ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-974

The Need for Knowledge Extraction: Understanding Harmful Gambling Behavior with Neural Networks

Chris Percy[2], **Artur S. d'Avila Garcez**[1], **Simo Dragičević**[2], **Manoel V. M. França**[1], **Greg Slabaugh**[1] and **Tillman Weyde**[1]

Abstract. Responsible gambling is a field of study that involves supporting gamblers so as to reduce the harm that their gambling activity might cause. Recently in the literature, machine learning algorithms have been introduced as a way to predict potentially harmful gambling based on patterns of gambling behavior, such as trends in amounts wagered and the time spent gambling. In this paper, neural network models are analyzed to help predict the outcome of a partial proxy for harmful gambling behavior: when a gambler "self-excludes", requesting a gambling operator to prevent them from accessing gambling opportunities. Drawing on survey and interview insights from industry and public officials as to the importance of interpretability, a variant of the knowledge extraction algorithm TREPAN is proposed which can produce compact, human-readable logic rules efficiently, given a neural network trained on gambling data. To the best of our knowledge, this paper reports the first industrial-strength application of knowledge extraction from neural networks, which otherwise are black-boxes unable to provide the explanatory insights which are crucially required in this area of application. We show that through knowledge extraction one can explore and validate the kinds of behavioral and demographic profiles that best predict self-exclusion, while developing a machine learning approach with greater potential for adoption by industry and treatment providers. Experimental results reported in this paper indicate that the rules extracted can achieve high fidelity to the trained neural network while maintaining competitive accuracy and providing useful insight to domain experts in responsible gambling.
Keywords. Neural Networks; Knowledge Extraction; Gambling; Problem Gambling.

1 INTRODUCTION

Responsible gambling is a recent and complex field of study that investigates how to best support gamblers, so as to reduce the harm that their gambling activity might cause [17, 20]. Account-based internet gambling revolutionized the amount of data available to identify early warning signs of potentially harmful behavior [12]. However, the quantity of data simultaneously opens up questions of how best to interpret the data and its results: specifically, how to transform raw gambling session data into meaningful, descriptive variables of behavior, called behavioral markers, and how then to relate those descriptive variables to an individual who is potentially at risk.

Through gambling platforms that permit individuals to self-exclude, it is intended that individuals might recognize 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 anonymized gambling data made available by industry leaders and research partners IGT, in this paper we explore how such behavior could be explained through the use of neural networks and knowledge extraction. We extend and apply for the first time the TREPAN knowledge extraction algorithm [4] to the problem of predicting self-exclusion from gambling. We then evaluate the knowledge extracted and its importance in the context of this industrial application, in relation to its fidelity to the neural networks from which it was extracted, but also in terms of interpretability in comparison with other machine learning methods such as Bayesian networks and random forests, as used in [15].

In [15], the performances of neural networks trained with backpropagation, random forests, logistic regression, and Bayesian networks were evaluated and compared systematically on the same IGT dataset used here. Although the test set accuracy of the regression model was the lowest (72%), this model may be preferred by industry leaders because of the importance assigned to interpretability. Random forests achieved the highest accuracy (87%), with Bayesian networks in second and neural networks third. But the random forests were very difficult to interpret, consisting of 200 decision trees of unlimited depth. Much simpler than the random forest, the neural network used was a single-hidden layer perceptron with 33 inputs, 17 hidden neurons and 2 outputs, one for self-excluding players and the other for the control group players. But with more than 500

[1] Department of Computer Science, City University London, United Kingdom. Emails: {a.garcez, manoel.franca, t.e.weyde, gregory.slabaugh.1}@city.ac.uk
[2] Data Analytics Group, BetBuddy London, United Kingdom. Emails: cwspercy@gmail.com, simo@bet-buddy.com

weights, the neural network was also a black-box. The Bayesian network, which used the K2 algorithm, included 360 separately defined conditional probabilities, and also failed to instigate useful insight about the problem when shown to industry experts. In contrast, the TREPAN extension proposed in this paper, when applied to the neural network model above, produced a compact decision tree with a low loss of accuracy that was easily interpretable by experts, as discussed in detail in what follows. As a result, as a baseline, in this paper, we also apply a decision tree directly onto the data, and confirm that the use of the neural network is not redundant. This will also be discussed in more detail in the next section.

