Domestic Scaled Company

Data: Tech Company Fundings (since 2020)

Minsoo Kwak

Contents

A1. Data: 2020년 이후 투자 받은 Tech company

	df_index	Company	Website	Region	Vertical	Funding Amount (USD)	Funding Stage	Funding Date
0	1	Internxt	https://internxt.com/	Spain	Blockchain	278940	Seed	Jan-20
1	2	Dockflow	https://dockflow.com	Belgium	Logistics	292244	Seed	Jan-20

A2. 원본 데이터 확인 (3575, 8)

원본 데이터 확인

	count	unique	top	freq
Funding Date	3575	19	May-21	332
Funding Stage	3575	22	Series A	951
Region	3563	72	United States	2034
Vertical	3575	143	B2B Software	632
Funding Amount (USD)	3575	969	10000000	105
Company	3575	3224	Internxt	4
Website	3575	3343	https://humaninterest.com/	4

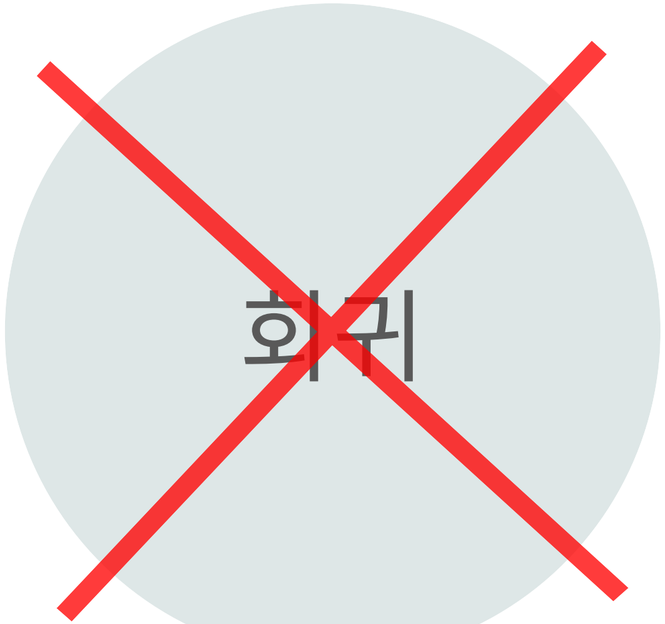
Warnings

다중공산성의 위험성

`Company` has a high cardinality: 3224 distinct values	High cardinality
`Website` has a high cardinality: 3343 distinct values	High cardinality
`Region` has a high cardinality: 72 distinct values	High cardinality
`Vertical` has a high cardinality: 143 distinct values	High cardinality
`Funding Amount (USD)` has a high cardinality: 969 distinct values	High cardinality
`df_index` has unique values	Unique

Selecting Model

Funding Amount(USD), Funding Date 를 제외한 feature들이 카테고리형임



EDA / Data Wrangling

B1. EDA 및 데이터 전처리_1

1. Column 명 변경

2. 결측치 확인 (가시적/ 비가시적)

```
[(x,df[x].isnull().sum()) for x in df.columns if df[x].isnull().any()]
```

```
[('Region', 12)]
```

```
df.dtypes
```

```
index                  int64
Company               object
Website               object
Region                object
Field                 object
Funding Amount(USD)   object
Funding_stage         object
Funding_date          object
dtype: object
```


int (정수형)

```
df.isna().sum() # 형 변환을 시켜줬더니 드러나지 않았던 결측치가 드러남
```

```
index                  0
Company                0
Website                0
Region                12
Field                  0
Funding Amount(USD)    9
Funding_stage          0
Funding_date           0
dtype: int64
```

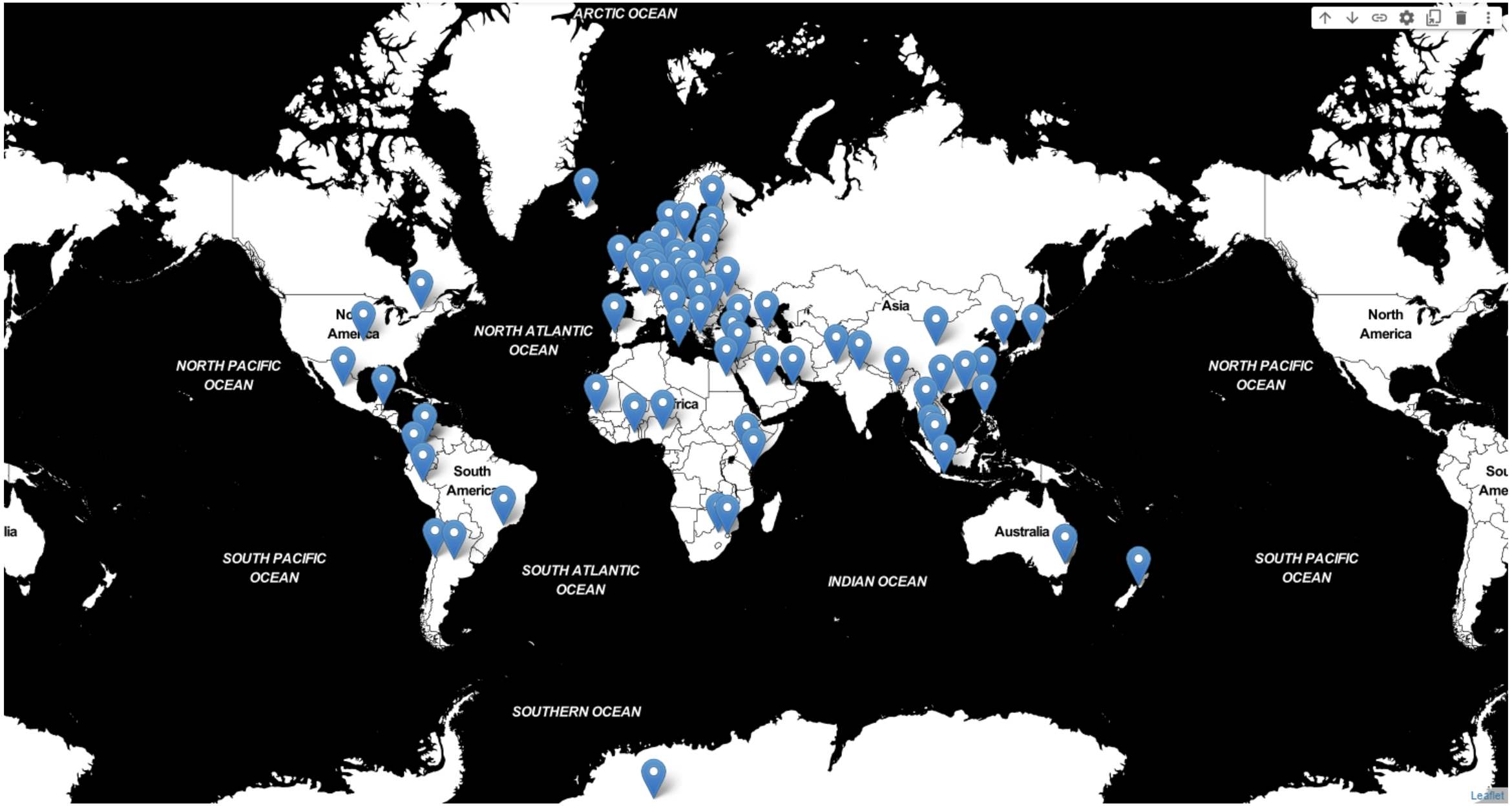
Funding Amount(USD) 에 있는 Unknown 항목 처리

결측치의 양이 적으므로 fillna()를 통해 결측치를 대체

	index	Funding Amount(USD)
count	3575.00	3566.00
mean	1788.00	57560141.00
std	1032.16	298197613.73
min	1.00	40000.00
25%	894.50	5000000.00
50%	1788.00	15496301.50
75%	2681.50	50000000.00
max	3575.00	16600000000.00

