Progress Report

신용관련 정보 데이터셋을 이용한

신용 점수 Classification 과 연체 일수 Regression

19 조

2022320062 정하연, 2022320090 조서윤, 2020170365 박가영

1. 문제 정의

1.1 문제 정의

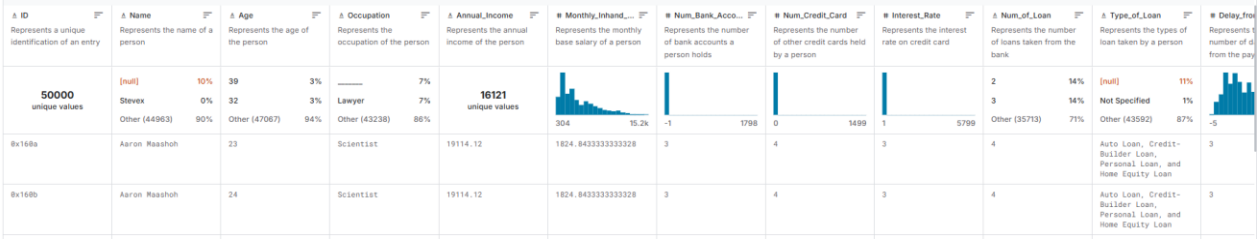
금융기관은 고객의 대출 상환 가능성을 정확히 예측함으로써 신용 리스크를 최소화하고, 동시에 고객에게 맞춤형 금융 서비스를 제공해야 하는 중요한 과제를 안고 있다. 그러나 각 고객의 재무 상황, 소득 구조, 지출 습관, 과거 상환 이력 등은 매우 다양하며, 단순히 신용등급이나 제한된 재무 지표만으로는 연체 위험을 충분히 평가하기 어렵다. 이러한 복잡한 환경에서는 전통적인 점수 기반 신용평가 방식만으로는 한계가 존재하며, 보다 정밀하고 데이터 기반의 분석이 요구된다.

이에 본 연구에서는 금융기관이 합리적으로 대출 승인 여부와 한도를 결정할 수 있도록, 머신러닝 기반의 분류(Classification) 모델을 개발하고자 한다. 분류 모델은 개인 고객을 연체 가능성 또는 신용 점수 수준에 따라 체계적으로 구분함으로써, 리스크 관리와 대출 심사 프로세스의 효율성을 동시에 개선할 수 있다. 아울러 연체 일수와 같은 연속적 위험 지표를 예측하는 회귀(Regression) 모델을 추가로 적용함으로써, 단순한 등급 분류를 넘어 보다 세밀한 신용평가 체계를 구축할 수 있다.
본 연구는 크게 두 단계(Task)로 구성된다.

- **Task1** 에서는 개인 고객의 금융 데이터를 기반으로 신용 점수를 세 단계로 분류하여 고객의 연체 위험을 정량적으로 평가한다. 이를 통해 금융기관은 대출 승인 여부와 한도를 보다 합리적으로 결정할 수 있다.
- **Task2** 에서는 고객의 연체 일수를 예측하여, 분류 모델에서 제공하는 범주형 평가를 보완하고, 개별 고객의 상환 패턴과 위험도를 보다 정밀하게 반영하는 신용평가 체계를 구축한다. 이 두 가지 Task 를 통합함으로써, 본 연구는 금융기관의 리스크 관리와 맞춤형 금융 서비스 제공을 동시에 지원할 수 있는 정밀하고 실용적인 예측 모델 개발을 목표로 한다.

1.2 데이터셋

사용할 데이터셋은 Kaggel 의 **Credit Score Classification** 이다.
데이터셋은 고객 인구통계, 소득/부채 등 재무지표, 그리고 결제/연체 패턴처럼 신용위험 평가에 핵심적인 요소들을 폭넓게 포함한다. 특히 *Credit_History_Age, Payment_Behaviour, Credit_Utilization_Ratio, Num_Credit_Inquiries* 등의 특성이 포함되어 있다.



