Understanding the Risk Profile of Gambling Behaviour through Machine Learning Predictive Modelling and Explanation

S. Dragicevic
Playtech plc
London, WC1V 6EA, UK
simo.dragicevic@playtech.com

A. d'Avila Garcez
City, University of London
London, EC1V 0HB, UK
a.garcez@city.ac.uk

C. Percy
Playtech plc
London, WC1V 6EA, UK
christian.percy@playtech.com

S. Sarkar
Playtech plc
London, WC1V 6EA, UK
sanjoy.sarkar@playtech.com

Abstract

The importance of providing algorithms and tools for experts to be able to analyse and explain black box learning models has been greatly acknowledged recently. This includes methods for opening the black box by providing descriptions of the entire learning model, methods for explaining individual cases, simplification and visualization methods to achieve comprehensibility, feature ranking methods and *teacher-student* methods which seek to create simpler and more interpretable, surrogate models within an acceptable loss of accuracy. In this paper, we propose a simple and efficient method for analysing how a chosen feature may influence the outcome of a classifier. The method produces curves which can reveal the trend of a feature value *all else being equal*, thus complementing feature ranking methods. Specifically, we show that it can help domain experts explore the directional impact (as opposed to the scale) of predictions as a feature value changes. Trends in the average values of a chosen feature are characterised w.r.t. a target outcome and visualized alongside error bands to support assessments of significance. The method is scalable and applicable in principle to other learning systems and domains. Differently from decision trees or rule-based approaches, it offers a visualization of feature value trends which can be used by domain experts for the iterative analysis and understanding of the learning system. We have applied this method, called feature risk curves, to the analysis of a real system used by Playtech Plc for the protection of online gamblers. The system predicts whether players may be at risk and recommends a break from the game. The curves have proved insightful in discussions with sector experts for understanding the model-level drivers of particular features towards gambling harm, providing nuance to the mainstream understanding of features like *night-play* or *declined deposits*. In particular, the use of error bands around the curves have helped us "know what we don't know", by drawing attention to the limits of what can be said about single features of player behaviour. Risk curves for consumer protection remain an active area of research for the team, with planned extensions listed at the end of the paper.

Keywords: Explainable AI, Knowledge Representation and Machine Learning, Gambling Harm Protection.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

1 Introduction and Motivation

Machine learning systems are now being used for automated decision making in areas such as security, finance, autonomous vehicles, robotics and healthcare. Yet the inner workings of state-of-the-art machine learning (ML) systems is frequently referred to as a black box. The size and complexity of the learned model calculations are considered to be beyond the capacities of humans to understand. Due to the recent impressive success of black box ML, in particular deep learning, in areas such as vision and speech recognition, many methods have been proposed which seek to provide explainability to the black box [4,5]. There are a number of reasons why explainability is considered increasingly important, including i) being able to provide a readable interface for humans who rely on ML applications to support decision making, ii) to test and debug ML applications to improve accuracy, iii) to better understand model blind spots and flaws, iv) to help ensure human biases are not encoded in ML models. It is worth noting, however, that the area goes back at least twenty years, when it was mainly focused on what was called knowledge extraction from neural networks [1]. In such early systems, the objective was to seek to extract interpretable symbolic rules or decision trees from trained neural networks.

Nowadays, the quest to develop explainable AI systems can be divided into two approaches: those seeking to provide (global) explanations of the entire system and those providing (local) explanations for individual predictions. TREPAN [1] is a global explanation system; the decision trees that it produces by querying the ML system - which in this case serves as an oracle - seek to represent the behaviour of the model as a whole in a simplified form. By contrast, as an example of a local explanation system, in [6] it is argued that single predictions can be explained by counterfactuals stating the minimum changes needed for an observation to change its classification. An example of a counterfactual explanation could therefore be: you would have received a loan, if your annual salary had been US$50,000 instead of the current US$42,000. In LIME [5], for example, local explanations are created by building a regression model that seeks to approximate the local input-output behaviour of the system. In [7], local *if-then* rules are created from data produced by querying the ML model.

While global methods may be costly computationally to extract from very large ML systems such as deep networks, local methods have been criticised for being difficult to evaluate: counterfactuals do not provide explanatory generalizations or an indication of a trend which one may obtain for example from the coefficients of a regression equation. LIME, on the other hand, does not measure its *fidelity* to the original ML system and has been shown to produce regression equations which are unrelated to the predictions of the underlying ML model [13].

