DOI 10.37882/2223-2966.2024.6-2.22

СИСТЕМНЫЙ ПОДХОД К КЛАССИФИКАЦИИ БОЛЬШИХ ДАННЫХ В КОРПОРАТИВНЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

A SYSTEMATIC APPROACH TO THE CLASSIFICATION OF BIG DATA IN CORPORATE INFORMATION SYSTEMS

A. Kasymov
A. Lysenko

Summary. Classification of big data in information systems has a critical rule in understanding the organization aims to make development in the organizations commerce. Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues of the companies, companies are seeking to develop means to predict potential customer to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn. this will enhance the organization activities, which will lead to more effective information system. Four classification algorithms were tested. Social Network Analysis (SNA) features are also extracted and used in classifications to enhance the information systems. The use of SNA enhanced the performance of the model of information system. The model experimented four algorithms: Decision Tree, Random Forest, Gradient Boosted Machine Tree «GBM» and Extreme Gradient Boosting «XGBOOST». However, the best results were obtained by applying XGBOOST algorithm. This algorithm was to get two classes as accurate classification as information system to be understood and informative.

Keywords: Big data classification, Customer churn, Machine learning algorithms, Social Network Analysis (SNA), Information systems.

Касымов Алексей Алексеевич
аспирант, Воронежский государственный технический университет
kasimlele@live.ru
Лысенко Алексей
Аспирант, Белгородский государственный аграрный университет имени В.Я. Горина

Аннотация. Классификация больших данных в информационных системах играет критическую роль в понимании целей организации, стремящейся к развитию в сфере коммерции. Отток клиентов является серьезной проблемой и одной из самых важных задач для крупных компаний. В связи с прямым влиянием на доходы компаний, они стремятся разработать методы для прогнозирования потенциального оттока клиентов. Следовательно, выявление факторов, увеличивающих отток клиентов, важно для принятия необходимых мер по его снижению. Это улучшит деятельность организации, что приведет к более эффективной информационной системе. Были протестированы четыре алгоритма классификации. Также были извлечены и использованы в классификациях признаки анализа социальных сетей (SNA) для улучшения информационных систем. Использование SNA повысило производительность модели информационной системы. Модель испытала четыре алгоритма: Decision Tree (дерево решений), Random Forest (случайный лес), Gradient Boosted Machine Tree (GBM) и Extreme Gradient Boosting (XGBOOST). Однако лучшие результаты были получены при применении алгоритма XGBOOST. Этот алгоритм позволил получить две классы с точной классификацией, что сделало информационную систему понятной и информативной.

Ключевые слова: классификация больших данных, Отток клиентов, Алгоритмы машинного обучения, Анализ социальных сетей (SNA), Информационные системы.

Введение

Область больших данных определяется как наборы данных, которые слишком велики или сложны, чтобы с ними можно было справиться с помощью традиционного прикладного программного обеспечения для обработки данных. они характеризуются четырьмя параметрами: объем, разнообразие, скорость и достоверность. Эти четыре переменные относятся к размеру данных, типам, которые организация обрабатывает, таким как структурированные, полуструктурированные и неструктурированные данные, полученные в рамках любого принятого проекта и за его пределами, скорость создания данных является отличительной чертой больших данных и, наконец, достоверность и надежность различных источников данных различаются. Например, социальные сети переполнены спамом, а веб-спам составляет более 20 % всего контента во Всемирной паутине. Аналогичным образом, потоки кликов с веб-сайтов и мобильный трафик сильно подвержены помехам. Кроме того, получение глубоких семантических знаний из текста во многих ситуациях остается сложной задачей, несмотря на значительные достижения в области обработки естественного языка [1, 2].

Цель

Основная цель данного исследования — показать, как интеграция методов классификации данных может улучшить управление данными в корпоративных информационных системах.

Методы

Системный анализ: В различных областях науки, информационных технологий и знаний сложность систем имеет большое значение. По мере усложнения систем традиционный метод решения проблем становился неэффективным. Системный анализ заключается в изучении бизнес-проблемы, определении ее целей и требований, а затем разработке наиболее оптимального решения для удовлетворения этих потребностей. системный анализ указан не в списке полей, а во всех полях, столько данных, сколько будет получено в поле, столько, сколько потребуется для получения системного анализа. например, в телекоммуникационных операторах, где насчитываются миллионы пользователей, и у каждого пользователя в течение одного дня могут быть звонки, сообщения, интернет-сессии, услуги, таким образом, у каждого пользователя может быть более 1000 выборок данных в день, для 20 миллионов пользователей это было бы сложнее в течение одного дня. и это обычный день [3–4–5].

