PREDICTIVE ANALYSIS OF KEY FACTORS THAT INFLUENCE HIGH RISK CROSS GAMBLING WITH CLASS BALANCED CATEGORICAL HYBRID FEATURE SELECTION

[1]Dr.T.R.Sivapriya,[2]C.Malarvizhi

[1]Associate Professor,[2]Assistant Professor
[1]Department of Computer Science,
[1]Lady Doak College, Madurai,India

Abstract :— Online gaming addiction greatly affects the social and psychological wellbeing of people. Early prediction of problem gambling plays a major role in the diagnosis and treatment of the problem gambler. Machine learning algorithms are found to be efficient predictors in finding the major factors that lead to problem cross gambling. Forecasting problem gambling behavior would mitigate the risk of unhealthy gambling behavior. Missing data and unbalanced real time datasets influence the classification accuracy to mine gambling behaviors. Preprocessing of dataset improves the performance of classifier. Three single Imputation and three hybrid techniques are compared by the accuracy provided by random forest. Class balancing of dataset improved the classification accuracy of random forest. Bayesian networks and random forest classifiers were used for prediction of problem cross gambling. Imputation followed by class balancing yielded an accuracy of 96.5% and AUC nearly 0.98 with random forest classifier. The improved accuracy would enhance early prediction for early intervention and medical care. The proposed categorical hybrid feature selection technique explored the optimal feature subset for the effective intervention in cross gambling that could be applied in similar online games, with an accuracy of 99.73.

***IndexTerms* - Problem gambling, classification, imputation, random forest.**

I. INTRODUCTION

Internet gambling has become a big addiction among various age groups of people. Regular betting habits and their neurobiological correlates have been reported from various research studies [1]. The importance of prediction for early intervention can bring about a great change in the lifestyle of a gamer. Machine learning has been widely applied to identify problematic gambling habits. There is an enormous rise in online gambling due to the technological boom. Online gambling has become an addiction among all age groups. Internet gaming disorder is identified as a major health problem affecting people all over the world. The impact of this addiction reflects in the personality or behavior of the people under online gaming addiction. Research shows that many of the students were addicted to gambling, and they were lacking in their educational progress [2]. As the peoples were addicted, they were not spending time with family members and the social activities. The addictions of Internet gambling were considered as doubly addicted as they were addicted to Internet usage first [3]. The impact of this addiction leads to social issues such as protection of vulnerability, Internet gambling in workplace and lacking in recent issues about the society [4]. The early invention of problem gamblers may lead them to recover from psychological issues and awareness about the societal concerns.

II. RELATED RESEARCH

Labrie et al.(2007) tracked the primary gambling behaviors of fixed odd and live action betting[5]. This study uncovered the types of Internet gambling such as casino games, poker playing and the population segments at greater or lesser risk for developing Internet gambling related problems. Laplante et al. (2008) analysed the online gambling participation and activity among a population of newly subscribed Internet bettors [6]. This study explored the implications for psychopathology and other tangible consequences of gambling-related problems. Braverman et al. (2010) identified behavioral markers for high-risk internet gambling [7]. During the first month of actual internet gambling on a betting site, betting patterns were identified to predict the problems related to gambling. KMeans clustering was used to identify a sub-group of high-risk gamblers. Limitations of this research was that it distinguished only a small proportion of the total sample and concluded that further research should be performed to analyze the high proportion of high-risk gamblers. Braverman et al. (2011) presented findings from the first taxometric study of actual gambling behavior to determine whether we can represent the characteristics of extreme gambling as qualitatively distinct or as a point along a dimension [8]. Ruscio'staxometric R program was used to produce taxomeric plots and perform all calculations. In their study, two taxometric procedures were applied (i.e., MAMBAC and MAXCOV) to three indicators of betting behavior, total money lost, total number of bets and total money wagered but failed to support a categorical Predictive analysis of key factors that influence high risk cross gambling with class balanced categorical hybrid feature selection Dr.T.R.Sivapriya, Lady Doak College, Madurai, Ms. Malarvizhi, Lady Doak College, Madurai understanding of excessive Internet Sports Gambling behavior. Philander (2014) identified high-risk online gamblers[9] and it expands the behavioral identification work by Braverman and Shaffer (2012) and LaBrie and Shaffer(2011). In this study, nine supervised learning models such as Step-wise logistic regression, Lasso/elastic-net logistic regression, Neural network(regression), Neural network(classification), Support vector Machines (Regression and Classification, one - Classification) and Random Forest regression and Classification) were used and concluded that the best SLM to predict the high-risk online gamblers for this dataset were the step-wise logistic and the GLM models [10]. Percy, Christian, et al., (2016) predicted online gambling self-excluders [11]. This study focused on achieving high accuracy while using Bayesian networks and creating datasets with a roughly equal number of self–excluders and control group gamblers (‘SMOTE’) and improvement of accuracy performance higher than 62-67% range in Philander (2014) while using random forest.

