

피처 엔지니어링 : 파생 변수

Feature Engineering : Transform

피쳐 엔지니어링이란?

특성 선택

차원 축소

파생 변수 생성

스케일링

피쳐 엔지니어링이란?

특성 선택

차원 축소

파생 변수 생성

스케일링



01

파생 변수란?

파생 변수란?



200만원 / 1인 가구



300만원 / 2인



400만원 / 4인



가구 총소득



200만원 / 1인 가구



300만원 / 2인



400만원 / 4인



소득 / 가구원 수



200만원



150만원

































100만원



사례 : 넷플릭스 영화 TOP 10

Most Watched Movies




List of the Most Watched Movies on Netflix in 2023 from January to June.

TITLE		TYPE	PREMIERE	GENRE	COUNTRY	HOURS
Search in 10899 titles...		Movies ▾	All ▾	All ▾	All ▾	TV Shows by se ▾
1.	 The Mother 	Movie	2023	Thriller		249,900,000
2.	 Luther: The Fallen Sun 	Movie	2023	Crime		209,700,000
3.	 Extraction 2 	Movie	2023	Action		201,800,000
4.	 You People 	Movie	2023	Comedy		181,800,000
5.	 Murder Mystery 2 	Movie	2023	Comedy		173,600,000
6.	 Your Place or Mine 	Movie	2023	Comedy		163,000,000
7.	 Glass Onion: A Knives Out Mystery 	Movie	2022	Crime		142,900,000
8.	 We Have a Ghost 	Movie	2023	Adventure		124,400,000
9.	 The Pale Blue Eye 	Movie	2023	Crime		120,500,000
10.	 AKA 	Movie	2023	Action		120,000,000

사례 : 넷플릭스 영화 TOP 10

영화 제목



























시청 시간

1.		The Mother N	249,900,000
2.		Luther: The Fallen Sun N	209,700,000
3.		Extraction 2 N	201,800,000

사례 : 넷플릭스 **한국** 영화 TOP 10

Most Watched Movies from South Korea




List of the Most Watched Movies from South Korea on Netflix in 2023 from January to June.

TITLE		TYPE	PREMIERE	GENRE	COUNTRY	HOURS
Search in 489 titles...		Movies ▾	All ▾	All ▾	Sout ▾	TV Shows by se ▾
1.	 Kill Boksoon 	Movie	2023	Thriller		68,500,000
2.	 JUNG_E 	Movie	2023	Science Fiction		47,100,000
3.	 Unlocked 	Movie	2023	Crime		42,500,000
4.	 Pinkfong & Baby Shark's Space Adventure	Movie	2019	Animation		22,300,000
5.	 Emergency Declaration	Movie	2022	Action		18,600,000
6.	 Carter 	Movie	2022	Action		11,900,000
7.	 20th Century Girl 	Movie	2022	Drama		11,900,000
8.	 Confidential Assignment 2: International	Movie	2022	Action		11,100,000
9.	 #Alive 	Movie	2020	Horror		10,700,000
10.	 Pinkfong Sing-Along Movie 2: Wonderstar Concert	Movie	2022	Animation		9,900,000



사례 : 넷플릭스 **한국** 영화 TOP 10

영화 제목

시청 시간

1.	 Kill Boksoon N	68,500,000
2.	 JUNG_E N	47,100,000
3.	 Unlocked N	42,500,000

넷플릭스는 어떤 것을 더 선호할까?




1.		Kill Boksoon N	68,500,000
2.		JUNG_E N	47,100,000
3.		Unlocked N	42,500,000

VS

1.		The Mother N	249,900,000
2.		Luther: The Fallen Sun N	209,700,000
3.		Extraction 2 N	201,800,000

넷플릭스는 어떤 것을 더 선호할까?



제작비

1.		Kill Boksoon N	68,500,000
2.		JUNG_E N	47,100,000
3.		Unlocked N	42,500,000

150억원

200억원

120억원





1.		The Mother N	249,900,000
2.		Luther: The Fallen Sun N	209,700,000
3.		Extraction 2 N	201,800,000




570억원

720억원

860억원

넷플릭스는 어떤 것을 더 선호할까?

