**검색 변화량과 주식지표를 활용한 주가 예측**

팀명 : Ajou 캐피탈

조원 : 김동현, 이은주, 전상범, 조규한, 홍소연

1. **프로젝트 개요**
   1. **문제 제시**

주가에 영향을 미치는 요소들에 대해 과거의 주식 관련 수치들 뿐만 아니라 그 외의 값도 영향을 줄 수 있다면 어떤 것이 있을 지에 대한 논의를 해보았다. 그러던 중 검색량이 주가와 영향이 있을 것이란 판단을 했고, 주 단위로 생성된 주가 지표와 검색량을 통해 다음 주의 주가 예측치를 실제 예측치와 비교하여 그 영향력을 보고자 하였다.

* 1. **목표**

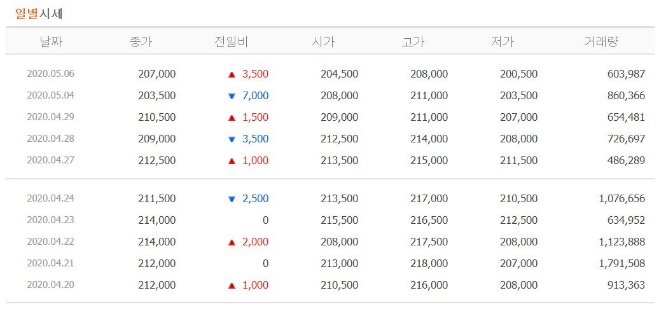
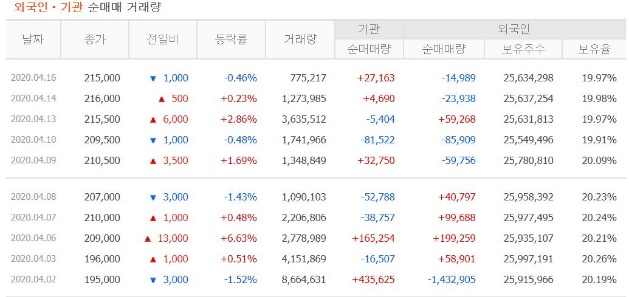
한 주 전의 주가 지표만 가지고 주가를 예측했을 때와 주가 지표에 검색량이 추가되었을 때의 주가 예측 중 무엇이 더 정답에 가까운지 알 수 있다. 또한 어떤 검색 지표가 영향을 많이 끼쳤는지 확인하고, 다양한 검색 방법(웹-구글, 유튜브, 네이버) 중 어떤 검색 방법이 특히 많이 영향을 끼치는지 탐색해본다. 이 과정에서 어떤 알고리즘이 가장 정답에 근접하게 예측해내는지도 확인해볼 수 있다.

* 1. **계획**

우선 주가 관련 데이터들을 얻어내기 위해 네이버 금융에서 시가총액이 높은 기업들의 2018년 정해진 기간의 주가 지표와 검색량을 총 6가지 분석 방법(multiple linear regression, Ridge Regression, SVR, Bagging, CNN, LSTM)으로 데이터 분석을 진행하고, 검색량 3가지의 포함여부에 따른 test score 분석 후 가장 높은 알고리즘과 포함 검색량을 뽑은 후 2020년 5월 주가에 적용해본다.

1. **데이터 분석**

먼저 다음 주 종가 예측을 위한 주가 지표로는 주어진 기간(2018.1.2~2020.4.30. 122주동안)동안의 각 기업의 한 주간 시가, 종가, 평균거래량, 평균가(=(시가 + 종가)/2), 외국인 순매수량, 기관 순매수량을 선택하였다. 주식시장이 월요일부터 금요일까지 열리는 관계로 5일 동안 각 지표마다 평균값을 사용하기로 하였다. 평균거래량, 외국인 순매수량, 기관 순매수량은 가 있었기 때문에 지표로 선택하였다. 이 지표들은 네이버 금융에서 제공하는 데이터를 크롤링하여 원하는 지표들 만을 추출하였다.

**그림 1. 셀트리온 시세 (네이버 금융) 그림 2. 셀트리온 외국인.기관 순매수량 (네이버 금융)**

검색량 지표로는 Google Trend에서 추출한 각 기업의 주어진 기간 동안의 구글 웹 검색량, 유튜브 검색량과 네이버 애널리틱스에서 추출한 네이버 웹 검색량을 사용하기로 하였다. Google 트렌드와 네이버 애널리틱스에서는 122주 중 가장 검색량 많았던 주의 검색량을 100으로 하였을 때의 상대적인 검색량만을 얻을 수 있었다.

