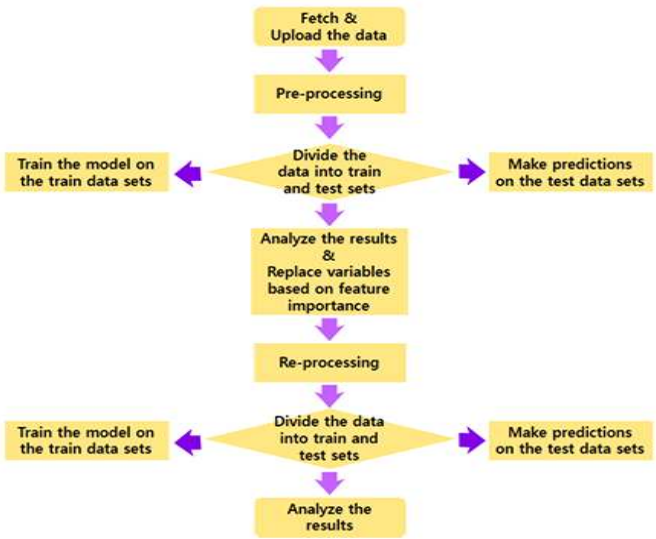
랜덤 포레스트 모델을 활용한 채권형펀드 성과 예측 연구

이 연구는 채권형펀드와 펀드 수익률의 예측에 관한 연구로써 해외에서는 학술 연구들이 많이 이루어지고 있는 것과는 달리 국내는 문헌이 부족한 상황에서 진행한 논문으로서 의의가 있다고 볼 수 있다. 연구에 사용된 시계열 분석은 가장 전통적인 수익률 예측 방법 중 하나다. 시계열 분석은 과거 데이터로부터 미래의 값을 예측하는 분석 방법으로, 자기회귀(AR) 모형, 이동평균(MA) 모형, AR과 MA 모형을 혼합시켜 자기회귀이동평균(ARMA) 모형이 있다. 현대 금융시장은 매우 복잡한 구조를 갖추고 있으며, 다양한 변수들에 의해 영향을 받고 있다. 학계에서는 이러한 구조를 파악하고 예측력을 높이기 위해 머신러닝 모형을 활용한 연구가 점점 더 활발해지고 있다. 이러한 기계학습 모형은 다양한 변수 간의 복잡한 패턴을 탐지하여 예측 정확도를 높일 수 있으므로 최근 금융시장에서는 매우 유용하게 활용되고 있는 방법론이다.

본 연구에서는 랜덤 포레스트 모형을 실험하기 위해 구글 코랩(google colab)을 사용하였다. 구글 코랩은 파이썬 스크립트를 작성하고 실행할 수 있는 편리한 도구이다. .랜덤 포레스트 모형을 구현하기 위해 사이킷런(scikit-learn), 데이터 조작 및 분석을 위해 판다스(pandas), 대규모의 다차원 배열을 쉽게 처리할 수 있도록 넘파이(numpy) 등 오프소스 파이썬 프로그래밍 언어용 라이브러리를 설치하였다.기계학습 실행에 앞서 전처리(pre-processing) 과정으로 데이터의 불균형(imbalance) 문제를 해결하기 위해서 모든 클래스의 개수를 맞춰 주었고, 데이터가 중복되고 불필요한 자료들은 삭제(drop)시켰다. 뿐만 아니라, 과적합(overfitting)과 훈련시간(training time) 단축을 위한 과정도 거쳤다



모델링 전에는 훈련세트(train data set)와 테스트세트(test data set)를 나누어 주는데, 본 연구 모형에서는 훈련세트와 테스트세트를 7:3의 비율로 조절 하였다. 또한 모델의 예측 능력을 높이기 위해서 하이퍼파라미터(hyperparameter)의 튜닝(tuning) 작업이 필요한데, Table 2은 본 모델에서 설정된 파라미터의 수치를 나타낸다.

Table 2. Set parameter values

Parameter	Value
n_estimators	100
max_depth	14
min_samples_split	2
min_samples_leaf	1
bootstrap	False
random_state	30

본 실험에서 결정 트리 개수는 100, 무작위로 선택할 난수의 개수는 30, 트리의 깊이는 14, 리프노드(leaf node)가 되기 위한 최소한의 샘플링 데이터 수는 1, 노드를 분할하기 위한 최소한의 샘플링 데이터 수는 2, 데이터 샘플링 중복 여부는 False를 가진다.

실험 결과 분석3.1 결정트리(decision tree) 분석2 1 False 30본 실험에서 결정 트리 개수는 100, 무작위로 선택할 난수의 개수는 30, 트리의 깊이는 14, 리프노드(leaf node)가 되기 위한 최소한의 샘플링 데이터 수는 1, 노드를 분할하기 위한 최소한의 샘플링 데이터 수는 2, 데이터 샘플링 중복 여부는 False를 가진다.

머신러닝 모형은 다양한 방법으로 예측 성능을 평가하는데, 본 연구에서는 분류 모델에 사용되는 평가지표를 사용하였다. Table 3은 실험 모형으로 테스트 데이터를 예측한 결과를 오차행렬(confusion matrix)로 나타낸 것이고, Table 4은 실험 모형의 예측성과를 표로 나타낸 것이다.

Table 3. Confusion matrix of prediction

		Predict	
		0	1
Actual	0	756	76
	1	78	769

Table 4. Evaluation metrics of result

	Precision	Recall	F1-score
0	0.91	0.91	0.91
1	0.91	0.91	0.91
Accuracy			0.9083

본 실험에서 정확도(accuracy)는 90.83%로 실제 데이터와 예측 데이터가 얼마나 일치하는지를 나타내는 수치로서, 이 연구의 모형은 뛰어난 예측 성과를 보여주고 있다. 정밀도(precision)와 재현율(recall)은 각각 91%로 데이터 분석에서 positive 데이터 세트의 성능을 평가하기 위한 지표이며, 분석 상황에 따라 이 중 하나를 선택하여 사용한다. F1 스코어는 정밀도와 재현율을 결합한 지표로, 두 값이 모두 균형을 이룰수록 높은 값을 가지며, 본 실험에서는 91%를 기록하였다