Final Project

온라인 리뷰가 자동차 판매량에 끼치는 영향

한국외국어대학교
경영학전공 202102858 임산별

목차

01	02	03	04
주제선정	분석방법	분석결과	기대효과

- 분석계획
- 데이터 수집
- 전처리
- 다중공선성 확인 및
- 상관관계 분석
- 다중선형회귀 분석

주제 선정 배경

소비자의 자동차구매 영향 요인

WOM
Word of Mouth
구전마케팅

일반적으로 WOM마케팅을 통해 상품 판매량 증대를 기대할 수 있다.
하지만 자동차는 WOM의 효과가 미미한 상품군임

따라서 자동차의 판매량을 예측할 수 있는 요인을 찾고자 하였고,
그 중 소비자 리뷰를 구매 영향 요인으로 검증하고자 함

자동차는 고관여 제품으로
'구전'만으로 상품구매를 결정하지 않음
정보탐색과정에서 다양한 정보를 수집하고 평가함

온라인 리뷰 데이터 활용

소비자의
리뷰 데이터

- 자동차 관련 온라인 리뷰는 차량 모델의 품질, 성능, 신뢰도를 평가하는 데 핵심적인 자료로 활용됨.
- 사용자 직접 생성 리뷰 데이터는 소비자가 구매 결정을 내릴 때 중요한 참고자료로 작용하며, 실제 판매량에도 큰 영향을 미침.

연구목표

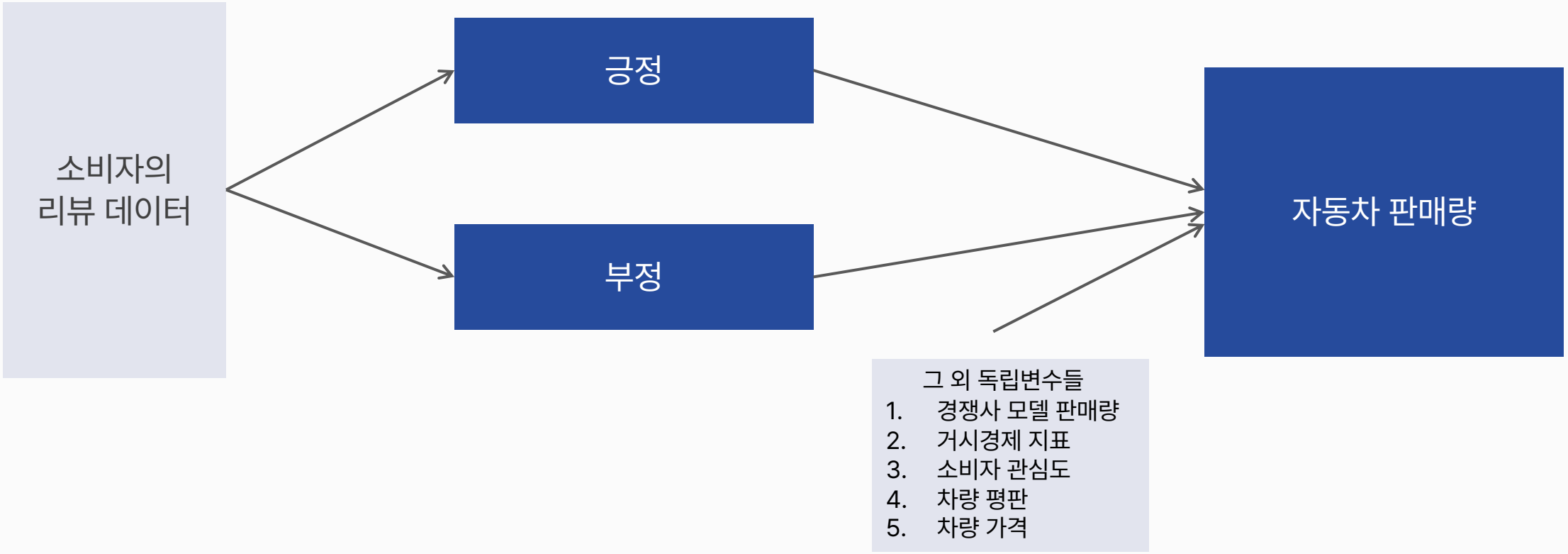
- 소비자의 자동차 구매에 영향을 끼치는 요인 분석
- 온라인 소비자 리뷰 데이터를 활용해 특정 차량 모델의 월별 판매량 예측.

