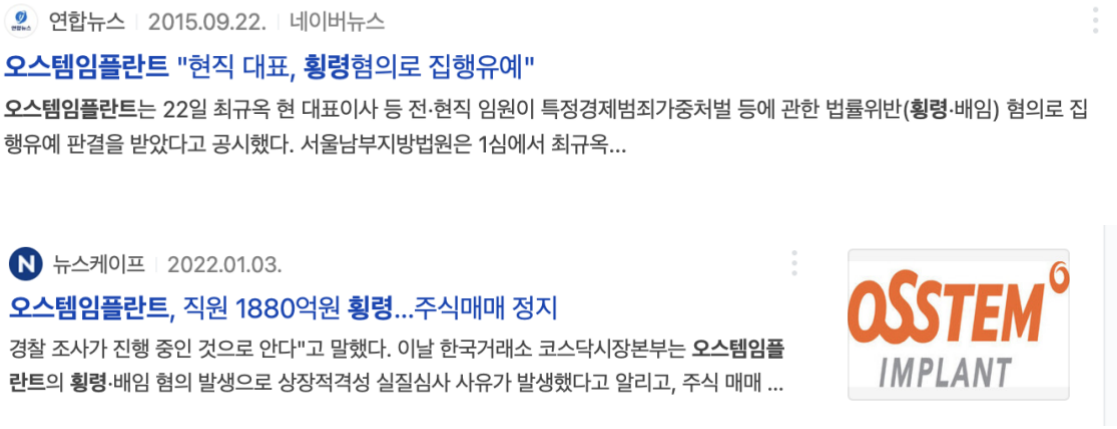
캡스톤 중간 발표

- 목차

1. 프로젝트의 배경과 설명

주식 거래 데이터 속에서 이상치를 지닌 날짜를 선별하여 기사를 이용해 종목별 주식 거래에 큰 영향력이 있었던 사건들을 보여주고자 합니다.



2. 이상치 모델

- 이상치 모델 사용 데이터

코스피 시장 거래 데이터 (코스피 거래량 회전율 , 코스피 일일 수익률)
코스피200 구성 종목 거래 데이터 (개별종목 거래량 회전율 , 개별종목 일일 수익률)
종목 거래량 회전율 : $\frac{\text{거래량}}{\text{종목 상장 주식수}}$

- 이상치

시장에 대한 관심도와 코스피200 구성 종목에 대한 관심도를 비교할 필요가 있습니다.
예) 코로나 이후 주식시장은 동학개미 운동으로 큰 관심을 받았습니다.
주식 시장이 관심을 받아 개별 종목이 관심을 많이 받을 수가 있기 때문에 주식시장 전체와 개별종목 관심 사이의 비교가 이상치탐지모형의 핵심입니다.

시장의 관심도 : 코스피 거래량 회전율 (가중거래량회전율), 코스피 일일 수익률
개별 종목 관심도 : 종목 거래량 회전율, 종목 일일 수익률

- 이상치 탐지를 위한 모델

1. Robust Regression

- 기존의 선형회귀의 loss function인 Least Square Method가 아닌 Huber loss function을 사용하여 추정 회귀선이 이상치에 영향을 덜 받도록 합니다.

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \le \delta, \\ \delta(|a| - \frac{1}{2}\delta), & \text{otherwise.} \end{cases}$$

- 종목 거래량 회전율과 코스피 거래량 회전율 사이 회귀모델
  종목 일일 수익률과 코스피 일일 수익률 사이 회귀모델
  두가지 모델의 예측값의 신뢰구간을 통해 신뢰구간을 벗어난 이상치를 탐지합니다.

2. Isolation Forest

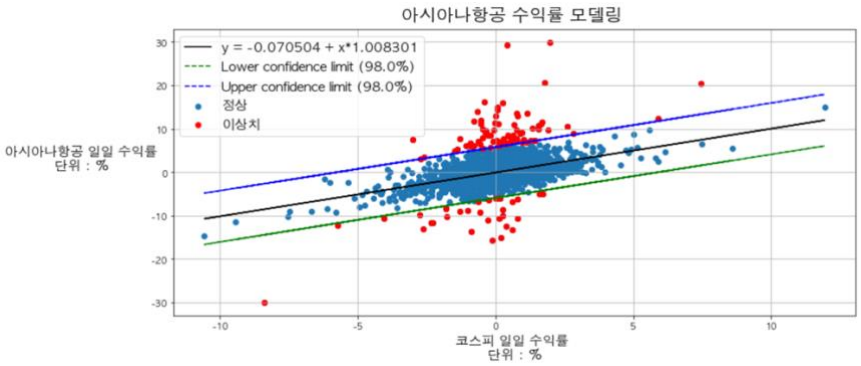
- Isolation Forest : Random Forest가 변수선택을 무작위로 한다면 Isolation Forest는 구분선을 무작위로 선정해 관측들을 고립시키는 모델으로 Robust한 모델이다.

