

Senior Safe

Dashboard를 활용한 시니어 맞춤형
종목 선별 시스템 구축

- 부실하체 트레이너
- 발표자 : 김도현
- 팀원 : 김지훈, 김도현, 김보윤, 오대한, 정승원

목차

Contents

01

Introduction

배경 | 목적 | 차별점 | 워크플로우 | 역할 분배

02

Data Set

데이터 수집 | 종속변수 라벨링 | EDA | 데이터전처리 | 피처엔지니어링

03

Modeling

모델링 | 성능평가

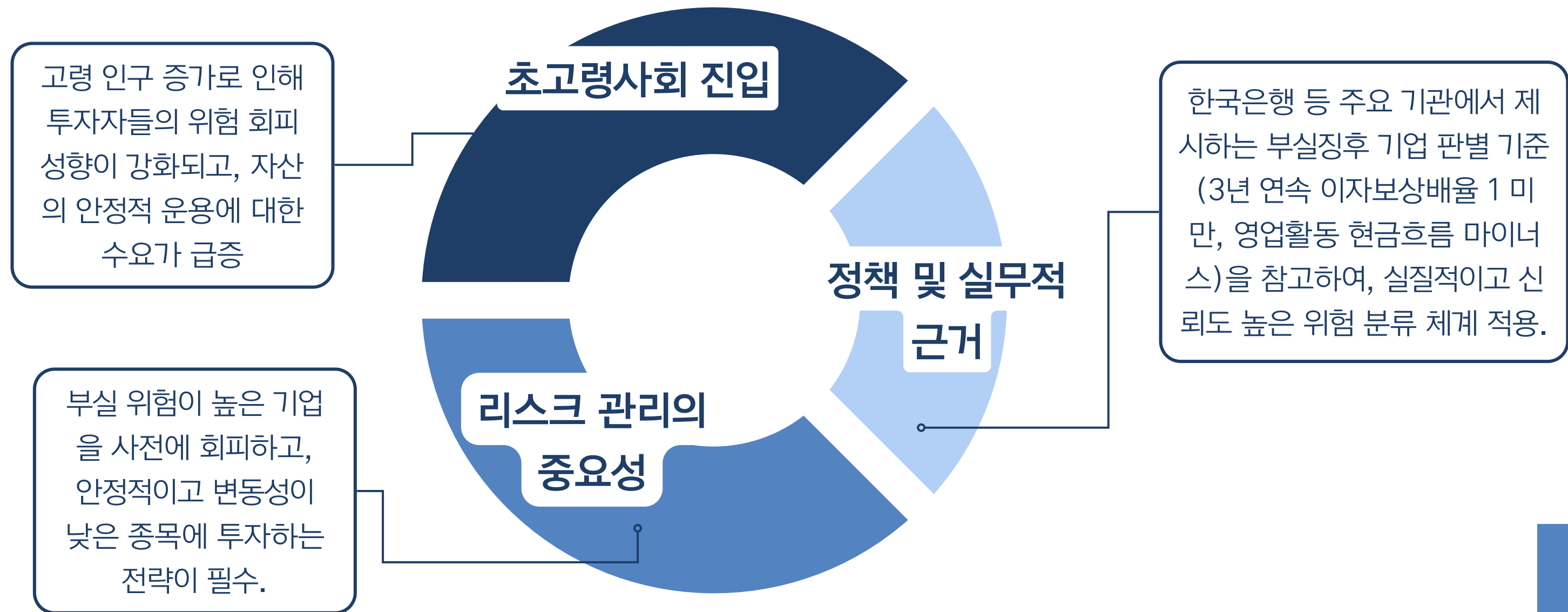
04

Utilization

필터링 | 백테스팅 | 대시보드 구축

Introduction

배경



Introduction

목표

부실 위험 회피 + 안정적·저위험 종목 추천

재무데이터 기반으로 부실
위험이 높은 기업을 체계적
으로 회피하고,
벤치마크 수익률을 이기면서
안정성까지 고려한 종목을
추천.

