프로젝트 요약 보고서

1. 데이터 분석 또는 모델링 배경 및 문제 설정 - 관측된 현상, EDA를 해본 결과 진단된 풀어야 할 문제 상황 또는 주어진 문제상황

차량 예약 서비스를 제공하는 Lomous.com의 설립자인 억만장자 사업가 엠제이 드마코는 그의 저서 '부의 추월차선' 에서 이렇게 말했다. *'인적 자원 시스템은 자원 중에 가장 다루기 힘든 자원이다.'* 사회생활 경험이 있는 성인이라면 이 말에 고개를 끄덕일 것이다.

사람에게는 모두 각자만의 성향이 있다. 성향이 비슷한 사람들끼리 모인다면 사람들을 관리하는 입장에서는 다소 수월함을 느낄 것이다. 하지만, 성향이 다른 사람들이 수십명도 아닌 수백명이 모여있는 곳이라면 어떨까?

본 프로젝트에서는 *사람에서 시작해서 사람으로 끝이 나는* HR에서 발생한 데이터들을 분석해 문제점들을 찾아보고, 그에 맞는 해결책을 제시해보고자 한다.

한국경영자총협회가 2013년에 발표한 대졸 신입사원의 1인당 교육비용은 8,630만원이다. (대기업 기준) 물가상승률을 고려한다면, 현재의 신입사원 1인당 교육훈련 비용은 더 증가했을 것으로 추정된다. 이는 *조기퇴직자가 증가할 수록 기업에 부담이 된다는 것을 시사*한다.

위의 정보를 상기시키면서 다음의 그래프를 보자. IBM에서 제공한 퇴직자 데이터를 가공하여 만든 '퇴직자의 근무 년수'를 나타내는 그래프이다. 막대가 0~10년에 밀집되어 있는 것을 보면 퇴직자들 중 조기퇴직자의 비율이 상당히 높음을 알 수 있고, 우리는 이 그래프를 통해서 조기 퇴직자들이 IBM의 경영에 있어 교육비용 등의 *금전적인 영향 등을 끼칠 수 있다는 문제*를 도출해 낼 수 있다.

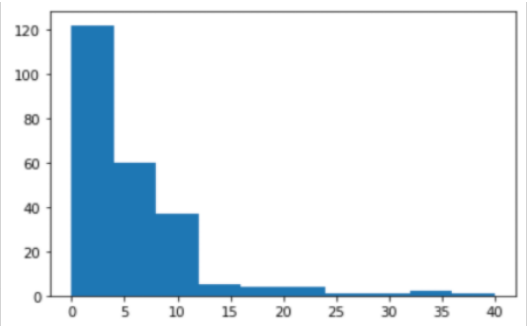

그림 1 IBM 퇴직군의 근속 년수

2. 해당 프로젝트를 진행하는 이유-이것을 해결함으로써 얻고자 하는 것 (+ Optional 해당 프로젝트를 취업 포트폴리오 관점에서 하는 이유)

본 프로젝트에서는 (조기)퇴직자들의 패턴을 분석하여 신입사원 채용시에 이를 활용하고자 한다. 또 신입사원 뿐만 아니라, 현재 근무 중인 임직원들중 퇴직자 패턴과 유사한 성향을 보이는, 즉 퇴직 가능성이 높은 직원들을 선발하여 제시된 해결책을 토대로 조기퇴직을 막고자 한다. 그 이유는 1에서 언급했던 바와 같이 조기 퇴직자 발생에 따른 기업의 경영적인 손실 부담을 낮추기 위해서 이다.

취업 포트폴리오 관점에서 진행하게 된 이유는 두 가지 관점에서 설명할 수 있다. 첫째, 우리 모두는 잠재적 사업가다. 본 프로젝트를 통해서 얻게 된 인사이트를 활용해 각자의 사업

혹은 타인의 사업체에서 든든한 조력자로서 활동할 수 있다. 단순 프로젝트로 끝나는 것이 아닌 것이다. 둘째, 연차가 쌓인 직장인이라면 팀장과 같은 리더급의 자리에 오르게 되어있다. 이 때 팀원 중 한 사람이 데이터 분석을 통해 도출해낸 (조기)퇴직자의 패턴을 보이는 사람이라면 무언가 조치를 취해야 한다. 만일, 본 프로젝트에서 제시한 해결책을 활용한다면 불필요한 퇴직을 막을 수 있는 가능성이 높아지며 기업의 손실도 방어할 수 있게 된다.

다시 말해, 리더십 역량을 키우기 위해서도 삶을 살아가면서 언젠가 유용하게 활용할 수 있는 인사이트를 얻기 위해서도 진행해야 하는 프로젝트라고 판단하였다. 이는 자신은 물론 타인 그리고 더나아가서는 기업에도 긍정적인 영향을 미칠 수 있기 때문이다.