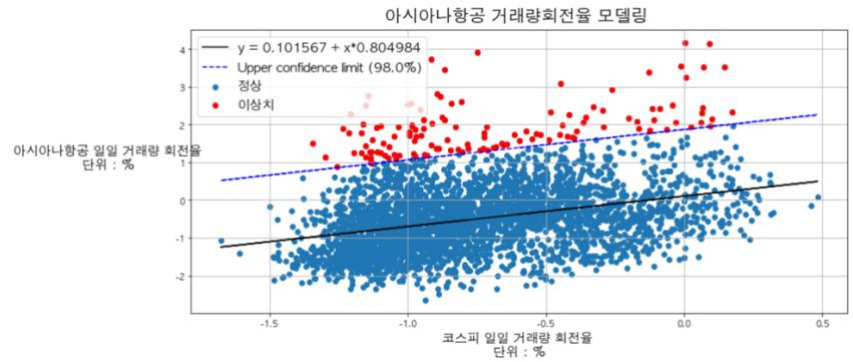
- 모델 input data : | 코스피 수익률 – 개별종목 수익률 |
  종목 거래량 회전율 / 코스피 거래량 회전율

3. 아시아나항공, 대한항공 이상치

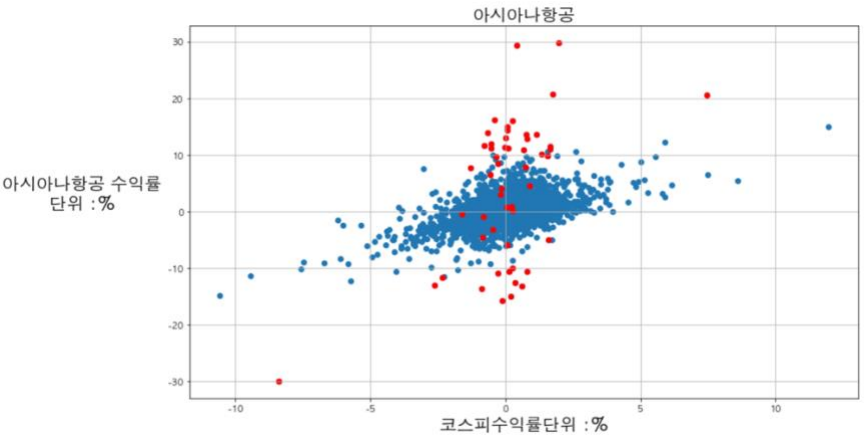
- 아시아나항공

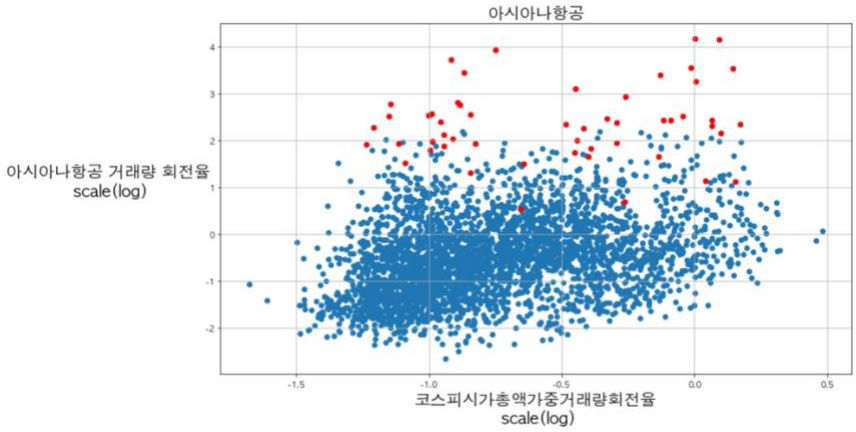
〈Robust Regression〉





