FIFA PREDICTION AND ANALYSIS BY USING DIFFERENT MACHINE LEARNING MODEL

REPORT SUBMITTED TO

SYMBIOSIS INSTITUTE OF GEOINFORMATICS

FOR PARTIAL FULFILLMENT OF THE M.Sc. DEGREE

BY

SAHIL ABBAS NAQVI

BATCH (2020-2022)

SYMBIOSIS INSTITUTE OF GEOINFORMATICS

SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)

5th FLOOR, ATUR CENTER, GOKHALE CROSSROAD

MODEL COLONY, PUNE – 411016

CERTIFICATE

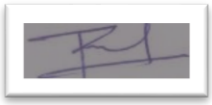
Certified that this report titled **FIFA PREDICTION AND ANALYSIS BY USING DIFFERENT MACHINE LEARNING MODEL**

is a

bonafide work done by Mr. **SAHIL ABBAS NAQVI**

at

Symbiosis Institute of Geoinformatics, Symbiosis International (Deemed University) Pune, under my supervision.



Supervisor (Internal)
Dr. Rajesh Dhumal

Symbiosis Institute of Geoinformatics

Symbiosis International (Deemed University)

INDEX

ACKNOWLEDGMENT

I am thankful to receive assistance from one person in making this project a success.

I take this opportunity to express my deep regards and gratitude to **Mr. Rajesh Dhumal (Professor)** for supporting me throughout the completion of the Project.

I would like to convey my gratefulness towards my dad for their kind and beloved cooperation and encouragement, which eventually helped me complete this Project.

Without your help, I surely will not be able to complete this Project. Thank you so much, sir, for helping me during rough and challenging times.

LIST OF FIGURES

PREFACE

As a part of this football universe and a huge fan of this most loveable sport globally, I had the appropriate knowledge of this domain to deepen and widen practical expertise in this wholesome concept of FIFA prediction and analysis by using different machine learning models, and I learned a lot.

Completing this Project helped me to know more about this concept and the football analytic world. While completing this project, I also learned about cooperation, coordination, and synergy of mind and body.

This report contains:

Exploratory Data Analysis: Investigate correlations, the importance of features to results, Hypothesis Testing.

Methodology: How I carried out this Project, which experiments I did.

Models: baseline model, logistic regression, random forest, gradient boosting tree, ADA boost tree, Neural Network. – Evaluation Criteria: F1, 10-fold cross-validation accuracy – Results and Conclusion

I hope you will find my report interesting. All constructive critics and any kind of feedback are cordially invited.

INTRODUCTION

In this work, I compared nine different modeling approaches for the soccer matches and goal differences on all international tournaments from 2005 – 2017, FIFA World Cup 2010 – 2014, and FIFA EURO 2012-2016. Within this comparison, while the performance of "**Win / Draw / Lose**" predictions shows little difference, "**Goal Difference**" prediction is quite favored to Random Forest and squad-strength based decision tree. We also apply these models in World Cup 2018, and again, Random Forest and Logistic Regression predict about 33% accuracy for "Goal Difference" and about 57% for "Win / Draw / Lose." However, a simple decision tree based on bet odd and squad strength is also comparable.

Objective: -

Feature Engineering: To determine who will be more likely to win a match, based on my knowledge, I came up with four main groups of features as follows:

1. head-to-head match history between 2 teams
2. recent performance of each team (10 recent matches), aka "form."
3. bet ratio before matches
4. squad strength (from FIFA video game)

PROJECT LIFECYCLE

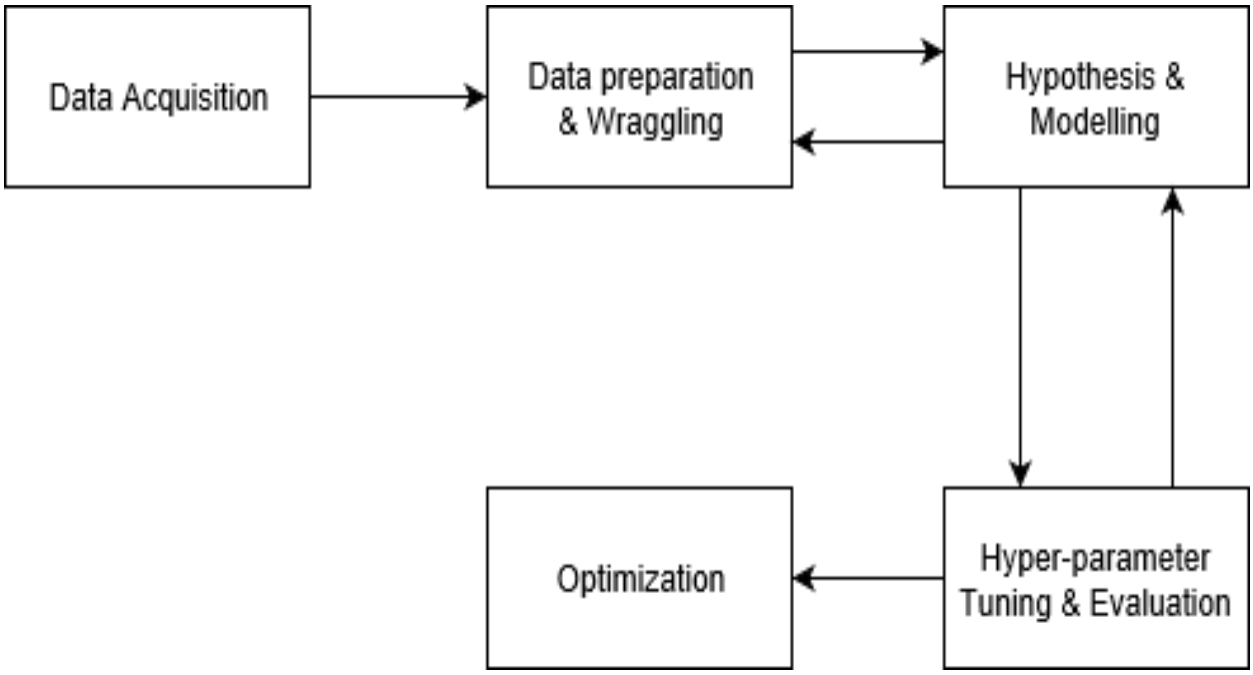


Fig.1 Project Lifecycle

DATA

The dataset is from all international matches from 2000 – 2018, results, bet odds, ranking, squad strengths.

1. FIFA WORLD CUP 2018
2. INTERNATIONAL MATCHES 1872-2018
3. FIFA RANKING THROUGH TIME
4. BET ODD
5. BET ODD 2
6. SQUAD STRENGTH – SOFIA
7. SQUAD STRENGTH – FIFA INDEX

Feature Selection: To determine who will be more likely to win a match, based on my knowledge, I came up with four main groups of features as follows:

1. **Head-to-head match history between 2 teams**. Some teams have few opponents who hardly win, no matter how strong they currently are. For example, the German squad usually loses / could not beat the Italian team in 90-minute matches.
2. **A recent performance of each team (10 recent matches), aka "form"** A team with "good" form, usually has a higher chance to win the next matches.
3. **Bet ratio before matches** odd bookmarkers already did many analyses before competitions to select the best betting odds, so why don't we include them.
4. **Squad strength (from FIFA video game).** We want a real squad strength, but these data are not free and not always available, so we use the strength from FIFA video games updated regularly to catch up with the natural strength.

Feature List: The feature list reflects those four factors.

- *Difference: team1 – team2
- *Form: performance in 10 recent matches

Feature Name	Description
team_1	Nation Code (e.g., US, NZ)
team_2	Nation Code (e.g., US, NZ)
date	Date of the match (yyyy – mm – dd)
tournament	Friendly, EURO, AFC, FIFA WC
h_win_diff	Head2Head: win difference
h_draw	Head2Head: number of draws
form_diff_goalF	Form: difference in “Goal For”
form_diff_goalA	Form: difference in “Goal Against”
form_diff_win	Form: difference in the number of wins
form_diff_draw	Form: difference in the number of draws
odd_diff_win	Betting Odd: difference bet rate for win
odd_draw	Betting Odd: bet rate for the draw
game_diff_rank	Squad Strength: difference in FIFA Rank
game_diff_ovr	Squad Strength: difference in Overall Strength
game_diff_attk	Squad Strength: difference in Attack Strength
game_diff_mid	Squad Strength: difference in Midfield Strength
game_diff_def	Squad Strength: difference in Défense Strength
game_diff_prestige	Squad Strength: difference in prestige
game_diff_age11	Squad Strength: difference in age of 11 starting players
game_diff_ageAll	Squad Strength: difference in age of all players
game_diff_bup_speed	Squad Strength: difference in Build Up Play Speed
game_diff_cc_pass	Squad Strength: difference in Chance Creation Passing

game_diff_cc_cross	Squad Strength: difference in Chance Creation Crossing
game_diff_cc_shoot	Squad Strength: difference in Chance Creation Shooting
game_diff_def_press	Squad Strength: difference in defense Pressure
game_diff_def_aggr	Squad Strength: difference in Défense Aggression
game_diff_def_teamwidth	Squad Strength: difference in defense Team Width

EXPLORATORY DATA ANALYSIS

There are few questions to understand data better.

IMBALANCE OF DATA

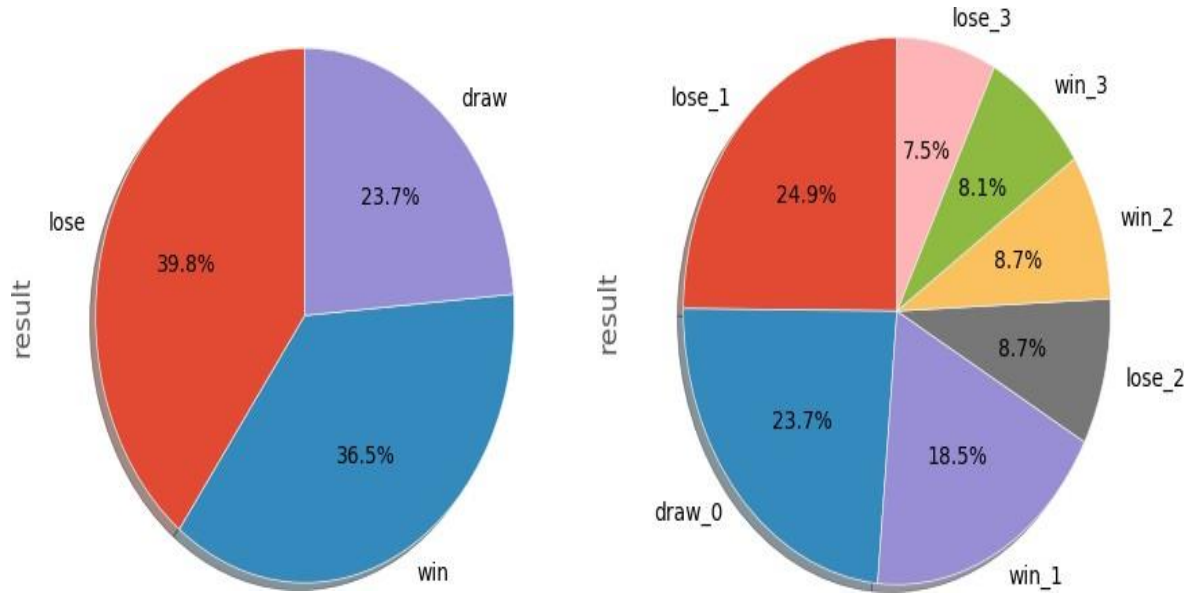


Fig. 2 Showing class imbalance

CORRELATION BETWEEN VARIABLES

First, we draw a correlation matrix of a large dataset that contains all matches from 2005-2018 with features groups 1, 2, and 3.

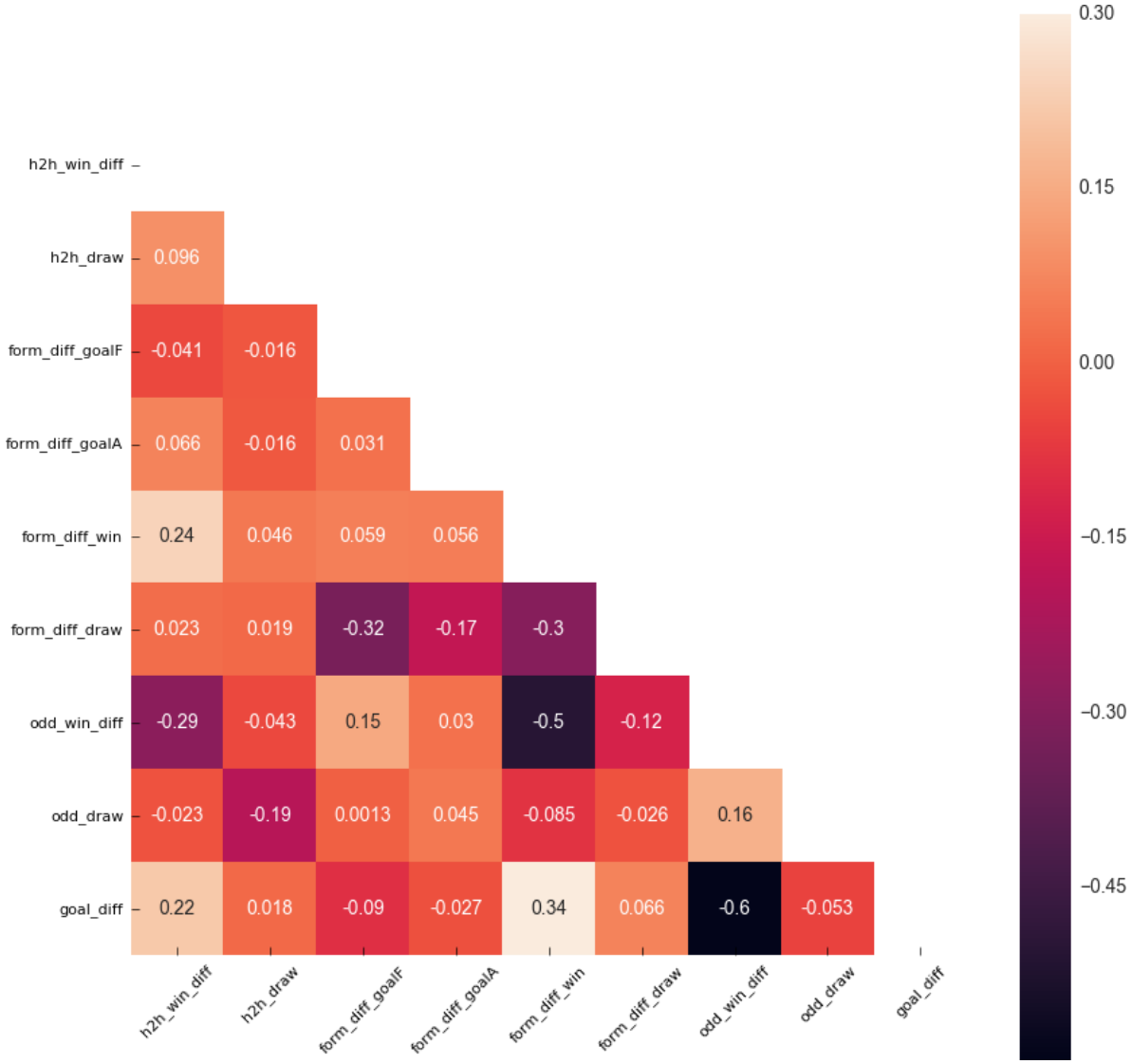


Fig.3 Correlation Matrix

In general, features are not correlated. “Odd_win_diff” is quite negatively correlated with “form_diff_win” (-0.5), indicating that the form of two teams reflects the belief of odd bookmarkers on winners. One more interesting point is that when the difference of bet odd increases, we would see more goal differences (correlation score = -0.6)

