Application of Data Mining and Machine Learning in Sports

Sankara Subramanian Venkatraman
School of Computing
National College of Ireland
Dublin, Ireland
x18179541@student.ncirl.ie

***Abstract*—This paper builds the classification and the regression model to accurately predict the result of games based on knowledge discovery in database. It evaluates the models using statistical methods. Logistic regression and k-nearest neighbor models are used to predict the result of a National Hockey League (NHL) game. The results of the models are compared using CrossTable() validation in k-NN model, and (R^2) and significance values in logistic regression. Sentiment analysis of tweets from a National Basketball Association (NBA) match using Naive Bayes (NB) to classify iPhone and android users. Data preprocessing such as data cleaning and transformation applied to the data were complex. The results are analyzed using accuracy, F-measure and Area Under Curve (AUC) value. To examine the winning position of horses using C5.0 decision tree and boosting model with regression tree in Hong Kong horse racing 2016. Finally, compare the correlation between variables, model results and explain the statistical significance using Mean Absolute Error (MAE) and Cohen's Kappa coefficient from confusion matrix. Alongside performance tuning performed in the models is also discussed.**

***Index Terms*—logistic regression; k-NN; c5.0 decision trees; regression trees; naive bayes; boosting model.**

I. Introduction

Massive amounts of data are being collected by modern organizations for mining the insights in the data to deliver better decisions. Machine learning models are built on historical data to extract the hidden pattern and predict future results using predictive analytics [1]. In this paper, various regression and classification models are applied to the datasets to identify the patterns and predict the results. Knowledge Discovery in Databases (KDD) methodology is applied to the datasets. As discussed earlier 3 datasets and 5 machine learning models are used in this paper to evaluate and predict the results.

Sports data mining has a large volume of data from each player, team, country and sports-wise. By applying machine learning models to the sports data we can predict the opportunity of a team winning a match and an athlete's past performance can help us to predict his/her chances of winning a gold medal in the upcoming Olympics.

The main objective of the National Hockey League[1] dataset is to predict the decision (Win or Lose) of upcoming games based on the historical data. The data has information about players, teams, games played, shots, saves and power play. The ultimate challenge of this dataset is that from previous studies the best model can predict the accurate result only 62% of the time. To improve the accuracy of the results, models such as kNN and logistic regression are applied and compare the best model which fits the dataset. Logistic regression is a simple and powerful model, which considers each predictor variable's impact and also provide statistical significance on the target variable [2]. Generally, logistic regression is used to predict dichotomous response variable.

Hong Kong horse racing result[2] dataset is analyzed using C5.0 classification tree and regression tree with boosting model. The gambling model was built on the assumption of the top 3 positions are considered as winners and the rest of the positions are treated as losers.

By default, C5.0 tree follows post pruning of decision tree. It allows the tree to intentionally over-fit the training data and later, prune the leaf nodes to an appropriate level. They are known as subtree raising and replacement respectively [3]. Regression tree uses recursive binary splitting, a heuristic or greedy approach. Instead of looking forward and picking the split, the tree-building process greedily splits the best at a particular step. The regression tree also uses subtree replacement like C5.0. It builds the tree so long as it decreases the Residual Sum of Squares (RSS) value on each split [4]. Finally choosing the best model which fits the data.

Cleveland Cavaliers and Golden State Warriors tweets[3] dataset contains information about tweets, source, created date, language, location, description, followers and country. Naive Bayes is applied to this dataset to classify the tweets are tweeted from an iPhone or Android mobile. There are about 40 languages in the tweets. But only English is chosen for this analysis. The assumption of NB is the features are independent and return MAP prediction. Naive Bayes has more advantages than other prediction models. The results are more accurate, works well for missing value data, easy to train, can handle very large datasets and the compactness is the simplicity of the model. Text analytics is performed on the dataset using Natural Language Processing (NLP)

[1]https://www.kaggle.com/martinellis/nhl-game-data

[2]https://www.kaggle.com/edwardckw/notebookd02724dac3/data

[3]https://www.kaggle.com/xvivancos/tweets-during-cavaliers-vs-warriors

II. Related Work

A. National Hockey League Dataset

*1) **National Hockey League**:* A study on National Hockey League (NHL) player's performance before and after a concussion. Six variables are taken into analysis and none of the attributes show significant interactions. There was no worse performance by any player even after 3 consecutive post-injury concussion was observed using advanced sports metrics. Using advanced laboratory approaches deficits of post concussions have been determined but have not been explained clearly. The reason for the failure of the analysis of post-concussion is due to neurologic deficit may vary after return to play, neurologic compensatory can mitigate the effect and the metrics are not sensitive to the decrement [5].

Another study on National Hockey League carrier model of Canadian ice-hockey players. This is a sports psychology study that focuses to devise NHL career model through interviews from 5 rookies, 5 veterans and 13 retirees. The dataset and the interviewers had undergone thematic analysis. Thematic analysis is to identify the patterns and is robust. This analysis is carried out in 6 steps and a 15-point checklist is prepared to ensure the data quality. Model was composed of various stages, demands, pathways, statuses, barriers and outcomes. The end results say it is good to make friendships, set expectations and assertiveness, and also it complements goal setting and cohesion [6].

*2) **k-Nearest Neighbour**:* Identification of high and low psychological potential archers (HPPA and LPPA) based on psychological coping skills using k-Nearest Neighbour method. 50 youths in the average age range of 16 to 18 are considered for the analysis. The variables assessed for k-NN are classified into 6 kernel functions. Cosine k-NN model provides better accuracy and low error rate in the prediction of HPPA and LPPA. They have used different evaluation techniques such as accuracy, sensitivity, specificity, precision, error rate and confusion matrix that was used for evaluating the k-NN model. The results demonstrate that findings of the investigation are non-trivial to sport managers and coaches to recognize the talents [7].

