Nuree Chung Follow
May 17 · 9 min read

EDA, 데이터 설명서에서 시작하기

EDA(Exploratory Data Analysis, 탐색적 데이터 분석)는 본격적인 모델링에 들어가기에 앞서, 반드시 선행되어야 하는 과정입니다. 그렇다면 EDA는 어디에서부터 시작해야 할까요? 데이터의 분포나 변수간 관계를 파악하기 위해 히스토그램, 산점도, 상관관계표 등 다양한 시각화 방법이 동원됩니다. 하지만 그 전에, 가장 먼저 해야 할 일이 있습니다.

Please read a data description first (Image Credit : https://www.keepcalm-o-matic.co.uk/p/keep-calm-and-please-read/)

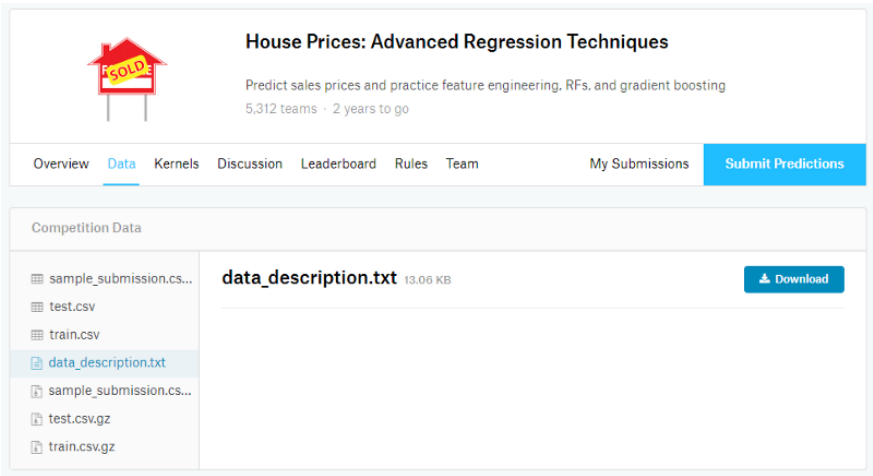
데이터 설명서를 잘 읽어보자

데이터 분석은 복잡한 수식이 들어간 멋진 모델링에서 시작되는 것도 아니고, 화려한 시각화에서 시작되는 것도 아닙니다. 시각화와 모델링을 할 수 있도록 데이터를 준비하는 과정, 즉 전처리를 하면서 데이터를

이해하는 작업이 필요합니다. 지루하고 따분한 과정이라고 생각할 수 있지만 데이터 꼼꼼히 들여다보며 각각의 변수가 어떤 의미인지, 어떤 방식으로 측정된 것인지, 명목형인지 수치형인지, 실제 의미대로 코딩되었는지 정확히 이해하는 이 과정은 데이터로부터 좋은 인사이트를 얻는 출발점이라고 할 수 있습니다.

물론 직접 데이터를 수집했다면 각각의 변수가 무엇을 나타내는지, 수치형 혹은 문자형으로 표현된 관측값이 무엇을 의미하는지 이미 알고 있을 수 있습니다. (그래도 시간이 지나면 거짓말처럼 내가 뭘 주워담은 건지 잊어버리게 되니 처음 데이터를 수집했을 때 데이터 설명서를 잘 써둬야 합니다.) 하지만 다른 사람이 수집한 데이터를 분석해야 할 때는 어떻게 해야 할까요? 예측모델 및 분석 대회 플랫폼으로 유명한 캐글(Kaggle)에 가보면 다양한 주제의 데이터를 구할 수 있습니다. 여기서 얻은 데이터로 모델링이나 시각화를 해보기 전, 가장 먼저 할 일은 바로 **'데이터 설명서(data description) 읽기'**입니다.

2년전 개최되어 지금까지 5,307팀이 참여한 "House Prices: Advanced Regression Techniques"라는 집값 예측 대회의 데이터를 봅시다.



Data 섹션에 들어가면 데이터셋과 함께 데이터 설명서(data_description.txt)를 찾을 수 있습니다

데이터 설명서를 잘 읽어보면 몇 가지 유형의 전처리가 필요하다는 점을 알 수 있습니다. 하나씩 살펴볼까요?

. . .

유형 1. 수치형으로 입력되어 있지만 실제는 명목형인 변수

```
MSSubClass: Identifies the type of dwelling involved in the sale.

        20      1-STORY 1946 & NEWER ALL STYLES
        30      1-STORY 1945 & OLDER
        40      1-STORY W/FINISHED ATTIC ALL AGES
        45      1-1/2 STORY - UNFINISHED ALL AGES
        50      1-1/2 STORY FINISHED ALL AGES
        60      2-STORY 1946 & NEWER
        70      2-STORY 1945 & OLDER
        75      2-1/2 STORY ALL AGES
        80      SPLIT OR MULTI-LEVEL
        85      SPLIT FOYER
        90      DUPLEX - ALL STYLES AND AGES
       120      1-STORY PUD (Planned Unit Development) - 1946 & NEWER
       150      1-1/2 STORY PUD - ALL AGES
       160      2-STORY PUD - 1946 & NEWER
       180      PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
       190      2 FAMILY CONVERSION - ALL STYLES AND AGES
```

MSSubClass 변수의 설명을 읽어봅시다. 입력된 수치는 아무런 의미가 없는 id 값임을 알 수 있습니다. 그럼 명목변수라고 생각하고 인코딩 해주면 될까요? 설명을 좀 더 자세히 읽어필요가 있습니다. 의미있는 변수인지 파악하기 위해 각 카테고리 별 빈도수 정보도 같이 볼까요?