The contribution of this paper is two-fold:

- We introduce feature risk curves as a high-fidelity method for analysing trends in the underlying ML system.
- We apply and evaluate the method to reducing harm from gambling showing that it is capable of producing descriptions of relevance to industry.

The feature trend analysis method for producing feature risk curves relies on querying the ML system, which serves as an oracle. It is therefore model agnostic and applicable to very large ML systems. Differently from TREPAN and other global methods, the querying that we use is feature specific, following a feature ranking. This makes our approach generally more scalable than global methods. The approach is global in the sense that it does not seek to explain individual cases (i.e. it is not local), but it does not seek to explain the entire system either, which is why it is scalable. It is therefore global but feature driven or *feature-specific global*. We say that it has fidelity by design because all the points plotted in the curves come from the underlying ML system through direct querying, with their outcomes therefore being the same as that of the ML system. As a result, fidelity with respect to a specific given test set becomes less relevant in this case. Instead, a human-in-the-loop approach is taken. By that we mean that our approach is intended to enable human experts to analyse application-relevant features and to derive insights from the analysis. It is correct therefore to assume that the approach proposed here will be of less value in general for applications such as image or speech recognition where the input features (e.g. pixels) may not be relevant for human analysis.

Application within gambling harm reduction: Supervised machine learning has become increasingly relevant recently as a method to help gambling operators better protect players from harm [9, 10]. Given the combination of public health concerns, corporate interests, regulatory pressures and

media attention at play, there has been significant pressure within the gambling industry to understand better how such ML models work [11, 12]. Of particular relevance to this paper is the desire to be able to describe which value ranges of specific player features the ML model interprets as being associated with higher levels of harm (ignoring interactions and co-variance with other player features). This has two sector-relevant applications: it helps to challenge and thus validate the model through a better understanding of its assessments, hopefully building up confidence that it is identifying the intended underlying issues rather than some superficial aspect of the data or function of the sample / modelling design. And it contributes to a broader discussion, along with other inputs, about which player behaviours may embody higher levels of risk, and hence it can contribute to the development of safe play guidelines and better intervention messaging for players.

The feature risk curves developed in this paper are applied to a supervised machine learning model developed by BetBuddy, a responsible gambling analytics subsidiary of Playtech plc. The model seeks to identify players whose recent playing behaviour showed a pattern closely matching that of "serious self-exclusion", defined as players who formally requested the relevant gambling operator not to let them play on that system for six months or more. Such "serious self-exclusion" is a proxy that captures a subset of players experiencing harm, typically those who have gambled on a site for several weeks, with a pattern-matching philosophy that identifies a separate subset of players who might be experiencing harm even if they would be unlikely to self-exclude.

In summary, we propose an oracle-based explainability method to produce risk curves that may apply to any black box classifier. We seek to understand the nature of the relationship between an individual feature and the predicted average outcome at model-level, not at the level of a single case, in a way that holds the values of other variables constant. When applied to an ML system for reducing harm from gambling and evaluated by domain experts, the proposed explainability method is shown to be useful, providing a quantitative lens and important nuance to challenging issues such as the proportion of play that takes place at night time and players' deposit frequency, and indicating the types of players for which self-exclusion-based models might be best.

In Section 2, we include the relevant background and related work that sets the objectives for the work. In Section 3, we describe the algorithm and charting approach to creating feature risk curves. In Section 4, we share two risk curves drawn from one of BetBuddy's prediction models as applied to a mass-market UK online bingo and slots provider, and a brief summary of the insights based on discussing these curves within the BetBuddy team and with five sector experts. In the conclusion, we discuss our plans for further work also to mitigate some of the limitations of the approach. This is an active area of research for us and interested researchers are invited to contact any of the authors with suggestions or comments.

2 Background and Related Work

Wachter et al.'s counterfactuals explain a single prediction by identifying 'close possible worlds' in which an individual would receive the prediction they desired [6]. A counterfactual explanation may involve changes to multiple features; an example might be: you would have received a loan if your salary had been US$47,000 (instead of US$40,000) and you had been employed for more than 5 years (instead of 3 years).