Sports-based k-NN classifier for ambient intelligence assistant. This system consists of a wireless sensor network (WSN) is connected to a mobile phone to monitor the heart rate in external conditions. Here k-NN model is used as a decision engine. It will fit the category based on the majority vote in case of classification and uses averaging for regression models. k-value is tuned between 1 to 30. For k=10, they got a success rate of 70%. Type-I error is more likely than type-II error (37% and 23%) respectively. The success rate can be tested with more predictor variables such as humidity and time elapsed [8].

*3) **Logistic Regression**:* Logistic regression of elite player's perception of football playing surfaces. 1129 players in 44 different countries were surveyed to check whether the elite players prefer to play in Natural or Artificial turf. An elite player's junior and senior-level training and played sessions are considered in the analysis. The questionnaire is subdivided into 6 sentiments and the players are asked to respond on 4 categories such as natural, artificial, gravel and indoor surfaces. From the questionnaire sentiments of C, E and F are analyzed using logistic regression. The models included 8 pitch properties and analyzed for 3 different sentiments. The models are evaluated using (R^2) and the significance value of the attributes [9].

A study on biased penalty calls in National Hockey League. The dependent variable is the penalty call for home or road team and the independent variables are the total difference between road and home penalties, total goal difference between home and road teams, time in the match and team strength based on season points. similarly, the gradient boosting model is also performed and compared with LR. The predicted probabilities are observed for each independent variable. From the results two things are observed first, the team with more penalties is less likely to call the next penalty. Second observation is the next penalty is more likely to call for the leading team [10].

A study on analysis of Table-Tennis (TT) as leisure-time in Poland using binary logistic regression. The total population of 12,406 people is considered for the analysis. The response variable is participation table-tennis "Yes" or "No". The significant predictor variables are gender, place of residence, age, income, being a student and children in the house. The result says 2.8% of the overall population is practicing TT and 6.6% is considering as leisure physical activity. From its undeniable benefits, it is promoted in various places such as schools, public gatherings and schools practiced as physical education activity [11].

B. Horse Racing Dataset

*1) **Horse Racing**:* The existence of parimutuel horse racing is one of the oldest forms of gambling. Betters will bet on the horse of their favorite choice. Monies will be collected by the racetrack and grouped as a pool. A small percentage of money is taken as profit by the racetrack. The remaining money will be shared amongst the winners. "Odds" is the terminology used to define the potential winnings of the bettor's [12].

A study to compare probabilistic seismicity forecasts based on parimutuel gambling. In this study, various statistical methods such as information gain, likelihood ratio and Molchan diagram are addressed. A spurious correlation between parimutuel to forecast is carried out in two ways head-to-head and round table. From this study, no statistically significant pattern is observed as the method is intuitive [13].

Decision tree and decision processes study in gambles and more natural decision tasks. In this process, they have experimented with the natural situation tasks decision can be generalized for gamble decisions. 32 participants from military, medical clinics and social organizations have participated in Austria. Half of the participants were trained to draw a decision tree during the process.

The experiment is carried out in identifying 4 decision tasks. 1 for gambling (GAMBLE) and 3 for non-gambling

include post office task (POST), investment task (INVEST) and Exploitation of natural resources task (RESOURCES). These 4 variables are considered as predictor variables and the dependent variables are decision-verbal and verbal-protocol data classified into 3 categories as Structure, Information and Processing. From the analysis, it concluded that decision-makers have invented new ideas in non-gambling tasks than gambling task. Also, the natural decision differs from gambling task is concluded [16].

*2) **C5.0 Decision Tree**:* A remarkable work on agricultural data analysis on the cloud using C5.0 advance decision tree for selecting the crops based on the soil fertility attributes. In this paper, they have developed an android mobile for farmers to suggest the crop selection on soil fertility. RHadoop RMapReduce and Rhdfs are integrated to handle a large volume of data. The concept of pruning trees is also incorporated to reduce tree size and increase memory usage. Finally based on pH, fertility level of soil and soil type the suitable crops are suggested to farmers [14].

A machine-learning study on analysis of Bangladesh cricket team One-Day International Match Data. In this paper, 5 different models such as C5.0 decision tree, Naive Bayes, SVM, Random Forest and k-NN are used to predict the result of the One-Day series played against South Africa. Various attributes such as Home-Away condition, toss, venue, day-night or day match and year are considered for the analysis. The models are conducted as experiments in 4 stages. Finally, model results are compared using CrossTable Validation. Except for SVM, all the models have classified the results as expected [15].

*3) **Regression Tree with Boosting**:* A study on identification of the best predictors of classifying winning and losing basketball in fast-paced and slower-paced games using classification and regression tree. Using k-means cluster the match is categorized into slower-paced and fast-paced matches. From the result, it is analyzed for fast-moving matches the CTR showed 5 significant factors that are influential in 4-stages trees with 14 contrasting node groups. For slower-paced games, CRT analysis showed 6 significant factors in 5-stages tree with 18 contrasting nodes [17].

Gradient Boosting Model (GBM) application to predict gender based on master athlete's motivation. A decision tree is refitted iteratively by GBM to residuals. It executed in a step-wise manner to minimize the loss function. In this study, different models are compared such as C5.0, J48, XGBoost, LightGBM and GBM. Out of 6 models, XGBoost and GBM produce better accuracy of 0.7012 and 0.7134 respectively. The models are evaluated using confusion matrix kappa value, accuracy, sensitivity and specificity are compared [18].

