

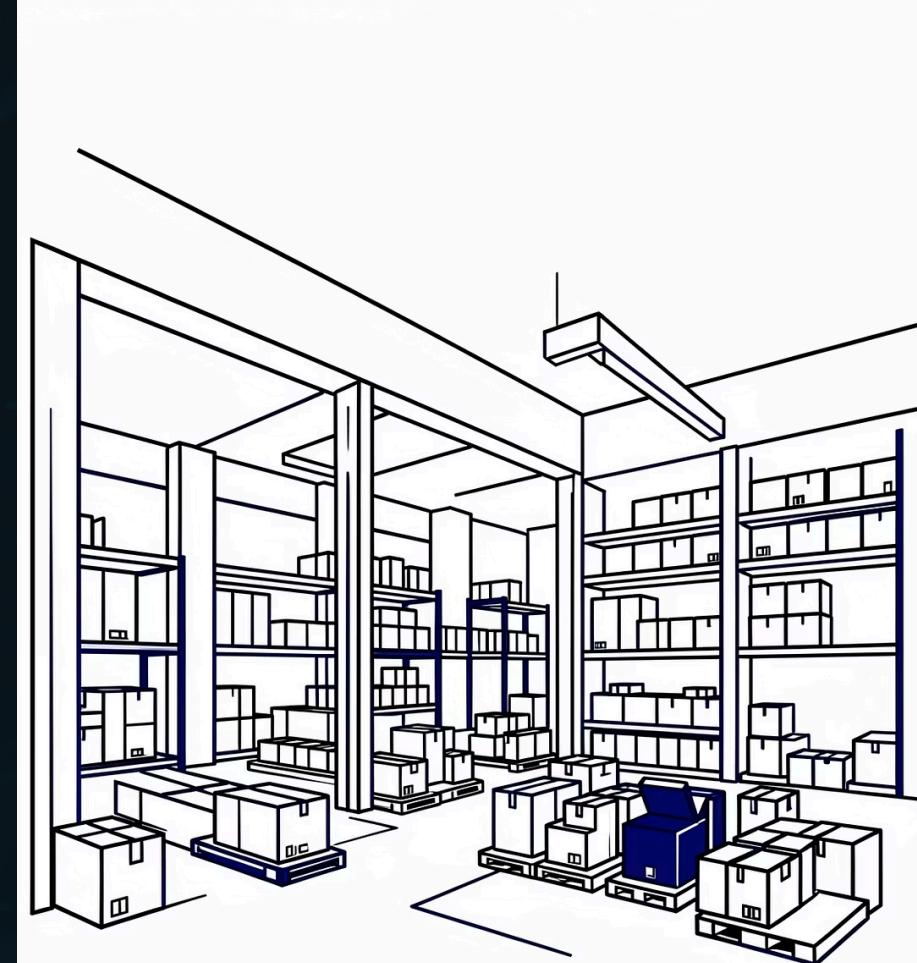
E-Commerce Shipping Analytics: 배송 지연 예측 모 델링

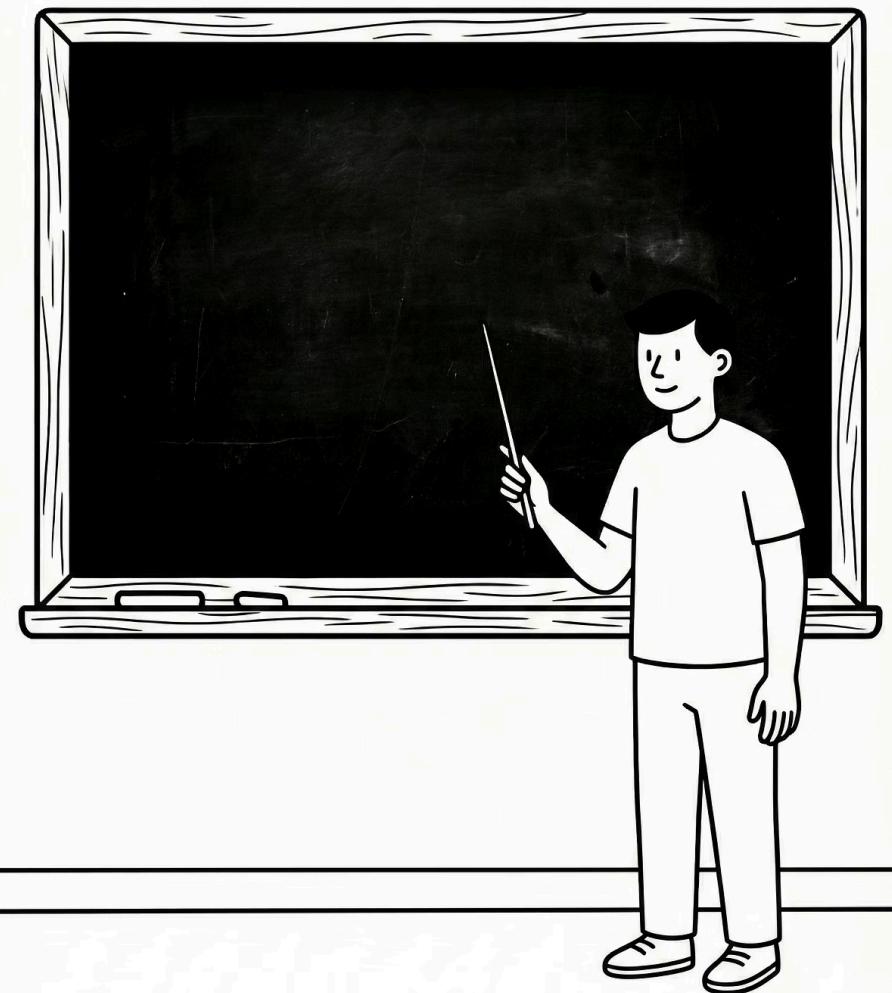
데이터 기반의 물류 최적화 및 고객 만족도 제고 전략

배송의 민족

2026년 2월 13일

박찬영, 이경욱, 지소연, 신우철





목차

1. 프로젝트 개요
2. 데이터 소개
3. 탐색적 데이터 분석 EDA
4. 전처리 및 특성공학
5. 모델링 및 성능 평가
6. 결론 및 제언



프로젝트 개요 및 목표

배경

E-커머스 시장이 급성장하면서 '배송 정시성'이 고객 경험의 핵심 요소로 부상하고 있습니다.

문제 정의

배송 지연은 고객 이탈의 주요 원인이며, 물류 비용 증가와 브랜드 신뢰도 하락을 초래합니다.

프로젝트 목표

머신러닝 모델을 통해 배송 지연 여부(1: 지연, 0: 정시)를 사전에 예측하고 선제적 대응 체계를 구축합니다.

기대 효과

배송 지연 최소화를 통한 고객 만족도 향상 및 운영 효율성 제고로 경쟁 우위를 확보합니다.

