Predicting crime during or after psychiatric care
Evaluating machine learning for risk assessment using the Danish patient registries
Trinhammer, M. L.; Merrild, A. C. Holst; Lotz, J. F.; Makransky, G.

Published in: Journal of Psychiatric Research

DOI: 10.1016/j.jpsychires.2022.06.009

Publication date: 2022

Document version Publisher's PDF, also known as Version of record

Document license: CC BY

Citation for published version (APA): Trinhammer, M. L., Merrild, A. C. H., Lotz, J. F., & Makransky, G. (2022). Predicting crime during or after psychiatric care: Evaluating machine learning for risk assessment using the Danish patient registries. Journal of Psychiatric Research, 152, 194-200. https://doi.org/10.1016/j.jpsychires.2022.06.009
Predicting crime during or after psychiatric care: Evaluating machine learning for risk assessment using the Danish patient registries

M.L. Trinhammer a,*, A.C. Holst Merrild b, J.F. Lotz c, d, G. Makransky a

a Department of Psychology, University of Copenhagen, Øster Farimagsgade 2A, 1353, Copenhagen, Denmark
b DTU COMPUTE, Technical University of Denmark, Building 324, 2800, Kongens Lyngby, Denmark
c The ROCKWOOL Foundation, Ny Kongensgade 6, 1472, Copenhagen, Denmark
d Department of Computer Science, University of Copenhagen, Universitetsparken 1, 2100, Copenhagen, Denmark

ARTICLE INFO

Keywords:
Computational psychiatry
Precision psychiatry
Forensic psychiatry
Statistical risk assessment
Machine learning

ABSTRACT

Background: Structural changes in psychiatric systems have altered treatment opportunities for patients in need of mental healthcare. These changes are possibly associated with an increase in post-discharge crime, reported in the increase of forensic psychiatric populations. As current risk-assessment tools are time-consuming to administer and offer limited accuracy, this study aims to develop a predictive model designed to identify psychiatric patients at risk of committing crime leading to a future forensic psychiatric treatment course.

Method: We utilized the longitudinal quality of the Danish patient registries, identifying the 45,720 adult patients who had contact with the psychiatric system in 2014, of which 474 committed crime leading to a forensic psychiatric treatment course after discharge. Four machine learning models (Logistic Regression, Random Forest, XGBoost and LightGBM) were applied over a range of sociodemographic, judicial, and psychiatric variables.

Results: This study achieves a F1-macro score of 76%, with precision = 57% and recall = 47% reported by the LightGBM algorithm. Our model was therefore able to identify 47% of future forensic psychiatric patients, while making correct predictions in 57% of cases.

Conclusion: The study demonstrates how a clinically useful initial risk-assessment can be achieved using machine learning on data from patient registries. The proposed approach offers the opportunity to flag potential future forensic psychiatric patients, while in contact with the general psychiatric system, hereby allowing early-intervention initiatives to be activated.

1. Introduction

Demand for forensic psychiatric services has seen an increase in Western European countries, occurring concurrently with reforms intended to downsize psychiatric hospital inpatient wards (Janson-Hart et al., 2011). This association is known as Penrose’s Law (Kalapos, 2016) and offers the hypothesis that deinstitutionalisation in the general psychiatric system is associated with the premature discharge of high-risk individuals (Lamb, 2015). Offering these citizens advanced mental healthcare while in contact with the psychiatric system is thus an increasingly pivotal task for Western European countries. Such enhanced treatment may ultimately lead to better quality of life for the individual, and the avoidance of future criminal episodes, with the myriad of benefits enjoyed by society in general.

Qualified risk-assessment prior to discharge from psychiatric care hereby emerges as a valuable instrument employed in order to identify individuals for whom a criminal trajectory is likely. The current approach to risk-assessment in psychiatry relies on a combination of structured interviews combined with clinical-decision making. Structured risk-assessment interviews remain time-consuming to administer (Viljoen et al., 2010) with questionable predictive power (Douglas et al., 2017). Such questionnaires have further been reported to be insufficiently specific, limiting their general usefulness (Large et al., 2011). Clinical judgement is additionally limited by decision-making bias (Murray and Thomson, 2010) exacerbated by large patient quantities and fewer resources. With the advent of advanced prediction techniques such as machine learning (ML), the opportunity to perform an initial risk-assessment to assist clinical decision making appears as an interesting path for the future psychiatric system. Current studies attempting such demarcation of risk via advanced statistical

* Corresponding author.
E-mail address: martin.lund.trinhammer@psy.ku.dk (M.L. Trinhammer).

https://doi.org/10.1016/j.jpsychires.2022.06.009
Received 8 January 2022; Received in revised form 1 June 2022; Accepted 6 June 2022
Available online 15 June 2022
0022-3956/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
techniques are mainly concerned with small population sizes, including only individuals who are already in secure- or forensic psychiatric hospitals (Watts et al., 2021; Wolf et al., 2018). A gap in the literature is therefore if statistical risk-assessment using large sample sizes enables the identification of this group of high-risk individuals before treatment in the forensic system is required. Such an approach could offer clinicians invaluable and easy-to-interpret estimates of which patients are in need of follow-up care, in an inexpensive and efficient manner (Passos et al., 2019).

