


빅데이터에 기반한 머신러닝을 통한 수산물 가격예측

Team 대홍단 왕감자

 이창우(팀장) Kamme10@naver.com

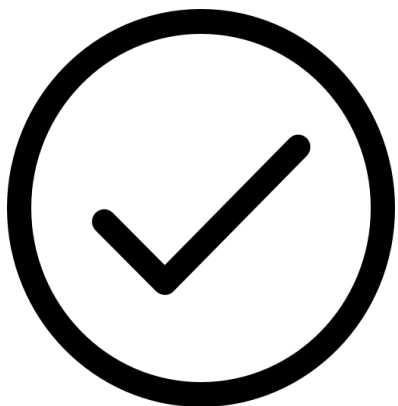
☒ 오승준 dhtmdwns1201@naver.com

☒ 배유진 bae102699@hanmail.net

☒ 손준영 sjy5448@naver.com



목차



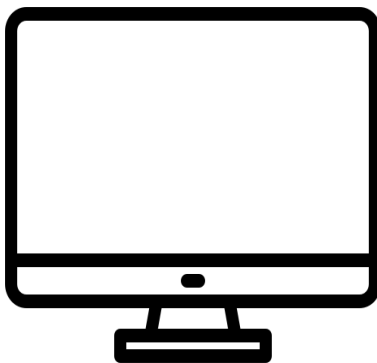
1. 문제 정의

1. 문제 정의
2. 분석 목표
3. 기본 제공 데이터



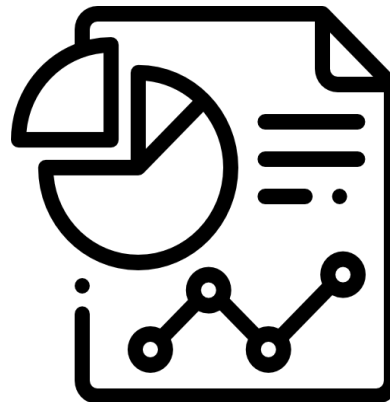
2. 데이터 전처리

1. 피처 전처리
2. LabelEncoding



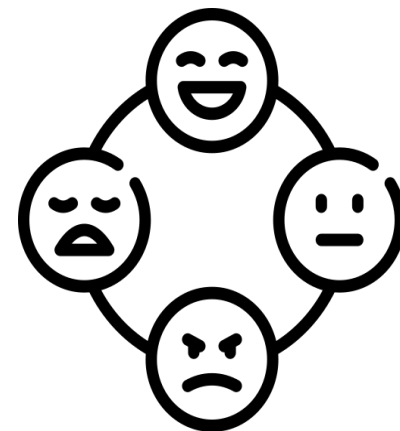
3. 모델 선정 및 훈련

1. 모델 선정 이유
2. 파라미터 선정 이유



4. 결과 분석

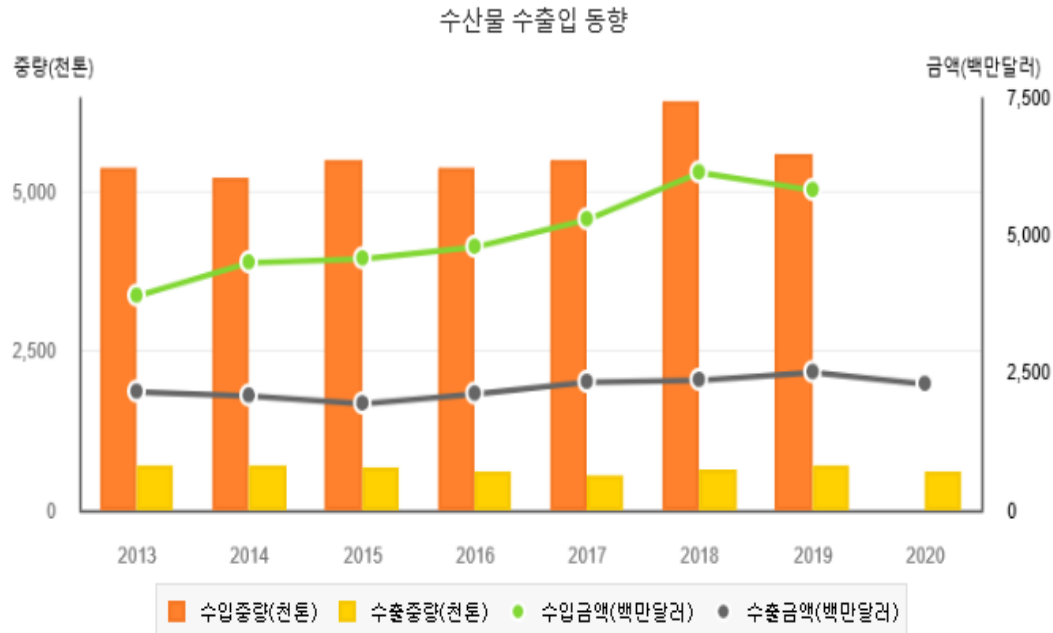
1. 결과 분석
2. 기대 효과



5. 한계점

1. 한계점
2. 참고 문헌

문제 정의



[단위 : 백만달러, 천톤]