분석 프로세스 (Pipe line)

데이터 품질 체크

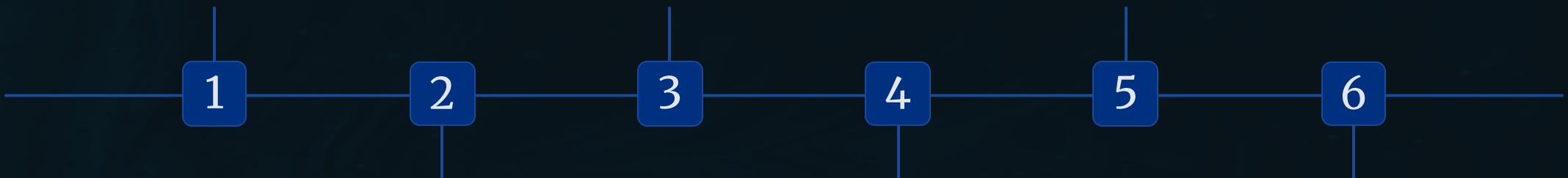
결측치 (Null)
중복값
데이터 타입 (수치형 / 범주형)

전처리

머신러닝 학습용 변환: 범주형 변수는 인코딩(Label/One-Hot) 처리하고, 수치형 변수는 모델 성능 향상을 위해 스케일링(Standardization)을 적용

모델링 (Modeling)

예측 모델 구축 및 고도화: Random Forest, XGBoost 등 다양한 알고리즘으로 베이스라인을 구축하고, Optuna를 활용해 최적의 하이퍼파라미터를 탐색



탐색적 데이터 분석 (EDA)

데이터 패턴 시각화 : 타겟 변수(지연 여부)의 비율을 확인 상관 관계를 시각적으로 분석하여 인사이트 도출

특성 엔지니어링

변수 최적화: 학습을 방해하는 이상치(Outlier)를 식별하여 처리하고, 분석에 유의미한 파생 변수를 생성하거나 불필요한 변수를 제거

평가 (Evaluation)

성능 검증: 정확도(Accuracy)와 F1-Score를 기준으로 모델을 평가하며, 혼동 행렬(Confusion Matrix)과 ROC-AUC를 통해 예측의 신뢰성을 검증합니다.

팀 역할 분담 (R&R)



이경욱

데이터 분석가 (Data Analyst)

주요 임무: EDA 및 인사이트 도출

- 변수별 분포 및 상관관계 분석
- 핵심 요인 시각화
- 피처 아이디어 제공
- 비즈니스 인사이트 작성



지소연

데이터 엔지니어 (Data Engineer)

주요 임무: 전처리 및 피처 엔지니어링

- 범주형/수치형 데이터 처리
- 파생 변수 생성
- 데이터 불균형 확인
- 파이프라인 구축



박찬영

머신러닝 모델러 (ML Modeler)

주요 임무: 모델 구축 및 튜닝

- 다양한 알고리즘 학습
- AutoGluon 활용
- 하이퍼파라미터 튜닝
- 성능 결과 정리



신우철

프로젝트 매니저 & QA

주요 임무: 총괄 및 검증

- 일정 및 GitHub 관리
- Feature Importance 검증
- 논리적 흐름 체크
- 팀 간 의견 조율

데이터셋 소개

변수명 (Variable)	설명 (Description)	데이터 타입 (Data Type)	결측치 (Missing Values)
ID	고유 고객 ID 번호	수치형 (Int64)	0 (0%)
Warehouse_block	물류 창고 구역 (A, B, C, D, E)	범주형 (Object)	0 (0%)
Mode_of_shipment	배송 운송 수단 (Ship, Flight, Road)	범주형 (Object)	0 (0%)
Customer_care_calls	배송 관련 문의 전화 횟수	수치형 (Int64)	0 (0%)
Customer_rating	고객 만족도 평점 (1: 최저 ~ 5: 최고)	수치형 (Int64)	0 (0%)
Cost_of_the_product	상품 가격 (USD 달러 기준)	수치형 (Int64)	0 (0%)
Prior_purchases	이전 구매 횟수	수치형 (Int64)	0 (0%)
Product_importance	상품 중요도 (Low, Medium, High)	범주형 (Object)	0 (0%)
Gender	고객 성별 (Male, Female)	범주형 (Object)	0 (0%)
Discount_offered	해당 상품에 제공된 할인율	수치형 (Int64)	0 (0%)
Weight_in_gms	상품 무게 (g 단위)	수치형 (Int64)	0 (0%)
Reached_on_time_Y_N	정시 도착 여부 (0: 정시, 1: 지연)	타겟/수치형 (Int64)	0 (0%)

E-Commerce Shipping Dataset

출처: Kaggle 공개 데이터셋

규모: 총 10,999건의 주문 데이터

- 변수 설명
- 수치형 / 범주형 데이터
- 결측치 : 0

주요 특징 변수 (Features)



고객 정보

- 고객 ID 및 성별
- 이전 구매 횟수
- 고객 등급 (Rating)



물류 정보

- 창고 블록 위치
- 배송 수단 (Ship/Flight/Road)
- 상품 무게 (Weight_in_gms)

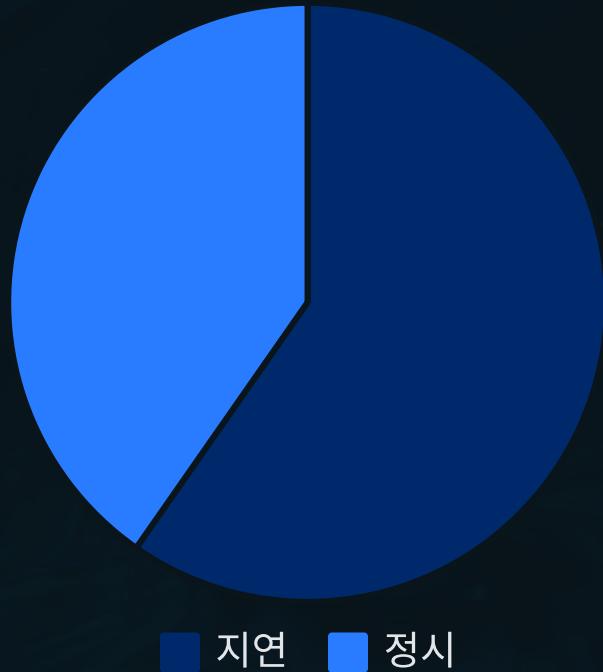


주문 정보

- 상품 가격 (Cost_of_the_Product)
- 할인율 (Discount_offered)
- 상품 중요도 등급

이러한 다양한 특징들을 종합적으로 분석하여 배송 지연 패턴을 파악하고 예측 모델의 정확도를 향상시킵니다.

타겟 변수 분포 (Class Balance)



