TALLINN UNIVERSITY OF TECHNOLOGY
Faculty of Information Technolog
Department of Computer Science

Nikita Tsykunov
182465IVSM

Early Detection of Online Gambling Addiction

Master's Thesis

Supervisor: Sven Nõmm, PhD

Tallinn 2020

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia Teaduskond
Tarkvarateaduse instituut

Nikita Tsykunov
182465IVSM

Internetis hasartmängusõltuvuse varajane avastamine

Magistritöö

Juhendaja: Sven Nõmm, PhD

Tallinn 2020

Declaration

I hereby certify that I am the sole author of this thesis and this thesis has not been presented for examination or submitted for defence anywhere else. All used materials, references to the literature and work of others have been cited.

Author: Tsykunov Nikita

May 18, 2020

Early Detection of Online Gambling Addiction

Abstract:

The growing popularity of online casinos raises concerns about gamblers' mental safety. To prevent customers to become addicted and to follow the law gambling companies are taking steps to help customers with early signs of gambling addiction. Unfortunately, all taken measures, such as quizzes, loss limits and emergency buttons are based on customer's willing to stop gamble, so that's why they are not helpful in detection or stopping of ad-dicted players. This can cause a problem for both player and gambling company. The most recent work in these fields is dated by 2013 and using the same dataset with a limited number of features and provided by the same company as in all papers before it. In opposite, in this paper was created a new dataset from actual data of gambling company. As a result, the model was made that managed to determine players with problem behaviour with accuracy level equal 80% that is higher than in previous works.

In this work approach based on using machine learning is used to determine if a player is addicted or not. During this work, features were found which characterize problem gamblers such as the number of deposits or sports bets variance. Random Forest algorithm was used for model training, which shows best results among other classification methods. Fisher Score algorithm was applied for future selection together with other methods to improve the accuracy of predictions.

The thesis is in English and contains 37 pages of text, 7 chapters, 13 figures, 3 tables.

Keywords:

online gambling, machine learning, random forest, addiction recognition

CERCS: T120

Internetis hasartmängusõltuvuse varajane avastamine

Annotatsioon:
Online kasiinode kasvav populaarsus muudab mängurite olukorra aina ohtlikumaks. Klientide kaitseks ja seaduse järgimiseks üritavad hasartmänguettevõtted varakult sõltlast tuvastada. Kahjuks kasutusele võetud tuvastusmeetmed, nagu küsitlused, kaotus limiidid ja paanika nuppud, põhinevad kliendi tahtel mängimist lõpetada, mis teeb neid ebaefektiivseks sõltlaste peatamiseks. See võib tekitada probleeme nii kliendile, kui ka ettevõttele. Viimased uurimised antud alal olid tehtud aastal 2013 ning kasutasid sama piiratud andmestiku, mis kõik eelnevad uuringd. Vastupidiselt on selles uuringus loodud uus andmestik, mis põhineb reaalse kasiino andmetel. Tulemusena oli loodud mudel mis suudab tuvastada sõltlase 80%-se täpsusega , mis on kõrgem kui kõikides eelnevates uuringutes.

Sellise tulemuse saavutamiseks oli kasutatud masinõpe, et otsustada kas tegemist on sõltlasega või mitte. Olid avastatud tunnused, mis iseloomustavad sõltuvuse probleemi, nagu tihedad sissemaksed või pannuste varieerumine. Masinõppe mudeli treenimiseks oli kasutatud Random Forest algoritm, mis näitab paremaid tulemusi võrreldes teiste klassifikatsiooni viisidega. Fisher Score algoritm koos teiste meetmetega oli rakendatud täpsuse täiustamiseks.

Lõputöö on tehtud Inglise keeles ja sisaldab 37 lehekülge, 7 peatükki, 13 pilti ja 3 tabelit.

Võtmesõnad:
online-hasartmängud, masinõpe, juhuslik mets, sõltuvustuvastus

CERCS: T120

Contents

List of Figures

List of Tables

1 Introduction

According to the most recent researches conducted in Estonia [17] number of people with problem gambling and pathological gambling is 3.1% and 3.4% respectively. These numbers are increased since previous similar research [1], where numbers were 2.6% and 2.4% respectively. Speaking about Europe in the general combined rate of problem gambling prevalence is from 0.5% to 5% depending on the year, country and measurement method. Although it's hard to say that online gambling causes this growth [13], but it may become a significant cause of the problem in case of not paying enough attention.

Online share of the total gambling market in Europe grew on 13% from 2013 to 2015 and gross gaming revenue (GGR) of online gambling in Europe grew on 20 billion euro from 2003 to 2017. That's show stable growth of the online entertainment field.

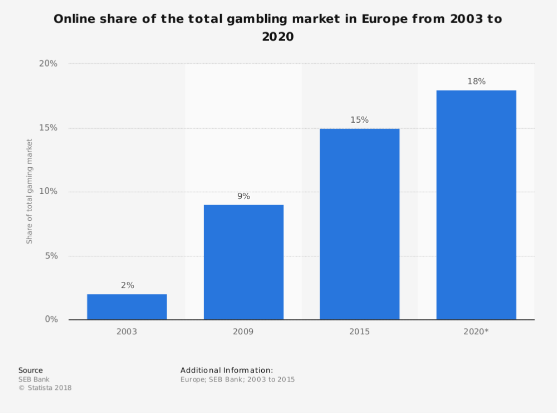


Figure 1. Online share of the total gambling market in Europe from 2003 to 2020. Source: statista.com

Problem gambling already has attention as countries by separate as an international organization. Such as problem gambling was classified by WHO as an addictive disorder(ICD-11) [29] according to DSM-5 in 2019. Sweden made conditions of getting the license for providing gambling service much stricter: the provider must always show the current time and several buttons to give to player ability to pause game session immediately in one click.

The main difference between online and on-ground casinos is that in online one player is not observable by anyone, so the only way that can stop playing is self-exclusion. Online casinos in Europe regulated by laws different from country to country, but all of them are following the European Union Commission Recommendation 2014/478/EU [9]. According to these recommendations, countries have open registries that people can use to ban themselves. Online casinos must check these registries and forbid the self-excluded customer to play. There was a study conducted to find the effectiveness of self-exclusion [8]. It shows that self-exclusion has an impact on gambling cravings in the medium term, but there must be also other conditions to help them to reduce the harm of gambling.

There are several studies which were conducted in the field of detecting the behavioural pattern of gamblers, but there are very few of them and they were conducted at the start of growth of the market. This makes them outdated because online casinos changed since then. Also dataset they working with is limited and predefined by the casino itself, that able to corrupt the result.

1.1 Background

Gambling addiction, or pathological gambling, was included in the list of mental disorders by the American Psychiatric Association in 2013 [22] and took place among drug alcoholism, drug dependency and tobacco addiction because it shares the same effects and brain reward system like a drug. Later, many other countries also started to recognize problem gambling as a public health issue, that needs to be regulated by authorities.

For classification purposes, the most used classification system is the Problem Gambling Severity Index(PGSI) [11]. PGSI is simple nine-item questionary designed for use with the general population rather than it's the nearest competitor - South Oaks Gambling Screen (SOGS), that is intended to be used in the clinical context.

1.2 Motivation

Online gambling has shown stable growth during the last couple of decades and expected this market will expand in the observable feature. With the growing number of gamblers around the world amount of gamblers with problems will be also increasing together with it. Although, European union, governments of many countries, a lot of international associations and organisations working in the field of problem gambling: all of them are aware of the growing problem and raises concerns about it, measures taken to prevent or mitigate this growth are not sufficient enough [25]. Most of them are targeted to customers who are already aware of the problem he has and wants to stop gambling, like self-exclusion, quizzes or loss limits, or very questionable in terms of effectiveness, e.g. not hideable on-site clock or session reminders, that give a quick overview of gambling history every hour or less.

From the other hand, machine learning already helps people int the very broad number of topic, starting with facial recognition to unlock the phone and ending up modelling Parkinson's disease [16]. Despite the number of studies about using machine learning for classifying customers, there are no pieces of evidence about practical usage of such approach for classification customers in gambling companies. It's clear, that such a strategy means a lot of unnecessary development for a company together with the lack of a clear framework for integration of machine learning logic into the business processes of a company. This reason together with a huge amount of mandatory development required by the gambling license is preventing online gambling companies from developing in this side of playing in online casinos and betting companies.

