Why DataVictors?

Final 0.72(179위)

DataVictors(한지민, 김용휘, 차유진)

전략

- 도메인 이해
- EDA & 전처리
- 모델 파악
- 경험적 접근
- 통계적 검증
- 기존 실험 결과
- 보완 계획

도메인 이해

- MQL(Marketing Qualified Lead) 선행 연구 자료 탐색

Financing lead triggers | Proceedings of the 19th ACM SIGKDD international conference on ...

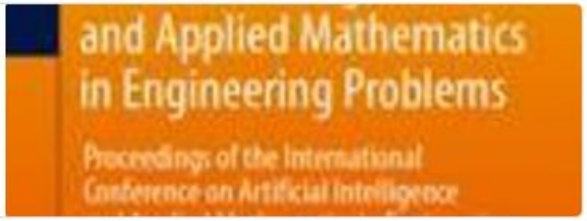
You will be notified whenever a record that you have chosen has been cited.

https://dl.acm.org/doi/abs/10.1145/2487575.2488190?source=post_page-----28e635bd5b1----------...

Prediction of Potential Bank Customers: Application on Data Mining

Banking is an important industry, where financial transactions are performed to meet our needs in our everyday lives. Today, banks are frequently used to meet all kinds of financial transactions. In line with the

https://link.springer.com/chapter/10.1007/978-3-030-36178-5_9?source=post_page-----28e635bd5b1----...



The state of lead scoring models and their impact on sales performance

Information Technology and Management - Although lead scoring is an essential component of lead management, there is a lack of a comprehensive literature review and a classification framework...

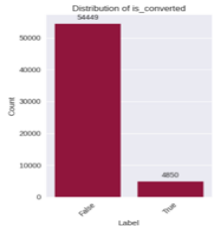
https://link.springer.com/article/10.1007/s10799-023-00388-w?source=post_page-----28e635bd5b1----...

EDA 및 전처리

- 효율적인 데이터 탐색을 위해 Kaggle, Dacon 참조

```
[expected_timeline] (450)
less than 3 months
nan
3 months ~ 6 months
9 months ~ 1 year
more than a year
6 months ~ 9 months
```

범주형 데이터 카테고리 출력 함수



데이터 구성 bar plot 시각화 출력 함수

- 타겟 변수(is_converted) 시각화로 data imbalance 발견
- 과정
  - (EDA) 상관분석 > 시각화 > 카테고리 확인
  - (전처리) 결측치 처리 > feature selection > 레이블 인코딩

모델 파악

- 의사결정 모델 특성
  - 숫자 데이터의 크기와 무관하게 분류됨
  - 데이터 불균형의 영향을 크게 받음
- 불균형 해소에 집중
  - SMOTE 기반의 oversampling 시도
  - undersampling 시도
- 전처리 이전의 데이터 활용 + undersampling -> 점수 향상

경험적 접근

- 목표는 **0.52** 달성 > 상위 **30%** 이내 달성 > **0.01점**이라도 올리기
- 0.52를 달성하기 위해 기존 skeleton code에 층만 깊게 쌓음 > **성공!**

- 선행 연구에서 공통적으로 가장 우수한 성능을 냈던 Random forest
- 가장 우수한 점수를 냈던 label encoding + 결측치 fillna(0)
- Random forest + 위 전처리 방식 + undersampling -> F1 0.66 **달성** (당시 상위 30% 이내)

- 추가적인 점수 향상을 위해 Optuna, Best model & Ensemble 활용하여 튜닝 -> 0.01점이라도 올리기 **성공!**

통계적 검증

- 목표 재설정(결승 진출)
- Impact 있는 점수 향상을 위한 선택 > 딥러닝 모델 활용
- 전처리
  - GPT로 범주형 요약> 원핫 인코딩 > 차원 축소 > 통계적 유의성 검정