스크린샷이(가) 표시된 사진

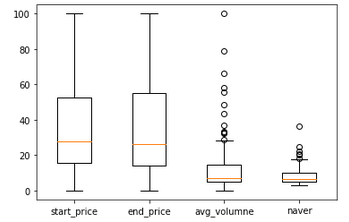
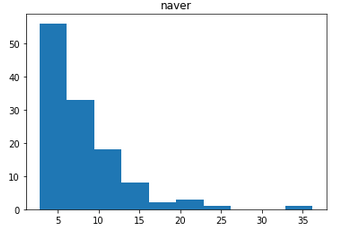
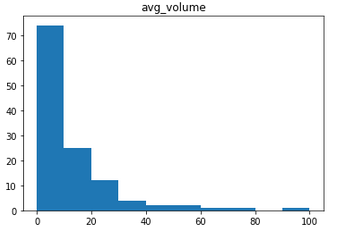
자동 생성된 설명 

**그림 3. 셀트리온 검색 구글 트렌드 그림 4. 셀트리온 csv**

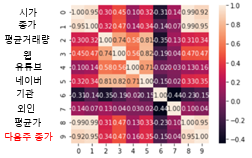
본 프로젝트에서는 2020.6.6 기준 시가총액 상위 10개 기업(현대차, 카카오톡, NAVER, 삼성전자우, 삼성전자, 셀트리온, SK하이닉스, LG화학, 삼성바이오로직스, 삼성SDI)을 종가 예측에 사용할 기업으로 선정하였다. 시가총액 상위 10개 기업들인 만큼 위에서 선정한 주가 지표와 검색량 지표에서 충분한 데이터를 얻을 수 있을 것으로 예상하였다.

각 기업의 데이터 분석을 위한 EDA를 진행하였다. 다음은 그 예시로 셀트리온에 대한 boxplot, histogram과 각 feature의 상관계수를 나타낸 heatmap이다.

**그림 5. 셀트리온 Boxplot, histogram**



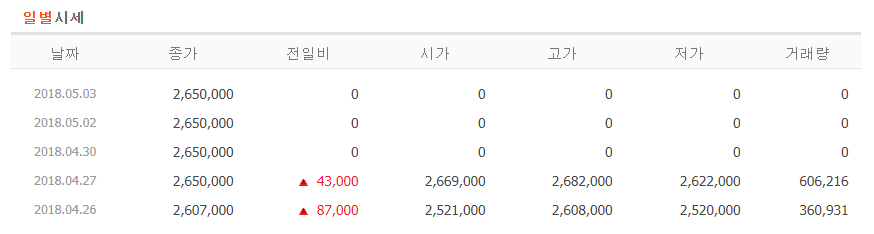
Boxplot과 Histogram을 통한 데이터 분석결과, 평균거래량과 네이버 검색량은 특정 날짜에 폭등하는 경향을 확인할 수 있었다. 하지만 boxplot에서 이상치로 나온 데이터 또한 주가 변동에 영향을 주는 의미있는 데이터라고 생각하였기 때문에 이상치를 제거하거나 대체하는 작업은 하지 않았다.



