Data challange

Andrea Maioli

September 26, 2022

Outline

1. Introduction
2. Data Analysis
   - Data exploration
   - Unbalanced data
3. Model design
4. Performance
5. Conclusions

Introduction

Data Analysis

Data exploration

Unbalanced data

Model design

Performance

Conclusions

Introduction

What's the problem?

- Multi class classification of Cardiotocographic data of Fetal Heart Rate (FHR)
- Classes:
  1. **Normal** (N)
  2. **Suspect** (S)
  3. **Pathologic** (P)
- 21 features describing measuraments of Cardiotocographic measured by **SisPorto** system.
- Click here to download data

Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Data Analysis

Data exploration

Features can be dived into 3 groups:

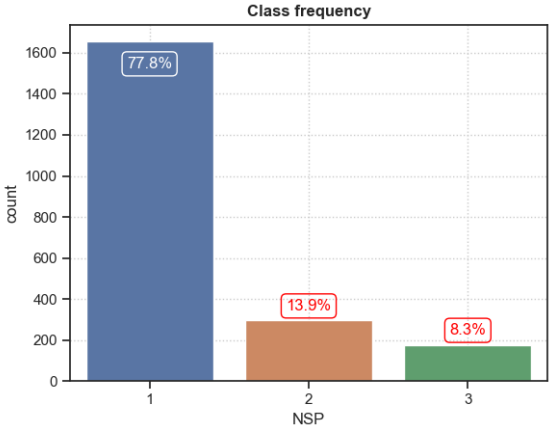
1. **Signal histogram properties**: Width, Min, Max, Nmax, Nzeros, Mode, Mean, Median, Variance, Tendency
2. **Signal global properties**: LB, AC, FM, UC, ASTV, mSTV, ALTV, mLTV
3. **Deceleration properties**: DL, DS, DP

All features are numerical except for **DS** that is binary.

Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Data exploration

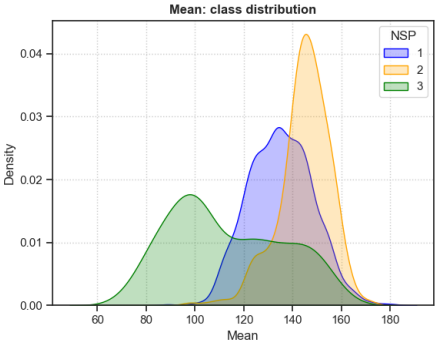
Class frequency

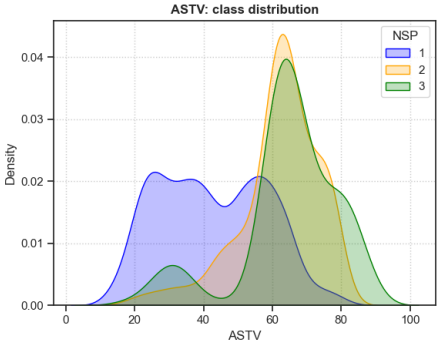


Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Features selection

Single variable distribution





Data challange

Andrea Maioli

Introduction

Data Analysis

Data exploration

Unbalanced data

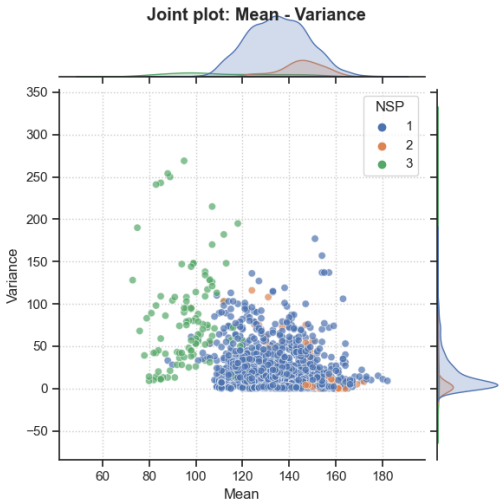
Model design

Performance

Conclusions

Features selection

Joint distribution



Features selection

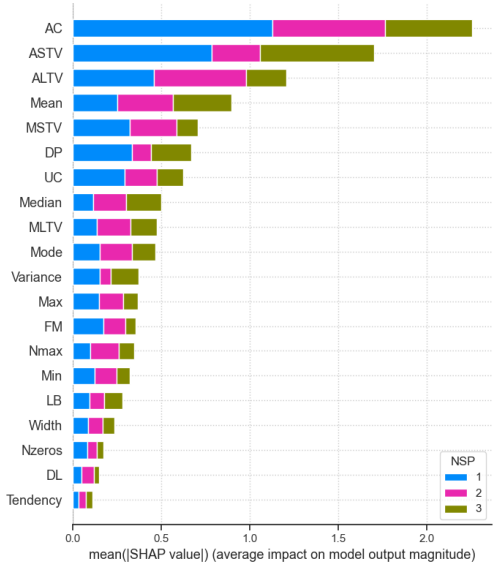
Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Conclusion

From graphical exploration we deduce that **AC**, **ASTV**, **ALTV**, **Mean**, **Variance**, **DP** will be the main regressors.

Features selection

Shapley value



Data challange

Andrea Maioli

Introduction

Data Analysis

Data exploration

Unbalanced data

Model design

Performance

Conclusions

Features selection

Conclusion

Shapley values confirm the graphical analysis. The main regressors will be: **AC, ASTV, ALTV, MSTV, Mean, Variance, DP**.

Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Unbalanced data

We must take care of unbalanced data to avoid models overfitting the most frequent class (**N**).

We adopt two different techniques:

- **Sample weight** applied to cost function.
- Oversampling via `SMOTE` algorithm.

Unbalanced data

Sample weight

We build a weight k_i for each class i to compensate data asymettry in the following way:

Formula:

$$c_i = \#\{\text{samples of class } i\}$$

$$K = \frac{1}{c_1} + \frac{1}{c_2} + \frac{1}{c_3}$$

$$k_i = \frac{K}{c_i}$$

Results:

$$k_1 = 0.06 \pm 0.002$$

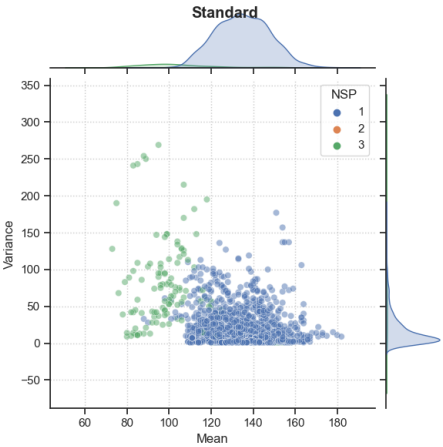
$$k_2 = 0.35 \pm 0.0016$$

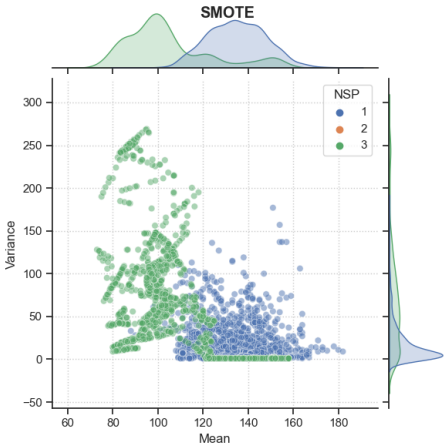
$$k_3 = 0.59 \pm 0.0013$$

Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Unbalanced data

SMOTE





Data challange

Andrea Maioli

Introduction

Data Analysis

Data exploration

Unbalanced data

Model design

Performance

Conclusions

Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Model design

Model design

Benchmarks

Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Benchmark models:

- `MostFrequent`: always predict most frequent class in training set.
- `SmartRandom`: predict extracting random class with a probability equal to frequency of the class in training set.

