

Appendix D. 데이터 정의 및 기타

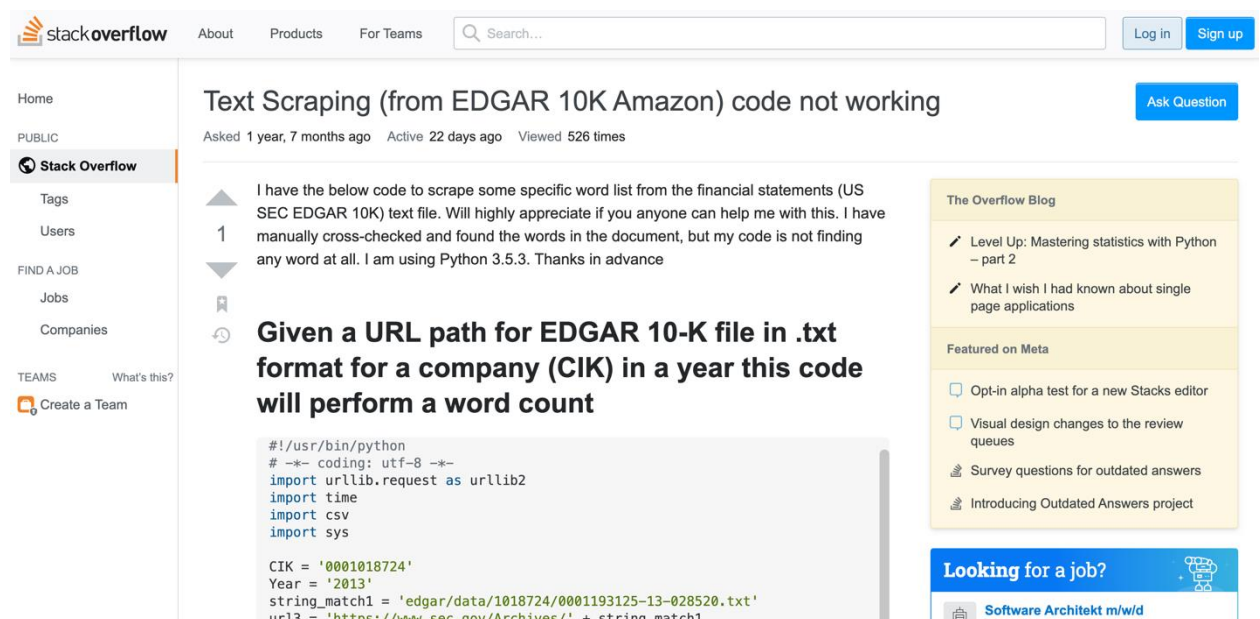
이번 부록에는 스택 오버플로 사이트 이용 팁, 자주 쓰이는 파이썬 기능 요약, 그리고 중급과 고급 프로젝트에 쓰인 데이터 정의 및 한영 변수명 전환표를 담았습니다.

D.1 스택 오버플로 이용 팁

책에서 종종 다음과 같이 변수에 숫자 형태가 아닌 값이 존재하는지 알아보는 코드를 선보인 바 있습니다.

```
# 숫자 형태가 아닌 SERIALNO 의 개수를 세는 파이썬 코드
mask = pd.to_numeric(df['SERIALNO'], errors='coerce').isna()
a = mask.sum()
print (a)
```

이 코드는 저자인 제가 스택 오버플로 사이트(www.stackoverflow.com)에서 여러 가지 대안을 검색해보고, 실제로 실행해본 것 중에서 가장 효율적인 코드를 채택한 것입니다. 참고로 스택 오버플로는 프로그래밍 언어와 관련해 질의응답을 하는 사이트입니다.



[그림 D-1] StackOverflow 화면 예시

다만, 스택 오버플로 사이트를 일부러 찾아가 질문을 할 필요는 없습니다. 구글에 본인이 궁금한 사항을 영어로 검색하면 이미 상당히 많은 검색 결과가 스택 오버플로에 있는 질의응답 내용일 것입니다. 때문에 다음과 같은 절차로 스택 오버플로를 이용하면 됩니다.

1. 구글 검색창에 간단한 영어로 질의합니다. 영어 단어 네다섯 개 정도면 충분합니다.
2. 가급적 관련된 라이브러리(예: pandas, sklearn 등)를 질문에 포함시킵니다.
3. 스택 오버플로 출처에서 나온 구글 결과 링크를 클릭합니다.
4. 1~2줄의 코드로 구성된 답변을 찾습니다.
5. for loop나 while loop로 구성된 답변은 프로그래밍 초보 시절엔 피하는 것을 권장합니다.
6. 최대한 데이터프레임(dataframe)을 대상으로 문제를 해결한 답변을 찾습니다.
7. 본인의 코드에 응용해봅니다.

