Predicting sports betting player suspensions by algorithm

potentials, limitations, and recommendations

Thomas Krause [1]   Vadim Kufenko[1]   Steffen Otterbach[1]

[1]Gambling Research Center, University of Hohenheim
Third Current Advances in Gambling Research (CAGR) Conference
June 28th and 29th
King's College London

Thursday 29th June 2023

Outline

Introduction and starting point

Data, Training and Estimation

Direct Compare of Pipelines

Best Pipeline and Model

Conclusions and consequences

Table of Contents

Introduction and starting point

Data, Training and Estimation

Direct Compare of Pipelines

Best Pipeline and Model

Conclusions and consequences

Introduction and starting point

- Ongoing liberalisation of the German online gambling market
- Increase in (online) sports betting in Germany (2021 conservatively approx. 10 billion euros)

- Other and new addiction potentials
- Mandate of the German State Treaty on Gambling 2021:

  [...] use an automated system based on scientific evidence and algorithms for the early detection of gamblers at risk of gambling addiction and of gambling addiction. (GlüStV 2021, translate by T.K.)

Opportunities for the prevention of addiction

- Extensive non-reactive data collection (safe-server infrastructure)
- Potential for early identification of gambling problems through ML models

Our guiding research questions are

- Which algorithms and data handling techniques are appropriate?
- Which (player) data should be used in the algorithms for this purpose?
- Which indicators and cut-offs are applicable for the early protection of at-risk and pathological gamblers?

Analytic approach

- Suspension events as a target variable
- Aggregation of process-generated behavioural data at daily, weekly and annual level
- ML-Estimations: potential and predictors
- Test of data handling procedures for rare event data (imbalanced)
- Each data-pipeline compared multiple modern models
- Hyperparameter-Search for the three best models in each pipeline
- Selection of best models in each pipeline

Pipelines

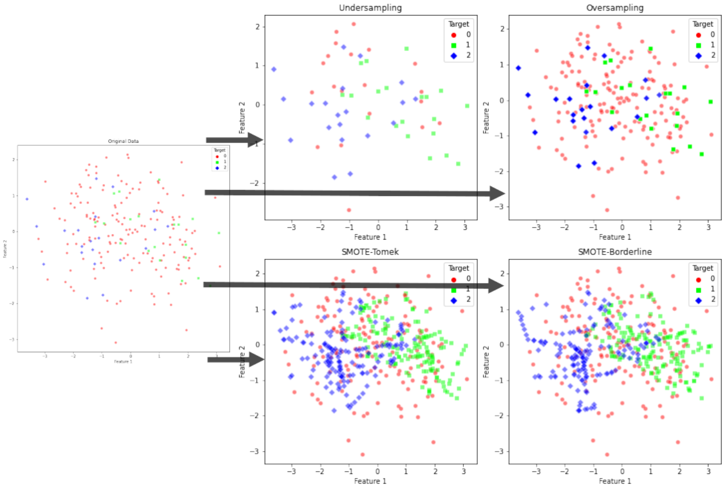


Figure: Raw Data to Pipelines

Table of Contents

Introduction and starting point

Data, Training and Estimation

Direct Compare of Pipelines

Best Pipeline and Model

Conclusions and consequences

Data

- Two biggest providers of sportbets in SH ($>$ 98% of a active sportbeters and 97% of all suspensions)
- Year 2020 to beginning of 2021 (26 459 active players)
- Aggregation of player-data, transactions data, bet data, results data in yearly, weekly, and daily time-intervals
- Totals, means, variations, shape, range and change of aggregated case data resulting in 399 features

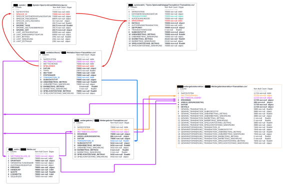
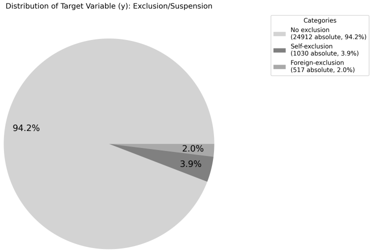
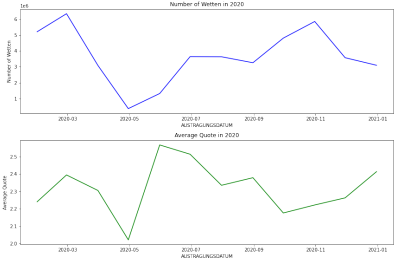


