



DATA ANALYSIS

PORTFOLIO

Soon Gyu Kwon

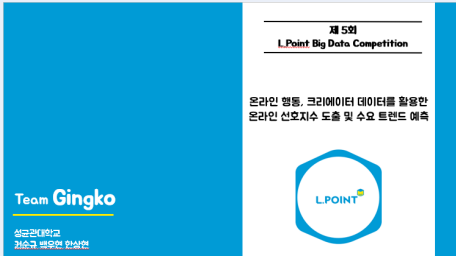
Contents



경제지표, 뉴스 기사, 인플루언서를 활용한
화장품 주가예측
Team Gingko

01 - 인플루언서를 활용한 화장품 주가예측

2018.06 ~ 2018.09



02 - 온라인 행동 데이터를 활용한 선호지수 개발

2018.11 ~ 2019.02



03 - 선수 유형 추천시스템

2018.12 ~ 2019.01



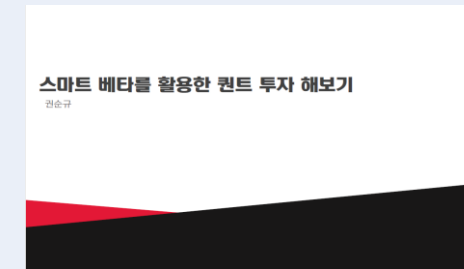
04 - 머신러닝을 활용한 동물 멸종 위기 등급 예측

2019.02 ~ 현재



05 - 비재무 데이터와 강화학습을 활용한 자동 매매시스템 구현

2019.02 ~ 2019.02



06 - 팩터 인베스팅

2019.09 ~ 현재



07 - 게임 이탈 유저 분석

2018.07 ~ 2018.09



08 - 맛집 추천시스템 개발

2018.09 ~ 2018.12

01

인플루언서를 활용한 화장품 주가 예측 2018.06 ~ 2018.09

(최종 등수 : 1등 / 176팀)

Summary

주가 변동 요인에 새로운 관점 제시 : 인플루언서

공신력 있는 뷰티유튜버의 활동 영역을 웹크롤링
뷰티유튜버의 영향력을 정량적으로 수치화

텍스트 마이닝을 통한 비정형 데이터 분석

화장품 주가에 영향을 많이 주는 국제 정세 관련
뉴스 기사를 수집하여 주가에 대한 영향력 분석



경제지표, 뉴스 기사, 인플루언서를 활용한
화장품 주가예측

Team Gingko



Data

경제 지표, 뉴스 텍스트, 인플루언서 (뷰티유튜버)
(출처 : Dart 전자공시, 웹크롤링)

Algorithm & 방법론

- 국내 화장품 생산 기업들의 주가 동향을 한 번에 나타내는 COSPI30지수 생성
- 뷰티유튜버 활동 정보 수치화
- LSTM 모델을 활용한 최종 예측
- Rolling-Window 기법을 통해 과적합 방지
- 화장품 대장주에 대한 투자 시뮬레이션 진행

Result

COSPI30지수 Test셋에 대한 결과

- RMSE: 17.15
- Accuracy: 63.7%

아모레퍼시픽, LG 생활건강 종목에 대한 2018년 7월 한 달 간의 투자 시뮬레이션 결과 각각 2.43% , 1.38% 의 수익률

* 미래에셋대우 제 2회 빅데이터 패스티벌 대상 수상

(주최 : 미래에셋 대우증권)

분석 의의

- 1) 온라인 인플루언서를 활용한 최초의 정량적 주가 예측 시도
- 2) 비정형 데이터를 수치화 하여 분석에 이용

▶ 뷰티유튜버 01 주제 소개



화장품을 리뷰하는 뷰티유튜버

우리는 '인플루언서'를 통한 뷰티 트렌드 형성이 화장품 구매에 관한 사람들의 심리 변화와 화장품 회사의 실적호전에 크게 영향을 줄 것으로 보았다.

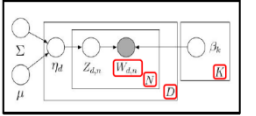
▶ 반응변수 | COSPI30 02 데이터 수집 및 전처리



한국거래소에서 산출하는 산업별 지수에 화장품 관련 지수 부재, "동일가중 방식" 사용해 직접 화장품종합주가지수(COSPI30)를 산출하였다.