영문으로 검색할 때는 문법이 맞지 않아도 상관없습니다. 'How to ~' 같은 문구도 넣어도 되지만, 더 짧게 쓰기 위해 아예 쓰지 않아도 상관없습니다. 예를 들어서 'How to select columns from a dataframe?'이라는 검색어를 쓰고 싶을 때는 'select columns from a dataframe'이라고만 입력해도 무방합니다. 물음표도 필요 없습니다. from a dataframe인지 from dataframe인지 헷갈리면 아무거나 입력해도 됩니다. 여기는 영어 문법을 체크하는 사이트가 아니므로 훌가분하게 입력하면 됩니다.

그리고 가급적 pandas, sklearn, numpy, matplotlib 등 관련된 라이브러리명이나 python, tensorflow, keras 등을 같이 입력해주면 좋습니다. 예를 들어 위의 질문을 'select columns from a dataframe pandas'로 바꾸면 더욱 좋습니다. 프로그래밍 언어명이나 라이브러리명을 질문에 병기하지 않으면 다른 프로그래밍 언어(예: R 언어)에 관련된 답변이 섞여 나와서 조금은 혼란스러운 답변 리스트가 검색되기 때문에 그렇습니다.

경험상 파이썬 관련 질의를 하면 구글에서 50~70퍼센트 정도는 스택 오버플로 출처의 답변이 검색됩니다. 이를 클릭해서 스택 오버플로의 질의응답 화면이 나타나면 그때부터 고민이 시작됩니다. 일단 원 질문자의 질문 사항을 읽는 것 자체가 시간이 오래 걸립니다. 이때 중요한 팁은 **질문에서는 제목과 예러문만, 답변에서는 코드만 읽어보자는 것**입니다. 그 이유는 다음과 같습니다.

파이썬 초보 학습자는 파이썬의 각종 용어 자체가 헷갈리기 때문에 질문이나 답변 글 전체를 읽으면 오히려 혼란이 가중됩니다. 예를 들어 이럴 때는 시리즈(series)를 쓰지 말고 튜플(tuple)을 써야 한다는 답변 글을 읽으면 시리즈는 뭐고 튜플은 또 뭔지 기억이 안 나서 아예 글 자체가 이해가 안 됩니다. 영어를 잘하지 못하는 것과는 또 다른 차원의 괴로움입니다.

때문에 질문한 사람의 글 제목만 읽어보고 (질문자의 코드가 이해되지 않아도 큰 문제는 안 됩니다) 이 사람이 질문하는 의도만 대강 파악한 상태에서 예러 구문만 파악해둡니다. 그리고 답변자의 코드만 읽어봅니다.

머신러닝/딥러닝 초보자 수준에서는 답변자의 코드가 5줄, 10줄 정도 되면 벌써 심적으로 엄청난 부담이 생깁니다. 다행인 것은 우리가 원하는 답변은 대부분 3줄 이상 넘어가지 않는다는 것입니다. 저자의 경험상 그렇습니다. 때문에 1줄, 2줄, 그리고 최대 3줄짜리 답변을 찾아서 코드만 읽어보는 것이 효율적입니다. 그리고 이해가 되면 본인의 코드에 응용하면 됩니다. 파이썬이 프로그래

밍 언어로서 가진 장점 중의 하나는 해석이 쉽다는 것입니다. 짧은 코드는 열심히 노력하면 웬만하면 해석할 수 있습니다. 텐서플로나 케라스도 다른 프로그래밍 언어에 비하면 그렇게까지 어렵지 않습니다.

그리고 다음 예시는 저자의 경험담입니다. 제가 MSBA 초기 단계에 for loop나 while loop로 구성된 스택 오버플로의 답변을 보면 답변자의 실력이 어마어마하게 뛰어난 것 같은 착각이 들었습니다. 그러나 loop 같은 반복 구문이 들어가면 프로그래밍 구문이 5줄, 10줄 넘는 것은 순식간입니다. 그리고 해석도 어렵습니다. 물론 해석이 불가능한 정도까지는 아닙니다만, 저자는 MSBA 초기 시절에 시간도 없고 해석도 어렵고 해서 for loop나 while loop가 포함된 답변은 최대한 피했습니다. 그런데 시간이 흘러보니까 파이썬 초기 학습 단계에서 이들 반복 구문을 피한 것이 대단히 좋은 습관이었다는 것을 나중에 알게 되었습니다.