In this study, we attempt the development of a supervised ML framework designed to perform clinically useful risk-assessment of individuals in the general psychiatric system, according to their future probability of requiring treatment in a forensic institution. Specifically, this is done by predicting if individuals will commit a crime leading to the issuing of court-ordered psychiatric admission (COPA) for a crime committed during or after an out-patient treatment course or after discharge from in-patient care. A COPA is the judicial verdict used in Denmark to assign mentally ill offenders to forensic psychiatric treatment. In addition to the quantitative findings in this paper, we also aim to analyse and discuss the clinical implications; including the effects of varying the risk-threshold used to classify patients (Chen et al., 2020; Douglas et al., 2017). The current study employs the Danish patient registries, hereby also investigating if satisfactory estimates can be achieved using readily-available and non-invasive data types (Bjerre-Nielsen et al., 2021), ensuring cheap and fast predictions.

2. Method

2.1. Data foundation

This study employs the Danish national psychiatric patient register, offering information on citizens in contact with Danish mental healthcare institutions. Each observation includes a pseudo-anonymous social security number enabling the linkage to socioeconomic and criminal justice registries.

2.2. Sample

The resulting study sample consists of 45,720 patients of whom 474 (1%) committed a crime leading to a COPA. The study sample contains all individuals between the age of 20 and 60 who were either admitted as a psychiatric in- or out-patient in the year of 2014. The subset of these patients who committed a crime, leading to the issuing of a COPA during or after out-patient treatment or after in-patient discharge, in either 2014 or 2015 constitutes class 1. The remaining patients constitute class 0. The aim of this study is thus to distinguish the two groups from one another, i.e. future recipients of COPA (minority group/class 1) vs. non-future recipients (majority group/class 0). For fitting the ML models, our dataset was split into two sets for training and evaluation: a 80% training set containing 36,576 patients, where 362 had received a COPA within the two year time span, and a 20% hold-out evaluation set containing 9,144 individuals with 112 COPA recipients.

2.3. Selection of predictors

The dataset comprises 39 predictors, all identified from research on characteristics of forensic psychiatric patients (Kirchebner et al., 2020; Soyka et al., 2007). The predictors are divided into three groups: socioeconomic, psychiatric history, and criminal history. The variables within each group are briefly described in the following paragraphs and further details can be found in the appendix. All categorical variables have been one-hot-encoded (Hastie et al., 2009).

2.4. Sociodemographic information

This group of predictors includes information regarding gender, age, the highest attained education, civil status, and area of residence. Age is included as four dummy variables each representing a decade from the 20’s to the 50’s. This is done in order to incorporate differences between age groups opposed to the effect of one year increase.

2.5. Psychiatric information

This group of predictors includes diagnoses as well as information regarding the patients’ previous treatment trajectory. The majority of patients included had multiple contacts with the Danish psychiatric system prior to their contact in 2014. In order to maximize the amount of information, the data from these contacts has been accumulated since 1995. This in turn means that the predictors loose temporal information, since the dataset is unable to account for when a given interaction occurred. In addition we compute the number of unique disorders that an individual has received over time. Regarding the treatment trajectory, we include the total amount of treatment courses an individual has had, the total number of days admitted to in-patient treatment and the total number of days admitted to out-patient treatment.

2.6. Criminal history

The information about an individuals’ criminal history is also accumulated since 1995. It includes the sum of convictions, the total number of criminal offences, the total number of sentences served in prison, the number of days sentenced to unconditional and conditional prison sentences, respectively, and the number of previous court-ordered psychiatric admissions, if applicable.

2.7. Missing values

Missing values were detected for 443 individuals. These were all members of the majority group and were dropped from the study sample. Missing values in the Danish patient registries often occur for short-term visitors and the overall influence of this group of people on the outcomes of this study can therefore be considered minuscule.

2.8. Sampling strategies

Since the minority group comprises just 1% of the total sample, the dataset is highly imbalanced, a property which is known to cause poor model performance (Fernández et al., 2018). Algorithmic and data level techniques have been investigated to address this issue. At the algorithmic level, SMOTE and One-Sided Selection have been attempted (Fernández et al., 2018), but due to poor performance compared to data level approaches, these were not included in the final analyses. Sampling strategies used at the data level are thus implemented on the training data with the purpose of changing its class distribution (Buda et al., 2018). This study compares two techniques, 1) Random Undersampling of the majority class, and 2) Random Oversampling of the minority class. When implementing the under sampling strategy, random observations without a COPA were randomly eliminated until a desired class balance was achieved. This ratio was tested in the range from 99:1 to 50:50. When implementing Random Oversampling, random observations with a COPA were duplicated until the desired class balance. This ratio was tested in the range from 99:1 to 95:5.

