**Predicting Online Gambling Self-Exclusion: An Analysis of the Performance of Supervised Machine Learning Models**

Christian Percy*, Manoel Franca, Simo Dragievic, Artur d'Avila Garcez

Christian Percy: Researcher,

Floor 1, 16-24 Underwood Street

London N1 7JQ, UK, 07733 063475, cwspercy@gmail.com – Corresponding author

Manoel Franca: Research Assistant, City University London, Northampton Square, London, UK, EC1V 0HB, 020 7040 5060, manoel.franca@city.ac.uk

Simo Dragicevic: Bet Buddy CEO,

Floor 1, 16-24 Underwood Street

London N1 7JQ, UK, 020 7193 6896., simo@bet-buddy.com

Artur d'Avila Garcez: Reader, City University London, Northampton Square, London, UK, EC1V 0HB, 020 7040 5060, a.garcez@city.ac.uk

**Acknowledgements:**

We would like to thank IGT for supplying the de-identified player data that was primarily used for this study, as well as the Transparency Project (www.thetransparencyproject.org), Division on Addiction, the Cambridge Health Alliance, a teaching affiliate of Harvard Medical School, which provided the second dataset reviewed. We would also like to thank Dr. Tillman Weyde and Dr. Gregory Slabaugh, both from City University London, for their support in completing this paper.

**Disclosure Statement**:

Simo Dragicevic is the founder of Bet Buddy and Chris Percy is the lead researcher for Bet Buddy, a UK software company focused on responsible gaming analytics.

Beyond those mentioned in the Disclosure Statement or the Acknowledgements, there are no further funding sources, constraints on publishing or competing interests among the authors.

**Abstract**

*As gambling operators become increasingly sophisticated in their analysis of individual gambling behaviour, this study evaluates the potential for using machine learning techniques to identify individuals who used self-exclusion tools out of a sample of 845 online gamblers, based on analysing trends in their gambling behaviour. Being able to identify other gamblers whose behaviour is similar to those who decided to use self-exclusion tools could, for instance, be used to share responsible gaming messages or other information that aids self-aware gambling and reduces the risk of adverse outcomes. However, operators need to understand how accurate models can be and which techniques work well. The purpose of the paper is to identify the most accurate technique out of four highly diverse techniques and to discuss how to deal analytically and practically with a rare event like self-exclusion, which was used by fewer than 1% of gamblers in our dataset. We conclude that balanced training datasets are necessary for creating effective models and that, on our dataset, the most effective method is the random forests technique which achieves an accuracy improvement of 35 percentage points versus baseline estimates.*

**Keywords:** Responsible gambling; Gambling self-exclusion; Machine learning; Supervised learning algorithms; Problem gambling; Internet gambling, Three-tier model

## Introduction

Online gambling operators and regulators are increasingly considering how best to support online gamblers, so as to minimise any harm that their gambling activity might cause. This trend can be seen in the activities of such recent conferences as the European Conference on Gambling Studies and Policy Issues held in Helsinki during September 2014 and the Responsible Gambling Trust Harm Minimisation Conference, held for the first time on December 2013 in London, and again in December 2014.

As Gainsbury (2011) and Philander (2013) note, the data collection made possible in account-based Internet gambling revolutionises the kind of analysis of gambling behaviour that is possible and opens up new ways to identify early warning signs of potentially harmful behaviour. However, the quantity of data simultaneously opens up questions of how best to interpret the data: specifically, how to transform raw gambling session data into meaningful, descriptive variables of behaviour (behavioural markers), and how then to relate those descriptive variables to an individual who is potentially at risk.

In this paper, we build on Philander (2013) and Dragicevic, Percy, Kudic, and Parker

(2013) to explore data analysis for a particular type of risk identification, that of self-exclusion. Through gambling platforms that permit individuals to self-exclude, it is intended that individuals might recognise that they are at risk of losing control during gambling sessions and instruct the gambling platform to deactivate or block their account for a certain period of time. Leveraging anonymised gambling data made available by our research partners IGT, we explore four different machine learning statistical techniques for predicting individuals that self-exclude (logistic regression, neural networks, Bayesian networks, and random forests – an explanation of each technique is presented in Table 1, Section 3).

As applied in this paper, these are "supervised learning" techniques, in that we present each method with known self-excluders and a known control group of non-self-excluders, along with various input variables that describe each gambler. Based on these known individuals, the techniques arrive at particular models – a series of mathematical rules based on the input variables to estimate the probability of any particular gambler being about to self-exclude at a particular point in time.

These techniques are chosen to represent a spectrum of complexity and interpretability. For instance, logistic regression model output provides a single value for each model input that can be simply described, whereas neural networks re-use and combine input variables through multiple analytical layers, which quickly results in complex patterns that are hard to interpret. Through this paper and future work, we are interested in exploring whether more complex models generate more accurate results and what implications there are for interpreting model output in a real world context. There are two important benefits of being able to predict self-exclusion events. The first lies in improved player protection. A common motivation, although not the sole motivation, for self-exclusion is concern over one's gambling behaviour and the potential for unhealthy levels of gambling. By identifying individuals whose play pattern approximates those who have subsequently chosen to self-exclude, or by identifying individuals in advance of a self-exclusion, the gambling operator can choose to share information or advice with the player that may support healthy engagement with the gambling platform. Alternatively, the operator may choose to restrict marketing activity or platform activities for that player for a certain period of time. The second benefit lies in more stable, long-term revenue flows to gambling operators, in terms of gamblers that might use their platform slightly less intensively than before, but do so with greater security and satisfaction.

On our dataset, random forests proved to be most effective technique, achieving a high level of accuracy, correctly identifying 87% of control group and self-excluder gamblers, versus a baseline performance of 52%. Bayesian networks performed almost as well on overall accuracy but were less consistent and less effective at identifying self-excluders relative to control group players. The primary contributions of this paper relative to earlier studies are:

(1) To demonstrate the value of including Bayesian networks in the suite of machine

learning tools that can achieve high accuracy on gambling-related events;

(2) To demonstrate the utility of a particular algorithm for creating datasets with a roughly equal number of self-excluders and control group gamblers ("SMOTE") such that models perform better than on unbalanced datasets; and

(3) To demonstrate accuracy performance on test datasets in the 72%-87% range, significantly higher than the 62%-67% range in Philander (2013), pointing towards a set of additional gambling behaviour variables that are worth assessing further.

The paper is structured as follows: The first section presents a literature review of at-risk gambling behaviours and supervised learning methods. The second section describes the data available and how it is transformed into behavioural markers. The third section describes the four machine learning techniques used and presents the results. The final section discusses the implications and limitations of our analysis, as well as avenues for further research.

## 1. Literature review

This section explains our approach for assessing at-risk gambling behaviours and sets out key works from the existing literature on predictive behaviours of self-exclusion and problem gambling and from the existing literature on machine learning in particular as applied to problem gambling.

### *1.1 Three-tier model for assessing at-risk gambling behaviours using player data*

Our approach to using data for understanding problem and at-risk gambling behaviours is the three-tier model, which is an evidence-based analytic approach that analyses data across these three tiers: exhibited, declared, and inferred behaviours. The three-tier model was first introduced by Dragicevic et al. (2013), which described the rationale for the model's approach to predicting harm in gambling. "Exhibited behaviour" compares an individual's gambling activity along a single parameter (such as frequency of gambling or loss chasing) against clinically-informed assessments of risk thresholds. "Declared behaviour" captures an individual's responses to risk assessment surveys. "Inferred behaviour", meanwhile, compares an individual's complete pattern of gambling activity against a reference set of known concerned gamblers' activity patterns. Dragicevic et al. (2013) focused on using data obtained from the gambling operator IGT to describe Internet gambling self-excluders in terms of their demographic and behavioural characteristics (a form of exhibited behaviour).

