<머신러닝 프로젝트 보고서>

2021101992 경영학과 배아람

[목차]

(Colab:https://drive.google.com/file/d/1w9GRoVKzVedmzyixLo0Jm2VbFYnFdB6E/view?usp=sharing)

1. 문제 정의

우리나라의 통신 산업은 세계적으로 높은 수준의 기술을 자랑하지만 여러 가지 문제점이 존재한다. 한국의 통신 요금은 타국에 비해 상대적으로 높고 다양한 요금제가 존재하여 소비자들이 본인에게 적합한 요금제를 찾기 어렵다는 단점이 존재한다. 또한 한국 통신 시장은 SKT, KT, LG U+의 3사에 의해 독점된 시장이며 가입자 중심의 서비스 산업이기 때문에 점유율이나 경쟁 구도의 변화가 상대적으로 적은 독점시장이여서 통신사 입장에서는 신규 고객 유치 비용이 매우 크다는 문제가 존재한다. 더불어 최근 위 3사의 통신망 중 사용하지 않는 통신망을 빌려서 사용하는 알뜰폰 사업으로 인해 소비자 요금제를 대폭 줄이는 서비스가 출시되어 잠재적 고객 뿐 아니라 기존에 해당 통신사를 사용하는 고객들도 유출되고 있는 상황이기에 기존 고객 유치가 매우 중요한 사항으로 부각되고 있다.

미국도 이와 다르지 않은 상황이다. 미국 통신 산업은 이미 굉장한 포화상태에 있으며, 우리나라와 비슷하게 AT&T, Verizon, T-Mobile등 대부분의 시장이 점유하고 있다. 더불어 최근 데이터 품질, 부가서비스 등 전반적인 서비스가 상향 평균화되어 제공 서비스 수준의 편차가 줄어들며 기존 고객 이탈율이 증가하고 있는 상황이다. 과거에 비해 고객들의 통신사 충성도가 낮아졌기 때문에 효율적인 마케팅 방안을 통해 고객이탈을 방지하는 것이 중요해지고 있는 상황이다. 때문에, 본 보고서에서 Telco Customer Churn 데이터를 통

해 고객 이탈에 대해 예측하고 해석을 통해 효율적인 마케팅 방안 및 고객 관리 방안을 제시해보고자 한다.

2. 데이터 소개 및 전처리

A. 데이터 소개

본 보고서에서 사용될 데이터는 통신사 고객을 유지하기 위해 행동을 예측하고 관련된 모든 고객 데이터를 분석하여 집중적인 고객 유지 프로그램을 개발하기 위한 데이터 셋이다. 해당 Raw 데이터는 7043개의 Customer과 21개의 Feature들을 갖는다. 우리는 고객 행동들을 통해 Target(Churn, 고객의 이탈 유무) 데이터를 예측하고자 한다.

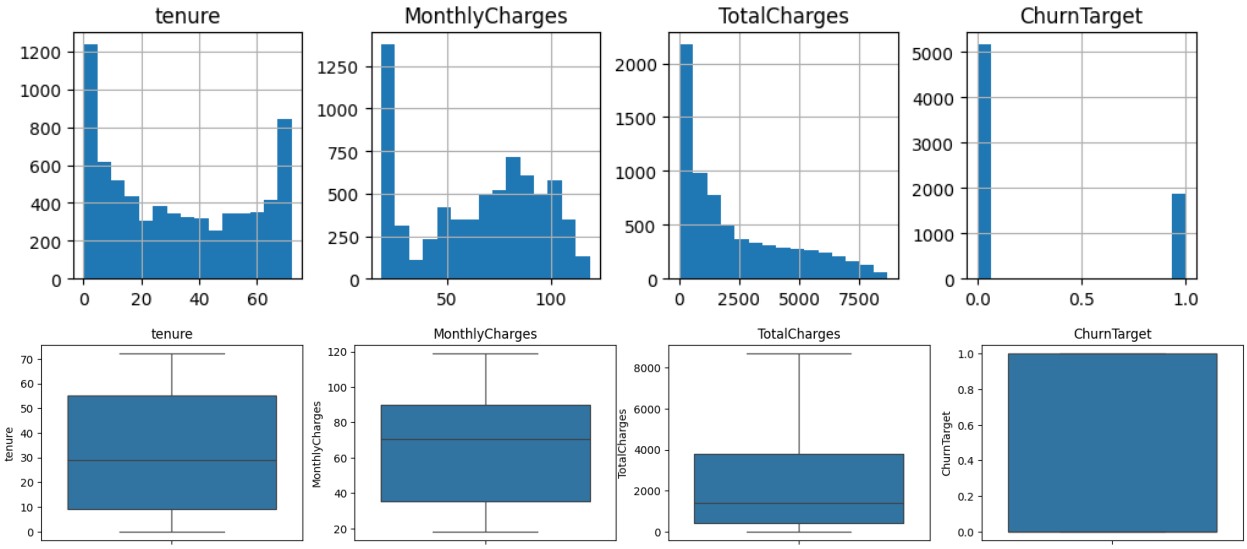
[Feature 설명 테이블]