▶ 입력변수 | 뉴스 텍스트 02 데이터 수집 및 전처리

1. 토크모deling: CMT Algorithm



TF-IDF 계산

단어별 문서에서 W라는 단어의 빈도수 * W라는 단어가 나온 문서의 개수-1

단어들이 가중치가 적용된 수치값을 부여해 중간값 이상의 단어들만 사용

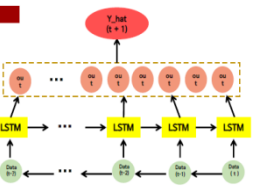
▶ 입력변수 | 인플루언서 지수 02 데이터 수집 및 전처리

1. 구독자와 조회수 기준으로 가장 영향력있는 뷰티유튜버 9명 선정
2. 9명의 일일 조회수를 더해 날짜별 "총 일일 조회수" 산출
3. 7일전부터 해당 일까지의 "일주일 누적 평균 조회수" 산출
4. 분산 안정화를 위해 로그를 취함

$\log("일주일 누적 평균 조회수")$

▶ 2. 모델링 03 COSPI30 모델링 및 모델 비교

LSTM 아란?



데이터의 장단기 종속성을 학습 할 수 있는 순환 신경망(RNN)의 변형 모델로, 기존 신경망과 달리 은닉층 내부에서 정보의 피드백이 재귀적으로 이루어짐

▶ 투자 시뮬레이션 04 개별 종목 모델링 및 투자 시뮬레이션

COSPI30 동향 고려



개별 종목 : 상승세

COSPI30 : 하락세

02 온라인 행동 데이터를 활용한 선호지수 개발

2018.11 ~ 2019.02
(최종 등수 : 3등 / 674팀)

Summary

온라인 수요 트렌드 예측

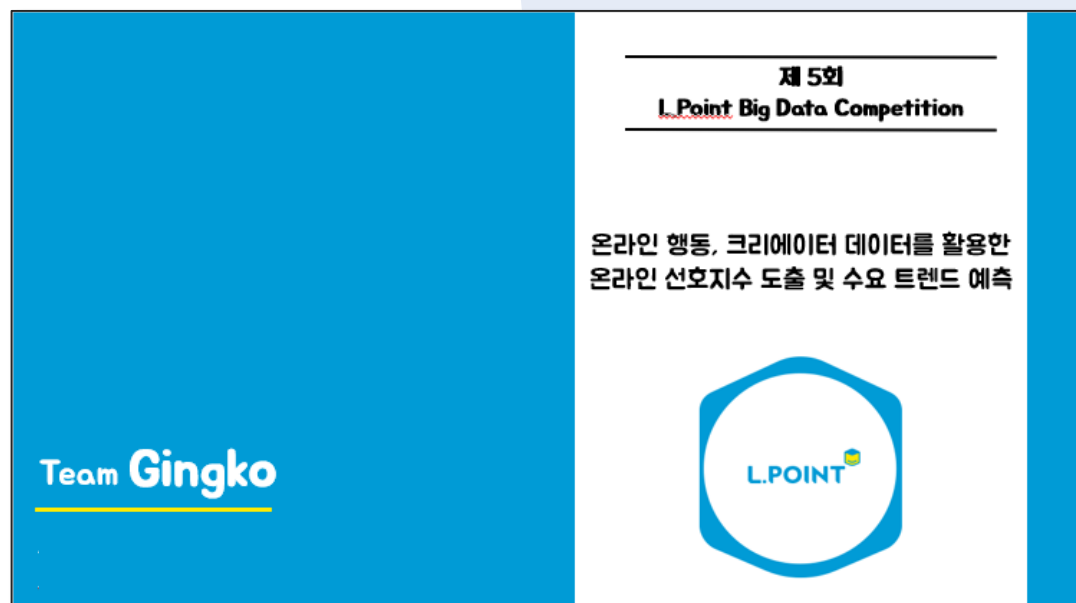
자체적으로 크롤링한 외부 데이터가
수요 트렌드에 미치는 영향력 분석

온라인 선호지수 개발

수요 트렌드에 영향을 미치는 변수를
1차원으로 축소하여 선호지수 개발

최종 서비스 제안

Anomaly Detection, 고객 유형 군집화, 유튜버와의 협력 등의
서비스 제안



Data

전국 60만 L.Point 이용 고객의 카드 기록 데이터

(출처 : 롯데멤버스 L.Point)

Algorithm & 방법론

- 롯데멤버스 기본 제공 데이터 이외에 쇼핑몰 리뷰 Text, 기상 변수 수집
- 비정형 데이터에 대해서 감성분석
- 수요트렌드 예측에 Dynamic Linear Model, ARIMAX, Prophet 모형 적합
- 주성분분석을 사용하여 온라인 선호지수 생성
- 생성된 선호지수와 판매량과의 비교를 통한 재고 관리, 고객 유형 군집화를 통한 유형별 맞춤 프로모션 서비스 제시