Category	Description	Frequency
20	1-STORY 1946 & NEWER ALL STYLES	1079
30	1-STORY 1945 & OLDER	139
40	1-STORY W/FINISHED ATTIC ALL AGES	6
45	1-1/2 STORY - UNFINISHED ALL AGES	18
50	1-1/2 STORY FINISHED ALL AGES	287
60	2-STORY 1946 & NEWER	575
70	2-STORY 1945 & OLDER	128
75	2-1/2 STORY ALL AGES	23
80	SPLIT OR MULTI-LEVEL	118
85	SPLIT FOYER	48
90	DUPLEX - ALL STYLES AND AGES	109
120	1-STORY PUD (Planned Unit Development) - 1946 & NEWER	182
150	1-1/2 STORY PUD - ALL AGES	1
160	2-STORY PUD - 1946 & NEWER	128
180	PUD - MULTILEVEL - INCL SPLIT LEV/FOYER	17
190	2 FAMILY CONVERSION - ALL STYLES AND AGES	61

MSSubClass 변수 카테고리 별 설명과 빈도수

여기서 파악할 수 있는 주요 정보는 다음과 같습니다.

먼저 건축연도, 층수, 스타일 유형 등을 기준으로 집을 분류하고 있습니다. 그런데 이 정보는 모두 YearBuilt, HouseStyle, BldgType 변수가 담고 있는 정보입니다.

```
YearBuilt: Original construction date


HouseStyle: Style of dwelling

       1Story   One story
       1.5Fin   One and one-half story: 2nd level finished
       1.5Unf   One and one-half story: 2nd level unfinished
       2Story   Two story
       2.5Fin   Two and one-half story: 2nd level finished
       2.5Unf   Two and one-half story: 2nd level unfinished
       SFoyer   Split Foyer
       SLvl     Split Level

BldgType: Type of dwelling

       1Fam     Single-family Detached
       2FmCon   Two-family Conversion; originally built as one-family dwelling
       Duplx    Duplex
       TwnhsE   Townhouse End Unit
       TwnhsI   Townhouse Inside Unit
```

중복되는 정보이고 정보가 뭉쳐져있기 때문에 이 변수보다는 정보를 따로 따로 담고 있는 세 변수를 사용하는 것이 타당해보입니다.

그럼 이 변수는 삭제하는 것이 나을까요? 다시 잘 읽어보면 Planned Unit Development(PUD) 라는 정보가 보입니다. 데이터 설명서를 읽어보면 다른 변수에 나타나지 않은 정보임을 알 수 있습니다. 구글에 검색해보니 여러 건물들로 구성된 단지를 조성하는 것 같네요. 집값 예측에 영향을 미칠 수 있는 변수일 확률이 높아 보입니다. 빈도수 또한 총 328개나 됩니다. 이 변수를 삭제하는 대신 PUD라는 새로운 변수를 만들어 PUD인 집과 아닌 집을 1,0으로 인코딩해 볼 수도 있을 것입니다.

또한 건축년도 1946년을 기준으로 집을 분류하고 있음을 알 수 있습니다. 우리는 잘 모르지만 실무적으로 보았을 때 이 시점을 기준으로 집을 분류하는 것이 의미 있을 수 있다는 유추를 해볼 수 있을 것입니다. 추후에 건축연도(YearBuilt)변수와 집값(SalePrice) 변수의 관계를 살펴볼 때 이를 기억해뒀다가 1946년을 기준으로 유의한 차이가 나타나는지 확인해볼 수 있겠습니다.

유형2. 명목형으로 입력되어 있지만 실제는 순서상의 의미를 가지는 변수

데이터 설명서를 쭉 읽다보면 집안 곳곳의 시설에 대한 Quality나 Condition을 나타내는 변수가 많이 보입니다. Ex, Gd, TA, Fa, Po 등문자로 입력되어 있는데 설명을 읽어보면 척도간 등급이 존재함을 알 수 있습니다. 이런 변수는 명목형으로 두었다가 더미변수로 바꿔주는 원핫인코딩(one-hot encoding)을 하는 것보다 수치형 데이터로 바꿔주는 것이 적절합니다. 순서에 대한 정보(ordinal information)가 사라지기 때문입니다. 가령 ExterCond 변수는 1~5점 척도로 측정되었다고 보고 바꿔줄 수 있겠네요.

```
ExterQual: Evaluates the quality of the material on the exterior

       Ex       Excellent
       Gd       Good
       TA       Average/Typical
       Fa       Fair
       Po       Poor

ExterCond: Evaluates the present condition of the material on the exterior

       Ex       Excellent
       Gd       Good
       TA       Average/Typical
       Fa       Fair
       Po       Poor
```

