[2020 금융 빅데이터 페스티벌] 과제설명자료

(과제1) 주식거래내역으로 매수 상위종목 예측

미래에셋대우

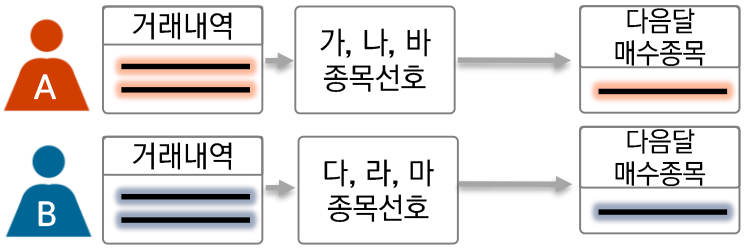
2020.08

MIRAE ASSET
미래에셋대우

목차

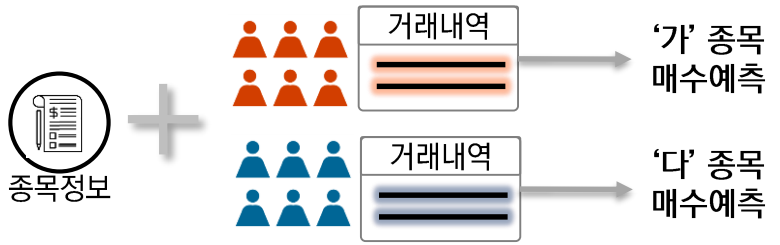
1. 과제 개요

과제 배경



개인의 **거래 종목 선호도**에 따라
다음 달 매수 종목이 다르지 않을까?

과제 접근 방향



선호 종목 유사성에 따라 사전에 분류된 그룹의
과거 거래내역과 종목 정보을 통해, 그룹 단위의
선호 종목을 예측하고자 함

※ 외부데이터 사용 가능

[과제] 과거 거래데이터를 이용해, 각 그룹에서 7월에 가장 많은 고객이 매수할 상위 3개 종목 예측

훈련 데이터 기간 : 2019년 7월 ~ 2020년 6월

예측기간 : 2020년 7월 1일~28일 (4주)

상위 3개 종목 : 2020년 7월에 그룹 내 매수 고객 수가 가장 높을 3개 종목

2. 분석 데이터 – (1) 거래내역 테이블(trade_train.csv)

: 2019년 7월부터 2020년 6월까지 거래한 고객들을 임의로 그룹화하여 거래내역을 통계 처리한(평균 및 중앙값) 테이블
각 그룹에서 월별로 거래한 종목들과 각 종목을 거래한 고객의 수, 거래된 가격(매수, 매도 단가)들의 중앙값을 포함함

데이터 상세

변수명	상세사항	샘플 예시
기준년월	YYYYMM	201907
그룹번호	임의로 고객군을 나눈 후 부여한 번호 (총 48개 그룹)	MAD01
그룹내고객수	그룹에 속하는 고객 수	288
종목번호	7자리 종목 식별 고유번호	A000660
그룹_매수여부	해당 기간에 그룹 내에서 종목에 대한 매수가 한번이라도 발생한 경우 (Y: 매수가 발생함/N: 매수가 발생하지 않음)	Y
그룹_매도여부	해당 기간에 그룹 내에서 종목에 대한 매도가 한번이라도 발생한 경우 (Y: 매도가 발생함/N: 매도가 발생하지 않음)	Y
매수고객수	해당 기간에 그룹 내에서 종목을 매수한 고객의 수	7
매도고객수	해당 기간에 그룹 내에서 종목을 매도한 고객의 수	17
평균매수수량	해당기간에 그룹내 각 고객의 일별 평균 매수 수량	19
평균매도수량	해당기간에 그룹내 각 고객의 일별 평균 매도 수량	234
매수가격_중앙값	해당기간에 그룹내 매수 단가들의 중앙값	74800
매도가격_중앙값	해당기간에 그룹내 매도 단가들의 중앙값	78500

2. 분석 데이터 – (2) 종목 테이블(stocks.csv)

: 2019년 7월부터 2020년 6월까지의 국내주식 종목들의 시가, 저가, 고가, 종가 테이블
2020년 7월 예측 대상 종목 후보가 되는 종목들이 별도로 태그 되어있으며, 각 종목의 산업정보 또한 제공됨

