Irresponsible gambling detection

Kiana Zeighami

Introduction

This essay discusses a method that detects irresponsible gambling behaviours. In particular, I focused on detecting the loss-chasing pattern, but a similar approach can be applied to the other patterns mentioned in the instructions. This problem can be defined as a classification problem, in such a way that a series of gambling behaviours will be classified as responsible or irresponsible.

In this essay, I will discuss a classification machine learning method to detect loss-chasing betting patterns amongst gamblers. For this purpose, I will first describe the required historical data and input features, then propose a method for preparing a training set, after that I will describe different machine learning techniques that can be applied to solve the problem and lastly, I will discuss evaluation metrics to estimate the accuracy of the machine learning models.

Historical data and feature selection

To analyse the irresponsible betting behaviour, first the relevant data should be collected. To achieve this, we can collect data for the following features for several years:

- Time series data for each individual for the following:
    - The bet amount in each day
    - The outcome of the bet in each day (win or lose)
- The betting frequency of each individual per month
- The average bet amount of each individual in the last year
- Personal information of each individual, such as age, gender, income and location

We consider a fixed window of 30 days to predict whether the loss-chasing pattern can be detected. Specifically, the input features of the model are:

input_features = [Age, Gender, Income, betting_frequency_month, average_bet_lastyear, bet_day1, bet_day2, …., bet_day30, bet_outcome_day1, …, bet_outcome_day30]

Note that some of the above-mentioned features such as age and gender can result in a model that discriminates based on age and gender, we should verify that it does not happen during the experiments.

Preparing the data set

After the data collection, the next step is to prepare a training data set which is used as a ground truth of our machine learning model. For this purpose, we will split our raw data into a training set, testing set and cross validation set. However, the main challenge is that initially our data is not labelled, which means that given each individual's betting pattern we do not have the information that whether it should be classified as responsible or irresponsible.

One method to tackle this issue is to consider the unsupervised machine learning methods, such as clustering. Ideally our model will split the data into 2 clusters, responsible and irresponsible. However, given the nature of our data this method may not be efficient. Because clustering algorithms group the data point by considering some distance metric (e.g., their Euclidean distance), and in this case

the data corresponding to responsible and irresponsible betting patterns, may not be clearly distinguishable. However, this method can be tested.

For the purpose of this essay, I will concentrate on the supervised classification methods and propose a method to label the training data set. This method consists of 2 steps:

- Designing a simple algorithm that flags the betting pattern of each individual that might exhibit the loss-chasing betting pattern
- Manually checking the flagged data and adjust the labels

The simple flagging algorithm will detect a potential loss-chasing behaviour. To achieve this, the algorithm accepts two input parameters: A time series that represents the bet amount over a month in each day and a time series that represents the loss amount over a month in each day. The algorithm computes the running average of the bet amount considering each 5 day interval, and the total loss amount in each 5 day interval. If the running average and the total loss amount are both increasing, it will flag the data. An alternative method can be estimating the correlation between the time series of bet amount and the time series of the loss amount. Using the Pearson correlation coefficient, we can flag the data corresponding to the correlation larger than a certain threshold.

To ensure that the data is labelled correctly, the flagged data will be manually analysed and adjusted. Note that if our flagging algorithm has a high accuracy we can use that as our main method. Also, after several months of experimentation, we can potentially obtain labelled data for the future iterations.

Classification method

After preparing our data set, the next step is choosing a suitable model. The choice of model is data dependent. If we have enough data, neural networks can perform well. Also, the other factor to consider is the amount of explainability required. If we are interested in explainable outputs, decision trees or logistic regression can be a better choice.

We can experiment with different machine learning techniques such as Logistic regression, Neural networks and Decision trees, and choose the one with the highest accuracy, using the evaluation metrics described in the next section.

After choosing the classification method we will train the model with the prepared training set. To achieve this, we will separate the target and input parameters. In this case, our target parameter is the responsible or irresponsible flag and the input parameters are all the other features.

Cross validation

We will use the cross-validation data set to determine the accuracy of the model. Since our data is likely to be imbalanced, meaning that the majority of betting patterns are responsible, we can use precision and recall metrics and F-score to evaluate the accuracy of our model. Precision metric calculates, of all the behaviour our model considers irresponsible, how many are actually irresponsible. This metric can be used to decrease the false positives. On the other hand, Recall metric calculates, of all the actual irresponsible behaviours, how many are classified correctly by our model. This metric helps us decrease the false negatives. Lastly, F-score represents both precision and recall as a single metric.

Another issue that arises due to having an imbalanced data set is that our model may not be able to learn the under-represented class accurately. To solve this issue, we can increase the weights of irresponsible gambling behaviours in our loss function.