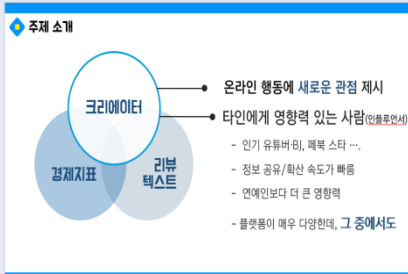
Result

RMSE(예측 오차) : 58.3141원

*제5회 L.Point Bigdata Competition 우수상 수상
(주최 : 롯데멤버스)

분석 의의

- 1) 외부 데이터를 직접 수집하여 온라인 수요 트렌드 분석
- 2) 분석 전개 논리성 확보 - 수집한 외부 변수의 유의성 입증 후, 이를 바탕으로 '온라인 선호지수', '서비스 제안'

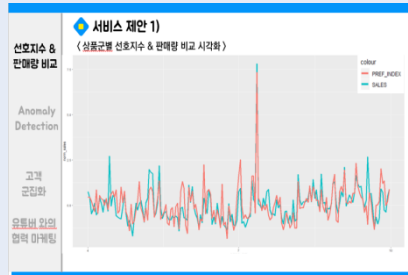
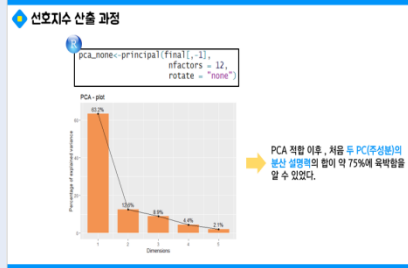
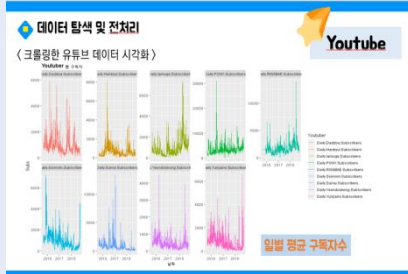


최종 모델 선정

최종 모델 비교 (변수 조합 별)

Input variables	Dynamic Linear Model	ARIMAX	Random Forest	XGBoost
날씨/요일	28.00103	14.1439	31.5424	29.4751
날씨/요일 + 온라인 검색/리뷰	31.2955	14.0183	35.9854	31.4847
날씨/요일 + 온라인 검색/리뷰 + 인플루언서	32.75278	13.9911	34.6424	28.2824

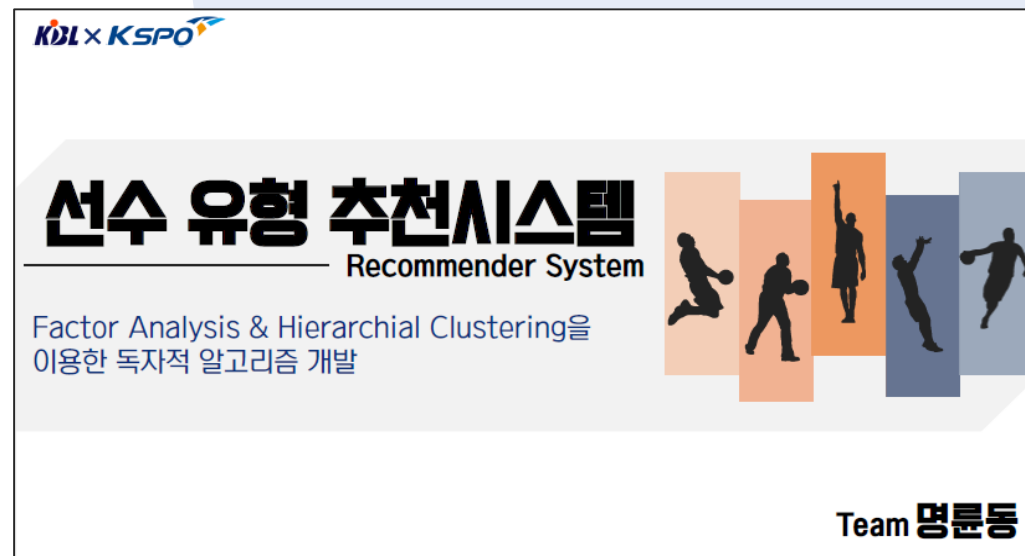
상품 소분류명: 물류관리 연구



Summary

KBL 관객 증대를 위해 관객 유형분류, 유형별 맞춤 전략 제시
농구를 잘 모르거나, KBL 관람 해본적이 없는 관중들에 대해
간단한 설문지를 통한 '선수 추천 알고리즘' 개발