주제 선정 배경

리뷰 데이터의 감성점수 활용

* 리뷰의 긍부정은 일반적으로 판매에 직접적 영향 (Qin et al., 2023)

소비자가 차량구매를 위한 정보처리 과정에서 수집하는 리뷰데이터를 긍/부정으로 인식함에 따른 판매량에 영향을 끼치는지 분석하고자함



분석방법_분석개요

데이터 수집

리뷰 데이터 수집

D1. cars.com
D2. Edmunds

그 외 변수 데이터 수집

- JD Power 점수
- 월간 미국 내 전체 인구 수
- 월간 Federal Funds Effective Rate (금리)
- 월간 Consumer Sentiment (소비자 심리 지수)
- 분기별 Median usual weekly nominal earnings.
- 월간 Unemployment Rate (실업률)
- 월간 Gas Price
- 월간 Competitor Sales (Honda CR-V, Chevrolet Equinox)
- 월간 Total Vehicle Sales
- 연간 RAV4 Price (MSRP)

데이터 전처리

리뷰 데이터 전처리

↓

리뷰 데이터 감성분석(긍/부)

↓

다중공선성 확인 : VIF 계산

↓

다중공선성 개선

↓

변수 생성 : PCA 결합

데이터 분석

상관관계 분석

↓

변수 표준화

↓

다중선형회귀 분석

분석결과

분석결과 평가 및 비즈니스 인사이트 도출

분석방법_데이터 수집

자동차 모델 선정

• 23년도 한 해 동안 미국에서 가장 많이 팔린 상위 10개 차종을 검토

=> TOYOTA RAV4

- 스테디셀러 모델 및 소비자 리뷰 데이터가 방대함
- Toyota RAV4가 23년도 미국 최다 판매 모델 4위를 기록(
  1-3 위 차량 모델은 판매량 데이터가 누락되어 획득 불가능함)

특정 한 사이트의 리뷰만 사용할 경우 리뷰의 경향이 편중될 수 있기때문에 리뷰 편향을 방지하고자 다양한 사이트의 리뷰데이터를 활용하고자함

데이터 수집

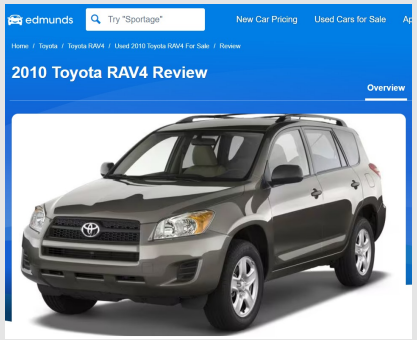
수집 방법 : 온라인 리뷰 사이트 크롤링

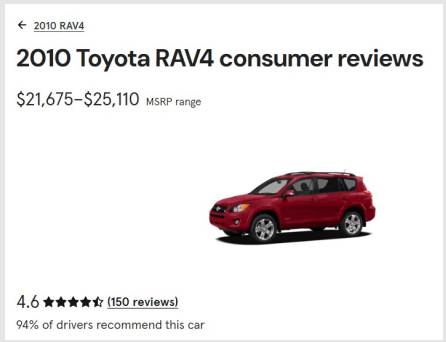
수집 사이트 후보군

사이트	선택 이유	크롤링 가능 여부
Reddit	미국의 최대 온라인 커뮤니티로, 자동차 페이지가 나누어져있 어 다양한 리뷰 데이터를 수집할 수 있음	2020년 이후 레딧의 크롤링 제한으로 크롤링 불가함
X(트위터)	전세계인이 사용하는 SNS로 대량의 리뷰 데이터들을 가지고 있음	크롤링 제한으로 불가함
Car.com	자동차 전문 리뷰 사이트로 신뢰도 높은 리뷰 데이터를 보유함	크롤링 성공
Edmunds	자동차 전문 리뷰 사이트로 신뢰도 높은 리뷰 데이터를 보유함	크롤링 성공

