Цель линейной регрессии — найти линию, которая минимизирует ошибку прогноза всех точек данных.

Важным шагом в любой модели машинного обучения является оценка точности модели. Показатели среднеквадратичной ошибки, средней абсолютной ошибки, среднеквадратичной ошибки и R-квадрата или коэффициента детерминации используются для оценки производительности модели в регрессионном анализе.

Средняя абсолютная ошибка представляет собой среднее значение абсолютной разницы между фактическими и прогнозируемыми значениями в наборе данных. Он измеряет среднее значение остатков в наборе данных.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

Where,

$\hat{y}$ – *predicted value of y*

$\bar{y}$ – *mean value of y*

Среднеквадратическая ошибка представляет собой среднее значение квадрата разницы между исходным и прогнозируемым значениями в наборе данных. Он измеряет дисперсию остатков.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

Среднеквадратическая ошибка — это квадратный корень из средней квадратичной ошибки. Он измеряет стандартное отклонение остатков.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

Коэффициент детерминации или R-квадрат представляет долю дисперсии зависимой переменной, которая объясняется моделью линейной регрессии. Это безмасштабная оценка, т. е. независимо от того, малы или велики значения, значение R-квадрата будет меньше единицы.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Скорректированный R в квадрате — это модифицированная версия R в квадрате, и он скорректирован с учетом количества независимых переменных в модели, и он всегда будет меньше или равен R^2. В приведенной ниже формуле n — это количество наблюдений в данных. k — количество независимых переменных в данных.

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1}\right]$$

Различия между этими показателями оценки

1. Среднеквадратическая ошибка (MSE) и среднеквадратическая ошибка штрафуют за большие ошибки прогнозирования по сравнению со средней абсолютной ошибкой (MAE). Однако RMSE широко используется, чем MSE, для оценки эффективности регрессионной модели с другими случайными моделями, поскольку она имеет те же единицы измерения, что и зависимая переменная (ось Y).
2. MSE — это дифференцируемая функция, которая упрощает выполнение математических операций по сравнению с недифференцируемой функцией, такой как MAE. Поэтому во многих моделях RMSE используется в качестве показателя по умолчанию для расчета функции потерь, несмотря на то, что его сложнее интерпретировать, чем MAE.
3. MAE более устойчив к данным с выбросами.
4. Меньшее значение MAE, MSE и RMSE подразумевает более высокую точность регрессионной модели. Однако желательным считается более высокое значение R квадрата.
5. Квадрат R и скорректированный Квадрат R используется для объяснения того, насколько хорошо независимые переменные в модели линейной регрессии объясняют изменчивость зависимой переменной. Значение R Squared всегда увеличивается с добавлением независимых переменных, что может привести к добавлению избыточных переменных в нашу модель. Однако скорректированный R-квадрат решает эту проблему.
6. Скорректированный квадрат R учитывает количество переменных-предикторов и используется для определения количества независимых переменных в нашей модели. Значение скорректированного квадрата R уменьшается, если увеличение квадрата R на дополнительную переменную недостаточно значительно.
7. Для сравнения точности различных моделей линейной регрессии RMSE является лучшим выбором, чем R Squared.

Вывод

Следовательно, если сравнивать точность прогнозирования между различными моделями линейной регрессии (LR), то RMSE является лучшим вариантом, поскольку его легко вычислить и дифференцируемо. Однако, если в вашем наборе данных есть выбросы, выберите MAE, а не RMSE.

Кроме того, количество переменных-предикторов в модели линейной регрессии определяется скорректированным R в квадрате, и выберите RMSE, а не скорректированный R в квадрате, если вы хотите оценить точность прогноза среди различных моделей LR.