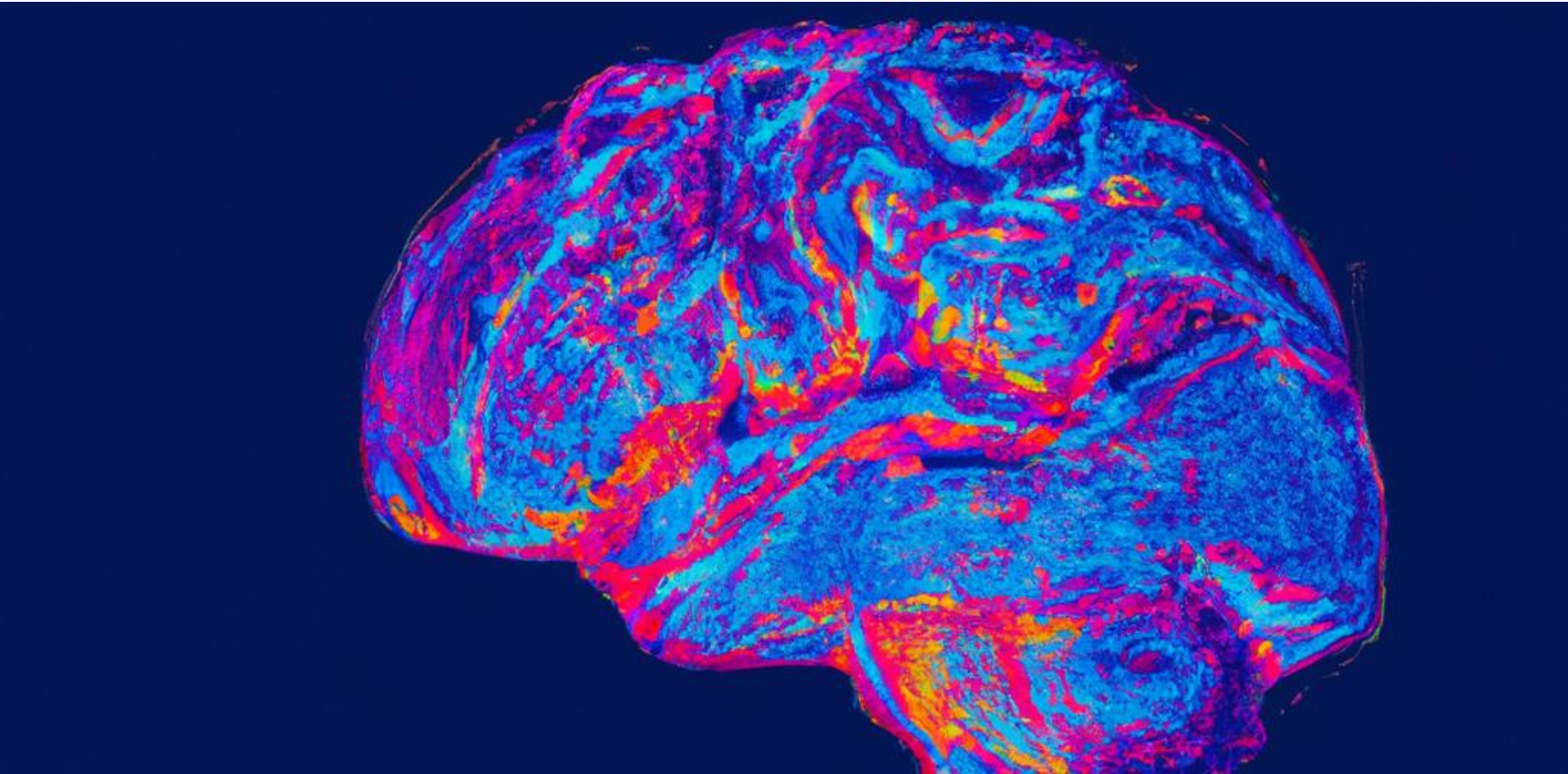
DETECT ONLINE PAYMENT FRAUD



ONLINE PAYMENT FRAUD

UNDERSTANDING THE DATASET

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	\
0	1	2	9839.64	C1231006815	170136.0	160296.36	
1	1	2	1864.28	C1666544295	21249.0	19384.72	
2	1	4	181.00	C1305486145	181.0	0.00	
3	1	1	181.00	C840083671	181.0	0.00	
4	1	2	11668.14	C2048537720	41554.0	29885.86	