분석방법_데이터 수집

데이터 수집





Cars.com과 Edmunds.com에서
2010년 1월부터 2023년 12월까지
14년 간의 RAV4 모델의 리뷰 4,261개를 크롤링

변수 설정 및 변수 데이터 수집

리뷰 데이터 외에 RAV4 판매에 영향을 끼치는 변수 설정 및 데이터 수집

차량 평판 데이터	가격적 요인 데이터
JD Power 점수	RAV4 Price (MSRP)

소비자 관심도 데이터
'Toyota RAV4(SUV)' 키워드의 월간 Google Trend 점수

거시 경제 지표	자동차 시장상황 데이터
미국 내 전체 인구 수 Federal Funds Effective Rate(금리) Consumer Sentiment(소비자 심리 지수) Median usual weekly nominal earnings (중위 주급) Unemployment Rate (실업률) 월간 Gas Price	Competitor Sales - Honda CR-V - Chevrolet Equinox Total Vehicle Sales

분석방법_데이터 수집

데이터 수집

데이터 명	출처	자료 형태	자료 크기
리뷰 데이터	Car.com	CSV	717KB
리뷰데이터	Edmunds.com		446KB
도요타 RAV4 소비자 관심도	Google Trend		RAV4 데이터로 합침 15.6MB
차량 평판 데이터	JD Power		
- 월간 미국 인구 수 - Median usual weekly nominal earnings - 월간 Unemployment Rate (실업률) - 월간 Gas Price - 월간 Federal Funds Effective Rate(금리)	자체 수집 후 데이터화		
월간 Consumer Sentiment	University of Michigan		
- Competitor Sales (Honda CR-V, Chevrolet Equinox) - Total Vehicle Sales	자체 수집 후 데이터화		
RAV4 Price (MSRP)	자체 수집 후 데이터화		

분석방법_전처리

리뷰 데이터 전처리

날짜, 리뷰 , 평점 칼럼만 포함
나머지 칼럼 삭제

평점 3점 미만 -> 부정적 리뷰 => 0
평점 3점 초과 -> 긍정적인 리뷰 => 1

명확한 리뷰 긍/부정 구분을 위해
3점에 해당하는 리뷰 242개 제외
제거 후 리뷰 : 4,019 개

	date	review	rating	sentiment
0	2023-01-20	I bought this car as my first and it never bro...	5.0	1
1	2022-09-22	Dependable, comfortable, no major mechanical i...	5.0	1

리뷰 텍스트 전처리

1. 소문자 변환
2. 특수문자 및 숫자 제거
3. 토큰화
4. 불용어 제거

Train-Test 데이터 Split

Train / Test 긍정,부정 비율

Size of Training Data 3214 | Size of Test Data 804

Training Data :
Positive Sentiment 94.1
Negative Sentiment 5.8

Testing Data :
Positive Sentiment 94.1
Negative Sentiment 5.8

분석방법_감성분석

머신러닝 기반 감성점수 예측

	review	sentiment_prediction
5399	Everything I hoped for Perfect car for my need...	1
377	We've driven several Camrys for the past 20 ye...	1

	year_month	sentiment_prediction
0	2010-01	1.0
1	2010-02	1.0

월별 리뷰 긍/부정 계산

변수 설정 및 변수 데이터 수집

```
sentiment_prediction
1     0.950971
0     0.049029
```

기존 데이터의 긍/부정 비율과 유사한 비율로 예측 정확도가 높음

```
Accuracy Score -  0.9601990049751243
ROC-AUC Score -  0.6994856516484443
```

Accuracy Score: 96%

⇒ 모델이 전체 데이터 중 96%를 정확히 예측함.

