**학습된 인공지능 모델**

1. **모델을 평가할 주요 지표 선정**

* Accuracy vs F1-score

: 모델을 평가할 주요 지표로 우리는 위 두 가지를 고민하였음. 우리 데이터셋을 분석해본 결과, target값의 비율이 0:1 = 48.66 : 51.34 로 균형을 이루고 있었음. 따라서 균형적인 데이터셋에 더 적합한 Accuracy를 평가지표로 사용하기로 결정하였음.

현재 우리가 확보한 데이터로 진행된 Kaggle 대회에 참여한 상위권 팀들의 Accuracy가 대략 0.82 이었기 때문에 우리는 목표 Accuracy를 0.82로 선정하였음.

**2. 모델 비교**

* **Logistic Regression**

|  |  |
| --- | --- |
| train\_accuracy | test\_accuracy |
| 0.8150 | 0.8133 |

* **Random Forest**

|  |  |
| --- | --- |
| train\_accuracy | test\_accuracy |
| 0.9665 | 0.8455 |

* **XGB**

|  |  |
| --- | --- |
| train\_accuracy | test\_accuracy |
| 0.8301 | 0.8300 |

* **CatBoost**

|  |  |
| --- | --- |
| train\_accuracy | test\_accuracy |
| 0.8588 | 0.8565 |

* **신경망**

|  |  |
| --- | --- |
| train\_accuracy | test\_accuracy |
| 0.8301 | 0.8300 |

3. **모델 비교 정리**

* Logistic Regression의 Accuracy는 train\_set의 경우 0.8150, test\_set의 경우 0.8133으로 우리의 목표치인 0.82를 넘기지 못했음.
* Random Forest 의 Accuracy는 train\_set의 경우 0.9665, test\_set의 경우 0.8455으로 과적합이 발생하였음.
* XGB의 Accuracy는 train\_set의 경우 0.8301, test\_set의 경우 0.8300으로 우리의 목표치인 0.82를 넘겼으며 train\_set과 test\_set의 결과가 균형적인 것으로 보아 좋은 성능을 보임.
* CatBoost의 Accuracy는 train\_set의 경우 0.8588, test\_set의 경우 0.8565으로 우리의 목표치인 0.82를 넘겼으며, 우리가 학습시킨 모델 중 가장 높은 Accuracy를 보임.
* 신경망의 Accuracy는 train\_set의 경우 0.8301, test\_set의 경우 0.8300으로 우리의 목표치인 0.82를 넘겼으며 train\_set과 test\_set의 결과가 균형적인 것으로 보아 좋은 성능을 보임.

4. **결론**

이번 프로젝트에서는 고객 이탈(Churn)을 예측하는 분류 모델을 구축하였음. 먼저 로지스틱 회귀를 사용하여 예측 모델을 만들어 보았고, 이후 랜덤 포레스트를 활용하여 feature importance를 분석했음. 그 결과, 주요 특성들 중 다수가 범주형 특성임을 확인함.

범주형 특성이 많은 데이터셋에 적합한 모델을 찾는 과정에서 CatBoost 모델을 알게 되었고, 이를 데이터셋에 적용한 결과 우수한 성능을 보여 최종 모델로 선정하였음.

이 프로젝트를 통해 CatBoost 모델의 강점을 확인하였고, 앞으로 다양한 데이터셋에 CatBoost를 어떻게 적용하면 좋을지 더 깊이 이해하기 위해 공부할 계획임. CatBoost 모델을 통해 향후 다양한 데이터셋에서도 좋은 성능을 기대할 수 있을 것으로 봄.