Horse Race Prediction Using Machine Learning Algorithms

Puja Chavan[1], Pankaj Kunekar[2], Prajakta Musale[3], Mrunal Shinde[4], Atharava Shinde[5], Nikhil Shinde[6]

puja.cholke@vit.edu[1], pankaj.kunekar@vit.edu[2], prajakta.musale@vit.edu[3], mrunal.shinde2111@vit.edu[4], atharva.shinde21@vit.edu[5], nikhil.shinde212vit.edu[6]
Vishwakarma Institute of Technology, Pune

Abstract. Horse racing has been one of the most enjoyable forms of gambling long since. Winning streaks are difficult to achieve without the right knowledge and tools which makes the betting aspect of this game intimidating. This system predicts the horse with the highest chance of winning under certain circumstances using machine learning technology. For this model, 6348 races and 79447 horse records were considered where a race had 14 horses each. This prevents the better from losing money to the immoral traps set by the bookie or the organization or due to his/her lack of knowledge and placing informed bets.

Keywords: Winning horse, machine learning, python, sci-kit learn, classification.

1 Introduction

Nobody has ever bet enough on a winning horse". For many years, horse race betting has aroused many emotions amongst the audience. Regal horses, excellent sportsmanship, competition and the possibility of winning bets and making fast money attract the public effectively. The beauty of this activity is that anyone can enjoy it without having to learn complicated terms and mechanisms.

To predict a winning horse, we can use machine learning. Thus, to make betting easier for the users our system takes into account different parameters that predict the highest chances of a horse winning a race. For the machine learning code with Python, we used Google Colab as it gives free access to GPUs (Graphics Processing units). Using the previous data, better suggestions could be provided to the users to place informed bets [1]. Horse Race Prediction is a solution for the prediction of the winning horse by taking into consideration different parameters such as the horse weight, jockey weight, horse's winning stats, etc. We are using machine learning, Python, and Scikit Learn. This system helps betters have fruitful returns from their betting pursuits.

As believed by our research and previous study, the relationship between the variation in odds and the winning horse was first done during this study [2]. The following paper discusses the various factors that affect the result of a horse race. A few of them are field conditions, jockey winning stats, previous records, the draw number, etc. There have been various researches conducted. Zanjan et al., to predict the winner of the New York Aqueduct Race used a machine learning technique [7].

2 Literature Review

2.1 Structuring Your Paper

Wai-Chung Chung, et al. [1] proposed a method for predicting Horse racing results in Hong Kong. They got their data set from Hong Kong jockey Club and using it for training SVM Based Committee Machine. Also, the accuracy of the prediction and the return rate was considered. The correct prediction ratio was found to be 7/10 for the Horse to win the Race. As a result, this method proved to be better than random forest and SVM in correctly predicting the outcome of the Race. For Hong Kong Horse racing the accuracy was high with 70.86%. Nishchay Nemdharry, et al. [2] predict the horse race winner at the Champ de Mars, which is organized by The Mauritius Turf Club (MTC). Collected around eight odds at different intervals. The module was trained using 232 races and tested on 27 races. This was the first study on the connection between odds and the rank of horses. There are many magazines and portals in Mauritius which provide plenty of information about horses. They tell us how we can also make a module or artificial network to guess the winner of horse races using the parameters such as jockey's weight and experience, stability, weighing capacity of the horse, age of the horse. But the result was around 7.4% so they came to know the outcome of horse races at Champs de Mars does not depend on horse racing odds.

Hsinchun Chen, et al. [3] highlight the importance of machine learning which involves using a computer algorithm to capture knowledge from data. A comparison is being made between three different human tracks using the machine learning technique. The author investigates a different problem-solving situation game-planning variable from eight competing dogs in the race. Approximately, 112 races were conducted every week. It was assumed that the behavioral capabilities of the participant were nearly stable which may improve or humble down. Further, a fair game is expected to be free from all the other external human factors. Amit Kumar Jain, et al. [4] proposed an approach for accurate RUL (Remaining Useful Life) prediction of milling cutters that are used in the industry such as CNC and VMC milling machines based on an Artificial Neural Network (ANN). For this model, they are considering various statistical features such as Average force, Standard deviation, Skewness, Amplitude of force, etc. for its

selection they are using stepwise regression. Statistics for including and excluding parameters in the model. This is a very effective technique as suggested by the authors. And also, for the training of the ANN model, a MATLAB neural network toolbox is used. As the cutter reaches the end of its life, the model successfully predicts the failures in impeding the milling cutters and allows us to take precautionary actions at the appropriate time.

