Отчет об учебной задаче

Обнаружение мошеннических транзакций

М.А. Кондахчан

Москва 2024

РЕФЕРАТ

Отчет 7 с., 12 рис., 7 источ.

ДАТАСЕТ, МОДЕЛЬ, МЕТРИКА, МЕТОД, ДАННЫЕ, НЕСБАЛАНСИРОВАННЫЙ

Объектом исследования являются методы машинного обучения, способные выявлять среди финансовых операций, совершенных с помощью кредитных карт, мошеннические с высокой точностью.

Цель работы – разработать алгоритм выявления мошеннических операций, основанный на методах машинного обучения, и провести сравнение использованных методов на предоставленном наборе финансовых операций.

Проводились исследования работы методов при различных параметрах. Также использовались 2 метода преобразования несбалансированных данных.

В результате выявлено, что методы показывают лучший результат при cost-sensitive learning, a random forest работает лучше XGBoost при проведенных экспериментах

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящей работе была рассмотрена задача обнаружения мошеннических финансовых операций. Исследование было проведено с использованием публичного набора данных. Данный набор данных является несбалансированным, в нем 492 мошеннических транзакции из 284807, что составляет 0.172%. Для выполнения задачи необходимо изучить подходящие модели, выбрать нужные метрики и затем написать программную реализацию.

1. ЗАДАЧА ОБНАРУЖЕНИЯ МОШЕННИЧЕСКИХ ФИНАНСОВЫХ ОПЕРАЦИЙ

1.1 ПОСТАНОВКА ЗАДАЧИ

Главной задачей является создать бинарный классификатор, который определяет финансовую операцию как легальную или мошенническую.

1.2 СПЕЦИФИКА ДАННЫХ

Датасет представлен в виде таблицы с метками классов, где 1 – мошенническая транзакция, 0 – легальная. Признаки являются только числовыми, и уже представлены в виде результата применения метода главных компонент, кроме времени и количества. Время – количество секунд между первой и текущей транзакциями, количество – сумма транзакции.

1.3 ОБЗОР ВОЗМОЖНЫХ РЕШЕНИЙ

Выбранные методы – деревья (xgBoost, random forest) приводятся в пример как одни из лучших моделей для работы с несбалансированными данными. Их преимущество заключается в возможности присваивать веса классам. Также, xgBoost корректирует веса при каждой итерации [2], а random forest сочетает в себе технику выборки и ансамблевое обучение, что приводит к уменьшению выборки класса, представленного в большем количестве, и построению деревьев на более сбалансированном наборе данных [7]. В большинстве своем именно эти модели упоминаются в статьях и на форумах [1] – [3].

Выбранные метрики pr-auc и roc-auc, для отображения результата классификации – матрица ошибок. Данные метрики лучше обеспечивают понимание эффективности модели, чем стандартные, поскольку дисбаланс по количеству примеров в классах приводит к вырожденности значений.

Для балансировки данных применен метод SMOTE, который заключается в создании экземпляров класса меньшинства на основе существующих, и взвешивание классов. Был выбран over-sampling метод, потому что он не приведет к потери информации класса большинства, а непосредственно SMOTE, поскольку не приводит к переобучению, создавая

новые экземпляры [4] – [6].

2. ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1 ОПИСАНИЕ РЕАЛИЗАЦИИ

При обработке данных отброшены все дублирующиеся записи.

Первыми были модели с взвешенными классами. Для XGBoost вес ошибки был рассчитан как отношение числа легальных к мошенническим транзакциям, а для Random forest – как отношение числа всех экземпляров к произведению числа классов и экземпляров класса. Применялся parameter tuning с помощью инструментов автоматической настройки гиперпараметров.

2.2 ПЛАН ПРОВЕДЕНИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

Сначала модели обучались с параметрами по умолчанию, после чего, применяя, определялись более подходящие в процессе tuning, и исследование результатов проводилось уже у модели с параметрами, выбранными таким образом. После того, как обучена модель, ее результат представлялся в виде матрицы ошибок (confusion matrix). Высчитывался roc-auc, pr auc, изображались кривые на графике.

Ниже приведены эти графики и матрицы для исследуемых моделей при разных способах работы с несбалансированными данными.

