WEEK7
: 데이터로 타이타닉 분석하기

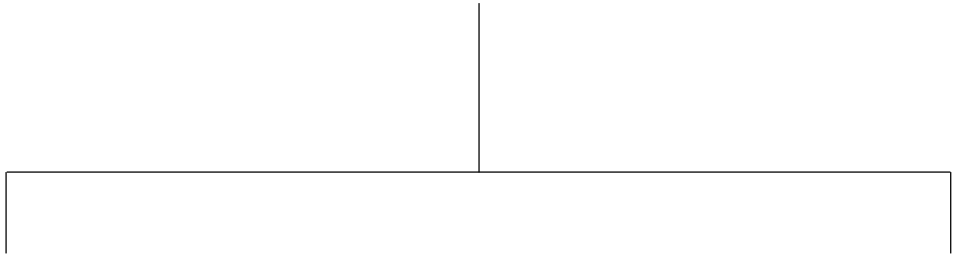
발표자 : 성신여자대학교 김지한

2020. 06. 04 (금)

동적 사이트와 정적 사이트의 차이는?

최초 : DATA1

접속 이후 : DATA2



정적데이터

DATA1 = DATA2

: 미리 저장된 파일 그대로
접속자에게 전달

동적데이터

DATA1 != DATA2

: 사용자의 요청에 따라 그때
그때 가공처리 및 생성

확인방법

1) 화면이 서서히 업로드 되어 loading시간이 김 (직감적)

2) 페이지 소스와 element tab을 비교해보면 정확히 알 수 있음.

Ycombinator : 페이지소스

```
<html lang="en" op="news"><head><meta name="referrer" content="origin"><meta name="viewport" content="width=device-width, initial-scale=1.0"><link rel="stylesheet"
href="news.css?6osDtdDRJS2fPerXRqDm">
        <link rel="shortcut icon" href="favicon.ico">
          <link rel="alternate" type="application/rss+xml" title="RSS" href="rss">
        <title>Hacker News</title></head><body><center><table id="hnmain" border="0" cellpadding="0" cellspacing="0" width="85%" bgcolor="#f6f6ef">
        <tr><td bgcolor="#ff6600"><table border="0" cellpadding="0" cellspacing="0" width="100%" style="padding:2px"><tr><td style="width:18px;padding-right:4px">
```

```
<html lang="en" op="news">
▼<head>
    <meta name="referrer" content="origin">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <link rel="stylesheet" type="text/css" href="news.css?6osDtdDRJS2fPerXRqDm">
    <link rel="shortcut icon" href="favicon.ico">
    <link rel="alternate" type="application/rss+xml" title="RSS" href="rss">
    <title>Hacker News</title> == $0
  </head>
▼<body>
  ▼<center>
    ▼<table id="hnmain" border="0" cellpadding="0" cellspacing="0" width="85%" bgcolor="#f6f6ef">
      ▼<tbody>
        ▶<tr>…</tr>
          <tr id="pagespace" title style="height:10px"></tr>
        ▶<tr>…</tr>
        ▶<tr>…</tr>
        </tbody>
      </table>
    </center>
    <script type="text/javascript" src="hn.js?6osDtdDRJS2fPerXRqDm"></script>
  </body>
</html>
```

Ycombinator : element tab

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" lang="ko" xml:lang="ko">
<!-- s:html:head -->
<head>
```

교보문고: 페이지소스

교보문고: element tab

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" lang="ko" xml:lang="ko">
  <!-- s:html:head -->
▶ <head>…</head>
… ▼ <body class="welcome"> == $0
  ▶ <iframe name="HiddenActionFrame" frameborder="0" width="0" height="0" style="display:none;" title="빈 프레임">…</iframe>
  ▶ <script>…</script>
    <!-- e:html:head -->
    <!-- 20111007 배너 설정시 추가 -->
  ▶ <script type="text/javascript">…</script>
  ▶ <div class="hidden">…</div>
  ▶ <div id="skip_to_content">…</div>
    <!-- #skip_content -->
    <!-- 상단 띠배너 -->
    <!-- jsp:include page="/prom/TopRibbonBanner.jsp"/ -->
    <!-- // 상단 띠배너 -->
    <!-- 상단 띠배너 -->
    <!-- *** s:(자동화)Ribbon배너 *** -->
    <!-- #####wel_RibbonBanner_START##### -->
    <!-- *** s:RibbonBanner 전용 Style *** -->
  ▶ <style type="text/css">…</style>
    <!-- *** //e:RibbonBanner 전용 Stylbe *** -->
    <!-- 상단 띠배너 -->
    <script type="text/javascript">
    var date_server = 20200601; // 개발: 서버날짜
    </script>
    <!-- *** s:RibbonBanner *** -->
    <!--
        띠배너 오더클릭
        &orderClick=dow / dox / doy
        띠배너 컬러 설정
```

