КЛАССИФИКАЦИЯ И КЛАСТЕРИЗАЦИЯ КЛИЕНТСКОЙ БАЗЫ ОРГАНИЗАЦИИ

А. О. Яблонская

anna.yablonskaya2002@yandex.ru;
Научный руководитель — А. Э. Малевич, кандидат физико-математических наук, доцент

Оценка степени благонадёжности клиента является важным фактором для принятия решений о дальнейшем финансовом сотрудничестве организации с клиентом. Это позволяет принимать более обоснованные решения при финансовом сотрудничестве, снижает риск финансовых потерь для организации, также может позволить определить потенциальные риски нарушения законодательства и принять соответствующие меры до возникновения проблем.

Ключевые слова: машинное обучение; степень благонадёжности; многоклассовая классификация; случайный лес; метрика F1-score; кластеризация; алгоритм Лейдена.

ВВЕДЕНИЕ

Организации в своих базах данных накапливают разнородную информацию о своих клиентах. Для решения конкретной задачи возникает необходимость разбить весь массив клиентов на отдельные категории. Сделать это можно, построив «модель» и обучив её, используя методы машинного обучения, на имеющихся в распоряжении организации данных о клиентах. Использование методов кластеризации в свою очередь позволяет выявлять скрытые или неочевидные взаимосвязи между клиентами.

ПРЕДОБРАБОТКА ИСХОДНЫХ ДАННЫХ

Классификацию клиентской базы продемонстрируем на примере решения задачи определения степени благонадёжности клиента. Требуется построить модель, которая, получив на входе информацию о клиенте, должна определить степень его благонадёжности: низкая, средняя или высокая.

Прежде чем приступать к построению модели, необходимо провести первичный анализ имеющихся в наличии данных, сформировать из них обучающий набор и выбрать наиболее адекватный данным алгоритм классификации. В рассматриваемом примере в силу сильной неоднородности исходных данных было принято решение разбить всех клиентов организации на несколько групп, различающихся по наличию информации о клиентах. В каждой группе клиенты изначально разбиты на три класса: низкая, средняя и высокая степень благонадёжности.

Соответственно для каждой группы необходимо было построить отдельную модель машинного обучения.

Для некоторых групп клиентов не хватало данных для обучения модели. Их необходимо было добавлять. Дополнять датасет абсолютно случайными данными нельзя, поскольку тогда могут появиться ложные зависимости в данных, что может привести к недообучению модели машинного обучения, или же за счёт ложной информации в дальнейшем модель будет плохо обрабатывать новых клиентов. Дополнительные данные были сгенерированы с учётом изначального распределения исходных данных.

Помимо того, при работе с данными оказалось, что они являются несбалансированными, а также некоторые признаки высоко коррелируют друг с другом.

КЛАССИФИКАЦИЯ

Для каждой группы клиентов необходимо строить отдельную модель машинного обучения. Было принято решение рассматривать несколько методов классификации. Лучшими оказались алгоритмы: бустинг (XGBoost), решающие деревья (Decision Tree) и случайный лес (Random Forest) [1]. Поскольку в данных имеется высокая корреляция, а также несбалансированность классов, наиболее подходящим алгоритмом в итоге оказался случайный лес, так как он в силу своих особенностей хорошо справляется с данными проблемами.

Случайный лес – это алгоритм, который комбинирует несколько решающих деревьев и принимает окончательное решение с помощью голосования. В случае классификации это большинство голосов. Алгоритм называется случайным лесом, поскольку он использует: (а) при построении деревьев – случайные выборки обучающего набора; (б) при разделении узлов – случайные подмножества признаков.

Для оценки качества моделей была выбрана метрика F1-score. Достигнутое качество построенных моделей приведено в таблице.

Результаты обучения моделей для трех групп клиентов

Группа / Модель	XGBoost	Decision Tree	**Random Forest**
группа1	75,4%	63,3%	**97,82%**
группа2	62,5%	54,2%	**98,37%**
группа3	73,7%	60,2%	**91,25%**

КЛАСТЕРИЗАЦИЯ

Применение методов кластеризации позволяет найти неочевидные связи между клиентами. В данной работе было использовано три метода кластеризации [2]: *k*-средних, иерархическая и графовая кластеризация алгоритмом Лейдена.

В случаях, когда количество кластеров заранее неизвестно, для его определения использовался метод силуэтов.

Также использовалась иерархическая кластеризация. В этом случае кластеры можно отобразить при помощи дендограммы (см. рис. 1). Она позволяет лучше увидеть разграничение кластеров и посмотреть, насколько кластеры отличаются друг от друга.

