AI (머신러닝, 딥러닝), 데이터분석

PORTFOLIO

2020.11.19

지원자 : 김서정

souljoin0229@gmail.com

CONTENTS

인적사항 및 주요경력

2020.08 서울시립대학교 공간정보공학과 졸업

프로젝트 수행 이력

2020.08 ~ 2020.09 서울시 강남구 스쿨존 교통사고 위험지역 예측모델 개발 프로젝트

2020.09 ~ 2020.10 자연어처리 기반의 KOSPI 및 YG엔터주가 예측 프로젝트

인적사항 및 주요경력

지원분야 데이터 분석 및 AI모델(머신러닝, 딥러닝) 개발

인적사항

이 름 김서정
생년월일 1996년 2월 29일 (만 24세, 여)
연 락 처 010-7301-6533
이 메 일 souljoin0229@gmail.com

학 력

2020.08 서울시립대학교 공간정보공학과 졸업
(학점 3.41/4.50)

2015.02 설월여자고등학교 졸업

지원각오 데이터와 함께 성장하는 인재가 되고 싶습니다!

프로젝트 수행 이력

프로젝트명	서울시 강남구 스쿨존 교통사고 위험지역 예측모델 개발 프로젝트
목적	머신러닝을 통해 서울시 강남구 스쿨존 내 교통사고 발생 위험 지역을 예측하고, 앱 또는 웹 사이트에서 해당 지역의 사고 위험도를 알림
수행기간	2020.08 ~ 2020.09
팀 구성 및 역할	4인 1개팀 / - 데이터 전처리 - 머신러닝 모델 설계 및 구현 - Map 웹사이트 구현
사용언어 및 도구	Python, Jupyter Notebook, Javascript & css, KAKAO MAP API, Pycharm
Github URL	https://github.com/kimseojeong6533/SZ-Wannabe/blob/master/%EC%8A%A4%EC%BF%A8%EC%A1%B4%EC%9C%84%ED%97%98%EC%98%88%EC%B8%A1_EDA_Modeling_Mapping.ipynb
비고	서울시와 ICT콤플렉스가 함께하는 2020 ICT콕 AI공모전 장려상 입상

산출물

https://youtu.be/OGY-zS1_KOE

프로젝트명	서울시 강남구 스쿨존 교통사고 위험지역 예측모델 개발 프로젝트
수행기간	2020.08 ~ 2020.09

수행 단계 및 방법

1. 데이터 전처리
2. 도로별 사고발생 횟수를 카운트하여 위험도 등급(Label값) 구분
3. Label값에 대한 각 Feature의 분포를 확인하여 공간적 상관성이 높은 Feature 추출
4. 앙상블트리모델 중 의사결정나무, 랜덤포레스트, 그래디언트 부스팅 회귀트리모델의 각 Train_Test_split_ratio, max_depth, n_estimator 등의 파라미터 별 성능 비교
5. 스쿨존 주소 기반의 Testset을 모델에 input하여 위험지역 도출
6. Javascript, kakaomap api를 이용해 위험지역을 시각화한 웹사이트 생성

프로젝트명	서울시 강남구 스쿨존 교통사고 위험지역 예측모델 개발 프로젝트
수행기간	2020.08 ~ 2020.09

수행 단계 및 방법

1. 데이터 전처리 (서울 열린 데이터광장의 공공데이터 이용)

- 강남구 데이터 외 삭제
- 사고 번호 삭제
- 사고 일시 → 연도/월/일/시로 분류
- 시군구, 도로명 → 위도, 경도변환 후 속성 추가(지오코딩 활용)
- 사고내용 → 사망자수, 중상자수, 경상자수, 부상신고자수 데이터 분류
- 사고 유형 -〉 차대차/차대사람으로 분류
- 법규 위반, 노면 상태 삭제
- 기상 상태 → 기타 항목 데이터 맑음, 흐림 데이터 양 비율로 랜덤하게 분류
- 도로 형태 → 기타 항목 제거, 교차로 및 단일로 항목만 분류