주요 인사이트

Taget 변수

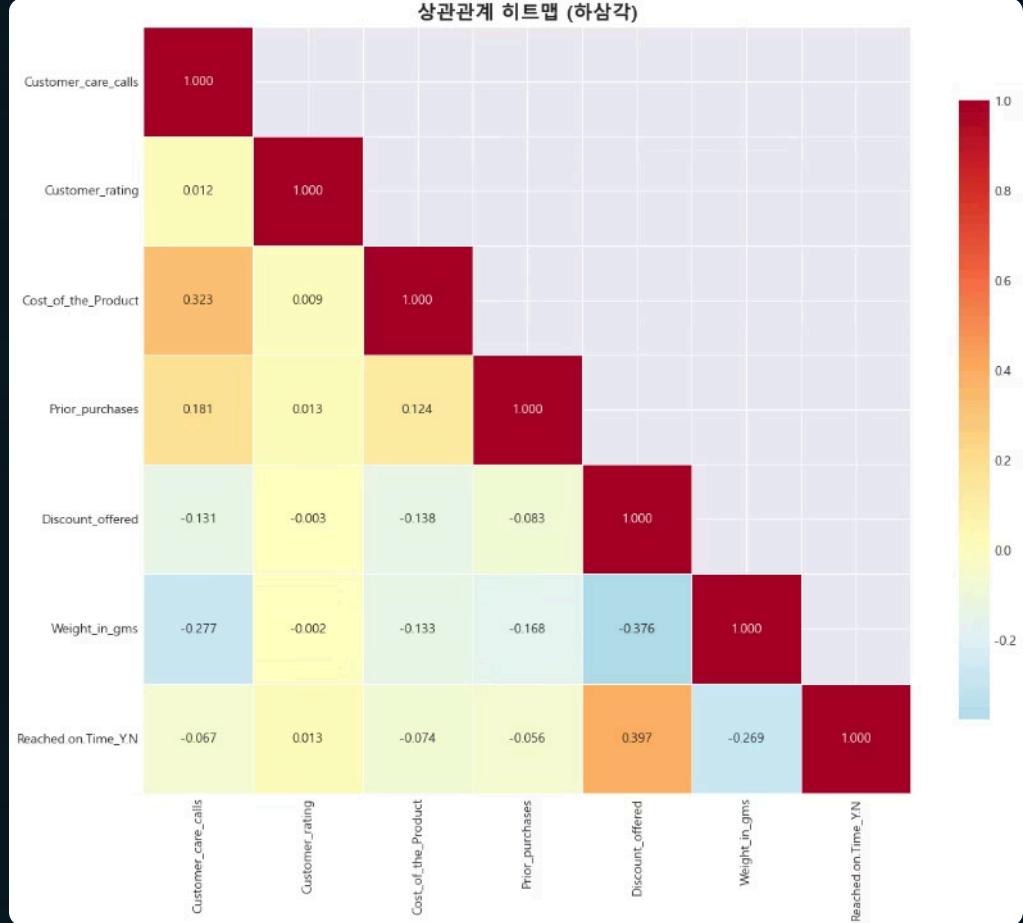
`Reached.on.Time_Y.N`

지연 59.7%

정시 40.3%

- 데이터가 비교적 균형 잡혀 있으나, 지연 비율이 더 높음
- 무조건 찍어도 60% 정확도이므로 모델은 이보다 훨씬 높아야 함

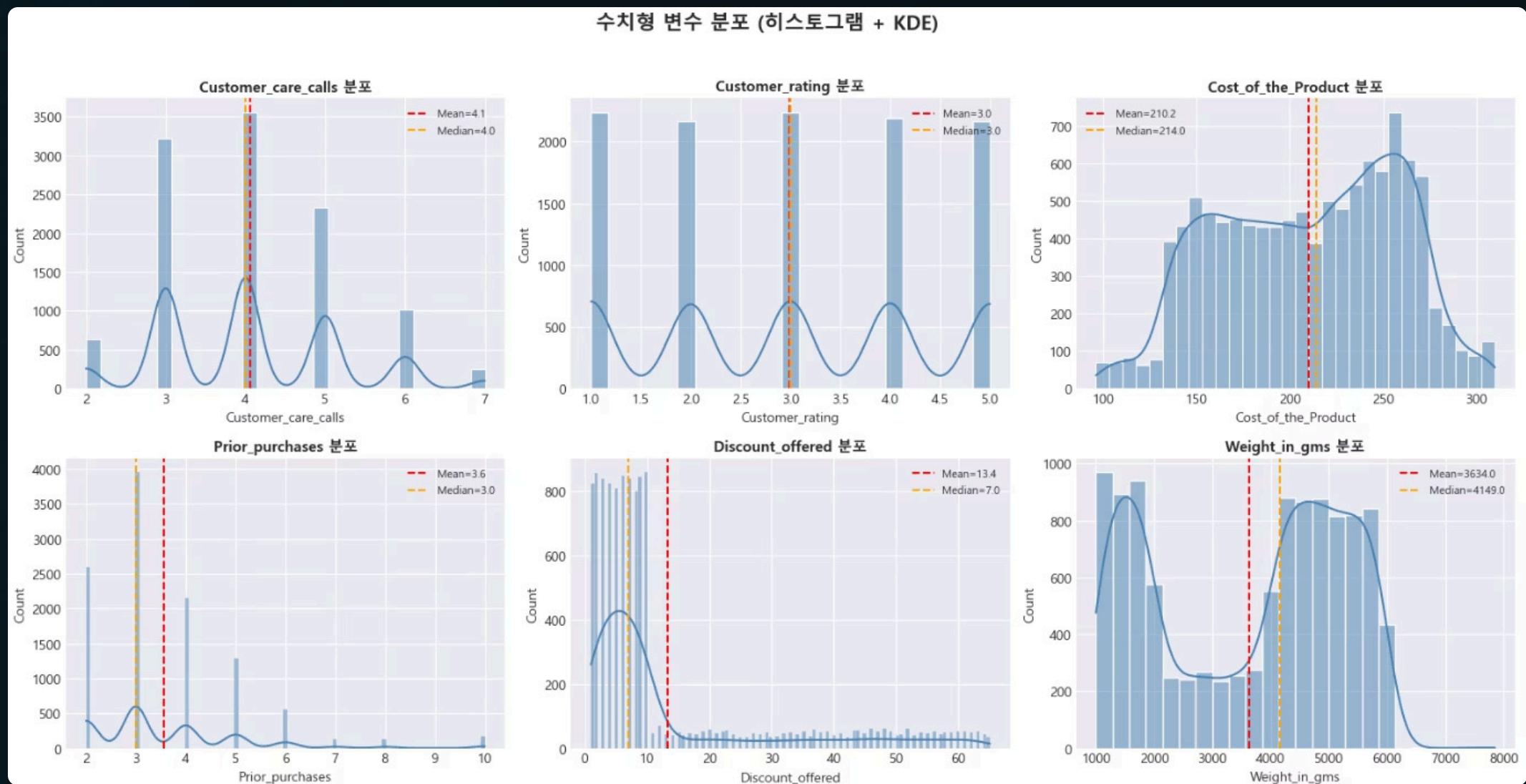
[탐색적 데이터 분석] 주요 변수 간 상관관계 분석



핵심 발견 (Key Findings)

- 🚩 할인율(Discount)의 경고 (Corr: 0.40): 배송 지역 여부와 가장 강한 양의 상관관계를 보입니다. 즉, "할인폭이 클수록 배송이 지연될 확률이 높다"는 강력한 징후가 발견되었습니다.
- 📦 무게(Weight)의 연관성 (Corr: -0.27): 무게 또한 배송 지역과 뚜렷한 음의 상관관계를 보입니다. 이는 무게가 가벼울수록, 혹은 특정 무게 구간에서 지역 이슈가 발생할 수 있음을 시사합니다.
- 📞 가격과 고객 문의 (Corr: 0.32): 상품 가격(Cost)이 높을수록 고객 센터 문의(Calls)가 많아지는 경향이 확인되었습니다.