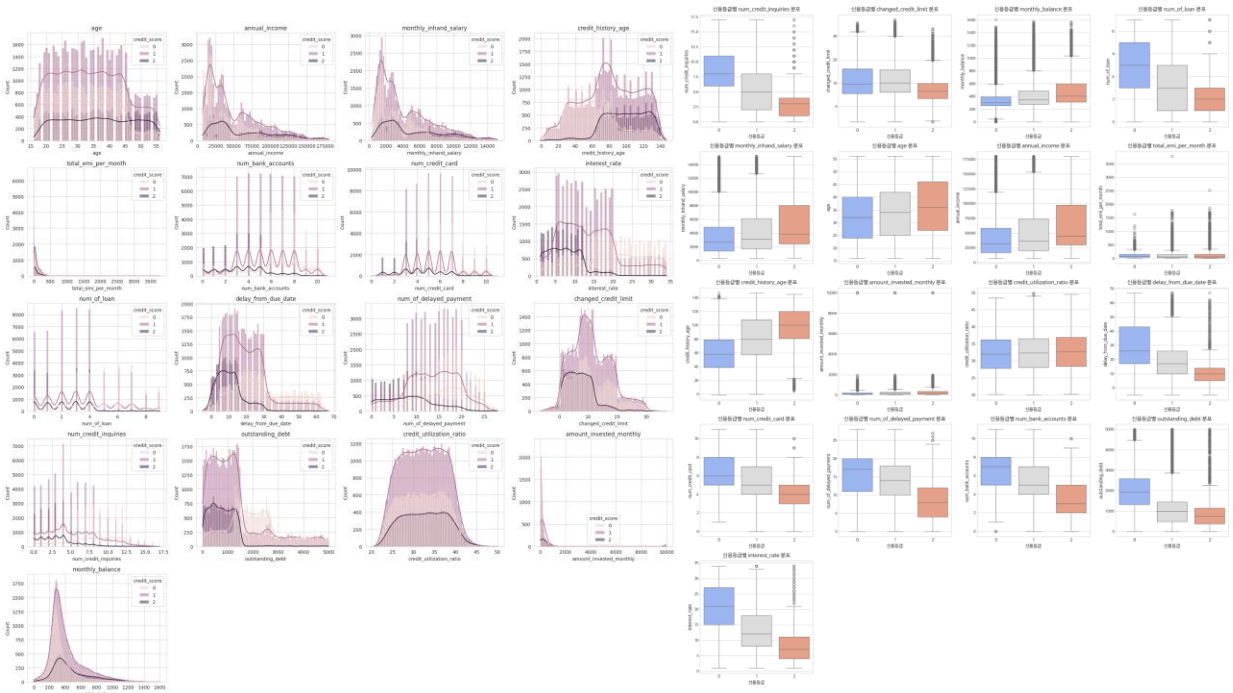
2. 데이터 분석 (EDA)

2.1. 타깃 변수 분석

Task 1 의 Target Variable ***Credit_Score*** 는 'Low', 'Standard', 'High'의 세 단계로 구성되어 있으며, 각 클래스의 분포를 살펴본 결과 순서대로 약 28%, 53%, 18%의 분포로 나타났다. 다소 불균형한 형태를 보이므로 모델 학습 시 클래스 편향에 유의해야 한다. Task 2 의 Target Variable ***Delay_from_due_date*** 는 대부분의 관측 시기가 5~30 일에 집중되어 있었고, 오른쪽으로 약간 치우침을 확인할 수 있었다.