2.9. Risk thresholding

When performing classification analyses, the standard approach is to assign observations with a predicted probability equaling or exceeding 0.5 to class 1, and assign those with a probability less than 0.5 to class 0. When working with highly imbalanced datasets, however, it is not given that this standard threshold is the optimal one (Fernández et al., 2018). To address this, our analyses will also investigate the models’ performance over varying classification thresholds, using 0.1, 0.3, 0.5, 0.7,
and 0.9 as the demarcation points. These findings will be used to qualify our discussion concerning the clinical implications for employing risk prediction tools in the general psychiatric system.

2.10. Machine learning models

A total of four machine learning models were trained: a conventional logistic regression model was included as a baseline; Random Forest (Breiman, 2001) since it is used in comparable studies (Chen et al., 2020; Watts et al., 2021) and has generally shown good performance (Fernández-Delgado et al., 2014); XGBoost which employs gradient boosting and is recommended for imbalanced datasets (Wang et al., 2020); and LightGBM (Ke et al., 2017) that recently proved very potent when dealing with tabular data (Borisov et al., 2021). For a detailed description to the model’s inner workings, we refer to the original articles (Breiman, 2001; Ke et al., 2017; Wang et al., 2020). All models were implemented in Python 3 (Pedregosa et al., 2011).

To determine the optimal set of hyperparameters for each model, a grid search with five fold cross-validation was applied to the training set. The sampling strategies were also applied within each fold to minimize the risk of overfitting. Generally, the grid search contained five to ten different values per hyperparameter; the range can be found in the appendix. For Random Forest, we tuned the tree depth, the maximum number features to consider before a split, the minimum number of observations allowed in a node before a split as well as in leaf nodes. Tuning the number of observations in the nodes suggests improved performance, particularly in situations where the sample size is large and the dimension is relatively low (Lin and Jeon, 2006).

For LightGBM and XGBoost we tuned comparable hyperparameters such as the maximum tree depth, and a range of parameters particularly suitable for highly imbalanced datasets, such as minimum samples in each child for LightGBM and the minimum sum of weight in child for XGBoost.

2.11. Evaluation metrics

Given the imbalanced class distribution, the proposed model can be expected to predict the majority of observations as true negatives, i.e. non-future recipients of COPA. As the conventional metrics (such as Accuracy and AUROC) are inflated by this, we instead validate our model according to the F1-Macro score. The F1 score is a weighted combination of precision (also known as the Positive Predictive Value, PPV) and recall (Lewis and Gale, 1994) with 1 being the best score achievable for a classifier and 0 the worst. The F1-Macro score is the arithmetic mean of the F1 score for each class in the data (Opitz and Burst, 2019). Recall, also known as sensitivity, is a measure of the share of positive cases that the model can actually identify, i.e. how many of the individuals who receive a COPA after discharge the model flags. Precision reports the amount of correctly identified cases, out of the sum of identified cases and is thus deflated by many false positives. When evaluating the results, the best model will be identified from the highest achievable F1-Macro score, a metric recommended for highly imbalanced datasets (Fernández et al., 2018). For evaluating the models’ performance on the evaluation set we also present the Negative Predictive Value (NPV), Specificity, Accuracy, Balanced Accuracy and the AUROC at different classification thresholds.

Table 1 and Table 2 presents the results from our ML models. Our study reports the LightGBM algorithm combined with Random Under-sampling (Table 1) as the best performing approach, yielding an overall F1-Macro score of 76%. The related, tree-based models, Random Forest and XGBoost, perform roughly equally well reporting an F1-Macro score of 75% and 74%, respectively. Minor, yet noteworthy differences between the three models are the recall and precision scores. The highest recall is achieved by the LightGBM model, though the Random Forest reports the highest precision score. Additionally, it should be noted that the very high rate of true negatives in our study inflates the NPV, the Specificity, the Accuracy and the AUROC which makes these metrics incomplete to solely rely on. Our study reports Random Undersampling as the best performing sampling strategy for all models in the battery.

The results reported in Tables 1 and 2 have been compared using the standard risk-threshold of 0.5, meaning that the model will classify a patient to the high-risk group (class 1), if the predicted probability equals or exceeds 0.5. In the context of this specific study, the adverse outcome of a false positive is less severe than that of a false negative. As such, to qualify adoption in a clinical setting, we investigated how the best performing LightGBM model would perform over varying risk-thresholds.