	2013	2014	2015	2016	2017	2018	2019	2020
	▲▼■	▲▼■	▲▼■	▲▼■	▲▼■	▲▼■	▲▼■	▲▼■
수입금액	3,895	4,506	4,556	4,791	5,268	6,125	5,794	5,621
수입중량	5,387	5,231	5,504	5,373	5,492	6,419	5,606	5,518
수출금액	2,151	2,067	1,924	2,128	2,329	2,377	2,505	2,306
수출중량	687	702	651	612	540	632	688	614



최신기사 정치 북한 경제 산업 사회 전국 세계 문화 라이프 연예 스포츠 오피니언 사람들 비주얼뉴스 재난포털 **재보**

#코로나19 #연론중재법 #프로배구 #현대건설 #GS칼텍스

뉴스홈 | 최신기사

2분기 수입곡물가격 상승 전망..."국내 식품물가도 따라오를 듯"

우리나라는 무역 의존도가 높은 **개방형 무역구조**를 지니고 있어
수입 물품의 공급량과 가격이 국내 물가 형성에 큰 영향을 미치고 있다.

→ 소비자의 합리적인 소비를 지원하기 위해 **수산물 수입가격**에 대한 정확한 예측이 필요하다.

- 1) e-나라지표.(2021년, 7월 28일).수산물 수출입 동향
- 2) 연합뉴스TV.(2021년, 4월, 16일). 2분기 수입곡물가격 상승 전망..."국내 식품물가도 오를 듯"



분석 목표

학습

2015년 12월 28일 ~ 2019년 12월 31일 데이터

검증

2020년 1월 6일 ~ 2020년 12월 28일 데이터

예측

2021년 1월 4일 ~ 2021년 6월 28일 데이터

-> 머신러닝 모델을 통한 수산물 단가 예측



제공 데이터

2015년 12월 28일 ~ 2019년 12월 31일 데이터

기준일	제품 구분	제조국	수출국	수입 용도	중분류	어종	상세어종	수입형태	평균단가
REG_DATE	P_TYPE	CTRY_1	CTRY_2	P_PURPOSE	CATEGORY_1	CATEGORY_2	P_NAME	P_IMPORT_TYPE	P_PRICE
2015-12-28	수산물	아르헨티나	아르헨티나	판매용	갑각류	새우	아르헨티나 붉은 새우	냉동	7.48
2015-12-28	수산물	바레인	바레인	판매용	갑각류	게	꽃게	냉동	2.92
2015-12-28	수산물	바레인	바레인	판매용	갑각류	게	꽃게	냉동,절단	3.36
2015-12-28	수산물	칠레	칠레	판매용	패류 멍게류	해삼	해삼	건조,자숙	18.26
2015-12-28	수산물	중국	중국	판매용	어류	서대 박대 페루다	서대	냉동	4.79