1.	 Kill Boksoon N	46만 시간/1억원	
2.	 JUNG_E N	24만 시간/1억원	
3.	 Unlocked N	35만 시간/1억원	

1.	 The Mother N	43만 시간/1억원	
2.	 Luther: The Fallen Sun N	29만 시간/1억원	
3.	 Extraction 2 N	23만 시간/1억원	

VOD 플랫폼 vs OTT 플랫폼



VOD 플랫폼 vs OTT 플랫폼



VOD 구입

4000원



넷플릭스
가입

13500원

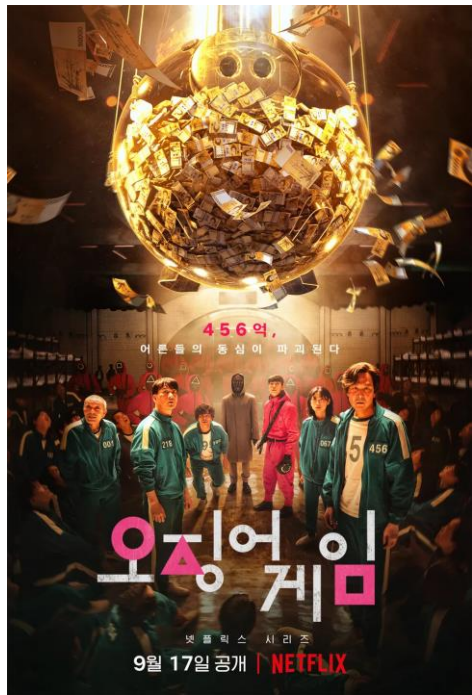
VOD 플랫폼 vs OTT 플랫폼



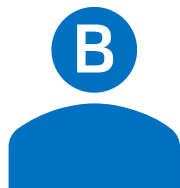
13500원



시청시간 점유율



오징어게임 총 시청
8시간 / 16시간



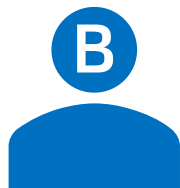
오징어게임 총 시청
8시간 / 40시간



시청시간 점유율



50%



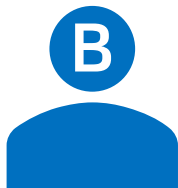
20%

오징어게임의 실질 기여금액



50%

6750원



20%

2700원

그 밖에 고려할 수 있는 것

최대 연속 시청시간 합

구독자들의 만족도, 콘텐츠의 흡입력 판단 지표

최초 공개 후 시청까지 걸린 시간

콘텐츠의 기대감을 반영하는 지표



파생 변수 생성 시 필요한 것



파생 변수 생성 시 필요한 것



도메인
지식

해당 분야에 대한
이해가 필요



활용가능
데이터

파생 변수 생성에
필요한 데이터가
활용 가능해야 함



A/B
테스트

모델을 생성 했을 때,
실질적으로 지표를
개선 시켰는지 확인

02

파생 변수를 만드는 다양한 방법

시계열 데이터 : 지난 기록

날짜	제품 판매량
2024/01	156
2024/02	245
2024/03	367
2024/04	255
2024/05	278



시계열 데이터 : 지난 기록

날짜	제품 판매량	지난달 판매량
2024/01	156	-
2024/02	245	156
2024/03	367	245
2024/04	255	367
2024/05	278	255



시계열 데이터 : 지난 기록

날짜	제품 판매량	지난달 판매량	작년 판매량
2024/01	156	-	123
2024/02	245	156	222
2024/03	367	245	454
2024/04	255	367	255
2024/05	278	255	278

