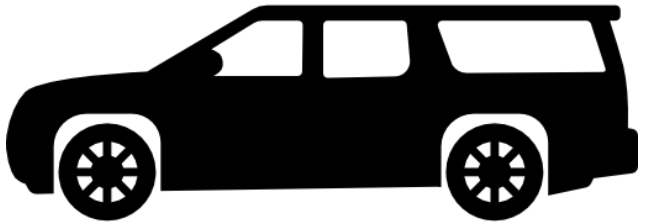
현대자동차 데이터분석 해카톤 - Task 2

(예측모델 결과서)



960723 김동욱

목차

과제 2

1. 문제 정의

★ 고객 별 6개월 내, "대차" 또는 "추가구매" 가능성(Score)에 대한 예측모델을 개발

→ "대차" 가능성을 선택해 모델링.

ㅁ 접근 방법

TASK 1에서 정의한 고객별 구매 데이터를 바탕으로 구매 유형을 "대차", "추가구매", "구매안함"의 세가지 중 하나로 라벨링 한 뒤, 이를 바탕으로 예측 모델을 학습.

ㅁ 구매 유형을 라벨링 하는 방법

주어진 데이터에서 6개월을 단위로 시점을 후행 (backward) 이동하며 고객별 파생변수(독립변수)를 수집하고, TASK 1에서 정의한 재구매유형을 참고해 종속변수를 생성.

<예시>

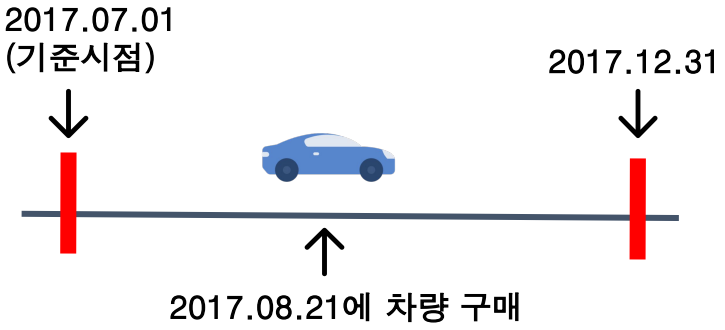
2017.07.01을 기준으로 할 때,

1. 해당 시점에서 파생변수를 수집 (수집할 파생변수의 종류는 추후 설명.)
2. 해당 시점을 기준으로 향후 6개월 이내 (2017.07.01 ~ 2017.12.31) 사이에 출고된 차량이 있는 고객은 6개월 이내에 구매를 한 고객으로 간주.
3. (2)에 해당하는 고객들을 제외한 나머지 고객들은 6개월 이내에 구매를 하지 않은 고객으로 간주.

과제 2

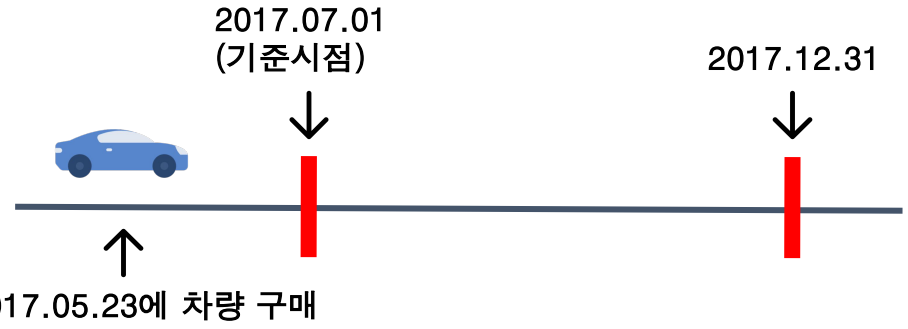
2. 샘플링 정의

예시 1) 6개월 이내에 구매내역이 있는 경우



→ 해당 차량의 재구매유형으로 라벨링
("대차" / "추가구매" 中 하나)

예시 2) 6개월 이내에 구매내역이 없는 경우



→ "구매안함"으로 라벨링

해당 과정을 시점을 6개월씩 후행 이동하며 반복해 총 5년간의 파생변수들을 수집한다.

ex) 2017.07.01 → 2017.01.01 → 2016.07.01 → … → 2012.01.01

- 이 경우, 동일한 고객이더라도 시점에 따라 수집되는 파생변수가 달라지기 때문에 개별 고객의 ID는 중요하지 않음.
- 5년이라는 기간을 설정한 이유는, 수집될 파생변수 중 "거주 주택 가격" 등의 고객 정보는 시점에 영향을 받는 변수 (time-dependent variable) 이기 때문에 너무 오래된 데이터는 모델링에서 제외하기 위함.
- 모델링은 "대차"와 "추가구매" 중 "대차"를 선택해 진행함.

