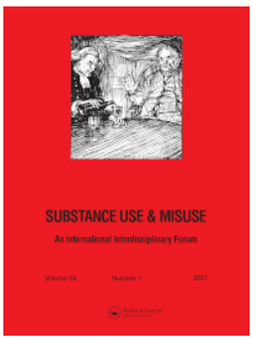


Substance Use & Misuse

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/isum20

Feature Extraction for Heroin-Use Classification Using Imbalanced Random Forest Methods

Matthew Beattie & Charles Nicholson

To cite this article: Matthew Beattie & Charles Nicholson (2021) Feature Extraction for Heroin-Use Classification Using Imbalanced Random Forest Methods, Substance Use & Misuse, 56:1, 123-130, DOI: 10.1080/10826084.2020.1843058

To link to this article: https://doi.org/10.1080/10826084.2020.1843058

Published online: 12 Nov 2020.

SUBSTANCE USE & MISUSE
2021, VOL. 56, NO. 1, 123–130
https://doi.org/10.1080/10826084.2020.1843058

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

Feature Extraction for Heroin-Use Classification Using Imbalanced Random Forest Methods

Anonymous for double blind review

Matthew Beattie and Charles Nicholson

Data Science and Analytics, University of Oklahoma, Norman, Oklahoma, USA

ABSTRACT

Background and aims: The National Survey on Drug Use and Health (NSDUH) contains a large number of responses and many features. This study aims to identify features from within NSDUH that are important in classifying heroin use. Proper implementation of random forest (RF) techniques copes with the highly imbalanced nature of heroin usage among respondents to identify features that are prominent in classification models involving nonlinear combinations of predictive variables. To date, methods for the proper application of RF to imbalanced medical datasets have not been defined. *Methods:* Three different RF classification techniques are applied to the 2016 NSDUH. The techniques are compared using scoring criteria, including area under the precision recall curve (AUPRC), to identify the best model. Variable importance scores (VIS) are checked for stability across the three models and the VIS from the best model are used to highlight features and categories of features that most influence the classification of heroin users. *Findings:* The best performing method was RF with random oversampling (AUPRC = 0.5437). The category of features regarding other drug use was most important (average *z*-scored VIS = 1.66) followed by age-of-first-use features (0.32). The most important individual feature was cocaine usage (*z*-scored VIS = 11.05), followed by crack usage (6.51). The most important individual feature other than specific drug use flags was the use of marijuana under the age of 18 (3.11). This study demonstrates a method for the use of RF in feature extraction from imbalanced medical datasets with many predictors.

KEYWORDS
NSDUH; heroin; random forest; imbalanced data

Introduction

The opioid epidemic

Over a century after the introduction of heroin as a medicine, opioid use, including heroin, has seen a marked increase, and with that increase has come a dramatic rise in overdose deaths. As shown in Figure 1, in 1999, there were a total of 8,050 opioid-related deaths in the United States, a mortality rate of 2.88 per 100,000 persons. By 2017, that figure had peaked at 47,600, a mortality rate of 14.62 per 100,000 persons, with the steepest rise having begun in 2013.[1]

The rise in opioid deaths has progressed in three overlapping phases (U.S. Centers for Disease Control, n.d.). The first wave began with increased prescribing of opioids such as oxycodone and hydrocodone in the 1990s. Abuse deterrent formulas of opioids helped to reduce their misuse (Wilkerson et al., 2016), but other abuse patterns arose. The second wave began in 2010 as heroin usage increased and caused a steep increase in overdose fatalities. The third wave began in 2013 and has seen a dramatic rise in deaths due to synthetic opioids such as fentanyl, which are often mixed with heroin without the knowledge of the user. Relative rarity of heroin use creates difficulty for researchers because even vast datasets are highly imbalanced between heroin users and non-users.

The National Survey on Drug Use and Health (NSDUH) is a dataset published annually by the Substance Abuse and Mental Health Services Administration (SAMHSA). This data is the result of extensive surveys tracking substance abuse and mental health in the United States since 1971 (U.S. Dept. of Health and Human Services, Substance Abuse and Mental Health Svcs. Admin., Ctr. for Behavioral Health Statistics and Quality, 2018). Heroin use among the adult respondents in the 2016 NSDUH data account for only 940 out of 42,625 observations. This extreme imbalance, in which only 2.2% of the cases are positively identified as heroin users, creates difficulties for statistical exploration of the phenomenon.

This study uses machine learning techniques to identify important predictors of heroin use from a large dataset containing a massive number of features. The investigation also closes gaps in feature extraction methods associated with highly imbalanced data.

CONTACT Matthew Beattie mjbeattie@ou.edu Data Science and Analytics, University of Oklahoma, 202 West Boyd St, Norman, OK 73019-0390, USA.

[1]Based upon US population of 279.0 million in 1999 and 325.5 million in 2017 (U.S. Census Bureau, 2017, 2020).

