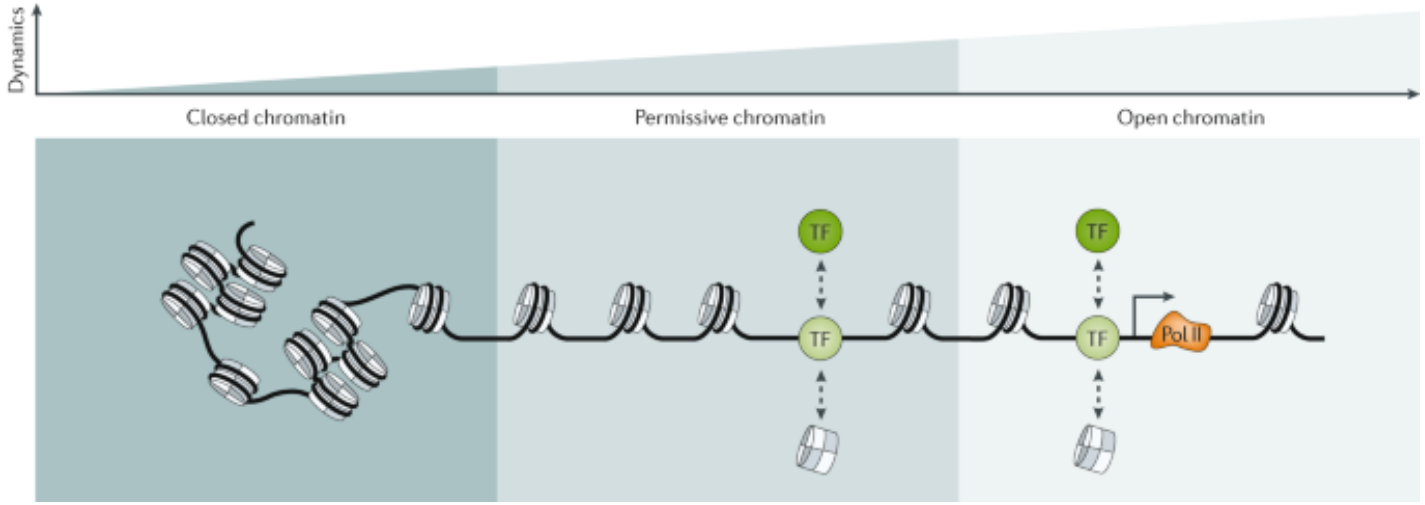
An interpretable classification of OCRs

Justin Bellavance

What's the problem?

(ATGACTAGCTACAGTGTACGA) -> does chromatin hide it?