3. 데이터셋 - 사용한(수집한) 데이터셋

Kaggle에서 제공되는 IBM HR Analytics Employee Attrition & Performance 데이터를 활용하였다. 독립변수 34개, 종속변수 1개의 데이터 셋이다. 총 1,400개의 raw data가 있으며 결측 데이터는 없었다.

4. 결과 및 액션 - 얻어낸 유의미한 정보. 정의했던 문제와 연관지어 원인을 찾거나 이에 대한 해결이 필요한 정보를 전달.

결정트리 모델(Baseline model, Accuracy 0.74)과 로지스틱 회귀모델(Accuracy 0.86)에서의 Important Feature의 경우 공통적인 항목이 *Overtime(초과근무)*임을 확인할 수 있었으며, 이 외에 결정트리 모델을 통해서 *MonthlyIncome(월 수입), TotalWorkingYears(경력기간), DailyRate(일 대비 급여 수준), HourlyRate(시간 대비 급여수준), Age(직원의 나이), YearsAtCompany(근속 연수) 항목이*, 로지스틱 회귀모델의 경우 *Department(업무 분야), PerfomanceRating(업무 성과), DistanceFromHome(집과의 거리), YearsSinceLastPromotion(마지막 프로모션), MaritalStatus(결혼 여부), Gender(성별)* 항목이 (조기)퇴직에 영향을 미치는 주된 항목임을 알 수 있었다.

이를 통해, (조기)퇴직자의 발생을 줄이기 위해서는 *초과 근무 경감*과 *급여에 대한 재정비가 필요*하다는 해결책을 제시할 수 있겠다. 이 외에도 (조기)퇴직자 그룹의 동향을 살펴보면 로지스틱 회귀모델을 통해 알 수 있듯이, 회사와 집과의 거리가 먼 직원들 중 퇴직자가 발생할 확률이 높았다. 결혼 여부의 경우 미혼-결혼-이혼의 순으로 퇴직자의 비율이 높았으며, 성별의 경우 여성보다 남성의 비율이 높았다. 이를 통해, 현재 근무 중인 임직원들 중 미혼 상태의 임직원과 성별이 남성인 임직원의 동향을 살펴볼 필요가 있다. 하지만, 이 둘의 상관관계가 어느 정도 높은지는 추가적인 분석이 필요하다.

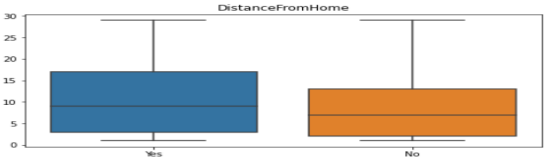

그림 3 퇴직군 그룹(Yes)과 대조군(No)의 집과의 거리

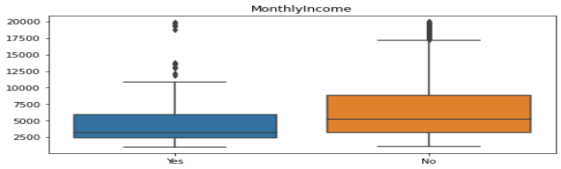

그림 2 퇴직군 그룹(Yes)과 대조군(No)의 월 수입

5. 분석 내용 - 분석 진행 방법 및 내용 요약 (시각화, 분석, 모델링 툴 또는 라이브러리 사용 확인 가능 시 기재)

가설 1. 퇴사율이 높은 직군은 업무 분화가 덜 되어 있을 것이다.

2023년 현재, 대한민국에서는 데이터 직군간의 경계가 모호하다는 의견이 분분하다. 데이터 직군이 국내에 도입된지 오랜시간이 지나지 않았고 기업마다 직무를 구분하는 기준도 조금씩 다르기 때문이다.

이 사실을 바탕으로 *'(조기)퇴직자가 가장 많이 발생한 Top3 직군의 업무 경계 또한 모호하여 업무의 세분화가 되어있지 않은 상태, 즉 업무의 과부하가 일어난 상태가 아닐까'* 하는 의문이 들었다. 조기 퇴직의 원인이 단순히 '급여가 적어서'의 문제가 아닐 수 도 있다는 관점이다.

이를 확인해보고자 조기퇴직자들이 가장 많이 밀집해있는 High-Top3직군(Laboratory Technician, Sales Executive, Research Scientist)과 조기퇴직자들이 가장 적은 직군 Low-Top2(Research Director, Manager)의 업무 만족도와 급여수준을 비교해 보았다.