3. Results

3.1. General results

Table 1 and Table 2 presents the results from our ML models. Our study reports the LightGBM algorithm combined with Random Under-sampling (Table 1) as the best performing approach, yielding an overall F1-Macro score of 76%. The related, tree-based models, Random Forest and XGBoost, perform roughly equally well reporting an F1-Macro score of 75% and 74%, respectively. Minor, yet noteworthy differences between the three models are the recall and precision scores. The highest recall is achieved by the LightGBM model, though the Random Forest reports the highest precision score. Additionally, it should be noted that the very high rate of true negatives in our study inflates the NPV, the Specificity, the Accuracy and the AUROC which makes these metrics incomplete to solely rely on. Our study reports Random Undersampling as the best performing sampling strategy for all models in the battery.

3.2. Risk thresholding

Table 3 presents the results for the LightGBM model when investigating a range of different classification thresholds. Risk-thresholds for the other models can be found in the appendix.

A trade-off can be identified between recall and precision. When the threshold for classification is high, i.e. 0.9, only patients with a very high probability for future forensic treatment is classified as class 1. In this case, the precision score is 0.96, meaning that the vast majority of flagged patients were true future offenders. This approach, however, omits far too many individuals reported in the low recall. Reversely, with a low threshold (0.1), many patients are assigned to the high-risk group as only a predicted probability larger than 0.1 is required. This approach identifies 81% of future offenders. The optimal classification threshold will be discussed next.

4. Discussion

4.1. Clinical usefulness

This study has demonstrated how a ML approach can offer useful risk-estimates of psychiatric patients’ probability for a future admission in the forensic psychiatric system. With a precision score of 57%, the results offer notable improvements to current structured interview approaches, as our model is correct in 57% of cases when assigning a patient to the high-risk group. When using conventional methods, 50%–99.7% of the patients categorized as high-risk would not go on to commit future crime, underscoring the highly sensitive character of these tools (Large et al., 2011). Further, when compared to current statistical approaches (Watts et al., 2021; Wolf et al., 2018) our model offers the ability to identify psychiatric patients before a crime leading
to forensic psychiatric treatment has been committed. Next to the myriad of benefits enjoyed by a host of societal actors, this offers vastly different mental healthcare trajectories for the individuals, as the type and quality of care expected in psychiatric services are of a wholly other kind than the care available at forensic clinics (Franke et al., 2020). Our approach additionally allows the development of early-intervention initiatives in the intersection between mental health care and crime, initiatives which has been invited by other authors (Ford, 2015). Such efforts could critically reform the current treatment procedures in the psychiatric system, as advanced treatment procedures can be targeted to the whole forensic psychiatric population, as some COPA-recipients are omitted by the model. Next to the still exploratory nature of setting an appropriate risk threshold, it can be argued that the impact of this approach on clinical evaluations by healthcare personnel, the misclassified individual would only be subjugated to further assessment before discharge. This creates consequentialist harms for the individual (Vayena et al., 2018) which are to be balanced with the benefits enjoyed by society in general (Douglas et al., 2017). Depending on the availability of further risk-assessment by clinicians, the results reported at the 0.1 threshold, could also be useful. This approach allows the identification of 81% of forensic psychiatric patients, though resulting in a precision score of 18%.

These reflections pertain to the still exploratory nature of setting an appropriate risk threshold. A range of different individual and societal outcomes are to be considered, adopting the usefulness of data-driven approaches, while not neglecting ethical concerns of privacy (Starke et al., 2021).

Another quality of the proposed model is the fact that the data foundation contains only already available, high-quality administrative registry data, arguably of a less invasive nature than bio metric information or results from questionnaires (Viljoen et al., 2010). Recently, scholars have argued how high-level administrative data types can, in some instances, perform equally well to personality- and behavioral data (Bjerre-Nielsen et al., 2021). These authors also suggest that incorporation of more privacy-intense data types may in fact be redundant.

When risk-assessment tools are to be implemented in psychiatric services, the opportunity to avoid highly invasive data types is beneficial in order to address privacy concerns.

5. Limitations

A number of limitations to the proposed approach deserve to be highlighted. First of all, our models are only designed to flag patients who committed crime leading to the issuing of a COPA. As such, the tool is unable to target the whole population of individuals committing any crime after a psychiatric treatment course. We choose to limit our sample to recipients of COPA, as one can only receive such a verdict if mentally incapacitated in the act of crime. This increases the probability that the subsequent act of crime is related to the mental state of the individual and thus that additional treatment in the psychiatric system could have avoided this incidence. If all acts of crime were included, it is possible that some or many of them would be completely unrelated to the mental state of the individual and additional treatment futile. In the same vein, the inclusion criterion for the study sample were psychiatric in- or out-patient treatment in 2014. Therefore, the tool is also unable to target the whole forensic psychiatric population, as some COPA-recipients in the years 2014–2015 did not have contact with the psychiatric system in 2014.

