International Gambling Studies

ISSN: 1445-9795 (Print) 1479-4276 (Online) Journal homepage: https://www.tandfonline.com/loi/rigs20

Identifying high-risk online gamblers: a comparison of data mining procedures

Kahlil S. Philander

To cite this article: Kahlil S. Philander (2014) Identifying high-risk online gamblers: a comparison of data mining procedures, International Gambling Studies, 14:1, 53-63, DOI: 10.1080/14459795.2013.841721

To link to this article: https://doi.org/10.1080/14459795.2013.841721

International Gambling Studies, 2014
Vol. 14, No. 1, 53–63, http://dx.doi.org/10.1080/14459795.2013.841721

Routledge
Taylor & Francis Group

Identifying high-risk online gamblers: a comparison of data mining procedures

Kahlil S. Philander*

William F. Harrah College of Hotel Administration, University of Nevada, Las Vegas, USA

(*Received 6 May 2013; accepted 2 September 2013*)

Using play data from a sample of virtual live action sports betting gamblers, this study evaluates a set of classification and regression algorithms to determine which techniques are more effective in identifying probable disordered gamblers. This study identifies a clear need for validating results using players not appearing in the original sample, as even methods that use in-sample cross-validation can show substantial differences in performance from one data set to another. Many methods are found to be quite accurate in correctly identifying player types in training data, but perform poorly when used on new samples. Artificial neural networks appear to be the most reliable classification method overall, but still fail to identify a large group of likely problem gamblers. Bet intensity, variability, frequency and trajectory, as well as age and gender are noted to be insufficient variables to classify probable disordered gamblers with arbitrarily reasonable accuracy.

Keywords: supervised learning algorithm; responsible gambling; classification; internet gambling

Introduction

The emergence of the Internet as a medium to provide gambling has transformed the way that operators can learn about their players' behaviour. Real-time monitoring of wagers, deposits, withdrawals and even mouse clicks is possible, which has opened a world of data mining possibilities. While early applications of this type of data has tended to focus on profit maximization similar to other forms of Internet commerce, the same data that allows operators to monitor behaviour and improve their product may also allow operators (and regulators) to better ensure their players' safety (Gainsbury, 2011).

This study expands on the behavioural identification work by Braverman and Shaffer (2012) and LaBrie and Shaffer (2011) by testing a set of supervised learning algorithms to determine which are effective in identifying probable disordered gamblers. By identifying the most reliable algorithms in problem gambler identification, or correspondingly eliminating the least useful algorithms, research focus can shift to deterministic data analysis, identifying more variables that may contribute to accurate player classification.

As a step in initial exploration of online player data, past studies have tended to focus on unsupervised clustering processes (e.g. Adami et al., 2013; Braverman & Shaffer, 2012; Dragicevic, Tsogas, & Kudic, 2011). While this information is important to understanding classes of player types, supervised learning algorithm (SLAs) may provide useful improvements in identification tasks, once behavioural markers of problem gamblers have been identified. SLAs, which can vary from explanatory models such as logistic regression to black box methods such as artificial neural networks, are more adept

*Email: kahlil.philander@unlv.edu

at pure risk prediction. They can factor in complex interrelationships and interactions between the independent variables to improve prediction. SLAs therefore reflect the class of algorithms that should be implemented in real-world responsible gambling programmes and software. However, empirical research has shown that even the best overall algorithms can be outperformed by others, depending on the particular type of data analysed (Caruana & Niculescu-Mizil, 2006). It is therefore important to specifically evaluate SLAs within the context of online player data.

Given the widespread adoption of online gambling in Europe, Australasia and other jurisdictions (Balestra & Krafcik, 2010; Philander & Abarbanel, 2013) and the recent interest in many North American jurisdictions in legalizing online gambling (Beasley & Chan, 2012), developing the capability to identify problematic gambling has become increasingly important. Foreign gaming operators, such as OLG (2011) of Canada or Svenska Spel (2010) of Sweden, have already made responsible gambling data analytics a priority in their online products. Developing a strong responsible gambling programme may also extend beyond an organizational (or governmental) desire for social responsibility. In the uncertain world of online gambling, responsible gaming measures have been shown to increase player trust in gaming sites (Wood & Griffiths, 2008), which may lead to increased patronage. Indeed, some sites have used their commitment to responsible gambling as part of their marketing campaigns, providing detailed information on their policies and research (e.g. bwin, 2012).