In practice, there are two important benefits of being able to predict self-exclusion events. The first is improved player protection. A common motivation, although not the sole motivation, for self-exclusion is concern over one's gambling behavior 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. For this to happen effectively, interpretability of results is important. Alternatively, the operator may choose to restrict marketing activity or platform activities for that player for a certain period of time. The second benefit is more stable, long-term revenue flows to gambling operators, since gamblers that might use their platform less intensively than before may do so with greater security and satisfaction.

Machine learning algorithms have only recently been applied to this field of study as a way of predicting potentially harmful gamblers [15, 17]. In this paper, differently from the related work report in the next section, we are interesting in revealing through knowledge extraction different aspects of harmful gambling behavior: which kinds of profiles fit into problematic gambling? Which attributes explain players who have such profiles? Such questions are motivated by survey and interview insight from a responsible gambling conference in Vancouver in which gambling operators, treatment providers and public policy officials set out the need for effective interpretation of such complex machine learning algorithms (New Horizons in Responsible Gambling, February 2016). We show that TREPAN can produce human-readable logic rules when applied to neural networks trained on gambling data. The rules obtained have high fidelity to the trained neural networks without much loss in accuracy. The decision tree produced by TREPAN, from which the rules can be read, additionally are found to summarize the key behaviors that are good predictors of self-exclusion, notably players who flagged highly on bet variability and intensity.

The rest of the paper is structured as follows: in Section 2, we present and discuss some of the most relevant literature regarding understanding gambling behavior and knowledge extraction from machine learning models. Section 3 introduces our methodology, discusses the changes made to TREPAN, and applies knowledge extraction to gambling behavior understanding using neural networks. Section 4 presents our empirical results comparing rule fidelity and accuracy to that of the neural networks but also with the direct application of decision trees to the data. Section 5 discusses the interpretability of our empirical results following direct feedback from industry leaders and gambling regulators, and concludes with the need for algorithm interpretation and directions for future work.

2 RELATED WORK

In current literature, despite significant research analyzing problem gambling more generally, studies using machine learning have been limited to prediction tasks, i.e. how well machine learning techniques can predict harmful gambling behavior. This paper builds on the literature by explaining the perceived value of knowledge extraction from black-box machine learning models and by adapting and applying a knowledge extraction technique to an industry dataset, and evaluating it both quantitatively and qualitatively through domain expert feedback.

Application to Clinical Analysis of Problem Gambling: In [2], a pathways model provides a framework with which to assess the effectiveness of machine learning models to support clinical analysis of problematic gambling behavior. The pathways model describes three possible pathways to gambling addiction: behaviourally conditioned problem gamblers (pathway 1), emotionally vulnerable problem gamblers (pathway 2), and antisocial, impulsive problem gamblers (pathway 3). In [8] it is argued that it is not possible to link pathway 2 (emotionally vulnerable problem gamblers) with data and behavioral insight extracted from game play. For example, while age is the variable used most in training the random forest predictions [15], there is no indication of its influence or dependence on other variables, or what value ranges are most relevant to the predictions. Similarly, behaviors associated with antisocial, impulsive problem gamblers (pathway 3), are also arguably very difficult to identify purely from analysing the patterns of play. However, in [8] it is argued that insights from behavioral data could provide evidence of gamblers at risk of becoming behaviourally conditioned problem gamblers (pathway 1), notably due to heavy or excessive gambling and loss chasing. For example, problem gamblers often fluctuate between regular, heavy and excessive gambling because of conditioning, distorted cognitions surrounding the probability of winning, or a series of bad judgments or poor decision-making. In [9] it is also noted that wager increase is an indicator of problem gambling behaviour.

Predicting Harmful Gambling Behaviour with Machine Learning: In [15], data obtained from the gambling operator IGT is used to describe internet gambling self-excluders in terms of their demographic and behavioral characteristics. Data analysis approaches and methods for improving the accuracy of predicting self-excluders are developed by hand towards inferred behavior models. Differently, this paper develops this by using artificial neural networks and TREPAN on the same IGT dataset to

describe, rather than predict, self-excluders through knowledge extraction.