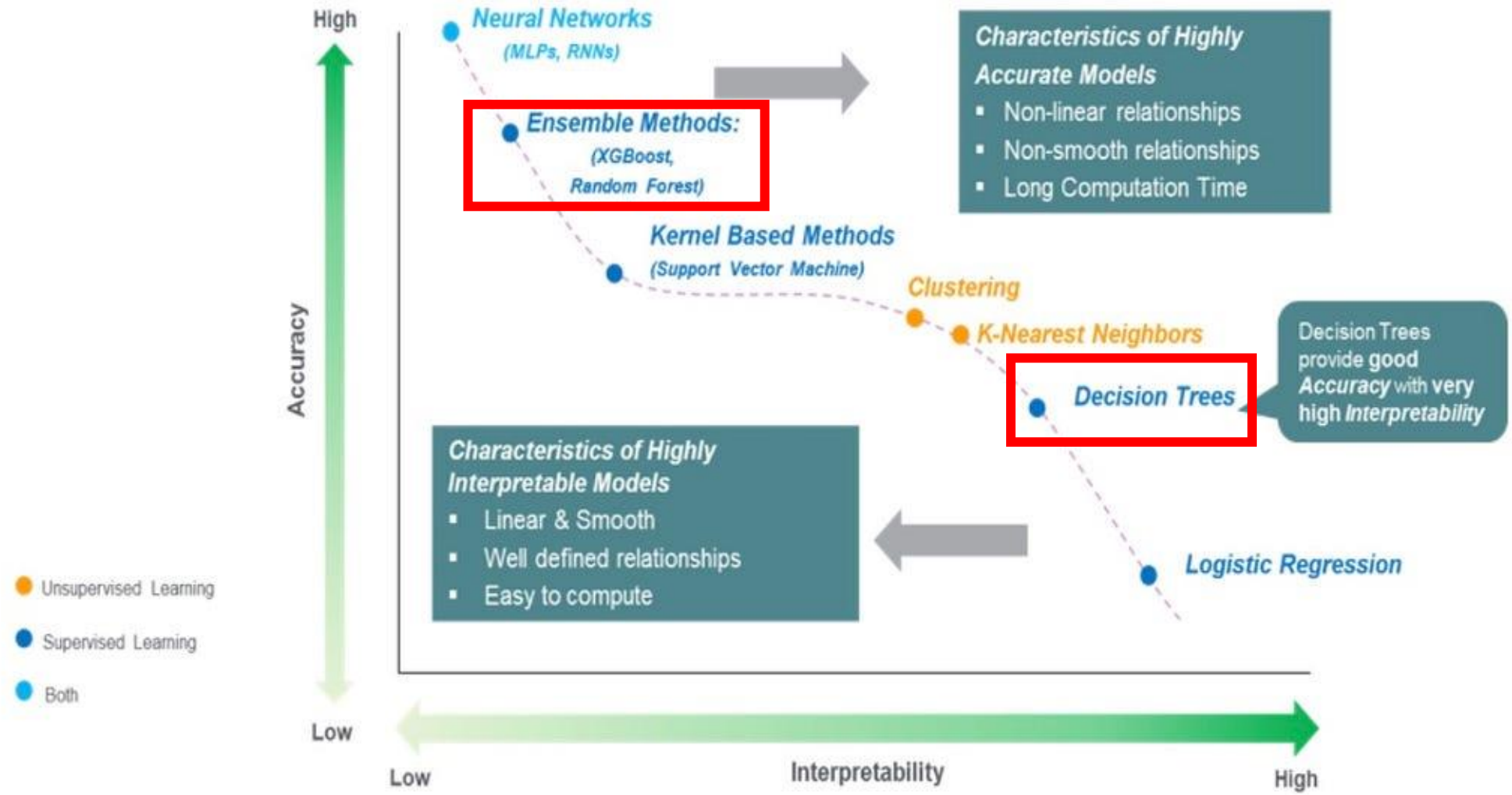
Very limited information based solely on DNA sequence

A lot of supplementary information available… but takes time to call

Deep learning can easily pick up on 'hidden' patterns without external help

- Perfect tool!


Deep Learning



by Sharayu Rane

Also I have no GPU :(

Goal: Beat Deep Learning Benchmark (not SOTA)

Method	Accuracy	F1
Pytorch CNN	68	66.1
TensorFlow CNN	68.8	72

https://github.com/ML-Bioinfo-CEITEC/genomic_benchmarks/blob/main/experiments/README.md#hyenadna

Methods (run 1)

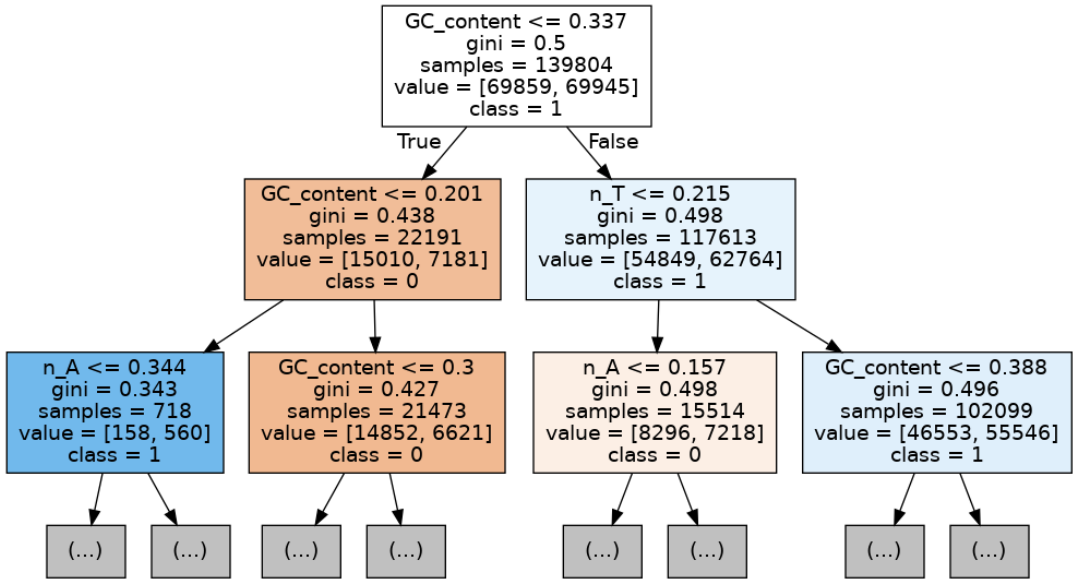
Try to think of all information I can possibly get from 4 letter alphabet DNA strands:

- A, C, T, G content
- Strand length
- CG%
- N content
- Contains start codon (ATG)
- Contains stop codon (TAG|TAA|TGA)

Use Decision tree then XGBoostClassifier (decision forest)

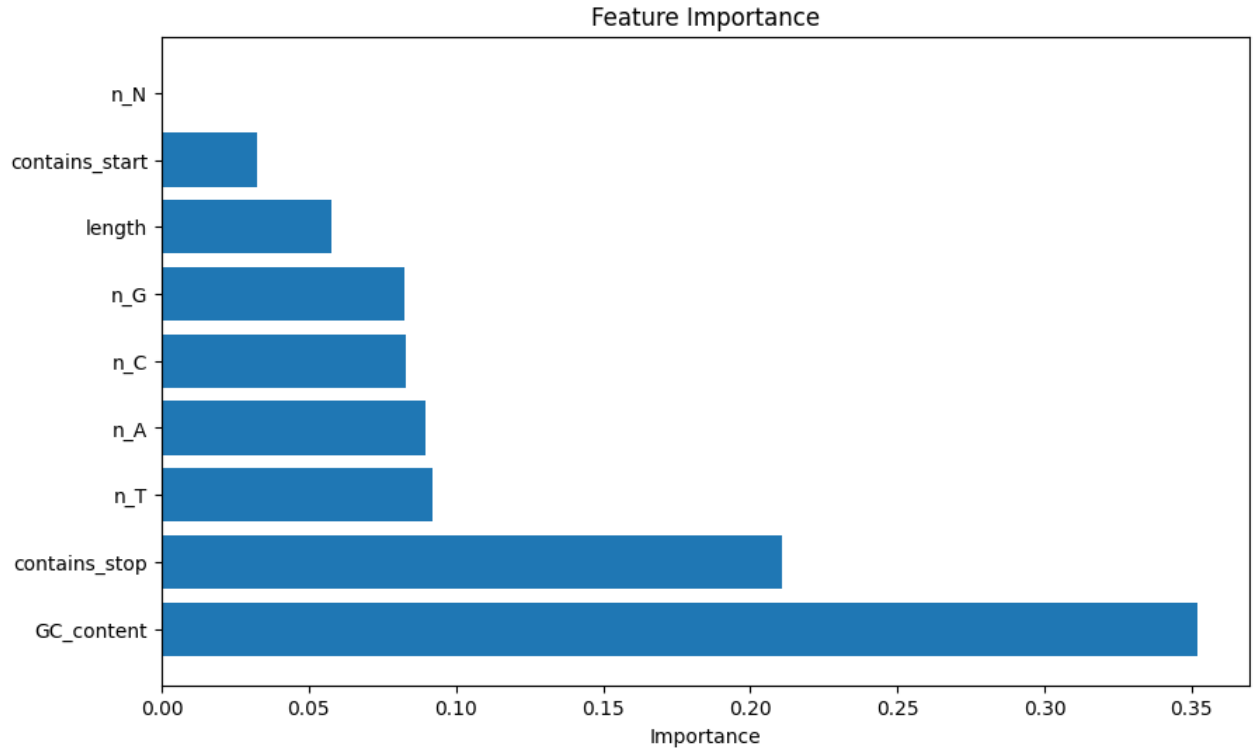
Results (Run 1)

