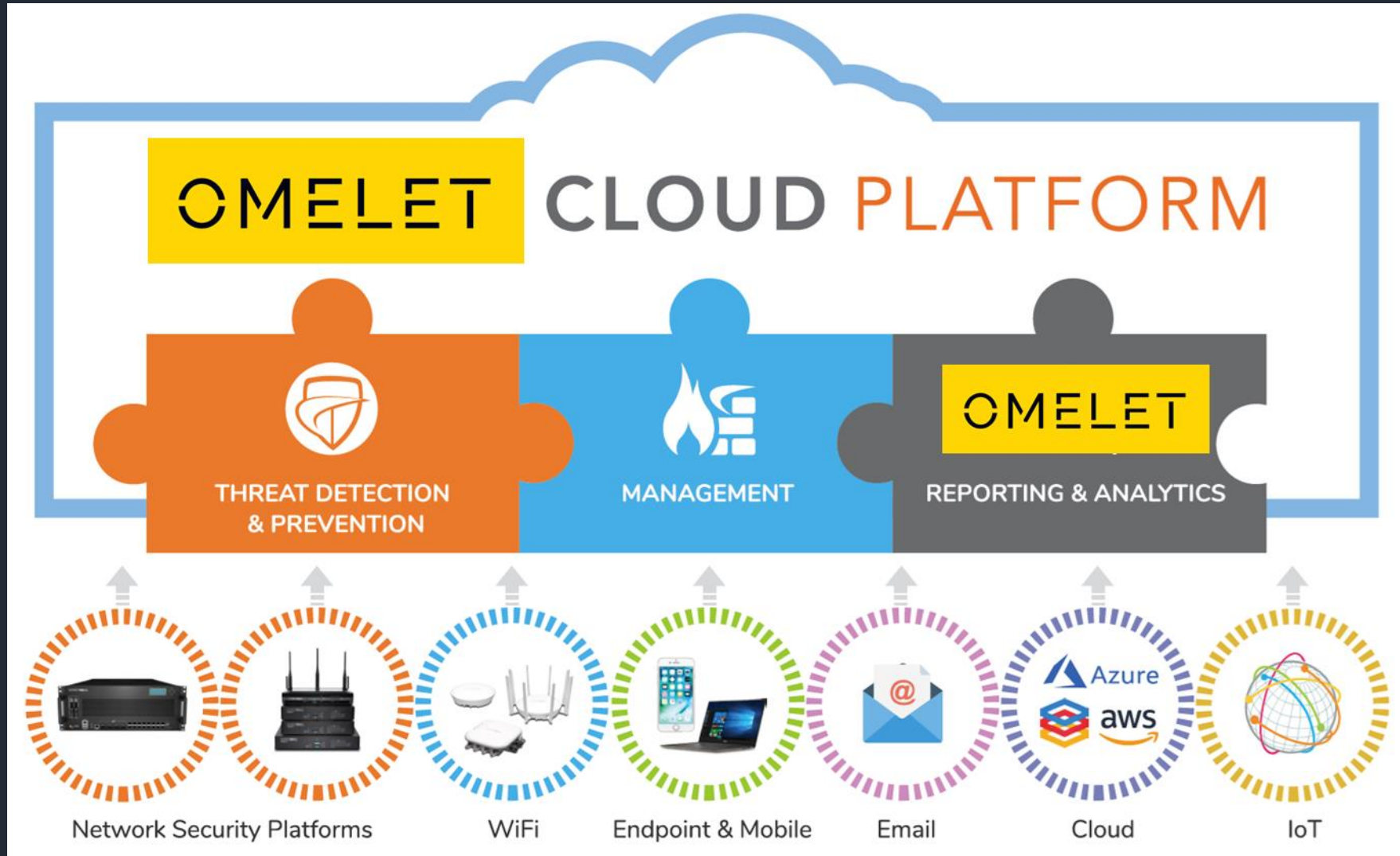
Model Explanation



Model Explanation



The Future of OMELET



THANK YOU !