Empirical studies

In recent years, the rise in player loyalty programmes that track player activity has created an opportunity to study problem gambling by using actual gambling activity. Some behavioural characteristics that were used to diagnose/identify problem gambler, including chasing losses or increasing bet sizes (American Psychiatric Association, 2000; Ferris & Wynne, 2001; Lesieur & Blume, 1987), can be monitored in a manner that is more reliable than past research tools such as survey self-report data (Hodgins & Makarchuk, 2003; Volberg, Gerstein, Christiansen, & Baldridge, 2001). Osborn, Skelt, Delfabbro, Nevile, and McMillen (2007) and Schellinck and Schrans (2007) conducted major empirical studies that examined observable signs of problem gambling in gaming venues. Schellinck and Schrans (2007) for example, analysed actual VLT (slot) play behaviour through players' use of a responsible gambling player card. Smaller-scale studies have also been produced, e.g. Livingstone (2005) or Hing and Nuske (2011).

Although play activity from player cards was largely confined to slot machines in land-based venues, seminal literature in the Internet domain by LaBrie, Kaplan, LaPlante, Nelson, and Shaffer (2008), LaBrie, LaPlante, Nelson, Schumann, and Shaffer (2007) and LaPlante, Kleschinsky, LaBrie, Nelson, and Shaffer (2009) has prompted an expansion of online play analysis.

The research conducted with player tracking data has sought to provide some insight into disordered gambling by identifying and describing the behavioural markers of gamblers who showed a pattern of high involvement with gambling. According to a study of online sports betting (LaBrie et al., 2007), 1% of sports gamblers displayed betting tendencies and patterns that were distinct from the rest of the sample, including a high number of bets, bets per day, amount of money per bet, total wagered and/or net loss. LaBrie et al. (2008) identified a distinct group of heavily involved Internet casino gamblers. This group represented the top 5% of their population sample in terms of total wagered. They gambled for a longer period of time, played on more days during their

active account period, and played longer sessions than the rest of the sample. In an analysis of online poker, Fiedler (2012) has shown that 1% of players generate 60% of site revenue, and that these players may be either at-risk players or professional gamblers.

In a recent study by Braverman, LaBrie, and Shaffer (2011) that used the same online sports betting population as LaBrie et al. (2007), the researchers were unable to support the view that the most involved Internet sports gamblers were a distinct category of gamblers that were qualitatively different from more involved recreational gamblers. However, they did raise the possibility that other factors, such as play duration and frequency, may justify a separate categorization of gamblers.

Dragicevic et al. (2011) analysed behavioural markers for high-risk Internet gambling using a data set of online casino gamblers. The researchers used k-means cluster analysis to identify four clusters of online gamblers that were distinguished by the same gambling risk factors studied by Braverman and Shaffer (2012). The factors were bet frequency, intensity, variability and trajectory. The authors concluded that the behaviour of highly intense gamblers can vary greatly and that problem gambling could be evident in three of the four clusters. Overall, they found it was difficult to assign any of the clusters to specific problem gambling clinical groups with a high degree of certainty. The researchers also stated that other statistical methodologies may be more effective than the k-means clustering process they use, and suggested that future research should explore more appropriate tools.

Using a step-wise logistic regression algorithm, Haefeli, Lischer, and Schwarz (2011) found that the relative number of email contacts per month and requests for account reopening are statistically significant predictors of self-exclusion by online gamblers at the $\alpha = 0.05$ level. Adami et al. (2013) expand on Braverman and Shaffer (2012), and found two new markers for identifying at risk gamblers – the 'sawtooth' oscillation between increasing wager size and rapids drops, and number of games played (gambling involvement). Adami et al. (2013) also note that the unsupervised k-means clustering process may be suboptimal.