This study is focused on assessing approaches and methods for improving the accuracy of predicting self-excluders in order to develop inferred behaviour (tier 3) methods from the three-tier model. For example, the behaviour described as "gambling trajectory" looks to see whether there is a strong trend of the player increasing his spend amounts over time; specifically we apply a linear regression analysis to the total amount bet on each day that the person gambled at least once, over the last 20 such active

gambling days. The gambling trajectory is one example of a behavioural trend of a player that is a clinically-relevant sign of risk and thus is also theorised to afford some insights as a direct, quantitative predictor variable in assessing how likely a player is to close his or her account.

Given that self-exclusion is a rare event, applying to less than 1% of the gambling cohort, it is important to consider what it means to predict an individual gambler's similarity to others that have self-excluded. Even a gambler who displays very similar gambling behaviour to those who self-exclude is nonetheless unlikely to self-exclude themselves, which motivates the value of identifying such gamblers under the "Inferred behaviour" principle: Some of these gamblers would perhaps benefit in the long run from a self-exclusion period or some other intervention, but do not do so for a variety of reasons, ranging from a lack of awareness of self-exclusion functionality or the need for a prompt before using it. However, other gamblers might display a similar pattern of behaviour to self-excluders and yet be gambling safely, given their personal circumstances and preferences. For this reason, we would caution operators against drawing absolute conclusions or making firm statements based on "inferred behaviour" risk scores. Nonetheless, such scores can provide a useful source of additional information to aid self-aware gambling or inform gambling operator choices, such as which marketing or responsible gambling messages to share with the gamblers.

This "inferred behaviour" approach should also be distinguished from predicting actual self-excluders several weeks or months in advance of their self-exclusion, which is a separate analytical question of interest and requires a different analytical approach (see Section 5 for more details).

### *1.2 Research using Internet player data to understand harm in gambling*

The dataset that is used for this study formed the basis of Dragicevic et al.'s (2013) study that analysed the gambling behaviour of Internet gamblers who self-excluded compared to a control group of non-self-excluders. The study demonstrated that self-excluders from the data sample were more likely to be young and male and were also more likely to adopt riskier gambling strategies and suffer greater losses. Other studies analysing actual Internet player data have also provided useful input into determining what could be good predictors of gambling harm.

There is support in the literature for a wide range of behavioural factors that contribute to the kind of adverse gambling activity which represents one of the motivations for self-exclusions. Johansson et al. (2009) identified 35 risk factors, not all of which are analytically available for this study. Cummins et al. (2009) argued that early wins influence more reckless betting patterns. Bet variability as a problem gambling risk factor can also be identified in the pathways model of Blaszczynski and Nower (2002). They identified three distinct groups of gamblers manifesting impaired control over their

gambling behaviour: (a) behaviourally conditioned problem gamblers; (b) emotionally vulnerable problem gamblers; and (c) antisocial, impulsive problem gamblers. In the first pathway (behaviourally conditioned problem gamblers), problem gamblers often fluctuate between regular/heavy and excessive gambling because of conditioning, distorted cognitions surrounding the probability of winning, and/or a series of bad judgments or poor decision-making. Ferris and Wynne (2001) also specify wager increase as an indicator of problem gambling behaviour. Hayer and Meyer (2011b) studied 259 gamblers from win2day.at showing that most who elected to self-exclude were males under forty. Subsequent research also revealed that many had gambled too much or spent too much time gambling, resulting in excessive financial losses.

LaBrie and Shaffer (2011) analysed the accounts of Internet sports gamblers who closed their accounts for reasons related to problem gambling. Having conducted some initial exploratory analysis, the authors compared 215 gamblers who closed their account for a self-declared concern over gambling-related problems, with 351 who closed their account due to self-declared loss of interest. The authors performed a multi-stage multivariate discriminant function analysis (MFDA), in which gamblers who were not distinctively and correctly classified were removed from the next stage of analysis. The purpose of this was ultimately to isolate a 'pure' group of self-declared problem gamblers, identified with 100% accuracy, and describe what makes that group distinctive.

The appropriate results for our study derive from the MFDA as applied to these two groups, prior to the removal of indistinct gamblers. This model achieved a 67% accuracy in identifying those who were no longer interested and 50% accuracy in identifying self-declared problem gamblers. The two most significant behavioural factors in this model were total winnings and total number of days between first and last bet. The ultimate model which achieved 100% accuracy on a subsample nonetheless points to additional key behaviours which are likely to be of value in identifying some individuals who might suffer from problem gambling, acknowledging that those who observe they have problems and self-exclude are an incomplete set of all those who might suffer from problem gambling. The four behaviours that correctly isolated this final set of self-declared problem gamblers from the final set of those who closed their account due to lack of interest were: i) placing more bets, ii) placing larger bets, iii) betting more frequently, and iv) betting intensely soon after enrolment.

Braverman and Shaffe (2010) have used an unsupervised learning technique (k-means clustering) to analyse player data in an attempt to understand what behaviours led to account closure. The study revealed that players characterized by high intensity and frequency of gambling and also by high variability of wager (bet) sizes during their first month of gambling were at higher risk than others to report gambling-related problems upon closing their accounts. Dragicevic, Tsogas, and Kudic (2011) replicated this method using a different data set and discovered that some of the typologies of gamblers were partially reproducible, notably high intensity gamblers, high frequency gamblers, and moderate gamblers. However the authors also concluded that the method, specifically

k-means clustering, was not optimal for predicting at-risk gamblers.

As well as assessing behaviours, research using Internet gambling data has attempted to assess whether game characteristics have an influence on problem gambling. For example, La Plante, Nelson, and Gray (2013) suggested that the importance of game characteristics as having a potentially causal relationship with the onset of problem gambling, particularly live action sports betting. Other research has indicated the certain game types, such as casino style games, which are continuous games like live action sport betting, can condition and reinforce gambling behaviour, and in that way could lead to the development of problem gambling behaviour (Turner, 2008). These studies point towards game characteristics as being potentially important pathways to predicting harm in gambling.

Wardle (2012) describes the characteristics of self-excluders who used an internet betting exchange. Self-excluders tended to be male, younger adults, heavily engaged in internet and offline gambling, and gambled with increased volume but experienced poorer returns. Two out of three self-excluders were problem gamblers, with the majority worried about the amount of time and money spent on gambling.

### *1.3 Using supervised learning methods to predict harm in gambling*

The aforementioned studies did not utilise the kind of supervised learning prediction models that are the subject of this paper. Schellinck and Schrans (2011) have assessed the use of supervised learning models in the context of predicting which gamblers could be at-risk of problem gambling. Their results suggest useful methods and techniques for building models that can predict gamblers at risk of harm. Whilst their research suggests useful general approaches and benchmarks for building such models, little is provided regarding specific techniques and variables that could prove to be good predictors of problem or at-risk gambling.

