제 17 회 SAS 분석 챔피언십

"Big Data 시대, Data Scientist 를 향한 도전"

주제 정의서(본분석)

July, 2019

SAS Korea

Contents

1. 분석 주제 정의

홈쇼핑 방송에서 판매실적에 영향을 주는 요인으로는 판매 상품, 방송 편성 등이 있습니다. 아래 제시된 분석 주제 하에 방송 편성과 관련된 다양한 데이터를 통계적으로 분석하여, 판매실적에 영향을 주는 요인을 파악하고 판매실적 예측을 위한 정확도 높은 모형을 구축합니다. 도출된 결과를 활용하여 판매실적 증대를 위한 아이디어를 제시합니다.

주제 1) 홈쇼핑 판매실적에 영향을 미치는 요인에 대한 탐색 및 시각화

주제 2) 홈쇼핑 판매실적 예측을 위한 파생변수 생성, 외부변수 발굴, 예측 모형 개발 및 판매실적 예측

주제 3) 분석결과를 이용하여 판매실적 증대를 아이디어 제시

2. 분석 요구사항

2.1 분석 개요

제공되는 자료는 최근 6 년간의 (13.01.01~19.04.30) 편성 시간표, PGM 편성정보, PGM 실적, 상품 정보 데이터로 구성되어 있습니다.

분석 과제는 홈쇼핑 방송 판매실적과 요인들에 대한 연관 관계를 규명하고 이를 통해 정확한 판매실적을 예측하는 것 입니다. 판매실적이 포함되어 있는 Train 데이터와 포함되어 있지 않은 Test 데이터를 제공합니다. Test 데이터에 대한 판매실적을 예측하여 그 결과를 제출하고, 판매실적 증대를 위한 아이디어를 제출하시면 됩니다.

효과적으로 분석을 하기 위해서는 제공된 자료들 외에도 개방된 공공 데이터 (공공데이터 포털, 서울 열린 데이터 광장, 기상자료개방포털 등)를 적극 활용해야 합니다. 날씨, 고속도로상황, 주가 등 판매실적에 영향을 미칠 것으로 예상되는 여러 정보를 복합적으로 분석에 적용하기 바랍니다.

또한 유의미한 파생변수를 생성하는 것이 중요합니다. 기존 변수 조합 또는 외부 변수와의 조합 등 여러 파생변수를 이용하기 바랍니다.

예측 모델링을 위해 필요한 데이터 가공작업(파생변수 생성)은 직접 수행하며, 통계분석 또는 머신러닝 기법에 근거한 예측 모델링 후 개발된 결과와 개발 과정, 결과 해석 등의 내용을 보고서 형식으로 제출합니다.

2.2 분석 내용

자료의 탐색적 분석을 수행하여 해당 분석 방법을 선택한 근거를 제시하고, 다음과 같은 3 가지의 주요한 분석을 포함해야 합니다.

① 홈쇼핑 판매실적에 영향을 미치는 요인에 대한 탐색 및 시각화

A. 방송 시간대, 요일(주중 / 주말), 휴일 여부 등 방송에 대한 정보, 상품별 특성 등을 포함한 다양한 요인을 탐색

B. SAS® Visual Data Mining and Machine Learning, SAS® Visual Statistics, SAS® Visual Analytics, 및 SAS 프로시져를 활용하여 탐색된 요인의 다양한 시각화 진행

② 홈쇼핑 판매실적 예측을 위한 파생변수 생성, 외부변수 발굴, 예측 모형 개발 및 판매실적 예측

A. 새로운 외부 변수를 추가하고 내/외부변수로부터 다양한 파생 변수를 생성하여 홈쇼핑 판매실적 예측 모형에 활용

B. SAS 가 제공하는 다양한 알고리즘 (특히 SAS® Visual Data Mining and Machine Learning 의 머신러닝 기법)을 이용하여 예측 모형을 개발하고 이를 이용하여 Test 데이터에 대한 판매실적 예측