FEATURE VECTOR

TARGET VARIABLE

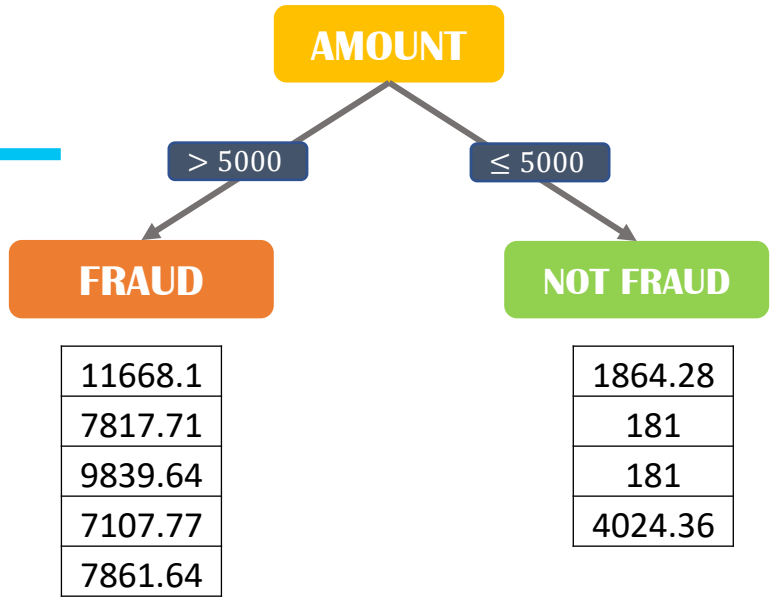
	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	M1979787155	0.0	0.0	No Fraud	0
1	M2044282225	0.0	0.0	No Fraud	0
2	C553264065	0.0	0.0	Fraud	0
3	C38997010	21182.0	0.0	Fraud	0
4	M1230701703	0.0	0.0	No Fraud	0

THEORY

WEAK LEARNERS

STUMP

amount
9839.64
1864.28
181
181
11668.1
7817.71
7107.77
7861.64
4024.36



11668.1
7817.71
9839.64
7107.77
7861.64

1864.28
181
181
4024.36

MATHS FOR STUMPS

$$Prediction = \log\left(\frac{Number\ of\ True\ Classifications}{Number\ of\ False\ Classifications}\right)$$

$$Probability = \frac{e^{\log(Pred)}}{1 + e^{\log(Pred)}}$$

WEAK LEARNERS

$$Y \sim Stump_1$$

$$Y \sim Stump_1 + Stump_2$$

$$Y \sim Stump_1 + Stump_2 + Stump_3 + Stump_4$$

$$Y \sim w_1 Stump_1 + w_2 Stump_2 + w_3 Stump_3 + w_4 Stump_4 \cdots$$

STUMPS HAVE DIFFERENT WEIGHTS

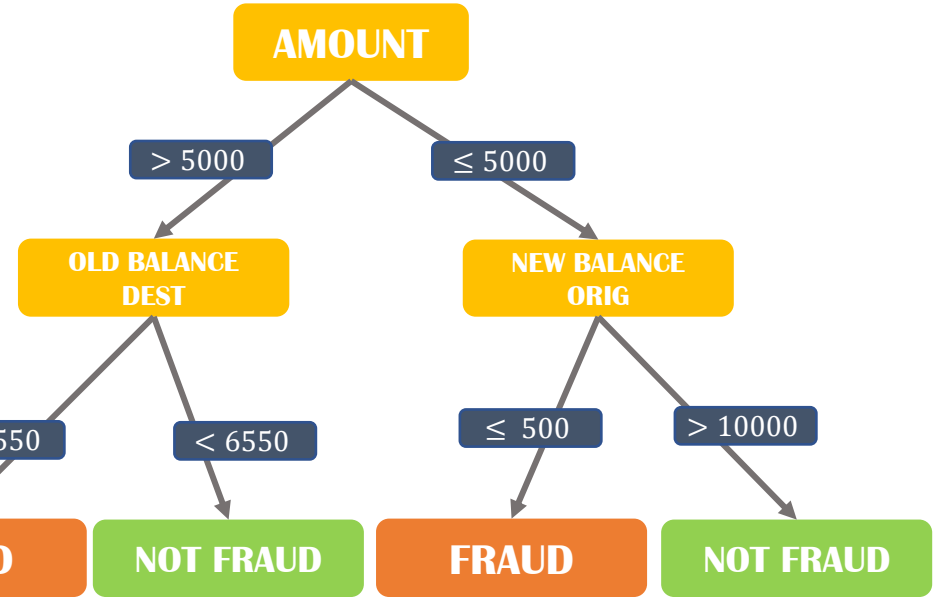
STUMPS ARE MADE USING PREVIOUS STUMPS' MISPREDICTIONS

REDUCE MISPREDICTIONS = ADJUSTING WEIGHTS

DECISION TREES

STUMP TREE

ADJUSTMENT TREE



GRADIENT BOOSTING DECISION TREES

FOREST OF TREES

$$Y \sim Stump + Tree_1 + Tree_2 + Tree_3$$

$$Y \sim Stump + LR_1 Tree_1 + LR_2 Tree_2 + LR_3 Tree_3 + \cdots$$
$$where\ 0 < LR < 1$$