프로젝트명	서울시 강남구 스쿨존 교통사고 위험지역 예측모델 개발 프로젝트
수행기간	2020.08 ~ 2020.09

수행 단계 및 방법

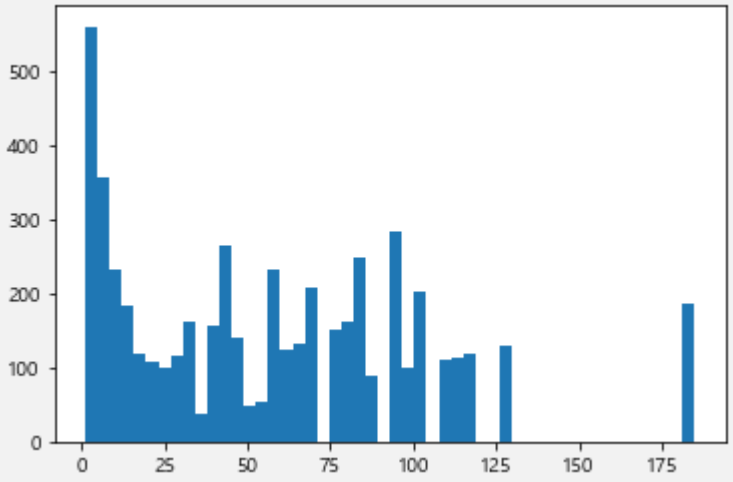
2. 도로별 사고발생 횟수를 카운트하여 위험도 등급(Label값) 구분

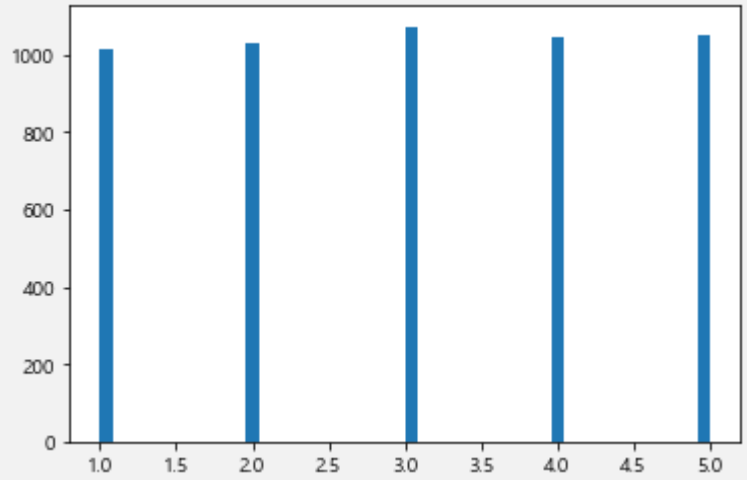
```
# 사고발생수 분포확인
x=np.array(df['사고발생수'])
print(x)
s = pd.Series(x)
s.describe()

n, bins, patches=plt.hist(x, bins=50)
```

```
num = pd.Series(x)

Q1 = num.quantile(.2)
Q2 = num.quantile(.4)
Q3 = num.quantile(.6)
Q4 = num.quantile(.8)
Q5 = num.quantile(1)
```