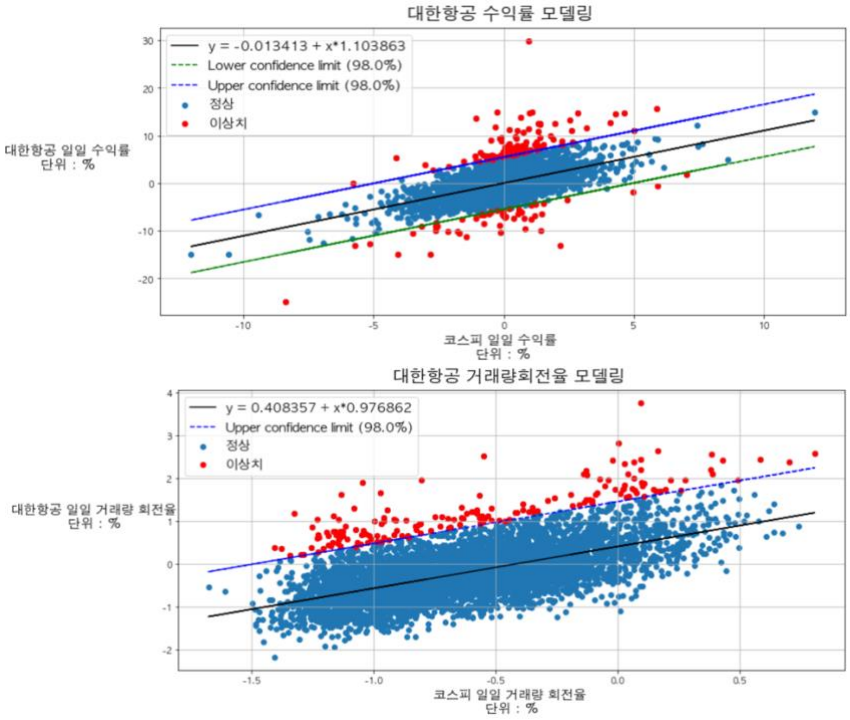
〈Isolation Forest〉



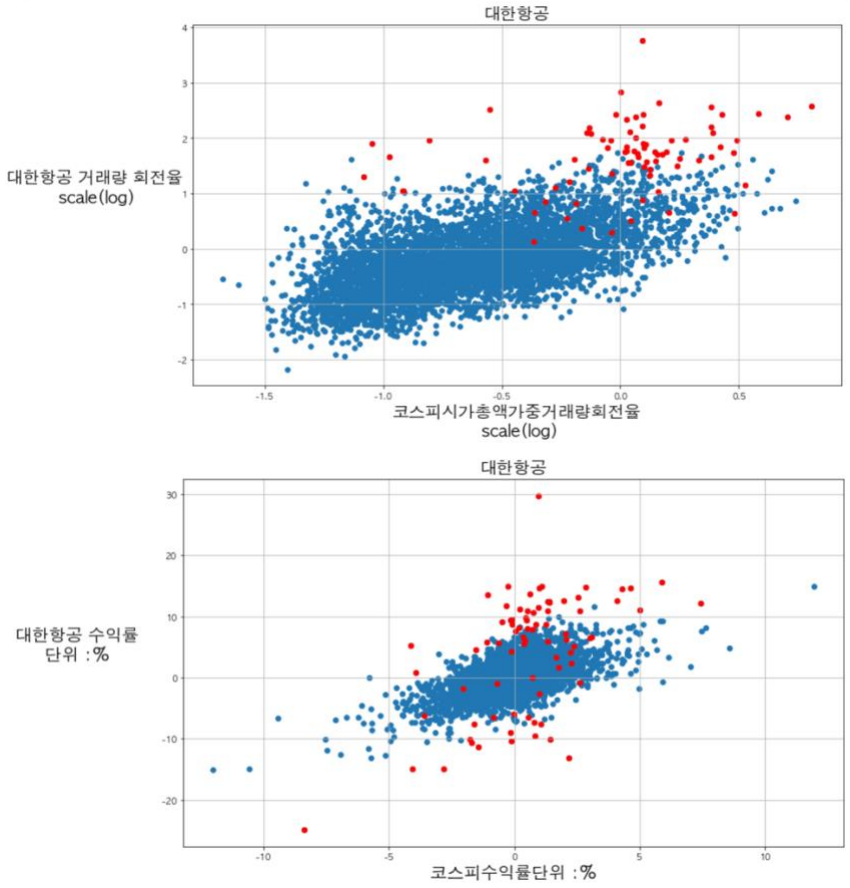


- 대한항공

〈Robust Regression〉



〈Isolation Forest〉



4. 뉴스 DB 구축

종목 별 뉴스를 정하는 기준

- 제목에 종목을 포함하는 것이 아니라, 내용에 종목의 이름이 들어가는 것으로 일단 정해보았다.