Результаты xgBoost с взвешенными классами:

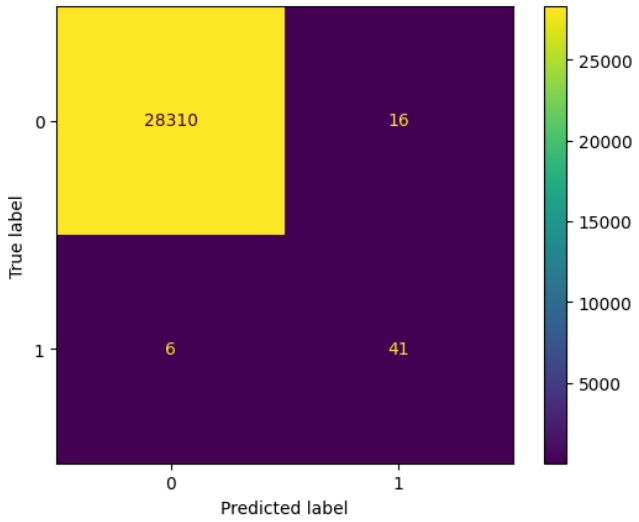

Рисунок 1 – матрица ошибок xgBoost, взвешенные классы.

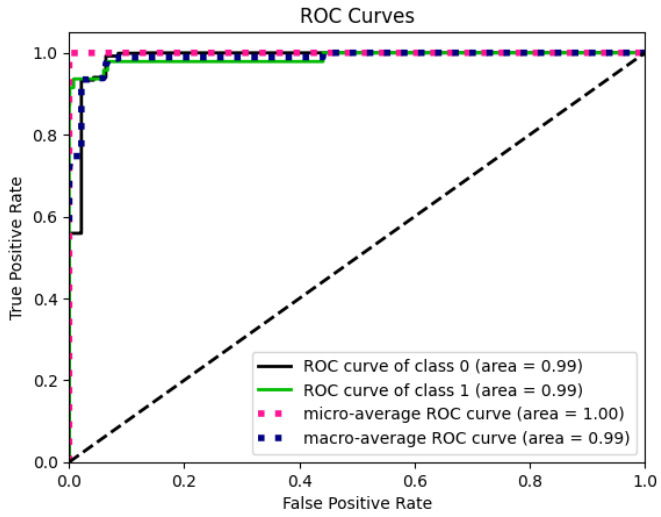

Рисунок 2 – ROC кривая xgBoost, взвешенные классы.

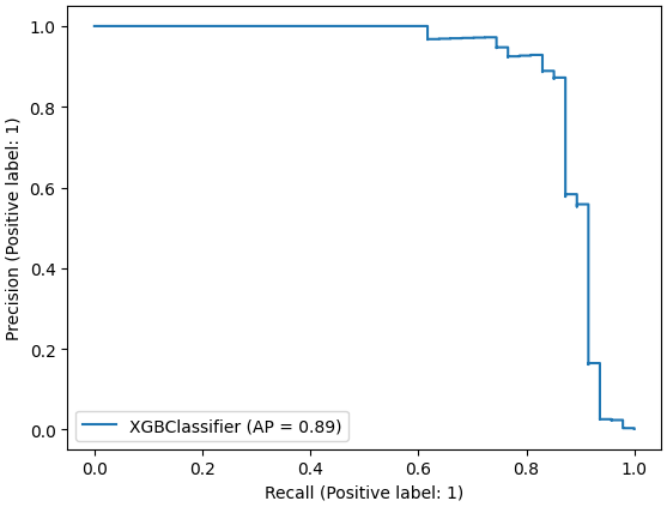

Рисунок 3 – PR кривая xgBoost, взвешенные классы.

Результаты xgBoost с SMOTE:

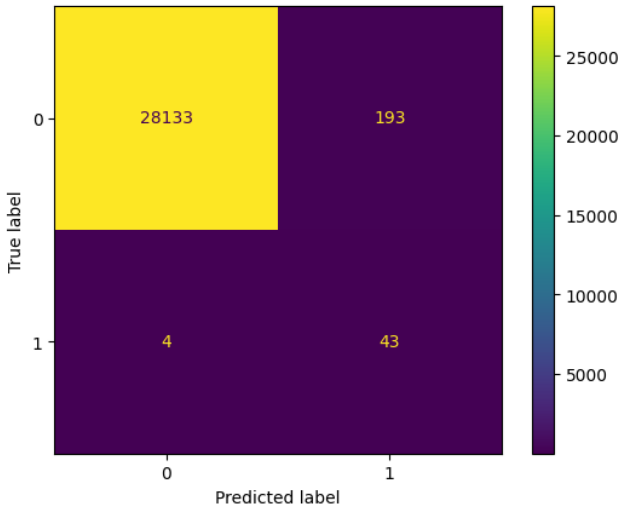

Рисунок 4 – матрица ошибок xgBoost, SMOTE.