Supervised machine learning models were evaluated in [20] in the context of predicting which gamblers could be at risk of problem gambling. Their results suggest useful but general methods and techniques for building models that can predict gamblers at risk of harm. While they propose benchmarks for building such models, specific techniques and the variables that could prove to be good predictors of problem or at-risk gambling are not investigated.

Building on the work from the live action sports betting dataset available from the Division on Addiction public domain, in [17] nine supervised learning methods are assessed 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, with results ranging from 62% to 67% with random forests the highest performing technique.

Knowledge Extraction from Neural Networks: Considerable interest and research was devoted to knowledge extraction around the turn of the century [1,4,13]. Recently, with a renewed interest in neural networks as a result of their successful application in a range of big data problems, the importance of knowledge extraction has also been highlighted, e.g. [5, 16, 18, 19, 21, 22]. In particular, there is a sense of knowledge extraction being needed to help organize the research in neural networks, though a better understanding of the strengths and limitation of the various models, but also to transfer knowledge from a source to a target domain in the context of transfer learning applications. In general, extraction methods continue to be classified as either decompositional (where the network is broken apart and its weight vectors are used by the extraction algorithm) or pedagogical (where the network or learning model is treated as an oracle to which queries are posed and answers are obtained). One of the early decompositional methods, TREPAN [4], is still nowadays one of the most successful extraction methods.

In [4], the TREPAN algorithm is proposed for the extraction of decision trees from trained neural networks. TREPAN is originally an M-of-N propositional tree inducer which uses a learned neural network as oracle to form a set of examples *S*, possibly distinct from the examples used to train the neural network, from which a decision tree is built recursively in the usual way, based on an information gain heuristic. M-of-N rules are of the form: if any M out of concepts $A_1,A_2,\ldots,A_n$ are *true* then concept B is *true*. In the next section, our approach for understanding gambling behavior relies on TREPAN with a few small but important modifications (discussed in the next section) to generate a decision tree which is then seen as a model for predicting gambling self-excluders.

As an example of a pedagogical approach, in [10, 13], a knowledge extraction algorithm is presented that is based on a partial ordering on the set of input vectors of the network, which is used to define a number of pruning rules and simplification rules that interact with such an ordering and allow sound knowledge extraction of rules in certain cases. Although provably sound, such a pedagogical extraction approach may generate far too many logical rules or take too long to compute in the case of large networks. In this paper, we take a more practical perspective.

The Need for Accountability and 'Human in the Loop' Oversight of Algorithms: In [6] it was stated that while algorithms can encode power to organizations, they also first can stand in tension with transparency. The types of questions and challenges addressed in [6] are: what is the basis for a prioritization decision? Is it fair and just, or discriminatory? What are the criteria built into a ranking, classification, or association, and are they politicized or biased in some consequential way? What are the limits to measuring and operationalizing the criteria used by the algorithm? How has the algorithm been tuned to privilege false positive or false negative errors? Does that tuning benefit one set of stakeholders over another? What are the potential biases of the training data used in a classifying algorithm? What types of parameters or data were used to initiate the algorithm?

By developing and being able to describe the machine learning algorithms used for predicting self-exclusion, answers to many of the questions posed above can be articulated. For example, the model input variables can be described and computations explained, model technical configurations can be defined, thresholds for flagging risk outlined, and limitations of the dependent variable as a proxy for predicting harm investigated. However, because of the complex nature of how algorithms use the data to obtain results, even a clear explanation of the above arguably still does not provide sufficient transparency as to why the algorithms produce the predictions they do.

In [6], it is further argued that autonomous decision-making is the crux of algorithmic power. Sometimes, though, the outcomes are important (or messy and uncertain) enough that a human operator makes the final decision. In the context of the gambling industry, this has important connotations. In [11], it is stated that 'human in the loop' algorithms enable industry and academics to understand, challenge, and improve models. In the context of the analysis of player communications data, for example, such as customer emails and online chats, operators can categorize the frequency, intensity, and complexity of such interactions to see if any gambling *red flags* emerge. In this example [11], adding a human-in-the-loop check after the algorithm's result has increased the chances of identifying harm correctly while at the same time not inconveniencing the core player base.

