**[1장]**

안녕하세요. 9조 발표 시작하겠습니다.

이번 프로젝트는 신용카드 회사의 고객들 세그먼트를 분류하는 게 목표였죠?

**(2초 정도 쉬고)**

저희도 이 목표에 맞게 데이터 분석 전 과정을 진행해봤습니다.

**[2장]**

발표의 순서는 다음과 같습니다.

**프로젝트 소개**  
**데이터 탐색 및 시각화**

**피처 선택 전략**   
**모델링 및 예측**

마지막으로 **결론**입니다.

**[3장]**

먼저 프로젝트 소개입니다.

**[4장]**

세그먼트를 정밀하게 분류해야 할 이유는 무엇일까요? **(조금 쉬고)**  
저희는 그 이유를 두 가지로 정리해봤습니다.

첫째, **고객 특성에 기반한 맞춤 전략을 수립하기 위해서**입니다.  
고객의 소비 패턴, 금융 활동, 채널 이용 행태를 종합적으로 분석함으로써 각 세그먼트에 최적화된 서비스와 마케팅 전략을 설계할 수 있습니다.

둘째, **방대한 고객 데이터를 세그먼트별로 사전 분류하여 잠재 리스크를 선제적으로 관리하고 운영 효율성을 극대화하기 위해서**입니다.  
이를 통해 세그먼트별 자동화된 대응 전략을 수립·운영함으로써 마케팅 비용을 절감하고 업무 프로세스를 간소화할 수 있습니다.

**[5장]**

저희는 앞서 말씀드린 두 가지 이유 외에도 세그먼트를 정밀 분류함으로써 **맞춤형 마케팅 전략 수립, 고객 이탈 방지 및 유지율 개선, 데이터 기반 의사결정 강화** 등의 다양한 기대 효과를 얻을 수 있다고 판단했습니다. 이와 같은 기대 효과를 바탕으로 본격적인 분석을 시작하겠습니다.

**[6장]**

데이터 탐색 및 시각화입니다.  
설명에 앞서, 이번 파트가 끝난 후 피처 선택 전략 섹션에 제시될 누적 상대 중요도 그래프에서 선별된 피처를 활용해 분석을 진행했다는 점을 먼저 말씀드립니다.  
그럼 분석 내용을 설명드리겠습니다.

**[7장]**

가입 경과 개월 수에 따른 세그먼트 분포 분석입니다.

세그먼트는 고객의 가입 경과 개월 수에 따라 뚜렷한 차이를 보였습니다.

**A, B**는 경과 개월 수가 길어질수록 비중이 증가하는 경향을 보였습니다.

**C, D** 같은 경우 약 75개월 전후 구간에서 가장 높은 분포를 나타냈습니다.

**E**는 가입 초기, 즉 경과 개월 수가 짧을수록 집중적으로 분포하는 모습입니다.

이를 통해 최근에 가입한 고객일수록 **E**로 분류될 가능성이 높고, 장기간 이용한 고객일수록 **A, B**에 분포할 가능성이 높다는 걸 알 수 있었습니다.

**[8장]**

카드 이용한도 금액 분포 분석 결과 입니다.  
**A**는 고한도 고객이 집중 분포하여 상위 등급 세그먼트라는 정체성이 뚜렷하게 드러났습니다.  
반면, **B, C, D**는 중·저한도 구간에 분포하고 있으며,  
**E**는 대부분 낮은 한도에 분포하고 있습니다.

**[9장]**

다음은 고객 세그먼트별 청구 금액 및 포인트 적립 특성 분석입니다.  
청구 금액 구간별 분포를 보면 **A와 B**는 7만원에서 15만원 구간과 15만원 이상 고액 구간에서 높은 비중을 차지하여 카드 사용이 활발한 고객층임을 알 수 있습니다.