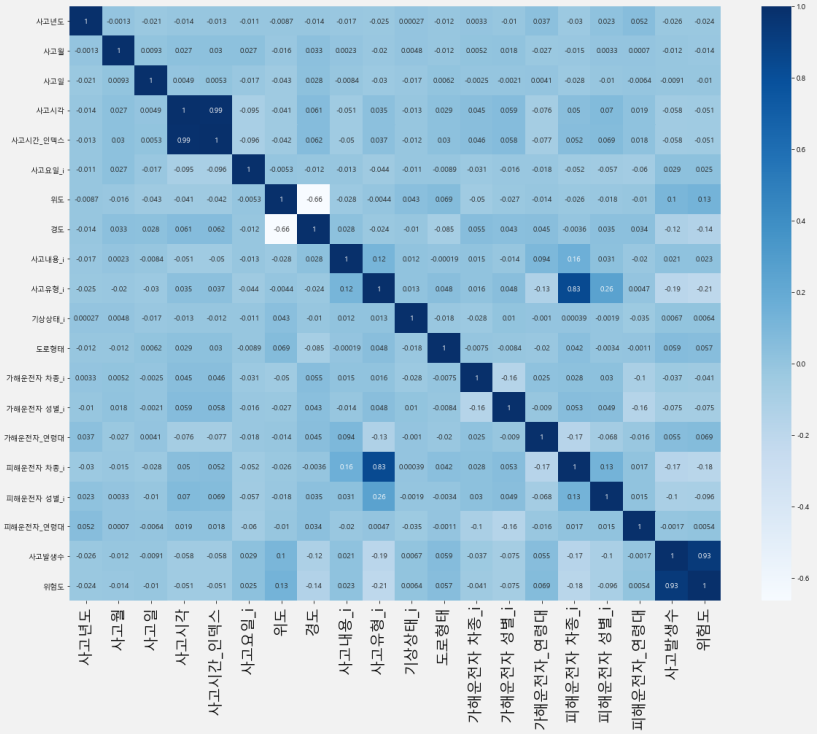




프로젝트명	서울시 강남구 스쿨존 교통사고 위험지역 예측모델 개발 프로젝트
수행기간	2020.08 ~ 2020.09

수행 단계 및 방법

3-1. Label값에 대한 각 Feature의 EDA 및 Input Feature 도출

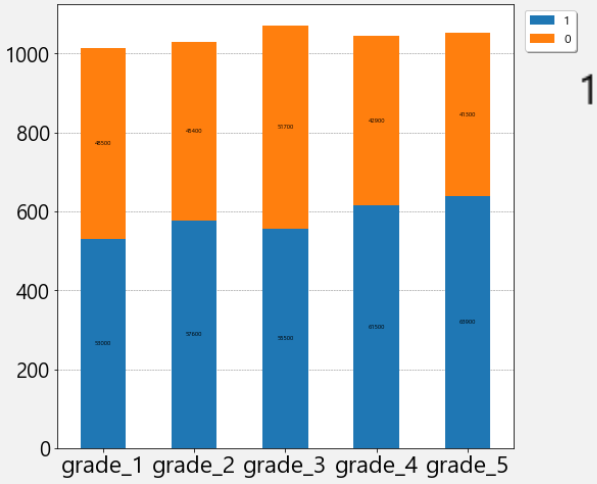


프로젝트명	서울시 강남구 스쿨존 교통사고 위험지역 예측모델 개발 프로젝트
수행기간	2020.08 ~ 2020.09

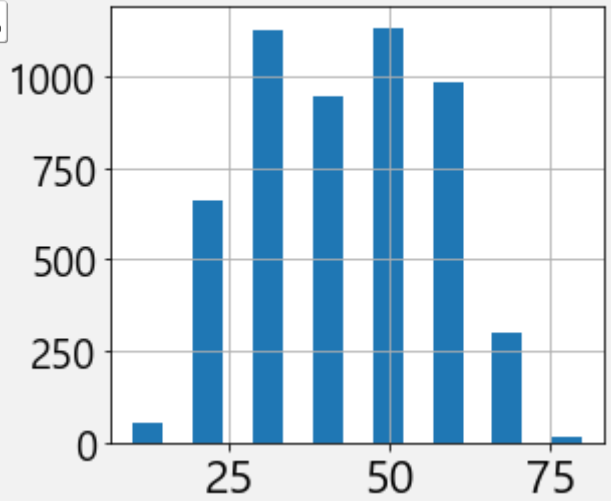
수행 단계 및 방법

3-2. Label값에 대한 각 Feature의 EDA 및 Input Feature 도출

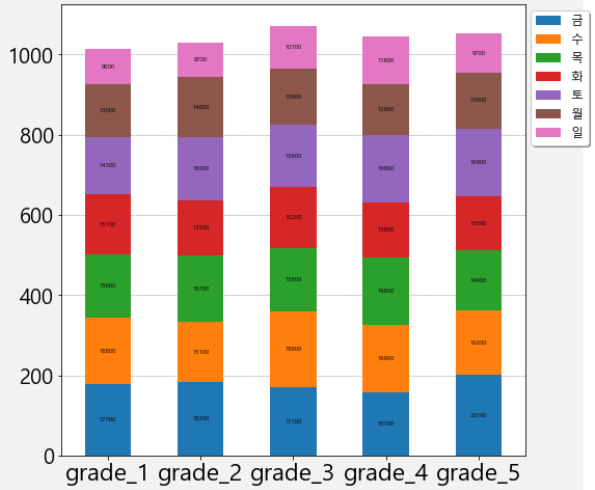
위험도별 도로형태 (0:단일로, 1:교차로)



피해운전자 연령대별 분포



위험도별 요일



➡ '위도','경도','사고시간_인덱스','사고요일_i','기상상태_i','도로형태','위험도'

4. 의사결정나무, 랜덤포레스트, 그래디언트 부스팅 회귀트리모델의 파라미터별 성능비교

```
# Gradient Boosting Regression Tree - 7 : 3

for i in range(1,6):
    X_train, X_test, y_train, y_test = train_test_split(data.iloc[:,:-1], data['위험도'], test_size=0.3, shuffle=True, random_state=0)
    gbrt = GradientBoostingClassifier(max_depth=i,random_state=111,n_estimators=33)
    gbrt.fit(X_train, y_train)

    print('Max_depth : {}'.format(i))
    print("훈련 세트 정확도: {:.3f}".format(gbrt.score(X_train, y_train)))
    print("테스트 세트 정확도: {:.3f}".format(gbrt.score(X_test, y_test)))
    print('특성 중요도 : ₩n', gbrt.feature_importances_)
    print()
```