투자자 맞춤형 종목 선별

투자자가 직접 위험률을 확
인하여 종목을 선택할 수 있
게 대시보드를 제공

Introduction

차별점

Why?

- 기존 단순 재무비율 스크리닝을 넘어, 부실 위험을 단계별 분류와 시계열·머신러닝 결합으로 **조기경보** 및 **실질적 투자성과** 극대화
- 초고령 사회의 투자환경 변화에 맞춘 실질적 투자자 보호 및 자산증식 솔루션 제공

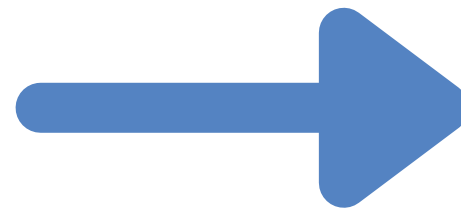
How?

- 공신력 있는 부실 판별 기준(한국은행 등) 기반의 라벨링 → 신뢰성 확보
- ML/통계모형과 시계열 검증 결합 → 실제 투자 환경 반영
- 저위험 종목 내 추가 필터링(베타, 배당 등) → 투자자 맞춤형 추천 실현

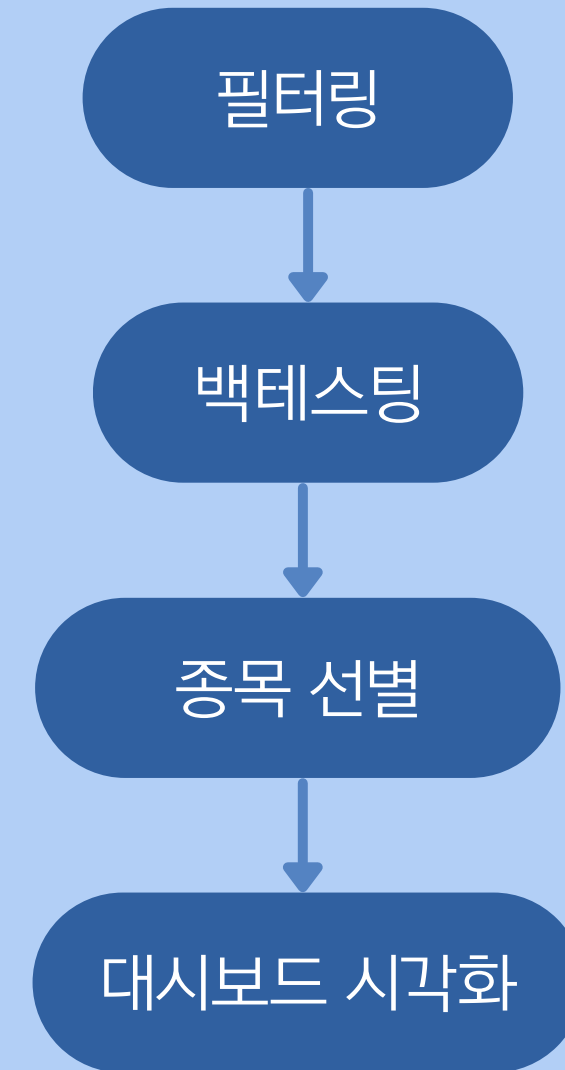
Introduction

워크플로우

위험도 분류모형 설계



분류모형 활용



Introduction

역할 분배



데이터 수집 및 전처리

EDA

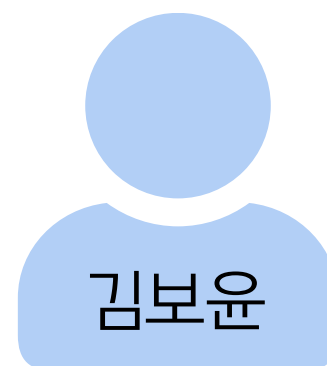
인사이트 도출



데이터 수집 및 전처리

변수 선정

데이터 시각화



데이터 수집 및 전처리

서브코더

PT 자료 준비



파생변수 선정

백테스팅

관련자료 수집



파생변수 선정

메인코더

대시보드 시각화

Data Set

데이터 수집

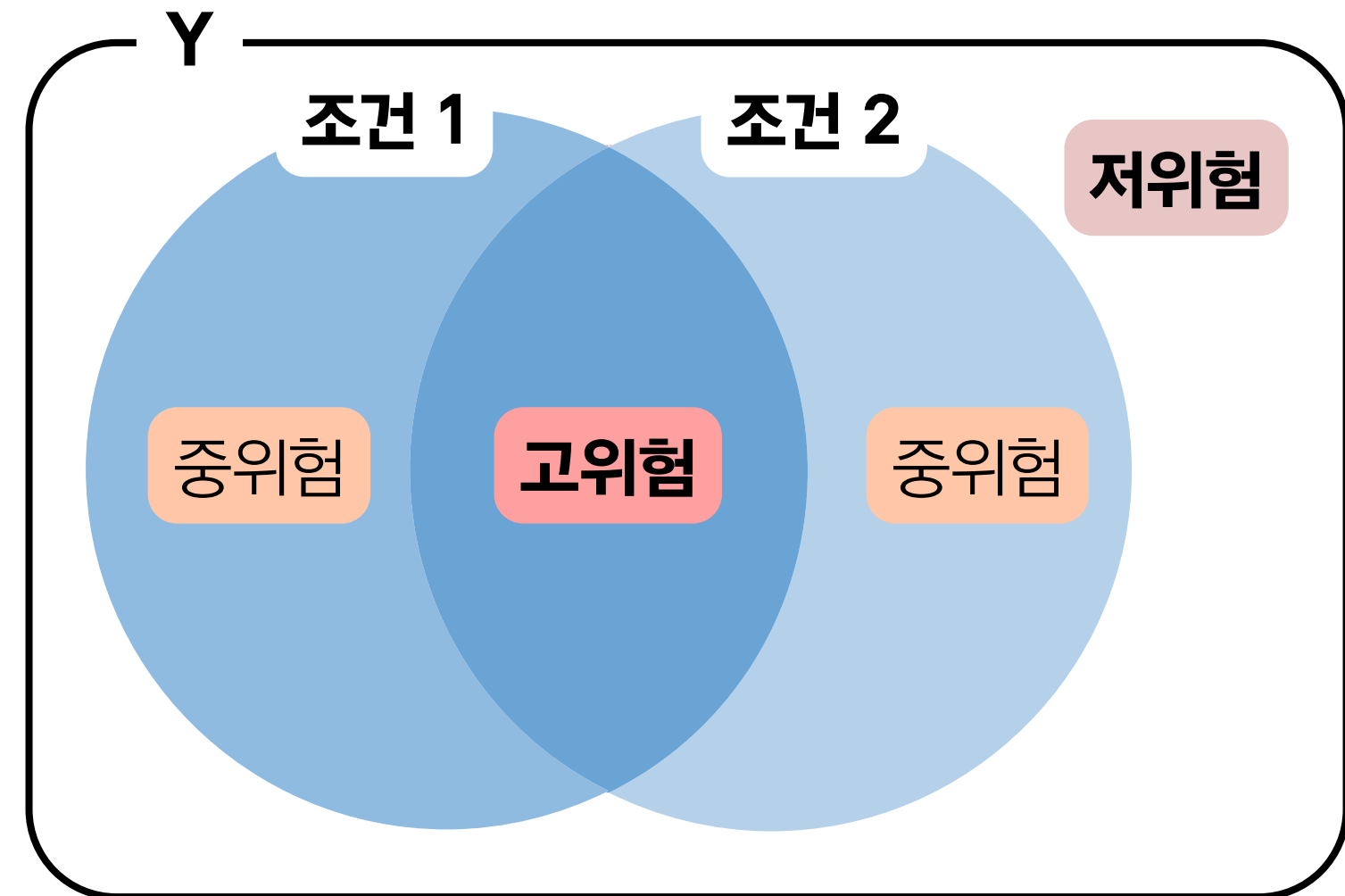


Data Set

종속변수 라벨링

종속변수 (Y)