customerID	고객 ID	**PhoneService**	폰 서비스 여부
gender	성별	**MultipleLines**	다중 회선 여부
SeniorCitizen	고령자 여부	**InternetService**	인터넷 서비스 여부
Partner	배우자 여부	**OnlineSecurity**	보안서비스 여부
Dependents	부양가족 여부	**OnlineBackup**	백업 여부
tenure	이용 기간	**DeviceProtection**	디바이스 보호 여부
TechSupport	기술 지원 여부	**MonthlyCharges**	월 청구 비용
Streaming	TV 스트리밍 여부	**TotalCharges**	전체 청구 비용
StreamingMovies	영화 스트리밍 여부	**Churn**	이탈 여부
Contract	계약 기간		
PaperlessBilling	전자 청구서 여부		
PaymentMethod	결제 방법		

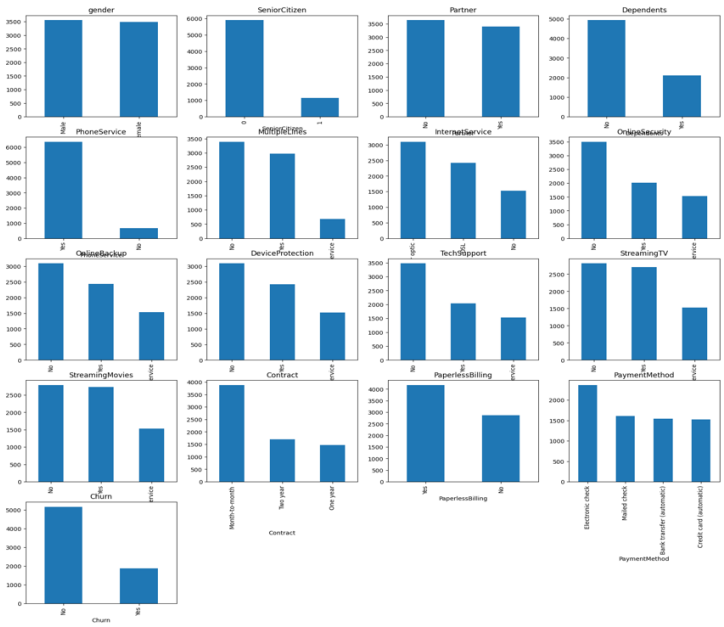
B. 데이터 전처리 및 시각화

해당 초기 데이터는 결측치가 존재하지 않고 SeniorCitizen, tenure, MonthlyCharges 데이터를 제외한 모든 데이터가 Object Type임을 확인할 수 있었다. 그중 TotalCharges 데이터가 수치형 데이터임에도 Object 객체로 설정되어 있음을 확인하여 이를 수치형 데이터로 변환해주고 변환이 불가능한 값은 0으로 보간하도록 했다. 더불어 SeniorCitizen 데이터는 범주형 데이터임에도 불구하고 수치형 데이터로 분류가 되어 있어 이를 Object 객체로 변환해주는 작업을 진행했다. 이후 상관관계 측정을 위해 Object 객체로 되어 있는 Churn 데이터를 수치형 데이터로 변환하여 ChurnTarget 데이터로 추가하는 작업을 진행하였다