이를 깨닫게 된 자세한 과정은 다음과 같습니다. 제 경험상, 대부분의 경우 파이썬의 판다스 라이브러리가 for loop나 while loop 구문이 들어간 파이썬 구문을 불필요하게 만듭니다. 판다스 라이브러리는 데이터 처리를 주로 데이터프레임 내에서 해결합니다. 데이터프레임 내에서 하는 데이터 처리 과정은 반복 구문을 사용하는 것보다 간단한 명령어 세트를 가지고 있습니다.

시중에 있는 파이썬 책 중에서 최근에 나온 책들은 간단한 코딩으로 구성되어 있는데, 수년 전에 나온 어떤 책들은 매우 긴 코드로 실려 있는 경우가 있습니다. 코드의 길이 차이는 판다스 라이브러리를 집중적으로 썼는가 아닌가에 크게 달려있습니다. 데이터프레임을 최대한 활용해서 문제를 풀다 보면 반복 구문의 사용이 불필요해지는 경우가 태반이기 때문입니다.

물론 for loop 혹은 while loop 구문이 꼭 필요할 때가 있는데, 그건 나중에 자연스럽게 필요성을 느끼게 되었을 때 검색해서 사용해도 늦지 않습니다. 때문에 초보자는 일단 for loop나 while loop 구문을 피하는 게 상책입니다. 머신러닝/딥러닝 데이터 처리에 쓰인 파이썬 구문이나 텐서플로 구문이 충분히 이해가 되면 초급을 벗어나 중급을 향해 달려가는 실력이 갖추었다고 볼 수 있습니다. 이때는 for loop나 while loop 구문도 두려움 없이 찾아보기 바랍니다.

여기에서 제시한 스택 오버플로 팁을 잘 숙지하고 활용해서 여러분의 머신러닝/딥러닝 프로그래밍 과정에 응용을 몇 번 해보면 스택 오버플로 검색 및 활용 과정이 그리 어렵게 느껴지지 않을 것입니다. 당분간은 답변에 있는 코드만 빨리 훑어보는 게 더 빠르고 효율적일 수 있습니다.

D.2 자주 쓰이는 파이썬 기능 요약

여기서는 책에서 자주 쓰이는 파이썬 기능을 요약해 두었습니다. 주로 데이터 처리와 탐색적 자료 분석에 쓰이는 파이썬 코드입니다. 파이썬 코드는 혹시나 외우지 못하더라도 괜찮습니다. 파이썬의 자주 쓰이는 기능들이 생각나지 않을 때 이 부록의 코드들을 찾아보면 큰 도움이 될 것입니다. 먼저, 데이터 파일을 불러오는 구문입니다.

```
# pd.read_csv()
# 원본 데이터 abc.csv 파일 불러오기
# pandas 와 numpy 라이브러리는 처음부터 같이 불러오는 것을 권장
import pandas as pd
import numpy as np
df = pd.read_csv('abc.csv')
```

