International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8119
Article Received: 10 July 2023 Revised: 30 August 2023 Accepted: 18 September 2023

Predicting Outcomes of Horse Racing using Machine Learning

Meenakshi Gupta*, Latika Singh
School of Engineering & Technology,
Sushant University, Gurgaon - 122018, Haryana,
E-mail:* meenakshi78gupta@gmail.com

Abstract: Machine learning with its vast framework is making its way into every aspect of modern society. The segment of betting sports particularly horse racing calls for the attention from a large spectrum of research community owing to its value to the stakeholders and the amount of money involved. Horse racing prediction is a complex problem as there are a large number of influencing variables. The present study aims to contribute in this domain by training machine learning algorithms for predicting horse racing results or outcomes. For this, data for a whole racing season from 2017 to 2019 of races conducted by Turf Club of India was considered which amounts to over 14,700 races. Six algorithms namely Logistic Regression, Random Forest, Naive Bayes, and k-Nearest Neighbors) k-NN were used to predict the winning horse for each race. Synthetic Minority Oversampling Technique (SMOTE) technique was applied to the imbalanced horse racing data set and the attributes of the horse race repository were analyzed. The results were compared with other sampling methods to evaluate the relative effectiveness of this method. The proposed framework is able to give an accuracy of 97.6% which is substantially higher when compared to other similar studies. The research can be beneficial to the stakeholders as well as researchers in the same area to do further analysis and experiments.

Keywords: Machine learning, imbalanced data, SMOTE, prediction, classification model, sports betting, horse racing.

I. INTRODUCTION

Due to the betting aspect and the volatility of racing, horse racing has been one of the most exciting and entertaining sports. The 2022 Grand National (UK) recorded 21% increase in betting volume from 2019 to reach a total trading sum of £92.8 million with over 50 million bets placed through its online platform[1] .For the 2023 Grand National event, over £1,000,000 is allocated in prizes alone at United Kingdom with 600 million people watching in over 140 countries and more than 3.5 billion US dollars in the United States [2]. The global online gambling market size was valued at USD 57.54 billion in 2021 and is expected to expand at a compound annual growth rate (CAGR) of 11.7% % from 2022 to 2030 [3]. Horse racing is a business that is primarily supported by betting on horses. Prediction in horse racing has long been considered as one of the research problem. It is a challenging problem because of numerous qualitative and quantitative variables. In the present research study, it is proposed to use Machine Learning (ML) algorithms to forecast the outcome of the horse races. In their study Allinson and Merritt[4]discussed horse racing prediction using neural networks based on multilayer perceptrons. Their model considered 200 horses of two years of age only,so there is a scope for more work in this direction by taking other age groups as well. In another study Hei et al[5] used two methods namely, Hope and Resheff[6]suggested a combined method of TensorFlow with Voting system to predict winner of the races accurately. The accuracy of their predictive model was 49% which shows a scope for improvement in this area. Schumaker and Johnson[7] used Support Vector Regression using sequential minimal optimization function in Weka to assess accuracy, precision, and nature of betting in Grey Hound racing which is quiet similar to horse racing as both involve similar uncertainties. Their methodology was adapted for discrete numeric prediction instead of classification and included a dataset of 1953 races that spanned 31 different race tracks. Another study by Williams and Li[8] applied four neural network algorithms and gave the best accuracy of 74% with Back-Propagation algorithm but it needed a longer training time and more parameter selections in predicting horse races in Jamaica using a dataset of 143 Jamaican horse races. Slightly better results were obtained by Davoodi and Khanteymoori[9] as they applied five different supervised neural network algorithms and gave an average prediction accuracy of 77% on a dataset of 100 races with a trade-off between more training time and higher accuracy. In his study Silverman[10] has performed a hierarchical Bayesian study of thoroughbred horse racing and identified which horses had the greatest speed. An analysis of the 36,006 observations from the 2973 different horse races on horse tracks in Hong Kong using Bayesian modeling was done and all of the horses' running speeds were estimated and the accuracy of 21.6% was achieved. However, machine learning approaches are more efficient when compared to this methodology of finding the greatest speed and predicting the winner of the race. The authors Padurath et al[11] assigned weights to the factors affecting a horse race and then developed a software in Java to make predictions about potential winners. The professional expert had predicted less (44%) than their statistical software approach(58%). But statistical approaches seemed to be less efficient compared to many ML classifiers like Logistic Regression classifier, Random Forest and K-NN models. In a study by Silverman and Suchard[12], two new approaches were introduced to the Conditional Logistic Regression in order to boost the system's efficiency: regularization parameters and frailty parameters. The historical data was gathered from 3681 races and it was shown that their mathematical model using a Graphical Processing Unit (GPU) was better than openMP solution by a factor of 2.68. In his research, Takahashi [13] investigated how age effects thoroughbred racehorse speed performance. As was reported in this study, horse begin to increase their average speed at an early half of four years and then their average speed begins to decline slightly at around latter half of four years and continues at that level for the rest of their adolescence. Only age was taken into consideration to assess the performance of racehorses' win-ability but there can be so many different attributes that may affect the outcome of a race which leaves a scope for further study. In the study by Bunker and Thabtah[14] examined ML approaches and Artificial Neural Network (ANN) frameworks in analyzing and predicting the outcomes of sports, in general and therefore asserts the usefulness of ML in sports predictions. Another ML based prediction analysis is done by Schumaker [15] for grey-hound racing. Their method is based on Support Vector Regression (SVR) which uses discrete values rather than classes and their system gave a best accuracy of 50.44% which indicates a scope of improvement.

