

Regression Tree

1. Cara kerja Algoritma

Secara garis besar, Regression Tree menggunakan struktur Tree yang memiliki cabang dan daun untuk membagi-bagi data. Regression Tree membagi data berdasarkan threshold yang ditentukan. Jika kurang dari threshold, data masuk bagian cabang kiri, jika lebih maka data masuk bagian cabang kanan. Pembagian terus dilakukan hingga semua data termasuk dalam sebuah kumpulan data yang terdefinisi sebagai daun.

Regression Tree memiliki beberapa variasi dalam penentuan threshold, saya menggunakan cara memilih threshold dari data unik pada dataset lalu dicari threshold yang menghasilkan nilai loss paling kecil.

Prediksi suatu nilai dilakukan dengan menelusuri pohon hingga daun yang fiturnya bersesuaian didapat. Jika ada lebih dari satu data pada daun, nilai prediksinya adalah rata-rata dari data pada daun.

2. Perbandingan Hasil Evaluasi Model

Model yang saya buat memiliki score yang sedikit lebih jelek dan waktu yang jauh lebih lama daripada model oleh scikit learn. Perbedaan waktu kalkulasi sebagian besar disebabkan oleh penggunaan dense array pada model saya sehingga ketika data masukan memiliki banyak fitur (one hot encoding dengan banyak kategori misalnya) proses kalkulasi akan memakan waktu jauh lebih banyak. Sedangkan model dari scikit learn menggunakan sparse matriks yang menghemat memori ketika matriks banyak memiliki angka nol (one hot encoding misalnya), kalkulasi diselesaikan jauh lebih cepat.

Perbedaan score disebabkan oleh perbedaan hyperparameter lanjutan, mengingat bahwa regression tree bisa memiliki banyak sekali hyperparameter yang tidak diterapkan di model saya.

```
Custom R2 Score: -0.6319391509616221
Sklearn R2 Score: -0.6357035013083547
Custom Mean Squared Error: 60718.48063415026
Sklearn Mean Squared Error: 60858.5383277803
```

3. Improvement

Untuk mempercepat waktu kalkulasi, model perlu menggunakan sparse matriks. Selain itu, penambahan hyperparameter yang lebih bervariasi juga bisa menambah fleksibilitas model dan mempermudah hyperparameter tuning untuk mendapatkan score yang lebih baik.