SW 최종 보고서

(Software Final Report)

Project Name	Design credit score classification system
Date	2023-12-09
작성자	32214684 최지원

<목차>

<표 목차>

<그림 목차>

1. 프로젝트 개요

1.1 프로젝트 명 및 기간

- 프로젝트 명 : Design credit score classification system
- 프로젝트 기간 : 2023.12.03 ~ 2023.12.10

1.2 프로젝트 소개

이 프로젝트는 확률 이론과 다양한 분류 방법을 활용하여 Credit Score Classification System 을 개발하는 것을 목표로 한다. Kaggle 에서 제공하는 신용과 관련된 데이터와 기계 학습 모델을 활용하여 고객의 신용 정보를 기반으로 Credit Score 을 예측하고 분류한다. 이 프로젝트를 통해 패턴인식과 관련된 지식을 익히며, 더불어 Credit Score 예측 모델 제작을 통해 금융 기관이나 다양한 업무 분야에서 신뢰성 있는 의사 결정을 내릴 수 있을 것으로 기대한다.

2. 배경지식

2.1 신용 정보

2.1.1 신용 등급이란

개인 신용 평점(등급)이란 신용조회화사가 향후 1 년 이내에 90 일 이상 장기 연체가 발생할 가능성을 통계적 분석 방법을 통하여 1~1000 점 (1~10 등급)로 수치화한 지표를 의미한다.
신용 등급이 높고 점수가 높을수록 신용상태가 우수하다. 등급 별 점수구간은 신용조회화사별로 상이함으로 개인 신용 등급은 신용 조회 회사에서 운영하는 사이트에 접속하여 확인해야한다.

2.1.2 신용 등급 별 특징

등급	구분	의미 및 특징
1~2 등급	최 우량	오랜 신용거래 경력과 다양하고 우량한 신용거래 실적을 보유하고 있어 부실화 가능성이 매우 낮음
3~4 등급	우량	활발한 신용거래 실적은 없으나, 꾸준하게 우량한 거래를 지속한다면 상위등급 진입이 가능하며 부실화 가능성은 낮은 수준
5~6 등급	일반	비교적 금리가 높은 금융업권과의 거래가 있는 고객으로 단기연체 경험이 있으며 부실화 가능성은 일반적 수준
7~8 등급	주의	비교적 금리가 높은 금융업권과의 거래가 많은 고객으로 단기연체의 경험을 비교적 많이 보유하고 있어 부실화 가능성이 높음
9~10 등급	위험	현재 연체 중이거나 매우 심각한 연체의 경험을 보유하고 있어 부실화 가능성이 매우 높음