Ayush Verma, et al. [5] designed a model that accurately predicts the prices of houses. According to customers' needs and requirements, they first analyze various parameters and used them to find an ideal price for the house. Also, their dataset contains all the required parameters that one needs to predict the house price. Using machine learning for prediction and analysis of the result they got, for this, they used linear regression, forest regression and Boosted regression for prediction. Further accuracy of the model can be increased by implementing Neural Network. Gaurav Meena, et al. [6] are predicting traffic flow information based on the factors affecting smooth traffic flow such as signals, accidents, repairing of roads or rallies and gives better way or path to follow for saving time. which also reduces major problems such as accidents or traffic jams. In this project, they used Machine learning, soft computing and some deep learning algorithms for analyzing the data of transportation systems. Also used Image Processing Algorithms for the recognition of traffic signals for more accuracy. In future scope, they further add that they are focusing on deep learning and genetic algorithm which are important in data analysis and gives higher accuracy than the existing project. Also, planes to make web server or an application. Some paper used the combination of machine learning and deep learning classifier to predict the seizure [11, 12].

Nisha Srinivas, et al. [7] paper predicts the age, race and gender of an individual using the CNN. The new and unique data set used was Wild East Asian Face Dataset (WEAFD). East Asian population's Statistics were predicted using the same data set. The dataset consisted of images captured in different environments. Gender prediction was a binary classification problem. Whereas, age and Race was Multi-class classification problem. CNN was used in order to predict Statistics about the population using a data set. Predicting the Race of one individual was the most challenging part, followed by the prediction of age and gender. Using different network designs and different methods of data augmentation the performance of the model can be significantly increased. Malay K. Ganai, et al. [8] puts forward an effective data race prediction method that has used lock recording-based incremental search on time-stamped locked histories. Record locking is the technique to prevent simultaneous access to data in a database, to prevent inconsistent results. This method can be easily understood with the 2 Bank clerk's example. Using this technique, the data is covered

properly and the prediction accuracy also increases. A specially designed algorithm is used that can store and reuse past search results. Thus, the cost of reasoning useful data and overlapping searching is reduced. This algorithm can handle hundreds and thousands of different events and predict known-unknown results.

3 Methodology

3.1 Flow of Proposed System

Fig.1. Flow of the Propsed System

3.2 Method

Fig. 1 shows the flow of the propsed system.
Step 1: Importing all libraries
The required libraries such as NumPy (easy working with arrays/multidimensional arrays, mathematical computations, open-source and easy to use), Pandas (fast and efficient for manipulating and analyzing data, easy handling of missing data), Matplotlib (data visualization and graphical plotting), Seaborn (It will be used to visualize random distributions), Scikit Learn (efficient tool for machine learning and includes models for classification, regression, clustering, etc.) are to be imported.

Step 2: Data set installation
https://www.kaggle.com/datasets/gdaley/hkracing
The horse racing data set from Hong Kong used for this project is downloaded from Kaggle. This data set has two files, the first being races.csv and the latter as runs.csv. The races.csv file has parameters such as race id, venue, the distance of the race, winning prize for the race also the type of race track surface and its condition, etc. The runs.csv file contains horse age, declared weight of the horse as well as the jockey, win odds, draw, won, etc.

Step 3: Data pre-processing

The two data set files, races.csv and runs.csv were merged based on the race-id column so the variables make meaningful data. The parameters chosen for this study did not have any missing values. The venue and going columns in the dataset had string inputs which were converted into numerical data by using an encoding feature provided by the Scikit-Learn library. A label encoder was used for the race venue and an ordinal encoder was used for going.