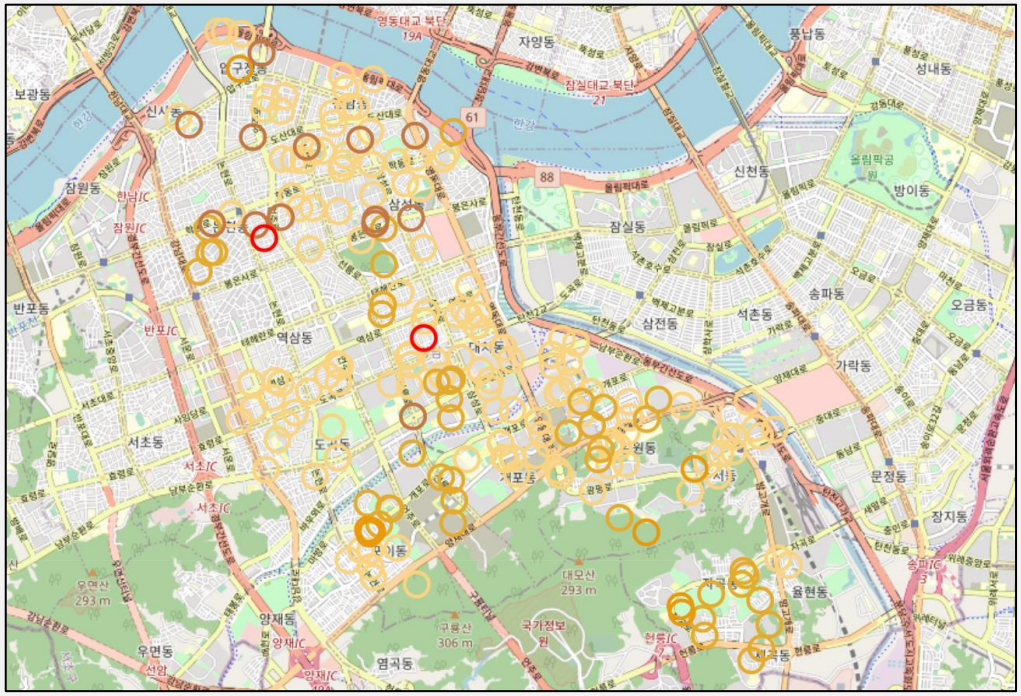
Trial condition	Max_depth	Random Forest (Train data, Test data) (n_estimators=33)	Decision Tree (Train data, Test data)	Gradient Boosting Regression Tree (Train data, Test data) (n_estimators=33)
5:5	1	0.320, 0.311	0.307, 0.336	0.489, 0.502
	2	0.416, 0.428	0.415, 0.430	0.698, 0.689
	3	0.459, 0.463	0.418, 0.435	0.746, 0.736
	4	0.531, 0.532	0.534, 0.544	0.802, 0.772
	5	0.632, 0.606	0.590, 0.602	0.824, 0.783
7:3	1	0.314, 0.339	0.314, 0.339	0.494, 0.504
	2	0.398, 0.409	0.382, 0.386	0.686, 0.689
	3	0.474, 0.474	0.406, 0.404	0.747, 0.737
	4	0.522, 0.519	0.489, 0.499	0.790, 0.776
	5	0.562, 0.555	0.553, 0.573	0.818, 0.788
8:2	1	0.332, 0.345	0.318, 0.334	0.812, 0.779
	2	0.416, 0.434	0.379, 0.397	0.812, 0.779
	3	0.456, 0.461	0.406, 0.405	0.812, 0.779
	4	0.529, 0.539	0.490, 0.497	0.812, 0.779
	5	0.594, 0.574	0.557, 0.579	0.812, 0.779