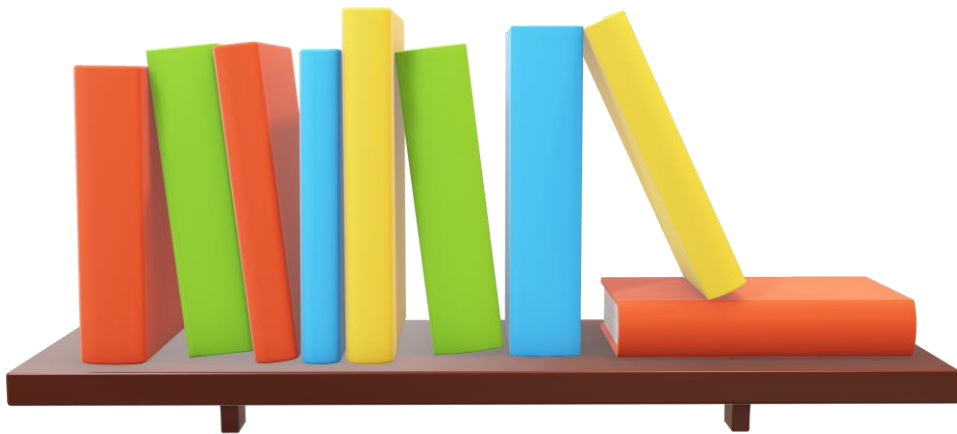


시계열 데이터 : 지난 기록

날짜	제품 판매량	지난달 판매량	최근 3개월 평균
2024/01	156	-	-
2024/02	245	156	-
2024/03	367	245	256
2024/04	255	367	289
2024/05	278	255	300



범주형 데이터 : 대표값으로 대체



타깃 변수

도서 구입에 사용하는 비용

범주형 데이터 : 대표값으로 대체

타깃 변수

도서 구입에 사용하는 비용

평균값

1학년

₩100,000

2학년

₩120,000

3학년

₩150,000

4학년

₩90,000

서열 척도?