C. NBA Tweet Dataset

*1) **Sentiment Analysis**:* Identification of sentiments specific to football tweets ate analyzed using SVM, Naive Bayes and Random forest Model. In this study, the data is collected from football fans Twitter tweets when they react to the match, goal-scoring, penalties and so on are analyzed from 3 different datasets. The sentiments are annotated into 3 categories positive, negative and neutral. In this work, 3 features are utilized Bag of Words (BOW), Part-of-Speech (POS) and Lexicons. The dataset is processed as testing and training data and the models are applied. For 3 datasets and 3 different features, 3 models are applied and their accuracy and F-score are tabulated and evaluated using Cross Table Validation. In this conclusion, SVM outperforms MNB and RF [19].

Estimation of unsupervised Sentiment-Analysis (SA) of informal text during public events based on language and location. This is a novel approach of combining geographical areas, the participating characters and the group belonging in public events. Unsupervised SA is carried out using Natural Language Processing (NLP) which evaluates people's sentiments, attitudes, emotions and attributes. It also used an unsupervised-lexicon based approach such as n-grams, stemming, tokenization and lemmatization. Two hypotheses are tested, the first one is a sense of group that is coherent to tweets posted at a particular time and place. Secondly, if the responses are typical to a particular action that particular users are not aligned. It is concluded that unsupervised SA does not make any assumption to the users [20].

A Big data sentiment analysis of twitter tweets during the 2014 Soccer world Cup using disposition theory. It states that the sentiment of sports spectators' behavior is affected by the supporting team position in the match. The data pre-processing is carried out in the same way as previous work. They used the NLP algorithm along with 7 lexicon emotion categories [21].

*2) **Naive Bayes**:* A comparison study on SVM Vs Naive Bayes techniques for 2013 FIFA confederation cup tweets. The data is collected using twitter API and the data is selected using a corpus that used DocumentTermMatrix (DTM). The sentiment is analyzed using 3 sentiments positive, neutral and negative. Various evaluation methods such as accuracy, sensitive, precision and F-score are used. From the conclusion, it is observed that SVM model performs better than Naive Bayes with an accuracy of 80.0% and 72.7% respectively. The positive and negative words are visualized using a word-cloud package [22].

Naive Bayes for classifying tweets with political motives containing hate speech. In this method, they have used WEKA for classification and finally Naive Bayes to calculate the accuracy of political motive hate speech. The hate speech is categorized as political motive, non-political motive and non-hate speech. An accuracy of 93.2% is achieved from Bayes algorithm [23].

III. KDD Methodology

KDD data mining methodology is followed through the paper[4]. Seven models are applied to 3 datasets and the observations are documented.

[4]http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

A. National Hockey League Dataset

Data Selection:

1) As discussed earlier, this dataset is sourced from kaggle website and the data source is reputed.
2) The data is downloaded from the website and stored in CSV format in the local machine.
3) The dataset contains 18 continuous predictor variables and the response variable is decision and it is dichotomous.
4) The dataset contains 24,646 records and the data quality is ensured in the below sections.

Data Cleaning and Pre-processing :

1) The data cleaning part is carried out by removing 1777 null values in the target variable.
2) Also, the condition of 'NA' value is examined to the dataframe and the value of **NA** is replaced with **0**.
3) The below figure shows the pre-processing steps followed in the dataset.

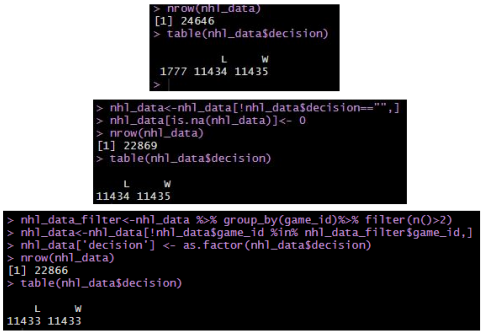

Fig. 1. Data Pre-processing

Transformation and Data Preparation:

1) The dataset is transformed using min-max normalization an User Defined Function (UDF) is created.

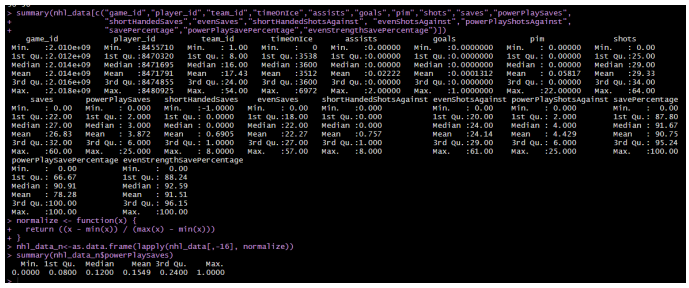

Fig. 2. Data Transformation

2) The independent attributes are normalized using UDF and the values are brought into a range of 0 to 1.
3) Using **dplyr** library, the dataset's duplicate record are removed which **group_by** game_id.
4) The dataset is split into 2 sets one for training the model and other for testing the model built.
5) 80% of the dataset is for training (18292) records and 20% is for testing (4574) records.
6) In the next step of KDD process, data mining models of k-NN and logistic regression are applied and the insights are explained in the evaluation section.

Data Mining Models:

k-Nearest Neighbor and logistic regression models are applied to classify the result and the methods are compared by CrossTable evaluation method in **gmodels** library.

1) k-Nearest Neighbor:

Min-Max Normalization:

- In k-NN model, the training dataset is passed into knn model's train value in the class package and the test dataset is passed for test value and the response variable which is stored in labels is passed to class factor vector and k value is tuned.
- The k-values are tuned from 5 to 200. In this dataset 5,10,100, 135, 151 and 200 are the values chosen for k-value.
- The reason for choosing **k = 151** is because the total records in the dataset is 22866 and the square root value is chosen as k-value. Similarly, the number of records in training dataset square root is **135**.
- The remaining k values of 5, 10, 100 and 200 are chosen random and to check whether the data can produce a better result for low-vales of k.
- The accuracy are observed using CrossTable validation for each values of k.