In order to resolve the issues raised in the aforementioned studies, the current study has attempted to develop prediction models that takes into consideration significant aspects of horse racing domain by applying ML modeling techniques.

The organization of paper is in 6 Sections. In section 1, a summary of research studies already reported for prediction in sports with emphasis on Horse Racing Dataset is provided. In section 2, a description of Horse Race dataset with all its features and constraints used for the study is provided. Section 3 describes the methodology adopted along with explanation of exploratory data analysis applied in this research. In section 4, the evaluation metrics that are considered in the study are discussed. Section 5 gives the details of results which include performance analysis along with the relevant Figures to support the findings. Finally, in section 6 conclusion and scope of future work is discussed.

II. DATASET AND PREPROCESSING

The dataset for this study was obtained from a live web site of a well known Horse Racing Company in India[16]. The data was in excel (csv) format and included 14,750 rows with 56 attributes. After identifying and eliminating rows and attributes with numerical ids and those which had no contribution towards racing analysis, namely, 'jursey_id', 'jursey_No', 'rail_id', 'horse_career_earnings', 'cup', 'race_day_no', 'day_race_no', 'second', 'third', 'forth', 'fifth', 'sixth', 'total', 'net_dist', 'race_class', 'rail_id' final 23 attributes (22 + 1 target variable) including the target variable 'position' corresponding to the race winner ('position'=0, 'position'=1) were finalized for ML analysis [Table-1].

Table 1.Major attributes of Horse Race dataset.

S. No.	Attributes	Description	Type of Data	Considered
1	Position	Position of horse in that race	Categorical	Yes
2	horse_seq	sequence of horse in a race	Nominal	Yes
3	horse_name	name of the horse	Nominal	No
4	age	age of horse	Continuous	Yes
5	trainer	name of trainer	Nominal	No
6	jockey	name of jockey	Nominal	no
7	weight	weight of jockey in Kg	Continuous	yes
8	allowance	allowance in weight limit for the jockey	Continuous	yes
9	draw	number on the stall door from where the horse starts its race	Continuous	yes
10	shoe	Aluminium(A) or Steel (S)	Categorical	yes
11	won_by	distance from the horse preceding this horse	Continuous	no
12	net_dist	distance from the winner horse	Continuous	no
13	horse_rating	rating of horse given by the club after the race on the basis of that days performance	Continuous	no
14	Odds	It reflects the amount of money bet on a horse; the more money that is invested, the shorter the odds (Amount of profit returned to the amount invested)	Continuous	no

15	finish_time	finish time in min: sec: millisec	Continuous	no
16	horse_ex_name	indentifying name of horse prior to current name	Nominal	no
17	horse_pedgree	name of breed of the horse	Nominal	no
18	Sex	sex of the horse	Categorical	yes
19	race_no_id	id of the race	Nominal	yes
20	race_no	race number of that day as in race_no_id	Nominal	no
21	race_date	date of race	Nominal	no
22	horse_rating_prev	rating of horse given by the club before the race on the basis of past performance	Continuous	No
23	horse_rating_after	rating of horse given by the club after the race on the basis of that days performance	Continuous	No
24	horse_running_position	horse rank at various distances in that race	Nominal	No
25	horse_eqip_1	equipments of horse	Categorical	no
26	race_eqip_2	equipments of horse especially for race	Categorical	no
27	race_name	name of race as declared by club	Nominal	no
28	distance	distanceof race in metres	Categorical	yes
29	prize	prize money in rupees	Continuous	no
30	race_class	one of the predefined classes	Categorical	no
31	race_fav_horse	favorite horse of the Race with minimum odds	Nominal	yes
32	race_incident	famous incident of the race as recorded in the club race listing	Nominal	no
33	penetrometer	Numercial value from one special instrument to suggest the track condition on the race day; lower value suggest hard surface and higer value suggest soft surface	Continuous	yes
34	track	Going Good, Going soft, Going firm, yielding	Categorical	yes
35	railing	(in mtr) the effect of false rails on tracks (with the help of rodometer) and rails are put so that there is no difference in actual distance covered on different tracks as the case may be because of oval nature of tracks.	Continuous	no
36	race_centre	city center of the race	Categorical	no
37	season	Season : Monsoon, regular, winter	Categorical	yes
38	club_name	Name of the Club	Categorical	yes
39	horse_id	horse id	Nominal	yes
40	trainer_id	trainer id	Nominal	yes
41	jockey_id	jockey id	Nominal	yes
42	Color	color of horse (9 values)	Categorical	yes
43	Dam	horse mother	Nominal	yes
44	Sire	horse father	Nominal	yes
45	owner_id	ID of the owner of the horse	Nominal	yes
46	body_weight	Weight of horse in Kg	Continuous	yes
47	jursey_id	Id of jockey jursey	Nominal	No
48	jursey_No	Number displayed on jockey jursey	Nominal	No
49	horse_career_earnings	total earnings of that horse	Continuous	No
50	Cup	name of the race cup	Nominal	No
51	Second	amount in rupees awarded for second place	Continuous	No
52	Third	amount in rupees awarded for third place	Continuous	No

53	Forth	amount in rupees awarded for forth place	Continuous	No
54	Fifth	amount in rupees awarded for fifth place	Continuous	No
55	Sixth	amount in rupees awarded for sixth place	Continuous	No
56	Total	amount in rupees awarded for total race	Continuous	No

