天47년(H) 558

Q

ARGO 모델 기반

김규범 정희철 안세현 김민지 김준서

1 주제 선정 배경

Q 주제 선정 배경은 무엇인가요?

- 주제 선정 배경
- Q ARGO 모델 소개
- 데이터 소개

Q ARGO 모델 소개

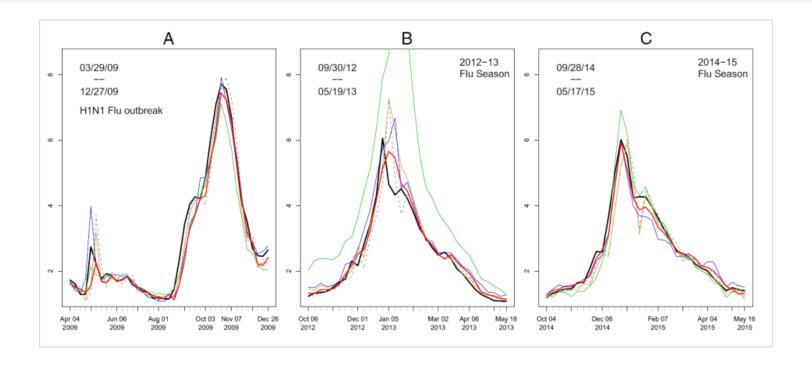
Auto Regression with GOole search data

$$y_t = \mu_y + \sum_{j=1}^{N} \alpha_j y_{t-j} + \sum_{i=1}^{K} \beta_i X_{i,t} + \epsilon_t$$

Estimation: Penalized method

$$\arg \min_{\mu_{y},\alpha,\beta} \sum_{t} (y_{t} - \mu_{y} - \sum_{j=1}^{52} \alpha_{j} y_{t-j} - \sum_{i=1}^{100} \beta_{i} X_{i,t})^{2} + \lambda_{\alpha} \|\alpha\|_{1} + \eta_{\alpha} \|\alpha\|_{2}^{2} + \lambda_{\beta} \|\beta\|_{1} + \eta_{\beta} \|\beta\|_{2}^{2}$$

Q ARGO 모델 소개



ARGO 모델: 상대적으로 간단한 모델로, 타 복잡한 모델에 비해서 실제 flu 상황을 나타내는 검은색 그래프의 추세를 잘 따르는 것을 확인 가능



2데이터 수집 과정

Q 데이터 수집 과정은 무엇인가요?

- Q Tesla 관련 데이터 크롤링
- Q Tesla 주가 데이터 수집
- Tesla 연관 데이터 수집

Q Tesla 관련 데이터 크롤링



Google Correlate의 서비스 중단



크롤링을 통해서
Tesla와 관련 있는
단어 100개
직접 선정

Q Tesla 관련 데이터 크롤링

https://www.reuters.com > article > us-tesla-lawsuit-zoo... •

U.S. self-driving car startup Zoox agrees to settle lawsuit ...

2020. 4. 14. — Zoox Inc said on Tuesday it had settled a lawsuit with **Tesla** Inc after admitting that some new hires from the electric carmaker were in possession of ...



브라우저를 자동화할 수 있는 Python의 Selenium을 통해 2년간 Tesla에 대한 제목과 내용 크롤링 진행

Q Tesla 관련 데이터 크롤링

제목과 내용을 크롤링한 데이터 생성

	Titles	Contents
1	'테슬라X비트코인' … 주목할 만한 5가지 핵심 포인트	직장인 이모(40)씨는 지난 5일 미국 전기차 업체 테슬라 주식을 …
2	Tesla partners with nickel mine amid shortage fears	Tesla has revolutionalized the auto industry, building cars…
3	Tesla Stock Forecast, Price&News	Elon Musk is still the chief executive of Tesla …
4	Top 4 2021Tesla Model 3 Features	Tesla, the luxury electric carmaker, said on Thursday …

Q 테슬라 주가 데이터 수집

2 데이터 수집

월요일 시가와 금요일 종가 수집



1 에이터 탐색

yfinance 라이브러리 이용



'데이터 수집 과정 '

최종 예측 데이터 선정 과정

3 회종 예측값 선정



종가-시가의 차이값 활용

Q 테슬라 주가 데이터 수집

1. '종가-시가 ' 를 최종 예측 값으로 선정

EX)