[표 2.1.2] 신용 등급 별 특징

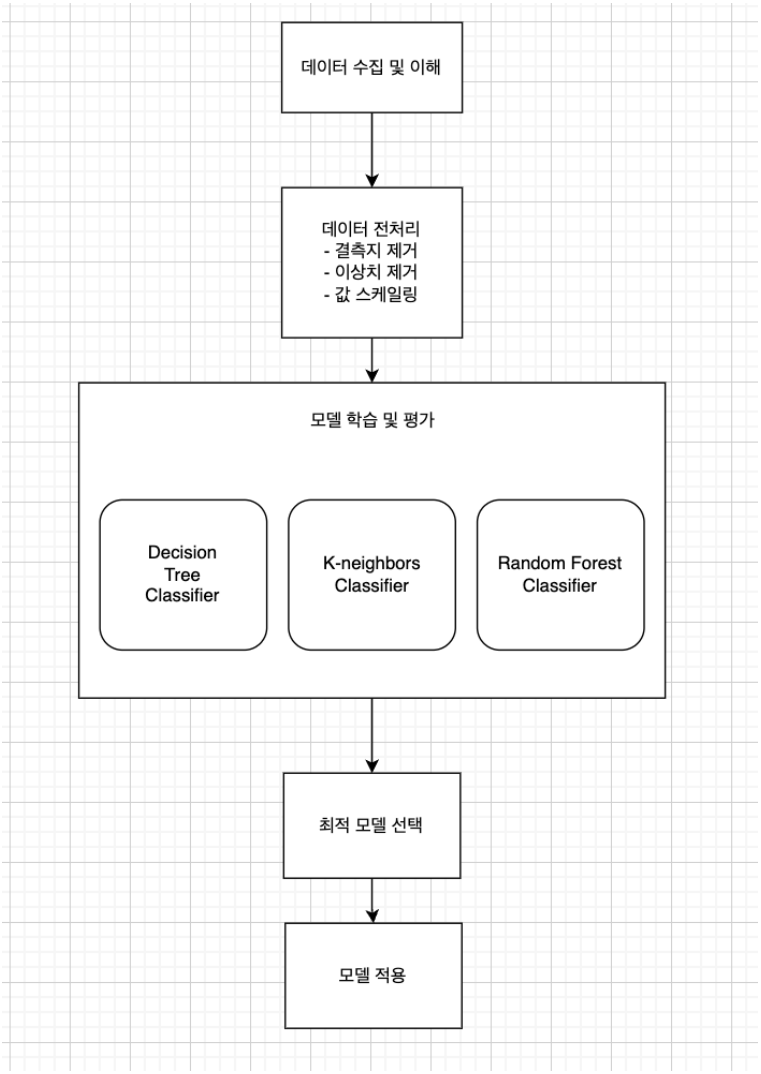
2.1.3 개인 신용 평가 반영 요소

긍정적 반영 요소	대출금 상환 이력
	신용카드 사용 금액 및 기간
	연체 상환 및 연체 상환 후 경과기간
	통신, 공공요금 성실 납부 실적
부정적 반영 요소	대출금 연체
	신규대출 및 대출건수 증가
	제 2 금융권 대출
	과도한 현금 서비스 이용

[표 2.1.3] 개인 신용 평가 반영 요소

3. 설계 및 아이디어 설명

3.1 설계 흐름



[그림 3.1] 설계 흐름

3.1.1 데이터 수집 및 이해

Kaggle 에서 제공하는 신용 관련 정보인 train data 와 test data 을 수집하였다. 데이터에는 어떤 특성들이 포함되어 있으며, 각 특성이 신용 스코어 예측에 어떤 영향을 미칠 수 있는지에 대해 이해하기 위해 데이터를 탐색했다.

3.1.2 데이터 전처리

누락된 값, 이상치, 또는 불필요한 특성을 처리하여 모델 학습에 적합한 형태로 데이터를 가공했다. 이 과정에서 데이터 스케일링과 인코딩 과정도 수행하였다. scikit-learn 라이브러리에서 제공하는 MinMaxScaler 을 사용하여 데이터 스케일링을 하였으며 범주형 데이터는 적절한 인코딩 기법을 적용하여 변환하였다.

3.1.3 모델 학습 및 평가

전체 데이터셋 중 80%는 모델 학습에, 나머지 20%는 모델 테스트에 사용하여 데이터를 랜덤하게 분할하였다. 세 가지 다른 분류 모델인 Decision Tree Classifier, K-Neighbors Classifier, Random Forest Classifier 을 활용하여 각 모델의 성능을 평가하였다.

3.1.4 최적 모델 선택

위에서 진행한 세 가지 분류 모델 중 테스트 데이터셋에서 가장 높은 정확도를 보인 모델을 선택하여 사용하였다. 테스트 결과, Random Forest Classifier, Decision Tree Classifier, K-Neighbors Classifier 순으로 accuracy 가 높게 나왔다.

3.1.5 모델 적용

Random Forest 을 활용한 설계한 classifier 모델을 새로운 test set 에 적용하여 결과를 확인하였다.

3.2 시스템 환경

개발 환경으로 Google Colab 을 사용하였다.

3.3 Classification(분류) 알고리즘

Classification 은 머신 러닝의 대표적인 지도학습 유형 중 하나로 주어진 data 을 클래스 별로 구별해 내는 과정이다. 데이터가 주어지면 학습된 모델을 통해 어느 label 에 속한 데이터인지 판단하고 예측하는데, 분류를 위한 다양한 알고리즘이 존재한다.

3.3.1 KNN (K-nearest neighbor)

데이터로부터 거리가 가까운 K 개의 다른 데이터의 레이블을 참조하여 분류하는 알고리즘이다. 거리를 측정할 때 유클리드 거리 계산법을 사용한다.

