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

|  |  |  |  |
| --- | --- | --- | --- |
|  | \* 해당란에 체크 표시 | | |
| **참가분야** | □ 혁신아이디어분야 □ 데이터분석분야 | | |
| **세부분야** | □ 퓨처스리그 □ 챔피언리그 \*데이터분석 분야에 한함 | | |
| **개인/팀여부** | □ 개인 □ 팀(구성원 명) | **개인/팀명** | 홍길동 |
| **대표ID** | (개인/팀장 홈페이지 접속 ID) | | |

※ **5장 내외로** 목차는 준수하여 자유롭게 작성

|  |  |
| --- | --- |
| **분석 주제명** | <NS shop + 편성표/분당 시청률 분석을 통한 상품별 매출 예측 및 코로나 확산 상황에서 매출 극대화를 위한 제언?> |
| **분석 배경** | 여행/숙박, 음식점, 문화/레저, 교육 등 수많은 업종이 예기치 못한 코로나 사태로 매출 부진에 허덕이고 있는 상황에서 비대면으로 거래가 이루어지는 온라인 쇼핑, 홈쇼핑 등은 예년에 비해 상승세를 타고 있다. 홈쇼핑 월별 판매액지수(출처: 통계청)를 살펴보면 코로나가 국내에 본격적으로 전파되기 시작한 2월 이후 3월 해당 지수가 눈에 띄게 증가한 양상을 볼 수 있다.[[1]](#footnote-1) 20년 6월 역시 작년에 비해 높은 판매액지수를 기록했다. 20년 1분기에 국내 코로나 신규 확진자(해외유입 포함, 1주일 단위) 숫자와 홈쇼핑 매출액 간 0.79의 상관계수를 보이기도 했다(출처: 하나금융연구소)  특히 NS홈쇼핑의 경우, 20년 상반기에 전년 동기 대비 매출(직매입 상품 판매와 전체 수수료 합계기준)과 영업이익이 각각 15%, 4% 증가했다(출처: 금융감독원 전자공시시스템) 판매한 상품 금액의 합계인 취급고는 9% 상승으로, 취급고를 공개하지 않은 롯데홈쇼핑을 제외한 CJ오쇼핑, GS홈쇼핑, 현대홈쇼핑보다 높은 증가율을 기록했다.  분기별로 나눠서 살펴보면 다음과 같았다. 20년 1분기에 전년 동기 대비 매출과 영업이익이 각각 20%가량 증가하는 성과를 거두었다. 코로나 사태로 언택트(untact) 소비가 선호됐기 때문으로 풀이된다(출처: 닐슨코리아 코로나19 임팩트 보고서)  온라인쇼핑과 홈쇼핑을 아우르는 소매유통업 경기전망지수는 20년 2분기 실적에 1분기와 달리 비관적이었다. (84로 기준치 100 미달 시, 악화 전망) (출처: 대한상공회의소) 실제로 NS쇼핑은 2분기에 전년 동기 대비 매출이 11% 증가했지만 영업이익은 10% 하락했다. 매출, 영업이익, 취급고 등 모든 부문에서 20년 1분기에 비해 2분기의 성적표가 아쉬운 것이 사실이다. 분기에 따른 시기상의 차이가 주는 효과가 아닌지 추가로 살펴봤다. 19년 2분기에 전년 동기 대비 매출과 영업이익이 각각 2%, -21%의 증감을 기록했다. 건강식품, 손소독제, HMR(가정간편식) 등 저수익 제품 편성 증가, tv 송출 수수료 인상과 긴급 재난지원금 사용제한에 따른 판매감소 등 요인을 고려하면 20년 2분기 영업이익이 19년 2분기와 비슷한 수준이라는 관계자의 말이다.  기존 상품군별 매출 실적도 변화를 맞이했다. 건강에 대한 관심이 높아지고 가정에서 끼니를 해결하면서 건강기능식품과 HMR(가정간편식) 등 식품 판매가 증가한 반면 ‘사회적 거리두기’ 등으로 외출을 자제하며 여행상품과 화장품 등의 항목은 예년에 비해 부진한 상황이다. NS홈쇼핑 히트상품 Top10에 재작년과 작년 2~3개 자리를 차지하던 식품브랜드가 20년 상반기에는 7자리를 차지했을 정도이다(출처: NS홈쇼핑) 같은 홈쇼핑 업계라도 20년 1분기 코로나 대응 전략에 따라 저마다 다른 실적을 냈는데 식품 취급 비중을 늘린 업체가 대체로 수익성이 상승했다고 한다. 