**C**도 3만원에서 7만원 구간을 중심으로 비교적 고르게 분포합니다.  
포인트 적립량을 나타낸 바이올린 플롯에서는 **B**가 가장 넓은 분포를 보이며

**A와 C**도 중상위권에서 활발히 적립하는 것으로 확인됩니다.  
반면 **D와 E**는 청구 금액과 포인트 적립 모두 낮아 소비 및 혜택 활용이 비활성화된 특성을 보입니다.

**[10장]**

이러한 소비 특성은 대출 및 카드론 잔액 추이 비교에서도 그대로 반영됩니다.  
**B**는 카드론, 현금서비스, 유이자 할부 등 모든 대출 항목에서 높은 잔액을 기록하며, 카드론 잔액이 최근까지 상승 추이를 보여 적극적인 금융 활동 성향이 두드러집니다.  
반면 **C, D, E**는 전반적으로 대출 잔액 수준이 낮고,

특히 **E**는 거의 활용하지 않았습니다.  
따라서 **C, D, E** 세그먼트는 리스크는 낮지만, 수익 기여도 역시 높지 않을 것입니다.

**[11장]**

이번에는 세그먼트별 이메일 마케팅 반응을 살펴보겠습니다.  
왼쪽 그래프는 당월 대비 최근 6개월간 이메일 마케팅 노출 건수 추이입니다.  
**B 세그먼트**는 당월 노출 건수가 3.2건에 불과하지만, 기간을 최근 6개월로 확장하면 27.5건으로 대폭 증가합니다.  
오른쪽 그래프는 **B 세그먼트** 내 이메일 수신 동의 비율을 보여주며, 70.1%가 수신에 동의하여 다른 마케팅 수단 대비 이메일 반응도가 높습니다.  
이를 종합하면, **세그먼트 B**는 이메일 마케팅 운영에 최적화된 타겟이라고 할 수 있습니다.

**[12장]**

이어서 이메일 마케팅 반응이 가장 높은 **B 세그먼트**의 주요 타깃층을 살펴보겠습니다.

왼쪽 파이차트는 이메일 노출 고객의 연령대 비율을 나타내며 **30대, 50대, 40대** 순으로 높은 비중을 차지합니다.  
오른쪽 그래프는 최근 6개월간 이메일 노출 고객을 연령대별 생애주기로 분류한 결과로 **30대, 40대, 50대의 자녀 성장기 고객**이 가장 큰 비중을 보입니다.  
이를 통해 **B 세그먼트**에서는 자녀 성장기인 30세에서 50세 고객을 중심으로 가족 맞춤형 혜택을 제안하는 전략이 이메일 마케팅 효과를 극대화할 수 있는 마케팅 방법이 될 것입니다 .

**[13장]**

이번에는 업종별 실사용 고객 비율 및 소비 성향 분석입니다.  
업종별 실사용 고객 비율과 소비 금액 모두에서 세그먼트 A가 압도적으로 높은 참여율과 소비 규모를 보여 핵심 소비층임을 확인했습니다.  
반면 세그먼트 E는 모든 업종에서 실사용 비율과 소비 수준이 전반적으로 저조하여 비활성 고객 비중이 높다고 판단됩니다.

**[14장]**

세그먼트별 신용카드 승인 금액 분포를 살펴보면

**A**는 가장 높은 중앙값과 넓은 분포 범위를 보여 활발한 소비를 하는 핵심 그룹임을 알 수 있습니다.

반면 **E**는 중앙값이 가장 낮고 분포도 좁아 소비 여력이 상대적으로 낮은 그룹으로 해석됩니다.

이어서 오른쪽 그래프에서는 고객의 총 승인 금액과 승인 건수를 기준으로 우량·중간·기초 등급으로 분류한 결과 그래프입니다.

**우량** 등급은 총 승인 금액이 20만 원 이상이거나 승인 건수가 300건 이상인 고객,

**중간** 등급은 총 승인 금액이 8만 원 이상 20만 원 미만인 고객,

**기초** 등급은 총 승인 금액이 8만 원 미만인 고객으로 정의했습니다.

해당 분류 결과 우량 등급 고객이 가장 높은 승인 금액을 기록했으며, 등급이 낮아질수록 승인 금액이 점진적으로 감소하는 경향을 확인할 수 있습니다.