III. DATA SET DESCRIPTION AND PROPOSED WORK

The data was collected from the Division on Addiction, Cambridge Health Alliance. The dataset provides the information about betting behavior of the bwin Internet casino subscribers who opened an account. Total numbers of instances are 4056, in which 1019 are in high risk and 3037 are in controlled level. Bwin interactive entertainment contains four types of products (Sports betting, Poker, Casino games, Soft games). The dataset contains the sum of bets, sum of stakes, Standard deviation (Variability) of stakes, Standard deviation (Variability) of bets and total active days during the first month for each of the four products. In addition, it contains the week frequency, weekend sum of stakes, weekend sum of bets, weekday sum of stakes, weekends sum to bets ratio and Week frequency trajectory for each product.

IV. PROPOSED METHODOLOGY

Contribution of the proposed research work is twofold. This work enhances the early detection of problem gambling by preprocessing for imputation of missing values and class balancing. With enhanced classification, categorical hybrid feature selection is implemented in R to find the best possible feature subset that should be taken into account for early diagnosis and treatment.

A. PREPROCESSING

Analysis of existing dataset shows an imbalance among the classes and having missing values for some of the features of the dataset. The overall missing values percentage in the dataset is 8.70. Missing percentage in the available dataset is depicted in the following Figure 1.

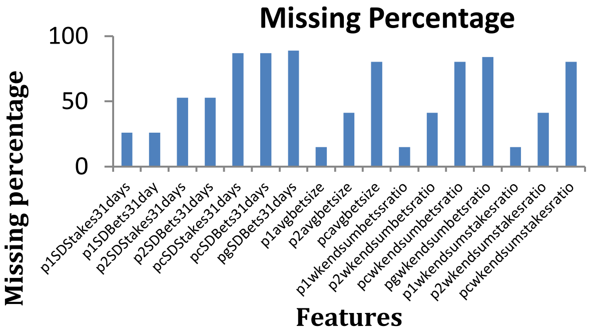

Figure 1: Missing Percentage for each feature.

B. IMPUTATION BY DATA ANALYSIS

The dataset must be preprocessed before training the data because the missing values in the dataset will affect the performance of the classifier [12]. The dataset was classified with Decision tree, Random forest and Bayesian. While training the dataset, it was observed that Random forest classifier yield the good prediction regardless of the missing values. The performance of the model was evaluated by the metrics such as accuracy, sensitivity and specificity. The accuracy of the model could not be evaluated absolutely as there were missing values in the dataset. While training with missing values, decision tree and Bayesian classifiers yielded comparatively lesser. The classification model must be designed to do the prediction with high accuracy. To improve the efficiency of a model, the dataset was subjected to imputation of the missing values. There are various methods to perform imputation. The dataset was imputed with six different combinations of imputation MICE, KNN and PCA and classified with Random forest as represented in Table 1.

Table 1: Accuracy with Random forest with imputation methods

	M1- (MICE)	M2- KNN	M3- PCA	M4- Hybrid (Mean +Ratio)	M5 missForest	M6- Mice+Ratio
Accuracy	86.05	78.8	78.55	74.03	76.01	78.39
Sensitivity	0.442	0.098	0.106	1	0.003	0.176
Specificity	1	1	1	0	0.009	1

MICE method creates multiple imputations for multivariate missing data. There are four methods used by this package to impute missing data. Predictive Mean Matching method was used as the features in the dataset contains numeric values. The intensity of the imputed data is shown in Figure 2. In this figure, blue represents the actual data and red represents the imputed data.

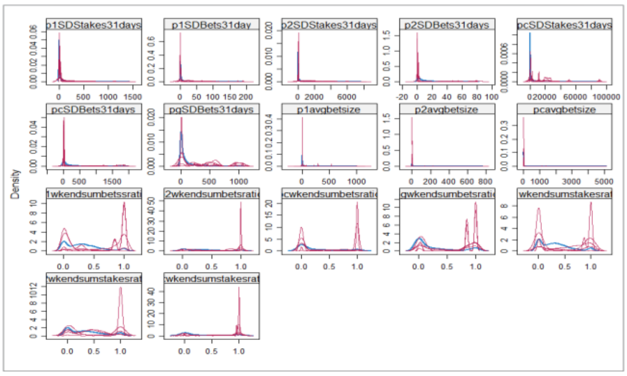

Figure 2. Imputed plot density for each attribute

KNN Imputation uses k-Nearest Neighbours approach to impute missing values. For every missing data to be imputed, it finds ‘K’ closest observations and calculates the weighted average. PCA method of computation is compared with KNN and MICE. The imputed datasets were trained with random forest classifier and it was observed that that random forest yielded higher accuracy of above 95% with the dataset imputed using MICE method. The accuracy is comparatively higher than the result reported with random forest without imputed dataset (Percy et al. 2016). This will be followed by class balancing.