KBL보다 미국 nba 농구를 즐기는 사람들을 위해 KBL선수와
nba선수의 유사도를 고려한 선수 추천 알고리즘 개발



“Factor Analysis, Hierarchial Clustering, 선수간 Euclidian distance 를 통한 선수 유형 추천시스템 고안”

Data

2015 – 2018 시즌 KBL 선수 기록 데이터
1980년~2018년 nba 선수 기록 데이터
(출처 : KBL , Kaggle)

Algorithm & 방법론

- 농구를 잘 모르는 사람도 본인의 성향에 맞는 선수를 추천받을 수 있는 방법을 고안 -> Factor Analysis를 통해 경기 기록 변수를 '새로운 이름의 변수'로 재정립 (ex : 턴오버, 스틸, 어시스트 -> 서포터형)
- 설문지를 고를 때마다 반복적 계층 군집분석을 수행, 5~6번 반복시 최종 선수 추천 완료
- KBL선수들과 nba선수들간의 경기 스탯을 normalizing하고, Euclidian Distance를 계산하여, 유사도를 바탕으로 선수 추천
(어떤 nba 선수를 좋아한다고 응답했을때, 그와 유사한 성향의 3명의 KBL선수 추천)

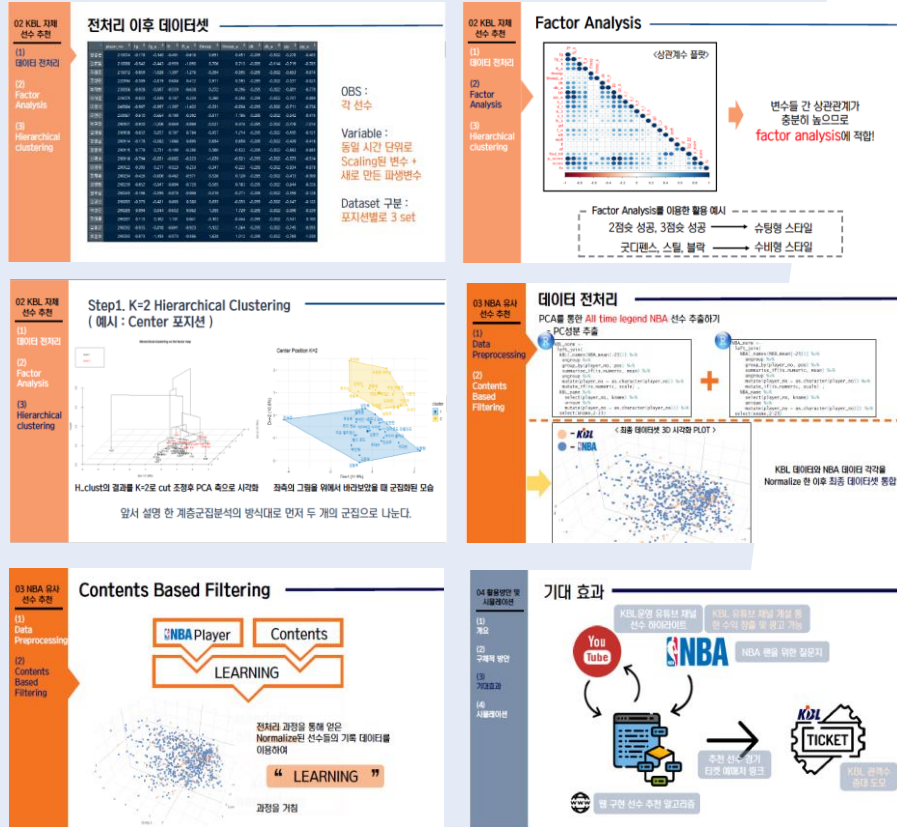
Result

Github Web Hosting을 이용한 구현
(https://ddeolddeorumi.github.io/who_do_you_like_basket/)

*제2회 프로농구 데이터 활용 경진대회 최우수상 수상
(주최 : KBL , 국민건강진흥공단)

분석 의의

- 1) Factor Analysis, 계층군집 분석 등의 통계적 분석 기법을 사용한 독자적 선수 추천 알고리즘 고안
- 2) 농구 선호층 및 비선호층 모두에게 상용화 가능한 선수 추천 방식 고안



04 머신러닝을 활용한 동물 멸종 위기 등급 예측

2019.01 ~ 현재

2019 한국과학창의재단 학부생 연구 프로그램 선정작

Summary

데이터 수집 및 전처리

IUCN redlist category에 등록된 7만여 개의 종에 대한 각종 데이터 수집
각 변수별 전처리 진행

동물 멸종 요인 분석 및 보호 정책 수립 제언

정보부족 분류 종 멸종 위기 등급 부여

IUCN에서 정보부족으로 멸종 위기 등급을 부여하지 못한 종들에
머신러닝을 활용해 위기 등급 부여



Data

각 종의 생물 분류(Phylum, Class, Order, Family), 생리학적 정보,
지리적 정보, 보전을 위한 노력, 상업적으로 활용된 정도 등