Building on the work from the live action sports betting dataset available from the Division on Addiction, Philander (2013) assessed nine supervised learning methods to determine which data mining methods are most effective at identifying disordered Internet sports gamblers. The supervised learning methods include logistic regression, regularized general linear models (GLM), neural networks, support vector machines (SVM) and random forests, on which more detail can be found in his paper. Philander (2013) presents results for training and test datasets; the relevant comparison for our paper and real world applications is performance on a testing dataset, in which model performance ranges from 62% to 67%, with random forests the highest performing technique.

## 2. Raw data and variable generation

This section sets out the gambling session data used in this study, how it is

transformed into behavioural variables for inclusion in the supervised learning models, and how over-identifying variables are treated.

### *2.1 Gambling session data*

We obtained de-identified Internet player data from IGT, an Internet gambling software provider for lotteries and commercial Internet gambling operators. IGT also provided basic demographic data on all players: gender, age and country of residence. In the context of this dataset, IGT provided gambling software for 10–15 clients during the data period, who each in turn might manage multiple gaming sites operating across multiple countries.

The cohort of self-excluding players was generated from 604 players who had self-excluded at least once during a period from April 2009 to July 2011, along with data on their gambling sessions in the months leading up to their first self-exclusion. We discount a further 98 individuals whose first self-exclusion period was specified as under 180 days, as such choices better reflect cooling off periods rather than a serious decision to restrict one's gambling behaviour long-term. Although specifying a cut-off point is an inexact science, six months is used regularly in the industry and is in line with the approach adopted by the UK's Gambling Commission.[1]

An approximate control group cohort of 871 players was generated, designed to be representative of the regular gamblers from among the 11,667 players who had gambled more than 10 sessions during the month of January 2009 on the basis of age, gender, players favouring poker or casino games, total amount bet, total number of sessions played and account status.[2]These 871 players only represent an approximate control group as we can only state that they had not self-excluded as of the time of data being downloaded, not that they might be at risk of self-exclusion or might have chosen to do so afterwards. For these players, gambling session data from January 2009 to December 2010 was made available by IGT. To avoid possible bias caused by gambling behaviour being affected by terminating all data collection for ongoing gamblers during the festive season, only data up to 1 November 2010 was entered into the calculation of behavioural variables. Nonetheless, the authors acknowledge a limitation of this analysis in that self-excluders stopped gambling on various dates throughout the year (with no clear pattern) whereas we take November 2010 as an artificial last day of analysis for the control group. For further details on sample selection, please see Dragicevic et al (2013).

In the calculation of behavioural variables, as described below, several players had not gambled for long enough to generate particular risk factors. For instance, they may have self-excluded within a week of opening their account or simply chosen to have

[1] http://www.gamblingcommission.gov.uk/FAQs/Problem-gambling/What-is-self-exclusion.aspx {accessed 4 June 2015, page last reviewed date given as July 2013}

[2] The self-excluders used in this study similarly gambled more than ten sessions in ~95% of months in which they gambled at all, discarding partial months in which they started or stopped. The comparison to the control group is nonetheless inexact and represents a limitation on our analysis: ~5% of self-excluders did gamble fewer than 10 sessions for over a quarter of their total months gambling.

stopped gambling from that account very early on within the date range of the data made available to us. Since the focus of this paper is on identifying medium-term trends in ongoing gambling behaviour, only those players with enough data to calculate all risk factors are included in the modelling: 176 self-excluders and a control group of 669. Individuals who choose to cease gambling or to self-exclude after only a short period of gambling play are worthy of separate analysis via a differently configured set of risk factors.

### *2.2 Derivation of gambling behaviour markers*

The raw data on gambling sessions as analysed in this paper incorporate the type of game played, the amount wagered and the amount lost or won in Euros, and the start and end time of each session. These session data are run through a series of algorithms, ranging from regression models to t-tests, to identify five different behaviour markers which are potential problem gambling 'risk factors'. Four of these risk factors (trajectory, frequency, intensity, and variability) are consistent behavioural markers analysed in previous peer-reviewed studies (see Braverman and Shaffer, 2010). We added an additional risk factor, session time, based on insights from Hayer and Meyer (2011b) to enable us to capture the effect of players who spend increasingly greater amounts of time during their real-money gambling sessions.

Across these five risk factors, we generate a total of 30 variables that capture the absolute level of activity, the statistical significance of a change in gambling behaviour and variables that capture the scale of any such change in behaviour. Combined with the three demographic variables (gender, age, country of residence), there are then 33 variables entered into the machine learning models.

The five risk factors are chosen based on the literature referred to above: how many days an individual typically gambles ("frequency"), the total amount of money an individual bets ("trajectory") and how often he or she places wagers on days when they are actively gambling ("intensity"), as well as the total time he/she spends online ("session time") and the extent to which the amount of money he/she gambles is volatile over time ("variability").

Two classification variables are also included in the analysis for each risk factor: a binary check for a significant quantitative change in behaviour (e.g. was there at least a 10% move in an adverse direction between comparison periods) and a risk factor rating based on the statistical significance of the calculated change (where 3 represents a p-value of 1% or lower, 2 is 1%-5%, 1 is 5%-10%, and 0 is anything higher; in this instance, a p-value can be thought of as the probability that you might observe a particular trend purely as a result of random variation, rather than as the result of a reliable underlying trend).

There are other risk factors that we plan to explore in future analyses, such as how individuals' losses increase or decrease over time, whether individuals top up their

account following a loss, and how they fund their gambling. Operating within certain data and programming constraints for this paper, we chose to focus on the five that are listed given their extensive use in previous peer-reviewed research.

### *2.3 Treatment of over-identifying variables*

An initial investigation of the data revealed that certain countries of residence were sufficiently rare among the gambler cohorts that they would almost uniquely identify a small number of self-excluders, i.e. the control group sample was insufficiently large to allow for such fine-slicing of the variables. Since this reflects the scarcity of data and relatively small sample size, rather than any underlying logic, such variables need treatment to prevent over-fitting bias within our predictive models. A hypothetical example would be a machine learning conclusion such as: All players based in Serbia are almost certain to be self-excluders, because there is only one player based in Serbia in our dataset and she self-excluded. While, by definition, such a conclusion cannot be disproved within the dataset, a broader understanding of the gambling market requires us to adjust for the input data. To adjust such variables, we apply the rules of thumb proposed by Agresti (2007) and Stokes, Davis, and Koch (1995) for logistic regression, and determined that we should aggregate the "country of residence" variable into Germany residents vs non-Germany residents.

Finally, we note that it is important to remove data that is available in retrospect but not in live data for identifying potential high risk gamblers: total days spent gambling, total amount bet, calendar date of first or last gambling sessions or the last login date. Such variables would accurately discriminate self-excluders in our dataset as they gambled for only part of the period being analysed (by definition, as they self-excluded at some point during the period) unlike the control group gamblers, but not in a way that can be applied to a real-life scenario. Indeed, models we constructed using these variables generated classification accuracy on our dataset during model testing in the 95%-98% range.

A list of the final 33 variables used in our modelling is presented in Appendix 1, along with sample descriptives that compare the control group cohort to the self-excluder cohort.

## 3. Modelling and results

This paper prepares predictive models using four methods implemented using WEKA software (Hall et al., 2009): logistic regression, Bayesian networks, neural networks and random forests. Three of these are used by Philander (2013); Bayesian networks is an additional method reviewed in this paper. We add Bayesian methods as the ability to unpick the conditional probabilities linking each variable to self-exclusion is helpful from an interpretation / communication perspective. Together they represent a

diverse set of widely used machine learning techniques spanning a broad spectrum of complexity and interpretability. Table 1 presents a comparison of the four algorithms employed.