Mon_Open	Fri_Close		
1001.51			
	1017.01		

변동률

 $\frac{1017.01 - 1001.51}{1001.51}$

= 0.0154

변화가 미미해 등락률을 체감하기 어려움

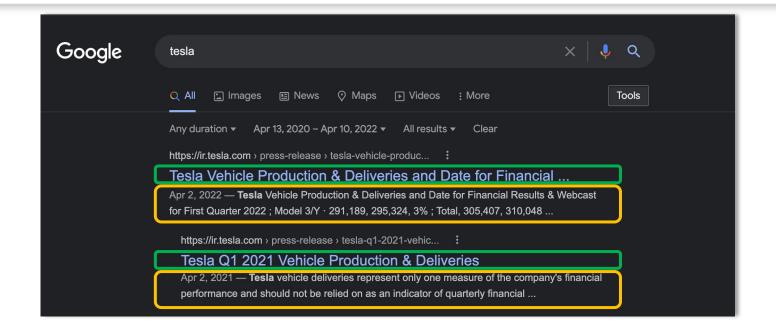
종가 - 시가

1017.01 - 1001.51

= 15.5

최종적으로 차이값을 예측값으로 선택

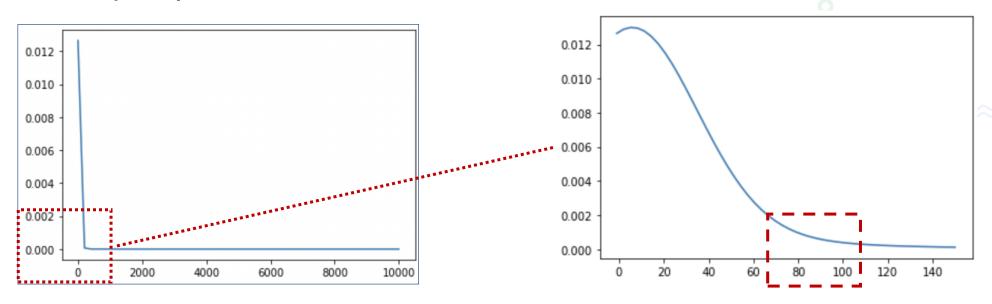
Q 텍스트 데이터 처리



2년 단위로 수집된 Tesla에 대한 **제목**과 내용 CSV 데이터를 형태소로 분해 → NLTK, Kkma로 자연어 처리 진행

Q 최종 데이터 생성

1. 빈도수 분석



Q 데이터 시각화

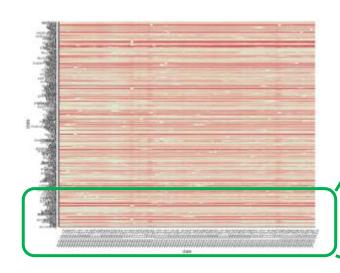
2. 최종 데이터 워드클라우드

```
uselike Beta Test day SEC still 관련시장 Wall used owners Cars week BMW Has Al기업 FSD buy futureSolar week BMW Has Al기업 FSD buy owners Cars week BMW Has All In Exposure FSD buy owners Cars week BMW Has All In Exposure FSD buy owners Cars week BMW Has All In Exposure FSD buy owners Cars week BMW Has All In Exposure FSD buy owners Cars week BMW Has All In Exposure FSD buy owners Cars week BMW Has All In Exposure FSD buy owners Cars week BMW Has All In Exposure FSD buy owners Cars week BMW 
           work Reviews With plant driver 자율주행 Germany including 모델 Cybertruck quarter according EvsThursday Fremont Bitcoin모델3MotorsAlex Used
             supplysales delivery Teslas energy thexa 0 Executive the Eventual thexa 0 Executive thexa 0 Executive thexa 0 Executive the Eventual thexa 0 Executive thexa 0 Executive the Eventual the Eventual thexa 0 Executive the Eventual the Eventual thexa 0 Executive the Eventual the Event
                             carmaker Monday Stock production one
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            investigation last shares biggest DA BESelf-Driving gixa0
             highfactory Model 3 Market Model Y
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          Stock<sub>two</sub> CEOGigaelectric 차량 Model S Musk
world convenient Be TSLA Model new Service Crypto EV On Berlin Service Crypto EV On B
                 batteries opportunity YouTube models
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                What P el Gigafactory YouTuber Supercharger
       review CNBC make value Nore crash profit Earnings Jump Schief SUV data Says Worth back Long Here As Set Subject of Subject Su
```

