

제 15회 SAS 분석 챔피언십 분석 결과 보고서

Team Name : SA354

고려대학교 정보통계학과

김문수 김태형 정래윤

목차

1. 프로젝트 개요

- 주제
- DATA 설명
- SAS code 설명

2. 회귀분석

- 데이터 분할
- 회귀 모델
- 검정

3. 결론

주제

여름 영화 시장에서 가장 선호되는 장르 중 하나는 '공포' 이다.

때문에 여름에 공포영화를 광고하는 경우가 많다.

이러한 광고는 공포영화 매출에 큰 영향을 끼칠 수 있다.

마찬가지로 유효한 독자층을 겨냥하여 마케팅 전략을 짜는 것은
매출 상승에 도움을 준다.

그래서 우리는

공포 장르의 영화에 나이, 성별, 지역 변수들이 얼마나 영향을 끼치는 지
초점을 두고 분석하였다.

DATA 설명

Table_Name	COLUMN_NAME	LABEL	DATA_TYPE	LENGTH	VALUE
고객정보 customer	pvs_gender_typ	성별	문자	1	M
	cus_age	나이	문자	3	46
	cus_bill_addr_1	주소	문자	101	서울 양천구 목5동
	sa_id	가입번호	문자	5	10001
컨텐츠 메타정보 contents_meta	genre_mid	중장르	문자	11	공포
	album_id	앨범ID	문자	15	M010987A25PPV00
시청이력 history	sa_id	가입번호	문자	5	10003
	album_id	앨범ID	문자	15	M01126IB10PPV00
상세시청이력 history_detail	sid	가입번호	문자	5	10001
	contents_id	앨범ID	문자	15	M011372036PPV00
	genre_mid	중장르	문자	14	공포
조회이력 search	sid	가입번호	문자	5	10001

- 필요한 변수들로 분석을 하기 위해 Data Handling
- 5개의 테이블을 하나의 테이블로 병합
- 가입번호(sa_id=sid), 앨범ID(album_id=contents_id)는 column_name은 다르지만 값(value)은 동일한 변수

SAS Code 설명

```
libname res '/data/sasv94/sasuser06';
```

```
proc sort data=res.customer;  
by sa_id;  
run;
```

```
proc sort data=res.history;  
by sa_id;  
run;
```

```
data res.test;  
merge res.customer res.history;  
by sa_id;  
run;
```

```
data res.his_detail;  
set res.history_detail;  
sa_id=sid;  
drop sid;  
run;
```

```
data res.contents_meta_2;  
set res.contents_meta;  
contents_id=album_id;  
drop album_id;  
run;
```

라이브러리 설정

Customer(고객정보) 테이블을
sa_id(가입번호) 변수로 정렬

History(시청이력) 테이블을
sa_id(가입번호) 변수로 정렬

Sa_id 변수를 기준으로 Customer와 History 테이블을
Merge 함수로 병합한 res.test 테이블을 생성

History_detail(상세시청이력) 테이블을
set함수로 불러온 후 가입번호 변수의 이름을
변경(sa_id=sid) 그리고 중복되는 변수(sid)를
drop으로 삭제 His_detail 테이블을 생성

Contents_meta(컨텐츠 메타정보) 테이블을
set함수로 불러온 후 앨범ID 변수의 이름을
변경(contents_id=album_id) 그리고 중복되는
변수(album_id)를 drop으로 삭제
Contents_meta_2 테이블을 생성

SAS Code 설명

```
proc sort data=res.his_detail;  
by contents_id;  
run;  
  
proc sort data=res.contents_meta_2;  
by contents_id;  
run;  
  
data res.test2;  
merge res.his_detail(in=aa) res.contents_meta_2;  
by contents_id;  
if aa;  
run;  
  
proc sort data=res.test2;  
by sa_id;  
run;  
  
data res.result;  
merge res.test res.test2;  
by sa_id;  
run;  
  
data res.search2;  
set res.search;  
sa_id=sid;  
drop sid;  
run;
```