This research suggests that there may be distinct behavioural profiles that allow differentiation between non-problem gamblers and patrons with gambling problems, and therefore there is an opportunity to employ responsible gambling intervention techniques using this data. However, the ability to use play data to predict problem gamblers with high reliability and validity has yet to be demonstrated in peer-reviewed literature. Most research analysing data mined gambling behaviour has focused on describing general behavioural tendencies and patterns amongst the total cohort, as opposed to providing evidence of pure prediction of risk.

In the next section, the SLAs used in this study are outlined, as well as their tuning parameters. Then, their comparative classification results are provided, including a discussion of their accuracy in training data and in hold-out testing data. Finally, a detailed discussion of the findings is provided in the conclusion, along with the implications for future player data and responsible gambling research.

Methodology

A set of nine supervised learning models are evaluated in this study to determine which data mining procedures are most effective in identifying probable disordered online sports gamblers. These SLAs are widely popular in the data mining literature. While a complete analysis of all classes of algorithms would be beyond the scope of any paper, the classes in this study provide a reasonable breadth of approaches to predictive modelling. They allow

for general conclusions to be drawn about their individual viability and general online player data modelling issues.

The SLAs and their related documentation are outlined in Table 1. A brief description of the SLA is provided, but for more detailed notes on the classification mechanism the reader is directed towards the source documentation by the SLA author. All of the algorithms are evaluated using the R computer language. R has been identified as the most popular software among data miners (Rexer, Allen, & Gearan, 2011), and also can be adapted as an input to low-level programming languages for real world application. The SLAs include logistic regression, regularized general linear models (GLM), artificial neural networks, support vector machines (SVM) and random forests. Where the models required parameter tuning, a wide range of values are provided and allowed to be optimized by the software (e.g. grid search) in order to reduce bias from researcher settings. These settings are specifically noted in Table 1.

These methods are all evaluated on the same analytic dataset that was used by Braverman and Shaffer (2012). The dataset is publicly available for research purposes from the Transparency Project, Division on Addictions (2012). The dataset includes a sample of virtual live action sports betting player accounts that fulfilled the following requirements:

- Opened an account with the Internet betting service provider *bwin* in February 2005
- Closed their account after 30 days, but before February 2007
- Had more than two active gambling days within the first month of account activation.

When closing their accounts, the 530 resulting players selected one of three available choices as a reason for the account closure: (1) having no further interest in gambling (48%); (2) being unsatisfied with the service (19%); or (3) due to gambling-related problems (33%). Using that information, a binary dependent variable was coded where '1' identifies the group of gamblers that closed their account due to gambling-related problems (self-exclude or SE) and '0' identifies the group of gamblers that closed their account due to any other reason (other). This approach treats the self-excluders as a proxy for problem/pathological gamblers, which is consistent with prior studies using this data set, but comes with validity limitations. The selection of this reason for account closure may indicate that the gambler has concerns about his/her play, but that he/she may not actually fit the clinical (or epidemiological) criteria for gambling-related problems. Similarly, account closers who selected other options may also be misclassified or simply selecting an option without due consideration to their response since they are ending their relationship with bwin. Account closers may also not provide a representative sample with which to generalize gamblers within the full player pool.

The supervised learning algorithms are provided with all variables in the data set with the exception of the predefined clusters from Braverman and Shaffer (2012), the player ID variable and the random variable. The remaining set of variables included gameplay variables, demographic variables (country, gender, and age at registration), and variables that were computed by Braverman and Shaffer (2012) and identified as predictive of likely problem gamblers: intensity, variability, frequency and trajectory.

The algorithms are trained on one sample from the dataset and validated on a randomly selected holdout sample to provide a better estimate of the SLAs predictive abilities. Approximately 70% of the sample is randomly selected into the training sample ($n = 369$) and the remaining 30% is used as the cross-validation testing sample ($n = 161$). More information on the data set is available in Braverman and Shaffer (2012), Dragicevic

Table 1. Algorithm overview and description.