과제 2

3. 독립/종속 변수 정의

★ **pandas의 groupby 메소드를 이용해 데이터로부터 다음과 같은 파생변수를 수집.**

<독립변수 - 수치형>

변수명	변수설명
PCE_PER_PYG	거주 주택의 평당 가격
승용_소형, 승용_준중형, 승용_중형, 승용_대형, 승용_고급, RV_준중형이하, RV_중형이상, 해치백, 스포츠카, 전기차	기준 시점 보유 차량의 등급별 개수
LPG, 가솔린, 디젤, 전기/하이브리드	기준 시점 보유 차량의 엔진 타입별 개수
하, 중, 중상, 상	기준 시점 보유 차량의 트림 등급별 개수
CAR_HLDG_DURATION_CONTID	기준 시점 보유 차량들의 평균 보유 기간
NUM_CONTACTS	기준 시점으로부터 6개월 전까지 전체 접촉 횟수
대면, 비대면	기준 시점으로부터 6개월 전까지 접촉 채널별 횟수
정비, 상담, 견적, 서비스	기준 시점으로부터 6개월 전까지 접촉 목적별 횟수

변수명	변수설명
CAR_HLDG_DURATION_TOTAL	전체 구매 차량들의 평균 보유 기간
CAR_HLDG_DURATION_MIN	전체 구매 차량들의 최소 보유 기간
CAR_HLDG_DURATION_MAX	전체 구매 차량들의 최대 보유 기간
CAR_HLDG_DURATION_FINISHED	보유 종료된 차량들의 평균 보유 기간
NUM_TOTAL_CARS	전체 차량 구매 횟수
NUM_CONTID_CARS	기준 시점 보유 차량 수
NUM_FINISHED_CARS	보유 종료 차량 수
CUS_AGE	기준 시점에서의 고객의 나이

과제 2

3. 독립/종속 변수 정의

★ **pandas의 groupby 메소드를 이용해 데이터로부터 다음과 같은 파생변수를 수집.**

<독립변수 - 범주형>

변수명	변수 설명
SEX_SCN_NM	고객의 성별
최초구매_포함	기준 시점 보유 차량 중 최초구매 차량 포함 여부

<종속변수>

변수명	변수 설명
TYPE_PURCHASES	구매 유형 (대차 / 구매안함)

→ 총 37개의 독립변수를 갖는 약 25만개 정도의 데이터로 모델링을 진행.

- “대차”와 “구매안함”의 라벨수는 각각 **12만개, 13만개** 정도로 비슷함.
- 사이킷런의 **train/test split** 함수를 사용해 약 **20%**의 데이터를 테스트셋으로 분리.

과제 2

4. 모델링

★ **pycaret 모듈**을 이용해 모델의 성능을 비교 **&** 하이퍼파라미터 튜닝을 진행 **(Auto ML)**

★ 트리 기반의 분류기를 사용해 앙상블 모델을 정의

- 트리 기반의 모델은 수치형 변수의 **scaling**이 불필요하고, **outlier**에 **robust**하며, 변수중요도를 시각화할 수 있는 장점과 더불어 **proba** 메소드를 이용해 구체적인 확률값을 산출할 수 있기 때문에 해당 과제의 모델링에 적합하다고 판단.
- 일반화 성능을 향상시키기 위해 앙상블 분류기를 사용.

<pycaret을 이용한 분류기 성능 비교>

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.8446	0.9214	0.8140	0.8752	0.8435	0.6896	0.6914	8.930
lightgbm	Light Gradient Boosting Machine	0.8442	0.9210	0.8106	0.8772	0.8426	0.6888	0.6910	1.434
xgboost	Extreme Gradient Boosting	0.8434	0.9201	0.8135	0.8734	0.8424	0.6872	0.6889	1.560
rf	Random Forest Classifier	0.8401	0.9153	0.8005	0.8779	0.8374	0.6807	0.6836	15.264
et	Extra Trees Classifier	0.8311	0.9086	0.7892	0.8704	0.8278	0.6629	0.6660	14.736

→ **XGBoost, LightGBM, RandomForest** 세가지 분류기를 하이퍼 파라미터 튜닝해 앙상블 분류기로 연결.