목차

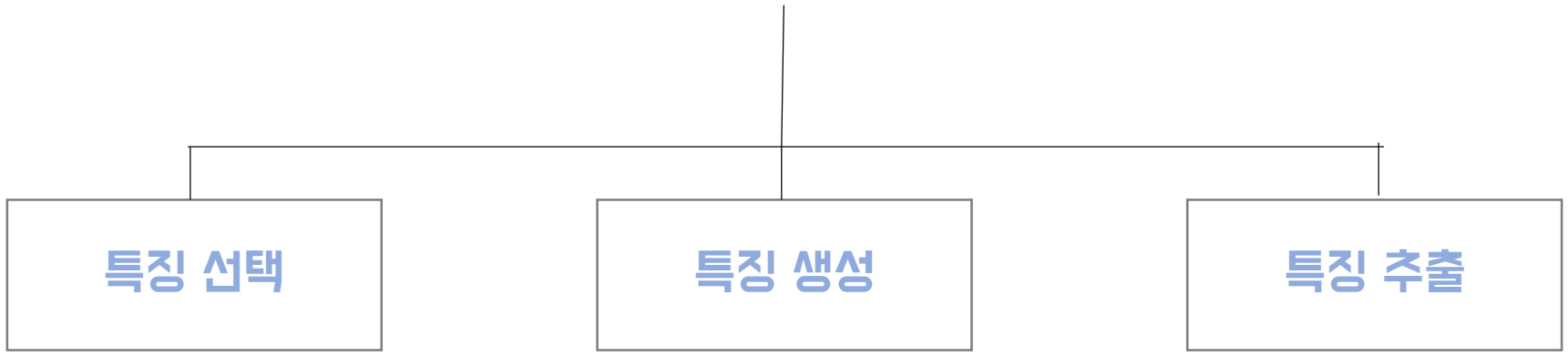
Feature Engineering

모델이 데이터의 특성을 잘 이해하고, 반영하여

그 성능이 향상될 수 있도록 특징을 생성, 수정, 가공 하는 것.

↓

컴퓨터가 이해하기 쉬운 데이터로 만들어주는 과정을 말함



Feature 선택

Feature 선택(특징선택)이란?

: 각 특징 별 중요도를 매기고, 분석결과에 큰 영향을 주는 특징을 선별하는 과정

PassengerId	Survived	Pclass	ame	Sex	ge	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	raund, Mr. Owe	male	2	1	0	A/5 21171	7.25		S
2	1	1	umings, Mrs. Jo	female	8	1	0	PC 17599	71.2833	C85	C
3	1	3	eikkinen, Miss.	female	6	0	0	STON/O2. 310128	7.925		S
4	1	1	utrelle, Mrs. Jac	female	5	1	0	113803	53.1	C123	S
5	0	3	llen, Mr. Willian	male	5	0	0	373450	8.05		S
6	0	3	loran, Mr. Jame	male		0	0	330877	8.4583		Q
7	0	1	cCarthy, Mr. Tir	male	4	0	0	17463	51.8625	E46	S

Feature 생성

Feature 생성(특징생성)이란?

: Feature Engineering 과정 중 대부분을 차지.

비어있는 데이터는 채워주고. 불필요한 터는 삭제해줌으로써 필터링을 용이하게 하기 위한 작업.