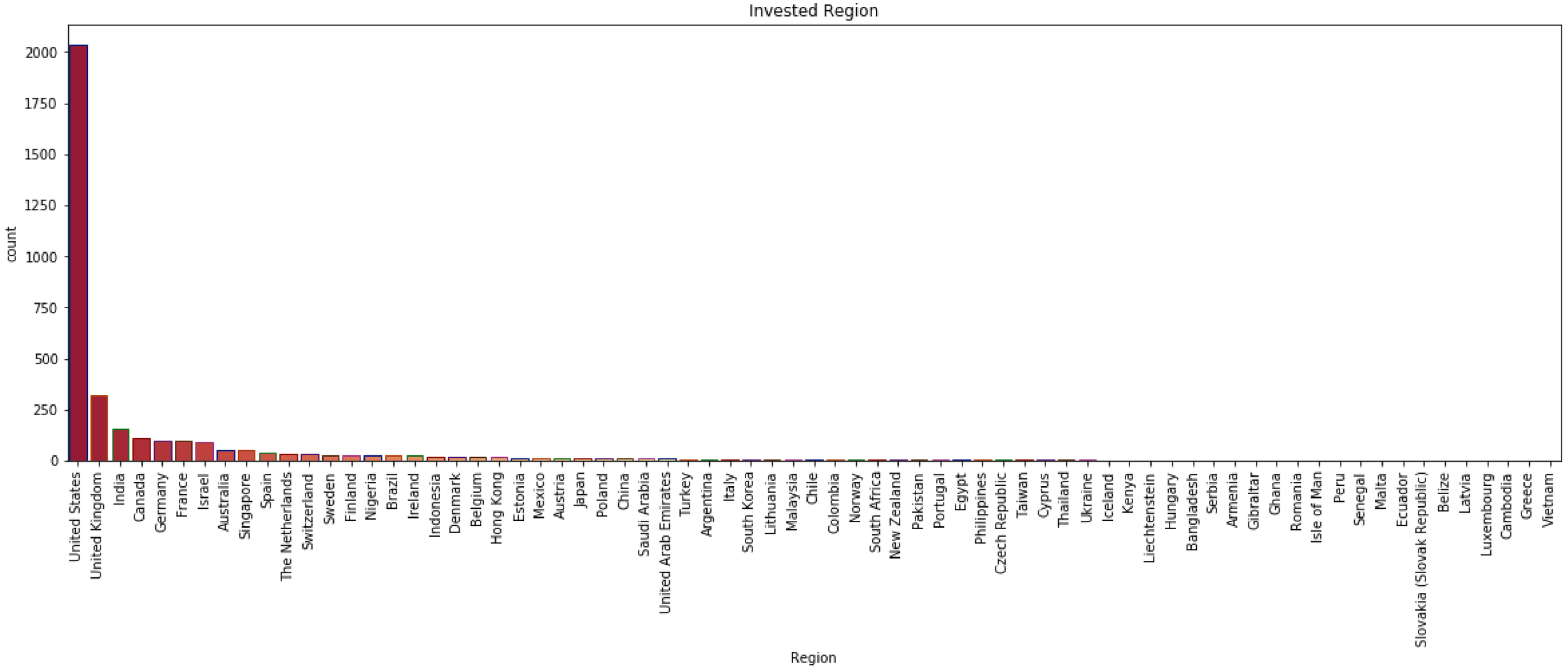
3. Funding Date 전처리
4. index, Website 열 제거

