**[데이터마이닝 보고서]**

3조 김현우 이다현 이호은 장지우 전현일

**[목차]**

**1. 분석 개요**

- 데이터 출처

- 분석 목적

- 통신사 데이터의 특징

- 데이터 소개

- 변수 소개

- 변수의 기초통계량 소개

**2. 데이터 전처리**

- 전처리 방안

- 결측치 / 이상치 / 파생 변수 추가 / encoding

- 전처리 방안의 문제점

- Data Leakage 현상 파악

**3. 모델링 방안**

- 모델링의 주요 흐름

- 사용 모델

- 데이터 분할 기준 및 활용 방안

- 데이터 분석 기본 Flow

- 주안점

- Double Modeling

- Pre-trained Feature Selection

**4. 최종 결과**

- 랭킹

- 최종 순위

- 모델 별 랭킹

- 데이터 전처리 방안 별 랭킹

- 모델링 방안 별 랭킹

- 성능

- 성능 평가 기준

- 사용자화 함수 제작

- 한계점

- 분석 목적에서의 한계

- 다양한 평가지표의 부재

- Data Leakage 컨트롤

- Etc

1. 분석 개요

(1) 데이터 출처

이번 분석에 사용한 데이터는 IBM Sample Data Sets에서 제작한 가상의 통신사 고객 데이터로, 특정 고객이 현재 가입되어 있는 통신사에서 이탈할 지 여부를 예측하는 문제를 해결하는 데에 사용됩니다. 데이터는 7043개의 관측값과 20개의 변수로 구성되어 있으며 KAGGLE에서 구할 수 있었습니다.

(2) 분석 목적

본 데이터는 미국 캘리포니아 주에 거주하는 주민을 대상으로 3분기에 수집되었음을 가정하고 있습니다. 정확한 고객의 이탈 시점을 예측하는 것이 통신사에게 실질적으로 도움이 되리라 예상하지만 원본 데이터에는 이탈에 대한 시계열 정보가 포함되어있지 않기 때문에 단순 이탈 여부만을 파악할 수 있습니다. 따라서 자료를 수집한 시점 직후에 이탈한 고객을 파악하지 못한다는 현실적인 한계에 봉착할 수 밖에 없습니다. 한편, 통신 산업과 같은 경우, 대부분의 국가에서 독과점 시장이 형성되어 있다고 합니다. 이 사실을 통해 저희는 새로운 고객의 유입 가능성이 현저히 낮으며 통신사는 기존 고객의 이탈에 중점을 두어 관리한다는 전략을 이해할 수 있었습니다. 이러한 판단을 통해 이번 분석 결과를 단기간에 반복적으로 적용하는 방안이 비록 비경제적일 수 있으나, 시계열 정보가 없다는 한계점을 해결하는 데에 도움이 될 수 있으리라 생각했습니다.

(3) 데이터 소개

(데이터 프레임 캡처 사진)

이번 분석에 사용한 데이터는 타겟 변수인 “Churn”변수를 포함해 총 20개의 변수와 7043개의 관측값으로 구성되어 있습니다. 20개의 변수는 크게 3가지 기준을 통해 분류할 수 있습니다. 고객의 인구통계(성별, 연령대, 배우자의 유무, 부양 가족의 유무)와 가입한 서비스의 종류(가입 개월 수, 핸드폰 서비스 가입 여부, 다중회선 여부, 인터넷 서비스 가입 여부, 온라인 보안 및 백업 서비스 가입 여부, 기기 보호 서비스 가입 여부, 기술 지원 서비스 가입 여부, TV 및 영화 스트리밍 서비스 가입 여부), 그리고 계약정보(가입 기간, 결제 방식, 디지털 청구 여부, 월별 요금, 총 요금)이 그 기준입니다.

데이터의 구조적 특징을 살펴보았습니다. 변수의 개수가 20개이지만 일부 변수는 존재하는 변수들의 선형결합으로 표현이 가능한 것이 하나의 특징입니다. 또한 17개의 변수가 범주형 변수이고, 3개의 변수만이 수치형 변수라는 것이 또 다른 특징이라고 파악했습니다. 이럴 경우 범주형 변수에 대해 인코딩을 진행하게 되면 데이터의 차원이 급격히 증가할 수 있고 사분위수를 기준으로 하는 이상치를 파악하는 데에 한계가 생길 수 있다고 예상했습니다.

(개별 변수 분포 사진)

위 사진은 개별 변수들의 분포와 타겟 변수인 Churn변수와의 관계입니다. 저희가 위의 그래프 중 모델링의 근거로 사용한 부분은 [가입 기간에 따른 이탈 여부] 입니다. 가입 기간 개월 수를 의미하는 변수 [Tenure]가 12이상일 때와 12이하일 때를 기준으로 구분하여 이탈 여부를 파악해보면 타겟 변수의 분포가 유의미한 차이가 존재하고 있음을 확인할 수 있습니다. 이를 통해 후에 모델링 과정에서 데이터를 특정 기준으로 세분화한 후 각각 모델을 적용해 결과값을 도출하는 과정을 진행할 수 있었습니다.