프로젝트명	서울시 강남구 스쿨존 교통사고 위험지역 예측모델 개발 프로젝트
수행기간	2020.08 ~ 2020.09

수행 단계 및 방법

5. 스쿨존 주소 기반의 Testset을 모델에 input하여 위험지역 도출

- 1단계 위험도
- 2단계 위험도
- 3단계 위험도
- 4단계 위험도
- 5단계 위험도



프로젝트명	서울시 강남구 스쿨존 교통사고 위험지역 예측모델 개발 프로젝트
수행기간	2020.08 ~ 2020.09

수행 단계 및 방법

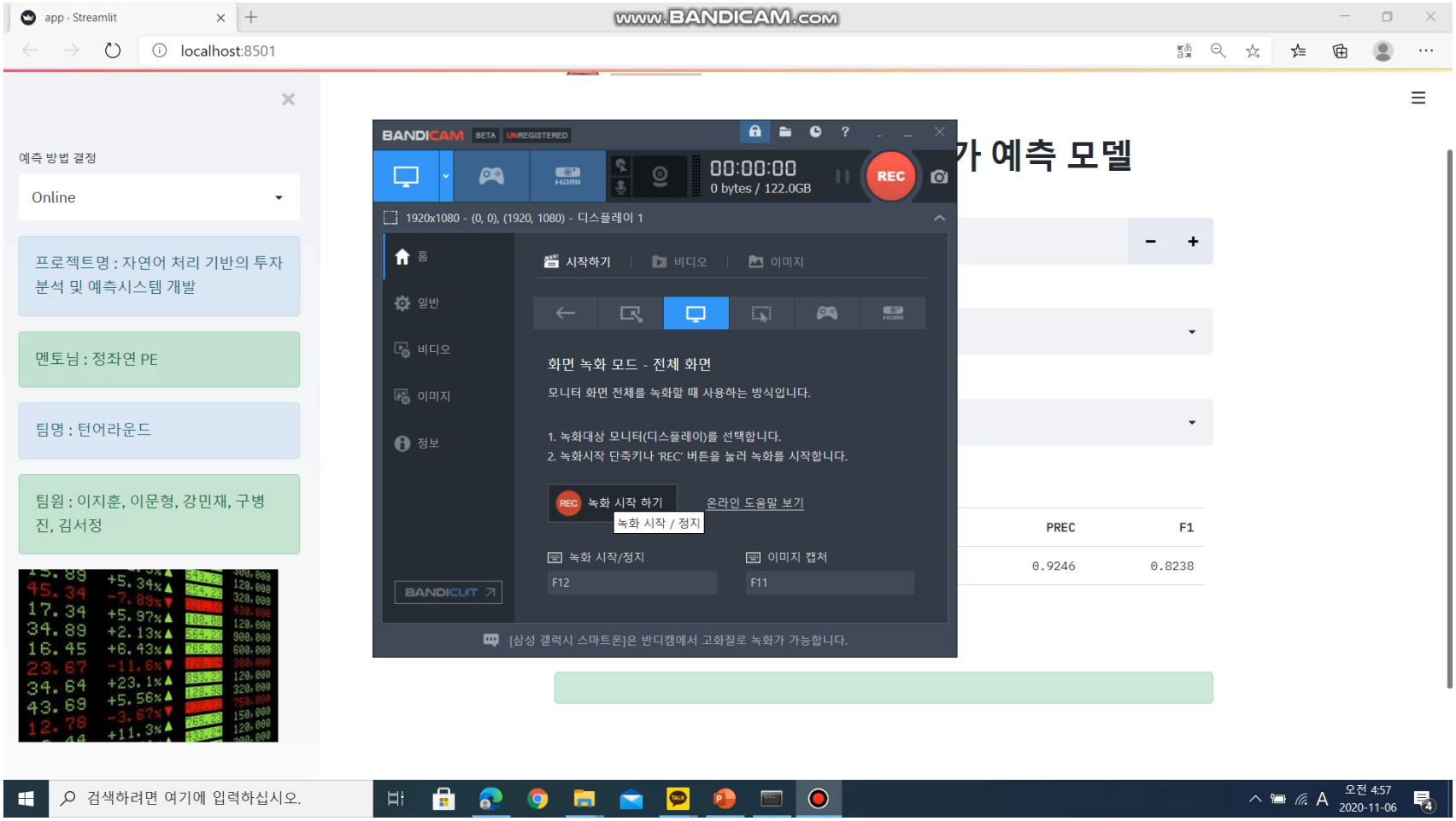
6. Javascript, kakaomap api, Pycharm를 이용해 위험지역을 시각화한 웹 생성

프로젝트 수행 이력

프로젝트명	자연어처리 기반의 KOSPI 및 YG엔터주가 예측 프로젝트
목적	경기에 영향을 받는 코스피와 테마주를 뉴스데이터와 prophet, AutoML, 강화학습 알고리즘을 활용하여 예측하고자 함
수행기간	2020.09 ~ 2020.11
팀 구성 및 역할	5인 1개팀 / - 학습데이터 전처리 - FBProphet을 사용하여 금융 시계열 분석 - Jupyter notebook을 이용한 전체 코드 통합
사용언어 및 도구	Python, Google Colab, Jupyter notebook, Pytorch, Tensorflow, Bert
Github URL	https://github.com/ejihoon6065/Project_TurnAround

