**FX Predict**

1조

1. 팀 프로젝트 주제와 서비스

최근 3년간 코로나 19, 우크라이나 사태 등 경제적 불안정한 요소들로 인해 안전 자산 중 하나에 해당하는 달러의 수요가 급격하게 증가했다. 이에 발맞춰 미국 연방도 금리를 올려 달러 가치를 더더욱 높이고 한화의 가치는 상대적으로 더 낮아졌다. 지속적인 미국의 금리인상과 인플레이션은 우리나라 원 달러 환율을 크게 증가시켰고, 2022년 끝내 2009년 금융위기 수준인 1300원을 넘기며 많은 경제 주체들에게 두려움을 남겼다.

원 달러 환율의 지나친 상승은 국내 경기의 위협으로 이어진다. 소비자 측면에서는 국내 물가가 오르고, 기업과 국가의 해외 부채 부담이 증가하며, 국 내외 투자자들이 국내에 투자한 자금을 회수하면서 경기가 전반적으로 위험에 빠진다. 따라서 환율의 급격한 인상을 방지하기 위해 정부는 금리나 통화량의 조절을 통해 환율 조정에 개입한다.

정부가 개입하는 만큼 환율은 시장의 원리 그 자체로 예측하기가 어렵다. 그럼에도 불구하고 경제주체는 환율이 급격한 상승 또는 하락에 대비하기 위해 환율을 미리 예측하기를 원한다.

그동안의 환율예측은 환율의 종가, 고가, 저가, 오픈 등 환율 자체만의 정보를 통해 이루어졌다. 하지만 우리의 프로젝트는 기존 환율 예측과 다른 방향으로 원 달러 환율을 예측하고자 한다. 첫째로, 다양한 경제 지표를 사용하여 환율을 예측하는 것이다. 단순히 환율의 추이로 환율을 예측하는 것이 아닌, 환율에 영향을 미치는 유의미한 지표를 찾아 이를 반영하고자 했다. 둘째, 서비스 이용자에게 매일 업데이트되는 환율 예측치와 추세를 제공하고자 했다.

1. 데이터 수집 및 전처리

우선 환율과 유의미한 지표를 찾기 위해 환율 스터디를 시작했다. 환율 기초 지식을 공부하고, 논문 약 35개와 각종 경제 신문기사를 참고하여 환율과 관련성 있는 지표들을 가져왔다. 경제적 요인으로는 물가지수인 GDP, 경제성장률. 정부지출, 국민총소득, 통화자료, 무역수지, 자본수지, 이자율 등의 정보를 가져왔고, 정치적 요인으로는 VIX, 주택 가격 지수, 뉴스 심리지수, 소비자 동향, 달러지수, 집권 정당 등을 조사했다. 추가로 코스피, 코스닥, 외국인 지분율 등 주요 주식시장 지표와 WTI 크루드 오일 가격과 금 가격 등의 주요 원자재 정보, 그리고 통화량 정보 등도 유의미하다 생각하여 가져왔다. 총 84개의 변수를 가지고 왔으며 주로 Fred, investing.com, 한국경제통계시스템 등의 사이트에서 정보를 얻었다.

엑셀 파일로 저장된 변수들의 상관관계 파악을 위해 모든 정보를 날짜별로 결합할 필요가 있었다. 따라서, 데이터의 날짜 정보를 통일하기로 했다. 우선 가져온 정보들이 날짜 별, 월별, 쿼터별로 빈도가 다양했기 때문에, 기준점을 정할 필요가 있었다. 쿼터 단위 자료를 모두 일별 자료로 늘리는 것은 큰 의미가 없을 것이라고 판단하여 월별 단위로 모든 데이터를 결합하기로 결정했다.

