|  |
| --- |
| **「2020 빅콘테스트」데이터 분석 계획서** |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | | |
| **참가분야** | □ 혁신아이디어분야 ■ 데이터분석분야 | | |
| **세부분야** | □ 퓨처스리그 ■ 챔피언리그 | | |
| **개인/팀여부** | □ 개인 ■ 팀(구성원 5명) | **개인/팀명** | 완판맨 |
| **대표ID** | gifrds1@gmail.com | | |

|  |  |
| --- | --- |
| **분석 주제명** | 새로운 시즌의 새로운 상품 완판을 위한 다각도 데이터 분석 |
| **분석 배경** | 제한된 시간 내에 상품을 판매해야 하는 홈쇼핑의 특성 상 시간을 효율적으로 사용하는 것이 중요하다. 따라서 시즌에 적절한 상품을 판매하는 동시에 고객들의 특성도 고려해야 하며 현재 경제 상황도 반영해야 할 것이다. 완판맨팀은 이를 해결하기 위해 취급액에 영향을 주는 내부요인과 외부요인들을 다각도로 분석하여 적절한 상품을 타겟 고객에게 제공하는 것을 목표로 한다. |
| **분석 내용**  **요약** | 2019년 NSshop+ 데이터 안의 내부적 요인(시계열 데이터와 상품군별 계절성 분석, 편성 횟수, 노출시간, 공휴일 및 황금 시간대, 상품명)을 분석하고, 외부데이터를 통해 외부적 요인(기상, 대기, 시청률, 코스닥 등)을 분석한다. |
| **분석방법 및**  **계획** | 1. **시계열 데이터 분석 및 상품군 별 계절성 분석**  * 상품군별 높은 취급액을 보인 월, 요일, 시간대 파악 * 시계열 예측 패키지인 prophet을 사용하여 상품군별 월, 주, 일별 계절성 파악  1. **월, 요일, 시간단위 상품군별 방송 편성 횟수 분석**  * 월, 요일, 시간단위로 상품군별 방송 편성 횟수를 구하여 상품군별 특정 날짜와 시간대의 1회 편성당 기대 취급액 계산  1. **방송 노출시간에 따른 취급액 변화 분석**  * 방송 시작 후 경과된 시간에 따라 변화하는 취급액 분석  1. **공휴일, 황금 시간대(Prime time) 등의 효과 분석**  * prophet을 통해 취급액이 공휴일과 황금 시간대 등에 의해 받는 영향을 파악  1. **상품명 분석**  * 데이터내용: NS mall의 카테고리 정보, 브랜드 정보 * 출처: NS mall 카테고리 분류 정보 및 하위 브랜드 크롤링 (<https://www.nsmall.com/>) * Doc2Vec 기법을 활용해 상품명과 가장 높은 유사도의 NS mall 상품군을 부여하여 예측 성능 향상 * Doc2Vec 기법을 활용해 상품과 상품간의 유사도를 파악, 가장 높은 유사도의 취급액을 참고하여 예측  1. **기상, 대기데이터 분석**  * 데이터내용: 강수량, 습도, 기온, 풍속 등 * 출처: 기상청 기상자료개방포털 (<https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36>) * 데이터내용: 미세먼지, 초 미세먼지 수치 등 * 출처: 서울 열린데이터 광장 ([https://data.seoul.go.kr/dataList/OA-2221/S/1/datasetView.do#](https://data.seoul.go.kr/dataList/OA-2221/S/1/datasetView.do)) * 일별 강수량, 기온, 불쾌지수가 취급액에 미치는 영향 분석 * 미세먼지, 초 미세먼지가 취급액에 미치는 영향 분석  1. **시청률 데이터 분석**  * 월별, 시간대별 시청 동향 분석 * 시청률의 특징을 활용한 황금 시간대 분석  1. **KOSDAQ 데이터 분석**  * 데이터내용: 일별 KOSDAQ 지수 * 출처: 한국거래소 (<http://www.krx.co.kr/main/main.jsp>)   - 사회적 이슈에 따른 경제 상황을 반영하기 위해 KOSDAQ 데이터 분석 |