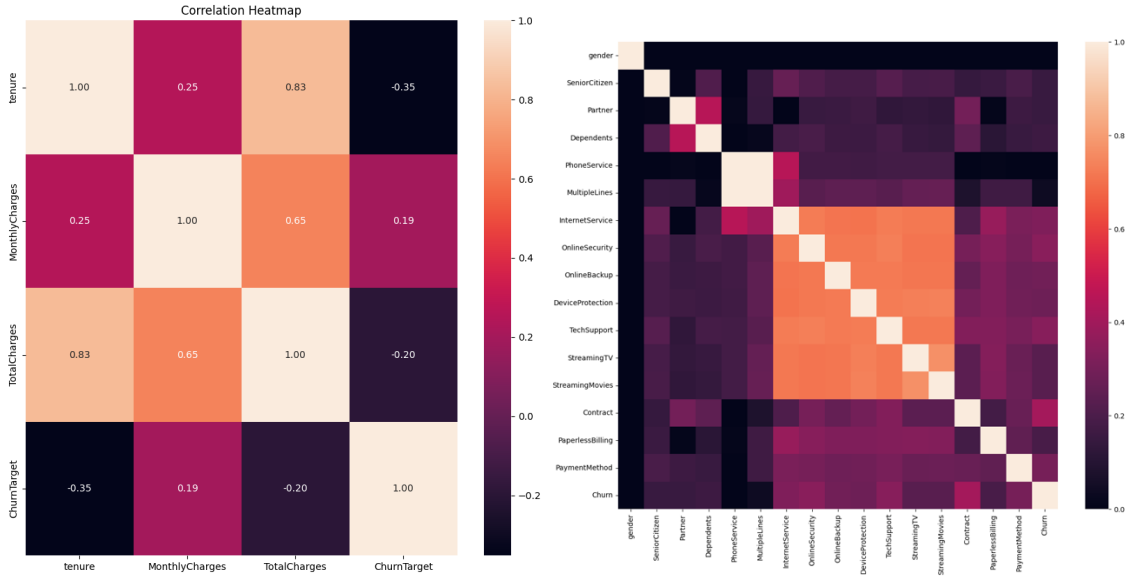
명목형 데이터와 수치형 데이터를 처리하는 과정을 진행한 이후, 수치형 데이터를 히스토그램과 박스플롯으로 시각화하는 작업을 진행했다. 아래 그래프를 보면 수치형 데이터는 분포가 고르게 분포되어 있음을 확인할 수 있다. 특히 tenure의 경우 오랜 기간 동안 서비스를 이용한 고객들이 많이 존재함을 확인하여 고객 충성도가 높은 고객이 대체적으로 많이 존재하는 것을 확인할 수 있다.



다음으로 범주형 데이터의 개수를 카운트하여 분포로 시각화하는 작업을 진행했다. 수치형 데이터와 마찬가지로 대부분의 데이터가 고르게 분포해있음을 확인할 수 있다. 주의깊게 볼 자료로 Contract 분포가 있었는데, 대부분의 고객들이 Month-to-month 계약이 되어 있어 60%가 넘는 고객들이 단기간에 변심으로 인해 언제든지 유출 고객으로 전환될 수 있음을 확인할 수 있다. 때문에, 해당 고객들을 대상으로 장기 계약을 유도하는 마케팅 방안이 필요함을 유추할 수 있다.



다음으로 수치형 데이터와 범주형 데이터의 변수간 상관관계를 분석하는 작업을 진행하였다. 수치형 데이터에서는 tenure 컬럼이, 범주형 데이터에서는 Contract, InternetService, OnlineSecurity 컬럼이 상대적으로 Churn target 데이터와 상관관계가 높음을 확인할 수 있다.



C. Feature Selection & One-Hot Encoding

최종적으로 예측에 사용할 변수를 각 수치형 변수, 범주형 변수의 Churn(target) 데이터와의 상관관계를 기준으로 결정했다. 상관계수의 Threshold를 0.3으로 설정하여 Churn과 0.3 이상의 상관관계를 갖는 변수들을 실제 예측에 사용할 Feature로 설정했다. 분석 결과, tenure, InternetService, OnlineSecurity, TechSupport, Contract, PaymentMethod 변수를 예측 모델에 사용할 변수로 선택하였다.