B2. Data wrangling 1. Region



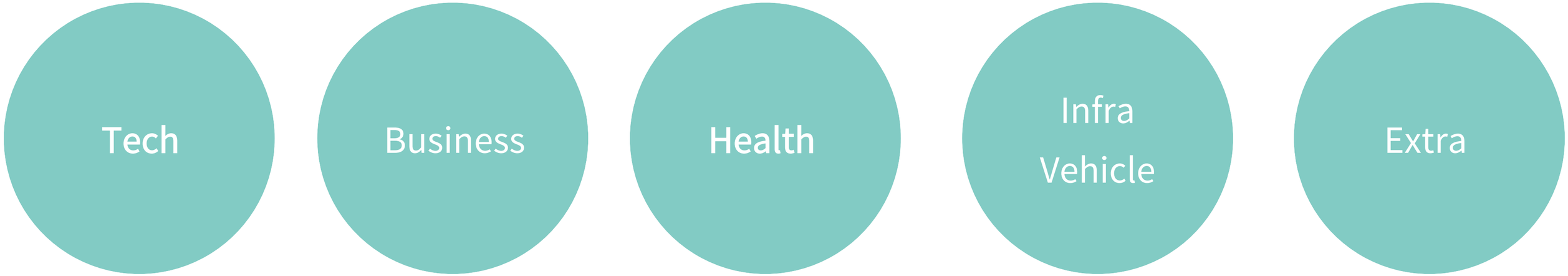
Invested Tech company by Region

미국이 2000을 넘는 수치로 가장 많음

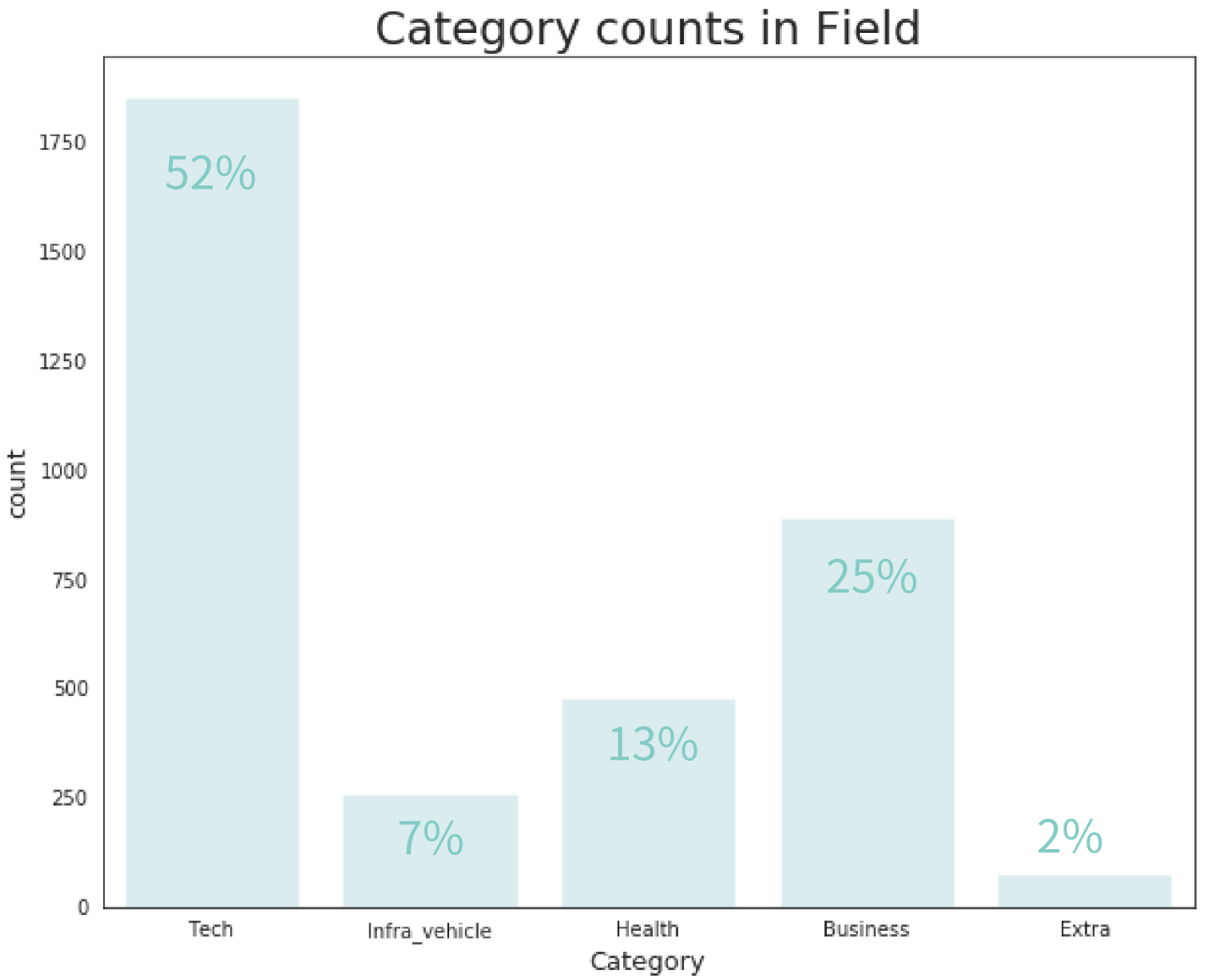


B3. Data wrangling 2. Field (Verticle)