Traditional studies regarding opioid abuse

Jones (2013) used the NSDUH data to show that heroin users tend to be non-Hispanic whites, have used cocaine or nonprescription opioids (NPOs) within the last year, live in larger cities, and have either no health insurance or rely on Medicaid. In a subsequent study, Jones et al. (2015) confirmed many of these factors while also showing a relationship between cocaine use, binge drinking, marijuana use, and other factors. Cerdá et al. (2015) used discrete time hazard models to evaluate NSDUH data across several years and found that heroin initiation is strongly related to prior abuse of NPOs. These studies follow the approach used by much drug-use research. The authors select a potential set of features and use traditional approaches, including *t*-tests and multivariable logistic regression, to determine the strength of the relationship between a predicted variable and a set of predictor features. While this approach is standard for hypothesis testing of a priori predictions, it is less effective in exploratory analyses, where many features and complex combinations of features need to be assessed for potential relevance. Without a robust exploration of a massive dataset such as NSDUH, researchers can miss unforeseen features and feature combinations that are important (Piper et al., 2011).

Random forest studies and medical data

Classification trees are effective tools for feature selection (Kuhn & Johnson, 2013), and the random forest (RF) method improves upon their performance (Breiman, 2001). Variable importance scores (VIS) from RF are used to identify features that drive classification. There have been a very limited number of prior studies using RF to explore opioid use disorders from NSDUH data. Han et al. (2020) compared several machine learning techniques to demonstrate the viability of predicting adolescent opioid misuse and found that RF performed best. They found that marijuana and tobacco use were important predictors. Wadekar (2020) used an RF method designed for use with imbalanced datasets and found that marijuana use before the age of 18 and mental illness were the two most important features in predicting opioid use disorder.

Improvements can be made upon both of these studies regarding the application of RF techniques. Han et al. (2020) did not make use of RF methods that are designed for imbalanced datasets, and the performance of its model as measured by the area under the precision recall curve (AUPRC) was only 0.172 – far below the optimum. Wadekar (2020) used the area under the receiver operating characteristic curve (AUROC) as the reported metric in their work, even though AUROC should not be used for highly imbalanced data (Davis & Goadrich, 2006). Moreover, both studies limit the features included in RF models based upon assumptions of the authors.

Other studies have explored the use of RF and classification rules on imbalanced medical datasets (Khalilia et al., 2011; Mena & Gonzalez, 2006; Zhu et al., 2018), but we were unable to find any that combined techniques for imbalanced data, proper scoring criteria, and consideration of how feature selection can vary across methods.

Goals of this study

This study combines imbalanced RF methods, proper scoring criteria, and a comparison of VIS across multiple models to accomplish two goals. First, to demonstrate an improved method for the application of RF to imbalanced multifeatured medical datasets. Second, to uncover individual features and groups of them that are important in classifying heroin use without an a priori restriction to a few candidates.

Methods

Random Forest as an exploratory analysis method

The RF method enables classification of an outcome variable as a function of a set of statistical predictors. The predictors do not imply temporal precedence to the outcome. Instead, presence of a set of values among predictors in an observation indicates the outcome variable's membership in a given class. RF combines ensembles of classification trees with random selection of predictors used to partition data into positive and negative classes (Breiman, 2001). Since the resultant classifications derive from an aggregate of predictions from many trees, RF often generates more stable classifications even with large and complex data. However, such ensemble machine learning models do not lend themselves to direct explanation or quantification of the relationship between predictive features and response. RF provides VIS. Variable importance for a given feature in a RF is an empirically derived quantity associated with the reduction in misclassification probabilities attributed to the presence of that feature in the model. Features with high VIS indicate predictors that most influence the correct classification of an observation.

Imbalanced RF methods

An imbalanced dataset is one where at least one of the classes of the outcome variable represents a disproportionately small percentage of observations. Unmodified application of RF to imbalanced data minimizes the overall classification error rate but can focus on correctly classifying the majority class at the expense of the minority. Chen et al. (2004) describe modifications to RF to improve minority classification. One is the incorporation of cost-sensitive learning, known as *weighted RF*. A weight is assigned to each class with the highest weight assigned to the minority class. The weights act as a misclassification penalty, and the penalty for misclassifying a member of the minority class is higher than those of other classes. Another is through sampling, where either the majority class is undersampled or the minority class is oversampled. These strategies can also be combined. Chen et al. (2004) compared these techniques and found that weighted RF and balanced RF outperformed

other measures. Hasanin et al. (2019) found that random undersampling in combination with feature selection improved the classification of imbalanced data.