그리고 다음은 16개의 자주 쓰이는 파이썬 명령어와 사용 예시를 알파벳 순으로 정리하였습니다. 단, df라는 명령어는 공통으로 쓰이므로 이를 제외한 명령문을 알파벳 순으로 정렬하였습니다.

```
# df.astype()
# 컬럼(변수) A의 dtype을 np.int64로 교체
df['A'] = df['A'].astype(np.int64)
```

```
# df.describe()
# 컬럼 A의 요약 통계 디스플레이
df['A'].describe()
```

```
# df.drop()
# 컬럼 C를 데이터프레임 df에서 제거
df.drop('C', axis=1, inplace=True)
```

```
# df.dtypes
# 컬럼(변수) A의 dtype 디스플레이
df['A'].dtypes
```

```
# df.fillna(0).astype()
# 컬럼(변수) A의 결측값을 0으로 메우고 dtype 교체
df['A'] = df['A'].fillna(0).astype(np.int64)
```

```
# df.head()
# 데이터프레임을 3행까지 디스플레이. 숫자가 없으면 5행까지 디스플레이
df.head(3)
```

```
# df.isna()
# 결측값을 갖고 있는 변수명(컬럼명) 찾기
df.isna().any()[lambda x: x]
```

```
# df.isnull().mean()
# 데이터프레임 내 모든 변수의 결측값 비율을 내림차순으로 디스플레이
df.isnull().mean().sort_values(ascending=False)
```

```
# df.isnull.sum()
# 컬럼 A의 결측값 개수 디스플레이
df['A'].isnull().sum()
```

```

# list(df.columns)
# 데이터프레임 df 내 모든 컬럼명 디스플레이
list(df.columns)

# pd.DataFrame()
# Series 형식인 데이터 d를 데이터프레임 df로 만들기
df = pd.DataFrame('d')

# df.replace()
# 컬럼 A의 점(.)을 결측값으로 표기
df['A'] = df['A'].replace('.', np.nan)

# df.shape
# 데이터프레임 df의 행과 열 차원 디스플레이
df.shape

# time.time()
# 런타임 디스플레이

import time
start = time.time()

# 런타임을 알고자 하는 프로그래밍 구문 삽입

end = time.time()
print(f"Runtime of the program is {end - start}")

# df.to_csv()
# 데이터프레임 df를 abc.csv로 저장
df.to_csv('abc.csv', index=False)

# df.value_counts()
# 컬럼 A의 값별로 개수를 디스플레이
df['A'].value_counts(dropna=False)

# 컬럼 A의 값별로 개수 비율을 디스플레이
df['A'].value_counts(dropna=False, normalize=True)

```

다음 두 파이썬 명령어는 데이터프레임을 병합(merging)하는 데 쓰입니다. 실은 병합 관련 명령어가 몇 개 더 있습니다. 그러나 `pd.concat()` 명령어가 거의 모든 종류의 병합을 처리할 수 있어서 가급적 이 명령어를 쓰면 좋습니다. `append()` 명령어는 가장 단순한 병합 명령어로서, 변수의 개수와 열 위치가 동일한 데이터프레임을 위아래 수직 방향으로만 병합할 수 있습니다.

```
# df1.append(df2)
# 데이터프레임 df1 과 df2 를 위아래(수직) 방향으로 병합(merging)하여 df3 로 저장
df3 = df1.append(df2)
```

```
# pd.concat(df1, df2, axis=1)
# 데이터프레임 df1 과 df2 를 위아래(수직) 방향으로 병합하여 df3 로 저장
df3 = pd.concat(df1, df2, axis=0)
```

```
# 데이터프레임 df1 과 df2 를 옆(수평) 방향으로 병합하여 df3 로 저장
df3 = pd.concat(df1, df2, axis=1)
```

마지막으로 다음 두 파이썬 명령어는 타깃 변수를 만들어내야 하는 상황이거나, imputation indicator(변숫값 중에 결측값이 있으면 1, 없으면 0을 표시하는 새로운 변수)를 만들어낼 때 필요한 명령어입니다. 이는 책과 부록에 있는 모든 프로젝트를 통틀어서 1~2회 정도밖에 쓰이지 않지만, 매우 중요한 코드이므로 여러분이 향후 편리하게 찾아볼 수 있도록 여기에 작성하였습니다.

```
# df.loc
# 기존 변수를 활용하여 새로운 타깃 변수 생성
df.loc[df['A'] >= 2, "Target"] = 1 # 변수 A 값이 2 이상이면 타깃 변수 Target 값은 1
df.loc[df['A'] < 2, "Target"] = 0 # 변수 A 값이 2 미만이면 타깃 변수 Target 값은 1
# 참고로 위의 예에서 숫자 2 는 A 값의 범위 내의 그 어떤 숫자로 대체해도 되나,
# 대체적으로 변수 A 의 분포가 정규분포에 가까우면 평균, 그렇지 않으면 중위수를 선택해서 입력하는 것이 무난함
```

```
# np.where()
df['A_indicator'] = np.where(df['A'].isnull(), 1, 0)
# SimpleImputer()의 add_indicator=True 기능으로 missing value indicator 를 자동으로
# 생성하지 않고, 데이터프레임에서 직접 missing value indicator 를 수동으로 작성할 때 쓰임
```

여기까지의 코드들이 얼핏 보기에 별거 아니게 보일 수도 있습니다. 그러나 책에서 다룬 프로젝트를 다시 들여다보면 이 파이썬 명령어들이 엄청난 빈도로 쓰였음을 확인할 수 있을 것입니다. 여기서 요약한 핵심 파이썬 코드를 기반으로 여러분이 유용하다고 판단하는 코드를 몇 개씩 늘려갔으면 하는 바람입니다. 여러분이 실전에서 고생해서 얻은 귀중한 코드들을 여기에 추가하면 금상첨화일 것입니다.