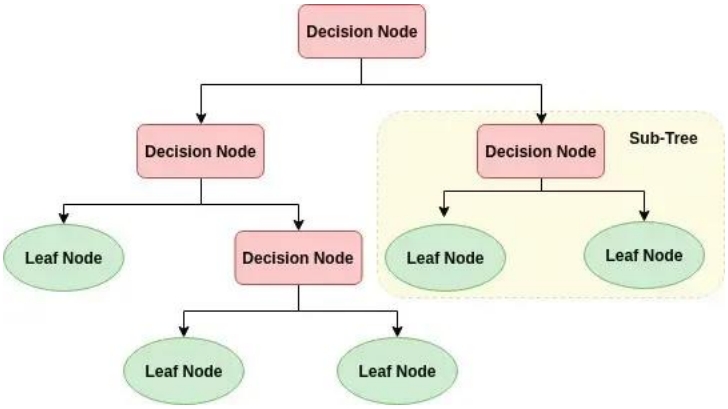
$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

[그림 2.2.1] 유클리드 거리 계산 식

3.3.2 Decision Tree

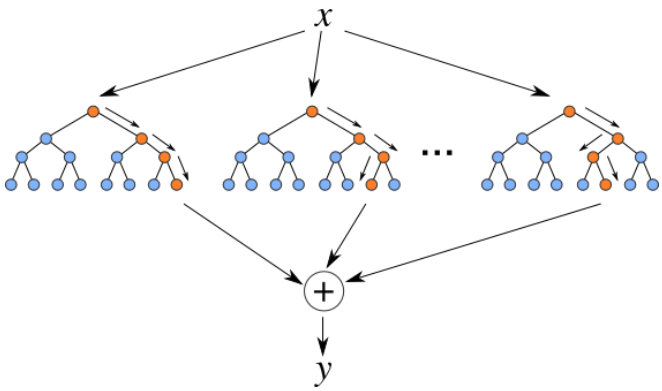
트리 기반의 분류 알고리즘으로 Decision node 에서 규칙에 따라 분할되며 각각의 서브 트리를 생성한다. 계속적으로 규칙에 따라 노드가 분할되며 최종적으로 Leaf Node 에서는 클래스 값을 가진다. 규칙 노드가 많아지게 되면 모델이 복잡해지고 과적합이 발생할 수 있기에 적절하게 트리를 구성해야 한다.



[그림 2.2.2] 의사결정 트리

3.3.3 Random Forest

Decision Tree 모델의 한계점을 극복하기 위해 개선 및 확장한 알고리즘이다. Decision Tree 는 트리의 높이에 따라 과적합으로 빠질 수 있는 가능성이 있으며 특이 값에 민감한 모델이 되기 쉽다. 랜덤 포레스트는 학습을 위한 데이터셋에서 데이터 일부를 뽑아 의사결정 트리를 만들고 각 의사결정 트리에게 같은 데이터셋을 입력으로 넣어 어떤 결과를 예측하는지 종합하여 판단한다.



[그림 2.2.3] Random Forest

4. 설계 제한 요소

경제성	금융기관이나 다양한 분야에서 이 모델을 통해 의사 결정에 도움이 될 수 있기에 유용한 소프트웨어이다.
미학성	본 설계 모델은 사용자를 위한 웹/앱 인터페이스를 구축하지 않았기에 사용자 편의성은 떨어진다.
윤리성	윤리, 보건 및 안전적으로 문제가 되지 않는 범위의 소프트웨어 이므로 윤리성을 충족한다.

생산성 및 내구성	필요시 입출력 데이터를 효과적으로 처리하고, 이상종료가 발생하지 않는 소프트웨어 시스템이므로 내구성이 뛰어나다.
산업 표준	산업 표준을 위반하지 않는다.

[표 4] 설계 제한 요소

5. 성능 평가

5.1 평가 방법

분류 모델을 평가하기 위해 Confusion Matrix 을 기반으로 Accuracy, Precision, Recall, F1 score 을 측정한다.

5.2 평가 데이터

수집한 train data 의 20%을 테스트에 사용하여 분류기의 결과와 실제 값을 비교한다.