유형3. 합쳐서 하나로 만들 수 있는 변수

두 변수로 나누어서 입력되어 있지만, 하나로 합쳐줄 수 있는 변수도 있습니다. 예를 들어 BsmtFullBath 나 BsmtHalfBath는 결국 지하실에 있는 화장실 수를 나타내는 변수로 하나로 합쳐도 무리가 없어보입니다. BsmtBathN 변수를 새롭게 만들어 BsmtFullBath가 1, BsmtHalfBath 가 1인 집은 1.5로, BsmtHalfBath만 1인 집은 0.5로 입력하는 것이죠. FullBath와 HalfBath 변수도 마찬가지입니다.

```
BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade
```

유형4. 쪼개서 나눌 수 있는 변수

반면에 하나의 변수로 측정되어 있지만, 둘로 나눌 수 있는 변수도 보입니다. 예를 들어 Fence 변수의 경우 Privacy와 Wood 변수를 따로 나눠서 FencePrivacy, FenceWood 등의 변수를 만들고, good 과minimum에 각각 점수를 부여 할 수 있을 것입니다.

```
Fence: Fence quality

       GdPrv    Good Privacy
       MnPrv    Minimum Privacy
       GdWo     Good Wood
       MnWw     Minimum Wood/Wire
       NA       No Fence
```

유형5. 결측값인지 0인지 헷갈리는 관측치

유형2의 변수들을 바꾸다보면, 상태 정보와 함께 해당 시설이 없는 경우 '시설 없음'이라는 카테고리가 있는 변수들이 있습니다. 예를 들어 GarageQual, BsmtQual, FireplaceQu등의 변수를 보면 No garage, No

basement, No fireplace 같은 해당 시설이 없음을 나타내는 카테고리가 보입니다. 데이터 설명서에 나오는대로 “NA”로 검색해보아도 해당 카테고리로 분류된 변수가 없는 것으로 나옵니다. 즉, 데이터 설명서에는 “no basement”라는 또 다른 척도가 사용된 것처럼 적혀있지만 데이터가 입력되는 과정에서 null로 입력된 것이죠. 정말 값이 누락된 것인지 (예: 다른 basement 변수는 값이 있는데 이 변수에만 값이 없는 경우) 지하실이 없는 것인지, 다른 변수들과 종합적으로 살펴볼 필요가 있겠습니다.

. . .

기초공사를 잘 하자

(Image Credit: http://www.jonomiller.com/year/2012/build-good-solid-foundation-800)

81개나 되는 변수가 있지만 본격적인 EDA 및 모델링에 들어가기에 앞서 변수 하나 하나를 살펴보며 제대로 인코딩해나가는 과정이 있어야 실제에 가까운, 더 많은 정보를 담은 데이터가 준비됩니다. 이렇게 변수의 유형을 나눠가며 인코딩하다보면 금방 데이터의 특성을 파악할 수 있을 것입니다.

아무리 화려한 집을 지어도 기초공사가 부실하면 쓰러지는 것처럼, 데이터 전처리를 제대로 하지 않는다면 멋진 모델을 만들더라도 무너지기 쉬울 겁니다. 데이터가 진짜 정보를 정확하게 담아내도록 해야 보다 정확한 모델링 결과를 얻을 수 있을 것이죠. 쓰레기가 들어가면 쓰레기가 나오기 마련이니까요.

Don't forget "Garbage in, garbage out"

자, 이제 본격적인 EDA를 해 볼 차례입니다. 다음 글에서는 기초통계량과 더불어 히스토그램, 산점도 등 시각화를 해보며 데이터를 좀 더 깊이 살펴보도록 하겠습니다.

감사합니다.

. . .

**Note: 본 글에서는 설명하고 있지는 않지만, 데이터 유형에 맞게 전처리 후 바로 바로 히스토그램 또는 박스 플롯 등을 통해 해당 변수의 분포나 특성을 파악해보는 것이 좋습니다.*

**Note: 탐색적 데이터 분석은 학습 데이터(train data) 만 가지고 해야 할까요? 아니면 홀드아웃 데이터셋(hold-out data set) 혹은테스트 데이터(test data) 와 합쳐서 해야 할까요? 데이터 전처리를 하는 지금 단계에서는 두 데이터를 합쳐서 같은 기준으로 데이터 전처리를 하는 것이 좋습니다. 이 단계에서 각 독립변수의 분포를 살펴볼 수도 있겠습니다. 하지만, 종속변수(이 데이터의 경우 SalePrice) 와 독립변수의 관계를 보는 EDA를 할 때는 테스트 데이터는 따로 떼어놓고, 들여다보지 않을 것을 권장합니다.*