제목	Z세대를 위한 대안신용평가모델 - 구직사이트의 합격자소서데이터를 중심으로

김효원, 이수정, 이은서

초록:

최근 비금융데이터를 활용하여 신용거래가 적은 씬파일러의 신용을 평가하는 대안신용평가모델이 각광을 받고있다. 일례로 신용대출 상환에 영향을 미친다고 알려져 있는 Big 5 성격(외향성, 성실성, 민감성 등)을 SNS 사용 이력에서 찾는 선행연구가 있다. 선행연구에 의하면 개방성, 외향성, 성실성,신경성(민감성)이 높다면 상환 가능성이 컸으며 개인의 성격과 밀접한 SNS 사용특성을 포함한 빅데이터 자료를 대출 심사 시에 적용하면, 금융 거래 기록이 없는 Thin-Filer에 대한 대출 실행 시 유용할 것이라는 결론을 도출했다.

이 연구에서는 사회초년생인 씬파일러를 타게팅한 대안신용평가모델을 만들기 위해 사회초년생들의 진솔한 내용을 반영하여 성격을 잘 파악할 수 있을 자기소개서 데이터를 활용했다는 점에서 선행연구와의 차별성이 있다. 자소서 데이터로 외향성, 성실성, 민감성 각각의 성격을 yes/no로 분류하는 multi-label multi classification 모델(LSTM, GRU, BERT 등)을 생성했으며, 월급, 자가유무 등 전통적인 금융 정형데이터를 활용하여 대출상환여부를 파악하는 머신러닝 모델(XGBoost, LGBM, Random Forest 등)을 만들어 결합한 대안신용평가 모델을 만들었다.

이 연구를 발전시켜 Z세대를 위한 대안신용평가모델을 만들고, 적절한 금융상품을 추천하여 기존 신용평가모델로 확보할 수 없었던 새로운 고객층을 확보할 수 있을 것이라 기대한다.

Ⅰ. 서 론

1.1 분석 배경

개인신용평점이란 NICE신용정보, KCB, SCI신용정보와 같은 개인신용평가회사(Credit Bureau, CB사)가 개인에 대한 신용정보를 수집한 후 이를 통계적 방법으로 분석하여, 향후 1년 내 90일 이상 장기연체 등 신용위험이 발생할 가능성을 수치화(1-1000점)하여 제공하는 지표를 말한다. 개인신용평점은 금융회사 등이 개인의 신용을 바탕으로 의사결정이 필요한 경우(대출실행, 카드개설 등)에 참고지표로 활용할 수 있다(NICE평가정보, 2018). 그런데 이러한 신용평가는 신용거래가 없는 경우 개인신용평점은 낮을 수밖에 없는 구조이다. 따라서 신용거래가 적은 사회초년생 등과 같이 신용평점이 낮은 고객들의 경우 대출 기회가 제한될 수밖에 없는 문제를 내포하고 있다.

이에 신용거래가 적은 사회초년생을 위한 대안신용평가모델을 만들어 사회초년생에게 필요한 금융 거래를 돕고자한다.

대안신용평가의 예시로는, 개인의 성향을 추론할 수 있는 SNS 사용특성이 신용대출 상환에 미치는 영향 요인을 분석한 연구자료가 있다. 경험에 대한 개방성, 성실성, 외향성, 친화성, 신경성 등의 성격 특성과 신용대출 상환의 관계성을 파악하고, 해당 성격 특성들을 SNS 사용 특성에서 찾아내는 것이다.

비슷하게, 텍스트 데이터에서 개인의 성향을 추론하여 대출 상환 위험 성격을 파악하고, 해당 성격을 반영하여 대안신용평가모델을 만들고자 한다. 이때 사회초년생들이 가장 많이 사용하며, 진솔한 내용이 담겨있기에 성격을 잘 파악할 수 있으리라 기대하는 기업 합격자의 자기소개서를 사용할 것이다.

1.2 선행 연구

본 연구 주제와 관련된 선행 연구 논문을 분석하고 요약하였다.