- 조건 1 : 3년 연속 **이자보상배율 1 미만**
- 조건 2 : 3년 연속 **영업활동현금흐름 0 미만**
- 고위험(2) : 두 조건 모두 해당
- 중위험(1) : 두 조건 중 하나만 해당
- 저위험(0) : 두 조건 모두 해당 안됨



※ 종속변수 (Y) 산출에 쓰인 변수 (이자보상배율, 영업활동현금흐름)는 X에서 제외하여 모형의 예측력과 해석력 확보 (동일 변수 반복 사용 방지)

Data Set

EDA

데이터 구조 및 기초 통계

- 전체 데이터 규모(행/열), 연도별·시장별(코스피/코스닥)·산업별 샘플 분포 파악
- 주요 변수(재무비율, 주가, 거시변수 등)별 기초통계량(평균, 중앙값, 표준편차, 최댓값/최솟값) 산출

변수 분포 및 시각화

- 주요 재무데이터, 시장지수, 거시변수의 전체 분포: 히스토그램, KDE, 박스플롯 활용
- 라벨별(고/중/저위험) 분포 비교: 그룹별 박스플롯·바이올린플롯 등 시각화
- 연도별, 산업별, 시장별 요약통계량 및 분포 변화 시각화

상관관계 탐색

- 피어슨/스피어만 상관계수 행렬로 변수 간 상관성 확인
- 변수 간 관계를 히트맵으로 시각화

Data Set

데이터 전처리

결측치처리

- 변수별 결측치 비율 집계 및 시각화
- 결측치가 5% 미만인 경우: 평균/중위수/이전값 등으로 대체
- 결측치가 5% 이상인 경우: 해당 변수 또는 관측치 제거

이상치처리

- 박스플롯, 히스토그램, 산점도 등으로 이상치 탐지
- 연속형 변수의 상하위 1~2% winsorizing(극단값 절단) 적용

다중공선성 점검

- VIF(Variance Inflation Factor) 고려

Data Set

데이터 전처리

데이터 불균형 진단 및 처리

- 라벨(고/중/저위험)별 샘플 비율 확인
- 불균형 심할 경우 SMOTE 등 오버샘플링, 언더샘플링, 클래스 가중치 조정 적용

시계열 데이터 처리

- 결측치/이상치 보정 후 시계열 정렬 및 연속성 점검
- 시계열 분할(Sliding/Expanding Window 등) 적용 준비

Data Set

피처 엔지니어링

라벨별 그룹 간 차이 검정

- 고/중/저위험 그룹별 주요 재무비율 평균 차이: t-test(이원), ANOVA(3그룹 이상)
- 정규성 미충족 시 Mann-Whitney U, Kruskal-Wallis 등 비모수 검정 활용
- 유의미한 차이 확인된 변수는 주요 피처로 우선 선정

변수 중요도 분석 및 선택

- 랜덤포레스트, XGBoost 등 모델 feature importance, SHAP 값 등 활용
- 정보이득(Information Gain), Gain Ratio 등 통계적 변수 선택 기법 병행

시계열 정상성 검정 및 변환

- ADF(Augmented Dickey-Fuller) 테스트로 선택된 변수들의 정상성 확인
- 비정상 시 차분(differencing) 등으로 정상화

Data Set

피처 엔지니어링

라벨별 그룹 간 차이 검정

- 고/중/저위험 그룹별 주요 재무비율 평균 차이: t-test(이원), ANOVA(3그룹 이상)
- 정규성 미충족 시 Mann-Whitney U, Kruskal-Wallis 등 비모수 검정 활용
- 유의미한 차이 확인된 변수는 주요 피처로 우선 선정

**위와 같은 체계적인 EDA 및 데이터 전처리 과정을 거쳐,
모델링에 활용될 변수의 품질과 신뢰도가 충분히 확보된 최종 데이터셋을 구축**

변수 중요도 분석 및 선택

- 랜덤포레스트, XGBoost 등 모델 feature importance, SHAP 값 등 활용
- 정보이득(Information Gain), Gain Ratio 등 통계적 변수 선택 기법 병행