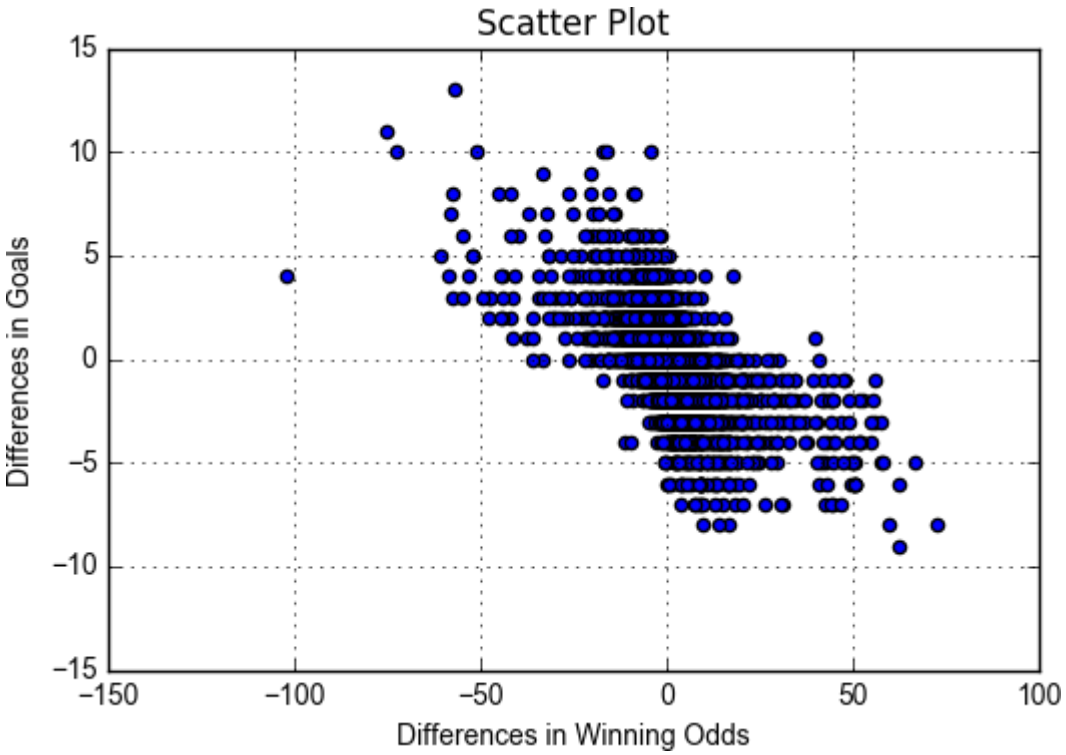


Fig.4 Winning Odds v/s Goals

Second, we draw a correlation matrix of the small dataset, which contains all matches from World Cup 2010, 2014, 2018, and EURO 2012, 2016

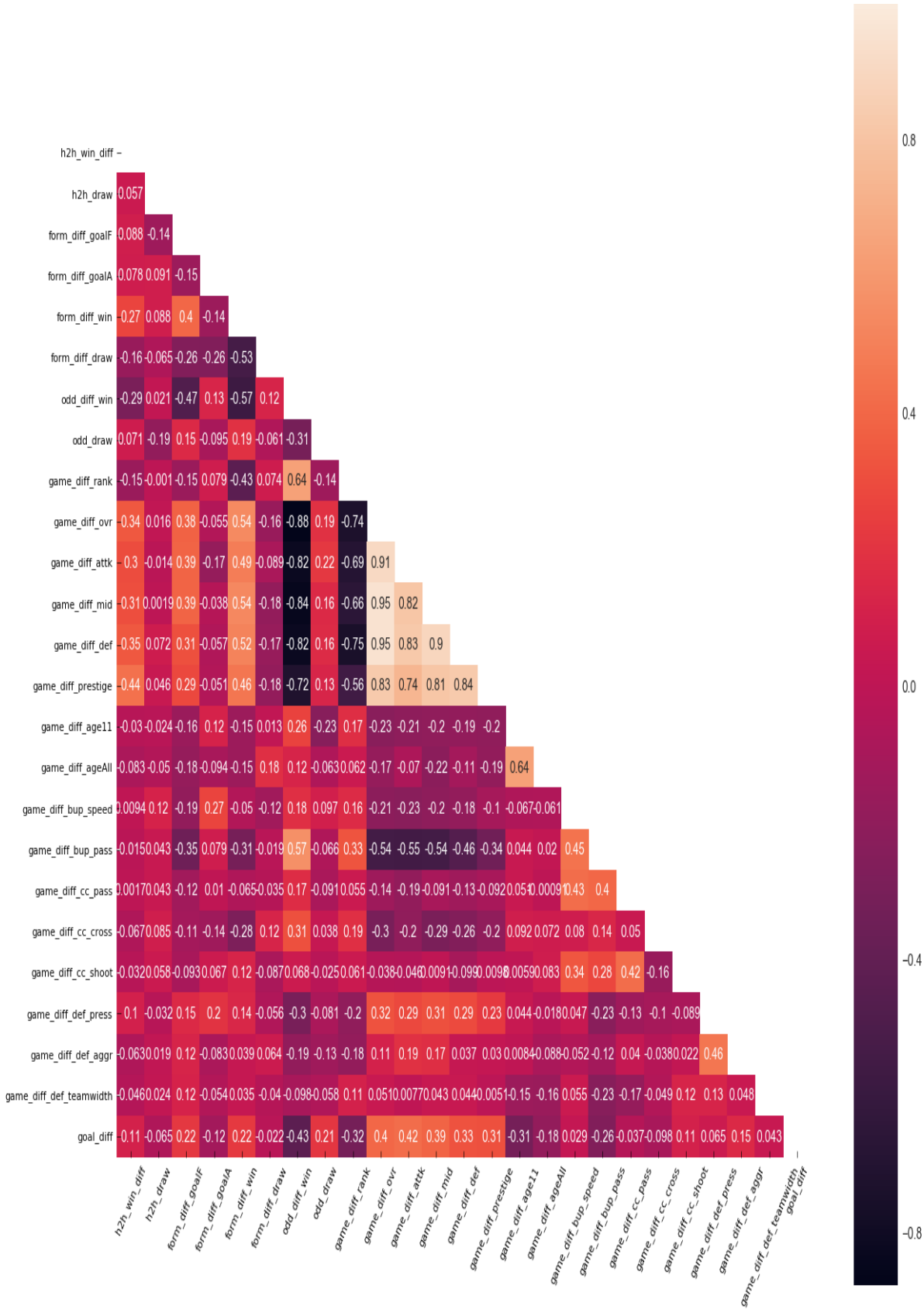


Fig.5 Correlation Matrix All

Overall Rating is just an average of “attack," “defense," and “midfield” index. Therefore, we see a high correlation between them. In addition, some of the new features of squad strength show a high correlation, for example, “FIFA Rank," “Overall rating," and “Difference in winning odd."

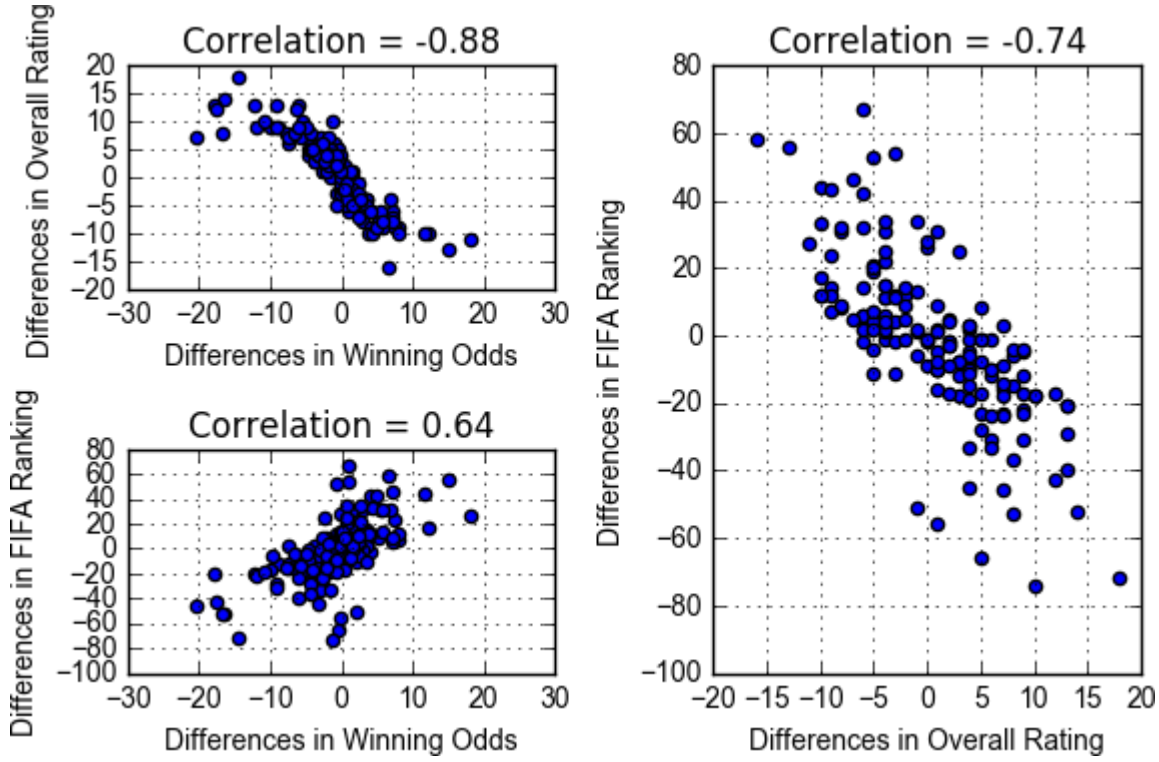


Fig.6 Rank v/s Rating

HOW HEAD – TO – HEAD MATCHUP HISTORY AFFECT THE CURRENT MATCH?

You may think when the head-to-head win difference is positive. The match result should be “Win” (Team 1 wins Team 2) and vice versa. When the head-to-head win difference is negative, the match result should be “Lose” (Team 2 wins Team 1). A positive head-to-head win difference indicates a 51.8% chance the match results end up with “Win," and a negative head-to-head win difference suggests a 55.5% chance the match results end up with “Lose."

Let us perform our hypothesis testing with two-sampled t-test Null Hypothesis: There is no difference of „h2h win difference“ between “Win” and “Lose” Alternative Hypothesis: There are differences of „h2h win difference“ between “Win” and “Lose."

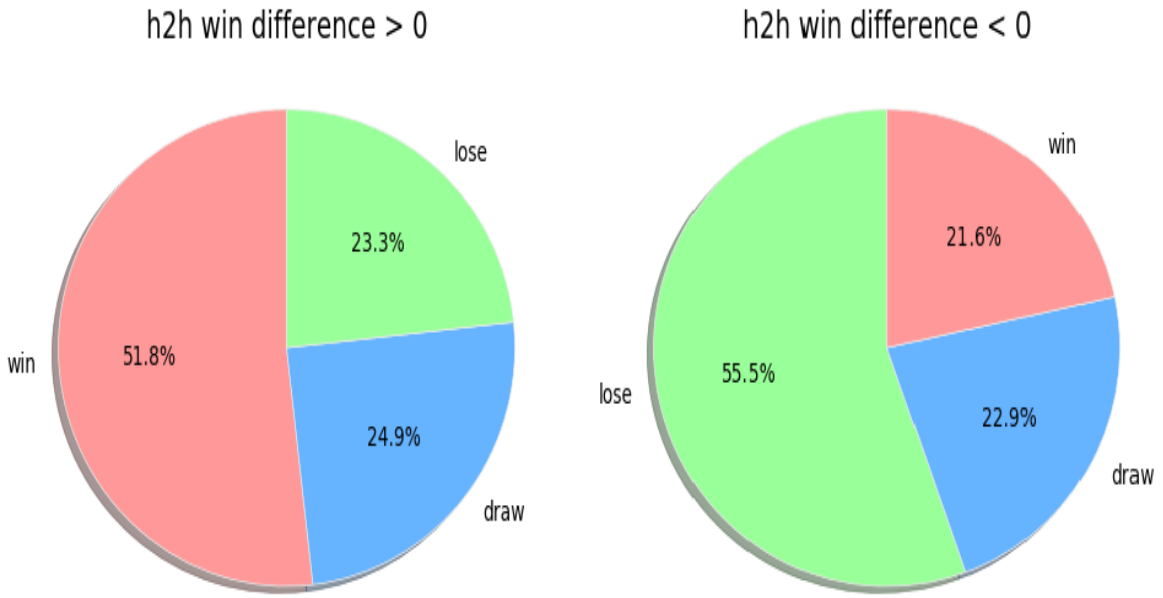


Fig.7

Head-to-Head positive win difference **Head to Head negative win difference**

Let us perform our hypothesis testing with two-sampled t-test Null Hypothesis: There is no difference of „h2h win difference“ between “Win” and “Lose” Alternative Hypothesis: There are differences of „h2h win difference“ between “Win” and “Lose."

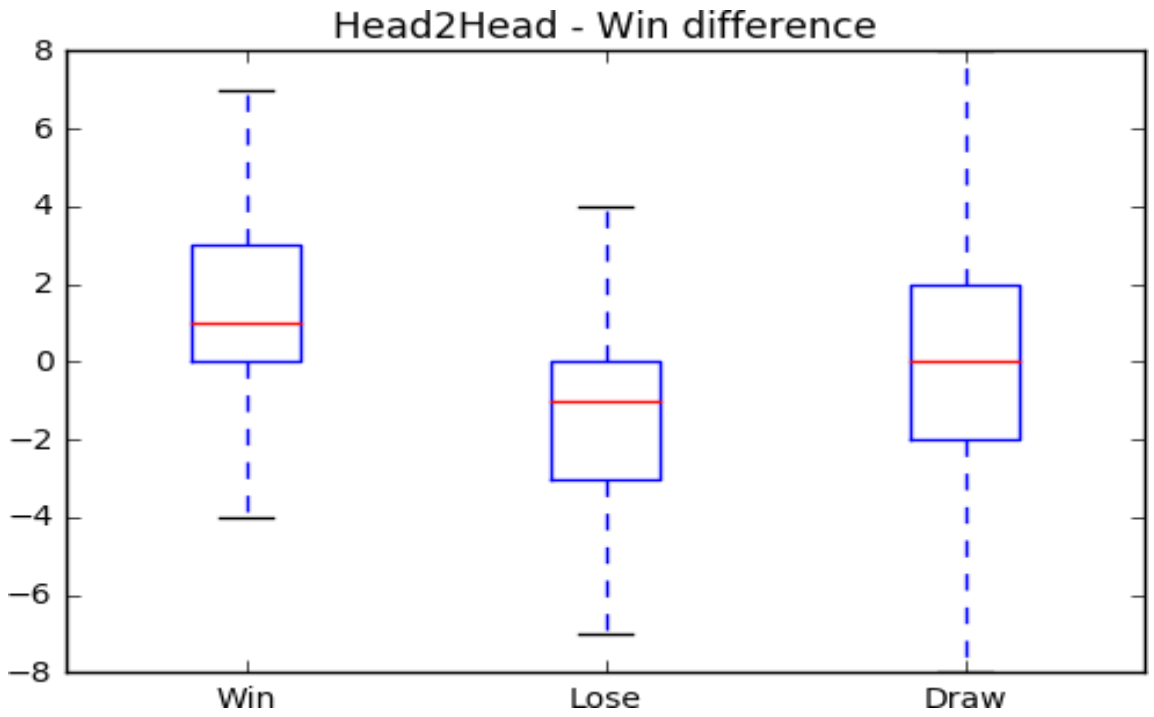


Fig.8 (Head-to-Head win difference Box Plot)

T-test between win and lose:
Ttest_indResult (**statistic=24.30496036405259**, **pvalue=2.503882847793891e-126**)

A very small p-value means we must reject the null hypothesis and accept the alternative hypothesis. Also, we can do the same procedure with win-draw and lose-draw.