「SNS 사용특성이 신용대출 상환에 미치는 영향에 관한 연구 (정원훈, 이재순), 2021」에서 SNS 사용특성(프로필 사진, 사용량)이 대출 상환에 미치는 영향을 살펴보고 신용평가에 활용될 수 있는지를 검증하였다. 이를 위해 SNS 대출상품인 T사 A대출 프로그램 이용자를 대상으로 이항로지스틱분석을 사용하였다. 분석 결과 첫째, 개방성, 외향성이 높은 '캐릭터·유머', 성실성이 강한 '취미 등 사회활동', 자기주관이 강하고 신경성이 높은 성향인 '사물·동물' 등을 프로필 사진으로 선택한 경우 상환 가능성이 컸다. 둘째, SNS상의 총 게시글 수가 많을수록 상환 가능성이 컸다. 셋째, 대출 기간이 길수록 연체 가능성이 커지는 것으로 나타났다. 따라서, 개인의 성격과 밀접한 SNS 사용특성을 포함한 빅데이터 자료를 대출 심사 시에 적용하면, 금융거래 기록이 없는 Thin-Filer에 대한 대출 실행 시 유용할 것이라는 결론을 도출했다.

「신용카드회사의 개인사업자 신용평가 업무에 관한 연구: 머신러닝 모델의 도입 (이건희, 이기환), 2022」에서 신용카드회사에 대한 개인사업자 신용평가 업무가 가능하게 됨에 따라 효율적으로 그러한 업무를 추진하는 방안을 연구했다. 따라서 선행 연구와 기존의 평가전문회사, 핀테크 기업의 사례를 중심으로 방안을 도출하였다. 선행 연구의 대부분이 신용평가와 부도 예측에서 머신러닝 기법의 활용을 주장하고 있다. 또한, 데이터의 수집과 활용에서도 재무 변수 중심의 은행의 심사 자료가 아니라 실시간의 즉각적인 대안 자료가 신용평가에 도움이 되는 체제로 가고 있다.

결론적으로 첫째, 신용평가 모델과 방법에서 머신러닝 기법을 도입하여 빅데이터를 처리할 수 있고 즉시 처리가 가능하도록 신용

평가를 수행해야 한다. 둘째, 대안 정보를 수집하고 집적하여 신용평가에 활용해야 한다. 개인사업자 대표의 개인신용정보, 모바일 사용 내역, 집세 연체 등 비금융정보의 수집과 활용이 필요하다.

「SNS 행동데이터가 신용평가에 미치는 영향 (이지형), 2021」에서 비금융정보 중 SNS 데이터를 활용하여 개인신용평가모형을 개발했다. 국내 핀테크사의 실제 데이터를 사용하였으며 전통적으로 인정받고 통용되는 개인신용평가모형 개발방법론(로지스틱회귀를 활용한 방법론)에 준용하여 모형을 설계 및 개발하였으며 개발된 모형의 안정성, 변별력을 계량적으로 검증하여 최종 모형을 만들었다.

결론적으로 첫째, 전통적인 금융 이력 데이터뿐만 아니라 핀테크 데이터도 금융 소비자의 신용도를 판단하는 데 의미 있고 중요한 평가 항목으로 활용 가능함을 증빙하였다. 둘째, 핀테크 스코어의 적용 타당성을 전통적인 개인신용평가모형 검증 방법론에 의해 통계적이고 계량적으로 증빙하였다. 셋째, 대안신용평가의 가장 중요한 목적인 금융이력 부족자를 대상으로 새로운 개인신용평가모형을 제시하였다.

「Deep Learning-Based Document Modeling for Personality Detection from Text (Navonil Majumder), 2017」에서는 James Pennebaker and Laura King&s stream-of-consciousness essay dataset(에세이를 쓰고, 각 에세이별 글쓴이의 성격 유형을 조사한 데이터)를 사용하여 에세이에서 성격유형을 예측하는 모델을 소개한다. 분석 프로세스는 데이터 전처리, 필터링, feature 추출, 분류 4단계로 구성된다.

1.전처리: 특수문자 제거, 소문자 통일

2.필터링: 성격 특성을 나타내지 않는 데이터 삭제

3.feature 추출: word2vec 임베딩

4.분류: CNN 모델로 성격 분류

결과적으로는 55-60 내외의 정확도를 보였다.