(출처 : IUCN webpage, Web scrapping)

1 주제 소개

학습을 위한 데이터

생물 분류에 관한 정보(Phylum, Class, Order, Family), 각 종의 생리학적 정보, 지리적 정보, 보전을 위한 노력 등

RedList 사이트에서 요청하여 받은 data + 웹 크롤링한 data

3 데이터 전처리

5) Entity Embedding

1차원 Embedding

2차원 Embedding

어린아, 블록버스터, 성인, 예술 영화

3 데이터 전처리

6) Word2Vec

rationale_v1	rationale_v2	rationale_v3
-1.75580364	1.105681166	1.307062313
-0.35833291	1.096288295	1.460741667
-0.506213207	1.946311559	1.572868203
-0.745429454	2.920664781	2.807926414
-1.568349966	2.407699666	0.93471211
-0.835355978	2.717569448	2.819040833
-1.579280961	2.54098318	1.853104935
-0.455085258	0.734643147	0.882727212

이후 각 단어별 tf-idf 스코어를 가중치로 사용하여 선행 결합

최종 3개의 변수 생성 !!

3 데이터 설명

변수명	설명	전처리 방향
criteriaVersion	해당 종 등급판정 책자 판	삭제
Language	해당 종 등급판정 책자 언어	삭제
habitat	종이 거주하는 지역 특성 설명	topic modeling: 10개의 topic
population	종의 인구적 특성	삭제
populationTrend	인구 증가 추세 [Decreasing/Increasing/Stable/Unknown]	유지
range	종이 거주하는 나라 및 도시의 설명	국가 이름에 대한 binary encoding: entity embedding을 통한 차원 축소 (177 -> 20)
useTrade	해당 종이 사용되는 방법	NotUsed: 사용방안 없음, Used: 상품화 가능, Collect: 수집용, byCatch: 추가적으로 채집됨
systems	종이 거주하는 생태계 [Freshwater/Marine/Terrestrial]	binary encoding
conservationActions	해당 종의 보호 방안	0: No action, 1: Yes action, NA: Unknown

3 데이터 전처리

4) Topic Modeling

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
Grass	Nest	Span	Year	River	Stream	Water	Mountain	Forest	Cave
Grassland	Bird	Style	Female	Water	habitat	Fish	Lowland	Plant	Stone
Habitat	Feed	Class	Male	Lake	Area	depth	Tree	Season	Inhabit
Soil	tree	Space	Size	Freshwater	Egg	coral	Forest	Fruit	hill

조림 동지 민물 바다 산 숲 동굴

4 모델링

GBM (Gradient Boosting Machine)

Model의 residual에 새로운 트리를 fit

이 과정에서 Gradient descent를 이용해 loss function (RMSE, classification error 등)을 최소화

$$y = f(x) + \epsilon_1 \rightarrow \text{Iteration 1}$$

$$= f(x) + g(x) + \epsilon_2 \rightarrow \text{Iteration 2}$$

$$= f(x) + g(x) + h(x) + \epsilon_3 \rightarrow \text{Iteration 3}$$

$\frac{\partial}{\partial \theta} l(y, \hat{y})$

$g(x)$ 는 ϵ_1 에 적합 시킨 모델
 $h(x)$ 는 ϵ_2 에 적합 시킨 모델

Algorithm & 방법론

- 데이터에 대해 다양한 전처리 방식 진행
 - Entity Embedding을 통한 차원 축소
 - Word2Vec 을 통해 numeric 벡터 임베딩
- Data Deficient 로 분류된 12000 여개의 종을 최종 Test 데이터로 분리.
- SMOTE 를 통해 불균형 데이터 문제를 해결.
- 나머지 Label로 모델 생성 및 10-fold CV 검증
- Classification에 뛰어난 성능을 보이는 다양한 모델 적합
- 계수 해석을 위한 Ordinal Logistic Regression, GAM 모델 적합

Result

Accuracy: 71.3%

[정보부족]에서 다른 등급으로 분류된 종에 대한 해석 진행 및 보호 정책 수립 제언

ex) 이 종의 ~한 특성이 있고, 멸종 위기 등급을 낮추기 위해 ~ 노력 필요

분석 의의

- 1) 수많은 동물들의 멸종위기단계를 일일이 추적해야 되는 기존 방식의 한계점 보완
- 2) 단순 개체수 파악이 어렵다는 이유로 [정보부족]으로 분류된 종들의 등급을 예측하여 효율적인 보호 정책 수립 가능