Scoring imbalanced RF results

A common metric for evaluating binary classification machine learning models is the AUROC. The AUROC plots the *true positive rate* (TPR) against the *false positive rate* (FPR) for all possible threshold values for a probabilistic classifier. These terms are now defined along with another important value, *positive predictive value* (PPV). A *true positive* (TP) is a positive case identified correctly by a classifier, whereas a *false positive* (FP) is a negative case misidentified by a classifier. Similarly, a *true negative* (TN) is a negative case identified correctly by a classifier, and a *false negative* (FN) is a positive case misidentified by a classifier. TPR, FPR, and PPV are defined in Equations (1)–(3).[2]

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{3}$$

For imbalanced datasets, AUROC can provide an overly optimistic view of model performance because the large number of negative cases biases FPR to small numbers. Davis and Goadrich (2006) point out that a better metric is the area under the precision-recall curve, which PPV on the *y*-axis and recall (S) on the *x*-axis. They also demonstrate that maximizing AUPRC will maximize AUROC, but the reverse is not always true.

Boyd et al. (2012) show that the upper bound for AUPRC can be less than 1.0, which is always the upper bound for AUROC, and depends on the degree of imbalance in the data. However, this does not mean that very low AUPRC values are necessarily acceptable. The maximum AUPRC of heroin use classification in the 2016 adult NSDUH dataset is 0.978.

Another summary metric for model performance is the *F-score*, which is given by:

$$F_\beta = \frac{(1+\beta^2)(\mathrm{PPV})(\mathrm{TPR})}{\beta^2(\mathrm{PPV}) + \mathrm{TPR}} \tag{4}$$

where $\beta > 0$ is a parameter used to select the relative importance of PPV and specificity. Most frequently, a lack of preference is used, and $\beta = 1$. In this case, the *F-score* is known as the *F1-score* – the harmonic mean of PPV and TPR. As a complement to AUPRC, we can characterize a model by its maximum *F1-score*, which ranges from 0 to 1:

$$F_1 = \frac{2(\mathrm{PPV})(\mathrm{TPR})}{\mathrm{PPV} + \mathrm{TPR}} \tag{5}$$

[2]In data science literature, PPV is commonly called *precision*, and TPR is referred to as *sensitivity* or *recall*.

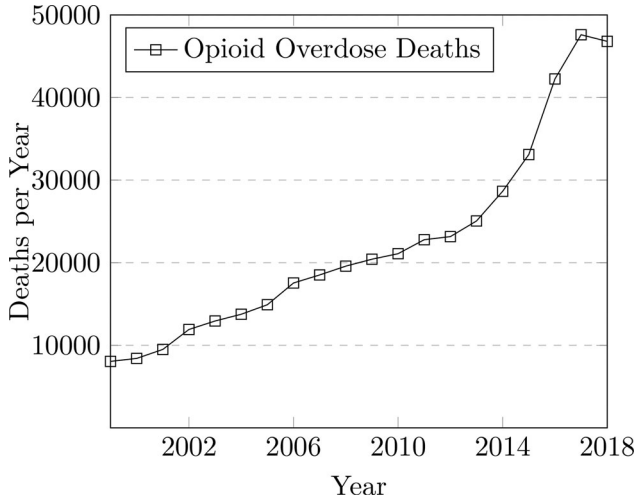


Figure 1. Opioid overdose deaths (Henry & Kaiser Family Foundation, 2020).

Interpreting variable importance scores

VIS can be unstable – multiple RF models can produce different VIS and different rankings of VIS results due to the stochasticity of the methods (Wang et al., 2016) or through biased selection of correlated variables or features with many different possible values (Hothorn et al., 2006). While there are methods for dealing with this instability that are beyond the scope of this study (Janitza et al., 2016), a literature search was unable to find studies that compare VIS across different imbalanced RF methods. It is likely that such variability does exist, and that researchers should compare scores between methods to determine similarities in outcomes. This work considers Pearson's correlation and Spearman's rank correlation coefficients to quantify the variable importance relationships between different models.

Data preparation

This study is performed on the 2016 National Survey on Drug Use and Health (NSDUH) dataset which contains 56,897 observations and 2,668 features (U.S. Dept. of Health and Human Services, Substance Abuse and Mental Health Svcs. Admin., Ctr. for Behavioral Health Statistics and Quality, 2018). Each observation in the dataset corresponds to a respondent, and the features of the observation are related to responses to questions in the survey. Features can be raw responses or can be fields that have been edited, recoded or imputed by the survey authors to improve the accuracy of feature values. This study adopts the NSDUH authors' recommendation that "where imputed or recoded variables are provided, users are encouraged to use them to produce estimates rather than raw or edited variables from the interview" (Center for Behavioral Health Statistics & Quality, 2018). To focus on adult respondents, all youth observations and variables that are exclusive to youths are removed. Redundant variables and metadata features such as imputation descriptions are also removed. Many features include legitimate missing values, indicating valid question skips. Others, such as height and weight, include invalid

Table 1. Feature categories.