**[15장]**

마지막으로 세그먼트별 유휴 징후, 위험등급, 경과월 매출 특성 분석입니다.  
고객 이탈 징후를 탐지하기 위해 **‘취소’, ‘미사용’, ‘실사용**’으로 범주형 변수를 생성했습니다. 이때 ‘**미사용’ 또는 ‘취소**’ 비중이 높을수록 해당 고객이 비활성 상태임을 나타내는 유휴 징후로 정의했습니다.

유휴징후 개수에 따라 고객을 고위험/중위험/저위험으로 분류 한 뒤 세그먼트 별 분포를 살펴봤습니다.

왼쪽 그래프에서 **E**는 유휴 징후 수의 중앙값과 상위값 범위가 모두 높아 휴면 기간이 긴 고객이 다수 포함된 그룹임을 알 수 있습니다.

반면 **A와 B**는 유휴 징후 수의 중앙값과 분포 폭이 낮아 지속적으로 활동하는 고객 비중이 높습니다.  
가운데 그래프에서는 고위험 등급 비율이 **E**에서 가장 높고 정상 등급 비율은 가장 낮아 리스크 관리가 시급함을 확인할 수 있습니다.  
오른쪽 그래프는 경과월 매출 분포를 보여줍니다.

최근 **3개월 이내** 활동한 고객은 승인 매출 중앙값과 분포 폭이 높았으며, **4~6개월 경과** 고객의 승인 매출은 크게 감소하여 사용 경과 기간이 길어질수록 매출이 감소하는 경향이 나타났습니다.

**[16장]**

이제부터 피처 선택 전략에 대해 말씀드리겠습니다.

**[17장]**

먼저 데이터 전처리 과정입니다.

데이터 시트별로 결측치 비율이 약 80~90% 정도인 컬럼과 한 가지 데이터값만 존재하는 컬럼은 제거했습니다.

그 후 날짜형 변수는 datetime 타입으로 변환하고, 범주형 변수는 인코딩을 통해 수치형으로 변경했습니다.

각 데이터 시트별 전처리를 완료한 뒤 모든 데이터를 병합하여 최종 데이터셋을 구성했습니다.

**[18장]**

최종 병합한 데이터셋을 기반으로 중요도에 따른 변수 선택을 진행했습니다.

XGBoost를 활용해 피처 별 누적 상대 중요도를 산출하고, 그래프에서 기울기가 완만해지는 엘보우 지점인 상위 90% 구간, 총 **205개**의 변수를 핵심 피처로 선별했습니다.

이어 VIF 계수를 확인하여 10 이상인 컬럼은 제거했고,

주요 정보를 최대한 보존 하기 위해 PCA 기반 차원축소를 적용하여 최종 **121개**의 피처로 모델링을 진행하게 되었습니다.

**[19장]**

앞서 설명드린 XGBoost 기반 누적 상대 중요도 그래프 입니다.

그래프에서 90% 지점이자 기울기가 완만해지는 구간에 해당되는 **205개**의 피쳐를 핵심 변수로 먼저 선별했습니다.

**[20장]**

피쳐 선별 및 VIF 제거 후 PCA를 진행한 그래프입니다.

데이터의 주요 정보를 설명하는 주성분을 중심으로 차원 축소를 수행하여 **121개**의 주성분으로 구성했습니다.

**[21장]**

이제 모델링 및 예측입니다.

**[22장]**

모델링 전략은 크게 두 가지로 진행했습니다.  
처음엔 전체 데이터를 대상으로 하나의 모델을 학습했으나 성능이 만족스럽지 않아,  
이후 세그먼트별로 단계적 이진 분류를 수행한 뒤 각 모델을 조합하는 방식으로 전환했습니다.  
두 전략 모두 CatBoost, XGBoost, LGBM을 활용해 성능을 비교했습니다.