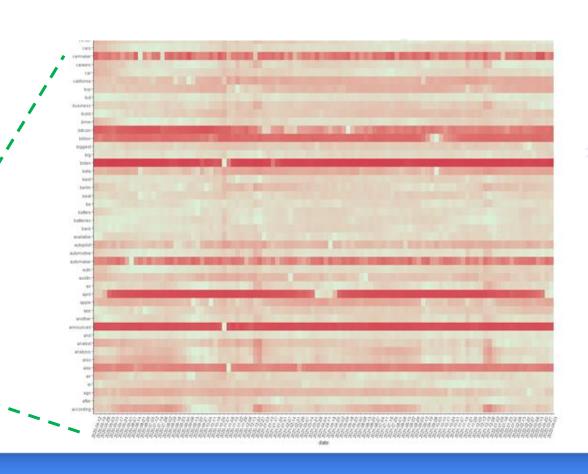
000

Q 데이터 시각화

3. 시각화



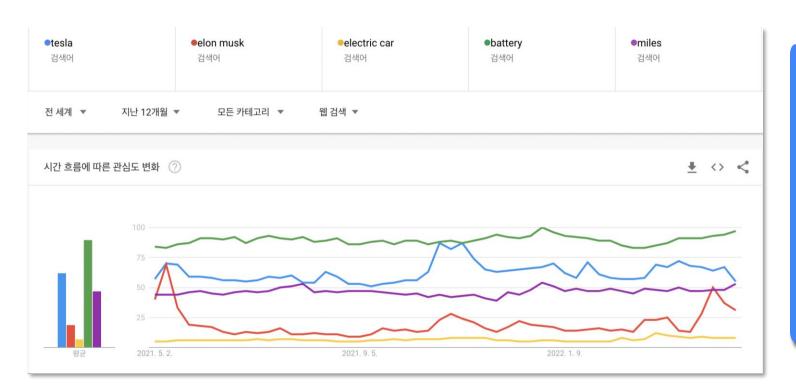
검색량 추세 시각화



$\bullet \bullet \bullet$

Q 최종 데이터 생성

4. 상관관계 분석



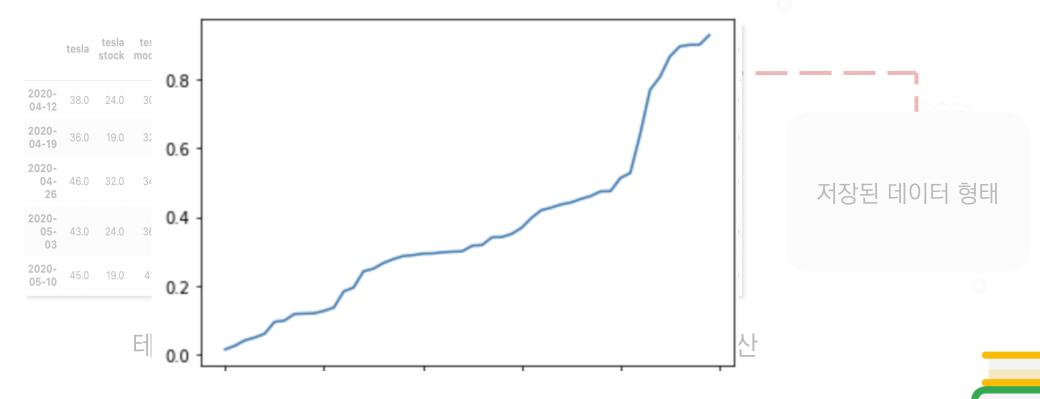
분석 과정

Tesla와 각 단어의 검색량 볼륨의 상관관계 분석 진행 . . .

