**В задачах МО ставится вопрос оценки результатов моделей. Без введенных критериев невозможно оценить “успешность” модели/сравнить разные алгоритмы.** Хотя существуют некоторые общие рекомендации к применению метрик для некоторых задач, конечное решение лежит на плечах аналитика. Для сравнения моделей по степени точности предсказаний используются метрики оценки. Рассм. Классификацию - одна из главных проблем в BI - распознавание лиц, медицинская диагностика, классификация текста.

**Бинарная классификация**. Данные поделены на два класса. Метрики основаны на использовании исходов модели: **истинно положительные (TP), истинно отрицательные (TN), ложно положительные (FP) и ложно отрицательные (FN)**. Для удобства эти числа изображают в виде таблицы, которую называют матрицей ошибок.

#выявление подозрения на заболевание. Если у пациента оно есть, то это положительный класс. Результатом работы модели может быть определение – следует ли «заподозрить» у пациента диагноз (true) или нет (результат false). Если модель верно определила и поставила положительный класс, это истинно положительный исход, если ставит отрицательную метку класса, тогда ложно отрицательный. И наоборот.

P = TP + FN сколько положительных результатов наблюдений.

N = TN + FP сколько отрицательных.

Разновидности прогностических ошибок могут иметь решающее значение. Ложноотрицательный результат в предсказании землетрясения обойдется дороже, чем ложноположительный.

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

**Accuracy (точность).** Наиболее простая и известная метрика. Количество правильно проставленных меток класса (истинно положительных и отрицательных) от общего количества.

Метрика не дает представления о том, где именно происходят ошибки прогнозирования. Она не учитывает соотношения ложных срабатываний модели, что может быть критическим, например когда стоит задача распознать все истинные случаи (диагноза). В некоторых задачах необходимо определить всех пациентов с диагнозом и можно пренебречь ложно положительными исходами, так как они могут отсеяться на следующих стадиях (после сдачи анализов).

Также точность не является хорошим показателем эффективности модели в случае, когда распределение классов не сбалансировано (один класс встречается чаще). Тогда, даже если модель прогнозирует все выборки как наиболее частый класс, мы получим высокую степень точности, которая вообще не имеет смысла (тк модель ничего не изучает, а просто все предсказывает как высший класс). #диагностика редких заболеваний.

**Precision** (тоже точность) показывает количество истинно положительных исходов из всего набора положительных меток. Важность метрики определяется тем, насколько высока для рассматриваемой задачи «цена» ложно положительного результата. #Если стоимость дальнейшей проверки наличия заболевания высока и мы просто не можем проверить все ложно положительные результаты, то стоит максимизировать данную метрику. precision

**Recall** («полнота»/«чувствительность») определяет количество истинно положительных среди всех меток класса, которые были определены как «положительный». Необходимо уделить внимание этой оценке, если в поставленной задаче ошибка нераспознания положительного класса высока. #выставление диагноза ка смертельной болезни.

Итого: в задаче предсказания злокачественности опухоли точность показывает, сколько из определённых нами как злокачественные опухолей действительно являются злокачественными, а полнота – какую долю злокачественных опухолей нам удалось выявить.

**F1-Score**. В том случае, если Precision и Recall являются одинаково значимыми, можно использовать их среднее гармоническое.  
Помимо точечных оценок, существует ряд **графических методов**.

**ROC** – график, показывающий зависимость верно классифицируемых объектов положительного класса от ложно положительно классифицируемых объектов негативного класса. Соотношение True Positive Rate (Recall) и False Positive Rate.

TRP доля положительных объектов, правильно предсказанных положит = TP/P.

FPR доля отрицательных объектов, неправильно предсказанных положит = FP/N.

Идеальное значение графика находится в верхней левой точке (TPR = 1, a FPR = 0). При этом кривая, соответствующая FPR = TPR является случайным гаданием. С помощью ROC — кривой можно сравнить модели, а также их параметры для поиска наиболее оптимальной (с точки зрения tpr и fpr) комбинации. #ищется компромисс между количеством больных, метка которых была правильно определена как положительная и количеством больных, метка которых была неправильно определена как положительная.

**AUC**. В качестве численной оценки ROC кривой принято брать площадь под этой кривой. AUC обладает статистическим смыслом: это вероятность того, что случайно выбранный экземпляр негативного класса будет иметь меньше вероятность быть распознанным как позитивный класс, чем случайно выбранный позитивный класс.

#сотовый оператор хочет научиться предсказывать, будет ли клиент пользоваться его услугами через месяц. На первый взгляд кажется, что задача сводится к бинарной классификации с метками 1, если клиент останется с компанией и 0.

Однако если задуматься, то окажется, что такие метки практически бесполезны. Компании скорее интересно упорядочить клиентов по вероятности прекращения обслуживания и в зависимости от этого применять разные варианты удержания: кому-то предложить скидку на следующий месяц.

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.

**Мульти-классификация**. Если в задаче классов больше, чем два. Одним из возможных способов обобщения является вычисление среднего метрики по всем классам. В качестве «положительного» класса берется вычисляемый, а все остальные — в качестве «отрицательного».

**Регрессионные модели** используются для прогнозирования непрерывных целевых значений (прогнозирования цен на жилье, прогноза погоды). Метрики, используемые для оценки этих моделей, должны иметь возможность работать с набором непрерывных значений (с бесконечным количеством элементов).

#### **MAE (**Ср. абсолютная ошибка/отклонение**)** измеряет среднюю сумму абсолютной разницы между фактическим значением и прогнозируемым значением, те штрафует за все ошибки одинаково.

MAE более устойчив к выбросам, чем MSE. Основная причина в том, что в MSE путем возведения в квадрат ошибок, выбросы доминируют в окончательной ошибке и влияют на параметры модели.

**Mean Squared Error (MSE)** находит среднеквадратичную ошибку между прогнозируемыми и фактическими значениями.Самая популярная метрика, используемая для задач регрессии. Измеряет среднюю сумму квадратной разности между фактическим значением и прогнозируемым значением для всех точек данных. Чем меньше MSE, тем точнее наше предсказание. Оптимум достигается в точке 0. По сравнению с средней абсолютной ошибкой, подчеркивает влияние больших ошибок по сравнению с меньшими. Является дифференцируемым, что позволяет более эффективно использовать для поиска экстремумов с помощью математических методов.

**Root Mean Squared Error (RMSE).** Корень от MSE. Ее легко интерпретировать, поскольку он имеет те же единицы, что и исходные значения (в отличие от MSE). Также она оперирует меньшими величинами по абсолютному значению, что может быть полезно для вычисления на компьютере.

Корень из среднеквадратичной логарифмической ошибки (Root Mean Squared Logarithmic Error, RMSLE). Помимо величины ошибки можно также принять во внимание и ее направление. 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.

Метрики не дают полной картины эффективности модели. Из-за переобучения модели, хорошо себя показавшие на уже имеющихся данных, могут не справиться с новыми. Чтобы этого избежать, всегда должны н использовать надлежащую процедуру валидации.

Невозможно дать каких-то четких гарантий и определить, какая из метрик лучше, выбирать и отдавать предпочтение стоит лишь исходя из опыта своего и других исследователей, потому что каждая задача имеют свою специфику и приоритеты.

Язык R библиотека caret Classification And REgression Training