3 METHODOLOGY

The IGT dataset is based on gambling behavior data made available by IGT from 2009 to 2011 for a sample of 669 control group players and 176 qualifying self-excluders who self-excluded for at least six months. The spin-level play data is manipulated in a number of ways to identify behavior markers that represent known aspects of risk, such as how much time gamblers spend online and how much they bet.

For full details of how the dataset was developed and behavioral markers generated, please see [15].

Our proposed approach for understanding gambling behavior through neural networks is composed of three steps: gambling data analysis, neural network training with backpropagation, and knowledge extraction using TREPAN. In the first step, gambling data analysis, an evaluation of variable relevance and redundancy is carried out on the IGT gambling data using the variable ranking approach mRMR [7]. Further, over-sampling was carried out using the SMOTE algorithm [3], which has been shown to perform better than other methods of dealing with dataset rebalancing. We have applied the optimal SMOTE level to achieve an approximately 50:50 split between control group and self-excluding cohorts. Then, network training is carried out using standard backpropagation with momentum and early stopping [15]. Finally, following network training, in order to perform knowledge extraction, a modified version of the TREPAN algorithm is applied to the trained neural network. We have adapted TREPAN in order to allow its efficient application to the domain of gambling behavior prediction, as follows: tree generation has been simplified with only the *maximum-size* criterion being used for stopping the process, and the search heuristic for best M-of-N split now takes into account the size of M by subtracting M/N from the original heuristic value for a given split. In this way, smaller values of M are preferred (larger values of M are penalized), leading to rules with fewer antecedents than the standard TREPAN and, hopefully, a better interpretability, as such rules should be easier to read.

In the above process, we are interested in evaluating the amount of accuracy loss which is expected when TREPAN is applied to produce an interpretable model from a neural network. The existing trade-off between accuracy and comprehensibility is well-known [1]. In this specific domain of application, gambling behavior, further work might also investigate whether other measures of relevance, in addition to accuracy, might be appropriate. As usual, we are also interested in evaluating the fidelity of the extracted model (decision tree) w.r.t. the neural network, as opposed to w.r.t. the data itself, as in the case of accuracy values. The next section contains the experimental results with such evaluations.

4 EXPERIMENTAL RESULTS

In this section, we present the results on knowledge extraction. We have used ten-fold cross validation and, below, we compare the performance of neural networks, TREPAN (including fidelity to the network), and decision trees obtained directly from the data.

Results show that extracted rules have highly competitive accuracy in comparison with the trained neural network, at around 79% vs an original 80%. A high fidelity rate of 87% of the TREPAN tree to the original neural network was achieved. Accuracy results of both the neural network and extracted TREPAN tree shown an improvement in relation to [17], suggesting that this approach is competitive in relation to previous analyses in the same domain.

In assessing whether extracting a decision tree from the neural network via TREPAN is a worthwhile process, we compare performance of the TREPAN tree with two decision trees created directly from the data, one with unlimited height and one restricted to the same height of the decision tree created by TREPAN. The decision trees were constructed in H2O with 10-fold cross validation, using the same approach to over-sampling as done for the neural networks.

Method	Accuracy [Fidelity for TREPAN]	Decision Tree Leaves	Decision Tree Height
Neural Network	79.8%	-	-
TREPAN	**78.8% [87.4%]**	**9**	**5**
Decision Tree (unlimited height)	77.3%	168	20
Decision Tree (max height 5)	75.1%	25	5

Table 1: Predictive accuracies, i.e. average test-set performance post SMOTE for the trained neural network and the extracted TREPAN tree in comparison with decision trees.

The results in Table 1 show an improvement in accuracy by TREPAN in comparison with both decision trees. More importantly, there is also a significant improvement in decision tree simplicity (and, hence, the likelihood of human readability) via the TREPAN approach. Even the decision tree with a restricted height of five layers, as produced naturally by the TREPAN algorithm, has a far more complicated decision structure, resulting in 25 separate routes that result in a label prediction, as opposed to the 9 leaves deployed by the TREPAN model. However, it is still necessary to determine whether the simpler structure of the TREPAN model is indeed human-readable in a meaningful sense and whether it can thus be subjected to human validation and application.