Also it was observed, that the data obtained was imbalanced (as mentioned in Table 2). Of the total 14,750 records, 13,179 records are for 'position'=0 (no win, majority class) and 1571 for 'position'=1 (win, minority class). After a stratified 70-30 train-test split of the binary class is done, 9229(70-30 split) and 1096(70-30 split) records are achieved for majority and minority class respectively (Table 2 above), that is minority class is just 11.8 % of the majority class thus emphasizing the skewness. To counter the problem of imbalanced dataset, data sampling method like oversampling, SMOTE technique by Huessein et al[17], Mishra et al [18] were applied and analysis was done further. First, an oversampling technique called random oversampling is applied as it balances the data by replicating the minority class samples, and so the number of samples in majority and minority class was 9229 and 6460 respectively (Table 2 below). After applying SMOTE sampling technique, the number of elements in majority and minority class was almost equal at 9245 and 9205 respectively(Table 2 below) and therefore this dataset was used for training of ML models.

III. METHODOLOGY

As mentioned in the previous section, it was essential to take care of the class imbalance problem which might lead to skewed results as difference in number of elements in balanced and imbalanced classes can definitely lead to biased results. A preprocessing technique called SMOTE (Synthetic Minority Oversampling Technique) was used to address the problem of imbalanced horse-racing data. SMOTE is a type of oversampling which uses k nearest nodes algorithmically to build the artificial data, and k=5 was used for balancing [19]. The SMOTE algorithm utilises the k-NN method to generate new samples by setting the minority class as set M. For each data point e ∈ M, the k nearest neighbors were determined. Then using one of these k neighbors (e_k), the vector between e and e_k was determined and multiplied by a random number between 0 and 1. The synthetic data point (e') was then obtained by adding it to the value of e. Thus new synthetic data are created with the help of interpolation between the positive instances that are near each other and a new minority class set M1 equivalent to majority class was constructed. The formula outlined as follows (Equation 1) explains how the new samples were generated. Here rnd (0,1) represents random numbers, with values between 0 and 1 as illustrated by Shrivastava et al[20]

$$e' = e + rnd\,(0, 1) * |e - e_k| \quad (1)$$

Table 2 compiles the distribution model of training dataset (of 70-30 train-test split) before and after applying the sampling techniques like Over Sampling and SMOTE technique as discussed earlier. Stratified cross-validation[21] of k fold with split=5 was applied and results are compared in Section 5.

Table 2. Distribution Model of Dataset before and after applying Sampling Techniques

Class Distribution Model for Training Data (70%)	Number of Minority class elements (70% train data)	Number of Majority class elements (70% train data)	%Ratio of Minority & Majority class to show skewness
Imbalanced Dataset(Original)	1096	9229	11.8%
Dataset after Over Sampling (using RandomOverSampler method)	6460	9229	69.8%
Dataset after SMOTE Technique	9205	9245	99.56%

Following ML classification algorithms are considered in this study to learn from given horse racing dataset.

3.1. Linear Regression

Linear Regression is a classification algorithm used to predict binary outcomes for categorical variables like 0 /1 or Yes/No given a set of independent variables as input and output [22]. The equation for Linear Regression is

$$\log[z/(1-z)] = a_0+a_1x_1+a_2x_2+...a_nx_n \quad (2)$$

where z is dependent variable (predicted output), x_1, x_2,..,x_n are independent (input) variables and a_0, a_1,a_2,...,a_n are constant coefficients. This equation gives the outcome of classification problem in the form of probability range from 0 to 1.

3.2.k-NN (k-Nearest Neighbours)

k-NN algorithm captures the idea of proximity. It is a type of supervised machine learning algorithm used to solve classification problems by determining the nearest data point based on the shortest distance amongst the (k) neighboring data

***Article Received:** 10 July 2023 **Revised:** 30 August 2023 **Accepted:** 18 September 2023*

points [22]. Dataset is separated into test data and training data and the category having the most frequency of data points is voted for.

3.3.Naive-Bayes

The Naive Bayes model works on the premise that the existence of a particular feature of a class is independent or unrelated to the existence of every other feature [23]. Bayes theorem gives for P($c|x$: how often c happens when x happens) from P(c: probability of c), P(x: probability of x) and P($x|c$: probability of x happening when c happens) calculation of the latter probability [24] as shown in equation 3

$$P(c|X) = (P(X|c)P(c)/P(X)) \qquad (3)$$

Here $X=(x_1,x_2,x_3,....,x_n)$ represent the features of the dataset

Replacing the denominator (because it remains unchanged) by proportionality, we get

$$P(c|x_{1,..,}x_n)\ \alpha\ P(c)\prod_{i=1}^{n} P(x_i|c) \qquad (4)$$

The equation to find the class with maximum probablity becomes

$$c = \max_c[P(c)*\prod_{i=1}^{n} P(x_i|c)] \qquad (5)$$

where $\max_c$ is a method to get the maximum value of the argument from a target function.

3.4.Random Forest

It is based on the concept of Decision trees and gives the prediction outcome based on the average of the output from its previous trees [25]. A subset of trees with randomly selected features is created at the split node.

To analyze the performance of ML models applied, various evaluation metrics are used as discussed in section 4 ahead.

