

Proposal Report

19조

2022320062 정하연, 2022320090 조서윤, 2020170365 박가영

1. 문제 정의

1.1 문제 정의

금융기관은 고객의 대출 상환 가능성을 정확히 예측하여 리스크를 최소화하고 맞춤형 금융 서비스를 제공해야 합니다. 하지만 고객별 재무 상황과 상환 습관이 다양하기 때문에, 단순한 신용등급이나 지표만으로는 연체 위험을 정확히 평가하기 어렵습니다. 이에 금융기관이 합리적으로 대출 승인 여부와 한도를 결정할 수 있도록, 분류(Classification) 모델을 개발하고자 합니다. 아울러 연체 일수라는 연속적 위험 지표를 예측(Regression)함으로써 신용평가 체계의 정밀성과 합리성을 더욱 강화할 수 있습니다. 본 연구는 다음 두 단계의 Task로 구성됩니다:

Task1: 개인 고객의 금융 데이터를 활용하여 신용 점수를 세 단계로 분류

Task2: 연체 일수를 예측하여 보다 정밀한 신용평가 체계 구축

1.2 데이터셋

사용할 데이터셋은 Kaggle의 Credit score classification입니다.

<https://www.kaggle.com/datasets/parisrohan/credit-score-classification>

2. 문제의 흥미 및 중요성에 대한 의견 제시

개인 신용 평가 모델 개발은 전통적인 금융 서비스에서 한 단계 더 나아가, 데이터와 IT 기술을 결합한 스마트 금융(FinTech)과 직결된다는 점에서 흥미로운 주제라 생각했습니다. 또한 이는 단순한 학문적 연구에 그치지 않고 실제 금융기관과 고객의 의사결정에도 직접적으로 기여할 수 있기에 중요합니다. 첫째, 금융기관의 신용 평가 효율성을 높여 금융 시스템 전반의 개선에 이바지할 수 있습니다. 둘째, 부정확한 신용평가 문제를 해결함으로써 대출 부실률을 감소시킬 수 있습니다. 셋째, 나아가 개개인에게 맞춤형 금융 서비스를 제공하는 기반을 마련할 수 있습니다.

3. ML 적용의 적합성

3.1. 사용할 ML 방법론 설명

3.1.1. Task1: Classification

고객의 나이, 직업, 연소득 등 기본 정보와 다양한 금융 데이터를 Feature로 하고 Credit_Score를 Label로 하여 Logistic Regression, Random Forest Classifier, XGBoost Classifier, FT Transformer를 이용해 신용 점수를 Good/Standard/Bad 3단계로 분류하는 Classification 작업을 수행합니다.

3.1.2. Task2: Regression

고객의 나이, 직업, 연소득 등 기본 정보와 다양한 금융 데이터를 Feature로 하고 Delay_from_due_date를 Label로 하여 Linear Regression, Random Forest Regressor, XGBoost Regressor, FT Transformer 모델을 활용한 Regression 작업을 수행합니다.

3.2. 해당 방법론 선택 이유

3.2.1. Task1: Classification

신용 점수 분류 작업에서는 해석력과 예측 성능의 균형이 중요합니다. Logistic Regression은 직관적인 해석이 가능해 베이스라인으로 적합하며, Random Forest는 비선형 관계와 변수 상호작용을 잘 포착합니다. XGBoost는 불균형 데이터와 노이즈 처리에 강점을 가지며, FT Transformer는 복잡한 패턴을 학습해 최신 FinTech 신용평가에서 새로운 가능성을 제시합니다.

3.2.2. Task2: Regression

연체 일수 예측 회귀 작업은 단순 모델부터 고도화된 모델까지 단계적으로 적용할 수 있습니다. Linear Regression은 해석력이 높아 기본 성능 비교에 적합하고, Random Forest Regressor는 비선형적 상호작용과 이상치에 강합니다. XGBoost Regressor는 노이즈가 많은 금융 데이터 처리에 효과적이며, FT Transformer는 범주형·연속형 변수를 동시에 학습해 복잡한 패턴을 포착하는 데 강점을 가집니다.

4. 프로젝트 수행 일정

- 9월: 문제 정의, 데이터 수집
- 10월: EDA, 데이터 전처리, 피쳐 엔지니어링, 모델 학습(Task1)
- 11월: 모델 학습(Task2), 모델 검증 및 성능평가
- 12월: 결과 분석(반사실적 해석, Task1과 Task2 신용등급 간 비교분석)

5. 팀 구성 및 역할 분담

1. 데이터 전처리 및 EDA: 정하연, 조서윤
2. 모델 학습 및 검증: 정하연(Task1), 조서윤(Task2)
3. 결과 분석 및 해석: 박가영

6. 예상되는 결과

본 연구를 통해 개발되는 모델은 금융기관이 고객의 신용 위험을 보다 정확히 평가할 수 있도록 지원할 것으로 기대됩니다. Task1의 분류 모델을 통해 고객을 신용 점수 단계별로 구분함으로써, 대출 승인 여부와 한도 결정을 보다 합리적으로 수행할 수 있을 것입니다. 더 나아가 Task2의 회귀 모델은 연체 일수를 예측하여, 신용등급 체계를 세분화하고 리스크 관리의 정밀성을 강화할 것으로 예상됩니다. 이를 통해 금융기관은 대출 부실률을 낮추고, 고객 맞춤형 금융 서비스 제공이 가능해질 것으로 기대됩니다.