To make this process more clear and to show the relevance of such development this thesis is focused on building a classifier for early detection of online gambling addiction. This work compares different approaches to building classifier in order to pick the most accurate method and to make ensure that company customer base won't experience a huge number of false-positive cases of classification.

1.3 Company in the case study

The company provided access to its data is StayCool OÜ, which owns web-site with the online casino and sports betting named Coolbet since 16th May 2016. Site has a huge client base about hundred thousands of people from several countries playing for several years.

Coolbet has several products on its platform including sports betting, casino and poker. All activities are carried out in accordance to license requirements and beside this company implemented additional checks to help customers play healthy.

Company is very concerned in the field of responsible gaming and interested in ways of improving client safety. Settings for loss limits with different durability, which can be easily changed only to more strict ones, are available for every customer. There are also limits with the same set of rules for depositing and wagering. Each client can enter himself into the self-exclusion registry with a couple of mouse clicks and cannot easily return if he changes his mind soon later. As a less strict measure, a client of the platform can exclude himself only from this site for 24 hours. All these measures together with round-the-clock multilingual support are targeted to prevent a customer from becoming addicted.

The company provided access only to gambling history of customers and according to GDPR, all data was anonymous. Each customer is associated with a unique identification sequence and so there are no clues pointed to the identity of the person behind gambling history data.

During the creation of this study author of it had ongoing work relationships with the company, which are didn't affect the workflow and results of this work in any case.

1.4 Related work

There are a lot of studies conducted in the field of problem gambling, but most of them are considering the problem only from a psychological perspective. One of the first paper that addressed the issue with the help of machine learning was the study of Braverman and Shaffer [5] in 2010. In this study, k-mean cluster analysis was applied to identify customers with similar first-month gambling behaviours. Dataset was compiled from data provided by Internet betting service provider *bwin.party* collected during February 2005 and consisted of 530 players who were actively involved in online betting. For identifying was used 4 main characteristics: frequency, intensity, variability and slope. As a result, four prime clusters were identified.

Next noticeable study was made by Dragicevic, Tsogas and Kudic in 2011. It is very similar to the previous one conducted by Braverman and Shaffer and it's explicitly declared in the work. But work is based on data received from Internet gambling software provider for lotteries and commercial Internet gambling operators *GTECH G2* and consisted of casino gambling history while Braverman used online betting data. research finished with similar results.

Gray, LaPlante and Shaffer conducted another study in 2012 targeted to identify behavioural characteristics of gamblers who triggered responsible gaming (RG) system [14]. Data for this study was also provided by *bwin.party*, but in this case, data was taken for the period between November 2008 and November 2009 and had information about all activities including casino and poker. Besides, data included information about RG events. For data was analyzed to find the correlation between features, including number and variance of stakes in different products, and the number of RG events. The study showed a significant correlation between nine variables and discriminant function. These variables are Active Betting Days; Duration; Bets per Betting Day; TotalBets; Euros per Bet; Total Stakes; Net Loss, for live-action sports betting; and Active Betting Days and Total Stakes, for fixed-odds sports betting. Later, in 2013, Braverman, LaPlante, Nelson and Shaffer [6] extended this study by creating models capable to predict the development of gambling-related problems among subscribers of *bwin.party* using the same dataset as in the previous paper. After classification results were compared to the classification made by the RG system of the company.

In 2013 Philander [23] made a comparison of data mining procedures and get accuracy for Random Forest algorithm equals 0.658. In this work, the target is to get higher accuracy and precision using a dataset with the most recent data.

This study is inspired by Braverman (2013), but for classification purposes, the Random Forest algorithm is used, suggested by Philander (2013). Another difference between them is a set of features extracted from most recent data and not provided by the company but gathered during this work independently, that excludes the possibility of manipulating with data before processing. Also, customers presented in the dataset have a different country of living in opposite to previous studies.

2 Problem statement

The goal of this study is to create a framework for classifying customers of the online betting platform as problem gamblers using their gambling history. To achieve this goal it needs to build a machine learning model using company data for training and testing purposes. After building a classifier, provided decisions must be explained to check correctness of classification and get clear understanding of affection of different features on classification process. All data will be retrieved from the company data warehouse including all needed features. All work naturally splits up to following steps:

1. Feature selection. Because a lot of features will be tested during model creation, six the most valuable features must be selected from the whole dataset for performance and accuracy reasons.

2. Model training and comparing results. With prepared data, several different models can be trained to compare results and pick the model with the most precise predictions. For this study was selected such models as random forest, logistic regression, SVM and a couple more.

3. Results explanation. After success training portrait of the typically responsible gambler and problem, the gambler can be visualised with the help of LIME or SHAP. This can help us to focus on the most significant signals of an addicted person.

As the outcome of this work, it's expected to have a working machine learning model that both can classify customer with high accuracy and also give an explanation of what is the most important feature characterizing problem gambling behaviour. Formally it needs to create a machine learning model using gambling history data. For this it needs to extract the most valuable data, choose algorithm giving the most accurate results for this type of data and apply it to extracted previously data. This will give the average portrait of a player having problem gambling behaviour.

3 Tools and methods

3.1 Tools

The initial feature set is based on previous studies but later was enriched by several more. During this study, several hypotheses are compared.

Python 3.7 is used in this study, because of its wide usage and modernity. Python de-facto is standard for scientific programming nowadays. Data is extracted from the organization's database using SQL (Postgres dialect) via database management tool called Adminer, which allows retrieve and observe information in the database with minimum efforts. For data processing was used such library as Pandas (version 1.0.3) which give the convenient way of storing and manipulating data, Scikit-learn (version 0.22.1) is used for machine learning and for displaying data Matplotlib (version 3.1.2) and Yellowbrick (version 1.1) was used. Yellowbrick is a library built on top of matplotlib specially designed to create reports from Scikit-learn models. Also, LIME (version 0.2) and SHAP (version 0.35) were used for explaining results.

3.2 Cross-validation

Cross-validation [27] is the technique for validating if the model able to predict the value for a testing set of data. Usually, for model training, an initial dataset is split into a training set which is used to train model and into testing set against which model is tested. Cross-validation can show if the model is overfitted and since can predict values only for the training set. Cross-validation consists of several iterations during which initial dataset is partitioned into complementary subsets and each subset is used in training separate model. Process of partitioning is shown in Figure 2.

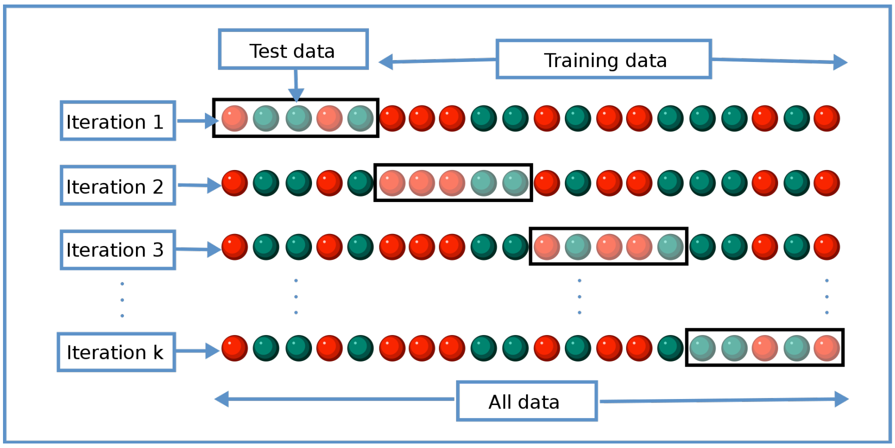


Figure 2. Cross-validation partitioning. Source: wikipedia.org

In this work was used k-fold cross-validation. This type of cross-validation the original sample is randomly partitioned into k equal parts. Out of these parts, one part is used for testing and other $k-1$ parts are used as a training set. Resulting accuracy is usually mean of k iterations.