(1) 전처리 방안

데이터의 기초통계량을 바탕으로 저희가 고안한 전처리 방법의 경우의 수는 총 32가지입니다. 우선 첫번째는 결측치 처리입니다. 7043개의 데이터 중 결측치는 고객이 지금까지 낸 총 요금인 [TotalCharges] 변수에서 11개 발견되었습니다. 이는 가입 기간이 1달 이하인 고객이 아직 요금을 지불하지 않았기 때문에 생긴 결측치라고 판단했으며, 전체 데이터 대비 그 개수가 적기 때문에 데이터를 삭제하는 것으로 결정했습니다.

두번째는 이상치 처리입니다. 수치형 변수에 대해 사분위수 기준 이상치를 찾는 알고리즘을 적용한 경우에 이상치를 발견할 수 없었습니다. 하지만 가입 기간 개월 수를 기준으로 데이터를 분할했을 때 타겟 변수의 분포가 상이함과 마찬가지로 가입 기간 수에 따른 부분집합에서는 상당 수의 이상치 데이터를 파악할 수 있었습니다. 이에 저희가 설정한 첫번째 이상치 기준은 [가입 기간 개월 수를 기준으로 한 부분집합에서의 이상치] 입니다. 데이터의 구조적 특징으로 다수의 범주형 변수가 있음을 파악했는데 이에 다른 수치형 변수에서는 이상치를 찾을 수 없었습니다. 하지만 고객 데이터의 특성 상 데이터 수집 과정에서 이상치가 생길 수 있다는 가정 하에 의미론적인 이상치를 찾았습니다.

데이터의 또다른 구조적 특징으로 변수간의 선형결합으로 표현될 수 있는 변수 조합이 존재함을 파악했는데 이에 선형결합의 결과값에서 크게 벗어나는 데이터를 이상치로 설정할 수 있었습니다. 예를 들어 가입 개월 수 X 월별 지불 금액 = 총 지금 금액의 식을 생각할 수 있는데 좌변의 값과 우변의 값에서 큰 차이가 생길 경우 이를 이상치로 파악한 것입니다. 따라서 저희의 두번째 이상치 기준은 [|가입 개월 수 X 월별 지불 금액 – 총 지금 금액|의 사분위수 기준 이상치] 입니다. 이상치를 삭제하는 것이 항상 좋은 결과를 장담할 수는 없기 때문에 모든 이상치 기준을 적용해 삭제하는 경우의 수 부터 이상치 기준을 적용하지 않는 경우의 수 까지 총 4가지 파생 데이터를 생성할 수 있었습니다.

다음은 인코딩입니다. 범주형 변수를 모두 OneHotEncoding할 경우 데이터의 차원이 40개 이상으로 증가하게 되고 학습에 악영향을 끼칠 수 있으리라는 판단에 이진분류로 변환 후 인코딩하는 방법론을 고안했습니다. 인터넷 서비스 가입 여부를 나타내는 InternetService 변수와 같은 경우, 가입하지 않은 데이터는 “No”라고 기록되어 있지만 가입한 경우 어떤 회사의 서비스에 가입했는지 기록되어 있습니다. 이를 인코딩할 경우 회사의 종류 수 만큼 데이터의 차원이 증가하기 때문에 No가 아닌 데이터를 모두 Yes로 변환한 후 인코딩하는 경우의 수를 생성했습니다. 인코딩은 모델에 적용하기 전 필수적으로 필요한 단계이기 때문에 단순 OneHotEncoding과 이진분류로 변환 후 OneHotEncoding으로 파생 데이터를 생성할 수 있었습니다.

이밖에도 데이터의 분포를 변환하는 분포 변환 과정(4가지)과 기존에 존재하는 변수간의 선형결합을 통해 새로운 변수(영상매체 선호도 변수 / PhoneService 가입 여부 변수 / 보안 민감도 변수 / 통신사 서비스 만족도 변수 / 고객 가입 서비스의 가성비 변수)를 생성해 데이터에 포함했습니다. 해당 변수를 포함시키는 경우의 수에 따라 32가지 경우의 수를 파생시킬 수 있었지만 현실적인 제약과 더불어 후에 변수선택과정을 진행하기 때문에 모두 데이터에 포함하는 방안으로 확정지을 수 있었습니다. 이에 생성된 데이터의 경우의 수는 32가지입니다. (이상치-4 X 인코딩-2 X 분포변환-4)