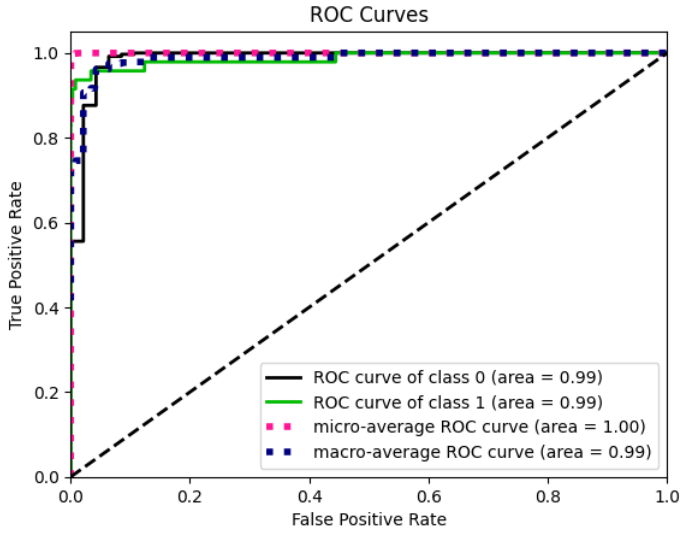

Рисунок 5 – ROC кривая xgBoost, SMOTE.

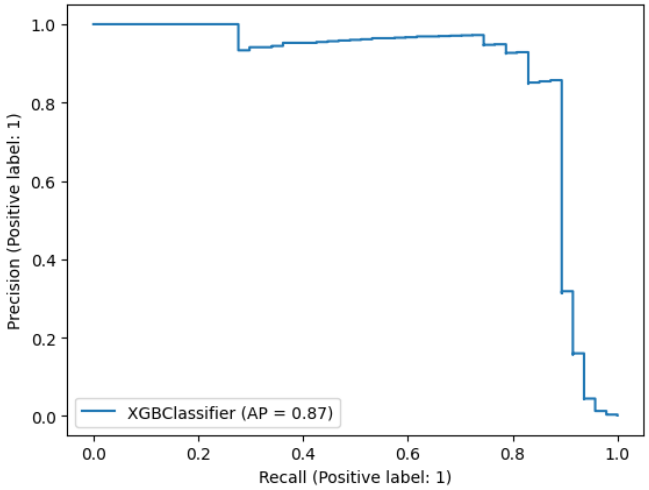

Рисунок 6 – PR кривая xgBoost, SMOTE.

Результаты Random Forest с взвешенными классами:

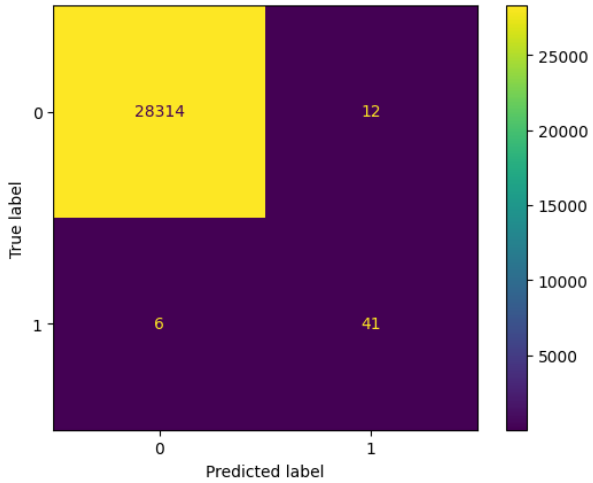

Рисунок 7 – матрица ошибок Random Forest, взвешенные классы.