  Different k-values Accuracy:

  – Accuracy of 0.64 for k=5
  – Accuracy of 0.67 for k=10
  – Accuracy of 0.68 for k=100
  – Accuracy of 0.67 for k=135
  – Accuracy of 0.67 for k=151
  – Accuracy of 0.67 for k=200

z-score Normalization:

- In z-score normalization, the z-score is calculated using the formula

$$\boxed{z = (x - \mu)/\sigma}$$

- The independent variables mean, standard deviation are calculated and substituted in the above formula to get the normalized data.
- Similarly, the data are split into training and testing dataset and the k-NN method is repeated.
- The new accuracy are observed using CrossTable validation for each values of k.

  Different k-values Accuracy:

  – Accuracy of 0.70 for k=5
  – Accuracy of 0.72 for k=10
  – Accuracy of 0.73 for k=100
  – Accuracy of 0.73 for k=135
  – Accuracy of 0.74 for k=151
  – Accuracy of 0.74 for k=200

- From the above two normalization, it is observed that z-score normalization provides better accuracy than min-max normalization for **k=151**.

2) Logistic Regression: The application of logistic regression is applied to the dataset from referencing the sites[5] [6].

- Above mentioned data cleaning, pre-processing and transformation is also applicable to this model.

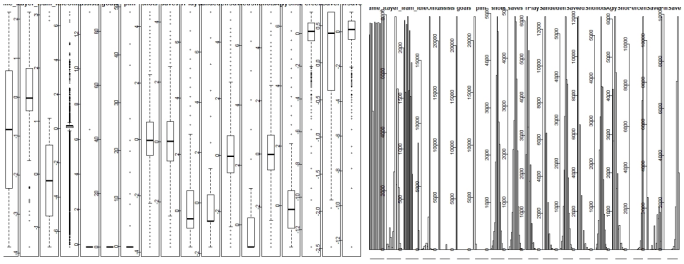

Fig. 3. Data Distribution using Histogram & Box Plot

- The above box and histogram chart represents how the independent variables are distributed among different values after z-score normalization.
- From the below figure, there are no missing values and correlation matrix between the predictor and target variables are observed using pairs function.

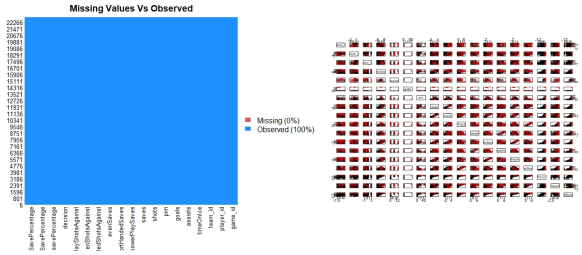

Fig. 4. Missing Value & Correlation Plot

- Logistic regression is applied to the same training and testing dataset using **glm** in caret package.

```
> glm.fit.train<- glm(decision~. -powerPlayShotsAgainst, data=nhl_log_train, family = "binomial")
> gm.fit.null <- glm(decision~ 1, data= nhl_log_train, family = "binomial")
> 1-logLik(glm.fit.train)/logLik(gm.fit.null)
'log Lik.' 0.2860997 (df=18)
> glm.probs.train <- predict(glm.fit.train, newdata = nhl_log_test, type="response")
> glm.pred.train <- ifelse(glm.probs.train>0.5, "W", "L")
> decision_test <- nhl_log_test$decision
> table(glm.pred.train,decision_test)
              decision_test
glm.pred.train    L    W
             L 1712  529
             W  575 1758
> mean(glm.pred.train==decision_test) # 0.758
[1] 0.7586358
>
```

Fig. 5. Logistic Regression Model Accuracy

- The above figure represents the accuracy and the CrossTable result which classified the results correctly, incorrectly and (R^2) value.

[5]https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/
[6]https://www.datacamp.com/community/tutorials/logistic-regression-R

Evaluation Method:

The above two data mining methods are evaluated using different methods. In this case, we are using CrossTable validation for evaluation. In the case of logistic regression, it is observed that out of 19 variables only 8 variables are statistically significant with R^2 value of 28.609%.

While comparing the accuracy of both the model's logistic regression outperforms k-NN model with better accuracy of 0.758 where as k-NN accuracy is 0.746.

Knowledge discovery:

From this dataset and models, we have discovered that both models have performed and provide better accuracy close to 0.75. In the case of logistic regression model can be improved by removing the insignificant variables and re-run the model which may increase the model overall performance provides as better R^2 value.

Similarly, the same can be followed in k-NN to check the accuracy after removing those insignificant attributes and compare the results for a better understanding. We conclude that logistic regression classifies the match decision better than k-NN model.

The logistic regression model can be predicted by the equation:

$$Y = \frac{\exp(174.9 + 0.0000000355x1 + 0.000940x2 + 1.255x3}{1 + \exp(174.9 + 0.0000000355x1 + 0.000940x2 + 1.255x3}$$

The model is built with predictor variables with 3 significant codes (***) i.e (p<0.0005).

B. Horse Racing Dataset

Data Selection:

1) This dataset is also sourced from kaggle website and 2 source files are merged to form a final dataset.
2) Horse and Race result dataset are merged using race_id.