시계열 정상성 검정 및 변환

- ADF(Augmented Dickey-Fuller) 테스트로 정상성 확인
- 비정상 시 차분(differencing) 등으로 정상화

Modeling

모델링

1. 다항 로지스틱 회귀 (Multinomial Logistic Regression)

- 다항 로지스틱 회귀는 범주형 종속변수가 3개 이상인 경우에 적합한 선형 분류 모델.
- 각 클래스에 속할 확률을 직접적으로 추정하며, 회귀계수의 해석이 명확해 변수와 결과 간의 관계를 통계적으로 이해하기 용이함.

통계적 강점

- 각 변수의 기여도를 계수로 정량화 가능
- 변수 간 독립성(낮은 다중공선성)만 확보되면 안정적 추정 가능
- IIA(irrelevance of irrelevant alternatives) 가정 하에서 클래스 간 선택 확률을 비교적 단순하게 모델링

Modeling

모델링

2. 다변량 판별분석 (Linear Discriminant Analysis, LDA)

- LDA는 각 클래스의 분포가 정규분포이고, 클래스 간 공분산이 동일하다는 가정 하에서 클래스 간 분리도를 극대화하는 선형 결정 경계를 학습함.
- 다중 클래스(3개 이상) 분류에 자연스럽게 확장 가능하며, 차원 축소와 분류를 동시에 수행

통계적 강점

- 클래스 간 평균 차이와 분산을 동시에 고려해 효과적인 분류
- 차원 축소(투영) 효과로 해석력 및 계산 효율성 우수
- 데이터가 선형적으로 분리 가능할 때 높은 정확도

Modeling

모델링

3. 랜덤 포레스트 (Random Forest)

- 랜덤 포레스트는 여러 결정트리를 배깅(bagging) 방식으로 결합하여 예측의 분산을 줄이고, 비선형적·복잡한 데이터 구조도 효과적으로 학습함

통계적 강점

- 변수 간 상호작용, 비선형성, 이상치에 강인함
- 변수 중요도 (Feature Importance) 산출로 해석력 제공
- 과적합 위험 감소 (앙상블 효과)

Modeling

모델링

4. XGBoost

- XGBoost는 경사하강법 기반의 부스팅(Boosting) 트리 모델로, 각 단계에서 오차를 줄여가며 강력한 예측 성능을 보입니다.
- 변수 중요도, 결측치 자동 처리, 정규화 등 지원.

통계적 강점

- 강한 비선형성, 변수 간 상호작용, 이상치/결측치 자동 처리
- 학습 과정에서 과적합 제어(정규화)
- 변수 중요도 산출로 해석력 제공

Modeling

모델링

5. LightGBM

- LightGBM은 XGBoost와 유사한 트리 기반 부스팅 모델이나, 더 빠르고 효율적으로 대용량·고차원 데이터를 처리함.
- leaf-wise 성장 방식으로 더 깊은 트리를 빠르게 생성, 높은 정확도를 달성

통계적 강점

- 빠른 학습 속도, 메모리 효율성, 대규모 데이터 적합성
- 희소 데이터, 범주형 변수 처리에 강점
- 변수 중요도 산출 및 해석 가능

Modeling

모델링

6. 앙상블 / 스택킹 (Ensemble / Stacking)

- 앙상블과 스택킹 기법은 여러 개의 서로 다른 머신러닝 모델의 예측을 결합해 단일 모델보다 더 높은 예측 성능과 일반화 능력을 확보할 수 있다.
- 스택킹은 특히 다양한 모델의 예측 결과를 메타 모델이 종합적으로 학습해, 데이터의 복잡한 패턴까지 효과적으로 반영할 수 있다.
- 여러 모델의 강점을 결합함으로써, 데이터의 다양한 특성과 불확실성에 유연하게 대응할 수 있다.