Классификация данных: когда наборы данных слишком велики, будет проще разделить эти данные на небольшие категории, социальные группы, организации, возможно, потребуется принять некоторые решения только для конкретных групп, для этого потребуется провести классификацию или кластеризацию данных. Классификация и системный анализ — это два интегрированных процесса, системы анализа необходимы для классификации, а классификация также необходима для глубокого анализа. На следующем рисунке показан анализ социальной сети после проведения классификации с использованием графических методов, чтобы определить, что есть группы, которые классифицируются по одному определенному признаку, но все группы имеют связи с другой группой.

Математические алгоритмы классификации

Деревья решений: Дерево принятия решений — один из самых известных мощных инструментов алгоритмов контролируемого обучения, используемых для классификации. Он создает древовидную структуру, подобную блок-схеме, где каждый внутренний узел обозначает тест по атрибуту, каждая ветвь представляет результат теста, а каждый конечный узел содержит метку класса. Он создается путем рекурсивного разделения обучающих данных на подмножества на основе значений атрибутов до тех пор, пока не будет выполнен критерий остановки, такой как максимальная глубина дерева или минимальное количество выборок, необходимых для разделения узла. Во время обучения алгоритм дерева решений выбирает наилучший атрибут для разделения данных на основе такого показателя, как энтропия или примесь Джини, который измеряет уровень примеси

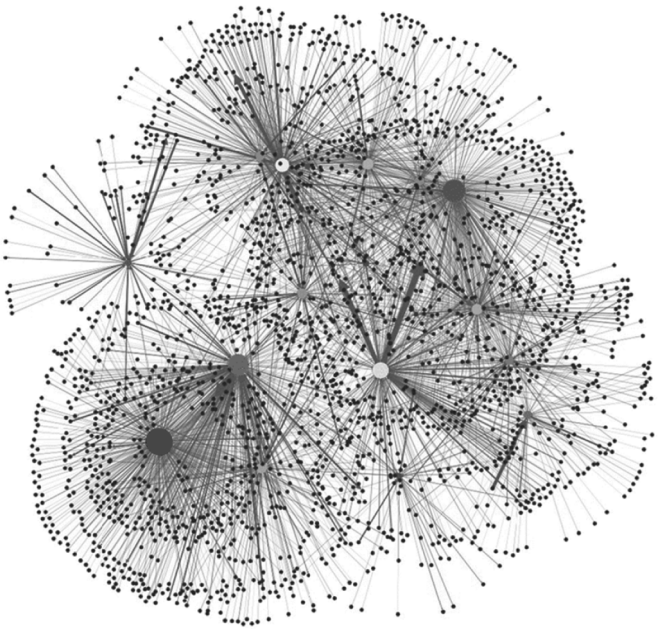

Рис. 1. Анализ социальных сетей [4]

или случайности в подмножествах. Цель состоит в том, чтобы найти атрибут, который максимизирует получение информации или уменьшает количество примесей после разделения [6].

Математически он представлен в основном структурой if и else с пороговыми значениями на многих уровнях, и эти пороговые значения неоднократно меняются во время обучения.

Случайный лес — очень полезный метод в машинном обучении. На этапе обучения он генерирует множество деревьев решений. Чтобы измерить случайное подмножество характеристик в каждом подразделении, для построения каждого дерева используется случайное подмножество набора данных. Поскольку в результате рандомизации каждое дерево становится более вариабельным, вероятность переобучения снижается, а общая эффективность прогнозирования повышается. В методе используется голосование (задачи классификации) для объединения результатов всех деревьев при прогнозировании. Этот совместный процесс принятия решений, основанный на анализе многочисленных деревьев, дает пример стабильных и точных результатов. Поскольку случайные леса могут обрабатывать сложные данные, сводить к минимуму переобучение и давать точные прогнозы в различных условиях, они часто используются для регрессии и классификации [7].

XGBoost (eXtreme Gradient Boosting) — это библиотека с открытым исходным кодом, используемая в машинном обучении и предоставляющая функциональность для решения задач, связанных с регуляризацией градиентного бустинга [8].