수치형 변수 요약 (Boxplot)

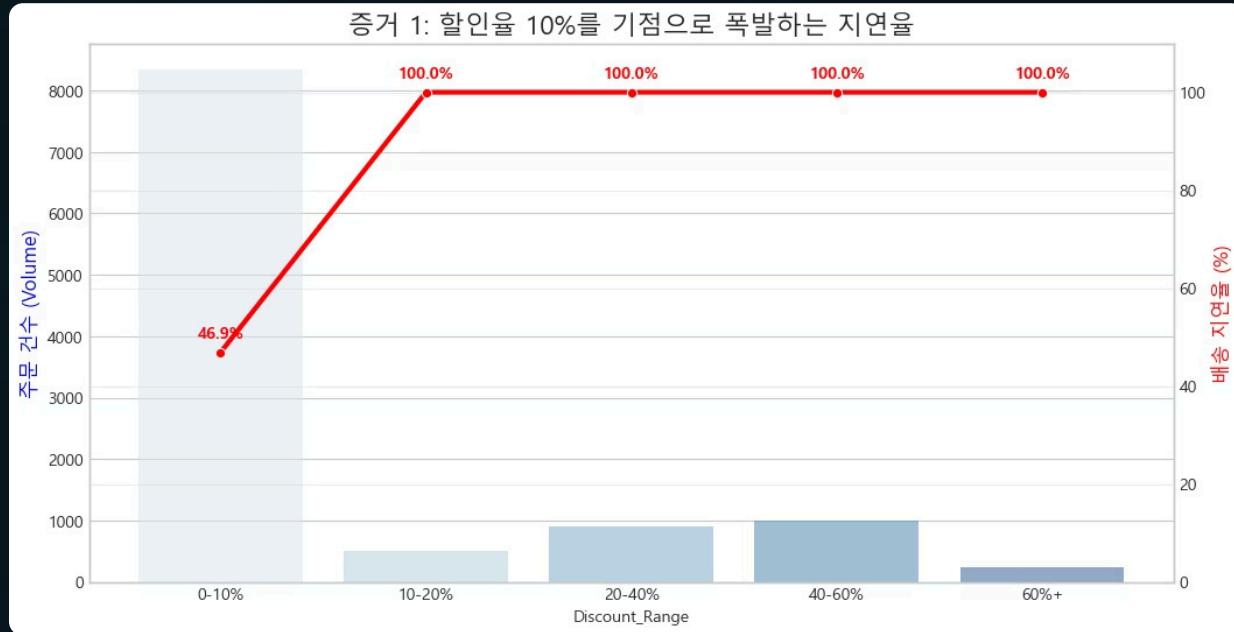


주요 인사이트: 이상치 (Outlier) 분석

수치형 변수들의 분포와 이상치 여부를 Boxplot을 통해 시각적으로 확인했습니다.

- 대부분의 변수에서 이상치가 관찰되었으며, 이는 데이터 전처리 단계에서 신중하게 다루어져야 할 요소입니다.
- 특히 **Discount_offered** 변수에서 극단적인 이상치들이 다수 발견되어, 모델의 성능에 큰 영향을 미칠 수 있으므로 별도의 처리 방안이 필요합니다.
- 이상치는 모델 학습 시 편향을 유발하거나 예측 성능을 저하시킬 수 있습니다.

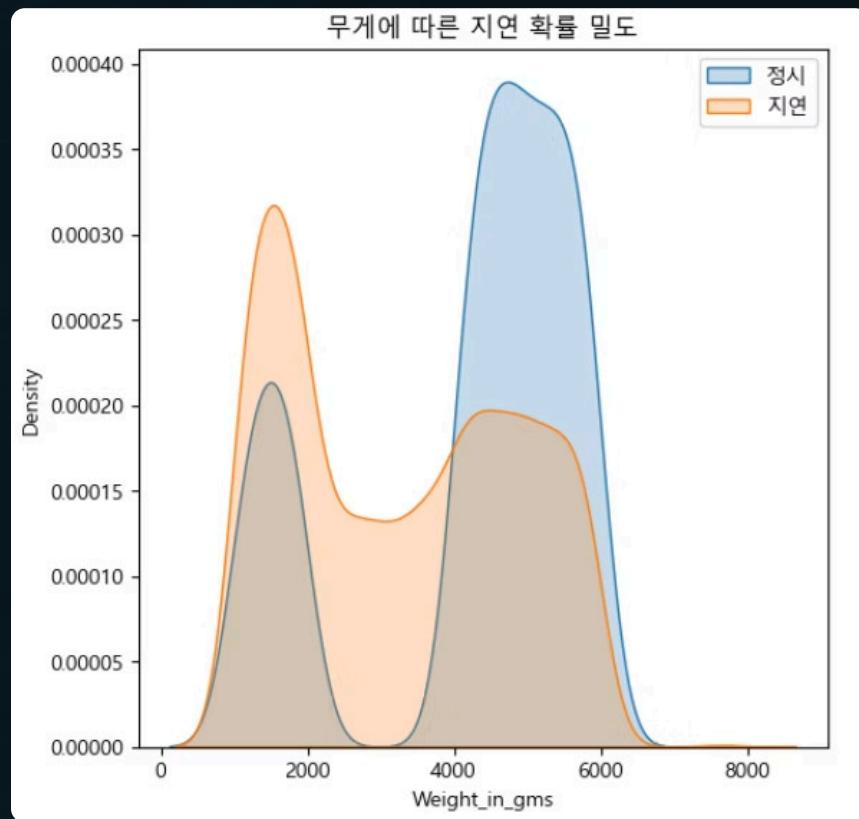
할인율(Discount)과 배송 지연의 상관관계 분석



비즈니스 인사이트

- **할인율 10% 미만 집중:** 전체 주문의 대다수가 0~10%의 낮은 할인율 구간에 밀집해 있습니다.
- **고할인율(High Discount)의 의미:** 10% 이상의 할인이 적용된 상품은 '재고 정리(Clearance)' 또는 '프로모션 상품'일 가능성이 높습니다.
- **가설 설정:** 재고 정리(Clearance) 또는 프로모션 상품 할인율이 높은 상품은 물류 우선순위가 밀려 배송이 지연될 확률이 높지 않을까?

상품 무게(Weight)와 배송 지연의 관계



핵심 인사이트

상품 무게별 배송 지연 집중 구간 발견

- **특정 구간 집중:** 1~2kg(경량) 및 4~6kg(중량) 구간에서 배송 지연 빈도 급증.
- **원인 추정:** 특정 무게군에 대한 물류 분류 오류 또는 특수 포장 프로세스상의 병목 현상 가능성.
- **전략 방향:** 해당 무게 구간의 프로세스 집중 점검 및 배송 지연 예측 모델 정교화.

상품 가격 분포 분석 (Cost Distribution Plot)