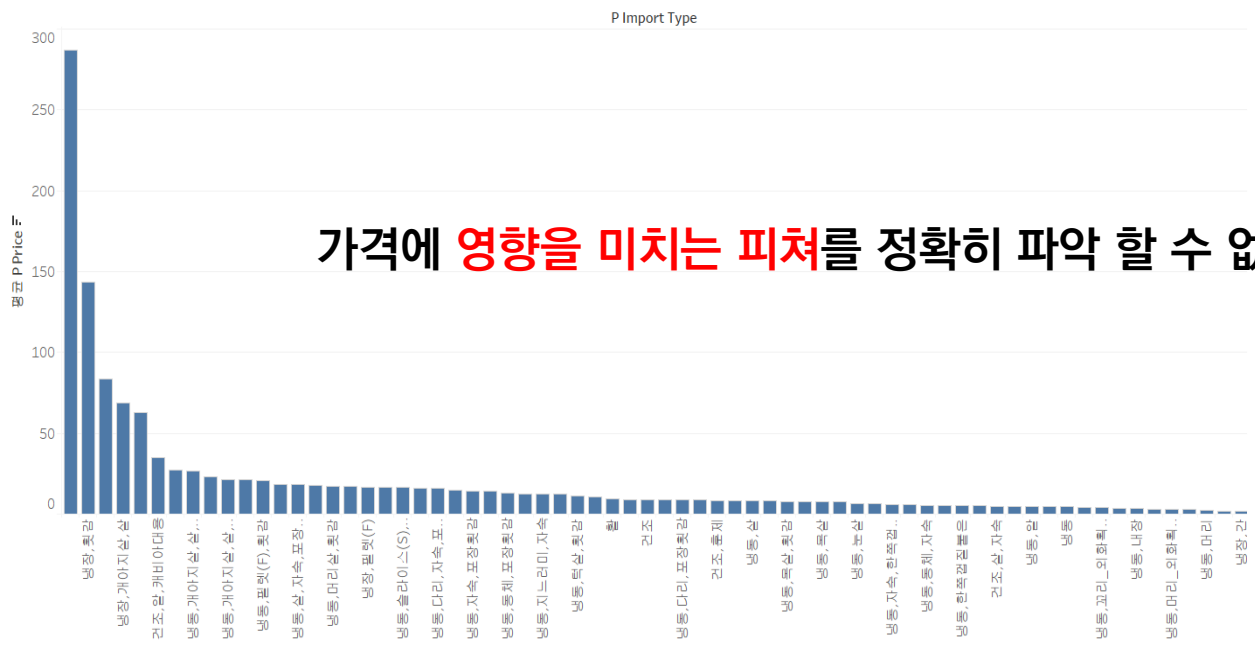


P_IMPORT_TYPE 피처 전처리

P_IMPORT_TYPE 피처는 총 78개로 개수가 방대

'냉동', '냉동,절단', '건조,자숙', '냉동,살,자숙', '냉동,살', '냉장,살', '염장,살',
'냉동,한쪽껍질붙은', '활', '냉장', '냉동,목살', '냉동,목살,횡감', '염장', '냉동,알',
'냉동,필렛(F),횡감', '냉동,포장횡감,필렛(F)', '냉동,동체', '냉동,참난', '냉동,살,포장횡감',
'냉동,자숙,포장횡감', '건조', '냉동,다리,포장횡감', '냉장,필렛(F)', '냉동,자숙,한쪽껍질붙은',
'냉동,필렛(F)', '냉동,살,자숙,포장횡감', '냉장,횡감', '냉동,다리', '냉동,슬라이스(S),포장횡감',
'냉장,포장횡감,필렛(F)', '냉동,동체,자숙', '냉동,지느러미', '냉장,필렛(F),횡감', '냉동,곤이',
'냉동,머리', '자숙,건조', '냉동,슬라이스(S),자숙,포장횡감', '냉동,슬라이스(S)', '냉동,자숙',
'냉동,자숙,절단', '냉동,간', '건조,알,캐비아대용', '냉동,볼살', '냉동,볼살,횡감', '냉동,개이지살,살',
'냉동,내장', '냉동,턱살', '냉동,집게다리', '냉동,외투막,살', '냉동,다리,자숙', '냉동,지느러미,자숙',
'냉동,머리살,횡감', '건조,필렛(F)', '절단,건조', '냉동,개이지살,살,포장횡감', '건조,살,자숙',
'냉동,턱살,횡감', '냉동,알,캐비아대용', '냉동,살,슬라이스(S),포장횡감', '냉동,머리살',
'냉동,자숙,절단,횡감', '염장,알,캐비아대용', '건조,훈제', '냉동,동체,포장횡감', '냉동,횡감',
'냉동,개이지살,살,횡감', '냉장,개이지살,살', '냉동,슬라이스(S),자숙', '냉동,난포선',
'냉동,동체,자숙,포장횡감', '냉동,머리_외화획득용', '냉장,간', '출기,염장', '건조,껍질',
'냉동,다리,자숙,포장횡감', '냉동,꼬리_외화획득용', '냉동,눈살', '건조,지느러미,자숙'],

P_IMPORT TYPE별 가격



P_IMPORT_TYPE 분할 필요



P_IMPORT_TYPE 피처 전처리

P_IMPORT_TYPE
[냉동, 살, 자숙]
[냉동, 목살, 찜]
[냉동, 살, 자숙]
[냉동, 살, 자숙]
[냉동, 필렛(F), 찜]
[냉동, 목살, 찜]
[냉동, 필렛(F), 찜]
[냉동, 포장찜, 필렛(F)]
[냉동, 살, 포장찜]
[냉동, 자숙, 포장찜]



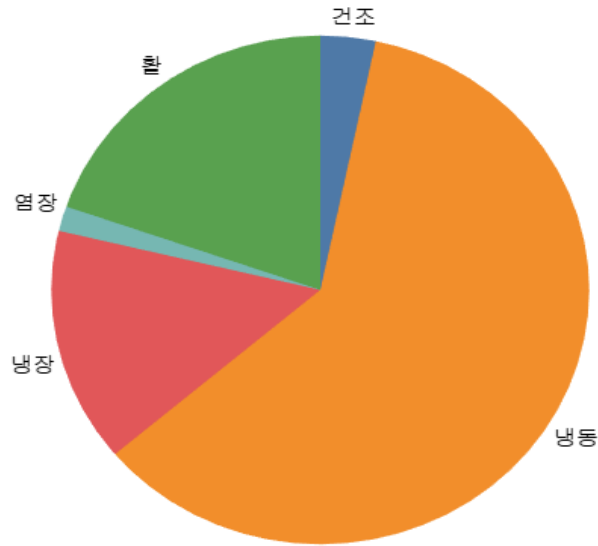
storage	part	process	sashimi	steaming_processing
냉동	살	None	None	자숙
냉동	목살	None	찜	None
냉동	살	None	None	자숙
냉동	살	None	None	자숙
냉동	None	필렛(F)	찜	None
냉동	목살	None	찜	None
냉동	None	필렛(F)	찜	None
냉동	None	필렛(F)	포장찜	None
냉동	살	None	포장찜	None
냉동	None	None	포장찜	자숙