Algorithm	References	Description	Other settings
Step-wise logistic regression	step: Ripley (2013)	Logistic regression using the general linear models function. A formula-based model is selected by AIC using a forward and backward step-wise search.	Maximum possible steps considered is set at 1000
Lasso/elastic-net logistic regression	glmnet: Friedman, Hastie, and Tibshirani (2009)	Lasso and elastic-net regularized generalized linear used to fit the entire lasso elastic-net regularization path for logistic regression models. The algorithm uses cyclical coordinate descent in a path-wise fashion.	10-fold cross validation used on training data; lambda chosen that provided the minimum mean cross-validated error
Neural network (regression)	nnet: Ripley (2009)	Fit single-hidden-layer neural network using a regression based categorization. A'1' denotes a person citing gambling problem and '0' denotes another reason for account closure with 0.5 as the cut-off level for prediction.	Model tuned by grid search: number of units in the hidden layer = [10, 25, 50]; parameters for weight decay [0, 0.2, 0.4]
Neural network (classification)	nnet: Ripley (2009)	Fit single-hidden-layer neural network using categorically based classification.	Model tuned by grid search: number of units in the hidden layer (10, 25, 50); parameters for weight decay (0, 0.2, 0.4)
Support vector Machines (eps Regression)	e1071: Dimitriadou et al. (2008)	A support vector machine is trained as a regression machine, using the eps-regression type. A'1' denotes a person citing gambling problem and '0' denotes another reason for account closure with 0.5 as the cutoff level for prediction.	Model tuned by grid search: gamma = $10^{[-6:-0]}$; cost of constraints violation = $10^{[-1:3]}$
Support vector Machines (c-Classification)	e1071: Dimitriadou et al. (2008)	A support vector machine is trained as a classification machine, using the C-classification type.	Model tuned by grid search: gamma = $10^{[-6:-0]}$; cost of constraints violation = $10^{[-1:3]}$

(*continued*)

Table 1. *(Continued)*

Algorithm	References	Description	Other settings
Support vector Machines (one-Classification)	e1071: Dimitriadou et al. (2008)	A support vector machine is trained as a classification machine, using the one-classification type for novelty detection.	Model tuned by grid search: gamma $= 10^{[-6:-0]}$; cost of constraints violation $= 10^{[-1:3]}$
Random Forest (regression)	Breiman, Cutler, Liaw, and Wiener (2012)	Implements Breiman (2001) random forest algorithm for regression. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.	Model tuned for the optimal number of predictors sampled for splitting at each node: number of trees used at the tuning step $= 50$; step factor $= 1$; improvement requirement in out of bag error for search to continue $= 0.5$
Random Forest (classification)	Breiman et al. (2012)	Implements Breiman (2001) random forest algorithm for classification. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.	Model tuned for the optimal number of predictors sampled for splitting at each node: number of trees used at the tuning step $= 50$; step factor $= 1$; improvement requirement in out of bag error for search to continue $= 0.5$

et al. (2011) and Adami et al. (2013). In order to provide a metric that is meaningful for actual implementation of these algorithms, the SLAs are compared based on their classification parameters: sensitivity, specificity, precision, accuracy, odds ratio and area under curve (AUC) (Sing, Sander, Beerenwinkel, & Lengauer, 2005).

Results

As shown in Table 2 the Random Forest algorithm performs exceptionally well in identifying the players' account closure reason within the training sample. The regression based method correctly classifies 99% of the 'self-excluding' (SE) players and 100% of the 'other' players, while the classification-based method correctly identifies all of the players. As shown in Table 3, while the Random Forest algorithms also perform well when identifying the 'other' players in the testing data sample, they perform quite poorly in identifying the SE players. The regression and classification algorithms' sensitivity is 7% and 6% respectively (93% and 96% type II errors respectively). Overfitting appears to be a concern with the Random Forest algorithms.[1] However, the RG algorithms do score highest among the classifiers in the testing sample on precision, accuracy and odds ratio.[2] Any application of this technique to real data should be extensively cross-validated, and

Table 2. SLA classification rate on training data.