「Personality Prediction Based on Text Analytics Using Bidirectional Encoder Representations from Transformers from English Twitter Dataset (Joshua Evan Arijanto), 2019」에선 트위터의 영어 텍스트 데이터셋을 이용하여 성격 예측 시스템을 구축하는 연구를 진행했다. 전처리된 성격 집합을 분류하기 위해 SVM, CNN, BERT 분류기를 이용하여 성능을 비교했고, BERT 모형을 적용했을 때 가장 높은 수행 점수를 얻는다는 것을 발견했다. 미리 훈련된 BERT를 사용하여 미세 조정 방식을 사용하면 더 나은 결과를 도출할 수 있었다. 딥러닝 모델은 많은 데이터를 필요로 하기 때문에 데이터셋이 부족하다는 것이 연구의 주요 한계로 보았다.

「Research on automatic labeling of imbalanced texts of customer complaints based on text enhancement and layer-by-layer semantic matching (Xiaobo Tang), 2021」에서는 고객 불만 데이터에 대해 deep label 과 shallow label로 나누어 자동으로 라벨링해주는 방법을 제시한다. BERT 모델로 텍스트를 shallow label로 분류하고, word2vec 기반의 semantic similarity를 파악하여 deep label을 붙인다. 이 과정을 통해 한 번에 deep label로 분류하는 것보다 정확성을 높였다. 학습한 데이터가 기업의 협조가 필요한 데이터이기 때문에 데이터가 부족하다는 것이 연구의 주요 한계였다.

이러한 선행 연구들을 보았을 때, 각종 MZ세대 대상 데이터에서 신용 성향을 파악하고 적합한 머신러닝 모델을 만들면 좋겠다는 결론을 도출하였다. 또한 데이터에서 성향을 파악하는 기존 논문의 정확도가 높지 않았기 때문에 정확도를 향상시키기 위한 여러 대안을 사용하는 것이 필요해 보인다.

Ⅱ. 본 론

2.1 활용 데이터

2.1.1 Loan Prediction Based on Customer Behavior 데이터

kaggle에서 찾을 수 있는 데이터로 이는 Univ.AI에서 주최하는 해커톤에서 제공한 것이다. 252000개의 행과 11개의 열로 구성되어 있으며, risk_flag의 1은 채무 불이행, 0은 채무 이행으로 신용평가를 이진분류로 진행할 수 있다.

2.1.2 big5 분류 데이터

아래 3개의 데이터들은 영어로 작성된 것으로 성격 분류(개방성, 외향성, 친화성, 민감성, 성실성)를 진행한 데이터이다. 이후 더 자세히 언급한 것이지만 이 데이터를 활용하여 모델을 만든 후 자소서 데이터의 라벨을 붙이는 방법으로 진행할 것이다.

(1) 에세이 데이터

Deep Learning-Based Document Modeling for Personality Detection from Text (https://github.com/SenticNet/personality-detection)의 데이터는 설문을 통해 해당 성격을 가진 사람들이 에세이를 작성하게 한 데이터로 총 2467개의 행으로 구성되어 있다.

(2) myPersonality 데이터

Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging (https://github.com/facebookresearch/fairseq/tree/main/docs)의 데이터는 facebook 데이터로 이는 해당 논문의 저자에게 메일을 보내 얻을 수 있었다. 이는 문장별로 성격분류가 되어 있으며 총 9917개의 행으로 구성되어 있다.

(3) github 데이터

성격분류를 예측하는 github (https://github.com/CMWENLIU/personality_prediction)에서 찾은 데이터는 한 사람이 작성한 글이 여러 열에 의해 작성되었으며 성격 분류 열의 값은 다른 데이터들과 달리 범주형이 아닌 연속형으로 점수화 되어 있다, 해당 데이터는 924개의 행으로 구성되어 있다.

2.1.3 합격 자기소개서 데이터

채용 정보를 알려주는 사이트인 <링커리어>와 <잡코리아>에서 볼 수 있는 취업에 성공한 사람들의 자기소개서 데이터이다. 현 시점(2022.09.22.) 기준 <링커리어>에선 2013년 상반기부터 2022년 상반기까지 총 16,438건의 합격 자기소개서를, <잡코리아>에선 2015년 하반기부터 2022년 상반기까지 총 7,320건의 합격 자기소개서를 열람할 수 있다. 본 연구에선 링커리어, 잡코리아에서 직접 크롤링한 약 23,000건의 합격 자기소개서 데이터를 이용한다.

