УДК: 33

DOI 10.21661/r-555930

С.С. Иванькова

ОЦЕНКА И АНАЛИЗ РИСКА БАНКРОТСТВА С ИСПОЛЬЗОВАНИЕМ DECISION TREE МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

Аннотация*: эффективное и заблаговременное прогнозирование банкротства компаний имеет важно значение для всех участников рынка. По мере развития информационного общества традиционные методы выявления банкротства становятся менее эффективными и более трудозатратными. Поэтому сочетание традиционных методов с современными моделями искусственного интеллекта может быть эффективно применено в современных экономических условиях. Основная цель данной статьи – оценить риск банкротства с помощью дерева решений, сравнить различные модели машинного обучения, определить наилучшую модель и соответствующий набор переменных для прогнозирования банкротства компаний.*

Ключевые слова*: банкротство, дерево решений, машинное обучение, прогнозирование банкротства, искусственный интеллект.*

Дерево решений – это метод контролируемого обучения (то есть вы объясняете, что такое входные данные, и каковы соответствующие выходные данные в обучающих данных), который можно использовать как для задач классификации, так и для задач регрессии, но, в основном, данный метод предпочтительнее для решения задач классификации. Это классификатор с древовидной структурой, где внутренние узлы представляют функции набора данных, ветви представляют правила принятия решений, а каждый конечный узел представляет результат.

В дереве решений есть два узла, которые являются Узлом принятия решений и Конечным узлом. Узлы принятия решений используются для принятия любого решения и имеют несколько ветвей, тогда как конечные узлы

являются результатом этих решений и не содержат никаких дополнительных ветвей. Решения принимаются на основе особенностей данного набора данных.

Дерево решений – это графическое представление всех возможных решений проблемы / задачи на основе заданных условий. Данный метод называется деревом решений, потому что, подобно дереву, оно начинается с корневого узла, который расширяется на дальнейшие ветви и создает древовидную структуру. Дерево решений просто задает вопрос и, основываясь на ответе (Да / Нет), дополнительно разбивает дерево на поддеревья.

Приведенная ниже диаграмма объясняет общую структуру дерева решений:

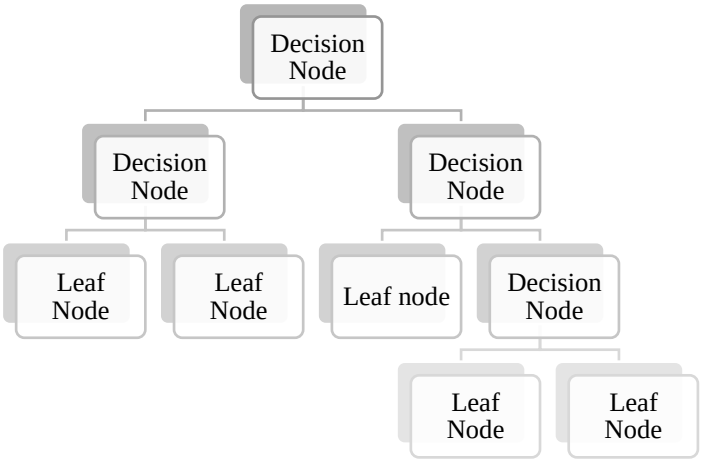

Рис. 1 Общая структура дерева решений

В дереве решений для прогнозирования класса данного набора данных алгоритм начинается с корневого узла дерева. Этот алгоритм сравнивает значения корневого атрибута с атрибутом записи (реального набора данных) и, основываясь на сравнении, следует за ветвью и переходит к следующему узлу. Для следующего узла алгоритм снова сравнивает значение атрибута с другими подузлами и движется дальше. Он продолжает процесс до тех пор, пока не достигнет конечного узла дерева. Полный процесс можно лучше понять, используя приведенный ниже алгоритм:

1) начните дерево с корневого узла, назовем его S, который содержит полный набор данных;

2) найдите лучший атрибут в наборе данных с помощью меры выбора атрибута (ASM);

3) разделите S на подмножества, содержащие возможные значения для наилучших атрибутов;

4) создайте узел дерева решений, который содержит наилучший атрибут;

5) рекурсивно создайте новые деревья решений, используя подмножества набора данных, созданного на шаге – 3. Продолжайте этот процесс до тех пор, пока не будет достигнута стадия, на которой вы не сможете дополнительно классифицировать узлы и называть конечный узел конечным узлом.

Алгоритм дерева решений всегда пытается максимизировать значение прироста информации, и узел / атрибут, имеющий наибольший прирост информации, разделяется первым. Прирост информации – это оценка изменений энтропии после сегментации набора данных на основе атрибута. Он вычисляет, сколько информации о классе предоставляет нам функция. В соответствии со значением прироста информации мы разделяем узел и строим дерево решений.