예측 모델에 사용할 변수중, tenure을 제외한 다른 변수들은 범주형 변수임을 확인하였다. 분석에 사용될 변수의 개수가 21개에서 6개로 줄어들었고 각 범주형 변수의 범주도 대부분 2~3개 수준이었기 때문에, 차원이 개수를 감안하여 범주형 데이터를 표현하기 위해 One-Hot Encoding 작업을 진행해주었다. 트리 모델과 KNN모델, Neural Net을 사용할 계획이기에 다중공선성을 고려하지 않고 모든 변수를 인코딩해주는 작업을 진행하였다.

3. 모델링

분석에 사용하는 모델로 Decision Tree, KNN, Random Forest, SVM, XGBoost, LightGBM, Neural Network(mlp)를 사용하여 예측 분석을 진행하였다. 분석의 첫번째 목표가

고객들의 이탈 유무를 예측하는 것이었기 때문에 대체적으로 성능이 잘 나오는 Tree기반 앙상블 모델을 사용하여 분석을 진행하였다. 각 모델들을 모든 파라미터를 고려하여 가장 성능이 좋은 파라미터를 채택하는 Gridsearch와 Cross Validation 중 가장 대표적인 K-Fold 방법을 이용하여 K는 5로 설정한 후 교차 검증을 통해 학습을 진행하였다.

1) Decision Tree(의사결정나무)

의사결정 나무는 지도 학습 모델 중 하나로, Tree 구조를 활용하여 엔트로피(Entropy)가 최소화되는 방향으로 데이터를 분류하거나 원하는 결과 값을 예측하는 방법을 말한다. 본 보고서에서는 max_depth, min_samples_split, min_samples_leaf 하이퍼 파라미터를 그리드 서치를 이용하여 설정해주었으며 각각 10, 10, 4의 최적의 파라미터로 학습을 진행하였다.

2) KNN(K-Nearest Neighbor)

KNN은 지도 학습 모델 중 하나로 데이터들을 K개의 가장 가까운 데이터와 비교하여 가장 가까운 속성에 따라 예측하고 분류하는 거리 기반 분류분석 알고리즘이다. 본 보고서에서는 n_neighbors, weights, metric 하이퍼 파라미터를 그리드 서치를 이용하여 설정해주었으며 각각 11, uniform, manhattan의 최적의 파라미터로 학습을 진행하였다.

3) Random Foreset(랜덤 포레스트)

Random Forest는 지도 학습 모델 중 하나로, 여러 개의 Decision Tree를 앙상블한 모델이다. 앙상블 모델은 단일 Decision Tree 모델보다 더 강력한 분류 모델을 구축하는 방법으로 Overfitting(과적합)을 줄이고 예측 성능을 향상시키는 데 효과적이다. 본 보고서에서는 n_estimators, max_depth, min_samples_split 하이퍼 파라미터를 그리드 서치를 이용하여 설정해주었으며 각각 100, 10, 10의 최적의 하이퍼 파라미터로 학습을 진행하였다.

4) SVM(Support Vector Machine)

SVM은 지도 학습 모델 중 하나로, 최대 마진 분리 초평면을 찾아서 두 부류 데이터 사이에 존재하는 여백을 최대화하여 데이터 포인트를 분류하며 커널 트릭을 통해 비선형 데이터를 효과적으로 처리할 수 있는 알고리즘이다. 본 연구에는 C(Regularization Parameter), kernel, degree 하이퍼 파라미터를 그리드 서치를 이용하여 설정해주었으며 각각 0.1, linear, 3의 최적의 하이퍼 파라미터로 학습을 진행하였다.

5) XGBoost

XGBoost는 성능과 효율성을 극대화한 강력한 경사 하강 부스팅 알고리즘이다. 과적합

방지를 위해 L1, L2 정규화 기법을 포함하고 있으며 빠른 학습 속도와 병렬 처리를 통해 대규모 데이터에서도 효율적으로 작동한다. 더불어 결측치 처리, 조기 종료 교차 검증 등 다양한 고급 기능을 제공하는 강력한 알고리즘이다. 본 보고서에서는 n_estimators, learning_rate, max_depth, min_child_weight 하이퍼 파라미터를 그리드 서치를 이용하여 설정해주었으며 각각 100, 0.1, 3, 1의 최적의 하이퍼 파라미터로 학습을 진행하였다.