STARTING TREE IS A STUMP TREE

NEXT TREES GENERATED USING MISCLASSIFICATIONS

ITERATE WITH A LEARNING RATE UNTIL CONVERGENCE

ABOUT MISCLASSIFICATIONS

$$Prediction = \log\left(\frac{Number\ of\ True\ Classifications}{Number\ of\ False\ Classifications}\right)$$

$$Probability = \frac{e^{\log(Pred)}}{1 + e^{\log(Pred)}}$$

$$New\ Probability = \frac{\sum Residual_i}{\sum[Previous\ Probability_i \cdot (1 - Previous\ probability_i)]}$$

GRADIENT BOOSTING DECISION TREES

START WITH A STUMP

GENERATE A STUMP TREE

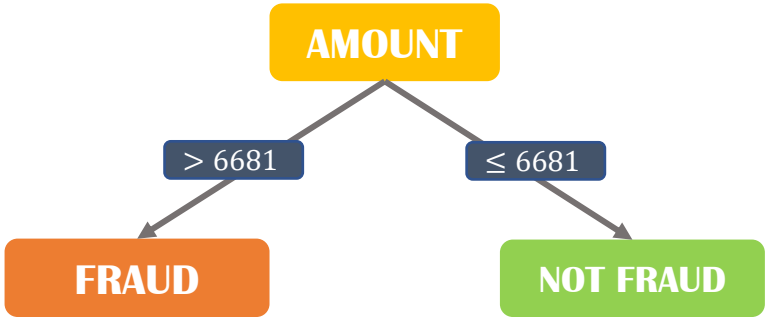
GENERATE TREE FROM MISPREDICTIONS OF STUMP TREE

GENERATE A TREE FROM PREVIOUS MISPREDICTIONS

REPEAT UNTIL
- MISPREDICTIONS ARE NEGLIGIBLE
- REACHED DESIRED NUMBER OF TREES

LGBM - IMPROVEMENTS

HISTOGRAM BASED SPLIT POINT SELECTION



amount
9839.64
1864.28
181
181
11668.1
7817.71
7107.77
7861.64
4024.36
5337.77

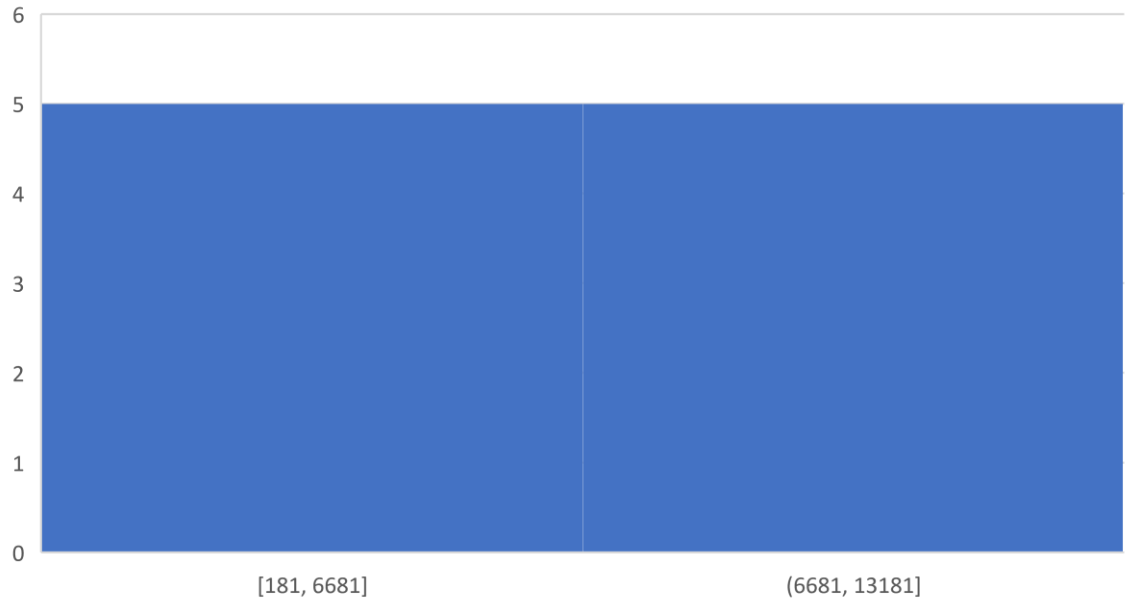
CONSTRUCTING HISTOGRAM

$$O(data_{count} * features_{count})$$

ITERATING TO OPTIMALITY

$$O(bins_{count} * features_{count})$$

BINS



EXCLUSIVE FEATURE BUNDLING

amount
9839.64
1864.28
0
181
11668.1
0
0
7861.64
0
5337.77

OldBalanceOrig
0
0
1200
0
0
1352.3
1292.1
0
7861.64
0

EFB
9839.64, 0
1864.28, 0
0, 1200
181, 0
11668.1, 0
0, 1352.3
0, 1292.1
7861.64, 0
0, 7861.64
5337.77, 0

CONSTRUCTING HISTOGRAM

$$O(data_{count} * exclusivefeatures_{count})$$

CODE

KAGGLE CHALLENGE

University of Southampton
Artificial Intelligence Society

Building a ML Model using Random Forest

Mon, Nov 7th 6pm

B100/3023