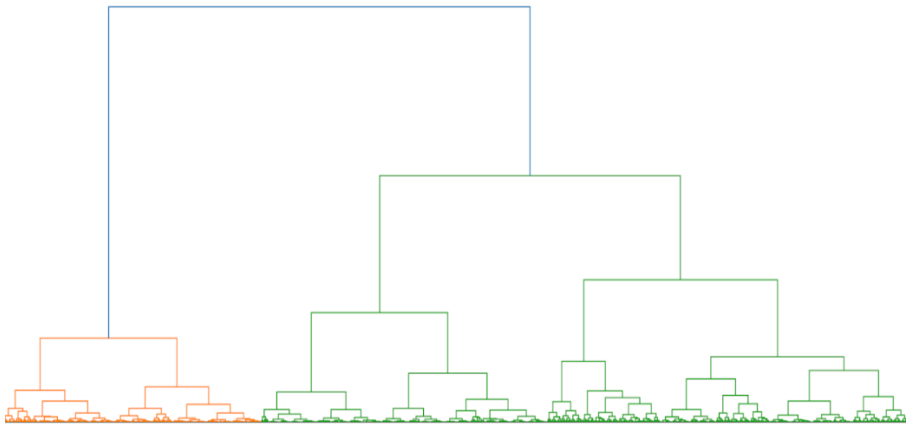

Рис. 1. Дендограмма, построенная на данных первой группы клиентов

Помимо того, был использован графовый метод кластеризации алгоритмом Лейдена [3]. Этот алгоритм разделяет узлы графа на непересекающиеся сообщества таким образом, чтобы максимизировать показатель модульности для каждого сообщества. Модульность в свою очередь максимизирует разницу между реальным количеством рёбер в сообществе и ожидаемым числом рёбер в сообществе.

Для данной задачи сначала необходимо из всех имеющихся данных составить граф. Вершинами графа должны быть клиенты, а рёбрами – любые связи между ними. Такие связи можно извлечь из информации о взаимодействиях между клиентами. Также можно связывать клиентов, исходя из каких-либо описательных критериев, например вида деятельности клиентов, их местоположения и т.п. (см. рис. 2).

ЗАКЛЮЧЕНИЕ

В итоге было построено несколько моделей машинного обучения для классификации клиентов по степени благонадёжности. Наилучшие результаты показала модель случайного леса (Random Forest). Также для нахождения новых скрытых взаимосвязей между клиентами были

применены методы кластеризации. Работа реализована на языке программирования Python с использованием библиотек: Pandas, numpy, Scikit-learn, Matplotlib и Seaborn.

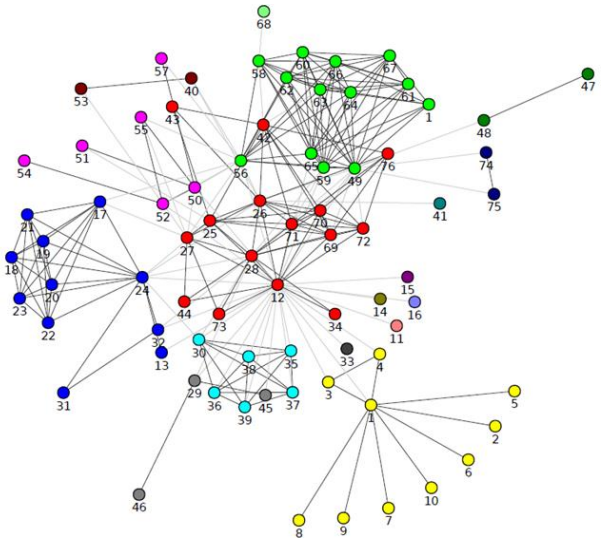

Рис. 2. Сообщества, выделенные в графе после применения алгоритма Лейдена

Библиографические ссылки

1. *Sruthi, E.R.* Understand Random Forest Algorithms With Examples (Updated 2023) [Электронный ресурс] / Analytics Vidhya, 2023. Режим доступа: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/ Дата доступа: 20.06.2023.
2. George Seif The 5 Clustering Algorithms Data Scientist Need to Know [Электронный ресурс] / Towards Data Science, 2018. – Режим доступа: https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68 – Дата доступа: 20.06.2023.
3. *Traag, V.A., Waltman, L. & van Eck, N.J.* From Louvain to Leiden: guaran-teeing well-connected communities // Sci Rep **9**, 5233 (2019). Режим доступа: https://doi.org/10.1038/s41598-019-41695-z – Дата доступа: 20.06.2023.