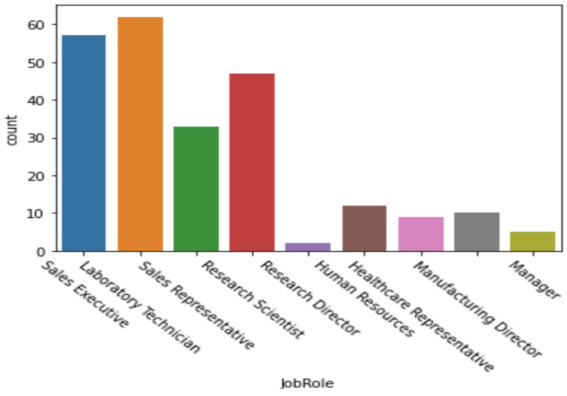

그림 4 IBM (조기)퇴직자들의 업무 종류

그 결과 놀랍게도 Low-Top2직군의 급여 수준은 각각 $19,250와 $12,000에서 시작되며 업무 만족도에 대한 응답의 경우 Research Director직군은 High, Medium 두 레벨만 존재했으며, Manager직군의 경우 Very High, High, Medium, Low 4가지의 응답이 있었으나 이 중 Medium항목의 응답이 가장 높았다. 그에 반해, High-Top3 직군의 경우 급여 수준이 각각 $1,000, $4,000, $1,000에서 시작되었으며, 업무 만족도에 대한 응답에서는 4가지 항목 모두 존재했으나 이중 High, Low의 비중이 가장 높았다.

업무에 관한 상세데이터의 부족으로 각 직군별 업무의 세분화 여부를 알 수 없었고 이에 가설에 대한 검증 결과는 불명확하지만, *업무의 대한 만족도*와 *급여 수준*이 직군간 퇴직율에 영향을 미침을 확인할 수 있었다.

가설 2. 달성해야 할 목표가 높아서 사기가 저하됐을 것이다.

각 직군의 평가제도에 대해서도 살펴볼 수 있어야 한다. 초우량기업의 경우 영업직군의 자신감 및 자기효능감 등을 충족시키기 위해 일부러 해당 직군의 70-80%의 인원이 달성할 수 있는 목표를 세운다고 한다. 임직원이 스스로에게 '낙오자'라는 이미지를 심지 않도록 하기 위함이다. 반면, 40-50%의 인원이 달성할 수 있는 목표를 세운다면 목표를 달성하지 못한 직원들의 사기 저하와 함께 영업 이익의 하락을 유발한다고 한다. [참고문헌 - 초우량기업의 조건]

이 사실을 바탕으로 IBM에서 *각 직군에게 내린 목표가 직군 내 40~50%에 달하는 인원만이*

달성할 수 있는 정도로 책정이 되어 있는지 , 이것이 퇴직에 영향을 미쳤는지 검증해보고자 한다.

우측의 막대그래프를 보자. *(조기)퇴직한 모든 직원의 업무 성과*를 그래프로 나타낸 것이다. 이를 통해서 알 수 있는 것은 퇴직군의 84.6%에 해당되는 사람이 매우 좋은 성과를 냈다는 것이며 이에 과도하게 높은 혹은 적은 인원만 달성할 수 있는 목표를 설정하지 않았음을 알 수 있다. 따라서 가설 2는 틀렸음을 확인할 수 있다.

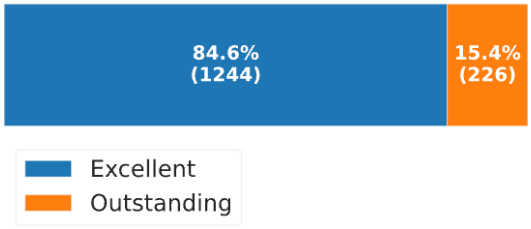

그림 5 (조기)퇴직자들의 업무 성과 비율

가설 3. 스톡옵션(주식매수청구권) IBM의 주가가 하락하여 퇴직률이 상승했을 것이다.
2022년 12월 네이버와 카카오뱅크는 경기 침체 및 증시 불황을 맞이했고 이로 인해 스톡옵션 가치가 급락을 했다. 이에 따라, 네이버의 경우 임직원 239명이 스톡옵션의 권리를 '포기'했다. 퇴사를 한 것이다.

위의 사례를 통해 IBM에서도 동일한 경우가 발생하여 (조기)퇴직자의 발생이 일어난 것이 아닌가 하는 가설을 세우게 되었다.