A overarching assumption of the chosen approach is further that any patient who is assessed to be in high-risk of forensic psychiatric

### Table 2

| Model | Sampling | F1   | Recall | Precision (PPV) | NPV | Spec. | Acc. | Balanced acc. | AUROC |
|-------|----------|------|--------|-----------------|-----|------|-----|---------------|-------|
| LR    | RO       | 0.62 | 0.19   | 0.4             | 0.99| 0.99 | 0.98| 0.59           | 0.77  |
| RF    | RO       | 0.72 | 0.31   | **0.83**        | 0.99| 0.99 | 0.99| 0.66           | 0.94  |
| XG    | RO       | 0.74 | 0.37   | 0.71            | 0.99| 0.99 | 0.99| 0.68           | 0.94  |
| LightGBM | RO | 0.75 | 0.45   | 0.6             | 0.99| 0.99 | 0.99| **0.72**       | **0.95** |

### Table 3

| Model | Thr | F1   | Recall | Precision (PPV) | NPV | Spec. | Acc. | Balanced acc. |
|-------|-----|------|--------|-----------------|-----|------|-----|---------------|
| LightGBM | 0.1 | 0.64 | 0.81   | 0.18            | 0.99| 0.95 | 0.95| 0.88          |
|       | 0.3 | 0.72 | 0.58   | 0.37            | 0.99| 0.99 | 0.98| 0.78          |
|       | 0.5 | 0.76 | 0.47   | 0.57            | 0.99| 0.99 | 0.99| 0.73          |
|       | 0.7 | 0.73 | 0.34   | 0.76            | 0.99| 0.99 | 0.99| 0.67          |
|       | 0.9 | 0.66 | 0.2    | 0.96            | 0.99| 0.99 | 0.99| 0.6            |
treatment is offered first additional assessment by clinicians and, if necessary, prolonged treatment opportunities. Scholars argue that risk-assessment without appreciable benefits for the individual is unethical (Douglas et al., 2017). As such, the ethically valid implementation of statistical risk-assessment techniques rests on sufficient opportunities for both additional clinical screening and treatment. Additionally, while the model remains a valuable first attempt to assess risk in the psychiatric system, external validation studies are needed. New developments in computational data methods can likely expand the current performance level of models. For instance, advanced techniques to overcome the problem of class-imbalance in psychiatric data have been proposed in the context of predicting in-patient readmission (Du et al., 2021a, 2021b) an outcome which is comparable in terms of complexity to the one presented here. Additionally, deep learning models used extensively elsewhere in mental healthcare research (Su et al., 2020) also constitute a promising quantitative trajectory for new studies. Such contributions are left for future research to address.

An additional factor which should be adequately studied before implementation is the effect of algorithmic assessments on clinicians decision-making process. As the valid application of this and related tools assumes that human decision-making is never subdued, it is crucial for these contributions to be explored in detail. While initial studies suggest that clinicians values and trusts risk-assessment from algorithms in a forensic psychiatric context (Zhong et al., 2021) more detailed and comprehensive qualitative studies are needed.

**Author statement**

**Martin Lund Trinhammer**: Conceptualization, Methodology, Formal analysis, Writing - Original Draft, Writing – Review & Editing.  
**Ane Cathrine Holst Merrild**: Methodology, Software, Data Curation, Formal analysis, Writing - Original Draft.  
**Jonas Færch Lotz**: Methodology, Software, Supervision, Writing – Review & Editing.  
**Guido Makransky**: Supervision, Validation.

**Acknowledgements**

The ROCKWOOL Foundation has supplied access to data from Statistics Denmark. The ROCKWOOL Foundation has no financial interests in this research.

**Appendix**

**Predictors**

| Socio-Economic | Type |
|----------------|------|
| Sex            | Binary (male - 1, female - 0) |
| Age_20s        | Binary |
| Age_30s        | Binary |
| Age_40s        | Binary |
| Age_50s        | Binary |
| Ethnicity_dk   | Binary |
| Single         | Binary |
| Urban          | Binary |
| Unemployed     | Binary |
| Education: max. 10th grade | Binary |
| Education: max. gym | Binary |
| Education: max. vocational | Binary |
| Education: max. Short | Binary |
| Education: max. Medium | Binary |
| Education: max. Long | Binary |
| Education: No Education | Binary |

**Psychiatric history**

| Type |
|------|
| Sum of contacts with the psychiatric system | Numeric |
| Total number of days hospitalised | Numeric |
| Total number of outpatient treatments | Numeric |
| Number of times diagnosed with: Mental retardation | Numeric |
| Number of times diagnosed with: Autism spectrum disorder | Numeric |
| Number of times diagnosed with: Substance abuse disorder | Numeric |
| Number of times diagnosed with: Schizophrenia | Numeric |
| Number of times diagnosed with: Organic disorders | Numeric |
| Number of times diagnosed with: Anxiety disorders | Numeric |
| Number of times diagnosed with: Affective disorders | Numeric |
| Number of times diagnosed with: Eating disorders | Numeric |
| Number of times diagnosed with: Personality disorders | Numeric |
| Number of times diagnosed with: Hyperkinetic disorder | Numeric |
| Number of times diagnosed with: Other | Numeric |
| Number of times diagnosed with: Selfharm | Numeric |
| Total amount of unique diagnoses received | Numeric |

