DBpia

XGBoost와 LightGBM을 이용한 안전 운전자 예측 성능 비교

Comparison of Safety Driver Prediction Performance with XGBoost and LightGBM

저자 (Authors)
장승일, 곽근창
Seung-Il Jang, Keun-Chang Kwak

출처 (Source)
Proceedings of KIIT Conference , 2019.6, 360-362(3 pages)

발행처 (Publisher)
한국정보기술학회
Korean Institute of Information Technology

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08750046

APA Style
장승일, 곽근창 (2019). XGBoost와 LightGBM을 이용한 안전 운전자 예측 성능 비교. Proceedings of KIIT Conference, 360-362

이용정보 (Accessed)
한신대학교
211.187.169.***
2020/01/11 19:29 (KST)

저작권 안내
DBpia에서 제공되는 모든 저작물의 저작권은 원저작자에게 있으며, 누리미디어는 각 저작물의 내용을 보증하거나 책임을 지지 않습니다. 그리고 DBpia에서 제공되는 저작물은 DBpia와 구독계약을 체결한 기관소속 이용자 혹은 해당 저작물의 개별 구매자가 비영리적으로만 이용할 수 있습니다. 그러므로 이에 위반하여 DBpia에서 제공되는 저작물을 복제, 전송 등의 방법으로 무단 이용하는 경우 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

Copyright Information
Copyright of all literary works provided by DBpia belongs to the copyright holder(s)and Nurimedia does not guarantee contents of the literary work or assume responsibility for the same. In addition, the literary works provided by DBpia may only be used by the users affiliated to the institutions which executed a subscription agreement with DBpia or the individual purchasers of the literary work(s)for non-commercial purposes. Therefore, any person who illegally uses the literary works provided by DBpia by means of reproduction or transmission shall assume civil and criminal responsibility according to applicable laws and regulations.

XGBoost와 LightGBM을 이용한 안전 운전자 예측 성능 비교

장승일*, 곽근창**

Comparison of Safety Driver Prediction Performance with XGBoost and LightGBM

Seung-Il Jang*, Keun-Chang Kwak**

요 약

본 논문은 머신러닝 기반 XGBoost와 LightGBM의 성능비교를 진행하였다. XGBoost 알고리즘은 GBDT(Gradient Boosted Decision Trees) 알고리즘을 기반으로 설계 되었다. 하지만 데이터가 클 경우 많은 시간이 소요 된다는 단점이 있다. LightBGM은 GBDT 알고리즘의 단점을 GOSS(Gradient Based One Side Sampling)와 EFB(Exclusive Feature Bundling)기법으로 보완하여 기존의 XGBoost보다 LightGBM방법이 뛰어난 성능과 빠른 처리속도를 보이는 것을 확인할 수 있었다.

Abstract

In this paper, we compare performance of XGBoost and LightGBM based on machine learning. The XGBoost algorithm is designed based on GBDT (Gradient Boosted Decision Trees) algorithm. However, there is a disadvantage in that it takes a lot of time if the data is large. LightBGM is complemented by GOSS (Gradient Based One Side Sampling) and EFB (Exclusive Feature Bundling), which shows that LightGBM has better performance and speed than existing XGBoost.

Key words
XGBoosting, LightGBM, machine learning, GBDT

Ⅰ. 서 론

최근 많은 기업들에서는 머신러닝 알고리즘을 활용해 고객들의 소비 패턴을 분석하여 소비자의 요구를 맞춤화 할 수 있는 콘텐츠를 창출함으로써 소비자의 구매 결정을 도와주는 가상 쇼핑 기능 까지 제공하고 있는 상황이다[1].

브라질에서 자동차 및 주택 보험 회사 중 하나인 Porto Seguro는 자동차 보험 청구 신청 확률을 보다 정확하게 예측하고 안전 운전자 예측 모델을 통하여 자사 고객에게 합리적인 보험금을 청구하고자 경진대회를 개최하였다.

본 논문은 Porto Seguro에서 제공한 약 140만개의 데이터 세트를 사용하여 XGboost 알고리즘과

* 조선대학교 전자공학부
** 조선대학교 저자공학부 교수

※ 본 논문은 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2017R1A6A1A03015496). 또한, 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음(2017-0-00137).

한신대학교 | IP:211.187.169.*** | Accessed 2020/01/11 19:29(KST)

LightGBM 알고리즘의 예측 성능을 비교 분석한다.

II. 본 론

데이터를 분석하고 예측하기 전에 전처리 과정을 거쳐 데이터 시각화하여 학습데이터 간의 특징들의 상관관계를 알아보는 과정은 필수적이다. 각 특징의 데이터 상관관계를 그림 1과 같이 나타낼 수 있다.

