안녕하세요. AI 15기 신희호입니다. 인도 지역에서 직원 이탈 여부 예측을 주제로 발표를 진행하겠습니다.

발표 순서는 다음과 같습니다.

문제 정의입니다.

직원들의 이탈은 회사 입장에서 큰 손실로 이어질 수 있습니다. 숙련된 직원들의 이탈도 기업의 핵심 인력이 이탈하는 것이므로 타격이 크지만, 신입 직원들의 이탈 또한 회사 입장에서는 타격이 큽니다. 회사는 신입 직원을 채용하여 그들의 육성을 위해 투자를 진행하는데, 그들이 육성 중에 이탈하면 그대로 손해를 입게 됩니다. 기존의 숙련된 직원들은 또 다시 신입 직원들을 교육해야 할 것이고, 교육이 충분하게 이루어질 때까지 숙련된 직원들에게 업무가 과중될 것입니다. 이는 숙련된 직원들의 불만으로까지 이어질 수 있습니다.

따라서, 어떤 특성을 가진 직원이 이탈할지 예측하고, 이를 바탕으로 직원의 이탈을 막기 위해 어떤 전략을 세우면 좋을지 살펴보고자 합니다.

첫번째로, 직원의 이탈에 영향을 가장 많이 주는 특성은 무엇인지,

두번째로, 직원의 이탈을 막기 위해 세울 수 있는 전략은 무엇인지 분석해보겠습니다.

가설설정입니다.

신입 직원이 가지고 있을 법한 특성을 중심으로 가설 설정을 진행하였습니다.

1. 최근 입사한 직원일수록 이탈 비율이 높을 것이다.

2. 경험이 적은 직원일수록 이탈 비율이 높을 것이다.

3. 나이가 어린 직원일수록 이탈 비율이 높을 것이다.

이렇게 세 가지 가설을 설정했습니다.

데이터설명입니다. 시간 관계 상 자세한 내용은 생략을 하고,

몇 가지 유의 사항에 대해 짚어보자면 PaymentTier 급여 수준은 1이 가장 높고, 3이 가장 낮습니다. EverBenched는 프로젝트에 참가하지 못했는지 여부입니다. Yes면 참가하지 못한 것이고, No는 참가한 것입니다. 또한, 예측해야하는 특성인 LeaveOrNot이 타겟이 됩니다. 그리고 Education에서 학사 학위는 1 석사 학위는 2 박사 학위는 3으로 설정했습니다.

데이터 전처리입니다.

결측치 이상치 모두 없었습니다. 중복행은 1889건이 발견되어 드랍하기로 결정했습니다. Age는 편의상 다음과 같이 4살 단위로 그룹화했습니다. ExperienceInCurrentDomain은 6,7년 경험에서 데이터 수가 부족해 드랍했고, 특성이름을 간단하게 Experience로 변경했습니다.

탐색적 데이터 분석입니다.

앞서 Train, Test 분리를 해주었습니다. Train은 모델이 학습할 데이터이고 Test는 모델이 예측할 미지의 데이터로 설정이 됩니다. 따라서 Train을 기반으로 Test에서 높은 점수를 받는 것이 중요합니다. 타겟의 균형이 잡혀있어야 보다 일반화된 예측이 가능해집니다. Train Test 각 세트에서 불균형이 심하지 않고, 세트 간의 불균형도 없는 수준입니다.

현재 알고 있는 데이터라고 설정된 Train만으로 탐색적 데이터 분석을 진행했습니다.

앞서 봤듯이 Train에서 1의 비율은 0.39였습니다. 불균형도가 심하지 않아 그대로 진행했습니다.

가설 설정과 관련된 변수 위주로 보면 2017년에 입사한 직원들이 가장 많고, 2번 연령대 그룹이 가장 많고, 2년 경험의 직원들이 가장 많은 것으로 확인됩니다.

특성들의 분포도입니다. 범주를 벗어나는 값들은 확인되지 않았습니다.

타겟 0과 1에 따라 분리하여 직원 수를 파악한 것입니다. 2018년에 이탈한 직원 수가 가장 많고, 연령대 2번 그룹이 가장 많이 이탈했고, 2년 경험의 직원들이 가장 많이 이탈했습니다.

이 그래프는 타겟 비율 그래프인데 1에 가까울수록 떠난 직원의 비율이 높은 것을 의미합니다. 2018년 입사한 직원들은 대부분이 나간 것을 확인할 수 있고, 연령대 1번 그룹이 가장 높은 비율을 보이고 있고, 3년 경력의 직원들의 비율이 가장 높은 것을 확인할 수 있습니다.