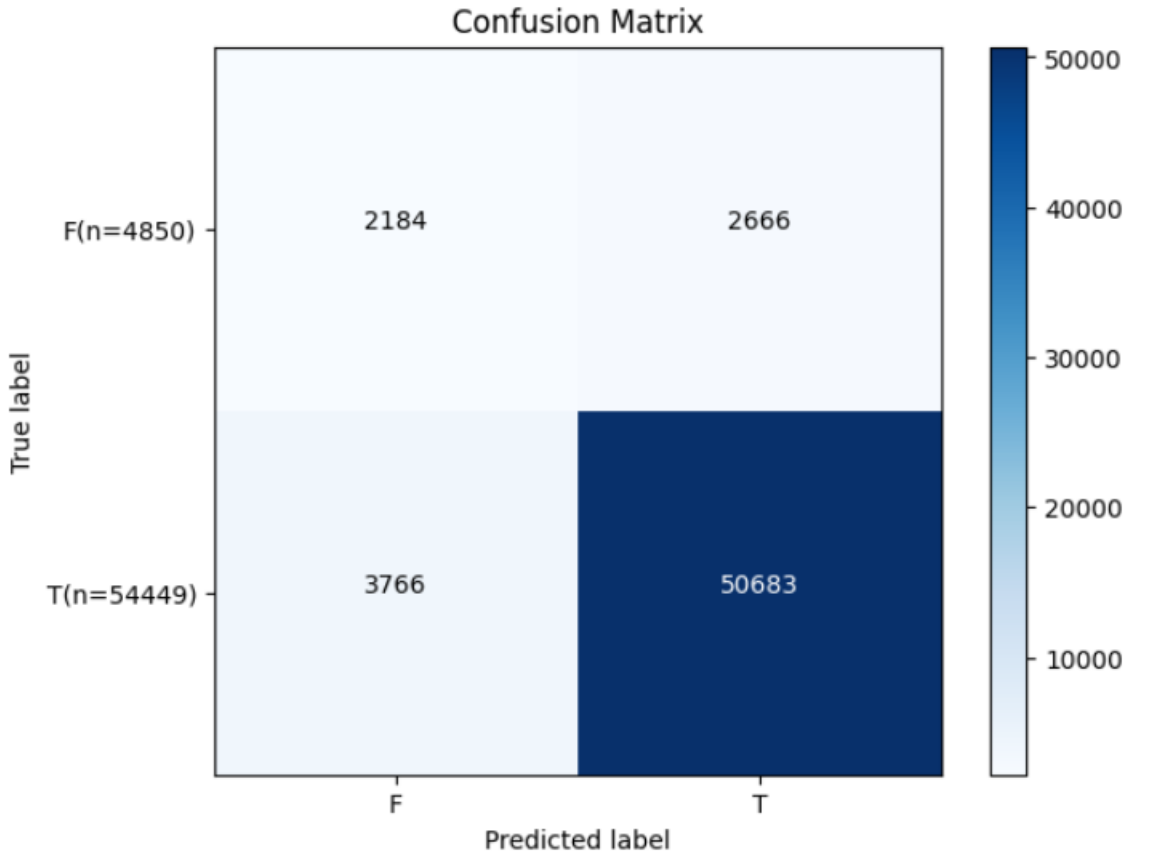
	enterprise	business_unit	product_category	inquiry_type	customer_type	business_area	customer_continent	customer_job
상관 분석	약한 상관관계	-	-	-	-	-	-	-
ANOVA	유의미	-	유의미	-	유의미	유의미	-	유의미

- Cost sensitive training
  - oversampling의 한계 > auto encoder, knn 모두 경계 불명확
  - 소수 데이터에 가중치를 크게 매기는 방식(penalized model)
- 모델 선정 등 레퍼런스 찾기

기존 실험 결과

- DeepOD(Anomaly Detection 모델) 라이브러리 활용

```
f1_0221_RoSAS() = 0.33
accuracy: 0.8915
precision: 0.3671
recall: 0.4503
F1: 0.4044
public f1-score = 0.36
```



보완 계획

모델 선정 → 전처리 → 유의성 검정 → 모델 튜닝 → 문제 해결