T-test between win and draw:
Ttest_indResult (**statistic=7.8385466293651023, pvalue=5.395456011352264e-15**)

T-test between lose and draw:
Ttest_indResult (**statistic=-8.6759649601068887, pvalue=5.2722587025773183e-18**)

Therefore, we can say **the history of head-to-head matches of two teams contributes significantly to the result.**

HOW DOES 10-RECENT PERFORMANCE AFFECT THE CURRENT MATCH?

We consider differences in “Goal For” (how many goals they got), “Goal Against” (how many goals they conceded), “number of winning matches," and “number of drawing matches." We performed the same procedure as the previous questions. From pie charts, we can see a clear distinction in the “number of wins” where the proportion of the “Win” result decreases from 49% to 25% while the “Lose” result increases from 26.5% to 52.3%.

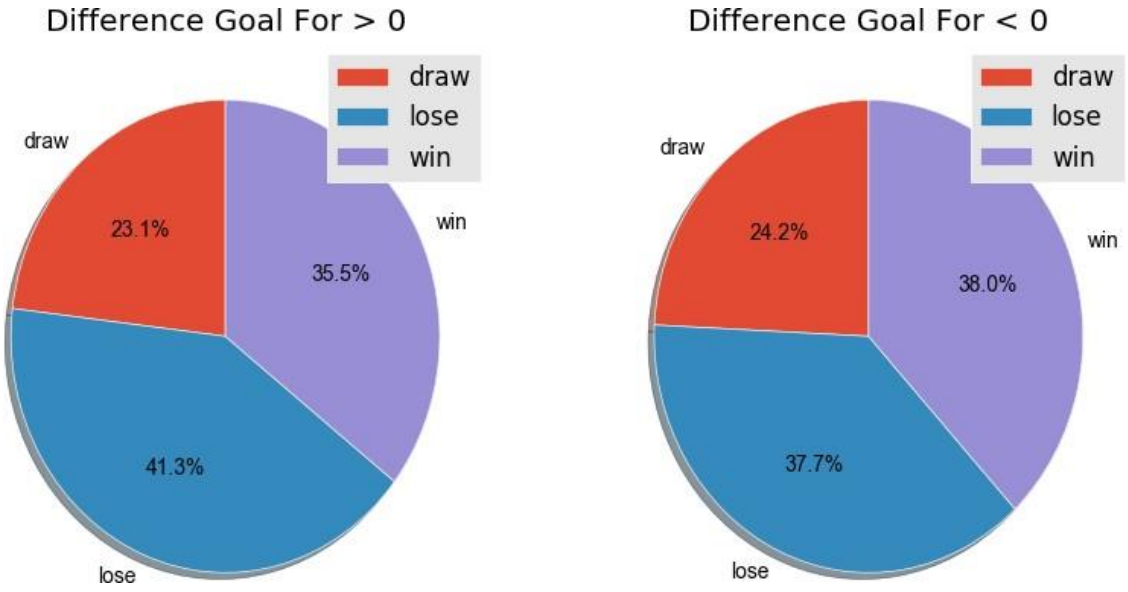


Fig.9

Form positive goal difference **Form negative goal difference.**

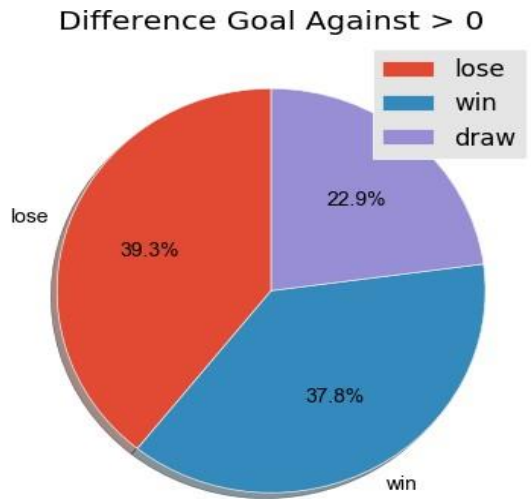


Fig.10

Form positive against goal **Form negative against the goal**

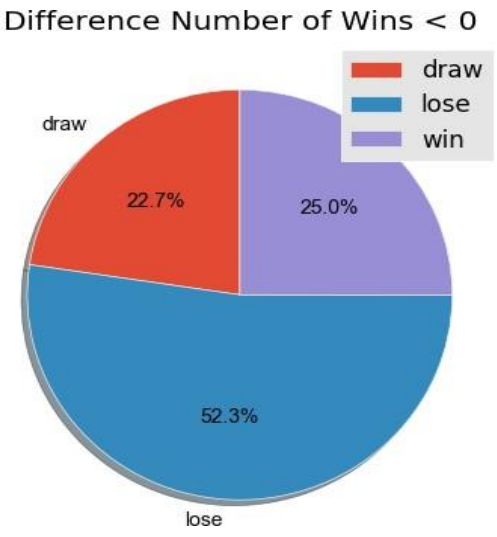
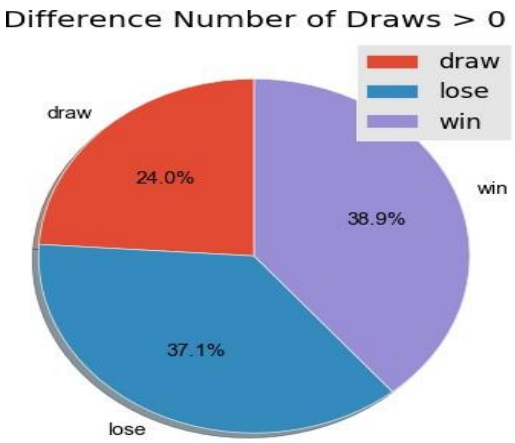


Fig.11

Form positive number of wins **Form positive number of wins.**



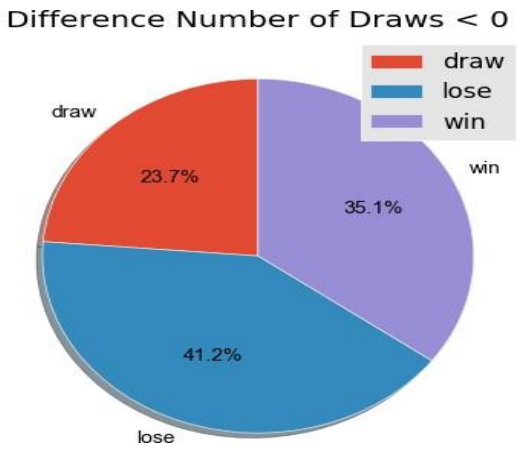


Fig.12

Form positive number of draws **Form negative number of draws.**

Pie charts are not enough. We should do hypothesis testing to see the significance of each feature.

Feature Name	**t-test between „win“ and „lose“**	**t-test between „win“ and „draw“**	**t-test between „lose“ and „draw“**
Goal For	pvalue = 2.50e-126	pvalue = 5.39e-15	pvalue = 5.27e-18
Goal Against	pvalue = 0.60	pvalue = 0.17	pvalue = 0.08
Number of Winning Matches	pvalue = 3.02e-23	pvalue = 1.58e-33	pvalue = 2.57e-29
Number of Draw Matches	pvalue = 1.53e-06	pvalue = 0.21	pvalue = 0.03

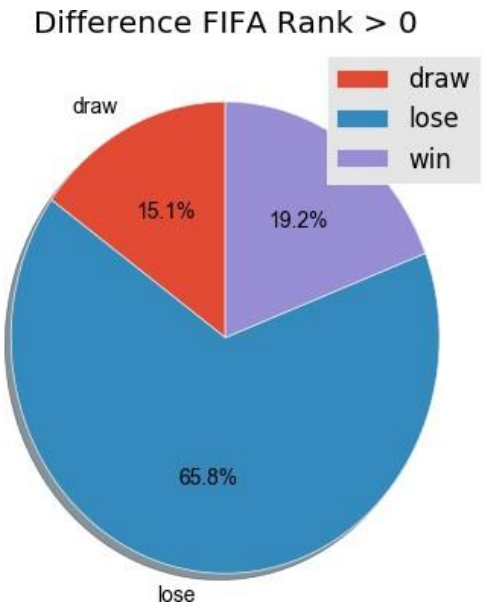
Table.1 T-TEST

We see many small values of p-value in cases of “Goal For” and “Number of Winning Matches." Based on the t-test, **we know the difference in “Goal For” and “Number of Winning Matches” are helpful features.**

DO STRONGER TEAMS USUALLY WIN?

We define stronger teams based on

- Higher FIFA Ranking
- Higher Overall Rating



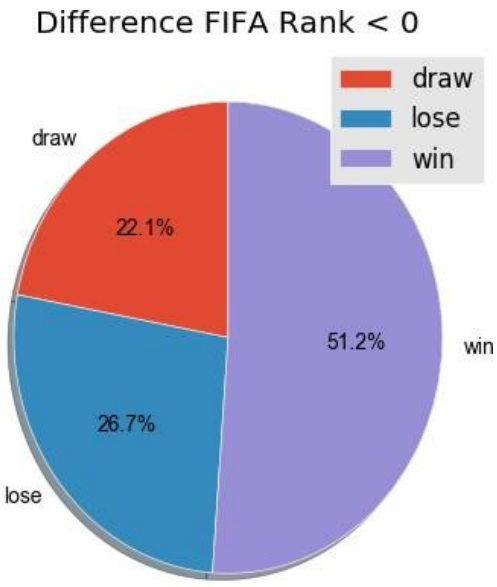
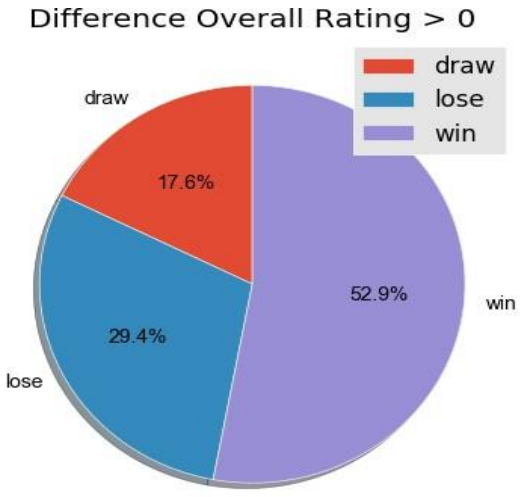


Fig.13

Positive FIFA ranking difference **Positive FIFA ranking difference**



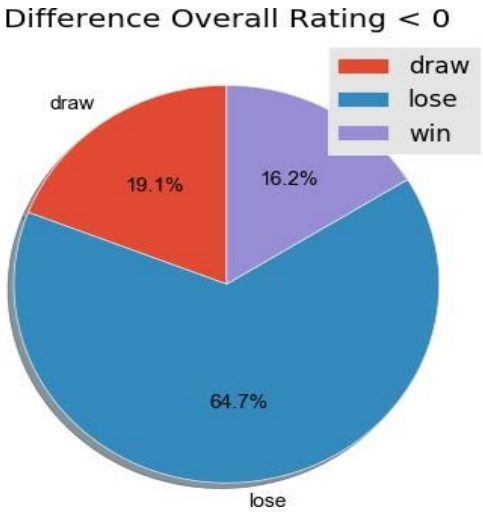


Fig.14

Positive overall FIFA ranking **Negative overall FIFA ranking**

Feature Name	t-test between „win“ and „lose“	t-test between „win“ and „draw“	t-test between „lose“ and „draw“
FIFA Rank	pvalue = 2.11e-10	pvalue=0.65	pvalue=0.00068
Overall Rating	pvalue = 1.53e-16	pvalue = 0.0804	pvalue = 0.000696

Table.2 T-TEST

DO YOUNG PLAYERS PLAY BETTER THAN OLD ONES?

Young players may have better stamina and more energy, while older players have more experience. We want to see how age affects match results.

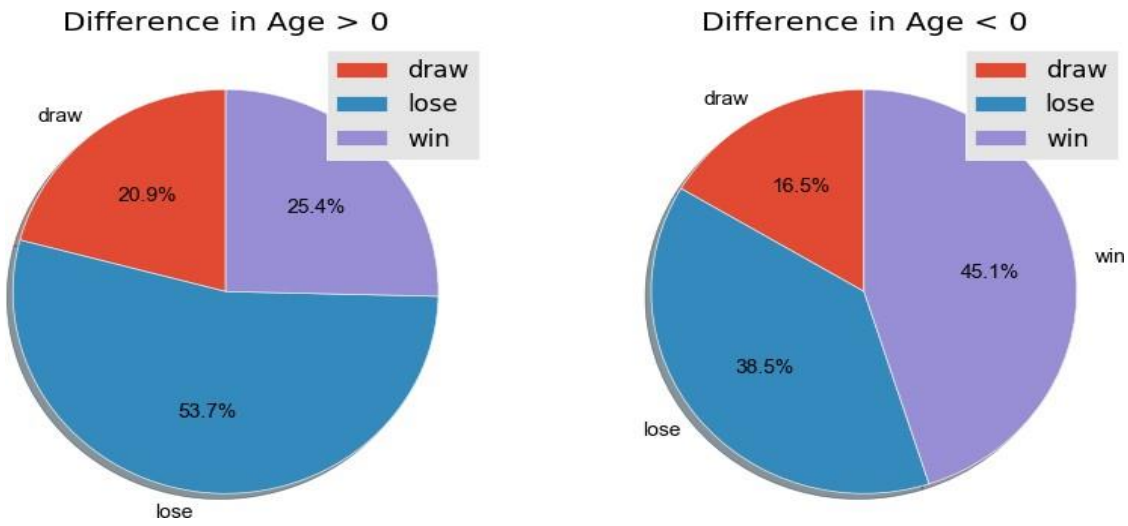


Fig.15

Game Age Difference **Game Age Difference**

Feature Name	t-test between „win“ and „lose“	t-test between „win“ and „draw“	t-test between „lose“ and „draw“
Age	pvalue = 2.07e-05	pvalue = 0.312	pvalue=0.090

Table.3 T-TEST

Based on the t-test and pie chart, we know that age contributes significantly to the result. More specifically, younger teams tend to play better than older ones.

