**(발표 대본)**

[도입부]

저희 조는 머신러닝 모델링을 통해 개인 금융고객의 신용등급을 정확하게 산정하였습니다. 또한 이에 더 나아가서 중신용자의 대출승인을 위한 잠재적 기준을 추가적으로 제안하려 합니다.

시작에 앞서 저희 프로젝트 진행 절차에 대해 소개하겠습니다. 데이터 탐색 => 데이터 전처리 => 모델링 => 3가지 class 등급 산정 => 클러스터링() => 중신용자를 위한 새로운 대출 잠재기준 제안 단계까지 진행하였습니다.

저희는 대출 심사에 영향을 미치는 여러 요인을 크게 3가지로 나누어 EDA를 진행하였습니다. 현재 고객이 부담하는 채무 부담, 고객이 갖고 있는 직업의 안정성과 연령, 연수입 등 고객 성향이 드러나는 고객 상황, 그리고 고객의 과거 신용등급을 반영한 이자율이라는 총 3가지 요인으로 분류했습니다.

먼저 target 값인 신용등급부터 살펴봤을 때, 각 등급별 비율이 2:5:3으로 standard 등급이 월등히 많은 불균형 데이터임을 파악했습니다. 다음으로, 이자율과 신용등급을 살펴보자면, **이자율이 클수록 저신용자 비율이 증가하고 있습니다.** 전체적으로 봤을 때 이자율별 신용등급 분포는 균일한 편입니다.

다음으로는 개인 신상 정보를 나타내는 개인 상황을 살펴보겠습니다. 연령, 연수입, 직업 등 각 지표별 신용등급 분포는 대부분 균일하게 나타났습니다. 추가적으로 변수별 특징을 살펴보자면, 연령대의 경우, 나이가 들수록 good과 poor 등급이 차지하는 비율이 증가하는 경향을 보인다는 것을 알 수 있으며, 연수입의 경우, 대다수가 낮은 연수입을 받고 있다는 것을 알 수 있는데, 이는 대출 절벽에 몰릴 가능성이 높은 그레이존 고객이 많음을 의미합니다.

고객의 채무 상환 의지를 보기 위해서 월별 상환금액과 리볼빙 여부를 살펴보았습니다. 먼저, 월별 상환금액과 신용등급 분포는 거의 균일한 분포를 보입니다. 또한, 리볼빙을 한 경우에 상대적으로 good 등급 비율이 적은 것을 알 수 있는데, 이를 통해 리볼빙을 사용하는 사람들은 주로 신용등급 관리가 어렵다는 사실을 파악하였습니다.

앞서 말씀드린 EDA 내용을 기반으로 각 요인별 중저신용자의 특징을 살펴봤을 때 높은 이자율, 낮은 경제지표, 그리고 리볼빙을 사용한다는 사실관계를 파악할 수 있었습니다. 전반적인 EDA 작업을 통해서, 과거 신용등급이 반영된 이자율뿐 아니라 채무 상환 의지도 신용등급 산정에 큰 영향을 미친다는 것을 알아냈습니다. 따라서 정확한 등급 산정 모델링뿐만이 아니라 다양한 관점에서 상관관계, 통계 분석이 필요합니다. 또한 본 프로젝트에서는 이렇게 데이터를 분석한 결과를 토대로, 추가적인 데이터 인사이트 도출까지 달성했습니다.

앞선 EDA 작업을 통해서 이상치가 의심되는 데이터 칼럼 세 개, age, annual income, outstanding debt를 확인하였고, 이상치를 자세히 살펴보기 위해 Boxplot EDA를 진행한 다음, 범위를 벗어나는 데이터에 대해 IQR 방식으로 이상치 처리를 진행했습니다.

결측치 처리는 수치형 변수와 범주형 변수로 나누어서 진행하였습니다. Payment Behaviour 같은 수치형 변수는 customer\_ID 로 그룹화한 후에 그룹별 중앙값으로 대체했습니다. Monthly inhand salary 와 같은 범주형 변수는 우선 customer\_ID 로 그룹화한 후에 그룹별 최빈값이 있다면 바로 대체하고, 최빈값이 없는 경우에는 중앙값으로 대체하였습니다. 이때 의미없는 데이터는 모두 NaN값으로 대체한 후에 미리 정해둔 결측치 처리 방법을 적용했습니다.