ROC-AUC Score: 69%

⇒ 모델이 긍정과 부정을 구분하는 능력이 69%로,
분류 성능이 평균보다 다소 우수함.

분석방법_다중공선성 분석

독립변수간 다중공선성 확인

#	Column
0	year_month
1	RAV4 Sales
2	Google Trend
3	JD Power
4	Population in Milion
5	Consumer Sentiment
6	Federal Funds Effective Rate
7	Median usual weekly real earnings
8	Unemployment Rate
9	Gas Price
10	Competitor Sales (Honda CR-V)
11	Competitor Sales (Chevrolet Equinox)
12	Total Vehicle Sales (In milions)
13	Price (MSRP)
14	sentiment_prediction

종속변수

독립변수

VIF 계산

High Multicollinearity (VIF >= 10):

	feature	VIF
0	Google Trend	91.0564
1	JD Power	2231.18
2	Population in Milion	24167.1
3	Consumer Sentiment	223.564
5	Median usual weekly real earnings	20847.9
6	Unemployment Rate	91.7945
7	Gas Price	114.292
8	Competitor Sales (Honda CR-V)	42.8972
9	Competitor Sales (Chevrolet Equinox)	32.2918
10	Total Vehicle Sales (In milions)	457.439
11	Price (MSRP)	2058.11
12	sentiment_prediction	43.1688

Moderate Multicollinearity (5 <= VIF < 10):

feature	VIF

Low Multicollinearity (VIF < 5):

	feature	VIF
4	Federal Funds Effective Rate	3.39046

VIF < 5

다중공선성이 거의 없는 것으로 간주

5 <= VIF < 10

다중공선성이 있을 수 있지만 심각하지는 않음

VIF >= 10:

다중공선성이 높음

⇒ **VIF가 10상인 변수가 13개로 다중공선성이 다소 높음**

⇒ **변수 제거 및 결합을 통해 다중공선성 감소작업이 필요함**

분석방법_다중공선성 분석

비슷한 정보를 담고있는 변수들 제거 :
변수들를을 제거 및 추가한 후 VIF계산을 진행하며 최종 변수들을 선정함

제거 변수

1. Population in Milion
2. Median usual weekly real earnings
3. JD Power
4. Unemployment Rate

PCA결합 변수 생성

[Price_Factor]
Gas Price + Price (MSRP)

[Consumer_Interest]
Goole Trend + Consumer Sentiment

[Competitor_Sales]
Competitor Sales (Honda CR-V) +
Competitor Sales (Chevrolet Equinox) +
Total Vehicle Sales (In milions)'

최종 변수 목록 및 VIF

	Variable	VIF
0	Federal Funds Effective Rate	2.368529
1	sentiment_prediction	1.713421
2	Consumer_Interest	1.779507
3	Competitor_Sales	1.543694
4	Price_Factor	1.329435

* Federal Funds Effective Rate
* sentiment_prediction

[새로운 변수]
- Consumer_Interest
- Competitor_Sales (Competitor Sales (Honda CR-V), Competitor Sales
- Price_Factor