월별로 데이터를 결합하려면 우선 일별 단위의 데이터와 쿼터 단위의 데이터를 월별 단위로 변경하는 작업이 필요했다. 따라서 일별 자료는 평균을 내어 월별 자료로 변경하고, 쿼터 단위의 자료는 결측 치를 보간법으로도 채우고, 기존 데이터를 한 쿼터에 해당하는 세 달에 똑같이 붙여 넣는 방법으로도 채웠다. 추후에 두 방법으로 결합한 데이터를 각각 선형모델에 넣어본 결과, 후자의 방법이 선형모델에서 더 좋은 결과값을 도출했다. 따라서 후자의 방법을 선택하기로 결정했다.

월별 단위로 각각의 84개의 변수들을 맞춰준 후, 이를 하나의 데이터셋으로 병합했다. 이 과정을 수행하기 위해 날짜 형식을 통일할 필요성이 있었다. 날짜 형식은 “년-월-일”, “년 월 일”, “년.월,일”등 다양했다. 이를 “년-월” 형식으로 통일하고 데이터 셋을 결합했다. 또한 일부 데이터에서 숫자에 ‘,’ 표시가 들어있어서 숫자가 아닌 문자로 인식되는 경우가 있었다. 따라서 ‘,’를 제거해주는 작업도 진행했다.

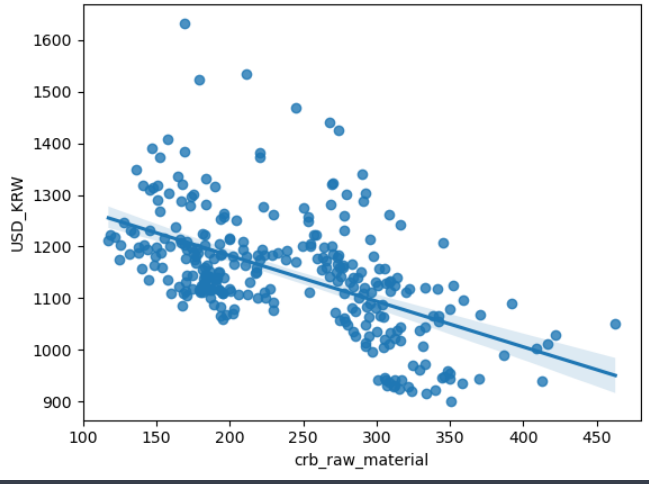
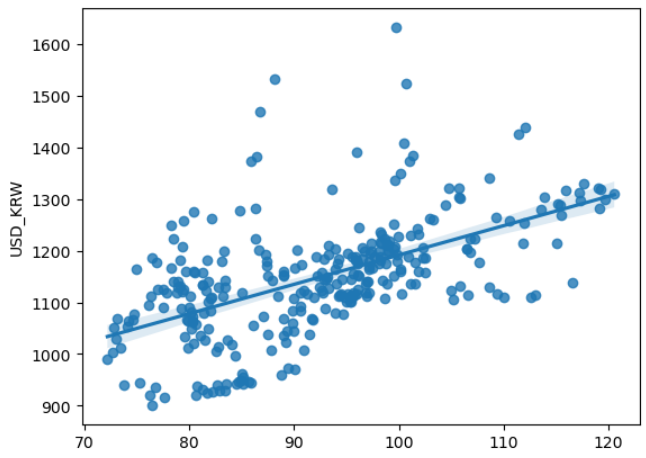
1. 변수들의 상관관계 파악

데이터 셋을 병합한 후 선형 회기 분석을 통해 변수들의 상관관계를 살펴보기로 했다. 사용했던 회기 분석 모델은 Linear Regression, Random Forest Regressor, Lasso, Ridge, Elastic 등 약 5개의 모델이 있다. 처음에는 회기 분석 모델들을 돌리면 가중치 값과 변수의 중요도가 나오기 때문에 이를 바탕으로 가장 유의미한 변수를 채택하여 LSTM 모델에 넣어보고자 하였다. 하지만 모델들마다 가중치 값과 중요도가 상이하게 나왔다. 따라서 선형 회기 모델을 보는 것으로는 상관관계를 알아보기에 역부족이라고 판단하였다.

