

'CTR 예측 시스템'을 활용한 '타행 여신 기업 유치 모델'들 의 비교 분석

팀명: SYM

팀장: 문상혁

팀원1: 강훈

팀원2: 황창조

2021 금융 데이터 경진대회 결과 제출 양식

□ 요약

금리 인상 정책의 시국 속 기업들의 연신 은행 변동의 가능성이 생기면서 타행 여신 기업 고객을 대상으로 하는 여신 사업 유치의 중요성이 높아지고 있다. 필연적으로 높은 유치 성공률과 효율성을 위하여 머신 러닝 모델의 도입이 각광받고 있다. 그 중 'CTR 예측 시스템'은 방대하고 다양한 형태의 데이터를 바탕으로 이진 분류/예측을 진행한다는 점에서 '타행 여신 기업 유치 모델'에 활용될 수 있으며 검증된 선행 연구들이라는 점에서 모델 탐색의 시간과 비용을 절약할 수 있다. 본 프로젝트는 널리 활용되는 'CTR 예측 시스템'들을 '타 은행 여신 고객 기업 대상 여신 사업 유치 성공 여부와 기업정보' 데이터에 적용, 성능을 비교 분석해 보았다. 데이터의 불균형을 'SMOTE NC' 샘플링으로 해소한 후, 'Logistic Regression, Decision Tree, Random Forest, Multi Layer Perceptron & bagging, Factorization Machine, Field-aware Factorization Machine, XGBoost' 총 7가지의 CTR 예측 알고리즘들을 비교 분석해 본 결과, 성능지표 'G-mean'을 기준으로 'Random Forest Classifier'가 0.844531의 값으로 가장 높은 성능을 보임을 알 수 있었다.

1. 주제

'CTR 예측 시스템'을 활용한 '타행 여신 기업 유치 모델'들의 비교 분석

'CTR (Click Through Rate) 예측 시스템'은 광고의 클릭율을 예측하기 위해 고안된 시스템들을 지칭하는 용어로 광고 클릭 여부 데이터를 바탕으로 특정 유저가 '광고를 클릭할 확률' 혹은 '클릭 여부'를 도출하는 것을 목적으로 한다. '타행 여신 기업 유치 모델'은 타 은행의 여신 상품을 사용중인 기업을 대상으로 특정 은행이 자사의 여신 상품 유치를 성공시킬 수 있는지 여부(또는 확률)를 도출하기 위한 머신 러닝 모델을 의미한다.

본 프로젝트에서는 '타행 여신 기업'들을 '광고 고객'에, '유치 성공 여부'를 '클릭 여부'에 대입해 'CTR 예측 시스템'에서 활용되는 머신 러닝 모델들을 '타행 여신 기업 유치 모델'이라는 관점에서 비교 분석하여 최고의 성능을 보여주는 분류/예측 모델을 찾고자 한다.

2. 배경 및 필요성

국내 은행의 수입 약 80%는 이자수익에서 나온다. 특히나 기업을 대상으로 하는 여신사업의 이

자수익 비중은 막대하다 할 수 있다. 그러나 기업 대상 여신시장은 한정되어 있기에 타 은행의 기업 고객 대상 여신 사업 유치가 필수 불가결하다. 특히나 코로나 시국의 영향인 제로금리를 견제하기 위한 금리 인상 정책이 진행됨에 따라 은행 별 여신 사업 상품에 차별화가 생기고 기업들의 여신 상품 선택에 변동이 생길 가능성이 있는 현재, 타행 여신 기업 대상 여신 사업 유치에 신경을 써야할 필요가 있다. 4차 산업혁명의 시대 속에서 이러한 여신 사업의 유치, 자세하게는 여신 사업 유치 기업 선정 심사에 관련된 부분은 머신 러닝 모델에 의한 자동화가 이루어지는 추세이다. 이는 여신 심사를 위해 고려해야하는 데이터가 매우 방대하고 다양하기에 더 높은 성공확률과 그에 따른 수익, 기업 탐색에 드는 시간과 비용 절감을 위해 필연적인 부분이다.