05 비재무 데이터와 강화학습을 활용한 자동 매매시스템 구현(인턴)

2019.02

Summary

1600개 상장 종목 재무데이터 수집

1600개 종목의 주가 및 재무제표(BPS,PER, PBR, 배당 등) 수집

Clustering with Prophet 적합 및 파생변수 생성

i) 모든 종목에 대해 시계열 예측 모형 적합 이후 MAPE값 산출,

ii) 증가, 재무제표 관련 다양한 파생변수 생성

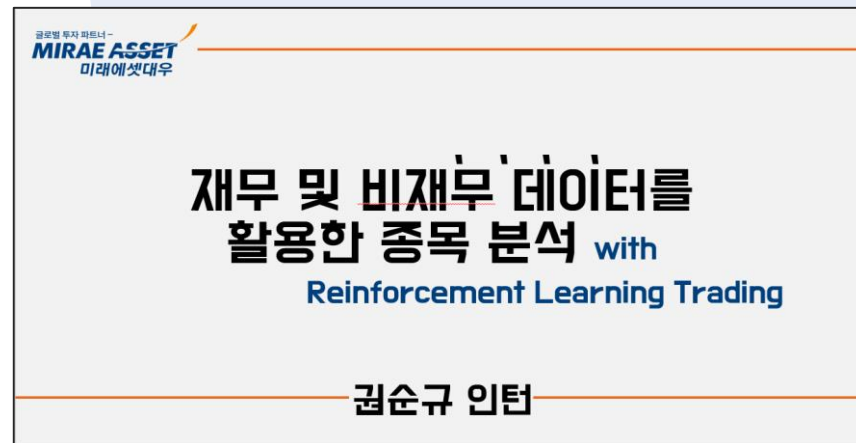
-> 둘을 결합하여 최종 클러스터링 진행

비재무적 요소에 취약한 군집의 종목들에 한하여 비재무 데이터 수집

제약회사 의약품에 대한 고객 후기, 이산화탄소 배출량 등

강화학습을 통한 자동 매매 시뮬레이션

최종 선정된 종목들에 한하여, 강화학습 알고리즘을 이용한 투자 시뮬레이션 진행



“재무 및 비재무 데이터를 활용한 종목분석
with 강화학습 자동매매시뮬레이션”

Data

1600개 상장 종목 가격, 거래량, 재무제표 데이터

20개 종목 비재무 데이터

(출처 : 웹 크롤링)

Algorithm & 방법론

- 1600개 종목 주가 및 재무제표 데이터 수집
 - 웹 크롤링 프로그램 'Octoparse' 를 사용하여 비재무 데이터 수집
 - PAM clustering 실시
 - 종가 예측에 Prophet 모형 적합
-
- 강화학습을 통한 투자 시뮬레이션
 Action : 주식 매매 State : 주식 가격 및 거래량 Reward : 손익률
 로 설정하여 2018년 7월 ~ 2018년 12월까지 트레이딩

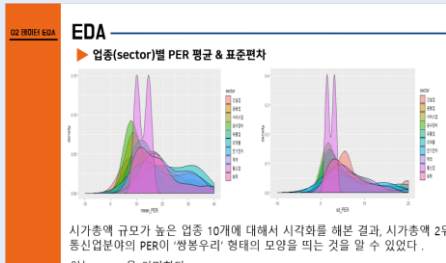
Result

비재무 데이터를 사용함에 따른 수익률 1.2 % 증가

비재무 데이터 수집 방안 제시

분석 의의

- 1) 클러스터링 기법을 통한 종목선정으로 논리성 부여
- 2) Octoparse 라는 프로그램을 사용하여 쉽고 빠른 비재무 데이터 수집
- 3) 강화학습을 통한 투자 시뮬레이션 진행



Summary

데이터 수집 (수정 종가, 재무제표)

네이버 금융 차트 통신기록 통해 수정종가 수집
FnGuide에서 재무제표 수집

데이터 전처리

행(row)은 각 종목, 열(col)은 각 팩터 형태로 정리

팩터 생성 및 결합

모멘텀, 변동성, 퀄리티, 벨류, 고배당, 소형주 팩터 생성 및 결합

백테스팅, 결과 해석

백테스팅 결과를 비교 및 해석

Data

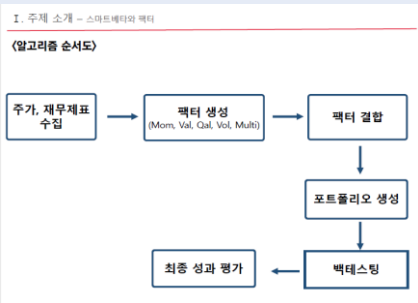
코스피, 코스닥 상장 전종목 수정종가, 재무제표(236개 항목)