C. CLASS BALANCING

The dataset was imbalanced dataset. In the dataset, the numbers of problematic internet gamblers were less than the responsible gamblers. For the purpose of model building, a balanced training dataset was created using SMOTE and ROSE. Random forest is the classifier used to evaluate datasets class balanced by SMOTE and ROSE. The number of trees used for random forest is 500 and the number of variables per split is 2. SMOTE yields good performance than ROSE. The balanced dataset created using ROSE yields eighty percentage accuracy and SMOTE yields 96.6% of accuracy.

D. Categorical Hybrid Feature selection:

The features are categorized into 6 categories of features such as Variance, Frequency, Betting, Trajectory, Games and Risk. The number of features in each categories are Variance of Bets and Stake (14), Frequency (57), Betting (16), Trajectory of wager (8), Games(6) and Risk Group of live action and fixed odd(3). The Redundant Feature elimination technique is used to select the significant features by removing the weakest features from each category. After removing the redundant and weak features from each subset, features were selected incrementally from each categorical subset with Random forest to identify a subset of features that contribute to increase the accuracy of the classifier. The most significant features from each subset were Variance of Bets and Stake (10), Frequency (22) and Betting (7) are combined and incrementally tested. The result is cross verified with a backward selection technique and has been compared with the literature. The subset of features selected under the category the betting characteristics, the randomness associated with the behavior and thereby improves the sensitivity (0.994) and accuracy (99.73) of the classification. The proposed method will select the optimal feature subset to identify problem gamblers in ‘N’ iterations. TS contains the entire feature subset, which is further subdivided into categories. Redundant features are eliminated from each category in during the first phase. In the second phase, features from each category are combined with each other incrementally the best predictive subset that could contribute in early prediction. The algorithm for the categorical hybrid feature selection is as follows:

```
Categorical Hybrid feature selection (N, TS, FSS)
N- total number attributes; TS - Total attributes; W- weightage; CN - Number of categories;
$C_i$- attribute in each category; $FC_i$- Relevant Feature Subset
Begin
Step 1.  Identify the number of categories in the dataset based on similarity of features
Step 2.  Divide feature subset into categories ; CN – Number of categories
Step 3.  For each Category from 1 to CN
                    For each attribute in $C_i$:
                    Identify features with equal weightage based on contribution to classification
          W=weightage of attributes
          End loop
Step 4. Elimination of attributes with very lower weightage below threshold
          For each Category from 1 to CN
                    For each attribute in $C_i$:
                    If $W(C_i)$ >  threshold (Average importance of each attribute)
                              $FC_i$<- $C_i$
                    End loop
          End loop
Step 5. #Incremental Combinatorial Featureselection(FC)
          FC- categories with selected attributes, i,j,k=1
          Repeat
                    $C_m$<- {$FC_i$} U{$FC_j$} (Union of  Ci and Cj)
                    Acc(k)<- Accuracy ($C_m$) ; k<-k+1
          Until all possible combinations are explored[End Loop]
Step 6 . Sort the Accuracy array and Combined feature subset
          For I in 1 to k
                    Decrementally sort Acc(k) and $C_m$ accordingly
          End loop
Step 7.FSS<- features subset Cm with highest accuracy

Return BestFeaturesubset FSS
End
```

V. RESULTS AND CONCLUSION

Prediction of problem gambling is essential for early intervention. The feature set available describes the entire characteristics of the online gambler describing the betting styles. The presence of missing values has greatly influenced the classification in earlier studies (Philander 2014, Percy et al. 2016). Hence in this research work, multiple imputation was performed to improve the sensitivity and specificity of the classifier. It was observed that among the classifiers, Random forest was robust providing an AUC of 0.657 even in the presence of missing values. With imputation the AUC increased to 0.953. However, sensitivity remained around 0.5. When the class balancing was implemented with SMOTE, Sensitivity increased to 0.91 and Accuracy to 0.96. With the optimum features, the AUC increased to 1, Sensitivity increased to 0.99 and Accuracy to 99.73 as given in Table 2.