The key problem with counterfactuals is that it fails to satisfy Woodward's requirement [8] that a satisfactory explanation of prediction Y should state a generalization relating X and Y. If an ML system assigned a probability of 0.75 of a client defaulting on a loan, stating the changes needed to salary and years of employment has explanatory value but falls short of being a satisfactory causal explanation according to Woodward [8] since it does not offer a generalization of X and Y. In this paper, although strict claims of causality are not made, we argue that the method proposed here is one step closer to satisfying Woodward's requirements by offering a more global perspective than counterfactuals, as detailed below.

With feature-specific global explanations, our objective is to be able to provide explanations of the form: given a target feature, e.g. average deposit per day, average loss per day, frequency of play at night-time, etc., if the feature values were below a certain threshold then the model would have reported an average self-exclusion probability a number of percentage points lower or higher than when the feature values fall in a different given range, assuming that no other features had changed (i.e. holding other features or player behaviours constant, even if they would normally co-vary with

the target feature). It will also be important to comment on the significance of this difference in value ranges relative to the proportion of the training set whose typical co-variance across other feature values might fail to dominate that difference.

Such analysis falls short of demonstrating causality for the following reasons: there is no hypothesised causal chain being tested, such that the co-varying relationship between the target feature and other features may be key for causal understanding and for model logic (rather than being held fixed as in this approach) and no insight is available with respect to change in individual outcomes given a change in individual behaviours, such as may be possible by experimental design or at least a player-level longitudinal analysis. However, the analysis may inform theoretical developments of possible causal chains, especially when combined with insights on the direct relationship between the target feature and the outcome variable outside of the model in a real-world environment, i.e. an environment in which other features are allowed to co-vary with the target feature.

Perhaps closest to the system proposed here is [2] which produces curves similar to the ones produced in this paper but which applies to regression tasks rather than classification tasks. Teacher-student models which seek to train e.g. a regression model from data sampled from a complex neural network or random forest have been recently considered inadequate for explainability (e.g. in the banking sector). As noted at the panel discussion at the NeurIPS 2018 workshop on AI in finance, it is the actual black box model that needs explaining, not a surrogate model (which in the above example would be the regression model).

As a well-studied global method, TREPAN seeks to simplify a neural network (or supervised ML system treated as an oracle) into a simplified decision tree in the form of M of N, N of N, 1 of N, or 1 of 1 rules. Such a global rule extraction method provides one way of generating a narrative out of an oracle but does not explicitly seek to explain the relationship between feature values and a target classification. In practice, only some TREPAN trees are simple enough to permit such reading [12].

Other related work includes a large number of recent visualization methods and local explanation methods such as LIME. As pointed out earlier, these are local while our intention with this paper is to provide a global method that is useful in practice. The reader is pointed to this recent survey for more information on related work [14].

3 Feature Risk Curves

The feature risk curve for a target feature X_i, out of k possible features, is generated by the following method, assuming we have a machine learning system ("m") which was trained from labelled data to map a vector X into a vector Y; m is said to generate prediction Y for observation X.

We have a total set of n observations each with input data $X = X_1, ..., X_k$ (and at least some with Y data providing the labelled outcomes used to train the original ML system).

1. Take the largest available set of features with their original feature values in X discarding the original labels Y;

2. For a target feature X_i, replace all of the original values of that feature across that set with the 1st percentile value of that feature (across the entire data set). Keep all other features at their original value. This creates a new vector for all n, which we can denote X^1. Note that X^1 creates an artificial set of observations that may be - at places - unrealistic in that there may exist dependencies between X_i and X_j meaning that certain vectors could not occur in reality or would only occur very rarely. The intention is not to create "realistic" artificial observations (in our case, gambling players) but to understand the significance of the target feature for the model, if it were able to vary independently;

3. Run X^1 through the model m. Record each probability value that the model assigns to a certain outcome Y^1 for each of the n observations (simple point estimates can be obtained by recording the percentage of the time that an outcome Y^1 is assigned, with an extension that weights the prediction with an indication of the model's confidence in that classification outcome);

4. From the n probability values for data set X^1, derive standard statistics for analysing distributions. By default, we derive the median, the 25th and 75th percentile values, and the 40th and 60th percentile values.

5. Repeat from (2) using the 2nd percentile value of X_i to create a new vector X for all n, X^2. Repeat until the 100th percentile value has been generated and descriptive distributional statistics have been recorded for the outcomes $Y^{1..100}$.