In assessing the readability of the TREPAN output, Figure 1 contains a visual representation of the entire TREPAN decision tree. While the extracted decision tree remains somewhat complex, it is possible to read it and infer certain conclusions from the split nodes, enabling further validation and exploration by industry professionals and domain experts, as follows:

The majority of self-excluders are identified due to either a weak flag on “Variability” or a strong flag on “Intensity” (at least a 22% increase in the number of bets placed). However, players flagging both on “Variability” and strongly on “Intensity” must also be men (approximately 80% of the sample are) and flagging high on “Frequency” to be assessed as self-excluders. This is a minor part of the decision tree as only 3% to 5% of the samples will flag on both “Variability” and “Intensity”. Otherwise, you are likely to be a control group player regardless of your score on

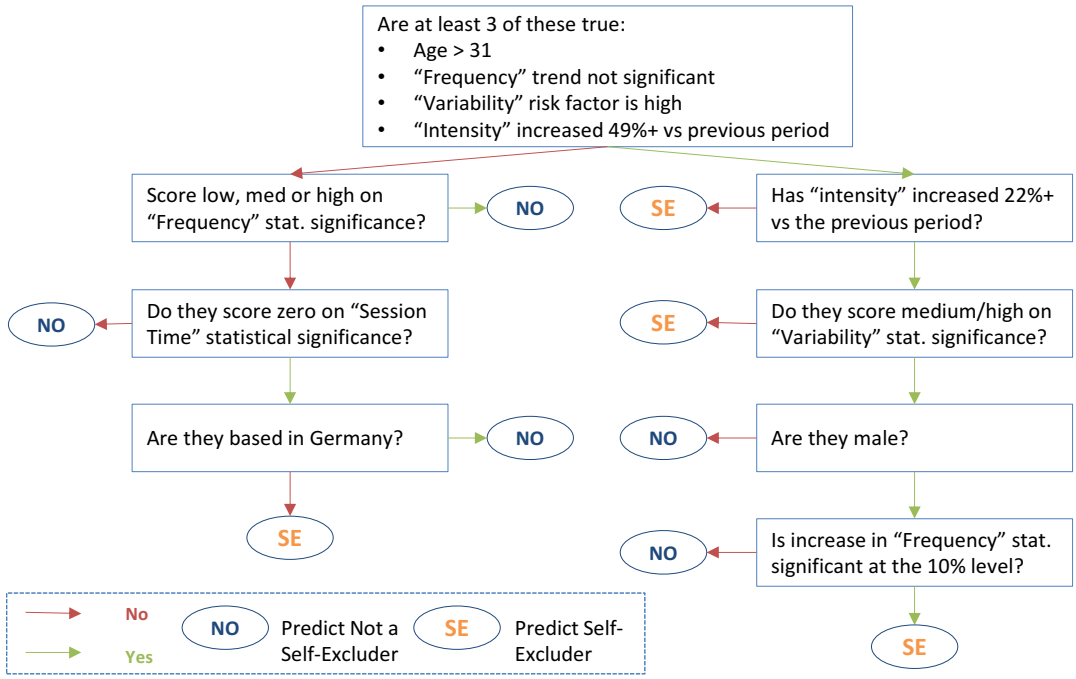

Figure 1: TREPAN decision tree. Phrases in double quotation marks refer to behavioral Risk Factors which were input to the neural network obtained by pre-processing the raw spin-level play data from IGT. For instance, "Trajectory" refers to a player's total amount of money wagered over an active gambling day and how it changes over time, "Variability" refers to the standard deviation in total amounts wagered over time, "Intensity" refers to the number of bets placed per day, "Frequency" refers to the share of calendar days on which gambling takes place, and "Session Time" refers to the time spent gambling online [15].

"Frequency", unless you are based outside Germany in which case you could be assessed as a self-excluding player, provided you have also flagged, even if only very mildly, on the "Session Time" risk factor (although some 75% of players in the sample were based in Germany).

The first two bullet points in the root node of the decision tree in Figure 1 suggest that "Variability" and "Intensity" are more important risk factors for identifying self-excluders than "Session Time" or "Frequency". These other risk factors still play a role in predicting self-exclusion, but it is a more nuanced role based on the interactions between risk factors and demographic circumstances. How risk factors work together to better isolate self-excluder behavioral patterns would benefit from further examination, considering that risk factors are (in theory at least) often correlated with each other, e.g. someone who spends more time online than before is also likely to be wagering more money in total and placing more bets.