수산물을 구분하는 방식을 참고하여 **5가지** 형태로 분할

-> 저장 방법(Storage), 부위(Part), 가공형태(Process), 찜(Sashimi), 자숙 유무(Steaming Processing)



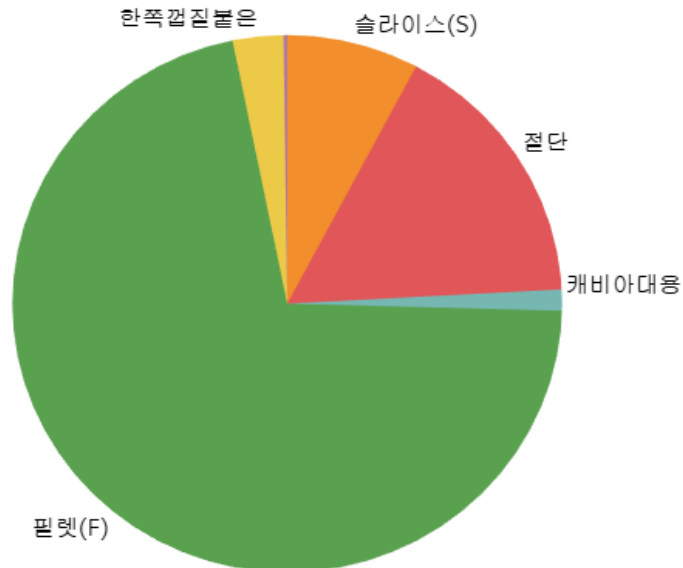
P_IMPORT_TYPE 피처 전처리



Storage(저장 형태)

: 수산물 보관 형태에 따른 분류

→ 건조, 냉동, 냉장, 염장, 활



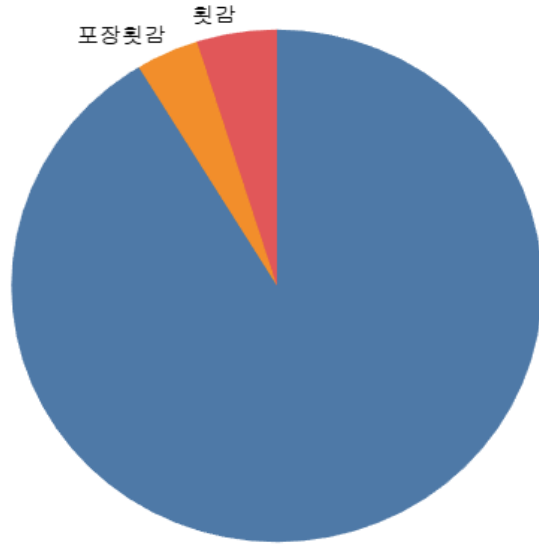
Process(가공 형태)

: 수산물 가공 방법에 따른 분류

→ 슬라이스, 절단, 캐비아대용, 필렛, 한쪽껍질붙은, 훈제



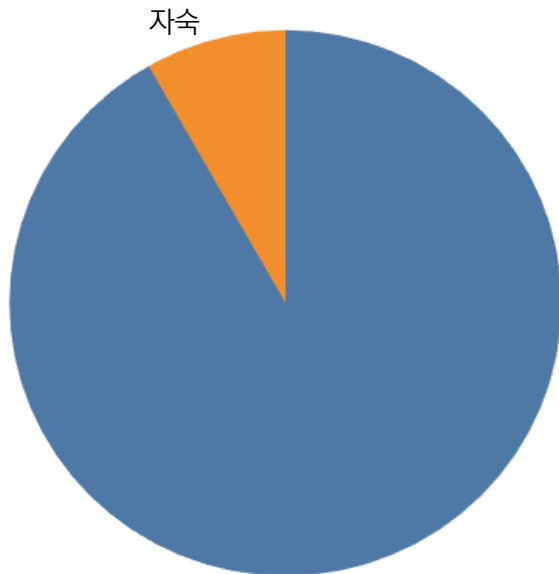
P_IMPORT_TYPE 피처 전처리



Sashimi(횟감)