2.2. 주요 변수 데이터 분포 및 이상치 탐색

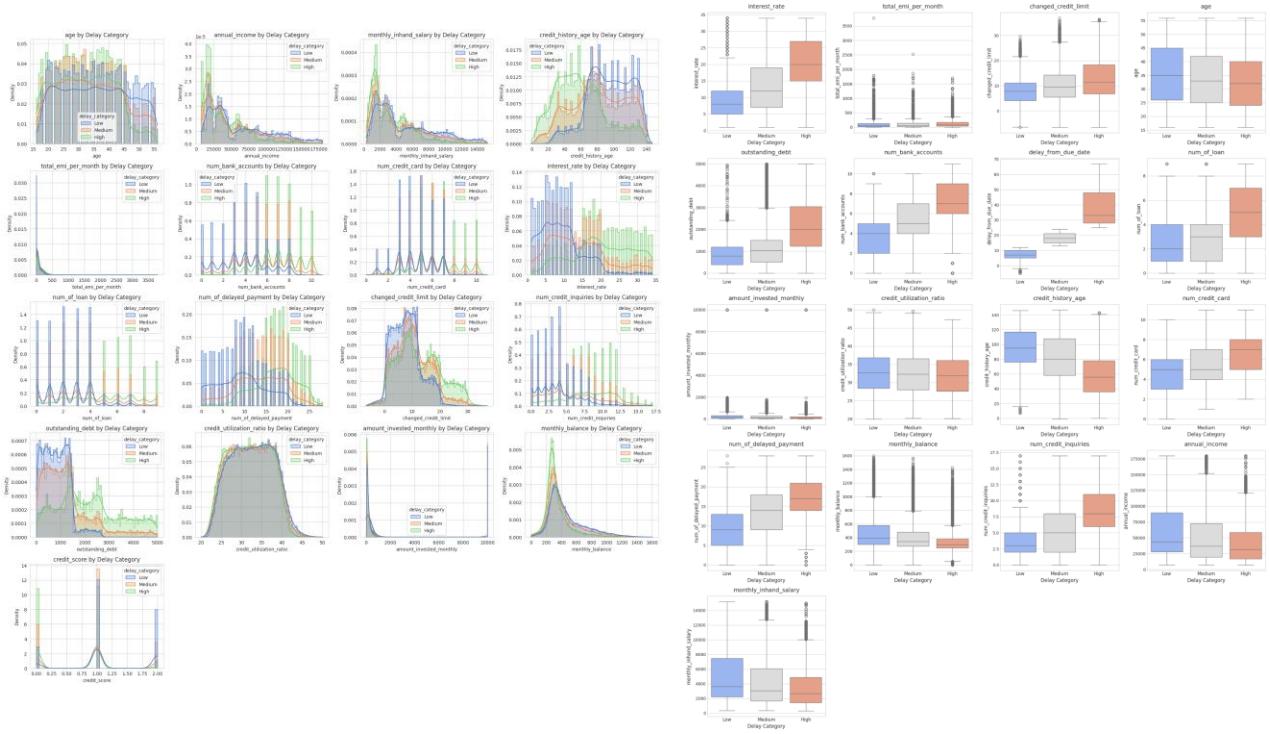
데이터셋의 주요 변수들에 대한 분포를 히스토그램과 박스플롯으로 나타내어 시각화해보았다. 이를 통해 변수별 데이터의 치우침과 이상치 존재 여부를 직관적으로 확인할 수 있었다. 이를 통해 변수별 데이터의 분포 특성과 함께, 두 가지 Task 의 Target Variable 이 데이터 전반에 걸쳐 어떤 경향성을 보이는지 확인할 수 있었다.



히스토그램에서는 각 수치형 변수에 대해 *Credit_Score* 를 hue 로 표현함으로써, 신용 점수 수준에 따라 변수값이 어떤 차이를 보이는지 시각적으로 비교하였다. 대부분의 변수에서 신용 점수 수준 Low(0)과 High(2)가 교차하는 지점이 발생하여, 의미를 도출할 수 있었다. 이러한 특성들이 해당 변수가 신용 등급 분류에 있어 유의미한 변수로 사용될 수 있음을 시사한다.

신용등급별 금융 특성을 박스플롯 기반으로 분석한 결과, 0 그룹은 *Num_of_Loan, Num_Credit_Card, Num_of_Delayed_Payment, Total_EMI_per_month, Num_of_Delayed_Payment, Delay_from_due_date* 이 상대적으로 높고, *Annual Income* 과 *Monthly_Balance* 은 낮아 재무 안정성이 낮은 것으로 나타났다. 일부 지표에서는 상향 극단치가 관찰되었다. 1 그룹은 저신용과 고신용 사이의