The influence of Germany-based players on the model is small, since it appears at the lowest level of the decision tree with only a small proportion of the samples still present to be differentiated at that node. Nonetheless, the presence of Germany in the model remains hard to interpret. It could be a quirk of the sample used for this analysis, or it could reflect some cultural or structural feature of online gambling in that market relative to other European gambling markets.

Industry Insight on the Importance of Algorithm Interpretability: While presenting a related paper at the 2016 New Horizons in Responsible Gambling conference [14], we polled the audience and conducted interviews to explore the importance of knowledge extraction and algorithm interpretability (see Table 2 for a summary of the polling data). By a combination of show of hands and smartphone-based electronic polling, the audience was made up of approximately 40% representatives from gambling operators and casinos, 25% regulators and policy officials, 25% academics and other attendees, and 10% clinicians and problem gambling treatment providers. Depending on the question, the sample size was 35 to 40 respondents.

Q: Which would you prefer: an algorithm that assesses problematic play that is 90% accurate which you cannot properly understand or explain OR one that is 75% accurate which you fully understand or explain?

	Total (incl. show of hands)	Gambling operators	Treatment providers	Regulators/ policy makers	Academics/ Others
Sample Size	40	8	2	6	6
Prefer 90% accurate	20%	13%	-	-	17%
Prefer 75% accurate	70%	75%	50%	67%	83%
It depends /Unsure	10%	13%	50%	33%	-

Q: Is it more important to have model-level interpretability or individual-level interpretability?					
	Total (incl. show of hands)	Gambling operators	Treatment providers	Regulators/ policy makers	Academics/ Others
Sample Size	35	8	2	5	7
Model-level	20%	25%	-	-	14%
Individual-level	26%	25%	-	-	29%
Both	46%	50%	100%	100%	43%
Neither /Unsure	9%	-	-	-	14%

Table 2: Audience Polling Results from New Horizons in Responsible Gambling Conference, Vancouver, February 2016. Note that audience members contributing via show of hands were not segmented into one of the four categories.

Respondents were asked whether they would prefer a responsible gambling assessment algorithm that provided a 90% accurate assessment of problem gambling risk that they could not explain or understand, or a model that provided a 75% accurate assessment that was fully interpretable and accountable. Only 20% chose the more accurate model, but 70% preferred to sacrifice 15 percentage points of accuracy for greater interpretability; 10% were uncertain or felt it depended on the circumstances, which means overall a significant majority for the interpretable model.[2] This pattern was broadly consistent among gambling operators, policy officials and treatment providers.

We also explored the value of two different types of interpretation for the audience, namely model-leicevel and individual interpretability. Model-level interpretability entails understanding which inputs and their values are most important for determining a prediction in the model. It does not change from gambler to gambler. Examples of model-level interpretability are coefficients in a logistic regression or the analysis of the TREPAN decision tree above. Such interpretability enables users to challenge, test and gain confidence in the model (since models and model development techniques are known to be imperfect) and understand its strengths and flaws. It might also enable policy makers to take industry-level action based on model insights (e.g. identify if some casino games are very high risk) or point towards ways to simplify the model and get similar accuracy, which might matter for real-time prediction systems or online learning.

The second type is individual-level interpretability, where one can explain specifically why a particular individual was given a particular risk assessment score, what factors contributed most to it, and what the individual might change to not gain such a risk assessment in future. This is important when one wants to explain a particular decision to someone, since providing a more detailed explanation might help gamblers accept the assessment and take action accordingly. An example of individual-level interpretability is the frequency analyses that can be done in the case of random forests, which are difficult to interpret as a whole but easy to analyse in individual cases by comparing the outcomes of the various decision trees.

In our survey, 46% of respondents said that both model-level and individual-level interpretability were important to them. Of those who said that only one type of interpretation was important, there was a slight preference for individual-level interpretability at 26% vs 20% for model-level. Only 8% were not sure or felt that neither was important. Each level had a significant majority voting for its importance.[3] All regulators and public policy officials indicated that both types of interpretability were important.