5.3 평가 결과

	precision	recall	f1-score	support
Bad	0.73	0.75	0.74	1391
Standard	0.77	0.79	0.78	2618
Good	0.65	0.53	0.58	707
accuracy			0.74	4716
macro avg	0.71	0.69	0.70	4716
weighted avg	0.74	0.74	0.74	4716

[그림 5.3] 평가 결과

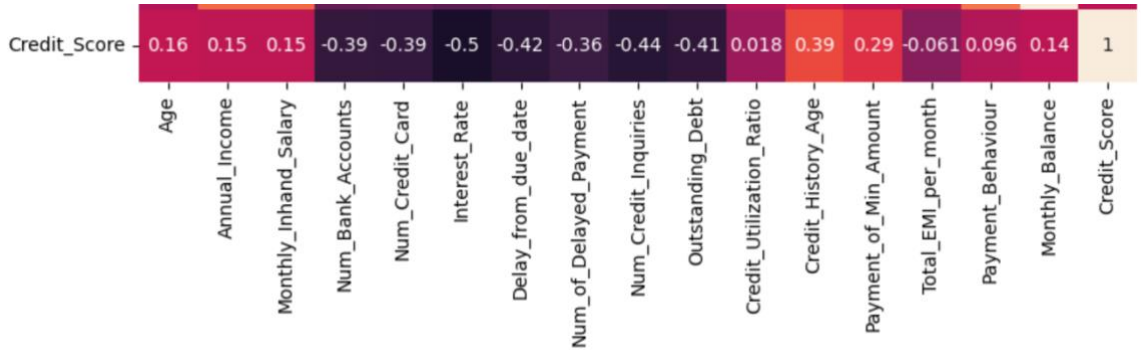
Random forest classifier 을 적용한 모델에 대해 scikit-learn 에서 제공하는 classification_report 함수를 사용하여 성능을 평가하였다.

먼저, 이 모델의 정확도는 74%이다. 각 class 별로 precision 과 recall 값을 통해 얼만큼 그 class 로 예측이 되었으며, 예측 된 것 중 실제로 동일한 값은 얼마인지 확인할 수 있다. F1-score 은 precision 과 recall 의 조화평균을 나타내며 support 는 각 클래스의 실제 샘플 수 이다.

종합적인 해석하자면, 해당 분류 모델이 Bad 및 Standard 클래스에 대해서는 상대적으로 좋은 성능을 보이지만, Good 클래스에 대해서는 정확도가 낮고 성능이 떨어지는 것으로 나타난다.

6. 토의 및 결론

6.1 feature 간 연관성



[그림 6.1] featurer 간 연관성

데이터의 feature 간의 상관관계를 분석하였을 때 Credit_Score(Target)에 가장 영향을 주는 features 는 Interest_rate(금리), 신용 조회 횟수, delay from due date, outstanding debt(현재 부채 수준) 등으로 확인할 수 있었다.

6.2 model 간 비교

Credit score 예측 문제에 대해 KNN, Decision Tree, Random Forest 모델을 바탕으로 데이터를 학습시킨 결과, 가장 성능이 우수한 모델은 Random Forest Classifier 을 활용한 것이었다.

6.3 모델의 한계점

Random Forest Classifier 을 활용하여 학습시킨 Classification 모델의 성능을 평가하였을 때, 74%의 정확도를 보였다. 하지만 해당 분류 모델의 한계점은 Good 클래스에 대해 상대적으로 성능이 떨어지는 것을 확인할 수 있었다.

6.4 고찰

Credit Score Classification 설계를 통해 확률 이론과 분류의 개념에 대한 깊은 이해를 얻었다. 다양한 신용 정보를 활용하여 신용 점수를 예측하는 분류 모델을 개발함으로써, 금융 기관 및 다양한 업무 분야에서 믿을 수 있는 의사 결정을 내릴 수 있을 것으로 기대한다.

그러나 앞서 언급한 모델의 한계로 인해 특정 클래스에 대한 성능이 낮게 나타나는 부분이 확인되었다. 이러한 한계를 극복하고 모델의 성능을 향상시키기 위해 후속 연구를 통해 추가적인 개선 및 최적화를 수행할 필요가 있다. 더 나아가, 향후 연구에서는 다양한 변수 및 데이터 특성을 고려하여 모델의 강건성을 향상시키고 다양한 상황에서의 적용 가능성을 탐색하는 것이 필요하다. 이를 통해 현업에서 높은 수준의 예측 성과를 달성할 수 있을 것으로 기대한다.