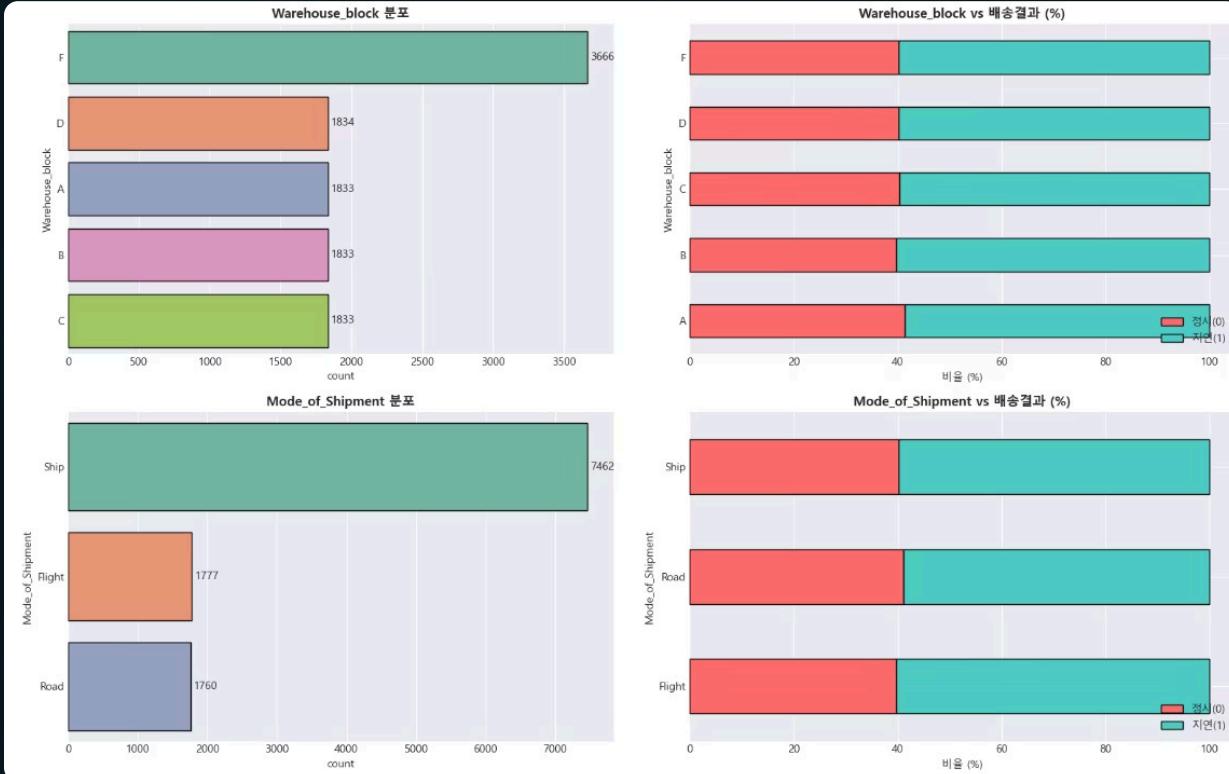


핵심 인사이트

이 그래프는 히스토그램과 밀도 곡선(KDE)을 결합하여 가격대의 흐름을 보여줍니다.

- 고른 분포:** 약 100달러에서 300달러 사이의 상품들이 특정 가격대에 치우치지 않고 일정하게 분포되어 있습니다.
- 해석:** 이는 데이터 내에서 '저가형'이나 '프리미엄' 상품이 골고루 포함되어 있음을 의미하며, 배송 지연의 핵심 변수가 가격보다는 다른 요인(무게, 운송 수단 등)일 가능성을 시사합니다.

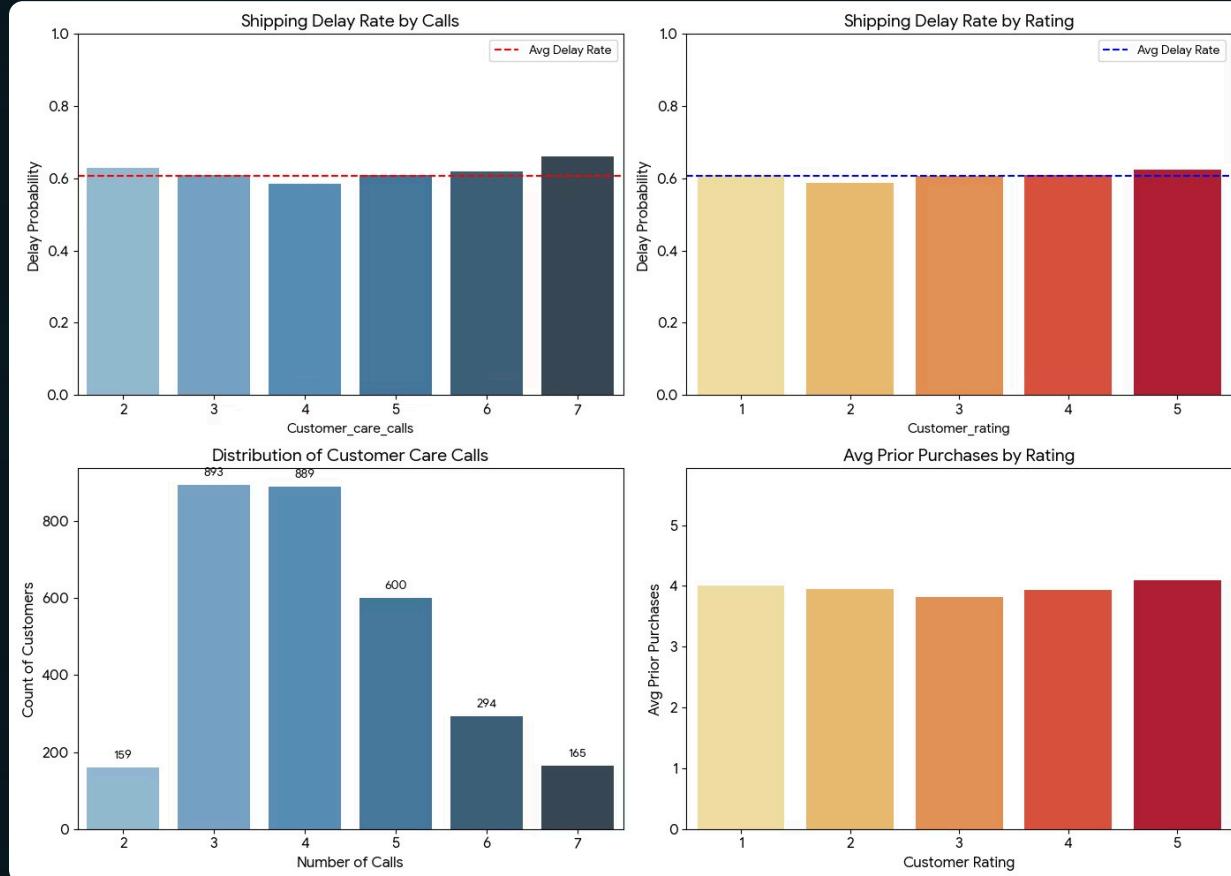
범주형 변수 분석 1: 창고(Warehouse) & 운송편(Shipment)



주요 인사이트

- 창고별 지연율**: 데이터 분석 결과, 창고 블록(A-F)별 배송 지연율에는 **유의미한 차이가 발견되지 않았습니다**. 이는 특정 창고의 운영 문제가 전반적인 지연에 큰 영향을 미치지 않음을 시사합니다.
- 운송 수단별 지연율**: 배송 수단(Ship, Flight, Road)에 따라서는 지연율에 다소 차이가 관찰되었으며, 이는 각 운송 수단의 특성 및 외부 요인(기상 조건, 교통량 등)의 영향을 받을 수 있음을 나타냅니다.
- 추가 분석 필요**: 창고 자체보다는 창고 내의 특정 프로세스나 운송 수단과의 조합에서 지연 원인을 심층적으로 탐색 할 필요가 있습니다.