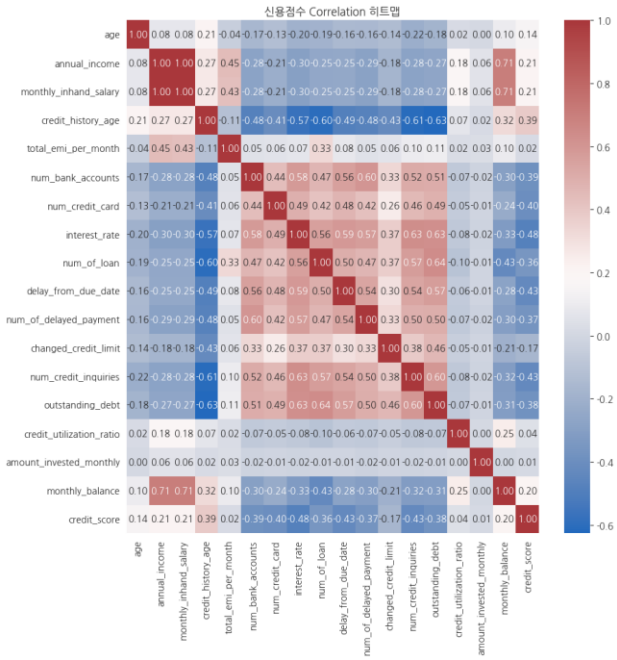
중간적 특성을 보이며, 전반적으로 균형 잡힌 재무 구조와 적정 수준의 부채·연체를 나타냈다. 2 그룹은 소득과 자산이 높고 신용거래 기간이 길며, 부채와 연체 수준이 낮아 안정적인 금융 상태를 유지했다. 이상치는 대부분 상향 형태로 확인되었으며, 하향 이상치는 거의 발견되지 않았다. 종합적으로, 신용등급은 금융 활동과 리스크 지표와 밀접히 연관되며, 등급이 낮을수록 부채와 연체가 증가하고, 등급이 높을수록 안정적 금융 패턴을 보이는 것으로 분석되었다.



연체일수 *Delay_from_due_date* 에 대한 히스토그램은 분포를 살피어 연체일수가 증가함에 따라 분포의 중심이 치우친 정도를 파악하여 회귀 모델에서 연체 위험도를 예측하는데 설명력을 가질 가능성을 확인하였다. *Age* 가 45-55 인 집단에서 연체일수가 낮은 집단이 많아지는 특징을 파악할 수 있었고, *Credit_History_Age* 가 클수록, *Annual_Income* 과 *Monthly_Inhand_Salary* 가 많을수록 안정적인 그룹으로 파악할 수 있었다. *Num_of_Delayed_Payment* 과 *Outstanding_Debt* 과 같이 직접적인 feature 뿐 아니라 *Num_Bank_Accounts, Num_Credit_Card, Change_Credit_Limit* 과 같이 간접적으로 보일 수 있는 feature 도 신뢰도 그룹에 영향을 준다는 경향성을 확인할 수 있었다.

2.4. 변수 간 관계 탐색

모든 수치형 변수의 상관관계를 계산하고, 히트맵을 만들어 시각화해보았다. 그 결과, *Monthly_Inhand_Salary* 와 *Monthly_Balnace* 간에는 0.71 의 매우 강한 양의 상관관계가 나타났으며, 이는 소득이 높을수록 월말 잔액이 많다는 것을 의미한다. 반면, *Credit_History_Age* 와 *Num_Credit_Inquires* 는 -0.61 의 음의 상관관계를 보여 신용 이력이 오래될수록 신용 조회가 줄어드는 경향이 있었다. 또한, 타깃 변수인 *Delay_from_due_date* 는 *Num_of_Delayed_Payment* 와 0.57 의 양의 상관관계를 보여, 연체 횟수가 많을수록 연체일수도 늘어나는 패턴이 확인되었다.



2.5. 데이터 전처리

원본 Credit Score Classification 원본 데이터셋에는 여러가지 데이터 품질 문제가 존재했다. 예를 들어, 일부 레코드에는 *Age* 가 음수로 표시되거나. *Occupation* 과 같은 범주형 변수에 결측치가 다수 존재했다. 이러한 이상치나 결측값은 모델 학습시 편향을 유발하거나 예측 안정성을 저하시킬 수 있기에, 이에 대한 데이터 정제 및 결측치 보정 과정을 수행하였다.