His_detail 테이블을
Contents_id(앨범ID) 변수로 정렬

Contents_meta_2 테이블을
Sa_id(가입번호) 변수로 정렬

Contents_id 변수를 기준으로 his_detail과
Contents_meta_2 테이블을
Merge 함수로 left join한 res.test2 테이블을 생성

Test2 테이블을 sa_id(가입번호) 변수로 정렬

Sa_id(가입번호) 변수를 기준으로 test와 test2을
Merge함수로 병합한 res.result테이블 생성

Search(조회이력) 테이블을
Set함수로 불러온 후 가입번호 변수의 이름을
변경(sa_id=sid) 그리고 중복되는 변수(sid)를
drop으로 삭제 search2 테이블을 생성

SAS Code 설명

```
proc sort data=res.search2;
by sa_id;
run;

data res.result2;
merge res.result res.search2;
by sa_id;
drop aa;
run;

data res.result2;
set res.result2;
if cus_age < 20 then age_in = '20대 미만';
else if cus_age < 30 then age_in='20대';
else if cus_age < 40 then age_in='30대';
else if cus_age < 50 then age_in='40대';
else age_in='50대 이상';
run;
```

Search2 테이블을 Sa_id(가입번호) 변수로 정렬

Sa_id(가입번호) 변수를 기준으로
Result와 search2 테이블을 Merge 함수로 병합하고
left join에 사용된 aa변수를 제거한
res.result2를 생성

Result2 테이블의 cus_age(나이) 이산형 변수를
age_in(나이대) 명목형 변수로 변경

SAS Code 설명

```
data res.TTA(keep=sa_id fear sex age10 age20 age30 age40 addr1 addr2 addr3 addr4 addr5 addr6);
set res.result2;

if genre_mid="공포" then fear=1;
else fear=0;

age10=0;
age20=0;
age30=0;
age40=0;
if cus_age < 20 then age10 = 1;
else if cus_age < 30 then age20=1;
else if cus_age < 40 then age30=1;
else if cus_age < 50 then age40=1;

sex=0;
if pvs_gender_typ = 'M' then sex=1;

addr1=0;
addr2=0;
addr3=0;
addr4=0;
addr5=0;
addr6=0;
if substr(cus_bill_addr_1,1,4) in ('경기','인천') then addr1=1;
else if substr(cus_bill_addr_1,1,4)='강원' then addr2=1;
else if substr(cus_bill_addr_1,1,4) in ('대전','세종','충북','충남') then addr3=1;
else if substr(cus_bill_addr_1,1,4) in ('부산','대구','울산','경북','경남') then addr4=1;
else if substr(cus_bill_addr_1,1,4) in ('광주','전북','전남') then addr5=1;
else if substr(cus_bill_addr_1,1,4)='제주' then addr6=1;

run;
```

나이변수와 주소변수를 가변수화하여
res.TTA 테이블 생성

성별변수의 값 'M', 'W'를 1, 0으로 변경

출력결과

	sa_id	fear	age10	age20	age30	age40	sex
1	10001	0	0	0	0	1	1
2	10001	0	0	0	0	1	1
3	10001	0	0	0	0	1	1
4	10001	1	0	0	0	1	1
5	10001	0	0	0	0	1	1
6	10001	0	0	0	0	1	1
7	10001	0	0	0	0	1	1
8	10001	0	0	0	0	1	1
9	10001	0	0	0	0	1	1
10	10001	0	0	0	0	1	1
11	10001	0	0	0	0	1	1
12	10001	0	0	0	0	1	1

[illegible]


공포 장르에 어떤 변수들이 얼마만큼의 영향을 끼치는 가?