Теперь мы реализуем дерево решений с помощью Python. Для этого мы будем использовать набор данных из Тайваньского экономического журнала по компаниям за период с 1999 по 2009 годы. Данные были взяты из открытого банка данных UCI Machine Learning Repository. Используя тот же набор данных, мы можем сравнить классификатор дерева решений с другими моделями классификации, такими как KNN, LogisticRegression, Random Forest.

Ниже приведены этапы построения моделей:

1) этап предварительной обработки данных;

2) подгонка алгоритма дерева решений к обучающему набору;

3) прогнозирование тестового результата;

4) проверка точности результата;

5) визуализация результата тестового набора.

В качестве анализируемого набора данных была взята выборка, состоящая из более 7000 записей, где каждая запись – это описание характеристик компании, и целевая переменная – является компания банкротом (1) или нет (0).

Исходный набор данных был не сбалансирован и сильно смещен в сторону финансовой стабильности, так как в нем 220 компаний банкротов, и 6599

компаний не-банкротов. Если обучить модель на этом наборе данных, прогноз будет также смещен в сторону финансовой стабильности. Данные были сбалансированы с помощью метода SMOTE (Synthetic Minority Over-sampling Technique).

Соотношение оптимальных параметров модели было подобрано с помощью метода GridSearchCV, и оказалось следующим: DecisionTreeClassifier (criterion='entropy', max_depth=4, min_samples_leaf=4).

К преимуществам алгоритма дерева решений можно отнести следующие:

Дерево решений использует внутреннюю логику принятия решений и рассматривается как алгоритм белого ящика, то есть полученные знания из набора данных могут быть легко извлечены в читаемой форме, что не является особенностью алгоритмов черного ящика, таких как нейронная сеть. Это также ускоряет время обучения дерева решений.

Дерево решений следует непараметрическому методу: оно не зависит от распределения и не зависит от предположений о распределении вероятностей. Может работать с данными высокой размерности с превосходной точностью.

Деревья решений могут полностью выполнять выбор признаков или проверку переменных. Они могут работать как с геометрическими, так и с числовыми данными. Кроме того, они могут решать проблемы с несколькими результатами или выходами.

В отличие от других алгоритмов классификации, при использовании деревьев решений нелинейные взаимосвязи между параметрами не влияют на производительность деревьев.

Далее представлены итоговые оценки построенных моделей.

Таблица 1

Итоговая оценка моделей

№	*Algorithm*	*Model Score*	*Precision*	*Recall*	*F1 score*	*ROC-AUC score*
1	Random Forest Classifier	90.69%	0.27	0.90	0.42	0.90
2	DecisionTree Classifier	87.17%	0.21	0.86	0.33	0.87

3	K Nearest Neighbour	85.26%	0.15	0.65	0.25	0.75
4	Logistic Regression	82.26%	0.02	0.06	0.02	0.46

Решая проблему выявления банкротства компании, мы заинтересованы в максимизации показателя Recall. Так как чем выше данный показатель, тем меньше компаний, которые будут потенциальными банкротами, мы пропустим при анализе. При этом мы «переплачиваем» за то, что выявляем банкротов больше, чем их оказывается на самом деле. Но это можно считать определенным преимуществом алгоритма, так как будет проведена дополнительная проверка таких компаний, что очевидно не повредит, а даже наоборот, поможет выявить слабые места и предотвратить банкротство заблаговременно.

Таким образом, сравнивая результаты эффективности четырех моделей (KNN, Logistic Regression, Decision Tree, Random Forest), можно сделать вывод, что наилучшие показатели эффективности показали модели Random Forest и Decision Tree. Объясняется это тем, что данные модели отлично подходят для решения задач классификации, особенно в условиях, когда имеется большое количество переменных, связь между переменными не является линейной и необходимо проанализировать большое количество данных в исходном датасете.

Список литературы

1. Decision Trees // Scikit-learn [Электронный ресурс]. – Режим доступа: https://scikit-learn.org/stable/modules/tree.html#tree (дата обращения: 29.01.2022).

2. Forests of randomized trees // Scikit-learn [Электронный ресурс]. – Режим доступа: https://scikit-learn.org/stable/modules/ensemble.html#forest (дата обращения: 29.01.2022).

3. Taiwanese Bankruptcy Prediction Data Set // Machine Learning Repository [Электронный ресурс]. – Режим доступа: https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction# (дата обращения: 22.12.2021).

Иванькова Светлана Сергеевна – студентка, Институт магистратуры ФГБОУ ВО «Санкт-Петербургский государственный экономический университет», Санкт-Петербург, Россия.