VIF 모두 5이하로 낮은 다중공선성을 가짐

분석방법_다중회귀 분석

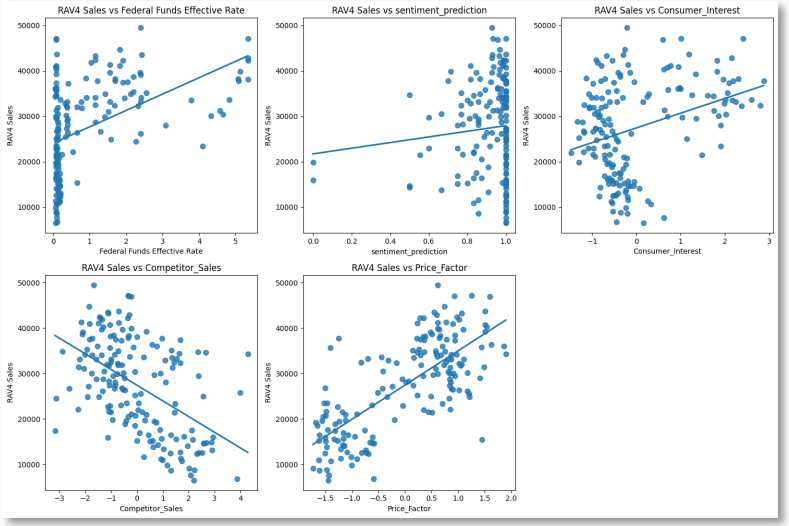
상관관계 분석

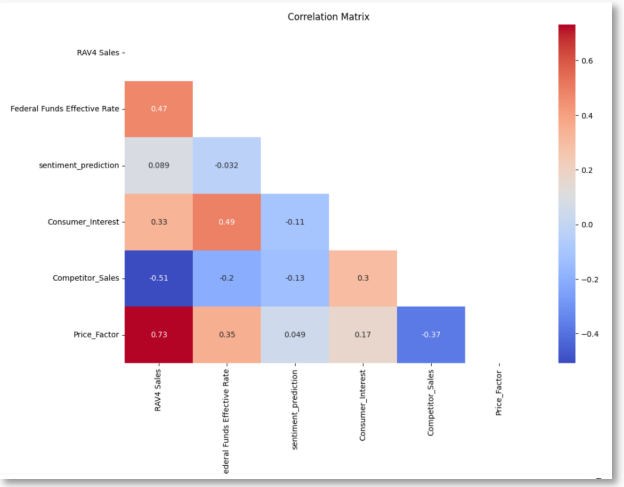
```
# 1개월 시간차 데이터 생성
data['RAV4 Sales'] = data['RAV4 Sales'].shift(-1)
```

종속변수와 모든 독립변수는 1개월 시차 적용하여
독립 변수를 통해 Toyota RAV4 차량의 익월 판매량을 예측하도록 설정

실제 비즈니스 환경에서 자동차 판매 예측 모델을 활용하기 위해서는 최소한 1개월 이후의 월별 판매량을 예측할 수 있어야 하며,
일부 독립변수는 월
말일 이후에 공시되기 때문