6) LightGBM(Light Gradient Boosting Machine)

LightGBM도 XGBoost와 마찬가지로 빠르고 효율적인 경사 하강 부스팅 알고리즘이다. 트리 기반 학습 알고리즘을 사용하여 빠른 학습 속도와 낮은 메모리 사용량을 자랑한다. 카테고리형 변수 처리를 개선하고 다양한 파라미터 튜닝을 통해 모델의 유연성과 성능을 극대화할 수 있는 알고리즘이다. 본 보고서에서는 n_estimators, learning_rate, max_depth, min_child_weight 하이퍼 파라미터를 그리드 서치를 이용하여 설정해주었으며 각각 100, 0.1, 3, 1의 최적의 파라미터로 학습을 진행하였다.

7) Neural Network(인공 신경망)

Neural Network는 인간의 두뇌 구조와 기능을 모방한 기계 학습 모델이다. 입력 데이터의 특징을 학습하여 복잡한 비선형 관계를 모델링할 수 있으며, 다층 구조화 활성화 함수를 통해 다양한 데이터의 패턴을 학습이 가능하다. 본 보고서에서는 alpha값만을 그리드 서치를 통해 0.1값을 도출하였으며, 활성화 함수는 relu, 최적화 알고리즘은 adam을 사용하여 학습을 진행하였다.

각 모델의 성능을 비교하여 아래와 같은 결과가 산출되었다. 모델마다의 F1 Score에 큰 유의미한 차이가 존재하지는 않지만, 대체로 XGBoost, LightGBM의 성능이 상대적으로 더 우수한 성능을 보이는 것을 확인할 수 있다.

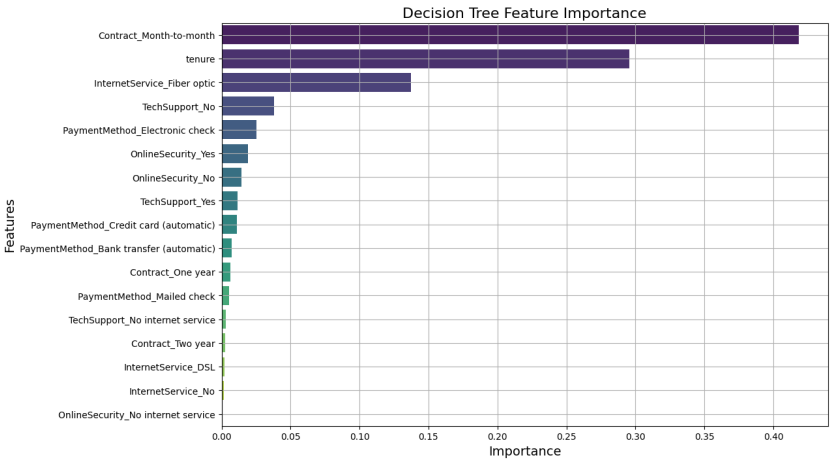
	F1 Score	Precision	Recall
Decision Tree	0.56	0.6	0.52
KNN	0.55	0.6	0.51
Random Forest	0.56	0.62	0.51
SVM	0.56	0.64	0.5
XGBoost	**0.61**	**0.67**	**0.56**
LightGBM	**0.60**	**0.66**	**0.56**
Neural Network	0.57	0.65	0.53

4. 모델 해석

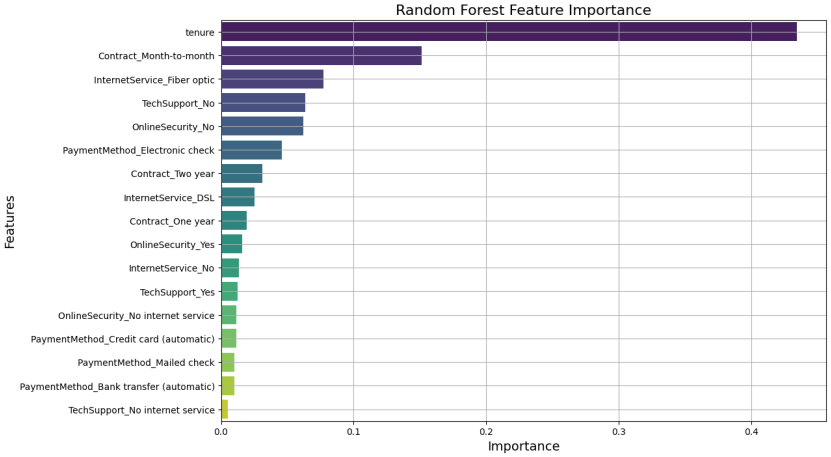
각 모델의 해석을 위해 상대적으로 성능이 뛰어난 트리모델(Decision Tree, Random

Forest, XGBoost, LightGBM)의 Feature Importance를 도출했다.