6. Chart the results such that the x-axis records 1st to 100th percentile values of X_i and the y-axis captures the distribution of the resulting classification probabilities at each of the 100 x-axis values. By default, we plot the median as the core curve to capture the directional implications of changing the target feature, bounded by two error bands: the 20th percentile band around the median (i.e. the range from the 40th percentile value to the 60th percentile value) and the interquartile range (from the 25th to the 75th percentile value) shown as two separate curves above and below the 20th percentile band.

This produces a curve which can be approximately interpreted as how much that feature influences the outcome's average probability for each value of that feature, all else being equal. In this way, it provides an indication of the *relative direction and shape of the impact* of a feature on a model prediction as that feature varies, to the extent that such an impact dominates, regardless of the values of other features (distributed according to the training data). This goes beyond and complements the *scale of impact* insights that can be obtained from feature ranking methods. Notice how feature ranking can be used in combination with feature curves whereby the ranking would indicate the curves to generate first for analysis, with the ranking order being used for the selection of target features X_i.

The two error bands help assess the significance of a movement in the median as a feature value varies (see Figure 1). If the 20th percentile bands at two different value ranges do not overlap, the change might be considered "weakly significant" in that the movement in that feature alone is sufficient to account for the difference in probability of classification given the typical co-variation of all other features for at least 20% of the training set. If the range percentile bands do not overlap at two different value ranges, it might be considered "strongly significant" in that the movement alone moves beyond 50% of the training set. The interpretation of different band sizes is indicative and subjective, with norms yet to be established as for p-values in different fields, but can serve a similar purpose in helping compare movements across value ranges both within and across features.

4 Practical Evaluation

A number of feature risk curves were produced from a "serious self-exclusion" prediction model trained for a UK mass market online bingo and slots provider.

By way of background on the trained model, each data sample in the training data set is a vector of 41 features representing playing, deposit, withdrawal and limits setting behaviour of one gambler. The target variable is binary, and represents whether a player is a serious self-excluder or not. A self-excluder is a player who has voluntarily taken a break from the game (deemed serious if the break is at least six months long), typically but not always as a result of being concerned at their play and the harm that they might be experiencing. The purpose of the ML system is player protection: to identify whether other players might have similar play habits to those who have self-excluded and therefore might benefit from support or interventions to manage their play, up to and potentially including a recommended self-exclusion by the system. The training data set is a 50:50 balanced data set with an equal number (3564) of data samples for serious self-excluders and other players who have not ever or not yet self-excluded (termed as *control group*). Originally, there were 891 serious self-excluders in the training data set. Synthetic data samples were generated using SMOTE [16] to make the number at par with the control group. The machine learning system which achieved the best performance is a random forest model with 100 trees, with the random forest outperforming neural networks, Bayesian networks and logistic regression in an earlier gambling sector prediction exercise [10] (we note though that the feature curves proposed here should apply to neural networks and other supervised ML models too). The random forest has a ten-fold cross-validation accuracy of 92%, and a hold-out test set accuracy of 76%, where the hold out set contains only data points not contained within the time range of the data used to build the model. This offers a fairer reflection of model performance post-deployment. Interested parties are welcome to contact the authors for more information about the ML modelling for consumer protection in gambling.

The results were discussed within the project team and were also reviewed, in simplified form focusing on the point estimates for self-exclusion, with five sector experts: A UK-based responsible gambling

academic with expertise in game design and game features; a compliance lead at Playtech for a business unit providing gambling services direct to consumers; an analytics manager at the Ontario Lottery and Gaming Corporation which provides gambling services directly to consumers, a member of the Responsible Gambling Council in Canada who is an expert in developing and evaluating responsible gambling training and treatment programmes; and a risk and compliance director at a major online gambling operator. All of our experts would be required, as part of their professional work, to either evaluate or deploy such machine learning models in real-world environments. Given space constraints, two curves are drawn out below which were discussed in our expert interviews: night-play ratio (Figure 1) and deposit decline ratio (Figure 2), although others were considered which are not shown here e.g. deposit frequency, play frequency, etc.