Group name	Feature count	Category description
drugflag	29	Indicator of use of a substance
recency	56	Recency of use of a substance
afuvals	18	Age of first use of a substance
reasons	13	Reason for use of a substance
risks	17	Risks ascribed to the use of a substance
dependency	34	Current or past dependency on a substance
spectopics	4	DUI, parole, or other legal issues
treatment	33	Current or past treatment for substance abuse [a]
healthvars	34	Pregnancy, height, weight, and incidence of various health problems
mentalhealthsvc	45	Current or past treatment for mental health conditions
social	8	Behavioral patterns
mentalhealth	6	Imputed mental health status and suicidal behavior
depression	7	Imputed past or present depression
demographics	8	Sex, race, education level, and other factors
milfamily	5	Membership of a household with military members
employment	1	Current employment status
household	3	Number and ages of household members
insurance	1	Presence of health insurance
income	3	Presence of income assistance and level of income
geographic	1	Location in the United States
Indian	1	Residence in an American Indian area[b]

[a]Other than heroin.
[b]U.S. Census defined.

missings. In these cases, *k-nearest neighbors imputation* is used to impute values.

RFs tend to be biased toward selecting input features with many potential values for splits (Hothorn et al., 2006). To address this, variables such as height, weight, and emergency room visits are transformed via range binning into a series of categorical variables. The vast majority of all other features are 0/1 flags. With these transformations, all variables considered for feature selection are categorical. Variables specific to the use of heroin are discarded, and needle usage was restricted to the use of needles with substances other than heroin.

The features are grouped into categories based upon the sections of the NSDUH survey. The feature groups are listed in Table 1 by order of appearance in the survey. After this preparation, the dataset available for RF consists of 42,625 observations, 320 features, and 21 feature categories. All 320 features are included in the RF classification model. The 42,625 observations are split into a training set of 34,100 observations and a test set of 8,525 observations.

Imbalanced data random forest modeling

The outcome variable of the RF models is a label associated with heroin usage. This variable takes on a positive value if the respondent indicates that she has ever used heroin in her lifetime and a negative value if she indicates that she has not. The study considers lifetime usage of heroin rather than current month or current year usage because the data does not allow us to determine a sequence of lifetime events leading up to usage. The NSDUH survey does not include temporal precision for any category of correlates other than use of other drugs. Furthermore, considering lifetime usage provides ample opportunity to demonstrate the proposed algorithmic approach on highly imbalanced data.

The statistical predictors of the classification label are the 320 features from the data preparation phase. For example, one feature is *Marijuana used under 18*, which is a member of the *afuvals*, or *Age of first use of a substance* group. The model determines those features whose values are most important in correctly classifying a respondent as ever having used heroin. The training set is used to tune the hyperparameters of the classifiers. The test set is used to evaluate the performance of the optimized models.

RF models from three different imbalanced data techniques are created and coded. The first is *balanced RF* (BRF), in which samples for tree construction are created from observations by sampling from the minority class of labels, in this case positive heroin use, and matching the size of that sample with a random selection from the majority class, negative heroin use (Chen et al., 2004).

The second method is *random oversampling* (ROS) *with RF*, where multiple copies of members of the minority class are added to the dataset to match the number of majority class members. The oversampled dataset is then passed to RF (Pedregosa et al., 2011). The third method is *random oversampling with weighted* (ROSW) *RF*, in which we combine ROS with *weighted RF*. In weighted RF, weights are assigned to each class, with a higher value for the minority class. The weights are incorporated into the splitting criterion for tree induction and in the voting for class determination (Chen et al., 2004).

There are five hyperparameters for the BRF and ROS models, and six for ROSW. *Estimators* are the number of classification trees in the RF. *Minimum split samples* are the lowest number of samples in a leaf node required to consider it for a split. *Minimum leaf size* is the smallest number of samples allowed in a leaf node. *Maximum features* are the number of features allowed to generate a tree. *Maximum depth* is the number of splits allowed in a tree. ROSW includes *Minority class weight*, the penalty associated with misclassification of the minority class. Candidate hyperparameters are found using a random search. A range of hyperparameters bracketing the candidates is created and included in a grid search.

Table 2. Grid-search optimized hyperparameters.

Hyperparameter	BRF	ROS	ROSW
Estimators	1200	800	800
Minimum split samples	2	2	2
Minimum leaf size	2	2	2
Maximum features	50	20	20
Maximum depth	60	None	None
Minority class weight	NA	NA	60:1

For each candidate set of hyperparameters, the models are trained using three-fold cross validation with 15 repeats each. The scoring criterion for hyperparameter selection during training is average PPV. The best performing BRF, ROS, and ROSW models from the training phase are then applied to the holdout test set of 8,525 observations. A comparison of AUPRC and maximum F1-score across the three models determines the one that best classifies the test set.

Variable importance score comparison

Pearson and Spearman's coefficients (ρ_p, ρ_s) determine whether there are any linear correlations in VIS or the ranks of VIS between the three methods. Ideally, there will be high similarity in VIS between models with similar AUPRC scores, which would indicate that variable importance is stable across different methodologies. If not, the need to use additional techniques in determining the true relevance of predictive features exists.

The most important features are listed to show which have the highest impact in classification. The total and average VIS of feature categories is produced to see which groups are most influential in predicting heroin use.