3.3 Random Forest and Decision Tree

Random forest algorithm [7] primarily used in this study is an algorithm based on decision trees. A decision tree is a pretty intuitive thing to understand. It's a binary tree each node of that consists of some distinct question and branches of this node are answers to it. Let's take a look at Figure 3. At this figure, an example of a decision tree can be seen, which is made using data from this study. The first line in the node set criteria by that dataset is split up further. To make it visible this example have the depth of tree equals 2. The full tree has a depth equal 19.

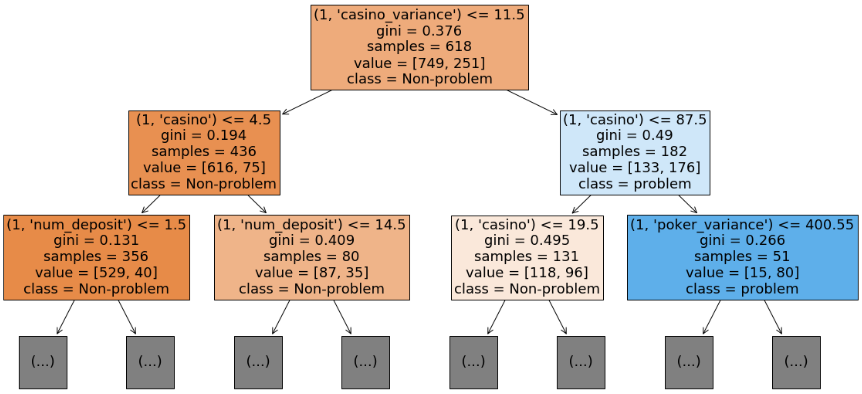


Figure 3. Decision tree

For huge dataset can be built thousands of decision trees with different level of confidence, but all of them are not fully correlated with real life. Random forest consist of a large number of individual decision trees and the most frequent outcome from all trees for some set of data becomes prediction. If features correlate with classification outcome, then most of the decision trees will give the correct answer and thus Random forest will give correct classification results. Example of how it works can be seen in Figure 4: out of six predictions five 1s and one 0s, therefore, predicted value is 1.

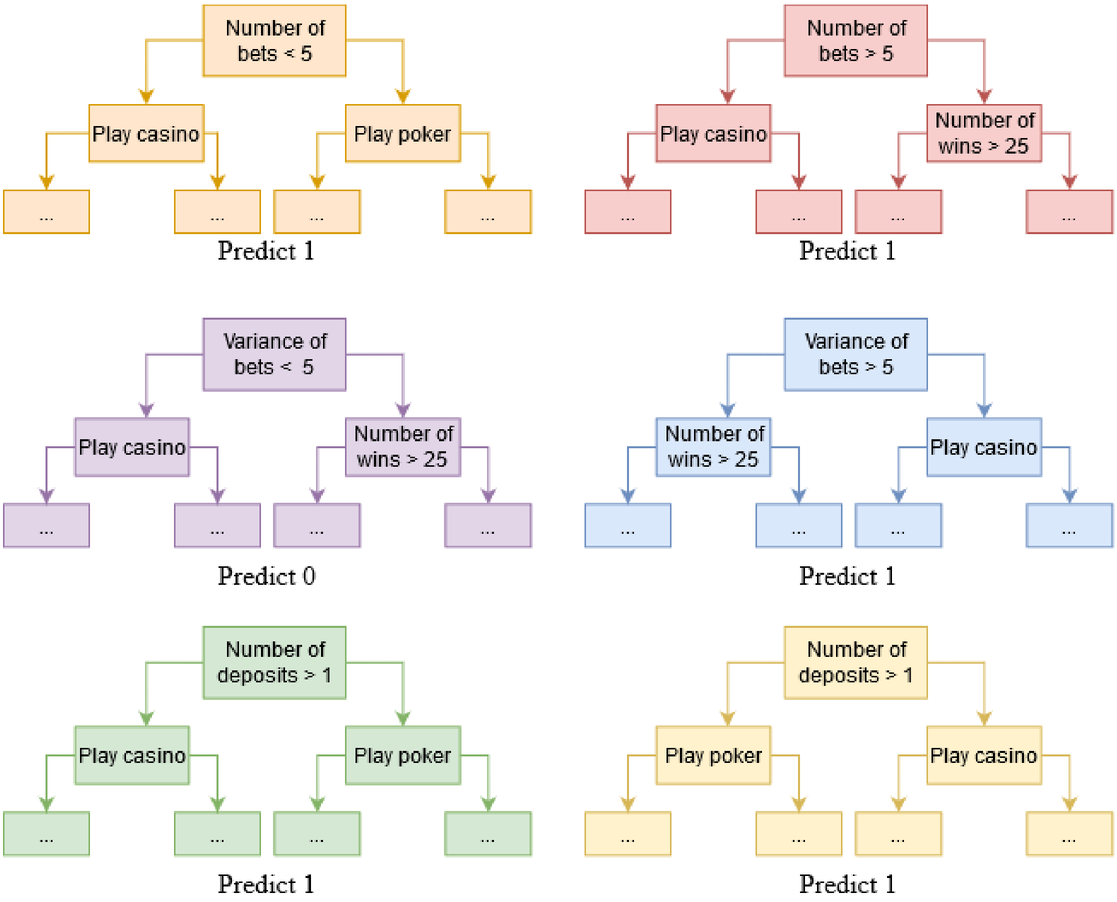


Figure 4. Random forest algorithm

3.4 AdaBoost

AdaBoost is a boosting classification algorithm created by Freund and Schapire in 1995 [12]. The core principle of AdaBoost (Adaptive Boosting) is to combine the results of many weak learners (i.e decision trees) trained on constantly changed set of data. Combination of results is happening through a weighted majority vote. This method is similar to the Random forest algorithm in terms of combining results of predictions from many weak learners.

The fundamental difference from other algorithms is the so-called boosting iteration that consists of applying weights $w_1, w_2, ..., w_N$ to each of the training samples. At the first iteration, those weights are set to $w_i = 1/N$ and later for each successful iteration weights are increasing for false predictions and decreasing for true prediction. Also, each learner gets a level of confidence so that in the next iteration learner with the highest level of confidence will get a sample with the highest weight. In other words, the best classifier will get the worst sample. Iterations continue until all samples in the training set will be classified correctly or the number of iterations will be exceeded.

In this work was used extended AdaBoost algorithm for multi-class case AdaBoost-SAMME.R created in 2009 [15].

3.5 KNN

KNN (k-nearest neighbours) is a classification algorithm of non-generalizing learning type [4]. It means, that it doesn't create a learning model, but simply stores the samples from training data. The classification process is a simple majority vote from *k* nearest neighbours of each point. Nearest neighbours for each point can have uniformly distributed weights or have weights based on the distance to the point that needs to be classified.

3.6 Logistic regression

Logistic regression uses a logistic function to give the probability of binary output [28]. It's an algorithm that relies on the output function received as a result of the logistic analysis. Definition of general logistic function has the form shown in equation (1).

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{1}$$

where $p(x)$ is the probability that the dependent variable equals a case, β_0 is the value of the criterion when the predictor is equal to zero and $\beta_1 x$ is the regression coefficient multiplied by some value of the predictor. e means the exponential function.

3.7 SVC

SVC (Support Vector Classifier) is a classifier that uses support vector machine (SVM) in classification purposes. The main goal of SVM is to find hyperplane with maximum margins in N-dimensional space (where N is the number of features). SVM uses the output of the linear function to determine whether it belongs to a class. For each sample, SVM decides to which side of hyperplane it goes. Figure 5 consists of a representation of two classes divided by a hyperplane.

