UDACITY

MACHINE LEARNING ENGINEER NANODEGREE

CAPSTONE PROJECT PROPOSAL

HANNES ENGELBRECHT

1 DECEMBER 2019

Abstract

This proposal sets out to identify a real world problem that can be solved by applying machine learning techniques covered in the Udacity Machine Learning Nanodegree. This proposal starts off with some background information on the problem that has been identified, after which a more detailed explanation of the problem is provided. Before arriving at a solution to this problem, I describe what data and tools are available to use during the creation of the solution. The proposed solution to the problem will be described in more detail along with any existing/benchmark solutions. The proposed solution's performance will be measured against the benchmark solution. Finally, the workflow of how the solution is developed and evaluated will be outlined at the end of this proposal.

Domain background

Machine learning in sports is an ever-growing industry with more and more raw data becoming available daily. According to Bunker and Thabtah [1], machine learning has shown promise when it comes to the domains of classification and prediction in sports, where club managers and owners desire accurate machine learning models to help formulate strategies to help them win matches, thereby maximizing profits and enhancing their club's reputation.

Although the above observations apply largely to North America and Europe, the same cannot be said of African countries (South Africa in particular). While tech companies in South Africa (SA) are aware of many of the techniques used by their US, Canadian and European counterparts, the application of such techniques in the SA sports industry remains surprisingly low. This begs the question: "If we have the tools to create value and insight from an increasing amount of data in South African sports, why are we not using it to do so?"

Upon investigating this topic, I came across multiple Kaggle data science challenges that relate to analytics in sport. Seeing that such a challenge is at an intersection of two of my passions, machine learning and sport, I decided to use one of these Kaggle challenges for my capstone project (*NFL Big Data Bowl [2]*). Although the chosen challenge applies to an American sport (the National Football League), the proposed solution may be applicable to SA sport as well. As mentioned above, the application of machine learning in SA sport is in its infancy and has much room for growth. If we, here in SA, can gain meaningful experience and investigate further possibilities in the application of machine learning in sports, a lot of value can be added. Hence, the two questions that this project sets out to answer are:

1. "Can I propose an accurate machine learning solution to the Kaggle NFL Big Data Bowl competition?" This is the primary focus of the project.
2. "What are the machine learning solutions that are already in use in other countries that we can utilize in SA sport?" Please note that this will be a peripheral focus of the project.

Problem statement

The Kaggle NFL Big Data Bowl competition, as set out on their website [3], reads as follows:

"American football is a complex sport. From the 22 players on the field to specific characteristics that ebb and flow throughout the game, it can be challenging to quantify the value of specific plays and actions within a play. Fundamentally, the goal of football is for the offense to run (rush) or throw (pass) the ball to gain yards, moving towards, then across, the opposing team's side of the field in order to score. And the goal of the defense is to prevent the offensive team from scoring.

…

In this competition, you will develop a model to predict how many yards a team will gain on given rushing plays as they happen."

The original goal of the challenge was to have competitors make predictions on live games *as they happen*. Seeing that the cut-off for this Kaggle data science challenge has expired and no new live game data will be available, predictions will be made on historical data as opposed to live data. This requires the provided dataset to be split into a training and testing set, the latter of which will be used to evaluate the model.

Datasets and inputs

A dataset has been provided by the NFL and can be found at https://www.kaggle.com/c/nfl-big-data-bowl-2020/data. This dataset contains 49 columns and roughly 500 000 rows. Each row in the dataset represents a single player's involvement in a single play for a specific game. This translates into roughly 500 'games' and 23 000 'plays'.

This dataset will be split into three separate datasets: one for training the model, one for hyperparameter tuning and one for evaluating model performance.

Solution statement

Neural Networks (NN's) and XGBoost models have recently received significant praise for performing well on myriad machine learning problems and are currently regarded as 'state of the art' models in the domains of data science and machine learning. These two models were initially the two strongest contenders for this project. However, we have 49 columns available and therefore have many features available from which a model can be derived. One of the challenges now becomes ensuring that the model does not overfit the data, which translates into ensuring that only the most significant features are selected for training purposes. NN's are known to overfit to the data when many features are available. We will therefore be making use of an XGBoost model to predict the number of yards that will be made on given rushing plays. For this purpose, I will be using Amazon Web Services' high level API to build the XGBoost model.

Benchmark model

Our XGBoost model will be compared against a simple linear regression model that will be trained and tested on the exact same training and testing data.

Evaluation metrics

For this project, the Root Mean Square Error (RMSE) will be used to evaluate the performance of both the XGBoost model and the (benchmark) simple linear regression model. The RMSE is defined as the standard deviation of the residuals [5], where a residual is a measure of the distance between the predicted and the actual yards gained/lost on given rushing plays. This definition can be seen below:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

where:

- n = total number of plays on which predictions are made
- y_j = actual yards gained/lost on the j'th play
- $\hat{y}_j$ = predicted yards gained/lost on the j'th play

Project design

The project can be broken down into the steps set out below:

1. Import the data
2. Exploratory data analysis – this step involves the following:
    a. Create a general summary of the data
    b. Assess the number of columns and rows
    c. Identify the scope of missing values
    d. Visualizing the features likely to be significant
    e. Identifying limitations of the data and how it affects model outcomes
    f. Identify which features would need to be transformed before they can be used in the model
3. Data cleaning
    a. Any missing values or faulty/dirty data will be cleaned in this step
4. Feature engineering – here we actually transform the features to be fed to the model. The following topics will be discussed in this section:
    a. Missing values and how to impute them
    b. Label encoding
    c. Feature scaling
    d. Correlation between features
    e. Feature selection, i.e. which features will actually be used in the model
    f. Split the data into training, validation and testing sets
5. Modeling
    a. Build an XGBoost Regressor
6. Train the model on the training dataset
    a. Hyperparameter tuning
    b. Select a tuned/optimized model
7. Test the selected model on the test dataset
8. Conclusion and recommendation
9. Write a blog post documenting the results and recommendations

References

[1] Rory P. Bunker & Fadi Thabtah (2017). A machine learning framework for sport result prediction.

Applied Computing and Informatics, 15(1), 27-33

[2] https://www.kaggle.com/c/nfl-big-data-bowl-2020 (2019)

[3] https://www.kaggle.com/c/nfl-big-data-bowl-2020/overview (2019)

[4] https://www.kaggle.com/c/nfl-big-data-bowl-2020/discussion (2019)

[5] https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d (2016)