Østfold University College

Poker Hand Induction: Multi-class classification of extreme imbalanced data with decision trees

Thomas Angeland, Kim-Andre Engebretsen & Mustafa Numic | Master in Applied Computer Science

21.02.2019

Problem statement

Train a classification tree to accurately assess poker hands of 5 cards from a 52 card deck.

Some info

- 52 unique cards in a deck.
- Each hand dealt contains 5 cards.
- 10 distinct, ranked poker hands.
- When order matters, There are 311 875 200 unique poker hands.
- 1 Royal Flush for every 649 740 hand dealt.

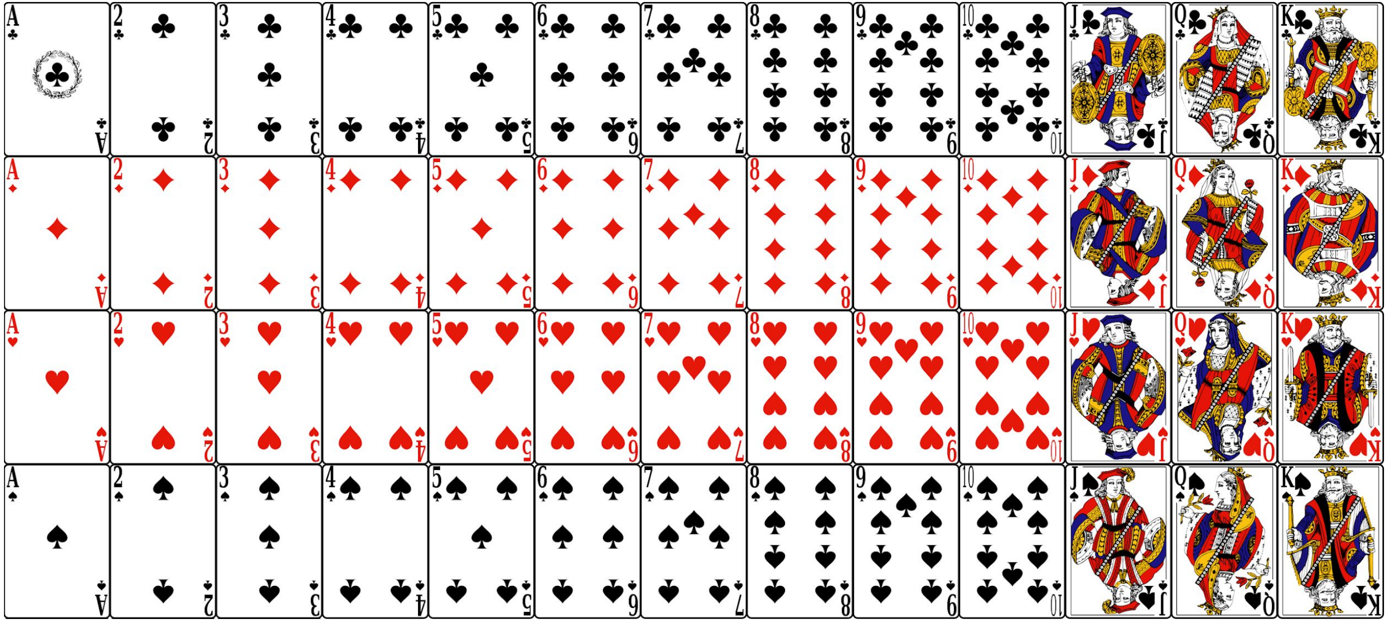
HAND	PROBABILITY	COMBINATIONS
Royal flush	0.00000154	480
Straight flush	0.00001385	4,320
Four of a kind	0.00024010	74,880
Full house	0.00144058	449,280
Flush	0.00196540	612,960
Straight	0.00392465	1,224,000
Three of a kind	0.02112845	6,589,440
Two pair	0.04753902	14,826,240
Pair	0.42256903	131,788,800
Nothing	0.50117739	156,304,800

Østfold University College

Cards (Predictors)

13 ranks

4 suits



Østfold University College

Poker hands (Classes)

