

1. 소개

안녕하세요. 코오롱모빌리티팀 발표를 맡은 이석재입니다.

지금부터 캡스톤 프로젝트 1차 중간발표를 시작하겠습니다.

저희 팀은 두 개의 프로젝트를 병행하고 있으며,
이번 발표는 프로젝트 1인,
중고차 판매가 예측 모델 구축에 대하여 진행하겠습니다.

2. 문제 정의

현재 코오롱 모빌리티의 중고차 판매 프로세스는
다음과 같은 두 가지 문제점을 가지고 있습니다.

첫째, / 딜러들은 내부 데이터 부재로 인하여,
엔카 웹사이트를 참고하며,
이는 시간적 비효율성을 만듭니다.

둘째로는, / 타사 광고가를 기준으로 한 가격 산정은
판매가와 가격 괴리를 야기합니다.

결론적으로, / 이 문제들로 인해
입찰 경쟁력이 떨어지고
고객 경험이 저하된다는 점을 발견 하였습니다.

이런 점들을 고려해서 저희 목표는
판매가 예측 모델을 구축해,
'매입가격 산정 기준'을 제공하여
기존 문제들을 해결하는 것입니다.

3. 데이터 개요

다음은 데이터 개요입니다.

코오롱 모빌리티로부터 받은 데이터는 / 엔카 데이터로써,
약 20만 행, 39개 컬럼, / 올해 1월부터 10월까지의 데이터를 제공받았습니다.

각 행의 의미는 개별 차량이며,
기본 차량 정보와 / 판매가 정보가 포함되어 있습니다.

저희 데이터의 특징은
제조사, 모델, 등급, 세부등급으로 나누어지는 / **계층적 데이터 구조**를 가지고 있습니다.

이로인해 **범주형 변수**가 매우 많고,
카디널리티가 높은 구조를 가지고 있습니다.

예들어 등급은 약 2,600개의 고유값을 갖고 있습니다.

또한 히트맵 분석에서는
동일한 **모델 계열**의 컬럼 간 / **강한 상관관계**를 가지고 있어,
다중공선성이 문제가 발생 할 수 있습니다.

이를 해결하기 위해 사전 연구 분석과 새로운 접근이 필요하였습니다.

4. 사전 연구 분석

선행 연구들을 분석한 결과,
중고차 가격 예측 연구는 크게 두 가지 방향으로 진행되고 있었습니다.

하나는 **차량 기본 스펙 중심 / 예측 방식**이며,
이는 어떠한 변수가 **예측의 정확도와 영향력**을 가지는데 집중한 논문들입니다.

여러 논문에서는 차량 제원 외에도 외생 변수들을 함께 사용합니다.
대표적인 예가 / 금리와 같은 / **거시경제 지표**인입니다.
실제로 이러한 변수들이 **중고차 가격 예측 성능**을 향상시킨다는 연구 결과도 확인했습니다.

다른 하나는, 세그멘테이션 기반 접근으로
차량을 특정 기준으로 나눠
판매가 예측 모델을 학습하는 방식입니다.

아래 장표에서 두 논문은 / 세그멘테이션에 대한 논문입니다.
첫번째 논문은
수치형변수를 K-means 통해 클러스터링하고
그 결과를 바탕으로 여러 모델들을 만들었다는 내용입니다.

두번째 논문은

범주형 변수를 K-modes를 통해 클러스터링하고 이를 하나의 모델에 하나의 파생변수로 추가하여 활용한 논문입니다.

기존 논문과 달리 저희의 차별점은

저희 데이터셋에 특성에 맞는 K-modes을 활용하되,
파생 변수로 사용한 논문과 달리
그 결과값을 활용해서 논문 1 처럼 각각의 고유 모델로 구축하며,
추가적으로 / 파생변수와 외생변수를 활용해서
만드는 것으로 최종 결정하였습니다.

5. Segmentation: K-modes 선택 이유

대부분의 기존 논문들은 K-means를 활용하여
수치형 중심으로 세그멘테이션을 진행하였습니다.

하지만 저희는 앞서 말씀드린것처럼
차량 데이터의 특성상, 범주형 컬럼이 매우 많고
카디널리티 또한 높기 때문에
수치형 중심의 K-means는 적합하지 않다는 판단을 하였습니다.

그래서 범주형 중심인 K-모즈를 채택하게 되었습니다.

K-모즈는 수치형 변수를 다루는 K-means와 달리
범주형 데이터의 빈도수와
불일치 수를 기반으로 클러스터링 하는 방식입니다.