IV. EVALUATION METRICS

To assess the performance of our prediction system, following three metrics are used:

(i) Accuracy Score (Accu) : It is the measure of performance model given in percentage.

$$\text{Accuracy} = \frac{(TrP+TrN)}{((TrP+FsP)+(TrN+FsN)} \qquad (6)$$

where, TrP= number of True Positives, TrN= number of True Negative, FsP= number of False Positives, FsN = number of False Negatives

(ii) F1 score : It is a performance metric for classification system especially useful for imbalanced class structure and is calculated using Equation 5.

$$F1 = \frac{2}{precision^{-1}+recall^{-1}} \qquad (7)$$

where precision $=\frac{TrP}{TrP+FsP}$ and recall$=\frac{TrP}{TrP+FsN}$

Here precision is defined as the actual correct prediction divided by total prediction made by model. Recall is calculated as the number of true positives divided by the total number of true positives and false negatives. A high F1 score will mean that both precision and recall are high and is therefore a desirable metric for the performance of the ML model.

(iii) The Receiver Operator Characteristic (ROC) : It is a (Probability) curve with range from 0 to 1. In this True Positive Rate (TPR) is plotted against False Positive Rate (FPR) .

TPR $=\frac{TrP}{TrP+FsN}$ and FPR $=\frac{FsP}{FsP+TrN}$

(iv) Area Under the Curve(AUC) : It is used to determine the performance of various ML models by viewing the Area under the ROC curve (AUC) for them. A higher value of AUC is desired as it signifies that the classifier distinguishes well between the positive class values and negative class values.

For this study, Python 3.8.3 is used to apply machine learning methodology. After loading and data preprocessing, the dataset is split into train and test data (70-30 split) using a random number so as to ensure random sample representation of the original problem dataset. Then different ML classifier models are applied and their performance efficiency is tested by using metrics as discussed earlier. The results are discussed in the following next section.

V. RESULTS

Exploratory data analysis was performed to fully comprehend the data before ML algorithms were applied. According to traditional subject experts, below mentioned features are considered important to predict the winner of the horse race.

FAVORITE HORSE

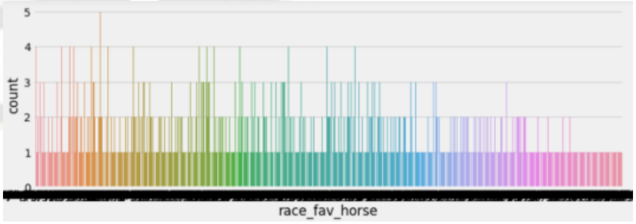


Fig. (1). Favorite horse

Favorite horse(race_fav_horse) is declared for every race by the respective race club based on its previous ratings. As per our analysis (Fig.(1)) 'Tapi' was declared as favorite 41 times but could actually win 5 times, followed by Paso Robles (5 out of 37 times).

JOCKEY

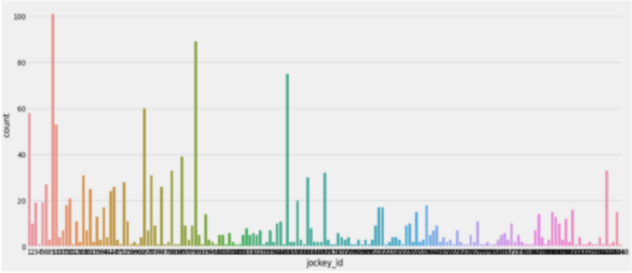


Fig. (2). Jockey ID

According to our dataset, Suraj Narredu (Jockey_Id=9) is a champion Jockey for the season in consideration and can be a deciding factor for placing bets for the winner in coming seasons as well (**Fig. (2)**).

TRAINER

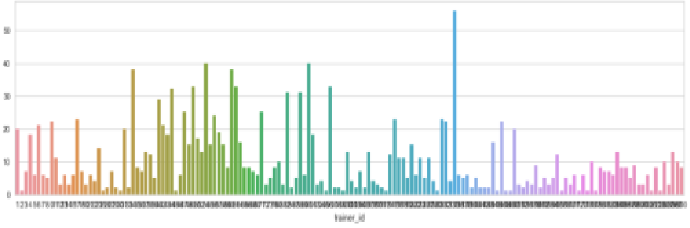


Fig. (3). Trainer ID

Trainer Vijay Singh-trainer_id:135 has recorded maximum wins(56) out of total 1571, followed by Magan Singh Jodha-trainer_id:90 and R H Sequiera- trainer_id:54 both with 40 wins each (**Fig. (3)**).