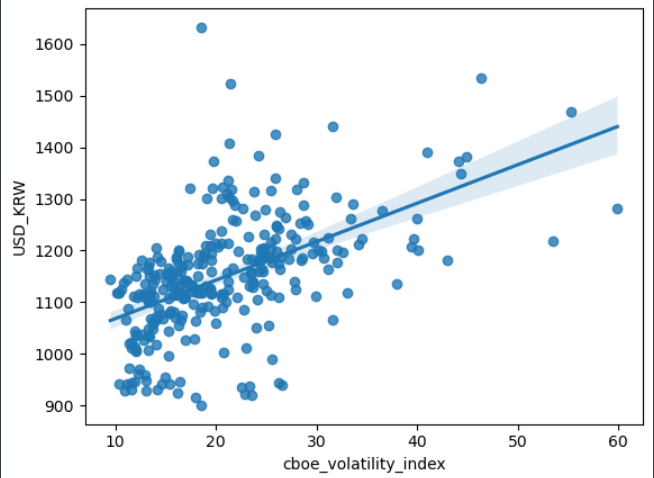
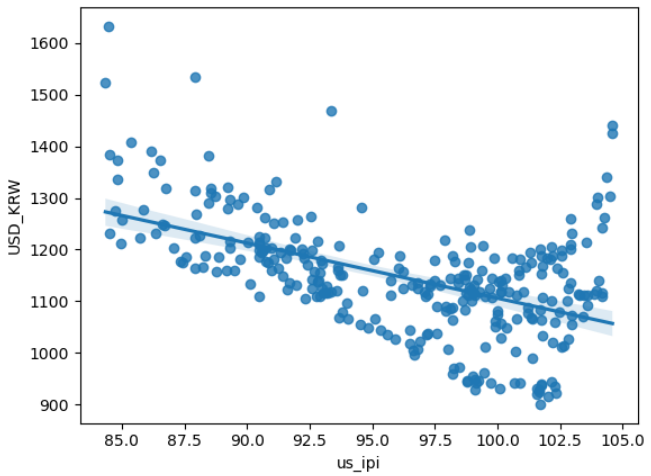
이후 corr() 함수를 통해 나온 상관계수의 값을 heatmap을 통해 그려서 변수들 간의 상관계수를 확인했으나, 선형 회기 분석과 마찬가지로 단순히 자료들 간의 관계를 최대한 선형으로 그려서 나온 상관계수 값으로 상관성을 판단할 수 없다는 사실을 알게 되었다. 따라서 가지고 있는 모든 84개의 변수를 원 달러 환율과 비교하여 시각화를 해보기로 하였다.

regplot()을 사용하여 84개의 변수 모두를 시각화 해보았다. 의외의 결과가 나왔는데, 흔히 생각하는 금값, 이자율, 금리 등의 경제 지표들은 원 달러 환율과 관련이 없는 형태로 나왔다. 추론해본 결과, 거시경제에서 변수들이 서로 영향을 주고받는 데에는 일정한 시간이 소요되기 때문에, 동시간으로 결합된 데이터 셋으로 상관관계를 나타내기 어렵다고 추론했다. 또한, 한 가지 요소 이외에 여러 가지 요소가 종합적으로 환율에 영향을 미치기 때문에, 단순히 한가지 지표가 환율과 선형관계를 이루기가 쉽지 않다고 생각했다. 마지막으로, 환율 자체는 국가의 정책에 영향을 많이 받기 때문에 케인즈 이론 등 여러 경제 이론에서 설명하는 변수들이 완전히 설명되기란 쉽지 않을 것이라고 추론했다.

