После того как модель построена, ее требуется оценить. Для сравнения моделей по степени точности предсказаний используются метрики оценки. Эти метрики определяют типы прогностических ошибок и штрафуют за них по-разному.

Рассмотрим три оценочные метрики, используемые чаще всего. В зависимости от целей нашего исследования, для того чтобы избегать ошибок специфического

типа, могут быть разработаны даже новые метрики. В связи с этим перечень приводимых в этой книге оценочных метрик ни в коем случае нельзя считать исчерпывающим. В приложении D рассмотрены другие примеры метрик.

Метрики классификации.

Процент верных прогнозов. Простейшая мера точности прогнозирования — это доля достоверно правильных предсказаний. Вернемся к примеру с гастрономическими покупками из табл. 1 Мы можем выразить результаты задачи по предсказанию покупки рыбы в таком утверждении: Наша модель с точностью 90 % пред-

сказывает, будет ли покупатель брать рыбу. Хотя эта метрика не так сложна для понимания, она не дает представления о том, где именно происходят ошибки

прогнозирования. Таблица 4 Матрица неточностей показывает точность

предсказаний о покупке рыбы

Матрица неточностей. Матрица неточностей (confusion matrix) дает представление о том, где наша модель прогнозирования преуспела и где она потерпела неудачу.

Посмотрите на табл. 4 Хотя общая точность модели составляет 90 %, она гораздо лучше предсказывает не покупки, чем покупки. Мы также видим, что число про-

гностических ошибок равномерно (по 5) распределилось между ложноположительными (FP, false positives) и ложноотрицательными (FN, false negatives).

Разновидности прогностических ошибок могут иметь решающее значение. Ложноотрицательный результат в предсказании землетрясения (то есть землетрясения не

ожидалось, но оно произошло) обойдется куда дороже, чем ложноположительный (землетрясение ожидалось, но не случилось).

Метрика регрессии.

Корень из среднеквадратичной ошибки (Root Mean Squared Error, RMSE). Поскольку при регрессии используются непрерывные числовые значения, то ошибки

обычно измеряют количественно, как разницу между предсказанными и реальными значениями, распределяя штрафы и исходя из величины ошибки. Корень из

среднеквадратичной ошибки — это популярная метрика регрессии, особенно полезная в случаях, когда мы хотим избежать крупных ошибок: каждая из них возводится в квадрат, что усиливает значимость такой ошибки. Это делает метрику крайне чувствительной к резко отклоняющимся значениям, за которые она штрафует модель.

Валидация

Метрики не дают полной картины эффективности модели. Из-за переобучения (см. раздел 1.3) модели, хорошо себя показавшие на уже имеющихся данных, могут не

справиться с новыми. Чтобы этого избежать, мы всегда должны подвергать модели оценке, используя надлежащую процедуру валидации. Валидация (validation) — это оценка того, насколько хорошо модель предсказывает новые данные. Тем не менее вместо ожидания новых данных для проверки модели мы можем разбить наш текущий набор данных на два сегмента. Первый выступит в роли нашего обучающего набора данных (training dataset), а второй послужит заменой для новой информации в качестве тестового набора данных (test dataset) для оценки точности прогностической модели. Лучшей моделью признается та, которая дает самые точные предсказания на тестовом наборе. Чтобы процесс валидации был эффективен, мы должны выбирать элементы для обучающего и тестового набора данных случайно и беспристрастно. Однако если изначальный набор данных мал, мы не

можем позволить себе роскошь отложить их часть для формирования тестового набора, поскольку тогда пришлось бы пожертвовать точностью, которая снижается

от сокращения доступного объема данных.

По этой причине, вместо использования двух различных наборов данных для испытания одного набора проверкой другим, мы можем обойтись изначальным набором, устроив перекрестную проверку — кросс-валидацию. Кросс-валидация (cross-validation) позволяет полностью задействовать данные путем разделения их набора на несколько сегментов для поочередной проверки модели. За одну итерацию все сегменты, кроме одного, используются для обучения модели, которая сама проверяется на последнем сегменте. Этот процесс повторяется до тех пор, пока каждый сегмент не отработает в роли тестового (рис. 3).

Поскольку для предсказаний на каждой итерации использовались разные сегменты, их прогнозы могут разниться. Приняв во внимание эту вариативность, мы можем дать более здравую оценку действительным прогностическим способностям модели. А в качестве итоговой оценки точности модели принимают среднее значение за все итерации. Если результаты кросс-валидации показывают, что прогностическая точность нашей модели невысока, мы можем вернуться к настройке параметров или обработать данные иначе.

Метрики оценки различаются по тому, как они определяют различные типы погрешностей прогнозирования и как штрафуют за них. В этом приложении представле но несколько наиболее типичных метрик в дополнение к рассмотренным в разделе 1.4.