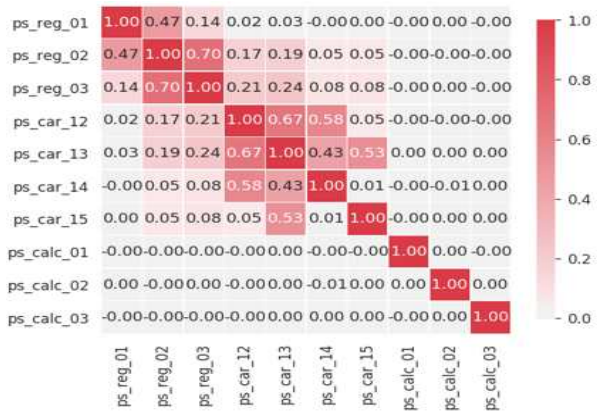

그림 1. 각 특징간의 상관관계 분석

Fig. 1. Analysis of correlation between each feature

본 논문에서는 자동차 보험 청구 신청확률의 목표 값이 매우 불균형하기 때문에 ROC AUC Curve와 유사한 지니계수(gini coefficient)를 사용하였다. 지니 계수는 다음과 같은 수식을 갖는다.

$$gini = 2 \times AUC - 1 \quad (1)$$

여기서 AUC가 갖는 값은 0~1 사이의 값을 갖게 된다. 실제로 gini가 갖는 값은 0~0.5 사이의 값을 갖게 되는데 AUC 값을 통해 gini 값을 좀 더 보기 좋게 만들기 위함이다. XGBoost 알고리즘은 GBDT(Gradient Boosted Decision Trees)라 불리는 기술을 사용한다. boosting은 약한 분류기를 하나의 세트로 정한 다음 정확도를 예측하는 방법이다. 또한 greedy 알고리즘을 사용하여 분류기를 찾고 분산처리를 통해 빠른 속도로 적합한 비중 파라미터를 찾는 알고리즘이다[2]. LightGBM 알고리즘은 GBDT의 단점인 데이터 처리시간을 GOSS(Gradient Based One Side Sampling)와 EFB(Exclusive Feature Bundling)기법으로 보완하여 기존 GBDT를 사용했던 XGBoost와 qGBRT 같은 알고리즘보다 월등한 속도 차이를 보여준다[3]. XGBoost와 LightGBM의 예측 성능을 비교하기 위해 참고한 코드는 Kaggle에 공유 되어진 Bert Carremans의 Data Preparation & Exploration Kernel을 참고하였다. 데이터 세트는 학습 데이터와 검증데이터로 나눠 각각 595,212개와 892,816개로 진행 하였고 입력 변수 개수는 57개이다. 출력은 안전 운전자의 보험가입 유무로 이진분류를 이용한 0과 1로 한다. 먼저 XGBoost 알고리즘의 최대깊이는 4,과적합을 방지하기 위한 부표본 값은 0.9로 하였다. XGBoost와 LightGBM의 성능은 지니계수를 사용하여 비교하였다. K-fold 값은 Light GBM과 동일하게 5이다. LightGBM 알고리즘의 최대 깊이는 4, 트리 모델의 복잡성을 위해 최소 가지수 값은 15로 하였다. 과적합을 방지하기 위한 부표본 값은 0.9, 학습률은 0.1로 하였다. 랜덤 포레스트(Random Forest)를 이용하여 성능을 표1과 같이 비교하였다.

표 1. 알고리즘 성능 비교 표

	지니계수	학습 시간(s)
LightGBM	0.2851	372.09
XGBoost	0.2797	1426.84
RF	0.2633	3004.12

III. 결 론

본 논문에서는 Porto에서 제공한 약 140만개의 데이터베이스를 가지고 LightGBM과 XGBoost을 이용한 안전 운전자예측을 지니계수를 통해 성능을 비교하였다. 실험결과, LightGBM이 XGBoost보다 지니계수 값이 높은 것을 확인할 수 있었다. 또한 XGboost의 단점이였던 대용량 데이터 학습시간을 GOSS와 EFB기법으로 보완한 LightGBM의 학습 시간이 4배정도 빠른 것을 확인할 수 있었다.

참 고 문 헌

[1] 정효주, “네이버 기계학습 활용 사례”, 대한전자공학회 국내학술대회, pp. 2025-2026. 2014.

한신대학교 | IP:211.187.169.*** | Accessed 2020/01/11 19:29(KST)

[2] T. Chen and C. Guestrin “XGBoost: A Scalable Tree Boosting System”, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

[3] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T. Lie. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. Advances in Neural Information Processing Systems 30 NIPS 2017.

한신대학교 | IP:211.187.169.*** | Accessed 2020/01/11 19:29(KST)