UDC 004.8:336:796.3:519.83
DOI: 10.20535/SRIT.2308-8893.2025.3.07

ANALYSIS AND FORECASTING OF THE FINANCIAL BENEFIT FOR THE TENNIS MATCH OUTCOMES BY MACHINE LEARNING METHODS

K. SHUM, N. KUZNIETSOVA

Abstract. Tennis is one of the most popular sports in the world, attracting considerable attention from casual fans and professional analysts. The application of machine learning methods enables the accurate prediction of match results, opening up opportunities for profit through betting on likely winners. This study evaluates the financial benefits of predicting tennis match outcomes by identifying an effective sports betting strategy. The study examines various machine learning methods and auxiliary algorithms, comparing them to select the best betting strategy for maximizing the user's potential profit. In the paper, the method and algorithm for determining effective sports betting strategies were developed. This algorithm and method were tested on tennis game datasets (for both women and men), and the best tennis betting strategy was identified. As part of the study, a software product has been developed to predict the outcomes of tennis matches.

Keywords: forecasting, machine learning, betting strategies, financial benefit.

INTRODUCTION

Tennis is a dynamic and unpredictable game, combining many factors that influence the course of events during matches: players' physical conditions, psychological state, chosen tactics, anthropometry, weather conditions, and more. Each of these aspects can be decisive in achieving the desired outcome. Thanks to this versatility, tennis ranks among the most popular sports globally, captivating a broad audience of fans, from casual spectators who enjoy the thrill of the game to professional sports analysts who study the game from a scientific perspective.

Match outcome prediction holds a special place among the various aspects of sports interest. As for the standard fan, a match result is typically a topic of discussion and emotional enjoyment. However, the prediction is practical for analysts and professional bettors who place wagers on sports events [1]. Knowing the likelihood of a player's victory not only allows for more informed betting to secure financial gain but also aids in developing strategies for long-term success. In this context, the betting process goes beyond simple gambling for many professional participants in the sports betting market. Predicting tennis match outcomes becomes a critical tool for making informed decisions, assessing risks, and evaluating potential benefits, ultimately supporting the financial growth of the bettor. It also gives us the possibility to solve such tasks as understanding behavior and forecasting the gamer's outflow [2]. The players who win are motivated to stay longer in the game while they understand the game's process and can also plan their own strategy and evaluate their financial benefits.

PROBLEM STATEMENT

This research was conducted to deeply understand the betting and gambling processes by applying modern techniques and approaches. It was first decided to try machine learning algorithms for evaluating and forecasting games' outcomes and for finding hidden dependencies. Then, based on these models, we can find the most important variables that could be interpreted as some key factors for winning on some side. It means that we can also take into account some preliminary information before making a bet. Next, the strategy of effective betting should be defined. For this reason, we will develop the algorithm for defining the most effective strategy that can be used by gamblers to maximize their profit as a result of the sports betting process.

MACHINE LEARNING METHODS AND AUXILIARY ALGORITHMS FOR THE GAME OUTCOMES PREDICTION

Machine learning methods

Machine learning is currently a powerful tool for solving various tasks across different fields of human activity. The significant potential and efficiency of machine learning methods and algorithms make this technology crucial in areas where traditional approaches may fall short. Predicting the outcomes of sports events is no exception. In this study, predictive models have been developed to predict the results of men's (hereafter, M) and women's (hereafter, W) tennis matches. These models are based on logistic regression (M and W), multilayer perceptron (W), random forest (M), and extreme gradient boosting (M).

Logistic regression is a method that models the relationship between a categorical target variable and a set of independent predictor variables. Although logistic regression is a classification algorithm, it is based on a linear regression model [3; 4]. To produce categorical outcomes, it transforms the continuous output of linear regression into a range between 0 and 1 (interpreted as the probability of belonging to a specific class) using the logistic function, also known as the sigmoid function [3], which can be described by the following formula:

$$\sigma(z) = \frac{1}{1+e^{-z}} \ ,$$

where $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$ is a linear combination of independent variables x_i and their coefficients β_i, $i = \overline{1,n}$, n is a number of predictor variables and β_0 is the intercept term.