Метрики классификации

Площадь под ROC-кривой, AUROC. AUROC (Area Under the Receiver Operating Characteristic Curve) - 3то метрика, позволяющая выбирать между максимизацией доли истинно положительных результатов и минимизацией доли ложноотрицательных результатов. Доля истинно положительных результатов (TPR) это доля правильно определенных положительных результатов среди всех положительных: TPR = TP / (TP + FN). Доля ложноположительных результатов (FPR) это доля неправильно определенных отрицательных результатов среди всех отрицательных:

FPR = FP / (FP + TN).

В самом крайнем случае можно пойти по пути максимизации доли истинно положительных результатов (TPR=1), определяя все значения как положительные. Хотя это полностью убирает ложноотрицательные результаты, это также значительно увеличивает число ложноположительных. Другими словами, необходимо равновесие между минимизацией ложноположительных и максимизации истинно положительных результатов. Этот баланс может быть визуализирован на ROC-криво (рис. 1).

Эффективность модели оценивается с помощью площади, охватываемой ROC-кривой, поэтому метрика и называется площадью под кривой ошибок (AUC). Чем точнее модель, тем ближе кривая к верхней левой границе графика. Идеальная модель продемонстрировала бы кривую при AUC = 1, что эквивалентно всей площади графика. В противоположность ей эффективность модели со случайным прогнозом была бы представлена диагональной пунктирной линией при AUC - 0,5.

На практике мы можем определить лучшую модель по тому, что она захватывает большую площадь AUC, а ее ROC-кривая использовалась бы для того, чтобы определить подходящий порог TPR и FPR, с которыми мы готовы смириться.

Теперь, когда ROC-кривая позволила нам выбрать тип ошибки, которого мы больше всего хотим избежать, можно применить штрафы ко всем ошибочным пред сказаниям с использованием такой метрики, как логарифмическая функция потерь (logarithmic loss metric).Логарифмическая функция потерь. При работе с бинарными и категориальными переменными предсказания обычно выражаются в виде вероятности того, что покупатель купит рыбу. Чем ближе вероятность к 100%, тем увереннее модель в том, что покупатель купит рыбу. Логарифмическая функция потерь использует эту уверенность модели для того, чтобы штрафовать за ошибочные прогнозы: чем выше уверенность, тем больше штраф.

На рис. 2 показано, что штраф резко увеличивается, как только модель достигает определенной степени уверенности. Например, если модель предсказывает, что покупатель купит рыбу с вероятностью 80%, но оказалось, что он не купил, то штраф составит 0,7 единиц. Если же модель была уверена на 99 %, то штраф увеличивается до 2 единиц. Из-за своей способности соотносить штрафы с уверенностью модели при прогнозе логарифмическая функция потерь широко используется в случаях, где ошибочные прогнозы весьма критичны.

Метрики регрессии

Средняя абсолютная ошибка (Mean Absolute Error, MAE). ПРОСТОЙ способ оценки моделей регрессии заключается в том, чтобы штрафовать за все ошибки одинаково. вычислив среднее отклонение между предсказанными действительным значением для всех элементов данных. Эта метрика называется средней абсолютной ошибкой.

Корень из среднеквадратичной логарифмической ошибки (Root Mean Squared Logarithmic Error, RMSLE). В разделе 1.4 мы описали метрику корень из среднеквадратичной ошибки (RMSE), которая увеличивает штрафы за большие ошибки. Но помимо величины ошибки можно также принять во внимание и ее направление, используя метрику корень из среднеквадратичной логарифмической ошибки (RMSLE). RMSLE используется в случаях, когда мы хотим избежать недооценки больше, чем переоценки, например, при предсказании спроса на зонты в дождливый день. Недооценка приведет к недовольству покупателей и упущенной выгоде, в то время как переоценка означала бы только лишние запасы.

ПРОЕЦНТ ВЕРНЫХ ПРОГНОЗОВ

МАТРИЦА НЕТОЧНОСТЕЙ

ИЕТРИКА РЕГРЕСИИ – КОРЕНЬ ИЗ СРЕДНЕКВАДРАТИЧНОЙ ОШИБКИ

ВАЛИДАЦИ – КРОСС

ПЛОЩВДЬ ПОД ROC КРИВОЙ

ЛОГАРИФМИЧЕСКАЯ ФУНКЦИЯ ПОТЕРЬ

МЕТРИКА РЕГРЕССИИ СРЕДНЯЯ АБСОЛЮТНАЯ ОШИБКА И КОРЕНЬ ИЗ СРЕДНЕКВ ЛОГАРИФМ ОШИБКИ

P (positive) — сколько у нас положительных ответов (результатов наблюдений)

N (negative) — сколько отрицательных ответов

Теперь посмотрим на прогнозы модели. Здесь тоже есть positive и negative, но их сразу же делят на четыре группы:

все positive-ответы (модель сказала "да") делят на:

TP (true positive) — модель ответил "да" и угадала

FP (false positive) — модель ответила "да" и ошиблась

все negative-ответы делят аналогично:

TN (true negative) — модель ответил "нет" и это было правильно

FN (false negative) — "нет" и это ошибка

P = TP + FN

N = TN + FP

Accuracy

Такая метрика уже лучше, чем ничего, но всё же, она очень плоха. Даже в моём примере (хотя я не подгонял специально числа) видно, что, с одной стороны, разумные модели имеют высокую точность, однако, побеждает по точности просто самая пессимистичная модель. Допустим вы хотите предсказывать землетрясения (какое-то очень редкое явление). Ясно, что по этой метрике всегда будет побеждать модель, которая даже не пытается ничего предсказывать, а просто говорит всегда "нет". Те же модели, которые будут пытаться говорить когда-то "да", будут иногда ошибаться в позитивных прогнозах и сразу же терять очки.

Precision, recall. sensitivity, hit rate, or true positive rate (TPR)

В этой метрике мы рассматриваем только P-случаи: когда в реальных наблюдениях было "да". И считаем, какую долю из этих случаев модель предсказала правильно.

Все случаи "нет" мы отбрасываем.

Recall сам по себе довольно бесполезен. Взгляните на результаты для нашей модели: модель, которая всегда тупо говорить "да" — безусловно побеждает. Фактически, recall пропорционален TP, если P — константа (напомню, что это просто количество ответов "да" в наших фактических данных).

У recall есть брат-близнец:

Specificity, selectivity or true negative rate (TNR)

Здесь верны все те же самые оговорки. Специфичность, фактически, пропорциональна TN.

Важно, так же, заметить, что если T и P сильно отличаются (как в примере с землетрясениями), то сравнивать recall и специфичность надо очень осторожно.

Precision aka positive predictive value (PPV)

Какая часть наших предсказаний "да" действительно сбылась:

Недостатки этой метрики аналогичны: она вообще никак не учитывает предсказания "нет". Из наших результатов видно, что побеждает модель, которая почти всегда говорить "нет". Она как бы снижает риск проиграть, выводя большую часть своих ответ за рамки рассмотрения.

У этой метрики есть аналогичный близнец

Negative predictive value (NPV)

Какая часть "нет"-предсказаний сбылась.

Как вы уже видели, каждая из этих метрик рассматривает только какое-то подмножество предсказаний. Поэтому их эффективность очень сомнительна.

Однако, их очень часто используют для двух вещей:

Во-первых, по ним можно судить о характере модели: какая часть предсказаний ей даётся лучше, а в чём она слаба. Во-вторых, из этих метрик можно собирать что-то полезное.

Давайте задумаемся, а что значит "одна модель лучше другой"? Единого ответа тут нет. В нашем примере с долгоносиком всё зависит от наших приоритетов.

Если мы хотим ни в коем случае не потерять урожай, то нам надо максимизировать TP любой ценой. Фактически, в предельном случае, мы можем выкинуть любые модели и просто опрыскивать дерево химикатами всегда.

Если мы хотим минимизировать применение ядов, то нам надо максимизировать TN. В предельном случае, нам просто надо никогда не опрыскивать дерево: потеря урожая для нас не так страшна, как безосновательное применение ядохимикатов.

В реальной же жизни, мы ищем некоторый компромисс. Во многих случаях он может быть совершенно чётко сформулирован, с учётом цен на химикаты, стоимости урожая, репутационных потерь и прочего.

[20 популярных метрик машинного обучения. Часть 1. Метрики классификации и регрессионной оценки (machinelearningmastery.ru)](https://www.machinelearningmastery.ru/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce/)

[Основные метрики задач классификации в машинном обучении (webiomed.ru)](https://webiomed.ru/blog/osnovnye-metriki-zadach-klassifikatsii-v-mashinnom-obuchenii/?)

[Оценка качества работы модели, Метрики оценки качества, Кросс-валидация - Программа для иерархической классификации веб-сайтов (studbooks.net)](https://studbooks.net/1173544/informatika/otsenka_kachestva_raboty_modeli?)

[Выбор метрики в машинном обучении (datalytica.ru)](http://blog.datalytica.ru/2018/05/blog-post.html)

**Доп питон**

[3.3. Метрики и оценки: количественная оценка качества прогнозов - scikit-learn](https://scikit-learn.ru/3-3-metrics-and-scoring-quantifying-the-quality-of-predictions/?)

В популярной Python-библиотеке Scikit-learn есть модуль metrics, который можно использовать для вычисления метрик в матрице ошибок.

[Метрики Accuracy, Precision и Recall для проверки моделей машинного обучения (pythonru.com)](https://pythonru.com/baza-znanij/metriki-accuracy-precision-i-recall?)