이름	역할	레벨	리포트	순서	제거	하한	상한
addr1	Input	BINARY	아니요		아니요	.	.
addr2	Input	BINARY	아니요		아니요	.	.
addr3	Input	BINARY	아니요		아니요	.	.
addr4	Input	BINARY	아니요		아니요	.	.
addr5	Input	BINARY	아니요		아니요	.	.
addr6	Input	BINARY	아니요		아니요	.	.
age10	Input	BINARY	아니요		아니요	.	.
age20	Input	BINARY	아니요		아니요	.	.
age30	Input	BINARY	아니요		아니요	.	.
age40	Input	BINARY	아니요		아니요	.	.
fear	Target	BINARY	아니요		아니요	.	.
sa_id	ID	Nominal	아니요		아니요	.	.
sex	Input	BINARY	아니요		아니요	.	.

- 종속(Target)변수 : fear -> 공포장르
- 독립(input)변수 : addr1-5, age10-40, sex

데이터셋 할당	
분석용(Training)	70,0
평가용(Validation)	30,0
검증용(Test)	0,0


- 분석용, 평가용 70대 30으로 설정 -> 데이터 분할에서 설정

방정식(Equation)	
주효과(Main Effects)	예
2요인 교호작용(Two-Factor Interactions)	마니요
다항식 항(Polynomial Terms)	마니요
다항식 차수(Polynomial Degree)	2
사용자 항(User Terms)	마니요
항 편집기(Term Editor)	

- 종속(Target)변수 : fear -> 공포장르
- 독립(input)변수 : addr1-5, age10-40, sex

Class 타겟(Class Target)	
회귀 유형(Regression Type)	로지스틱 회귀
연결함수(Link Function)	로짓(Logit)

- 종속변수인 fear가 0 또는 1인 binary 이므로 로지스틱 회귀 분석을 사용한다.
- 이것을 사용하기 위해 범주형이었던 독립변수들을 가변수로 변환 하였다. (슬라이드 8&9)

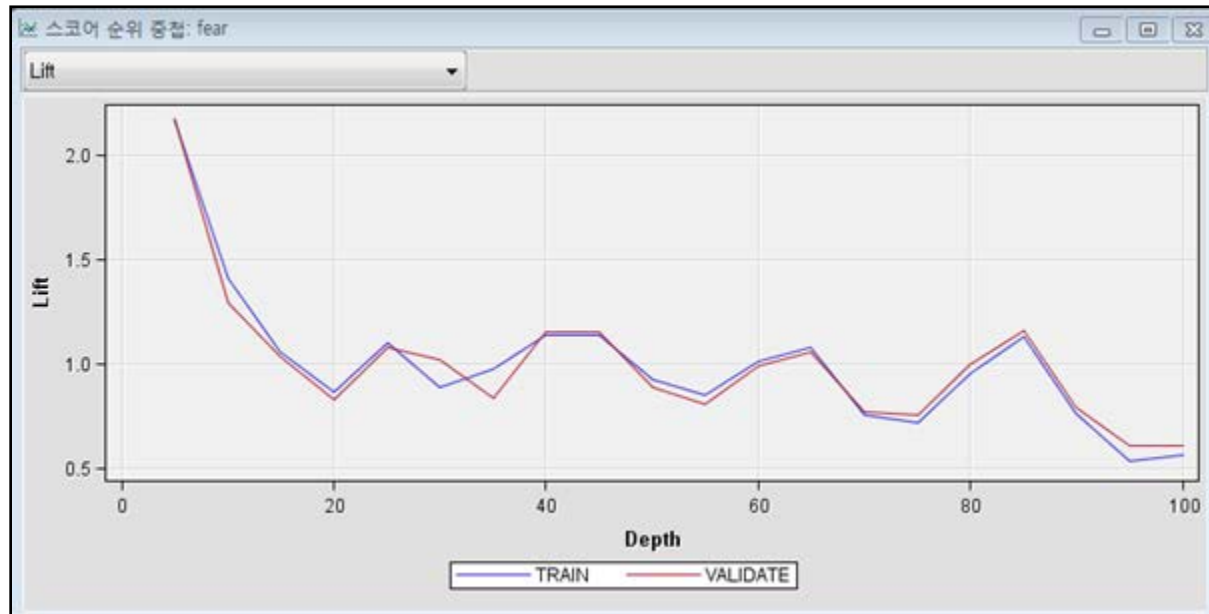
모델 선택	
모델 선택(Selection Method)	단계별 선택
선택 기준(Selection Criterion)	기본
선택 옵션 기본값 사용(Use Default Selection Options)	<input checked="" type="checkbox"/>
선택 옵션(Selection Options)	

- 종속변수 `fear`에 영향을 끼치는 최적의 독립변수들을 알아 내기 위해 `stepwise` 변수선택법을 사용함.