산출물

https://youtu.be/aDm0r-_bh3I



프로젝트 수행 이력

프로젝트명	자연어처리 기반의 KOSPI 및 YG엔터 주가 예측 프로젝트
수행기간	2020.09 ~ 2020.11

수행 단계 및 방법

1. 데이터 크롤링

1-1. KRX, **Yahoo Finance, Investing.com** 등 주가 및 투자보조지표 등의 데이터를 크롤링하여 2018. 01. 01 ~ 2020. 10. 26 주가데이터 확보

1-2. 한국경제 신문의 경제,국제부문의 2018. 01. 01 ~ 2020. 10. 26 뉴스기사 타이틀(1일당 50개 기사)을 크롤링

2. 자연어처리 (TF계산)

2-1. Mecab 형태소분석기를 이용해 텍스트 데이터를 정제

2-2. 텍스트의 Unigram,Bigram단어들의 TF를 구하고 Top1000의 단어를 뽑아 복합명사 일체화 및 단일 글자 처리

2-3. 한자를 한글로, 고유명사를 각 도메인으로 단어 치환

2-4. 코스피 등락률을 확인하여 전일 기사의 라벨링에 반영

3. AutoML, Prophet 등을 활용한 등락 예측

3-1. **코스피**, YG**주가에 대한 회귀모델, 분류 모델 생성**

프로젝트 수행 이력

프로젝트명	자연어처리 기반의 KOSPI 및 YG엔터 주가 예측 프로젝트
수행기간	2020.09 ~ 2020.11

1. 크롤링

```
# 종합지수 (코스피) 차트 데이터
kospi_ = stock.get_index_ohlcv_by_date(start_date__, end_date__, "1001")
kospi_.columns = ['Open','High','Low','Close','Volume']

# 코스피 투자자별 공매도 거래량
kospi_short_sell_volume = stock.get_shorting_investor_volume_by_date(start_date__, end_date__, "KOSPI")
kospi_short_sell_volume.columns = ['kospi_inst_volume','kospi_indi_volume','kospi_fore_volume','kospi_etc_

# 코스피 투자자별 공매도 거래대금
kospi_short_sell_value = stock.get_shorting_investor_price_by_date(start_date__, end_date__, "KOSPI")
kospi_short_sell_value.columns = ['kospi_inst_value','kospi_indi_value','kospi_fore_value','kospi_etc_valu
```

2018-10-31	허성무 창원시장, 김동연 경제부총리에 "산업 체질 개선 위한 전략산업" 지원 요청	[대출규제에 금리인상 압박까지 下] 전문가 "빚테크 전략 재정비해야"	[대출규제에 금리인상 압박까지 上] 11월 대출시장 '한파' 예고	신세계그룹, 1조 유치 확정…2023년 매출 10조 '한국판 아마존' 키운다	북미권역 본부 본격 가동하는 현대차…돌파구 찾나	KT&G, 릴 100만대 판매 기념 '보상판매 이벤트' 실시	롯데제과, 빼빼로 공식 캐릭터 '빼빼로 일레븐' 선보여	'공유주방' 시장 갈수록 커진다…주방으로 들어온 공유경제	'존경받는 기업인' 늘린다…중기부, 中企 사람중심 문화 확산 지원	...	'유인 드론' 시험비행 쉬워진다…건축물 점검에도 드론 허용	[2018 수입차 결산] 한경닷컴이 뽑은 '수입 올해의 차'에 티구안	KFC, 올해 13번째 신규 매장 신촌역점 오픈	DGB금융 편입 하이투자증권 출범…"금융투자 톱10 도약"	손보사 빅4, 모바일 등 기우편 서비스 도입…"비용절감 톡톡"	인민은행 "1달러=7위안 막아라"…홍콩서 3조원 채권 발행	증 유 삼 재 심 입 분 회 공 2 진

프로젝트 수행 이력

프로젝트명	자연어처리 기반의 KOSPI 및 YG엔터 주가 예측 프로젝트
수행기간	2020.09 ~ 2020.11

2. 자연어처리

TF(Text Frequency) : Uni-gram

Word	TF
미국	214
금융	202
한국	186
경제	177
기업	152
달러	146
LG	127
일본	123
정보	123
구조	113

TF(Text Frequency) : Bi-gram

Word	TF
구조조정	95
정보통신	51
LG전자	38
시스템개발	36
금융기관	34
천만달러	31
벤처기업	29
한국경제	28
금리인하	28
금융위기	26