Model design

Models

- `Logistic`
- `LogisticWeight`: logistic model trained on weighted data
- `CatBoost`
- `CatBoostWeight`: Boosted Tree model trained on weighted data
- `CatBoostSMOTE`: Boosted Tree model trained on augmented data via `SMOTE` algorithm.

Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Model design

Models

Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Note

- `StandardScaler` has been applied on top of `Logistic` models due to L^1 regularisation.
- Early stopping have been appliead to `Boosted Tree` models to avoid overfitting.
- `RandomUnderSample` have been applied on top of `SMOTE` algorithm as suggested in the original paper.

Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Performance

Performance

Metrics... which one?

Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Nice to have properties:

- **Recall** score of class **P** as high as possible.
- Minimization of **False Negative** of classes **P** predicted as **N**.
- Reasonable **Precision** of classes **N** and **S** to avoid too much **False Positive**.

Performance

Metrics... which one?

Custom performance criteria

1. Choose the 3 models with best average **F2** score.
2. Among selected models choose the one with the highest **Recall** score of class **P**.

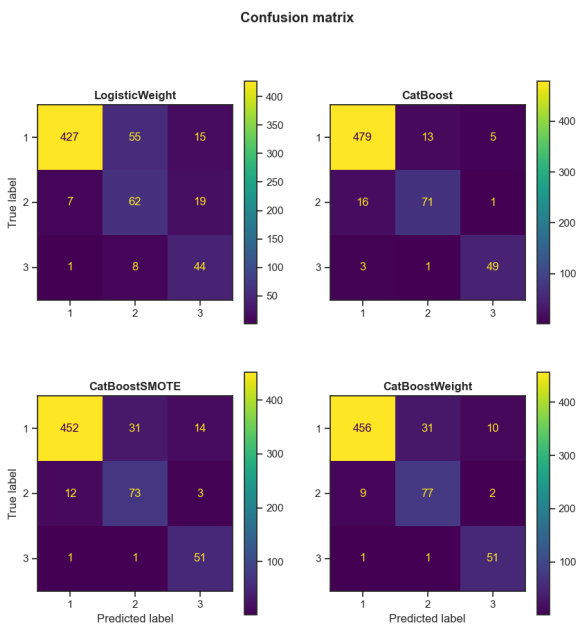
Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Performance

Performance computed using `StratifiedCrossValidation` with `cv=5`:

Model	F2 Score	Recall class P
MostFrequent	0.315	0
SmartRandom	0.322	0.094
Logistic	0.712	0.678
LogisticWeight	0.755	0.785
CatBoost	**0.801**	0.795
CatBoostWeight	**0.797**	0.812
CatBoostSMOTE	**0.797**	**0.818**

Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions


Confusion matrix
LogisticWeight
CatBoost
CatBoostSMOTE
CatBoostWeight
True label
Predicted label
427 55 15
7 62 19
1 8 44
479 13 5
16 71 1
3 1 49
452 31 14
12 73 3
1 1 51
456 31 10
9 77 2
1 1 51
1 2 3
400 350 300 250 200 150 100 50

Data challange

Andrea Maioli

Introduction

Data Analysis

Data exploration

Unbalanced data

Model design

Performance

Conclusions

Conclusions

Conclusions

- Non linear models are necessary to fully exploit this problem and overperform linear model.
- Both data augmentation and sample weight contribute to performance improvement, in particular in term of **Pathologic Recall**.

Best model

Even if `CatBoostSMOTE` is the best model according to our criteria **I would choose** `CatBoostWeight` **as best model** because the two models have very close performance but the first one has a random component not fully controllable.

Data challange
Andrea Maioli
Introduction
Data Analysis
Data exploration
Unbalanced data
Model design
Performance
Conclusions

Data challange

Andrea Maioli

Introduction

Data Analysis

Data exploration

Unbalanced data

Model design

Performance

Conclusions

Fin.