DAM, SIRE

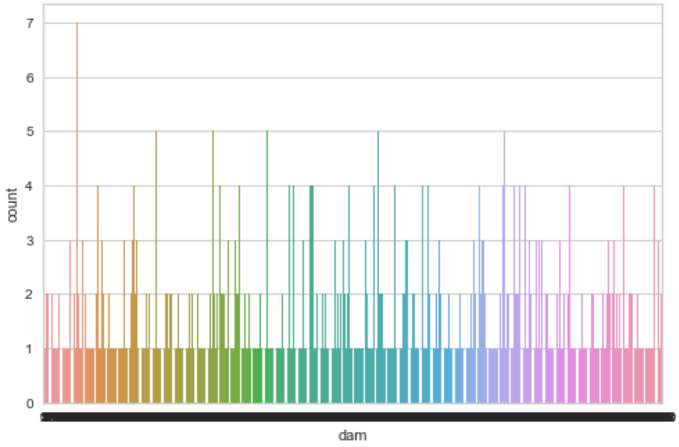


Fig. (4). Dam Id

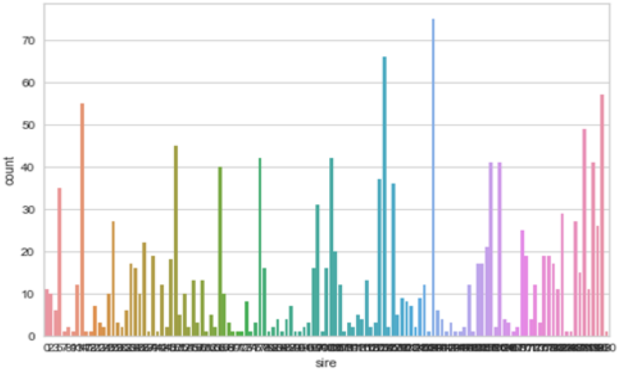


Fig. (5). Sire Id

Knowledge about the sire and dam are significant factors as each play a different role in determining how the racehorse will run. A male horse is tagged a "sire" after one of his offspring wins a registered race and determine the distance and surface where the runner will be most effective (underfoot conditions) while the female family (dam) determine quality of run of the foal [26],[27]. A stakeholder can pick their favorite horses with respect to speed and stamina by identifying the different pedigree of sire and dam. However, much attention is paid to the sire line[28]. As shown in our analysis (Figure4, Figure 5), sire - Phoenix Tower has appeared 75 times for the winning horse as compared to dam - Mink Mitten appearing 7 times for the winning position.

OWNER

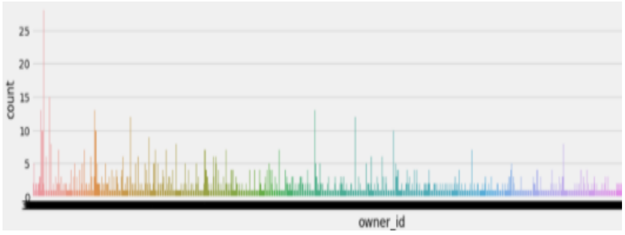


Fig. (6). Owner Id

As is apparent from the graph analysis above (Figure 6), owner _id 3558 (DR M.A.M. Ramaswamy, Chennai) corresponds to the maximum number or winning positions in our data. He was actually considered the Racing Moghul of India and has many national and international records to his credit[29].

BODY WEIGHT

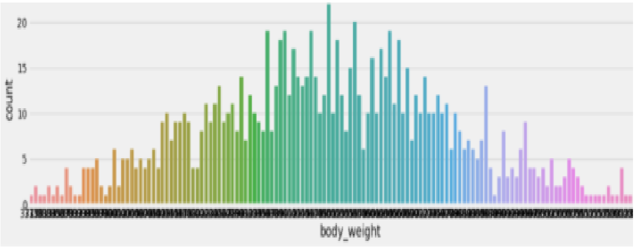


Fig. (7). Body Weight

Article Received: *10 July 2023* ***Revised:*** *30 August 2023* ***Accepted:*** *18 September 2023*

A thoroughbred horse that's primarily used for racing would be expected to weigh in at around 500kg on average, ranging from about 400kg to roughly 600kg. As shown in Figure 7 above the maximum number of wins are recorded for range of 440 kg to 450 kg (101 out of toal 147 wins). So this tells us about the desired optimum body weight of the competing horses for betting which can prove very benefitting for the stakeholders.

WEIGHT

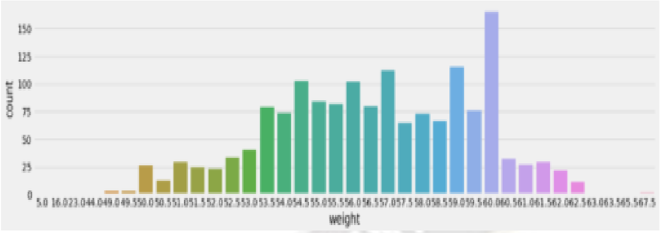


Fig. (8). Weight

This is the weight of the jockey in the race. Each horse in a race has to carry a certain amount of weight that is assigned on the basis of their gender, age and past performance. For fair competition all jockeys must weigh out before and after a race to make sure they (including the saddle) are the right weight. If a jockey is lighter than the weight the horse has to carry, the difference in weight will be made up by thin lead weights in a special saddle cloth. The less weight a horse carry, more are his chances of winning. So the jockey with the closest weight so as to match the assigned weight has more chances of winning. According to subject experts, stakeholders want the jockeys to be at 60 kg weight because adding dead weight of lead saddle (in case of jockey weighing less) is less preferable than having the same weight skilled jockey. This is confirmed by our analysis (FIG. (8)) showing maximum (165 wins) with jockeys of weight 60 Kg followed by 59 Kg (115 wins).

SHOE

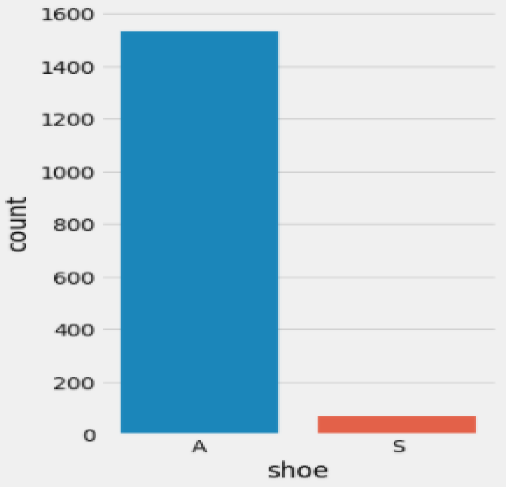


Fig. (9). Shoe

Steel shoes provide comfort and support but at the expense of speed whereas aluminum is as sturdy as steel and gives a boost to the speed also[30]. Horse owners in India prefer aluminum shoes for their horses[30]. Our graph analysis in Figure 9 also supports the psychology of stakeholders as Aluminum shoe horses have won 1532 times out of total 1571 wins in dataset even though the ratio of aluminium to steel is 59:41 in the dataset implying that aluminum is the preferred choice for shoe.