**그림 6. Feature 간 상관계수 Heatmap**

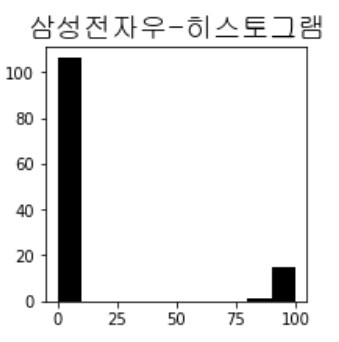
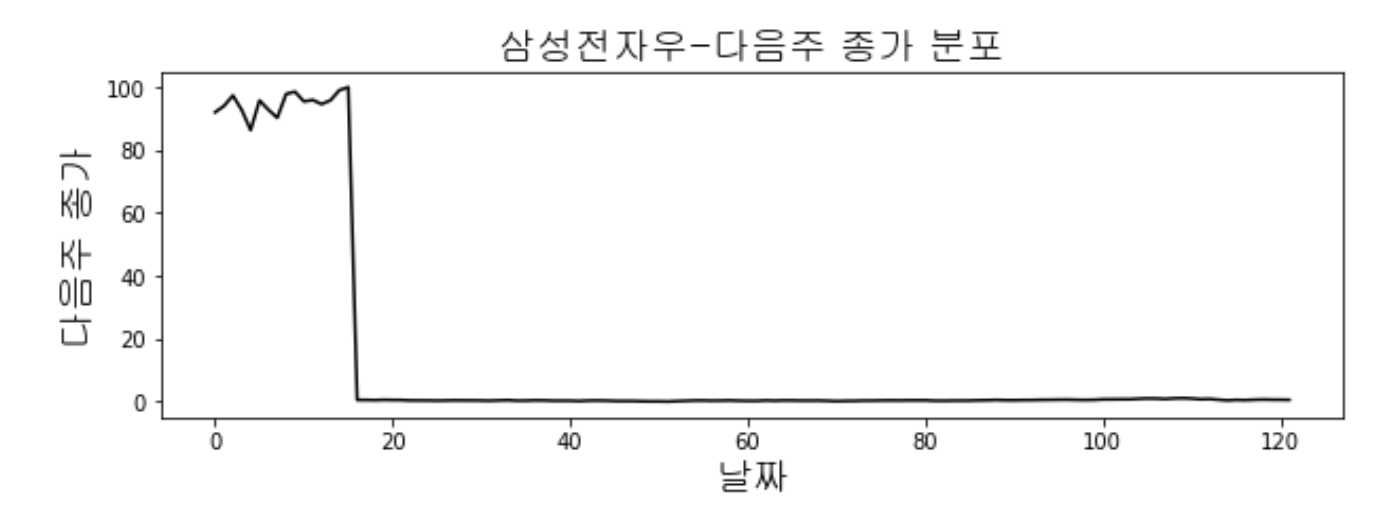
Heatmap 확인 결과, 셀트리온의 target date인 다음 주 종가와 가장 상관관계가 높은 feature는 이번 주 종가(0.95), 평균가(0.95), 시가(0.92), 구글 웹 검색량(0.47), 네이버 검색량(0.35), 평균거래량(0.34), 유튜브 검색량(0.16), 기관(0.15), 외인(0.04) 순이었다. 구글 웹 검색량과 네이버 검색량이 영향을 많이 끼칠 것으로 예상했던 평균거래량과 기관, 외인 순매수량보다 높은 것을 확인할 수 있었다.

1. **데이터 전처리**

**3-1. Data 전처리**

**그림 7. 삼성전자 2018년 4월 시가**

그림7의 전일비, 시가, 고가, 저가, 거래량이 모두 0으로, 주가 분할로 인하여 3일동안 거래가 정지되어 데이터 변화가 없다. 또한 이 시점을 기준으로 주가의 1/5로 거래되어 종가의 변동이 커진다. 특히 시가총액 2위 기업인 SK하이닉스와 시가총액이 무려 5배 차이가 났고 다른 선정 기업들에 비해 대략 10배 가량 차이가 났기 때문에 삼성전자 기업의 데이터는 삭제하여 총 9개의 회사에 대해 연구하기로 한다.



**그림 8. 삼성전자우 다음주 종가 분포**

2018년 4월 ~ 2018년 5월 변동 폭이 크기 때문에 2018년 4월까지의 데이터를 삭제한 뒤 분석을 진행한다. Log 스케일에 관해서도 생각해 보았으나 시계열 데이터이므로 앞부분을 삭제하는 것이 적합하다 판단하여 삭제하였다.

**3-2. Feature scale**

- 구글 트렌드

웹, 유튜브 데이터 : 수치는 특정 지역 및 기간을 기준으로 차트에서 가장 높은 지점 대비 검색 관심도를 나타낸다. 값은 검색 빈도가 가장 높은 검색어의 경우 100, 검색 빈도가 그 절반 정도인 검색어의 경우 50, 해당 검색어에 대한 데이터가 충분하지 않은 경우 0으로 나타난다.

- 네이버 검색량

네이버 사이트 상의 조회수를 구글 트렌드에서 얻은 데이터 형식과 동일하게 표현한다.

- 주가 지표

시가,종가, 평균거래량, 평균가 : 시가, 종가와 평균거래량의 범위 차이가 너무 컸기 때문에 min-max scaler를 이용하여 범위를 0~100 사이로 맞추었다.