Results

Hyperparameter selection

After tuning the models, we find that all three approaches, BRF, ROS, and ROSW, perform best with many underlying classification trees (at least 800 per forest) that are large (at least 60 levels). For the best ROSW model, the minority class was best weighted 60 times higher than the majority class. The final hyperparameters for the models are shown in Table 2.

Validation

The best models for each of the three methods are applied to the test dataset and their AUPRCs and F1-scores are compared to determine the best performing method. All three of the models perform similarly – the highest AUPRC, which is from the ROS model, is only 3.9% higher than that of the lowest, which is from the BRF model. Likewise, the maximum F1-score from the ROS model is only 6.5% higher than that of the BRF model. These metrics are summarized in Table 3, and the precision-recall curve for ROS, the best model, is shown in Figure 2. The curves for BRF and ROSW exhibit similar shapes to that of ROS (Figure 3).

Table 3. Model validation scores.

Metric	BRF	ROS	ROSW
AUPRC	0.5234	0.5437	0.5247
Maximum F1-score	0.5040	0.5369	0.5187
Optimal threshold	0.8511	0.2010	0.2100

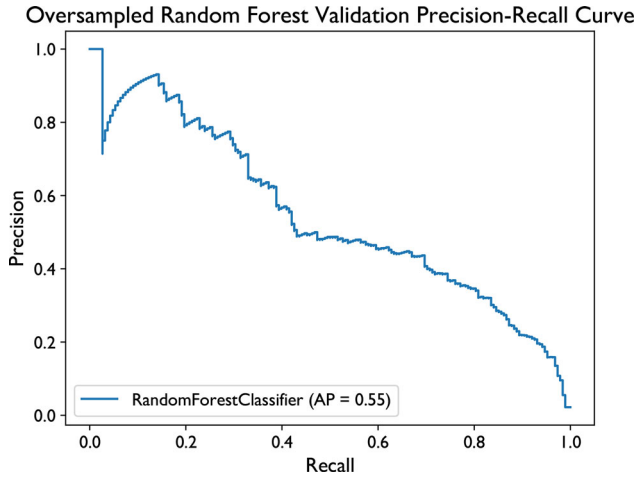


Figure 2. Precision-recall curve for random oversampled (ROS) RF.

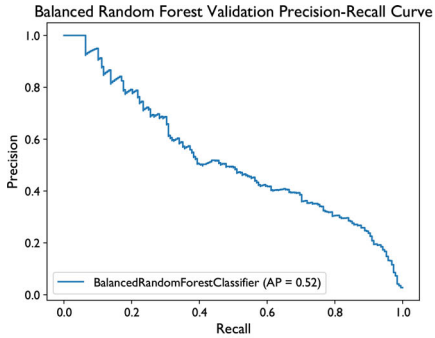
The F1-score curve for the ROS model (Figure 4) shows a skew toward low threshold parameters, with a peak of 0.5369 at the threshold value 0.2010. At this optimal threshold value, the ROS model correctly identifies 69.7% of the positive class observations, and incorrectly identifies only 2.0% of the negative ones (Table 4).

Variable importance scores

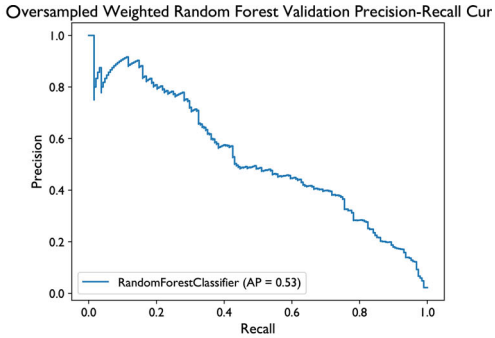
There are differences in the VIS between the three methods (Figure 5). BRF and ROS have a great deal of similarity, with $\rho_p = 0.96$, but ROSW differs somewhat significantly from the other two methods. A comparison of the ranks of the scores shows general agreement across the methods. BRF and ROS are nearly identical at $\rho_s = 0.99$, while the lowest correlation is between ROS and ROSW, $\rho_s = 0.89$.

Given the agreement in feature ranks across the three models, we focus our attention on the VIS from ROS, the best performing model. The top 20 most important features (Table 5) are dominated by drug use flags, and the VIS for cocaine-use is almost twice as high as the next most important feature. The most important non-drug-flag feature is Marijuana use under the age of 18. The first feature not directly related to substance use or availability to appear in the list is education level, the 31st ranked VIS with a z-scored importance of 0.24.

The results for total and average VIS by group are shown in Table 6. The total VIS for drug use scores is very high. This is in part due to the large number of features in this category, but is also driven by high individual scores within the category. Averaging VIS scores takes away the effect of the numbers of features within a group. Average VIS scores for drug use flags remain the highest among all groups, followed by those for age-of-first use features. Employment



(a) Balanced Random Forest



(b) Random Oversampling with Weighted Random Forest

Figure 3. Precision-recall curves for BRF and ROSW.