해당 데이터셋은 *Customer_ID* 로 구분된 12500 명의 사람의 1 월부터 8 월까지의 신용 정보로 구성되어 있다. 이를 바탕으로 대부분의 결측치는 예측이 가능했기 때문에 데이터 일관성을 유지하기 위해 문맥 기반 Imputation 을 진행했다. *Age* 에서 "24_"처럼 숫자에 문자가 섞인 값은 정규식으로 불순문자를 제거한 뒤 정수/실수형으로 강제 캐스팅했다. *Payment_Behavior, Credit_Mix* 등은 공백과 기호를 정리하고 희소한 레이블은 의미 기반으로 병합하여 카테고리를 표준화했다.

또한 모델 학습에 불필요한 식별자 속성 *ID, Customer_ID, Name, SSN* 을 제거하였으며, 범주형 데이터 *Unique_Loan_Types* 와 *Occupation* 은 One-Hot Encoding 을 통해 새로운 컬럼을 통해 표현함으로써 Logistic Regression 과 Linear Regression 에 적합한 형태로 변환했다. 이 외의 모든 범주형 데이터는 수치형으로 변환하여 모델이 이해할 수 있도록 표현하였다. 연속형 변수는 normalization 을 통해 scaling 하여 특정 속성이 다른 속성보다 과도하게 영향을 주지 못하도록 했다. 이러한 과정을 통해 데이터셋의 신뢰성을 확보하기 위해 노력하였고, 결과적으로 모델의 학습 안정성과 일반화 성능을 향상시킬 수 있었다.

3. 시도한 방법론 소개

3.1 Task 1: Classification

본 연구에서는 개인 고객의 금융 데이터를 활용하여 신용 점수를 세 단계(0: 낮음, 1: 중간, 2: 높음)로 분류하기 위해 **로지스틱 회귀(Logistic Regression)** 모델을 구축하였다. 로지스틱 회귀는 해석력이 높고 계산 효율이 뛰어나며, 각 특성이 신용 등급에 미치는 영향을 선형적으로 추정할 수 있다는 장점을 가진다. 데이터 전처리 과정에서는 결측치 처리, 범주형 변수의 원-핫 인코딩(One-Hot Encoding), 그리고 수치형 변수의 표준화(Standardization)를 수행하여 모델 입력의 일관성과 학습 안정성을 확보하였다. 모델은 학습용 데이터로 훈련한 후 테스트 데이터에 대해 성능을 검증하였으며, 정밀도(Precision), 재현율(Recall), F1 점수(F1-Score), 그리고 ROC-AUC(Receiver Operating Characteristic–Area Under Curve) 등의 분류 성능 지표를 활용하여 평가하였다. 이러한 평가지표를 통해 모델의 분류 정확도뿐만 아니라 불균형 데이터에 대한 판별력과 신뢰도를 종합적으로 검증하였다.

3.2 Task 2: Regression

본 연구에서는 개인 고객의 금융 데이터를 활용하여 연체 일수(Delay_from_due_date)를 예측하기 위한 회귀 기반의 신용평가 모델을 구축하였다. 데이터 전처리 단계에서는 먼저 결측치 존재 여부를 점검하여 데이터의 품질을 확보하고, 수치형 변수에 대해서는 표준화(Standardization)를 적용하여 변수 간의 스케일 차이를 보정하였다. 또한, 범주형 변수는 원-핫 인코딩(One-Hot Encoding) 기법을 활용하여 모델이 비정형 데이터를 수치적으로 인식할 수 있도록 변환하였다. 이후 예측 모델로는 해석력이 높고 계산 효율이 우수한 **선형 회귀(Linear Regression)**를 사용하였으며, 학습용 데이터로 모델을 훈련한 뒤 검증용 데이터에 대해 성능을 평가하였다. 모델의 성능 평가는 평균절대오차(MAE), 평균제곱근오차(RMSE), 결정계수(R^2) 등의 지표를 활용하여 수행하였으며, 이를 통해 회귀 기반 신용위험 예측 모델의 적합성과 안정성을 검증하였다.

4. 중간 결과 및 해석

4.1 Task 1: Classification

모델의 전체 정확도(Accuracy) 는 0.74, ROC-AUC 점수(ROC-AUC Score) 는 0.88 로 나타나, 전반적으로 양호한 분류 성능을 보였다.정밀도(precision), 재현율(recall), F1 점수를 종합적으로 보면 다음과 같은 특징이 있다.