그렇다면 어떤 머신 러닝 모델을 사용해야 하는가? 'CTR (Click Through Rate) 예측 시스템' 은 광고의 클릭율을 예측하기 위해 고안된 시스템들을 지칭하는 용어로 광고 클릭 여부 데이터를 바탕으로 특정 유저가 '광고를 클릭할 확률' 혹은 '클릭 여부'를 도출하는 것을 목적으로 한다. '타행 여신 기업'들을 '광고 고객'에, '유치 성공 여부'를 '클릭 여부'에 대입해 보면 'CTR 예측 알고리즘'들이 추구하는 방향성이 '타행 여신 기업 유치 모델'과 부합하는 것을 확인할 수 있다. 'CTR 예측'이 가장 트렌디한 binary classification 연구의 일종이며, 여신 심사 데이터와 같은 방대하고 다양한 형태의 데이터를 다루는 분야라는 점을 생각한다면, 우리는 'CTR 예측'에서 검증된 머신 러닝 모델들로 실험해봐야 할 모델들을 한정할 수 있을 것이다.

따라서 본 프로젝트는 'CTR 예측'에서 널리 활용되는 머신 러닝 모델들의 적용 및 성능 비교 분석을 통해 은행의 타행 여신 기업 대상 여신 사업 유치 성공률을 높여주고 기업 탐색과 모델 탐색에 필요한 시간과 비용 절약을 도모하고자 한다.

3. 아이디어 제안 및 분석 결과

3.1 서론

'타행 여신 기업 고객 대상 여신 사업 유치 모델'에 적합한 머신 러닝 모델의 탐색을 위한 시간과 비용을 줄이기 위하여 본 프로젝트에서는 검증된 연구 분야인 'CTR 예측 시스템'에서 널리 활용되는 머신 러닝 알고리즘들만을 탐색해볼 예정이다. 데이터의 불균형을 샘플링으로 해소한 후, 'Logistic Regression, Decision Tree, Random Forest, Multi Layer Perceptron & bagging, Factorization Machine, Field-aware Factorization Machine, XGBoost' 총 7가지의 CTR 예측 알고리즘들을 비교 분석하여 가장 성능이 좋은 모델을 도출하고자 한다.

3.2 데이터 구성 및 전처리

3.2.1 데이터 구성

본 프로젝트는 우리은행에서 제공되는 '타 은행 여신 고객 기업 대상 여신 사업 유치 성공 여부와 기업정보' 데이터를 활용한다. 3번째 열부터 94번째 열까지의 92개 열은 '기업별 여신, 수신, 이체, 업체, 재무, 입금, 채널 정보'들로 이루어져 있으며 설명변수로 활용한다. 92개의 기업 정보 데이터 열은 숫자 1 또는 2로 구성된 8개의 nominal 데이터 열과 continuous 데이터를 1~4 사이의 숫자로 범주화 한 84개의 ordinal 데이터 열로 이루어져 있다. 두번째 열의 'TARGET'이라 명칭 되어 있는 '기업별 여신 사업 유치 성공 여부'는 종속변수로 활용한다. 이는 '실패=0', '성공=1'로 구성된 이진 데이터이며 '0'에 해당하는 기업의 수가 118720개, '1'에 해당하는 기업의 수가 '815'개로 상당히 불균형한 데이터이다. 'Fig.1'은 종속변수 'TARGET'의 불균형을 보여주는 그래프이다.

3.2.2 데이터 전처리

-통합 전처리: 데이터에 결측치는 없었으며, 숫자 1, 2로 이루어진 8개의 nominal 데이터 열을 숫자 0, 1로 이루어진 binary 데이터 열로 전환했다.

-Train set, Test set 구성: 지도 학습 모델에 사용되는 데이터는 크게 학습을 위한 Training set과 학습된 모델의 평가를 위한 Test set으로 구성된다. 본 프로젝트에서는 119534개의 기업 데이터 중 70%를 Training set, 30%를 Test set으로 구성하였다.

-Sampling: 종속변수 내 두 집단의 비율 차이가 아주 큰 경우에 작은 집단은 큰 집단으로 오분류 되는 경우가 많은 반면, 전반적인 오분류율은 작으므로 분류 성능은 좋게 나타나는 경우가 대부분이다. 하지만, 이런 불균형데이터의 분류는 작은 집단을 큰 집단으로 잘못 분류하면 그 반대 상황보다 훨씬 더 큰 손실(loss)을 가져온다. 따라서 두 집단의 비율을 비슷하게 맞추는 sampling을 적용한 후에 분류 방법론을 적용해야 한다. Sampling 방법 중 작은 집단을 큰 집단의 크기에 맞추어 반복 추출하는 방법을 'Over-sampling' 이라 한다. 'SMOTE'는 이 중 Over-sampling에 해당하는 방식으로 아래와 같은 방식으로 작동한다.