숫자로 이루어진 특성과 이탈 여부와의 상관관계가 어떤지 보여주는 표입니다. 절댓값이 높을수록 상관성이 높고, 양수면 양의 상관성, 음수면 음의 상관성을 갖습니다. JoiningYear는 양의 상관성을, PaymentTier와 Age는 음의 상관성을 보이는데 상관성 자체는 낮은 것으로 확인됩니다.

모델링 및 모델 평가입니다.

평가의 기준이 되는 기준 모델을 통해 파악한 결과 정확도가 0.61로 확인되었습니다.

사용할 모델은 예측 성능이 좋다고 알려진 다음 3가지를 선정하여 비교했습니다.

해당 모델들은 의사결정나무 기반으로 Boosting이라는 기술을 사용합니다.

보다 일반적인 모델을 만드는 것이 중요하기 때문에 부스팅은 약한 모델을 여러 개 결합하여 결과를 종합하는 것입니다. 이 과정에서 잘못 예측한 데이터를 잘 분류할 수 있도록 분류 성능을 점차 개선하는 알고리즘이 포함되어있습니다.

숫자형 변수가 아닌 City, Gender, EverBenched는 Ordinal Encoding 기술을 사용하여 문자에서 숫자로 변경을 해주었습니다.

성능 지표로 기준 모델과의 비교를 위해서는 정확도를, 모델 간 비교를 위해서는 F1을 사용했습니다. F1은 정밀도와 재현율이라는 지표를 모두 중요시할 때 사용하는 지표입니다. 정밀도는 떠날 것이라고 예측한 직원들 중에서 실제로 떠난 직원들의 비율이고, 재현율은 실제 떠난 직원들 중에서 떠날 것이라고 예측한 직원들의 비율입니다. 이 두 가지가 모두 중요하다고 판단하여 F1을 성능 지표로 삼았습니다.

KFold 기법을 통해 모델 비교 및 성능 개선에서 검증 세트를 따로 두지 않고 성능 평가를 진행하였습니다.

이를 바탕으로 모델 간 비교를 먼저 실시하였고, 그 결과 모델 중에서 LightGBM이 F1 성능에서 높은 수치를 보였습니다. 기준 모델과 비교했을 때 정확도는 0.17 가량 높아졌습니다. LightGBM의 수치를 조정 후 결과를 보면, 0.05, 5% 가량 정확도와 F1이 높아졌습니다. 이 모델을 기반으로 Test 세트 결과를 확인해보니 다음과 같은 성능을 보였습니다. 거의 차이가 없었고, 과적합한 모델도 아니라고 판단했습니다.

특성 중요도는 해당 특성이 타겟을 예측하는데 얼마나 영향을 주는지에 관한 수치입니다. 가설과 관련하여 JoiningYear가 가장 높고, Experience, Age 순으로 파악됩니다.

PDP 플롯을 통해 특정 특성의 값을 변화시킬 때 모델의 예측 양상이 어떻게 변하는지 파악할 수 있습니다.

JoiningYear에서 2012년부터 2017년까지 큰 차이 없다가 2018년에 이탈 비율이 높아집니다.

Experience는 각 연차에서 큰 차이가 없는 것으로 파악됩니다.

Age는 나이가 적은 연령대 그룹일수록 이탈이 높은 것으로 확인됩니다.

결론입니다.

먼저 가설 검증입니다.

최근 입사한 직원일수록 이탈 비율이 높을 것이고, 경험과 이탈 비율은 관계가 없으며, 나이가 어린 직원일수록 이탈 비율이 높을 것입니다.

가장 최근 2018년에 입사한 직원들에 대해 심층적인 설문조사가 필요합니다.

이들은 거의 모든 신입이 퇴사하였으며 이들이 퇴사를 결정한 이유를 기반으로 신입 채용 과정을 보완해야 합니다.

나이가 적은 직원 그룹에서 이탈 비율이 높기 때문에 이들의 관심사 파악을 통해 회사에 정착할 수 있도록 유인이 필요합니다.

추가로, 여성의 이탈 비율이 높기 때문에 직장 내 부당한 대우가 있지는 않는지 직무 특성 상 여성의 고충이 많은지 등을 조사해볼 필요가 있습니다. Payment Tier2 직원의 이탈 비율이 높기 때문에 급여 만족도 등에 대한 조사가 필요합니다. 근무지 Pune의 이탈 비율이 높기 때문에 Pune 지역 근무지 시설이나 근무지 내 복지 환경 등을 조사할 필요가 있습니다.

보완해야할 점은 구체적인 임금이나 만족도 설문조사 항목을 추가해 분석하면 좀 더 의미 있는 분석이 될 것이라 생각하지만, 설문조사의 신뢰도는 생각해보아야 하는 문제일 것입니다. 이상으로 발표 마치겠습니다. 감사합니다.