현대차 'N'에 자극받은 토요타, 고성능 소형 확대

↓ Mecab

현대차 자극 토요타 성능 소형 확대

프로젝트 수행 이력

프로젝트명	자연어처리 기반의 KOSPI 및 YG엔터 주가 예측 프로젝트
수행기간	2020.09 ~ 2020.11

3. 사용한 모델

	Regression Model	Classification Model	Reinforcement Learning
KOSPI	Prophet AutoML (TheilSen, Linear Regression, Ridge) -> Ensemble	AutoML (Gradient Boosting, Linear Discriminant, Ridge) -> Ensemble NLP (Bert, LSTM, AutoML)	
YG	Prophet AutoML (Linear Regression, RANSACR) -> Ensemble	AutoML (Gradient Boosting, Decision Tree, LightGBM) -> Ensemble NLP (Bert, LSTM, AutoML)	A2C (value & policy network LSTM)

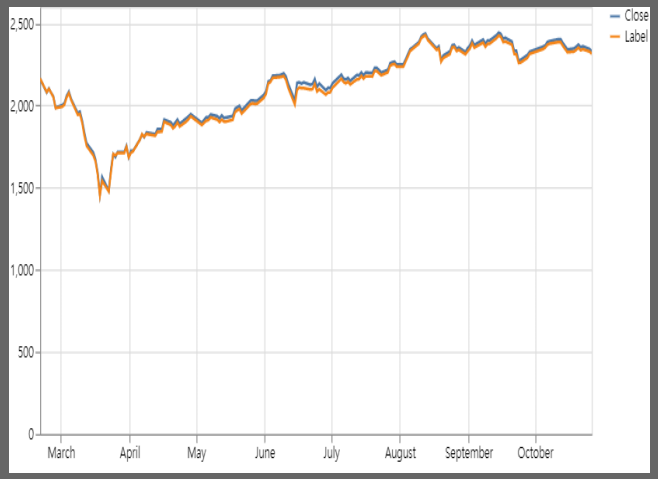
KOSPI Model

Regression Model

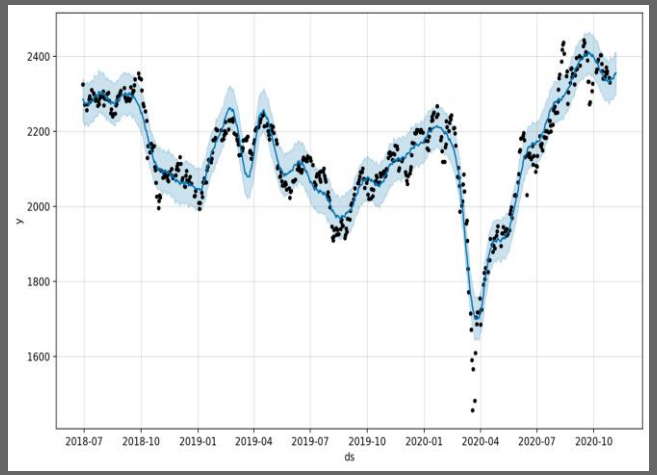
Test Metric Scores

	MAE	MSE	RMSE	R^2
Prophet	85.415	15,005	122.4969	-5.5035
AutoML	16.1176	311.3388	17.6446	0.9942

Actual & Prediction Graph



AutoML



Prophet

KOSPI Model

Classification Model

Test Metric Scores

	Accuracy	ROC AUC	Recall	Precision	F1
AutoML	0.9152	0.9081	0.8676	0.9219	0.8939
NLP Bert	0.8008	0.8205	0.7429	0.9246	0.8238
NLP LSTM	0.8005	0.8069	0.7645	0.8729	0.8151
NLP AutoML	0.8088	0	0.859	0.8171	0.6068