DRAW

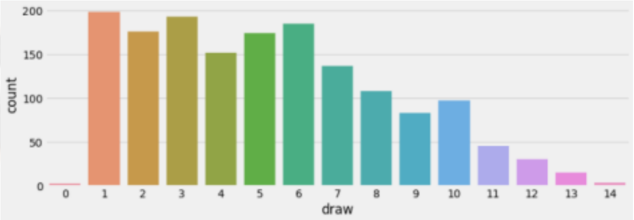


Fig. (10). Draw Vs No of Races

A horse stall number or 'draw' is randomly allocated for each horse just before the start of a particular race. A lowest number indicates that the stall is near to the inside railing and thus have to travel less distance than those with higher values. The analysis in Figure 10 shows that draw '1' tops the list with 198 out of total 1571 wins.

SEX

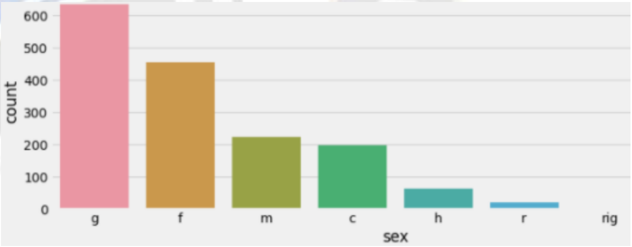


Fig. (11). Sex

Colt is a young male horse, yearling is a horse (either male or female) between the ages of one and two, a young female is called filly and a mare when she is adult. An adult horse when castrated is called gelding, a rig or ridgling on other hand is partially castrated[31]. A gelding supersedes all (634) followed by filly (456), trailed poorly by ridgling (4) in the last (Figure 11). However there is a scope of study in this area, as authors have found almost no research papers on the said topic.

Of the total 14,750 records, 13,179 records are for 'position'=0 (no win, majority class) and 1571 for 'position'=1 (win, minority class). After a stratified 70-30 train-test split of the binary class is done, 9229(70-30 split) and 1096(70-30 split) records are achieved for majority and minority class respectively (Table 2 above), that is minority class is just 11.8 % of the majority class thus emphasizing the skewness. Oversampling technique called random oversampling is

DOI: https://doi.org/10.17762/ijritcc.v11i9.8119
***Article Received:** 10 July 2023 **Revised:** 30 August 2023 **Accepted:** 18 September 2023*

applied as it balances the data by replicating the minority class samples, and so the number of samples in majority and minority class was 9229 and 6460 respectively (Table 2 above). After applying SMOTE sampling technique, the number of elements in majority and minority class was almost equal at 9245 and 9205 respectively(Table 2 above) and dataset was used for training of ML models. Results are compiled and analyzed using Table 3. Case 1 refers to the results achieved by applying the six ML models mentioned above using SMOTE technique with cross validation=5 and Case 2 when no sampling technique is applied to the dataset.

To apply Linear Regression we used the Grid Search technique with hyper-parameter tuning. A grid-search looks for all combinations of hyper-parameters into the model individually, and then returns the set of parameters which gives the best result [32]. The accuracy achieved with Linear Regression with grid-search hyper-parameters without using SMOTE (case 2) was 89%, ROC=50%, but F1 score was 0%, which implied that the results were biased because of imbalanced dataset and not even a single winner was predicted correctly. Random Forest Classification gave ROC=51% and F1 score of 5% and accuracy of 89%. Naive Bayes gave an accuracy of 88.8%, ROC=50.3%, but with no improvement in F1 score (3%) and only 6 winners were predicted correctly. Then SMOTE based prediction model (case 1 as described above) were applied and it gave best performance with the Random Forest Classifier (ROC = 97.6% and F1 score of 92.9% and accuracy of 93.1% followed by k-NN (ROC = 85.5%, F1 = 79%), Naive Bayes (ROC= 66.4%, F1=66%) and Linear Regression(ROC = 56%, F1=56.5%). The results are compiled in Table 3. This analysis clearly emphasized the usefulness of SMOTE technique to predict correctly the horse race winner.

Table 3 presents the ML classifiers performance using evaluation metrics discussed above . The results shown in case 1 represent the evaluation metrics obtained by applying SMOTE technique respectively. The evaluation metrics in case 2 denotes the performance without using SMOTE technique on dataset. From the inspection of this table, it can be seen that Accuracy, F1 Score, and ROC are evidently improved in case 1 than in case 2 for all classifiers. Also, it is noteworthy, that even though the accuracy without SMOTE (case 2) for the ML classifiers is high (Random Forest, Linear Regression (both showing 89%) followed by K-NN and Naive Bayes marginally, the F1 scores of the same are very less and are evident of the skew-ness arising out of imbalanced nature of the data.

Table 3. Performance Metrics for ML Prediction Algorithms Model of Dataset

ML Algorithm Applied	Using SMOTE (CASE 1)			Without using SMOTE (CASE 2)		
	Accu	F1	ROC	Accu	F1	ROC
K-NN	76.3	79.0	85.5	88.5	4.0	50.4
Naive Bayes	59.2	66.0	66.4	88.8	3.0	50.3
Random Forest	93.1	92.9	97.6	89.0	5.0	51.0
Linear Regression	54.6	56.5	56.0	89.0	0.0	50.0

Figure 12 summarizes the comparative ROC performance of all classifiers in two cases as discussed above.