2.2 분석 방법론

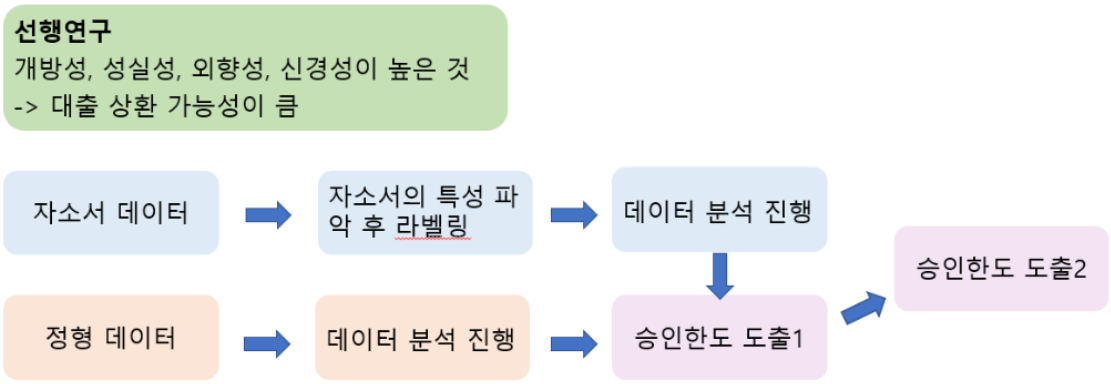

그림 1. 분석 플로우

분석 프로세스는 데이터 형성, 데이터 전처리, 모델링 3단계로 구성된다.
1.데이터 형성:
정형데이터: kaggle 수집 (https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior)
비정형 데이터: 선행 연구 데이터, 자소서 데이터 수집
2. 데이터 전처리
정형 데이터: 데이터 불균형 해소, 범주형 데이터 catboost encoding
비정형 데이터: 번역, 특수문자 제거, 소문자 통일, okt 토큰화, word2vec 임베딩
성격 특성을 나타내지 않는 데이터 삭제
3. 모델링
정형데이터: XGBoost, LGBM, AutoML 모델로 성격 열을 추가하여, 대출 상환 예측
비정형데이터: LSTM, GRU, BERT 모델로 성격 예측

2.2.1 데이터 형성
선행 연구에 따르면 개방성 및 외향성, 성실성, 신경성에 있어 강한 성향을 띠면 대출 상환 의지가 강함을 알 수 있다.
성격의 분류 중 외향성은 &대인 관계의 범위&를 나타내는 것으로, 외향적인 사람은 사교적이고 자신감이 넘치며 적극적이다. 2016년 ㈜핀테크가 개발한 페이스북과 카카오스토리 등의 SNS 활동 내역을 분석해 개인 신용등급을 평가하는 소셜 신용평가시스템에서도 외향성과 개방성은 대출 상환에 긍정적인 요소로 작용하였다. 성실성은 &목표에 관심을 집중시키는 정도&를 나타내는 것으로, 성실한 사람은 체계적이고 책임감 있게 일을 하며 성과 지향적이다. 신경성을 대변하는 민감성은 &다양한 환경에 대해 반응하는 정도&를 나타내는 것으로, 민감한 사람은 변화에 대해 민감하게 반응하거나 변덕이 심한 편이다. 대출 측면에 있어 외향성과 성실성은 긍정적인 성격으로, 민감성은 부정적으로 판단할 수 있다.
이에 의거하여 라벨링을 진행하였다.

2.2.2 데이터 전처리
우선 자소서를 크롤링한 데이터엔 라벨링이 없기에 영어로 작성된 데이터들(에세이,myPersonality,Github)을 한국어로 번역한다. 해당 데이터들에는 성격 분류 라벨링이 되어 있기에 분류 모델을 만들 수 있다. 이 모델은 1차 모델로 진행될 것이며, 성능이 좋지 않다면 데이터 증강 기법을 활용하여 기본 모델의 예측력을 높인 다음 자소서 문단의 라벨을 붙이고자 한다.

2.2.2.1 정형 데이터 전처리
원핫 인코딩을 적용할 시에는 정형데이터 열의 개수가 11에서 411로 매우 커져 이후의 모델링 과정에서 차원의 저주 현상이 쉽게 발생할 수 있다. 이에 열의 수를 증가시키지 않는 catboost encoding을 사용하여 범주형 변수에 변화를 주었다.