84개의 변수 중 환율과 유의미한 상관관계를 보인 변수는 총 네 개이다. 유의미한 상관관계를 판단할 때, 변수와 원 달러 환g율 간의 산점도와 선형 관계를 그래프로 나타냈을 때, 산점도가 선형 그래프를 기준으로 큰 분산을 그리지 않고 그려지는지를 중점적으로 보았다. 총 네개의 변수는 다음과 같다.



<달러지수> < 원자재지수(crb)>

 <미국 산업 생산자 지수> <VIX>

위 네 가지 변수 중 미국 산업 생산자 지수를 제외한 나머지 세개의 변수들은 일별 데이터 확보가 가능하기 때문에, 일별데이터를 확보하여 LSTM 모델을 학습시키기로 결정하였다. 이후, LSTM 모델을 학습시키면서, VIX지수를 제외하고 난 후 LSTM 모델의 성능이 더 좋아진다는 것을 깨달았다. VIX 그래프 분포를 살펴보면, 분포가 고르지 않고 앞쪽으로 몰려 있기 때문에 원 달러 환율을 예측하기에 적합하지 않다고 판단하였다. 따라서 최종 변수로 달러지수와 원자재지수를 선택하기로 하였다.

원자재 지수가 원달러 환율과 높은 상관관계를 가진다고 예상하지 못하여 crb에 대해 조사한 결과, 원자재지수(crb)는 원유, 천연가스, 귀금속 등 19개의 주요 상품 수요와 인플레이션 일부를 반영하는 지수이기 때문에, 국가 간의 상대적인 물가를 비교하여 수치로 나타내는 ‘환율’과 유의미한 상관관계가 있을 수 있다고 판단하였다.

1. 환율 예측 모델

4-1) 변수 크롤링 및 데이터 병합

가장 먼저 환율 예측 모델에 들어갈 정보인 원 달러 환율, 달러지수, 그리고 원자재지수(crb) 자료를 investing.com 에서 1일 1회 크롤링한다. 크롤링한 정보를 기존 데이터셋과 병합하여 하나의 큰 데이터 셋을 완성한다.

4-2) LSTM 모델 파라미터 조정

LSTM모델은 어떤 파라미터 값을 주는가에 따라 모델의 성능에 큰 차이가 난다. 따라서 최적의 파라미터를 찾는 것이 중요하다. 먼저, 학습 데이터를 전체 데이터셋의 90프로를 줌으로써 모델이 비교적 최근 데이터까지 학습할 수 있도록 하였다. 또한, 너무 많은 과거 데이터를 학습하거나 먼 미래를 맞출수록 모델의 성능이 떨어진다는 것을 알게 되었다. 따라서 과거 10일 데이터를 학습한 후 1~2일의 예측 값을 보는 것이 가장 적합하다고 판단했다. 따라서 이용자는 향후 몇일의 예측치를 볼 것인지 입력할 수 있으나, 먼 미래를 볼수록 오류가 증가하므로, 하루나 이틀치의 예측치를 입력하기를 권장한다.

4-3) 데이터 전처리

4-1) 에서 완성한 데이터 셋의 평균과 표준편차가 차이가 많이 나기 때문에 데이터셋의 어깨를 맞추기 위해 표준화를 해준다. 또한 ‘split data’라는 사용자 정의 함수로 학습 데이터와 평가데이터를 각각 전체 데이터셋에서 과거 데이터 (10)일과 미래 데이터 (1~2일)를 인덱싱하여 분리해준다.