C. 예측 모형을 통해 판매실적(target)에 영향을 주는 주요 요인을 제시하고, 이에 대한 도출 과정, 개발된 예측 모형에 대한 해석 및 Test 데이터에 적용한 결과를 제시

③ 분석결과를 이용하여 판매실적 증대를 위한 아이디어 제시

A. 참신한 활용 방안을 자유롭게 제시하되, 홈쇼핑 사업의 특성을 고려.

B. 방송 직전 갑자기 편성을 바꾸는 것은 업체에 대한 갑질로 보일 여지가 있음. 일주일 전에 편성한다는 원칙을 지켜야 함.

C. 과학기술정보통신부의 기준 상 방송 편성 시간의 70% 이상은 중소기업의 상품 운영 해야함. 극단적인 상품 포트폴리오 구성을 하지 말 것.

D. 경쟁사 관련 데이터를 활용하는 방안은 담합으로 보일 여지가 있음.

[NOTE]

- 자료 분석 실시 전, 사용할 통계적 기법에 대한 이해와 장단점 파악이 선행되어야 합니다.
- 방송 판매실적과 요인 간의 관계는 단순한 관계를 제시하는 것이 아닌 분석 결과를 통한 논리적 근거와 함께 제시하여야 합니다.
- Train, Validate, Test, Score Set 을 참가자가 직접 선정하여 예측 모형의 성능 평가 및 검증 과정이 필수적으로 포함되어야 합니다.
- 모든 과정은 SAS 소프트웨어를 이용합니다. 데이터 처리는 SAS® Enterprise Guide® 및 SAS® Studio 를 예측 모델링은 SAS® Visual Data Mining and Machine Learning 이용을 추천합니다.

- 자료에서 유용하게 사용될 부분(일부 혹은 전체), 파생변수 개발, 통계 기법의 선정은 참가자 스스로 결정합니다.
- 효율적인 계산과 자원 배분이 필요합니다. (CPU, 메모리, 저장 공간 등)

3. 본분석에 대한 가이드

참가자는 분석작업을 위해 다양한 SAS 프로시저를 사용할 수 있고, 특히 예측 모델링을 이용하는 경우에는 SAS 모델링 방법론인 SEMMA (Sampling, Exploration, Modification, Modeling, Assessment) 프로세스에 의해 진행해야 합니다. 수행된 분석 과정 중에는 최소한 다음의 단계들이 포함되어야 합니다. (단, Sampling 단계는 제외할 수 있습니다.)

3.1 데이터 획득

① 서버 공용 폴더(/data/ldhome/vuser00/share)에 본분석용 데이터가 업로드 되어 있습니다. 서버 접속 방법은 웹사이트 '자료 제출' 메뉴의 'SAS 소프트웨어 사용 가이드'에서 다운로드 가능합니다.

② 주어진 공용 폴더(/data/ldhome/vuser00/share)의 데이터는 읽기 전용이므로 각 팀에 할당된 저장 공간으로 복사한 후 분석에 활용합니다.

③ 분석의 전 과정에서 주어진 데이터셋에 대한 활용은 모두 참가자가 판단하여 결정합니다.

3.2 데이터 분할 및 적용

예측모델링 시 사용되는 분석용 데이터셋은 분할하여 사용하되 분할 방법 및 비율은 참가자가 결정합니다.

3.3 데이터 정제

① 이상치 처리: 처리 대상의 이상치 선정 근거를 토대로 이상치를 처리합니다.

② 데이터 변환: 변환 대상에 대한 선정 근거를 토대로 변환합니다.

③ 파생 변수 생성: 파생변수 생성에 대한 근거로 가설을 제시하며, 생성 로직을 함께 제시해야 합니다.

3.4 예측 모델링

SAS 에서 제공하는 클라우드 환경(Web 기반) 하에서 다양한 모델링(머신러닝 포함) 방법을 사용합니다.

3.5 주의 사항

① 분석에 사용되는 SAS 소프트웨어 활용 시 아래 내용을 반드시 숙지하기 바랍니다.

[주의] 제공되는 SAS 소프트웨어는 본 대회를 위한 목적 이외에 사용하거나 참가자 당사자 이외의 다른 사람에게 제공할 수 없으며, 이로 인해 발생되는 법적인 불이익은 참가자 본인이 감수 해야 합니다.

② 모든 분석은 SAS 에서 제공하는 Web 기반의 클라우드 환경 하에서 수행하여야 합니다. 단, SAS® Enterprise Guide®는 참가자 개인 PC 에 설치한 후 서버에 연결하여 사용할 수 있습니다.

③ 하나의 서버를 여러 팀이 사용하므로 원활한 분석을 위하여 반드시 아래 사항을 지켜주세요.