Data Cleaning and Pre-processing:

1) The dataset containing valid final_position values from 1 to 14 are taken and rest of the values are filtered using **droplevels**.

```
> # Data Cleaning
> library(dplyr)
> finishing_positions <- c("1","2","3","4","5","6","7","8","9","10","11","12","13","14")
> horse_race_result <-horse_race_result %>% filter(horse_race_result$finishing_position %in% finishing_positions)
> prop.table(table(droplevels(horse_race_result$finishing_position)))

           1           10           11           12           13           14            2            3            4            5            6            7            8
0.08040458 0.07696499 0.07444490 0.06879172 0.03388503 0.02853835 0.08016619 0.08002997 0.07968942 0.07972347 0.07989375 0.07965536 0.07931481
0.07849748
> horse_race_result <- merge(horse_result, race_result, by = "race_id", all.x = TRUE)
> prop.table(table(droplevels(horse_race_result$finishing_position)))

           1          1 DH           10          10 DH           11          11 DH           12          12 DH           13           14            2
7.821248e-02 3.975221e-04 7.486666e-02 2.650147e-04 7.241528e-02 1.325074e-04 6.691622e-02 6.625369e-05 3.296121e-02 2.776029e-02 7.798059e-02
          2 DH            3            3 DH            4            4 DH            5            5 DH            6            6 DH            7            7 DH
4.637758e-04 7.784808e-02 6.625369e-04 7.751681e-02 1.159439e-03 7.754994e-02 5.962832e-04 7.771557e-02 4.637758e-04 7.748369e-02 2.650147e-04
           8            8 DH            9            9 DH         DISQ          DNF           FE           PU          TNP           UR           WV
7.715242e-02 3.975221e-04 7.635737e-02 2.650147e-04 3.312684e-05 2.318870e-04 3.312684e-04 9.606784e-04 1.987611e-04 8.612979e-04 1.527147e-02
          WV-A           WX          WX-A
3.378938e-03 5.300295e-04 3.312684e-04
```

Fig. 6. Horse Final Position

2) The null and NA values are converted to 0 for all the attributes of the dataframe by creating a UDF.
3) Only C5.0 decision tree can handle **factors with multiple levels**.

Transformation and Data Preparation:

1) The final_position is converted to integer from factor.
2) Then the position 1 to 14 is classified into 2 factors. Position (1, 2, 3) is treated as "Winner" and 4 to 14 is considered as "Loser".

```
> horse_race_result$finishing_position <- as.integer(horse_race_result$finishing_position)
> horse_race_result['result'] = ifelse(horse_race_result$finishing_position %in% c(1,2,3),"Winner","Loser")
> horse_race_result$result <- as.factor(horse_race_result$result)
> table(horse_race_result$result)

 Loser Winner
 25556   4633
```

Fig. 7. Transformation Decision Tree

3) 90% of the data is used for training and 10% is used for testing the model built.

Data Mining Models:

C5.0 advanced regression tree and regression tree using rpart and M5P models are applied. The result is evaluated using Mean absolute error (MAE) and the model's performance is tuned using gradient boosting.

1) C5.0 decision tree:

- C5.0 decision tree from C50 library is used to predict the decision tree.
- In first run of the tree, seed is set to 123.
- race_id is removed and 18 variables are considered for analysis.
- The tree size is 25 and only 5 attributes are contributing to the trees.

  Attribute Usage:

  - length_behind_winner - 100%
  - running_position_4 - 92.24%
  - running_position_3 - 83.85%
  - running_position_5 - 73.42%
  - running_position_6 - 57.35%
- The error rate is 0.1% which incorrectly classify only 8 records 19 records in False positive and 9 in False negative in training dataset.
- The model built is predicted with testing dataset which contains 6697 records which incorrectly classifies 3 records in false positive and 5 in false negative.
- The accuracy is 0.998 and kappa statistics is 0.9955.

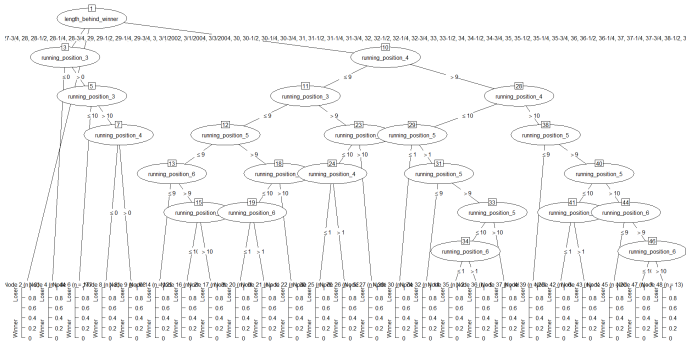

Fig. 8. Classification C5.0 Boosting Model

- The C5.0 model is boosted by increasing the number of trials=10.
- After boosting the tree size is reduced to 14.3 from 25.
- This model produces 100% accuracy and 11 attributes contributes to the trees.
- While predicting the model with test data it incorrectly classifies 44 false positive with accuracy of 0.993 and kappa statistics of 0.9748.

2) Regression Tree:

Data pre-processing and Transformation:

In the regression tree, all the factors are converted to an integer and the character attributes such as horse_name, horse_id, jockey and trainer are removed. Only 15 attributes are considered for this model. The response attribute result "Winner" or "Loser" is converted to an integer as 1 and 2.

rpart regression tree:

- In this regression tree also 90% is used for training and remaining 10% is used for testing the model built.
- rpart regression tree from **rpart** library is applied and the numeric digits is controlled by digits parameter which is there in the decision tree diagram.
- Also fallen.leaves, type and extra are the other parameters also applied in the model.
- The leaf node at the bottom of the tree will be aligned when using fallen.leaves=TRUE.