3.8 LIME

LIME is based on the work presented in this paper [26]. LIME stands for Local Interpretable Model-Agnostic Explanations, which means that it gives understandable explanations for any black-box model. LIME split model to several pieces, find likelihood

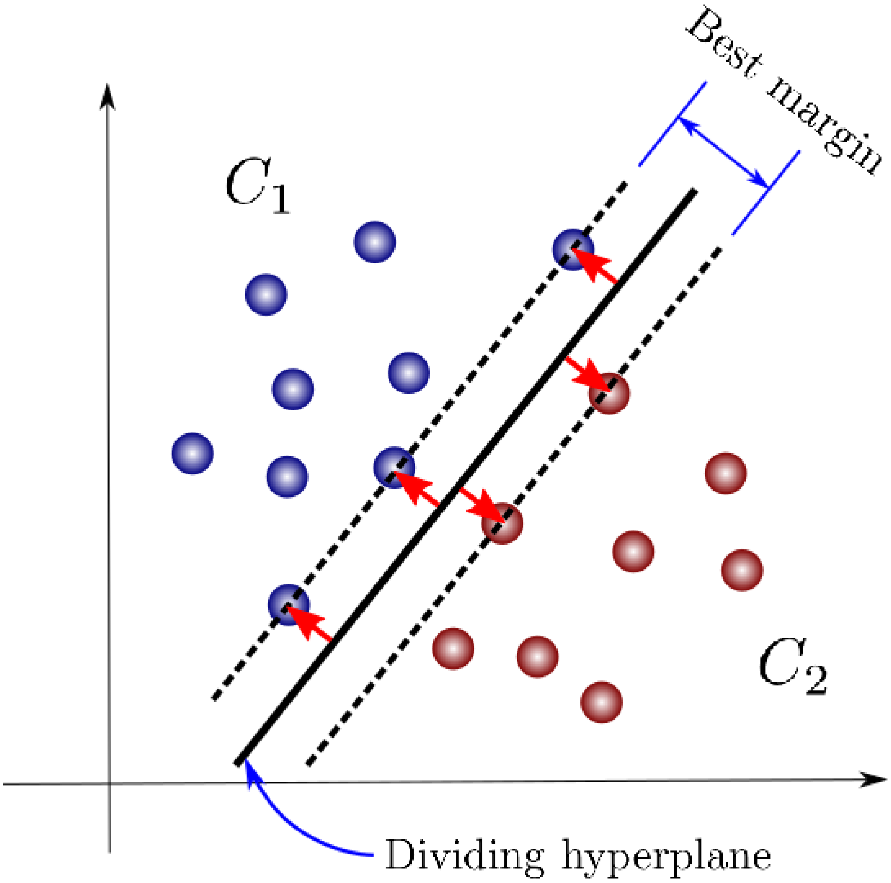


Figure 5. Two classes divided by optimal hyperplane. Dashed lines are called support vectors. Source: towardsdatascience.com

for each piece separately and then gives the weight of each piece. It helps to understand why the model gave this or that prediction. As an example, the explanation of Google's pre-trained Inception neural network can be seen in Figure 6. In which figure Google's model classified original picture of the frog as "tree frog" as most likely case, also two other options "pool table" and "balloon" are among answer given by it. LIME explains why it's happened. In practice, LIME helps to understand if a model can be trusted.

3.9 SHAP

SHAP (SHapley Additive exPlanation) is another explainer giving an understanding of how features affect on final classification. Shap is similar to LIME in terms of finding local approximations, but under the hood, it uses the classic Shapley values from game theory and their related extensions [18] [19]. This library unifies some other methods

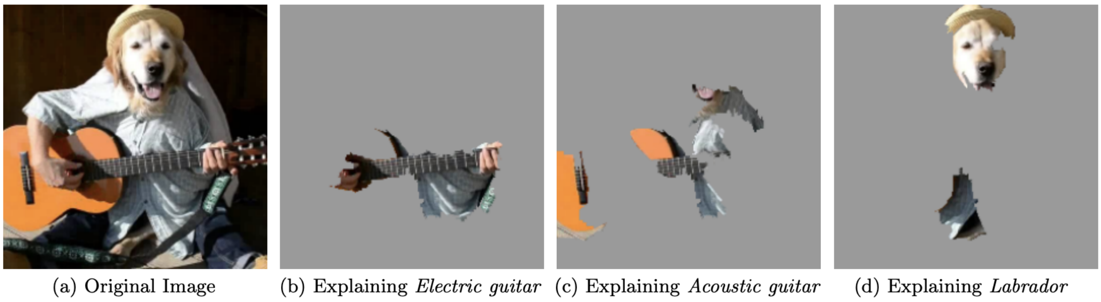

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 6. Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" (p = 0.32), "Acoustic guitar" (p = 0.24) and "Labrador" (p = 0.21). Source: [26]

to produce Shapley values such as LIME, DeepLift, Layer-Wise Relevance Propagation and Classic Shapley Value Estimation and is targeted to improve the accuracy of explanation made by these methods.

4 Implementation

4.1 Feature selection

A dataset with 41 feature was obtained after data acquisition and its transformation to a more convenient format. These features describe gambling history of 9397 customers and consist of the following data:

1. Number of bets in the casino, poker and sports.
2. Variance of bets in the casino, poker and sports.
3. Number of active days.
4. Number of activities customer participated.
5. Country of living.
6. Number of deposits.
7. Number of withdrawals.
8. Number and variance of bets together with a number of deposits and withdrawals throughout all gambling history split by quartiles by date.
9. Label denoting if a gambler is counted as addicted or not.

There was no opportunity to get medical data about the customer to check if someone actually has behavioural disorder connected with gambling, therefore for ground trust data was accepted combination of indirect signs such as registration in one of the self-exclusion registry, a closing account with reason connected with problem gambling or contacting customer support via email or phone about problem connecting with gambling.

This amount of features is redundant and hard to process in whole. In order to reduce the number of features without losing any important data feature selection based on Fisher‘s score [3] was made according to (2). Fisher‘s score describes the discrimination power of the features in one class and unlikeliness of features in separate classes - the higher score indicates better suitability for classification usage.

$$F = \frac{\sum\limits_{j=1}^{C} p_j(\mu - \mu_j)^2}{\sum\limits_{j=1}^{C} p_j s_j^2} \tag{2}$$

where C is the number of different classes (two in case of the present study) , p_j - proportion of the elements in class j, μ - mean value for the entire sample, μ_{j} - mean value of the class j and s_j is the standard deviation of the class j. Higher values of the Fisher's score indicate higher discriminating power of the feature.

Result of the algorithm for most significant features is presented in Table 1 in descending order.

Table 1. Top 10 features and corresponding Fisher scores.

Feature name	**Fisher‘s score**
Country	0.0448
Poker bets variance	0.0275
Sport bets variance	0.0243
Number of sport bets	0.0162
Number of deposits 1 qt.	0.0138
Number of deposits 4 qt.	0.0132
Number of deposits 2 qt.	0.0122
Number of deposits 3 qt.	0.0116
Number of withdrawals 2 qt.	0.0093
Sport bets variance 4 qt.	0.0092

At Figure 7 distribution for three features with best Fisher scores can be seen. Every number on country axis indicates different country. Every green dote represents one customer who classified as a problem gambler. As at can be seen all dataset separated between countries and variances in poker and sports bets show good correlation. There is an obvious necessity in separation dataset by country and excluding countries with a too low number of entries or representing only one class. A class should be understood as a subset representing one of the variants of classification.

At figure 8 distribution for three features with best Fisher scores excluding country can be seen. All three features show good correlation inside the feature group and with classification labels. This means that these features are suitable for final classification.

Although features show low Fisher‘s score by separate, combined they give score equals 0.044. It shows that a single feature has low affection on the result but as a group of features it can be used to classify customers. To calculate Fisher‘s score for a group of features was found Euclidean distance between rows of the dataset and then equation (2) was used to calculate resulting Fisher‘s score for group of features.