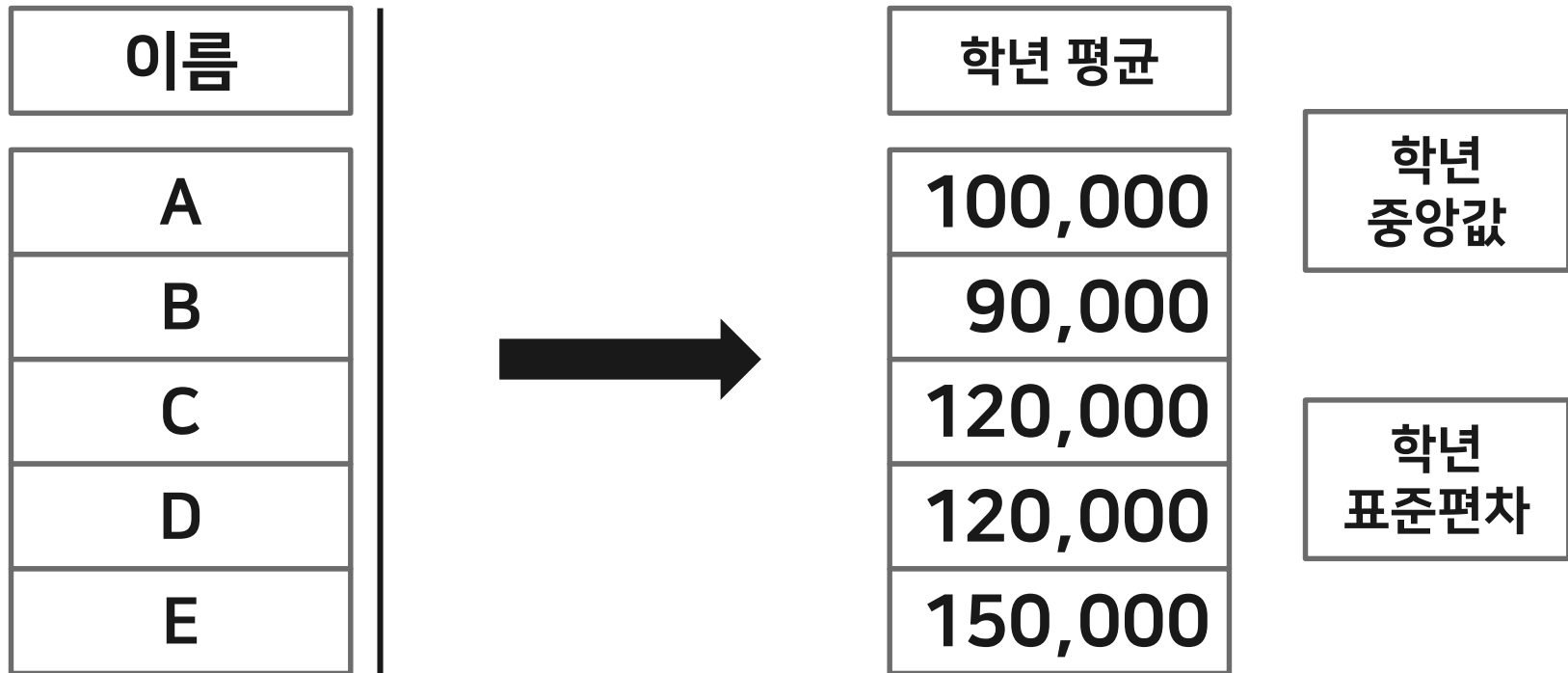


범주형 데이터 : 대표값으로 대체

이름	학년
A	1
B	4
C	2
D	2
E	3



범주형 데이터 : 대표값으로 대체



중복 데이터 : 개인 식별 ID 생성



데이터

신용카드 연체 기록

중복 데이터 : 개인 식별 ID 생성

데이터

신용카드 연체 기록

타깃 변수

카드 연체 여부 예측하기



- 데이터의 각 레코드는 한 장의 카드를 나타냄
- 한 명이 여러 장의 카드를 사용한다면, 여러 번 나타남

중복 데이터 : 개인 식별 ID 생성

월 소득	₩3,200,000	가족 구성원 수	3
거주지	서울 광진구	신용점수	920

한 개인을 특정할 수 있는 피처만 선택

~~신용카드 발급일자~~ ~~신용카드사~~ ~~신용카드 이용금액~~

중복 데이터 : 개인 식별 ID 생성

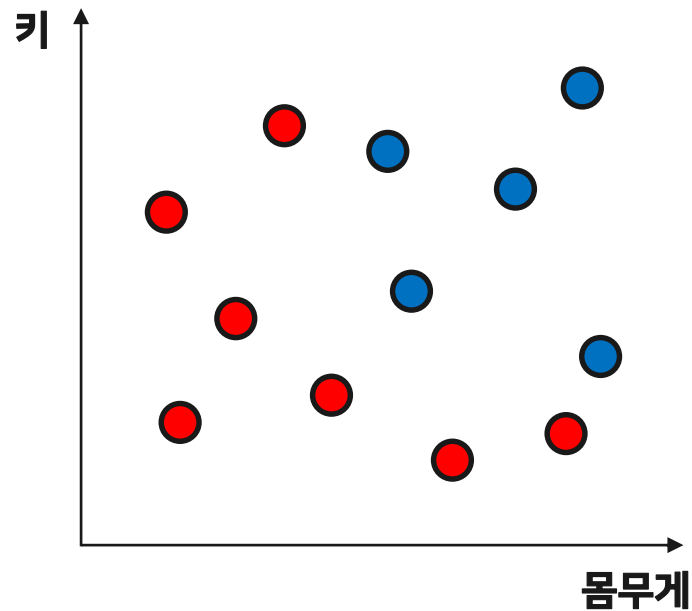
식별 ID

3200000_서울 광진구_3_920

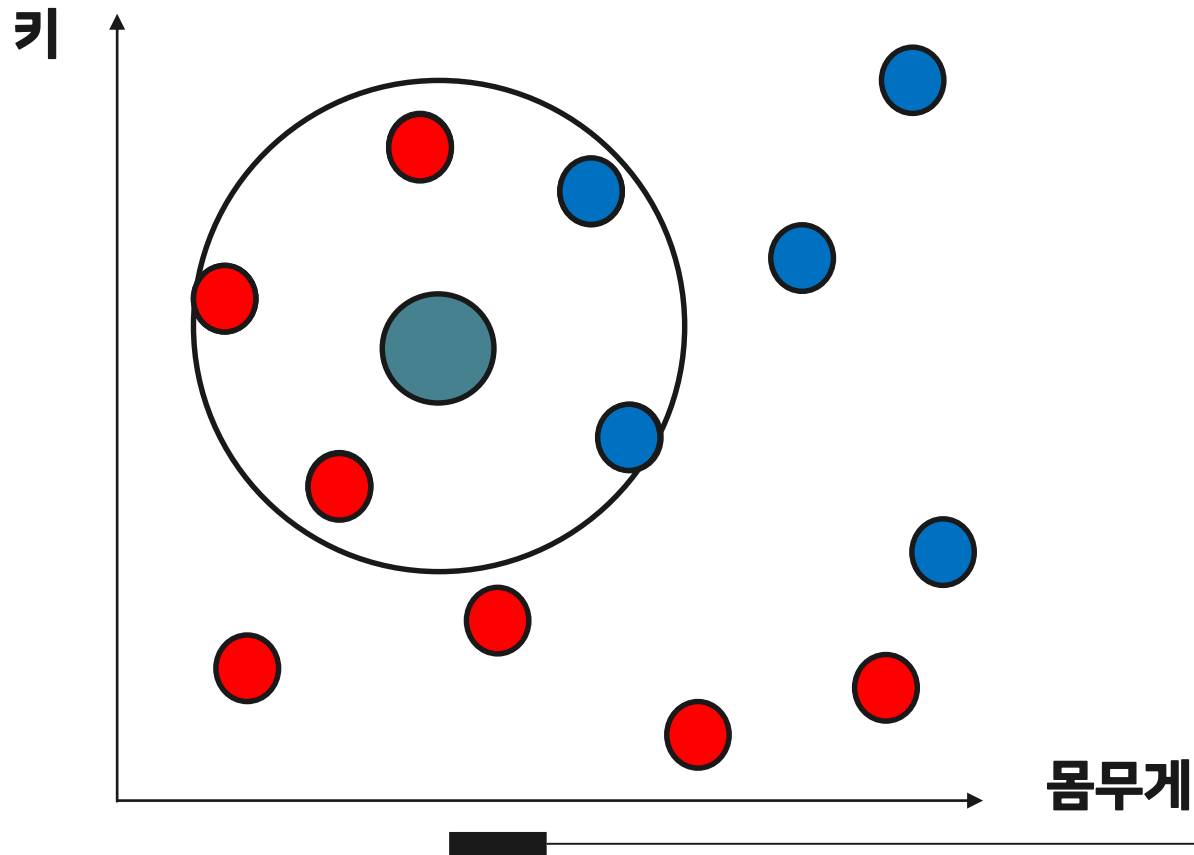
한 개인에 대하여 고유하게 식별할 수 있음