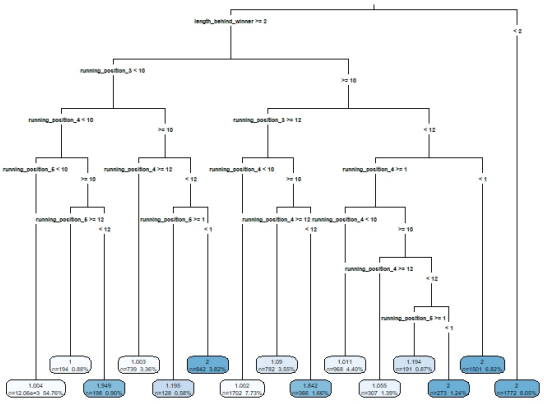

Fig. 9. Regression Tree Model

M5P regression tree:

```
LM num: 75
result =
        -0 * horse_number
        + 0 * actual_weight
        + 0 * declared_horse_weight
        - 0 * draw
        + 0 * length_behind_winner
        - 0 * running_position_1
        - 0 * running_position_2
        - 0.0004 * running_position_3
        - 0.0024 * running_position_4
        - 0 * finish_time
        + 0 * win_odds
        - 0.0503 * running_position_5
        + 0.0001 * running_position_6
        - 0 * race_distance
        + 1.7216

Number of Rules : 75
```

Fig. 10. M5P Regression Model

- As part of improving model performance M5P regression tree is also applied to the dataset.
- M5P is available in RWeka library and the number of rules is 75.

- The above figure represents the linear model of the M5P regression tree. Correlation and MAE of the trees are explained in the evaluation methods.

Gradient Boosting Model:

- Boosting approach is applied to this dataset using gbm library which stands for gradient boosting model.
- It helps us to boos the regression tree build earlier using rpart and provides better results by iterating the regression model.
- As we are using binary classification, we will use **distribution="bernoulli"** and **n.trees=5000** and **interaction.depth=4** [24].

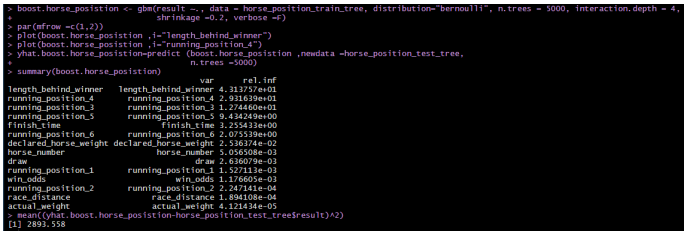

Fig. 11. Gradient Boosting Model

Evaluation Method:

1) In this dataset 4 machine learning algorithm are applied to check their performance with different measure.
2) Firstly, C5.0 Vs Regression Tree using rpart is evaluated using confusion matrix.

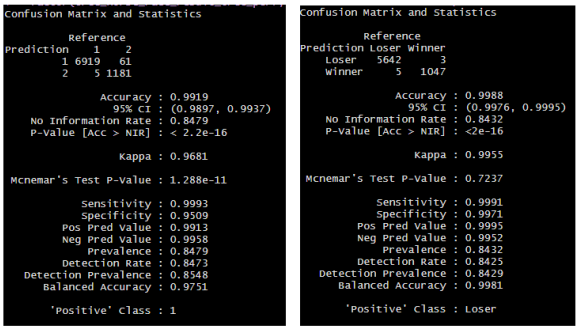

Fig. 12. Confusion Matrix for C5.0 & Regression Tree

3) From the above **confusion matrix** it is clearly evident that both the models classifies equally good.
4) The left side of the figure is the result of regression and the right side for C5.0 decision tree.
5) Looking at the kappa statistics and accuracy C5.0 is best model for this dataset with 99.8% accuracy.
6) When comparing the value of rpart regression tree with M5P tree the model is evaluated using **Mean Absolute Error** (MAE) value and **correlation coefficient (r) value**.
7) From the above figure the correlation between the M5P predict and target variable is 0.98 and in case of rpart it is 0.96 which shows **Strong correlation**.

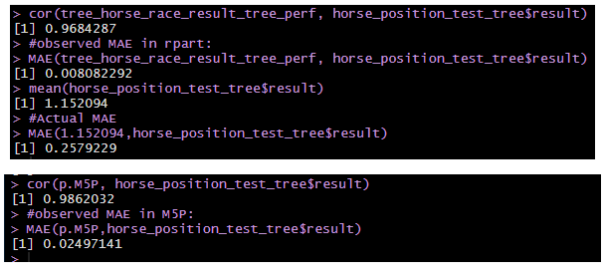

Fig. 13. MAE for rpart & M5P

8) The actual MAE value is 0.2579 but the rpart observed value is too with 0.00808 but M5P has a better value of 0.0249 but still not a great value.

Knowledge Discovery:

1) In case of C5.0 decision tree no need to remove the variables with more number of factors.
2) But for rpart and M5P regression tree the character variables should be removed or converted to numeric or converted to a certain factor level.
3) This behaves same as C5.0 regression tree because the top 4 variables in GBM and attribute contribution in C5.0 shows a similarity.
4) From the above models, it is found that C5.0 is the best model for Horse Racing dataset.

C. NBA Tweet Analysis

Data Selection:

1) In this dataset twitter tweets of an NBA match between Cleveland Cavaliers and Golden State Warriors.
2) A lot of pre-processing and transformation steps are carried out to perform Naive Bayes.
3) This dataset is also sourced from kaggle, a lot of sentiment analysis was already performed in the dataset. So, we will see a novel approach of classifying the type of the mobile phone (iPhone or Android) from which the tweets are posted.

Data Cleaning and Pre-processing:

1) The dataset contains 43 attributes but only 3 attributes text, language and source are used in this analysis.

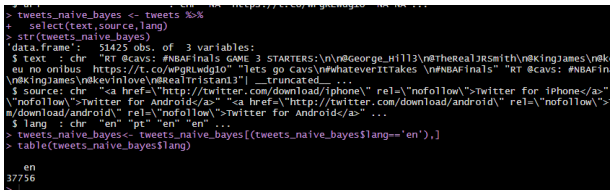

Fig. 14. Naive Bayes Cleaning & Pre-processing

2) Also this analysis is restricted to English language out of 40 different language tweets.
3) Out of 51, 425 records only 37756 records are selected after applying the language filter.

***Transformation and Data Preparation**:*

1) Using **mutate** function in dplr library a new binary column mobile_phone is created based on the text in it.
2) Initially, the values are classified into 3 categories ”iPhone”, ”Android” and ”others”. Later ”others” category is filtered.