Then the probability of an object's belonging to a specific class can be represented as:

$$P(y=1) = \sigma(z)\,; \quad P(y=0) = 1 - \sigma(z)\,.$$

Random forest is an ensemble machine learning method that combines the predictions of multiple decision trees to improve the model's accuracy and stability. The trees are constructed on random subsets of data from the training set, and random subsets of features are used to reduce the correlation between the trees. The final prediction $\hat{y}$ is determined by majority voting, making the method robust to overfitting and effective for various tasks [5], and is determined by the following formula:

$$\hat{y} = Mode\{h_i(x)\}, \quad i = \underline{1, n} \ ,$$

where $h_i(x)$ is a prediction of the i-th tree, and n is a number of trees in the forest.

Extreme Gradient Boosting (XGBoost) is an ensemble machine learning method that implements gradient boosting with decision trees. XGBoost uses an iterative approach, where the key idea is to build an ensemble of decision trees, with each subsequent tree sequentially correcting the errors of the previous ones, thereby improving the model's overall accuracy [6]. If the prediction for the i-th sample after $k-1$ iterations is represented as $\hat{y}_i^{(k-1)}$, then at the k-th iteration, the prediction value will be updated using the following formula:

$$\hat{y}_i^{k)} = \hat{y}_i^{(k-1)} + \eta h_k(x_i),$$

where η is the learningn rate, which determines how strongly each tree influences the final prediction.

A *multilayer perceptron* (MLP) is a type of artificial neural network consisting of several layers: an input layer, one or more hidden layers, and an output layer. Each layer contains neurons that take a weighted sum of input data from the previous layer, apply an activation function to it, and pass the result to the next layer. Weighted sum is counted using the following formula:

$$z_j^{(l)} = \sum_{i=1}^{n} w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)},$$

where $z_j^{(l)}$ is the activation of neuron j in layer l; $w_{ij}^{(l)}$ is the weight connecting neuron i in the previous layer to the neuron j; $a_i^{(l-1)}$ is the activation of neuron i in the previous layer; $b_j^{(l)}$ is the bias of neuron j.

The training process is repeated over several iterations until the model converges to an optimal solution. Due to the architecture, MLP can model complex nonlinear relationships between data [7].

Auxiliary algorithms

The *time discounting method* is an approach that assigns greater significance to newer data and less to older data [8]. The idea is to apply weights relative to the time between events. In the study, this method is used to predict player statistics in the men's division, as the prediction of match performance is based on the values of relevant statistical variables. Weights are applied using an exponential function $W(t)$:

$$W(t) = \min(f^t, \ f),$$

where t represents the time in months between the scheduled match and a previously played match, and f is the discount factor, which can range from 0 to 1. The discount factor determines the extent of time discounting and is set by the researcher. The smaller the value of f, the lower the weight given to older matches.

The *data filtering algorithm* is the process of selecting subsets from a large dataset based on specified criteria. Initially, the selection conditions are defined (for example, these may be the values of variables), after which the corresponding samples are formed, which allows the efficient extraction of the most relevant data for further analysis and use.

BUILDING THE MACHINE LEARNING MODELS

For this study, we decided to use real data and develop our models for the prediction of tennis match outcomes using Python, along with relevant machine learning and data processing libraries (Sklearn, Pandas, Numpy, etc.). For this reason, we used two different datasets for men's and women's games. The dataset from user JeffSackmann's GitHub repository [9] was used to predict men's matches, and the dataset from the Tennis-data website [10] was utilized for women's matches. In both cases, the records began at the start of 2010, with a total of 153.959 and 37.731 games recorded, respectively.

For both datasets, initial preprocessing was carried out: the properties and specifics of each variable were analyzed, missing values were handled, irrelevant records and variables were removed, and the data was transformed into a format suitable for future models. Each dataset was duplicated, and the corresponding player columns were swapped to balance the number of positive and negative classes (1 for the first player's victory, 0 for loss). As a result, the final training datasets contained 209.116 and 47.816 records for men and women, respectively.

