|  |  |  |  |
| --- | --- | --- | --- |
| **2021 혁신성장 빅데이터 분석 프로젝트 요약서** | | | |
| **팀 명** | ADAM | **팀 장** | 김기영 |
| **팀 원** | 김민송, 박소연, 이동렬 | | |
| **주 제 명** | 호텔 예약 수요 데이터의 분석을 통한 다양한 경영 인사이트 도출 | | |

|  |  |  |
| --- | --- | --- |
| 구분 | 프로젝트 상세내용 | |
| 프로젝트명 | 호텔 예약 수요 데이터의 분석을 통한 다양한 경영 인사이트 도출 | |
| 구현 목표 | 1. 예약 취소자들의 특성 파악 및 예약 취소 고객 예측해내는 분류 모델 만들기 2. 변수 선택 작업과 시각화 작업(feature 선택) 3. 변수간의 상관도 히트맵 파악 4. 어떤 시즌에 고객들이 많이 방문하는가에 대해 알아보고 시각화 5. 숙박을 할떄 얼마나 오래 머무는지에 대해 알아보고 시각화 6. 어느달에 취소율이 높게 나타나는지를 알아보고 시각화 7. 예약 경로별 취소 고객 분석 및 시각화 8. 예약 상태에 대한 pi-그래프 그리기 9. 예약한 룸타입에 관한 변수 선택하지 않은 이유에 대해 시각화하여 입증 10. 입금 형태별 취소 고객 분석 11. 어떠한 분류기 모델을 활용할지 찾아보고 예측 | |
| 구현 내용 | 1-1. 모델을 설계하는데 어떤 변수를 선택해야할까? 어떤 변수가 모델의 성능을 높일 수 있을까?  1-2. 어떤 식으로 데이터 스케일링을 진행할 것인가?  1-3. 어떤 분류기 모델을 사용할 것인가? | 2-1. 예약 취소자들의 특성을 통해 예약 취소 고객 예측해내는 분류 모델 만들기  2-2. 다음 해의 예약자 수와 다음 해의 매출을 예측해보기  (회귀모델을 적용하여) |
| -어떤 변수를 선택하는지에 따라 모델의 성능이 달라진다  -모델링에 사용할 수 없는 변수는 제거하기  -변수의 속성을 변경하기  -후진제거법을 활용하기  -rf모델링 후 변수 중요도 순으로 특성을 제거해보기  -최종적으로 결정한 내용: 특성들 간의 상관관계가 크지 않고 모듈을 활용한 방법을 몇 가지 적용하여 보았으나 의미있는 수준의 성능 개선이 이루어지지 않았음  =>최종적으로 우리가 생각했을때 유의미하다고 판단되는 변수를 선택하여 일반 사람들에게 설명력이 있으면서도 어느 정도의 성능을 이끌어 낼 수 있는 모델을 구현하고자 함    -target값의 분포가 비슷하지 않기에 모델의 일반화 성능을 높이기 위해 오버샘플링 진행하기 | - 모델 예측의 해석 관련하여 EDA 및 시각화의 필요성(feature 선택 및 청중 설득을 위하여)  ->일부 중요한 특성들이 취소에 미치는 영향력이 있을 것이다. 를 설득하기 위해, 일부 특성별로 취소가 발생하는 현상을 관찰할 수 있었다. 는 흐름에서 시각화가 일부 필요할 듯.  -적용한 분류기 모델의 종류  ->트리계열 분류기  ->최근접 이웃 모델  ->서포트 벡터 머신  ->lr  ->GridSearch  ->교차검증과정 |
| 상세 설명 | **1. 주제 선정/개요**  1-1) 선정 배경  -’호캉스’와 같은 신조어가 생길 정도로 최근 호텔을 비롯한 숙박 산업의 수요 급증  -호텔 산업에서의 치명적인 No-Show와 취소를 예측해보다 나은 서비스 창출이 기대됨  -호텔 경영에 도움이 되는 다양한 인사이트 도출 가능하다고 판단함  1-2) 솔루션  데이터를 분석, 목표에 부합하는 유의미한 결과를 도출해내기 위한 모델 설계시 어떤 변수선택이 유의미하고 모델 성능을 높일 수 있는지에 대해 파악  또한 변수 선택과 데이터 스케일링, 어떤 분류기 모델을 사용할지를 파악한 후 시각화 과정을 거쳐서 결과 해석  **2. 분석 및 기능 설계/개발**  2-1) 데이터 수집 및 전처리  -주사용 데이터: hotel\_bookings.csv  -데이터수집사이트: Kaggle  (<https://www.kaggle.com/jessemostipak/hotel-booking-demand>)  -데이터 개요: 포르투갈 소재 호텔 2 곳의 2015-2017 3년간의 방문 고객 데이터  -분석설계: 데이터 키워드   1. Kaggle(Hotel booking demand) 2. 호텔 예약자들의 취소자들의 특성 파악 3. 호텔 예약자들의 취소 분류 예측   2-2) Google Colab을 이용한 데이터 가공  2-3) 모델 설계과정에서  모델의 성능을 높일 수 있는 변수 파악  2-4) 모델 설계과정 중 데이터 스케일링 및 시각화  2-5) 시각화 작업에 대한 해석  2-6) 이때에 어떠한 분류기 모델을 활용할지 알아보기 위해 적용  2-7) 분류기 모델 결과 해석 및 활용  **3. 프로젝트 결과 요약**  **(각 모델에 대한 평가 결과)**  연구1. 최종 특성 선택 과정   1. 가장 직관적으로 취소에 영향이 있을 것이라고 생각되는 데이터들을 선정. 2. 모델에 반영하기 쉬운 이산형 데이터를 우선적으로 고려하였음. 3. 범주형일 경우, unique 값이 너무 많은 특성은 탈락시킴.   연구2. 수학적 분류기 모델  (수학적 분류기 모델들은 standard scaling된 값을 기준으로 함)   1. 