	Training data					
	Sensitivity	Specificity	Accuracy	Precision	AUC	Odds ratio
Step-wise Logistic	0.0413	0.996	0.683	0.833	0.519	10.647
GLM LASSO/elasticnet	0.0579	0.972	0.672	0.500	0.515	2.114
Neural Network – Regression	0.2397	0.960	0.724	0.744	0.600	7.502
Neural Network – Classification	0.3471	0.923	0.734	0.689	0.635	6.408
SVM – eps Regression	0.8099	1.000	0.938	1.000	0.905	∞*
SVM – C-Classification	0.0909	1.000	0.702	1.000	0.546	∞*
SVM – one-Classification	0.1488	0.843	0.615	0.316	0.496	0.937
Random Forest – Regression	0.9917	1.000	0.997	1.000	0.996	∞*
Random Forest – Classification	1.0000	1.000	1.000	1.000	1.000	∞*

*The odds ratio produces ∞ values or all cut-offs corresponding to False Negative = 0 or False Positive = 0, which limits its usefulness for comparison in some instances.

even then may not produce satisfactory prediction results, but it may be a useful approach for avoiding false positives while still identifying some true positives.

Similar to the Random Forest models, the support vector machine eps regression method performed well in the training data with 81% classification accuracy of the SE players, but performed poorly with the hold-out sample (9% sensitivity). The SVM one-classification method produced better results in the testing data. It correctly classified 24% of SE players, albeit with a higher Type I error rate (18%). The reason for this improvement may be due to the fact that SVM one-classification is designed for novelty detection (Dimitriadou, Hornik, Leisch, Meyer, & Weingessel, 2008). Therefore, it may be a worthwhile algorithm to test in full player pools where the occurrence of disordered gambling is even rarer. The SVM C-classification method underperformed other support vector machines based on the AUC and odds ratio metrics, and may be a suboptimal SLA for future player classification tasks.

Table 3. SLA classification rate on testing data.

	Testing data					
	Sensitivity	Specificity	Accuracy	Precision	AUC	Odds ratio
Step-wise Logistic	–	0.991	0.652	–	0.495	–
GLM LASSO/elasticnet	0.018	0.981	0.652	0.333	0.500	0.963
Neural Network – Regression	0.164	0.896	0.646	0.450	0.530	1.690
Neural Network – Classification	0.291	0.811	0.634	0.444	0.551	1.764
SVM – eps Regression	0.091	0.906	0.627	0.333	0.498	0.960
SVM – C-Classification	–	1.000	0.646	–	0.491	–
SVM – one-Classification	0.236	0.821	0.621	0.406	0.529	1.417
Random Forest – Regression	0.073	0.972	0.665	0.571	0.522	2.693
Random Forest – Classification	0.055	0.972	0.658	0.500	0.513	1.981

The artificial neural network classification method did not classify SE players in the training data as well as other methods, but it did outperform all other methods in the classifying SE players within the testing data. However, this improvement in classification also occurred in tandem with only 81% specificity. This compares unfavourably with, for example, the Random Forest specificity of 3%. It appears that when classification accuracy is improved by any of the SLAs in the testing data, it is accompanied by an increase in the Type I error rate. Among the classifiers, the neural network (classification) model and the random forest (regression) model appear to be the preferred SLAs. The neural network model produces the highest AUC score while correctly identifying the most likely problem gamblers, and the random forest model produces the highest odds ratio while producing $<3\%$ false positives.

When comparing the similarly high rates of type II errors in both the training and testing data sets using the LASSO/elasticnet method (GLM), the issue of out-of-sample generalizability becomes clearer. The GLM fits a 10-fold cross-validation model when using the training data. This method penalizes results that are poor fits for prediction outside of selected data, and therefore should make similarly strong (or weak) predictions in the training data and the testing data. Effectively, the SLA uses bootstrapping methods to penalize overfitted models while the model is being fitted. As such, this algorithm should alleviate overfitting issues that appeared to characterize the step-wise logistic regression (4% sensitivity in the training data but 0% sensitivity in the testing data). The GLM model did perform marginally better than the step-wise regression in predicting likely problem gamblers (a 2% prediction rate as opposed to 0%), but it is clear that given the large rate of type II errors, predicting the outcomes of out of sample players is difficult; regardless of whether models are cross-validated during the estimation process or whether they are manually validated by the researcher ex-post. Effectively, the GLM results within the training data are representative of the same problem that faced the Random Forest and SVM algorithms when applied to the testing data.[3]