IS A SHORT PASS BETTER THAN A LONG PASS?

A higher value of “Build Up Play Passing” means “Long Pass," and a lower value means “Short Pass," value in the middle means “Mixed-Type Pass."

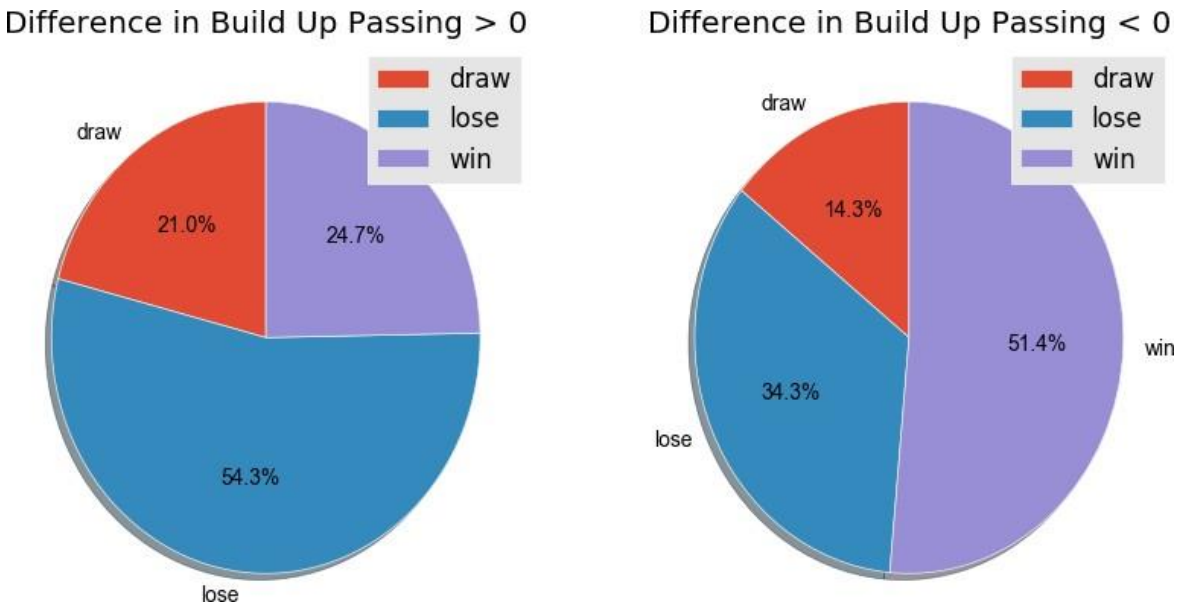


Fig.16

Build-up pass difference (Short) **Build-up pass difference (Long)**

Feature Name	t-test between „win“ and „lose“	t-test between „win“ and „draw“	t-test between „lose“ and „draw“
Age	pvalue = 1.05e-07	pvalue = 0.0062	pvalue = 0.571

Table.4 T-TEST

Based on the t-test and pie chart, we know that age contributes significantly to the result. More specifically, teams that reply on “Longer Pass” usually lose the game.

HOW ARE LABELS DISTRIBUTE IN REDUCED DIMENSIONS?

For this question, we use PCA to pick two first principal components which best explain the data. Then we plot data in a new dimension.

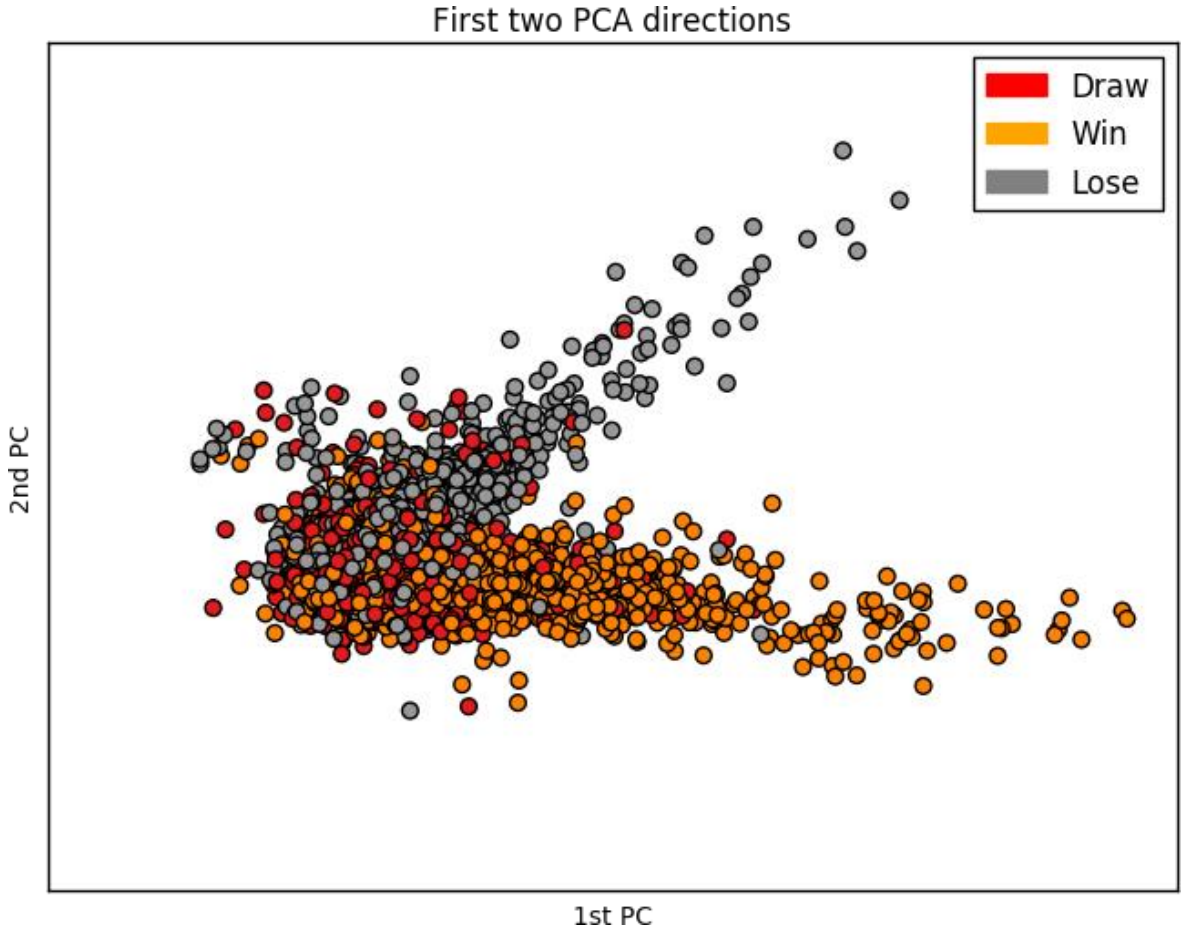


Fig.17 Principal Component Analysis

While “Win” and “Lose” are while separate, “Draw” seems to be mixed between other labels.

METHODOLOGY

Our main objectives of prediction are “Win / Lose / Draw” and “Goal Difference." In this work, we do two main experiments. For each experiment, we follow this procedure.

- Split data into 70:30
- First, we perform “normalization” of features, convert category to the number.
- Second, we perform k-fold cross-validation to select the best parameters for each model based on some criteria.
- Third, we use the best model to predict 10-fold cross-validation (In which we reserve nine folds for training and 1-fold for testing) to achieve the mean of test error. This error is more reliable.

Experiment 1: Build classifiers for “Win / Lose / Draw” from 2005. Because feature “Bet Odds” is only available after 2005, we only conduct experiments for this period.

Experiment 2: Build classifiers for “Goal Difference” for “World Cup” and “UEFA EURO” after 2010. The reason is that features of “Squad Strength” are not always available before 2010. Some national teams do not have a database of squad strength in FIFA Video Games. We know that tackling prediction with regression would be challenging, so we turn “Goal Difference” into classification by defining labels as follows:

Team A vs. Team B

- “win_1”: A wins with one goal differences
- “win_2”: A wins with two-goal differences
- “win_3”: A wins with three or more goal differences
- “lose_1”: B wins with 1 goal differences
- “lose_2”: B wins with 2 goal differences
- “lose_3”: A wins with three or more goal differences
- “draw_0”: Draw

Experiment 3: In addition, we want to test how our trained model in **Experiment 2** predicts the “Goal Difference” and “Win/Draw/Lose” of World Cup 2018.

MODELS

Baseline Model: In the EDA part, we already investigate the importance of features and see that odd, history, form, and squad strength are significant. We divide features into three groups: odd, h2h-form, squad strength, and build "Baseline Models" based on these groups. To keep the baseline model simple, we set hyper-parameter of Decision Tree maximum depth = 2, maximum leaf nodes = 3