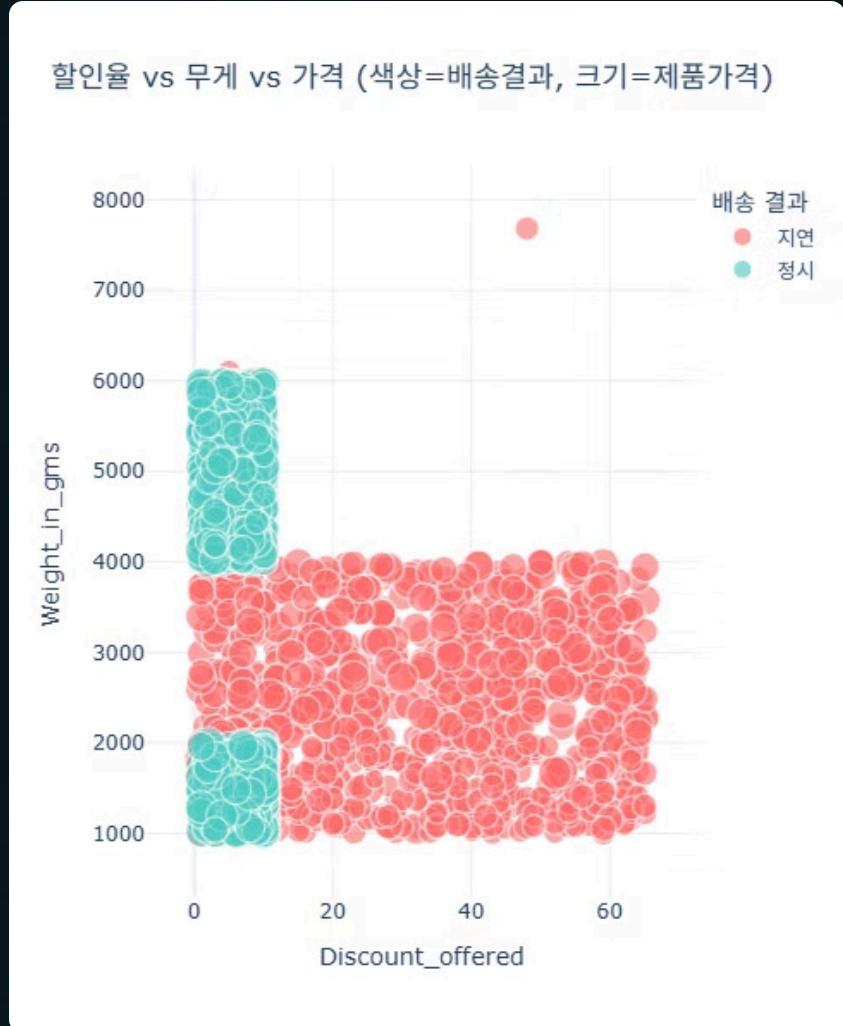
범주형 변수 분석 2: 고객 등급 & 문의 횟수



주요 인사이트

- 고객 문의 횟수(Customer_care_calls)와 지연:** 문의 전화가 많은 고객의 주문이 배송 지연될 확률이 높은지 분석합니다. 이는 고객의 불만 수준이나 주문의 복잡성과 연관될 수 있습니다.
- 고객 등급(Customer_rating)과의 연관성:** 고객 등급이 배송 지연에 어떤 영향을 미치는지 살펴봅니다. 높은 등급의 고객에게는 더 빠르고 정확한 배송 서비스가 제공되는지, 혹은 등급과 지연 간에 유의미한 관계가 있는지 확인합니다.
- 비즈니스 시사점:** 만약 문의 횟수가 많은 고객의 지연율이 높다면, 이들의 주문을 사전에 감지하고 추가적인 관리를 통해 불만을 완화할 수 있는 전략을 수립할 수 있습니다.

다면량 분석: 무게와 할인율에 따른 배송 지연

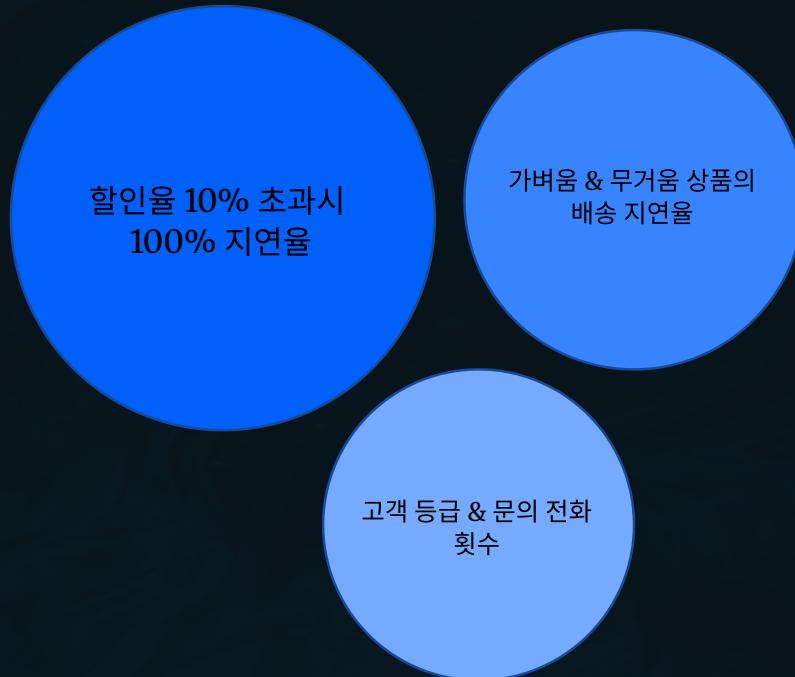


주요 인사이트

배송 지연의 결정적 임계값(Threshold) 발견

- 🚩 마의 10% 구간 (High Risk): 할인율이 10%를 초과하면 무게와 상관없이 100% 지연 발생.
- ⚖️ 무게의 함정 (Weight Trap): 할인율이 낮더라도 (10% 미만), 앞서 Slide 12에서 확인했듯 1~2kg(경량) 및 4~6kg(중량) 화물에서 지연이 빈번하게 발생합니다.
- 💡 시사점: "할인율 10% 이상"과 "중간 무게 (2~4kg)"가 배송 지연의 핵심 예측 변수임이 입증됨.

탐색적 데이터 분석 및 주요 발견



데이터 탐색을 통해 배송 지연에 영향을 미치는 핵심 요인들을 식별했습니다.

"탐색적 데이터 분석(EDA) 결과, 배송 지연은 무작위로 발생하는 것이 아니라 명확한 패턴을 따르고 있음을 확인했습니다.

1. 할인율 10% 초과 여부가 지연의 가장 강력한 결정 요인입니다.
2. 무게는 특정 구간 (가벼움2,000 이하 / 무거움4,000 이상)에서 지연 리스크를 높입니다.
3. 반면, 고객 등급이나 문의 횟수는 지연과 무관했습니다.