종속변수와 독립변수 간 산포도를 통한 선형 관계 의미 분석





Federal Funds Effective Rate(금리)와 종속변수 : 약한 양의 상관관계

Price Factor와 종속변수 : 강한 양의 선형 관계

Sentiment prediction과 종속변수

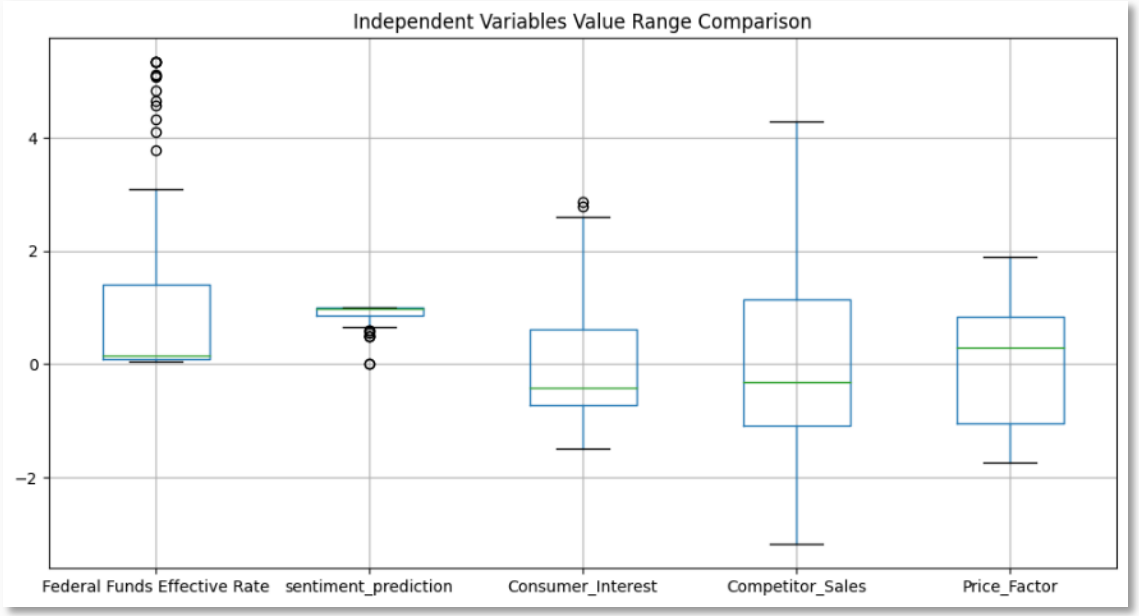
: 매우 약한 선형 관계를 보이며 , sentiment prediction 대부분의 값이 1에 몰려 있어, 종속 변수와 유의미한 상관관계 없음

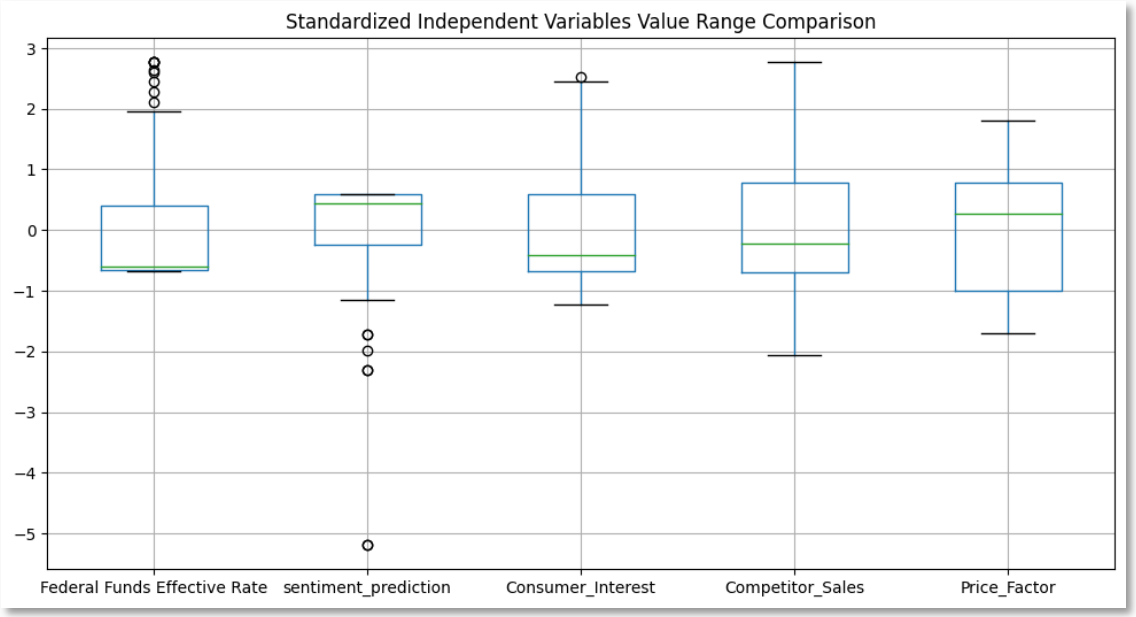
Consumer Interest와 종속변수 : 약한 양의 상관관계.