기관, 외인 순매수량 : 기관과 외인이 매도량이 더 큰 경우 (-)값을 가진다. 위의 주가지표처럼 min-max scaler를 이용할 경우, 매도량이 더 큰 경우를 구분할 수 없기 때문에 N(0,1)을 따르는 정규분포로 값들을 scaling 하였다.

다음은 현대차의 feature scale을 거친 후의 데이터프레임이다.



**그림 9. 현대차 데이터프레임**

1. **모델링**

Regression data split 방법 : Train data와 Test data를 9 : 1로 나누었다. 데이터가 크지 않아 5-fold cross validation을 이용해 하이퍼파라미터 조정을 하였다.

NN data split 방법 : 시계열 데이터의 특징을 이용하기 위해 Train data와 Test data를 시간 순으로 9 : 1로 나누었다.

**4-1. Multiple linear regression**

Multiple linear regression은 종속 변수 y와 한 개 이상의 독립 변수 X와의 선형 상관 관계를 모델링하는 회귀 분석 기법이다. 프로젝트 설명 변수로써 기존의 주가를 예측하는 대표적인 특성인 시가, 종가, 평균거래량, 기관 등락률, 외인 등락률, 시가와 종가의 평균가 6가지 feature에 더불어 ‘검색량’ 이라는 feature가 다음주 종가를 예측하는 정도를 살펴보기 위해 multiple linear regression을 학습 모델로 선정하였다. MLR을 통해 구글 웹, 유튜브, 네이버 검색량이 각각 다음주 주가 예측에 얼마나 영향을 미치는지 확인해 볼 수 있다.

**4-2. Support vector regression**

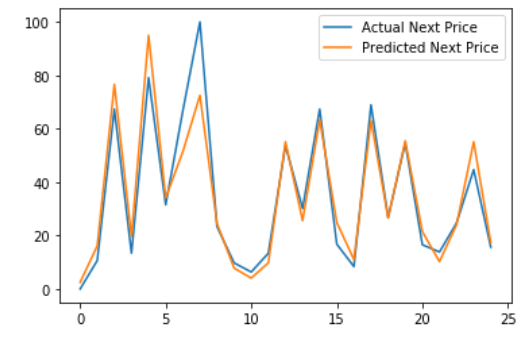
Support vector regression은 Support vector machine의 Regression용 알고리즘이다. Multiple linear regression과 비슷하게 종속 변수 y와 한 개 이상의 독립 변수 X와의 상관 관계를 모델링하는 회귀 분석 기법이다. 이 알고리즘에서 데이터를 고차원 공간에 매핑하기 위해 커널 기법이라는 것을 사용하는데, 실제로 데이터를 확장하지 않고 확장된 특성에 대한 데이터 특성에 대한 데이터 포인트들의 거리를 계산한다. 이번 프로젝트에서는 원래 특성의 가능한 조합을 지정된 차수까지 모두 계산하는 다항식 커널과 가우시안(Gaussian) 커널이라고도 불리는 RBF(radial basis function) 커널, 그리고 linear 커널 이렇게 3가지 커널을 비교한 후 가장 정확도가 높게 나온 커널로 test set에 적용했다.

**4-3. Bagging regression(using Lasso)**

Bagging regressor는 기본이 되는 단일 알고리즘에서 서로 다른 샘플링이 된 dataset을 학습한 뒤, 이에 대한 평균을 산출해내는 방법이다. 단일 model을 활용하여 예측했을 때 보다 분산을 줄일 수 있는 효과를 볼 수 있어 성능을 향상시킬 수 있다. 기본 모델로 사용한 Lasso 모델은 비선형성 문제와 다중 공선성 문제를 해결하는 모델로, L1 norm을 패널티로 사용한다. 높은 예측률과 쉬운 해석력을 가지고 있어 선택하였다. Lasso의 하이퍼파라미터 alpha값과 샘플링을 얼마나 진행할 것인지 n\_estimator는 cross validation 을 이용하여 score가 가장 높게 나온 값으로 설정하였다. 9개의 기업 대부분은 alpha = 0.005, n\_estimator = 10(default) 으로 나왔고, 일부 기업은 n\_estimator = 100으로 설정하였다.