**Criminal history**

| Type |
|------|
| Total number of convictions received | Numeric |
| Total number of criminal offences | Numeric |
| Total number of prison sentences | Numeric |
| Total number of days convicted to unconditional prison | Numeric |
| Total number of days convicted to conditional prison | Numeric |
| Previous Treatment Sentence | Binary |

**Response variable**

| Type |
|------|
| Treatment Sentence | Binary |
Hyperparameters

Table 5

| Model | Hyperparameter       | Range             | Function                           |
|-------|----------------------|-------------------|------------------------------------|
| RF    | Estimators           | Fixed at 500      | No. of trees in the forest         |
|       | Max depth            | [5,50], no upper limit | Depth of individual trees          |
|       | Max features         | [2,30]            | No. of features for split          |
|       | Min samples leaf     | [1,5]             | No. obs. in leaf nodes             |
|       | Min samples split    | [2,5]             | No. obs. in splitting-nodes        |
| XGBoost| Eta                  | [-7,1]            | Contribution from each estimator   |
|       | Subsample            | [0.2,1]           | Fraction of observations for each tree |
|       | Min child weight     | [0.5,5]           | Min. sum of weight in child        |
|       | Column by tree       | [0.2,1]           | Fraction of column for each tree   |
| LightGBM| Num. leaves     | [3,6,9,12]        | No. of leaves                      |
|       | Min child samples    | [5,10,15,20]      | Min. samples in each child         |
|       | Max depth            | [1,10]            | Depth of individual trees          |

Risk thresholding

Table 6

| Undersampling | Thr. | F1   | Recall | PPV  | NPV  | Specificity | Acc.  | Balanced acc. |
|---------------|------|------|--------|------|------|-------------|-------|---------------|
| LR            | 0.1  | 0.58 | 0.51   | 0.11 | 0.99 | 0.95        | 0.94  | 0.73          |
|               | 0.3  | 0.64 | 0.36   | 0.35 | 0.99 | 0.99        | 0.98  | 0.67          |
|               | 0.5  | 0.63 | 0.25   | 0.28 | 0.99 | 0.99        | 0.98  | 0.62          |
|               | 0.7  | 0.63 | 0.22   | 0.36 | 0.99 | 0.99        | 0.99  | 0.61          |
|               | 0.9  | 0.64 | 0.2    | 0.47 | 0.99 | 0.99        | 0.99  | 0.62          |
| RF            | 0.1  | 0.57 | 0.82   | 0.11 | 0.99 | 0.92        | 0.92  | 0.87          |
|               | 0.3  | 0.72 | 0.64   | 0.34 | 0.99 | 0.98        | 0.98  | 0.81          |
|               | 0.5  | 0.75 | 0.44   | 0.61 | 0.99 | 0.99        | 0.99  | 0.72          |
|               | 0.7  | 0.7  | 0.25   | 0.88 | 0.99 | 0.99        | 0.99  | 0.62          |
|               | 0.9  | 0.58 | 0.09   | 1.00 | 0.99 | 1.00        | 0.99  | 0.54          |
| XG            | 0.1  | 0.65 | 0.72   | 0.2  | 0.99 | 0.96        | 0.96  | 0.84          |
|               | 0.3  | 0.72 | 0.54   | 0.39 | 0.99 | 0.99        | 0.98  | 0.77          |
|               | 0.5  | 0.74 | 0.44   | 0.54 | 0.99 | 0.99        | 0.99  | 0.72          |
|               | 0.7  | 0.73 | 0.37   | 0.64 | 0.99 | 0.99        | 0.99  | 0.68          |
|               | 0.9  | 0.72 | 0.29   | 0.87 | 0.99 | 0.99        | 0.99  | 0.65          |
| LightGBM      | 0.1  | 0.64 | 0.81   | 0.18 | 0.99 | 0.95        | 0.95  | 0.88          |
|               | 0.3  | 0.72 | 0.58   | 0.37 | 0.99 | 0.99        | 0.98  | 0.78          |
|               | 0.5  | 0.76 | 0.47   | 0.57 | 0.99 | 0.99        | 0.99  | 0.73          |
|               | 0.7  | 0.73 | 0.34   | 0.76 | 0.99 | 0.99        | 0.99  | 0.67          |
|               | 0.9  | 0.66 | 0.2    | 0.96 | 0.99 | 0.99        | 0.99  | 0.6            |