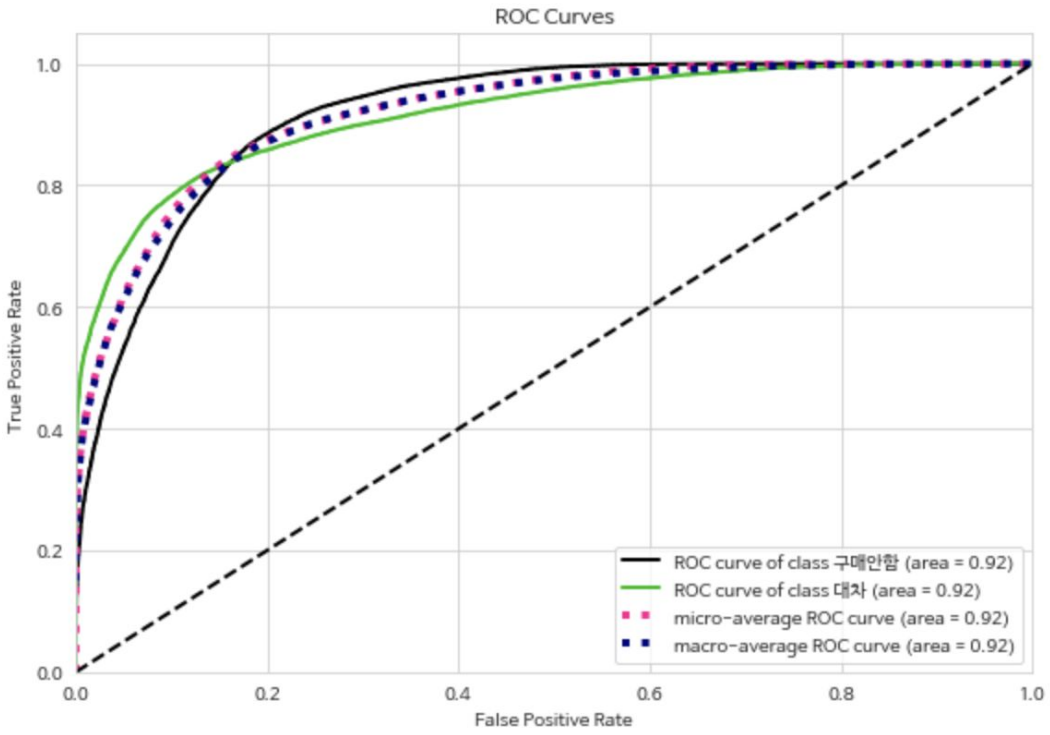
과제 2

5. 평가지표 선정 및 모델 평가

★ 평가지표로 정밀도(precision)과 재현율(recall)의 조화 평균인 **F1 Score**를 사용.

- **5-Fold Cross Validation**을 바탕으로 측정

★ **ROC** 곡선을 이용해 시각화.



	precision	recall	f1-score	support
구매안함	0.813656	0.874995	0.843211	24991.000000
accuracy	0.842412	0.842412	0.842412	0.842412
macro avg	0.843661	0.843405	0.842408	51603.000000
weighted avg	0.844603	0.842412	0.842383	51603.000000
대차	0.873665	0.811814	0.841605	26612.000000

최종적으로 약 0.84 정도의 F1 Score를 달성.

- "대차" 클래스의 정밀도가 약 **87%**로 "구매안함" 클래스의 정밀도 (약 **81%**) 보다 높다.
  → 마케팅적 관점에서 차량을 구매하지 않을 고객들보다는 차량을 구매할 고객들을 잘 예측하는 것이 중요하므로 좋은 결과로 판단됨.

과제 2

6. 모델 활용 방안

★ 모델링에 사용된 독립변수만 정의할 수 있으면, 새로운 고객에 대해서도 해당 고객의 향후 **6개월** 이내에 대차 가능성을 구체적인 확률값으로 예측 가능.

→ 모델링에 사용하지 않은 약 **20%** 정도의 테스트 데이터셋을 바탕으로 **6개월** 이내 대차 확률을 예측.

<훈련된 모델을 이용한 6개월 이내 "대차 가능성" 예측 예시>

	구매안함_확률	대차_확률	실제결과
0	0.647	0.353	구매안함
1	0.617	0.383	구매안함
2	0.229	0.771	대차
3	0.698	0.302	구매안함
4	0.768	0.232	구매안함
...	...	...	...
51598	0.664	0.336	구매안함
51599	0.214	0.786	대차
51600	0.190	0.810	대차
51601	0.646	0.354	구매안함
51602	0.609	0.391	구매안함

<독립변수 전처리 파이프라인 설명>

- "평당 주택 가격"은 시점의 영향을 받으므로, 새로운 고객에 대한 데이터가 주어졌을 시 **2017년을 기준으로 보정**할 필요성이 있음.
- 범주형 자료인 "고객의 성별"만 원핫 인코딩을 통해 전처리. 나머지 변수는 별도의 전처리 불필요.

→ 전처리 파이프라인의 복잡도가 낮음!

과제 2

7. 한계점 및 발전 방향

★ 사용한 독립변수 중 차량의 보유기간 및 보유 차량수와 관련한 변수가 상대적으로 높은 중요도를 보임.
→ 반면, 고객의 개인 정보와 관련한 변수는 분류기의 성능에 핵심적이지 않음.

★ 본 결과서는 “대차 가능성” 하나만을 모델링 했으나, “추가구매”, “대차”, “구매안함” 세가지의 클래스를 분류할 수 있는 **Multi-class classifier**를 모델링 할 수도 있음.
→ 단, **trial & error** 의 결과 현재 가용한 독립변수 만으로는 만족할 만한 성능이 나오지 않았음.

★ 고객의 경제력과 개인적 특성에 대한 좀 더 핵심적인 독립변수를 추가하면 분류기의 성능 향상 및 앞서 언급한 **Multi-class classification** 문제도 해결 가능할 것이라고 예상.
(ex. 기혼여부, 가족구성원 수, 연봉 등)