신선식품에 대한 고객들의 니즈를 확인할 수 있는 대목이다. NS 홈쇼핑은 ‘세계 최초의 식품 전문 홈쇼핑’으로 2004년에 창립한 만큼 방송시간의 60% 이상을 식품류로 편성하고 있는데 이는 사회적 위기 속 기회라고 할 수 있다. NS홈쇼핑은 업계에서 유일하게 식품안전연구소를 운영하고 자체 품질보증(QA) 시스템을 개발하는 등 농수산 식품에 대한 안정성과 고객의 신뢰성 확보를 위해 노력해왔다. NS홈쇼핑만의 차별화된 강점을 이용하면 코로나 사태로 인한 위기 속에서 돌파구를 찾을 수 있을 것이다.  (이에 홈쇼핑 매출에 영향을 끼친다고 알려진 날씨 변수, 프라임 타임 등 시간 변수, 그 밖에 EDA 과정을 통해 생성해낸 변수들과 코로나로 인한 파문까지 종합적으로 고려해서 매출을 예측하였다.) |
| **분석 내용**  **요약** | 홈쇼핑 데이터는 feature 가 매우 적은 간단한 데이터였지만, 그렇기 떄문에 예측의 정확도를 올리기 위해선 다양한 feature 들을 추가해야했다. 우선 EDA 과정에서는 품목명에서 브랜드, 남/여 상품 등등 의 정보를 뽑아내었고, 시간에서는 월 / 시간/ 분 등의 데이터를 뽑아내었고, 평균 방송분 등등의 변수를 통해서 기존에 있는 변수들의 특성을 최대한 분해해서 모델이 해석할 수 있도록 FEATURE 를 늘렸다.  그다음에는 외부데이터로 강수 / 온도 / 경제지수 등을 활용하여 변수를 늘렸다.  변수를 늘린 후에 X 데이터를 모델이 해석 가능하게 가공한 후에, multilinear 부터 시작하며 randomforest, xgboost , lgb 등의 모델들을 적합시켜 본 뒤에 그중에 제일 값이 잘 나온 모델을 선택하려고 한다. |
| **분석방법 및**  **계획** | **최종적으로 택한 모델링은 DNN 모델이였다. Relu 만을 적용한 5`6층의 깊이의 간단한 모델에서도 MAPE 는 42가 나왔다. regression 의 모델중 제일 나았던 baysian optimization 을 적용한 lgb 보다 mape 가 2 높은 46 정도의 값이 나왔다는 것을 고려해보면, DNN 모델을 이용하는 것이 타당해 보였다.**  **만일 9월 말까지 계속 eda , optimization 을 통하여도 dnn 의 mape 를 넘지 못한다면 DNN 을 쓸 계획이다.**  **다양한 activation function 과 구조를 통해서 최대한 서로 닮지 않은 3개 정도의 모델을 형성한 뒤(이래야만 편향을 줄일 수 있으니까), 세개 모델을 조합해서 예측값을 형성할 계획이다.**  **하지만 random forest 모델은 최소한 feature importance 처럼 모델 해석의 여지가 어느정도 있는 반면에 DNN 모델은 너무 복잡해 모델 자체적으로 어떤 변수가 중요하거나 어떤 영향이 있는지 말해주지 못하는 점이였다. 이 점을 Lime 또는 shap 패키지의 모델 해석을 이용해 보완할 계획이다. Lime 을 통해서는 local 변수에 대한 어느정도의 해석이 가능하므로 우리가 target 으로 삼은 시간대 및 품목에 대한 상세한 예측이 가능할 것이다. 이를 통해서 어떤 시간대에 어떤 것이 영향을 끼쳣고 어떻게 optimization 해야하는지 알 수 있을것이다.** |
| **분석결과 활용 및 시사점** |  |

※ 제출자료는 최종 출품작 평가시 활용될 수 있음

1. 불변지수 기준 2월 113->130, 3월 134->146, 6월 123->139 [↑](#footnote-ref-1)