: 횟감 유무에 따른 분류

→ 횟감, 포장횟감



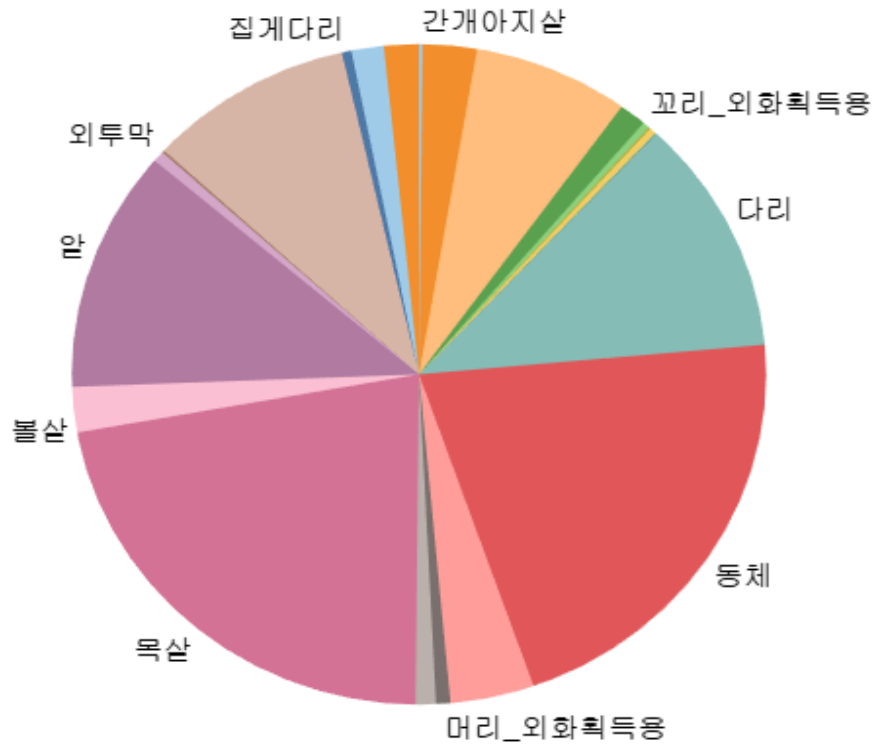
Steaming Processing(자숙)

: 김으로 찌서 익힌 것과 익히지 않은 것

→ 자숙



P_IMPORT_TYPE 피처 전처리



Part(부위)

: 수산물 부위에 따른 분류

→ 간, 개아지살, 곤이, 껍질, 꼬리_외화획득용, 난포선, 내장, 눈살, 다리, 동체, 턱 살, 머리, 머리_외화획득용, 머리살, 목살, 볼살, 알, 외투막, 지느러미, 집게다리, 창난



REG_DATE 피처 전처리

년도, 월, 날짜 분리

REG_DATE
2015-12-28
2016-01-04
2016-01-11
2016-01-18
2016-01-25
...
2019-12-02
2019-12-09
2019-12-16
2019-12-23
2019-12-30



YEAR	MONTH	DAY
2015	12	28
2016	01	04
2016	01	11
2016	01	18
2016	01	25
...
2019	12	02
2019	12	09
2019	12	16
2019	12	23
2019	12	30

REG_DATE를 **YEAR, MONTH, DAY**로 나눈 후
컬럼 추가

모델에 사용하기 편한 형태로 변환하기 위해 REG_DATE 피처를 각각
YEAR, MONTH, DAY로 나누어서 컬럼을 추가



REG_DATE 피처 전처리

년도, 월, 날짜 분리

REG_DATE	YEAR	MONTH	DAY	주차
2015-12-28	2015	12	28	1
2016-01-04	2016	01	04	2
2016-01-11	2016	01	11	3
2016-01-18	2016	01	18	4
2016-01-25	2016	01	25	5
...
2019-12-02	2019	12	02	1
2019-12-09	2019	12	09	2
2019-12-16	2019	12	16	3
2019-12-23	2019	12	23	4
2019-12-30	2019	12	30	5

데이터는 **매주 월요일** 데이터이므로
각 월을 DAY를 기준으로 1~5주차로 나누어 컬럼 추가

매월마다 같은 주차이지만 day가 다른 경우가 있어
'주차' 컬럼을 추가했다.



REG_DATE 피처 전처리

주차 column 추가

REG_DATE	P_TYPE	CTRY_1	CTRY_2	P_PURPOSE	CATEGORY_1	CATEGORY_2	P_NAME
2015-12-28	수산물	아르헨티나	아르헨티나	판매용	갑각류	새우	아르헨티나불운새우
2015-12-28	수산물	바레인	바레인	판매용	갑각류	게	꽃게
2015-12-28	수산물	바레인	바레인	판매용	갑각류	게	꽃게
2015-12-28	수산물	칠레	칠레	판매용	패류 멍게류	해삼	해삼
2015-12-28	수산물	중국	중국	판매용	어류	서대 박대 페루다	서대