Step 4: Training the model

ML algorithm was trained by feeding the dataset in this step. The data set selected will be divided into 80% training and the 20% testing data set randomly and the training data set will be used to train the model. The accuracy of different classification models has been tested including dummy classifier, logistic regression, decision tree, random forest and support vector machine (SVM). Fig. 3 shows the architecture of proposed model.

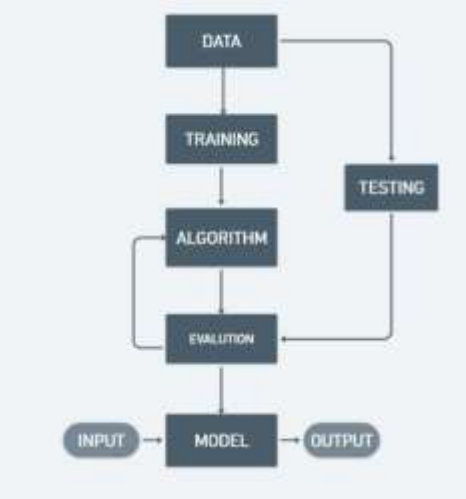

Fig. 3. Architecture

Step 5: Testing and checking the output
The testing part of the data set was used to test the ML model by supplying the testing data set as an input to the model. After testing, the accuracy of the classification model was compared.

4 Result and Discussions

The ML model was able to predict the winning horse from a race based on the data set that was used. From the predicted results we also came to know which factors were the most important for winning and should be considered before placing any bets. Fig. 4 shows the count plot of Won.

As the dataset had more "not winning" horses than the "winning" horses, the 'Dummy Classifier' always predicted "not winning" horses which gave an accuracy of 92% without actually having trained the model. Table 1 shows the comparion table.

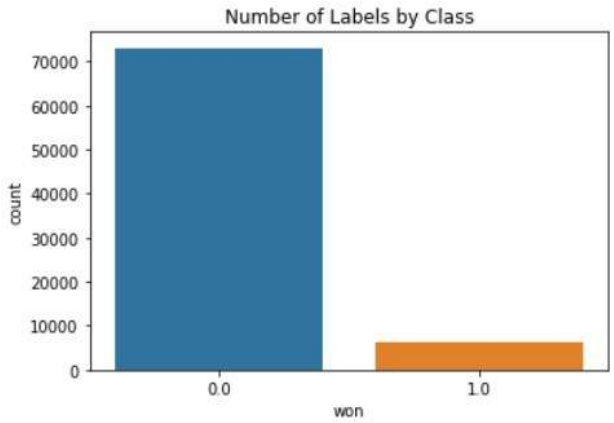

Fig. 4. Count plot: Won

Table 1. Comparison Table

Classification Model	Accuracy	Precision	Recall	F1 score
Dummy Classifier	0.92	0.46	0.50	0.48
Logistic Regression	0.92	0.46	0.50	0.48
Decision Tree	0.85	0.55	0.55	0.55
Random Forest	0.92	0.62	0.52	0.51

Support Vector Machine	0.92	0.46	0.50	0.48

For “Logistic Regression” which is the trained classifier, the accuracy was found to be the same as that of the dummy classifier. Thus, for comparison of this classification problem, accuracy was not a suitable option. A basic model that always predicted "not win" was seen to have great accuracy, reducing the accuracy of predicting a win. Precision, Recall and F1 score could have been used for comparison between models. For “Decision Tree”, the accuracy was found to be 85% with a precision of 55% as compared to a precision of 46% for the Dummy Classifier and Logistic Regression. For “Random Forest”, accuracy was 92% along with a high precision of 62% and for “Support Vector Machine” (SVM) the accuracy and precision were found to be the same as that of Logistic Regression.

After comparing all classification models “Random Forest” was found to be the best algorithm for this classification problem because of its high precision and high accuracy. Fig. 5 shows the Confusion Matrix.

