Summary

The approach taken for this prediction task consists of the following steps:

1. Data Collection and feature engineering
2. Predicting outcomes and score of individual matches
3. Simulation and Results

Firstly, data is collected from the popular sports website espn as well as transfermarkt. The data is then converted into features which are then used for modeling. Two models are trained, one to predict the outcome and the other to predict the score given the features of two teams. Lastly a simulation is run based on the tournament structure of the world cup as well as the outcome probabilities predicted by the model. Lastly, once the finals and third place playoff outcomes have been predicted, the score for each team is then predicted by the second model.

The code for this project can be found at:
https://github.com/nykznykz/fifa-world-cup-2018-prediction/

Data Collection and Feature Engineering

Despite several datasets being provided by the hosts of the competition, we opted to collect our own data as we had a set of features in mind. The features selected would have to be able to be backdated as well as available for the 2018 world cup. In the end, these are the features for each team that we chose to collect:

1) Average number of wins in the previous 5 international games
2) Average number of draws in the previous 5 international games
3) Average goal difference in the previous 5 international games
4) Average age of the squad at the time of the match
5) Average goals per appearance of the squad in the last year
6) Average assists per appearance of the squad in the last year
7) Average yellow cards per appearance of the squad in the last year
8) Total minutes played in the last year

A web scraper was written using the Python framework Selenium. The espn site was scraped for each world cup game for the last 6 world cups. We did not go further back as the there were an unacceptable number of missing features past that. For each game, the past five games and their results were scraped and processed into features 1-3 for both teams. The lineups were also stored.
Next, the website transfermarkt was scraped. For each player in each team's lineup, the number of appearances, goals, assists, yellow cards and minutes were obtained for the year leading up to the world cup as well as his age (e.g. A player's 2013 data was obtained for 2014's world cup). Again, the data is processed into features 4-8.

Thus, in total we have 16 features (8 per team) to predict the outcome of each match.

Predicting Outcomes and Score of Individual Matches

There are two types of machine learning tasks here:

1) Predicting the probabilities of team1_win, team2_win, draw
2) Predicting Score
    2.1) Predicting team1_score
    2.2) Predicting team2_score

For task 1, the Gradient Boosted Tree classifier was used. The dataset was split into train, dev and test set. GridsearchCV was used to tune hyperparameters. The result was an accuracy of 51% for the 3 class-accuracy problem which is not very impressive but still better than the baseline accuracy of 33% from random guessing. This is to be expected given the difficulty of predicting the outcomes of games since even a favoured team can end up with a draw instead of a win.

For task 2, the Gradient Boosted Tree Regressor was used. Two models were created to predict the respective scores of team1 and team2.

As an additional note, the models from task 2 will be used mainly to predict the scores of the finals and third place playoffs only. To determine which teams win, the model from task 1 will be used. This is because the model from the classification task is able to output probabilities which makes it easier to run simulations compared to the deterministic nature of the regression output. Furthermore, it is easier to predict the outcome of the game rather than exact scores, which makes it more reliable as a sorting mechanism.

Simulation and Results

A simulator was built to observe the tournament outcomes over a large number of experiments to determined the top 4 spots. The simulator, similar to the actual tournament, comprises of:

1) Round Robin
2) Ro16
3) Ro8
4) Ro4
5) Finals and Third place playoffs

For each game in the simulation, the features are retrieved put together for the model trained earlier to give a prediction of probabilities. These probabilities are used to randomly decided the winner. (i.e. A team with 0.6 predicted probability to win will be predicted as the winner 6 times

out of 10) The teams play according to the tournament format, with teams being eliminated or proceeding to the next round. Eventually, the top four teams are observed.

The simulator was run for 100,000 simulations and the results were tallied to determine the best choices for the finals and third place playoffs as well as who wins:

Finals: Brazil vs Belgium (Brazil wins)
Third Place: Peru vs Portugal (Peru wins)

Next, the scores of these matches were predicted from the scoring model and the results are:

Finals: Brazil vs Belgium (Brazil wins 1-0)
Third Place: Peru vs Portugal (Peru wins 2-1)

Epilogue

At the time of writing Peru was eliminated from the group stages. A brief look at the results show that both the games Peru lost were close (0-1) with Peru leading in possession. Examining the features for Peru showed that they had great performance for their previous 5 games leading up to the world cup (4 wins and a draw), perhaps giving the model confidence in their ability to win against most opponents in their group.