Działanie algorytmu:

1. Przygotowanie danych treningowych i testowych, które zawierają ciągi czasowe oznaczone jako zwykłe lub dane które są anomaliami.
2. Stworzenie sieci enkodera-dekodera z warstwami LSTM. Warstwa enkodera powinna przetwarzać ciągi czasowe na wektory, a następnie warstwa dekodera powinna tworzyć oryginalne ciągi na podstawie wektorów cech.
3. Dodawanie warstwy DROPOUT do sieci, aby zwiększyć jej zdolność do uogólniania i zmniejszyć nadmierne dopasowanie się do danych uczących (Tutaj należy pobawić się parametrami wejściowymi sieci).
4. Trenowanie sieć na danych treningowych. Minimalizowanie błędu rekonstrukcji ciągów czasowych przez dekoder. W czasie treningu należy monitorować błąd na danych testowych, aby móc wykryć, kiedy sieć zaczyna zbyt mocno dopasowywać się do danych treningowych. Fajnie jest dodać earlyStoper który sam wykryje przetrenowanie
5. Po zakończeniu treningu należy obliczyć błąd rekonstrukcji dla danych testowych i porównać go z progiem podanym w kodzie. Jeśli błąd jest większy niż próg, to można uznać, że ciąg czasowy jest anomalią.

Jeśli dostaliśmy satysfakcjonujące wyniki to koniec, a jeśli nie to należy:

Zmodyfikować próg, aby uzyskać dobrą wydajność wykrywania anomalii. Można to zrobić na podstawie miary F1-score.

1. Zapisać nauczoną sieć.
2. Ostatecznie należy przetestować wykrywanie anomalii na nowych, nieznanych danych, aby sprawdzić, czy algorytm działa poprawnie. Jeśli anomalie zostaną wykryte, należy na nie spojrzeć, aby upewnić się, że są one rzeczywiście poprawną anomalią, a nie błędem w danych lub w algorytmie.

Jak działa F1:

Miara F1 jest miarą oceny jakości wykrywania anomalii w algorytmie opartym na enkoderze LSTM z użyciem dropout. Miara ta jest wyznaczana jako średnia harmoniczna między precyzją (precision) a czułością (recall) i określa, jak dobrze algorytm radzi sobie z wykrywaniem anomalii.

W algorytmie zastosowana jest funkcja (detect\_anomalies), która na podstawie błędu rekonstrukcji obliczonego przez sieć enkodera-dekodera wyznacza anomalie. Wartość progu błędu rekonstrukcji, powyżej którego dane są uznawane za anomalie, jest ustalana na podstawie wartości threshold.

Miara F1 jest obliczona na podstawie ilości prawdziwie pozytywnych (TP), fałszywie pozytywnych (FP) i fałszywie negatywnych (FN) wykrytych anomalii. Wartości te są obliczane za pomocą funkcji sklearn.metrics.f1\_score z biblioteki scikit-learn.

Normalizacja danych:

Normalizacja danych wejściowych jest ważnym krokiem podczas szkolenia sieci LSTM, ponieważ pozwala na uniknięcie problemów związanych z różnicami w skali i rozkładzie danych. Aby znormalizować dane wejściowe do sieci LSTM, należy skorzystać z jednej z poniższych metod:

1. Normalizacja z-score: W tej metodzie każda wartość danych wejściowych jest przekształcana w wartość z-score, która określa, jak dana wartość różni się od średniej wartości i jej odchylenia standardowego. Można użyć tej metody, stosując następujący wzór:

X\_normalized = (X - mean(X)) / std(X)

gdzie X to wektor wartości danych wejściowych, a mean(X) i std(X) to odpowiednio średnia i odchylenie standardowe wektora X.

1. Normalizacja min-max: W tej metodzie każda wartość danych wejściowych jest przekształcana w wartość z przedziału [0, 1], co pozwala zachować informacje o proporcjach między wartościami. Można użyć tej metody, stosując następujący wzór:

X\_normalized = (X - min(X)) / (max(X) - min(X))

gdzie X to wektor wartości danych wejściowych, a min(X) i max(X) to odpowiednio najmniejsza i największa wartość wektora X.

W celu wykrywania anomalii w ciągu czasowym można zastosować technikę Dropout. Dropout jest techniką, która polega na losowym usuwaniu neuronów podczas treningu sieci, co pozwala uniknąć przeuczenia. Wartości wejściowe do warstwy Dropout powinny być przeskalowane do wartości z przedziału [0, 1] przy użyciu jednej z powyższych metod normalizacji, aby zapewnić, że losowe usuwanie neuronów nie zaburzy proporcji między wartościami.

Jak olabelować dane:

Clustering: Można zastosować algorymty grupowania, aby podzielić dane na grupy i przypisać etykiety dla każdej grupy. Grupy, które zawierają punkty danych uznawane za anomalię, mogą być oznaczone jako takie, a pozostałe grupy jako normalne.

Z wykorzystaniem metryk odległości: Można obliczyć odległość każdego punktu danych od średniej wartości w określonym przedziale czasowym, a następnie zastosować określony próg, aby przypisać etykietę anomali lub normal. Na przykład, punkty, które są odległe o więcej niż określoną wartość standardową od średniej wartości, mogą być uznane za anomalię.