GBM: Этот алгоритм поэтапно строит аддитивную модель; он позволяет оптимизировать произвольные дифференцируемые функции потерь. На каждом этапе деревья регрессии n_classes_ подбираются по отрицательному градиенту функции потерь, например, двоичная или многоклассовая логарифмическая потеря. Бинарная классификация — это особый случай, когда индуцируется только одно дерево регрессии [9].

Существует слишком много алгоритмов классификации данных, таких как k-ближайшие соседи, наивные байесовские алгоритмы, алгоритмы глубокого обучения, нейронные сети и т.д.

Практическое применение

Компании-оператору связи необходимо внести некоторые усовершенствования, чтобы привлечь новых клиентов и предотвратить их уход, это называется «прогнозированием оттока». Планируется разделить клиентов на несколько классов в соответствии с их лояльностью, класс с минимальной лояльностью — это классы, которые могут заинтересовать компанию, это позволит создать информационную систему, классифицирующую клиентов на основе лояльности.

Другой информационной системой может быть классификация, основанная на рабочих местах, таких как класс учащихся, инженерный класс, класс преподавателей и т.д.

Информационная система также может быть основана на классификации в зависимости от интересов, таких как спорт, музыка, искусство и т.д. эта информационная система позволит компании легко понять клиента и его потребности. Каждый класс получит от компании несколько реальных действий, которые будут точно соответствовать целевому классу.

Проблемы классификации

Некоторые основные проблемы классификации больших данных для информационных систем возникают в процессе обучения, поскольку процесс классификации состоит из нескольких основных этапов:

Подготовительные данные: это основной момент, поскольку необходимо получить данные в виде сбалансированных классов, где у каждого класса есть образцы в виде других классов, это необходимые сбалансированные данные. Несбалансированные данные приведут к проблемам, которые называются недостаточным или избыточным соответствием. предположим, что существует только два класса, и рационы этих двух классов составляют 80 % и 20 %, это позволит представить данные так, как будто все они относятся к первому классу.

Разработка функций или атрибутов означает, как получить новые функции из существующих, чтобы сделать информационные системы более точными. Новые функции могут быть получены путем выполнения некоторых вычислений, таких как получение данных. Пороговое значение данных также может быть использовано для получения новых функций.

Типы атрибутов: это означает, что существуют числовые и категориальные атрибуты, категориальные атрибуты необходимо преобразовать в числовой тип, используя некоторые функции библиотек языка программирования (python), такие как индексация.

Практическая часть

Данные взяты из[1]. Основная цель — протестировать классификацию с использованием описанных выше четырех алгоритмов.

Необходимо определить некоторые параметры:

Ложно положительное — это ошибка в бинарной классификации, при которой результат теста ошибочно указывает на наличие состояния (например, заболевания, когда заболевание отсутствует), в то время как ложно отрицательное — это противоположная ошибка, когда результат теста ошибочно указывает на отсутствие состояния, которое на самом деле присутствует. Это два вида ошибок в бинарном тесте, в отличие от двух видов правильного результата (истинно положительный и истинно отрицательный). В медицине эти ошибки также известны