The most important features for prediction by using statistical methods were selected, such as:

- **For men:** twenty significant predictor variables were selected, including tournament seeding numbers, differences in height, ranking, ranking points, as well as percentage differences in various statistical indicators (e.g., first serve percentage, percentage of points won on return, etc.).
- **For women:** five significant predictor variables were selected, including differences in ranking points, age, and differences in the win odds set by the Pinnacle bookmaker and between maximum and average odds from other bookmakers.

The next stage involves constructing machine learning models using the methods mentioned earlier. To determine their best parameters, the grid search algorithm was applied. This algorithm selects the combination of the most effective features from a given set that ensures the highest model performance. The results obtained are presented in Tables 1–5.

Table 1. Parameters of the logistic regression model (women)

Parameter	Description	Value
test_size	The proportion of the dataset that is used for testing the model	0.1
solver	The method for determining the optimal model weights that minimize the loss function	liblinear
fit_intercept	The presence of a bias term in the model equation	False
C	Regularization strength	4.25
penalty	The type of regularization used to control the model's overfitting	L2

Table 2. Parameters of the multilayer perceptron model (women)

Parameter	Description	Value
test_size	The proportion of the dataset that is used for testing the model	0.1
solver	The method for determining the optimal model weights that minimize the loss function	lbfgs
activation	Activation function of the hidden layer	relu
alpha	L2-regularization strength	0.005
hidden_layer_ sizes	Number of neurons in the hidden layers	(100,)
learning_rate	Learning rate for weight updates	constant

Table 3. Parameters of the logistic regression model (men)

Parameter	Description	Value
test_size	The proportion of the dataset that is used for testing the model	0.1
solver	The method for determining the optimal model weights that minimize the loss function	newton-cg
fit_intercept	The presence of a bias term in the model equation	True
C	Regularization strength	1
penalty	The type of regularization used to control the model's overfitting	None

Table 4. Parameters of the random forest model (men)

Parameter	Description	Value
test_size	The proportion of the dataset that is used for testing the model	0.15
n_estimators	The number of trees in the forest	100
criterion	The function to measure the quality of a split	log_loss
max_features	The number of features to consider when looking for the best split	None
min_samples_leaf	The minimum number of samples required to be at a leaf node	2

Table 5. Parameters of the XGBoost model (men)

Parameter	Description	Value
test_size	The proportion of the dataset that is used for testing the model	0.15
n_estimators	The number of trees (iterations) of the model	100
learning_rate	Learning rate	0.05
max_depth	The maximum depth of each tree	6
reg_alpha	L1-regularization parameter	0.5
reg_lambda	L2-regularization parameter	1.5

After building the models with the specified parameters, their performance was evaluated on the training and validation datasets using standard classification quality metrics. The results are presented in Tables 6–10.

Table 6. Evaluation of the logistic regression model (women)

Sample	Quality metric					
	Accuracy	Precision	Recall	F1 Score	Roc Auc	Loss
Training	0.68966	0.68897	0.70124	0.69496	0.76102	0.57987
Validation	0.68341	0.68355	0.68202	0.68277	0.75474	0.58704

Table 7. Evaluation of the multilayer perceptron model (women)

Sample	Quality metric					
	Accuracy	Precision	Recall	F1 Score	Roc Auc	Loss
Training	0.69175	0.68903	0.70954	0.69894	0.7599	0.5803
Validation	0.68290	0.68239	0.68323	0.68274	0.7535	0.58745

Table 8. Evaluation of the logistic regression model (men)

Sample	Quality metric					
	Accuracy	Precision	Recall	F1 Score	Roc Auc	Loss
Training	0.98278	0.98292	0.98253	0.98272	0.99790	0.05599
Validation	0.98145	0.98146	0.98144	0.98145	0.99789	0.05618

Table 9. Evaluation of the random forest model (men)

Sample	Quality metric					
	Accuracy	Precision	Recall	F1 Score	Roc Auc	Loss
Training	0.98345	0.98345	0.98376	0.98381	0.99763	0.07793
Validation	0.98363	0.98362	0.98366	0.98349	0.99789	0.07484