Regarding parameter settings for each predictive model, a small subsample of training data (10%) has been taken for fine-tuning to arrive at the values presented in Table 1. Where no parameter is specified, we used the default values provided by WEKA.

Each technique applies supervised learning with ten-fold cross validation (Geisser, 1993). In ten-fold cross-validation, the dataset is split randomly into ten sets. Removing one of each of the ten sets of data in turn, 90% of the data is used as a training sample to generate a model which is then tested on the remaining 10% as a pure test dataset. The model results, such as measures of goodness of fit, are then averaged across all ten experiments in order to reduce data order bias and provide more significant results. This is a different technique to Philander (2013), who adopts a single train/test split for the majority of his experiments.

Since our dataset contains fewer self-excluders than control group gamblers, the dataset is imbalanced, which can lead to unintended behaviour, e.g. classifying all gamblers as non-self-excluders, which may superficially maximise the accuracy of the prediction on the dataset, even though it is not useful in a real world context. To correct for this, we over-sample the dataset of self-excluders via the SMOTE algorithm which has been shown to perform better than other methods of dealing with dataset rebalancing (Russell and Norvig, 2003; Ha and Bunke, 1997; Bowyer, Chawla, Hall, and Kegelmeyer, 2002; Nakamura et al.,2013). In this paper, we apply the optimal SMOTE level to achieve an approximately 50:50 split between control group and self-excluding cohorts, within the constraints that SMOTE must enhance dataset sizes by multiples of 50%.

Table 2 presents the results of our analysis, showing both pre-SMOTE dataset and the post-SMOTE balanced dataset for comparative purposes.

## 4. Discussion

In the actual dataset, we observe higher performance on specificity metrics than sensitivity metrics. In other words, overall model accuracy is best achieved when setting parameters that correctly isolate control group gamblers rather than self-excluders. This is exactly to be expected since the actual dataset is weighted in favour of the control group. Since this weighting is relatively high at around 4:1, it is unsurprising that all models perform with similar levels of accuracy and do not markedly outperform a trivial majority class model on accuracy (i.e. in which all players are assumed to be non-self-excluders). For this reason, applying SMOTE or some other mechanism to balance the dataset is essential.

Not only does dataset balancing strongly reduce the bias towards specificity, it

also enables the supervised learning algorithms to react more accurately to differences in behaviour between the two groups, just as observed by previous authors applying SMOTE to datasets in other fields of study (see Section 3). This improvement in performance is particularly clear when we compare the machine learning methods to a baseline performance of majority classification, i.e. in which we trivially allocate all players to the larger class of non-self-excluders in the control group. Additionally, as expected, SMOTE results in better area under ROC curve results, since this is a measurement of how well each algorithm is able to classify each class and a balanced dataset requires the models to fully utilise the information contained in both the control group cohort and the self-exclusion cohort.

The highest performing model in terms of both overall accuracy and area under ROC is random forests, with 87% of self-excluders and control group players correctly classified and over 0.9 of ROC area. The Bayesian networks model follows very closely behind with 86% accuracy, but at a higher standard deviation indicating that its performance varied more widely across the different datasets used for cross-validation and the model's performance is thus relatively unstable.

It is also noteworthy that Bayesian networks performs better in specificity then sensitivity, with no such variation to be observed among random forests. Bayesian networks thus might be of higher value where the primary concern is to avoid unduly alarming gamblers that they might be at risk. Indeed, we note that global prevalence surveys suggest, albeit subject to their design limitations, that most people gamble without significant problems (Williams, Volberg, and Stevens, 2012) and operators will be concerned with a balanced approach that also minimizes distress caused to non-problem gamblers.

Finally, we note that the poor relative performance by logistic regression on the balanced dataset is most likely due to the restricted set of relationships applied by this implementation of logistic regression between the input variables and the outcome variables. In other words, the higher performance by the three other, more flexible models suggests that, in a complex real-world setting, there are more optimal links between, say, how often someone gambles and their probability of choosing to self-exclude than a linear input into a logistic function.

We wish to extend the comparison with Philander's 2013 paper, particularly as his dataset is publicly available whereas ours is available only under a confidentiality agreement with our corporate research partner. On that basis, we applied our set of four machine learning methods to the same dataset as used in his paper. The results are presented in Table 3.

Although we do not expect identical results to Philander (2013) as there are different ways of parametrising the machine learning models and due to our use of ten-fold cross-validation, it is reassuring that the results on the training dataset (where all the datapoints are used to construct the model) are very similar. The results for logistic regression and neural networks are within a few percentage points. On random forests, the results are very close for total accuracy, but with higher sensitivity and lower specificity.

On the testing dataset, we see a slight uptick in performance: we obtain accuracy levels of 64%-70% across all models, whereas Philander obtains a range of 62%-67%. This slightly higher accuracy is likely due either to variation in the random sample selection for the training set, model parameter choices or due to the use of SMOTE to correct for the 1:2 imbalance in the data sample.

On this dataset, there is no large difference between the four different machine learning methods we apply, although random forests and Bayesian networks remain strong performers when considering a preference for high accuracy and low standard deviation. On the other hand, in terms of sensitivity, specificity and area under ROC curve, we notice significant differences between the SMOTE and non-SMOTE results. By attempting to learn with an imbalanced dataset, the non-SMOTE results display significant differences between sensitivity and specificity, which indicates that these learning models are biased towards a certain class, which is also reflected in poorer performance in the area under the ROC curves.

Finally, we note that our machine learning models performed better on the IGT dataset than the Philander (2013) dataset, with accuracy on test datasets in the 72%-87% range, significantly higher than the 62%-67% range. The higher model performance is unlikely to be solely down to the use of SMOTE or a different approach to train/test partitioning, as the uptick in performance as applied to Philander's own dataset is only 1-3 percentage points. Instead, the potential value is likely to lie in the additional predictor variables, unless there is some idiosyncrasy in our dataset of European gamblers that makes the players inherently easier to categorise.

For instance, we conducted different behavioural analysis on the IGT dataset, including the focus on how statistically significant are the breaks in players' behavioural trends, as well as calculations on players' intensity and volatility of gambling activity. A thorough comparison of the variable performances lies outside the scope of this paper given the technical complexity of such analyses but, as discussed below, knowledge extraction from the machine learning models and an assessment of key predictor variables will be the subject of a follow-up paper.

## 5. Limitations and further research

As well as data and technique limitations mentioned earlier and in the referenced literature, we would particularly highlight the complexity involved in interpreting self-exclusion events. Hayer and Meyer (2011a,b) have found that only 76% of a sample of

self-excluders were problem or pathological gamblers and Wardle (2012) found similarly that not all online sports bettors who self-excluded were problem gamblers. This is due in part to different motivations for self-exclusion, which we are unable to disaggregate with the currently available data, ranging from a desire to control problem gambling behaviour to a player expressing dissatisfaction with the service, disinterest in gambling more generally, or simply curious about what would happen if they did self-exclude for a temporary period, potentially securing better treatment from the gambling operator. It is also due to the large amount of unobserved features that define the context of a gambler's activity, including, for instance, their other gambling, both off-line and on other Internet sites, their changing financial circumstances and any co-morbidities such as alcohol use.