While this polling evidence presents a clear picture of importance of interpretability, and indeed a willingness to sacrifice some accuracy in favour of clearer models, subsequent interviews with treatment providers identified an insightful way to use both types of models to better protect gamblers. The most accurate model, even if as a black-box, can be used to most accurately identify which gamblers require a responsible gambling intervention, for instance with a responsible gambling message about the player setting a limit, a cessation of email marketing by the operator, or a conversation over a cup of tea in the case of venue-based players. In the case of messaging or a conversation, while the black box can best identify which gambler would benefit from such an intervention, the less accurate but more understandable model can be used to determine the content of these interactions, thus helping the player appreciate and come to terms with their patterns of behaviour and determine what action they might take in the future. The assumption in this approach is that, even with discrepancies in accuracy, both models are fundamentally exploring and interpreting the same underlying features of a person's gambling behavior. Alternatively, we can see that, provided descriptions of a gambler's behavior are relevant and easy to understand, they can be used by professionals in conversations or interventions, even if they are not the most precise predictors of the associated machine learning model.

5 DISCUSSION AND CONCLUSION

In the neural network used here to classify gambling self-exclusion, each input can affect the network's output via 34 different possible routes with an overall model parametrized by over 500 weights. Such a complex set-up, while fully determined mathematically, constitutes a black box from the point of view of human readability. While the model

[2] We compared votes for 'more accurate' vs. those 'explainable' + 'undecided' in a one-sided Binomial test (n=40), which yielded $p<0.01$ with null hypothesis of random choice with equal probabilities.

[3] For each level, we applied a one-sided Binomial test (n=35), i.e. to 'both' + 'individual' vs. 'model' + 'undecided' votes yielding $p<0.01$, and to 'both' + 'model' vs. 'individual' + 'undecided' votes yielding $p<0.05$ against random choice with equal probabilities.

produces both a predictive assessment of each player and a view of how accurate that assessment is likely to be on average, it fails to permit challenge and validation by domain experts and it fails to allow users to explain risk assessments to gamblers, which might offer the potential for such explanations to lead to changes in behavior.

The reduced form decision tree generated by applying a small variation of the TREPAN knowledge extraction algorithm resolves these concerns with only very minimal loss of overall accuracy (approx. 1 percentage point) and with 87% overall fidelity to the neural network model. The reduced form tree remains somewhat complex, with a height of 5 and 8 distinct decision points at split nodes, showing interactions between the input risk factors that are not necessarily straightforward to interpret. Nonetheless, the reduced form model is human readable and can be translated into a series of statements that are meaningful to domain experts. Simultaneously, it is possible to trace the route of an individual gambler through the decision tree to identify at which points they become more or less likely to be assessed as a self-excluder. Industry leaders and regulators have also indicated that it would be possible to use a simplified, interpretable model, such as the TREPAN decision tree, alongside a more complicated, more accurate model, in order to best serve gambling clients.

However, in [15] random forests were the highest performing method on the same dataset used here, with a lower standard deviation in accuracy across the ten-fold cross-validation and an even balance between sensitivity and specificity, unlike the unhelpful bias towards specificity in the case of the neural network. On this basis, we suggest that a TREPAN decision tree derived from this neural network may not be an optimal approach and its human-readable conclusions should not be treated as fixed. We hope that further work will apply the TREPAN approach to models such as random forests as well. As a result, this paper demonstrates the industry need for different forms of model interpretability and knowledge extraction, alongside the effectiveness of TREPAN as a tool in enabling this requirement to be met.

In what concerns the challenge of using the knowledge extraction results obtained here in a clinical capacity, it is important to note that the players' likelihood of self-excluding was analyzed here by modelling their play data. One could not assess, for example, whether any of the gamblers in our study suffered from poor coping or problem solving skills, or negative family experiences, or suffered from behavioral problems such as substance abuse, which are important elements in pathways 2 and 3 [2]. In order to enable that, one would have to augment the model with data which is much more difficult to obtain, requiring interviews and observations from gambling operator staff in a physical casino or retail environment, for example. That would enable a more detailed assessment, however, of any negative consequences our players may suffer as a result of their gambling behaviors.