What also emerges from the logistic and GLM analysis is that results from more conventional regression models (e.g. logistic regression) can be improved by using models that allow for complex relationships among the variables (e.g. ANNs, SVMs and Random Forests). The best SLMs from each of those classes of models outperformed the step-wise logistic and the GLM models based on AUC and the odds ratio in the testing sample.

Conclusion

As described in detail by Gainsbury (2011), access to player account data is one of the greatest opportunities for the field of responsible gambling in recent decades. While several private companies have begun to develop enterprise solutions in this area, ensuring that the academic community continues to research and publicly disseminate reliable findings and limitations is a critical aspect of continued development for all stakeholders.

In order to effectively advance the field of responsible gambling data mining, progress in two areas is important:

(1) Expanding the behavioural marker literature by developing theory from the fields of psychology, sociology and economics into new quantitative metrics; and
(2) Improving the efficiency of existing models through the application of more robust statistical and computer science based designs.

In pursuit of the latter goal, this study evaluated a set of supervised learning algorithms to determine which data mining procedures may be more effective in identifying probable

disordered gamblers. It was found that specific SLAs can generally improve classification accuracy over more conventional techniques.

Breiman's random forest algorithm was found to be the most accurate method to classify players in training data, but was noted to perform poorly for identifying likely problem gamblers on hold-out testing samples. Although artificial neural networks appeared to be the most useful classification method in identifying likely disordered gamblers in the hold-out sample, many people with gambling-related problems were still left unidentified, and false positives also become a concern. Bet intensity, variability, frequency and trajectory, as well as age and gender, are noted to be insufficient variables to reasonably predict likely problem gamblers in this data set. Studies that identify new behavioural variables that can be built using data mining techniques, such as those by Adami et al. (2013), will be crucial to the refinement and successful implementation of future responsible gambling data mining algorithms.

This study identified a clear need for hold-out testing in these types of data mining studies that classify gamblers based on historical play data. Even methods that use bootstrapping-based cross-validation showed higher classification accuracy in training data than in testing data. Given that most research in this field has been done using unsupervised clustering methods without hold-out samples, existing models may need much refinement before their results can be reasonably used to predict out-of-sample problem gamblers.

This dataset examined a sample that is not representative of the general online gambling population. If modelling is conducted on a more representative sample, the small population of disordered gamblers may bias some algorithms and lead to underestimation of their probabilities (King & Zeng, 2001). In that case, it may be necessary to change the choice of estimation method, at the expense of overall classification accuracy. For example, one-classification SVMs could be used in lieu of eps regression SVMs, or the untested rare events logistic regression could be used in lieu of typical logit models.

When classification accuracy was improved by any of the SLAs in the testing data, it was generally accompanied by an increase in the number of false positives. Unless future behavioural markers disproportionally improve one algorithm over another, choice of which algorithm to implement may have to be a function of the acceptable sensitivity to false positive ratio. As a final contribution, this study provides a baseline level of classification accuracy by which future studies of online gambling behaviour and problem gambling can be evaluated.

The findings from this study rely on the accuracy of the data found in Braverman and Shaffer (2012). Anomalies in the original data set would limit the applicability of this study's results, as do limitations in the validity of the dependent variable. Self-excluders may have concerns about their play but they may not actually fit diagnostic criteria for gambling-related problems. There also is no scale of severity. Other populations may also respond differently from the population in this study. As more progress is made in this area, it will be important to assess the validity of these instruments or conduct further research with well-validated instruments, such as the problem gambling severity index.

The choice of algorithms in this study was designed to be representative of popular algorithm classes, but is by no means comprehensive. Discriminant analysis, boosting, bagging and other classes of algorithms could also be explored for prediction accuracy. As new methods continue to be developed, responsible gambling researchers should continue to explore how they might improve player behaviour modelling.