Table 10. Evaluation of the XGBoost model (men)

Sample	Quality metric					
	Accuracy	Precision	Recall	F1 Score	Roc Auc	Loss
Training	0.98396	0.98501	0.98293	0.98396	0.99834	0.04733
Validation	0.98359	0.98267	0.98455	0.98361	0.99842	0.04601

After analyzing the results obtained, it can be concluded that the models predicting women's matches perform at an acceptable level but are slightly worse than those predicting men's matches. The models for men's tennis show excellent values across all metrics. However, it is important to note that their performance may decline due to the necessity of applying the time discounting method to predict statistics for future matches. None of the developed models exhibit signs of overfitting, as the quality metrics for both the training and validation datasets are very close.

To facilitate the process of predicting matches and to provide a straightforward interpretation of the results, a web interface was developed to allow users to interact with the developed models easily, input the necessary data for predictions via the keyboard, and modify it if needed. Separate prediction pages for men and women were implemented, with their interfaces shown in Figs. 1 and 2. Additionally, a database containing historical match records for men was created.

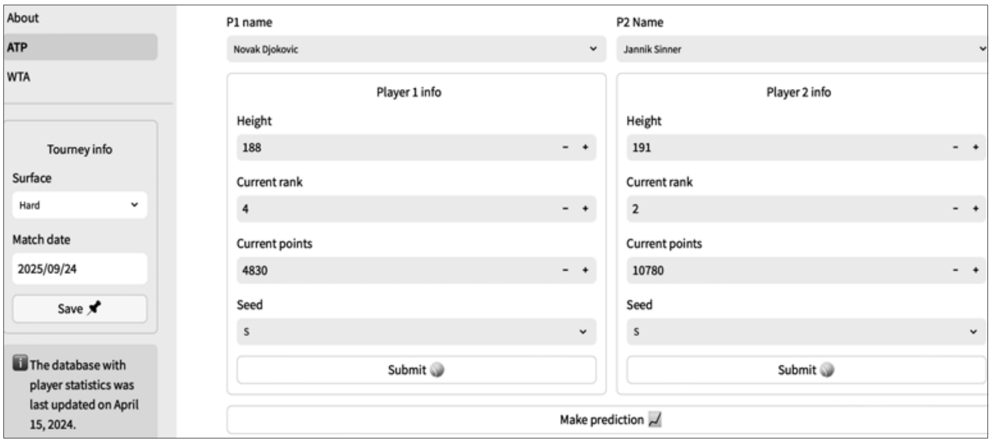

Fig. 1. Prediction page for men's match outcomes

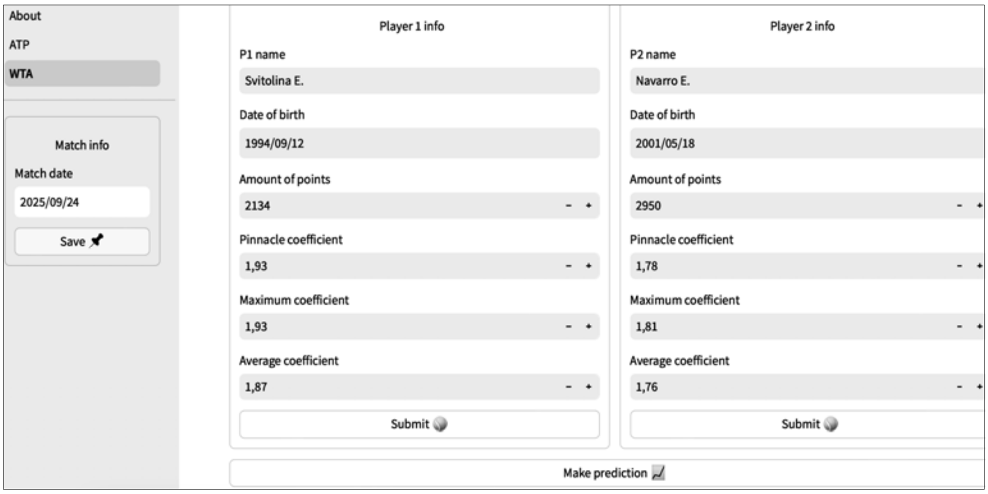

Fig. 2. Prediction page for women's match outcomes

ALGORITHM FOR DETERMINING THE BEST BETTING STRATEGIES

To determine an effective betting strategy, a method based on the ROI (return on investment) metric as the prior indicator of a bettor's success and, accordingly, the target metric of the built predictive model's effectiveness was developed and applied. An important condition is that each bet must be evenly distributed with an identical amount.