1. **Odd-based model:**

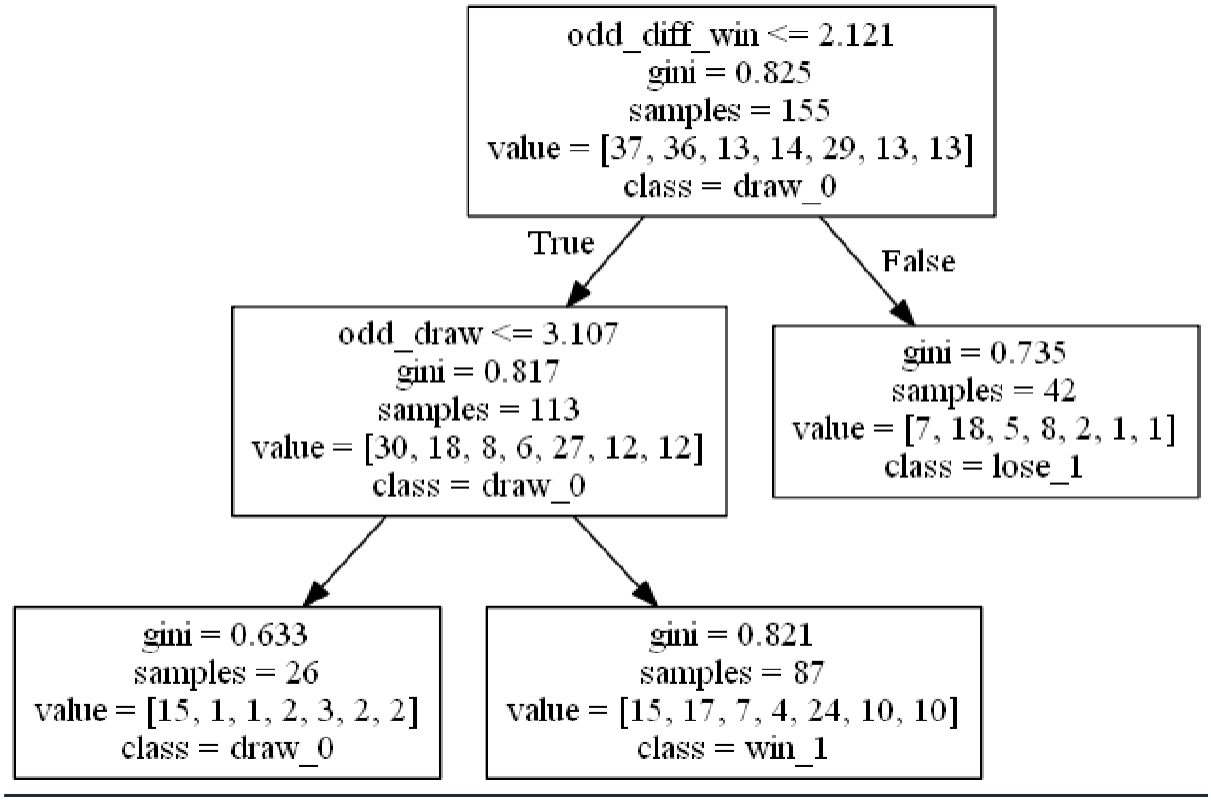


Fig. 18 Tree Odd-based model

2. **History-Form-based model:**

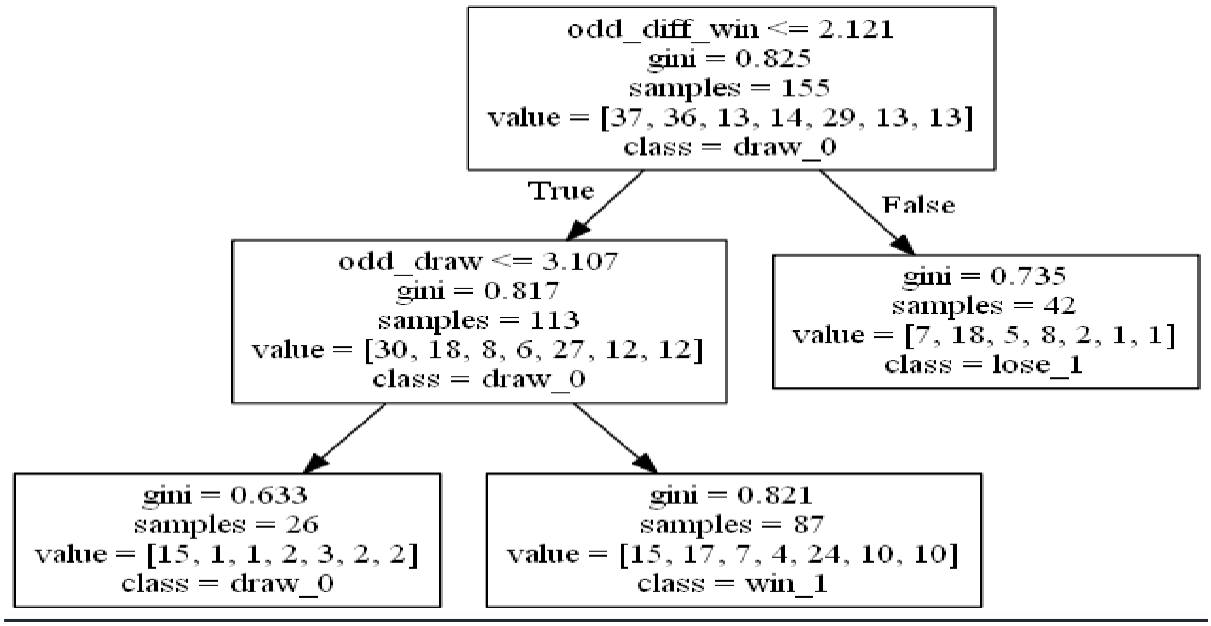


Fig. 19 Tree Head-to-Head (Form) based model

3. **Squad-strength based model:**

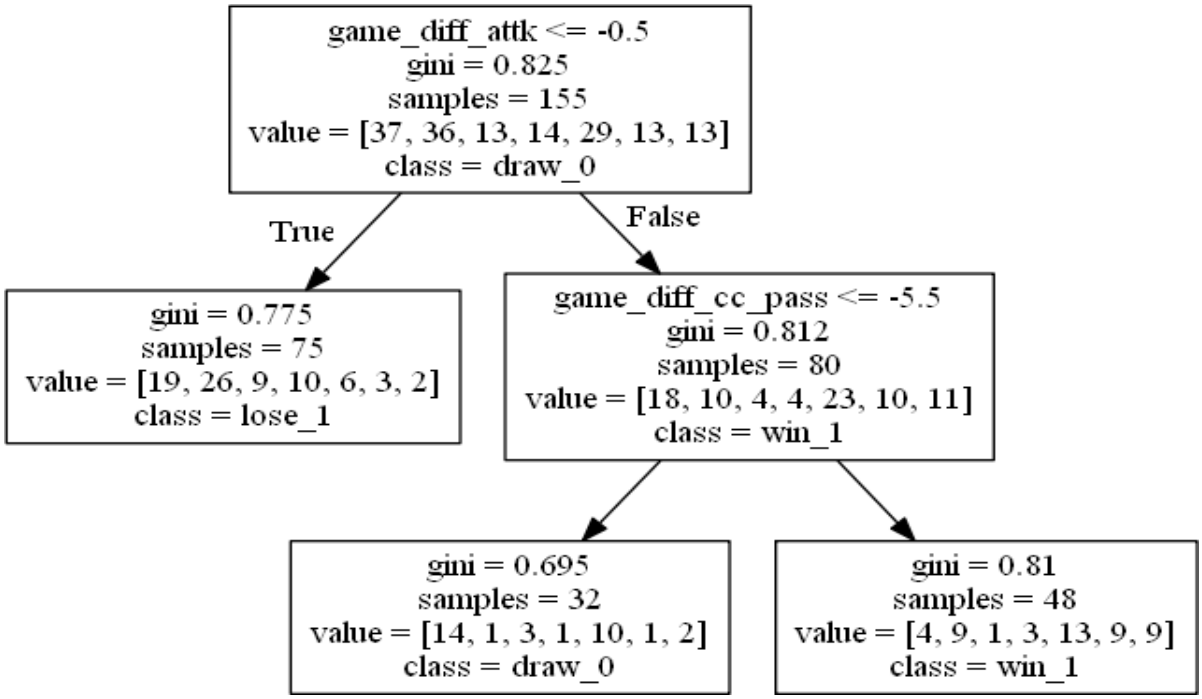


Fig. 20 Tree Squad Strength-based model

Enhanced Model: To beat the baseline models, we use all features, and several machinealgorithms as follows

1. Logistic Regression
2. Random Forest
3. Gradient Boosting Tree
4. ADA Boost Tree
5. Neural Network
6. LightGBM

EVALUATION CRITERIA

Models are evaluated on these criteria, which are carried out for each label “win," “lose," and “draw."

- **Precision:** Among our prediction of “True” value, how many percentages did we hit? In that case, the higher value we get, the better prediction it is.
- **Recall:** Among actual “True” values, how many percentages do we hit? In that case, the higher value we hit, the better prediction we get.
- **F1:** A balance of Precision and Recall. In that case, the higher value we get, the better prediction it is. There are two types of F1 Scores.
    - **F1-micro:** compute F1 by aggregating True Positive and False Positive of each class.
    - **F1-macro:** compute F1 for each class independently and then take the mean of all classes.

If we have a different-class classification setup, micro-mean is most preferable if you doubt that there might be a class disparity (i.e., you may have many different examples of one class than that of other classes). In this case, we should stick with F1-micro.

- **10-fold cross-validation accuracy:** Mean of accuracy for each cross-validation fold. This is a reliable estimation of test error of model evaluation (no need to split to train and test)
- **The area under ROC:** For indulging in binary classification experiment, True Positive Rate // False Positive Rate for all thresholds.

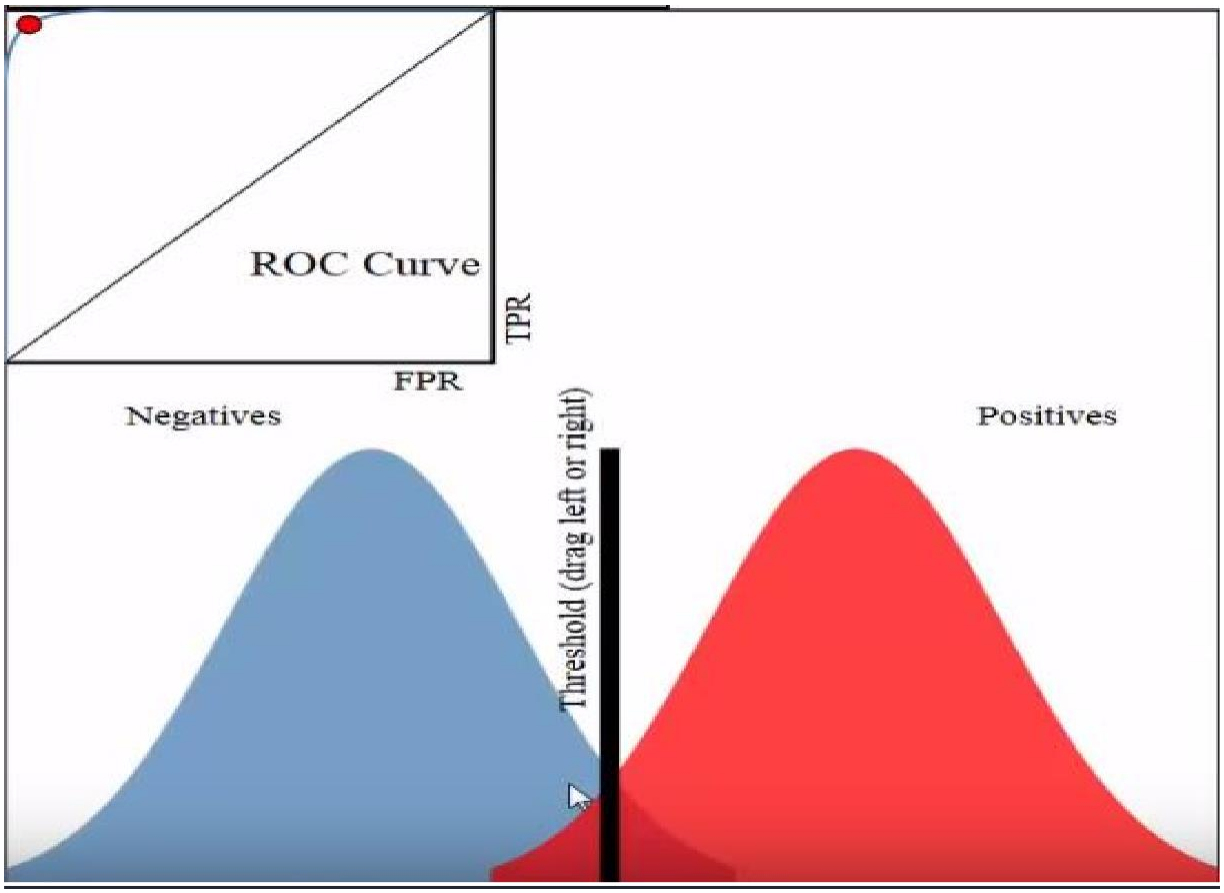


Fig. 21 ROC Introduction

RESULTS

Experiment 1 "Draw / Lose /Win"

1. **Odd-based Decision Tree:**

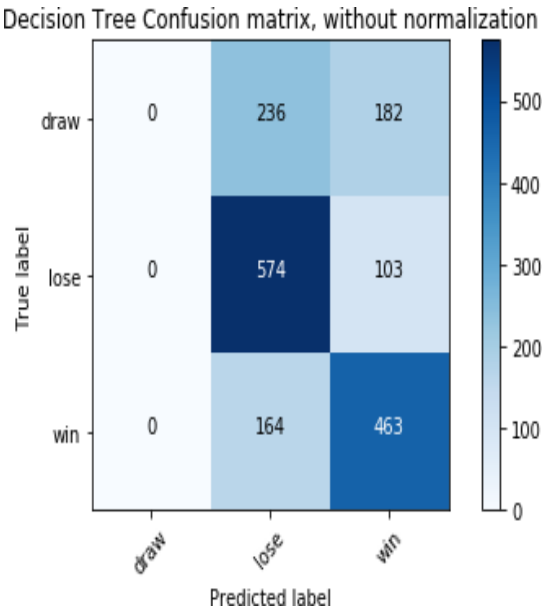


Fig. Confusion Matrix Odd

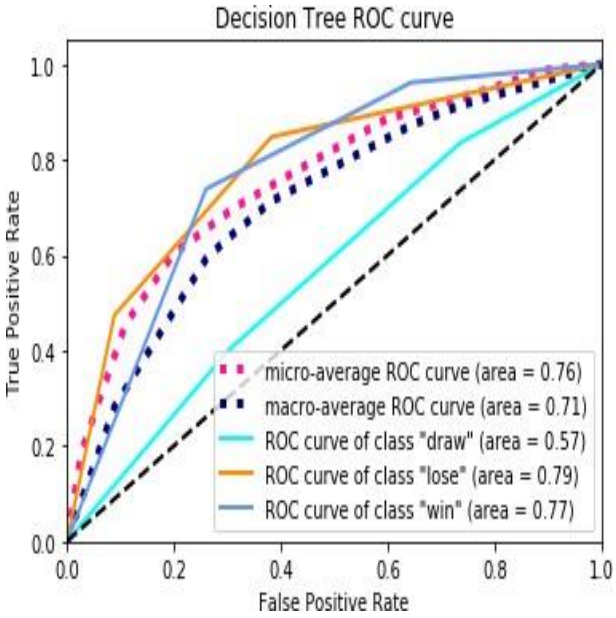
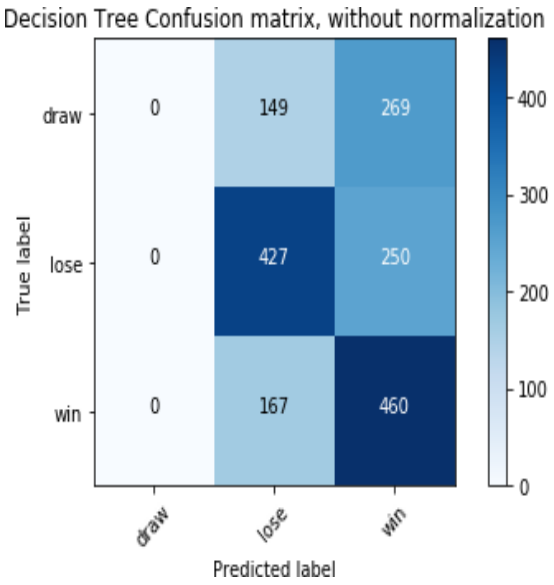


Fig. ROC curve Odd

The odd-based decision tree model confusion matrix and ROC curve are shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **60.22**, and from the ROC curve area was found to be **0.76** micro-average.

2. h2h-Form-based Decision Tree:



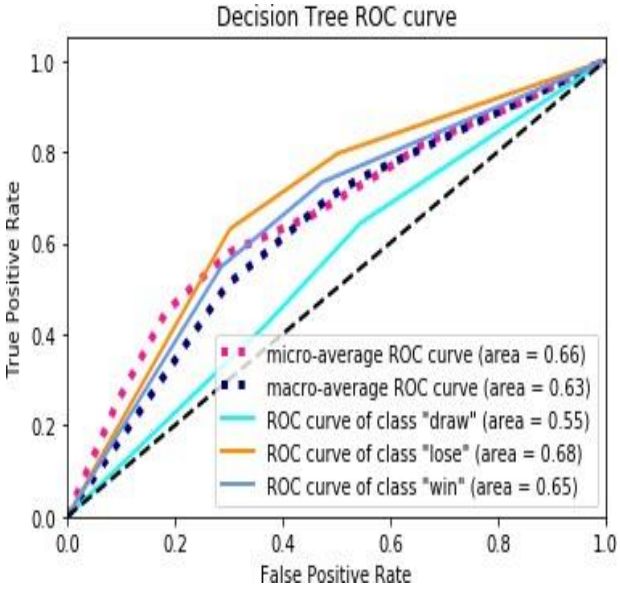
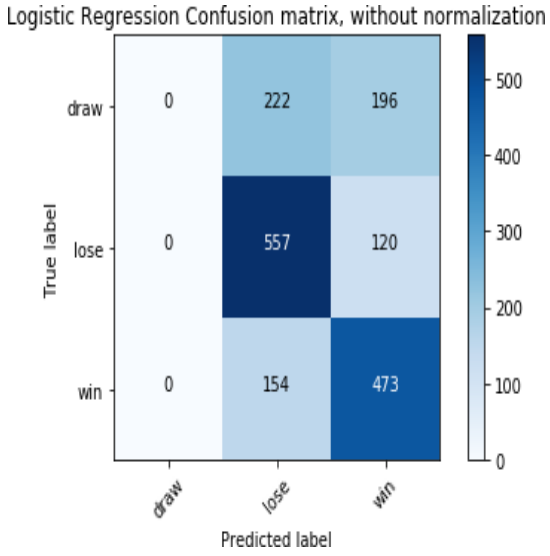


Fig. Confusion Matrix Head-to-Head **Fig. ROC curve Head-to-Head**

H2H form-based decision tree model confusion matrix and ROC curve are shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **51.52**, and from the ROC curve area was found to be **0.66** micro-average.

3. Logistic Regression:



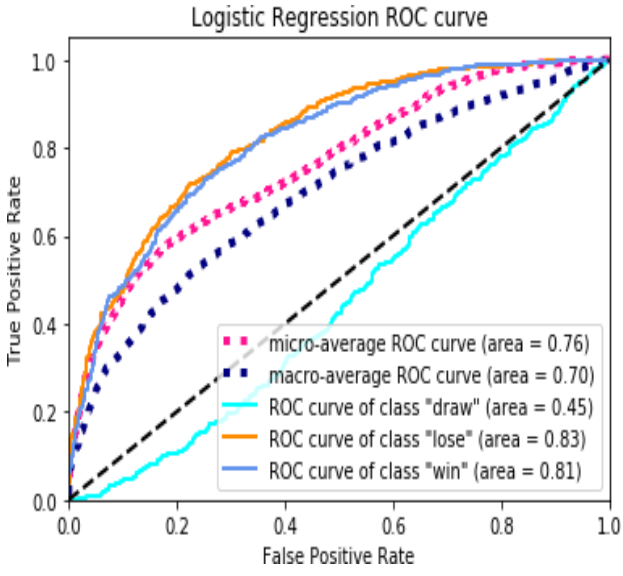
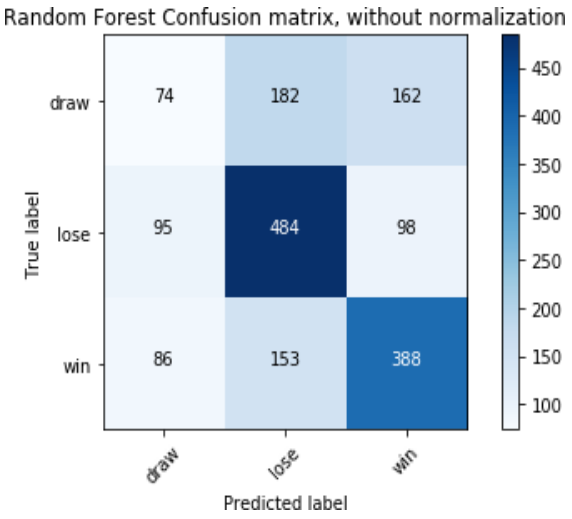


Fig. Confusion Matrix Logistic Regression **Fig. ROC curves Logistic Regression.**

The logistic Regression model confusion matrix and ROC curve are shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **59.87**, and from the ROC curve area was found to be **0.76** micro-average.