-

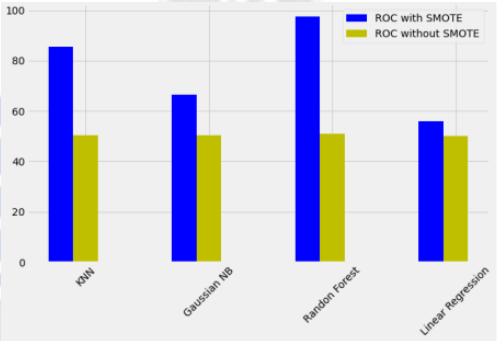


Fig. (12). ROC Values for ML classifiers

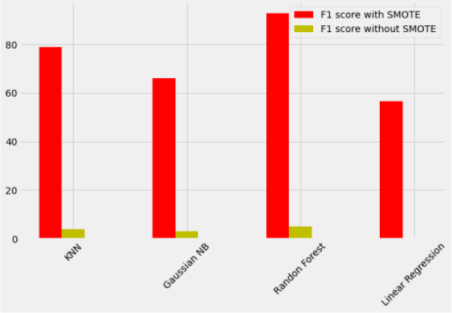
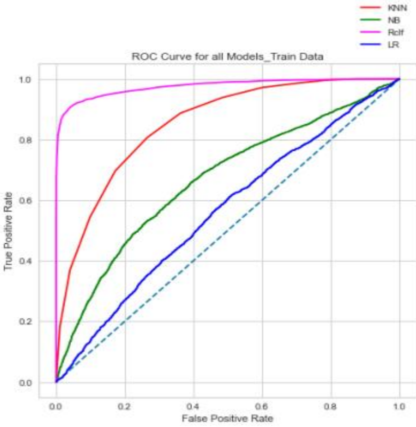


Fig. (13). F1 scores of ML classifiers

F1 scores of all the above discussed algorithms are given in Table 3 above with Random Forest (92.9) taking the lead followed by K-NN (79.0), Naive Bayes (66.0) and Linear Regression (56.5) in that order. The F1 scores in Figure 13 shows the comparison of all classifiers for both the cases as described above. The ROC-AUC curve in Figure 14 below annotates the above analysis and clearly shows that Random Forest Classifier has given the best result with ROC of 97.6% followed by K-NN (85.5%) which performed better than Naive Bayes and Linear Regression classifiers (66.4% and 56% respectively) after applying SMOTE sampling technique.

Article Received: *10 July 2023* ***Revised:*** *30 August 2023* ***Accepted:*** *18 September 2023*



s

Fig. (14). AUC-ROC curve for ML classifiers with SMOTE technique

In predicting the winner of the horse by several feature picks, the authors have not found any research that has made a comparison examination of the results of different classification systems based on SMOTE technique. In conclusion, it can be stated that Random Forest with SMOTE technique outperformed all other classifiers. This research has also highlighted the imbalanced nature of horse racing data domain and SMOTE as a technique to counter the same. This research may be expanded to several aspects. One may grow the training set by gathering additional real time data and continuing adding and testing many more functionalities for their overall impact on the result of a race. By adding new features, and advanced feature selection methods, the feature set may be enhanced. Also, further analysis with feature selection with SMOTE technique can be done so as to get more insight and accuracy from the data.

VI. CONCLUSION

Through this study it is aimed to analyze Machine Learning modeling techniques to predict the horse race winner accurately. The above mentioned ML framework predicted the winner with an accuracy of 97.6% and has outperformed the results reported by previous predictive models. This paper also analyzes the available literature in ML modeling, focusing on the Algorithms for prediction of sport results, issues with imbalanced dataset, behavior of sampling techniques and ML framework using SMOTE. This research can be helpful to fellow researchers as well as it can be used by stakeholders like punters, horse race managers, horse owners, club owners etc. to make more profit using the results and analysis. This research work can also be extrapolated to other domains for prediction which have imbalanced data like healthcare, financial sector, education etc.

REFERENCES

[1] “Grand National prize money set at £1m as large crowds expected at Aintree.” https://sbcnews.co.uk/retail/2022/01/13/grand-national-prize-money-set-at-1m-as-large-crowds-expected-at-aintree/ (accessed Jan. 25, 2023).

[2] D. Lange, “• Horse racing track market value US 2021 | Statista.” https://www.statista.com/statistics/1017245/us-horse-racing-tracks-market-size/ (accessed Feb. 19, 2022).

[3] “Global Sports Betting Market Size & Growth Report, 2030.” https://www.grandviewresearch.com/industry-analysis/sports-betting-market-report# (accessed Jan. 25, 2023).

[4] N. M. Allinson and D. Merritt, “Successful prediction of horse racing results using a neural network,” in IEE Colloquium on Neural Networks: Design Techniques and Tools, 1991, pp. 1–4.

[5] L. C. Hei, C. L. Wai, and S. B. P. M. R. Lyu, “Research in Collective Intelligence through Horse Racing in Hong Kong”.

[6] I. L. Tom Hope, Yehezkel Resheff, “Learning Tensorflow,” J. Chem. Inf. Model., vol. 53, no. 9, pp. 1689–1699, 2013, Accessed: Jan. 25, 2023. [Online]. Available: https://www.oreilly.com/library/view/learning-tensorflow/9781491978504/

[7] R. P. Schumaker and J. W. Johnson, “An investigation of svm regression to predict longshot greyhound races,” Commun. IIMA, vol. 8, no. 2, p. 7, 2008.