Competitor Sales와 종속변수 : 음의 선형 관계 보임 -> 경쟁 모델 판매량과 RAV4 판매량 반비례 함을 의미

분석방법_다중회귀 분석

변수 표준화





RAV4 판매량 예측을 위해 여러 독립 변수들을 사용. 각 변수의 값 범위를 확인하기 위해 박스 플롯을 작성. 그 결과, 변수 간 값의 범위 차이가 크게 나타났으며, 이는 모델의 성능에 영향을 줄 수 있음을 확인

이에 따라 변수 간의 균형을 맞추기 위해 표준화 실시.
StandardScaler를 사용하여 각 변수의 평균을 0, 표준편차를 1로 조정
표준화된 데이터의 박스 플롯을 작성하여 결과를 시각화한 결과, 모든 변수들이 평균 0, 표준편차 1의 분포를 가지는 것을 확인

분석방법_다중회귀 분석

회귀 분석 모델간 비교

산포도를 통해 독립변수 모두 종속변수와
강하지 않은 선형 관계임을 확인함
단순한 선형 회귀 분석만으로는 정확한 예측이 어려울 수 있어
다양한 선형 회귀 모델과
비선형 모델을 사용해 정확한 예측을 하고자함.

랜덤 포레스트 | Lasso | SVR

Linear | Ridge

성능 해석 지표

평균 제곱근 오차 (RMSE)
: MSE의 제곱근으로, 해석이 더 쉬운 형태로 변환한 것
값이 낮을수록 모델 성능이 좋음을 의미

평균 절대 오차 (MAE)
: 예측 값과 실제 값의 절대 오차의 평균을 측정
값이 낮을수록 모델의 예측 정확도가 높음을 의미

결정 계수 (R^2)
:모델이 데이터를 얼마나 잘 설명하는지를 나타내며
1에 가까울수록 모델의 설명력이 높음을 의미한다.

회귀 모델 결과 값 비교

Model	MSE	RMSE	MAE	R-squared
Random Forest	2.647479e+07	5145.366	4034.784	0.770
Ridge Regression	2.956134e+07	5437.034	4248.548	0.744
Lasso Regression	2.965996e+07	5446.096	4249.502	0.743
Linear Regression	2.966202e+07	5446.285	4249.488	0.743
Support Vector Regressor	1.276601e+08	11298.678	9305.622	-0.108

랜덤 포레스트 모델은 모든 수치 (MSE, RMSE, MAE, R^2 모두)에서
다른 모델들보다 우수한 성능 보임
SVR 모델 음의 결정 계수(R^2) 값을 보여, 데이터에 대한 설명력이 매우 낮음
릿지 회귀, 라쏘 회귀, 선형 회귀 모델: 거의 유사한 성능을 보였지만,
랜덤 포레스트보다 성능이 떨어짐
=> 랜덤 포레스트를 모델로 선정