그 외 처리를 간단히 말씀드리자면, 라벨 인코딩 기법을 활용하여 occupation, payement behavior 등의 범주형 변수에 적용하였으며, 대표적 오버샘플링 기법인 SMOTE 기법을 적용하여 기존 데이터가 갖고 있던 불균형 문제를 해결하였으나, 모델 성능을 비교한 결과, 오버샘플링이 성능에 큰 차이를 보여주지 않아서 결론적으로는 적용시키지 않았습니다.

저희는 pycaret이라는 AutoML 오픈소스를 활용한 결과 가장 좋은 성능을 보였던 ExtraTrees를 선택해 모델링을 진행하였습니다. ExtraTree는 RandomForest보다 좀 더 무작위성으로 각 특성을 분할하는데, 이러한 특성으로 인해 속도면에서 장점을 가지며, 데이터 과적합 문제를 회피해주기에 단일 트리보다 좋은 성능을 보인다고 생각되어 선택하였습니다.

Extra Tree를 사용한 개인고객 신용등급 분류 모델의 성능은 약 0.82 로 나왔습니다. k-fold 교차검증을 실시해 모델을 검증하였으며, 오버샘플링과 튜닝 여부가 성능에 큰 차이를 보이지 않아 최종 결과에서 제외했습니다. 변수 중요도를 보면, credit mix와 outstanding debt 요소가 신용 등급 평가에 영향이 큰 변수임을 알 수 있습니다. 일반적으로 생각했던 나이나 직업 등 개인의 신상이나 경제 지표와 관련 있는 변수들은 의외로 신용등급에서 큰 중요도를 보이지 않았습니다. 신용등급 모델링을 진행한 후에, 데이터 인사이트를 얻기 위해 저희는 추가적인 작업을 진행했습니다.

15까지는 함

16, 17은 도입부에 있던 우리 주제(추가적 인사이트 도출 부분)

은행의 주요 수익원인 예대마진율이 연속 감소 추이를 보이고 있습니다. 이는 특히 대출 규모의 연속 감소세 때문인데요, 그래프를 살펴보면, 전반적인 시중은행과, 하나은행 역시 예대마진율 부분에서 연속 감소 추세를 보이는 걸 알 수 있습니다. 이러한 가운데 인터넷 뱅크는 전통 금융권에서 등한시한 중저신용자 즉, 그래이존 고객에게 공격적으로 대출 마케팅을 선보이며 파이를 가져가고 있습니다. 이에 전통 금융권에서도 보다 더 정확한 신용등급 조회로 대출디폴트 리스크를 줄여야 할 뿐만 아니라, 이러한 중저신용자인 그레이존 고객들 중에서도 신용 위험이 적은 고객을 골라내어 적극적으로 대출 마케팅을 펼쳐야 할 필요성이 대두되고 있습니다.

앞서 진행한 과정을 토대로, 중신용자를 k-means clustering 으로 다시 이진분류하는 모델을 만들었습니다. 중신용자 중에서도 상대적으로 신용이 좋은 고객과 그렇지 않은 고객을 분류해내기 위함입니다. 이러한 클러스터링 작업에서 k-means clustering을 적용한 이유는 수렴성이 보장될 뿐만이 아니라 사전에 특정 변수에 대한 역할 정의가 필요하지 않아서 알고리즘 적용이 비교적 편리했기 때문입니다.

지금까지 클러스터링으로 중신용자를 중고신용자, 중저신용자로 분류했다면, 이제는 어떠한 요인들때문에 이러한 이진분류가 가능했는지를 통계 분석해보았습니다. 먼저 T 검정을 수행했을 시에 total EMI per month, outstanding debt, delay from due date, credit utilization ratio 변수들은 모두 유의미하다는 결론을 얻었습니다. 즉 중신용자들을 재분류하는데 이 변수들이 크게 영향을 미친 것이라고 해석할 수 있습니다.

다음으로 카이제곱 검정을 수행했을 시에는 num of delayed payment와 num credit inquiries변수는 유의미하지 않았습니다. 즉 중신용자들을 재분류하는데 별로 영향을 미치지 못 한 것입니다.

클러스터링 분석 후 통계 분석까지 진행한 결과, 주로 EDA에서 채무 부담 즉, 채무 상환 의지로 정의한 변수들이 크게 영향을 미쳤습니다. 따라서 이러한 변수들로 구성된 새로운 대출 승인 기준을 제안합니다.