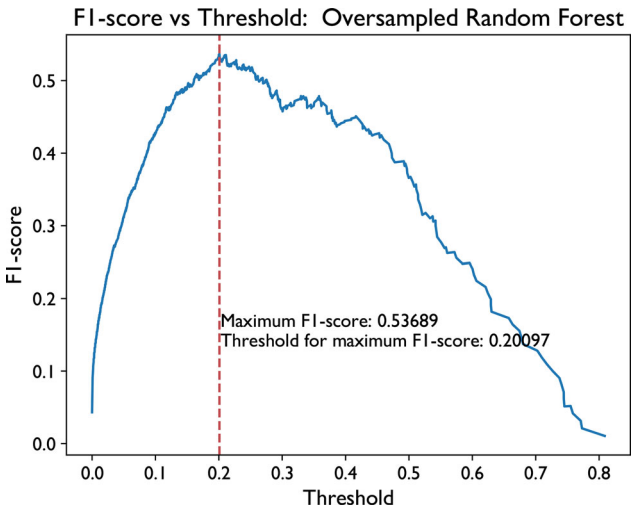


Figure 4. F1-score curve for random oversampled (ROS) RF.

Table 4. Confusion matrix for random oversampled (ROS) RF.

		Predicted heroin use	
		Negative	Positive
Actual heroin use	Negative	8168	169
	Positive	57	131

status, demographic features, and household size complete the top five groups.

Discussion

Process for using random forest on medical datasets

The BRF, ROS, and ROSW models have similar AUPRC performance, and the ROS model performs slightly better than the other two. It is a subjective matter to declare a model to be a good one based upon AUPRC. The ROS AUPRC score of 0.5437 is lower than the theoretical maximum of 0.978, but is much higher than that of Han et al. (2020). This is likely due to the fact that our models included many more features than Han et al. (2020), whose a priori reduced feature set lacked drug-use flags. A comparison to Wadekar (2020) is not possible because that study did not report AUPRC scores.

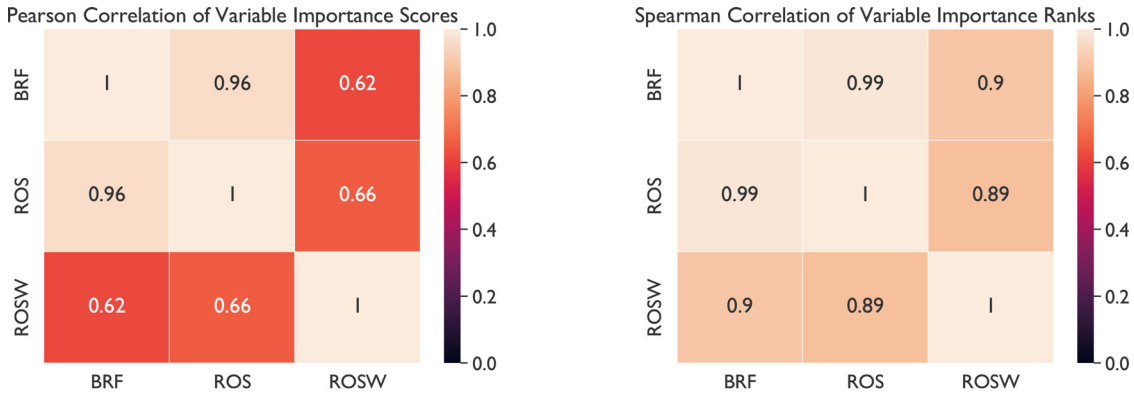
The models produce different sets of VIS. Two of the models, BRF and ROS, have very similar scores while ROSW has quite different ones. However, the feature VIS rank orders of all three models are similar, indicating a level of stability among the models as they combine predictors. Since VIS scores can vary between models, comparing the outcomes of different methods using the Spearman's rank correlation coefficient is a good practice. If the correlation between different outcomes is low, one should hesitate to claim the importance of a set of features.

Medical studies often involve imbalanced data, and the use of RF for feature extraction can be highly effective in such cases. RF must be used carefully on imbalanced data, yet instructions for how to do so don't appear to be in the literature. This study presents guidelines to follow when using RF on medical data:

- RF is a good method for classification or feature extraction in preparation for other methods, and there are specific algorithms and sampling techniques that should be used with imbalanced datasets.
- There appears to be no dominant imbalanced RF method, so comparing the performance and output of a few models should be done.
- Proper scoring criteria, notably AUPRC, must be used instead of AUROC when evaluating RF performance on imbalanced data.
- The instability of VIS should be measured by using Pearson and Spearman's correlation coefficients to compare the output of different RF techniques.