4. Random Forest:



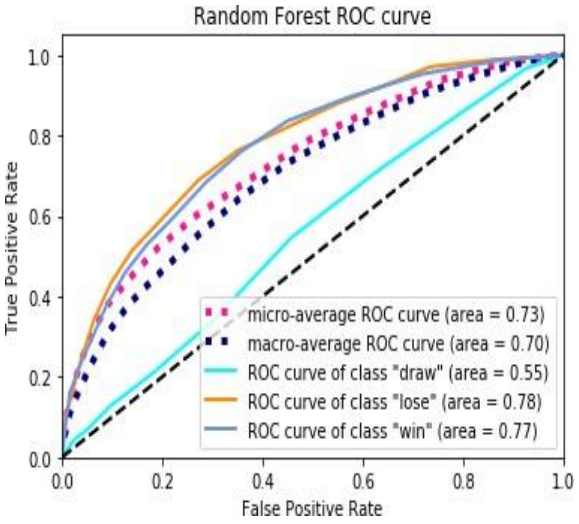
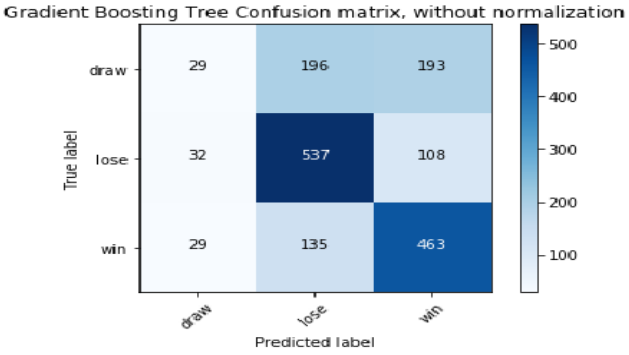


Fig. Confusion Matrix Random Forest **Fig. ROC curve Random Forest**

Random Forest model confusion matrix and ROC curve are shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **55.92**, and from the ROC curve area was found to be **0.74** micro-average.

5. Gradient Boosting tree:



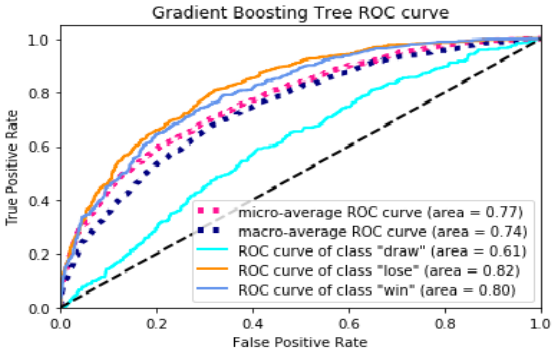
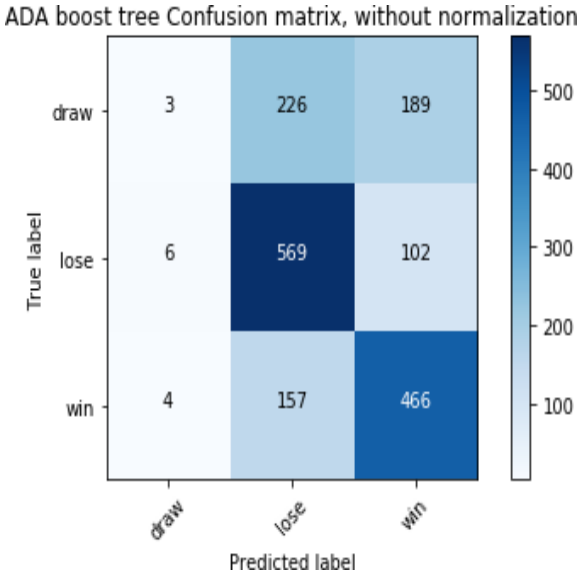


Fig. Confusion Matrix Gradient Boosting Tree **Fig. ROC curve Gradient Boosting Tree**

Gradient Boosting Tree model confusion matrix and ROC curve are shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **59.47**, and from the ROC curve area was found to be **0.77** micro-average.

6. ADA boost tree:



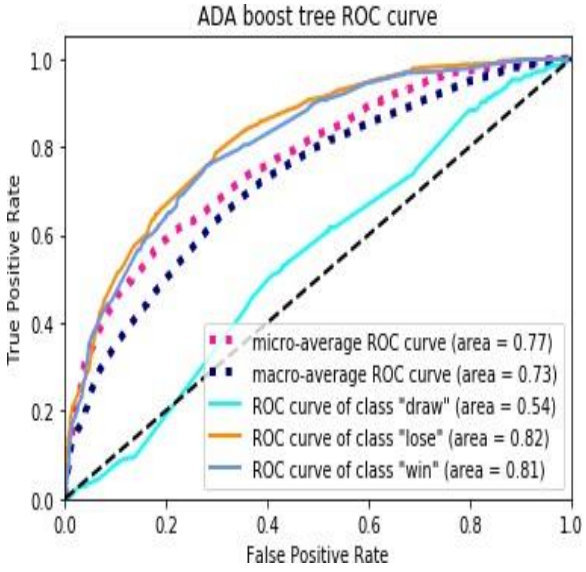
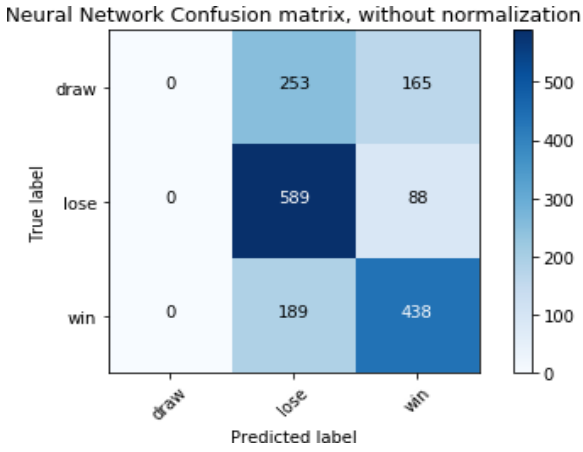


Fig. Confusion Matrix ADA boosts Fig. ROC curve ADA boost.

ADA Boost Tree model confusion matrix and ROC curve are shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **60.22**, and from the ROC curve area was found to be **0.77** micro-average.

7. Neural Network:



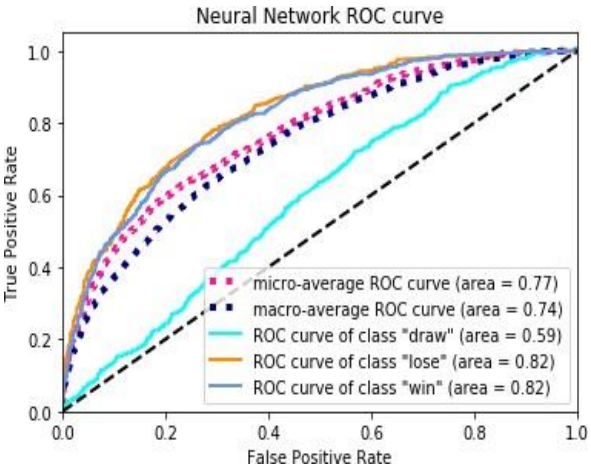


Fig. Confusion Matrix NN **Fig. ROC curve NN**

The Neural Network model confusion matrix and ROC curve are shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **58.36**, and from the ROC curve area was found to be **0.77** micro-average.

8. Light GBM:

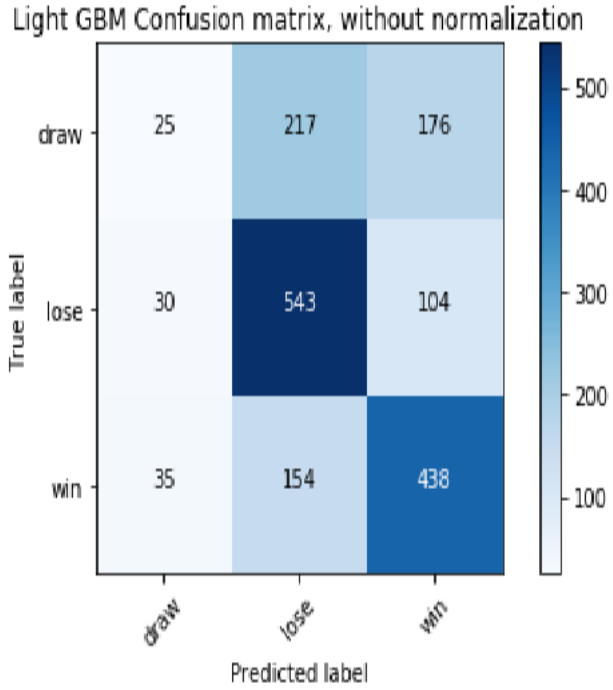


Fig. Confusion Matrix Light GBM

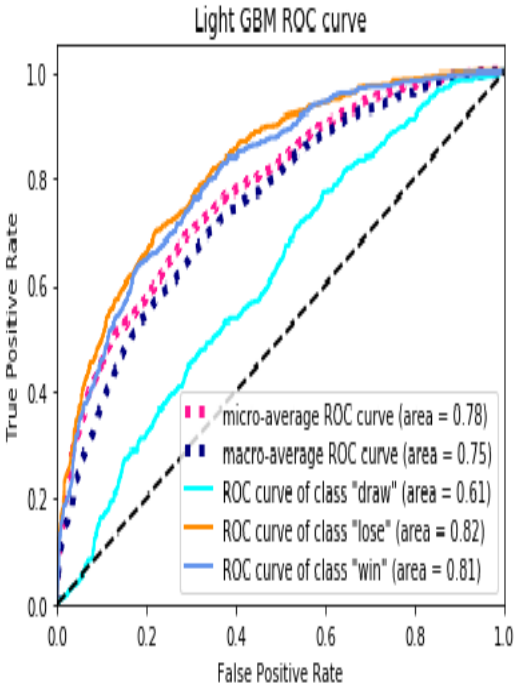


Fig. ROC curve Light GBM

Light GBM model confusion matrix and ROC curve are shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **60.28**, and from the ROC curve area was found to be **0.78** micro-average.

Model	10-fold CV accuracy (%)	F1 - micro average	AUROC - micro average
Odd-based Decision Tree	59.28	60.22	0.76
H2H-Form based Decision Tree	51.22	51.52	0.66
Logistic Regression	59.37	59.87	0.76
Random Forest	54.40	55.92	0.74
Gradient Boosting tree	58.60	59.47	0.77
ADA boost tree	59.08	60.22	0.77
Neural Net	58.96	58.36	0.77
LightGBM	59.49	60.28	0.78

Table.5 Model Accuracy Table

Experiments 1 show slight improvement between enhanced and baseline models based on three evaluation criteria: 10-fold cross-validation, F1, and Area Under Curve. A simple Odd-based Decision Tree is enough to classify Win/Draw/Lose. However, according to the confusion matrix in Appendix of experiment 1, we see that most of the classifiers failed to classify the "Draw" label; only Random Forest and Gradient Boosting Tree can predict the "Draw" label, 74 hits and 29 hits, respectively. Furthermore, as we mentioned, there is not much difference of classifiers in other criteria, so our recommendation for classifying "Win / Draw / Lose" is **"Gradient Boosting Tree"** and **"Random Forest."**

Experiment 2 "Goal Difference"

1. **Odd-based Decision Tree:**

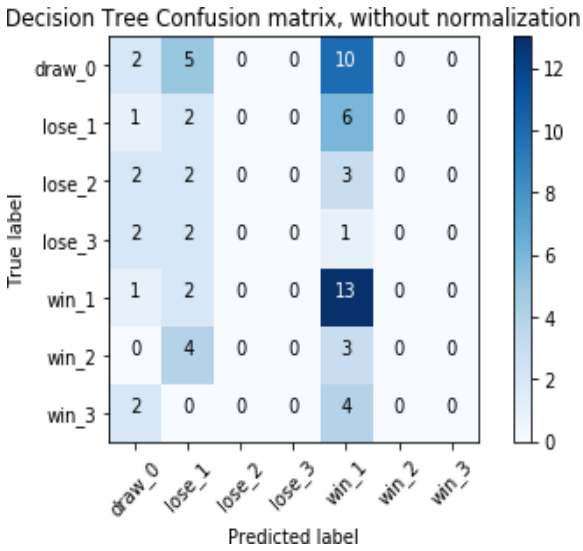


Fig. Confusion Matrix Odd

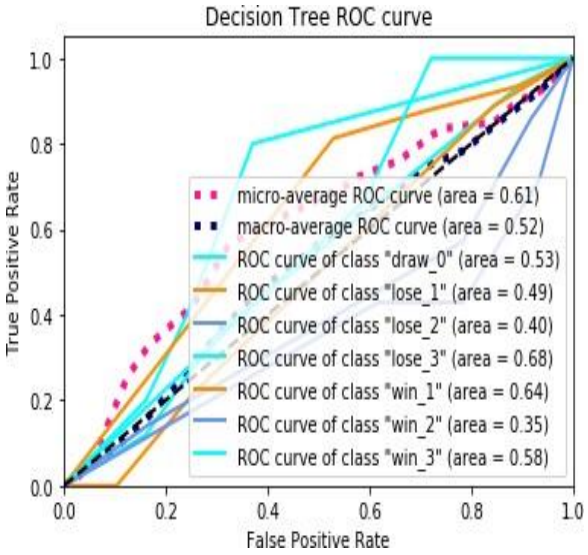


Fig. ROC curve Odd

For Goal Difference, odd-based decision tree model confusion matrix and ROC curve are shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **25.37**, and from the ROC curve area was found to be **0.62** micro-average.