Q 최종 데이터 생성



상관관계를 오름차순으로 정렬한 Plot



Q 수집된 데이터 시각화

수집된 영어 데이터

Are SEC First set It Buy BetaShop SaysbigNow ago Chief May day Varial auto toxa0 Jump Private Market Safety serview may Bidden Sepaces million Bitcoin update Reviews US COULD delivery Cryptouse features system Executive companies Twitter Self-Driving EV How could year what Motors panasonic Varial auto toxa0 Jump Private Market Safety years frow for the Spaces Million Bitcoin update Reviews US COULD Toyota Flying Europestilitweets plans BMW Choi Shares school Suv would Tuesday would reside the Warrender of Shares Shool Suv Would Tuesday would reside the Warrender of Shares Shool Suv Would Tuesday would reside the Warrender of Stransition deliveries FSD Car's Supercharger Inearnings Battery Vehicles Monday BE Elon time Model Y He battery Company Toyota Flying Europestilitweets plans BMW Choi Shares School Suv Would Tuesday Vehicles Texas driving maderange another what Motors panasonic Venices Self-driving Wednesday Vehicles Texas driving maderange another would year What Motors panasonic Venices Self-driving Wednesday Vehicles Texas driving maderange another would year What Motors panasonic Venices Self-driving Wednesday Vehicles Texas driving maderange another would year what Motors panasonic Venices Self-driving Wednesday Vehicles Texas driving maderange another would year what Motors panasonic Venices Self-driving Wednesday Vehicles Texas driving maderange another who would year would reside the Work of New York of Self-driving Wednesday Vehicles Texas driving maderange another would year who would year would reside you would year would residence of the year Self-driving Wednesday Vehicles Texas driving maderange another who would year who would year first New opportunity element of the year Self-driving Wednesday Vehicles Texas driving maderange another who would year who would year would year would year who would year who would year who would year

수집된 한글 데이터

대부분의 단어가 겹치는 것을 확인 -> 한국어 데이터를 제외



Q 최종 데이터셋

 $\bullet \bullet \bullet$

최종 데이터셋

Date	Tesla	Electric	Charging	Supply	Pltr stock	Like
2020-04-13	39	64	66	97	l 0	96
2020-04-20	36	68	67	100	l 0	99
2020-04-27	46	71	71	99	0	100
2020-05-04	43	73	75	99	0	99

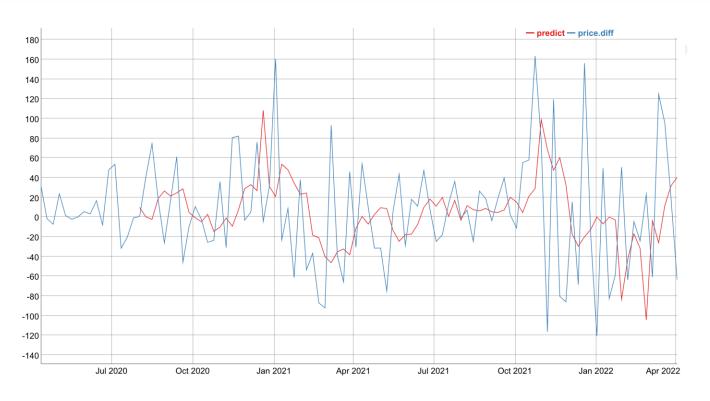
Q 모델 생성

INPUT MODEL OUTPUT

단어 100개의 트렌드
Lag terms ARGO 예측값

초기 모델에 들어간 데이터는 단어 100개의 트렌드와 lag terms ARGO model에 넣어 성능 확인

Q 실제 변동폭과 예측 변동폭



예측치는 크게 엇나가며 추세는 시점이 뒤로 밀리는 현상을 보임

. . .

Q 새로운 데이터 수집

저조한 성능의 원인 : 데이터 측면에서 두 가지 가능성

데이터의 부족

2년 데이터 **↓**6년 데이터

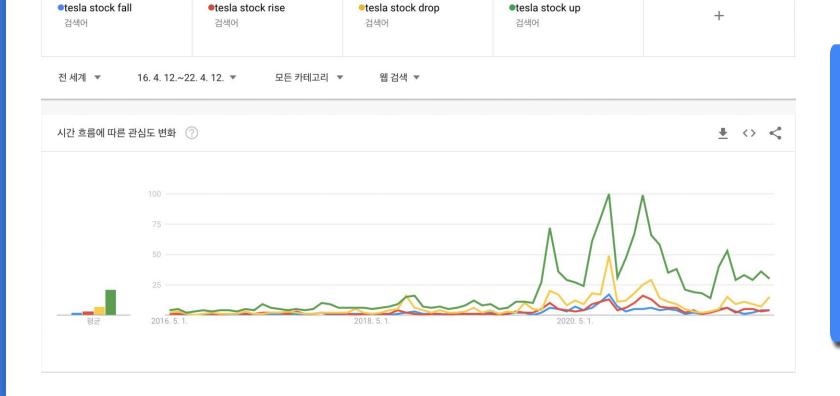
검색어 구체화

"TESLA"