Ex) 다른 경쟁사의 영향으로 인해서, 다른 대안 종목의 가격이 내려가면, 그 뉴스도 종목 거래에 영향을 준 사건이라 볼 수 있다.

어느 시기부터 크롤링을 해야 할까?

→ 회사의 상장일 기준으로부터 시작.

만약 상장일이 2000 년 1 월 1 일 전이라면, 2000 년부터 시작해도 될 것이라고 판단하였습니다.

원래는 전 종목 전체 뉴스 DB 를 구축하려고 하였으나 의미 없는 뉴스 삭제 및 효율적인 전처리를 위해 뉴스 전체를 구축하기보다는 관련도(정확도)순으로 검색을 하여, 의미 있는 DB 를 구축하는 것이 낫다고 생각 하였습니다.

	Title	Contents	Date
1	故 조양호 회장 '마지막 비행'...정·재계 조문 행렬	[앵커]미국에서 별세한 故 조양호 한진그룹 회장의 운구가 오늘 새벽 국내에 도착했습...	2019-04-12
2	조양호 회장 빈소 첫날...각계 인사 조문	[컨슈머타임스 전은정 기자] 고(故) 조양호 한진그룹 회장의 빈소가 차려진 첫날인 ...	2019-04-12
3	"안타깝다..." 故 조양호 회장 빈소에 각계 조문 행렬(종합2보)	원태·현아·현민 3남매 문상객 맞아...文 대통령 등 조화 보내 애도 고인과 친분 있는...	2019-04-12
4	"항공에 평생 바친 분" 조양호 회장 빈소 첫날...조문 행렬 이어져	12일 오전 서울 서대문구 신촌 세브란스 병원 장례식장에 마련된 고(故) 조양호 한...	2019-04-12
5	[4월13일자] 비즈니스포스트 아침의 주요기사	자동등록방지 코드 확인 - 200 자까지 쓰실 수 있습니다. (현재 0 byte /...	2019-04-12
...	...	...	...
2771	대한항공 기내식 하루 8만4936식 생산 '역대 최대'	조유진 기자 tint@asiae.co.krAD[아시아경제 조유진 기자]은 지난달 3...	2016-08-02
2773	KT, 대한항공과 신규 데이터 로밍서비스 출시	ZZZ 0 KT는 대한항공과 신규 로밍서비스 3종을 1일 출시했다./제공=KTKT가...	2016-08-02
2774	하루에만 8만5000식...대한항공 기내식 역대 최대 기록	[사진= 대한항공 기내식 생산 현장 모습. 제공=대한항공][헤럴드경제=정태일 기자]...	2016-08-02
2776	대한항공, 하루 기내식 생산량 역대 최고치 기록	▲ 대한항공 기내식 생산 현장 (제공: 대한항공)8만 4936식 생산량 달성... 19...	2016-08-02
2778	'여름 휴가 절정' 대한항공 기내식 역대 최대 기록 경신	대한항공 기내식 조리사들이 기내식을 만들고 있다.[사진제공=대한항공]대한항공 기내식...	2016-08-02

현재 크롤링 및 뉴스 DB 구축 코드는 구현 완료하였으며, 뉴스 DB 구축 중에 있습니다.

5. 추후 일정

- 6 월 중 기존 이상치 탐지와 다른 이상치 탐지 모형들 개발 후 모형 최적화를 뉴스데이터와 함께 진행할 예정입니다.

- 6 월 1 - 2 주 : 크롤링 마무리 및 뉴스 데이터 전 처리

- 뉴스 요약 모델(남은 방학 기간 동안 모델 개발 및 Tableau 공부)

- 구글 딥러닝 기술 **"bert"** 적용 예정

  - 최종 시각화 및 추가 서비스 개발 완료(최종 마무리)

Bert 란?

구글에서 개발한 NLP(자연어처리) 사전 훈련 기술이며, 특정 분야에 국한된 기술이 아니라 **모든 자연어 처리 분야에서 좋은 성능을 내는 범용 Language Model** 입니다.

한국어 버전의 BERT 인 KoBERT 모델이 있는데, 이 모델을 활용하면 더 효과적으로 task 를 완료할 수 있을 것이라고 생각합니다.