(출처 : 웹 크롤링)

스마트 베타를 활용한 퀀트 투자 해보기

권순규





II. 데이터 - 수집

구성해야할 데이터셋의 최종 형태

	팩터1	팩터2	팩터3	팩터4	팩터5
삼성전자					
SK하이닉스					
LG화학					
현대차					
삼성바이오					
LG화학					
POSCO					
삼성바이오로직스					
...					
삼성바이오					
SK하이닉스					
LG화학					
현대차					
삼성바이오					

*row : 종목
col : 팩터



III. 팩터 생성하기

3) 퀄리티 팩터 (Quality)

: 해당 기업이 재무적으로 우량한가? 매출액이 높은가? 영업이익이 높은가?

```

    # 팩터 생성
    R <- data.frame(
      ROE = R[, "ROE"],
      GPA = R[, "GPA"],
      CFO = R[, "CFO"],
      EPS = R[, "EPS"],
      EPS_P = R[, "EPS_P"],
      EPS_M = R[, "EPS_M"],
      EPS_B = R[, "EPS_B"],
      EPS_F = R[, "EPS_F"],
      EPS_Y = R[, "EPS_Y"],
      EPS_T = R[, "EPS_T"],
      EPS_L = R[, "EPS_L"],
      EPS_S = R[, "EPS_S"],
      EPS_D = R[, "EPS_D"],
      EPS_O = R[, "EPS_O"],
      EPS_A = R[, "EPS_A"],
      EPS_M = R[, "EPS_M"],
      EPS_B = R[, "EPS_B"],
      EPS_F = R[, "EPS_F"],
      EPS_Y = R[, "EPS_Y"],
      EPS_T = R[, "EPS_T"],
      EPS_L = R[, "EPS_L"],
      EPS_S = R[, "EPS_S"],
      EPS_D = R[, "EPS_D"],
      EPS_O = R[, "EPS_O"],
      EPS_A = R[, "EPS_A"]
    )
  
```

ROE, ROA, GPA, CFO, 장기차입금, 유동자산, 매출총이익 등 여러가지를 계산했으나, 최종적으로는 ROE, ROA, GPA, CFO 네 가지 지표만 사용!

III. 팩터 생성하기

6) 투자자본 수익률 팩터 (Return on Capital)

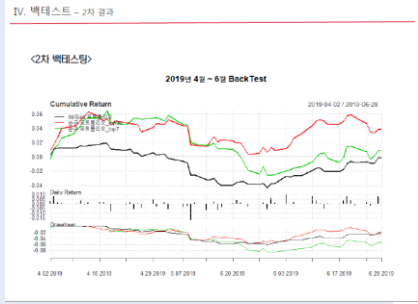
: (기업의 수익) / (투자한 자본)

투자자본 수익률 = $\frac{\text{이자 및 법인세 차감전이익}}{\text{투자자본}}$

투자자본 = $\frac{\text{당기순이익} + \text{법인세} + \text{이자비용}}{(\text{유동자산} - \text{유동부채}) + (\text{비유동자산} - \text{당기상각비})}$

```

    # 팩터 생성
    R <- data.frame(
      ROE = R[, "ROE"],
      GPA = R[, "GPA"],
      CFO = R[, "CFO"],
      EPS = R[, "EPS"],
      EPS_P = R[, "EPS_P"],
      EPS_M = R[, "EPS_M"],
      EPS_B = R[, "EPS_B"],
      EPS_F = R[, "EPS_F"],
      EPS_Y = R[, "EPS_Y"],
      EPS_T = R[, "EPS_T"],
      EPS_L = R[, "EPS_L"],
      EPS_S = R[, "EPS_S"],
      EPS_D = R[, "EPS_D"],
      EPS_O = R[, "EPS_O"],
      EPS_A = R[, "EPS_A"]
    )
  
```



Algorithm & 방법론

- R을 활용한 전 종목 수정 종가, 재무제표 데이터 수집
- 팩터 생성 및 결합
- 각 종목 팩터에 대한 순위(z score)를 계산하여 포트폴리오 형성
- 고배당 팩터 기준으로 상위 10개 종목 최종 선정