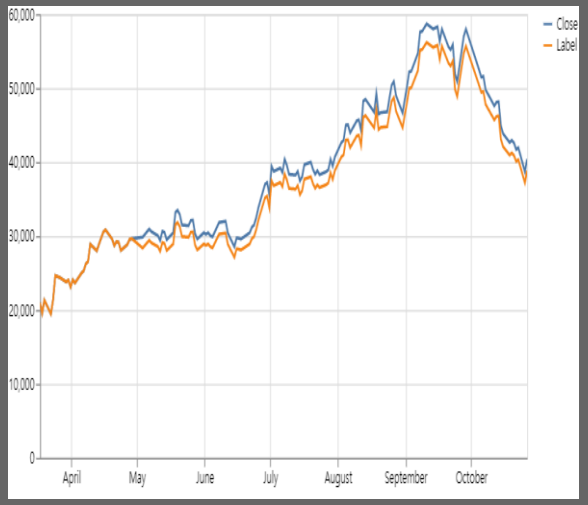
YG Model

Regression Model

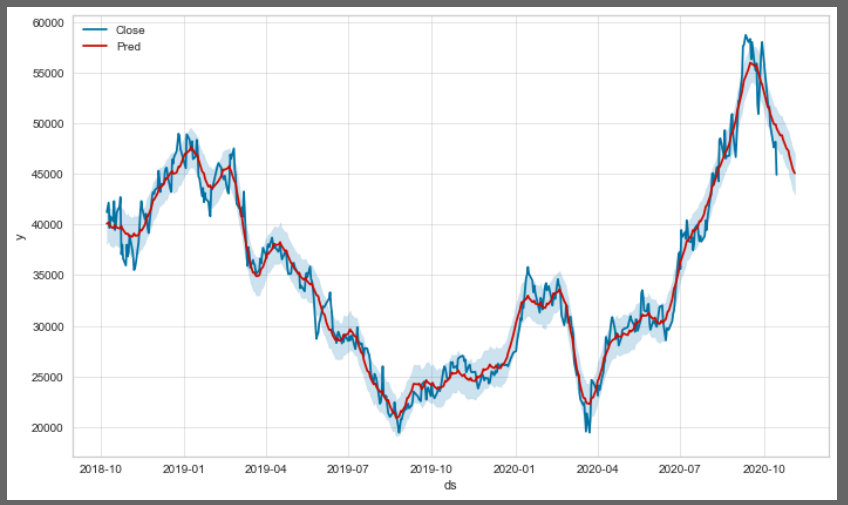
Test Metric Scores

	MAE	MSE	RMSE	R^2
Prophet	1686.042	4,849,041	2202.053	0.9422
AutoML	1,491.8399	2,835,922.830	1,684.3167	0.9724

Actual & Prediction Graph



AutoML



Prophet

YG Model

Classification Model

Test Metric Scores

	Accuracy	ROC AUC	Recall	Precision	F1
AutoML	0.8844	0.8782	0.8387	0.8814	0.8595

결론

선행지표, 보조지표 및 뉴스 데이터를 활용

↓

예측 정확도 상승가능

↓

선행지표, 기술적 분석 및 뉴스 분석을
주식투자 의사결정에 활용가능

프로젝트명	서울시 강남구 스쿨존 교통사고 위험지역 예측모델 개발 프로젝트
수행기간	2020.08 ~ 2020.09

프로젝트명	자연어처리 기반의 KOSPI 및 YG엔터 주가 예측 프로젝트
수행기간	2020.09 ~ 2020.11

해당 프로젝트들을 통해 얻게 된 역량

1. Python을 이용한 시각화,데이터 분석, 모델링 등 프로그래밍 역량

2. 데이터 크롤링 및 텍스트 데이터 전처리 역량 강화

3. 시계열 데이터 모델링, 머신러닝 및 딥러닝 등 AI관련 지식

4. Github 등 협업 프로젝트 플랫폼 경험

비전 및 핵심역량

데이터를 기반으로
유의미한 정보를 제공하며 가치를 발견하는 개발자

머신러닝에 대한 이해

- 인공지능 교육훈련을 통한 핵심역량 배양
  (혁신성장 청년인재 사업 중 실무중심의 인공지능 개발자 3기 과정 11월 수료예정)
- MOOC 등을 활용한 온라인 교육 참여
- 서울시와 ICT콤플렉스가 함께하는 2020 ICT콕 AI공모전 장려상 입상

유관 프로젝트 수행 경험

- 서울시 강남구 스쿨존 교통사고 위험 지역 예측 프로젝트 수행(2020)
- 자연어처리 기반의 KOSPI 및 YG엔터 주가 예측 프로젝트 수행(2020)

다양한 개발 툴, Git 경험

- Pytorch, Tensorflow, Bert모델 경험
- Jupyter notebook, Google colab을 활용한 Python 프로그래밍
- Pycharm, Javascript를 이용한 웹 개발
- Github 주소 : https://github.com/kimseojeong6533

Thank You