"TESLA STOCK"

•••

Q 새로운 데이터 수집



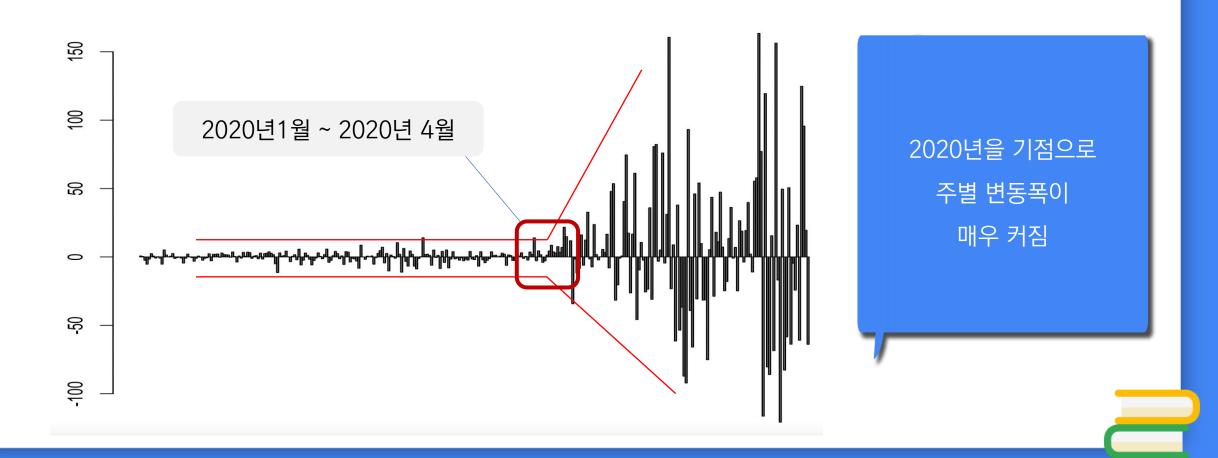
연관 검색어

Tesla stock fall/rise, Tesla stock drop/up 등

테슬라 주가 등락을 나타내는 검색 어들의 연관 검색어들을 추가 수집

000

Q 테슬라 주가 주별 변동폭

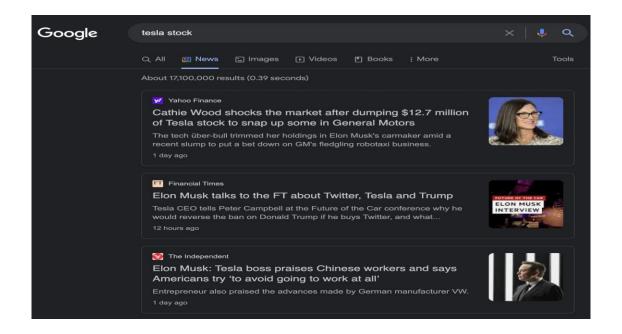


Q 데이터의 변화

esla motors stock Solar city stock Nike stock Giga factory Nikola stock Bitcoin Dodge coin Meta stock Xpeng stock Elon musk