The selected model is the model trained in the last step (Step 6). It consists of the following effects:

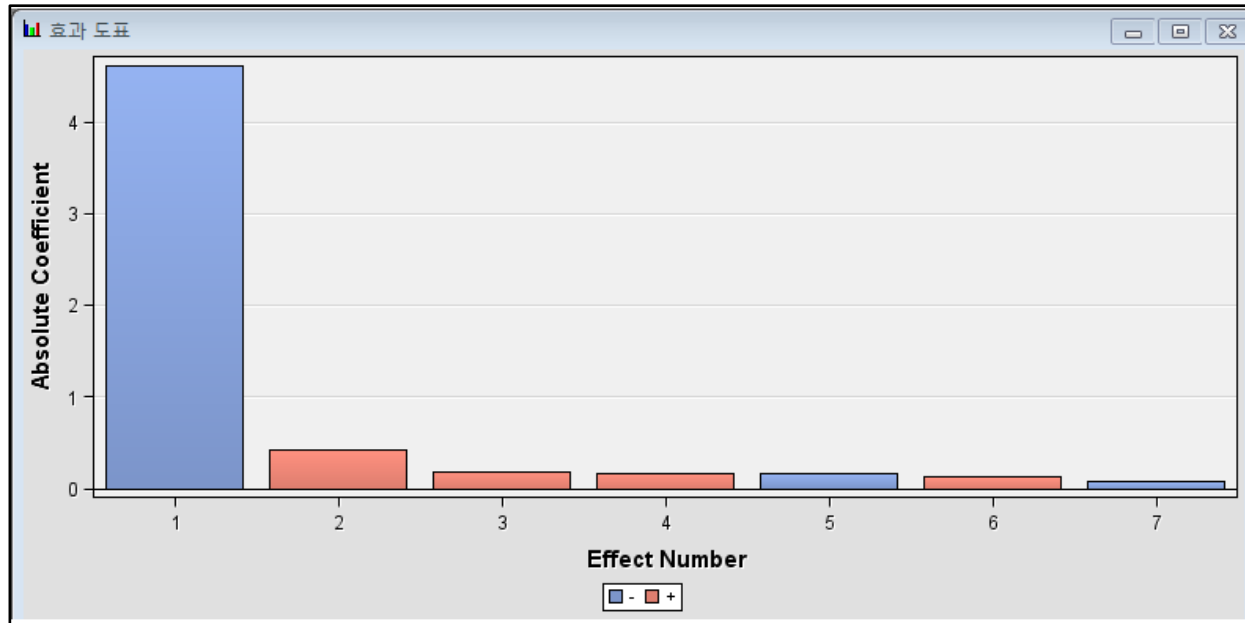
```
Intercept  addr4  age10  age20  age30  age40  sex
```

- 단계별 변수선택법을 통해 총 11개의 독립변수 중 5개의 변수로 만든 회귀모형이 가장 적합하다는 것을 알 수 있다.



[Lift Chart]

- Train 데이터와 validate 데이터의 차이를 보여줌
- 앞쪽이 경사가 가파르므로 좋은 결과임을 알 수 있다.
- Train 데이터와 validate 데이터의 큰 차이는 없다.