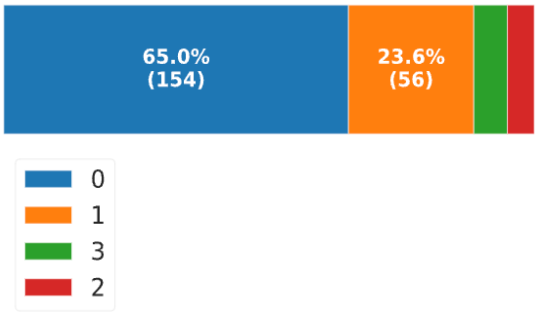

그림 6 IBM 퇴직군의 스톡옵션 보유 레벨

하지만 데이터 분석 결과, 퇴직자들의 스톡옵션 보유 레벨을 확인해보면, '거의 소유하고 있지 않음' 과 '낮은 레벨로 소유하고 있음' 의 합이 높았다. 따라서, 스톡옵션을 소유하고 있었으나 가치가 하락하여 (조기)퇴직을 했다라는 가설보다는 '소유하고 있지 않음' 혹은 '낮은 레벨로 소유하고 있음' 자제가 원인이 되었음을 추론해볼 수 있다. 이에 가설 3은 틀린 것으로 확인됐다.

6. 개선 필요한 점 - 부족했던 점이나 추가로 있었으면 더 좋았을 데이터 등 개선을 위한 고민의 결과
뉴스 기사와 참고 문헌을 탐독하며 HR에 관한 도메인 지식을 쌓고자 했다. 또 당연한 가설과 당연한 해결책이 제시되지 않도록 퇴직자를 하나의 데이터로만 바라보는 것이 아니라, 한 명의 개인으로서 그 심리를 읽어보려는 시도를 했다.

하지만 본 프로젝트를 마치고 나니 *데이터(분석)에 대한 접근법*이 엉뚱했던 것은 아닌가 하고 의문을 갖게 되었다. 그 이유는 다음과 같다. 금번에 사용한 데이터셋은 데이터 과학자가 인공으로 만들어낸 데이터셋이다. (이 사실은 본 프로젝트가 마무리 되기 3일전에 알게 되었다.) 이에 *처음부터 데이터셋을 통해 말하고자 하는 바가 미리 정해져있던 것은 아닐까*하는 생각이 들었다. 그리고 이것들은 본 프로젝트에서 피하고자 했던 *당연한 가설*과 *당연한*

*해결책*인 것 같다는 추론을 하게 만들었다. 동일한 데이터셋을 사용한 다른 분석가들의 포트폴리오를 참조해보니 모두가 비슷한 가설과 이에 따른 비슷한 해결책을 제시했기 때문이다.

*'당연한 것을 당연하게 받아들이려 하지 않다보니 엉뚱한 가설이 나왔다.'*가 본 프로젝트의 핵심적인 실책점이며 역량 증진의 필요성을 느끼게 하는 뼈 아픈 교훈이 되었다.

또 본 프로젝트에 온전한 시간을 쏟지 못했기에 '용두사미'의 느낌을 지울 수가 없다. Colab과 python이 아닌, SQL을 활용한 BigQuery와 Tableau를 활용해보려 시간을 투자했으며, 이커머스 데이터셋에 대한 미련을 버리지 못해 프로젝트 기간 중에 타 유료 교육매체를 통해 그로스 해킹에 대한 전문 지식을 쌓거나, [그로스 해킹, 린분석, 디맨드, 빅데이터 사람을 읽다] 와 같은 서적을 읽는 등 시간을 할애했다. 여기에 이커머스 관련 데이터셋으로 변경을 시도하는 둥 무모한 행위까지 저질렀다. IBM HR 데이터셋에 온전히 집중하지 못한 점을 반성한다.

금번 프로젝트를 통해서 현재 어느 수준에 있는지 인지하게 되었으며, 그동안 배운 내용을 단순 지식이 아닌 자신의 것으로 만들어 활용해야 하는 필요성을 강하게 느꼈다. 또 도메인의 중요성도 체감하게 되었다. 다음 프로젝트는 그동안 축적해왔던 이커머스 관련 지식을 토대로 진행을 해보려 한다. 끝.

참고자료

[조선일보] 휴지조각 된 '잭팟'... 스톡옵션 포기하고 줄퇴사
https://www.chosun.com/economy/industry-company/2023/02/06/ZJEX3U5HBVBTBAETB2YK6ZALXM/
[jobsN] <퇴사공화국 ⑤> "퇴사율을 낮춰라", 기업들 이탈 막기 사활
https://m.post.naver.com/viewer/postView.nhn?volumeNo=16721795
[초우량 기업의 조건] - 톰 피터스, 로버트 워터먼 지음
[부의 추월차선] - 엠제이 드마코 지음

사용 툴

pandas, pandas_profiling
numpy
matplotlib.pyplot
seaborn
sklearn.pipeline - make_pipeline
sklearn.model_selection - train_test_split, GridSearchCV, learning_curve
sklearn.preprocessing - LabelEncoder, RobustScaler
sklearn.metrics - accuracy_score, roc_curve, roc_auc_score, recall_score, f1_score, confusion_matrix, precision_score, classification_report
sklearn.linear_model - LogisticRegression
sklearn.tree - DecisionTreeClassifier, export_graphviz