Fig. 15. Naive Bayes Transformation using Mutate

3) The mobile_phone is converted to factor and only text and the mobile_phone is passed to the model.

***Data Mining Models**:*

Naive Bayes is applied to this dataset and the data processing is carried out using corpus package in text-mining **tm** library.

- Naive Bayes is applied and the dataframe tweets are converted to 32,182 documents using VCorpus.

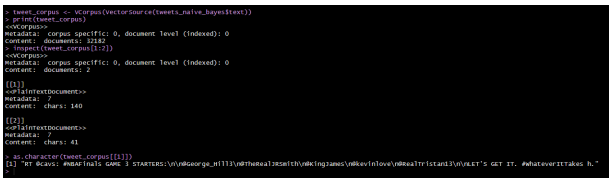

Fig. 16. Naive Bayes Document Creation using VCorpus

- The next step is remove punctuation, stop-words, numbers, and white-spaces. All the text are converted to lower case using lower function using **tm_map**.

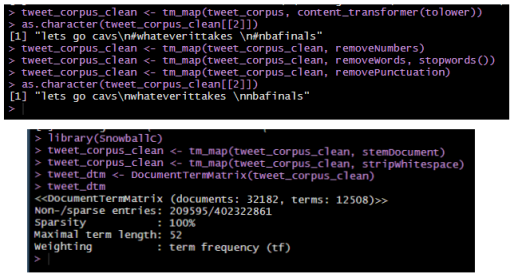

Fig. 17. Remove punctuation, Stop words, white space and numbers

- In the next step, Stemming, white space removal and conversion of document to DocumentTermMatrix (DTM) is carried.
- The 75% of the document is trained and remaining 25% is used for testing the model.
- In both, test and training dataset 66-67% are iPhone and remaining 32-33% are Android. It shows the documents are split equally.
- Using Wordcloud package, the words with more than 3 times and 40 top words are shown in wordcloud for better visualization.

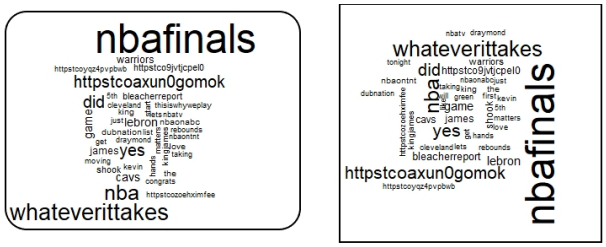

Fig. 18. Wordcloud Android Vs iPhone

- The squared one is for iPhone and the edges in the circle represent android.
- In the next stage, an UDF for dictionary is created and corpus train and test data is passed into the dictionary and only character vectors are stored in a new DTM.
- An another UDF is created to check the counts and store the counts in a new variable for all the words in train and test.
- In e1071 package, naiveBayes training dataset is passed and the dataset is predicted using test dataset.

***Evaluation Method**:*

1) In this model along with cross table validation, sensitivity, specificity, accuracy, precision, f-score are calculated manually using the formulae.

   Confusion Matrix Factor:

   - Accuracy - 0.65
   - Sensitivity - 0.47
   - Specificity - 0.71
   - Precision - 0.34
   - F-score - 0.40

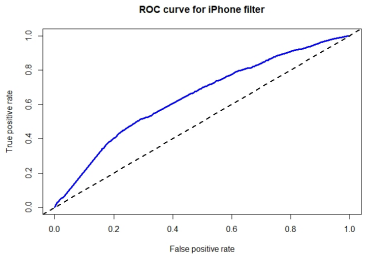

Fig. 19. AUC ROC

2) Also, Area Under Curve (AUC) ROC curve is also predicted using ROCR library. The probability is predicted using predict function and **type=raw**.
3) Next the predicted probability is passed to prediction function and target attribute is given in the labels.
4) ROC curve is plotted against True Positives Vs False Positives. Value of 0.6430 is observed for AUC.

Knowledge Discovery:

From this dataset, it is observed that the model can be fine-tuned to some extent using other NLP models such as SVM or Neural network. The evaluation method shows it has an accuracy of 65% which is a weak model and the AUC value is very minimal with a value of 0.6430 which usually in the range of 0.5-1. Previously there was no analysis is carried out to find the device based on the tweet comments. It helped me to learn how text mining works and classifies the result with the help of the Sparse matrix. Also, the Bayesian network model is improved by adding a Laplace value of 1 increased the ROC value by 0.003.

IV. Conclusion & Future Scope

From this project, we have applied the KDD model for all the dataset and the process of data selection, cleaning, transformation, model applied, evaluation methods and knowledge discovered is explained in brief.

In the first dataset on NHL, from the previous work, it is mentioned that only 62% times the predicted results were correct from the best model. By applying k-NN and logistic regression we achieved by improving the model performance to 74.6% and 75.8% respectively. In the future, we can apply a boosting model such as random forests to improve its performance by more than 76%.

From the second dataset on Hong Kong Horse Racing, we have evaluated the model performance using different methods such as confusion matrix which explains kappa statistics, F-score, accuracy and also MAE is also evaluated which explains which model is close the actual MAE in case of rpart and M5P regression trees. In the future scope of this dataset association mining, bagging model or logistic regression can be tested and check how the model performance with C5.0 advanced decision tree which classifies the result closely to 99%. One interesting observation found in c5.0 ADT is that after applying boosting value to 10 it classifies 100% in the training dataset but predicting with test dataset it incorrectly classifies 44 in false-negative which is more than without applying boosting model.

In the NBA tweets dataset, AUC is around 65% because of the split in the dataset 67% in iPhone category and 33% in Android category because of the bias the AUC value is not improvising even after Laplace value is increased. In the future scope of the dataset, we will try to remove the bias. Also, the dataset is limited to the English language which can be extended to other languages. Other complex models such as Neural Network and SVM can also be performed.