| **분석결과 활용 및 시사점** | 1. **시계열 데이터 분석 및 상품군 별 계절성 분석**  * 월 단위 상품군별 취급액 비율을 분석한 결과 1월과 8월에 추석의 영향을 받아 ‘농수축’ 상품군의 취급액이 증가하는 것을 볼 수 있었다. 이를 통해 공휴일이 취급액에 영향을 준다고 판단할 수 있었다. * 6월에는 ‘속옷’ 상품군의 취급액이 비율이 가장 높게 나타났고, 여름이 되면서 기온과 습도가 높아져 판매량이 늘었다고 판단했다. 이를 통해 기온데이터와 불쾌지수 데이터를 분석하는 것이 도움이 될 것으로 판단된다. * 시간 단위 상품군별 취급액을 분석한 결과 각 시간대별로 주로 팔리는 상품군을 알 수 있었고, 방송 편성 최적화를 진행할 때 활용할 예정이다. * 이 외에도 각 월, 시간 별로 선호되는 상품군이 존재할 것이며 이는 편성기준 결정과 취급액에 큰 영향을 끼칠 것이라 예측되기 때문에 시계열 분석을 통해 계절성을 파악하면 최적화 방안 제시에 큰 도움이 될 것이다. 또한 feature로 활용하여 6월에 다른 상품군 보다 계절성이 특히 높았던 상품의 경우 가중치를 두어 모델에 반영할 예정이다.  1. **월, 요일, 시간단위 상품군별 방송 편성 횟수 분석**  * 각 상품군별로 편성 횟수를 분석한 결과 평성 횟수가 많아질수록 해당 상품군의 매출액이 오르는 결과를 볼 수 있었다. 하지만 증가하는 매출액이 상품군 별로 달라 편성효율을 계산할 필요성을 느꼈다. 또한 편성효율을 통해 높고 낮은 효율의 편성안을 발견할 수 있을 것이고, 이 편성효율에 따라 방송을 편성하면 매출을 향상시킬 수 있을 것으로 보인다.  1. **방송 노출시간에 따른 취급액 변화 분석**  * 대부분의 상품이 방송 시작 후 마감으로 갈수록 취급액이 높아지는 것을 확인할 수 있었고, 이를 반영하여 모델을 생성할 때 누적 노출시간을 고려하면 더 좋은 성능을 낼 것으로 기대된다.  1. **공휴일, 황금 시간대(Prime time) 등의 효과 분석**  * 공휴일에 대한 적절한 가중치를 모델에 적용시켜 예측 시 관련 근거로 사용할 예정이다. 공휴일, 주말과 황금 시간대는 매출이 주로 몰리며 편성 시에도 중요하게 고려되기 때문에 반드시 분석이 필요하다고 판단된다. 이를 통해 Feature를 만들고 편성 최적화 방안에서 활용 예정이다.  1. **상품명 분석**  * 기존 상품군을 세분화 하여 대, 중, 소 분류로 나누었다. 새롭게 나누어진 상품군으로 데이터 분석을 진행할 예정으로 더 성능이 좋은 feature들을 추출할 수 있을 것으로 예상된다. * 한 상품의 취급액과 가장 유사한 상품명과의 취급액의 비율이 평균적으로 2.5정도로 측정된다. 이를 feature로 사용하면 어느 정도 취급액의 기준점을 잡아주는 역할을 할 것으로 기대된다.  1. **기상, 대기데이터 분석**  * 분석 결과 비가 오래 내리거나 기온이 평균보다 높거나 낮으면 취급액이 증가하였다. 이를 통하여 기온과 날씨에 따라 그날의 취급액이 평균보다 높을지 낮을지 예측할 수 있다. * 불쾌지수, 미세먼지, 초 미세먼지의 경우 취급액 예측에는 큰 차이가 없었으나 feature 중요도는 높게 관측 되었다. 해당 요소들이 영향을 주는 경우만 따로 추출해 세부적으로 분석하여 상황에 따라 적용할 수 있도록 조정하면 예측 성능이 향상될 것으로 기대된다.  1. **시청률 데이터 분석**  * 월별 시청률 동향을 살펴본 결과 더운 여름철 혹은 추운 겨울철과 같이 야외활동이 줄어드는 시기에 대체로 시청률이 높았고, 야외 활동이 늘어나는 4~6월의 경우 시청률이 낮았다. 또한, 월별 그래프를 분석한 결과 인접한 월 사이에는 시청률 그래프가 비슷한 것으로 보아 시청률도 계절성을 띄고 있는 것을 확인할 수 있었다. * 주중과 주말(공휴일포함)을 비교했을 때 주말의 경우 주중보다 점심과 저녁 사이의 시청률 차이가 더 크게 나타났다  1. **KOSDAQ 데이터 분석**  * 2020년 코로나와 관련된 주요 이슈들 중, ‘국내 첫 감염 || 2020-01-20’, ‘신천지(그룹감염1차) ||2020-02-20’, ‘WHO에서 팬더믹 선언 || 2020-03-11’, ‘사회적거리두기 첫 시행 || 2020-03-22’ 와 같은 시점에 코스닥 지수가 떨어지는 것을 확인할 수 있었으며, KOSDAQ지수가 사회 이슈를 대체로 잘 표현하고 있음을 알 수 있었다. 또한, KOSDAQ지수를 기준으로는 2020년 6월의 경우, 코로나로 인한 경제 악화에서 어느정도 벗어난 상태임을 알 수 있었다. 2020년 6월도 현재 시점에서 보면 과거이기에 KOSDAQ데이터를 직접적으로 이용은 가능하지만, 문제의 최종목표가 미래의 최적의 편성을 예측하는 것이기 때문에 미래의 편성을 예측하기 위해서는 현재가 아닌 미래의 경기지표가 필요함을 느끼게 되었다. 이를 반영하기 위해 선행, 동행, 후행 지수를 반영할 예정이다. |