Let S_0 represent the bettor's (player's) initial capital.

$$S_0 = s\,t,$$

where s is the amount of a single bet (always the same), and t is the bettor's tolerance for losses, i.e., the number of consecutive bets he is willing to lose before ceasing to follow the strategy. The tolerance is determined by the bettor and can be adjusted during the betting process.

Then, S_i is the player's current capital after the i-th bet has been placed.

$$S_i = S_{i-1} + \mathrm{I}_i\, s\, coef_i - s\ ,$$

where $coef_i$ is the coefficient of the i-th bet, $i = \overline{1,n}$, and n is the total number of bets, while I_i is the indicator of the success of the i-th bet, which is determined as follows:

$$\mathrm{I}_i = \begin{cases} 1, & \text{if the bet won} \\ 0, & \text{otherwise} \end{cases}.$$

Let P_i be the player's profit after calculating the i-th bet:

$$P_i = S_i - S_0\,.$$

Let ROI_i be the percentage of winnings from each bet, calculated after the i-th bet, averaged over the distance:

$$ROI_i = \frac{P_i 100}{s\,i},\ i = \overline{1,n}\,.$$

Then, the betting strategy is considered effective if the following conditions are met:

1. $S_i > s,\ i = \overline{1,n}$. This condition means that the player's current capital must always be greater than the amount of one bet to be able to place it.

2. $ROI_i > 0$ при $i = \underline{h,n}$. Here, n is the total number of bets placed, and h is the minimum number of bets determined as the calculation threshold for profit, which can be adjusted by the player. This condition means that after the h-th bet, the return on investment (ROI) must always be greater than 0, demonstrating the strategy's stability and profitability over the long term.

3. $\sum_{j=0}^{t-1} 1(ROI_{k+j} < ROI_{k+j-1}) < t$,

where $k = \underline{1, n-t+1}$ for $n > t-1$, $ROI_0 = 0$, and $1(\cdot)$ is an indicator defined as follows:

$$1(\cdot) = \{1, \text{ if } ROI_{k+j} < ROI_{k+j-1}\ 0, \text{ otherwise} .$$

This condition means that the player is willing to tolerate no more than t consecutive lost bets (tolerance for loss).

The algorithm for determining an effective strategy using the described method is presented in the form of a flowchart in Fig. 3.