**4-4. Ridge Regression**

Ridge regressor는 평균 제곱 오차를 최소화하면서 회귀계수 벡터의 L2 norm을 제한하는 기법이다. 평균 제곱 오차 뿐만 아니라 L2 norm 또한 최소화하는 것이 Ridge regressor의 목적이다. Lasso regressor에 비해 변수 간 상관관계가 높아도 좋은 성능을 보이면 크기가 큰 변수를 우선적으로 줄이는 방식이다. 단순한 Linear regression 보다 훈련 데이터를 잘 예측하고 추가 제약 조건을 만족시키고자 Ridge regressor를 이용해보았다. Lasso regressor와 마찬가지로 하이퍼파라미터 alpha값과 n\_estimator를 cross validation을 이용하여 score가 가장 높게 나온 값으로 설정하였다. 9개의 기업에서 alpha 값 0.01을 사용하였고, n\_estimator는 10으로 설정하였다. 다음은 r2\_score가 0.904가 나온 셀트리온의 test set의 실제 가격과 예측 가격 그래프이다.



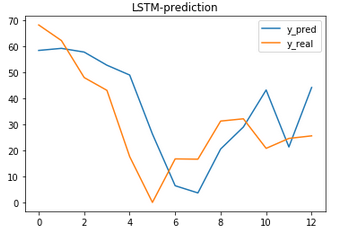
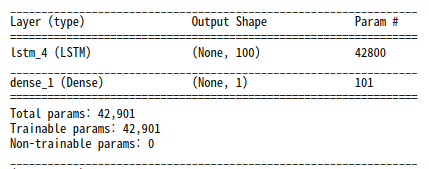
**그림 9. 셀트리온 주가 변동 그래프**

**4-5. LSTM**

LSTM(Long short-term memory)은 RNN에서 vanishing gradient problem을 극복하기 위해 고안된 알고리즘이다. 본 연구에서 매주 주가 데이터와 검색량 데이터로 구성된 시계열 데이터를 사용하였기에 다른 알고리즘보다 RNN 기반의 알고리즘을 적용하면 더 좋은 예측이 가능할꺼라 예상하여 LSTM을 적용해보았다.

시계열 데이터인 점을 이용하려면 여러 주(week)의 데이터를 한번에 input으로 넣어야하는데 실제로 해보니 한 주(week)씩만 input으로 넣는 것이 가장 결과가 좋게 나왔다. 또한 output size를 100으로 했을 때 가장 좋게 나왔다. 하지만 가장 좋게 나온 결과도 다른 알고리즘에 비하면 너무 낮아 LSTM을 제대로 활용하려면 훈련 데이터 양이 훨씬 많아야 할 것 같다. 다음은 사용모델과 test 데이터에 대한 모델 적용 결과 그래프이다.

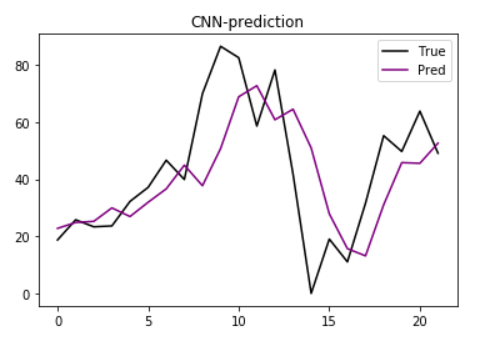
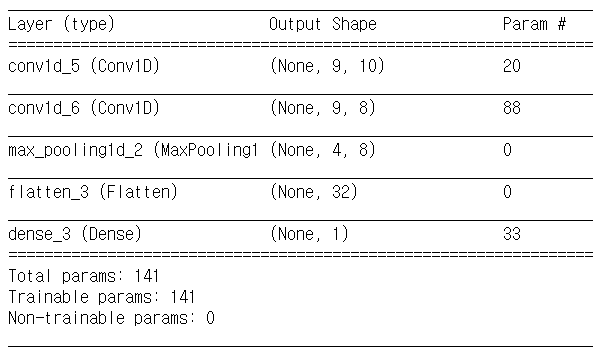
사용모델



**4-6. CNN**

이미지를 인식하고 분류하는데 효과적인 알고리즘으로 알려져 있는 합성곱 신경망(Convolutional Neural Network, CNN)은 여러 분야의 분류 및 예측 문제에 널리 응용되고 있다. 본 연구에서는 시계열로 구성되어 있는 데이터를 입력값으로 사용하여 다음주 종가를 예측하는 예측기로써 합성곱 신경망을 적용하였다. 이는 시계열 데이터를 보고 주가 지수를 정확한 수치로 예측하는 기계학습 알고리즘을 개발하기 위한 과정으로 볼 수 있다. 검색량을 feature로 사용하는지 유무에 따라 더 좋은 예측이 가능한 지 판단해보고자 해당 알고리즘을 선택하였다.