10♥	J♥	Q♥	K♥	A♥	9: Royal Flush
5♠	6♠	7♠	8♠	9♠	8: Straight Flush
A♥	A♣	A♦	A♠	2♥	7: Four of a kind
A♠	A♦	A♣	K♥	K♠	6: Full House
2♥	4♥	6♥	8♥	K♥	5: Flush
5♥	6♣	7♦	8♠	9♥	4: Straight
A♣	A♥	A♠	2♦	7♣	3: Three of a kind
K♦	K♣	Q♥	Q♠	J♦	2: Two pair
A♣	A♥	9♠	8♦	7♣	1: One Pair
A♥	8♠	6♦	4♣	2♥	0: Nothing in hand

Dataset

Title: Poker Hand Data Set

Abstract: Purpose is to predict poker hands

Source: https://archive.ics.uci.edu/ml/datasets/Poker+Hand

Robert Cattral (cattral@gmail.com)

Franz Oppacher (oppacher@scs.carleton.ca)

Carleton University, Department of Computer Science

Intelligent Systems Research Unit

1125 Colonel By Drive, Ottawa, Ontario, Canada, K1S5B6

Data Set Characteristics:	Multivariate
Attribute Characteristics:	Categorical, Integer
Associated Tasks:	Classification
Number of Instances:	1025010
Number of Attributes:	11
Missing Values	0
Date	2007-01-01

Dataset

In a hand with *n* cards, each card is denoted by a S[n] (suit, 1-4) value and C[n] (rank, 1-13) value.

CLASS is assigned a value between 0 to 9 and represents how good the hand is.

Example:

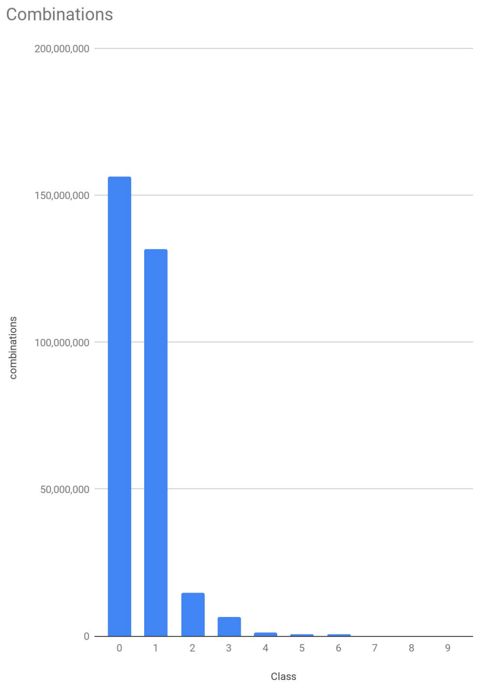
S1	C1	S2	C2	S3	C3	S4	C4	S5	C5	CLASS
3	8	1	7	3	3	2	4	3	11	0
3	9	4	12	4	6	2	10	2	3	0
2	11	2	6	2	5	4	7	3	2	0
4	4	3	13	3	6	2	7	1	12	0
2	12	4	4	2	6	2	2	4	8	0
4	4	4	12	3	2	2	13	4	6	0
3	7	3	13	3	11	1	3	2	2	0
4	11	3	12	3	7	1	7	4	3	1
4	3	4	1	2	2	3	11	2	3	1
2	8	2	1	4	12	4	10	1	12	1
4	10	2	12	4	1	2	1	3	9	1
2	4	1	5	2	10	4	11	4	10	1

Dataset

Problem: Class distribution is very imbalanced.

Attempted to solve this by;

- generating a new dataset programmatically;
- extracting a stratified sample from the generated dataset;
- using under- and over sampling;
- using early stopping, i.e using evaluation set; and
- using sample weights.



Preprocessing

The final solution does not use any methods of preprocessing (yet).

- *Scaling*, *centering* and *Box-Cox* in R yielded worse result than with no pre-processing.
- For similar reasons, PCA was not used, and we also observed 100% usage of predictors in both C5.0 and XGBoost.

Østfold University College

Preprocessing

“The SMOTE approach can improve the accuracy of classifiers for a minority class.”