(2) 전처리 방안의 문제점

저희가 진행한 전처리는 Data Leakage현상을 고려하지 못했다는 한계점이 존재합니다. 이는 이상치 처리와 분포 변환 과정에서 발생했다고 파악했습니다. 저희가 정의한 Data Leakage현상은 Independent한 Test Data의 정보가 학습 과정으로 유출되었음을 의미합니다. 이러한 현상을 방지하기 위해서는 train data와 test data로 구분 후 전처리 과정을 각각 진행했어야 했는데 분포 변환과 같은 경우 전체 데이터에 대해 적용한 이후에 train과 test데이터로 분할했기 때문에 문제가 발생했고 결국 해당 방법을 적용하지 못했습니다. 이상치와 같은 경우에도 해당 순서를 유지해야 하지만 이상치 데이터를 삭제 후 분할하게 되었습니다. Test Data에 이상치가 존재하는가에 대해서는 이견이 존재할 수 있는데 이를 토대로 저희가 삭제한 이상치 데이터가 수집 과정에서 생긴 오류라고 판단한 후 분석을 이어갈 수 있었습니다.

위와 같은 문제점 때문에 분포 변환에서 파생되는 경우의 수를 제외하고 총 8가지의 데이터를 생성한 후 모델링에 적용할 수 있었습니다.

3. 모델링 방안

(1) 모델링의 흐름

다음은 모델링입니다. 저희가 이번 분석에서 사용한 모델은 Boosting계열의 CatBoost, XGBoost, LightGBM과 Bagging계열의 RandomForest 그리고 LogisticRegression입니다. 8개의 데이터에 대해 5가지 모델과 4가지 모델링 방법론을 적용해서 총 160개의 모형으로 결과를 비교할 수 있었습니다. 모든 모델링은 데이터를 4분할 한 후 진행했습니다. 데이터를 12:4:4:5의 비율로 분할하고 이를 각각 A, B, C, D로 명명하고 D데이터는 test data로 고정했습니다. 해당 비율을 고수한 이유는 train과 test의 비율을 4:1로 유지하기 위함입니다. A, B데이터는 하이퍼파라미터 튜닝을 하는 데에 이용되었고, C데이터는 모형의 랭킹을 산출 하는 데에 사용되었습니다. D데이터는 최종 모형의 성능을 파악하는 데에 사용되었습니다.

(모델링 flow사진 첨부)

모델링은 전진선택 및 하이퍼파라미터 튜닝을 중심으로 진행됩니다. 전진선택에서 변수와 하이퍼파라미터는 A데이터를 통해 학습하여 B데이터로 검증한 후 도출한 [Accuracy]를 기준으로 선택됩니다. 하이퍼파라미터 튜닝에는 Optuna 모듈을 이용했으며 Optuna는 특정 범위 내에서 최적의 하이퍼파라미터 조합을 반환하는 역할을 합니다. 전진선택과 하이퍼파라미터 튜닝 과정을 마치게 되면 “7번 데이터 – CatBoost – 3번 모델링” 의 최적의 변수 조합과 파라미터 조합이 생성됩니다. 해당 조합에 대해 A + B 데이터로 학습을 시킨 후 C 데이터를 통해 검증하면 accuracy\_i가 생기게 되며 이 accuracy\_i를 기반으로 모형 간의 순위를 도출할 수 있습니다. 그리고 해당 과정은 데이터 shuffle을 포함하는 100번의 반복을 포함합니다. 최종 랭킹을 산출하는 데에 쓰이는 지표는 다음과 같습니다.

(2) 주안점

이번 모델링 과정에서 주안점을 두었던 부분은 Double Modeling과 Pre-Trained기반 전진선택입니다. 모델링은 4가지 경우의 수가 존재하는데, 이는 Naïve Modeling vs Double Modeling과 Pre-Trained 전진선택 vs Naïve 전진선택에서 파생됩니다. 본 분석에서는 이를 각각 NN(Naïve Feature Selection – Naïve Modeling), ND(Naïve Feature Selection – Double Modeling), FN(Pre-Trained Feature Selection – Naïve Modeling), FD(Pre-Trained Feature Selection – Double Modeling)으로 명명했으며 모델 명(Lo – LogisticRegression, R – RandomForest, L – LightGBM, X – XGBoost, C – CatBoost)와 함께 LoNN, LoND, LoFN, LOFD으로 기입했습니다.

Double Modeling은 tenure변수를 기준으로 데이터의 부분집합을 생성한 후 진행됩니다. 모든 과정을 두 부분집합이 독립적으로 수행한 후 정확도의 가중평균을 모형의 최종 정확도로 사용하게 됩니다. 이는 tenure변수의 특정 값을 기준으로 타겟 변수의 분포가 상이하게 존재했기 때문에 해당 특성을 이용해보고자 고안되었습니다. Tenure가 alpha값 이하인 데이터프레임의 데이터 개수를 i, 정확도를 accuracy\_a로 명명하고 alpha값 초과인 데이터프레임의 데이터 개수를 j, 정확도를 accuracy\_b라고 명명했을 때 Double Modeling의 최종 accuracy는 다음 수식을 따르게 됩니다.

