# Report on ways to predict irresponsible gambling behaviors

Over the last twenty years, the online gambling business has expanded globally and technologically and has allowed customers to play 24 hours a day, seven days a week from home, work, or public spaces. This explosive growth created a plethora of new data to explore wagers, withdrawals, deposits, open rates and even mouse clicks. All this data can be applied in data mining creating a world of possibilities to study.

One of the subjects that concern more online gambling companies is to guarantee a responsible gaming. Among other areas, this concept is competent of protecting vulnerable customers and delivering a fair game experience in overall.

There are certain indicators that would be able to reveal change in the gambling behaviors. This set of indicators include variability of bet size, intensity and frequency of betting and trend of wagered money, fluctuations between intervals of increasing wager size followed by rapid drops and the total number of different games played per day by the same gambler.

There were many studies that propose a model that is able to detect high-risk online gamblers by using different supervised learning algorithms. Algorithms such as Neural Networks, Random forest, SVM that might perform better when classifying likely problem gamblers in training data. We could also use unsupervised learning techniques such as clustering, anomaly, and outlier detection.

In order to figure out the anomalies associated I would choose to apply a Principal Component Analysis (PCA) base model to the dataset. Having considered the dimension of the data and the number of variables that influence the data, using this approach we can transform the data into new dimensions by still accounting for its variability.

We can divide the new data into two different components namely normal and anomalous. In order to detect different behaviors in the data, an anomaly threshold could be defined and to evaluate the method, time-series analysis and user-based analysis can be considered.

In time series analysis we aim to capture the activity of all the users within a specific amount of time, and in the user-based analysis, we can test each user for any suspicious behavior. Both these methods can be scanned for outliers. We can determine something is wrong with user if he registers an anomaly score higher or equal to say 90%.

We could also use Outlier detection joined with visualization techniques as they can be extremely suitable for detecting anomalies in the data.