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2011). SMOTE: Synthetic Minority Over-sampling Technique. Journal Of Artificial Intelligence Research. https://doi.org/10.1613/jair.953

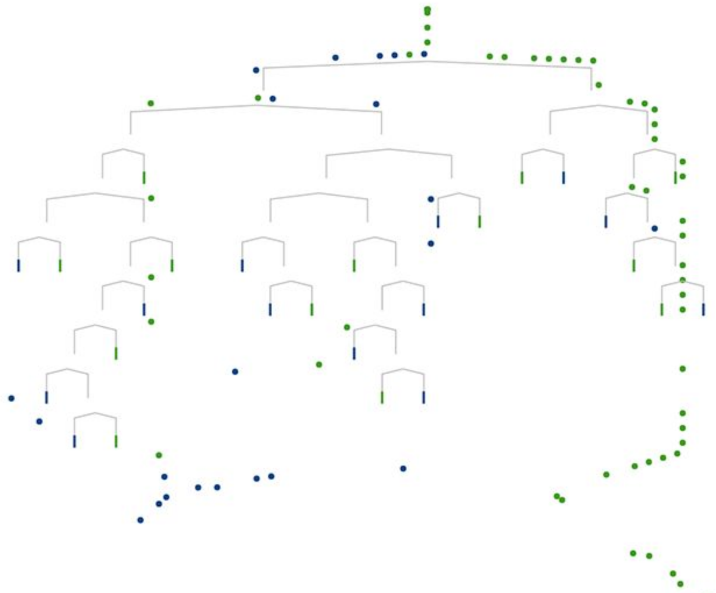
It didn’t for this problem.

Neither oversampling, SMOTE, undersampling or other custom sampling methods yielded any improvements in our testing.

More testing is needed in order to fully disclose it.

Classifiers

- C5.0
- AdaBoost
- RandomForest
- XGBoost
- LightGBM
- CatBoost



Global parameters

- Sample size: 2% stratified sample of population, 6 237 504 unique poker hands.
- Sample distribution:
  - 20 % train set (0.4 % of population, 1 247 500)
  - 10 % validation set (0.2 % of population, 623 750)
  - 70% test set (1.4 % of population, 4 366 252)
- Training with sample weights
- Parameters:
  - Number of trees / iterations: 250-2000
  - Depth: 4-10
  - Learning rate: 0.23-0.8
  - Loss function: multiclass log loss, MultiClassOneVsAll
- Increasing tree depth and iterations usually yielded better accuracy, but worse geometric mean, meaning the modell overfitted the dominant classes.

Østfold University College

Evaluation metrics

Accuracy is not a good indicator for multi-classification on imbalanced data. Geometric mean is.

G_{mean} reflects the ability to classify positive samples and negative samples at the same time.

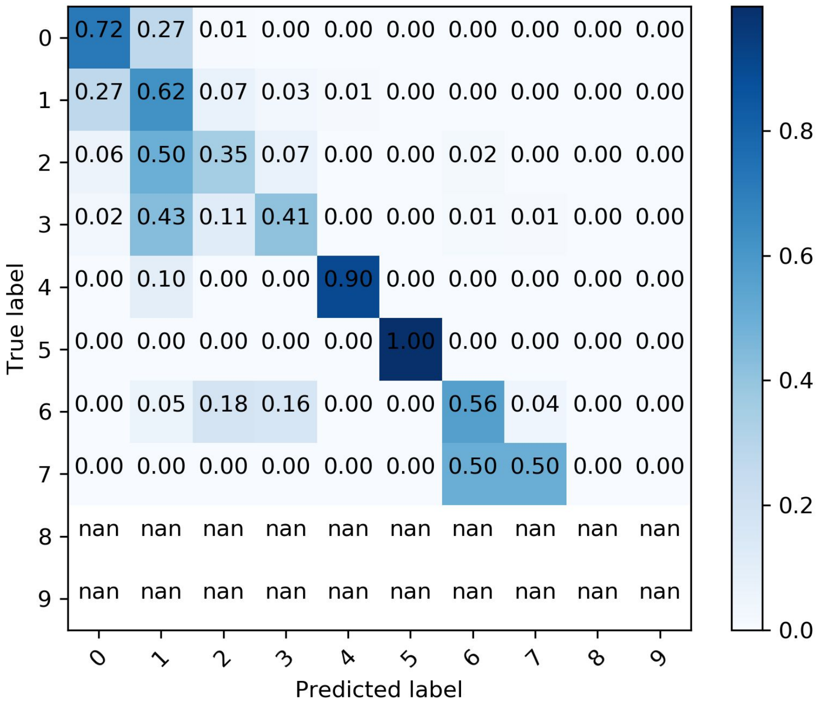
$$G_{mean} = \sqrt{R_{t+} \cdot R_{t-}},$$

where R_{t+} represents true positive rate, which is calculated by $R_{t+} = N_{TP} / N_{P}$. R_{t-} represents true negative rate, which is formulated as $R_{t-} = N_{TN} / N_{N}$.