오늘 < 2016.01 > ☐ 음력 ☐ 손없는날 ☒ 기념일

일	월	화	수	목	금	토
27	28	29	30	31	1 신정	2
3	4	5	6 소한	7	8	9
10 음 12.1	11	12	13	14	15	16
17	18	19	20	21 대한	22	23
24 음 12.15	25	26	27	28	29	30
31	1	2	3	4	5	6

2016년의 1월 첫째주로 처리

2015년의 REG_DATE는 2015-12-28 하루밖에 없음



LabelEncoding

CTRY_1	CTRY_2	P_PURPOSE	P_NAME	storage
아르헨티나	아르헨티나	판매용	아르헨티나붉은새우	냉동
바레인	바레인	판매용	꽃게	냉동
바레인	바레인	판매용	꽃게	냉동
칠레	칠레	판매용	해삼	건조
중국	중국	판매용	서대	냉동
...
러시아	러시아	판매용	왕게	활
중국	중국	판매용	낙지	활
러시아	러시아	판매용	대게	활
중국	중국	판매용	바지락	활
노르웨이	노르웨이	판매용	연어	냉장



CTRY_1_le	CTRY_2_le	P_PURPOSE_le	P_NAME_le	storage_le
48	49	3	172	1
30	29	3	48	1
30	29	3	48	1
66	66	3	265	0
65	64	3	158	1
...
16	15	3	187	4
65	64	3	49	4
16	15	3	64	4
65	64	3	120	4
9	9	3	178	2

모델에 문자형은 사용할 수 없기 때문에 **LabelEncoding**을 사용해 문자형을 **정수형**으로 변경



모델 선정 이유



XGBoost 장점

뛰어난 예측 성능

과적합 규제(Overfitting Regularization)

Tree pruning(트리 가지치기)

자체 내장된 교차 검증

결측치 자체 처리



CatBoost 장점

범주형 변수 처리 기법의 개선

빠른 학습 시간

모델 튜닝의 간소화

최적화된 파라미터



모델 선정 이유



과적합 규제

같은 P_TYPE, 월, 주차를 묶고 그 중 가장 많은 제조국,수출국, P_PURPOSE 조합을 3개로 묶어 사용했다.

이는 최빈값으로 전체 데이터를 대표했기 때문에 과적합 위험이 있어 과적합 규제가 있는 XGBoost를 사용했다.



결측치 자체 처리

기존에 P_IMPORT_TYPE을 5개 피처로 분할해서 사용하였다. 이 과정에서 하나의 피처에 해당하지 않는 값들이 모두

결측치로 나타나 결측치를 자체 처리할 수 있는 XGBoost를 사용했다.



성능

STEAMING_PROCESSING FEATURE에 CatBoostClassifier를 사용한 이유

Steaming_Processing 피처의 경우 CatBoost 모델을 사용했을 때 더 좋은 성능을 나타냈다.



모델 훈련 과정

수산물 이름, MONTH, 주차가 같은 것들끼리 묶은 후
제조국, 수출국, 수입목적 3가지 조합 중 가장 많은 조합 TOP3를 이용해 TEST 데이터 생성

P_NAME = 오징어, MONTH = 12, 주차 = 1의
제조국, 수출국, 목적 조합의 수

CTRY_1	CTRY_2	P_PURPOSE	
중국	중국	판매용	14
칠레	칠레	판매용	11
페루	페루	판매용	9
대만	대만	판매용	4
중국	중국	자사제품제조용	3
페루	페루	자사제품제조용	3
뉴질랜드	뉴질랜드	판매용	2
대한민국	중국	판매용	2
미국	미국	판매용	1

REG_DATE	P_NAME	CTRY_1	CTRY_2	P_PURPOSE	P_NAME
2020-01-06	오징어	페루	페루	판매용	오징어
2020-01-06	오징어	칠레	칠레	판매용	오징어
2020-01-06	오징어	중국	중국	판매용	오징어



모델 훈련 과정

카테고리를 순서대로 예측

카테고리를 독립적으로 예측하면 그 날짜에 **없는 값**과 **없는 조합**이 나올 수 있기 때문에 카테고리를 순차적으로 예측하였다.

							XGBClassifier	XGBClassifier	XGBClassifier	XGBClassifier	CatBoostClassifier
P_NAME	CTRY_1	CTRY_2	P_PURPOSE	YEAR	MONTH	주차	⇒ storage_le	⇒ part_le	⇒ process_le	⇒ sashimi_le	⇒ steaming_processing_le
오징어	중국	중국	판매용	2021	1	1	1	9	0	0	0
오징어	페루	페루	판매용	2021	1	1	1	9	0	0	1
오징어	대만	대만	판매용	2021	1	1	1	0	0	0	0
오징어	중국	중국	판매용	2021	1	2	1	9	0	0	0
오징어	페루	페루	판매용	2021	1	2	1	9	0	0	1
...
흰다리새우	태국	태국	판매용	2021	6	9	1	15	0	0	0
흰다리새우	에콰도르	에콰도르	판매용	2021	6	9	1	0	0	0	0
흰다리새우	베트남	베트남	판매용	2021	6	10	1	15	0	0	1
흰다리새우	태국	태국	판매용	2021	6	10	1	15	0	0	0
흰다리새우	에콰도르	에콰도르	판매용	2021	6	10	1	0	0	0	0



모델 훈련 과정

가격예측

카테고리를 독립적으로 예측하면 그 날짜에 **없는 값**과 **없는 조합**이 나올 수 있기 때문에 카테고리를 하나씩 순서대로 예측하였다.