사용 모델



그래프 관찰 결과 비슷하게 예측함을 알 수 있지만, regression 결과보다 R2 score가 0.31로 현저히 작게 나왔다. 모델의 층을 더 쌓아 보았지만 score가 음수가 나오는 경우가 있고 최대 0.34로 비슷한 결과가 나옴을 확인했다.

1. **결과**

**5-1. 회사별 모델 결과**

각 회사별로 5개의 모델을 학습시켜 가장 높게 나온 R2 score의 값과 사용한 Feature를 기록한다.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **사용 Feature** | | | | | | | |
| **All** | **웹** | **유튜브** | **네이버** | **웹+유튜브** | **웹+네이버** | **유튜브 + 네이버** | **Base** |

\* All : 기존 주가 지표 (시가, 종가, 평균가, 평균거래량, 기관, 외인) + 검색량 (구글 웹, 유튜브, 네이버)

\* Base : 기존 주가 지표 (시가, 종가, 평균가, 평균거래량, 기관, 외인)

\* Max : r2 score가 가장 높게 나왔을 때의 사용 feature

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SK하이닉스 | | | | | | | | | | |
|  | **Multi** | | **SVR** | | **Bagging** | | **Ridge** | | **LSTM** | |
| **Max** | **0.9414** | | **0.8563** | | **0.867** | | **0.857** | | **0.548** | |
| **All** | **0.9414** | | **0.8404** | | **0.854** | | **0.857** | | **0.196** | |
| **Base** | **0.8520** | | **0.8517** | | **0.851** | | **0.853** | | **0.167** | |
| 삼성바이오로직스 | | | | | | | | | | |
|  | | **Multi** | | **SVR** | | **Bagging** | | **Ridge** | | **LSTM** |
| **Max** | | **0.9318** | | **0.9613** | | **0.9542** | | **0.952** | | **0.513** |
| **All** | | **0.6294** | | **0.9506** | | **0.9535** | | **0.951** | | **0.479** |
| **Base** | | **0.4081** | | **0.9613** | | **0.9501** | | **0.952** | | **0.363** |
| 삼성전자우 | | | | | | | | | | |
|  | | **Multi** | | **SVR** | | **Bagging** | | **Ridge** | | **LSTM** |
| **Max** | | **0.9987** | | **0.9965** | | **0.9964** | | **0.993** | | **0.254** |
| **All** | | **0.9890** | | **0.9965** | | **0.9913** | | **0.993** | | **0.209** |
| **Base** | | **0.9915** | | **0.9965** | | **0.9952** | | **0.676** | | **0.173** |
| 네이버 | | | | | | | | | | |
|  | | **Multi** | | **SVR** | | **Bagging** | | **Ridge** | | **LSTM** |
| **Max** | | **0.9981** | | **0.9983** | | **0.9979** | | **0.998** | | **0.256** |
| **All** | | **0.9960** | | **0.9981** | | **0.9956** | | **0.996** | | **0.015** |
| **Base** | | **0.9957** | | **0.9983** | | **0.9970** | | **0.998** | | **0.256** |
| 셀트리온 | | | | | | | | | | |
|  | | **Multi** | | **SVR** | | **Bagging** | | **Ridge** | | **LSTM** |
| **Max** | | **0.9564** | | **0.8935** | | **0.8974** | | **0.894** | | **0.394** |
| **All** | | **0.9011** | | **0.8876** | | **0.8895** | | **0.893** | | **0.172** |
| **Base** | | **0.9445** | | **0.8911** | | **0.8782** | | **0-894** | | **-0.143** |
| **LG화학** | | | | | | | | | | |
|  | | **Multi** | | **SVR** | | **Bagging** | | **Ridge** | | **LSTM** |
| **Max** | | **0.8297** | | **0.8692** | | **0.9058** | | **0.882** | | **0.239** |
| **All** | | **0.6873** | | **0.8395** | | **0.8884** | | **0.882** | | **0.058** |
| **Base** | | **0.2202** | | **0.8329** | | **0.7931** | | **0.799** | | **0.115** |
| **삼성SDI** | | | | | | | | | | |
|  | | **Multi** | | **SVR** | | **Bagging** | | **Ridge** | | **LSTM** |
| **Max** | | **0.9915** | | **0.9968** | | **0.9918** | | **0.992** | | **0.157** |
| **All** | | **0.9261** | | **0.9968** | | **0.9897** | | **0.991** | | **-0.130** |
| **Base** | | **0.9851** | | **0.9967** | | **0.9903** | | **0.992** | | **-0.059** |
| **현대차** | | | | | | | | | | |
|  | | **Multi** | | **SVR** | | **Bagging** | | **Ridge** | | **LSTM** |
| **Max** | | **0.9348** | | **0.9275** | | **0.9288** | | **0.929** | | **0.311** |
| **All** | | **0.8750** | | **0.9191** | | **0.9116** | | **0.912** | | **0.101** |
| **Base** | | **0.8829** | | **0.9120** | | **0.9202** | | **0.919** | | **0.311** |
| **카카오** | | | | | | | | | | |
|  | | **Multi** | | **SVR** | | **Bagging** | | **Ridge** | | **LSTM** |
| **Max** | | **0.9487** | | **0.9327** | | **0.9336** | | **0.929** | | **0.327** |
| **All** | | **0.9274** | | **0.9234** | | **0.9295** | | **0.927** | | **0.101** |
| **Base** | | **0.9263** | | **0.9237** | | **0.9280** | | **0.927** | | **0.327** |