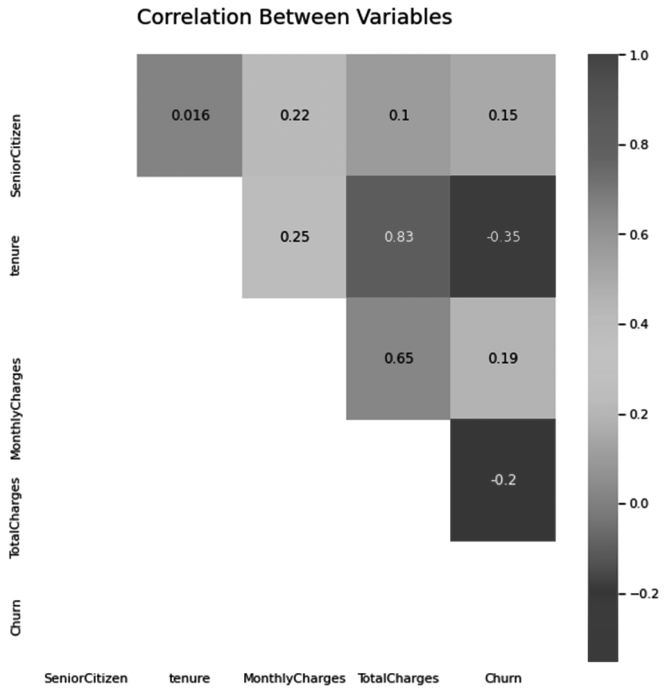

Рис. 2. Корреляция между оттоком и другими атрибутами

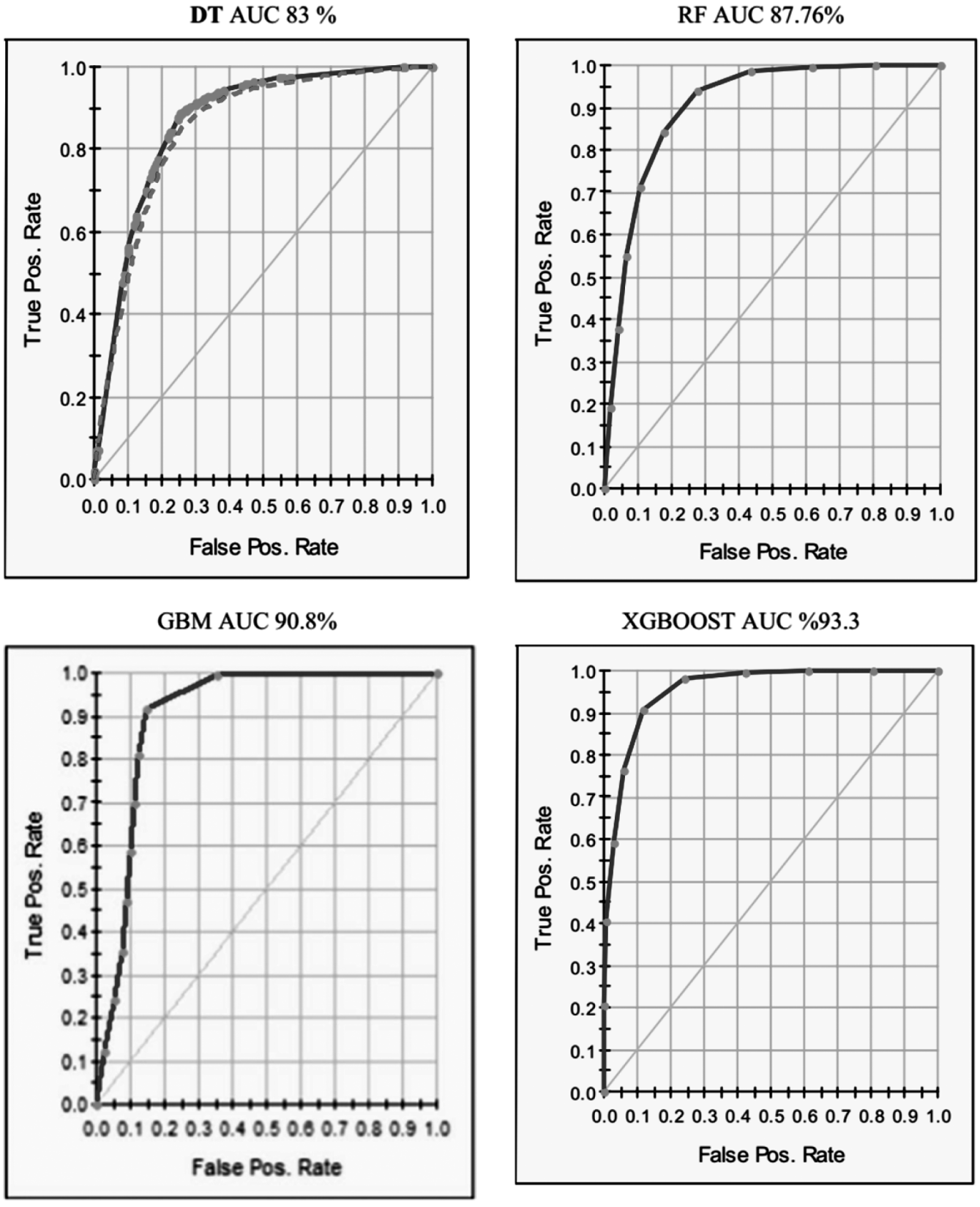

Рис. 3. Кривые точности

как ложноположительный (или ложноотрицательный) диагноз, а в статистической классификации как ложноположительная (или ложноотрицательная) ошибка.