The higher the geometric mean, the better ability of the classifier to recognize positive class and negative class samples at the same time.

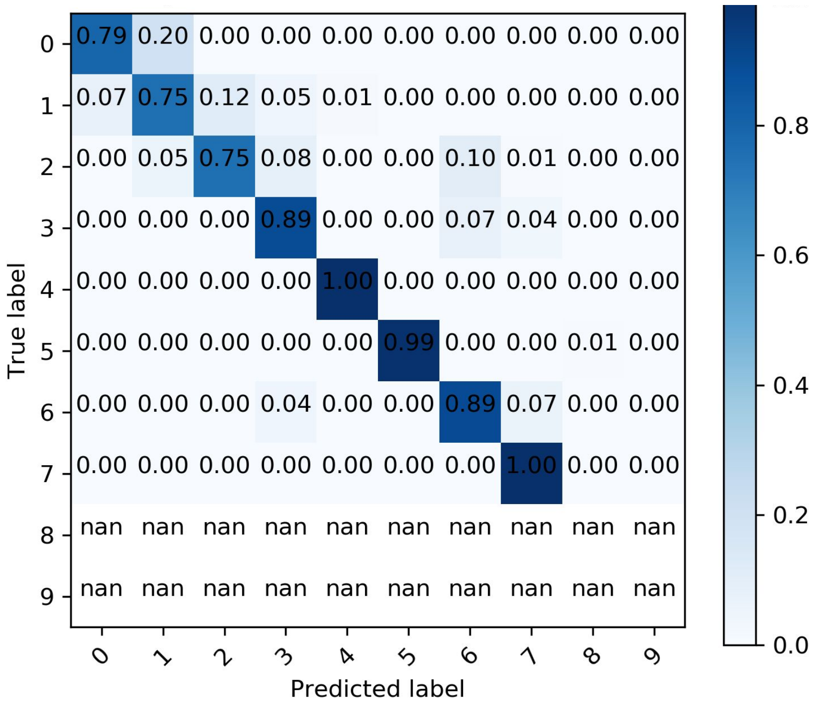
Li, F., Zhang, X., Zhang, X., Du, C., Xu, Y., & Tian, Y.-C. (2018). Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets. Information Sciences, 422, 242–256. https://doi.org/10.1016/J.INS.2017.09.013