모델링으로 새로운 피처 만들기

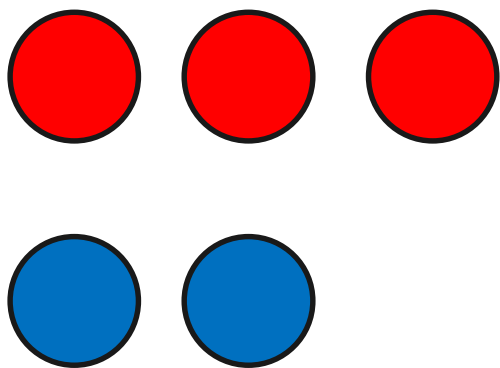
K - 최근접 이웃 (KNN)



모델링으로 새로운 피처 만들기



모델링으로 새로운 피처 만들기



$$\frac{2}{5}$$

predict_proba

0.4

모델링으로 새로운 피처 만들기

이름	키	몸무게
A	182	85
B	162	54
C	158	53
D	167	64
E	177	73

모델링으로 새로운 피처 만들기

이름	키	몸무게	KNN
A	182	85	1.0
B	162	54	0.4
C	158	53	0.2
D	167	64	0.6
E	177	73	0.8

피쳐 엔지니어링이란?

특성 선택

차원 축소

파생 변수 생성

스케일링

파생 변수 생성 후에

특성 선택

차원 축소

스케일링

03

파생 변수가 중요한 이유

**모델을 바꾸고,
하이퍼파라미터를
튜닝하는 것만으로는,**



**성능 향상에
한계가 있기 때문!**



AutoML의 등장

모델 선택

하이퍼파라미터 튜닝



AutoML의 등장



OPTUNA





9기 박영원 / 정다연