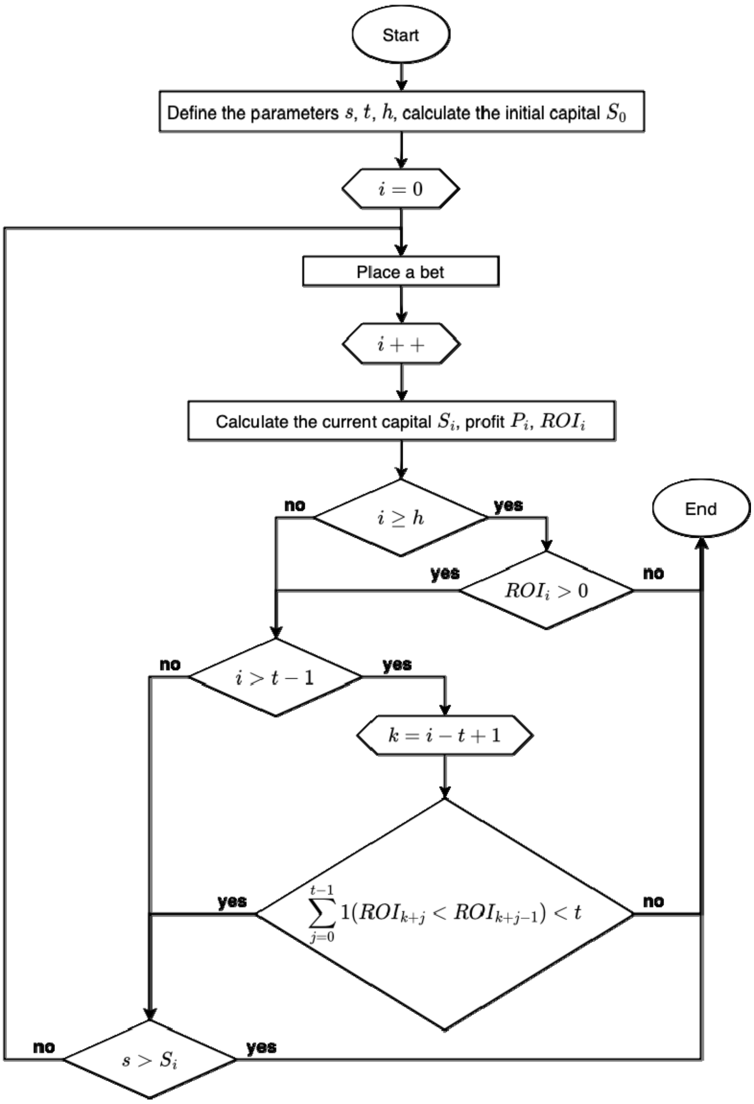

Fig. 3. Flowchart of the algorithm for determining the effectiveness of a betting strategy

DETERMINING THE BEST BETTING STRATEGIES BY DEVELOPED ALGORITHM

New datasets of predicted matches were created using the models and algorithms previously developed to determine potentially successful betting strategies. A sample of 239 predicted matches for women and 346 for men was compiled. The prediction of players' statistics for men was performed in two algorithm variations: for all court surfaces and only for a selected surface. As a result, six predictions were made for each men's match.

A filtering process was applied to the obtained data to exclude specific categories of games, thereby increasing the scope for identifying the most suitable conditions for profitability. For women, filters were considered for the minimum probability of a player's victory and the minimum odds. For men, filters included the current form (i.e., the number of matches played in the last 60 days) and the minimum odds. Tables 11–12 present the number of successful (profitable) strategies for each model.

Table 11. Successful strategies (women)

Model	Number of profitable strategies	Total number of strategies	Percentage of profitable strategies
Logistic regression	44	99	44
Multilayer perceptron	34	99	34

Table 12. Successful strategies (men)

Model	Number of profitable strategies	Total number of strategies	Percentage of profitable strategies
Logistic regression (all surfaces)	24	144	17
Random forest (all surfaces)	33	144	23
XGBoost (all surfaces)	0	144	0
Logistic regression (selected surface)	18	99	18
Random forest (selected surface)	0	99	0
XGBoost (selected surface)	7	99	7

The tables reveal that the XGBoost model with the algorithm for predicting players' statistics across all surfaces and the random forest model with the algorithm for predicting players' statistics on a selected surface did not show any profitable betting strategies.

Tables 13–14 present the most successful and effective strategies for each model with the parameters $s = 100$, $t = 5$, $h = 10$.

Table 13. Most successful effective strategies for the women's division

Model	Minimum probability threshold	Minimum coefficient threshold	Prediction ratio (correct predictions / total predictions)	Percentage of correct predictions	Increase in initial capital (%)	ROI (%)
Logistic regression	0.65	1.35	22/23	96	154.8	33.65
Multilayer perceptron	0.65	1.35	27/28	96	201.8	36.04

T a b l e 14. Most successful effective strategies for the men's division

Model	Minimum number of matches played threshold	Minimum coefficient threshold	Prediction ratio (correct predictions / total predictions)	Percentage of correct predictions	Increase in initial capital (%)	ROI (%)
Logistic regression (all surfaces)	12	1.5	14/24	57	57.4	11.96
Random forest (all surfaces)	12	1.25	19/30	63	158.6	26.43
Logistic regression (selected surface)	8	1.6	12/19	63	82	21.58
XGBoost (selected surface)	8	1.55	14/23	61	104.2	22.48

Based on the results, it can be concluded that the two best strategies for obtaining financial gains from betting on tennis match outcomes are:

- **For women:** the multilayer perceptron model, as its strategy has a higher percentage increase in initial capital and ROI.
- **For men:** the random forest model with the algorithm for predicting players' statistics across all types of courts, as it has the highest ROI and percentage increase in initial capital.

