УДК 004.852

ОЦЕНКА ЧУВСТВИТЕЛЬНОСТИ МЕТРИК ЗАДАЧ КЛАССИФИКАЦИИ

Жуков Никита Романович, *студент 1 курса института управления и АПК, ФГБОУ ВО РГАУ−МСХА имени К. А. Тимирязева, e-mail: dayzwaitik123@yandex.ru*
Научный руководитель – Харитонова Анна Евгеньевна, *к.э.н, доцент, доцент кафедры статистики и кибернетики, ФГБОУ ВО РГАУ−МСХА имени К. А. Тимирязева, e-mail: kharitonova.a.e@rgau-msha.ru*

Аннотация. *Для задач классификации машинного обучения существует большое число метрик, характеризующих качество построенных моделей, при этом интерес представляет различия в их значениях при разном формировании исходных данных. В результате исследования получены данные, позволяющие сделать выводы о том, как сбалансированность входных категорий и доля тестовых данных влияют на такие метрики как accuracy и F1.*
Ключевые слова: *машинное обучение, метрики качества, accuracy, F1, задача классификации.*

Машинное обучение – это раздел информатики, позволяющий на основе данных обучить модель таким образом, чтобы можно было делать прогнозы для новых значений. При этом для моделей машинного обучения выбор правильной метрики имеет решающее значение при оценке.

Целью исследование было оценивание различий между метриками качества классификации при различных условиях формирования исходных данных. Достижение поставленной цели предполагает решение следующих задач:

1) выбрать метод классификации;
2) изучить основы его применения;
3) изучить и проанализировать метрики качества моделей классификации;
4) применить модель для классификации при равномерном распределении меток и оценить различия в метриках;
5) применить модель для классификации при неравномерном распределении меток и оценить различия в метриках;
6) сравнить полученные результаты;
7) сделать выводы.

В качестве метода классификации была выбрана логистическая регрессия, так как данная модель отлично подходит для решения поставленных задач. В основе ее применения лежит отнесение метки к какому-то

классу с определенной долей вероятности [1].

По результатам проверки модели выводится матрица ошибок. При этом матрица ошибок строится по новым данным, которые не участвовали в построении модели (тестовая выборка). Матрица ошибок представляет собой таблицу с 4 различными комбинациями прогнозируемых и фактических значений.

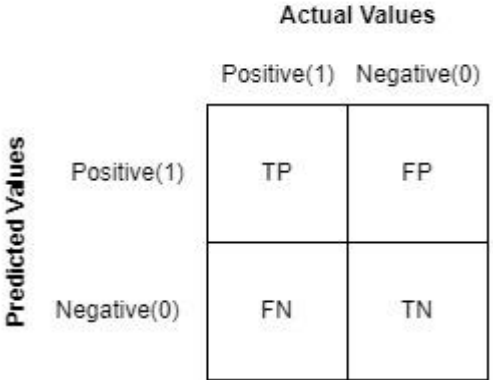

Рисунок 1 – Матрица ошибок

TP(True Positive) − число верно предсказанных положительных значений;

TN(True Negative) − число верно предсказанных отрицательных значений;

FP(False Positive) − число неверно предсказанных положительных значений;

FN(False Negative) − число неверно предсказанных отрицательных значений.

Используя матрицу ошибок, мы можем вывести такую *метрику* как accuracy. Эта метка показывает долю верных предсказаний. Следующие 2 метрики являются оценкой качества работы алгоритма для каждого из классов. *Precision* показывают долю объектов, названных классификатором положительными и при этом действительно являющимися положительными (precision=TP/ =(TP+FP)). *Recall* же показывает какую долю объектов положительного класса из всех объектов положительного класса нашел алгоритм(recall=TP/(TP+FN)). Метрика, объединяющая две предыдущие называется *F1*-метрикой [2].

Построение моделей, расчеты и графики получены и реализованы на языке программирования Python с использованием сторонних библиотек таких как numpy, pandas, sklearn, scipy, matplotlib. Данные для построения модели взяты с сайта kaggle [3]. Задачей предсказание является определение вышла ли компания в прибыль или нет на основе статистик акций в этот день (открытие, закрытие, минимум, максимум). Для начала модель строилась на данных, где распределение меток-классификаторов равномерно. Строилось по 200 моделей для разных соотношений тестовых и

тренировочных данных: 10, 20, 30, 40 % доля тестовых данных от всего набора. Для каждого набора предсказанных данных были рассчитаны метрики *F1-метрика* и *accuracy.*

При проверке каждого набора на нормальное распределение тест Шапиро-Уилка дал положительный результат. Так как все выборки метрик распределены нормально, был проведен t-тест, чтобы выяснить, различаются ли средние значения выборок. Тесты были проведены между выборками F1-score между 10 и 20 % test-size, 20 и 30 % test-size, 30 и 40 % test-size. Аналогично были проведены тесты со значениями метрик accuracy. В результате t-теста, проведенному между выборками соответствующих метрик, получены результаты, говорящие о статистически значимой разнице между средними в выборках.

В результате проведенного исследования по данным с равномерным распределением классов можно отметить:

1) с увеличением размера тестовой доли данных значения обеих метрик уменьшаются;

2) значение accuracy больше чем F1-score во всех проведенных испытаниях;

3) вариативность F1-score больше, чем у accuracy.

Следовательно, увеличение тестовой выборки уменьшает значения метрик, при чем F1-score изменяется больше, чем accuracy.

Далее модель строилась на данных, где метки-классификаторы распределены неравномерно. Построение моделей аналогично шагам, описанным выше. Выборки со значением метрик

Стоит отметить, что при проведении t-теста, не было выявлено статистически значимых различий между средних генеральных совокупностей выборок с 10 и 20 % test-size для метрик F1-score и accuracy.

По результатам проведенных исследований по данным с неравномерным распределением классов можно сделать следующие выводы:

1) при увеличении test-size значения метрик ведут себя не линейно;

2) значение accuracy больше чем F1-score;

3) вариативность обоих метрик крайне мала.

Сравнивая проведенные исследования можно отметить, что значения метрик accuracy выше F1-score в каждом исследовании, но в испытаниях с неравномерным распределением классов отличие между accuracy и F1-score больше. В исследовании с равномерным распределением классов значение метрики F1-score выше, чем во втором.

В целом можно сделать вывод, что при неравномерном распределении меток-классификаторов качество оценок моделей ниже и разность между метркиами больше. Таким образом, для оценок качества построенных моделей не следует опираться лишь на одну из метрик особенно при неравномерном распределении классификационных признаков. Оценка моделей должна вестись комплексно по всем показателям.

Библиографический список

1. **Рашка, С.** Python и машинное обучение: машинное и глубокое обучение с использованием Python, scikit-learn и TensorFlow 2 / С. Рашка, В. Мирджалили / 3-е изд.: Пер. с англ. − СПб. : ООО «Диалектика», 2020. − 848 с.

2. **Лабинцев, Е.** Метрики в задачах машинного обучения / Е. Лабинцев. Режим доступа: https://habr.com/ru/company/ods/blog/328372/ Дата обращения 01.10.2022).

3. **Kaggle** Apple revenue from 1980 to 2022 Режим доступа: https://www.kaggle.com/datasets/meerashareef/apple-revenue-from-1980-to-2022 (Дата обращения 01.10.2022).