References

[1] B. M. N. John D. Kelleher, Aoife D'Arcy, "Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies," 2015, pp. 1–3.

[2] C. Chalitsios, T. Nikodelis, V. Panoutsakopoulos, C. Chassanidis, and I. Kollias, "Classification of Soccer and Basketball Players' Jumping Performance Characteristics: A Logistic Regression Approach," Sports, vol. 7, no. 7, p. 163, 2019.

[3] B. Lantz, "Machine Learning with R," 2015, p. 135,136.

[4] M. Akinkunmi, "Introduction to Statistics Using R," in Synthesis Lectures on Mathematics and Statistics, vol. 11, no. 4, 2019, pp. 305–309.

[5] T. A. Buckley et al., "Concussion and national hockey League player performance: An advanced hockey metrics analysis," J. Athl. Train., vol. 54, no. 5, pp. 527–533, 2019.

[6] R. C. Battochio and N. Stambulova, "Coping resources and strategies of Canadian ice-hockey players: An empirical National Hockey League career model," Int. J. Sport. Sci. Coach., vol. 14, no. 6, pp. 726–737, 2019.

[7] Z. Taha, R. M. Musa, A. P. P. Abdul Majeed, M. R. Abdullah, M. M. Alim, and A. F. A. Nasir, "The application of k-Nearest Neighbour in the identification of high potential archers based on relative psychological coping skills variables," IOP Conf. Ser. Mater. Sci. Eng., vol. 342, no. 1, pp. 0–7, 2018.

[8] P. López-Matencio, J. Vales Alonso, F. J. González-Castaño, J. L. Sieiro, and J. J. Alcaraz, "Ambient intelligence assistant for running sports based on k-NN classifiers," 3rd Int. Conf. Hum. Syst. Interact. HSI'2010 - Conf. Proc., pp. 605–611, 2010.

[9] A. Owen, A. C. Smith, P. Osei-Owusu, A. Harland, and J. R. Roberts, "Elite players' perceptions of football playing surfaces: a mixed effects ordinal logistic regression model of players' perceptions," J. Appl. Stat., vol. 44, no. 3, pp. 554–570, 2017.

[10] T. B. S. David Beaudoin, Oliver Schulte, "Biased Penalty Calls in the National Hockey League," pp. 366–372, 2016.

[11] E. Biernat, S. Buchholtz, and J. Krzepota, "Eye on the ball: Table tennis as a pro-health form of leisure-time physical activity," Int. J. Environ. Res. Public Health, vol. 15, no. 4, pp. 1–11, 2018.

[12] N. Silverman and M. A. Suchard, "Predicting Horse Race Winners Through Regularized Conditional Logistic Regression With Frailty," J. Predict. Mark., vol. 7, no. 1, pp. 43–53, 2012.

[13] J. D. Zechar and J. Zhuang, "A parimutuel gambling perspective to compare probabilistic seismicity forecasts," Geophys. J. Int., vol. 199, no. 1, pp. 60–68, 2014.

[14] S. Rajeswari and K. Suthendran, "C5.0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud," Comput. Electron. Agric., vol. 156, no. December 2018, pp. 530–539, 2019.

[15] M. M. Rahman, M. O. F. Shamim, and S. Ismail, "An Analysis of Bangladesh One Day International Cricket Data: A Machine Learning Approach," 2018 Int. Conf. Innov. Sci. Eng. Technol. ICISET 2018, no. October, pp. 190–194, 2018.

[16] O. Huber and A. Kühberger, "Decision processes and decision trees in gambles and more natural decision tasks," J. Psychol. Interdiscip. Appl., vol. 130, no. 3, pp. 329–339, 1996.

[17] M. A. Gómez, S. J. Ibáñez, I. Parejo, and P. Furley, "The use of classification and regression tree when classifying winning and losing basketball teams," Kinesiology, vol. 49, no. 1, pp. 47–56, 2017.

[18] J. Walsh, I. Heazlewood, and M. Climstein, "Application of gradient boosted trees to gender prediction based on motivations of masters athletes," Model Assist. Stat. Appl., vol. 13, no. 3, pp. 235–252, 2018.

[19] S. Aloufi and A. El Saddik, "Sentiment Identification in Football-Specific Tweets," IEEE Access, vol. 6, pp. 78609–78621, 2013.

[20] M. Fernández-Gavilanes, J. Juncal-Martínez, S. García-Méndez, E. Costa-Montenegro, and F. Javier González-Castaño, "Differentiating users by language and location estimation in sentiment analisys of informal text during major public events," Expert Syst. Appl., vol. 117, pp. 15–28, 2019.

[21] Y. Yu and X. Wang, "World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets," Comput. Human Behav., vol. 48, pp. 392–400, 2015.

[22] A. L. F. Alves, C. De S. Baptista, A. A. Firmino, M. G. De Oliveira, and A. C. De Paiva, "A comparison of SVM versus naive-bayes techniques for sentiment analysis in tweets: A case study with the 2013 FIFA confederations cup," WebMedia 2014 - Proc. 20th Brazilian Symp. Multimed. Web, pp. 123–130, 2014.

[23] R. Reza El Akbar, R. N. Shofa, M. I. Paripurna, and Supratman, "The Implementation of Naïve Bayes Algorithm for Classifying Tweets Containing Hate Speech with Political Motive," ICSECC 2019 - Int. Conf. Sustain. Eng. Creat. Comput. New Idea, New Innov. Proc., no. December 2018, pp. 144–148, 2019.

[24] M. Akinkunmi, "Introduction to Statistics Using R," in Synthesis Lectures on Mathematics and Statistics, vol. 11, no. 4, 2019, pp. 305–309.