2. **h2h-Form-based Decision Tree**:

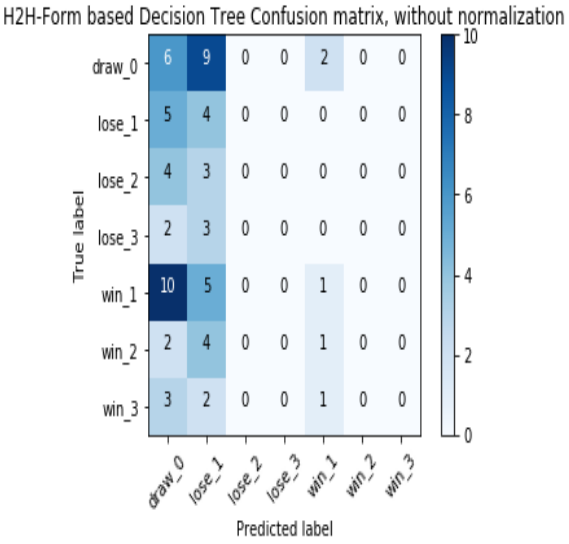


Fig. Confusion Matrix Head-to-Head

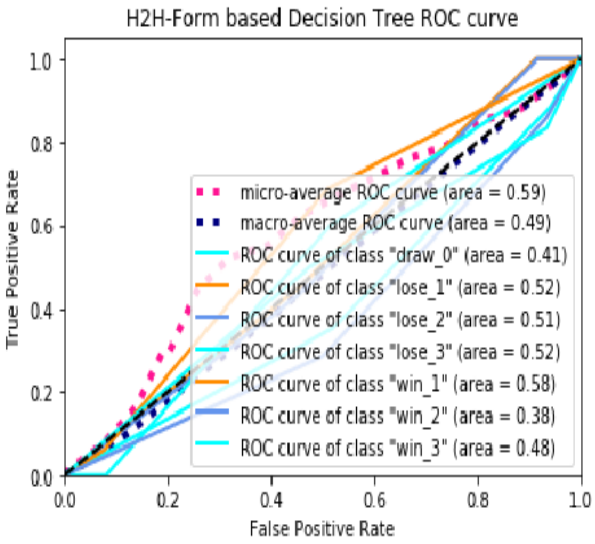
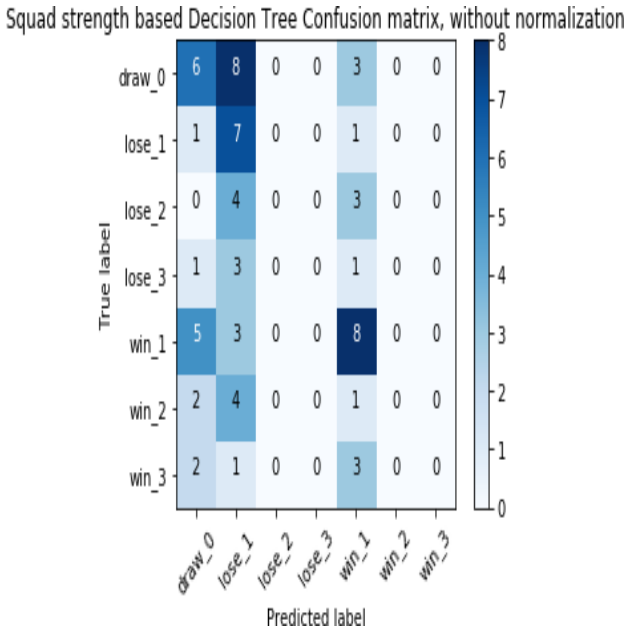


Fig. ROC curve Head-to-Head

For Goal Difference H2H form-based decision tree model confusion matrix and ROC curve is shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **18.94**, and from ROC curve area was found to be **0.59** micro-average.

3. squad-strength-based Decision Tree:



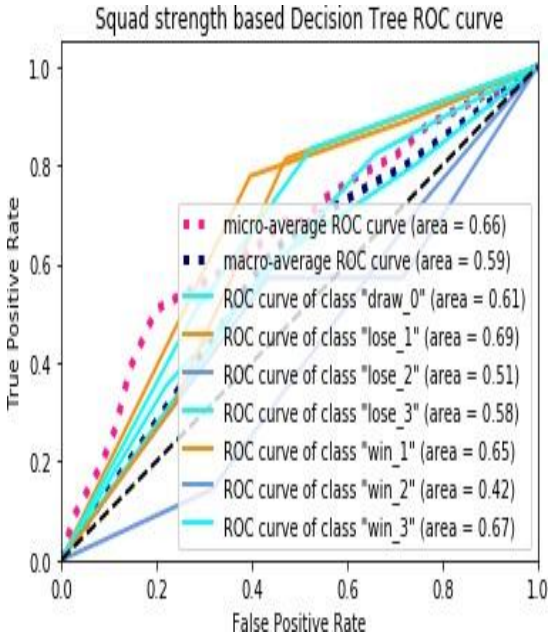
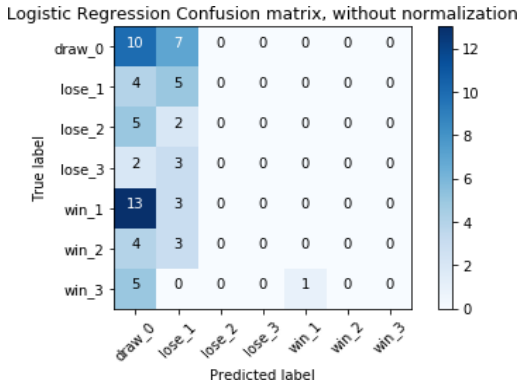


Fig. Confusion Matrix squad-strength **Fig. ROC curve squad-strength.**

For Goal Difference H2H form-based decision tree model confusion matrix and ROC curve is shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **31.34**, and from the ROC curve area was found to be **0.66** micro-average.

4. Logistic Regression:



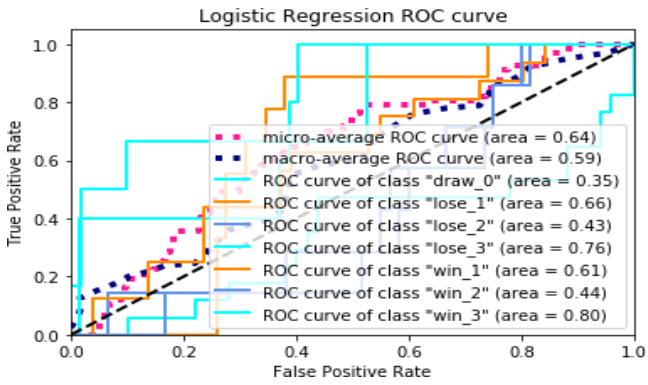
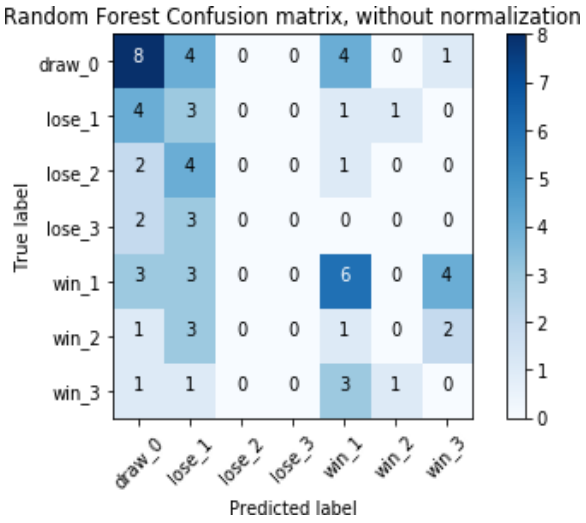


Fig. Confusion Matrix Logistic Regression **Fig. ROC curves Logistic Regression.**

For the Goal Difference Logistic Regression model, the confusion matrix and ROC curve are shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **22.38**, and from ROC curve area was found to be **0.64** micro-average.

5. Random Forest:



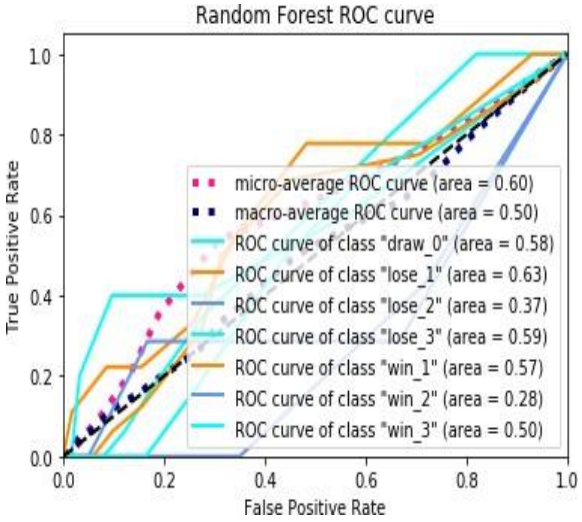
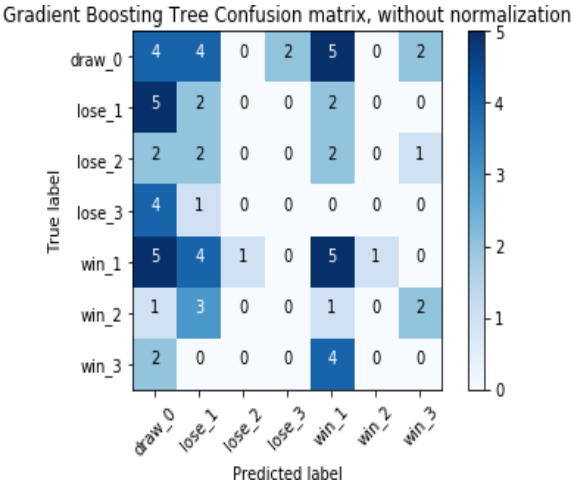


Fig. Confusion Matrix Random Forest **Fig. ROC curve Random Forest**

For the Goal Difference, the Random Forest model confusion matrix and ROC curve are shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **25.37**, and from the ROC curve area was found to be **0.60** micro-average.

6. Gradient Boosting tree:



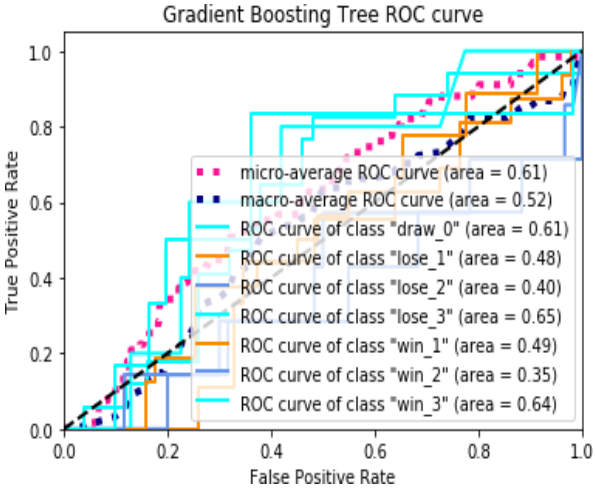
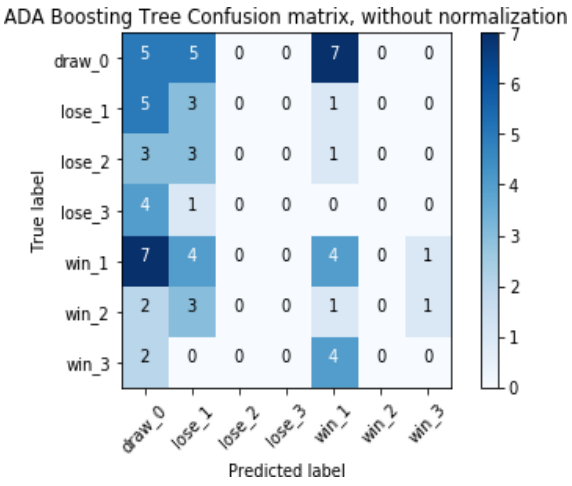


Fig. Confusion Matrix Gradient Boosting Tree Fig. ROC curve Gradient Boosting Tree

For the Goal Difference Gradient Boosting Tree model, the confusion matrix and ROC curve are shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **16.42**, and from the ROC curve area was found to be **0.58** micro-average.

7. ADA boost tree:



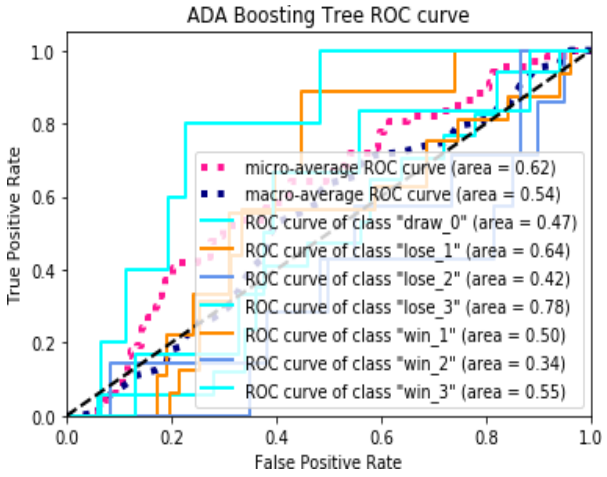
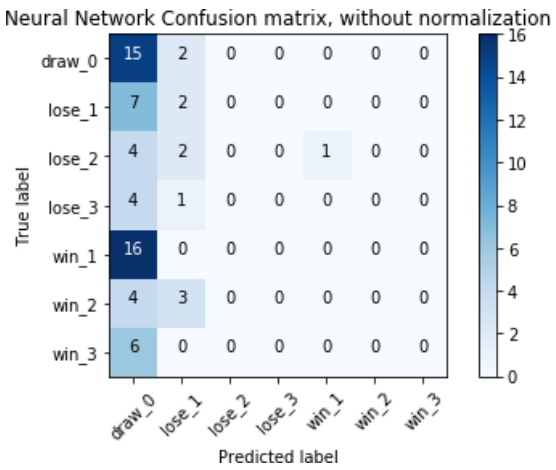


Fig. Confusion Matrix ADA boosts Fig. ROC curve ADA boost.

For Goal Difference, ADA Boost Tree model confusion matrix and ROC curve are shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **16.41**, and from the ROC curve area was found to be **0.59** micro-average.

8. Neural Network:



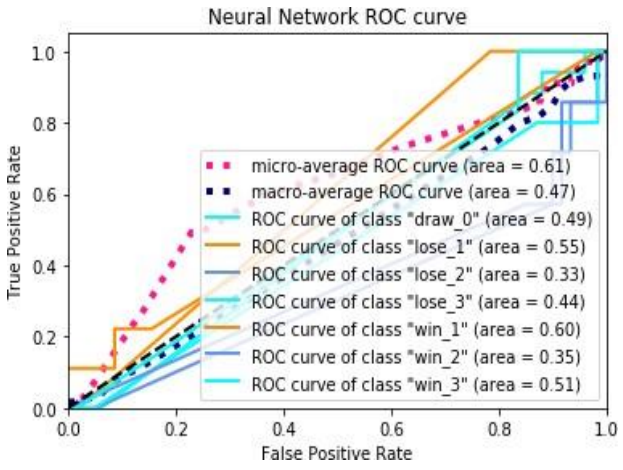


Fig. Confusion Matrix NN

Fig. ROC curve NN

For the Goal Difference Neural Network model confusion matrix and ROC curve is shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **25.37**, and from the ROC curve area was found to be **0.63** micro-average.

