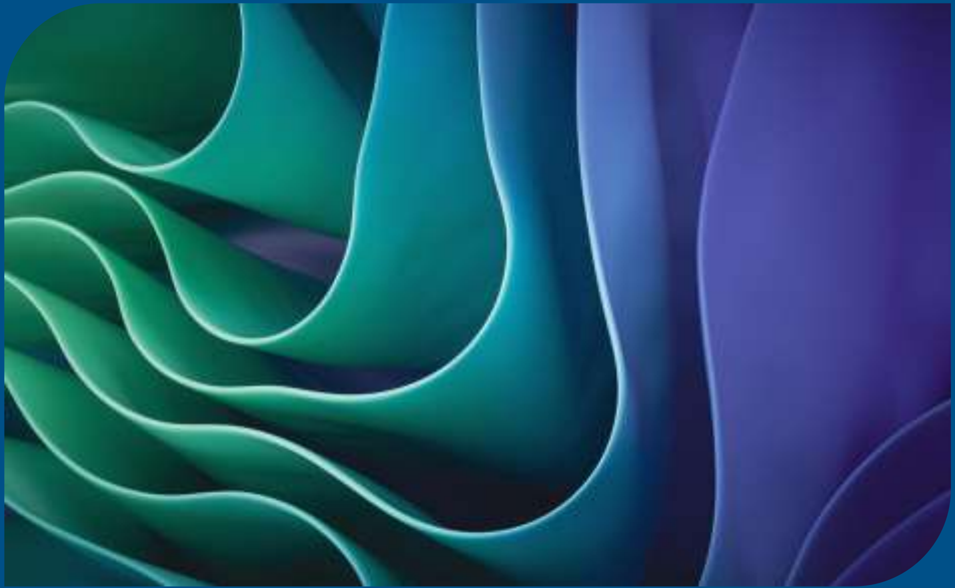
A Cost-Sensitive Learning Approach for Handling Imbalanced Data in Decision Trees

Aidan Lee, Sami Saleh
Period 4 ML 1



1. What is the Problem?
2. Solution
3. Dataset
4. Methods
5. Results/Conclusion

1. What is the Problem?
2. Solution
3. Dataset
4. Methods
5. Results/Conclusion

Class imbalance is a big issue when it comes to classification using decision trees. How can we combat this problem?

1. What is the Problem?
2. Solution
3. Dataset
4. Methods
5. Results/Conclusion

Cost-Sensitive Learning

- Cost-sensitive Learning
  - Assigns different "costs" to misclassification errors
- Cost-matrix with a cost-sensitive Gini Index algorithm for Decision Trees

COST MATRIX

		Predicted class	
		Positive	Negative
Actual class	Positive	$C(+,+)$	$C(-,+)$
	Negative	$C(+,-)$	$C(-,-)$

1. What is the Problem?
2. Solution
3. Dataset
4. Methods
5. Results/Conclusion

Original Dataset: Credit Card Fraud Detection

- 284,807 instances (transactions)
  - 492 frauds (0.172%)
  - ~577:1 Class Ratio
- 30 Features
  - 28 features “V1, V2, …V28”: normalized and unknown info to protect privacy
  - “Time”: How long after first instance transaction happened
  - “Amount”: How much the transaction was
- “Class”: 0 (non-fraudulent) or 1 (fraudulent)

Preprocessing

- Removed “Time”: not valuable info
- Discretized “V1-V28” and “Amount”: Three bins
- Numerical to Nominal in WEKA
- 70-30 train-test split

1. What is the Problem?
2. Solution
3. Dataset
4. **Methods**
5. Results/Conclusion

Methods

0	1
500	0

Cost of 500 for classifying 1 as 0 and 1 for classifying 0 as 1. Others are 0 since they are correct classifications

- Decision Tree
- Gini Impurity
- 0 means node is pure, optimal is lowest Gini Index
- Log Computations in InfoGain are expensive
- Incorporate SMOTE (Strategic Minority Oversampling Technique)
  - Better majority-minority class ratio (20:1 vs 577:1)

$$GiniIndex = 1 - \sum_j (p_j^2) \quad \text{where } p_j \text{ is the probability of class j.}$$

$$Cost\ Sensitive\ Gini = \sum_{i=1}^{n} p_i * (\sum_{j \neq i} Cost(i, j) * (1 - p_i))$$

Cost(i, j) is the cost of misclassifying an instance of class i as class j

1. What is the Problem?
2. Solution
3. Dataset
4. Methods
5. **Results/Conclusion**

Precision, Recall, and F1-Score

Accuracy is not sufficient since dataset is very large and imbalanced

- Precision: $Precision = \frac{TP}{TP + FP}$
- Recall: $Recall = \frac{TP}{TP + FN}$
- F1-Score: $F1 = \frac{2*Precision*Recall}{Precision+Recall}$

Table: Results

Model Type	Precision	Recall	F1-Score
DT	0.713	0.755	0.733
CSL-DT	0.701	0.701	0.701
DT w/ SMOTE	0.974	0.862	0.915
CSL-DT w/ SMOTE	0.974	0.862	0.915

Conclusion

- Cost-sensitive Decision Trees performed similar or even worse than regular Decision Trees
- F1-Score of 0.915 suggests potential for cost-sensitive learning
- Model showed overfitting, with higher precision, recall, and F1-Score in training than testing
  - Likely caused by limiting max depth to reduce computation time
- Future Work: Explore different cost values, test on multiple imbalance datasets, reduce overfitting

Works Cited

Aznar, Pablo. “Decision Trees: Gini vs Entropy ⋆ Quantdare.” *Quantdare*, 13 Dec. 2020, quantdare.com/decision-trees-gini-vs-entropy/.

Brownlee, Jason. “Cost-Sensitive Decision Trees for Imbalanced Classification.” *MachineLearningMastery.Com*, 20 Aug. 2020, machinelearningmastery.com/cost-sensitive-decision-trees-for-imbalanced-classification/.

Mienye, Ibomoiye, and Yanxia Sun. “Performance Analysis of Cost-Sensitive Learning Methods with Application to Imbalanced Medical Data.” *Informatics in Medicine Unlocked*, Elsevier, 3 Aug. 2021, www.sciencedirect.com/science/article/pii/S235291482100174X#bib28.

Thai-Nghe, Nguyen. *Cost-Sensitive Learning Methods for Imbalanced Data*, www.ismll.uni-hildesheim.de/pub/pdfs/Nguyen_et_al_IJCNN2010_CSL.pdf. Accessed 28 Jan. 2025.