Acknowledgements

This paper utilized data from the Transparency Project (http://www.thetransparencyproject.org), Division on Addiction, the Cambridge Health Alliance, a teaching affiliate of Harvard Medical School.

Funding

Funding for this study was provided by the Conrad N. Hilton Foundation.

Notes

1. As a matter of vocabulary, Random Forests do not overfit (Breiman, 2001) but since each tree grows to maximum size they perform poorly when generalized to the testing data, creating a similar issue as the typical overfitting problem.
2. $$\text{Odds ratio}: \frac{(\text{true positive})\cdot(\text{true negative})}{(\text{false positive})\cdot(\text{false negative})}$$
3. Despite the use of the 10-fold cross-validation technique, the training/testing data partition remains in order for an equal comparison with the other algorithms.

Notes on contributor

Kahlil Philander is a Visiting Assistant Professor at the William F. Harrah College of Hotel Administration, University of Nevada, Las Vegas. He is also the Director of Research at the International Gambling Institute, University of Nevada, Las Vegas. His research interests include the economics of gambling, gambling policy, and responsible gambling.

References

Adami, N., Benini, S., Boschetti, A., Canini, L., Maione, F., & Temporin, M. (2013). Markers of unsustainable gambling for early detection of at-risk online gamblers. *International Gambling Studies*, *13*, 188–204.

American Psychiatric Association. (2000). In M. B. First (Ed.), *Diagnostic and statistical manual of mental disorders* (4th ed, pp. 671–674). Washington, DC: American Psychiatric Publishing.

Balestra, M., & Krafcik, C. A. (Eds.). (2010). *Casino city's internet gambling report*. Newton, MS: Casino City Press.

Beasley, D., & Chan, E. (2012). Analysis: U.S. casinos, vendors eye big online-poker stakes. Retrieved from http://www.reuters.com/article/2012/01/06/us-usa-gambling-idUSTRE8051K320120106

Braverman, J., LaBrie, R. A., & Shaffer, H. J. (2011). A taxometric analysis of actual internet sports gambling behavior. *Psychological Assessment*, *23*, 234–244.

Braverman, J., & Shaffer, H. J. (2012). How do gamblers start gambling: Identifying behavioural markers for high-risk internet gambling. *The European Journal of Public Health*, *22*, 273–278.

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

Breiman, L., Cutler, A., Liaw, A., & Wiener, M. (2012). randomForest: Breiman and Cutler's random forests for classification and regression. *R package version, 4.6–7*.

bwin. (2012). *Responsible gaming at bwin*. Retrieved from https://home.bwin.com/page.aspx?view=aboutus&path=/respGaming

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning, Carnegie Mellon University, Pittsburgh, PA, USA* (pp. 161–168). New York: ACM.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2008). Misc functions of the Department of Statistics (e1071), TU Wien. *R Package*, 1–5.

Division on Addiction. (May 17, 2012). *How Do Gamblers Start Gambling: Identifying Behavioural Markers for High-risk Internet Gambling*. Medford, MA: Division on Addiction, The Transparency Project [database distributor].

Dragicevic, S., Tsogas, G., & Kudic, A. (2011). Analysis of casino online gambling data in relation to behavioural risk markers for high-risk gambling and player protection. *International Gambling Studies*, *11*, 377–391.

Ferris, J., & Wynne, H. (2001). *The Canadian problem gambling index: User manual*. Toronto, ON: Canadian Centre on Substance Abuse.