2020년 이전엔 없지만 2020년 이후에 있는 단어들과 2020년 이전엔 있지 만 2020년 이후에 없는 단어들을 구분하고 전처리

Q 뉴스 본문 크롤링



Google News에서 검색어 "tesla stock" 결과 수집

• • •

Q 수정된 최종 데이터

자연어 처리

토큰화 어간추출 원형복원

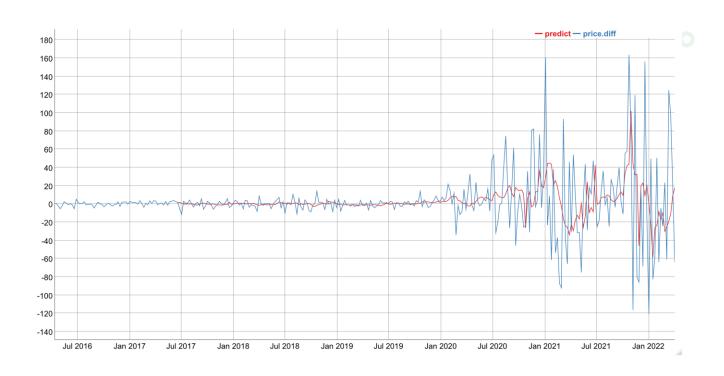
빈도수 추출

임계 빈도수를 넘긴 단어들에 한해 구글 트렌드 데이터 수집

상관관계 추출

Tesla stock의 검색량 추세와 가장 상관관계가 높은 단어 100개 추출

Q 실제 변동폭과 예측 변동폭



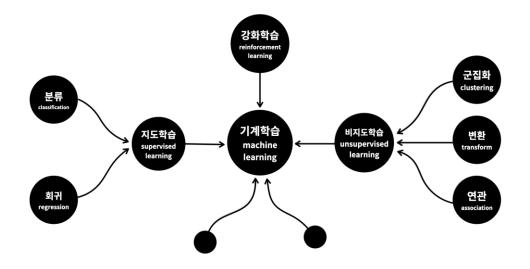
추세를 더 잘 쫓아가는 모습을 보이지만 여전히 변동폭 예측에는 한계가 있음



. . .

Q 비교할 모델 선정

ARGO 모델의 성능을 평가하기 위해 성능을 비교할 모델이 필요



모델 훈련에 사용된 데이터셋

X

313주 간의
Tesla stock과 **상관관계**가 높은 **100개** 단어의 **검색량 트렌드** 데이터



313주 간의 Tesla 주식 가격의 변동폭

모델 훈련에 사용된 데이터셋

X

313주 간의
Tesla stock과 **상관관계**높은 **100개** 단어의 **검색량 트렌드** 데이터



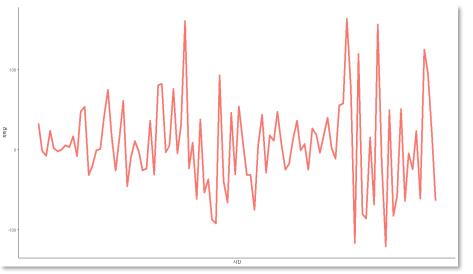
313주 간의 Tesla 주식 가격의 변동폭



Train set : 272주

Test set : 41주

예측 변수를 변동폭으로 설정





Tesla **주가 변동폭**이라는 **수치형** 변수 예측을 목표로 모델링 진행

예측 변수를 변동폭으로 설정

	실제 변동폭	ARGO 예측 값		
L				
2022/02/06	-63.79	-16.13		
2022/02/13	-4.59003	-13.907		
2022/02/20	-24.26	-15.918		
2022/04/04	-63.890015	10.18199462		



수치를 봤을 때 예측 결과가 좋지 않아 보임

. . .

Q 비교할 모델 선정

예측 변수를 범주로 설정

상승

하락

주식 가격 예측에 대한 대부분의 선행연구는 이진분류로 성능을 평가하는 것 참고

 $\bullet \bullet \bullet$

예측 변수를 범주로 설정

급락 하락 상승 급등



최종 예측 변수

수치형

Tesla 주식 가격의 한 주간 **변동폭**

범주형

Tesla 주식 가격의 변동폭을 0을 기준으로 나눈 2개 범주

범주형

Tesla 주식 가격의 변동률을 표준화 후 사분위수를 기준으로 나눈 **4개 범주**

Q 최종 성능 비교 : RMSE

ARGO > Smoothing Spline > XGBoost > Random Forest > LGBM

	ARGO	Random Forest	XGBoost	LGBM	Smoothing Spline
RMSE	32.69610	66.98938	66.4967	68.6638	43.1539

Q 최종 성능 비교 : Accuracy

2진 분류: LGBM 〉 KNN 〉 XGBoost 〉 ARGO 〉 Logistic Regression 〉 SVM

4진 분류: XGBoost 〉 LGBM 〉 ARGO 〉 Logistic Regression 〉 SVM 〉 KNN

	ARGO	XGBoost	LGBM	KNN	Logistic Regression	SVM
2진 분류	48.5%	53.6%	57.5%	55%	48%	46%
4진 분류	31.7%	36.6%	32.5%	22.5%	26%	24%