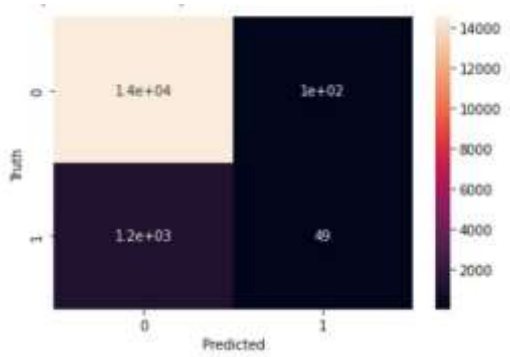

Fig. 5. Confusion Matrix

The results from this particular study can be used effectively in various other important fields such as education, finance, industry, medicine, management, games and businesses.

5 Future Scope

Hosting the model on the web along with creating a safe betting platform for the users. More accuracy on the prediction. Implementation of varied sports with a larger dataset of different places. The way ML has been implemented for predicting the winner of a horse race, it can be used similarly for various sports.

6 Conclusion

In this research, an ML model is created for horse race winner prediction using python and Sklearn. We developed a horse race prediction for predicting which horse will win. Training the model is the first half of our model and testing the trained model is as follow. Thus, the system can be used for real-time prediction of the winning horse.

References

1. Wai-Chung Chung, Chuan-Yu Chang Chien-Chuan Ko, "An SVM Based Committee for prediction of Hong-Kong Horse Racing", 2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media).
2. Nishchay Nemdharry, Trivartsingh Ramjeawon, Harrykesh Ramma and Ritesh Mungroo, "Impact of the Variation of Horse Racing Odds on the outcome of Horse Races at the Champ de Mars", University Of Mauritius, 2017.
3. Hsinchun Chen, Peter Buntin Rindc, Linlin She, Siunie Sutjahjo, Chris Sommer, Daryl Neely, University of Arizona, "Expert Prediction Symbolic Learning and Neural Working an Experiment on Greyhound Racing"
4. Amit Kumar Jain and Bhupesh Kumar Lad, "Speed Milling Cutters based on Artificial Neural Network", International Conference on Robotics, Automation, Control and Embedded Systems – RACE 2015 18-20 February 2015, Hindustan University, Chennai, India.
5. Ayush Verma, Abhijit Sharma, Sagar Doshi and Rohini Nair, "House Price Prediction Using Machine Learning and Neural Networks", International Conference on Inventive Communication And Computational Technologies, 2018
6. Gaurav Meena, Deepanjali Sharma, Mehul Mahrishi. "Traffic Prediction for Intelligent Transportation System using Machine Learning", 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE-2020), 07-08 February 2020, (IEEE Conference Record 48199).
7. Nisha Srinivas, Harleen Atwal, Derek C. Rose, Gayathri Mahalingam, Karl Ricanek, Jr. and David S. Bolme, "Age, Gender, and Fine-Grained Ethnicity Prediction using Convolutional Neural Networks for the East Asian Face Dataset", 2017 IEEE 12th International Conference on Automatic Face & Gesture Recognition.
8. Malay K. Ganai, "Efficient Data Race Prediction with Incremental Reasoning on Time-Stamped Lock History".
9. Giriesh Hegde, Vishwanath R Hulipalled, J.B. Simha, "A Study on Agriculture Commodities Price Prediction and Forecasting".
10. Gabriel Rushin, Cody Stancil, Muyang Sun, Stephen Adams, Peter Beling, "Horse Race Analysis in Credit Card Fraud—Deep Learning, Logistic Regression, and Gradient Boosted Tree", University of Virginia.
11. Puja A. Chavan, Sharmishta Desai, "Effective Epileptic Seizure Detection by Classifying Focal and Non-focal EEG Signals using Human Learning Optimization-based Hidden Markov Model", Biomedical Signal Processing and Control, Volume 83, 2023, 104682, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2023.104682.
12. Puja Chavan, Dr. Sharmishta Desai, "A Review on BCI Emotions Classification for EEG Signals Using Deep Learning", Recent Trends in Intensive Computing, ADVANCES IN PARALLEL COMPUTING" IOS Press Publisher, December 2021, pp. 544-551, doi:10.3233/ APC210241.