**[23장]**

세그먼트 별로 나눠 개별 학습을 진행한 이유를 설명드리겠습니다.

전체 데이터를 한 번에 학습할 경우, **세그먼트 E**의 비중이 지나치게 높아 모델이 **E**에 편향될 위험이 큽니다.

반면 **A, B**는 샘플 수가 적어 학습 데이터가 부족하고, 이로 인해 오분류 비율이 높게 나타났습니다.

또한 **C와 D**는 경계가 명확하지 않아 서로를 혼동하는 경우가 많았습니다.

따라서 각 세그먼트 별로 개별 모델을 학습함으로써 샘플 불균형과 구간 혼동 문제를 완화하고, 예측 정확도를 높이고자 했습니다.

**[24장]**

**CatBoost** 결과입니다.

이때 클래스 불균형을 완화하기 위해 오버샘플링과 클래스 가중치를 적용했으나, 성능 향상은 미미했습니다.  
이에 CatBoost뿐 아니라 XGBoost, LGBM, 랜덤포레스트를 혼합한 앙상블을 시도했으나, 최종 정확도는 약 **0.86**으로 단일 CatBoost 모델과 유사한 수준에 그쳤습니다.

**[25장]**

다음은 **LGBM** 결과입니다.

앞서 설명했던 단계적 이진 분류 및 모델 조합 방식을 적용했으며, 최종 정확도는 **0.87**을 기록했습니다. CatBoost보다는 소폭 향상되었으나 XGBoost보다는 다소 낮은 성능을 보였습니다.

**[26장]**

마지막으로 **XGBoost** 결과입니다.

XGBoost 역시 세그먼트별 단계적 이진 분류 전략을 통해 각 모델을 조합하여 학습했으며, 정확도는 **0.88**을 기록했습니다.  
이를 통해 XGBoost 기반 전략이 모든 모델 중 가장 우수한 예측 성능을 보였음을 확인했습니다.

XGBoost가 가장 우수한 예측 성능을 기록하여 최종 모델로 선정했습니다.

**[27장]**

결론입니다.

**[28장]**

발표 초반에 언급드린 기대 효과를 바탕으로 저희 분석 결론을 정리하겠습니다.

먼저 저희의 목표였던 세그먼트 예측 모델은 정확도 **0.88**로 마무리하게 되었습니다.

기대 효과였던 맞춤형 마케팅 전략 수립 가능은 일부 세그먼트 한정으로 구분이 가능했습니다.

고객 이탈 방지 및 유지율 개선 같은 경우 고위험 고객인 **E**는 높은 정확도로 예측이 가능하다는 걸 알 수 있었습니다.

마지막 데이터 기반 의사 결정 강화는 전체 흐름은 파악이 가능하지만 **A와 B**에서 정확도가 너무 낮아 한계점을 보였습니다.

**[29장]**

이제 지금까지 저희가 프로젝트를 하면서 생겼던 문제점들에 대해서도 한번 정리해봤습니다.

초기 전처리와 분석 방향성 설정을 확실하게 하지 못해 전체적으로 진행이 지연되고 중복된 작업을 발생시켰습니다.

전처리 단계에서 과도한 시간 소모가 발생하면서 모델링 단계에 충분한 리소스를 할당하지 못해 모델 튜닝 및 추가 실험을 제대로 수행하지 못했습니다.

이러한 문제를 통해 초기 전처리 로드맵과 피처 설계 기준을 명확히 수립해야 한다는 점과, 유연한 실험 계획을 세워 모델링 단계에서도 충분한 리소스를 확보해야 한다는 교훈을 얻었습니다.

**[30장]**

이것으로 저희 발표 마무리하겠습니다.

들어 주셔서 감사하고 이제 질문 받도록 하겠습니다.