1월 월간 보고서

1. 기간 : 2020년 1월 2일 ~ 2020년 1월 31일

2.장소 : DNA 본사

1. OHLC를 활용한 LSTM 모델 생성.

- 단순 **Close**만 사용한 모델보다, **OHLC**를 전부 사용한 모델이 정확도면에서는 성능이 좋은 것으로 나타남. OHLC를 그대로 사용할경우 **Correlation**이 매우 높아 **다중공선성**의 문제 등이 우려됨. 따라서 **PCA**나 **새로운 지표**들을 통하여 다중공선성 문제 해결이 필요해 보임.

2. DGX Station을 활용한 모델 심화 구조

- **DGX**를 사용하면, 심화된 구조의 딥러닝 모형을 사용할 수 있고, 빠른 학습시간이라는 장점이 있음. 하지만, 성능이 기존의 다른 모델들에 비해 현저하게 뛰어난 것은 아님.
- 일반적으로 요리사가 뛰어난 것보다는 재료 자체가 좋은 식재료이면 기본은 하듯이, DGX를 활용한 모델 최적화도 중요하지만, **데이터가 더 중요한 것**으로 여겨진다.
- 또한 시계열 모델에서 성능이 뛰어나다고 알려진 LSTM이나 GRU 모형을 사용하는 것도 좋지만, 단순히 코드 사용법만 이해하는 것이 아닌 **수학적인 알고리즘 방식을 이해하는 것도 모델 최적화를 위해선 중요한 부분**이라 판단 됨.

3. 전체 종목 모델에 대한 이상값 확인

- 전체 2500여개 종목을 하나의 종목으로 간주하고, 통합 모델 생성.
- **5일전 데이터들로 5일 뒤의 종가를 예측**하는 모형.
- KOSPI, KOSDAQ 기준 상위, 하위 10, 1주 당 가격 50만 이상 우량주, 1주 당 가격 1000원 이하 소형주 등을 이상값으로 판단.
- 각 이상값 종목들에 대하여 모델 성능 비교. -> **전체 평균과 큰 차이를 보이는 종목은 없었음.**
- 이렇게 전체 종목에 대하여 모델을 만들 경우 어떤 종목이 오더라도 일반화를 잡을 수는 있지만, 개별 종목에 대하여 핏팅한 모델보다는 성능이 조금씩은 떨어진다.

4. 중간 과정 정리

- **4.1** 중간 과정 정리
- **4.2 향후 모델 진행 방향**
  - **4.2.1 Regression -> Classification** : 단순 주가 예측에서 -> 변화율을 분류
  - **4.2.2 Feature Engineering** : 각 종목별 특성을 반영할 수 있는 Feature 추가 ex) 시장구분(코스피,코스닥), 거래정지여부, 관리구분, 락구분, 시장경보구분, 불성실공시지정여부, 증거금, 신용증거금, ETF, 소속
  - **4.2.3 Feature Engineering** : 기술적 지표 구현
  - **4.2.4 Algorithm Optimization** : 다양한 기법의 머신러닝, 딥러닝 알고리즘 적용 및 최적화

업무 5 : 환율 프로젝트

환율 관련 프로젝트

- **5.1. DB 구축**
  - 한국수출입은행의 Open API를 활용하여 2010년 ~ 2019년까지의 환율 데이터 DB 구축.
  - DB Code
- **5.2 EDA**
  - 탐색적 데이터 분석을 통한 데이터 특성 파악.
  - EDA Code
- **5.3 Many to One Model**
  - 선형회귀를 이용해 many - to - one Modeling
  - 예측 시점이 뒤로 갈수록 RMSE가 매우 높아짐.
  - INPUT 5 DAYS -> FEATURE 1 DAYS
  - INPUT 10 DAYS -> FEATURE 5 DAYS
  - INPUT 20 DAYS -> FEATURE 20 DAYS
  - INPUT 120 DAYS -> FEATURE 120 DAYS
- **5.4 Virtual Data Model**
  - 가상 데이터를 사용한 Modeling
  - ex) 2019년 10월 ~ 2019년 12월의 60개의 데이터를 이용하여 2020년 1월 1일을 맞춤. 2019년 10월 ~ 2019년 12월의 59개의 데이터와 가상의 1월 1일 데이터를 이용해 1월 2일을 맞춤. 2019년 10월 ~ 2019년 12월의 58개의 데이터와 가상의 1월 1일, 2일 데이터를 이용해 1월 3일을 맞춤. ..... 이렇게 60일까지 예측
  - 1년까지 가기에는 성능이 매우 떨어지나, 3개월(시장일 기준 60일) 정도는 흐름을 파악할 수 있음.
  - Virtual Model
- **5.5 Many to Many Model**
-
  - 과거 60개의 데이터를 이용해 1~60일을 각각 60개의 모델로 예측
  - 10일이 지난 시점부터 예측력이 현저하게 떨어짐.
  - 1 ~ 60 Model
- **5.6 예정**
  - (예정) PCA 진행
  - (예정) 알고리즘 변경 : LightGBM, XGBoost
  - (예정) 하이브리드 신경망 (SVM + NN)
  - (예정) 데이콘 우승자 코드 확인
  - (예정) 오토인코더
  - (예정) Self-Attention
  - (예정) 모든 데이터를 전일 대비 등락률로
  - (예정) CNN, DNN 적용 고민...
  - (예정) 주기를 변경...일주일 10일 2일 5일...
  - (예정) 모델은 최대한 빠르고 쉽게 단층 구조...