Classifier/ Performance	Without imputation				With imputation				Proposed method With imputation and class balancing			
	AUC	Acc	Sen	Spe	AUC	Acc	Sen	Spe	AUC	Acc	Sen	Spe
Bayesian	0.515	24.2	0.201	0.828	0.506	29.75	0.925	0.087	0.599	98.2	1	0
Random forest	0.623	69.2	0.6	0.5	0.953	86.05	0.442	1	1	96.6	0.918	1
Decision tree	0.512	75.2	0.993	0.003	0.859	90.96	0.6	0.5	1	100	1	1

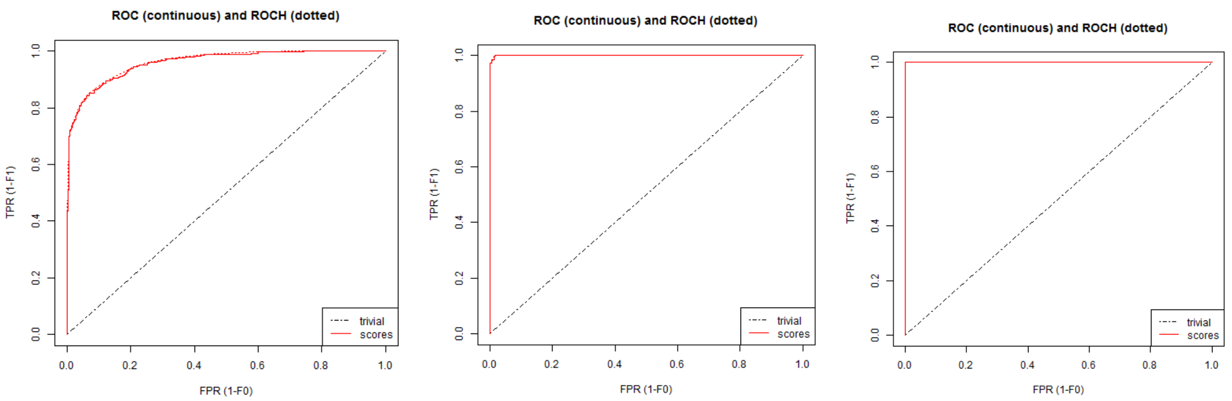

Fig 3.(a) AUC of Random forest with Fig 3.(b) AUC of Random forest with Fig 3.(c) AUC with optimal feature subset
imbalanced dataset balanceddataset

The accuracy of the model was enhanced with balanced imputed dataset as shown in Figure 3.b. and the accuracy of optimal features selected for early invention of highrisk gamblers is shown in Fig. 3.c. Optimal feature subset correlated with Richard et al. [1]. The optimal feature subset contains the betting behavior, sum of stake and total number of active days of gambling. In data preprocessing, as the missing values of the features were imputed and increased the minority class, the performance of the classifier was improved. The categorical hybrid feature selection finds the key factors of internet gambling. Hence, the performance of the classifier was improved. The proposed research work deals with the data preprocessing, class balancing and categorical hybrid feature selection was done with cross gaming features and is not restricted to single game, the obtained optimal feature subset is applicable to any given online game with similar characteristics. Further research could analyze the psychological problems related to gamblers in country wise, gender wise behavior problems.

REFERENCES

[1] D. J. Kuss, H.M. Pontes, and M. D. Griffiths,“ Neurobiological correlates in Internet Gaming Disorder: A systematic literature review,” *Frontiers in psychiatry* ,9, 2018.

[2] M.Griffiths, and A. Barnes, “ Internet gambling: An online empirical study among student gamblers,” *International Journal of Mental Health and Addiction*, 6(2), 194-204, 2008.

[3] M. D. Griffiths, and J. Parke, “The social impact of internet gambling,” *Social Science Computer Review*, 20(3), 312-320, 2002.

[4] M. Griffiths, “Internet gambling: Issues, concerns, and recommendations,”. *CyberPsychology and Behavior*, 6(6), 557-568, 2003.

[5] R. A. LaBrie., D. A. LaPlante, S.E. Nelson, A. Schumann, and H.J. Shaffer, “Assessing the playing field: A prospective longitudinal study of internet sports gambling behavior,” *Journal of Gambling studies*, *23*(3), 347-362, 2007.

[6] D. A. LaPlante., A. Schumann., R. A. LaBrie., and H.J. Shaffer, “Population trends in Internet sports gambling,” *Computers in Human Behavior*, *24*(5), 2399-2414, 2008.

[7] 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, 2010.

[8] J. Braverman, R. A. LaBrie, and H. J. Shaffer, ”A taxometric analysis of actual internet sports gambling behavior,” *Psychological Assessment*, *23*(1), 234, 2011.

[9] K. S. Philander, and T. MacKay,“Online gambling participation and problem gambling severity: is there a causal relationship?,” *International Gambling Studies*, *14*(2), 214-227, 2014.

[10] K. S. Philander, “Identifying high-risk online gamblers: a comparison of data mining procedures,” *International Gambling Studies*, *14*(1), 53-63, 2014.

[11] C. Percy, M. França, S. Dragičević, and A. d’AvilaGarcez, “Predicting online gambling self-exclusion: an analysis of the performance of supervised machine learning models,” *International Gambling Studies*, *16*(2), 193-210, 2016.

[12] P. Schmitt, J. Mandel, and M. Guedj, “A comparison of six methods for missing data imputation,”. *Journal of Biometrics & Biostatistics*, *6*(1), 1, 2015.