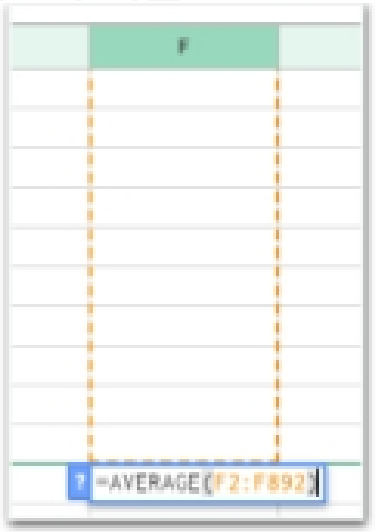
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen	male	22	1	0	A/5 21171	7.25		
2	1	1	Cumings, Mrs. Joh	female	38	1	0	PC 17599	71.2833	C85	
3	1	3	Heikkinen, Miss. L	female	26	0	0	STON/O2. 310128	7.925		
4	1	1	Futrelle, Mrs. Jacq	female	35	1	0	113803	53.1	C123	
5	0	3	Allen, Mr. William	male	35	0	0	373450	8.05		
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Tim	male	54	0	0	17463	51.8625	E46	

Feature 생성 - age 카테고리화

1. A-Z로 정렬 후 빈칸 채워주기
 : 머신러닝을 하기 위해선 빈칸이 있으면 안됨.

3	Svensson, Mr. J	male	5
1	Barkworth, Mr. A	male	5
3	Moran, Mr. Jame	male	
2	Williams, Mr. Ch	male	
3	Masselmani, Mr	female	
3	Emir, Mr. Farred	male	
3	O'Dwyer, Miss. E	female	
3	Todoroff, Mr. Lal	male	
1	Spencer, Mrs. W	female	
3	Glynn, Miss. Ma	female	

해당 열 (age)를 클릭한 후 ,
오른쪽 하단에 sum -→ average로 바꿔주면 쉽게 구할 수 O

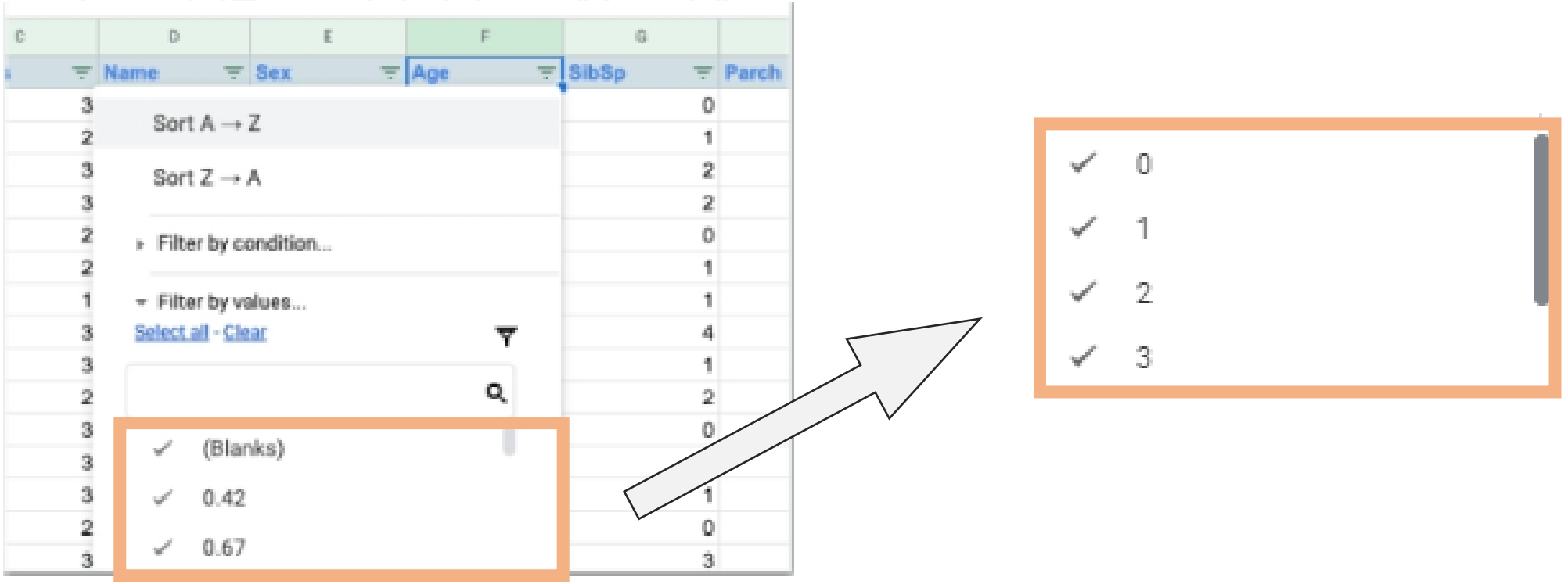


=AVERAGE(F2:F892)

2. 구해준 Average값을 위의 빈칸에 채워준다.