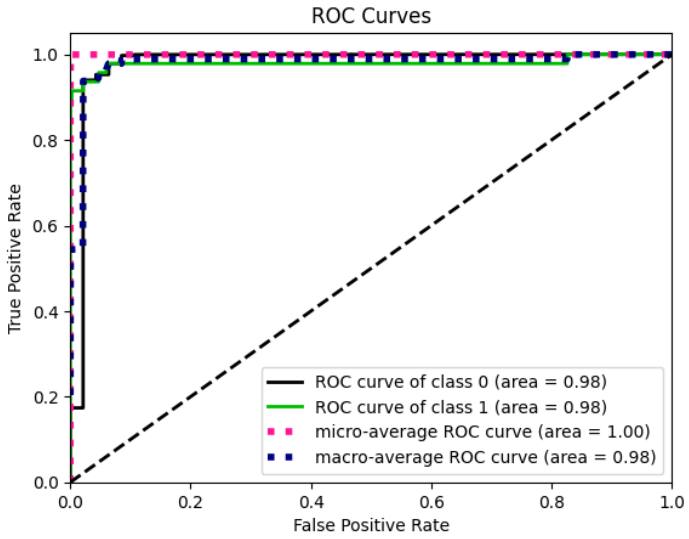

Рисунок 8 – ROC кривая Random Forest, взвешенные классы.

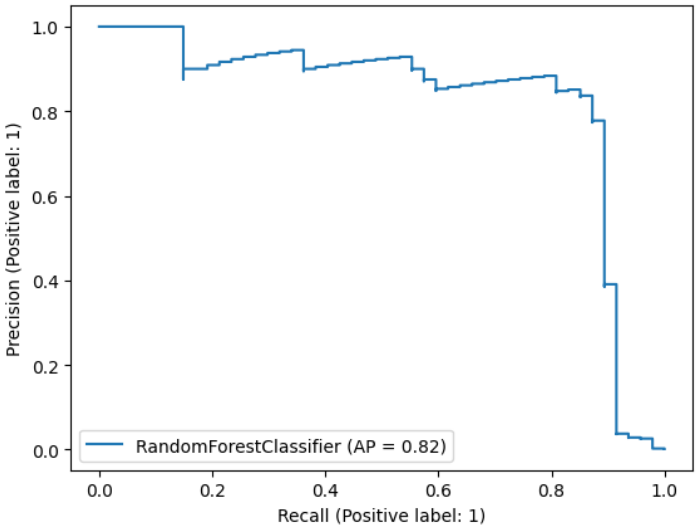

Рисунок 9 – PR кривая Random Forest, взвешенные классы.

Результаты Random Forest с SMOTE:

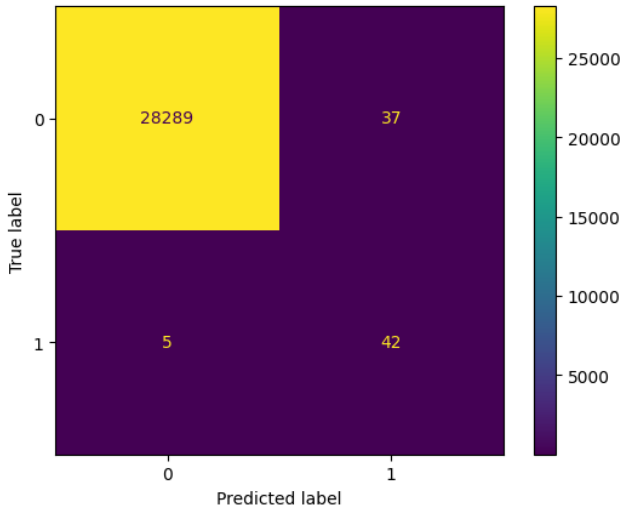

Рисунок 10 – матрица ошибок Random Forest, SMOTE.

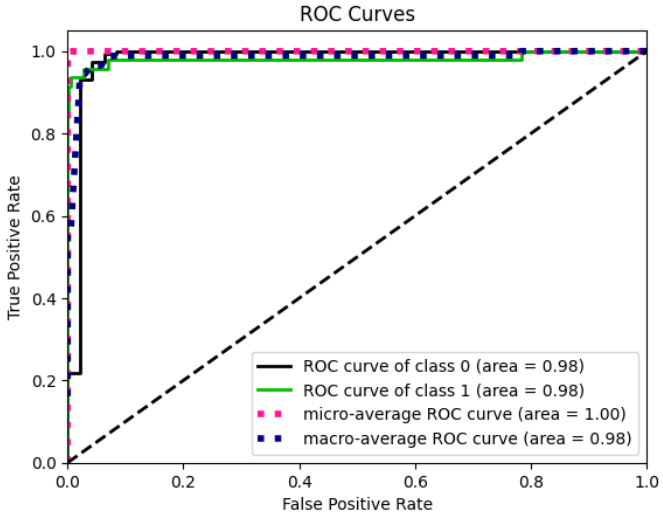

Рисунок 11 – ROC кривая Random Forest, SMOTE.