2.2.2.2 비정형 데이터 전처리
영어로 작성된 데이터들을 한국어로 번역한 후 전처리를 진행하였다.

(1) 에세이 데이터 (2467 행)
성격과 관련된 감정 단어가 들어가 있지 않으면 제거한다. 라벨링을 수기로 진행할 때, 앞뒤 문맥을 따져 성격 라벨링을 하는 경우도 존재하지만 대부분 성격과 관련된 감정단어의 존재 유무로 라벨링 하는 경우가 대다수이다. 따라서 감정단어가 존재하지 않으면 성격을 분류하기 어려운 낮은 질의 데이터라고 가정하여 제외하였다.

(2) myPersonality 데이터 (9917행)
영어로 된 문장이므로 파파고를 사용하여 한글로 번역해주었다. 페이스북 데이터 특성상 존재하는 이모티콘(:),XD)과 특수문자를 제거했다.

(3) github 데이터 (외향성: 2993행, 성실성: 2953행, 신경성: 2820행)
성격 라벨링이 범주가 아닌 점수로 연속형 값을 가지므로 점수가 평균보다 높으면 y, 평균보다 낮으면 n으로 바꿔주었다.
tokenizer로는 okt, 임베딩은 word2vec을 사용하였다.

(4) 합격 자기소개서 데이터 (11799행)
자기소개서 답변에서 불필요한 요소(글자수 등)와 특수문자를 제거했다. 답변에서 성격이 드러나는 '장점', '단점', '장단점'과 관련된 1477개의 문항을 선별한 후, 각 답변을 문장 단위로 분리하였다. 문장마다 감정 단어의 포함 여부를 확인하고, 감정 단어가 없는 문장은 성격을 분류하기 어려운 데이터라 판단하여 제거하였다.

2.2.3 모델링
2.2.3.1 정형데이터모델링
정형데이터(대출 상환 여부를 예측하는데 사용되는 금융 관련 정형데이터)로 대출상환여부를 예측하기 위한 머신러닝 모델을 만들었다. 수업 시간에 수강한 Random Forest 모델(AutoML 라이브러리를 통해 도출된 최선의 모델)과 Random Forest 모델의 단점을 개선한 모델인 XGBoost와 LGBM 모델을 생성했다.

(1) XGBoost
XGBoost는 앙상블 부스팅 기법의 한 종류이다. 이전 모델의 오류를 순차적으로 보완해 나가는 방식으로 모델을 형성하는데, 이전 모델에서의 실제값과 오차(loss)를 훈련 데이터를 투입하고, gradient를 이용하여 오류를 보완하는 방식을 사용한다. XGBoost의 분류 라이브러리 중 하나인 XGBClassifier 모델을 이용해 데이터 분류를 진행하였다. hyperparameter를 조절한 모델에 데이터를 학습시킨 결과 accuracy 0.8665의 성능을 내는 것을 확인하였다.

(2) LGBM (Light Gradient Boosting Machine)
다른 알고리즘들이 트리를 수평으로 확장하는 것에 반해 트리를 수직으로 확장한다. 이는 left-wise tree growth로 최대 delta loss가 증가하도록 잎의 개수를 정한다. 이로 인해 leaf-wise 알고리즘은 다른 level-wise 알고리즘보다 낮은 loss를 달성할 수 있도록 한다. 하지만 데이터의 크기가 작은 경우 leaf-wise는 train data set에 과적합(overfitting) 되기 쉬우므로 max_depth를 줄여줘야 한다.

Grid search cv, random search cv를 통해 하이퍼 파라미터 튜닝을 진행할 수 있으며 변수 중요도 역시 확인할 수 있다.

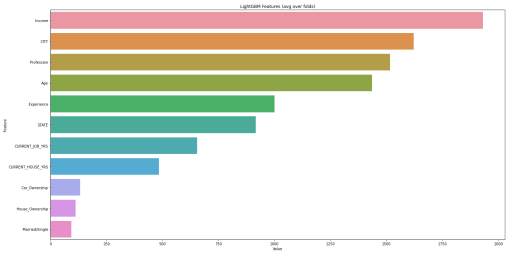

그림 2. Feature Importance

(3) AutoML

AutoML(Automated Machine Learning)은 인공지능을 만들기 위한 별도의 인공지능을 사용하여, 모델링의 전체 또는 일부를 자동화하는 기술이다. AutoML은 모델링 단계 중 일부 또는 모든 단계를 자동화하여 모델의 예측 결과 정확도를 떨어뜨리지 않고, 통계적 지식이나 컴퓨터 프로그래밍 지식이 없는 사람이라도 손쉽게 머신러닝을 이용할 수 있게 한다. AutoML 라이브러리인 pycaret 라이브러리를 사용하여, 자동 전처리, 매개변수 최적화, 모델 선정 과정을 진행했다.