AdaBoost 2% sample - 20/10/70 - geometric mean: 0.690570

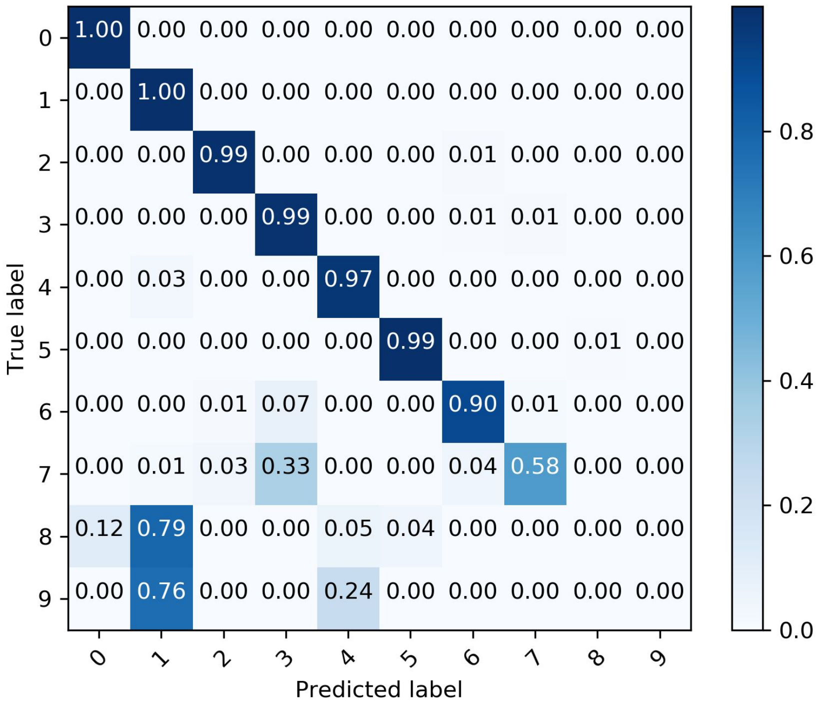


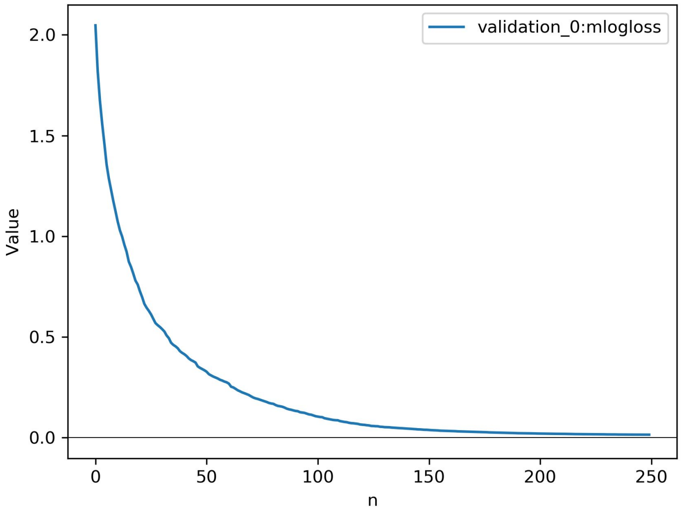
Østfold University College

RandomForest 2% sample - 20/10/70 - geometric mean: 0.826486

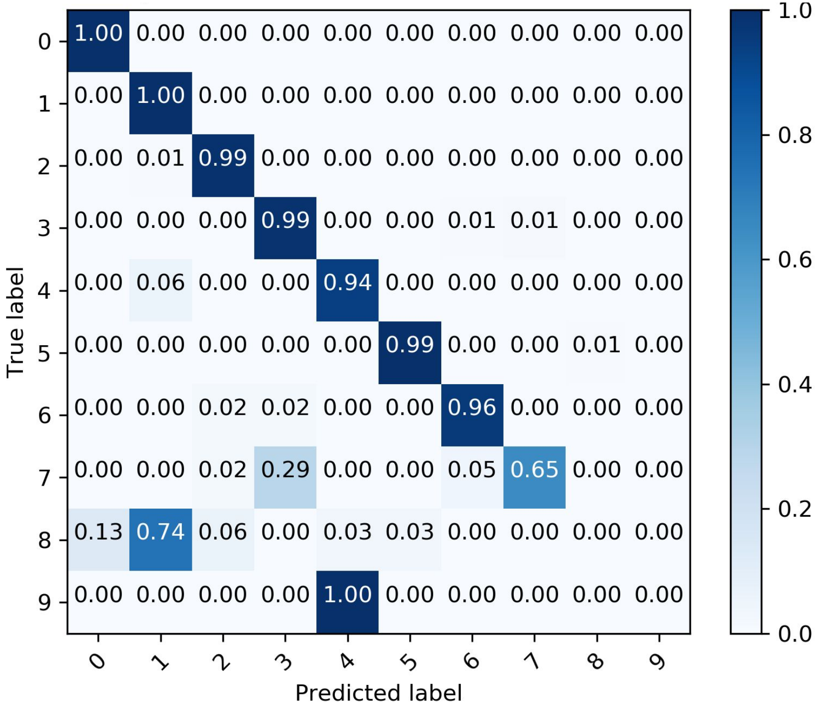


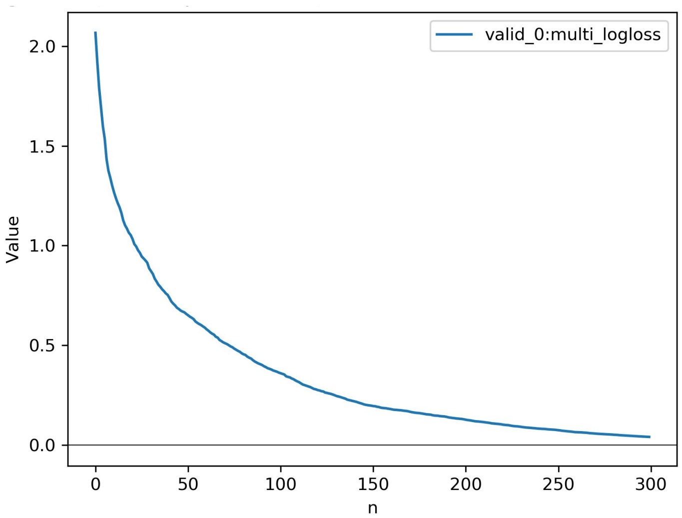
XGBoost 2% sample - 20/10/70 - geometric mean: 0.861604



