Ministério da Educação

Instituto Federal de Educação, Ciência e Tecnologia de São Paulo

Campus Campinas

Especialização em Ciência de Dados

Introdução à Ciências de Dados

Mateus Guilherme Fuini

Porto Seguro Data Challenge

Atividade prática 2

2022

Objetivo

O objetivo deste documento é explicar os principais pontos da solução aplicado no algoritmo base disponibilizado para a elaboração da atividade 2 denominada Porto Seguro Data Challenge.

Pontos a considerar

Para a elaboração da atividade foram testados 3 algoritmos de classificação executados na base balanceada e não balanceada.

O primeiro algoritmo implementado foi o XGBoost (eXtreme Gradient Boosting), desenvolvido por Chen e Guestrin [1] na Universidade de Washington, que usa a estrutura de aumento de gradiente em seu núcleo. A escolha deste algoritmo foi motivada pelos resultados apresentados por Omar, 2018 [2].

O KNN (K Nearest Neighbor), K-vizinho, foi o segundo algoritmo implementado. Proposto por Fukunaga e Narendra [3], considerado um dos classificadores mais simples de ser implementado. Este algoritmo consiste em determinar o rótulo de classificação de uma amostra baseado nas amostras vizinhas advindas de um conjunto de treinamento.

A Regressão Logística foi o último algoritmo a ser testado. Este modelo logístico mostra a relação entre os recursos e, em seguida, calcula a probabilidade de um determinado resultado.

Cada um dos algoritmos recebeu a base de dados desbalanceada e balanceada. Para o balanceamento foi utilizado a técnica proposta por Chawla et al [4] SMOTE (Synthetic Minority Over-sampling Technique), que consiste em criar observações intermediárias entre dados parecidos.

Resultados

A seguir é mostrada a tabela com os resultados obtidos:

Algoritmo	Acurácia	F1-score
XGBoost não balanceado	0,8640	0,6245
XGBoost balanceado	0,9132	0,9120
KNN não balanceado	0,8206	0,4837
KNN balanceado	0,8122	0,8339
Regressão Logística não balanceado	0,8322	0,4752
Regressão Logística balanceado	0,7633	0,7562

Também foi gerado o arquivo para a submissão na plataforma kaggle, que apresentou os seguintes resultados:

Algoritmo	Score privado	Score público
XGBoost não balanceado	0,61445	0,63957
XGBoost balanceado	0,64608	0,65071
KNN não balanceado	0,47554	0,49268
KNN balanceado	0,46238	0,45418
Regressão Logística não balanceado	0,479	0,45431
Regressão Logística balanceado	0,54752	0,55315

Através dos dados apresentados pode-se verificar que o resultado com os dados balanceados apresentaram melhores resultados que os dados não balanceados. Outro ponto que chamou a atenção é a superioridade dos resultados utilizando o algoritmo XGBoost.

Bibliografia

- [1] Chen, T., Guestrin, C. (2016), XGBoost: A Scalable Tree Boosting System., in Balaji Krishnapuram; Mohak Shah; Alexander J. Smola; Charu Aggarwal; Dou Shen & Rajeev Rastogi, ed., 'KDD', ACM, , pp. 785-794 .
- [2] Omar, K.B.A. XGBoost and LGBM for Porto Seguro's Kaggle Challenge: A Comparison. Distributed Computing Group, ETH Zurich.
- [3] Fukunaga, K.; Narendra, P. M. A branch and bound algorithm for computing k-nearest neighbors. IEEE Transactions on Computers, v. 100, n. 7, p. 750–753, 1975.
- [4] Chawla , N. V., Bowyer ,K. W., Hall, L. O., Kegelmeyer, W. P., SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research, 321-357, 2002