머신러닝 팁!

저자의 경험상, 시각화 및 머신러닝/딥러닝 모델 구축 시 프로그래밍 구문은 필요할 때마다 관련 책을 참조해서 찾아 써도 큰 무리가 없습니다. 이 분야는 코딩이 좀 복잡하고 길 뿐, 대부분 동일한 코드를 재활용해서 사용하기 때문입니다. 즉, 시각화 및 머신러닝/딥러닝 모델 구축 부분의 코드를 못 외운다고 좌절하지 않아도 됩니다. 필요할 때 인터넷이나 책에서 찾아 쓰면 충분히 대응할 수 있습니다.

D.3 데이터 정의 및 한영 변수명 변환표

책 4장의 중급 프로젝트에서 쓰인 전체 데이터의 정의는 다음과 같습니다.

Variable Name	Description	Role	Measurement Level	Reason for Rejecting
ACCESS	인터넷 접속 Access to the Internet 1 .Yes, by paying a cell phone company or Internet service provider 2.Yes, without paying a cell phone company or Internet service provider 3 .No access to the Internet at this house, apartment, or mobile home	Input	Nominal	
ACR	부지 크기 Lot size 1. House on less than one acre 2. House on one to less than ten acres 3. House on ten or more acres	Input	Nominal	
ADJHSG	집값 조정 계수(인플레이션 반영) Adjustment factor for housing dollar amounts	Rejected	Interval	An adjustment factor,
ADJINC	소득 조정 계수(인플레이션 반영) Adjustment factor for income and earnings dollar amounts	Rejected	Interval	An adjustment factor,
AGEP	나이 Age	Input	Interval	
BATH	욕조/샤워 시설 유무 If the property has a bathtub or shower 1. Yes 2. No	Input	Binary	
BDSP	침실 수 Number of bedrooms	Input	Interval	
BLD	주택 타입 Units in structure * 8장에서는 2번과 3번 타입만 분석 2. One-family house detached 3. One-family house attached * 데이터에 포함되어 있는 타 주택 타입 1. Mobile home or trailer 4. 2 Apartments 5. 3-4 Apartments 6. 5-9 Apartments 7. 10-19 Apartments 8. 20-49 Apartments 9. 50 or more apartments 10. Boat, RV, van, etc.	Input	Nominal	
BROADBND	휴대전화 데이터 플랜 보유 여부 Cellular data plan for a smartphone or other mobile device 0. No 1. Yes	Rejected	Binary	Rejected as we have ACCESS column

CONP	월 주택 수선 비용 Monthly condo fee	Input	Interval	
COW	근로자 타입 Class of worker 1. Employee of a private for-profit company or business. 2. Employee of a private not-for-profit organization 3. Local government employee 4. State government employee 5. Federal government employee 6. Self-employed in own not incorporated business, professional practice, or farm 7. Self-employed in own incorporated business, professional practice or farm 8. Working without pay in family business or farm 9. Unemployed and last worked 5 years ago or earlier or never worked	Input	Nominal	
ELEP	월 전기료 Monthly electricity cost	Input	Interval	
ENG	영어 구사 능력(세부 측정) Ability to speak English 1. Very well 2. Well 3. Not well 4. Not at all	Rejected	Nominal	Rejected as we have LANX column
FENGP	영어 구사 능력(간단 분류 측정) Ability to speak English 0. No 1. Yes	Rejected	Binary	Rejected as we have LANX column
FER	최근 12개월간 아이 출생 여부 Gave birth to child within the past 12 months 0. No 1. Yes	Rejected	Binary	Rejected as we have FPARC column
FESRP	고용 상태 Employment status recode 0. No 1. Yes	Input	Binary	
FKITP	부엌 설비 완비 여부 Complete kitchen facilities 0. No 1. Yes	Input	Binary	
FMARHYP	최근 1년간 결혼 유무 Year last married 0. No 1. Yes	Rejected	Binary	Rejected as we have MAR column
FMARP	결혼 유무 Marital status 0. No 1. Yes	Rejected	Binary	Rejected as we have MAR column
FPARC	가족 구성 Family presence and age of related children 1. With related children under 5 years only 2. With related children 5 to 17 years only 3. With related children under 5 years and 5 to 17 years 4. No related children	Input	Nominal	