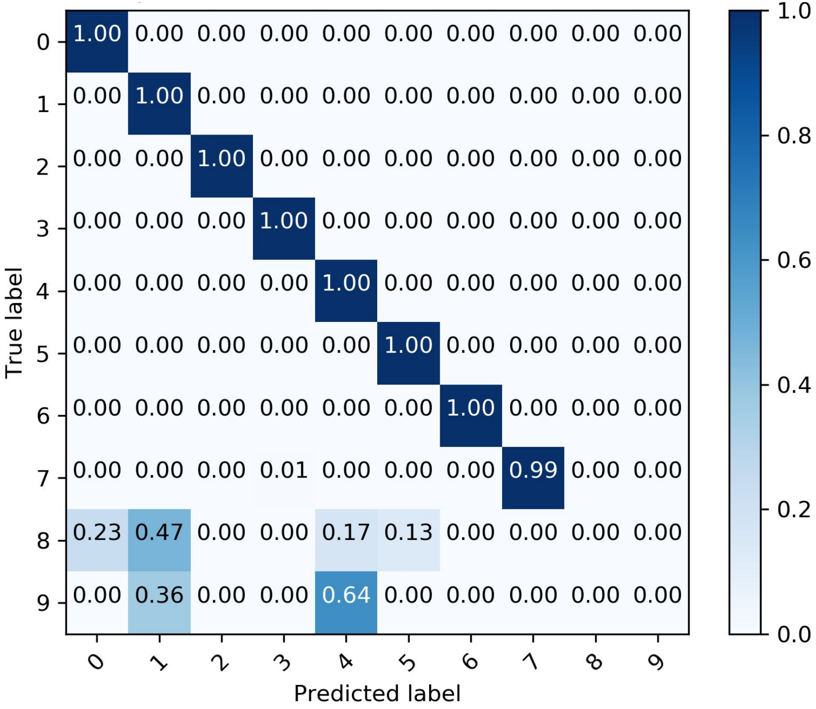


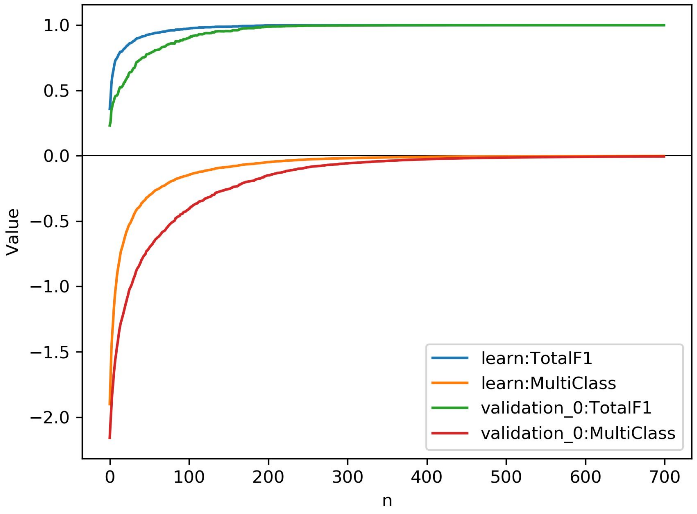
LightGBM 2% sample - 20/10/70 - geometric mean: 0.866858





