

Секция «Искусственный интеллект и машинное обучение в цифровой экономике»

**Применение методов машинного обучения для выявления фактов
фальсификации финансовой отчетности**

Научный руководитель – Ларин Александр Владимирович

Елизарова Юлия Михайловна

Студент (магистр)

Национальный исследовательский университет «Высшая школа экономики» - Нижний Новгород, Факультет экономики, Кафедра экономической теории и эконометрики, Нижний Новгород, Россия

E-mail: elizarova-julia7@yandex.ru

На сегодняшний день применение технологий искусственного интеллекта затрагивает многие сферы жизни общества - образование, здравоохранение, маркетинг, логистику, управление человеческими ресурсами и др. Она из наиболее популярных областей для реализации алгоритмов машинного обучения - сфера экономики и финансов. На данный момент широкую популярность набирает автоматизация процессов оценки финансовой отчетности у аудиторских компаний при помощи создания систем, основанных именно на применении алгоритмов машинного обучения. Целью настоящей работы является тестирование наиболее популярных алгоритмов машинного обучения на основании данных финансовой отчетности и аудиторских заключений различных компаний, торгующихся на американских фондовых биржах. Тестирование основано на оптимизации гиперпараметров каждого из алгоритмов для выявления того, какой из них даст максимальную точность оценки при проверке на оптимальной комбинации гиперпараметров.

Существует достаточно большое количество исследований, посвященной проблеме выявления фактов фальсификации финансовой отчетности на основании некоторых показателей, являющихся так называемыми «красными флагами», которыми в эконометрических исследованиях часто называют необычно высокие показатели финансовой деятельности организации, низкий денежный поток, выполнение и перевыполнение годовых планов относительно выручки и прибыли и тд. [3]

Кроме того, основной проблемой всех исследований, рассматривающих именно методы машинного обучения, является отсутствие результатов реальных вычислительных экспериментов, в которых алгоритмы полностью адаптируются к имеющимся данным за счет оптимизации гиперпараметров и лишь после этого сравниваются между собой, что дает более точные результаты. [1], [4]

Задача выявления фактов мошенничества с отчетностью относится к подкатегории задач классификации, использующей обучение с учителем, где необходимо идентифицировать категориальные метки классов для каждого из образцов. Стоит отметить, что в нашей задаче будет присутствовать бинарная классификация, где 1 отвечает за положительный класс (отсутствие факта мошенничества), а 0 за отрицательный класс (наличие факта мошенничества).

В качестве объясняющих переменных будут использованы 34 показателя, оценивающих результаты финансово-хозяйственной деятельности компании, а также индексы динамики изменения показателей отчетности. Для выявления наилучшего алгоритма машинного обучения будет произведено сравнения результатов применения таких линейных методов, как метод ближайших соседей, дерево решений, метод опорных векторов, логистическая регрессия, а также ансамблевых методов - случайный лес, градиентный бустинг. Как известно, у каждого алгоритма существует определенный набор гиперпараметров, которые

могут быть заданными как по дефолту, так и найдены при помощи специальных оптимизационных алгоритмов. Следовательно, для сравнения нескольких алгоритмов машинного обучения необходимо оптимизировать гиперпараметры каждого из них, найти наилучшую комбинацию гиперпараметров, которая даст максимальное значение по определенной метрике качества классификации, и выбрать на основании этого тот алгоритм, который наилучшим образом описывает целевую зависимость.

Существует несколько различных методов оптимизации гиперпараметров моделей. Наиболее популярные среди них - это random search, grid search, а также оптимизация с использованием эвристических алгоритмов перебора. Последний способ является наименее ресурсоемким, так как в отличие от классического поиска по заданной сетке параметров он не использует все возможные комбинации гиперпараметров, а производит поиск внутри заданного гиперпараметрического пространства за указанное количество итераций, что позволяет существенно сократить время оптимизации. Для реализации описанного процесса оптимизации была использована библиотека Hyperopt. [2] В таблице 1 представлены результаты вычислений.

Таблица 1

Результаты тестирования алгоритмов

Алгоритм обучения

Гиперпараметр

Способ задания диапазона

Диапазон

Значение параметров в наилучшей комбинации

Precision

ROC-AUC

F1

DecisionTreeClassifier

max_depth

hp.choice

range(1,20)

11

0,725

0,634

0,845

max_features

hp.choice

range(1,5)

4

KNeighborsClassifier

n_neighbors

hp.choice

range(1,50)

31

0,893

0,78

0,901

p

hp.choice

[1, 2]

1

```
LogisticRegression
C
hp.loguniform
0.01, 1
1,493
0,926
0,741
0,833
SVC
C
hp.loguniform
0, 20
3,985
0,931
0,798
0,952
kernel
hp.choice
['linear', 'rbf', 'poly', 'sigmoid']
linear
gamma
hp.quniform
0, 20
5,95
RandomForestClassifier
max_depth
hp.choice
range(1,20)
17
0,851
0,67
0,834
max_features
hp.choice
range(1,5)
3
n_estimators
hp.choice
range(1,20)
13
criterion
hp.choice
['gini', 'entropy']
entropy
GradientBoostingClassifier
max_depth
hp.choice
range(1, 4, 1)
1
```

0,927
0,851
0,899
learning_rate
hp.quniform
0.01, 0.5, 0.01
0,45
subsample
hp.quniform
0.1, 1, 0.01
0,83

Сравнение результатов работы алгоритмов проводилось при помощи вычисления значений нескольких метрик качества на вложенной кросс-валидации (вычисление гиперпараметров производилось во внутреннем цикле, а оценка метрик на тестовых данных во внешнем цикле). По итогам алгоритм SVM показал наилучшее значение по нескольким метрикам, именно он подойдет для проверки финансовой отчётности на наличие фактов мошенничества.

Источники и литература

- 1) Apparao G. (2009) Financial Statement Fraud Detection by Data Mining. International // Journal of Advanced Networking and Applications. V. 01(№3), pp. 159-163
- 2) Bargstra J., Komer B., Yamis D. (2014) Hyperopt: A Python Library for Model Selection and Hyperparameter Optimisation // Proc. of the 13th Python in Science Conf.
- 3) Beneish M. (1999) The Detection of Earning Manipulation // Financial Analysts Journal. № 5. pp. 22—36.
- 4) Ravisankar P., Ravi V., Raghava G., Bose I. (2011) Detection of Financial Statement Fraud and Feature Selection using Data Mining Techniques // Decision Support Systems. №50. pp. 491-500