Figure 1: Feature risk curve for Night-Play Ratio

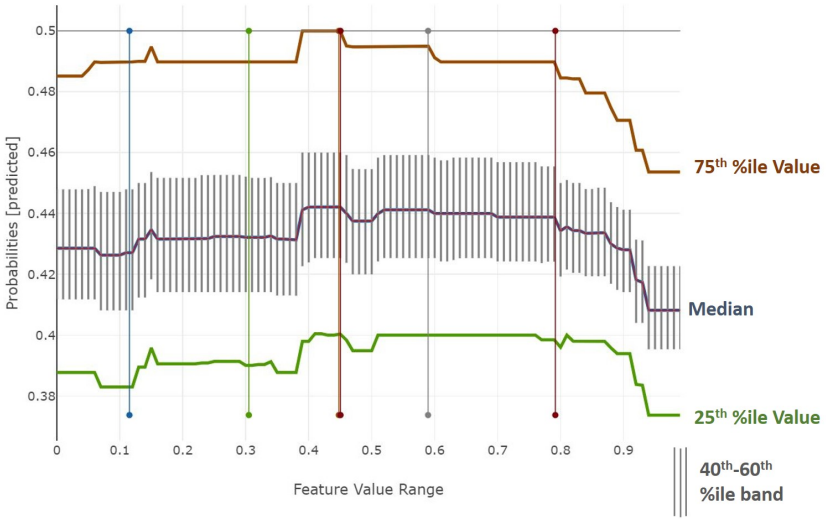
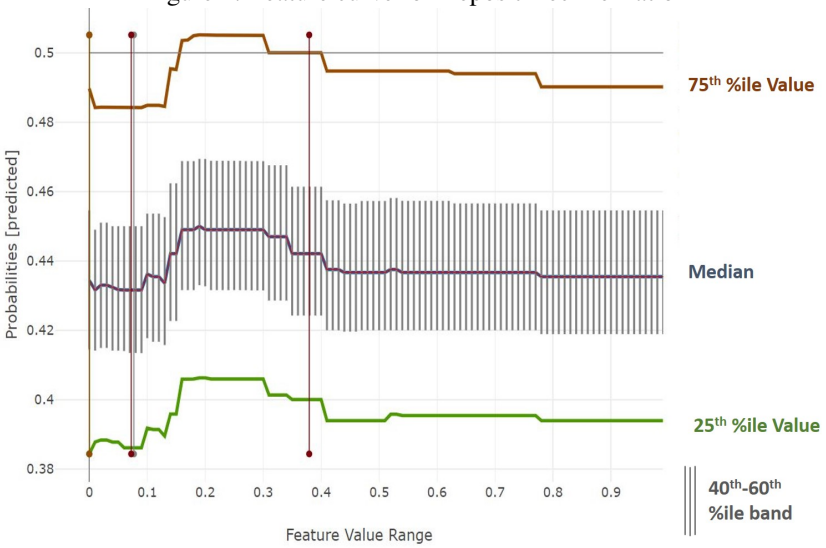


Figure 2: Feature curve for Deposit Decline Ratio



The thin vertical lines show the distribution of feature values in the data set (blue, green, brown, grey and dark crimson show the 5th, 25th, 50th, 75th and 95th %ile values respectively, with the mean shown by a dark brown line).

Night-play ratio analysis and commentary. In the first curve (Figure 1), Night-Play Ratio - the proportion of online gambling session time that players did between midnight and 8am in the 90 days prior to a self-exclusion or non-self-exclusion event - is typically seen as a risk factor for harmful

gambling. A high night time play ratio can suggest that a player is at greater risk of harm as most recreational players gamble during the evenings and weekends (if they work) or during day time. Whilst there may be genuine reasons for gambling at night (e.g. shift work or one-off events like the Superbowl), a higher than average ratio is thought in general to increase risk due to potential for impaired decision making while tired and negative impacts on work and family life. Most industry discussion presents this simplistically and directly, as can be seen in the large-scale analysis conducted for the UK sector by PWC in 2017 [15] in which insights are drawn from a logistic regression model to say that "time of day" (i.e. night-play) has a negative statistical relationship to problem gambling (c.f. [15], Table 16) and problem gamblers are "more likely [...] to bet late at night" ([15], p.44).

This feature risk curve supports the mainstream assessment to an extent, but adds important nuance to it. The model sees an increase in average self-exclusion probability if players go from never playing at night to playing a little bit, but not significant by the error bands chosen for this analysis. The highest probability values occur from 38% to 68%, holding other behaviours around betting and depositing constant. However, after this level, average self-exclusion probability declines and in a way that is weakly significant compared to the peak value ranges. By the time someone is playing almost entirely at night (in the top 5% of the value range, marked by the dark crimson vertical line, and especially from 94% night-play and above), the model sees less self-exclusion than in day-players.

