1. 소개
안녕하세요. 코오롱모빌리티팀 발표를 맡은 이석재입니다.

지금부터 캡스톤 프로젝트 1차 중간발표를 시작하겠습니다.

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

---

2. 문제 정의

현재 **코오롱 모빌리티의 중고차 판매 프로세스**는
다음과 같은 두 가지 문제점을 가지고 있습니다.

첫째, / 딜러들은 내부 데이터 부재로 인하여,
엔카 웹페이지을 참고하여, 판매가를 예측합니다.
이는 시간적 비효율성을 만듭니다.

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

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

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

3. 데이터 개요
다음은 데이터 개요입니다.

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

또한 기본 차량 정보와 / 판매가 정보가 포함되어 있습니다.

여기서 중요한 파트는 상관계수 히트맵입니다.
데이터셋에서 강한 상관관계를 가지고 있는 것을 확인 할 수 있습니다.

이는 저희의 데이터가 제조사, 모델, 등급, 세부등급으로 나누어지는
계층적 데이터 구조를 가지고 있기 때문입니다.

이로 인해 **다중공선성**이 문제가 발생 할 수 있습니다.

이를 해결하기 위해 사전 연구 분석과 새로운 접근이 필요하였습니다.

ㅡ----------

4. 사전 연구 분석

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

하나는 금리 등 거시경제 지표 외생변수를 활용하여,
중고차 가격 예측의 정확도를 높이는 방향이고,
다른 하나는 차량을 특정 기준으로 나누는
세그멘테이션을 통해 정확도를 높이는 방향입니다.

세그멘테이션을 활용한 주요 두 논문 중 하나는 수치형 변수를
K-means를 통해 클러스터링하여 여러 모델을 만들었고,
다른 하나는 범주형 변수를 K-modes를 통해 클러스터링 하고
하나의 파생변수로 추가하였습니다.

기존 논문과 달리 저희의 차별점은
데이터셋 특성에 맞는 K-modes를 활용하되,
하나의 파생변수로 추가한 논문과 달리
그 결과를 활용하여 각각의 고유 모델로 구축하는 것입니다.

5. Segmentation: K-modes 선택 이유

이번 장표에서는 세그멘테이션 전략에 대해서
좀더 설명해드리겠습니다.
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.

K-모즈는
범주형 데이터의 빈도수와
불일치 수를 기반으로 클러스터링 하는 방식으로
본 프로젝트의 데이터 특성에 가장 적합하다는 판단 되었습니다.

ㅡ----------

6. 추가 Feature 생성

세그멘테이션을 하기 앞서
아래 3가지 변수를 비닝으로 전처리 하였습니다.

배기량은 500cc 단위로 구간화하여
범주형 변수로 변환하였습니다
이는 국내 자동차세 구간 구조와 일치합니다.

판매속도지수는 장표의 등식처럼
광고 시작일과 판매일 간의 차이로 만들어집니다.
만들어 지수는 평균화 넣었습니다.
이 변수는 각 차량의 인기와
판매 회전 속도를 **반영**할 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

다음은 클러스터 개수 선정 기준입니다.
세그멘테이션에서 중요한 것은 최적의 클러스트 개수를 정하는 것 입니다.

합리적인 값을 값을 찾기 위해
저희가 검증한 방식으로
엘보우와 실루엣 스코어를 측정하였습니다.
보시는 보는것 처럼
그래프를 통해서
숫자 육이 최적의 클러스트 수라고 판단하였습니다.

(기존 Elbow - k-means(sume of sequared error) 플래토가 시작되는 지점)
(기존 실수엣 스코어 - 유킬리안 유클리디언 거리 기반이지만 / hmaming distance을 상용)
(0.1~ 0.25 약한 군진 (실무 점수))

ㅡ----------

8. 클러스터 해석 결과

6개 클러스터을 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 클러스터으로 이루어졌고,
그안에서 배기량 기준으로 자연스럽게 분리되어 있습니다.

또한
고가 차량 / 대형차 / 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
아래와 같은 변수들을 선정하였습니다.

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

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

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

파생 변수로는
사고의 심각도를 손상 부위에 따라 랭크화 시켰습니다.

ㅡ----------

10. 최종 모델 구조

결론적으로 최종 모델은 아래와 같은 구조로 설계하였습니다.
1) 높은 카다닐티 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 방식**'을
다양하게 적용하고
클러스터 품질 검증 강화를 시도해볼 예정입니다.
이상으로 발표를 마치겠습니다.

이어 파트 2 발표를 이어 진행하겠습니다.