XGBRegressor												P_PRICE
P_NAME	CTRY_1	CTRY_2	P_PURPOSE	YEAR	MONTH	주차	storage_le	part_le	process_le	sashimi_le	steaming_processing_le	
오징어	중국	중국	판매용	2021	1	1	1	9	0	0	0	2.567956
오징어	페루	페루	판매용	2021	1	1	1	9	0	0	1	5.952095
오징어	대만	대만	판매용	2021	1	1	1	0	0	0	0	3.664923
오징어	중국	중국	판매용	2021	1	2	1	9	0	0	0	2.567956
오징어	페루	페루	판매용	2021	1	2	1	9	0	0	1	5.952095
...
흰다리새우	태국	태국	판매용	2021	6	9	1	15	0	0	0	11.268864
흰다리새우	에콰도르	에콰도르	판매용	2021	6	9	1	0	0	0	0	6.688584
흰다리새우	베트남	베트남	판매용	2021	6	10	1	15	0	0	1	8.638298
흰다리새우	태국	태국	판매용	2021	6	10	1	15	0	0	0	11.268864
흰다리새우	에콰도르	에콰도르	판매용	2021	6	10	1	0	0	0	0	6.688584



모델 훈련 과정

같은 날짜, 이름 별로 가격 평균

REG_DATE	P_NAME	CTRY_1	CTRY_2	P_PURPOSE	CATEGORY_2	P_PRICE
2021-01-04	오징어	중국	중국	판매용	오징어	2.567956
2021-01-04	오징어	페루	페루	판매용	오징어	5.952095
2021-01-04	오징어	대만	대만	판매용	오징어	3.664923
2021-01-11	오징어	중국	중국	판매용	오징어	2.567956
2021-01-11	오징어	페루	페루	판매용	오징어	5.952095
...
2021-06-21	흰다리새우	태국	태국	판매용	새우	11.268864
2021-06-21	흰다리새우	에콰도르	에콰도르	판매용	새우	6.688584
2021-06-28	흰다리새우	베트남	베트남	판매용	새우	8.638298
2021-06-28	흰다리새우	태국	태국	판매용	새우	11.268864
2021-06-28	흰다리새우	에콰도르	에콰도르	판매용	새우	6.688584

4.0616583824157715

3.5651605129241943,
3.5651605129241943,
3.5651605129241943,
3.703249216079712,
3.703249216079712,
3.703249216079712,
3.703249216079712,
3.703249216079712,
3.703249216079712,
3.703249216079712,
3.703249216079712,
3.703249216079712,
3.6874351501464844,
3.7066452503204346,
3.7066452503204346,
3.7066452503204346,
3.7066452503204346,
3.6307477951049805,
3.6307477951049805,
3.6307477951049805,
3.7548553943634033,
4.61310338973999,
3.6307477951049805,
3.6307477951049805,
3.6307477951049805,
3.7548553943634033,
12.983365058898926,
14.986671447753906,
12.983365058898926,
12.983365058898926,
14.121719360351562,
14.121719360351562,
12.983365058898926,
12.983365058898926,
12.983365058898926,
12.983365058898926,
12.983365058898926,
12.983365058898926,
12.983365058898926,
12.983365058898926,
12.983365058898926,
13.114311218261719,
13.114311218261719,