위의 결과 표를 살펴보면 Multiple linear regression, SVR, Bagging, Ridge regression 기법을 활용한 모델은 높은 정확도로 다음주 종가를 예측하였으나, LSTM은 예측의 정확도 자체가 높지 않게 나타난다. 대체적으로 기존 주가지표만을 feature로 사용했을 때보다 모든 검색량 정보를 예측 변수로 추가하였을 때 더 높은 예측 정확도를 보이고, 모든 검색량 정보를 feature로 사용했을 때보다 특정 검색량 정보를 예측 변수로 사용하였을 때 더 높은 정확도를 보인다. 특히, SK하이닉스는 구글 웹과 유튜브 검색량, 삼성전자우는 네이버 검색량, LG화학은 유튜브와 네이버 검색량, 삼성 SDI, 현대차, 카카오는 유튜브 검색량이 주가 예측에 많은 영향을 주고 있음을 확인할 수 있다.

**5-2. 한계점**

* 2018년 1월부터 2020년 4월까지 122주의 데이터를 사용했으나 LSTM, CNN과 같은 NN모델에서는 데이터가 불충분하여 regression 보다 좋은 성능을 내지 못하였다. 이전 데이터를 가져와서 데이터의 양을 늘려준다면 더 나은 성능을 낼 것이라 예측한다.
* 전체적으로 검색량이 더 정확한 주가 예측에 도움을 주기는 하지만, 기업 별로 어떤 검색 데이터와 어떤 모델을 적용하였을 때 정확한 예측이 가능한지 각기 다른 결과를 보이고 있다. 따라서 구체적인 검색 데이터와 예측 모델을 일반화하기에 어려움이 있다. 더 많은 기업을 대상으로 예측 모델을 생성해본다면 다른 기업의 주가 예측에 일반화하기에 더욱 수월해질 것이다
* 구글 트렌드, 네이버 검색량과 같은 검색 데이터는 이미 정제된 데이터로만 존재하기에 절대적인 수치를 얻을 수 없는 한계가 있다.

**5-3. 기대효과**

프로젝트 결과, 기업 ‘삼성전자우’의 주가는 네이버 검색량의 영향을 많이 받고, 기업 ‘현대차’와 ‘카카오’의 주가는 유튜브의 검색량의 영향을 많이 받는 등 위 결과를 해당 기업의 주식 투자에 활용할 만한 요소가 발견되었다. 따라서 주식 투자를 원하는 기업이 있다면 위와 같은 방법으로 검색량을 추가 변수로 사용한 예측 모델을 활용하여 투자에 더 좋은 결과를 얻어낼 것이라고 기대해볼 수 있다.

뿐만 아니라, 변동폭이 크고 그 원인을 예측하기 어려운 현재 주식 시장의 경향성에 따라, 기존에 알려진 주가 지표만으로는 사실상 정확한 주가 예측이 어렵다. 이런 실정에서 ‘검색량’이라는 유의미한 예측 변인을 뽑아냄으로써 조금 더 정확하게 주가를 예측하는 데에 기여해볼 수 있다.