통계적 강점

- 다양한 모델의 예측을 결합해 단일 모델보다 더 높은 정확도와 신뢰도 확보
- 모델 간 약점을 상호 보완해 과적합 위험 감소, 새로운 데이터에 대한 예측력 강화
- 다양한 알고리즘(선형, 비선형, 트리 기반 등) 조합으로 데이터의 복잡한 패턴 효과적 반영

Modeling

성능 평가

평가지표

- 정확도 (Accuracy)
- 재현율 (Recall), 정밀도 (Precision)
- F1-score 등 다양한 분류 성능 지표 활용

Utilization

필터링

★ 각 위험 종목군에서 추가 필터링 ★

- **베타(β)**: 시장 변동성에 대한 민감도(저변동/고변동 분류)
- **안정성**: 수익률 변동성, 재무 건전성 등
- **배당**: 배당수익률 등

WHY?

위험등급별 투자성과 분석

- 고·중·저위험 분류별로 **과거 수익률, 변동성, 배당수익률** 등 실질적 투자성과와의 연관성 분석
→ 투자자 성향별 종목 추천의 신뢰성 확보

투자자 성향 반영

- 대시보드에서 투자자 성향(위험회피/수익추구 등) 입력 시,
- 저위험+저변동+고배당, 저위험+고변동 등 맞춤형 종목 추천

Utilization

백테스팅

저위험 종목군 선정 근거

- 베타, 안정성, 배당 등의 지표를 반영해 필터링된 저위험 종목군을 대상으로 백테스팅을 실시
- 국고채 3년물 이상의 수익률을 목표로, 절대적 안정성과 실질적 투자 대안으로서의 성과를 검증

슬라이딩 윈도우(sliding window) 방식

- 과거 5년 데이터를 학습(포트폴리오 구성) 후, 직후 구간에서 실제 투자성과를 검증합니다.
- 이후 윈도우를 일정 간격만큼 이동시키며 반복 수행
- 전략의 일관성과 시장 환경 변화에 대한 적응력을 평가