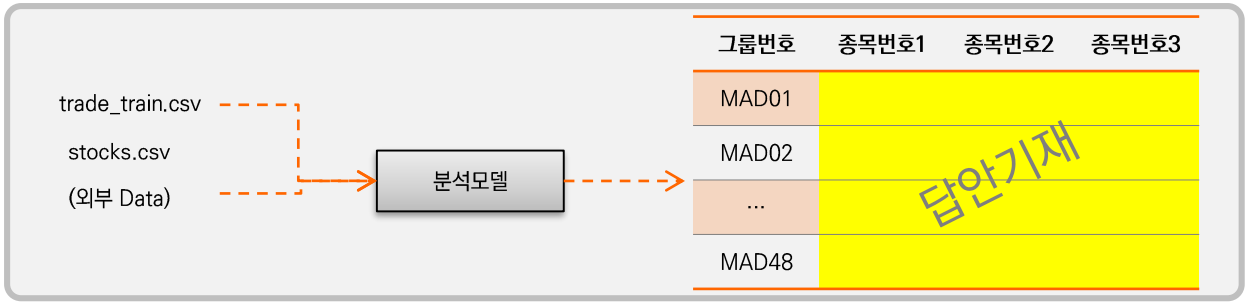
데이터 상세

변수명	상세사항	샘플 예시
기준일자	YYYYMMDD	20200701
종목번호	7자리 종목 식별 고유번호	A000020
종목명		동화약품
20년7월TOP3후보군	20년 7월에 가장 많은 고객이 매수할 세가지 종목 후보군 (총 135개 종목)에 속하는 종목 (Y: 속함, N: 속하지 않음)	N
시장구분	코스피/코스닥	코스피
산업구분_대분류		제조업
산업구분_중분류		의료용 물질 및 의약품 제조업
산업구분_소분류		의약품 제조업
종목시가	해당 일자에 최초로 체결된 거래 가격	9680
종목고가	해당 일자에 가장 높게 체결된 거래 가격	9840
종목저가	해당 일자에 가장 낮게 체결된 거래 가격	9680
종목종가	해당 일자에 마지막으로 체결된 거래 가격	9710
거래량	거래소에서 거래된 주식 수의 총합	31341
거래대금_만원단위	거래소에서 거래된 주식의 총대금	30581

3. 답안 제출 방법

파일의 구성

구분	학습용 데이터셋1	학습용 데이터셋2	제출용 파일(답안파일)
파일명	trade_train.csv	stocks.csv	answer_sheet.csv
데이터 건수	30,200	286,061	48
기간	2019년 7월 ~ 2020년 6월	2019년 7월 ~ 2020년 7월	2020년 7월 (1일 ~ 28일)



제출 방법 및 주의사항

- 예측한 타겟 값을 답안 파일에 적어 프로그래머스 내 “파일제출” 페이지에 제출
- 답안 파일에 모든 값이 빠짐없이 입력하되, ID값을 임의로 변경 금지
- 답안 파일에 변수명(ID, target)과 파일의 형식(csv) 임의로 변경 금지
- 답안 파일 내 Top3 매수 종목은 반드시 종목코드의 오름차순으로 제출 (평가 방법 참고)

4. 평가 방법

예선

- 평가지표 수치에 기반한 정량적 평가

본선이후

- [본선] 모델의 독창성, 논리적 타당성 등을 고려한 정성적 평가
- [결선] 모델의 전반적 평가, 업무 활용성, 발표 능력 등을 고려한 종합적 평가

1. stocks.csv의 〈20년7월Top3후보군〉변수에서 Y로 기입된 135개 종목에서 정답지 예측
2. 그룹별 매수 종목의 Top3 종목 간 **순서에 관계없이** 포함 여부로 채점
3. 종목번호1 ~ 종목번호3 **입력 순서는 종목번호 오름차순** 으로 제출

   ex) MAD99그룹의 매수 종목 순위를

   (종목번호1, 종목번호2, 종목번호3) = (A999999, A777777, A888888)로 예측하였더라도,

   (종목번호1, 종목번호2, 종목번호3) = (A777777, A888888, A999999)로 입력
4. 정답지에서 빈도가 낮은 종목은 **평가 시 가중 배점**
5. 예선 결과는 코드 내 오류 등을 검수하는 최소한의 정성적 평가가 이루어진 후 발표됩니다. (10월 16일)

5. FAQ

Q. '20년7월Top3후보군'은 어떤 변수인가요?

A. 훈련데이터를 이용하여 7월 상위 매수종목을 예측할 때, 예측할 종목들의 범위입니다. 실제 7월 상위 매수종목들을 포함한 총 135개 종목들이 후보 군으로 태그 되어 있습니다.

Q. 상위 3개종목의 기준은 무엇인가요?

A. 각 그룹 내에서 20년 7월에 가장 많은 고객이 매수한(매수고객수(unique)가 높은 순) 3개 종목입니다. 거래대금, 거래량 등의 기준이 아닙니다.

Q. 사용 가능한 외부데이터의 범위가 어떻게 되나요?

A. 출처가 명확하고 법적 제약이 없는 경우에 가능합니다. 크롤링의 경우에는 코드를 함께 제출해주세요. 본선 진출 시 제출할 보고서에 사용한 데이터의 출처 및 링크를 모두 명시해야 합니다.

Q. 거래데이터가 실제 값과 다른 것 같아요. 왜 이런 건가요?

A. 2020 금융 빅데이터 페스티벌에서 제공한 데이터는 개인정보법 개정 시행령에 따라 익명처리 적정성 평가를 받은 데이터입니다. 이 과정에서 실제 고객 정보를 추정할 수 없도록 비식별화 처리 하였기때문에, 실제 값과 차이가 있습니다. 다만, 분포는 유사하므로 이 점 고려하셔서 분석을 진행하시면 됩니다.

Q. 종목테이블의 결측치는 어떻게 이해하면 될까요?

A. 매매거래정지나 관리종목으로 지정된 경우 등 종목별 상이한 이유로 결측이 발생할 수 있습니다. 학생들의 다양한 결측 처리 기법을 기대해보겠습니다.

감사합니다

MIRAE ASSET
미래에셋대우