- 143개 영역을 5개 범주로 카테고리화
- 사용하지 않을 Field(Verticle) 제거

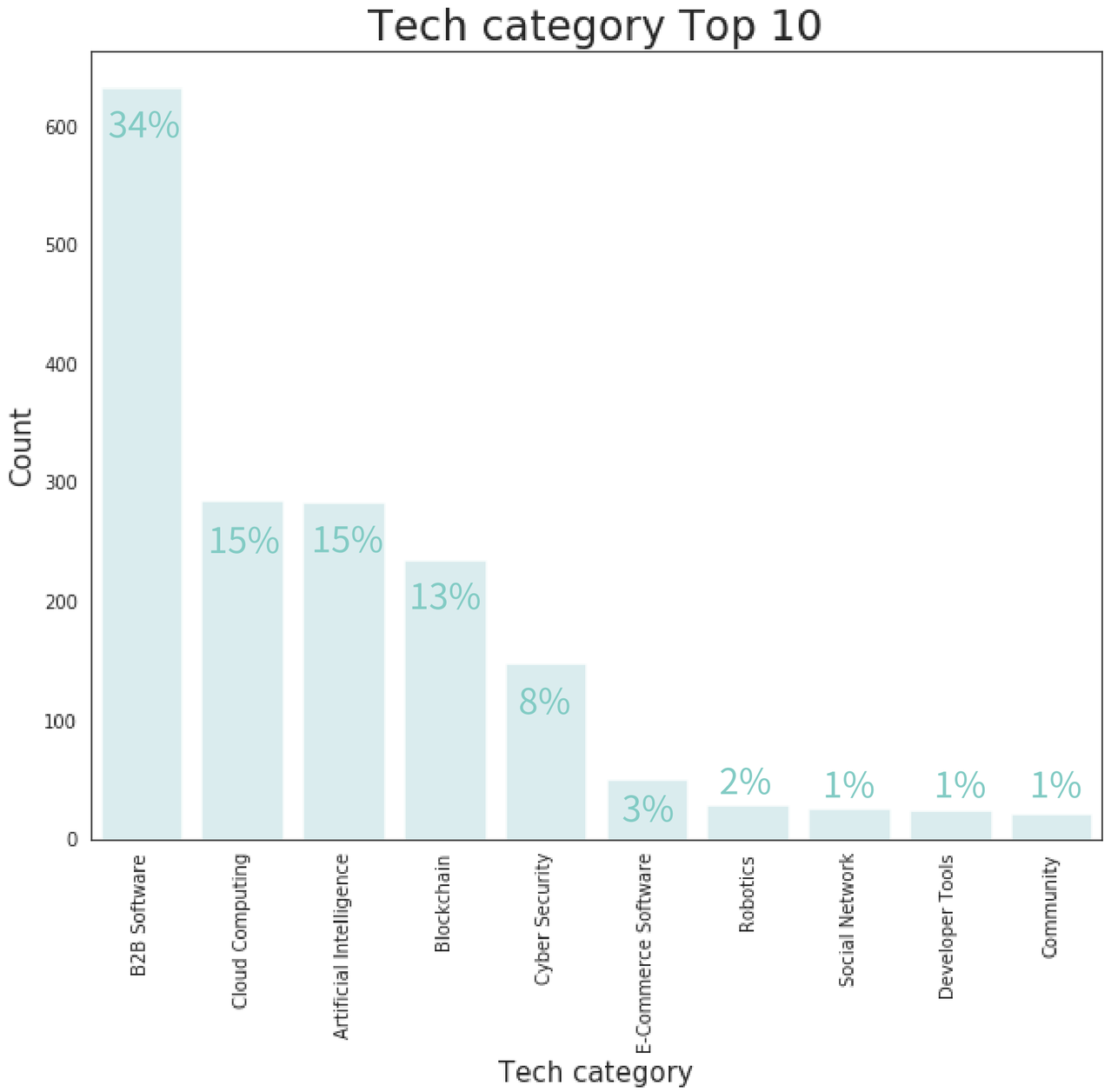


Categorized Field(Verticle) 분포



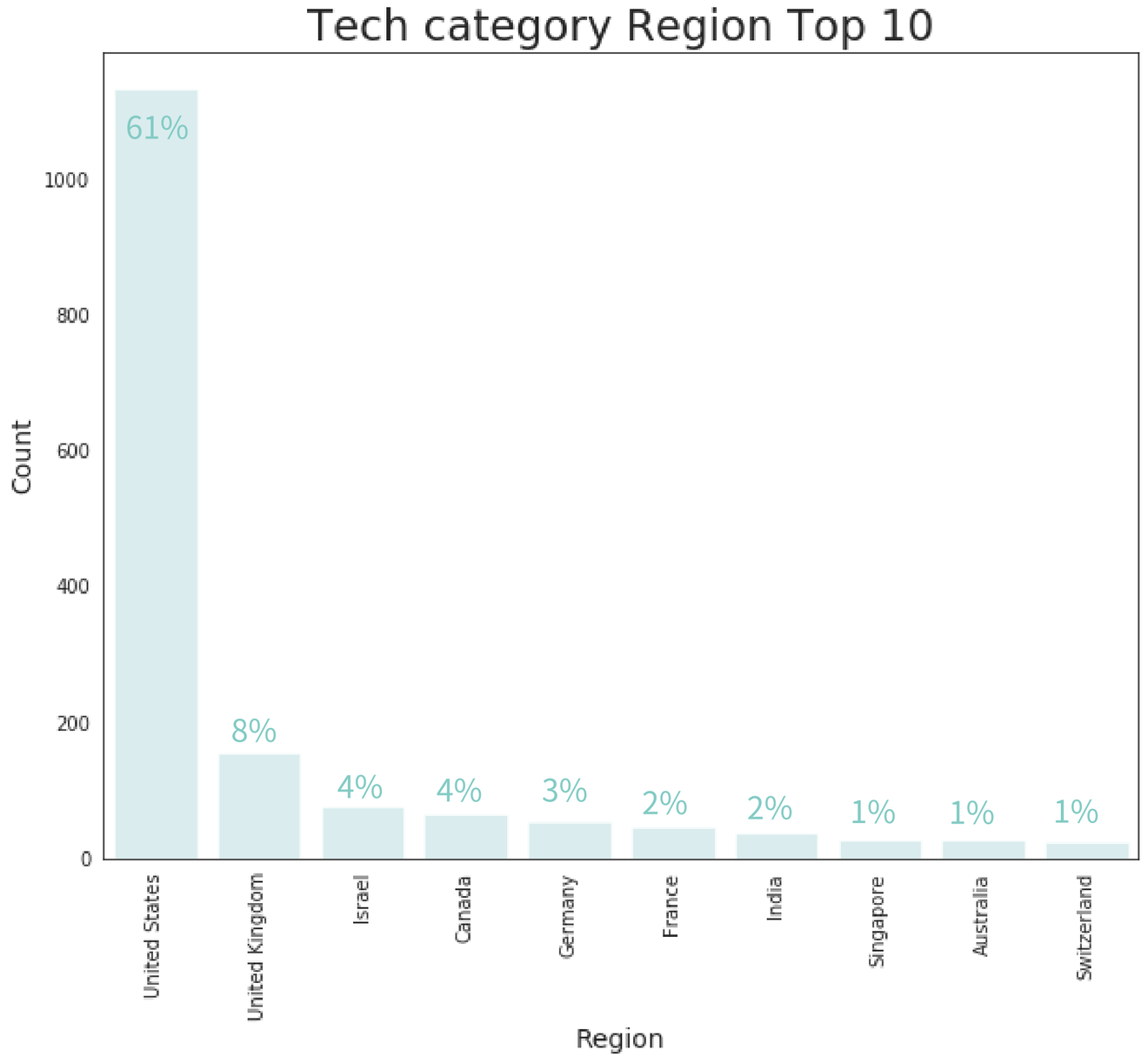
Tech Category Top 10

	index	Field
0	B2B Software	632
1	Cloud Computing	285
2	Artificial Intelligence	283
3	Blockchain	235
4	Cyber Security	147
5	E-Commerce Software	50
6	Robotics	29
7	Social Network	26
8	Developer Tools	24
9	Community	20



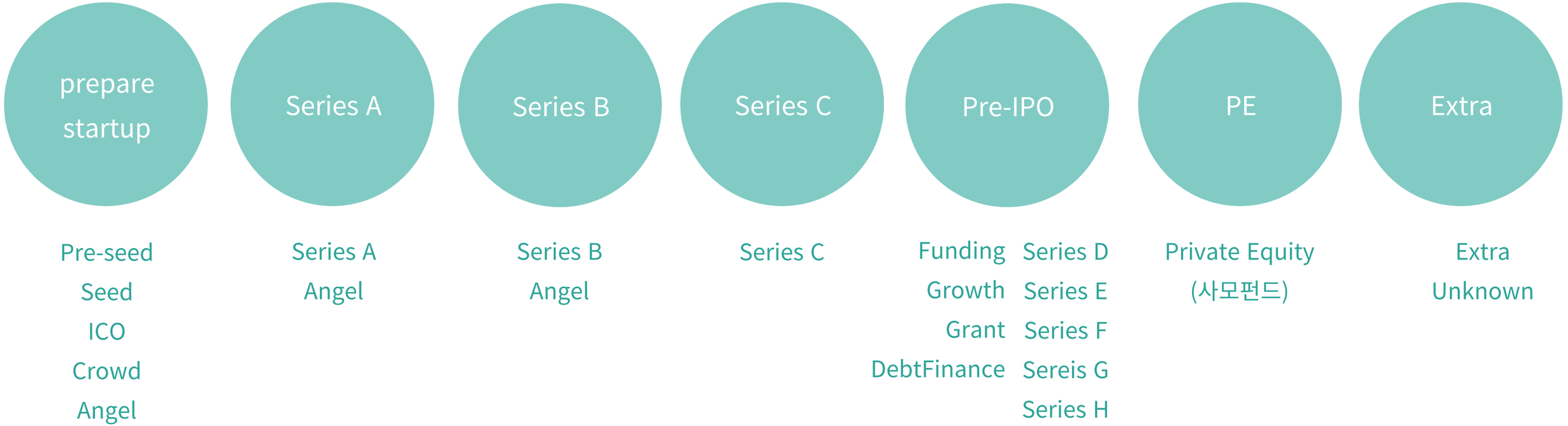
High category In Region Top 10

	index	Region
0	United States	1130
1	United Kingdom	153
2	Israel	75
3	Canada	65
4	Germany	53
5	France	46
6	India	37
7	Singapore	27
8	Australia	26
9	Switzerland	22