4.2 Model training

According to research conducted in 2013 by Philander in the field of identifying high-risk online gamblers [23], Random Forest shows the best accuracy. However, this re-

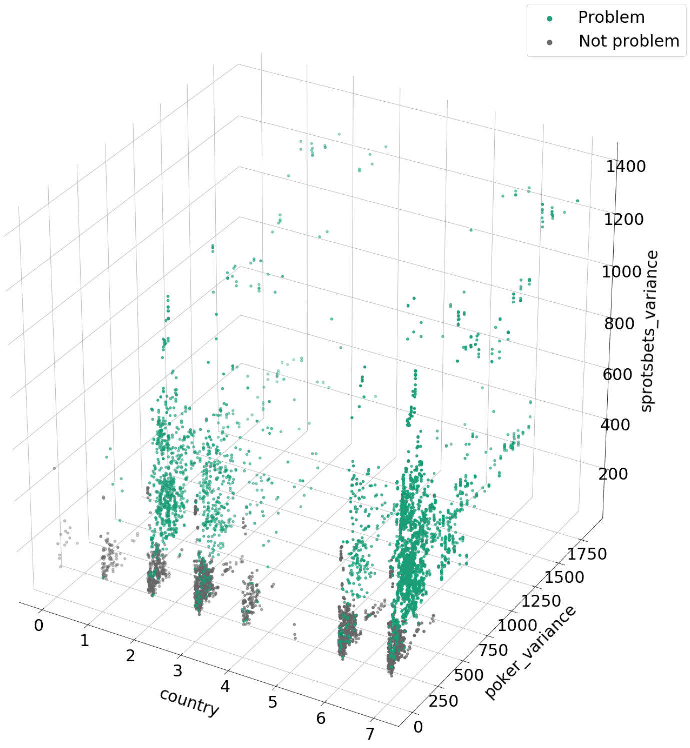


Figure 7. Scatter plot representing the distribution of countries and poker and sports bets variance for the problem and non-problem gamblers. The plot shows the distribution of customers among countries. For each country, non-problems gamblers give the lower variance of bets in poker and sports.

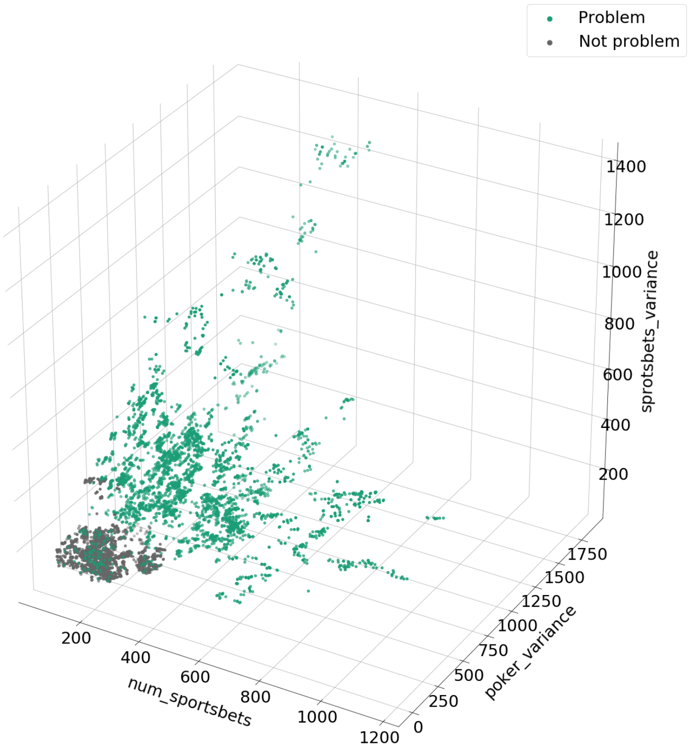


Figure 8. Scatter plot representing the distribution of the number of sports bets and poker and sports bets variance for the problem and non-problem gamblers. Almost all non-problem gamblers concentrated in the corner with low variance pf bets in poker and sports together with a low number of bets in sport.

search has lack classifying algorithms such as Decision Tree and Logistic Regression. So it was decided to compare the accuracy of several classification algorithms to pick one with the highest accuracy. The comparison was made using 5-fold cross-validation with training data and the results of it shown in Table 2. For each folding was received accuracy, precision, recall, ROC AUC and F_1 score and calculated mean for every characteristic for every classifier.

These 5 metrics allow us to examine the quality of the resulting model. Accuracy shows us the proportion of correct predictions) among the total number of cases examined [20]. Precision is the metric reflecting proportion of correct positive prediction among of all prediction. Precision should always be used together with recall, that is the proportion of correct positives that were classified correctly. A system with high recall but low precision returns many results, but most of its predicted labels are incorrect. A system with high precision but the low recall is just the opposite, returning very few results, but most of its predicted labels are correct. ROC is a probability curve and AUC represents the capability of the model to distinguish between classes [10]. And finally, F_1 score is the harmonic mean of precision and recall, metric, that combine them both in one measure [24].

Settings for every classifier are default given by Scikit-learn package listed below:

1. Random Forest. The number of decision trees: 100, quality of split criteria: 'gini'.

2. Ada boost. The number of estimators: 50, boosting algorithm: SAMME.R.

3. Decision tree. Quality of split criteria: 'gini'.

4. KNN. The number of neighbours: 5, the algorithm used to compute the nearest neighbours: 'auto' (the most appropriate algorithm will be chosen from BallTree, KDTree or brute-force based on values passed to the fitting method), distance metric: 'minkowski'.

5. Logistic Regression. The norm used in the penalization: 'l2', the tolerance for stopping criteria: 1e-4, the algorithm to use in the optimization problem: 'lbfgs', the maximum number of iteration: 100.

6. SVC. Kernel type: 'rbf', tolerance for stopping criteria: 1e-4.

According to Table 2, the Random Forest algorithm shows the best results among others. Together with many numbers of works about the Random Forest algorithm, it was chosen for later usage.

After choosing the algorithm separate classifier for each country was trained and classification report was created with the help of the Yellowbrick library. For training country datasets were split to train and test dataframes, where the proportion of the test set is 0.3. Results of training of models are in Figure 9. All country having only one

Table 2. Comparrison of classification algorithms.

Algorithm name	**Accuracy**	**Precision**	**Recall**	**ROC AUC**	**F_1 score**
Random Forest	0.79 +/-0.00	0.77 +/-0.00	0.79 +/-0.01	0.85 +/-0.00	0.78 +/-0.01
Ada boost	0.78 +/-0.00	0.75 +/-0.01	0.80 +/-0.01	0.84 +/-0.01	0.77 +/-0.00
Decision tree	0.71 +/-0.01	0.69 +/-0.01	0.67 +/-0.02	0.70 +/-0.01	0.68 +/-0.01
KNN	0.60 +/-0.00	0.57 +/-0.00	0.55 +/-0.02	0.63 +/-0.01	0.56 +/-0.01
Logistic regression	0.62 +/-0.02	0.67 +/-0.03	0.38 +/-0.02	0.66 +/-0.03	0.48 +/-0.02
SVC	0.60 +/-0.01	0.65 +/-0.02	0.29 +/-0.02	0.66 +/-0.01	0.40 +/-0.02

class or having less than 5 entities were filtered out due to impossibility of building correct classifier.

At Figure 9 every subplot represents different country. Title of subplot consists of country code written in by ISO 3166-1 alpha-2. Subplot itself consists of heatmap each row of that represents the problem and non-problem classes of gamblers and columns represent precision, recall and F_1 score respectively.

The support level for each country indicates that each subset is imbalanced and so classification results are inaccurate. In order to mitigate dataset imbalance, each classifier was retrained with a weight set for each class. For simplification weight calculation was delegated to Scikit-learn package by adding setting class_weight='balanced' to Random Forest classifier. This setting calculates class weight according to (3).

$$w_j = \frac{n}{kn_j} \tag{3}$$

where w_j is the weight to class j, n is the number of observations, n_j is the number of observations in class j and k is the total number of classes.

Process of model training was repeated already with the new setting which helps to mitigate consequences of dataset imbalances. Results for datasets with calculated weight are in Figure 10. F_1 score significantly improved for subsets which have imbalance ratio lower or equal of 1:3.

4.3 Explanation of results

One of the goals of this work is to find more valuable signs of an addicted gambler. the most convenient method to do it is to apply methods giving a human-readable explanation of decisions made by classificator. There is no agreement within the machine learning community about the necessity of model explanation or what is human-readable explanation is [21]. In the other hand, modern machine learning models are too complex and thus it's too hard to interpret by a human.