To visualize the change in ROI from betting according to the most successful effective strategies, we generated the graphs shown in Fig. 4 and 5, illustrating the effectiveness of the chosen strategies and models for women's and men's tennis matches, respectively.

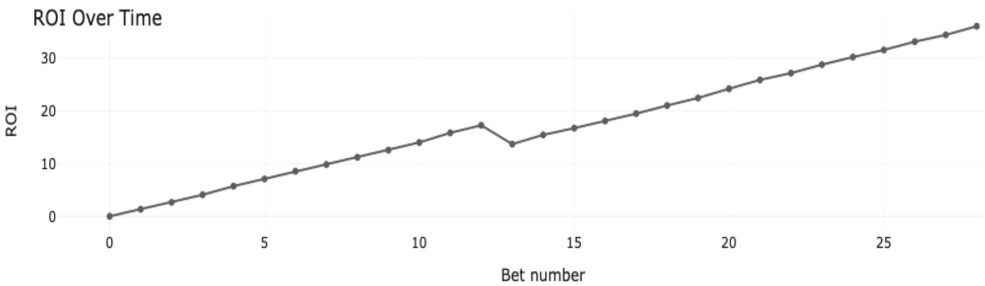

Fig. 4. Change in ROI for the most successful effective strategy based on the multilayer perceptron model for women's tennis matches

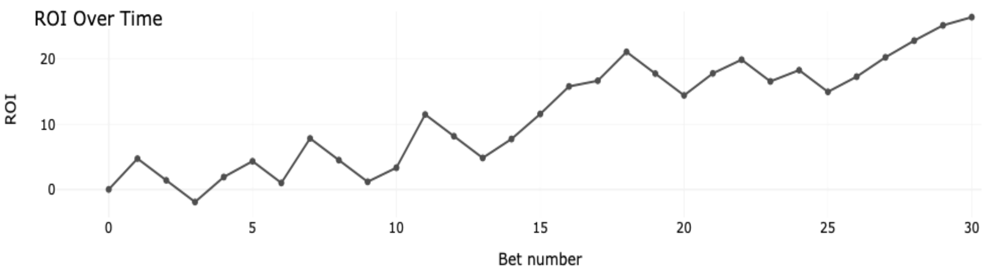

Fig. 5. Change in ROI for the most successful effective strategy based on the random forest model with statistical prediction algorithm across all types of courts for men's tennis matches

CONCLUSIONS

The first part of the conducted study was dedicated to the implementation of various machine learning methods and auxiliary algorithms for predicting the outcomes of tennis matches. The second part aimed to determine the best strategies for obtaining financial benefit from sports betting. For this work the real data both for men's and women's tennis matches were selected, processed and analyzed. Five machine learning models were developed based on logistic regression, multilayer perceptron, random forest, and extreme gradient boosting methods. Men's tennis results forecasting is based on players' statistics as predictor variables. Therefore, an algorithm that uses the time discounting method was applied, enabling the statistics forecasting for future matches based on the player's historical games. Forecasting of outcomes were made on new datasets to determine the best betting strategies. Based on the results obtained, using a filtering algorithm and the developed method for assessing strategy effectiveness, the most successful and effective betting strategies were identified for use in sports betting to maximize user profits.

A web interface was created to facilitate the use of the developed models and provide a clear interpretation of the obtained results. This interface allows users to easily manipulate input data for prediction by entering it via the keyboard or, if necessary, modifying it. In future research, we will focus on studying and using background information received from the key variables as well as modifying and proposing more different strategies for the players based on their attitude and risk tolerance.