파생변수 생성

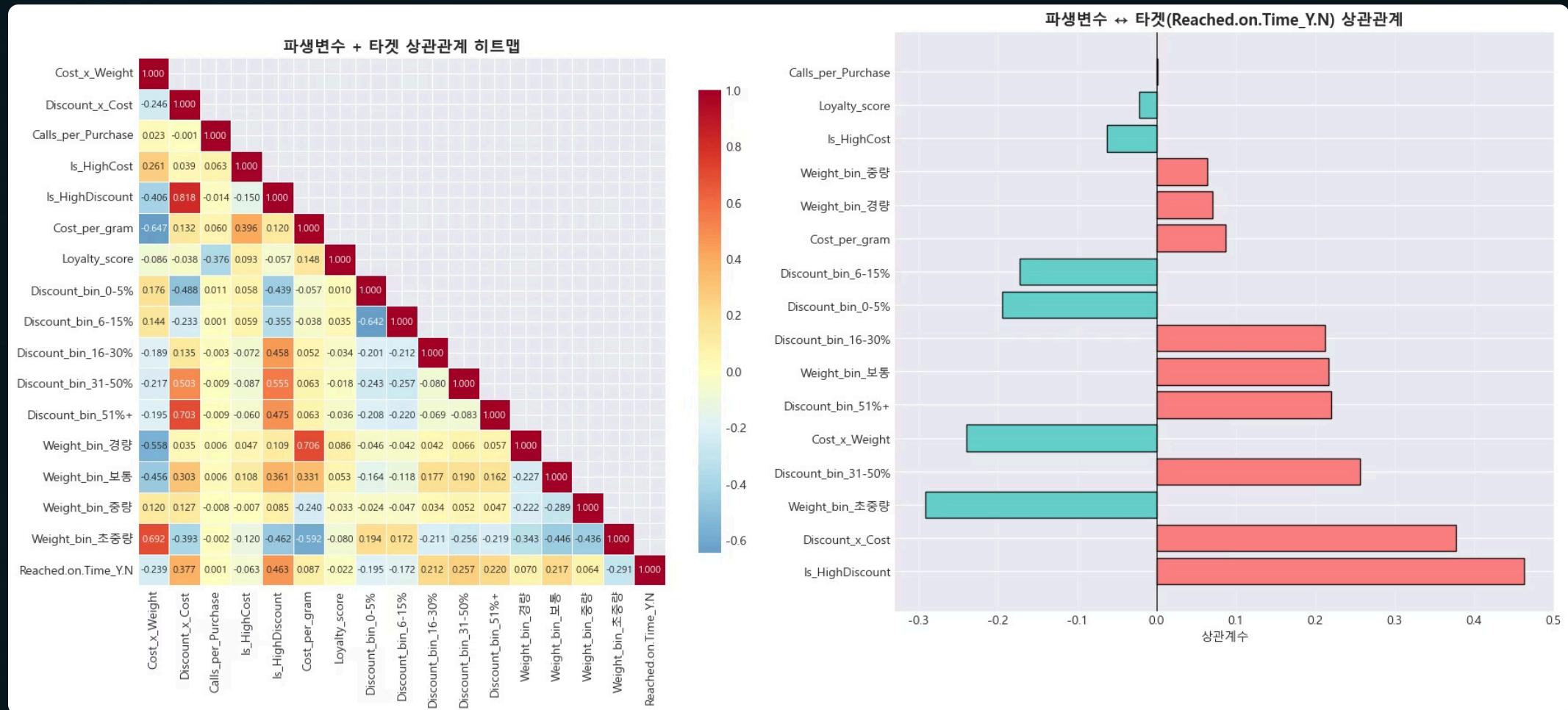
- **Discount_bin** (구간화): 할인율을 5개 구간으로 나눔
- **Weight_bin** (구간화): 무게를 구간별로 나눔
- **Cost_x_Weight** (교호작용): 가격 × 무게
- **Discount_x_Cost** (교호작용): 할인율 × 가격 (실제 할인 금액)
- **Cost_per_gram** (비율): 가격 ÷ 무게 (g당 가격)
- **Calls_per_Purchase** (비율): 문의 횟수 ÷ 구매 횟수
- **Is_HighCost** (플래그): 가격 상위 구간 여부 (0/1)
- **Is_HighDiscount** (플래그): 할인율 상위 구간 여부 (0/1)
- **Loyalty_score** (지수): 평점과 재구매 횟수 조합 점수

파생변수(Feature Engineering) 생성 전략



파생변수 유효성 검증 (Correlation Analysis)

어떤 변수가 배송 지연을 가장 잘 설명하는가?



RISK FACTOR

Is_HighDiscount (Corr: 0.46)

단순 할인율보다 우리가 만든 '고할인 여부 (Flag)' 변수가 배송 지연을 예측하는 제1의 지표임이 증명됨.

Discount_x_Cost(할인 금액) 또한 0.38로 매우 높은 설명력을 가짐.

SAFETY FACTOR

Weight_bin_초중량 (Corr: -0.29)

'초중량' 구간 변수가 강한 음의 상관관계를 보임. 즉, 아주 무거운 화물은 별도 관리 프로세스로 인해 오히려 정시 도착 확률이 높음을 발견.

LOW IMPACT

Customer Variables

Loyalty_score, Calls_per_Purchase는 상관계수가 0에 수렴. 고객의 충성도나 문의 빈도는 배송 프로세스에 영향을 주지 않음 (Feature Selection 시 제외 고려).

결론: Feature Engineering을 통해 생성한 변수들이 기존 변수보다 타겟(지연 여부)을 더 명확하게 설명하고 있음

타겟 변수와 유의미한 파생변수 설명

파생변수	상관계수	해석
Is_HighDiscount	+0.463	할인율 10% 초과 시 배송 지연 확률이 뚜렷하게 증가 (가장 강력)
Discount_x_Cost	+0.377	실질 할인 금액이 클수록 지연 경향
Weight_bin_초중량	-0.291	초중량(4,500g+) 상품은 오히려 정시 도착 경향
Discount_bin_31-50%	+0.257	높은 할인 구간일수록 지연
Cost_x_Weight	-0.239	가격×무게가 큰(크고 비싼) 상품은 정시 도착 경향
Discount_bin_51%+	+0.220	51% 이상 할인은 지연과 관련
Weight_bin_보통	+0.217	보통 무게 구간에서 지연 경향
Discount_bin_16-30%	+0.212	중간 할인 구간도 지연과 양의 관계
Discount_bin_0-5%	-0.195	할인이 거의 없으면 정시 도착
Discount_bin_6-15%	-0.172	낮은 할인도 정시 도착 경향

탐색적 데이터 분석 (EDA) 주요 발견 요약



할인율의 핵심 영향

할인율이 **10%**를 초과하는 경우, 배송 지역 확률이 **100%**에 달하여 가장 강력한 예측 변수임을 확인했습니다.



무게별 지역 패턴

1~2kg (경량) 및 **4~6kg** (중량) 구간에서 배송 지역이 집중적으로 발생하며, 특정 무게군에 대한 물류 처리 개선이 필요합니다.



고객 정보의 낮은 영향

고객 ID, 성별, 고객 등급, 문의 횟수 등은 배송 지역과 유의미한 상관관계를 보이지 않아 모델링 시 제외를 고려합니다.

이러한 분석 결과는 모델링에 사용할 변수 선정 및 파생변수 생성 전략 수립에 중요한 기반이 됩니다.

전처리 전략 (Preprocessing Strategy)

1단계: 결측치 처리 (Missing Values)

현황: 결측치 없음 (No Missing Values)

조치: 데이터 품질 우수

효과: 추가 처리 불필요

3단계: 스케일링 (Scaling)

- Standard Scaling: 평균 0, 표준편차 1로 정규화
- 대상: 수치형 변수 (Weight, Cost, Discount, Calls 등)

목적: 학습 속도 최적화 및 모델 수렴 가속화

2단계: 범주형 인코딩 (Categorical Encoding)

- One-Hot Encoding: 명목형 변수 (Warehouse, Shipment_mode, Gender)
- Label Encoding: 순서형 변수 (Product_importance)

목적: 머신러닝 모델이 이해할 수 있는 수치형으로 변환

전처리 과정을 통해 학습 속도를 최적화하고 모델 성능을 극대화합니다.

전처리 완료

구분	컬럼명	변환 후 타입	비고
타겟	Reached.on.Time_Y.N	bool	True=지연, False=정시
수치형(원본)	Customer_care_calls	uint8	0~255 범위 다운캐스팅
	Customer_rating	uint8	1~5 범위
	Cost_of_the_Product	uint8 / uint16	상품 가격
	Prior_purchases	uint8	이전 구매 횟수
	Discount_offered	uint8	할인율
	Weight_in_gms	uint16	상품 무게
원핫인코딩	Warehouse_block_A ~ _F	int8	창고 블록 (5개 더미)
	Mode_of_Shipment_Flight / _Road / _Ship	int8	배송 수단 (3개 더미)
	Product_importance_High / _Low / _Medium	int8	상품 중요도 (3개 더미)
	Gender_F / _M	int8	성별 (2개 더미)

참고: 원핫인코딩 시 `drop_first=False`로 설정하여 모든 범주를 유지하였습니다. 다운캐스팅을 통해 메모리 사용량이 약 70~80% 절감되었습니다.

모델링 전략 및 실험

베이스라인 모델 비교

8개 알고리즘을 동일한 조건에서 실험하여 성능 벤치마크를 수립했습니다.