Discussion with experts interpreted this as the distinction between night-play when it is typically a regular, stable habit (i.e. when night-play is the clear majority of play; e.g. shift workers) and night-play when it represents evening sessions leaking later into the night as players struggle to stop (e.g. “I’m on a roll, I’ll just play one more hour” or “I didn’t mean to lose that; I’ll play a bit longer and win it back”). The latter set of players may sometimes play longer than planned, being more tired for work the following day or stressed at having lost self-control, which may be less frequently an issue for the former group of consistent night-players.

Deposit decline ratio. The second curve (Figure 2) covers the proportion of deposit attempts in the last 90 days that were declined by the player’s payment provider. This is considered a clear indicator of risk, reflecting losing control of spending, potentially losing track of how much money they are losing such that their account has run out of money or credit (benign interpretations exist here but are thought to be less common, e.g. a bank rejecting a transaction that looks suspicious just because you are on holiday at a new country, entering a security code incorrectly as a slip rather than an indication of an erratic emotional state, or where players have a separate bank account for gambling that they run down as a form of loss limit control).

This curve provides only weak support for this mainstream interpretation. Around 50% of players have never had a deposit declined (denoted by the brown vertical line falling on the y-axis) and it happens only rarely for a further 25% (up to the grey vertical line). The model interprets this level of deposit decline as low risk, holding other features constant. Once deposits are declined more than around one time in six (c. 18%), we see a modest increase in self-exclusion probability, c. 1.5%pts of probability, but not statistically significant by the proposed error bands. Further increases in deposit decline rate from this point do not lead the model to estimate further increases in self-exclusion risk; if anything, the average risk declines for the c. 5% of players (approaching and beyond the dark crimson vertical line) for whom declines occur most often, e.g. 35/40%+, although we should highlight there is no material variation by the minimum measure of significance suggested in this paper (variance beyond the 40th-60th %ile range across players at each feature value). Two aspects of this curve were counter-intuitive to our interviewees: first that high deposit decline rates do not relate to continuing increased risk; and second that the initial increase in risk is so modest.

With regard to the first counter-intuitive element, our domain experts considered that it may highlight an insight on the kinds of risk identified in this model by virtue of its design. Serious self-exclusion may be less effective at identifying harm among very committed players, some of whom may be addicted to their gambling habit and would not consider self-exclusion (hence why the model does not expect them to self-exclude). In its discussions with operators, BetBuddy typically emphasises the need to monitor such high-intensity players via more manual processes or via threshold triggers or to train models specifically for this player segment by collecting other proxies for player harm. With regard to the second counter-intuitive element, further interrogation suggests this may reflect the limited role of deposit decline ratio in the model. The deposit decline ratio ranks 12th in terms of the cross-entropy based feature importance metric in the model. It may or may not be important as a single feature if no other features were available (the curve alone cannot inform that), but other

features contain more information that differentiates serious self-exclusion. Effectively, once we control (approximately, via the method in this paper) for details like a player's average bet size, time spent gambling, amount deposited, frequency of play and so on, whether or not deposits are occasionally declined adds only a limited amount of additional insight to the model. This is a nuance not explicitly considered in much industry discussion of transaction declines. Nonetheless, given the strong priors of the sector, further research is recommended.

5 Conclusions, Limitations and Next Steps

Feature risk curves have proved insightful in discussions with sector experts to help understanding the directional implications of model-level drivers of particular features in a gambling harm ML model developed for a mass-market UK online bingo and slots operator. While acknowledging that it is only an initial discussion on example curves rather than a full analysis of the model, it was clear that the feature curves introduced here can add significant opportunity for model challenge and issue exploration with domain experts as compared to feature ranking methods which alone do not provide insights into directionality or the shape of the relationship.

The feature risk curves as included here allow for the extraction of informal knowledge in a systematic way, drawing on the curve values and the error bars. For instance:

- A night-play ratio of 94% and above sees a reduction of 3%pts on the average self-exclusion prediction probability, holding other player features constant, by comparison with the peak self-exclusion probability at a night-play ratio of 38%-68% significant against an error bar of 20% of data points around the median.
- A deposit-decline ratio of 18%-35% sees a 1 to 1.5%pts increase on the average self-exclusion probability, holding other player features constant, compared to an otherwise generally lower ratio and, to a lesser extent, a higher ratios which is nevertheless not significant against an error bar of 20% of data points around the median).