Important features in heroin use prediction

RF inherently addresses both linear and non-linear predictor-response relationships, generates accurate classifications, and produces VIS values that quantify the impact of the features on correct classification. RF reveals the importance of features that might otherwise be overlooked with traditional methods. For example, first use of cocaine under the age of 18 has a higher χ^2-test value ($\chi^2 = 3484$, $p = 0.0000$) than marijuana use under the age of 18 ($\chi^2 = 1502$, $p = 0.0000$), yet has a much lower z-scored VIS (1.03 vs 3.11). The RF model recognizes that combinations of features with under-18 marijuana use explain heroin use to a



(a) Variable score correlation (b) Variable rank correlation

Figure 5. Correlations of variable importance scores and ranks among three RF methods.

Table 5. Feature variable importance scores.

Feature	*z*-Scored VIS	Group name
Cocaine – ever used	11.05	drugflag
Crack – ever used	6.51	drugflag
Methamphetamine – ever used	4.30	drugflag
Pain reliever – ever misused	4.25	drugflag
LSD – ever used	3.79	drugflag
Psychotherapeutics – ever misused	3.40	drugflag
Marijuana used under 18	3.11	afuvals
Ecstasy – ever used	2.56	drugflag
Marijuana – ever used	2.21	drugflag
Cigarette use in past year	1.53	recency
Cigarette use in past month	1.24	recency
Hallucinogen used under 18	1.22	afuvals
Inhalants – ever used	1.04	drugflag
Cocaine used under 18	1.03	afuvals
Daily cigarette use under 18	0.97	afuvals
Heroin easy to obtain	0.94	risks
Cigarette used under 18	0.91	afuvals
PCP – ever used	0.86	drugflag

greater degree than combinations with under-18 cocaine use.

Substance use under the age of 18, particularly marijuana use, is *much* more important in classifying heroin use than any feature group other than drug-use flags. Wadekar (2020) reached the same conclusion regarding opioid use disorder. Wen et al. (2015) found that the implementation of medical marijuana laws increases marijuana initiation in subjects aged 12–20 years old. These findings should cause policy makers to carefully consider unintended consequences when extending the availability of drugs considered not overly risky, such as marijuana, in their communities.

Average VIS for feature categories better illustrates patterns in the data than consideration of individual features. In particular we find that non-heroin drug use is the most important feature category in the model. While this is not a surprising result, it is imperative to note that ignoring such a dominant feature category can lead to invalid models. There are non-a priori ways to factor out drug-use flags from classification models. One is to sequentially pull out features, such as cocaine use, that have been well studied and stop when model performance falls below an acceptable AUPRC threshold. This would allow one to not ignore the impact of other drug use while focusing our attention on other features.

Table 6. Average variable importance scores by group.

Group name	Total *z*-scored VIS	Average *z*-scored VIS	Category description
drugflag	46.39	1.66	Indicator of use of a substance
afuvals	5.76	0.32	Age of first use of a substance
employment	0.07	0.06	Current employment status
demographics	0.42	0.05	Sex, race, education level, and other factors
household	0.05	0.03	Number and age of household members
income	0.01	0.00	Presence of income assistance and level of income
geographic	−0.02	−0.02	Location in the United States
social	−.40	−0.05	Behavioral patterns
risks	−1.16	−0.07	Risks ascribed to the use of a substance
recency	−5.30	−0.11	Recency of use of a substance
healthvars	−3.80	−0.13	Pregnancy, height, weight, and incidence of various health problems
reasons	−2.10	−0.16	Reason for use of a substance
spectopics	−0.73	−0.18	DUI, parole, or other legal issues
insurance	−0.19	−0.19	Presence of health insurance
milfamly	−1.09	−0.22	Membership of a household with military members
depression	−1.53	−0.22	Imputed past or present depression
mentalhealth	−1.32	−0.22	Imputed mental health status and suicidal behavior
mentalhealthsvc	−12.50	−0.28	Current or past treatment for mental health condition
specuse	−2.02	−0.29	Needle and special drug use
dependency	−9.68	−0.29	Current or past dependency on a substance
treatment	−10.51	−0.32	Current of past treatment for substance abuse
indian	−0.32	−0.32	Residence in an American Indian area

Implications

Techniques such as RF have become easy to use due to modern computing and software capabilities. However, they cannot be blindly applied, especially to medical data. Medical datasets are commonly imbalanced, and this study provides a rigorous method to leverage RF for appropriate feature selection. Also, this study shows a relationship between marijuana use under the age of 18 and heroin use. This conclusion has been reached by other studies and

suggests a danger of marijuana use that needs to be better understood.

The high correlation between early use of other substances with lifetime heroin use suggests some intervention strategies. For example, care providers in rehabilitation facilities could consider requiring more frequent post-discharge testing or education on protocols relating to medical heroin intervention for patients who have a history of early age drug use.

Finally, further studies should focus on the temporal aspect of heroin use. With the fields in the NSDUH survey, a researcher can construct a sequence of the drugs used by a respondent prior to heroin use. Unfortunately, the structure of the survey does not allow for the sequencing of other factors, such as onset of health problems or the occurrence of a life event like job loss. Given the scale and regularity of the survey, researchers would benefit from the incorporation of questions that can help construct a timeline of factors leading to heroin use.

