Capstone Proposal:

"Win/Lose prediction in Soccer games based on Team Line-up"

Domain Background:

Sports Betting Industry: Develop an algorithm to determine win/lose percentage based on roster selection and formation, rather than team general expected results.

Machine learning is heavily being used for establishing betting odds and win-lose percentages. This paper actually introduces a similar approach to the one suggested, while focusing on Tottenham Hotspur Football club data in the early 2000s.

This paper also provides a solid starting point for establishing a proper framework using Neural Networks.

There is also a couple of interesting blog posts and discussion boards such as Link1, Link2 and Link3, that provide relevant and up-to-date information of the different approaches used in the current betting industry.

Problem Statement:

Win/Lose Percentage is usually based on recent team performance and historical encounters between teams. The objective of this project is to develop an algorithm that takes other factors into perspective, focusing on team selection and lineup in matches.

On brief this algorithm will first identify the players of each team that has usually the greatest impact on the result, and accordingly will predict the W/L outcome based on the team lineup.

It is therefore a regression task with an Win/Lose ratio as an output.

Data Sets and Inputs:

This project will use the European Soccer Database available on Kaggle.

This data set should have all the information required for the project.

It consists of 7 different data sets:

1. Country: 11 x 2 (rows x column)
2. League: 11 x 3 (rows x column)
3. Match: 26k x 115 (rows x column)
4. Player: 11.1k x 7 (rows x column)
5. Player_Attribute: 184k x 42 (rows x column)
6. Team: 299 x 5 (rows x column)
7. Team_Attribute: 1458 x 25 (rows x column)

The data includes:

- Data gathered from 11 different European countries in the seasons between 2008-2016
- Player's and Team's attributes sourced from EA Sports data
- Team line up and formations of each game
- Betting odds from 10 different providers
- Detailed match statistics:
    - Date
    - Goals
    - Shots off/on target
    - Cards
    - Possession

The target value will rely heavily on the Match dataset since it includes most of the information needed, along with the Team and Player dataset to support the prediction.

It will represent the values home_team_api_id, away_team_api_id (the win and lose values) and will try to predict these values. Additionally, the plan is to establish a win/lose percentage similar to the betting odds included in the 'Match' data set as well.

Solution Statement:

Line-ups and team selections play a huge role in predicting the outcome of any anticipated game. Missing players in key positions in the formation may alter any expected results. Thus, determining these key players in each position that greatly influence the final results, would help get a better and more accurate results.

This Algorithm will be able to first predict the most valuable players in each team and thus a better prediction of a game's outcome based on the lineup of both team.

Benchmark Model:

The project will use the available data of the betting odds available in the same data set to use as a benchmark to measure the efficiency of the developed algorithm.

Moreover, I believe a simple linear regression model dictating a 50/50 Win-Lose ratio could serve as a great baseline for this algorithm as a sanity check for the developed algorithms

Evaluation Metrics:

The Results will be represented in Win/Draw/Lose Percentage similar to the percentages usually provided by Sports betting companies

Since this is considered a Regression prediction algorithm "R Square" will most probably be used to evaluate and measure the model's overall performance.

Project Design

The solution will be developed in several steps.

First the key players in each position will be identified based on their contribution to the game. As a starting point, the consider will only consider games won/drawn/lost as the main factor for getting this job done.

This information will then used to first establish the Win/Lose value already included in the dataset enhance the percentages provided by the betting agencies to provide better insights and proper predictions.

Since the data is already labeled, the algorithm will use the usual regression rather classification, the usual algorithms we learned during the course.

In a later phase, the algorithm could get more into details by rating the players based on their direct contribution, such as goal scored, assists, goal conceded etc.