9. Light GBM:

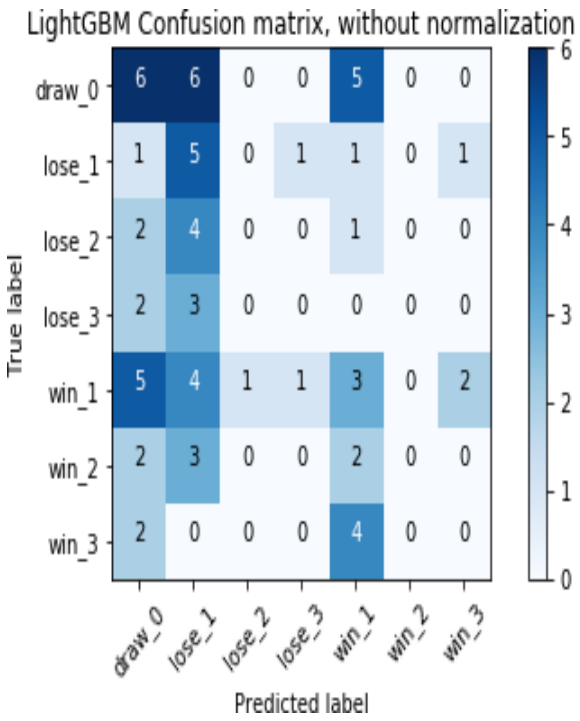


Fig. Confusion Matrix Light GBM

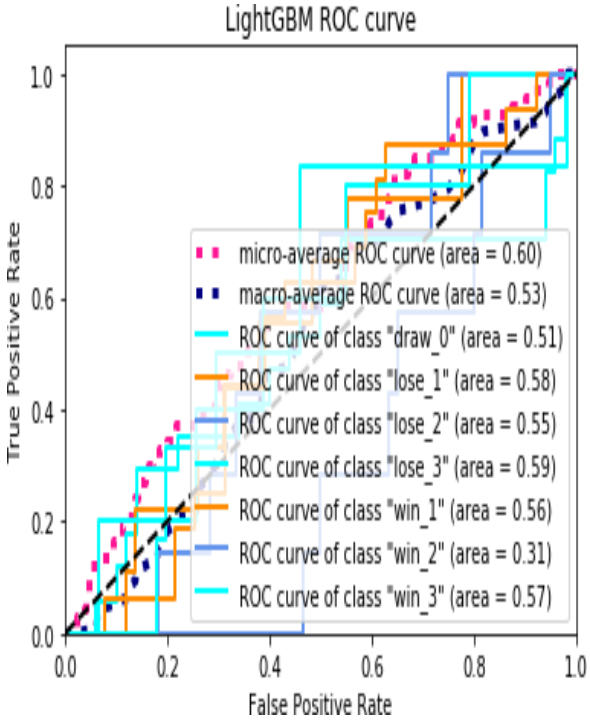


Fig. ROC curve Light GBM

For Goal Difference Light GBM model confusion matrix and ROC curve is shown above. From the above confusion matrix, we can easily calculate our F1 score, and in this case, it came out to be **20.89**, and from ROC curve area was found to be **0.57** micro-average.

Model	10-fold CV accuracy (%)	F1 - micro average	AUROC - micro average
Odd-based Decision Tree	26.41	25.37	0.62
H2H-Form-based Decision Tree	16.74	18.94	0.59
Squad-strength-based Decision Tree	31.64	31.34	0.66
Logistic Regression	21.39	22.38	0.64
Random Forest	25.36	25.37	0.60
Gradient Boosting tree	27.27	16.42	0.58
ADA boost tree	26.92	16.41	0.59
Neural Net	22.42	25.37	0.63
LightGBM	25.62	20.89	0.57

Table.6 Model Accuracy Table

In experiment 2, the "Squad Strength" based Decision Tree tends to be superior to other classifiers. We can see that 10-fold cross-validation accuracy in the squad-strength-based decision tree is enormous compared to others. On the other hand, the performance of the H2H form-based decision tree tends out to be worst. So, by comparing all the models, we can say that **Squad-strength based decision tree** is more suitable in the case of goal difference calculation.

Experiment 3 "Goal Difference" and "Win/Draw/Lose" in World Cup 2018

Model	"Goal Difference" Accuracy	"Win/Draw/Lose" Accuracy (%)	F1 - micro average
Odd-based Decision Tree	31.25	48.43	31.25
H2H-Form based Decision Tree	25.00	34.37	25.00
Squad strength-based Decision Tree	28.12	43.75	28.12
Logistic Regression	32.81	57.81	32.81
Random Forest	32.81	56.25	32.81
Gradient Boosting tree	21.87	45.31	21.87
ADA boost tree	28.12	51.56	28.12
Neural Net	20.31	35.94	20.31
LightGBM	32.81	56.25	32.81

Table.7 Model Accuracy Table

By looking at the model accuracy table, it is visible that **Logistic Regression** must be the best choice while going for "Goal Difference" and "Win/Draw/Lose" prediction.

Apart from Logistic Regression **Random Forest** also performed well, there is not much difference between both models. So, it is solely our choice between both the models, but we must give preference to **Logistic Regression.**

CONCLUSION

In conclusion, odd-based features from bet bookmarkers are reliable to determine who is the winner of matches. However, it is horrible to find out whether matches end up with a draw result. Instead, the Ensemble method like Random Forest and Gradient Boosting tree is superior in this case. Squad index from FIFA video games provides more information and contributes significantly to predicting "Goal Difference." Other complex machine learning models show not much difference from simple odd-based or strength-based trees; this is reasonable because the amount of data is limited, and a simple decision tree can provide an easy solution.

SIGNIFICANCE OF THIS WORK

Sports analytics is quite common these days as an enormous number of websites provide the data, but only data is not sufficient. Raw data cannot provide us the hidden information.

So, from this work, we optimize our working area, and as per the KDD process, we extract some refined information that is knowledge. Moreover, the knowledge we extracted in this Project can be helpful for legal betting in India and overseas as we all know that football is the best and most recognized sport worldwide and have a substantial economic market. So, our study can be used for legal betting activities in the field of football analytics.

Furthermore, one of the most crucial significances of this work is that football managers (coaches) around the globe can use this study so that they can create or manage their squad accordingly, and by doing that, they remain one step ahead of the rest and eventually expect better results out of it.

FUTURE WORK

There are so many unanswered Questions apart from this study, and the deeper we dig more knowledge we get. Football analytics is like an ocean full of knowledge. So, it depends on us to what extent we seek to refine information from the whole lot.

I have listed some of my question statements on which the future work will have a more focused and better solution. Anyone can use this study and get a wholesome idea to answer these question statements listed below.

HOW DOES CROSSING PASS AFFECT MATCH RESULT?

HOW DOES CHANCE CREATION SHOOTING AFFECT MATCH RESULTS?

HOW DOES DEFENCE PRESSURE AFFECT MATCH RESULTS?

HOW DOES DEFENCE AGGRESSION AFFECT MATCH RESULTS?

HOW DOES DEFENCE TEAM WIDTH AFFECT MATCH RESULT?

If you are a football enthusiast, you must opt for these question statements and try to createmore question statements like this and answer them all. Your contribution to the field of football analytics will always be remembered.

REFERENCES

- Journal article

Rory Bunker, Tabatha (2019). Titled – "A machine learning framework for sportsresults from prediction and analysis."

- Online document

Understanding t-test (Student's t-test - Wikipedia)

- Article by sci-kit-learn

Confusion-matrix (Understanding confusion – matrix in a simpler manner)

- Article by sci-kit-learn

Precision and Recall (Understanding the concept of Precision and recall most readily)

- Article by sci-kit-learn

Receiver Operating Characteristics (ROC) (Understanding the concept of Receiver Operating Characteristic)

- Article by sci-kit-learn

Model evaluation (Metrics and scoring for model evaluation: "understanding the concept of quantifying the quality of predictions in the experiment")

- Article by sci-kit-learn

("How to tune the hyper-parameters of an estimator")

- Article by sci-kit-learn

Understanding Validation curves ("plotting scores to evaluate models")

- Online document

American odds v/s Decimal odds ("Betting odds formats explained" (pinnacle.com))

- Online document

European championship 2016 results and its historical odds ("Soccer Europe Archive" (oddsportal.com))

WORLD CUP 2018 RESULTS

Now the model is applying for World Cup 2018 in Russia with **simulation time = 100 000**.

Result Explanation:

Team A vs. Team B **(only valid until 90th minute)**

- "win_1": A wins with one goal differences
- "win_2": A wins with two goal differences
- "win_3": A wins with three or more goal differences
- "lose_1": B wins with 1 goal differences
- "lose_2": B wins with 2 goal differences
- "lose_3": A wins with three or more goal differences
- "draw_0": Draw

MATCH WEEK – 1

Round	Team A	Team B	win 1	win 2	win 3	lose 1	lose 2	lose 3	draw 0
1	Russia	Saudi Arabia	63.575	0.703553	4.18531	3.69347	3.66436	0.787403	23.3641
1	Egypt	Uruguay	0.198	0.349999	0.0109999	55.9517	7.7846	23.7667	11.9075
1	Morocco	Iran	11.884	5.77031	3.80533	36.6331	1.16133	2.5285	38.1824
1	Portugal	Spain	11.807	2.48671	1.85873	10.2633	0.786792	3.96592	68.8066
1	France	Australia	44.519	0.62872	6.04127	4.03893	3.11828	0.230865	41.3918
1	Argentina	Iceland	67.258	1.35509	10.9515	1.95145	2.62786	0.335718	15.4949
1	Peru	Denmark	25.507	5.44761	1.25761	32.4685	1.18723	0.217857	33.8786
1	Croatia	Nigeria	70.421	4.56179	2.67699	1.52282	1.97244	4.19859	14.6225
1	Costa Rica	Serbia	0.702	0.248998	0.120999	59.2594	1.81291	1.07633	36.7538
1	Germany	Mexico	72.845	3.75027	2.93875	2.32415	1.06413	0.0479602	17.0079
1	Brazil	Switzerland	15.949	3.25848	17.3837	1.51145	22.7523	1.07035	38.0374
1	Sweden	South Korea	27.126	5.96538	4.43353	14.1697	7.21727	0.592651	40.4589
1	Belgium	Panama	35.138	1.87034	20.2095	5.5868	0.943407	0.185882	36.031
1	Tunisia	England	0.138	0.106	0.0589999	67.7748	2.76412	4.6537	24.4805
1	Colombia	Japan	67.829	0.885399	2.4723	2.89294	2.63105	0.787396	22.4776
1	Poland	Senegal	31.383	1.2776	0.656785	14.0553	2.72871	7.63917	42.2246

MATCH WEEK – 2

Round	Team A	Team B	win_1	win_2	win_3	lose_1	lose_2	lose_3	draw_0
2	Russia	Egypt	23.182	1.7256	3.47613	43.9865	1.05124	2.40623	24.1377
2	Portugal	Morocco	53.267	10.1746	11.6296	13.8406	2.22103	0.154859	8.67908
2	Uruguay	Saudi Arabia	58.848	4.14456	6.39497	2.70312	9.73798	0.0649469	18.0762
2	Iran	Spain	1.694	0.0739987	0.0239996	21.5036	4.89686	67.7349	4.04812
2	Denmark	Australia	66.577	5.45137	1.01727	14.9661	2.00224	0.142871	9.81715
2	France	Peru	84.295	0.353702	2.4859	1.47172	4.68885	0.343679	6.34706
2	Argentina	Croatia	49.233	4.27989	4.58555	13.4312	8.14018	2.80576	17.4896
2	Brazil	Costa Rica	21.231	47.8039	11.3831	2.00439	2.71876	0.0849277	14.7394
2	Nigeria	Iceland	6.104	1.27892	6.23154	29.249	1.98115	1.38438	53.7402
2	Serbia	Switzerland	8.955	1.74484	0.610935	19.8997	4.02075	1.58644	63.1547
2	Belgium	Tunisia	29.115	2.98713	4.77247	11.6057	15.2476	0.381757	35.853
2	South Korea	Mexico	1.086	0.176998	0.182998	73.3609	3.73421	3.84698	17.5905
2	Germany	Sweden	58.987	3.38101	8.93443	1.63583	14.6323	0.0279755	12.3712
2	England	Panama	62.028	0.301813	5.88933	1.21817	0.421707	0.191866	29.923
2	Japan	Senegal	1.711	0.899985	0.227994	72.4649	0.486634	3.09266	21.0954
2	Poland	Colombia	54.125	2.32574	0.656629	13.7711	3.83428	2.17038	23.0853

MATCH WEEK – 3

Team A	Team B	win_1	win_2	win_3	lose_1	lose_2	lose_3	draw_0
Saudi Arabia	Egypt	2.076	8.975	4.005	43.398	1.338	2.528	37.68
Uruguay	Russia	41.698	2.667	6.917	2.859	1.984	0.323	43.552
Iran	Portugal	0.428	0.178	0.019	86.287	1.007	1.983	10.098
Spain	Morocco	79.205	1.164	4.747	0.295	0.21	0.891	13.488
Australia	Peru	13.139	3.929	0.367	29.428	3.436	1.591	48.11
Denmark	France	1.519	1.278	0.064	60.62	1.047	2.85	32.622
Iceland	Croatia	5.963	1.485	0.478	22.636	1.375	1.598	66.465
Nigeria	Argentina	0.248	3.059	4.27	49.139	7.498	4.259	31.527
Mexico	Sweden	59.267	1.943	0.312	1.262	0.094	1.535	35.587
South Korea	Germany	0.004	0.065	0.004	82.62	3.554	10.279	3.474
Serbia	Brazil	0.58	0.243	0.904	26.705	30.196	4.788	36.584
Switzerland	Costa Rica	36.93	17.482	18.43	6.769	1.502	0.058	18.829
Japan	Poland	5.142	2.505	2.045	50.9	2.153	6.212	31.043
Senegal	Colombia	35.225	1.151	13.16	9.381	11.23	2.516	27.337
England	Belgium	14.926	2.104	6.952	19.563	13.204	1.338	41.913
Panama	Tunisia	13.052	1.963	0.46	24.808	9.657	9.055	41.005

ROUND OF 16

Team A	Team B	win_1	win_2	win_3	lose_1	lose_2	lose_3	draw_0
France	Argentina	2.436	0.627	16.497	39.563	0.792	0.933	39.152
Uruguay	Portugal	5.552	1.451	2.394	59.14	2.157	1.551	27.755
Spain	Russia	37.712	2.847	7.222	3.439	2.217	0.482	46.081
Croatia	Denmark	16.486	16.731	1	23.01	0.491	1.408	40.874
Brazil	Mexico	50.348	12.074	3.426	5.794	5.49	0.003	22.865
Belgium	Japan	72.257	0.384	19.971	1.76	0.75	0.002	4.876
Sweden	Switzerland	21.975	3.374	0.159	34.893	1.631	3.043	34.925
Colombia	England	4.387	0.526	0.013	67.295	0.188	7.62	19.971

QUARTER-FINALS

Team A	Team B	win_1	win_2	win_3	lose_1	lose_2	lose_3	draw_0
Uruguay	France	4.542	0.136	0.148	75.609	0.258	3.051	16.256
Brazil	Belgium	67.634	7.757	10.894	4.5	1.314	0.287	7.614
Sweden	England	0.381	0.06	0.231	52.197	0.753	2.708	43.67
Russia	Croatia	18.606	2.613	0.502	30.869	9.509	4.731	33.17

SEMI-FINALS

Team A	Team B	win_1	win_2	win_3	lose_1	lose_2	lose_3	draw_0
France	Belgium	23.117	2.246	5.304	20.852	4.498	0.117	43.866
Croatia	England	0.309	0.22	0.185	49.05	0.57	0.327	49.339

THIRD PLACE AND FINAL

Team A	Team B	win_1	win_2	win_3	lose_1	lose_2	lose_3	draw_0
Belgium	England	81.536	0.577	0.013	3.951	0.412	0.283	13.228
France	Croatia	48.662	0.104	2.572	0.304	0.018	0.021	48.319