- 클래스 0(신용 낮음): 재현율 0.78, F1-score 0.76 으로 안정적인 분류 성능을 보였다.
- 클래스 1(신용 중간): 정밀도는 0.86 으로 높았으나 재현율이 0.68 로 낮아, 일부 중간 등급 고객을 다른 등급으로 잘못 분류하는 경향이 있었다.
- 클래스 2(신용 높음): 재현율이 0.87 로 매우 높아 실제 신용이 높은 고객을 잘 찾아냈으나, 정밀도는 0.56 으로 낮아 오분류가 다소 발생했다.

혼동 행렬을 통해 확인한 결과, 특히 클래스 1 과 클래스 2 사이의 경계에서 오분류가 빈번했다. 이는 두 등급 간 데이터 특성이 유사하거나, 선형 모델의 한계로 인해 비선형적 결정 경계를 충분히 학습하지 못한 데에 기인한 것으로 판단된다. 그럼에도 불구하고, 높은 ROC-AUC 점수(0.88)는 모델이 전체적으로 클래스 간 구분 능력을 비교적 잘 학습했음을 시사한다.

Classification Report:

	precision	recall	f1-score	support
0	0.74	0.78	0.76	1374
1	0.86	0.68	0.76	2575
2	0.56	0.87	0.68	886
accuracy			0.74	4835
macro avg	0.72	0.77	0.73	4835
weighted avg	0.77	0.74	0.74	4835

4.2 Task 2: Regression

모델 평가 결과, 평균절대오차(MAE)는 약 4.24 일, 평균제곱근오차(RMSE)는 6.20 일, 결정계수(R^2)는 0.825 로 나타났다. 이는 모델이 전체 연체 일수 변동의 약 82.5%를 설명할 수 있음을 의미하며, 회귀 기반 예측 모델로서는 상당히 우수한 성능을 보인 것으로 해석된다. MAE 와 RMSE 모두 상대적으로 작은 값을 보이고 있으며, RMSE 가 MAE 보다 다소 높은 것은 일부 관측치에서 큰 오차(극단적인 연체값)가 존재함을 시사한다. 그러나 두 지표 간의 차이가 과도하지 않다는 점에서, 모델이 대체로 안정적인 예측력을 유지하고 있음을 알 수 있다.

또한, 결정계수(R^2)가 0.8 을 상회한다는 점은 모델이 고객의 신용 특성과 연체 행동 간의 관계를 효과적으로 포착하고 있음을 보여준다. 즉, 고객의 소득 수준, 부채 규모, 신용 이용률, 연체 이력 등의 주요 금융 특성이 연체 일수 예측에 유의미하게 기여하고 있음을 확인할 수 있다. 이는 단순히 고객을 '양호/위험'으로 구분하는 이분법적 분류를 넘어, 실제 상환 행태를 수치적으로 예측할 수 있는 정량적 신용평가 체계의 가능성을 제시한다는 점에서 의미가 있다.

```
=== Linear Regression Performance ===
MAE:  4.2377
RMSE: 6.2025
R² :   0.8251
```

5. 개선이 필요한 부분과 개선 계획

향후 Task1(Classification)와 Task2(Regression) 모두에서 모델 성능 향상을 위해 데이터 분석 기반과 모델 기반 피처 선택을 병행하여 중요 피처만 선별하는 개선을 계획하고 있다. 또한 현재는 모든 범주형 변수에 One-Hot Encoding 을 적용하고 있으나, 향후에는 모델 특성에 맞춰 인코딩 방식을 최적화할 계획이다. 구체적으로, 트리 기반 모델에는 Label Encoding 을, 선형/신경망 모델에는 기존처럼 One-Hot Encoding 을 적용할 예정이다.