< AutoML 최종 모델 >

model	accuracy	auc	recall	prec.	f1
랜덤포레스트	0.8996	0.9384	0.5365	0.6079	0.5699

2.2.3.2 비정형 데이터 모델링

자소서 데이터(비정형) 데이터에서 성격을 예측하기 위해, 에세이,깃헙,myPersonality 데이터로 train set을 구축하고 자소서로 test set을 구축하여 텍스트 데이터에서 성격을 예측하는 딥러닝 모델을 생성했다. 사용한 모델은 LSTM, GRU, BERT 모델이 있다.

(1) LSTM (Long Short-Term Memory Network)

LSTM은 기존 RNN의 기억 소실 문제를 해결한 것으로 RNN의 hidden state에 cell-state를 추가함으로써 gradient vanishing problem을 해결하고자 하였다.

외향성, 성실성, 민감성을 분류하는 multi-classification을 수행하는 모델 3개를 생성하였고 이의 metric은 accuracy를 사용하였다.

(2) GRU

GRU는 LSTM을 간소화한 모델로, LSTM과 비교해서 학습할 파라미터가 더 적은 것이 장점이다. 전처리한 에세이, myPersonality, github 데이터에 대해 외향성(y/n),성실성(y/n),민감성(y/n)을 분류하는 3가지 이진분류 GRU 모델을 만들었다. 데이터가 균형적이기 때문에, metric은 accuracy를 사용했다.

성격	epoch	accuracy
외향성	5	0.504
성실성	5	0.507
민감성	5	0.516

(3) BERT

전처리한 에세이 데이터에 대해 외향성(y/n), 성실성(y/n), 민감성(y/n)을 각각 분류하는 3개의 이진 분류 모델을 생성했다. 한국어를 다루는 모델이기 때문에 한국어 텍스트로 미리 학습된 'bert-kor-base' 모델을 선택했고, 텍스트 데이터 예측에서 최적의 알고리즘을 찾아주는 AutoGluon의 TextPredictor 모듈을 사용하여 test set을 학습시켰다. y와 n의 비율이 균형적인 데이터이기 때문에 metric은 accuracy를 사용했다.

성격	accuracy	log loss
외향성	0.5313	0.6919

성실성	0.5388	0.6967
민감성	0.5412	0.6897

2.2.3.3 정형데이터 비정형 데이터 매핑

(1) 매핑 목적
자기소개서로부터 예측한 성격을 통해 대출 승인 여부 판단

(2) 매핑 문제점
두 데이터의 공통된 column 부재한다. 즉, 데이터의 출처가 다르기 때문에 (동일한 이용자에 대한 정보가 아니기에) 정확한 매핑은 불가능하다. 원 데이터의 출처가 다른 지금 이 문제를 완벽하게 해결할 수는 없지만 이 문제로 인한 성능 저하를 최대한 줄이는 방법을 찾았다.

(3) 매핑 방법
매핑 방법 - 불필요한 항목 제거 후 매핑하기

- 자기소개서 작성자 모두 직업이 없는 20대 초중반의 Z세대이자 씬파일러로 가정 (*씬파일러: 최근 2~3년간 대출이나 신용카드, 연체 내역 등이 없는 자)
- 금융 데이터에서 Age값(나이)이 20대가 아닌 행 제거
- 제거 후 남은 금융 데이터와 자기소개서 데이터를 임의 매핑

Ⅲ. 결 론

3.1 연구 결과
3.1.1 정형/비정형 데이터 분류 모델 선택

(1)정형 데이터 분류 모델: random forest

모델	accuracy
AutoML	0.8996
(Random Forest)	
LGBM	0.8938
XGBoost	0.8665

(2) 비정형 데이터 분류 모델 : BERT

모델	accuracy [E/C/N]
BERT	[0.5313/0.5388/0.5412]
LSTM	[0.5176/0.5179/0.5214]
GRU	[0.504/0.507/0.516]