In this paper, we have evaluated how well neural networks can be used to classify but more importantly describe and therefore explain self-excluding gamblers. Understanding gambling behavior can lead to gambling studies that develop new techniques and models capable of managing and helping gambling operators or treatment providers handle problematic players. As our model shows, there are a number of factors that might indicate problematic gambling behavior, and we highlight the value of further investigation on why and how these factors can be used in practice to support gamblers.

ACKNOWLEDGEMENTS

This work is partially supported by Innovate UK and the EPSRC under grant number EP/M50712X/1.

REFERENCES

[1] 1. R. Andrews, J. Diederich, and A. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks, Knowledge Based Systems, 8(6), 373-389, 1995

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

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

[4] Craven, M., and Shavlik, J. (1996). Extracting Tree-Structured Representations of Trained Networks. *Advances in Neural Information Processing Systems*, 37–45.

[5] M. G. Augasta and T. Kathirvalavakumar, Reverse engineering the neural networks for rule extraction in classification problems, Neural Processing Letters, vol. 35, no. 2, pp. 131–150, 2012.

[6] Diakopoulos, N. (2013). Algorithmic Accountability Reporting: On the Investigation of Black Boxes (Available at http://towcenter.org/wp-content/uploads/2014/02/78524_Tow-Center-Report-WEB-1.pdf, last accessed 2016/03/08)

[7] Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, *3*(2), 185–205.

[8] Dragičević, 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.

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

[10] Franca, M. V. M., Zaverucha, G., and Garcez, A. S. D. (2014). Fast relational learning using bottom clause propositionalization with artificial neural networks. *Machine Learning*, *94*(1), 81–104.

[11] Dragičević, S. (2016). Human-in-the-Loop Machine Learning and Responsible Gambling Analytics (available at: http://bet-buddy.com/blog/human-in-the-loop-machine-learning-and-responsible-gambling-analytics/, last accessed 2016/03/22)

[12] Gainsbury, S., Wood, R. (2011). Internet gambling policy in critical comparative perspective: The

effectiveness of existing regulatory frameworks. *International Gambling Studies,11(*3), 309-32.
[13] Garcez, A. S. D., Broda, K., and Gabbay, D. (2001). Symbolic knowledge extraction from trained neural networks: A sound approach. *Artificial Intelligence*, *125*(1-2), 155–207.
[14] Percy, C. (2016). Can 'BlackBox' responsible gambling algorithms be understood by users? A real-world example. New Horizons in Responsible Gambling conference paper, Vancouver, February 2016.
[15] Percy, C., Franca, N., Dragičević, S., and Garcez, A. S. D. (2016). Predicting online gambling self-exclusion: an analysis of the performance of supervised machine learning models. *International Gambling Studies*. DOI:10.1080/14459795.2016.1151913.
[16] K. Odajima, Y. Hayashi, G. Tianxia, and R. Setiono, Greedy rule generation from discrete data and its use in neural network rule extraction, Neural Networks, vol. 21, no. 7, pp. 1020–1028, 2008.
[17] Philander, K. S. (2013). Identifying high risk online gamblers: a comparison of data mining procedures. *International Gambling Studies.* DOI: 10.1080/14459795.2013.841721
[18] Karim, A. and Zhou, S. X-TREPAN: a multi class regression and adapted extraction of comprehensible decision tree in artificial neural networks. arXiv:1508.07551, Aug, 2015.
[19] R. Setiono, B. Baesens, and C. Mues, "Recursive neural network rule extraction for data with mixed attributes," IEEE Trans. Neural Networks, vol. 19, no. 2, pp. 299–307, 2008.
[20] T. Schellinck and T. Schrans (2011). Intelligent design: How to model gambler risk assessment by using loyalty tracking data. *Journal of Gambling Issues: 26*(1), 51-68.
[21] A. Hara and Y. Hayashi, Ensemble neural network rule extraction using RE-RX algorithm. International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia, June 10-15, 2012, 2012.
[22] T. A. Etchells and P. J. G. Lisboa, Orthogonal search-based rule extraction (OSRE) for trained neural networks: a practical and efficient approach, IEEE Trans. Neural Networks, vol. 17, no. 2, pp. 374–384.