Result

2017년 2분기, 2018년 2분기 백테스팅 결과
-> 벤치마크 포트폴리오(주식60 채권40) 대비 4.25%, 5.76% 아웃퍼폼

분석 의의

- 1) 직접 수집한 데이터를 통한 퀀트 투자
- 2) 팩터간 비중 조절에 관한 방법론 추가 논의 필요

Summary

게임 유저 이탈 시점 예측

유저들의 게임시간, 거래 횟수, 시간대별 접속 여부, 길드 활동, 현금 결제 내역 등의 주어진 데이터로 예측 모형 적합

이탈 여부를 결정하는 주요 변수 확인

머신러닝 해석 알고리즘을 통한 주요 변수 해석

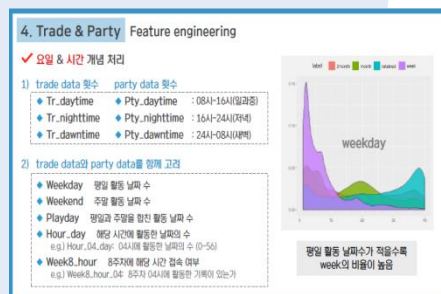
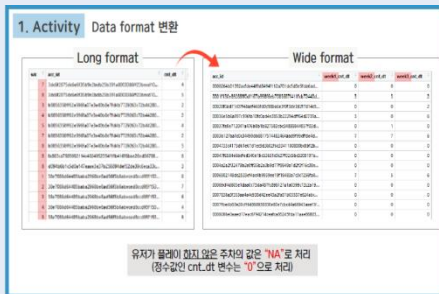


Data

블레이드앤소울 게임 유저 8주간의 활동 데이터

Obs: 14만 개 (Train set: 10만 개 / Test set: 4만 개)

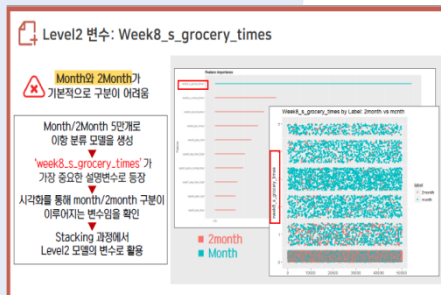
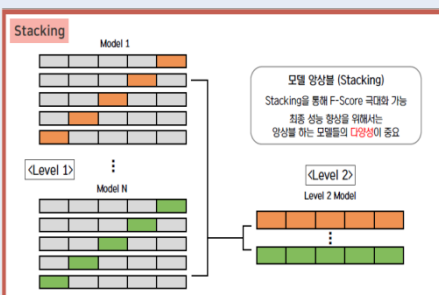
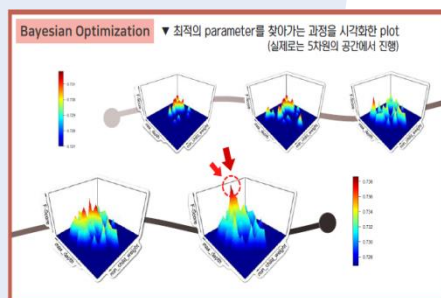
(출처 : NC소프트)



알고리즘 비교

	Regression	Neural Network	Ensemble	Random Forest	Extra Trees	XGBoost
성능	X	O	O	O	Δ	O
설명력	O	X	Δ	Δ	O	Δ
속도	Δ	X	O	Δ	O	O
이상치 영향	큼	큼	적음	적음	적음	적음
결측치 자동처리				X	X	O
정교함 (parameter)				Δ	X	O

→ XGBoost를 주 모델로 선정



Algorithm & 방법론

- 합리성, 일관성, 정보 극대화를 기준으로 데이터 전처리
- Bayesian Optimization을 통해 모형의 파라미터 튜닝 효율 극대화
- XGBoost, RandomForest, Extra Trees를 앙상블하여 최종 모델 생성
- PDP를 이용하여 이탈 요인 해석

Result

F1-Score (예측 정확도) : 73.6%

분석 의의

- 1) 파라미터 차원이 많은 모델의 효율적인 최적화
- 2) 반복적인 Cross Validation과 시각화를 통한 과적합 방지
- 3) PDP를 이용하여 블랙박스 모델에 대해서도 계수 해석
- 4) 타겟 마케팅을 위한 이탈 위험 유저 예측 및 이탈 요인 분석

08 맛집 추천시스템 개발

2018.09 ~ 2018.12

Summary

개인화 맛집 추천 알고리즘 개발

기존의 '다이닝코드', '식신' 과 같은 인기순, 협찬순의 맛집 추천이 아닌 머신러닝 추천 알고리즘 연구를 통한 개인화 서비스 제공

Data

유저별 식당 평가 데이터
식당 소개 및 기본 정보 데이터
(출처 : 식신 사이트 웹크롤링)

Algorithm & 방법론

- 협업 필터링과 콘텐츠 기반 필터링을 앙상블 하여 하이브리드 필터링 사용





KWON
SOON GYU

CONTACT

E-mail purestar0509@gmail.com

Phone 010 – 7551 – 0164