Decision Tree



58.9% accuracy

XGBoost



60.5% accuracy

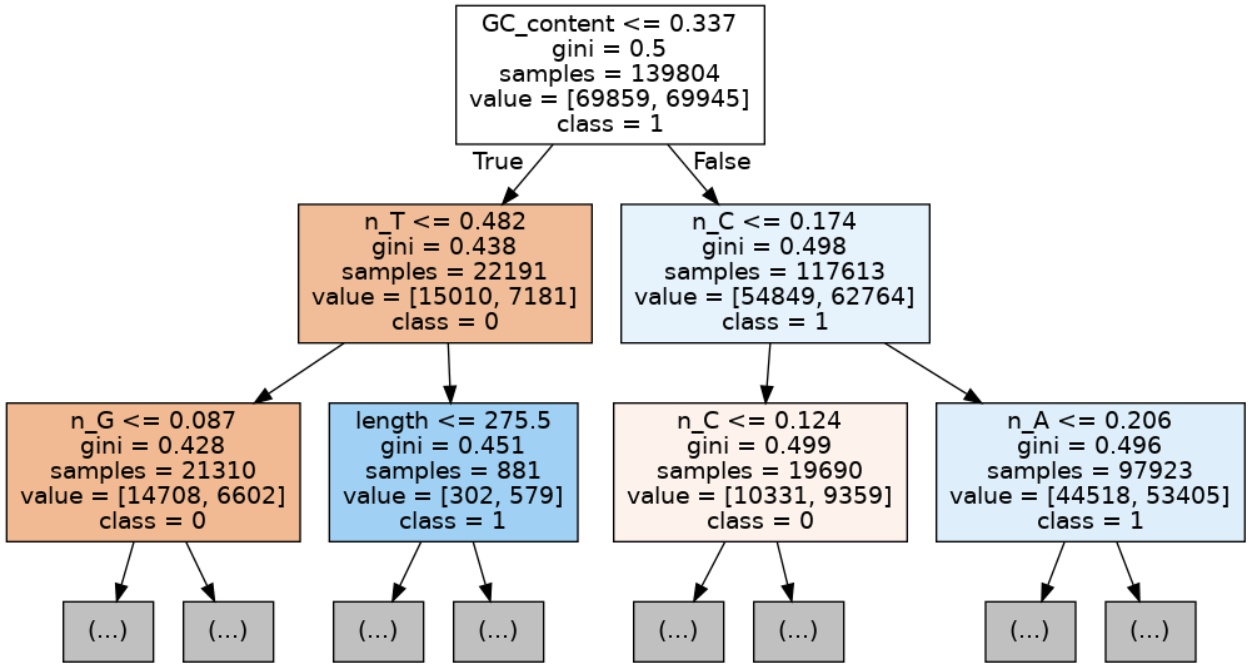
Methods (after feature engineering)