[8] J. Williams and Y. Li, “A case study using neural networks algorithms: horse racing predictions in Jamaica,” in Proceedings of the International Conference on Artificial Intelligence (ICAI 2008), 2008, pp. 16–22.

[9] E. Davoodi and A. R. Khanteymoori, “Horse racing prediction using artificial neural networks,” Recent Adv. Neural Networks, Fuzzy Syst. Evol. Comput., vol. 2010, pp. 155–160, 2010.

[10] N. Silverman, “A hierarchical bayesian analysis of horse racing,” J. Predict. Mark., vol. 6, no. 3, pp. 1–13, 2012.

[11] S. Pudaruth, N. Medard, and Z. B. Dookhun, “Horse Racing Prediction at the Champ De Mars using a Weighted Probabilistic Approach,” Int. J. Comput. Appl., vol. 72, no. 5, 2013.

[12] N. Silverman and M. Suchard, “Predicting horse race winners through a regularized conditional logistic regression with frailty,” J. Predict. Mark., vol. 7, no. 1, pp. 43–52, 2013.

[13] T. Takahashi, “The effect of age on the racing speed of Thoroughbred racehorses,” J. equine Sci., vol. 26, no. 2, pp. 43–48, 2015.

[14] R. P. Bunker and F. Thabtah, “A machine learning framework for sport result prediction,” Appl. Comput. Informatics, vol. 15, no. 1, pp. 27–33, Jan. 2019, doi: 10.1016/j.aci.2017.09.005.

[15] R. P. Schumaker, “Machine Learning the Harness Track: A Temporal Investigation of Race History on Prediction,” J. Int. Technol. Inf. Manag., vol. 27, no. 2, pp. 2–24, 2018.

[16] “The Best Indian Horse Racing information site for live Results, Live Odds,form guide, selection, Videos, Photos, Reviews.” https://www.inhorseracing.com/blog (accessed Jan. 25, 2023).

[17] A. S. Hussein, T. Li, C. W. Yohannese, and K. Bashir, “A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE,” Int. J. Comput. Intell. Syst., vol. 12, no. 2, p. 1412, 2019.

***Article Received:** 10 July 2023 **Revised:** 30 August 2023 **Accepted:** 18 September 2023*

[18] S. Mishra, P. K. Mallick, L. Jena, and G.-S. Chae, “Optimization of skewed data using sampling-based preprocessing approach,” Front. Public Heal., vol. 8, p. 274, 2020.

[19] “SMOTE — Version 0.10.1.” https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html (accessed Jan. 25, 2023).

[20] S. Shrivastava, P. M. Jeyanthi, and S. Singh, “Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and boosting,” http://www.editorialmanager.com/cogentecon, vol. 8, no. 1, Jan. 2020, doi: 10.1080/23322039.2020.1729569.

[21] C. Soto Valero, “Predicting win-loss outcomes in MLB regular season games-a comparative study using data mining methods,” Int. J. Comput. Sci. Sport, vol. 15, no. 2, pp. 91–112, 2016, doi: 10.1515/IJCSS-2016-0007.

[22] D. Coomans and D. L. Massart, “Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-Nearest neighbour classification by using alternative voting rules,” Anal. Chim. Acta, vol. 136, pp. 15–27, 1982.

[23] PN Tan, M.Steinbach, A.Karpatne, and V.Kumar, “Introduction to Data Mining.” https://www.pearson.com/en-us/subject-catalog/p/introduction-to-data-mining/P200000003204/9780137506286 (accessed Jan. 25, 2023).

[24] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, “AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes,” J. Supercomput., vol. 77, no. 5, pp. 5198–5219, 2021.

[25] G. Biau and E. Scornet, “A random forest guided tour,” Test, vol. 25, no. 2, pp. 197–227, 2016.

[26] “Understanding the Types and Classes of Horse Races.” https://www.liveabout.com/understanding-the-types-and-classes-of-horse-races-1880414 (accessed Jan. 25, 2023).

[27] “Why is pedigree important in horse racing? | myracing.” https://myracing.com/guides/guide-to-racing/pedigree-important-horse-racing/ (accessed Jan. 25, 2023).

[28] “Thoroughbred breeding theories - Wikipedia.” https://en.wikipedia.org/wiki/Thoroughbred_breeding_theories (accessed Jan. 25, 2023).

[29] “M.A.M. Ramaswamy: King of the course - India Today.” https://www.indiatoday.in/magazine/sport/story/19830930-mam-ramaswamy-indias-biggest-racehorse-owner-771050-2013-07-17 (accessed Jan. 25, 2023).

[30] Kumar Sharan, “News Horse Racing - Aluminium or steel: What is your shoe? - by Sharan Kumar - Racing India’s first and foremost website on horse racing India.” https://www.racingpulse.in/code/stpageprint.aspx?pgid=36194 (accessed Jan. 25, 2023).

[31] “Colt (horse) - Wikipedia.” https://en.wikipedia.org/wiki/Colt_(horse) (accessed Jan. 25, 2023).

[32] J. Wong, T. Manderson, M. Abrahamowicz, D. L. Buckeridge, and R. Tamblyn, “Can Hyperparameter Tuning Improve the Performance of a Super Learner?: A Case Study,” Epidemiology, vol. 30, no. 4, p. 521, Jul. 2019, doi: 10.1097/EDE.0000000000001027.