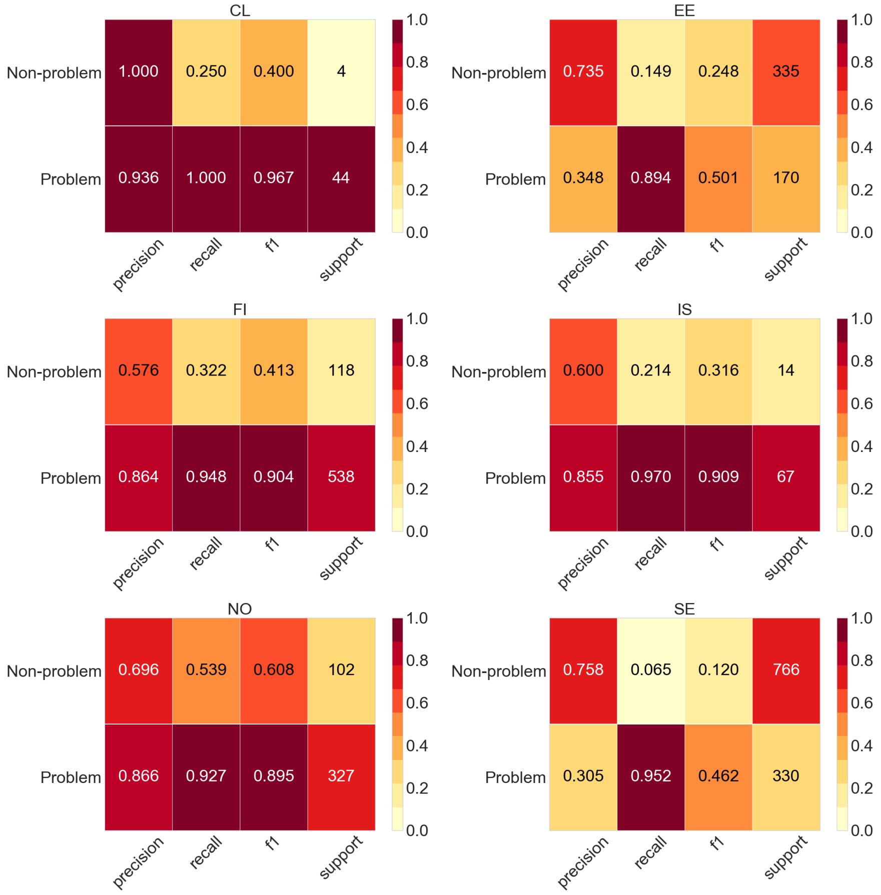


Figure 9. Classification reports by country

To get understandable explanations were used such libraries as LIME and SHAP. Both of them evaluates the power of influence on classification results using the method of local approximation but using slightly different techniques to achieve it.

To simplify the analysis of results explanation for examination was taken countries with highest F_1 score and support level. In this case, dataset representing Norway and Finland suits the best.

Random forest algorithm can be self-explanatory because it's based on large numbers of decision trees each of them able to show why was made this or that classification. In the other hand number of trees and features are also raise a problem of explanation, since it's a huge amount of work and every tree can give slightly different decision path. A random sample of three decision trees can be seen in Figure 11. Although depth of shown trees is only 2, it's already hard to find a correlation between them. The Random Forest model used in this model consists of one hundred trees with a depth equal 19

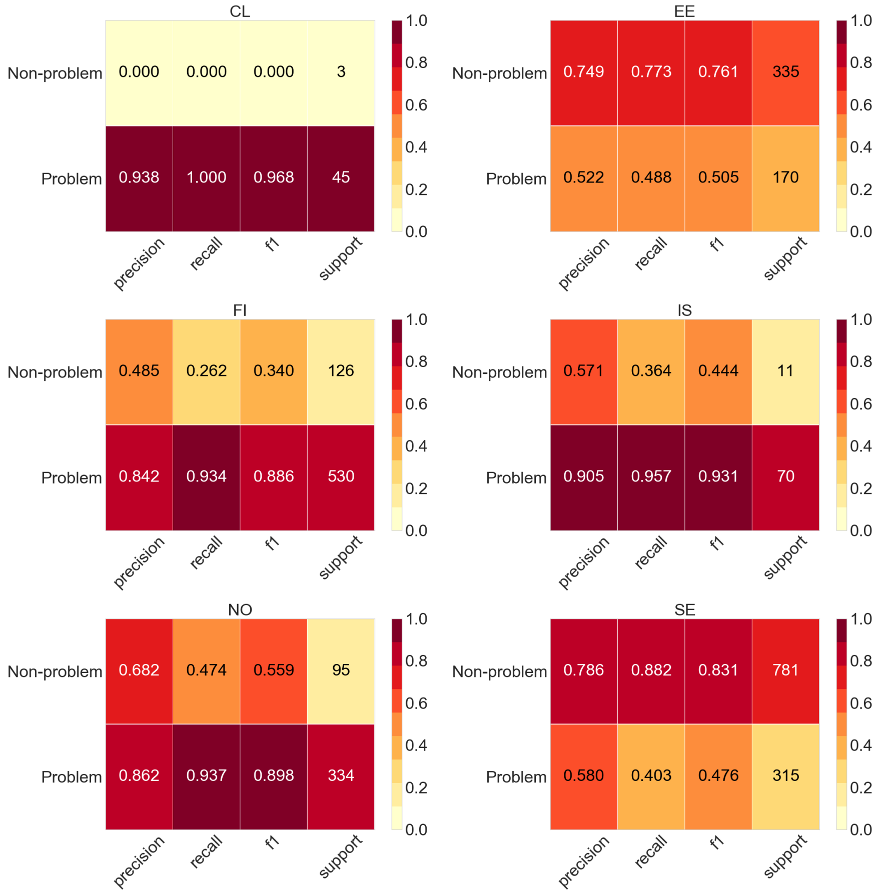


Figure 10. Classification reports by country for classifier with weighted classes

(depth of a decision tree is the length of the longest path from a root to a leaf). Thus decision trees are good explainers then it needs to deal with very few of them, what is not this case.

To show which features are more important and find the difference between true and false predictions dataframe with predictions was split into two parts according to the type of predictions and each part was explained separately. The separation was made by comparing predicted probabilities for both outcomes with the actual value of classification.

Although they work the same, LIME is more suitable for explaining single instances and can’t process the whole dataset in a single run. So for visualisation of classification results using LIME every instance of the dataset was explained separately and mean of numerical values of explanation was found. For LIME was used discretization feature that split up every feature to several subranges and analyzes them separately. Shap, on