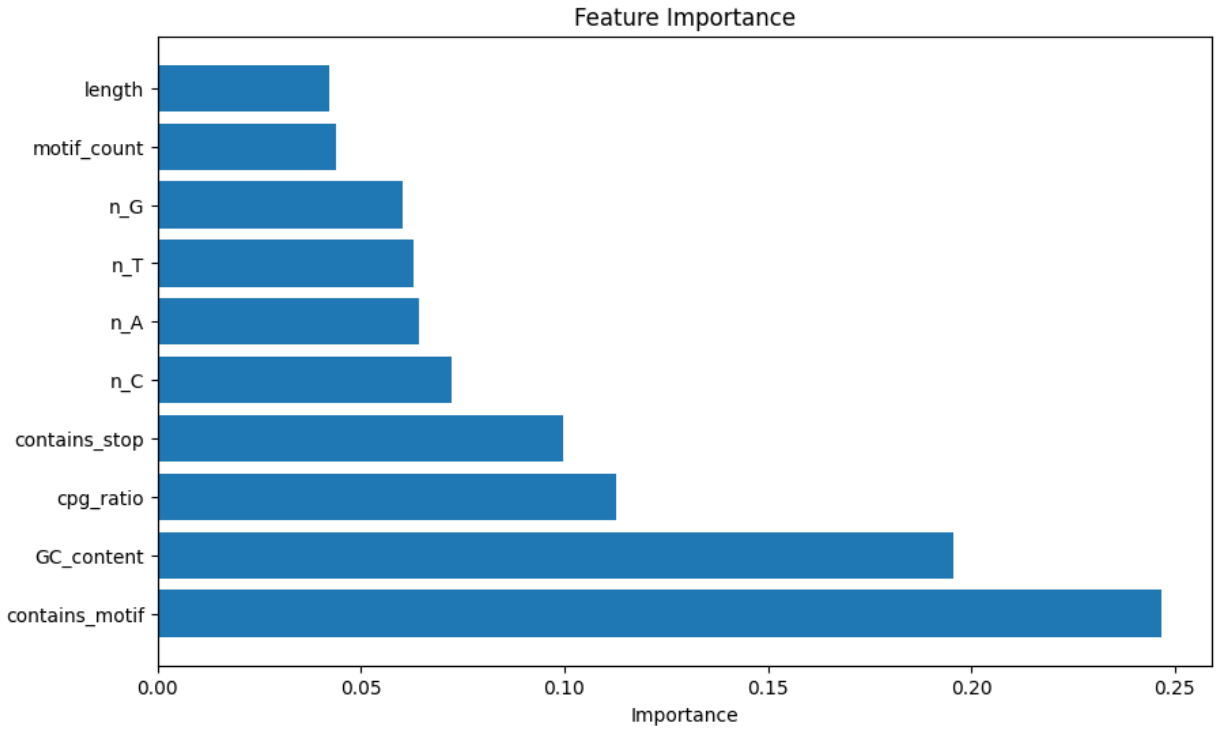
Parameters Now

- GC content
- CpG islands
- Length
- # of TF binding motifs (from JASPAR)
    - (99% contained a motif)