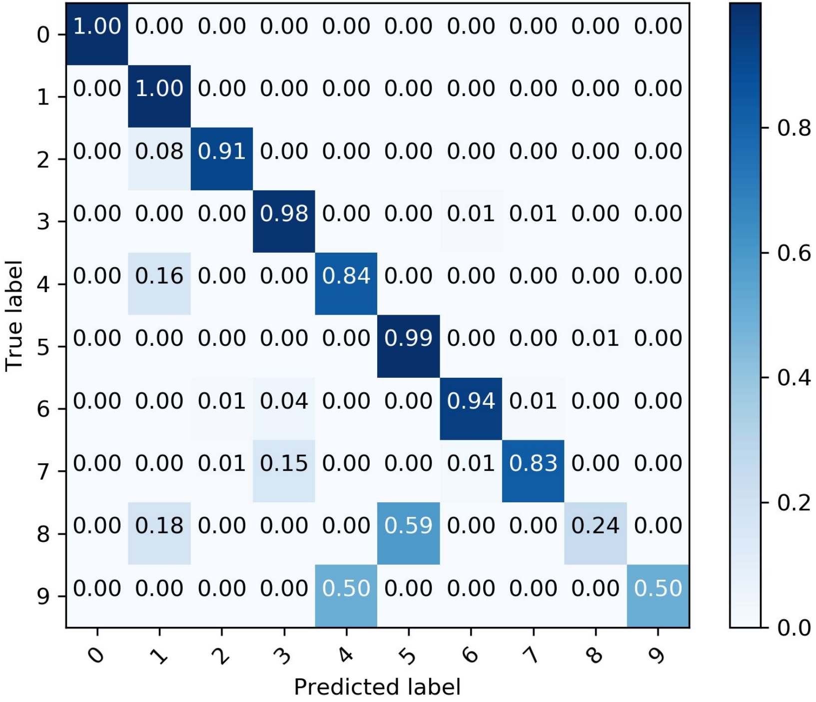
CatBoost 2% sample - 20/10/70 - geometric mean: 0.893551