서포트 벡터 머신 분류기 SVC   훈련 데이터 점수 : 0.7790054749471478  평가 데이터 점수 : 0.7998157299606332  장점: 오버피팅 별로 안 나면서도 성능이 준수함.  단점: 모델이 너무 무거움.(시간이 오래 걸림.)   1. 최근접 이웃 모델 KNN   훈련 데이터 점수 : 0.869486655945648  평가 데이터 점수 : 0.7869587067593601  장점: 매개변수 값 따로 주지 않고서도 잘 돌아감.  단점: 오버피팅, 이것도 모델이 무거움.(시간이 오래 걸림.)\   1. 로지스틱 회귀   훈련 데이터 점수 : 0.7401837026210012  평가 데이터 점수 : 0.7609096239216014  장점: 훈련/평가 데이터 점수간의 차가 크지 않으면서도, 다른 수학적 분류기 모델들에 비해 빠르게 돌아감.  단점: 매개변수값 조정해도 성능 항상 제자리.  매개 변수의 영향을 생각보다 많이 받지 않음 일반화가 덜된 것 같다  다른 방법론으로 접근하여 과적합이 발생하지 않게 할수도 있을 것 같다.  데이터 전처리 과정에서의 | 연구2. 트리 계열 분류기(dt, rf, gbrt)   1. 결정트리   train set score is 0.990  test set score is 0.784  장점: 높은 분류예측 성능.  단점: 오버피팅 심하게 발생   1. 랜덤포레스트   train set score is 0.990  test set score is 0.812  장점: 높은 분류 예측 성능. dt보다는 오버피팅 덜 발생함.(하지만 그래도 오버피팅 문제 심함.)  단점: 오버피팅.   1. 그래디언트 부스팅   train set score is 0.689942881811842  test set score is  0.765516374905771  장점: 오버피팅 발생 x  단점: 다른 트리 계열 분류기 모델보다 성능 낮음.  n\_estimator 값 300까지 올려도 성능이 높지 x  연구3. DT, RF모델의 매개변수값 조정   1. RF모델의 max\_depth가 12~15 정도에서 가장 준수한 성능을 보이는 것을 확인 가능.   dt보다 rf 모델이 근소하게 더 높은 test 성능을 보임.  1-1. max\_depth = 15, 나머지 값은 default로 두고 랜덤포레스트 모델 분류예측 진행 했을 시 성능  -train set score is 0.798  -test set score is 0.774   1. cross\_val\_score활용한 교차 검증 검수 확인   -교차 검증 점수 : [0.7856876 0.78983853 0.78389373 0.79032794 0.78742217]  -교차 검증 점수 평균: 0.7874339931413028  2-1. 교차 검증 점수의 편차가 크지 않으며  일반화도 어느 정도 잘 되고 있다는 판단이 가능함.   1. dt 도 마찬가지.   그리드 서치 결과 선택된 criterion = ‘gini’, ‘splitter = ‘best’ 로 두고,  max\_depth는 위 그래프에서 확인된 훈련/테스트 점수 차이가 크지 않은 12로 두고 돌렸을 때,다음과 같은 결과가 도출됨  -train set score is 0.777  -test set score is 0.763  교차 검증 점수 : [0.76800465 0.76812918 0.76309672 0.77065172 0.77102532]  교차 검증 점수 평균: 0.7681815181915658  3-2. 랜덤포레스트와 비슷함. dt가 훨씬 가벼운 모델이라는 점을 고려했을 때, 상당히 좋은 모델이라고 할 수 있음.  ->이에 대한 결론: df, rf, svc를 best 취소 예측 분류 모델로 선정(단, svc는 비용 대비 효율이 떨어지므로 dt, rf 모델이 최적 모델로 여겨짐  **4. 분석도구**  -PYTHON, 머싱러닝  **5. 활용방안 및 기대효과**  호텔 예약 취소 예측 분류모델이 호텔 경영을 하는데 있어 도움을 줄 수 있을 것이라 예상된다.  취소를 어느정도 할지 예측이 가능해진다면 이에 대비하여 호텔측에서 각종 서비스를 제시하여 이를 방지함   1. 예약 취소 확률이 높은 고객에게 예약 확인 연락을 더 자주 넣기. 2. 예약 취소 확률이 높은 고객에게 재방문 할인 쿠폰 등 반드시 방문할 수 있도록 하는 유인책 제시하기. 3. 예약 취소를 막기 위해, 예약 취소 확률이 높은 고객에게 호텔 무료 조식 등의 이벤트 제시하기.   **6. 프로젝트 보완점**  -매개 변수의 영향을 생각보다 많이 받지 않으며 일반화가 덜 이루어졌음  -다른 방법론으로 접근하여 과적합이 발생하지 않게 할수도 있을 것 같다. |

|  |  |  |
| --- | --- | --- |
| 각 팀원 역할 |  |  |
|  |  |
| 이름-담당역할-업무내용-분석프로그램 등등  김기영(조장) 발표  김민송(팀원)  박소연(팀원)  이동렬(팀원)  각자 역할을 구분지어 하기 보다는 서로서로 도와가며 데이터 추출에서부터 데이터 처리과정을 함꼐 수행했습니다. | |

양식은 팀 프로젝트에 따라 세부항목 및 내용 변경가능하오니, 수정하여 작성해 주세요.

PPT 사진 첨부 및 상세 작성