- 종속변수에 독립변수들이 얼마나 영향을 끼치는 지 보여주는 효과 도표
- 왼쪽에서부터 상수, 나이 10대, 20대, 30대, 경상도, 40대, 성별
- 양수와 음수를 감안 했을 때, 나이에 따라서는 10대 -> 40대 순으로 공포영화를 선호한다.
- 나이와 같은 맥락으로 지역 면에서는 서울에 비해 경상도에서 공포영화를 지양하는 것을 알 수 있다.
- 마지막으로 성별 변수가 음수인 것으로 보아 남자를 1 여자를 0으로 설정 했기에 여자가 공포영화를 조금 더 선호하는 것을 알 수 있다.

1. 회귀 모형 전체의 최대우도 검정

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	-4.0706	0.0139	86361.93	<.0001	0.017

* 카이스퀘어 p-value가 0.001보다 작으므로 이 회귀모형은 좋은 모형임을 알 수 있다.

2. Beta 검정

Likelihood Ratio Test for Global Null Hypothesis: BETA=0					
-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq	
53848.336	53848.336	0.0000	0	.	

* 카이스퀘어 p-value가 0.001보다 작으므로 이 회귀모형은 좋은 모형임을 알 수 있다.

1. 각 변수들의 최대우도 검정

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	-4.6273	0.0635	5316.74	<.0001	0.010
addr4	0 1	-0.1675	0.0155	117.38	<.0001	0.846
age10	0 1	0.4174	0.0521	64.22	<.0001	1.518
age20	0 1	0.1856	0.0208	79.38	<.0001	1.204
age30	0 1	0.1681	0.0203	68.85	<.0001	1.183
age40	0 1	0.1217	0.0193	39.79	<.0001	1.129
sex	0 1	-0.0729	0.0142	26.43	<.0001	0.930

* 카이스퀘어 p-value가 0.001보다 작으므로 각각의 독립변수들이 종속변수에 유의한 영향을 끼침을 알 수 있다..

2. 오즈비

- 이 오즈비는 가변수의 기준점에 비해 각 개체들에 얼마만큼의 영향을 끼치는 지를 알 수 있다.
- Addr(주소) 의 기준은 서울
- 나이 의 기준은 50대
- 성별의 기준은 여성
- 각각의 개체의 기준이 1 이므로 1보다 크면 기준보다 각 개체가 종속변수의 영향을 더 끼침을 알 수 있다.
- 만약 1보다 작은 값을 가지면 기준점이 종속변수에 더 유의한 영향을 끼치는 것을 알 수 있다.

Effect	Point Estimate
addr4 0 vs 1	0.715
age10 0 vs 1	2.304
age20 0 vs 1	1.450
age30 0 vs 1	1.399
age40 0 vs 1	1.276
sex 0 vs 1	0.864

결론

$Fear = -4.6273 - 0.1675 * addr4 + 0.4174 * age10 + 0.18568 * age20 + 0.1681 * age30 + 0.1217 * age40 - 0.0729 * sex$

위의 회귀 모형을 통하여 공포 장르에 특정 변수가 얼마나 영향을 끼치는 지를 알 수 있었다.
나이 성별 지역 세 가지 변수 중에서 **나이 변수는 나머지 변수들에 비해 유의**한 결과를 보인다.

나이 변수 중에서 **10대 연령층이 공포물을 가장 선호** 한다는 것을 알 수 있다.

이 분석 결과를 통하여 LG U+ 에서 내년 여름 공포 영화의 판매전략을
더 효율적으로 세울 수 있을 것 이다.

영화 시장의 경우 한 가지 장르를 선호하면 그 장르의 영화를 여러 번 시청하는 것이
일반적이기 때문에 위의 결과를 **여성이 남성보다 공포물을 더 선호**한다는 결과와 결합시켜

10대 여성에게 취향에 맞는 공포영화를 추천해주는 전략을 세워

공포영화의 판매율을 극대화 시킬 수 있을 것이다.

위의 예시 이외에도 취약한 판매 층을 끌어 들일 수 있는 맞춤 전략을 세우는 등

이 연구 결과로 여러가지 판매전략을 만들 수 있을 것이라 생각한다.