

Home Credit – Credit Risk Model Stability

팀원: 최찬우, 정원우

목차

1. 프로젝트 개요
2. 데이터 분석 및 전처리
3. 모델 선정 및 구현
4. 대회 결과

1. 프로젝트 개요

문제 소개:

신용 기록이 거의 없는 고객은 대출이 거부되는 경우가 많다.

프로젝트 목표:

대출 상환 불이행 가능성을 예측하여

신용 기록이 부족한 고객에게 대출 기회를 제공하는 모델을 개발

2. 데이터 분석 및 전처리 (데이터 분석)

case_id: 각 신용 케이스에 대한 고유 식별자이다. 관련 테이블을 기본 테이블과 결합할 때 이 ID가 필요하다.

date_decision: 대출 승인 여부에 대한 결정이 내려진 날짜를 나타낸다.

WEEK_NUM: 집계를 위해 사용되는 주 번호이다. 테스트 샘플에서는 WEEK_NUM이 훈련 데이터의 마지막 WEEK_NUM 값부터 연속적으로 이어진다.

MONTH: 집계 목적으로 사용되는 월을 나타내는 열이다.

target: 특정 신용 케이스(대출)에 대해 일정 기간 후 고객이 연체했는지 여부에 따라 결정되는 목표 값이다.

num_group1: depth=1 및 depth=2 테이블에서 case_id의 과거 기록을 위한 인덱싱 열이다.

num_group2: depth=2 테이블의 case_id 과거 기록을 위한 두 번째 인덱싱 열이다. num_group1과 num_group2의 순서는 중요하며, 특징 정의에서 명확히 설명된다.

2. 데이터 분석 및 전처리 (데이터 분석)

데이터 셋은 유사한 그룹에 대해 다음과 같은 표기법을 사용

P: DPD(Days past due) 변환

M: 카테고리 마스킹

A: 금액 변환

D: 날짜 변환

T: 미정의 변환

L: 미정의 변환

2. 데이터 분석 및 전처리 (데이터 전처리)

[Pipeline 클래스]

set_table_dtypes(df): 데이터프레임(df)의 각 열을 순회하면서 특정 조건에 따라 데이터 유형을 변환한다.

handle_dates(df): 데이터프레임의 날짜 열을 처리한다.

filter_cols(df): 특정 조건에 따라 열을 필터링하여 제거한다.

2. 데이터 분석 및 전처리 (데이터 전처리)

[Pipeline 클래스]

set_table_dtypes(df): 데이터프레임(df)의 각 열을 순회하면서 특정 조건에 따라 데이터 유형을 변환한다.

case_id, WEEK_NUM, num_group1, num_group2 열의 경우: Int64로 변환한다.

date_decision 열의 경우: Date 타입으로 변환한다.

열 이름의 마지막 문자가 "P" 또는 "A"인 경우: Float64로 변환한다.

열 이름의 마지막 문자가 "M"인 경우: String으로 변환한다.

열 이름의 마지막 문자가 "D"인 경우: Date 타입으로 변환한다.

2. 데이터 분석 및 전처리 (데이터 전처리)

[Pipeline 클래스]

handle_dates(df): 데이터프레임의 날짜 열을 처리한다.

각 열을 순회하며, 열 이름의 마지막 문자가 "D"인 경우 다음 작업을 수행한다:

현재 열의 날짜 값을 "date_decision" 열의 날짜 값과 뺀다.

두 날짜 사이의 총 일수를 계산한다.

처리 후, "date_decision"과 "MONTH" 열을 데이터프레임에서 제거한다.

2. 데이터 분석 및 전처리 (데이터 전처리)

[Pipeline 클래스]

filter_cols(df): 특정 조건에 따라 열을 필터링하여 제거한다.

각 열을 순회하며, "target", "case_id", "WEEK_NUM" 열을 제외한 나머지 열에 대해 결측치 비율을 계산한다.
결측치 비율이 70%를 초과하는 열을 삭제한다.

다시 각 열을 순회하며, "target", "case_id", "WEEK_NUM" 열을 제외한 나머지 열 중 문자열 타입의 열에 대해 고유 값의 수를 계산한다.

고유 값의 수가 1이거나 200을 초과하는 열을 삭제한다.

2. 데이터 분석 및 전처리 (데이터 전처리)

[Aggregator 클래스]

num_expr(df): 데이터프레임에서 수치형 특징을 추출한다. (P,A)

date_expr(df): 데이터프레임에서 날짜 관련 특징을 추출한다.(D)

str_expr(df): 데이터프레임에서 문자열 특징을 추출한다.(M)

other_expr(df): 데이터프레임에서 기타 특징을 추출한다(T,L)

count_expr(df): 데이터프레임에서 카운트 관련 특징을 추출한다.(num_group)

get_exprs(df): 이전 함수들로부터 받은 모든 표현식을 집계하여 종합적인 특징 추출 표현식 리스트를 만든다.

2. 데이터 분석 및 전처리 (데이터 전처리)

[기타 처리 함수]

feature_eng(df_base, depth_0, depth_1, depth_2): 기본 데이터프레임(df_base)과 여러 추가 데이터프레임(depth_0, depth_1, depth_2) 세트를 대상으로 피처 엔지니어링을 수행한다.

to_pandas(df_data, cat_cols=None) : 데이터프레임을 Pandas 데이터프레임으로 변환하고, 선택적으로 지정된 열을 범주형 데이터 타입으로 변환한다.

reduce_mem_usage(df) : 데이터프레임의 모든 열을 순회하며 데이터 유형을 수정하여 메모리 사용량을 줄이는 기능을 수행한다.

3. 모델 선정 및 구현

LightGBM (Light Gradient Boosting Machine):

- 마이크로소프트에서 개발.
- 특징: 리프 중심 트리 성장, 빠른 학습 속도, 낮은 메모리 사용량, 결측값 및 범주형 변수 자동 처리.

CatBoost (Categorical Boosting):

- 안덱스에서 개발.
- 특징: 범주형 변수 효율적 처리, 과적합 방지를 위한 순열 기반 부스팅, 높은 예측 성능, 빠른 학습 속도.

Voting Model:

- 여러 모델을 결합하여 예측 수행.
- LightGBM 모델 5개와 CatBoost 모델 5개를 사용하여 예측 평균을 도출.

3. 모델 선정 및 구현

[Sub.csv 생성]

LightGBM 및 CatBoost 모델 각각 5개씩 사용하여 test 데이터에 대해 확률 예측값을 저장

저장된 예측 값들의 평균 도출

Sub.csv에 임시 저장.

3. 모델 선정 및 구현

[Submission.csv 생성(최종)]

Train데이터의 target 값: 0 , Test데이터의 target 값: 1 설정

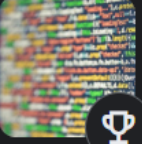

Train데이터에 Test데이터 결합

기본 파라미터의 LightGBM 모델을 위의 Train데이터로 학습

Test 데이터에 대해 모델이 예측한 확률이 threshold(0.996)보다 낮다면 확률 조정
=> 기존 확률 * threshold - 0.05

이를 통해 훈련 데이터와 테스트 데이터의 분포 차이 고려

4. 대회 결과





Home Credit - Credit Risk Model Stability

Create a model measured against feature stability over time

Featured · Code Competition · 3856 Teams · 17 days ago

565/3856



3856개 팀 중 565위 달성.