

Business Requirements

## Rationale Statement



Anmol Panna (100722039)

Sharang Ahluwalia (100700203)

Abdul Rahman Khan (100719011)

Levi Nair (100733132)

Course Name: Capstone Term II

Course Code: AIDI 2005

Facilitator: Marcos

Date: 24 Jan 2020

**Business Requirements**

**Executive summary:**

Predicting the result of a sport game has been at focus for years but humans are still superior at predicting sport outcomes. Although, there are companies which use data to predict the outcomes but at the root, they still have human experts to manually track the games. The proposed model is based purely on historical data of over 20,000 football matches played between 2008 and 2018. The model would be considered a success if it predicts the outcome with a precision of 70%. The algorithm integrates Logistic Regression Random Forest, GaussianNB, KNN and SVM and CNN and selects the best parameters to avoid over fitting and local minima problems thereby increasing the accuracy. This model aims to get better prediction results as compared to the one’s being used by the bookmakers.

**Introduction:**

Predicting the outcome of football games based on past information is a non-standard predictive task because of the nature of the game outcome, Also because of high uncertainty involved. The outcome of the game could be observed as Goal-difference or further aggregated to reflect whether the game ended as a win home win, away win or draw. And thus, standard Linear Modelling perform poorly on these predictions. So, a problem of this nature is highly intriguing because of Level of luck or uncertainty involved. Non- linear models geared towards estimation of probability of ordered outcomes such as points or wins would be used for this purposed model. A similar approach was implemented previously by Lechner and Osaka as well. “Lechner and Okasa (2018) propose to adapt classical random forest estimation, which is known to have excellent predictive performance to the problem of predicting probabilities of ordered categorical outcomes,

such as the win-draw-loss problem of a football game”. (Knaus, 2018) But this model will adapt more than just Random Forest estimation. Tackling this problem would not just open-up avenues to understand non-standard predictive tasks but also improve machine predictions for highly uncertain tasks. Also, getting better odds than Bookmakers is always beneficial.

**Rationale Statement:**

Our models will evidently predict the outcome of the football game based on historical data obtained from Kaggle.com. A combination of KNN, SVM, Random Forest, Naïve Bayes, XG Boost and AdaBoost will try to solve the problem of overfitting, zero probability and look to decrease computational cost.

The idea behind designing such a project is to check whether it is possible to increase the accuracy of current sports prediction models without the need of human intervention. If it is possible, then figure out which techniques and algorithms can be used to challenge the existing models being used by huge betting companies, thereby increasing our market value and gain attention of the media in the long run. Our model would assist our target audience who want to maximize their investment with the least amount of risk involved.

**Problem Statements:**

This model aims at finding out whether a structure of Random Forrest, KNN, SVM, Naïve Bayes, XG Boost and AdaBoost and other algorithms can accurately predict the outcome of a football game by analyzing the following:

• Is the form of a team important while predicting the outcome of a match?

• Does the team selection make a significant impact on the team’s performance?

• Is there an advantage to the team playing at home?

**Alignment:**

We are going to make a website and show the prediction for the results of the game and we are going to do that

By using several machine learning algorithms and test them and make sure everything is working perfectly.

**KPI:**

1. Predicting the output for the games and making it possible for people to bet on the games.
2. Making the website reliable by showing accurate results.