4-4) LSTM 모델 구축 및 학습

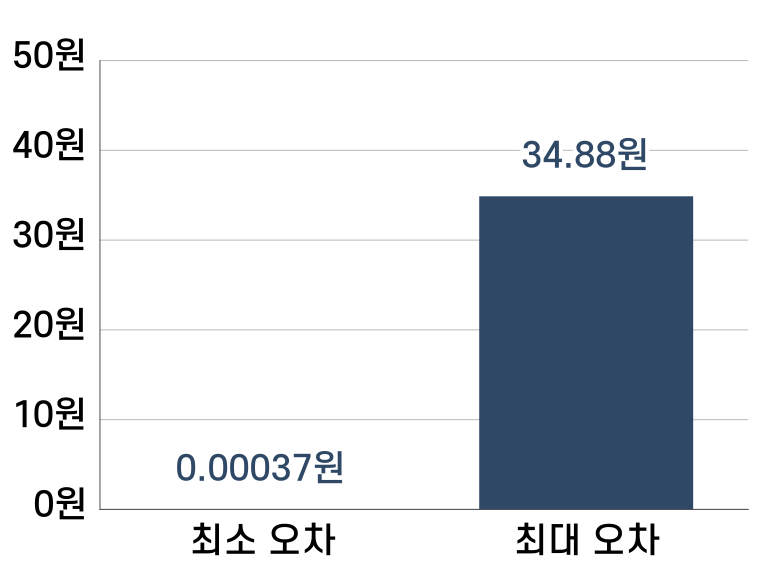
LSTM을 구동할 때, 학습 데이터에 지나치게 최적화되는 과적합을 방지하고, 변수들의 최적의 가중치 값을 저장하고자 했다. 따라서 EarlyStopping을 통해 검증데이터의 오차가 커지는 순간 모델의 학습을 멈추는 제어를 걸어주었다. 하지만 검증 데이터의 오차가 커진 이후에도 검증데이터의 오차가 더 감소할 가능성이 있으므로, patience값을 주어서 학습이 바로 멈추지 않고, 일정 기간 모델이 더 학습할 수 있도록 해주었다.

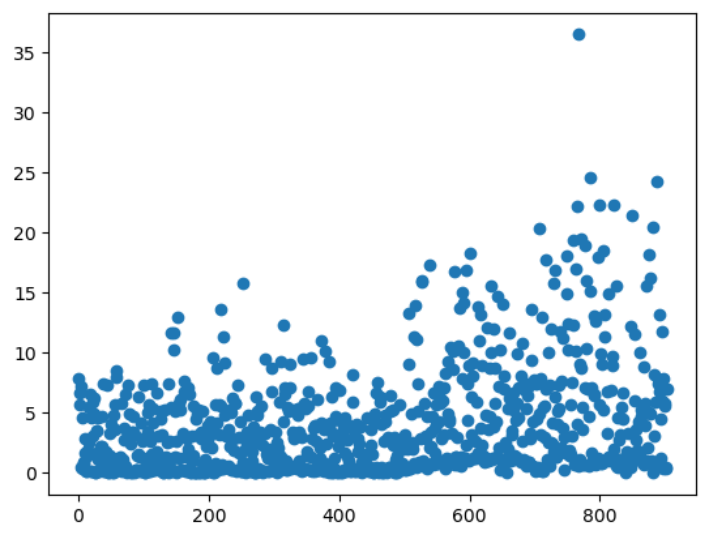
또한 LSTM을 구동할 때, ModelCheckpoint를 통해 검증데이터의 오차가 최소가 될 때 변수들의 가중치 값을 지속적으로 업데이트 하여 저장할 수 있도록 하였다.

4-5) 예측

학습이 완료된 모델에 저장한 최적의 가중치를 재 로드 한 후, 이 모델을 바탕으로 다음날 혹은 이틀 후의 환율까지 예측하도록 하였다. 예측치를 도출하기 위해 최근 10일간의 달러지수, 원자재지수(crb) 그리고 원 달러 환율 데이터를 표준화를 한 후, 3차원 배열로 만들어 모델 안에 넣어준 후, 표준화를 다시 해제해주면 예측치를 얻을 수 있다.