More generally, the analysis of the curves shown here challenged some simplistic mainstream interpretations in the problem-gambling community, adding important nuance to previous understanding. First, increasing night-play is only problematic up to a point and only weakly so, provided other play behaviours remain stable. Beyond that point, consistency of play at night appears to be a protective factor, at least with respect to serious self-exclusion risk. Second, more frequent declined deposits on their own (i.e. with no other changes in behaviour) do not necessarily indicate much greater probability of serious self-exclusion - the trend is neither monotonic nor significant with respect to either tight or generous error bars. The declined-deposit curve, in line with other curves like frequency of play, also point to the possibility that models based on serious self-exclusion are effective at identifying a subset of players at risk, but not others, with important implications for developing and implementing gambling risk analytics. A number of other curves were produced which are not shown here due to space constraints.

The technical approach as presently described has a number of possible improvements that we intend to explore in further work. For instance, (i) analysing how well individual or linear combinations of curves can capture the overall predictive insights from the model; (ii) analysing how much individual observation predictions change as the feature varies (as opposed to the average prediction across all artificial observations), providing more insight into flat segments of the curve; and (iii) extending beyond a univariate perspective to consider, for instance, how night play ratio and declined deposit ratio interact with each other (and with other features, like play frequency or volume) to inform the model's predictions and designing a grid search to identify potentially interesting combinations.

Within a gambling sector application, we also look forward to applying this technique to models trained on other data sets and combining it with other sources of insight to inform a broader, sector-wide discussion on the levels of risk typically implied by certain player behaviours and the guidelines and personalised player interventions that might be designed as a result.

References

[1] M Craven, JW Shavlik. Extracting tree-structured representations of trained networks, Advances in neural information processing systems, NIPS 1996.

[2] R. Caruana, P. Koch, Y. Lou, M. Sturm, J. Gehrke, N. Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. KDD'15, Aug. 2015, Sydney, Australia.

[3] S. Sarkar, T. Weyde, A. d'Avila Garcez, G. Slabaugh, D. Dragicevic, C. Percy. Accuracy and interpretability trade-offs in machine learning applied to safer gambling. CEUR Workshop Proceedings, 1773, NIPS Workshops, Dec 2016.

[4] N. Frosst, G. Hinton. Distilling a Neural Network Into a Soft Decision Tree. https://arxiv.org/abs/1711.09784, CEX Workshop at AI*IA, 2017, Bari, Italy.

[5] Building Trust in Machine Learning Models (using LIME in Python). https://www.analyticsvidhya.com/blog/2017/06/building-trust-in-machine-learning-models

[6] S. Wachter, B. Mittelstadt, C. Russell. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. https://arxiv.org/abs/1711.00399, 2017.

[7] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, F. Giannotti. Local Rule-Based Explanations of Black Box Decision Systems. May 2018. https://arxiv.org/abs/1805.10820.

[8] J. Woodward. Making Things Happen: A Theory of Causal Explanation. Oxford Studies in Philosophy of Science, 1st Edition, 2003.

[9] A. Hassanniakalager, P. Newall. A machine learning perspective on responsible gambling. Behavioural Public Policy, 1-24. doi:10.1017/bpp.2019.9, 2019.

[10] C. Percy, M. França, S. Dragicevic, A. d'Avila Garcez. Predicting online gambling self-exclusion: an analysis of the performance of supervised ML models, International Gambling Studies, 16:2, 2016.

[11] Responsible Gambling Algorithms Roundtable, 13 July 2016 at City University London. www.bet-buddy.com/media/1190/responsible-gambling-algorithms-roundtable-1-august-2016-final.pdf

[12] C. Percy, A. d'Avila Garcez1, S. Dragicevic, M. Franca, G. Slabaugh, T. Weyde. The Need for Knowledge Extraction: Understanding Harmful Gambling Behavior with Neural Networks. In Proc. ECAI: pp. 974–981, doi: 10.3233/978-1-61499-672-9-974, 2016.

[13] A. White, A. d'Avila Garcez. Towards Providing Causal Explanations for the Predictions of a Deep Network. In Proc. Human-like Computing Machine Intelligence Workshop MI21-HLC, June 2019.

[14] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi and F. Giannotti. A Survey Of Methods For Explaining Black Box Models, http://arxiv.org/abs/1802.01933, Aug 2018.

[15] PWC. Remote Gambling Research: Interim report on Phase II. Report for gambleware. Aug 2017.

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