- 기준 Sample(소수 집단 내 임의의 한 sample)과 거리(유클리드 거리)가 가까운 k 개의 Sample(KNN)을 찾는다. 이 k 개의 Sample 중 랜덤하게 1 개의 Sample 을 선택한다. 이 Sample 을 KNN Sample 이라 명명한다.
- 새로운 Synthetic Sample 은 아래와 같이 계산한다.
- $X_{new} = X_i + (X_k - X_i) * \delta$

'SMOTE NC'는 'nominal'과 'continuous'가 혼재되어 있는 데이터를 사용하는 경우에 쓸 수 있는 'SMOTE' 방식의 Over-sampling 기법이다. 본 프로젝트에서 사용하는 데이터가 ordinal 데이터와 nominal 데이터가 혼재되어 있다는 점, ordinal 데이터의 원본이 continuous 데이터라는 점을 고려하여 데이터의 순서관계와 원본 데이터의 연속성을 반영하기 위해 'SMOTE NC' 샘플링 기법을 사용하였다. 'Fig.2'는 'SMOTE NC' 샘플링 기법을 적용한 후 데이터의 종속변수에 대한 불균형 해소를 보여주는 그래프이다.

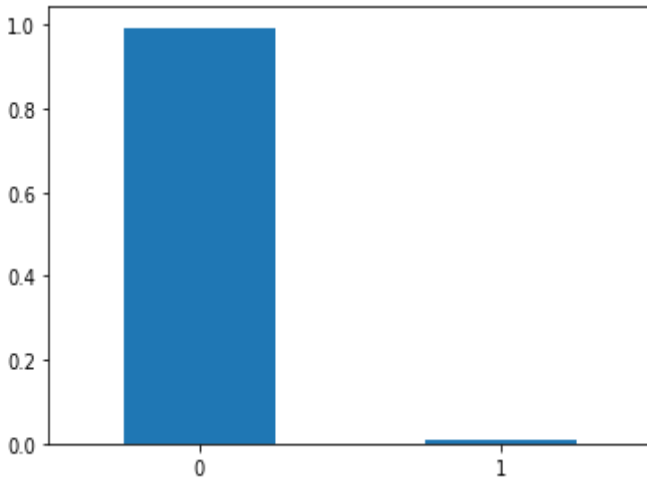


Fig.1 'TARGET'의 불균형

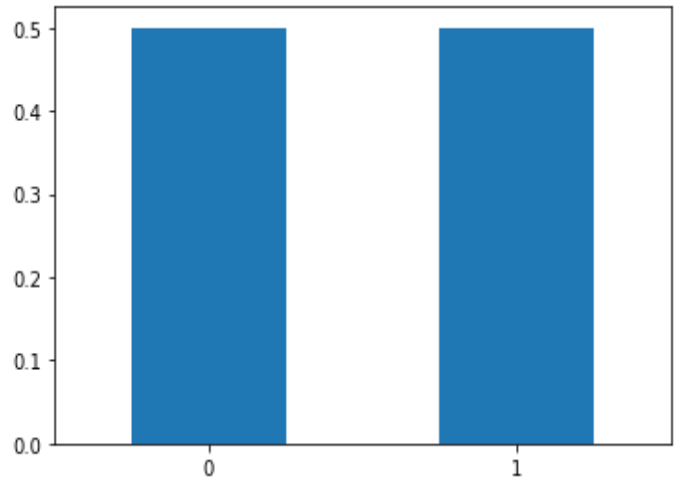


Fig.2 'SMOTE NC' 후 'TARGET'의 불균형 해소

3.3 분류/예측 모델 설계

3.3.1 Logistic Regression

Logistic Regression은 선형 회귀 방식을 분류에 적용한 알고리즘으로 가장 많이 쓰이는 방법론 중에 하나이다. 각각의 클래스의 사후확률 odds비가 설명변수 X 와 선형관계를 가진다고 가정하며, 아래와 같은 모형 하에서 회귀계수에 대한 추정치는 IRLS(Iterative Reweighted Least Squared)를 이용하여 최대 우도 추정량(MLE)을 일반적으로 사용한다.

$$\log \frac{P(G=k | X=x)}{P(G=K | X=x)} = \beta_{k0} + \beta_k^t x, \quad k = 1, \dots, K-1$$

3.3.2 Decision Tree

Decision Tree는 머신러닝 알고리즘 중 직관적으로 이해하기 쉬운 알고리즘으로 의사결정규칙 (decision rule)을 도표화하여 분류(Classification) 또는 예측(Regression)을 수행할 수 있다. 정보의 '불순도'를 기반으로 부모 노드(Parent node)에서 자식 노드(Children node)로 분기가 일어나며 어떤 규칙을 통해 분기가 일어났는지 쉽게 알 수 있고 시각화 또한 쉽게 가능하다. Hyper parameter로는 최

대 깊이 4, 분기가 일어나는 노드의 최소 샘플의 수 50, 리프노드(leaf node)의 최소 샘플의 수를 25로 설정하였다.