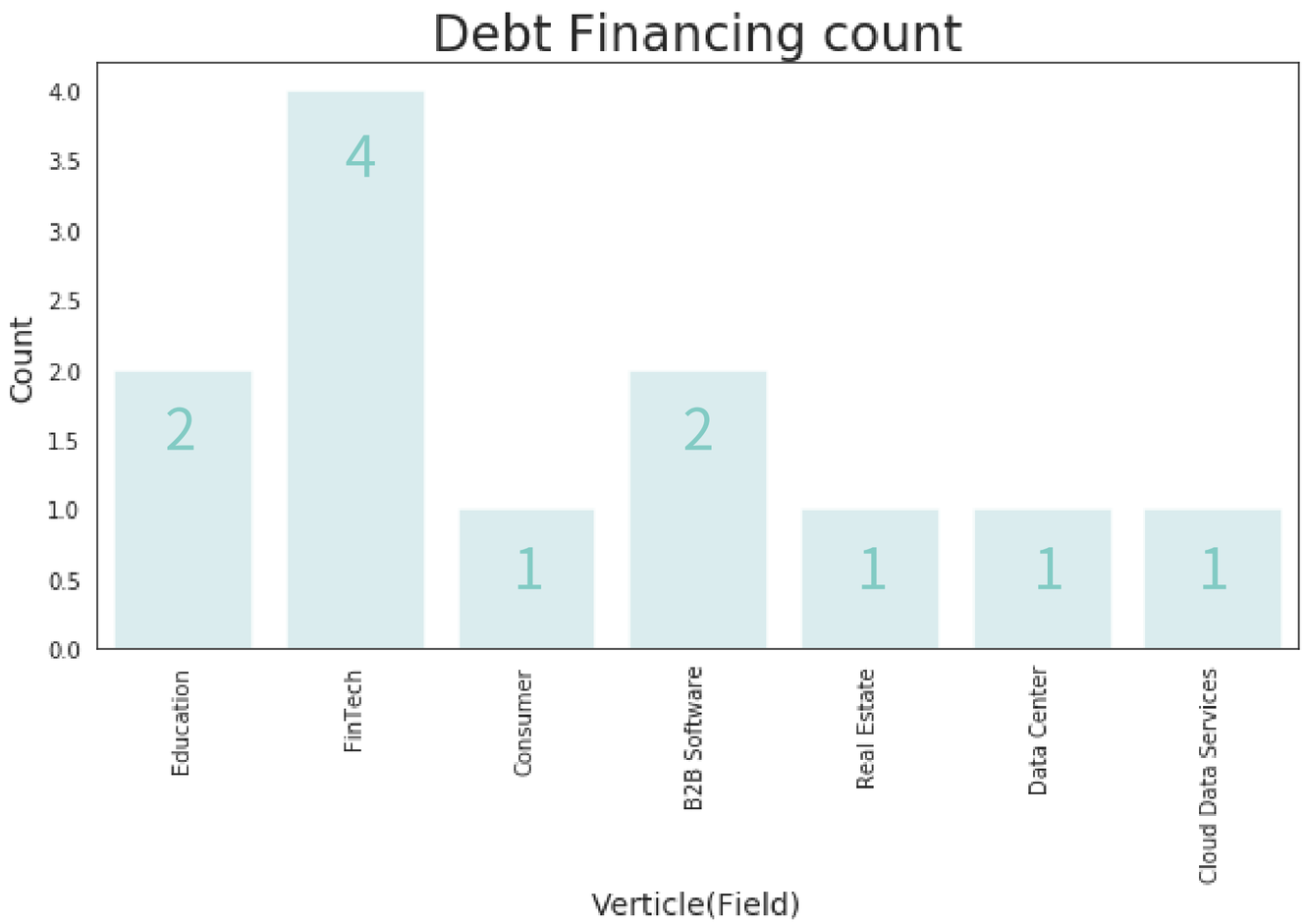


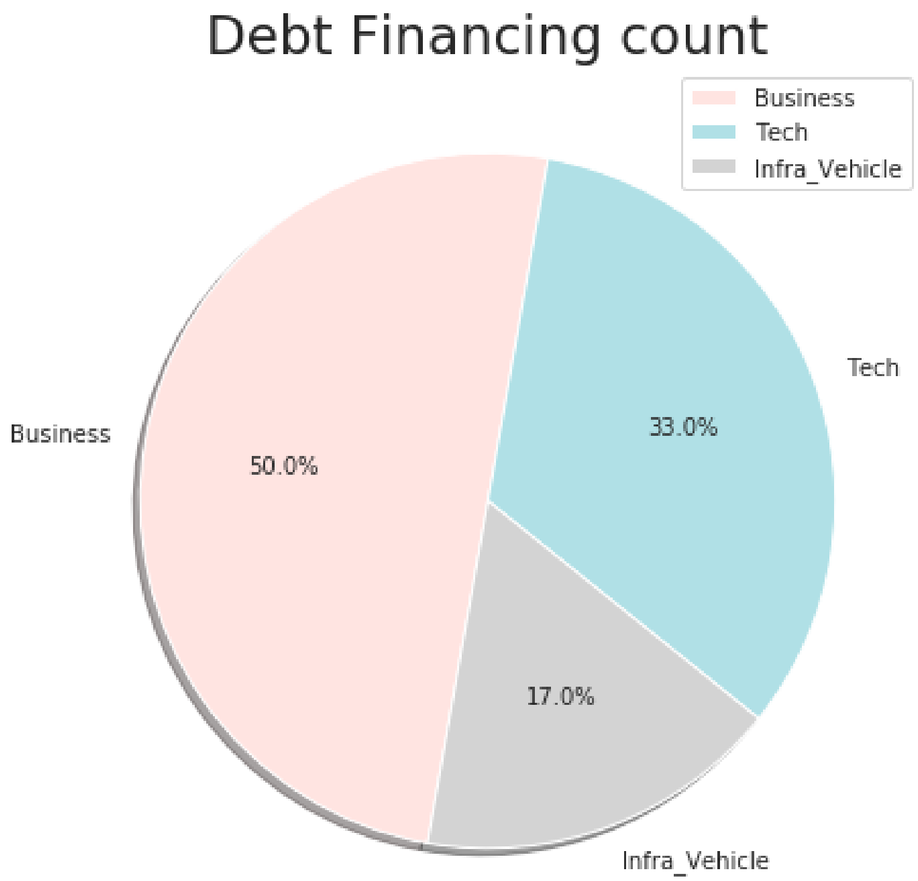
B3. Data wrangling 3. Funding stage

- 20개의 영역을 7개의 범주로 카테고리화하여 invest_pace feature 생성
- 다중공산성 방지를 위해 Funding_stage 제거