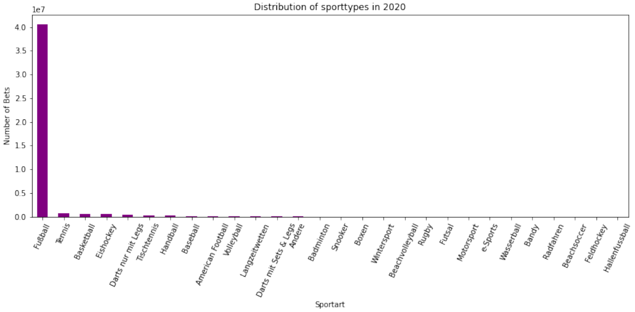
Figure: Data, Structure and Connections

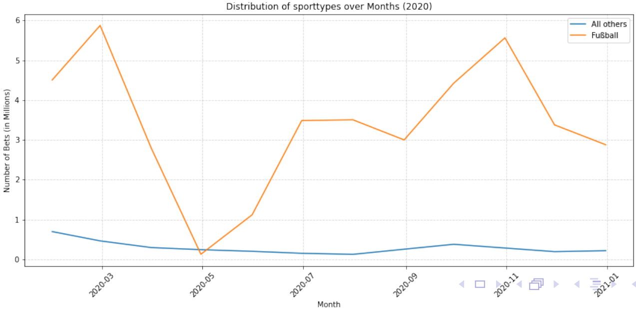
Descriptive Data





Descriptive Data





Challenges for Data-Analysis

- Variation in provider labels (even wrong use of variables)
- Erroneous information
  - betting odds < 1
  - placement of bets despite present suspension
  - Overcoverage: NON-online-Players in data
- Missing unblocking events

Model Training and Estimators

- Train/Test-Split: **75/25**
- Feature-Space-Reduction: **Boruta**
- Hyperparameter Search: **Optuna** (Akiba et al. 2019) with Tree Structured Parzen Estimator (**TPE**) with Asynchronous Successive Halving Algorithm (**asha**) at 200 Iterations
- Estimators: **rf, ada, et, lightgbm; gbc, xgboost, catboost**
- Rebalancing Data-Pipelines: random **undersampling**, random **oversampling**, SMOTE-**TOMEK**, SMOTE-**Borderline**
- **F1**-**Score** as main scorer

Table of Contents

Introduction and starting point

Data, Training and Estimation

Direct Compare of Pipelines

Best Pipeline and Model

Conclusions and consequences

Pipeline Performance Comparison

Table: Comparison of ML model performance

	Pipeline			
Metric	**Unders.**	**Overs.**	**Tomek**	**Borderline**
Model-Class	GBM	XGB	LGBM	XGB
Accuracy	0.77	0.93	0.94	0.94
Precision(macro)	0.38	0.47	0.51	0.55
Precision(weighted)	0.92	0.91	0.92	0.92
Recall (macro)	0.51	0.42	0.40	0.43
Recall (weighted)	0.77	0.93	0.94	0.94
F1-Score (macro)	0.39	0.43	0.42	0.43
F1-Score (weighted)	0.83	0.92	0.92	0.94
AUC-ROC (OvR)	0.747	0.817	0.826	0.816

Table of Contents

Introduction and starting point

Data, Training and Estimation

Direct Compare of Pipelines

Best Pipeline and Model

Conclusions and consequences

Best Pipeline and Model: SMOTE-Borderline with XGBoost

```
              precision    recall  f1-score   support

           0       0.95      0.99      0.97      6228
           1       0.46      0.26      0.33       258
           2       0.24      0.04      0.07       129

    accuracy                           0.94      6615
   macro avg       0.55      0.43      0.46      6615
weighted avg       0.92      0.94      0.93      6615
Accuracy: 0.9399848828420257
AUC ovr: 0.8162012258005035
Average precision score, micro-averaged over all classes: 0.97
Average precision score, macro-averaged over all classes: 0.46
Average precision score, weighted-averaged over all classes: 0.94
Average precision score, samples-averaged over all classes: 0.97
```