3.1.2 정형,비정형 데이터 매핑

id	...	cCON	cEXT	cNEU	Risk_Flag
1		y	n	n	0

3.1.3 대안신용평가모델 구축
(1) 타겟층: 직업이 없는 20대 초중반
(2) 신용평가 과정
Step1 - 사용자로부터 자기소개서를 입력받는다.
Step2 - 자기소개서로부터 외향성, 성실성, 민감성 여부를 예측한다.
Step3 - 외향성, 성실성, 민감성에 따른 대출 승인 여부를 판단한다.
Step4 - 사용자에게 대출 승인 여부를 보여준다.

3.1.4 모델 성능 평가

비정형 데이터에 대해 언어 모델을 생성하기 위해서 Github[1]을 참고하였다. Neural Network의 rnn 계열인 LSTM, GRU와 Transformer인 BERT를 사용하였다.

1) SenticNet, Deep Learning-Based Document Modeling for Personality Detection from Text. 2019.02.22(GitHub - SenticNet/personality-detection: Implementation of a hierarchical CNN based model to detect Big Five personality traits, 2022.12.11)

매핑한 데이터에 대해 LGBM, XGBoost, AutoML로 대출 승인 여부 판단한 결과 이전 정형데이터만으로 모델링 했을 때보다 성능이 낮아졌다.

모델	매핑 전	매핑 후
LGBM	0.9166	0.9150
XGBoost	0.9058	0.9016
AutoML (Random Forest)	0.8595	0.8514

모델별, 매핑 전후 accuracy

하지만 임의로 매핑한 결과이기에서 숫자는 의미가 없고, 차후 성격 데이터와 대출 데이터를 이런 방식으로 결합하여 성능을 예측할 수 있을 것이다.

지도 학습의 문류 문제를 해결하기 위한 tree 알고리즘으로 부터 발전된 LGBM과 XGBoost를 사용하였으며 AutoML로 선택된 randomforest역시 tree기반이다. Randomforest는 여러 tree들을 형성하고 모델의 분산을 줄여주며, bagging의 단점, 모델의 공분산을 해결하기 위해 형성된 알고리즘이다.

3.2 활용 방안

금융 이력이 부족한 Z세대의 신용 향상 서비스를 제공하여 과소평가된 Z세대의 신용을 적절하게 판단할 수 있다.

3.3 제안

매핑된 금융데이터와 비금융 데이터(자소서)를 수집할 수 없어, 임의로 금융 데이터와 비금융 데이터를 결합한 모델을 만들었다. 신용평가사에 데이터를 요청하여 같은 사용자를 대상으로 두 종류의 데이터를 결합하고, 실제 결과를 예측할 수 있다면 결과를 평가하기에 더 적합할 것이다.

또한 자소서 데이터를 auto labeling을 진행할 수 있을 것이다.

「Distribution-Balanced Loss for Multi-Label Classification in Long-Tailed Datasets (2021) 」에서는 auto labeling을 할 때, 클래스 불균형으로 tail 클래스가 80%를 차지함에도 head 클래스에 과적합되어 잘못 라벨링되는 현상을 방지하는 방법을 소개한다. 이 논문에서는 기존 binary cross entropy loss을 개선한 distribution-balanced loss를 사용하는 것이 특징이다. distribution-balanced loss는 꼬리 클래스가 잘 반영되게 하기 위해서 re sampling과 negative tolerant regularization을 사용하여 BCE loss를 개선시켰다.

「Hybrid Feature Selection Technique for Multi-label Text Mining (유인준), 2018」에서 멀티 레이블 텍스트 분류를 위해 진화적 탐색 기법과 새로운 지역적 탐색 기법을 결합하여 효율적인 하이브리드 특징 선별 기법을 제안하였다. 새로운 지역적 탐색 기법을 위해 멀티 레이블 텍스트 데이터의 특성을 고려한 새로운 스코어 함수를 생성하였고, 이를 이용해 진화적 탐색 기법이 유망한 특징을 선별하도록 결합하였다. 7개의 비교 알고리즘과 함께 분류 정확도를 평가한 결과 본 연구에서 제안한 알고리즘이 더 효율적이고 우수한 결과를 도출하였다.