AUC: Одним из важных аспектов машинного обучения является оценка модели. У вас должен быть какой-то механизм для оценки вашей модели. Именно здесь на первый план выходят показатели производительности, которые дают нам представление о том, насколько хороша модель. Если вы знакомы с некоторыми основами машинного обучения, то вы, должно быть, сталкивались с некоторыми из этих показателей, такими как точность, прецизионность, отзыв, auc-roc и т.д., которые обычно используются для задач классификации. В этой статье мы подробно рассмотрим один из таких показателей — кривую AUC-ROC.

На рисунке 2 показана корреляция между оттоком и другими атрибутами.

Результаты и обсуждение

Результаты AUC алгоритмов классификации показаны ниже на рисунке 3.

Результаты классификации для прогнозирования оттока приведены в таблице ниже.

Таблица 1.

Точность результатов четырех алгоритмов

	XGBOOST	GBM	Случайный лес	Деревья решений
Нормальные атрибуты	86.3	81.2	69.1	71
Социальные атрибуты	79.3	70.3	72.5	75.2
Нормальные и социальные	95.5	88.9	92.99	88.2

Информационная система может быть усовершенствована за счет использования новых атрибутов, которые извлекаются в качестве социальных атрибутов из существующих, эти результаты позволят создать надежную классификацию, что позволит организации эффективно совершенствовать информационную систему

Выводы

Исследование демонстрирует, насколько важна категоризация больших данных для совершенствования бизнес-информационных систем. В исследовании рассматривается проблема оттока клиентов, которая является серьезной проблемой для крупного бизнеса, поскольку напрямую влияет на доходы. Это подчеркивает, насколько важно предвидеть возможный отток клиентов. Выявляя причины, приводящие к оттоку клиентов, фирмы могут принимать соответствующие меры для решения проблемы, что улучшает общую работу организации и приводит к созданию более эффективных информационных систем. Деревья принятия решений, случайные леса, машинные деревья с градиентным ускорением (GBM) и экстремальное ускорение градиента (XGBOOST) — вот четыре метода классификации, которые были опробованы в этом исследовании. Для дальнейшего повышения эффективности информационных систем в процесс категоризации были также включены характеристики, полученные на основе анализа социальных сетей (СНС). Как алгоритм, обладающий наилучшей точностью при разделении данных на две отдельные категории, XGBOOST зарекомендовал себя как наиболее эффективный метод разработки системы информации, которая является одновременно понятной и поучительной. Этот тщательный метод подчеркивает, насколько важно интегрировать исследования социальных сетей и использовать передовые алгоритмы категоризации, чтобы максимально повысить функциональность и производительность корпоративных информационных систем.

ЛИТЕРАТУРА

1. Ульман Д., Лесковец Ю., Раджараман А. Анализ больших наборов данных. — Litres, 2022.
2. Гойкхман Ш., Илиопулос А., Леви Е. Балансировка нагрузки для больших баз данных в оперативной памяти. — 2018.
3. Власов А.И., Карпунин А.А., Новиков И.П. Системный анализ технологии обмена и хранения данных blockchain // Современные технологии. Системный анализ. Моделирование. — 2017. — №. 3 (55). — С. 75–83.
4. Al-Molhem N.R., Rahal Y., Dakkak M. Social network analysis in Telecom data // Journal of Big Data. — 2019. — Т. 6. — №. 1. — С. 99.
5. Ahmad A.K., Jafar A., Aljoumaa K. Customer churn prediction in telecom using machine learning in big data platform //Journal of Big Data. — 2019. — Т. 6. — №. 1. — С. 1–24.
6. Пичугин О.Н., Прокофьева Ю.З., Александров Д. М. Деревья решений как эффективный метод анализа и прогнозирования //Нефтепромысловое дело. — 2013. — №. 11. — С. 69–75.
7. Пальмов С.В., Денискова А.О. Случайный лес: основные особенности // Наука сегодня: теоретические и практические аспекты. — 2017. — С. 51–53.
8. Chen T., Guestrin C. Xgboost: A scalable tree boosting system // Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. — 2016. — С. 785–794.
9. Ayyadevara V.K., Ayyadevara V.K. Gradient boosting machine // Pro machine learning algorithms: A hands-on approach to implementing algorithms in python and R. — 2018. — С. 117–134.