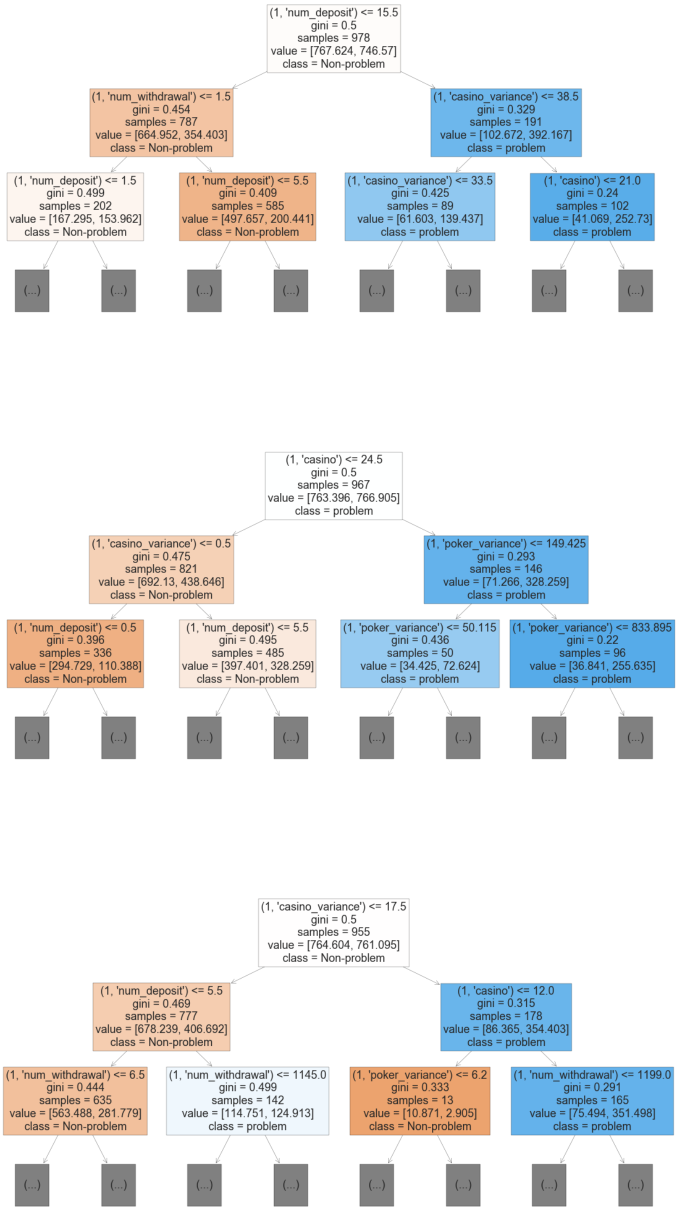
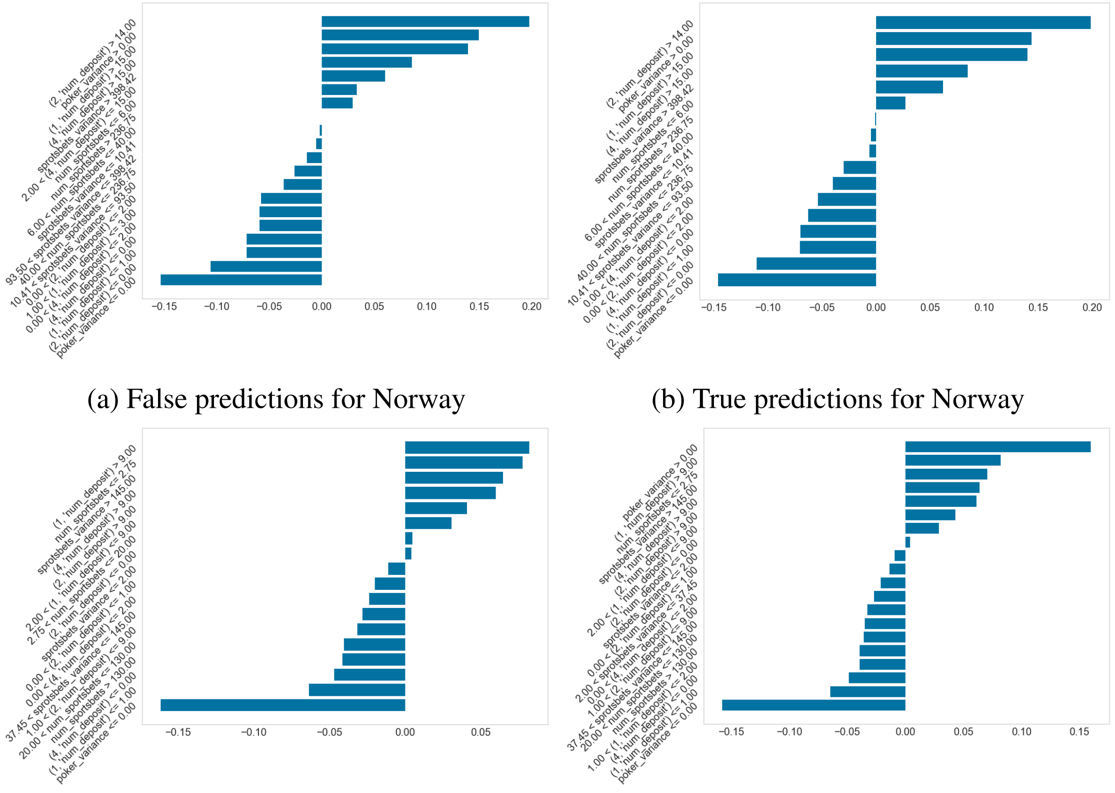


Figure 11. Random sample of three decision trees

the other hand, is easier to use and allow to visualise the whole dataset with a summary plot. Results of explanation are in Figure 12 and Figure 13.At figures showing results of LIME explanation bars at the left side of plot show power of features to classify a customer as not having problems and bars at the right power of features to classify a customer as a problem gambler.



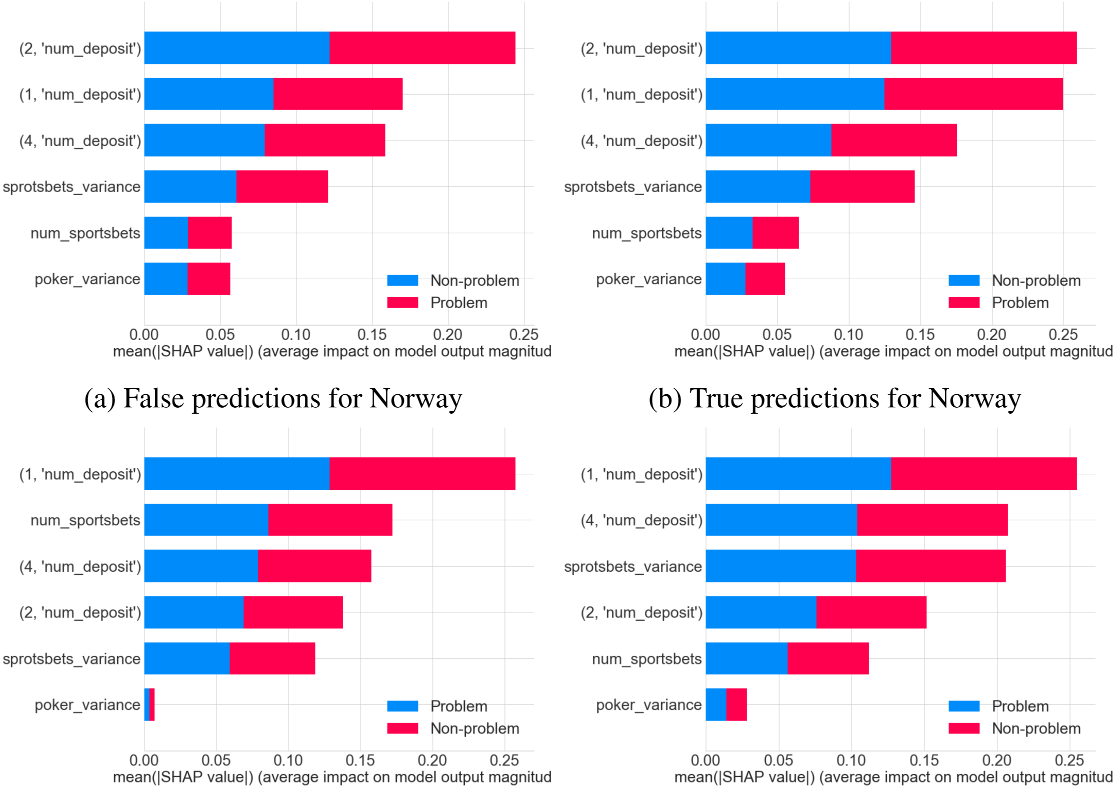
(a) False predictions for Norway

(b) True predictions for Norway

(c) False predictions for Finland

(d) True predictions for Finland

Figure 12. LIME results explanation for Norway and Finland



(a) False predictions for Norway

(b) True predictions for Norway

(c) False predictions for Finland

(d) True predictions for Finland

Figure 13. SHAP results explanation for Norway and Finland

5 Results

Accuracy level of classification is differ by countries, but in general it shows goods results. F_1 score for classification as problem and non-problem gambler for every country presented in Table 3.

LIME show no significant difference between false and true predictions and also between positive and negative for Norway. Both of them show a similar distribution between predictions. For Finland, it also shows a very similar distribution, but for true prediction, poker variance plays a very significant role. The only difference is that for false predictions LIME managed to split features into more ranges using discretize feature.

According to LIME, the most critical features characterizing problem gambler in Norway are the number of deposits in the first, second and fourth quartiles of gambling history with the number of deposits higher than 15, 14 and 15 relatively. Also any stakes in poker influence a lot. For Finland, these features are poker bets variance, number of deposits during the first quartile higher than 9, number of sports bets equals or lower

Table 3. Resulting F_1 scores by country.