5.1 Task 1: Classification

본 연구에서는 개인 고객의 금융 데이터를 활용하여 신용 점수(0: 낮음, 1: 중간, 2: 높음)를 분류하는 모델을 구축하였다. 초기 모델로 로지스틱 회귀(Logistic Regression)를 적용하였으며, 이 모델은 각 특성의 영향력을 직관적으로 해석할 수 있고 계산 효율이 높다는 장점이 있다. 그러나 변수 간

비선형 관계나 복잡한 상호작용을 충분히 반영하지 못한다는 한계가 있으며, 실제 금융 데이터에서는 여러 요인이 복합적으로 작용하는 비선형적 구조가 흔하게 나타난다.

따라서 이러한 한계를 보완하고 예측 성능을 향상시키기 위해, 보다 표현력이 높은 앙상블 및 딥러닝 기반 분류 모델을 순차적으로 적용할 계획이다. 우선, Random Forest Classifier 를 도입하여 다수의 결정트리를 결합함으로써 변수 간 비선형 관계와 복잡한 상호작용을 정교하게 학습할 수 있도록 할 예정이다. 이어서, XGBoost Classifier 를 활용하여 학습 과정에서 잔여 오차를 반복적으로 보정(Boosting)함으로써 예측 정확도를 높이고, 과적합(overfitting)을 방지할 계획이다. 마지막으로, 표형(tabular) 데이터에 특화된 FT Transformer(Fully Tokenized Transformer)모델을 적용하여, 기존 트리 기반 모델이 포착하기 어려운 고차원적 피처 상호작용과 비선형 패턴을 효과적으로 반영할 예정이다.

이러한 단계적 모델 개선 과정을 통해, 단순한 선형적 분류를 넘어 비선형적이고 복합적인 고객 특성 반응을 반영하는 정밀한 신용 점수 예측 체계를 구축하는 것을 목표로 한다. 향후 각 모델의 성능은 정확도, 정밀도, 재현율, F1-score, 다중클래스 ROC-AUC 등 다양한 평가 지표를 기준으로 비교·분석하여 최적의 분류 모델을 선정하고, 실무 적용 가능한 신용평가 시스템의 신뢰성과 효율성을 극대화할 계획이다.

5.2 Task 2: Regression

본 연구에서는 선형 회귀(Linear Regression) 모델을 활용하여 연체 일수(Delay_from_due_date)를 예측하였다. 해당 모델은 변수 간 관계를 해석하기 용이하고 계산 효율이 높다는 장점이 있으나, 비선형적 관계나 변수 간 상호작용(Interaction)을 충분히 반영하지 못한다는 한계가 있다. 실제 금융 데이터의 특성상, 고객의 연체 행동은 단일 변수의 영향보다는 여러 요인이 복합적으로 작용하는 비선형적 구조를 띠는 경우가 많다.

따라서 이러한 한계를 보완하고 예측 성능을 향상시키기 위해, 앙상블 및 딥러닝 기반 회귀 모델을 추가로 적용할 계획이다. 우선, Random Forest Regressor 를 도입하여 다수의 의사결정나무(Decision Tree)를 결합함으로써, 변수 간의 비선형 관계와 복잡한 상호작용을 보다 정교하게 학습할 수 있도록 할 예정이다. 다음으로, XGBoost Regressor 를 활용하여 학습 과정에서의 오차를 반복적으로 보정(Boosting)함으로써 예측 정확도를 향상시키고, 과적합(overfitting)을 제어할 계획이다. 마지막으로, 최근 구조화 데이터(tabular data)에서 높은 성능을 보이는 FT Transformer(Fully Tokenized Transformer) 모델을 적용하여, 전통적 회귀 모델이 포착하지 못하는 고차원적 피처 상호작용과 비선형 패턴을 효과적으로 반영할 예정이다.

이러한 단계적 모델 개선 과정을 통해, 단순한 선형 추정에 기반한 신용위험 평가를 넘어, 비선형적이고 복합적인 고객 특성 반응을 반영하는 정밀한 예측 체계를 구축하는 것을 목표로 한다. 향후 각 모델의 성능은 MAE, RMSE, R^2 등의 회귀 평가 지표를 기준으로 비교·분석하여, 최적의 예측 모델을 선정하고 신용평가 시스템의 실효성을 극대화할 계획이다.

Dataset: https://www.kaggle.com/datasets/parisrohan/credit-score-classification
Reference: https://www.kaggle.com/code/iremnurtokuroglu/credit-score-detailed-comprehensive-eda