3.3.3 Random Forest

Random Forest는 대표적인 배깅(bagging) 알고리즘으로 여러 개의 Decision Tree 분류기가 전체 데이터에서 배깅 방식으로 각자의 데이터를 샘플링해 개별적으로 학습을 수행한 뒤 최종적으로 모든 분류기가 보팅(Voting)을 통해 예측 결정을 하게 된다. Random Forest는 이런 임의의 설명변수의 개수(m)과 bootstrap의 샘플 수와 같은 tuning parameter가 있지만 별도의 tuning없이도 상당히 좋은 결과를 보여준다. Decision Tree 분류기의 개수를 20개, 최대 깊이를 5로 설정하고 분석을 진행하였다.

3.3.4 Multi Layer Perceptron (MLP)

MLP는 인공 신경망 구조 중 하나로서 초기 가중치를 임의의 값으로 정의하고 예측 값의 활성화 함수 반환 값과 실제 결과 값의 활성화 함수 반환 값이 동일하게 나올 때까지 가중치의 값을 계속 수정하는 perceptron을 여러 층으로 구성한 알고리즘이다. 이를 구현하기 위해 'scikit learn'에서 제공하는 'MLPClassifier' 패키지를 사용했다. MLP의 구조는 학습을 진행하기 위한 10개의 node를 가지는 hidden layer 2층으로 구성하였으며, 활성화 함수는 'logistic', 최적화 함수는 'sgd(stochastic gradient descent)'를 적용했다. 가중치 규제에 대한 'L2 penalty'에 해당하는 'alpha'는 0.01, 기울기 학습율을 조정하는 'Learning rate'는 0.1를 적용하였다. 학습 방법을 결정하기 위한 'batch_size'와 'max_size'는 각각 32와 500으로 설정하였다.

3.3.5 Mlp with bagging

Bagging이란 bootstrap aggregating의 줄임말로 통계적 분류와 회귀 분석에서 사용되는 기계 학습 알고리즘의 안정성과 정확도를 향상시키기 위해 고안된 일종의 앙상블 학습법의 메타 알고리즘이다. MLP의 경우 weight의 초기값에 따라 예측 결과의 변동성이 크기 때문에 변동성을 줄이기 위하여 bagging 알고리즘을 사용하였다. 이를 구현하기 위해 'scikit learn'에서 제공하는 'BaggingClassifier' 모델을 사용하였다. Base_estimator의 경우 상기의 MLPClassifier를 사용했으며 동일한 파라미터를 적용하였다. base estimators의 개수를 의미하는 'n_estimators'의 경우 500으로 설정하였고 이는 2000으로 늘렸을 때도 동일한 결과를 보여주었다. 'max_features'의 경우 X의 학습시킬 최대 features의 개수를 의미하는데 우리 데이터의 경우 target을 제외한 feature가 총 92개이므로 100개로 설정해주었다. 이 외의 파라미터는 모두 default값을 사용하였다.

3.3.6 Factorization Machine (FM)

FM 모델은 설명 변수간 상호작용을 고려하는 모델로 factorized 상호작용을 이용하여 피쳐 벡터 x 의 값 사이에 있는 가능한 상호작용들을 모델화 하는 알고리즘이다. FM 모델의 수식은 아래와 같다.

$$\hat{y} = w_0 + \sum_{i=1}^n w_i x_i + \sum_{i=1}^n \sum_{j=i+1}^n (\vec{v}_i \cdot \vec{v}_j) x_i x_j$$
$$\hat{y} = w_0 + \sum_{i=1}^n w_i x_i + \sum_{i=1}^n \sum_{j=i+1}^n (v_i \cdot v_j) x_i x_j$$

이를 구현하기 위해 'xlearn'에서 제공하는 'fm' 패키지를 사용했다. 이 패키지에서 입력 받을 수 있는 형식으로 데이터를 전환하기 위하여 'DMatrix Transition'이 적용됐다. 기울기 학습율을 조정하는 'Learning rate (lr)'는 0.03, 'regularization'을 위한 'lamda'는 0.0001이 사용됐다. 쓰이는 데이터가 불균형 데이터인 점을 고려하여 'evaluation metric'을 나타내는 'metric'은 auc를 사용했다.