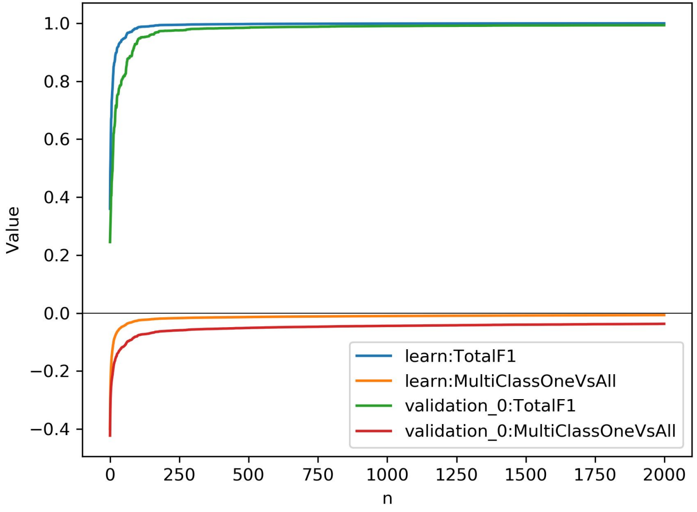




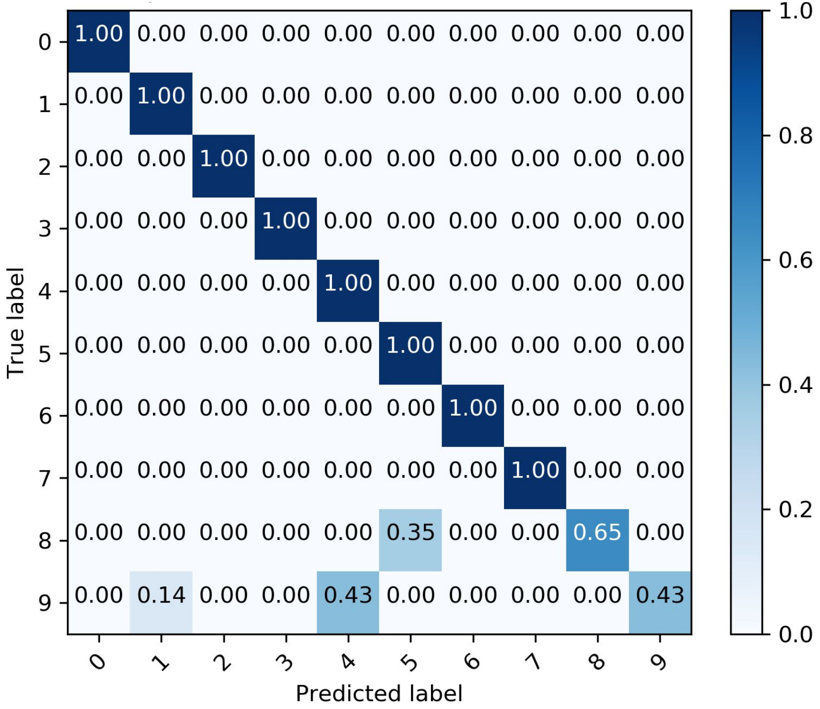
Østfold University College

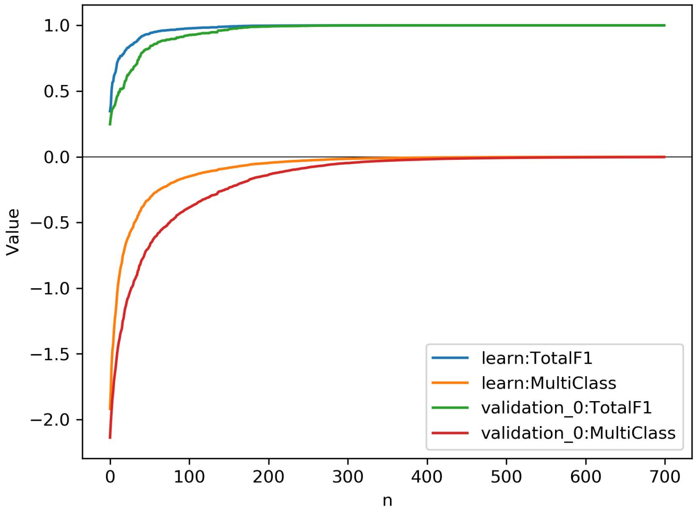
CatBoost 2% sample - 20/10/70 - geometric mean: 0.906385





CatBoost 10% sample - 20/10/70 - geometric mean: 0.952491





Østfold University College

Summary

Model	Sample size (%)	Sample distr. (%)	Accuracy	Geometric mean
C5.0	2	20/-/70	0.738	-
AdaBoost	2	20/-/70	0.660446	0.690570
RandomForest	2	20/-/70	0.779012	0.826485
XGBoost	2	20/10/70	0.998940	0.861604
LightGBM	2	20/10/70	0.998926	0.866858
CatBoost	2	20/10/70	0.999940	0.893551
CatBoost	2	20/10/70	0.993507	0.906385
CatBoost	10	20/10/70	0.999987	0.952491

Results by others

- Sihota (2015): 65% with AdaBoost.
- Kristian, Calvin & Ding (2017): 70% with Random Forest
- Hamelg (2014): 76.9% with Random Forest
  - increased the number of trees to 2000.
  - Concludes that generating new training examples or increasing the number of trees to more than 2000 could improve accuracy.
- Competition at Kaggle.com
  - https://www.kaggle.com/c/poker-rule-induction/discussion
  - Multiple competitors claims 100% accuracy.
  - As no method or code are disclosed, victory can be cheated.
  - Not restricted to decision trees.

Østfold University College

Suggestions for future development

- Feature engineering: Order doesn’t matter. Sort the data. Merge poker hands that contain the same cards, but different order. Population decreases to 2568960 hands, i.e. 0.8237% of the population size. Cheating?

- Use custom classifiers or custom loss functions that focuses primarily on multi-class classification imbalance.

  - Bi, J., & Zhang, C. (2018). An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme. Knowledge-Based Systems, 158, 81–93. https://doi.org/10.1016/J.KNOSYS.2018.05.037
  - Li, F., Zhang, X., Zhang, X., Du, C., Xu, Y., & Tian, Y.-C. (2018). Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets. Information Sciences, 422, 242–256. https://doi.org/10.1016/J.INS.2017.09.013
  - Oh, K., Jung, J.-Y., & Kim, B. (2018). Imbalanced classification of manufacturing quality conditions using cost-sensitive decision tree ensembles AU - Kim, Aekyung. International Journal of Computer Integrated Manufacturing, 31(8), 701–717. https://doi.org/10.1080/0951192X.2017.1407447
  - Du, J., Vong, C.-M., Pun, C.-M., Wong, P.-K., & Ip, W.-F. (2017). Post-boosting of classification boundary for imbalanced data using geometric mean. Neural Networks, 96, 101–114. https://doi.org/10.1016/J.NEUNET.2017.09.004
  - Kim, M.-J., Kang, D.-K., & Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. Expert Systems with Applications, 42(3), 1074–1082. https://doi.org/10.1016/J.ESWA.2014.08.025

Code repositories (ours)

Generating new data set (C#):

https://github.com/Alvtron/PokerData

R code:

https://github.com/Alvtron/ITI43210-Machine-Learning

Python code:

https://github.com/Alvtron/PythonMachineLearning

Østfold University College