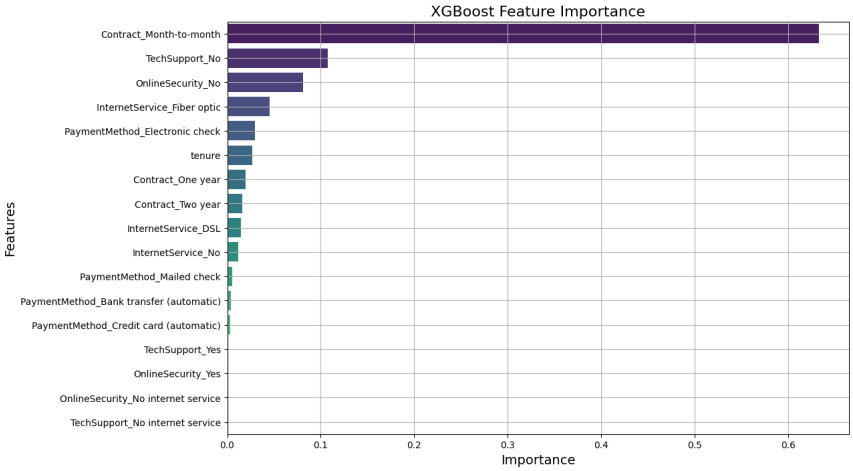
가장 먼저 Decision Tree의 경우 Contract_Month-to-month와 tenure Feature가 모델 해석에 대해 상대적으로 높은 중요도를 가지고 있음을 확인할 수 있다.



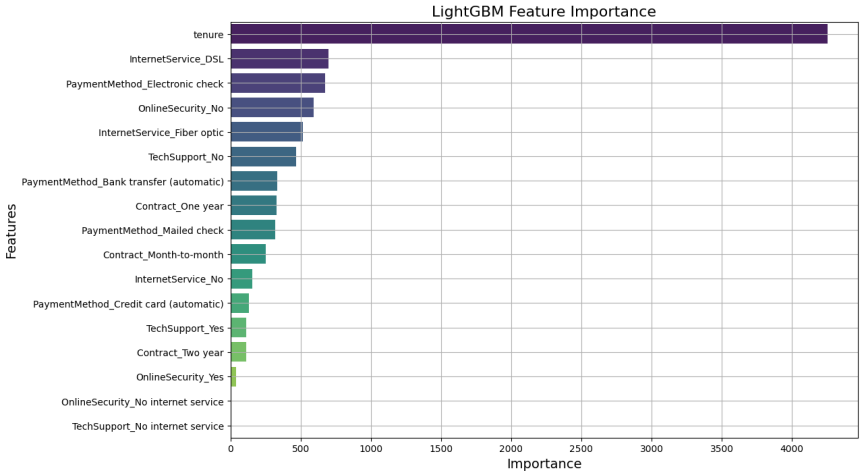
그 다음 Random Forest에서도 마찬가지로 tenure과 Contract_Month-to-month Feature가 상대적으로 분석에 큰 영향을 미쳤음을 확인할 수 있다.



그 다음 XGBoost에서는 압도적으로 Contract_Month-to-month 변수가 분석에 매우 중요한 변수로 사용이 되었음을 확인할 수 있다.