Fiedler, I. (2012). The gambling habits of online poker players. *The Journal of Gambling Business and Economics*, *6*(1), 1–23.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package Version, 1*.
Gainsbury, S. (2011). Player account-based gambling: Potentials for behaviour-based research methodologies. *International Gambling Studies*, *11*, 153–171.
Haefeli, J., Lischer, S., & Schwarz, J. (2011). Early detection items and responsible gambling features for online gambling. *International Gambling Studies*, *11*, 273–288.
Hing, N., & Nuske, E. (2011). Assisting problem gamblers in the gaming venue: An assessment of practices and procedures followed by frontline hospitality staff. *International Journal of Hospitality Management*, *30*, 459–467.
Hodgins, D. C., & Makarchuk, K. (2003). Trusting problem gamblers: Reliability and validity of self-reported gambling behavior. *Psychology of Addictive Behaviors*, *17*, 244–248.
King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*, 137–163.
LaBrie, R. A., Kaplan, S. A., LaPlante, D. A., Nelson, S. E., & Shaffer, H. J. (2008). Inside the virtual casino: A prospective longitudinal study of actual internet casino gambling. *The European Journal of Public Health*, *18*, 410–416.
LaBrie, R. A., LaPlante, D. A., Nelson, S. E., Schumann, A., & Shaffer, H. J. (2007). Assessing the playing field: A prospective longitudinal study of internet sports gambling behavior. *Journal of Gambling Studies*, *23*, 347–362.
LaBrie, R., & Shaffer, H. J. (2011). Identifying behavioral markers of disordered Internet sports gambling. *Addiction Research & Theory*, *19*, 56–65.
LaPlante, D. A., Kleschinsky, J. H., LaBrie, R. A., Nelson, S. E., & Shaffer, H. J. (2009). Sitting at the virtual poker table: A prospective epidemiological study of actual internet poker gambling behavior. *Computers in Human Behavior*, *25*, 711–717.
Lesieur, H., & Blume, S. (1987). The south oaks gambling screen (SOGS): A new instrument for the identification of pathological gamblers. *American Journal of Psychiatry*, *144*, 1184–1188.
Livingstone, C. (2005). Desire and the consumption of danger: Electronic gaming machines and the commodification of interiority. *Addiction Research & Theory*, *13*, 523–534.
OLG. (2011). Pre-qualification #1112-110 data analysis tools & services for responsible gaming. Retrieved from http://www.merx.com/English/Supplier_Menu.asp?WCE=Show&TAB=3&ZPORTAL=MER&XState=7&id=PR233860&print=Y&src=osr&ForceLID=&HID=&hcode=mPo3cLBxGQRIGY2AQNGHhw==
Osborn, A., Skelt, L., Delfabbro, P., Nevile, M., & McMillen, J. (2007). *Identifying problem gamblers in gambling venues*. Melbourne, VIC: Gambling Research Australia.
Philander, K. P., & Abarbanel, B. L. L. (2013). Determinants of Internet Poker Adoption. *Journal of Gambling Studies*. Advance online publication. doi:10.1007/s10899-013-9382-9.
Rexer, K., Allen, H., & Gearan, P. (2011). 2011 data miner survey summary. Retrieved from http://www.rexeranalytics.com/Data-Miner-Survey-Results-2011.html
Ripley, B. (2009). nnet: feed-forward neural networks and multinomial log-linear models. R package, version 7.3-1.
Ripley, B. (2013). (R Core Team Ed.), *R: A Language and Environment for Statistical Computing Reference Index* (Version 3.0.0, pp. 1660–1662). Vienna: R Foundation for Statistical Computing.
Schellinck, T., & Schrans, T. (2007). Assessment of the behavioral impact of responsible gaming device (RGD) features. Retrieved from http://www.nsgc.ca/images/uploads/Focal%20Research%20Report%20_2_.pdf
Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: Visualizing classifier performance. *R. Bioinformatics*, *21*, 3940–3941.
Svenska Spel. (2010). Svenska Spel strengthens its responsible gaming. Retrieved from http://media.svenskaspel.se/en/2010/03/22/svenska-spel-strengthens-its-responsible-gaming/
Volberg, R. A., Gerstein, D. R., Christiansen, E. M., & Baldridge, J. (2001). Assessing self-reported expenditures on gambling. *Managerial and Decision Economics*, *22*, 77–96.
Wood, R. T., & Griffiths, M. D. (2008). Why Swedish people play online poker and factors that can increase or decrease trust in poker web sites: A qualitative investigation. *Journal of Gambling Issues*, *21*, 80–97.