Best Pipeline and Model: SMOTE-Borderline with XGBoost

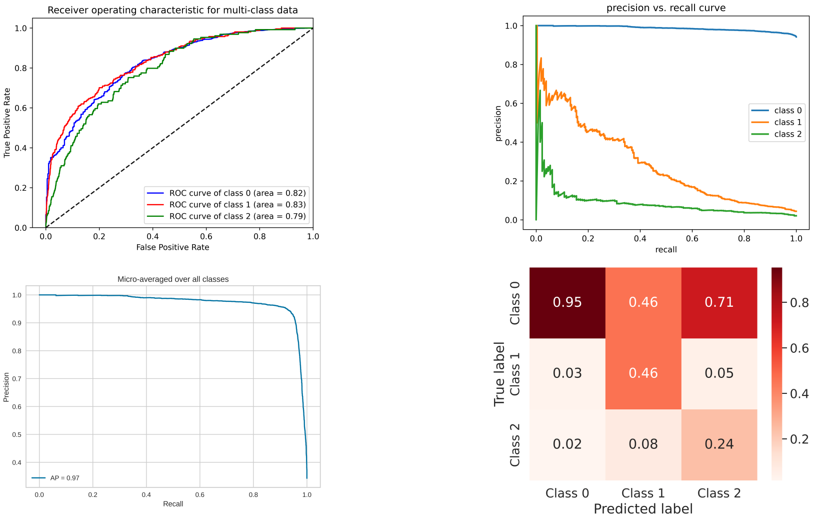


Figure: SMOTE-Borderline

Best Pipeline and Model: SMOTE-Borderline with XGBoost

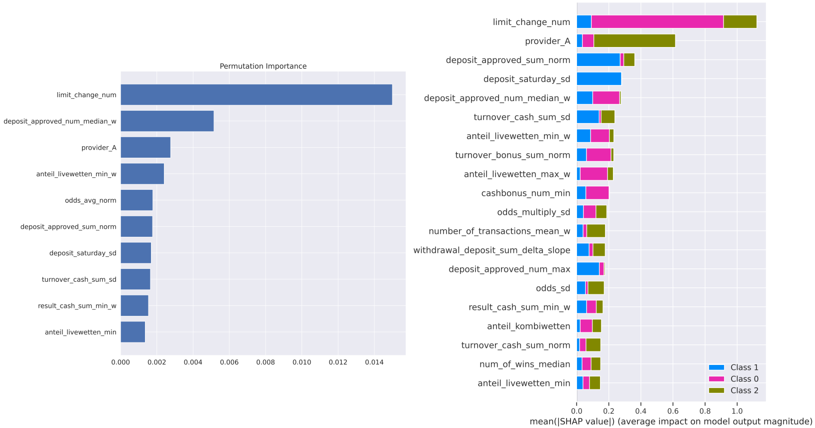


Figure: SMOTE-Borderline

Table of Contents

Introduction and starting point

Data, Training and Estimation

Direct Compare of Pipelines

Best Pipeline and Model

Conclusions and consequences

Conclusion for the prediction of suspension events

- SMOTE-Borderline with gradient boosting (XGBoost or LightGBM) are advisable for prediction
- Many false-positive cases and why this is plausible and to be expected
- Problems of the target variable for our predictions
- Third-party suspension (literal translation of German term: Foreign Exclusion)
  1. does not follow a reconstructible logic
  2. differs greatly between providers

Political consequences and pathways for gambling supervision.

- better data oversight is needed
  - Uniform labels are needed
  - Unblocking has to be documented (seems now to be the case)
  - Implausible values must be compulsorily checked for an effective monitoring of operators
- Additional datapoints are needed for an effective "automated system":
  - Assessment of PGSI (etc.)
  - Documentation of communication between operator and user

Contact

Dr. Thomas Krause
thomas.krause@uni-hohenheim.de
University of Hohenheim,
Gambling Research Center
Schwerzstraße 46,
70599 Stuttgart
https://gluecksspiel.uni-hohenheim.de

UNIVERSITY OF HOHENHEIM

Best Pipeline and Model: SMOTE-Borderline with XGBoost

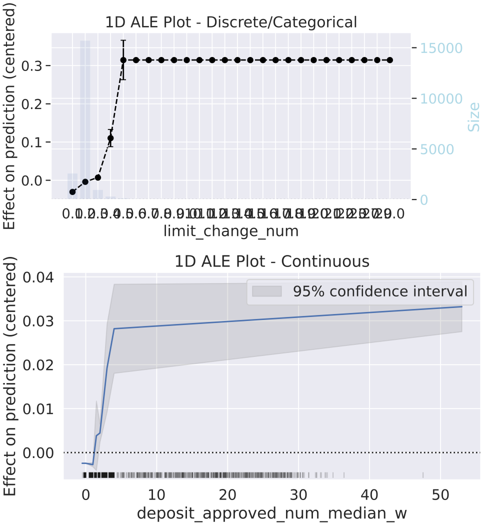


Figure: SMOTE-Borderline