랜덤 포레스는 비선형적 특성을 처리하는데 강점을 가지며,
변수 중요도를 제공하여 분석결과를 쉽게 해석할 수 있음

분석결과

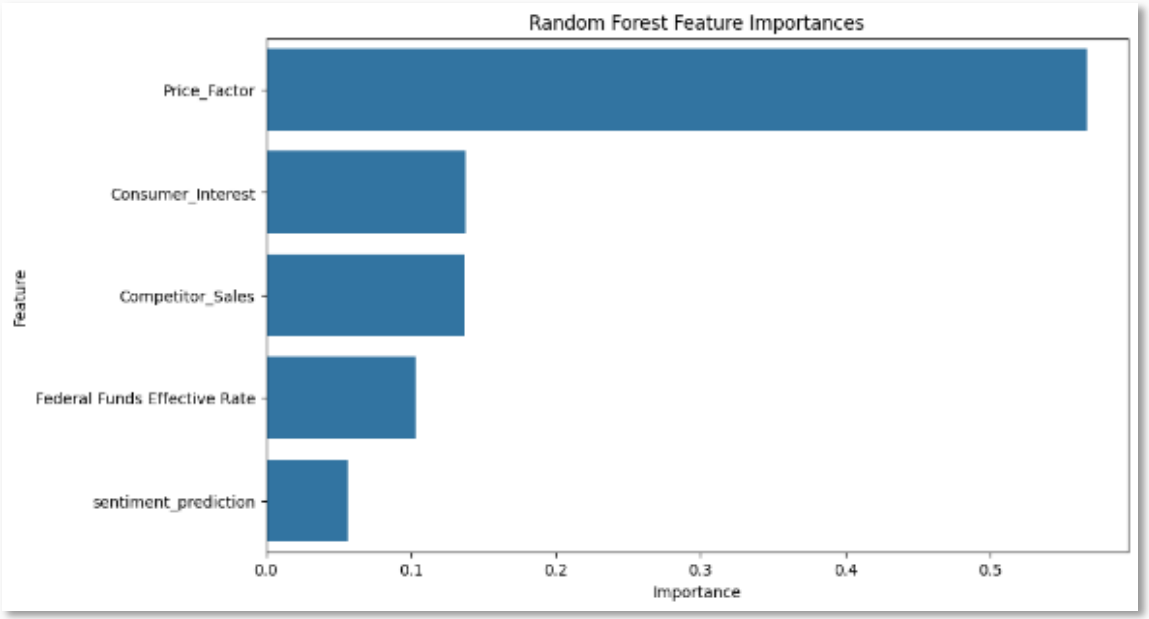
각 변수들의 상대적 중요도 분석 설명

각 변수들의 상대적 중요도 분석 설명

Senitment prediction (소비자 감성)은 가장 낮은 중요도

이는 **소비자 온라인 리뷰가 월별 자동차 판매량에 미치는 영향이 상대적으로 미미함을 시사**

	Feature	Importance
4	Price_Factor	0.579628
3	Competitor_Sales	0.137869
2	Consumer_Interest	0.135780
0	Federal Funds Effective Rate	0.099211
1	sentiment_prediction	0.047512



<Random Forest 분석 시 독립 변수들의 Importance 수치>

1. 회귀 모델의 보편성을 보장할 수 없음

본 연구는 미국 내의 특정 차량 모델을 대상으로 진행되어, 해당 차량 판매량 예측 모델이 과적합 (Overfitting)의 문제를 지닐 수 있음.
이와 더불어 구글 트랜드는 미국 등 구글의 시장 점유율이 독보적인 일부 국가에서만 유효하게 적용되는 독립변수라는 점에도 주목하여야 함

2. 감성분석 결과 긍정이 극단적으로 높은 값을 가짐

감성 분석 결과 긍정 및 부정 비율은 각각 95.3% 와 4.7%로 기록. 이를 통해 소비자 리뷰 데이터를 크롤링한 커뮤니티가 해당 차량에 우호적인 경향성을 가져 편향되었다고 추측 가능. 가능하다면, 후속 연구에서는 X (이전 twitter) 등의 대형 소셜 미디어와 같이 많은 수의 유저를 갖추거나 JD Power와 같이 많은 수의 객관적인 패널 데이터를 지닌 사이트의 데이터를 활용하는 방안을 고려

3. 본 연구에서는 모든 독립 변수가 종속 변수와 인과 관계를 가질 것이라고 가정하였으나, 이는 실제와 차이가 있을 가능성이 多

예를 들어 TOYOTA RAV4 모델의 판매량과 경쟁 차량 판매량 및 미국 내 전체 차량 판매량은 실제로는 명확한 인과 관계를 가지지 않을 가능성이 多

4. 판매량은 소비자 수요를 온전히 반영하지 않음

자동차 판매량은 재고 상황에 따라 결정되는데, 재고가 없다면 수요가 존재하더라도 자동차 판매량은 0을 기록할 수 있기 때문