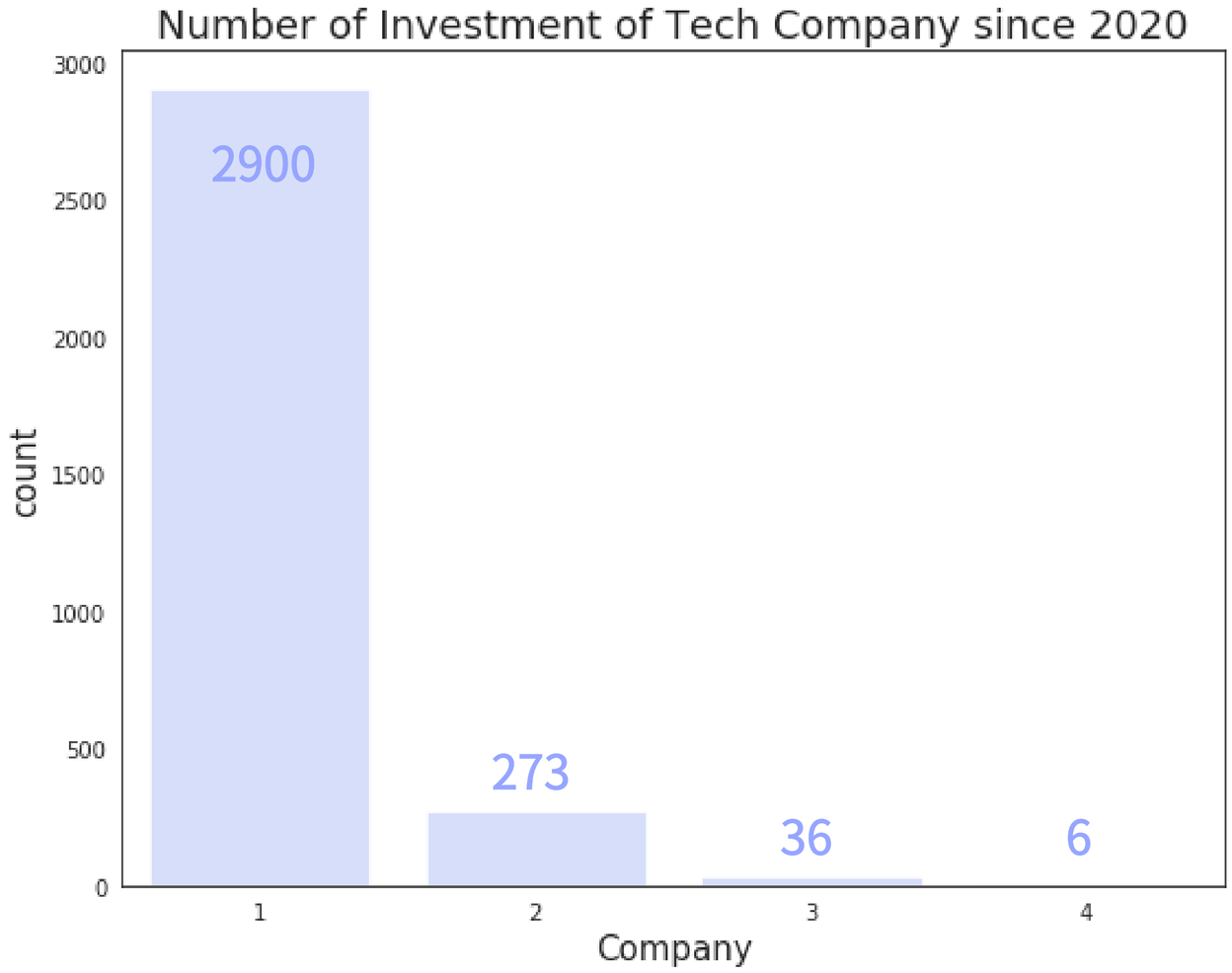
Debt Financing Fund





+ Company : insight

- 대부분의 기업이 한 개 영역에서 투자 받았음



Targeting

Series C 범주의 min 이상

- 국내에서 규모 형성
- 자체 수입으로 기업 영위 가능

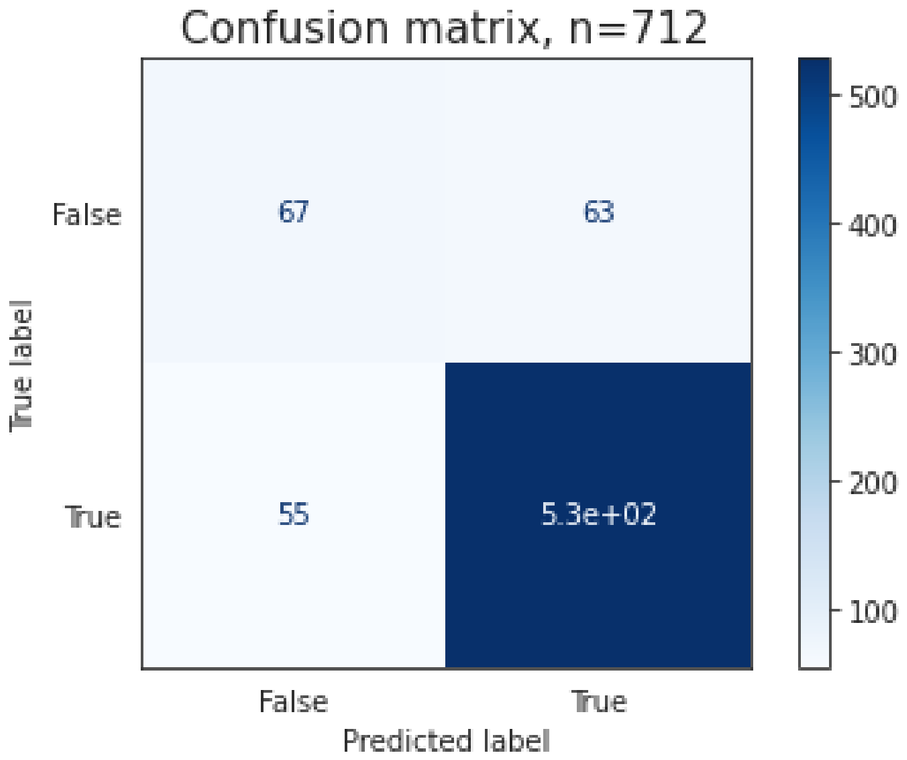
C1. Target 분포

- 다중공산성 방지를 위해 Funding Amount(USD) 항목 제거

True	0.82
False	0.18

C2. Data Leakage 여부 확인

- Data Leakage 확인을 위해 데이터셋을 Test, Val set으로 분리
- Oridinal Encoder와 결정트리를 이용하여 파이프라인 형성
- 검증정확도 83%
- 타겟의 분포가 True 82%로 불균형
- Confusion matrix, classification report, roc-auc curve 통해 확인



- TP가 굉장히 높지만, 다른 부분은 높지 않음
- Target의 범주 비율이 높은 비중으로 불균형 했던 것의 결과

```
from sklearn.metrics import classification_report
y_pred_pre = pipe.predict(X_val_pre)
print(classification_report(y_val_pre,y_pred_pre))
```

	precision	recall	f1-score	support
False	0.55	0.52	0.53	130
True	0.89	0.91	0.90	582
accuracy			0.83	712
macro avg	0.72	0.71	0.72	712
weighted avg	0.83	0.83	0.83	712

-정밀도(Precision) : True 범주 High, False 범주 Low

-재현율(Recall) : True 범주 High, False 범주 Low

- 본 모델은 재현율(Recall)이 더 중요

```
from sklearn.metrics import roc_auc_score
auc_score = roc_auc_score(y_val_pre, y_pred_pre)
auc_score
```