A further limitation lies in model interpretation. It is not straightforward to identify which variables are driving the predictive results (and hence to query them or use them as part of productive interventions and discussions with potentially at-risk gamblers). When discussing clinical risks and possible interventions, the lack of interpretative power raises research, legal and practical concerns. For instance, is it appropriate to flag up someone's behaviour as 'at risk' without being able to tell them in detail why this is the case? This field of model interpretation is termed Knowledge Extraction and is focused on opening the 'black box' of algorithms, providing explanations for the behaviour of the sub-symbolic networks, and offering symbolic descriptions which are crucial in computer science to enable compositionality, system maintenance and evolution, and the consolidation of knowledge.

Finally, we should note that our data is drawn from a single backoffice gaming platform provider, IGT, and it is possible that unique features of that platform have influenced the results. Similarly, all the gambling players within the survey are in Europe, and the majority from Germany, and this may also influence results in an unknown fashion.

In terms of next steps, we see four broad areas for this research to be further developed, across the range of constraints in any empirical research journey from data availability to real world impact. The first is to apply the methods to larger and more varied datasets, including other outcome variables of key interest to addiction clinicians, such as problem gambling self-test scores, which would address some of the limitations related to using self-exclusion as a single proxy for harm in online gambling.

The second is to explore whether different parameter choices for the gambling behavioural markers or the inclusion of additional behavioural markers result in higher performing models e.g., assessing whether previous uptake of social responsibility tools, such as limit setting, result in better predictors. For instance, we might consider whether the trajectory of increasing amount bet is a more significant predictor if measured over 40 active gambling days or 20 active gambling days.

The third, and the focus of the authors' next paper with this dataset, is in the area of Knowledge Extraction, specifically to unpack the different models in terms of how individual variables relate to model performance. It will be interesting to observe not only

which gambling behavioural markers are the strongest predictors of self-exclusion or other outcomes of interest but also how common the strongest predictors are between the different machine learning methods. All machine learning models are ultimately the result of rule-based operations and numeric calculations on input variables and as such the impact of different variables can be computed, synthesised and summarised. However, large networks, particularly those with many predictor variables and complex inter-relations between predictor variables, require significant analysis in order to draw meaningful conclusions.

Fundamental research in the area of knowledge extraction (D'Avila Garcez, Broda, and Gabbay, 2002, D'Avila Garcez, Lamb, and Gabbay, 2009), carried out at City University in collaboration with many researchers in the UK, Europe and worldwide, has proved strong correspondences between a number of artificial intelligence network models and different types of formal logic from other disciplines (D'Avila Garcez, Lamb, and Gabbay, 2006, D'Avila Garcez, Gabbay, Ray, and Woods, 2007). Such correspondences offer a sound mechanism for simplifying and describing in a compact way how these complex models work and provide a foundation for further work on this dataset (Franca, Zaverucha, and D'Avila Garcez, 2014).

As researchers working closely with gambling operators, we are also excited by the potential of the fourth area for further research - testing these models in a live environment. How early can self-exclusion propensity or problem gambling risk scores be usefully identified (we note that gambling behaviour up to the day of self-exclusion is applied in this paper)? Is it possible to identify at-risk behaviour early enough to support effective interventions, such as player messaging or website interface and marketing adjustments? Are we able to quantify this difference in intervention success, perhaps by comparing a natural experiment of players who opt-in to certain programs or by only making certain programs available to different sets of players? Finally, we observe that these four different areas of research can be pursued in tandem and we look forward to seeing and supporting the progression of this research agenda.

**References**

Agresti, A. (2007). *An Introduction to Categorical Data Analysis (2nd ed)*. Wiley

Blaszczynski, A., & Nower, L. (2002). A pathways model of problem and pathological gambling. *Addiction*, 97, 487–499.

Bowyer, K., Chawla, N., Hall, L., Kegelmeyer P. (2002). SMOTE: Synthetic Minority Over sampling Technique. *Journal Of Artificial Intelligence Research,* Volume 16, pages 321-357.

Braverman, J., & Shaffer, H.J. (2010). How do gamblers start gambling: Identifying behavioural markers for high-risk Internet gambling. *European Journal of Public Health.*

Doi: 10.1093/eurpub/ckp232

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 1 (October 2001), 5-32.

Cooper, G. and Herskovits, E. (1991). A Bayesian method for constructing Bayesian belief networks from databases. In *Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence* (UAI'91), D'Ambrosio, Smets, and Bonissone (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 86-94.

Cummins, L.F., Nadorff, M.R., & Kelly, A.E. (2009). Winning and positive affect can lead to reckless gambling. *Psychology of Addictive Behaviors*, 23, 287–294.D'Avila Garcez, A. S., Broda, K., and Gabbay, D. (2002). *Neural-Symbolic Learning Systems: Foundations and Applications, Perspectives in Neural Computing*. Springer.

D'Avila Garcez, A. S., Lamb, L. C., and Gabbay, D. M. (2006). Connectionist Computations of Intuitionistic Reasoning. *Theoretical Computer Science Journal.* 2006 vol 358, no. 1, pp. 34 - 55

D'Avila Garcez, A. S., Lamb, L. C. (2006). A Connectionist Computational Model for Epistemic and Temporal Reasoning. *Neural Computation Journal.* 2006 Jul;18(7):1711-38

D'Avila Garcez, A., Gabbay, D. M., Ray, O. & Woods, J. (2007). Abductive reasoning in neural-symbolic learning systems. *Topoi: An International Review of Philosophy*, 26(1), pp. 37-49

D'Avila Garcez, A. S., Lamb, L. C., and Gabbay, D. M. (2009). *Neural-Symbolic Cognitive Reasoning*. *Cognitive Technologies,* Springer.

Dragicevic, S., Percy, C., Kudic, A., & Parke, J. (2013). A Descriptive Analysis of Demographic and Behavioral Data from Internet Gamblers and Those Who Self-exclude from Online Gambling Platforms. *Journal of Gambling Studies* Advance online publication. doi: 10.1007/s10899-013-9418-1

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 (3), 377-391.

Ferris and Wynne (2001). *The Canadian Problem Gambling Index: Final report.* Ottawa, ON: Canadian Centre on Substance Abuse

França, M., Zaverucha, G., and D'Avila Garcez, A.. (2014). Fast relational learning using bottom clause propositionalization with artificial neural networks. *Machine Learning*, 94, 1, 81-104.

Freedman, D. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press.

Gainsbury, S. (2011). Player account-based gambling: Potentials for behaviour-based research methodologies. *International Gambling Studies*, 11, 153–171.

Geisser, S. (1993). *Predictive Inference*. New York, NY: Chapman and Hall.

Ha, T. and Bunke, H.. (1997). Off-line, Handwritten Numeral Recognition by Perturbation Method. *Pattern Analysis and Machine Intelligence,* vol. 19/5, pp. 535-539.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann P., Witten, I. (2009). *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, Volume 11, Issue 1. WEKA.

Hayer, T., & Meyer, G. (2011a). Self-exclusion as a harm minimization strategy: Evidence for the casino sector from selected European countries. *Journal of Gambling Studies*, 27(4), 685–700.

Hayer, T., & Meyer, G. (2011b). Internet self-exclusion: Characteristics of self-excluded gamblers and preliminary evidence for its effectiveness. *International Journal of Mental Health and Addiction*, 9, 307–596.