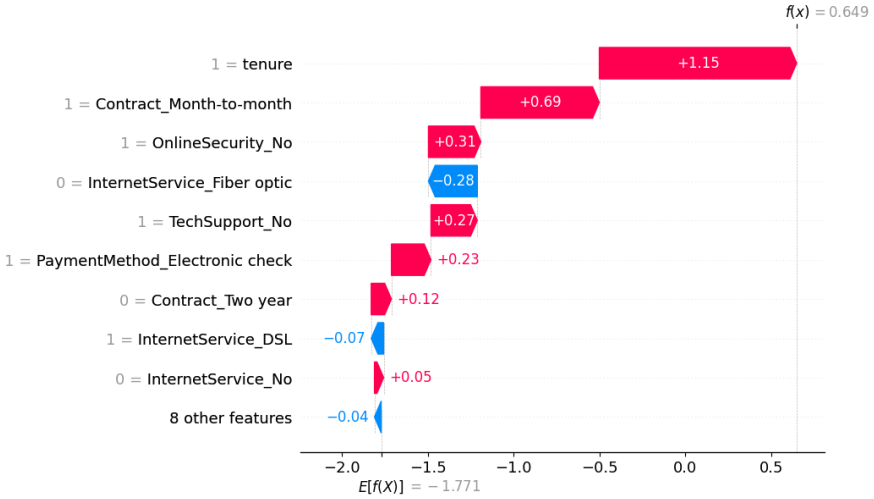
마지막으로 LightGBM에서는 압도적으로 tenure 변수가 분석에 매우 중요한 Feature로 사용이 되었음을 확인할 수 있다.



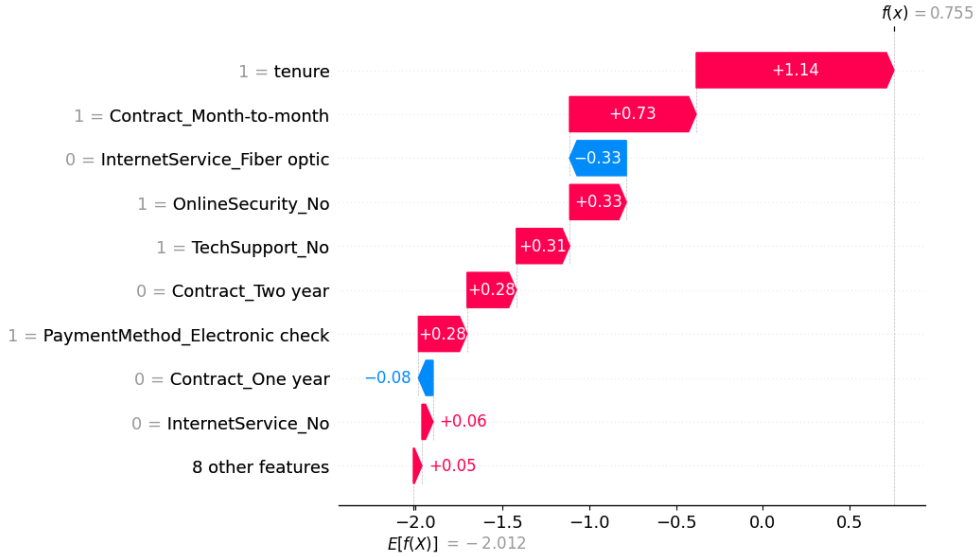
위 4개의 모델 Feature Importance를 종합해보면 통신사 서비스 이용 기간을 의미하는 tenure Feature와 통신사 서비스 계약기간 중 1달 계약을 한 고객들을 의미하는 Contract_Month-to-month Feature가 전반적으로 고객의 이탈 여부에 중요한 영향을 미쳤음을 확인할 수 있다.

이번에는 가장 성능이 좋은 XGBoost와 LightGBM에 한하여, SHAP를 사용하여 각 Feature의 기여도를 계산하여 해당 모델의 예측을 설명하고자 한다. SHAP은 기계 학습 모델의 예측을 설명하는 수학적 방법으로 해당 모델에 가장 중요한 변수를 찾고, 각 변수들이 모델 예측에 미치는 영향을 확인할 수 있는 기법이다. 앞선 Feature Importance들이 단순히 예측 분석에 영향을 미치는 변수를 찾는 과정이었다면, SHAP은 해당 예측에 대해 각 변수가 얼마만큼 어느 방향으로 영향을 미쳤는지를 파악할 수 있는 방법이다.

