* 안녕하십니까 섹션 2 프로젝트 발표를 맞은 AI 부트캠프 15기 박준영입니다. 이번 프로젝트에서는 머신러닝 분류 모델을 이용해서 ABC BANK의 고객 이탈을 예측하는 모델을 만들고 결론을 내겠습니다.

* 먼저 목차입니다. 프로젝트를 진행하는 상황을 설정하고 문제를 정의하겠습니다. 그리고 데이터 셋을 확인하고 EDA 진행을 보여드린 후 분류 모델링의 결과를 통해 해석과 결론을 도출하면서 발표를 마무리하겠습니다.

* 상황설정을 말씀드리면 저는 ABC뱅크의 데이터 직군의 신입이고 고객이탈에 대한 데이터 분석을 비데이터 직군인 직원들에게 발표하는 상황입니다.

* 본격적으로 문제를 풀어보도록 하겠습니다. 먼저 은행에서 고객이탈이 왜 큰 문제인지를 알아야 합니다. 기본적으로 은행은 수익구조는 예금과 대출금리의 차액을 얻는 예대마진 구조와 그 이외에 직접 투자를 통해 얻는 비이자 수익이 있습니다. 대출과 직접투자의 기반이 되는 자금은 고객이 은행에 예치한 돈이 됩니다. 따라서 은행에서 고객의 이탈은 수익구조의 기반이 무너지는 것이므로 제때 방지하지 않는다면 은행의 존폐의 위기가 올 수 도 있습니다.
또한 이탈한 고객은 이탈한 채로 있는 것이 아니고 돈의 저장을 목적으로 우리의 경쟁 은행으로 유입되게 됩니다. 이는 우리 은행의 점유율과 산업경쟁력 부분에서 약점으로 작용할 수 있습니다.
따라서 저는 머신러닝의 분류 알고리즘을 활용하여 고객 이탈 모델을 만들고 이탈에 주요한 영향을 미치는 특성을 분석하겠습니다. 그래서 어떤 고객을 대상으로 무슨 상품을 만들지 / 그리고 어떤 고객에게 상품 관련 알림을 보내서 이탈을 방지하고 신규 가입을 유도할 것인지에 대한 문제를 풀어보겠습니다.

* 제가 이용할 데이터 셋은 10000명의 고객정보를 12개의 특성을 통해 보여주는 데이터이고 예측할 타겟 특성은 고객 이탈을 나타내는 특성입니다. 데이터에 결측치와 중복치는 없었습니다.

* 데이터의 12가지 특성은 다음과 같으며 고객 고유 번호는 중복치가 없기에 삭제를 해주어도 무방해서 삭제를 했습니다. 또한 나이에서 가입 유지 년도를 빼 가입 당시 나이 특성을 추가하고 미국의 FIFO신용점수를 기준으로 고객의 신용점수를 5개의 신용등급으로 나누어서 신용등급 특성을 추가하였습니다.

* 성별, 신용카드 유무, 정회원가입 유무 와 같이 두 개의 결과로 나오는 특성에 따른 이탈수치를 보면 다음과 같습니다. 정회원일 때 유지비율이 높고 이탈율은 낮은 걸 볼 수 있습니다.

* 특성 별로 서로 영향을 주는 관계를 확인했을 땐 특성 간 특별한 상관관계는 나타나지 않고 독립적인 것을 확인했습니다.

* 타겟인 이탈유무 특성의 비율은 어느정도 불균형이 존재하기에 모델을 만들면서 타겟 데이터 불균형 처리가 필요할 것으로 보입니다.

* 데이터를 살펴보았으면 이제 활용할 분류 모델을 알아보겠습니다. 제가 사용할 모델은 로지스틱 회귀 모형과 랜덤 포레스트, XG부스트 모델입니다.

* 모델의 평가지표는 AUC 점수를 이용할 것입니다. AUC 점수를 간단하게 설명드리면 이탈고객을 이탈료 예측하는 확률을 TPR, 유지고객을 유지로 예측하는 확률을 FPR이라고 할 때 TPR이 높고 FPR이 낮으면 좋은 모델입니다. 각 수치를 통해서 만든 곡선을 ROC곡선이고 AUC 점수는 이 곡선의 아래 면적이고 0.8이상이면 좋은 모델이라고 할 수 있습니다.

* 저는 세 모델의 AUC점수를 비교하여 최적을 모델을 선택할 것입니다.