김효원(Hyowon Kim)
성균관대학교 데이터사이언스융합전공
※관심분야: 머신러닝, 자연어 처리

이수정(Soojeong Lee)
성균관대학교 데이터사이언스융합전공/통계학과
※관심분야: 정형&자연어 데이터 분석

이은서(Eunseo Lee)
성균관대학교 데이터사이언스융합전공/통계학과
※관심분야: 신용평가모델, NLP 모델

구분	공통 작업	개인 작업
김효원	데이터 수집,정형 데이터, 비정형 데이터 전처리	에세이 데이터 전처리, BERT, xgboost 모델링, 데이터 매핑
이수정		github 데이터 전처리, 잡코리아 크롤링, LSTM, lgbm 모델링, 데이터 매핑
이은서		기획, myPersonality 전처리, 링커리어 크롤링, GRU, AutoML 모델링

선행연구 비교테이블

	논문명	published in	연구주제	특징	dataset	한계점	시사점	선택된 feature
1	소비자의 성격이 영화 채널 선택 요인과 채널 태도에 미치는 영향	한국콘텐츠학회 논문지 2019, vol.19, no.7, pp. 348-359 (12 pages)	소비자의 성격이 영화 채널 선택 및 태도 파악	-big5 성격을 바탕으로 소비자 설문조사 진행	설문조사 데이터	영화 선택에 미치는 소비자의 특성은 소비자 성격 외, 소비자의 성별, 연령, 소득수준 등이 있으나 이를 고려하지는 못함	실제로 극장과 온라인 채널을 통해 영화 콘텐츠를 제공하는 기업이 소비자를 대상으로 어떤 마케팅 전략을 실행해야 하는지에 대한 방향제시	성격 특성 정의
2	Deep Learning-Based Document Modeling for Personality Detection from Text	IEEE Computer Society	에세이 성격유형 예측	big5성격 유형으로 분류, CNN 모델	에세이별 글쓴이의 성격유형 데이터	정확도 55%	글에서 big5 성격 추출하는 방향 제시	big5 성격 분류 모델
3	Personality Prediction Based on Text Analytics Using Bidirectional Encoder Representations	INTERNATIONAL JOURNAL of FUZZY LOGIC and INTELLIGENT SYSTEMS Vol.21 No.3	트위터의 영어 텍스트를 이용한 성격 예측 시스템	다양한 분류기 사용 (SVM, CNN, BERT)	트위터의 영어 텍스트	딥러닝 모델 성능 향상을 위한 데이터 부족	더 큰 데이터셋을 수집하기 위한 준지도학습 방식 추천	미리 훈련된 BERT 모델

	from Transformers from English Twitter Dataset							
4	Research on automatic labeling of imbalanced texts of customer complaints based on text enhancement and layer-by-layer semantic matching	nature scientific reports	BERT와 word2vec를 이용한 자동 라벨링	라벨을 shallow와 deep으로 나눠 단계적으로 라벨링함	고객 불만-태그 텍스트	데이터 부족 (기업과 협업 필수), 더 발전된 모델 사용 가능	한번에 깊은 단계로 분류하는 것이 아니라, 더 얕은 라벨로 분류하고 깊은 라벨로 분류하여 정확도 높임	bert와 word2vec을 사용한 자동 라벨링
5	SNS 사용 특성이 신용대출 상환에 미치는 영향에 관한 연구	인문사회 21	SNS 사용 특성과 신용점수의 상관관계	이항로지스틱모형 (binary logistic regresiion)사용	T사 A대출 프로그램 대출이용자 756명의 프로필 사진, SNS 사용량(총 게시글 수)	이미지 분석이 아닌 자의적 분류에 의한 7가지 유형,	SNS 데이터를 통해 성격을 분류하고 성격과 신용을 연결함	성격 특성이 신용에 미치는 영향
6	SNS 행동 데이터가 신용평가에 미치는 영향	이지형	SNS 데이터를 이용한 개인신용평가모델 개발	로지스틱 회귀분석 사용, 검증지표 PSI, CAR 사용	카카오페이 모바일 통합 데이터	롤레이트 분석/빈티지분석 불가	핀테크사의 데이터만으로 평가항목 구성, 핀테크 스코어 안정성, 변별력 검증	핀테크 스코어