Country	Problem $\mathbf{F_1}$ **score**	Non-problem $\mathbf{F_1}$ **score**
Chile	0.968	0.000
Iceland	0.931	0.444
Norway	0.898	0.559
Finland	0.886	0.340
Estonia	0.505	0.761
Sweden	0.476	0.831

2.75, sports bets variance higher than 145 and number of deposits during the fourth quartile higher than 9.

Shap doesn't have to discretize features so its results are more general, but show the same picture. Number of deposits in the first two quartiles together with the fourth quartile are the best influencers in classifying both problem and non-problem gamblers for Norway customers. There only a couple of notable differences with LIME results. First one is that variance in poker stakes affects very low. Secondly, Shap shows little difference between true and false predictions: number of deposits in the first quartile of gambling history affects more on classification for true predictions.

A completely different picture is existing for Finland. First of all, for false predictions, the number of sports bets plays a much more significant role and sports bet variance affect classification result noticeably less than for true predictions. For true predictions, features also have different order comparing to the same classifications made for Norway. Number of deposits during the first and fourth quartiles together with sports bets variance have the prevailing role in the case of the classification.

6 Discussion

This work was aimed to get higher accuracy of customer classification and to find the most significant features affecting this classification. To achieve this result as a part of this study separate machine learning model for each country was created. F_1 score for models varies from 0.476 to 0.968 for different countries, but overall, quality of prediction is high. Low F_1 scores for Estonia and Sweden can be explained by imbalanced dataset used for these countries. In these datasets number of non-problem gamblers are higher than the number of gamblers with problem behaviour, so precision is higher for classification customer as non-problem. Initial dataset was created with equal parts of the problem and non-problem gamblers and further separation by country were made during data exploration. This problem was mitigated by using weighted dataset.

To get explanations of classifier outcome were applied the most popular explainers LIME and SHAP. Both of them show a different set of features that affects the classification results the most. The number of deposits during first, second and fourth quartiles for Norway and number of deposits during first and fourth quartiles together with sports bets variance for Finland. These differences can be explained by the personal preferences and different personal characteristics of the citizens of these two countries. Another observation is the influence of sports bets variance that is also different for Finland and Norway. Assumption can be made, that influencing of this feature depends on numbers of national sports and country championships available in sports betting part of the platform.

The casino doesn't show any noticeable affection on final classification results. It can be connected with that fact, that Coolbet positions itself as a sports betting platform and thus the majority of customers don't play at all or play casino games very rarely.

Despite positive results achieved by this work, there are some limitations. The first thing that can be improved is to get more reliable ground trust data. In this work, there was no opportunity to get data about customers who have problem gambling behaviour, because it's medical data that can't be gathered freely. The possible solution to overcome this limitation can be adequate questionnaires made by qualified in this field people and involving a huge part of the customer base.

Another limitation that can affect the final results is a single source of data. Customers of online casino can play on different platforms simultaneously, which give an incomplete picture of the gambler. The study, conducted between several online casino companies inside one country can give much more information about the gambler. limiting data by one country make sense because customers from different providers can be easily identified by the unique code of citizen.

The method used in this work gives good results in the case of classifying gamblers, however, if combine it with the method suggested by Adami [2], than prediction can be more accurate.

Future extension of this work, connected with the dependence of portrait of problem gambler from its country of origin can show some interesting connection between them.

7 Conclusions

The outcome of this paper demonstrates the strong correlation between several behavioural markers and problem gambling behaviour, that can be used for classifying customers on early stages of his gambling history to help him when it's still possible and has maximum impact on problem behaviour developing. Most of the markers are connected to a number of deposits made by the customer at the start or the end it gambling history. During this work was demonstrated the possibility of classification gamblers using their gambling history.

Demonstrated classification accuracy is higher than in previous works in this field. Explanation of results using LIME and SHAP found correlatoin between gambler customer's country of living and set of signs affecting the most. This aspect of gambler classification didn't get enough attention during studies conducted before.

Using this approach customers of online gambling platforms can be classified as a problem gambler with a high level of confidence in order to prevent harm from uncontrolled gambling. Such system implemented in every online betting platform can drastically decrease number of people affected by online casinos and online totalisators.

References

[1] *Elanike kokkupuuted hasart- jaõnnemängudega [Gambling prevalence in Estonia].* Faktum Uuringukeskus, 2004.

[2] Nicola Adami, Sergio Benini, Alberto Boschetti, Luca Canini, Florinda Maione, and Matteo Temporin. Markers of unsustainable gambling for early detection of at-risk online gamblers. *International Gambling Studies*, 13(2):188–204, aug 2013.

[3] Charu C Aggarwal. *Data mining: the textbook.* Springer, 2015.

[4] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, aug 1992.

[5] J. Braverman and H. J. Shaffer. How do gamblers start gambling: identifying behavioural markers for high-risk internet gambling. *The European Journal of Public Health*, 22(2):273–278, jan 2010.

[6] Julia Braverman, Debi A. LaPlante, Sarah E. Nelson, and Howard J. Shaffer. Using cross-game behavioral markers for early identification of high-risk internet gamblers. *Psychology of Addictive Behaviors*, 27(3):868–877, sep 2013.

[7] Leo Breiman. *Machine Learning*, 45(1):5–32, 2001.

[8] J. Caillon, M. Grall-Bronnec, B. Perrot, J. Leboucher, Y. Donnio, L. Romo, and G. Challet-Bouju. Effectiveness of at-risk gamblers' temporary self-exclusion from internet gambling sites. *Journal of Gambling Studies*, 35(2):601–615, jul 2018.

[9] Margaret Carran. Review of the implementation of selected provisions of european union commissionrecommendation 2014/478/eu across EU states. `https://www.egba.eu/uploads/2018/12/181206-Consumer-Protection-in-EU-Online-Gambling-EBGA-Report-December-2018.pdf`, December 2018. Accessed: 13.05.2020.

[10] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, jun 2006.

[11] Jacqueline Ann Ferris and Harold James Wynne. *The Canadian problem gambling index.* Canadian Centre on Substance Abuse Ottawa, ON, 2001.

[12] Yoav Freund and Robert E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Lecture Notes in Computer Science*, pages 23–37. Springer Berlin Heidelberg, 1995.

[13] Sally M. Gainsbury. Online gambling addiction: the relationship between internet gambling and disordered gambling. *Current Addiction Reports*, 2(2):185–193, apr 2015.

[14] Heather M. Gray, Debi A. LaPlante, and Howard J. Shaffer. Behavioral characteristics of internet gamblers who trigger corporate responsible gambling interventions. *Psychology of Addictive Behaviors*, 26(3):527–535, sep 2012.

[15] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class AdaBoost. *Statistics and Its Interface*, 2(3):349–360, 2009.

[16] Anna Krajuškina. Modelling parkinson's disease with gait analysis approach. mathesis, Tallinn University of Techology, 2019.

[17] Stella Laansoo and Toomas Niit. *Estonia*. Springer New York, dec 2008.

[18] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, jan 2020.

[19] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[20] C. E. Metz. Basic principles of roc analysis. *Seminars in nuclear medicine*, 8:283–298, October 1978.

[21] Sven Nõmm. Lecture 13 of machine learning course in tallinn university of technology by s. nõmm. `https://courses.cs.ttu.ee/pages/ITI8565#Lecture_13_Trace.2C_explain_and_interpret`, April 2020.

[22] Nancy M. Petry, Carlos Blanco, Randy Stinchfield, and Rachel Volberg. An empirical evaluation of proposed changes for gambling diagnosis in the dsm-5. *Addiction (Abingdon, England)*, 108:575–581, March 2013.

[23] Kahlil S. Philander. Identifying high-risk online gamblers: a comparison of data mining procedures. *International Gambling Studies*, 14(1):53–63, oct 2013.

[24] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[25] Christine Reilly. Responsible gambling, aug 2019.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”: Explaining the predictions of any classifier.

[27] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, jan 1974.

[28] Juliana Tolles and William J. Meurer. Logistic regression. *JAMA*, 316(5):533, aug 2016.

[29] World Health Organization. International statistical classification of diseases and related health problems (11th ed.). `https://icd.who.int/`, 2019.