3.3.7 Field-aware Factorization Machine(FFM)

FM의 개량된 버전인 Field-aware FM 모델은 User, Item, Tag와 같이 3개의 Latent Vector를 사용하고 두개의 Latent Vector를 사용하는 FM 모델에 비하여 뛰어난 성능을 보여주었다. 3개의 데이터를 input으로 받기 때문에 csv파일은 직접 적용이 불가능하고 LIBSVM의 데이터 포맷을 사용하게 된다. 따라서 우리 코드는 먼저 csv를 LIBSVM파일로 변환하는 과정을 거치고 이후 학습을 진행하게 된다. 이를 구현하기 위해 'xlearn'에서 제공하는 'ffm_model' 모델을 사용하였다. 'task'의 경우 binary classification으로 사용하였고 학습에 큰 영향을 주는 파라미터인 'learning rate', 'regular lambda', 'k'는 Factorization Machine에 대한 기존 연구들을 참고, 파라미터 후보군을 만든 뒤 우리 데이터에서 가장 높은 성능을 보여주는 파라미터를 적용하였다. 'evaluation metric'의 경우 우리 데이터가 불균형 데이터이기 때문에 accuracy가 높은 것이 큰 의미가 없다고 판단 AUC를 평가 지표로 사용하였다. 최적화 함수는 'sgd'를 사용하였다.

3.3.8 XGBoost(eXtra Gradient Boost)

XGBoost는 트리 기반의 앙상블 학습에서 가장 각광받고 있는 알고리즘이다. XGBoost는 GBM(Gradient Boosting Machine)에 기반하고 있지만 GBM의 단점인 느린 수행 시간 및 과적합 규제(Regularization) 부재 등의 문제를 해결해서 매우 각광을 받고 있다. 일반적인 분류와 회귀 영역에서 뛰어난 예측 성능을 발휘하며 자체적으로 교차검증, 가지치기(pruning), 결측치 처리 등의 기능을 가지고 있다. 최대 깊이를 8, 학습률을 0.1, 손실함수는 이진분류일 경우 적용하는 logistic, 검증에 사용되는 함수는 logloss를 사용하였다.

3.4 분석 결과 및 성능 평가

3.4.1 성능 평가 지표

분류방법론에서는 적합한 모형이 분류를 얼마나 정확하게 했는지를 보기 위해서 여러가지 성능 평가 지표를 사용한다. 본 프로젝트는 타겟 데이터의 불균형 문제를 고려하여 모델 성능 평가를 위해 다음과 같은 지표들을 고려한다.

	Predicted 0	Predicted 1
True 0	TN(True Negative)	FP(False Positive)
True 1	FN(False Negative)	TP(True Positive)

Table. 1 오차행렬(Confusion Matrix)