Pre-Trained Feature Selection은 전진선택 과정에 진입하기 직전에 사용됩니다. 개의 변수에 대해 전진선택법을 적용하는 컴퓨팅 문제를 해결하기 위해 고안된 방법입니다. 변수의 개수가 인 원본 데이터프레임에 인덱스와 같이 학습에 도움이 되지 않는 임의의 변수 를 추가하여 변수의 개수를 로 증가시킵니다. 변수는 단순히 인덱스를 나타내고 있기 때문에 올바른 모델링 과정을 거쳤다면 학습에 사용되어서는 안되는 변수입니다. 하지만 변수간의 조합을 토대로 학습에 미친 영향을 파악하는 Permutation Importance지표를 개의 변수에 대해 산출하면 변수보다 적은 영향을 미치는 변수들이 있음을 경험적으로 파악할 수 있었습니다. 이에 변수보다 적은 영향을 끼치는 변수 개를 삭제하고 개의 변수로 전진선택을 진행하는 과정이 Pre-Trained Feature Selection 방법입니다. 해당 과정의 결과를 신뢰할 수 있도록 100번의 반복 과정을 거친 후, Permutation Importance의 평균을 토대로 변수를 선택하였고 모든 과정은 임의의 하이퍼파라미터 100개의 조합을 적용했습니다.

4. 최종 결과

- 랭킹

- 최종 순위

- 모델 별 랭킹

- 데이터 전처리 방안 별 랭킹

- 모델링 방안 별 랭킹

- 성능

- 성능 평가 기준

- 사용자화 함수 제작

- 한계점

- 분석 목적에서의 한계

- 다양한 평가지표의 부재

- Data Leakage 컨트롤

- Etc

4. 최종 결과

(1) 랭킹

(최종 순위 차트 추가)

(모델 별 순위 차트 추가)

(데이터 전처리 방안 차트 추가)

(모델링 방안 차트 추가)

160개의 모형 순위는 [그림]과 같습니다. 상위 10개의 모형 중 NN모델링 방안이 8개 FN 모델링 방안이 2개 포함되었음을 통해 Pre-Trained모형의 효과가 생각보다 미비함을 알 수 있었지만 FN모형이 1등을 차지하여 이 점은 고무적이라고 생각했습니다. 또한 [그림]을 보면, 상위 10개 모형 중 7번 데이터가 4개, 6번 데이터가 2개 포함되었는데 6번과 7번 데이터는 각각 이상치의 두 기준을 모두 적용한 데이터로서 이상치 삭제의 효과를 체감할 수 있었습니다. 마지막으로 [그림]을 보면 극 상위권은 부스팅 계열 모델들이 차지하고 있으나, LogisticRegression이 10위권 내에 다수 존재하고 있음을 알 수 있었습니다.

(2) 성능

160개의 모형 중 1등을 차지한 CFN\_7의 최종 성능을 도출하고, 이를 활용해 최적의 파라미터와 정확도를 파악한 후 간단한 함수를 제작했습니다.

(함수 그림 추가)

(3) 한계점

다음은 분석의 한계점입니다. 우선 분석 목적에서의 한계점을 꼽을 수 있습니다. 이탈 여부보다 중요한 것이 이탈 시점이라고 판단했으나, 시계열 데이터의 부재로 인해 해당 결론을 도출하지 못했습니다. 이는 곧, 특정 데이터가 데이터 수집 기간에는 이탈하지 않았으나 수집 기간이 종료된 이후 바로 이탈하는 상황과 같이 현실적인 문제를 고려하지 못한다는 점에서 한계점을 보일 수 있습니다. 두번째는 다양한 평가지표의 부재입니다. 이번 분석에서는 모형을 선택하고 평가하는 기준으로 오직 정확도만 사용했는데 Class가 Imbalanced했던 만큼 다양한 지표를 통한 다양한 해석이 필요했다고 생각했습니다. 세번째는 Data Leakage control의 실패입니다. 데이터를 분할한 후 해당 과정을 거치며 더욱 다양한 데이터의 경우의 수를 파생하는 것이 분석의 정확도를 높이는 데에 도움이 되었으리라 판단됩니다. 이밖에도 불균형 데이터에 대한 oversampling 방법론을 적용하지 않은 것과 데이터의 부분집합을 생성하는 기준점을 임의로 선택한 지점에서 한계가 드러났다고 평가했습니다.