REFERENCES

1. Gianluca Di Censo, Paul Delfabbro, Daniel King, "Examining the Role of Sports Betting Marketing in Youth Problem Gambling," *Journal of Gambling Studies*, vol.40, pp. 2005–2025, 2024. doi: http://dx.doi.org/10.1007/s10899-024-10347-x
2. N. Kuznietsova, P. Bidyuk, "Forecasting of Financial Risk Users' Outflow," *IEEE First International Conference on System Analysis & Intelligent Computing (SAIC)*, Kyiv: 08–12 October, 2018, pp. 250–255. Available: https://ieeexplore.ieee.org/abstract/document/8516782
3. A.T. Fernandes, D.B. Figueiredo Filho, E.C. da Rocha, W. da S. Nascimento, "Read this paper if you want to learn logistic regression," *Revista de Sociologia e Política*, vol. 28, no. 74, 2020. doi: https://doi.org/10.1590/1678-987320287406en
4. N.V. Kuznietsova, Z.S. Chernysh, "Regression models application for analysis and forecasting of the financial activity quality indicators of the company," *System Research and Information Technologies*, no. 2, pp. 67–81, 2020. doi: https://doi.org/10.20535/SRIT.2308-8893.2020.2.05
5. M. Brown, *Applied Predictive Analytics*. 2023. [Online]. Available: https://core.ac.uk/download/568285513.pdf
6. C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, Aug. 2020. doi: https://doi.org/10.1007/s10462-020-09896-5
7. K.Y. Chan et al., "Deep neural networks in the cloud: Review, applications, challenges and research directions," *Neurocomputing*, vol. 545, p. 126327, Aug. 2023. doi: https://doi.org/10.1016/j.neucom.2023.126327

8. M. Sipko, W. Knottenbelt, *MEng Computing -Final year project Machine Learning for the Prediction of Professional Tennis Matches*. 2015. Available: https://www.doc.ic.ac.uk/teaching/distinguished-projects/2015/m.sipko.pdf
9. "GitHub - JeffSackmann/tennis_atp: ATP Tennis Rankings, Results, and Stats," *GitHub*. Available: https://github.com/JeffSackmann/tennis_atp
10. "Tennis Betting | Tennis Results | Tennis Odds," *Tennis-data.co.uk*, 2019. Available: http://www.tennis-data.co.uk/alldata.php

Received 24.12.2024

INFORMATION ON THE ARTICLE

Kyryl I. Shum, ORCID: 0009-0009-7503-3249, Educational and Research Institute for Applied System Analysis of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: shumkirillid@gmail.com

Nataliia V. Kuznietsova, ORCID: 0000-0002-1662-1974, Educational and Research Institute for Applied System Analysis of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: natalia-kpi@ukr.net

ПРОГНОЗУВАННЯ РЕЗУЛЬТАТІВ ТЕНІСНИХ МАТЧІВ І АНАЛІЗ ФІНАНСОВИХ ВИГОД / К.І. Шум, Н.В. Кузнєцова

Анотація. Реалізовано програмний продукт, який дозволяє прогнозувати результати тенісних матчів, розроблено метод визначення ефективних стратегій спортивних ставок. Теніс є одним із найпопулярніших видів спорту у світі, який привертає значну увагу як звичайних уболівальників, так і професійних аналітиків. Використання методів машинного навчання дає змогу ефективно прогнозувати результати матчів, що відкриває можливості для отримання прибутку від ставок на ймовірних переможців. Мета дослідження – оцінювання фінансової вигоди від прогнозування результатів тенісних ігор через пошук ефективної стратегії спортивних ставок. Розглянуто різні методи машинного навчання і допоміжні алгоритми та виконується їх порівняння з метою вибору найкращої стратегії укладання ставок задля максимізації потенційного прибутку користувача. Об'єкт дослідження – прогнозування результативності тенісних матчів. Предмет дослідження – моделі, методи машинного навчання та допоміжні алгоритми прогнозування результативності тенісних ігор. Результатом дослідження є визначення найкращої стратегії тенісного беттингу.

Ключові слова: прогнозування результатів тенісних ігор, машинне навчання, спортивний беттинг, стратегії ставок.