이는 의미가 겹치는 범주형 정보 처리에 유리하며
본 프로젝트의 데이터 특성에 가장 적합하다는 판단 되었습니다.

6. 추가 Feature 생성 (Segmentation 이전 처리)

세그멘테이션을 하기 앞서
아래 3가지 변수를 비닝으로 전처리 하였습니다.

배기량은 배기량이 100~500 단위 간격으로
명확히 구분되는 특성을 보였고,
이는 국내 자동차세 구간 구조와 일치합니다.
따라서 500cc 단위로 구간화하여
범주형 변수로 변환하였습니다.

판매속도지수는 장표의 등식처럼
광고 시작일과 판매일 간의 차이로 만들어집니다.
판매속도지수는 모델의 특성상
직접 쓸수 없기에
대표차종별 평균값을 사용하였습니다.
이 변수는 각 차량의 인기와
판매 회전 속도를 반영할 수 있어
세그멘테이션에서 유용한 요소로 판단하였습니다.

(3) 신차가격 Binning

신차가격은 옵션값까지 포함된 고유 정보이기 때문에
가격대 구간으로 범주화하여 Segmentation 변수로 활용하였습니다.

7. 최적 클러스터 개수 선정

다음 최적의 클러스터 개수를 찾기 위해
저희가 검증한 방식으로
엘보우와 실루엣 스코어를 측정하였습니다.
Elbow Method의 Cost 그래프와
Silhouette Score 측정한 결과
보시는 바와 같이
클러스터 수의 기준이
6이 최적이라고 판단하였습니다.
(시간 남으면 하락, 상승 기준 설명)

8. 클러스터 해석 결과

6개 클러스터를 분석한 결과 다음과 같은 특징이 도출되었습니다.

내부의 차트는 각 클러스터를 대표하는 특성을 나타내는 지표입니다.
저희데이터 셋은 상당수가 국산 대중차로 이루어졌습니다.
따라서 보시는바와 같이 국산 대중차 중심으로 3개의 차종 클러스터로 이루어졌고, 그안에서 배기량 기준으로 자연스럽게 분리되어 있습니다.

또한
신차가격 7천만 원 이상의 고가 차량,
대형차 중심,
경차 중심 클러스터으로
클러스터 구조가 실제 차량 특성과 매우 잘 일치하는 것을 확인하였습니다.

9. 예측 모델을 위한 Feature 선택

이어 판매가 예측을 위한 지도학습 모델에서는 아래 변수들을 선정하였습니다.

핵심 변수로서는 사전 변수 중요도에서 높은 값을 가진 차량 연식 그리고 주행거리입니다.

외생 변수로는
거시경제 영향을 반영하는 금리와
세대교체로 인한 감가를 적용하기위해
차량 세대 정보를 추가하였습니다.

파생 변수로는
손상부위 기반으로 사고의 심각도를 랭크화하여 추가하였고

색상의 경우
흰색.검정.취색 등 인기 색상은 유지하였고
다른 원색 계열은 통합하여 카디널리티를 줄이기 위해 노력했습니다.
이를 통해 실제 현업에서 인기 색상이 아닐 경우 감가하는 것을 반영할것으로 기대합니다.

10. 최종 모델 구조

결론적으로 최종 모델은 아래와 같은 구조로 설계하였습니다.

- 1) 높은 카다널티 범주형 변수는 K-modes로 Segmentation으로 처리하고,
- 2) Segmentation된 결과를 활용하여
- 3) 수치형 변수 중심(연식·주행거리·신차가격 등)으로 지도학습을 수행하여
- 4) 목표 변수, 즉 최종 판매가 예측하는 구조로 설계하였습니다.

11. 평가 계획 및 향후 방향

사용하게 될 평가지표는

모델의 설명력을 볼 수 있는 R-squared

가격 오차를 알수 있는 RMSE

%단위로 오차를 제공하는 MAPE를 사용할 계획입니다.

모델링은

Linear Regression

Gradient Boosting

Ensemble

까지 확장해 비교할 예정입니다.

향후 개선 방향은 다음과 같습니다.

K-modes 이외의 다른 범주형 군집 방식도 실험하고

범주형 간 거리 정의 방식을 다양하게 적용하고

클러스터 품질 검증 강화를 시도해볼 예정입니다.

이상으로 발표를 마치겠습니다.

이어 파트 2 발표를 이어 진행하겠습니다.