정확도(Accuracy) = $\frac{TP+TN}{TP+TN+FP+FN}$

정밀도(Precision) = $\frac{TP}{TP+FP}$

재현율(Recall) = $\frac{TP}{TP+FN}$

F1 Score = $2 * \frac{Precision*Recall}{Precision+Recall}$

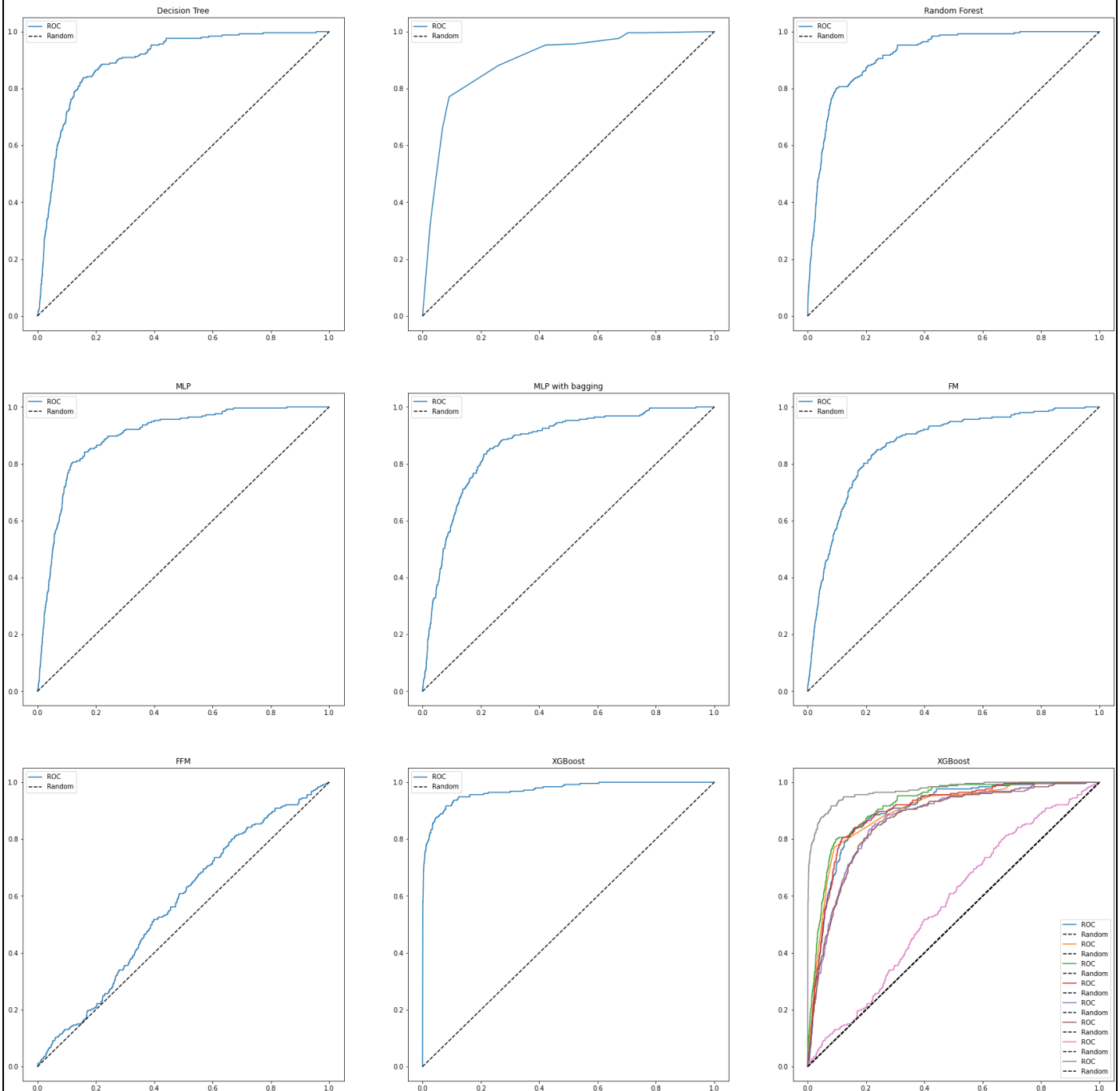
오차행렬을 통해 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 Score의 값을 도출할 수 있으며 이들은 이진분류 성능 평가 지표로써 자주 활용되는 지표들이다. 정확도는 실제 데이터에서 예측 데이터가 얼마나 같은지를 판단하는 지표이다. 직관적으로 모델 예측 성능을 평가할 수 있는 지표이지만 불균형한 데이터 세트에서 Positive에 대한 예측 정확도를 판단하지 못한 채 Negative에 대한 예측 정확도만으로도 분류의 정확도가 매우 높게 나타나는 수치적인 판단 오류를 일으키게 되므로 정확도 외에 지표들을 복합적으로 평가할 필요가 있다. 정밀도는 예측을 Positive로 한 대상 중에 예측과 실제 값이 Positive로 일치한 데이터의 비율을 뜻하고 재현율은 실제 값이 Positive인 대상 중에 예측과 실제 값이 Positive로 일치한 데이터의 비율을 뜻한다. F1 Score는 정밀도와 재현율을 결합한 지표로 정밀도와 재현율이 어느 한쪽으로 치우치지 않는 수치를 나타낼 때 상대적으로 높은 값을 가진다.

이외에 FPR(False Positive Rate)의 변화에 따른 TPR(True Positive Rate)의 변화를 나타내는 곡선인 ROC 곡선(Receiver Operation Characteristic Curve)과 이에 기반한 AUC(Area Under Curve) 스코어, TPR와 TNR(True Negative Rate)의 기하평균인 G-mean을 모델 성능 평가 및 비교 지표로 활용할 예정이다. 그중 G-mean은 불균형 데이터 세트에서 모형의 성능을 평가하는 데 많이 이용되어 왔기 때문에 가장 중요한 평가 지표로 삼는다.