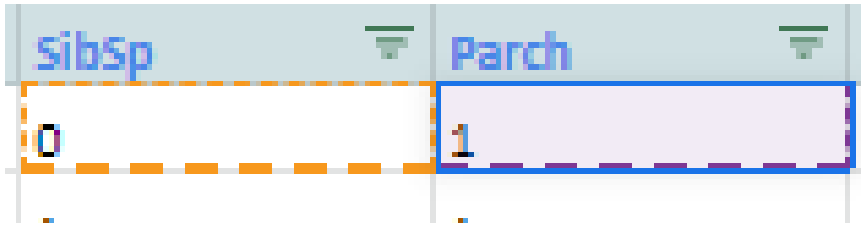
Feature 생성 - age 카테고리화

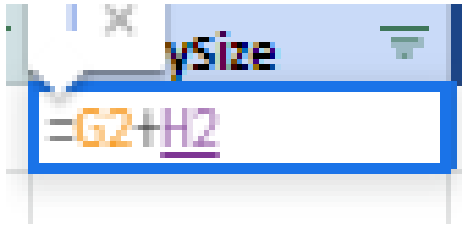
3. 다시 A-Z로 정렬 후 빈칸 채워주기
   : 아래에 위치한 29.6까지 다시 정령해주기 위해.



Feature 생성 - SibSp 와 Parch 합치기

1. 형제 수 (SibSp) + 부모님 수 (Parch)
: 불필요하게 분류된 특징들을 하나로 합해준다.





Feature 생성

필요 없는 열 삭제해주기. (삭제 대신 색깔로 표시)

C	D	E	F	G	H	I	J	K	L
Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
3	Thomas, Master. A	male	0	0	1	2625	8.5167		C
2	Hamalainen, Mast	male	0	1	1	250649	14.5		S
3	Baclini, Miss. Hele	female	0	2	1	2666	19.2583		C
3	Baclini, Miss. Euge	female	0	2	1	2666	19.2583		C
2	Caldwell, Master. A	male	0	0	2	248738	29		S
2	Richards, Master.	male	0	1	1	29106	18.75		S
1	Allison, Master. Hu	male	0	1	2	113781	151.55	C22 C26	S
3	Panula, Master. Ei	male	0	4	1	3101295	39.6875		S
3	Johnson, Miss. Ele	female	0	1	1	347742	11.1333		S
2	Becker, Master. Ri	male	0	2	1	230136	39	F4	S
3	Nakid, Miss. Maria	female	0	0	2	2653	15.7417		C
3	Goodwin, Master.	male	0	5	2	CA 2144	46.9		S
3	Dean, Master. Ber	male	0	1	2	C.A. 2315	20.575		S
2	Mallet, Master. An	male	0	0	2	S.C./PARIS 2079	37.0042		C
3	Palsson, Master. G	male	0	3	1	349909	21.075		S
3	Rice, Master. Euge	male	0	4	1	382652	29.125		Q
3	Andersson, Miss. E	female	0	4	2	347082	31.275		S

테스트해보기

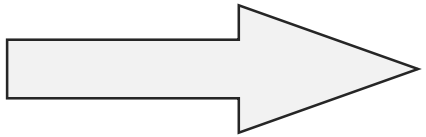
성별	점수
male	20
female	65

나이 변환값	점수
0	45
1	30
2	20
3	35
4	30
5	25

가족 수	점수
0	0
1	10
2	20
3	30

특징	가중치
성별	0.6
나이	0.3
가족 수	0.1

생존 기준	50



정답 수	700
오답 수	191
점수(예측성공률)	78.56%

성별	점수
male	30
female	75

나이 변환값	점수
0	45
1	30
2	40
3	35
4	30
5	20

가족 수	점수
0	0
1	10
2	20
3	30

특징	가중치
성별	0.7
나이	0.3
가족 수	0.1

생존 기준	50

Kaggle에서 체점받기

180...	Yunhyung Seo	0.76555	1	2d
180...	bgy1060	0.76555	1	2d
180...	jeongjin Kim	0.76555	2	2d
180...	지한	0.76555	2	1m

Your Best Entry ↑

Your submission scored 0.76555, which is not an improvement of your best score. Keep trying!