* 모델 분석에 앞서 기준모델을 먼저 설정해줍니다. 데이터를 학습 검증 평가로 각각 8:2 비율로 나누고 타겟 데이터의 최빈값으로 기준모델의 AUC점수를 확인했을 때 0.5임을 알 수 있습니다.

* 본격적으로 첫 번째 모델은 로지스틱 회귀 모델입니다. 회귀를 사용하여 데이터가 어떤 범주에 속할 확률을 예측하고 그 확률에 따라 가능성이 더 높은 범주로 분류하는 모델입니다.
 타겟 데이터가 불균형한 것은 클래스 웨이트 설정을 통해 조정해주고 학습 데이터 학습 후 먼저 검증 데이터 성능을 확인해보면 0.7정도의 점수가 나왔습니다.

* 평가 데이터의 ROC커브를 그리고 평가데이터의 AUC점수를 확인했을땐 0.723임을 알 수 있습니다. 이후 분석에서 ROC커브는 생략하겠습니다.

* 다음은 랜덤포레스트 모델입니다. 훈련 과정에서 구성한 다수의 결정 트리로부터 평균 예측치를 출력해서 동작하는 모델로서 이 역시도 클래스 웨이트로 타겟 데이터 불균형을 조정해주겠습니다.
 또한 랜덤포레스트 모델은 다양한 하이퍼 파라메타를 통해 과적합을 방지합니다. 하이퍼 파라메터는 최적의 모델을 구현하기 위해 설정하는 변수라고 생각해주시면 됩니다. 랜더마이즈 서치 방식으로 최적의 파라미터를 설정하고 재학습을 했습니다.

* 그 결과 검증과 평가 데이터의 점수가 나왔고 0.86의 평가 데이터 점수가 나옴을 확인 할 수 있습니다.

* 마지막으로 XG부스트입니다. 부스팅 기법은 약한 예측 모형들의 에러에 가중치를 두고, 순차적으로 다음 학습 모델에 반영하여 강한 예측모형을 만드는 것인데 XG부스트가 가장 잘 활용되는 모델입니다.
 스케일 포스 웨이트를 통해 타겟불균형을 조정하고 역시 랜더마이즈서치를 통해 최적의 파라메터를 설정합니다.

* 또한 얼리스타핑이란 방법을 통해서 학습을 하더라고 성능이 더 이상 증가하지 않으면 정지하면서 신뢰성을 높이겠습니다. 그 결과 평가 데이터 점수는 0.85임을 알 수 있습니다.
 랜덤포레스트와 XG부스트의 점수가 거의 동일하게 나오므로 저는 XG부스

트를 기준으로 분석을 하기로 하였습니다.

* XG부스트를 기준으로 순열중요도 기법으로 주요한 영향을 미치는 특성을 정해보겠습니다. 순열중요도는 각 특성을 무작위로 배열해서 중요도를 계산하는 방법입니다.

* 그 결과 가입 상품 번호, 나이, 정회원 가입 이 세가지 특성이 고객 이탈에 주요한 영향을 미치는 것으로 보입니다.

* 가입 상품 번호를 보면 2번 상품의 이탈율이 낮고 3,4 상품의 이탈율은 높게 영향을 줍니다.

* 나이대로 보게 되면 40대 중반 이후 이탈율이 높게 나타납니다.

* 또한 정회원으로 가입한 고객의 이탈율이 낮게 나타남을 알 수 있습니다.

* 정리하면 다음과 같습니다.

* 계약 상품 종류에 대해서 좀 더 살펴보면 1,2는 인기와 유지비율이 준수하지만 3,4 는 가입률이 적고 이탈율은 유지비율 보다 높으므로 고객만족도가 현저히 떨어짐을 알 수 있습니다. 따라서 차라리 3,4 상품을 폐기하면서 유지비를 절약하고

* 그 예산으로 4,5O 대 중장년층을 대상으로 하는 새로운 상품 개발이 필요해 보입니다.

* 또한 고객의 정회원 가입 유도는 모든 영업과정에 필수가 되어야 할 것입니다.

* 최종적으로 보면 코로나 시기와 모바일 금융의 일반화로 50대 이상의 대부분의 고객들이 모바일 금융서비스를 이용하는 것을 활용하여 서비스를 제공하고자 합니다. 3,4 상품을 보완하는 상품을 개발하고 50대 이상 정회원이 아닌 고객에게 앱 알림을 통해 정회원 가입시 혜택이 주어지는 새로운 상품 알림을 발송하여서 이탈율을 줄이고 신규 가입율을 높일 수 있을 것이라 결론을 내게 되었습니다.