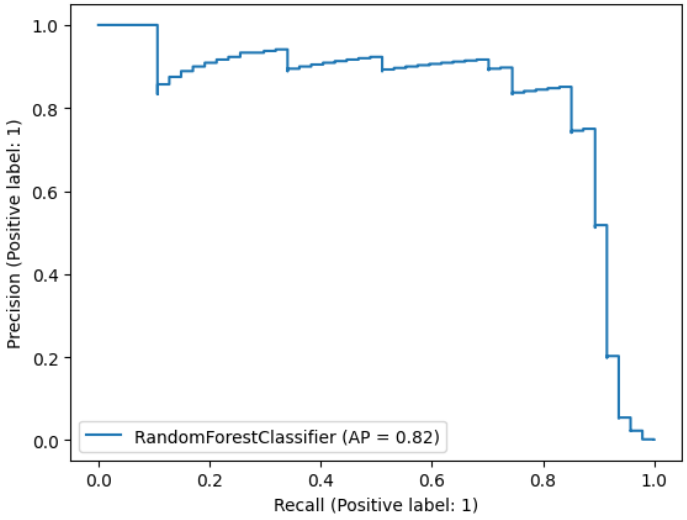

Рисунок 12 – ROC кривая Random Forest, SMOTE.

Сравнивая методы работы с несбалансированными данными для одних и тех же моделей, заметно, что XGBoost гораздо точнее классифицировал при взвешенных классах, чем при ресемплинге. Хоть значения под roc кривой почти идентичны, исходя из матрицы ошибок видно, что большое количество (193) транзакций было определено как мошеннические. Random forest также стал классифицировать менее точно. Но при этом в обоих случаях число ложноположительных результатов снизилось. Исходя из предметной области можно сказать, что такой результат имеет и положительную сторону, поскольку лучше определить легальные транзакции как мошеннические, чем наоборот.

Сравнивая модели друг с другом, для случая взвешивания классов, результаты почти идентичны по всем показателями, но значения немного лучше у random forest – меньше ложноотрицательных определений. При ресемплинге разница сильнее и также в пользу random forest.

Значения roc-auc при этом всегда велики и около 1, а сами кривые выглядит почти как прямой угол, значения pr-auc больше 0,8, кривые также имеют схожий с прямым углом вид, что говорит о высокой точности при классификации.

ЗАКЛЮЧЕНИЕ

Рассмотрены методы машинного обучения xgBoost и Random forest. Предложено использовать 2 способа для балансировки данных: ресемплинг с генерацией экземпляров класса меньшинства и взвешивание классов, и настройку гиперпараметров. Испытаны обе модели с приведенными способами балансировки и достигнуты результаты, что применение алгоритма Random forest с взвешенными классами классифицирует финансовые операции наиболее точным образом.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. 7 Techniques to Handle Class Imbalance in Machine Learning / [Электронный ресурс]. – Режим доступа: URL: https://medium.com/@data.pilot/7-techniques-to-handle-class-imbalance-in-machine-learning-eb1297419ec9 (дата обращения 09.3.2024).
2. Imbalanced data: best practices / [Электронный ресурс]. – Режим доступа: URL: https://rihab-feki.medium.com/imbalanced-data-best-practices-f3b6d0999f38 (дата обращения 11.3.2024).
3. Quora / [Электронный ресурс]. URL: https://www.quora.com/Which-machine-learning-algorithms-tend-to-perform-better-on-highly-imbalanced-datasets (дата обращения 09.3.2024).
4. 10 Techniques to Solve Imbalanced Classes in Machine Learning (Updated 2024) / [Электронный ресурс]. – Режим доступа: URL: https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/#The_Metric_Trap (дата обращения 09.3.2024).
5. Machine Learning Resampling Techniques for Class Imbalances / [Электронный ресурс]. – Режим доступа: URL: https://towardsdatascience.com/machine-learning-resampling-techniques-for-class-imbalances-30cbe2415867 (дата обращения 11.3.2024).
6. What Is Imbalanced Data and How to Handle It? / [Электронный ресурс]. – Режим доступа: URL: https://www.turintech.ai/what-is-imbalanced-data-and-how-to-handle-it/ (дата обращения 09.3.2024).
7. Surviving in a Random Forest with Imbalanced Datasets/ [Электронный ресурс]. – Режим доступа: URL: https://medium.com/sfu-cspmp/surviving-in-a-random-forest-with-imbalanced-datasets-b98b963d52eb (дата обращения 05.04.2024).