G- mean = $\sqrt{TPR * TNR}$

3.4.2 모델 성능 평가 및 비교 분석

Fig.3 ROC curve



Class		0	1	Accuracy	AUC score	G-mean
Logistic Regression	Precision	0.998519	0.037963	0.851760	0.896220	0.836919
	Recall	0.851971	0.822134			
	F1 Score	0.919442	0.072575			
Decision Tree	Precision	0.998861	0.023434	0.740023	0.824539	0.792311
	Recall	0.739019	0.881423			
	F1 Score	0.849515	0.045655			

Random Forest	Precision	0.998447	0.047277	0.883996	0.917420	0.844531
	Recall	0.884548	0.806324			
	F1 Score	0.938053	0.089317			
MLP	Precision	0.998284	0.027926	0.800619	0.865908	0.803446
	Recall	0.800579	0.806324			
	F1 Score	0.888567	0.053983			
FM	Precision	0.998022	0.029853	0.821672	0.861932	0.795979
	Recall	0.822034	0.770751			
	F1 Score	0.90152	0.05748			
FFM	Precision	0.998220	0.028001	0.803045	0.571881	0.800745
	Recall	0.803078	0.798419			
	F1 Score	0.890079	0.054105			
XG Boost	Precision	0.997222	0.684444	0.995259	0.971162	0.779411
	Recall	0.998006	0.608696			
	F1 Score	0.997614	0.644351			

Table.2 성능지표

Logistic Regression은 Class 1에 대한 재현율이 0.8221로 비교적 높은 수치를 보여주고 있으나 Class 1에 대한 정밀도는 0.0379의 수치를 보여주었다. Class 1에 대한 정밀도가 재현율에 비해 현저히 떨어지는 수치를 보여 F1 Score가 0.0725로 낮지만 정확도와 AUC Score, G-mean은 타 모델들에 비교적 높은 수치들을 보여준다. Logistic Regression의 G-mean은 다른 모델들의 값들과 비교했을 때 0.8369로 두번째로 높은 수치이다.

Decision Tree는 Class 1에 대한 재현율이 다른 모델들과 비교했을 때 0.8814로 가장 높으나 그와 동시에 0.0234의 가장 낮은 정밀도를 보여준다. 이로 인하여 Class 1에 대한 F1 Score 또한 0.0456으로 가장 낮다. 정확도는 0.74로 전반적인 예측 성능은 가장 낮다고 할 수 있다.

Random Forest는 앞선 두 모델들과 비슷하게 Class 1에 대한 높은 재현율, 낮은 정밀도를 보여준다. 정확도와 AUC Score는 각각 0.8839, 0.9174로 다른 모델들과 비교상 두번째로 큰 수치이다. G-mean은 가장 높은 수치인 0.8445를 보여준다.

MLP와 FM은 여타 대부분의 모델들과 비슷하게 Class 1에 대한 재현율이 정밀도에 비해 현격히 높으며 전체적인 수치상 준수한 성능을 보여준다. FFM 또한 Class 1에 대한 재현율과 정밀도가 비슷한 양상을 띄고 있으나 AUC Score가 0.5718로 가장 낮은 수치를 보여준다.

XG Boost는 앞선 모델들과는 다르게 Class 1에 대한 정밀도가 0.6844, 재현율이 0.6086으로 수치상 큰 차이가 없다. 이로 인해 F1 Score 값이 0.6443으로 가장 높으며 정확도와 AUC Score값이 0.9952, 0.9711로 전반적인 성능은 가장 좋다고 할 수 있다. 하지만 우리 주제에서 가장 유심하게 봐야 할 수치들인 Class 1에 대한 재현율과 G-mean 에는 가장 낮은 값을 보여준다.