FSCHP	재학 여부 School enrollment 0. No 1.Yes	Input	Binary	
FTAXP	연 재산세 납부 여부 Property taxes (yearly amount) 0. No 1.Yes	Input	Binary	
GASP	월 가스비 Gas (monthly cost)	Input	Interval	
HINCP	가계 소득 House income	Input	Interval	
HHL	가정내에서 쓰는 언어 Household language 1. English only 2. Spanish 3. Other Indo-European languages 4. Asian and Pacific Island languages 5. Other language	Input	Nominal	
HHT	가계 타입 Household/family type 1. Married couple household 2. Other family household: Male householder, no spouse present 3. Other family household: Female householder, no spouse present 4. Nonfamily household: Male householder: Living alone 5. Nonfamily household: Male householder: Not living alone. 6. Nonfamily household: Female householder: Living alone 7. Nonfamily household: Female householder: Not living alone	Input	Nominal	
INSP	연간 화재/상해/침수 보험료 Fire/hazard/flood insurance (yearly amount)	Rejected	Interval	Rejected as INSP got influenced by VALP_B1.
LANX	가정에서 쓰는 영업 외 타 언어 사용 유무 Language other than English spoken at home 1. Yes, speaks another language. 2. No, speaks only English.	Input	Binary	
MAR	결혼 유무(세부 분류) Marital status 1. Married 2. Widowed 3. Divorced 4. Separated 5. Never married or under 15 years old	Input	Nominal	
MV	현재 거주하는 집에 이사들어온 시기 When moved into the house or apartment 1.12 months or less 2. 13 to 23 months 3. 2 to 4 years 4. 5 to 9 years 5. 10 to 19 years 6. 20 to 29 years	Input	Nominal	

	7. 30 years or more			
NPF	가족 구성원 수 Number of persons in family	Rejected	Interval	Rejected as we have NRC column
NRC	가계내 (관련) 어린이 수 Number of related children in household	Input	Interval	
PERNP	총 개인소득(earning) Total person's earnings	Rejected	Interval	Rejected as we have HINCP column
PINCP	총 개인소득(income) Total person's income	Rejected	Interval	Rejected as we have HINCP column
PUMA	퍼블릭 유즈 마이크로데이터 지역 코드 Public use microdata area code	Rejected	Nominal	Rejected as we are only considering data from Oklahoma
R18	가계내 18세 이하 유무 Individuals under 18 years old in household 0. No person under 18 in household 1. 1 or more persons under 18 in household	Input	Binary	
R65	가계내 65세 이상 유무 Individuals of 65 years and over in household 0. No person 65 and over 1. 1 person 65 and over 2. 2 or more persons 65 and over	Input	Nominal	
RAC1P	상세 인종 코드 Recoded detailed race code 1. White alone 2. Black or African American alone 3. American Indian alone 4. Alaska Native alone 5. American Indian and Alaska Native tribes specified; or American Indian or Alaska Native, not specified and no other races 6. Asian alone 7. Native Hawaiian and Other Pacific Islander alone 8. Some Other Race alone 9. Two or More Races	Input	Nominal	
RMSP	방 수 Number of rooms	Input	Interval	
RT	기록 타입 Record Type H. Housing Record or Group Quarters Unit P. Person Record	Rejected	Binary	Rejected due to the inner join since we have person's table on the left.

RWAT	온수 및 냉수 시스템 보유 유무 If the property has hot and cold running water 1. Yes 2. No 9. Case is from Puerto Rico, RWAT not applicable.	Input	Binary	
SCH	재학 여부 School enrollment. 1. No, has not attended in the last 3 months 2. Yes, public school or public college 3. Yes, private school or college or home school	Input	Nominal	
SCHG	재학 학년 Grade level attending 1. Nursery school/preschool 2. Kindergarten 3. Grade 1 4. Grade 2 5. Grade 3 6. Grade 4 7. Grade 5 8. Grade 6 9. Grade 7 10. Grade 8 11. Grade 9 12. Grade 10 13. Grade 11 14. Grade 12 15. College undergraduate years 16. Graduate or professional school beyond a bachelor's degree	Rejected	Nominal	Rejected as we have SCHL column
SCHL	학업 성취 Educational attainment 1. No schooling completed 2. Nursery school, preschool 3. Kindergarten 4. Grade 1 5. Grade 2 6. Grade 3 7. Grade 4 8. Grade 5 9. Grade 6 10. Grade 7 11. Grade 8 12. Grade 9 13. Grade 10 14. Grade 11 15. 12th grade - no diploma 16. Regular high school diploma 17. GED or alternative credential 18. Some college, but less than 1 year 19. 1 or more years of college credit, no degree 20. Associate's degree 21. Bachelor's degree 22. Master's degree 23. Professional degree beyond a bachelor's degree 24. Doctorate degree	Input	Nominal	