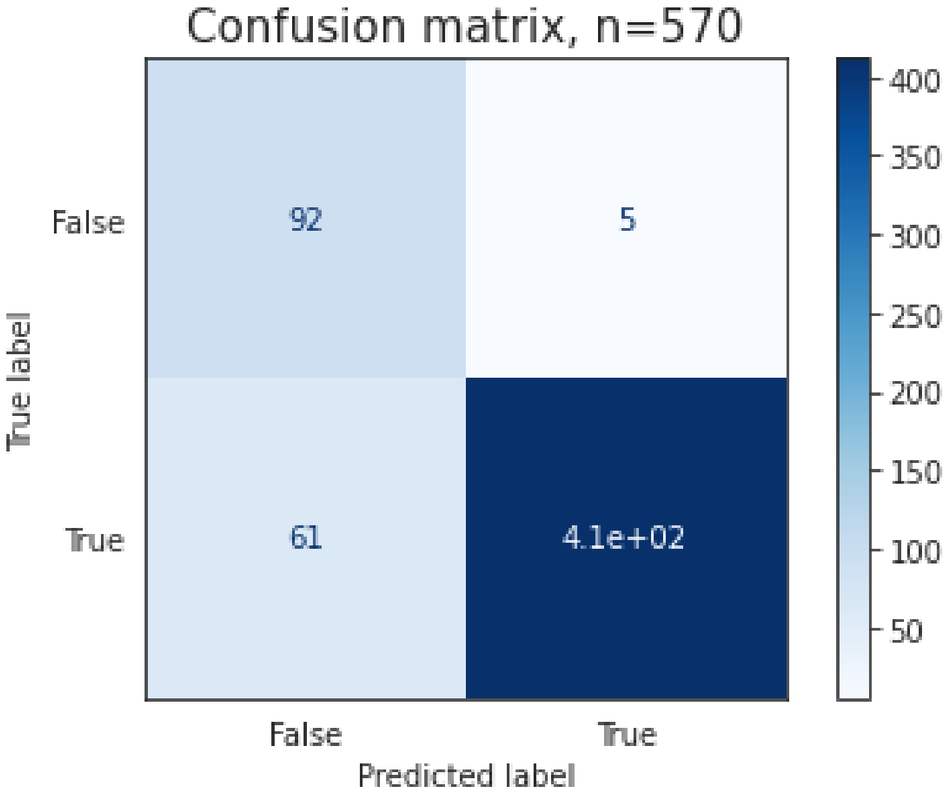
```
0.7104414485857784
```

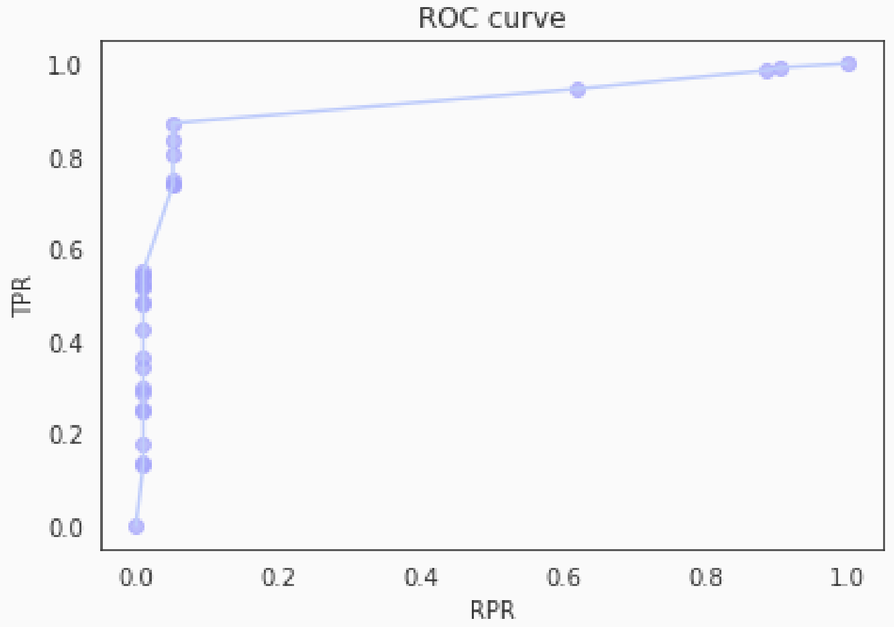
C3. Modeling

- Logistic Regression
- XG boost model

Logistic Regression Model & XG boost Model

- 검증 정확도: 88%
- AUC score : 0.91



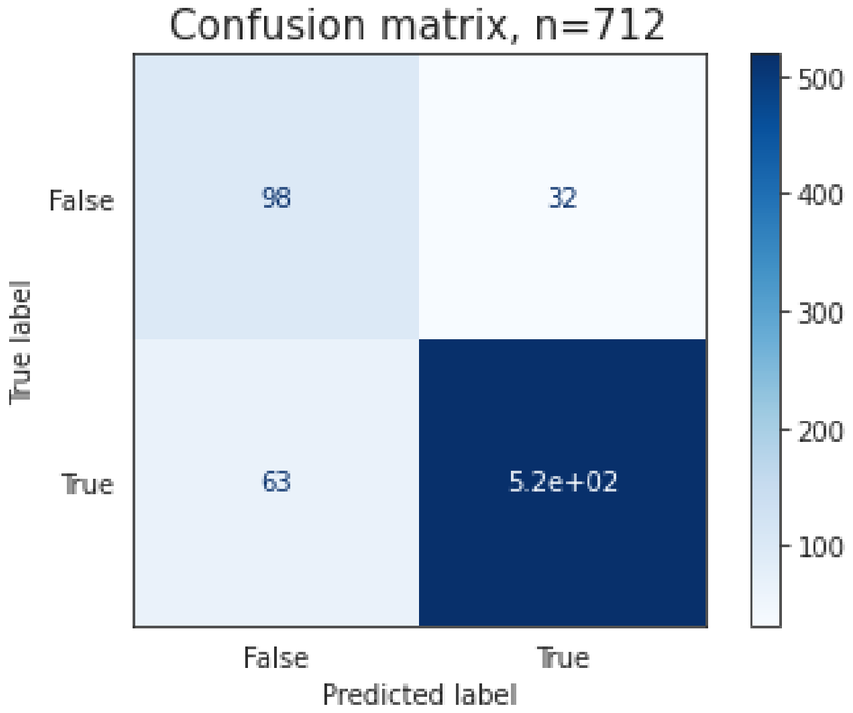


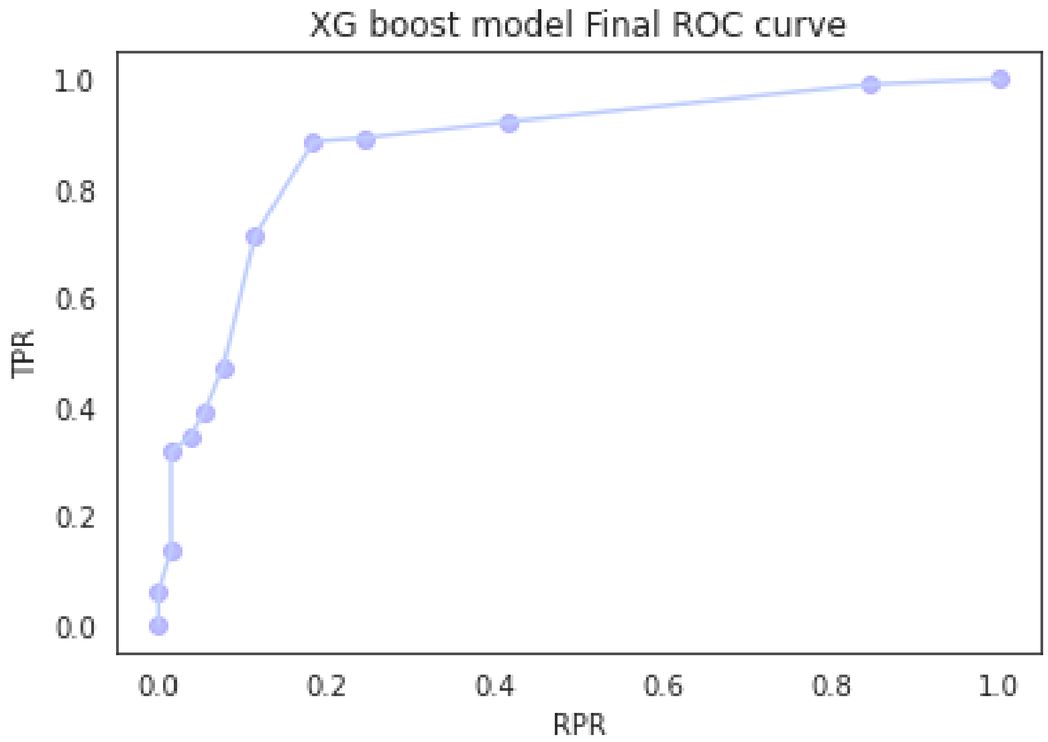
	precision	recall	f1-score	support
False	0.60	0.95	0.74	97
True	0.99	0.87	0.93	473
accuracy			0.88	570
macro avg	0.79	0.91	0.83	570
weighted avg	0.92	0.88	0.89	570

- 두 모델의 결과가 동일했다.
- 하이퍼파라미터를 조절해봤지만 변화가 없으므로 모델 학습을 진행하였다.