Disclosure of interest

The authors report no conflict of interest.

Funding

This study was self-funded by the authors.

ORCID

Matthew Beattie http://orcid.org/0000-0001-7623-6894
Charles Nicholson http://orcid.org/0000-0002-7023-8802

References

Boyd, K., Costa, V. S., Davis, J., & Page, C. D. (2012). Unachievable region in precision-recall space and its effect on empirical evaluation [Paper presentation]. Proceedings of the 29th International Conference on Machine Learning, ICML 2012, 1, 639–646.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Center for Behavioral Health Statistics and Quality (2018). *2016 National Survey on Drug Use and Health Public Use File Codebook*.

Cerdá, M., Santaella, J., Marshall, B. D., Kim, J. H., & Martins, S. S. (2015). Nonmedical prescription opioid use in childhood and early adolescence predicts transitions to heroin use in young adulthood: A national study. *The Journal of Pediatrics*, *167*(3), 605–612.e2. https://doi.org/10.1016/j.jpeds.2015.04.071

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. Department of Statistics, University of California, Berkeley, CA. Technical Report 666.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves [Paper presentation]. Proceedings of the 23rd International Conference on Machine learning – ICML '06, New York, New York, USA, ACM Press, pp. 233–240.

Han, D. H., Lee, S., & Seo, D. C. (2020). Using machine learning to predict opioid misuse among U.S. adolescents. *Preventive Medicine*, *130*, 105886. https://doi.org/10.1016/j.ypmed.2019.105886

Hasanin, T., Khoshgoftaar, T., Leevy, J., & Seliya, N. (2019). Examining characteristics of predictive models with imbalanced big data. *Journal of Big Data*, *6*(1), 69.

Henry, J, Kaiser Family Foundation (2020). *Opioid overdose deaths by race/ethnicity*. Retrieved from https://www.kff.org/other/state-indicator/opioid-overdose-deaths/

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. https://doi.org/10.1198/106186006X133933

Janitza, S., Tutz, G., & Boulesteix, A. L. (2016). Random forest for ordinal responses: Prediction and variable selection. *Computational Statistics & Data Analysis*, *96*, 57–73. https://doi.org/10.1016/j.csda.2015.10.005

Jones, C. (2013). Heroin use and heroin use risk behaviors among nonmedical users of prescription opioid pain relievers. United States, 2002 to 2004 and 2008 to 2010. *Drug and Alcohol Dependence*, *132*(1-2), 95–100. https://doi.org/10.1016/j.drugalcdep.2013.01.007

Jones, C., Logan, J., Gladden, R., & Bohm, M. (2015). *Morbidity and mortality weekly report vital signs: Demographic and substance use trends among heroin users – United States, 2002–2013*.

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, *11*(1), 51.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer New York.

Mena, L. & Gonzalez, J. A. (2006). Machine learning for imbalanced datasets: Application in medical diagnostic Vol. 34.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., & Grisel, O. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Piper, M., Loh, W.-Y., Smith, S., Japuntich, S., & Baker, T. (2011). Using decision tree analysis to identify risk factors for relapse to smoking. *Substance Use & Misuse*, *46*(4), 492–510. https://doi.org/10.3109/10826081003682222

U.S. Census Bureau (2017). *Intercensal estimates of the United States population by age and sex, 1990–2000*. Retrieved from https://www.census.gov/

U.S. Census Bureau (2020). *2017 National population projections tables: Main series*. Retrieved from https://www.census.gov/

U.S. Centers for Disease Control. (n.d.). *Understanding the epidemic, drug overdose (CDC injury center)*. Retrieved from https://www.cdc.gov/drugoverdose/epidemic

U.S. Dept. of Health and Human Services, Substance Abuse and Mental Health Svcs. Admin., Ctr. for Behavioral Health Statistics and Quality (2018). *National survey on drug use and health 2016 (NSDUH-2016-DS0001)*. Retrieved from https://datafiles.samhsa.gov/

Wadekar, A. S. (2020). Understanding opioid use disorder (OUD) using tree-based classifiers. *Drug and Alcohol Dependence*, *208*(January), 107839–107819.

Wang, H., Yang, F., & Luo, Z. (2016). An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinformatics*, *17*(1), 1–18. https://doi.org/10.1186/s12859-016-0900-5

Wen, H., Hockenberry, J., & Cummings, J. (2015). The effect of medical marijuana laws on adolescent and adult use of marijuana, alcohol, and other substances. *Journal of Health Economics*, *42*, 64–80.

Wilkerson, R., Kim, H., Windsor, T., & Mareiniss, D. (2016). The opioid epidemic in the United States. *Emergency Medicine Clinics of North America*, *34*(2), e1–e23.

Zhu, M., Xia, J., Jin, X., Yan, M., Cai, G., Yan, J., & Ning, G. (2018). Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access.*, *6*, 4641–4652. https://doi.org/10.1109/ACCESS.2018.2789428