SERIALNO	주택별 개인 시리얼 넘버 Person serial number of housing unit	Rejected	Nominal	Rejected since SERIALNO does not impact the outcome of our analysis.
SEX	성별 Gender 1. Male 2. Female	Input	Binary	
SRNT	렌탈 유닛 Specified rental unit 0. A single-family home on 10 or more acres. 1. A single-family home on less than 10 acres or any other type of building, including mobile homes, with no regard to acreage	Rejected	Unary	Rejected from being unary
TYPE	주택 타입 Type of housing unit 1. Housing unit 2. Institutional group quarters 3. Noninstitutional group quarters	Rejected	Unary	Rejected from being unary
VALP	주택 가격 Property value	Rejected	Interval	Replaced by VALP_B1
VALP_B1	주택 가격 이진 분류 Binary target derived from VALP (property value) 0. Less than the median value of VALP 1. More than or equal to the median value of VALP	Target	Binary	

다음은 부록에 있는 고급 프로젝트에 필요한 데이터 한영 변수명 변환표입니다. 범주형 변수들의 값에 대한 추가 정의는 고급 프로젝트를 진행하는 부록에 설명해두었습니다. 아울러 고용량 관련 변수명과 기업 간 전략적 제휴 관련 변수명은 각기 70여 개와 40여 개에 육박하나, 결측값이 50%를 초과하는 변수가 대부분이어서 모델에 투입된 변수를 제외하고는 아래 표에서 제외하였습니다.

한글 변수명	영문 변수명
행정구역(시도))	Area
재무구조_자산_총계	Asset1
재무구조_자산_당좌자산	Asset2
재무구조_자산_재고자산(일반업종)	Asset3
재무구조_자산_현금및예치금(금융보험업종)	Asset4
재무구조_자산_유가증권(금융보험업종)	Asset5
재무구조_자산_대출채권(금융보험업종)	Asset6
재무구조_자산_투자자산	Asset7
재무구조_자산_유형자산	Asset8
재무구조_자산_무형자산	Asset9
기업간거래_직접수출액_거래액	B2B_export1
기업간거래_직접수출액_관계회사거래액	B2B_export2
기업간거래_직접수입액_거래액	B2B_import1
기업간거래_직접수입액_관계회사거래액	B2B_import2
기업간거래_구입액_거래액	B2B_purchase1
기업간거래_구입액_관계회사거래액	B2B_purchase2
기업간거래_매출액_관계회사거래액	B2B_revenue
재무구조_자본총계	Capital1
재무구조_자본잉여금	Capital2
자본금중외국자본비율	CapitalRatio1
기업체패널키	CompanyID
성과보상제도_스톡옵션	Compensation1
성과보상제도_우리사주	Compensation2
성과보상제도_연봉제	Compensation3
성과보상제도_성과급	Compensation4
경영실적_매출원가,금융보험관련비용	Cost1
비용내역_영업비용_감가상각비	Cost10
비용내역_영업비용_세금과공과	Cost11
비용내역_영업비용_광고선전비	Cost12
비용내역_영업비용_대손상각비	Cost13
비용내역_영업비용_기타영업비용	Cost14
비용내역_영업외비용_계	Cost15