Table 7

| Oversampling | Thr. | F1   | Recall | PPV  | NPV  | Specificity | Acc.  | Balanced acc. |
|--------------|------|------|--------|------|------|-------------|-------|---------------|
| LR           | 0.1  | 0.58 | 0.39   | 0.12 | 0.99 | 0.96        | 0.96  | 0.68          |
|              | 0.3  | 0.62 | 0.23   | 0.28 | 0.99 | 0.99        | 0.98  | 0.61          |
|              | 0.5  | 0.62 | 0.19   | 0.4  | 0.99 | 0.99        | 0.99  | 0.59          |
|              | 0.7  | 0.6  | 0.14   | 0.37 | 0.99 | 0.99        | 0.99  | 0.57          |
|              | 0.9  | 0.59 | 0.12   | 0.45 | 0.99 | 0.99        | 0.99  | 0.56          |
| RF           | 0.1  | 0.6  | 0.79   | 0.14 | 0.99 | 0.94        | 0.94  | 0.86          |
|              | 0.3  | 0.75 | 0.46   | 0.54 | 0.99 | 0.99        | 0.99  | 0.73          |
|              | 0.5  | 0.72 | 0.31   | 0.83 | 0.99 | 0.99        | 0.99  | 0.66          |
|              | 0.7  | 0.65 | 0.18   | 1.00 | 0.99 | 1.00        | 0.99  | 0.59          |
|              | 0.9  | 0.51 | 0.02   | 1.00 | 0.99 | 1.00        | 0.99  | 0.51          |
| XG           | 0.1  | 0.71 | 0.51   | 0.38 | 0.99 | 0.99        | 0.98  | 0.75          |
|              | 0.3  | 0.72 | 0.38   | 0.57 | 0.99 | 0.99        | 0.99  | 0.69          |
|              | 0.5  | 0.74 | 0.37   | 0.71 | 0.99 | 0.99        | 0.99  | 0.68          |
|              | 0.7  | 0.73 | 0.33   | 0.8  | 0.99 | 0.99        | 0.99  | 0.66          |
|              | 0.9  | 0.7  | 0.26   | 0.88 | 0.99 | 0.99        | 0.99  | 0.63          |
| LightGBM     | 0.1  | 0.64 | 0.79   | 0.19 | 0.99 | 0.96        | 0.95  | 0.88          |
|              | 0.3  | 0.75 | 0.64   | 0.41 | 0.99 | 0.99        | 0.99  | 0.82          |

(continued on next page)
Table 7 (continued)

| Over sampling | Thr. | F1     | Recall | PPV | NPV | Specificity | Acc. | Balanced acc. |
|---------------|------|--------|--------|-----|-----|-------------|------|---------------|
| 0.5           | 0.75 | 0.45   | 0.6    | 0.99| 0.99| 0.99        | 0.99 | 0.99          |
| 0.7           | 0.73 | 0.31   | 0.88   | 0.99| 0.99| 0.99        | 0.99 | 0.99          |
| 0.9           | 0.66 | 0.2    | 1      | 0.99| 1   | 0.99        | 0.99 | 0.6           |

References

Bjørner-Nielsen, A., Kassamik, V., Lassen, D.D., Lehmann, S., 2021. Task-specific information outperforms surveillance-style big data in predictive analytics. Proc. Natl. Acad. Sci. USA 118 (14). https://doi.org/10.1073/pnas.2002581118.

Borsov, V., Leemann, T., Settler, K., Haag, J., Paweleczky, M., Kannei, G., 2021. Deep Neural Networks and Tabular Data: A Survey. https://doi.org/10.48550/arrow/2110.01889. Preprint arXiv:2110.01889.

Breiman, L., 2001. Random Forests. Mach. Learn. 45 (1), 5–32.

Buda, M., Maki, A., Mazurkowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Network. 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011.

Chen, Q., Zhang, James, Y., Barnett, E.J., Lichtenstein, P., Jokinen, J., D’Onofrio, B.M., et al., 2020. Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: a machine learning study using Swedish national registry data. PLoS Med. 17 (11) https://doi.org/10.1371/journal.pmed.1004316.

Du, G., Zhang, J., Li, S., Li, C., 2021a. Learning from class-imbalance and heterogeneous data for 30-day hospital readmission. Neurocomputing 420, 27–35. https://doi.org/10.1016/j.neucom.2020.08.064.

Du, G., Zhang, J., Ma, F., Zhao, M., Lin, Y., Li, S., 2021b. Towards graph-based class-imbalance learning for hospital readmission. Expert Syst. Appl. 176, 114791 https://doi.org/10.1016/j.eswa.2021.114791.

Fernández, A., García, S., Galcer, M., Prats, R.C., Krawczyk, B., Herrera, F., 2018. Learning from Imbalanced Data Sets, vol. 10. Springer, Berlin.

Fernández-Delgado, M., Ceradas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15 (1), 3133–3181.

Ford, L., 2015. First-episode psychosis in the criminal justice system: identifying a critical intercep for early intervention. Harv. Rev. Psychi. 23 (3), 167–175.