Hosmer, D. & Lemeshow, S. (2000). *Applied Logistic Regression (2nd ed.).* Wiley

Johansson, A., Grant, J.E., Kim, S.W., Odlaug, B.L., & Gotestam, G.K. (2009). Risk factors for problematic gambling: A critical literature review. *Journal of Gambling Studies*, 25, 67 92.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

Mitchell, T. (1997). *Machine Learning* (1 ed.). McGraw-Hill, Inc., New York, NY, USA.

Nakamura, Munehiro, Kajiwara, Yusuke, Otsuka, Atsushi, Kimura, Haruhiko (2013). LVQ-SMOTE - Learning Vector Quantization based Synthetic Minority Over-sampling Technique for biomedical data. *BioData mining*, 6. p. 16.

La Brie, R., & Shaffer, H. (2011). Identifying behavioral markers of disordered Internet sports gambling. *Addiction Research and Theory,* 2011, Vol. 19, No. 1 , Pages 56-65

LaPlante, D. A., Nelson, S. E., & Gray, H. M. (2013, August 5). Breadth and Depth Involvement: Understanding Internet Gambling Involvement and Its Relationship to Gambling Problems. *Psychology of Addictive Behaviors.* Advance online publication. doi: 10.1037/a0033810

Lauritzen, S. & Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems. In *Journal Royal Statistics Society B*, 50(2), 157-194.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Philander, K., International Gambling Studies (2013): Identifying high risk online gamblers: a comparison of data mining procedures. *International Gambling Studies*, DOI:

10.1080/14459795.2013.841721

Quinlan, J. (1986). Induction of Decision Trees. *Mach. Learn.* 1, 1 (March 1986), 81-106.

Rebane, G. and Pearl, J., The Recovery of Causal Poly-trees from Statistical Data, *Proceedings, 3rd Workshop on Uncertainty in AI,* (Seattle, WA) pages 222–228, 1987

Rumelhart, D., Hinton, G. and Williams, R.. (1985). *Learning internal representations by error propagation*. No. ICS-8506. California University of San Diego La Jolla, Institute for Cognitive Science.

Russell, S. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2 ed.). Pearson Education.

Schellinck. T. and Schrans, T. (2011). Intelligent design: How to model gambler risk assessment by using loyalty tracking data. *Journal of Gambling Issues*, Issue 26, pp. 51-68. doi: 10.4309/jgi.2011.26.5

Spirtes, P.; Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review* 9 (1): 62–72.

Stokes, M.E., Davis, C.S., Koch, G.G. (1995). *Categorical Data Analysis Using the SAS System*. Cary, NC: SAS Institute Inc

Turner, E.N. (2008). Games, gambling and gambling problems. In Zangeneh, Blaszczynski, & Turner (Eds.), *The Pursuit of Winning: Problem Gambling Theory, Research and Treatment* (33-64). Berlin, Heidelberg, New York: Springer.

Wardle, H. (2012). *Understanding self-exclusion—Profile, processes and improvements: Evidence and implications from a research study of online betting exchange users*. Paper presented at meeting of Responsible Gambling Council Discovery 2012 Conference.

Williams, R.J., Volberg, R.A. & Stevens, R.M.G. (2012). *The Population Prevalence of Problem Gambling: Methodological Influences, Standardized Rates, Jurisdictional Differences, and Worldwide Trends.* Report prepared for the Ontario Problem Gambling Research Centre and the Ontario Ministry of Health and Long Term Care. May 8, 2012

Xuan, Z., & Shaffer, H. (2009). How do gamblers end gambling: Longitudinal analysis of Internet gambling behaviors prior to account closure due to gambling related problems. *Journal of Gambling Studies,* 25 (2), 239-252.

*Table 1: Overview of machine learning algorithms*

| **Method / References** | **Summary** | **Relevant strengths** | **Key WEKA Parameters** |
|---|---|---|---|
| Logistic regression<br>*Freedman, 2009*<br>*Hosmer and Lemeshow, 2000* | Simple logistic regression relates each input variable directly to the output variable with no mapping of inter-input variable relationships. The variables are mapped onto a logistic curve to provide a probability that any given player is a self-excluder, with a specific cut-off point chosen to optimise accuracy or some other error function. | Minimal computational cost<br>Easy statistical interpretation via odds ratios and p-values<br>But: Less flexibility in modelling a variety of relationships | Link function: Binomial logistic<br>Classification cut-off value: 0.5 |
| Bayesian networks<br>*Pearl, 1998, 2000*<br>*Mitchel, 1997*<br>*Cooper and Herskovits, 1991* | Bayesian networks produce a structured map of the main ways that all input variables and the output variable relate to each other based on their conditional probabilities.<br>First, the method identifies an efficient structure for which variables affect which others most strongly, leading up to the output variable itself (self-exclusion). Second, the conditional probabilities for each link are estimated and added to the structure. | Can handle bias due to correlations between available input variables and missing input variables<br>Can mitigate over-fitting risk<br>Can incorporate expert /prior knowledge | Parents: Unlimited<br>Score type: entropy<br>Type: SimpleEstimator<br>Algorithm: K2<br>No prior knowledge assumed |
| Neural networks<br>*Mitchell, 1997*<br>*Rumelhart, Hintel, and Williams 1985* | Neural networks are organised in layers, made up of a number of interconnected nodes, which contain activation functions (typically linear or sigmoid) which are trained based on the inputs of the previous layer. The first layer will contain the values of all the input | Can be applied to data where the relationship between variables is only very vaguely understood<br>Works well with high data | Number of hidden neurons: sqrt(#inputs)<br>Momentum: 0.1<br>Learning rate: 0.05<br>Decay factor: 0.999 |

| | | | |
|---|---|---|---|
| | variables. | dimensionality | Learning rule: Backpropagation |
| Random forests<br>*Quinlan, 1986*<br>*Russell and Norvig, 2003*<br>*Breiman, 2001* | A random forest represents an ensemble of many decision trees, where each individual decision tree seeks to classify a gambler into self-excluder or not based on the values of a subset of input variables (e.g. gender=male, age<30, gambled>twice a week etc.) | Typically produces high performing models on given datasets, but computationally intensive and can be at particular risk of over-fitting | Max depth: unlimited<br>Number of trees: 200<br>Number of features used for random selection: 3 |

*Table 2: IGT dataset - Model performance results – Correct classification, % (standard deviation)*

| Method | Metric | Pre-SMOTE dataset | Balanced dataset (post-SMOTE) |
|---|---|---|---|
| Logistic regression | **Overall accuracy** | **80% (1.91)** | **72% (2.66)** |
| | Sensitivity | 0.15 (0.08) | 0.70 (0.07) |
| | Specificity | 0.97 (0.02) | 0.74 (0.05) |
| | Area under curve | 71% (0.07) | 80% (0.03) |
| Bayesian networks | **Overall accuracy** | **78% (5.08)** | **86% (3.47)** |
| | Sensitivity | 0.31 (0.10) | 0.77 (0.08) |
| | Specificity | 0.90 (0.06) | 0.94 (0.03) |
| | Area under curve | 77% (0.06) | 93% (0.02) |
| Neural networks | **Overall accuracy** | **80% (1.52)** | **77% (3.40)** |
| | Sensitivity | 0.10 (0.05) | 0.73 (0.06) |
| | Specificity | 0.98 (0.01) | 0.80 (0.06) |
| | Area under curve | 70% (0.06) | 85% (0.03) |
| Random forests | **Overall accuracy** | **80% (3.26)** | **87% (2.31)** |
| | Sensitivity | 0.21 (0.08) | 0.87 (0.05) |
| | Specificity | 0.95 (0.03) | 0.87 (0.04) |
| | Area under curve | 79% (0.06) | 94% (0.02) |