아래는 XGBoost를 기준으로 SHAP Explainer를 생성한 그래프이다. 특성 그래프의 숫자는 예측값에 기여하는 양을 보여준다. 빨간색은 특성 값이 예측값을 증가시키는 경우고 파란색은 특성값을 감소시키는 경우를 의미한다. 때문에, 여기서는 tenure이 고객의 이탈에 영향을 미쳤으며 그 예측값을 1.15만큼 증가시켰다고 해석할 수 있으며, Contract_Month-to-month도 고객의 이탈에 영향을 미쳤으며 그 예측값을 0.69만큼 증가시켰다고 해석할 수 있다. 추가로 유의미한 수치로 InternetService_Fiber Optic이 고객 이탈 방지에 영향을 미쳤으며 그 예측값을 0.28정도 감소시킴을 확인할 수 있는데, 이는 InternetService까지 구독이 사람들의 통신사 서비스 이탈을 방지하는 결합 상품의 효과를 가져다 주는 것을 유추해볼 수 있다.



lightGBM의 SHAP explainer도 마찬가지로 tenure이 1.14, Contract_Month-to-Month가 0.73만큼 고객의 이탈 예측값을 증가시켰음을 확인할 수 있으며, InternetService_Fiber optic이 고객 이탈 방지 예측값을 0.28정도 감소시킴을 확인할 수 있다.



5. 결론

분석 결과, XGBoost와 LightGBM 모델이 다른 모델에 비해 상대적으로 높은 성능을 보였으며, 이를 통해 Telco Customer Churn 데이터에서 두 가지 주요 인사이트를 도출하고자 한다. 더불어 이에 따른 마케팅 방안도 함께 제안한다.

첫 번째 인사이트는 tenure와 Contract Feature가 고객 이탈에 중요한 영향을 미친다는 점이다. tenure는 통신사 서비스 이용 기간을, Contract_Month-to-month는 한 달 단위의 통신사 서비스 계약을 의미한다. 이탈할 것으로 예측되는 단기 고객들에게는 지속적인

프로모션을 제공하여 이탈을 방지할 필요가 있을 것이라고 생각이 된다. 또한, 계약 만료가 다가올 때 사전 공지를 통해 추가 혜택을 제공하여 계약 연장을 유도할 필요가 있다. 단기 계약 고객의 이탈률이 높은 점을 고려할 때, 장기 계약을 유도하는 전략도 중요하다. 장기 계약 고객에게는 추가적인 혜택을 제공하여 더 많은 가치를 느끼도록 하는 마케팅 방안을 제안할 수 있다.

두 번째 인사이트는 InternetService가 고객 이탈 방지에 긍정적인 영향을 미친다는 점이다. 이는 통신사의 결합상품과 연관 지어 볼 수 있다. 고객들은 일반적으로 동일 통신사의 Phone, TV, Wifi, 인터넷 등 다양한 서비스를 함께 이용하는 경향이 있다. 따라서, 결합상품의 다양화와 혜택 강화를 통해 고객이 서비스를 가입할 때 결합상품을 선택하도록 유도하는 것이 필요하다. 이는 고객의 이탈 가능성을 줄이고, 서비스 만족도를 높이는 데 기여할 것이라고 생각한다.

이러한 분석을 통해 고객 이탈 가능성이 높은 고객을 사전에 파악하고, 이들을 집중적으로 관리하는 프로그램을 구축할 수 있고, 마케팅 비용을 절감함에 더하여 장기적으로 충성도 높은 고객을 확보하여 통신사 매출을 증대에 기여할 수 있을 것이라고 생각한다. 앞으로도 지속적인 데이터 분석과 고객 관리를 통해 통신사 서비스의 경쟁력을 강화하고, 고객 만족도를 높이는 Data Driven한 분석 방식에 주력한다면 좋은 성과를 얻을 수 있을 것이다.

Reference

- https://www.kaggle.com/datasets/blastchar/telco-customer-churn
- https://www.hani.co.kr/arti/economy/it/1121317.html
- https://zzinnam.tistory.com/entry/SHAP-value%EC%97%90-%EB%8C%80%ED%95%9C-%EA%B0%84%EB%8B%A8%ED%95%9C-%EC%86%8C%EA%B0%9Cwith-Python
- http://www.goodkyung.com/news/articleView.html?idxno=234857
- https://brunch.co.kr/@b047a588c11b462/61