Utilization

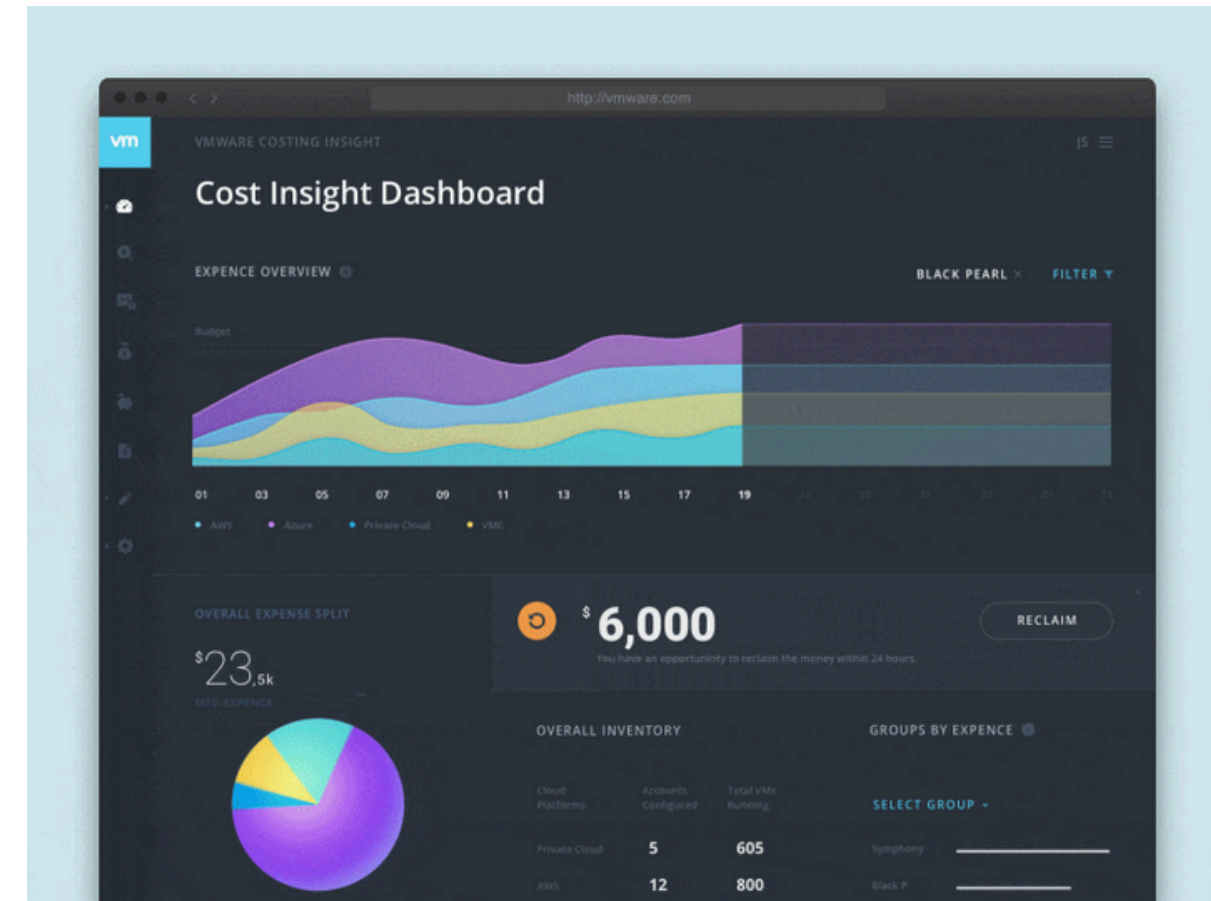
대시보드 구축

투자자 성향 입력 및 진단 기능

- 위험 선호도(저위험/고위험), 변동성 선호(저변동/고변동), 배당 선호 등 투자자 성향 설문 입력
- 입력값 기반으로 최적의 포트폴리오 추천 로직 자동 적용

위험등급별 종목 필터링 기능

- 고/중/저위험 등급별로 종목 자동 분류
- 저위험군 내에서 베타(변동성), 배당수익률, 재무건전성 등 추가 필터링 옵션 제공
- 저위험-저변동, 저위험-고변동 등 맞춤형 조합 선택 가능




대시보드 예시



Time Table

타임테이블

Plan A 
Plan B 

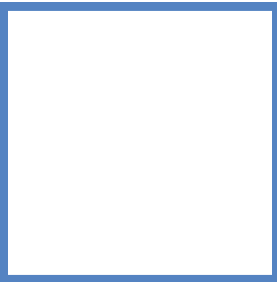
1주차 (06/02~06/09)	2주차 (06/10~06/16)	3주차 (06/17~06/23)	4주차 (06/24~06/30)	5주차 (07/01~07/08)
논문 리뷰 준비	프로젝트 기획			
		데이터 수집/EDA		
			모델링	
				백테스팅
				대시보드 제작

06/09 (월)
논문발표

06/13 (금)
기획발표

06/24 (화)
중간발표

07/08 (화)
최종발표



Reference

참고문헌

- 이자보상배율 취약기업 증가 배경 및 시사점_한국은행_금융안정보고서(202106)
- 지현미 (2015). 낮은 현금기준 이자보상배율이 영업이익의 가치관련성에 미치는 영향. Journal of The Korean Data Analysis Society, 17(6), 3197 - 3210.
- 고령화 사회, 경제성장 전망과 대응방향_KDI정책포럼 제273호(2019-02)(2019. 4. 18)
- 박희정, 강호정 (2009), “로지스틱회귀분석을 이용한 코스닥기업의 부실예측모형 연구” 한국콘텐츠학회논문지, 제9권제3호
- 김종훈, 박규일, 김민철 (2011), “상장폐지기업의 재무적 특성과 예측에 관한 연구” 회계연구, 제16권 제2호, pp.125-142

감사합니다

Thank you

■ 부실하체 트레이너 ■ 발표자 : 김도현

김지훈 김도현 김보윤 오대한 정승원