*Note: Standard accuracy is the number of players correctly classified as a percentage of the total sample size, sensitivity is the share of self-excluders correctly classified "(i.e. the true positive rate: correctly identified self-excluders divided by the total number of self-excluders in the sample)", specificity is the share of the control group correctly classified "(i.e. the true negative rate)" and area under ROC curve is the measurement of how far each algorithm is from being a random classifier", where a higher number represents a more effective model".*

*Table 3: Philander 2013 dataset - Model performance results – Correct classification, % (standard deviation)*

| Method | Metric | Train accuracy (w/o SMOTE) | Train accuracy (SMOTE) | Test accuracy (w/o SMOTE) | Test accuracy (SMOTE) |
|---|---|---|---|---|---|
| Logistic regression | **Overall accuracy** | **68%** | **69%** | **66% (2.18)** | **65% (5.71)** |
| | Sensitivity | 0.2 | 0.66 | 0.01 (0.04) | 0.62 (0.10) |
| | Specificity | 0.92 | 0.72 | 0.98 (0.04) | 0.67 (0.07) |
| | Area under curve | 0.69 | 0.75 | 0.52 (0.05) | 0.68 (0.05) |
| Bayesian networks | **Overall accuracy** | **68%** | **74%** | **65% (3.47)** | **70% (4.40)** |
| | Sensitivity | 0.14 | 0.75 | 0.11 (0.06) | 0.71 (0.09) |
| | Specificity | 0.94 | 0.73 | 0.92 (0.06) | 0.69 (0.08) |
| | Area under curve | 0.65 | 0.82 | 0.56 (0.05) | 0.75 (0.05) |
| Neural networks | **Overall accuracy** | **84%** | **86%** | **64% (4.57)** | **64% (6.07)** |
| | Sensitivity | 0.59 | 0.86 | 0.12 (0.05) | 0.62 (0.07) |
| | Specificity | 0.96 | 0.87 | 0.90 (0.06) | 0.66 (0.09) |
| | Area under curve | 0.91 | 0.94 | 0.61 (0.07) | 0.71 (0.04) |
| Random forests | **Overall accuracy** | **100%** | **100%** | **62% (5.52)** | **67% (6.52)** |
| | Sensitivity | 1 | 1 | 0.19 (0.06) | 0.66 (0.09) |
| | Specificity | 1 | 1 | 0.84 (0.07) | 0.69 (0.10) |
| | Area under curve | 1 | 1 | 0.58 (0.09) | 0.76 (0.06) |

*Note 1: Standard accuracy is the number of players correctly classified as a percentage of the total sample size, sensitivity is the share of self-excluders correctly classified "(i.e. the true positive rate: correctly identified self-excluders divided by the total number of self-excluders in the sample)", specificity is the share of the control group correctly classified "(i.e. the true negative rate)" and area under ROC curve is the measurement of how far each algorithm is from being a random classifier", where a higher number represents a more effective model". Note 2: Philander (2013) dataset is the Analytic Dataset (High Risk) under the project How Do Gamblers Start Gambling: Identifying Behavioural Markers for High-risk Internet Gambling with the dataset made available by Division on Addictions via www.thetransparencyproject.org. The dataset covers 530 gamblers who closed their account, 33% of whom self-declared that they did so for reasons of problem gambling, as opposed to, for instance, losing interest in the site, which provides a control group for comparing against self-identified problem gamblers.*

**Short biographical note for authors**

Chris Percy is Bet Buddy's lead researcher and has over 8 years' experience in statistical modelling and longitudinal research covering disciplines such as internet gambling and problem gambling, supply chain optimization, financial trading back-testing and public policy on unemployment and education initiatives. He has experience in applying a wide range of data mining techniques, with particular expertise in logistic regression. Chris' research has been published in numerous peer-reviewed journals, including the Journal of Gambling Studies and the Journal of Education and Work, and he has presented his gambling-related research at the European Association for the Study of Gambling 2014 Conference.

Manoel Franca is currently a Research Assistant at City University London, working on a collaborative project with Bet Buddy on predicting harm in gambling. Manoel's speciality on this project is applying multiple forms of machine learning and analytic models, including logistic regression, artificial intelligence, SVMs, and Bayesian logic, to solve complex prediction problems. During his PhD, his research has focused towards neural-symbolic integration, with First-Order Logics and Artificial Neural Networks. His interests are neural-symbolic systems, inductive logic programming and relational learning, and his research has been published in peer-reviewed journals.

Simo Dragicevic is founder and CEO of Bet Buddy, a UK software company focused on responsible gaming analytics. Bet Buddy's responsible gaming analytics platform has been implemented in Europe and Ontario, Canada. Simo is an expert in the domain of data analytics and gambling research, having also collaborated with organizations such as GamCare and acted as an expert advisor to The Responsible Gambling Trust in the UK. Prior to founding Bet Buddy Simo was a Practice Lead at Accenture and Director of Barclays Bank.

Dr. Garcez is a Reader in Neural-Symbolic Computation at City University London. He is a Fellow of the British Computer Society. Dr. Garcez has an established track record of research in Neural Computing, Artificial Intelligence, Data Mining, and Computer Science Logic. He has co-authored two books has over 100 peer-reviewed publications with 1040 citations. Dr. Garcez is editor-in-chief of the Artificial Intelligence and Neural Computation books series. His research has been funded by the Royal Society, the European Union and the Nuffield foundation. He has acted as reviewer for the Royal Academy of Engineering, BBSRC, and EPSRC.