A. SAS® Visual Analytics 및 SAS® Visual Data Mining and Machine Learning 에서 사용하는 데이터는 총합 메모리기준 20G, 디스크 Storage 는 100G 내로 사용하여야 함

B. 각 테이블당 크기는 4G 를 넘지 않도록 철저히 관리하여 사용하여야 함

C. 메모리 총합 20G 초과로 인하여 메모리 작업이 불가할 경우 사용하지 않는 테이블을 디스크에 저장 한 후 메모리에서 삭제하여 메모리를 확보할 수 있음

④ 만일의 경우를 대비하여 생성된 SAS 프로그램 및 SAS® Enterprise Guide®로 수행한 Project 는 주기적으로 로컬 환경에 백업을 하여 분석에 지장이 없도록 합니다.

⑤ 서버 및 분석 과정에서 나타나는 기본적인 오류 대한 해결 방법은 웹사이트 '자료 제출' 메뉴의 'SAS 소프트웨어 사용 가이드'를 참고하시길 바랍니다.

4. 본분석에 사용되는 데이터

4.1 데이터 개요

본 과제에서 사용되는 데이터 활용 시 아래 내용을 반드시 숙지하기 바랍니다.

[주의] 제공되는 데이터는 본 대회를 위한 목적 이외에 사용하거나 참가자 당사자 이외의 다른 사람에게 제공할 수 없으며, 이로 인해 발생되는 법적인 불이익은 참가자 본인이 감수 해야 합니다.

4.2 데이터 셋 구성 및 설명

① 분석용 데이터셋과 변수에 대한 설명은 웹사이트 '자료 제출' 메뉴의 '데이터 정의서'에서 다운로드 가능합니다. 추가되는 내용은 웹사이트를 통해 업데이트 될 예정이니 공지사항을 지속적으로 확인하도록 합니다.

4.3 주의 사항

① 서버에 올려져 있는 데이터 파일은 로컬 환경으로 다운 받을 수 없습니다.

② 공공데이터 포털, 서울 열린 데이터 광장, 기상자료개방포털 등 개방된 공공데이터 활용 시, 추가 데이터에 대한 상세 출처를 반드시 명시하여야 합니다.

5. 참고

① 모든 주요한 공지사항은 공식 웹사이트(https://www.sas-analytics.co.kr/) 에서 안내됩니다.

② SAS 제품 가이드는 SAS 사용자 커뮤니티인 MYSAS(http://mysas.co.kr/)의 'SAS 활용하기' 메뉴와 SAS SCHOOL 에서 제공되는 교육 책자를 참고해주세요.

③ 분석 결과물은 '심사요강(본분석)' 2.1 의 내용을 숙지하여 가이드라인에 맞게 제출하여야 합니다.

④ SAS 소프트웨어는 7 월 19 일 오전 9 시부터 사용 가능하며 8 월 30 일 오후 2 시에 사용이 종료됩니다. 종료 되기 전까지 '심사요강(본분석)'의 2.1 을 참고하여 폴더를 정리해 주세요.

제 17 회 SAS 분석 챔피언십

"Big Data 시대, Data Scientist 를 향한 도전"

심사요강(본분석)

July, 2019

SAS Korea

Contents

1. 1 차 심사 기준

제 17 회 SAS 분석 챔피언십의 1 차 심사는 제안서를 기준으로 복수의 심사위원이 서면 심사합니다.

1.1 1 차 심사 대상물

① 1 차 제안서

→ "팀번호.pdf" PDF 파일로 제출합니다.

1.2 1 차 심사 대상 평가 기준 및 배점

① 홈쇼핑 매출에 영향을 미치는 원인과 징후에 대한 탐색 및 시각화 방안 **(30%)**

A. 분석 주제와 목적에 대한 적합성

B. 탐색 및 시각화 방안 타당성

② 매출 예측을 위한 파생변수 생성 및 모델링 방법론 **(40%)**

A. 파생변수의 다양성 및 참신성

B. 분석 및 예측 모델링 전 과정의 타당성

③ 예상되는 분석결과와 그에 관한 활용방안 아이디어 **(30%)**

1.3 1 차 심사 일정

① 제출 마감일: 2019 년 7 월 12 일(금) 오후 2 시

② 제출 방법: 공식 웹사이트 '자료 제출' 메뉴

③ 결과 발표: 2018 년 7 월 18 일(목) 오후 6 시 예정

2. 2 차 심사 기준

제 17 회 SAS 분석 챔피언십의 2 차 심사는 1 차 합격자를 대상으로 하며 복수의 심사위원이 분석 결과를 서면으로 심사합니다.