Frankle, J., Speiser, O., Dudek, M., Streb, J., 2020. Clinical ethics support services are not well established in forensic psychiatry as in general psychiatry. Front. Psychiatr. 11, 186. https://doi.org/10.3389/fpsyt.2020.00186.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, vol. 2. Springer, New York, pp. 1–758.

Jansman-Hart, E.M., Seto, M.C., Crocker, A.G., Nicholls, T.L., Côté, G., 2011. International trends in demand for forensic mental health services. Int. J. Forensic Ment. Health 10 (4), 326–336. https://doi.org/10.1080/14999013.2011.625591.

Kalapos, M.P., 2016. Penrose’s law: methodological challenges and call for data. Int. J. Law Psychiatr. 49, 1–9. https://doi.org/10.1016/j.ijlp.2016.04.006.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Liu, T.Y., 2017. Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30, 3138–3146.

Kircheiner, J., Günter, M.P., Lau, S., 2020. Identifying influential factors distinguishing recidivists among offender patients with a diagnosis of schizophrenia via machine learning algorithms. Forensic Sci. Int. 315, 110435. https://doi.org/10.1016/j.foresci.2021.110435.

Lamb, H.R., 2015. Does deinstitutionalization cause criminalization?: the penrose hypothesis. JAMA Psychiatr. 72 (2), 105–106.

Large, M.M., Ryan, C.J., Singh, S.P., Paton, M.B., Nielsen, O.B., 2011. The predictive value of risk categorization in schizophrenia. Harv. Rev. Psychiatr. 19 (1), 25–33.

Lewis, D.D., Gale, W.A., 1994. A Sequential Algorithm for Training Text Classifiers. Springer, London, pp. 3–12. SIGIR 94.

Lin, Y., Jeon, Y., 2006. Random forests and adaptive nearest neighbors. J. Am. Stat. Assoc. 101 (474), 578–590. https://doi.org/10.1198/016214506000001230.

Murray, J., Thompson, M.E., 2010. Clinical judgement in violence risk assessment. Eur. J. Psychol. 6 (1), 128–149. https://doi.org/10.3964/j.iejpy.2010.6.1.175.

Opitz, J., Burs, S., 2019. Macro F1 and Macro F1 arXiv preprint arXiv:1911.03347.

Passos, L.C., Mwangi, B., Kapczinski, F., 2019. Personalized Psychiatry: Big Data Analytics in Mental Health. Springer International Publishing.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Soyka, M., Graz, C., Bottender, R., Dirschfeld, P., Schoeh, H., 2007. Clinical correlates of later violence and criminal offences in schizophrenia. Schizophr. Res. 94 (1–3), 89–98. https://doi.org/10.1016/j.schres.2007.03.027.

Starke, G., De Clercq, E., Borgwardt, S., Elger, B.S., 2021. Computing schizophrenia: ethical challenges for machine learning in psychiatry. Psychol. Med. 51 (15), 2515–2521. https://doi.org/10.1017/s0033291721000168.

Su, C., Xu, Z., Pathak, J., Wang, F., 2020. Deep learning in mental health outcome research: a scoping review. Tranl. Psychiatry 10 (1), 1–26.

Vayena, E., Blasimme, A., Cohen, I.G., 2018. Machine learning in medicine: addressing ethical challenges. PLoS Med. 15 (11) https://doi.org/10.1371/journal.pmed.1002668.

Vijoen, J.L., McLachlan, K., Vincent, G.M., 2010. Assessing violence risk and psychopathy in juvenile and adult offenders: a survey of clinical practices. Assessment 17 (3), 377–395. https://doi.org/10.1177/1053810310372000183.

Wang, C., Deng, C., Wang, S., 2020. Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. Pattern Recogn. Lett. 136, 190–197. https://doi.org/10.1016/j.patrec.2020.05.035.

Watts, D., Moulden, H., Mamak, M., Upfold, C., Chaimowitz, G., Kapczinski, F., 2021. Predicting offenses among individuals with psychiatric disorders—A machine learning approach. J. Psychiatr. Res. 138, 146–154. https://doi.org/10.1016/j.jpsychires.2021.03.026.

Wolf, A., Fanshawe, T.R., Sariaslan, A., Cornish, R., Larsson, H., Fazel, S., 2018. Prediction of violent crime on discharge from secure psychiatric hospitals: a clinical prediction rule (FoVox). Eur. Psychiatr. 47, 88–93. https://doi.org/10.1016/j.eurpsy.2017.07.011.

Zhong, S., Yu, R., Cornish, R., Wang, X., Fazel, S., 2021. Assessment of violence risk in 440 psychiatric patients in China: examining the feasibility and acceptability of a novel and scalable approach (FoVox). BMC Psychiatr. 21 (1), 1–9. https://doi.org/10.1186/s12888-021-03115-3.