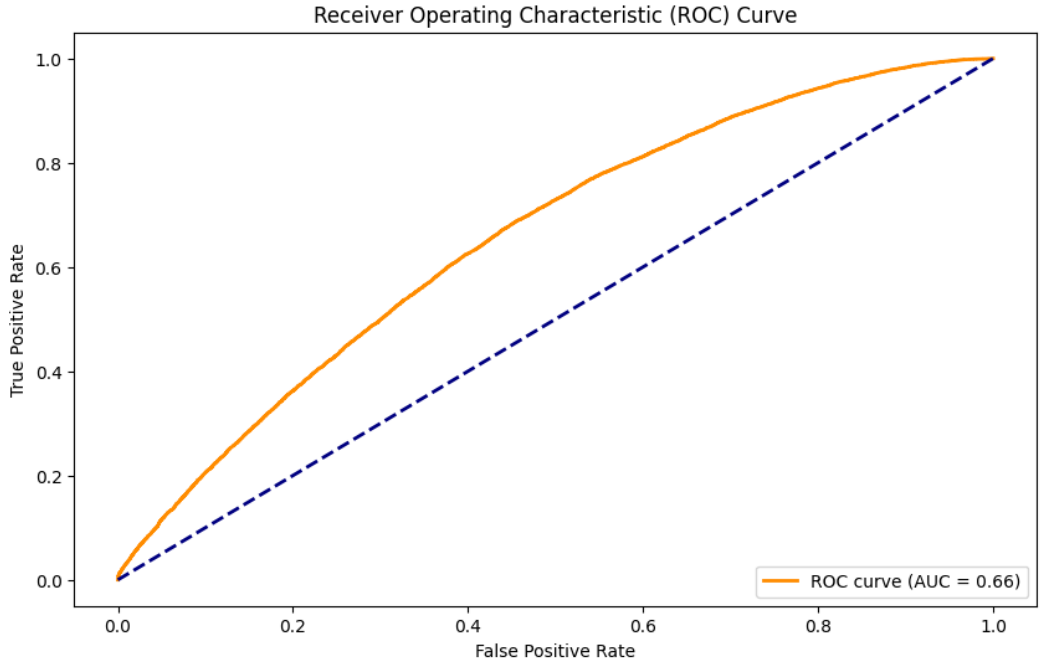
BayesSearchCV for hyperparameter tuning

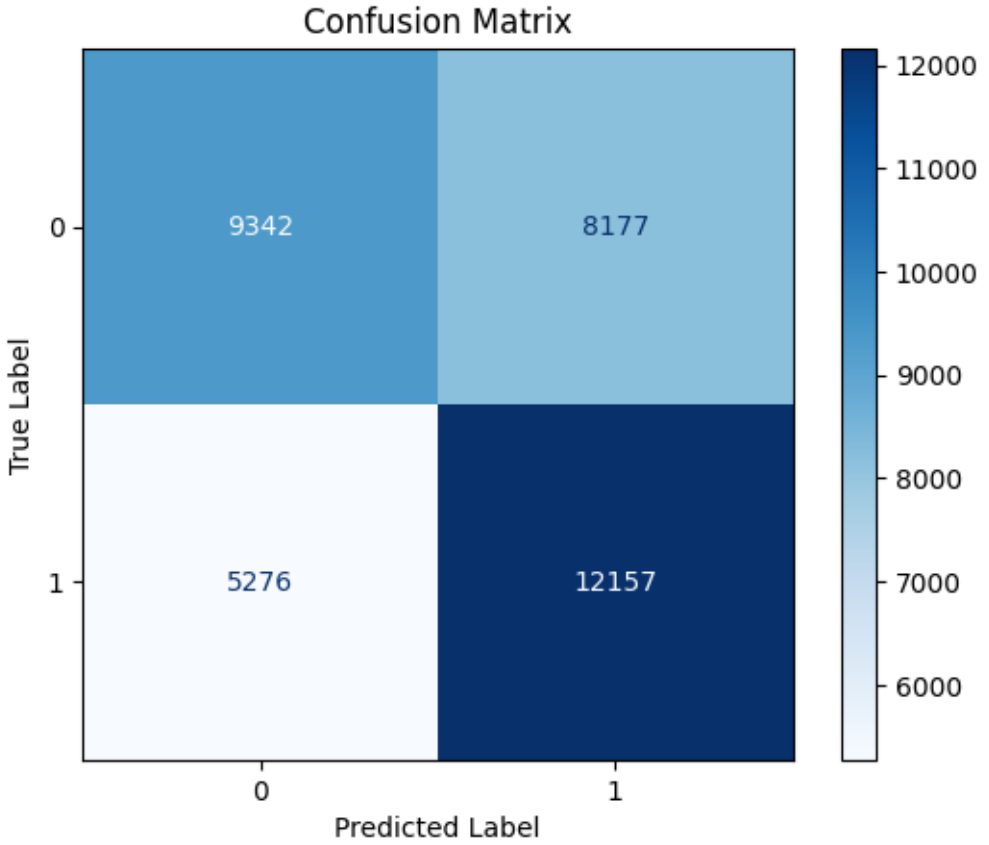
result


GC_content <= 0.337
gini = 0.5
samples = 139804
value = [69859, 69945]
class = 1
True
False
n_T <= 0.482
gini = 0.438
samples = 22191
value = [15010, 7181]
class = 0
n_C <= 0.174
gini = 0.498
samples = 117613
value = [54849, 62764]
class = 1
n_G <= 0.087
gini = 0.428
samples = 21310
value = [14708, 6602]
class = 0
length <= 275.5
gini = 0.451
samples = 881
value = [302, 579]
class = 1
n_C <= 0.124
gini = 0.499
samples = 19690
value = [10331, 9359]
class = 0
n_A <= 0.206
gini = 0.496
samples = 97923
value = [44518, 53405]
class = 1
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

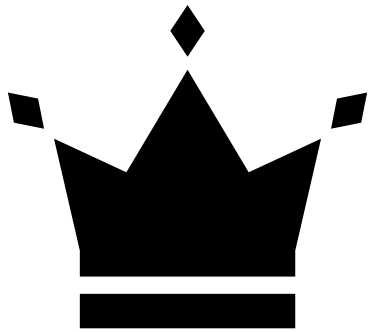

Feature Importance
length
motif_count
n_G
n_T
n_A
n_C
contains_stop
cpg_ratio
GC_content
contains_motif
0.00
0.05
0.10
0.15
0.20
0.25
Importance

result


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.66)


Confusion Matrix
True Label
Predicted Label
9342
8177
5276
12157
0
1
12000
11000
10000
9000
8000
7000
6000

Method	Accuracy	F1
Pytorch CNN	68	66.1
TensorFlow CNN	68.8	72
XGBoostClassifier	61.5	NA



Deep Learning