2.1 2 차 심사 대상물

① 분석 결과 보고서

→ "팀번호_분석결과보고서.pdf" PDF 파일로 제출합니다.

→ 파일의 크기는 10M 를 넘을 수 없습니다.

→ Appendix 로 사용된 코드 리스트와 간단한 설명을 포함합니다.

② 분석에 사용된 SAS 프로그램 및 SAS® Enterprise Guide® 로 수행한 Project

A. SAS® Studio (SAS® Visual Analytics, SAS® Visual Statistics, SAS® Visual Data Mining and Machine Learning)에서 작성한 프로그램

→ SAS® Studio 의 파일에 "팀번호_프로그램" 폴더를 생성하여 폴더안에 프로그램 파일들을 저장합니다. (위치: 파일/data/ldhome/vuser00)

→ 개별 프로그램명은 "팀번호_작업번호_작업명.sas"으로 생성합니다.

(예: SA000_01_데이터 생성하기.sas)

B. SAS® Enterprise Guide® 로 수행한 Project

→ SAS® Studio 의 폴더에 "팀번호_EG 프로젝트" 폴더를 생성하여 폴더안에 프로그램 파일들을 저장합니다. (위치: 폴더/홈/data/ldhome/vuser00)

→ 개별 프로젝트명은 "팀번호_작업번호_작업명.epg"으로 생성합니다.

(예: SA000_01_데이터 생성하기.epg)

2.2 2 차 심사 대상 평가 기준 및 배점

① 분석 결과 보고서: 기본분석 요건 **(70%)**

A. 분석 주제에 대한 전반적인 이해력

B. 분석 데이터 준비 과정 / 내용의 분석 주제와 목적에 대한 적합성

- 데이터 탐색, 이상치 탐지, 처리 과정, 다양한 파생 변수 생성

C. 분석 결과를 도출하는 분석 전 과정의 타당성

D. 분석 결과의 활용방안에 대한 적절성 및 창의성

② SAS 소프트웨어로 수행한 분석 과정 **(30%)**

A. 사용한 분석 방법의 다양성

B. 주어진 SAS 소프트웨어를 이용한 데이터 처리에서 모델 평가까지의 전 과정 처리 수행능력

C. SAS® Visual Data Mining and Machine Learning 내 기능 사용에 대한 창의성

2.3 2 차 심사일정

① 제출 마감일: 2019 년 8 월 30 일(금) 오후 2 시

② 제출 방법: 웹사이트와 클라우드를 통해 제출

③ 결과 발표: 2019 년 9 월 5 일(목) 오후 6 시 예정

2.4 주의 사항

심사평가 대상의 최소 요건을 만족시키려면 반드시 '분석요건(상기 2.2)' 에 대한 결과를 제출해야 합니다.

3. 3 차 심사 기준

제 17 회 SAS 분석 챔피언십의 3 차 심사는 2 차 합격자를 대상으로 하며 복수의 심사 위원에게 프리젠테이션합니다.

4.1 3 차 심사 대상물

① 프리젠테이션을 위한 보고서 (파워포인트)

A. 2 차 분석 결과 보고서를 기반으로 프리젠테이션에 적당하도록 수정 및 보완하여 파워포인트로 제출합니다.

B. 분석 주제 수행 결과 및 분석 결과의 활용 전략을 포함합니다.

C. SAS® Drive 내 '데이터 탐색 및 시각화'를 이용하여 분석 결과를 시각화하고 그 결과물(동영상, 라이브 데모, 화면 캡쳐)을 활용합니다.

4.2 3 차 심사 대상 평가 기준 및 배점

① 프리젠테이션 및 보고서 내용 **(70%)**

A. 분석 수행 전 과정에 대한 적절성

B. 분석결과 활용의 적절성 및 창의성

C. 프리젠테이션 발표력 및 질의응답 적절성

D. 분석결과 시각화의 창의성

② 2 차 평가 결과 **(30%)**

4.3 3 차 심사 일정

① 제출 마감일: 2019 년 9 월 11 일(수) 오후 2 시

② 제출 방법: 공식 웹사이트에 추후 공지 예정

③ 프레젠테이션 및 시상식: 2019 년 9 월 19 일(목) 오후 1 시