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine
- K-Nearest Neighbors
- Naive Bayes

고도화 전략

Tree-based Ensemble 모델에 집중하여 XGBoost, LightGBM, CatBoost를 중점 실험했습니다.

하이퍼파라미터 최적화

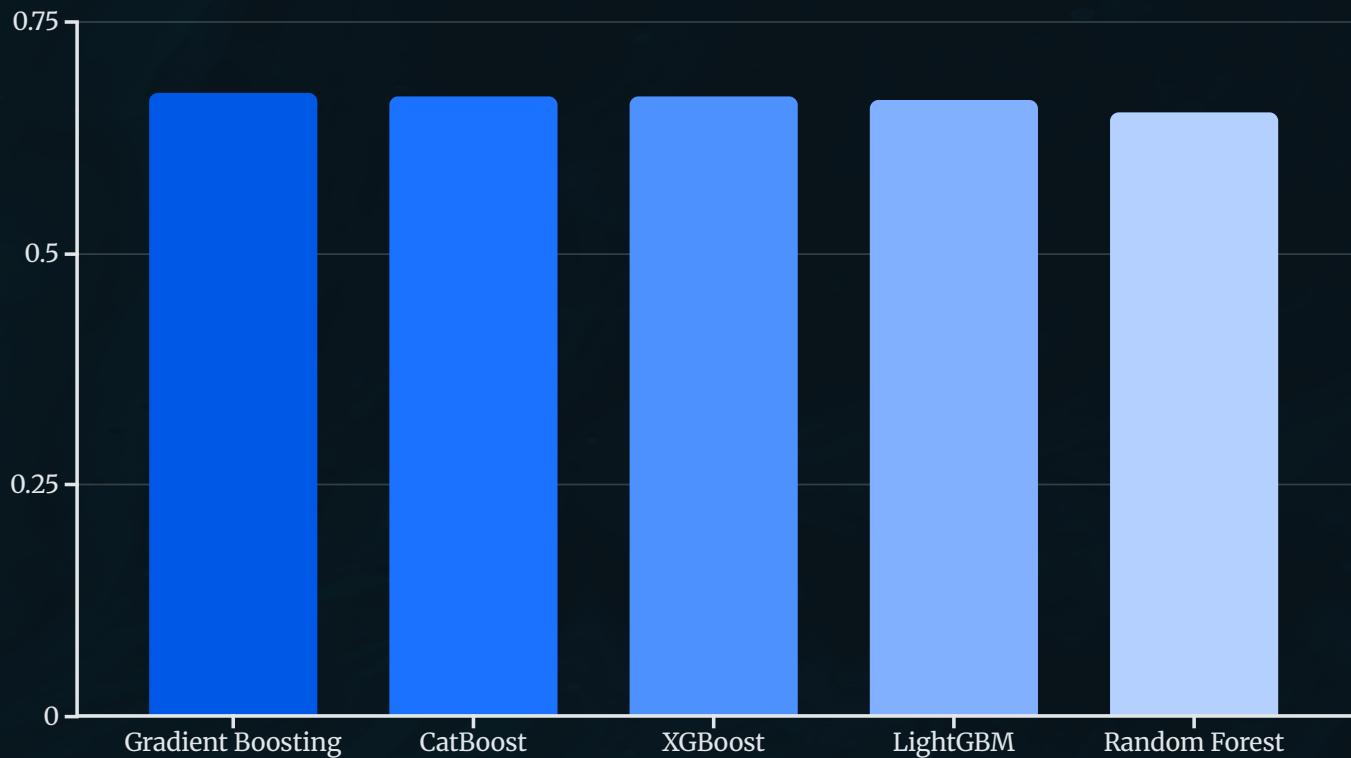
Optuna 베이지안 최적화

- 학습률(Learning Rate): 0.01-0.3 탐색
- 트리 깊이(Max Depth): 3-10 범위
- 최소 샘플(Min Samples): 동적 조정
- 정규화 파라미터 자동 튜닝

앙상블 전략

Soft Voting Classifier로 여러 모델의 예측 확률을 결합하여 안정성을 높였습니다.

최종 모델 성능 평가



평가 지표

F1-Score: 불균형 데이터에서 Precision과 Recall의 조화 평균

ROC-AUC: 이진 분류 모델의 전반적인 성능 지표

실험 결과

Gradient Boosting 모델이 F1 0.6738로 최고 성능을 기록했으며, ROC-AUC 0.7475로 균형 잡힌 분류 능력을 보였습니다.

최종 선정

안정적인 예측 성능과 변수 중요도(Feature Importance)를 통한 설명력을 종합적으로 고려하여 Gradient Boosting을 최종 모델로 선정했습니다.

비즈니스 인사이트 및 실행 전략



할인 프로모션 최적화

높은 할인율(15% 이상) 적용 시 주문 폭주가 예상되므로, 사전에 물류 용량 (Capacity)을 확보하고 배송 인력을 증원해야 합니다.



중량 화물 관리

3kg 이상 화물에 대해 별도의 Fast Track을 운영하거나, 전담 배송 파트너사와 협력하여 처리 속도를 개선합니다.

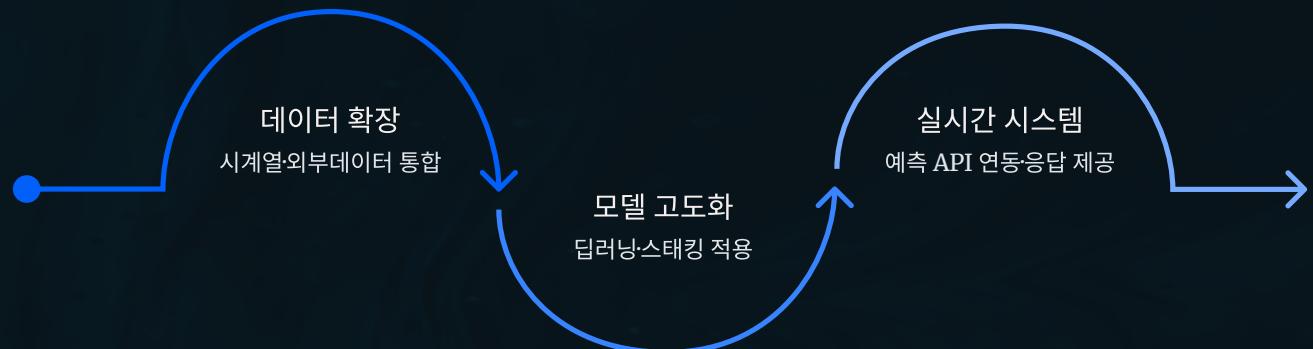


예측 기반 사전 조치

모델 예측 결과를 활용하여 지연 가능성이 높은 주문에 대해 고객에게 사전 안내 및 보상 프로그램을 적용합니다.

- **한계점:** 약 1만 건의 샘플 수 한계 및 날씨, 교통 등 외부 변수 부재로 인해 실제 운영 환경에서는 추가적인 데이터 통합이 필요합니다.

향후 개선 방향 및 Q&A



데이터 확장

시계열 데이터(주문 일자, 요일별 패턴)와 외부 데이터(날씨, 교통 상황, 공휴일)를 통합하여 예측 정확도를 향상시킵니다.

모델 고도화

딥러닝(LSTM, Transformer) 및 Stacking 양상을 기법을 도입하여 복잡한 비선형 패턴을 더욱 정교하게 포착합니다.

실시간 시스템 구축

예측 모델을 주문 시스템에 API로 연동하여 고객에게 실시간 예상 배송일을 안내하고 투명성을 제고합니다.

감사합니다

질의응답 시간을 갖도록 하겠습니다.