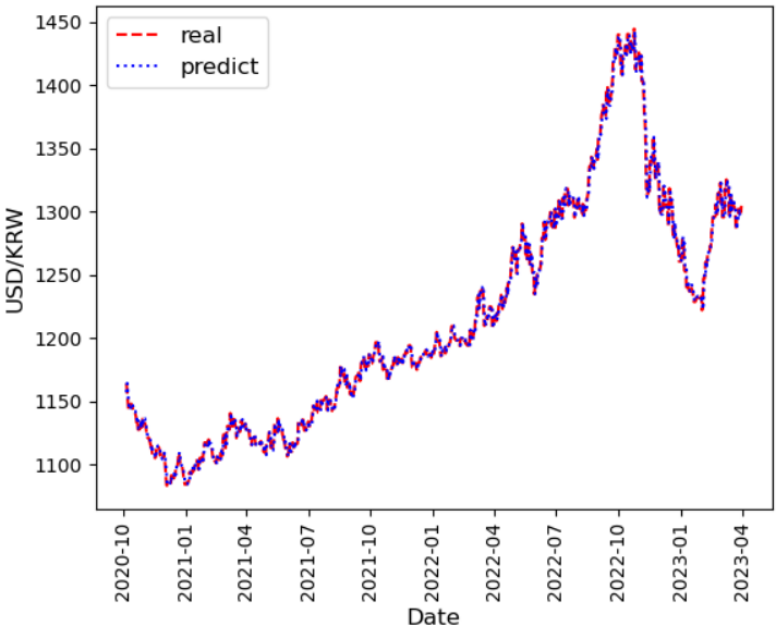
4-6) 오차 및 성능 평가

 모델의 성능을 평가하기 위하여 모델 파라미터를 조정할 때 마다 검증데이터의 오차 평균을 구했다. 우선 전처리 된 검증데이터의 독립변수를 model.predict 안에 넣어서 예측치를 산출한다. 산출된 예측치와 실제값의 차이를 구해 평균을 낸 후 표준화를 해제해주어 오차가 약 몇 원이 나오는지 살펴보았다.



평균적으로 약 3.8원의 오차가 났으며, 최대 오차가 약 34원, 최소오차가 약 0원 정도의 성능을 보였다.

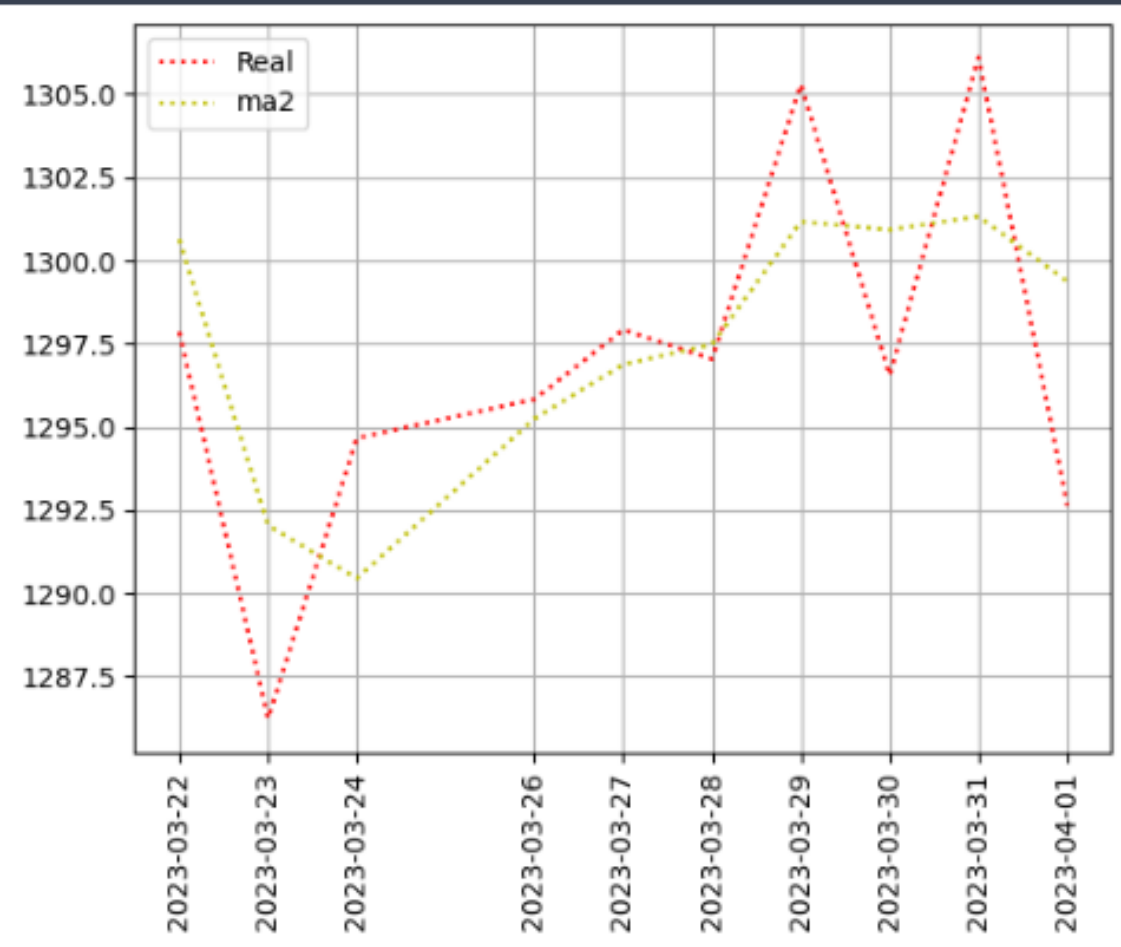
검증데이터의 예측값과 실제 값을 시각화 한 결과는 다음과 같다:



1. 이동평균선

적은 오차로 환율을 예측한다고 해서 환율의 상승과 하락 등 추세를 정확하게 예측할 수 있다고 할 수 없다. 따라서 보다 정확한 추세를 예측하고자 이동평균선을 추가로 그리기로 했다. 본 모델은 약 1~2일의 단기 환율의 예측하기 때문에, 실제 환율과 LSTM에서 구한 예측 값을 포함한 자료를 바탕으로 2일의 단기 이동평균선과 실제 환율을 그래프로 그려서 환율의 추세를 구하도록 했다.

통상적으로 단기 이동평균선과 실제 값 선이 만나는 지점에서 추세가 변동한다. 따라서 예측날과 그 전날의 데이터를 살피어 단기 이동평균선과 실제 선이 만나는 지점을 찾도록 했다. 하지만 이틀 간의 데이터로 변동하는 지점을 발견하기에 부족하다고 판단하여 코드로는 예측 날로부터 약 3일 전까지 살피어서 단기 이동평균선과 실제 환율 선이 교차하는 지점을 발견하도록 구현했다.



위 그래프와 같이 사용자가 원하는 기간만큼의 단기 이동평균선과 실제 환율선을 시각화 할 수 있도록 코드로 구현하여 보다 정확한 추세를 알 수 있도록 하였다.

1. 성과 및 보완점

본 프로젝트는 단순히 환율 데이터의 추세로 환율을 예측하는 것이 아닌 다양한 변수를 토대로 보다 정확히 환율을 예측했다는 점에서 기존 환율 예측 모델과 차이점이 있다. 또한, 환율 예측치와 함께 환율 추세 또한 시각적으로 알려줌으로써 서비스 이용자로 하여금 필요한 환율 정보를 제공한다.

2차 프로젝트에서는 LSTM 이외에 다른 예측 모델들도 만들어 서로의 오차를 보완하게끔 하여 더 정확한 환율 예측치를 제공하고자 한다. 또한 단순 이동평균선 이외에도 최근 데이터에 더 많은 가중치를 주는 가중이동평균선 등을 추가로 그려 더 정확한 환율 추세를 예측하고자 한다.