RMSE



평균제곱근 편차

1.22985019112614



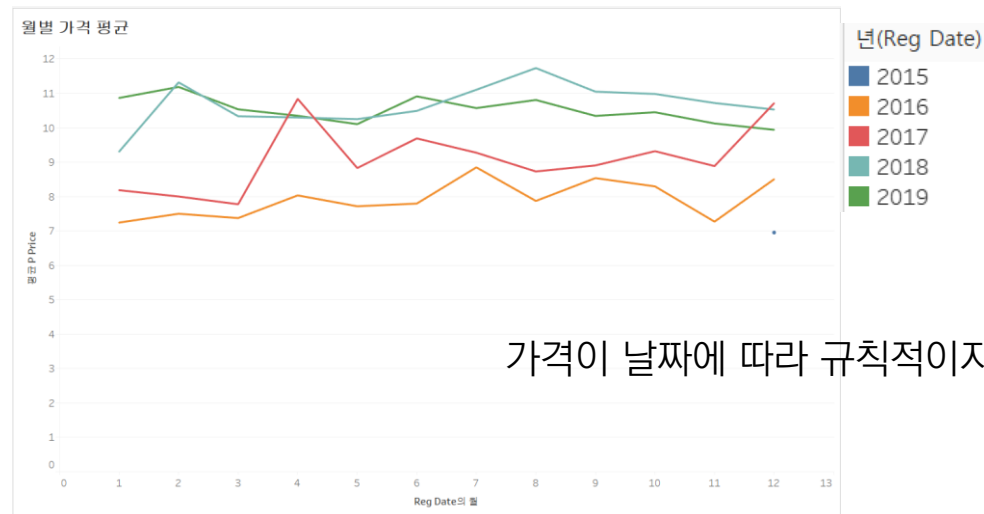
모델 훈련 과정

FEATURE IMPORTANCE 확인

```
주차 0.05021935  
MONTH 0.011416837  
YEAR 0.0050181183  
CTRY_1_le 0.07469623  
CTRY_2_le 0.017206483  
P_NAME_le 0.015486821  
storage_le 0.12454991  
part_le 0.006913515  
process_le 0.23939598  
sashimi_le 0.32472345  
steaming_processing_le 0.12651905  
P_PURPOSE_le 0.0038542484
```

IMPORTANCE가 가장 낮은 FEATURE는 P_PURPOSE

날짜와 가격의 상관관계 확인

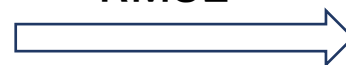


P_PURPOSE FEATURE 제거

+

날짜를 나타내는 주차, MONTH, YEAR FEATURE 제거

RMSE

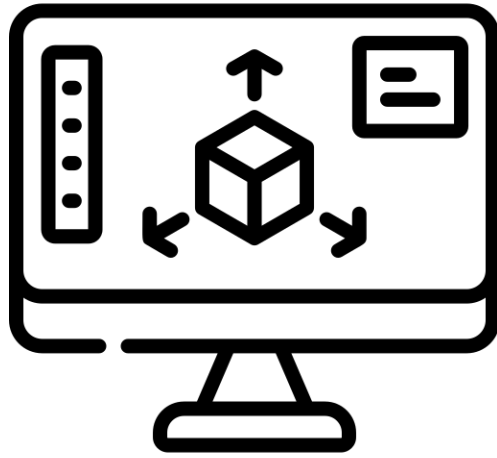


평균제곱근 편차

1.0651528904195824



기대 효과



수산물 수입가격 예측모델



수산물 수입 **의사결정** 및 수산물 **가격 선정**



수산물 수입 **가격 조정**



소비자의 **합리적인** 소비

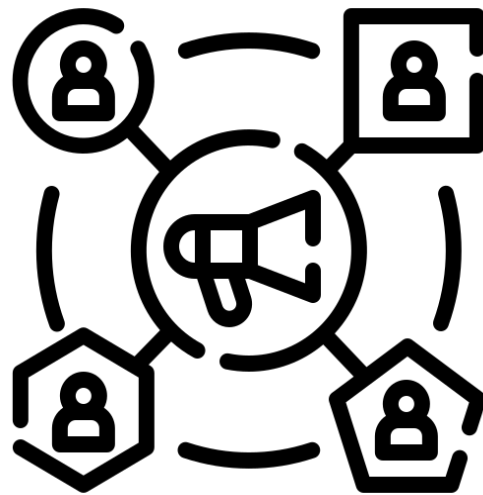


한계점 및 향후 계획



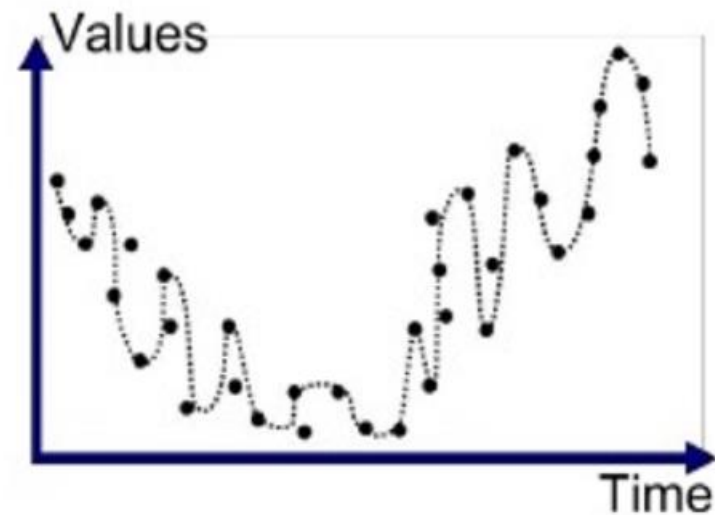
외부 데이터의 추가 활용

(ex) 날씨 데이터,
물가 상승률 데이터 등)



모델의 확장

수산물에 더해 더 다양한 분야에서의
수입가격 예측 모델의 활용



과적합 가능성

데이터를 선정해서 활용해
과적합의 가능성 존재