3.5 결론

타행 여신 분류 예측에 있어서 유치 성공에 대한 데이터가 희소한 측면이 있어서 불균형 문제가 있다. 이러한 문제를 해결하기 위해 SMOTE-NC 알고리즘을 활용해 오버샘플링을 진행한 후 분석을 진행하였고 CTR예측에 활용되는 다양한 머신러닝 알고리즘들을 비교 분석하였다. 불균형 데이터에서 모델을 평가하는데 가장 중요하게 활용되는 G-mean을 최우선 순위에 두고 다른 지표들까지 종합적으로 비교한 결과 Random Forest Classifier를 최종 모델로 선정하였다. Random Forest는 다른 모델들에 비해 가장 높은 G-mean 값을 보여주었고 그 다음으로 Logistic Regression, MLP, FFM, FM, Decision Tree, XGBoost 순으로 좋은 수치를 보여주었다. 정확도는 XGBoost, Random Forest, Logistic Regression, FM, FFM, MLP, Decision Tree 순으로 높고 AUC Score는 XGBoost, Random Forest, Logistic Regression, MLP, FM, Decision Tree, FFM 순이다. XGBoost가 정확도와 AUC Score값에서 Random Forest에 비해 더 좋은 수치를 보여주나 이 프로젝트에서 가장 중요하게 고려하는 G-mean값이 가장 낮으므로 최종 모델로 선정하지 않았다.

4. 기대효과

도출된 최고 성능 모델을 도입함으로써 얻을 수 있는 기대효과는 크게 3가지이다. 첫째, 모델 탐색에 드는 시간과 비용을 절약할 수 있다. 본 프로젝트는 'CTR 예측'에서 검증된 모델들을 대상으로 '타행 여신 기업 유치 모델'들을 비교 분석해서 최고 성능의 모델을 도출하였기에 최적의 모델이라 할 수 있다. 둘째, 높은 성능의 모델을 기반으로 유치 가능성을 예측함으로써 높은 유치 성공률을 기대해 볼 수 있다. 성능지표 'G-mean'은 민감도와 특이도의 기하평균으로 종속변수의 소수집단에 대한 예측 성공율에 민감하다. 최고 성능 모델로 도출된 'Random Forest Classifier'는 'G-mean' 값이 '0.844531'으로 나왔기에 유치 성공에 대한 예측율이 높다고 할 수 있다. 따라서 유치 성공률이 높을 것이라 기대할 수 있다. 마지막으로 기업 탐색에 드는 시간과 비용을 절약할 수 있다. 최고 성능의 "모델 적용을 통해 각 기업별 여신 사업 유치 성공확률 혹은 성공 여부가 도출되기에 성공으로 예측되는 혹은 성공 확률이 높은 기업에만 여신사업의 마케팅과 시간, 비용 투자를 한정할 수 있다.

5. 활용 데이터

2021금융데이터경진대회에서 제공하는 타행여신유치모델 개발을 위한 데이터

데이터설명 : 타행여신유치모델 개발을 위한 다양한 변수와 값(레이블) 제공

데이터 카테고리 : 당기자산총금액, 부채총금액, 차입총금액 당기매출금액, 매출원가금액,

금융비용금액, 부채비율, 수신계좌잔액, 최근6개월전체수신평균 등 많은 변수와 값 제공

데이터 기간 : 2019년 1년치

6. 참고자료

-Comparison of resampling methods for dealing with imbalanced data in binary classification problem (Geun U Park, Inkyung Junga, Division of Biostatistics, Department of Biomedical Systems Informatics, Yonsei University College of Medicine)

-Classification Analysis for Unbalanced Data (Dongah Kima, Suyeon Kanga, Jongwoo Songa, Department of Statistics, Ewha Womans University)

-파이썬 머신러닝 완벽 가이드 (지은이:박찬규, 위키북스, 발행일 2020.12.03)

-분류모형을 이용한 여신회사 고객대출 분석에 관한 연구.한국데이터정보과학회지(김태형, 김영화, 2013, 24(3),411-425).


-Context-Aware Ad Contents Scheduling over DOOH Networks based on Factorization Machine (Van Hoang Nguyen, Thanh Binh Nguyen, Sun-Tae Chung, 2019, 멀티미디어학회논문지,22(4),515-526.)

-Factorization Machine을 이용한 추천 시스템 설계 (정승윤, 김형중, 2017, 한국디지털콘텐츠학회 논문지, 18(4), 707-712)

※ 작성 시 유의사항

- 분량 제한 없음 /글자 폰트 크기 11 포인트(한글 및 워드로 작성)
- 도표, 이미지 등 활용 가능
- 설명을 위한 추가자료 첨부 가능
- 제출 시 표지를 함께 제출하되, 식별이 가능하도록 참가자 명(팀의 경우 팀장 및 팀원명)을 작성하여 제출

2021 년 09월 06일

참가자(대표자) (인  서명)

금융보안원 귀중