C4. 최종 모델 학습 (XG boost model)

- 검증 정확도: 86%
- AUC score : 0.875





기준모델 vs. 최종모델

- 정확도, precision, recall, f1-score 등 모든 영역에서 향상됨

기준모델 (Decision Tree)

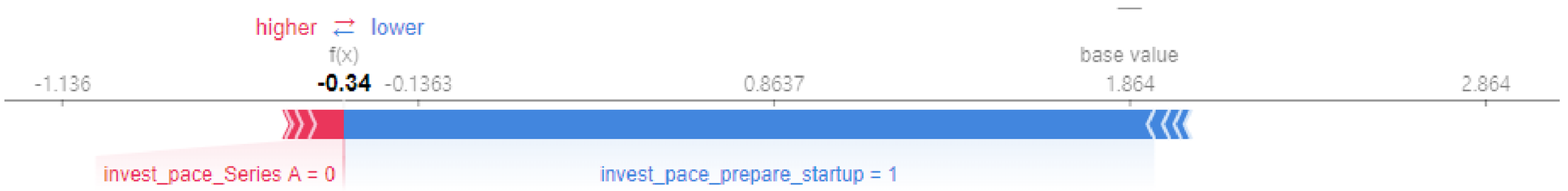
	precision	recall	f1-score	support
False	0.55	0.52	0.53	130
True	0.89	0.91	0.90	582
accuracy			0.83	712
macro avg	0.72	0.71	0.72	712
weighted avg	0.83	0.83	0.83	712

최종모델 (XG boost)

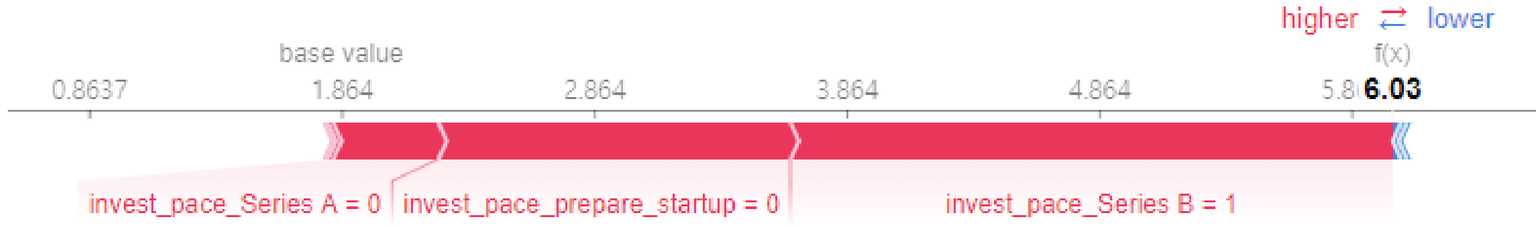
	precision	recall	f1-score	support
False	0.60	0.95	0.74	97
True	0.99	0.87	0.93	473
accuracy			0.88	570
macro avg	0.79	0.91	0.83	570
weighted avg	0.92	0.88	0.89	570

C5. 모델 해석_1. SHAP

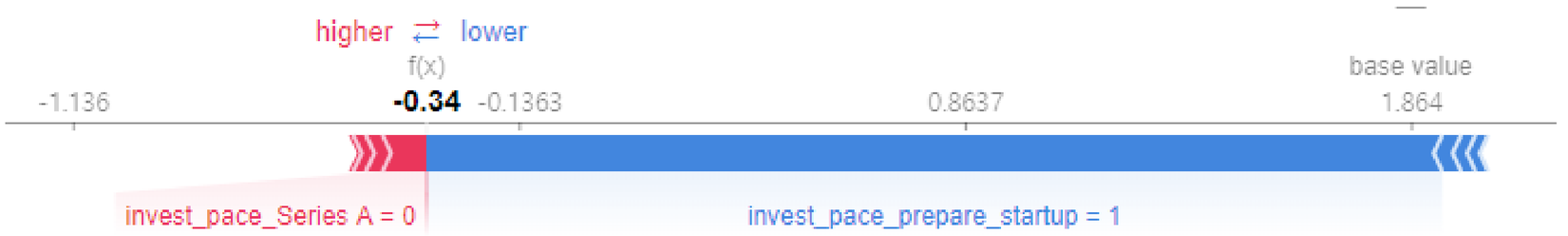
row1



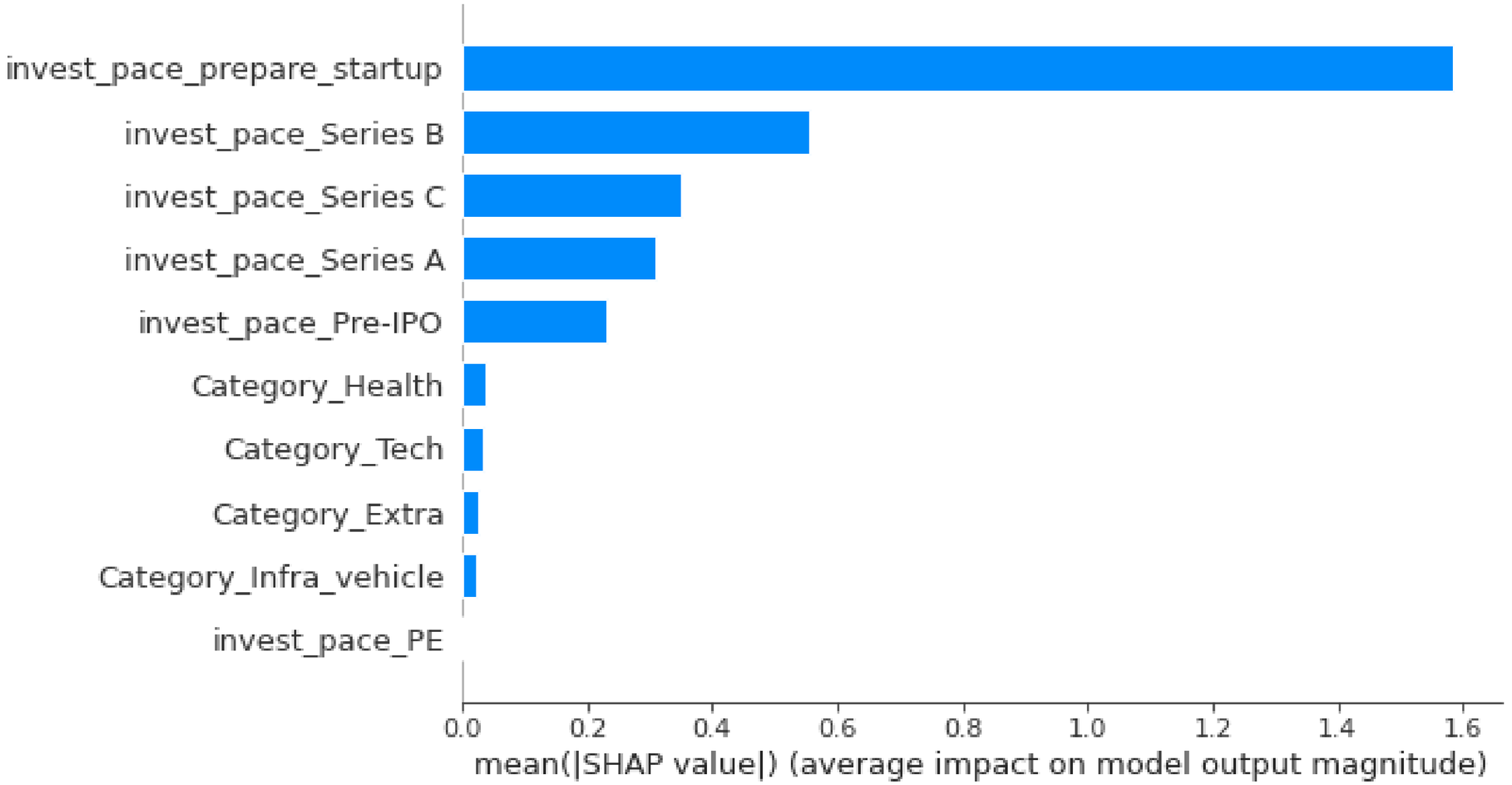
row2



row3



모델 해석_2. SHAP Value



모델 해석_3. 예측값 비교

- 예측확률과 실제값을 반영해서 데이터프레임 생성
- 이후 모델에 사용했던 데이터프레임에 합쳐 예측 확인