Appendix 1: Sample descriptive statistics

| | Self-Excluder cohort | | | Control Group | | |
|---|---|---|---|---|---|---|
| Variable | Min | Avg (mean) | Max | Min | Avg (mean) | Max |
| ***Trajectory risk factor*** | | | | | | |
| Statistical significance (1-P-value) | 0.0 | 0.2 | 1.0 | 0.0 | 0.2 | 1.0 |
| Slope coefficient | -1,771 | 19 | 3,904 | -8,172 | -50 | 1,790 |
| Average amount bet per gambling day | 45 | 4,761 | 45,660 | 1 | 2,259 | 71,600 |
| Increase in average bet per day | -2.2 | -0.1 | 2.3 | -2.4 | -0.1 | 2.6 |
| Dummy variable for increase >10% | 42% did increase >10% | | | 35% did increase >10% | | |
| Statistical significance category (higher = higher risk/significance) | 0.0 | 0.1 | 3.0 | 0.0 | 0.2 | 3.0 |
| ***Frequency risk factor*** | | | | | | |
| Statistical significance (1-P-value) | 0.0 | 0.5 | 1.0 | 0.0 | 0.4 | 1.0 |
| Current Frequency (share of gambling days in last 30 days) | 0.0 | 0.4 | 0.9 | 0.0 | 0.3 | 1.0 |
| Prior Frequency (share of gambling days in 30 days prior) | 0.0 | 0.3 | 0.9 | 0.0 | 0.3 | 1.0 |
| Increase in Frequency between periods | -0.9 | 1.3 | 21.3 | -1.0 | 0.6 | 15.7 |
| Dummy variable for increase >10% | 59% did increase >10% | | | 55% did increase >10% | | |
| Statistical significance category (higher = higher risk/significance) | 0.0 | 0.8 | 3.0 | 0.0 | 0.5 | 3.0 |
| ***Intensity risk factor*** | | | | | | |
| Statistical significance (1-P-value) | 0.0 | 0.2 | 1.0 | 0.0 | 0.2 | 1.0 |
| Current Intensity (number of bets placed over the last 10 gambling days) | 1.0 | 14.1 | 396.3 | 1.0 | 7.1 | 257.9 |
| Prior Intensity (number of bets placed over the prior 10 gambling days) | 1.2 | 15.0 | 366.9 | 1.0 | 8.5 | 467.3 |
| Increase in Intensity between periods | -0.8 | 0.3 | 29.5 | -1.0 | 0.3 | 75.5 |
| Dummy variable for increase >10% | 32% did increase >10% | | | 33% did increase >10% | | |
| Statistical significance category (higher = higher risk/significance) | 0.0 | 0.1 | 3.0 | 0.0 | 0.1 | 3.0 |
| ***Session time risk factor*** | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Statistical significance (1-P-value) | 0.0 | 0.3 | 1.0 | 0.0 | 0.2 | 1.0 |
| Slope coefficient | -36.5 | -0.3 | 15.1 | -31.1 | -0.9 | 37.9 |
| Average session time per gambling day | 0.0 | 1.5 | 13.2 | 0.0 | 1.2 | 8.6 |
| Increase in average session time | -2.7 | 0.0 | 2.9 | -2.6 | -0.1 | 2.9 |
| Dummy variable for increase >10% | 41% did increase >10% | | | 33% did increase >10% | | |
| Statistical significance category (higher = higher risk/significance) | 0.0 | 0.1 | 3.0 | 0.0 | 0.2 | 3.0 |
| ***Variability risk factor*** | | | | | | |
| Statistical significance (1-P-value) | 0.0 | 0.3 | 1.0 | 0.0 | 0.3 | 1.0 |
| Amount bet standard deviation (last 10 gambling days) | 55 | 6,405 | 79,275 | 1 | 2,430 | 80,927 |
| Amount bet standard deviation (10 gambling days prior) | 56 | 5,611 | 80,449 | 1 | 3,322 | 193,602 |
| Increase in standard deviation | -0.9 | 0.9 | 33.8 | -1.0 | 1.0 | 102.2 |
| Dummy variable for increase >10% | 44% did increase >10% | | | 42% did increase >10% | | |
| Statistical significance category (higher = higher risk/significance) | 0.0 | 0.1 | 3.0 | 0.0 | 0.1 | 3.0 |
| ***Demographic variables*** | | | | | | |
| Gender | 74% male | | | 83% male | | |
| Age in 2010 | 22 | 44 | 79 | 21 | 51 | 100 |
| Germany-based dummy variable | 74% Germany-based | | | 82% Germany-based | | |

Appendix 2: Logistic regression model output

| | Non-SMOTE dataset | | Post-SMOTE dataset | |
|---|---|---|---|---|
| **Variable** | Odds ratio | P-value | Odds ratio | P-value |
| Intercept | 1.86 | 0.37 | 0.06 | 0.00 |
| ***Trajectory risk factor*** | | | | |
| Statistical significance (1-P-value) | 5.98 | 0.04 | 2.87 | 0.06 |
| Slope coefficient | 1.00 | 0.15 | 1.00 | 0.03 |
| Average amount bet per gambling day | 1.00 | 0.10 | 1.00 | 0.04 |
| Increase in average bet per day | 0.96 | 0.87 | 0.88 | 0.53 |
| Dummy variable for increase >10% (1=Yes,0=No) | 0.36 | 0.04 | 0.57 | 0.08 |
| Statistical significance category | 0.97 | 0.93 | 1.48 | 0.08 |
| ***Frequency risk factor*** | | | | |
| Statistical significance (1-P-value) | 0.21 | 0.05 | 2.54 | 0.00 |
| Current Frequency (share of gambling days in last 30 days) | 0.13 | 0.03 | 0.76 | 0.60 |
| Prior Frequency (share of gambling days in 30 days prior) | 6.27 | 0.07 | 0.74 | 0.59 |
| Increase in Frequency between periods | 0.89 | 0.04 | 0.88 | 0.00 |
| Dummy variable for increase >10% (1=Yes,0=No) | 0.21 | 0.01 | 2.26 | 0.00 |
| Statistical significance category | 1.36 | 0.07 | 0.82 | 0.04 |
| ***Intensity risk factor*** | | | | |
| Statistical significance (1-P-value) | 1.44 | 0.69 | 0.07 | 0.00 |
| Current Intensity (number of bets placed over the last 10 gambling days) | 0.98 | 0.06 | 0.98 | 0.00 |
| Prior Intensity (number of bets placed over the prior 10 gambling days) | 1.00 | 0.46 | 1.00 | 0.47 |
| Increase in Intensity between periods | 1.08 | 0.10 | 1.07 | 0.05 |
| Dummy variable for increase >10% (1=Yes,0=No) | 1.18 | 0.76 | 12.72 | 0.00 |
| Statistical significance category | 0.80 | 0.44 | 1.44 | 0.11 |
| ***Session time risk factor*** | | | | |

| | | | | |
|---|---|---|---|---|
| Statistical significance (1-P-value) | 0.20 | 0.06 | 0.21 | 0.00 |
| Slope coefficient | 0.98 | 0.48 | 0.97 | 0.10 |
| Average session time per gambling day | 0.84 | 0.03 | 0.84 | 0.01 |
| Increase in average session time | 1.17 | 0.52 | 1.16 | 0.39 |
| Dummy variable for increase >10% (1=Yes,0=No) | 1.34 | 0.57 | 1.65 | 0.06 |
| Statistical significance category | 1.63 | 0.11 | 1.30 | 0.18 |
| ***Variability risk factor*** | | | | |
| Statistical significance (1-P-value) | 1.60 | 0.50 | 1.09 | 0.86 |
| Amount bet standard deviation (last 10 gambling days) | 1.00 | 0.02 | 1.00 | 0.00 |
| Amount bet standard deviation (10 gambling days prior) | 1.00 | 0.01 | 1.00 | 0.00 |
| Increase in standard deviation | 1.00 | 0.94 | 1.00 | 0.79 |
| Dummy variable for increase >10% (1=Yes,0=No) | 1.04 | 0.92 | 1.35 | 0.30 |
| Statistical significance category | 0.92 | 0.75 | 0.86 | 0.42 |
| ***Demographic variables*** | | | | |
| Gender (1=Male, 0=Female) | 0.77 | 0.24 | 2.12 | 0.00 |
| Age in 2010 | 1.05 | 0.00 | 1.06 | 0.00 |
| Germany-based dummy variable (1=Germany, 0=Non-German) | 1.65 | 0.03 | 1.71 | 0.00 |
| | *AIC = -1.93; n = 845* | | *AIC = -1.71; n = 1285* | |

Note 1: This model output is created for illustrative purposes using the full sample size to generate propensity of being control group cohort. The purpose of this paper is to compare model predictive power, not assess which individual model inputs have most predictive power (see Section 5 for more detail). We caution readers against interpreting individual variables in the above table, due to covariance between related variables.

Note 2: For the three other machine learning models, the presentation of analysis at an input variable level is more complex, because each input variable appears many times in the output models of each of Bayesian Networks, Neural Networks and Random Forest models, resulting in the "black box" moniker applied to such methods (see Section 5 for more detail).