비용내역_영업외비용_이자비용(일반업종)	Cost16
비용내역_영업외비용_기부금	Cost17
경영실적_판매비와일반관리비	Cost2
경영실적_영업외비용	Cost3
비용내역_영업비용_계	Cost4
비용내역_영업비용_재료비(일반업종)	Cost5
비용내역_영업비용_금융비용(금융보험업종)	Cost6
비용내역_영업비용_보험비용(금융보험업종)	Cost7
비용내역_영업비용_인건비	Cost8
비용내역_영업비용_임차료	Cost9
디자인권_보유건수	Design1
디자인권_보유중활용건수	Design2
디자인권_활용중자사개발건수	Design3
EBiz시스템_없음	EBizSystem1
EBiz시스템_기타여부	EBizSystem10
EBiz시스템_(3 to 10 모두 0이면 0 아니면 1, variable newly made)	EBizSystem11
EBiz시스템_(3 to 10 합계 수치, variable newly made)	EBizSystem12
EBiz시스템_기업의자원관리	EBizSystem2
EBiz시스템_온라인교육	EBizSystem3
EBiz시스템_지식관리	EBizSystem4
EBiz시스템_인적자원관리	EBizSystem5
EBiz시스템_전자입찰	EBizSystem6
EBiz시스템_고객관리	EBizSystem7
EBiz시스템_공급망관리	EBizSystem8
EBiz시스템_전자상거래	EBizSystem9
종사자수_상용(여)	Emp_female
종사자수_상용(남)	Emp_male
종사자수_상용(남+여)	Emp_total
종사자수_상용_본사본점_본사기능(남+여)	emp6
종사자수_임시및일용근로자(남+여)	emp66
자본금(출자금)	EquityShareCapital
프랜차이즈가맹점_유치여부	Franchise1
프랜차이즈가맹점_향후추진방향	Franchise2
산업분류(대)	IndCategory1
산업분류(중)	IndCategory2
무형자산당기취득액	ITAssetC
재무구조_부채_총계	Liability1
재무구조_부채_유동부채	Liability2

재무구조_부채_비유동부채	Liability3
신규진출산업분류 1	NewEntry1
신규진출산업분류 2	NewEntry2
신규사업진출검토여부	NewEntry3
경영실적_법인세차감전순이익 (Profits)	NProfitB4Tax
외부위탁업무분야_없음	Outsourcing1
외부위탁업무분야_재무회계경리법무	Outsourcing10
외부위탁업무분야_중사자교육연수관리	Outsourcing11
외부위탁업무분야_중사자복리후생	Outsourcing12
외부위탁업무분야_기타여부	Outsourcing13
외부위탁업무분야_생산	Outsourcing2
외부위탁업무분야_정보처리	Outsourcing3
외부위탁업무분야_디자인제품기획	Outsourcing4
외부위탁업무분야_조사마케팅	Outsourcing5
외부위탁업무분야_연구개발	Outsourcing6
외부위탁업무분야_운송배송보관등 물류	Outsourcing7
외부위탁업무분야_경비청소시설관리	Outsourcing8
외부위탁업무분야_인사노무관리	Outsourcing9
외주비	OutsourcingCost
국외진출여부	Overseas1
국외진출방식_지사?지점?출장소	Overseas2
국외진출방식_공장	Overseas3
국외진출방식_법인투자	Overseas4
국외진출계획_계획없음	Overseas5
국외진출계획_지사?지점?출장소	Overseas6
국외진출계획_공장	Overseas7
국외진출계획_법인투자	Overseas8
국외진출계획_기업이전	Overseas9
모회사_유무	ParentCompany1
모회사_출자비율	ParentCompany2
모회사_국가분류부호	ParentCompany3
모회사_산업중분류	ParentCompany4
특허권_보유건수	Patent1
특허권_보유중활용건수	Patent2
특허권_활용중자사개발건수	Patent3
경영실적_매출액,영업수익 (Revenue or Sales)	RevProfit1
경영실적_영업외수익	RevProfit2
연구개발비_자체	RNDcost1

연구개발비_위탁	RNDcost2
연구개발비_위탁_관계회사비용	RNDcost3
연구개발비_수탁	RNDcost4
연구개발비_수탁_관계회사비용	RNDcost5
기업간전략적 제휴여부	SAlliance1
주식시장_상장형태	StockMktListing
자회사,관련회사_보유유무	Subsidiary1
자회사,관련회사_수_국내	Subsidiary2
자회사,관련회사_수_국외	Subsidiary3
자회사,관련회사_투자액_국내	Subsidiary4
자회사,관련회사_투자액_국외	Subsidiary5
유형자산당기취득액	TAssetC1
유형자산당기취득액_토지	TAssetC2
유형자산당기취득액_건물	TAssetC3
유형자산당기취득액_구축물	TAssetC4
유형자산당기취득액_기계장치	TAssetC5
유형자산당기취득액_선박차량운반구	TAssetC6
유형자산당기취득액_기타	TAssetC7
유형자산당기취득액_건설중인자산	TAssetC8
상표권_보유건수	TradeMark1
상표권_보유중활용건수	TradeMark2
상표권_활용중자사개발건수	TradeMark3
실용신안권_보유건수	UMR1
실용신안권_보유중활용건수	UMR2
실용신안권_활용중자사개발건수	UMR3
조사기준년도	Year