Modelling Answer

Salman Hanif - 13523056

Apa itu hold-out validation dan k-fold validation

Hold-out validation dan k-fold validation keduanya biasa digunakan ketike meengevaluasi kinerja dari sebuah model machine learning. Ini adalah metode untuk membagi dataset kepada bagian-bagian untuk nantinya digunakan untuk melatih dan mengevaluasi model.

Hold-out Validation

Pada metode hold-out validation, dataset langsung dibagi menjadi 2 bagian, yaitu data train dan test (Xtrain, ytrain, Xtest, ytest). Biasanya pembagian dilakukan secara acak, mana saja baris data yang masuk train/test dengan rasio data secara umum 70-80% untuk train dan 20-30% untuk test.

Kelebihannya pembagian data lebih mudah dilakukan dan cepat, tetapi di sisi lain kinerja model jadi tergantung pada bagaimana dataset dibagi secara acak tadi. Sebagian dataset yang tidak digunakan untuk train juga disayangkan.

K-Fold Validation

Pada metode ini, K adalah jumlah folds dari data. jadi maksudnya teknik ini membagidataset menjadi k bagian yang sama besar. Pada pelatihan modelnya, setiap iterasi, satu fold digunakan sebagai data test dengan k-1 sisanya sebagai data train. Proses ini akan diulangi sebanyak k kali, sehingga semua fold pernah mencoba jadi data test.

Hasil model akan lebih stabil dan tidak ada dataset yang terbuang karena tidak digunakan untuk train, hanya saja memerlukan waktu dan kompleksitas yang lebih besar.

Kondisi yang membuat hold-out/k-fold lvalidation ebih baik dari salah satunya

Secara teori, jelas K-fold yang memanfaatkan dataset dengan maksimal dapat dianggap lebih baik dari Hold-out. Tapi pada praktiknya, ada kondisi di mana

hold-out lebih baik.

Pada dataset yang jumlahnya sangat besar, dengan jumlah fitur sampai ratudan, baris sampai puluhan atau bahkan ratusan ribu. Jelas K-Fold akan memakan waktu yang jauh lebih lama, dan membutuhkan sumber daya komputasi yang sangat besar karena proses perulangannya. Tentu saja di situasi ini, pendekatan hold-out jauh lebih efektif. Data yang besar pun tidak dikhawatirkan menyebabkan model starving atau kekurangan data. Singkatnya, jika data besar, ingin cepat, dan pc kentang, hold-out lebih diutamakan.

Pada kondisi lain dengan dataset yang lebih kecil, jelas K-Fold mengungguli Hold-out karena potensi bias data tidak ada pada pada K-fold.

Apa itu Data Leakage

Sesuai namanya, data leakage adalah kebocoran data. Maksudnya, informasi dari data yang seharusnya tidak tersedia untuk model (terutama dari data validasi atau data uji) secara tidak sengaja bocor ke data latih **(**data validasi bocor ke data train), sehingga model jadi seperti melakukan praktik “menyontek” yaitu dia tau jawaban/prediksi pada data yang dia ketahui di data train.

Dampak data leakage terhadap kinerja model

Data leakage ini akan membuat model berkinerja baik secara artifisial pada proses validasi, tapi saat praktik di dunia nyata tidak akan sama dengan performa validasinya. Kayak orang nyontek, pas ujian nilainya bagus, tapi belum tentu bisa kerja :v.

Kinerja dan prediksi dari model yang data leakage, dia akan menyesuaikan dengan dataset yang sudah diketahui. Akibatnya, model jadi tidak benar-benar belajar dari pola data, tetapi lebih ke menghapal. Nanti ketika diberikan data baru, model akan kesulitan melakukan prediksi yang tepat/akurat karena memang model jadi gagal menangkap pola data.

Solusi untuk mengatasi permasalahan data leakage

Model yang sudah di-train dengan data leakage akan sulit untuk diperbaiki, lebih baik buat model baru dengan data baru. Data yang sudah bocor pun sulit untuk ditangani. Tentunya istilah **mencegah lebih baik daripada mengobati** tepat di sini, jadi untuk mengatasi permasalahan data leakage dalam machine learniing, SOP atau pipeline selama pengolahan data adalah yang harus diperkuat.

Pisahkan data dengan benar, perlakukan data validasi sebagai data suci yang tidak boleh disentuh selama melatih model. Lakukan pra-permosesan data di lingkupnya sendiri, tidak mengganggu data validasi.